A Study on Information Visualization

Ali Mohammad

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Applied Mathematics and Computer Science

Eastern Mediterranean University
September 2022
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Prof. Dr. Nazım Mahmudov
Chair, Department of Mathematics

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Asst. Prof. Dr. Mehmet Ali Tut
Supervisor

Examining Committee

1. Prof. Dr. Rashad Aliyev
2. Asst. Prof. Dr. Hüseyin Lort
3. Asst. Prof. Dr. Mehmet Ali Tut

ABSTRACT

Information visualization is a process of visualizing numeric and categorical data to assist humans with superior apprehension of data which can be accomplished by data visualization. This thesis investigates the relation and dissimilarities between information and data visualization and the tools and programming languages' libraries to provide depictions of data, importance of visualizing information and its applications in real world problems. The distinction between the terms, charts and graphs will also be elucidated. Furthermore, in the thesis we have illustrated a various array of datasets using different types of charts and graphs such as violin, funnel and radar charts. In the end, this thesis demonstrates that in the realm of visualization each chart and plot has one or more diverse characteristics and is utilized based on the sort of the dataset and number of variables.

Keywords: information visualization, data visualization, charts, graphs, plots, data, dataset, Python programming language, Jupyter notebook, plotly library, numeric variables, categorical variables, distribution, comparison, composition, relationship

ÖZ

Bilgi görselleştirme, veri görselleştirme ile gerçekleştirilebilecek üstün veri kavrayışında insanlara yardımcı olmak için sayısal ve kategorik verileri görselleştirme sürecidir. Bu tez, veri tasvirlerini sağlamak için bilgi ve veri görselleştirme ile araçlar ve programlama dillerinin kütüphaneleri arasındaki ilişkiyi ve farklılıkları, bilgiyi görselleştirmenin önemini ve gerçek dünya problemlerindeki uygulamalarını araştırmaktadır. Terimler, çizelgeler ve grafikler arasındaki ayrım da açıklanacaktır. Ayrıca, tezde keman, huni ve radar çizelgeleri gibi farklı türde çizelgeler ve grafikler kullanarak çeşitli veri kümeleri dizisini gösterdik. Sonuç olarak, bu tez, görselleştirme alanında her bir grafiğin bir veya daha fazla farklı özelliğe sahip olduğunu ve veri kümesinin türüne ve değişken sayısına göre kullanıldığını göstermektedir.

Anahtar Kelimeler: bilgi görselleştirme, veri görselleştirme, çizelgeler, grafikler, çizimler, veriler, veri kümesi, Python programlama dili, Jupyter not defteri, plotly kitaplığı, sayısal değişkenler, kategorik değişkenler, dağıtım, karşılaştırma, kompozisyon, ilişki

TO MY MOTHER

ACKNOWLEDGEMENT

Words cannot express my appreciation to my supervisor, Assit. Prof. Dr. Mehmet Ali Tut who was not only a supervisor but also a solicitous mentor and a trustworthy friend. Without his guidance I could not have accomplished my thesis, I would like to express my deepest gratitude to him.

Any hardship would not have been overcame without her. I would love to thank my lovely mother for being a motivation and a real supporter in my life. I will always be indebted to her.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

Information visualization points out to the utilization of computer-supported, interactive visual depiction of numerical and non-numerical abstract sets of data to augment human cognition. Information visualization is capable of generating a magnificent difference, transforming numbers, digits, ratios, diversities and other sort of dull numerical data into innovative visualization causes any kind of information look intriguing. Information visualization are often constructed with the consideration of an audience in mind and designed to exhibit influential information that they need to comprehend. In other words information visualization is a successful method to share observations in a straightforward form for both non-professionals and experts.

In this chapter some fundamental concepts of both information visualization and data visualization and questions regarding them will be described and answered.

1.1 What Is Information Visualization?

Information visualization is the process that provides a computer program an array of preparations and guidance for the abstraction of huge amounts of unorganized body of sophisticated data, with the objective of turning unprocessed information into perceptible form [29]. Information visualization and visual analytics perform a key role in modern business intelligence practices, carried out by the help of visual information software and graphic designers.

1.2 Why Is Information Visualization Important?

Nowadays many people working in contemporary organizations are requesting access to data, preparing the representation of the data in a digestible format. Business users require a procedure to illustrate data and interact with it in an intuitive way [30]. With the aid of information visualization Decision-makers handle the data with less hardship and as a result express value to the whole organization. Information visualization is vital skill today as more enterprises seek to digitally convert and produce data an essential resource across the organizations. Being capable of presenting data in a relevant way for others to find out in these days with ever-growing amounts of data has become necessary for a business to remain competitive.

1.3 Types of Information Visualization

Users correlate distinct values, demonstrate the bigger picture, track trends in the data, and recognize various relationships between variables with the assistance of information visualization tools [29]. The following visualization formats are the most frequently applied for these goals:

- Line chart
- Histogram
- Bar graph
- Stacked bar graph
- Network graph
- Column chart
- Pie chart
- Box plot
- Violin plot
- Bubble chart

- Scatter plot
- Dual-axis chart
- Stream graph
- Choropleth map
- Chord diagram
- Sankey diagram
- Hex map
- Ridgeline plot
- Voronoi polygon diagram
- Tree map
- Heat map
- Circle packing

1.4 How to Organize Information Visually

The primary move in the information visualization development is the formation of information necessities of the target user group through qualitative research, which can be implemented to decide the most efficient method for information organization [32]. The procedure of organizing visual information is moderately linear and consistent:

Describe the problem: Set up the problem the information visualization will solve by asking what the user requires and how they will work with it. Sophistication should be chosen based on the skill level of the user.

Elucidate the data to be depicted: Data ought to be categorized to demonstrate the manner in which it is mapped – quantitative data, original data, or categorical data.

Explain the dimensions needed to demonstrate the data: Dimensions and attributes should be sketched out to decide the sorts of analysis which can be administered – univariate analysis, bivariate analysis or multivariate analysis.

Explain the structures of the data: Data connections are ordinarily structured as either linear, hierarchical, spatial, temporal or networked connections.

Define the interplay expected from the visualization: Determine how much interaction the user will request from the information visualization to discover which model will be the most beneficial – static models, changeable models or manageable models.

1.5 Information Visualization Examples

The following are some instances of diverse uses of information visualization across various industries [29]:

- **Scientific studies**

  Scientific studies is a multidisciplinary research field that looks for settling scientific proficiencies in philosophical, social and historical frameworks.

- **Data mining**

  Data mining is a practice utilized by organizations to transform unprocessed data into meaningful information. With the help of tools and software companies can acquire the patterns in large amounts of data and learn regarding their customers to develop more efficient approaches, boost sales and diminish expenses.

- **Digital libraries**

  A digital library is a set of documents in arranged electronic form, accessible on the internet or on compact disks and read-only memories [33].

- **Healthcare**

  The exercise or business of supplying medical aids. Various sets of services provided by a country or an organization for the treatment of the physically and the mentally ill [34].

- **Market research**

  It is the practice of detecting the feasibility of a novel service or product through research administered with the potential customers. Market research makes an organization or a company able to identify the target market and achieve feedbacks from consumers [35].

- **Manufacturing**

  Any industry that produces goods and items from raw materials by the application of labor or machinery and that is usually accomplished with a division of labor. The most important manufacturing industries are the ones that produce petroleum related products, ships, computers, cars, aircrafts, and heavy machinery [36].

- **Crime mapping**

  Crime incident patterns can be analyzed, mapped and visualized by analysts in law enforcement agencies with the help of crime mapping. Mapping crime is an essential factor of crime analysis.

- **Financial Evaluation**

  It is a method of classifying businesses, budgets and other financial transactions to decide their appropriateness and accomplishment.

1.6 Introduction to Data Visualization

The approaches we arrange and details that we make pictures of in our minds are modifying fast and becoming more complex every day. Because of the increase of social media, the availability of portable devices and digital services, data can be found for any human action working with technology. The generated information are very crucial and assist use to analyze styles and identify patterns. Therefore, data visualization can be a powerful method to demonstrate the end-users with understandable details [37].

1.7 A Succinct History of Data Visualization

Prior to the seventeenth century, data visualization was primarily utilized to represent land marks, towns, highways, and resources on maps. Better representations were required as the demand for more accurate mapping and physical measurement expanded [38].

A Flemish astronomer named Michael Florent Van Langren is thought to have created the first visual depiction of statistical data around 1644. He provided a one-dimensional line graph which depicts the difference in longitude between Toledo and Rome at the time of the estimate, as well as the names of the astronomers who made the estimation. While Van Langren might have supplied this information in the form of a table, it is the usage of the graph that effectively conveys the vast range of estimations [38, 39].

Thematic mapping first appeared in the 18th century. Near the end of the century, attempts were made at thematic mapping of geology, economic, and medical data. At this period, abstract graphs of functions, measurement error, and empirical data collecting were all introduced [39].

William Playfair, commonly regarded as the originator of many of today's most famous graphs, was born during this time period (line, bar, circle, and pie charts). During this time, several statistical chart types were invented, including histograms, time series plots, contour plots, scatterplots, and others. From 1565 to 1820, a graph by Playfair (1821) depicts the price of wheat, weekly pay, and the ruling king during a two-hundred-and-fifty-year period [38, 39].

Friendly points out to the second part of the nineteenth century as the "Golden Age of Statistical Visuals." John Snow's map of cholera plagues in the London epidemic of 1854 and Charles Minard's 1869 chart demonstrating the number of men in Napoleon's 1812 notorious Russian campaign army, with army location denoted by the X-axis and severe cold temperatures showed at points where frostbite did take a fatal toll are two well-known instances of data visualization from that era. The Rose chart was a novel representation devised by Florence Nightingale that this period gave us [38, 39].

Several elements contributed to this "Golden Age" of statistical graphing: the industrial revolution, which gave rise to modern business; official government statistical offices to support an increasingly aware and global populace; and a growing appreciation for the significance of quantitative data in social planning, medicine, military, industrialization, commerce, and transportation. Statistical Theory also presents a method for making sense of immense datasets [38].

In the early twentieth century, the rising tendency toward statistical visualizations encountered a little stumbling block. Friendly refers to this time period as the "contemporary dark ages" of data visualization. Statisticians are more obsessed with precise figures and believe that visuals are extremely imprecise. While innovation in the area may have shifted away from data visualizations, the public's awareness of data visualization grew over this time period. Charts and graphs of all varieties were swiftly becoming widespread in textbooks, commercial applications, science, and government [38].

Friendly refers to the second part of the twentieth century as the "rebirth of data visualization," which was brought about by the advent of computer processing. Computers enabled statisticians to collect and store data in ever-increasing amounts, as well as swiftly and readily display the information. During the 1960s and 1970s, scholars like as John W. Tukey in the United States and Jacques Bertin in France pioneered the science of information visualization in the fields of statistics and mapping, respectively. Edward Tufte's foundational book, The Visual Display of Quantitative Information, was published in the early 1980s and is currently utilized in university courses for data visualization and statistical analysis. Tufte additionally introduced us to the sparkline, which depicts the basic form of a trend in a little amount of area [38].

Over the last three decades, the subject of data visualization has exploded into dozens, if not hundreds, of specialized areas. Dashboards and data discovery tools, scorecard apps, analytics suites, and a variety of other software tools allow corporations, academics, and people to examine their data in new and more creative ways. Notable writers and lecturers of the contemporary age, such as Alberto Cairo, Stephen Few, and Colin Ware, continue to enhance and elevate the science and art of data visualization [38].

We are living in a fascinating yet complicated era for data visualization. Even as we reveal novel methods for gathering, aggregating, analyzing, and visualizing data, we are also identifying some new and considerable social concerns related to data privacy violations and possible abuse - both purposeful and inadvertent. As we enter the information era, it’s both exhilarating and worrisome to think about what the future holds for us as people and as a nation. Randy Bachman summed it up best: "We ain’t seen nothing yet!" [38].

1.8 What is Data Visualization?

Data visualization refers to the conversion of information into a perceptible framework, such as a chart, a graph or a map, to make data uncomplicated for the human brain to grasp. The main purpose of visualizing data is to make it straightforward to classify patterns, outliers and trends in huge data sets. The term is usually utilized instead of other terms in this field, comprising information visualization, mathematical diagrams and graphics of information. Data visualization is one of the steps of the data science procedure, which introduces that after data has been gathered, refined and formed, it has to be visualized for decisions and conclusions to be made and reached [37, 40]. Visualization is essential to state-of-the-art analytics for similar reasons. It becomes necessary to visualize the outputs to observe result at the moment that a data scientist provides advanced predictive analytics or machine learning algorithms.

1.9 Why Is Data Visualization Important?

Nowadays visual contents perform and essential role in our lives. A research administered by SHIFT Distributive Learning has illustrated that humans normally process pictures 60000 times faster than a text or a table and that our brains perform a superior job of remembering them in the future. The study has discovered that after three days, the analyzed studies preserved between 10% and 20% of written or spoken information, in comparison with 65% of visual information [37, 41].

Data visualization can be applied to all sorts of business because it demonstrates data graphically, so normal audience and experts will be able to comprehend the data in a straightforward way. This process aids the business to recognize the areas where they should enhance [30].

Other advantages of data visualization involve the following:

- The capability to comprehend information swiftly.
- Enhance intuitiveness and make quicker decisions.
- A boosted ability to conserve the users' enthusiasm with information they can comprehend.
- Remove the demand for data scientists because data is more reachable and comprehensible.

1.10 Information Visualization vs. Data Visualization

There is always a frequent question among many people about the contrast between information and data visualization. Data visualization is a presentation of unprocessed data, information visualization shows the data visually in the framework of business logic. It is analogous to ask what the difference between information and data is. Information is data which has an explanation attached to it. Information visualization normally comprises more of a story around the data instead of a totally quantitative depiction. It is necessary to notice the difference between the two terms is often confusing and many people wield them interchangeably [29].

1.11 Relation between Information Visualization and Data Visualization

Data visualization and information visualization are associated with each other in different ways. The production of data visualization is often a forerunner to the creation of information visualization. Perhaps an analyst first applies visual techniques to spot noticeable points to be communicated. Once the creator specifies the message, the data visualization may be utilized again, as is, for information visualization objectives [31]. Considering the demands of decision-makers, circumstances and preferences might increase the conviviality and understanding of visualization. While data visualization permits the audience to examine, alter, analyze and attempt what-if schemes, an information visualization is the outcome of such analysis. In practice, there is a slope from data visualization to information visualization, depending on how much the decision-makers are willing to analyze for themselves.

1.12 Data Visualization and Big Data

Big data is gaining reputation day by day and projects of data analysis have made visualization more necessary than ever. Organizations are frequently utilizing machine learning to collect huge amounts of data that can be troublesome and slow to dig through and understand. By the assistance of visualization we can accelerate this and demonstrate information to business owners in an unambiguous way to understand. Big data visualization often moves beyond the commonplace approaches utilized in ordinary visualization, such as histograms and pie charts. It utilizes more sophisticated representations, such as heat maps and streamgraph. Powerful computer systems are required for big data visualization to gather unprocessed data, prepare it and transform it into visual representation that end-users are able to use to draw insights fast [30].

1.13 How Data Visualization Is Being Applied

- **Decision-making**

  Decision-making is a method that data visualization can be applied in. Nowadays businesses are able to observe the different patterns of attributes fast so that they are capable of describing a massive amount of data in a straightforward way and can reach results from that information [40].

- **Examining the market**

  We have the ability to analyze the market by giving all the information of various markets to the data visualization. It will deliver a clear intuitiveness into where should we concentrate more consideration and where we are failing. We can take a clear figure of chances within the markets with the help of charts and graphs [40].

- **Relationships and patterns**

  If the sophisticated data is represented visually, it will start to make sense. There will be no identification regarding the correlation of two attributes without data visualization. By spotting these connections and their patterns we will be able to obtain an idea of where to concentrate on to achieve the objectives [40].

- **Trends**

  This factor is one the most substantial usage of data visualization. We will not be able to make any anticipation without past and present data. Trends over time will demonstrate presently where we are and from this trend [40].

1.14 The Data Visualization Science

The science of data visualization derives from comprehension of how humans collect and analyze information. Daniel Kahn and Amos Tversky [42] cooperated on research that specified two diverse approaches for collecting and refining information [30].

The first system concentrates on idea processing that is quick, automatic and unconscious. This approach is commonly utilized in routine life and aids achieve:

- reading the text on a sign
- solving easy math questions, like 1+1
- recognizing where a sound is coming from
- riding a bicycle
- demonstrating the dissimilarity between colors

The second system focuses on gradual, reasonable, calculating and uncommon idea processing. This approach is applied to one of the following circumstances:

- mentioning a phone number
- answering complicated math questions, such as 654 × 137
- demonstrating the contrast in meaning between multiple signs standing abreast
- comprehension of sophisticated social clues

1.15 Data Visualization Tools

There is an array of various methods that data visualization tools can be utilized. The most frequent application is as a business intelligence (BI) reporting software tool. Users can install visualization tools to create automatic dashboards that trace company accomplishment. Many business sections put data visualization into action to track their own initiatives. For instance, a commerce team might implement the software to supervise the performance of an email campaign [30].

The well-known terms in big data tools market comprise Microsoft, IBM, SAP and SAS. Some others suggest big data visualization software; famous names in this market involve Tableau, Qlik and Tibco. While Microsoft Excel sustains to be a trendy tool for data visualization, others have been developed that support more complex capabilities:

IBM Cognos Analytics

IBM Cognos Business Intelligence is an integrated business intelligence suite that runs on the web. It features capabilities for reporting, analytics, scorecarding, and event and metric monitoring. The program is made up of multiple components that are meant to fulfill a company's various information needs. IBM Cognos Framework Manager, IBM Cognos Cube Designer, and IBM Cognos Transformer are instances of IBM Cognos components [43].

Qlik Sense and QlikView

Qlik is a company that sells a business analytics platform. The software company was formed in Lund, Sweden in 1993 and is currently headquartered in King of Prussia, Pennsylvania. Qlik Replicate and Qlik Sense, both tools for business intelligence and data integration, are the company's major products [44].

Microsoft Power BI

Microsoft's Power BI is an interactive data visualization software application intended most often at business intelligence. It's a component of Microsoft's Power Platform. Power BI is a set of software services, apps, and connectors that work together to transform disparate data sources into logical, visually immersive, and interactive insights. Data can be retrieved from a database, a webpage, or structured files such spreadsheets, CSV, XML, and JSON [45].

Oracle Visual Analyzer

Oracle Data Visualization makes it simple to analyze your data and discover valuable insights. It comprises easy-to-use visualization and data management capabilities to help you evaluate data in your Oracle Business Intelligence repository as well as data outside the repository [46, 47].

SAP Lumira

SAP Lumira is a visual intelligence platform for creating and visualizing data stories. It was previously known as a Visual Intelligence tool, and it did allow you to visualize data and create stories to convey graphical details about the data [48].

SAS Visual Analytics

SAS Visual Analytics is a comprehensive analytics visualization platform that enables you to spot patterns and correlations in data that were previously undetectable.

Tibco Spotfire

TIBCO Spotfire software is one of the most comprehensive analytics solutions available, allowing anybody to explore and visualize new data discoveries through immersive dashboards and sophisticated analytics. Spotfire analytics provides scalable features such as predictive analytics, geolocation analytics, and streaming analytics. And, with Spotfire Mods, you can create customized analytic apps quickly, often, and at scale [49].

Zoho Analytics

We can analyze our data, generate spectacular data visualizations, and uncover hidden insights with Zoho Analytics, a self-service BI and data analytics platform [50].

D3.js

D3.js (short for Data-Driven Documents), a JavaScript framework for creating dynamic, interactive data visualizations in web browsers, is sometimes known as D3. Scalable Vector Graphics (SVG), HTML5, and Cascading Style Sheets (CSS) are all employed. It is the successor to the Protovis framework from before. Its progress was observed in 2011, with the release of version 2.0.0 in August of that year. D3 was transformed from a single library to a collection of smaller, modular libraries that may be used individually with the release of version 4.0.0 in June 2016 [51].

Jupyter

The purpose of Project Jupyter is to "create open-source software, open standards, and services enabling interactive computing spanning dozens of programming languages." Fernando Pérez and Brian Granger split it out from Ipython in 2014. The name Project Jupyter is a play on the three primary programming languages supported by Jupyter: Julia, Python, and R, as well as a nod to Galileo's notebooks documenting the discovery of Jupiter's moons. Jupyter Notebook, JupyterHub, and JupyterLab are interactive computing tools created and sponsored by Project Jupyter. NumFOCUS is a financial supporter of Jupyter [52].

MicroStrategy

MicroStrategy Incorporated is a business intelligence (BI), mobile software, and cloud-based services provider located in the United States. The company, which was founded in 1989 by Michael J. Saylor, Sanju Bansal, and Thomas Spahr, creates software to analyze internal and external data and develop mobile apps. It is a publicly traded firm based in Tysons Corner, Virginia, which is part of the Washington metropolitan region. SAP AG Business Objects, IBM Cognos, and Oracle

Corporation’s BI Platform are among its major business analytics rivals. Saylor is the company’s CEO and chairman of the board of directors [53].

Google Charts

Google Charts is a Web application that generates graphical charts from data provided by users. The user enters data and a formatting specification in JavaScript embedded in a Web page, and the service responds with a chart picture [54].

Chapter 2

CHARTS

2.1 Introduction

In any data science project, one of the most important tasks is to demonstrate your data in order to explore and explain your results. At the start of a project, visualizing your data may help you better comprehend it and spot trends and patterns [12].

After the accomplishment of a study or utilizing several machine learning models, data visualization is capable of assisting us explain our discoveries more effectively at the conclusion of the project [12].

An effective data visualization strategy may make or destroy the project. If we put in a lot of time and effort into analyzing and modeling our data but end up selecting the wrong chart type to display the conclusions, the audience will not comprehend the work put in or how to utilize the results [12].

Charts are an important component of working with data because they allow us to compress enormous volumes of information into an easy-to-understand format. Data visualizations may provide insights to those who are viewing the data for the first time, as well as explain discoveries to others who will not be able to access the raw data. There are several chart kinds available, each with its own set of applications. Choosing which chart type is ideal for the task at hand is sometimes the most challenging component of developing a data visualization [9, 12, 55].

2.2 The Phase before Opting For the Right Chart

Before we begin looking at chart kinds, we should ask ourselves five essential questions regarding our data. These questions will assist us to comprehend the data much better and, as a result, select the best chart type to depict it [55].

2.2.1 What Is The Story Your Data Is Aiming to Transfer?

The first thing we should know regarding our data is what story it is attempting to tell. Why and how was this information gathered? Is our data being gathered in order to identify trends? To compare several options? Is there any distribution? Or is it used to examine the link between several value sets? [12, 55]

Understanding the genesis story of our data and what it is attempting to provide can make selecting a chart style much easier for us.

Charts are classified into four main categories [1, 21, 55, 56]:

- **Comparison**

  To compare one or more datasets, comparison charts are employed. They can compare objects or demonstrate changes throughout time.

- **Relationship**

  Relationship charts are utilized to demonstrate a link or correlation between two or more variables.

- **Composition**

  Composition charts are used to show how different aspects of a whole change over time.

- **Distribution**

  Distribution charts are used to depict how variables change over time, assisting in the identification of outliers and patterns.

2.2.2 Who Will You Present Your Results to?

After we've determined the story behind our data, we'll require to establish who we'll be presenting your findings to. If we're researching stock market patterns and planning to submit our results to some businesspeople, we may utilize a different chart type than if we're giving our findings to novice investors. As a result, we must first understand our audience in order to select the ideal chart type to utilize when presenting our data to them [12, 55].

2.2.3 How Large Is Your Data?

The magnitude of our data will have a big impact on the style of graphic we employ. Some charts are not intended to be used with large datasets, whilst others are ideal for huge data [12, 55].

Pie charts, for example, function best with a limited number of datasets; but, if we're utilizing a large number of datasets, a scatter plot will make more sense.

We must choose a chart type that is appropriate for the magnitude of our data and properly depicts it without being cluttered.

2.2.4 Have You Identified The Data Type?

There are several forms of data, including descriptive, continuous, qualitative, and categorical data. Some chart kinds can be eliminated by using the type of data. For example, if we have continuous data, a bar chart may not be the ideal option; instead, use a line chart [12].

Similarly, if we have categorical data, a bar chart or a pie chart can be a smart choice. We should generally avoid using a line chart with categorical data since continuous categories are impossible to have.

2.2.5 How Do The Dissimilar Components of Your Data Associate With Each Other?

Finally, consider how the various pieces of our data relate to one another. Is our data order determined by a factor such as time, size, or type? Isn't this a rating based on some variable? Or is it a relationship between distinct variables? [12]

Is our data a time-series – that is, does it vary over time? Is it a distribution, or is it more of a distribution?

The link between the numbers in our collection may make deciding on which chart type to employ a little easier.

2.3 What Are Graphs and Charts?

Graphs and charts are depictions of data connection that are intended to make the data comprehensible and simple to recall. Graphs and charts are widely utilized to depict trends, patterns, and links between collections of data. Certain sorts of data might receive advantage from graphs, while others benefit from charts. The chart or graph we opt for will commonly be determined by the essential points we want others to comprehend from the data we've gathered [57].

2.4 Graphs vs. Charts: What Is The Difference?

The primary distinction between a chart and a graph is that a chart is intended to graphically depict data, but a graph is intended to aid in the observation and recognition of the relationship between various linked data. A chart is commonly used in an analytic presentation, whereas a graph is utilized to visualize data connections. A graphical depiction of a data set that is constructed utilizing symbols is called a chart. Charts are categorized into a variety of shapes and sizes, such as a histogram, a pie chart, or a scatter plot. An expanded, enormous data set can be transformed into a coherent and promptly noticeable graphic [57].

Graphs are commonly utilized in the subject of mathematics. This diagram depicts facts in a graphical format so that the mathematical link may be clearly identified. A graph's two axes represent categories and numbers, and it aids in visualizing the relationships between the many features described in the supplied data [57, 58].

2.5 Tools

In this thesis we have made use of different tools and technologies in order to write codes to generate the charts. The tools are:

- **Python**

  Python programing language is high-level, object-oriented with dynamic semantics that is interpreted. Its powerful built-in data structures, together with dynamic binding and dynamic typing, make it tempting especially for usage as a glue or scripting language to link current elements together. Python's concise, easy-to-learn syntax prioritizes readability, lowering software maintenance costs. Python has support for packages and modules, which promotes program modularity and reusability of code. The Python interpreter and substantial standard library are free to utilize and share either in source or in the form of binary for all leading platforms [59].

- **Plotly library**

  One of the most advantageous graphing libraries of Python is "plotly".The plotly library is an interactive, open-source plotting toolkit that supports over 40 different chart types for statistical, financial, geographic, scientific, and three-dimensional use-cases [60, 61].

- **Jupyter**

  Jupyter or the Jupyter Notebook is a free and open-source web tool that lets you generate and share documents with visualization, live code, equations, and narrative prose. Data cleansing and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and many more applications are possible. Jupyter Notebook (previously known as Ipython Notebooks) is an interactive web-based computing environment for authoring Jupyter notebook papers. Depending on the context, the term "notebook" refers to a variety of items, most notably the Jupyter online application, Jupyter Python web server, or Jupyter document format [62].

Chapter 3

VISUALIZATIONS

3.1 Introduction

In this chapter we will be discussing the charts that have been utilized and illustrated. The application, advantages and disadvantages of each chart will be explained in order to have a solid comprehension concerning how to choose and use them appropriately based on a dataset. In the appendix there is a table which demonstrates a summary of frequent and uncommon charts in the field of visualization. All the charts here have also an interactive version, in the explanation of each chart the features in the interactive mode have been mentioned so, some familiarities with the environment of the Jupyter notebook and Python programming language are expected from the readers. A sample code is given in page 43, there are many attributes in each graph object that users can easily customize the charts for instance the colors, the size of the charts, changing the width of grids, modifying axes and axes' ticks.

We have plotted 15 charts plus some of their variations generated with Python language. 3D Line chart, 3D scatter plot, bar chart, box plot, bubble chart, contour plot, funnel chart, heatmap, histogram, pie chart, radar chart, 2D line chart, 2D scatter plot, stacked bar chart and violin chart. The charts will be discussed in this chapter with details. Basic charts, scientific charts, statistical charts and financial charts are all comprised in the thesis. For instance line chart is categorized as a basic chart, histogram is a statistical chart, funnel chart is utilized when progress or finance is our concern and radar chart is considered as a scientific chart.

3.3 Line Charts

Line charts is one of the most often used charts in the world. The line chart is simple to read, has clear depiction, and may be produced with a pen and a piece of paper. Lines connect data values to represent values over a continuous time, following trends and patterns [1].

Line graphs allow users to keep track of alteration over short and extended time periods. As a result, these graphs are useful for detecting minor changes.

Line graphs can be used to analyse changes for many groups over the same time period. They're also useful for determining how various groups interact with one another. This sort of graph might be used by a company to study sales rates for diverse goods or services over time [2].

A Spaghetti plot is another sort of a line plot that displays many lines at once. With more than a few (approximately 5) groupings, this type of graphic becomes extremely difficult to read and consequently gives little insight into the data [3].

3.3.1 3D Line Chart

The chart in figure 3.1 is a three dimensional line chart. Three dimensional (3D) line charts are not as common as the two dimensional (2D) line chart. The considerable difference between the two types of line charts is the number of variables. In the above line graph there are 2 numeric variables which represent year and average rainfall on x-axis and z-axis respectively. The categorical variable represents the months in a year on y-axis. The graph demonstrates the average rain fall in millimeters in the United Kingdom from 2010 to 2019 month by month for each year [63]. In the first line chart we might not observe the number of lines because of two reasons, the decent perspective of the chart and lack of points on each line to display the changes take place on each individual line. It is obvious that the line chart in figure 3.1 is not appealing and informative to both professionals and regular audience. The big disadvantage of a line chart is when we have many lines together and the chart can create a “spaghetti plot”. In figure 3.2 by adding color points to each change on each line we are capable to make it more impressive and readable.

In data visualization colors are key factors for graphs. The color bar shown in the graph is a scale to give the audience a different experience of reading data. The representation of colors in the color bar is continuous in order to comprise both integers and decimals. There is still confusion and complexity in the line chart in figure 3.2. An adjustment in the perspective of the graph can play an important role in the readability of the graph. Figure 3.3 is the finalized version of our 3D line chart. In the Interactive version of our 3D line chart the user is able to hover over each point to see the corresponding year, month and average rain fall.

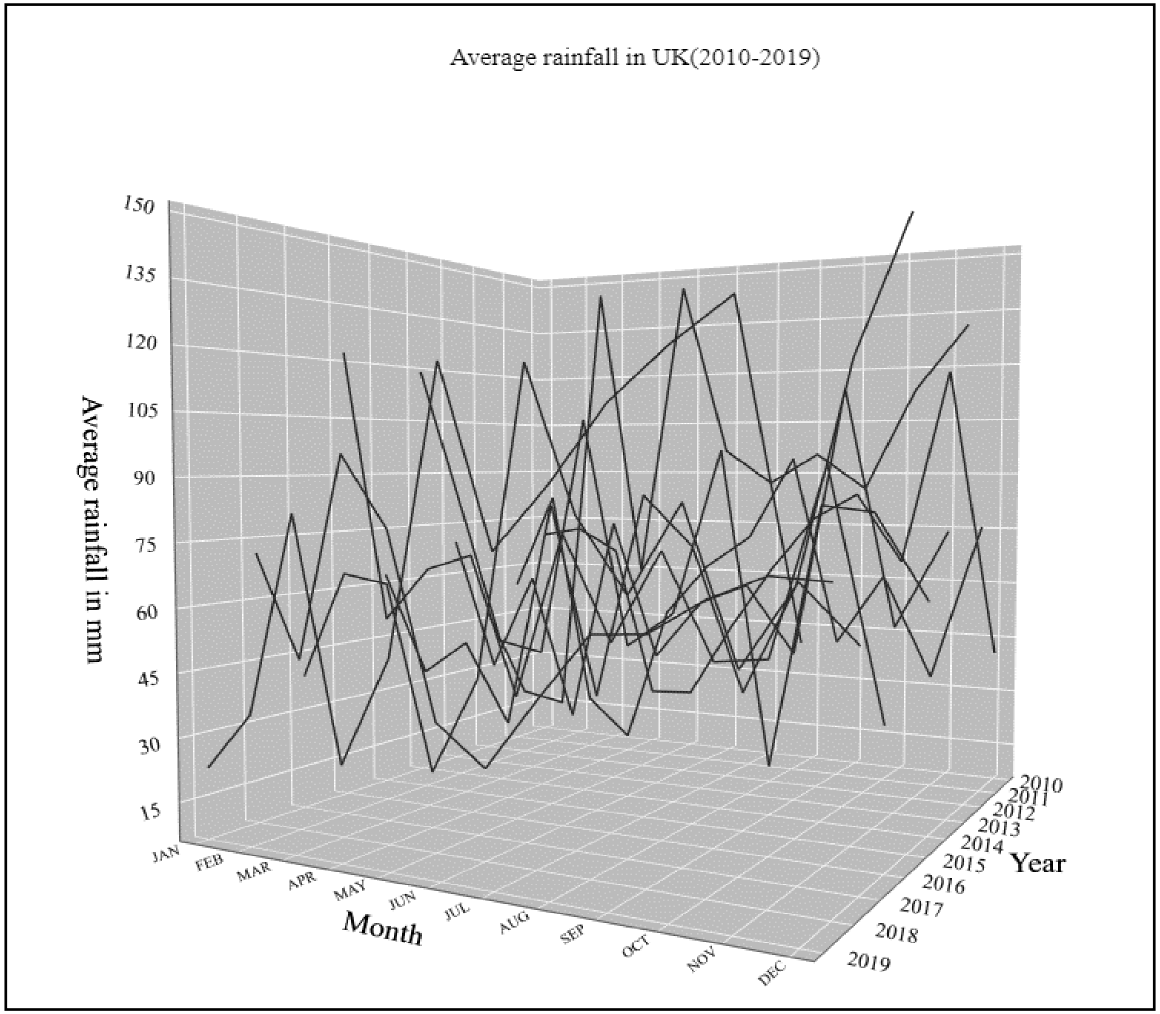


Figure 3.1: 3D line chart without data points

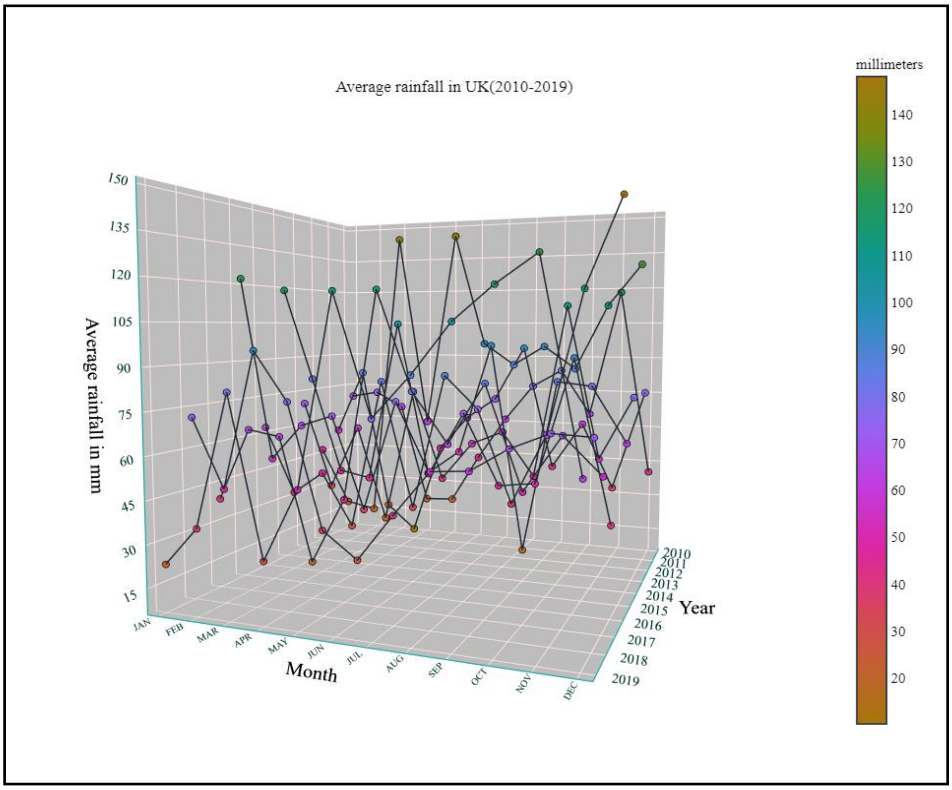


Figure 3.2: 3D line chart with color data points

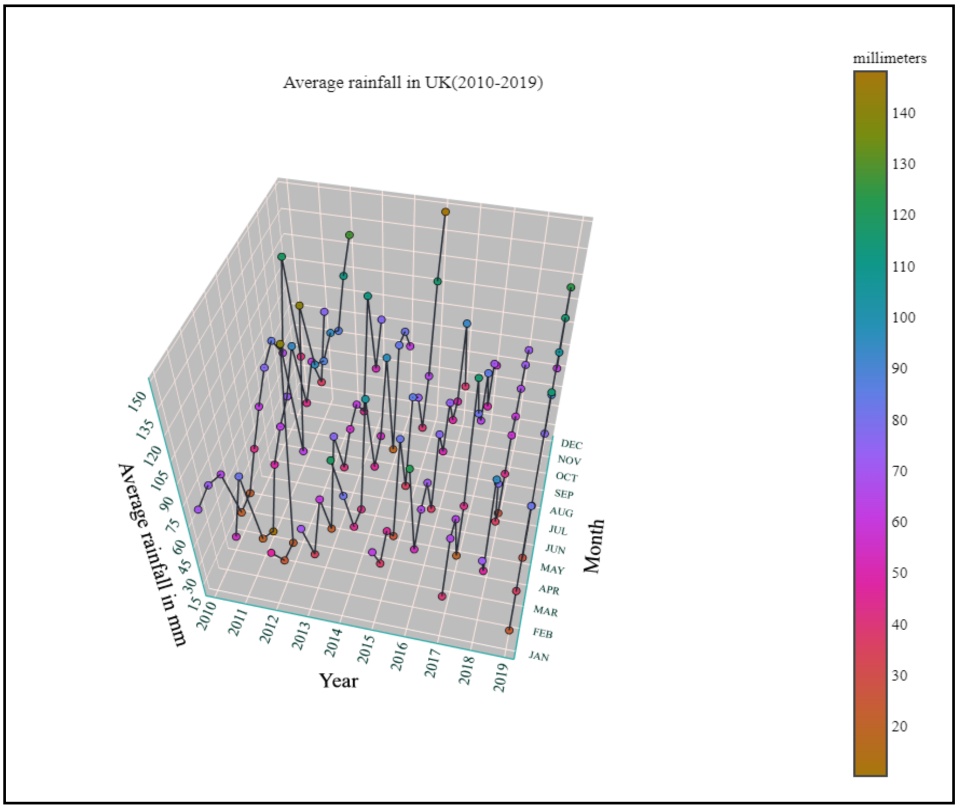


Figure 3.3: 3D line chart with color data points from top-view perspective

3.3.2 Line Chart with One Line

We have discussed a line chart with multiple lines in a 3D coordinate system in figure 3.3 which illustrated the average rainfall in the United Kingdom. The line chart in figure 3.4 is the common line chart in a 2D coordinate system having one single line which describes the bitcoin price from October 2013 to April 2022 [64]. The line charts are definitely one of the simplest sorts of charts to be read. The line in the chart is different from the other lines that we might have seen because of its style which is generated by dashes and dots. The style of our line is not related to the data or does not change our data, it is only a part of visualization process. The points which are on the line can build a thorough chart. The reader is able to locate the change happens in a specific period using the data points on the line. As it can viewed there are not frequent fluctuations from October 2013 to April 2017, but gradually some changes in price can be seen until September 2020. The line ascends suddenly from October 2020 to March 2021. There are two unforeseen decreases when the line plummets from March 2021 to May 2021 and from October 2021 to January 2022, respectively.

Normally line charts are drawn without points, but in our example we would like to have a more comprehensive and different line chart. There is one crucial note that should be taken into account is if we have a large amount of data and add data points to the line, we might cover the line chart which may create a scatter plot instead. Both x and y axes have sufficient number of ticks to provide a fairly definite value. Numerous modifications can be applied to the line chart such as changing the size and the color of data points, the thickness, the style and the color of the line, the background color, and the colors of axes and the style of the gridlines. The chart provides the user to hover over the line or the data point to see the time and the price. Zooming in on a specific segment of the line will make us able to observe the points and their corresponding periods accurately.

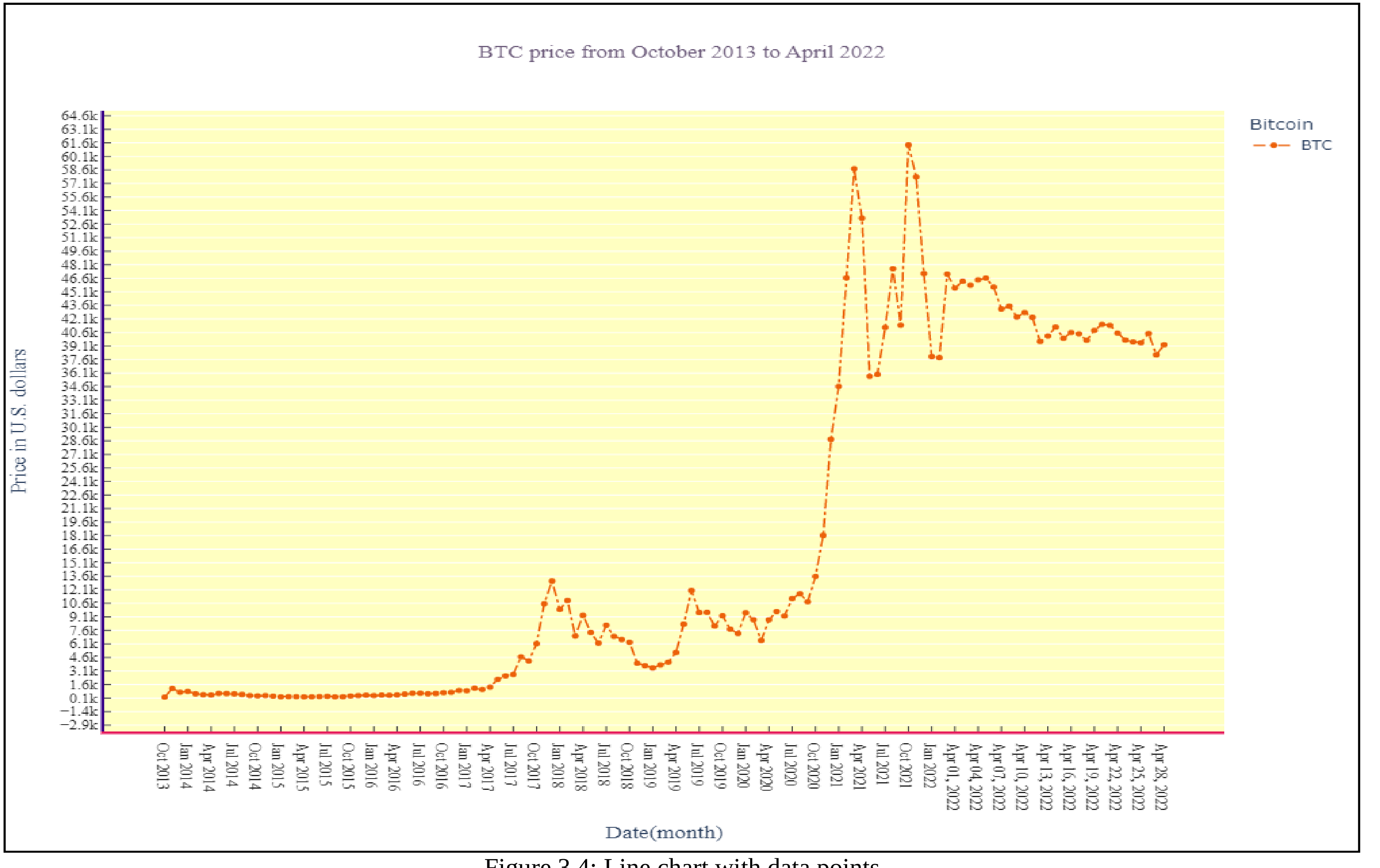


Figure 3.4: Line chart with data points

3.2.3 Line Chart with Multiple Lines

The line chart plotted in figure 3.5 describes the stocks data of three popular companies, Amazon, Apple and Facebook from 2016 to 2021 [65]. Line chart is one of the common types of charts in the realm of statistics and data visualization. A line chart can be read without any difficulty by any user at any level. In our line chart we have three different lines which implies there is no limit on the number of lines in a line chart. We need to be concerned regarding the number of lines not to have a complicated and chaotic chart which is normally called a spaghetti chart. The line chart depicts the changes in stock data over a continuous time for three companies. It is totally obvious that we do not have many fluctuations in values for Apple and Facebook companies but, the changes in values for Amazon Company are noticeable. The line chart has two variables, one numeric and the other one is a combination of numeric and categorical because we have shown year and month together. The axes x and y represent the time and the maximum stock price, respectively. Having both horizontal and vertical gridlines assists the readers to simply identify the corresponding value of the desired time. On each line we might add data points to make the line chart more complete. When we have many data points there should not be points on the lines because the points will be closely next to each other without any gap. As a result adding data points first requires a reasonable number of data.

We might have seen line charts with a shaded areas below each line as depicted in figure 3.6, shading the area underneath each line is not a necessity, but can provide more elegance to the chart. Although the ticks on x axis are combination of months and years, but we can add more ticks to the axis. The background color, the color of axes, the thickness of gridlines and the colors of lines can be altered using our Python code. When the user clicks on each line in the legend, the line will be invisible from the chart temporarily to compare the remaining lines without having any confusion. By hovering over the lines the reader can observe the time and the maximum stock price of the corresponding company. When the user zooms in on a part of a line they can see the changes more precisely.

Figure 3.5: Line chart with multiple lines

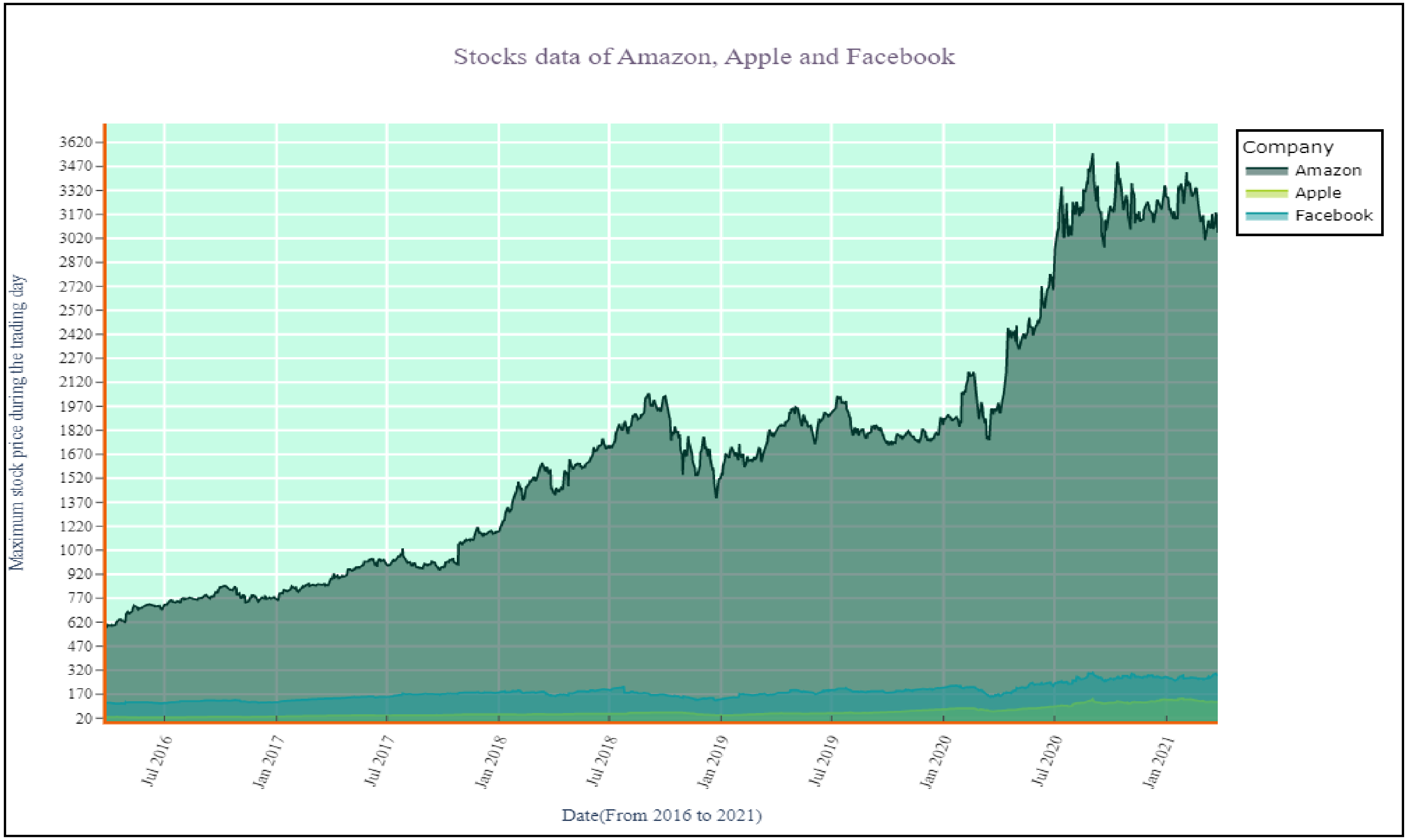


Figure 3.6: Line chart with multiple lines and shaded areas underneath

3.3 Scatter Plots

The scatterplot is one of the most frequent visualizations used to show correlations between two variables—one of the variables is displayed along a horizontal axis, and the other along a vertical axis. In the constructed space, the individual observations are plotted. In contrast to a bar chart, the scatterplot axes do not always have to begin at zero, especially if zero is not a conceivable value for the data series. We need to consider two critical factors to avoid using scatter plot, if we have a really small or huge datasets and the variables are not correlated [1, 4].

3.3.1 3D Scatter Plot

The 3D scatter plot or scatter chart in figure 3.7 describes the average rainfall in the United Kingdom [63]. Unlike a two dimensional scatterplot we can have more attributes shown in a 3D scatterplot as it was demonstrated in figure 3.3. Normally in scatter plots we do not observe color points scattered on the plot, but in the field of data visualization the primary purpose is to make the data easy to read and more comprehensive compared to old-fashioned visualization techniques. In the 3D scatterplot there is a horizontal color bar which gives a color scale corresponds to each point on the plot. As it can be seen the audience is able to identify the highest and the lowest values according to the color scale given on top of the plot. A note that we need to take into account is the colors appeared in the color bar can be another set of continuous colors depending on the preference. There are 3 axes in the plot, x-axis, y-axis and z-axis representing years, months and the amount of average rainfall respectively. The correlation of three variables is displayed in the 3D scatter plot which cannot be achieved using a 2D scatterplot. We have some numbers below the color bar called “ticks” can be altered to be more accurate with less gap between each pair of numbers. The size of points in our 3D scatterplot must be defined rationally based on the number of data that we have. Having large points on the plot can lead to a chaotic and unreadable plot due to numerous overlaps that hide some other points on the plot. In the interactive version of the plot the user is capable of hovering each data point to see its value, change the perspective of the plot and zoom in on a specific part of the plot to see more gaps between points.

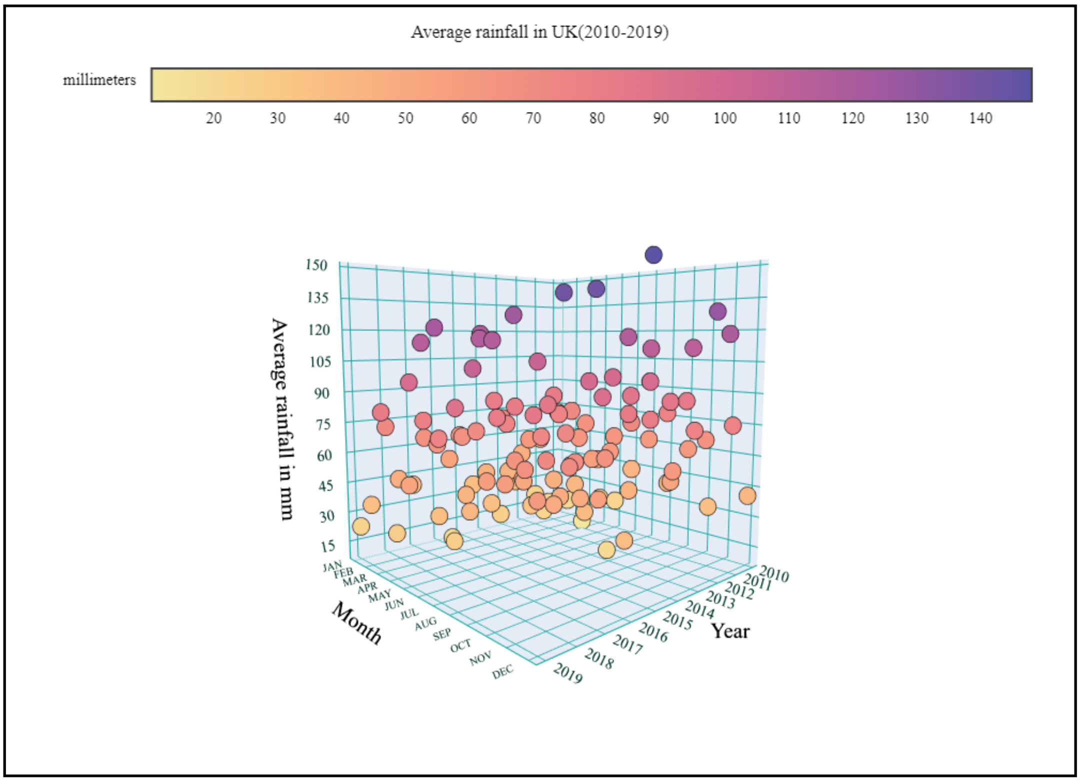


Figure 3.7: 3D scatter plot with color points

3.3.2 2D Scatter Plot

Demonstration of relationship between two or more variables can be done by one of the popular charts called scatter plot. The plot in figure 3.8 is a 2D scatter plot fairly comparable to the 3D scatter plot shown in figure 3.7. The apparent difference between them is in a 2D scatter plot we have all the points distributed on a 2D coordinate system, whereas there is z axis in the 3D scatter plot in figure 3.7. In this example the chart illustrates the relationship among three numeric variables height, weight and body mass index briefly BMI of 500 people [66]. The x and y axes represent weight in kilograms and height in centimeters, respectively. The BMI is expressed as a color of each data point. The colors are utilized here are continuous because we need to consider the decimals as well. The color bar next to the plot guides the readers to figure out the relationship between the colors and values. The points are relatively cover the entire plot which can indicate there is not density in the plot. According to the color bar we can comprehend that most people have BMI more than 4 and there are few people with BMI less than 1. The number of ticks on each axis can be either extended or reduced depending on the accuracy that we need to give to our plot. Based on the preference the color set which used in this plot can be changed into other set of continuous colors. The thickness and the color of borderline around each data point can be modified by the Python code. The user is able to zoom in on some points to have fewer points to analyze or slide to either right or left side by bringing the mouse cursor to the x axis. The y axis can also give the feature to the user to slide up or down the plot. By hovering over each data point the user will be capable of observing the values of x and y axes as a tuple and the corresponding BMI value.

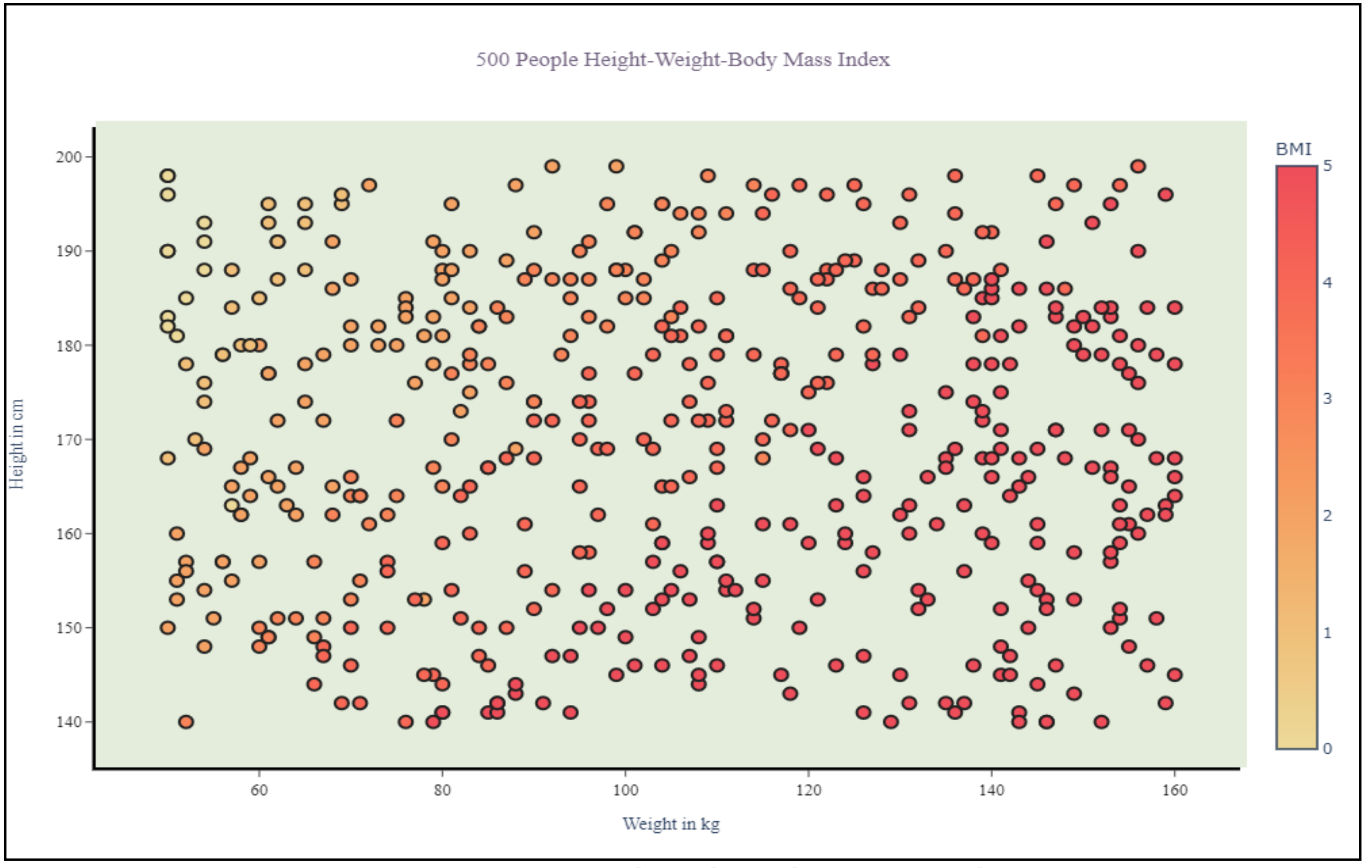


Figure 3.8: 2D scatter plot with 500 data points in color

3.4 Bar Charts

In bar and column charts the height or length of the rectangular bars demonstrates the value of your data and is one of the most recognizable data visualizations. The rectangles can be positioned along the vertical axis such that the bars lie horizontally (called a bar chart) or vertically (often addressed as column chart) [1]. Bar graphs can be used to compare data across groups or to follow changes over time. Bar graphs are especially effective when there are significant changes or when comparing one group to another [2]. Despite its great efficiency, bar plot is frequently considered as a dull way to show information. There are a few variants such as lollipop (figure 3.9) [6] and circular bar plot (figure 3.10) [7] that allow for more eye-catching numbers while maintaining bar plot accuracy. We should remember that bar charts and histograms are totally different from each other, a histogram only comprises numeric variable and illustrate its distribution [5].

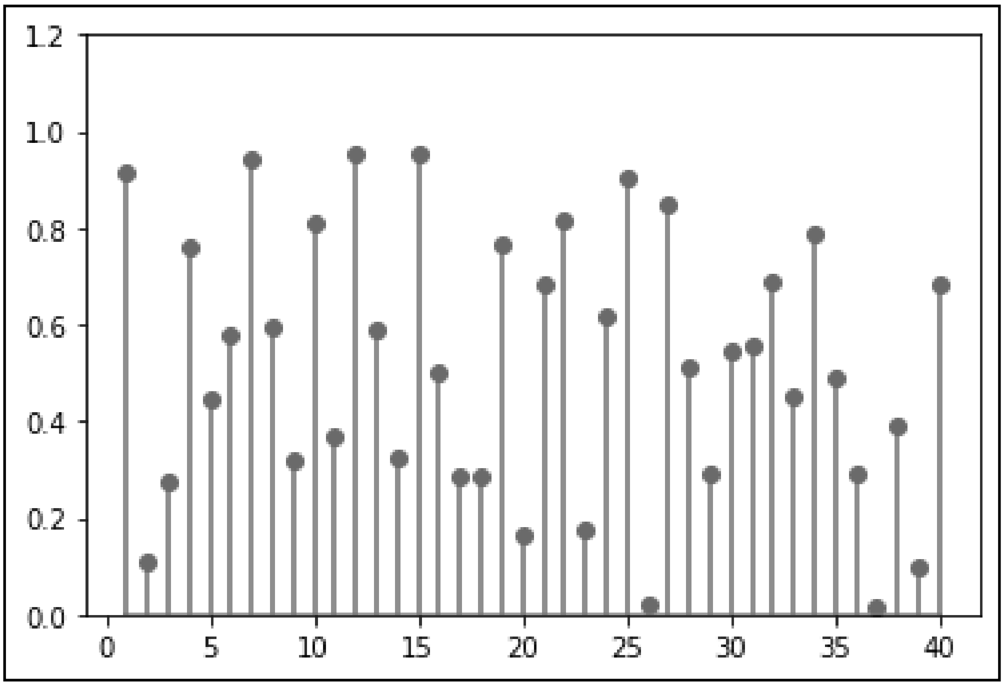


Figure 3.9: Lollipop chart, a variation of bar plot

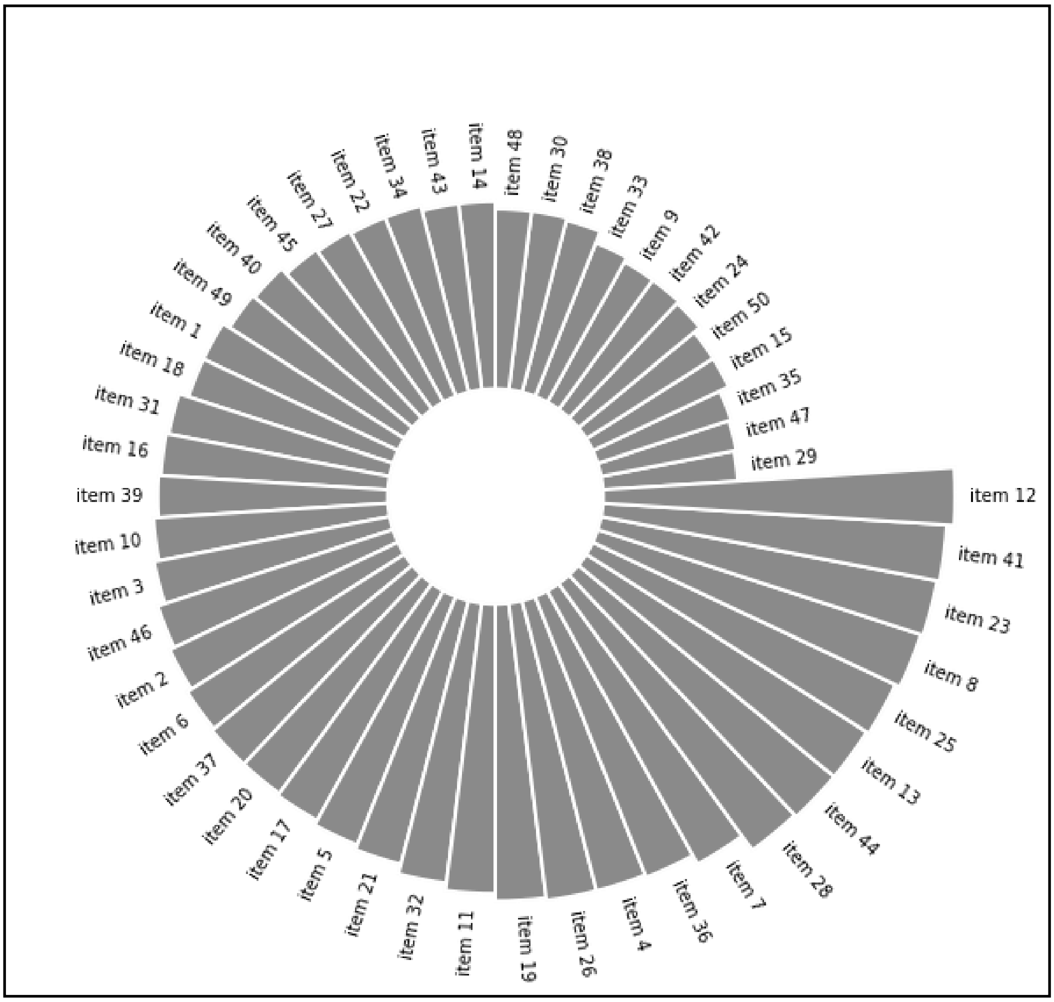


Figure 3.10: Circular bar chart, a variation of bar plot

3.4.1 Bar Chart with More Than 10 Bars

The bar chart or column chart in figure 3.11 illustrates OPEC oil price annually from 1960 to 2022 [67]. The bar chart in figure 3.11 contains two numeric variables, year and price. An appropriate comparison between each year can be made by simply looking at the years that we want to consider. In our bar chart the number of years on the x-axis is fixed, but we can add more intervals on the y-axis which depicts the price of each barrel in U.S. dollars. We can include more ticks to the y-axis to have a more precise price for each barrel. The gap between each bar is essential because the gap enables us to distinguish the difference between a bar chart and a histogram. The color, the width, the line color of each bar can be altered based on our preference. In the chart we can effortlessly observe that the price of one barrel was significantly cheaper before the year 1974. The notable fluctuations occur after the year 1990. One factor that we ought to deal with in bar charts is that visualizing a huge amount of data with a bar chart cannot be a decent idea to implement because there will be many bars to analyze. The size of the gap between each bar in our bar chart is sufficient, by increasing or decreasing the gap irrationally we might create a confusing bar chart.

At first glance, the audience will notify that there is no grid either horizontally or vertically which is commonplace in all bar charts. But with a simple modification in the related Python code of the chart we can add the grids to our plot depending on the requirements. The ticks on the x-axis might seem unusual because of their angles, but we can rotate them in different directions or make them vertical as ticks we see in other bar charts. One trivial yet important part regarding the bar charts is the initial point always starts at zero. In our bar charts the initial point of each bar is number zero. In the Jupyter notebook environment, one can hover over each bar to see the corresponding year and price per barrel. Zooming in on some bars is another feature that the user can benefit from. The Python code mentioned in page 43 can generate the bar chart, some familiarity with Python programming language can help the end users to customize the bar chart according to their needs such as changing the color of bars, modifying the width of axes, increasing or decreasing the number of ticks on each axis. There are more advanced customization options that users can apply by getting help from the documentation of plotly library [61].

```
import pandas as pd
import plotly.graph_objects as go
oil=pd.read_excel('Opec_Oil_price.xlsx')
fig=go.Figure()
fig.add_traces(go.Bar(x=oil.Year,y=oil.Price,marker_color='#00221C',name='oil barrel', marker_line=dict(width=1.2,color='#028476'),orientation='v'))
fig.update_layout(height=700,width=1000,plot_bgcolor='#c6fce5',
title_text='OPEC oil price annually 1960-2022', xaxis_title='Year',
yaxis_title='In U.S. dollars per barrel', title_font_family="Oswald",
title_font_color='#6c5b7b',title_x=0.5,title_font_size=20,
xaxis=dict(showline=True,showgrid=False,showticklabels=True,
linecolor='#A5CFE3', tickangle=-70,linewidth=2.5, ticks="outside",
tick0=1960,dtick=1, tickfont=dict(family='Oswald',size=11,color='#212121'),
zeroline=True,gridwidth=2,gridcolor='white',
title=dict(font_size=14,font_family='Oswald')),
yaxis=dict(showline=True,showgrid=False,linecolor='#A5CFE3',linewidth=2.5,
tickfont=dict(family='Oswald',size=11,color='#212121'),showticklabels=True,
ticks="outside",tick0=0,dtick=10,title_font=dict(family='Oswald')),
showlegend=True,bargap=0.3)
```

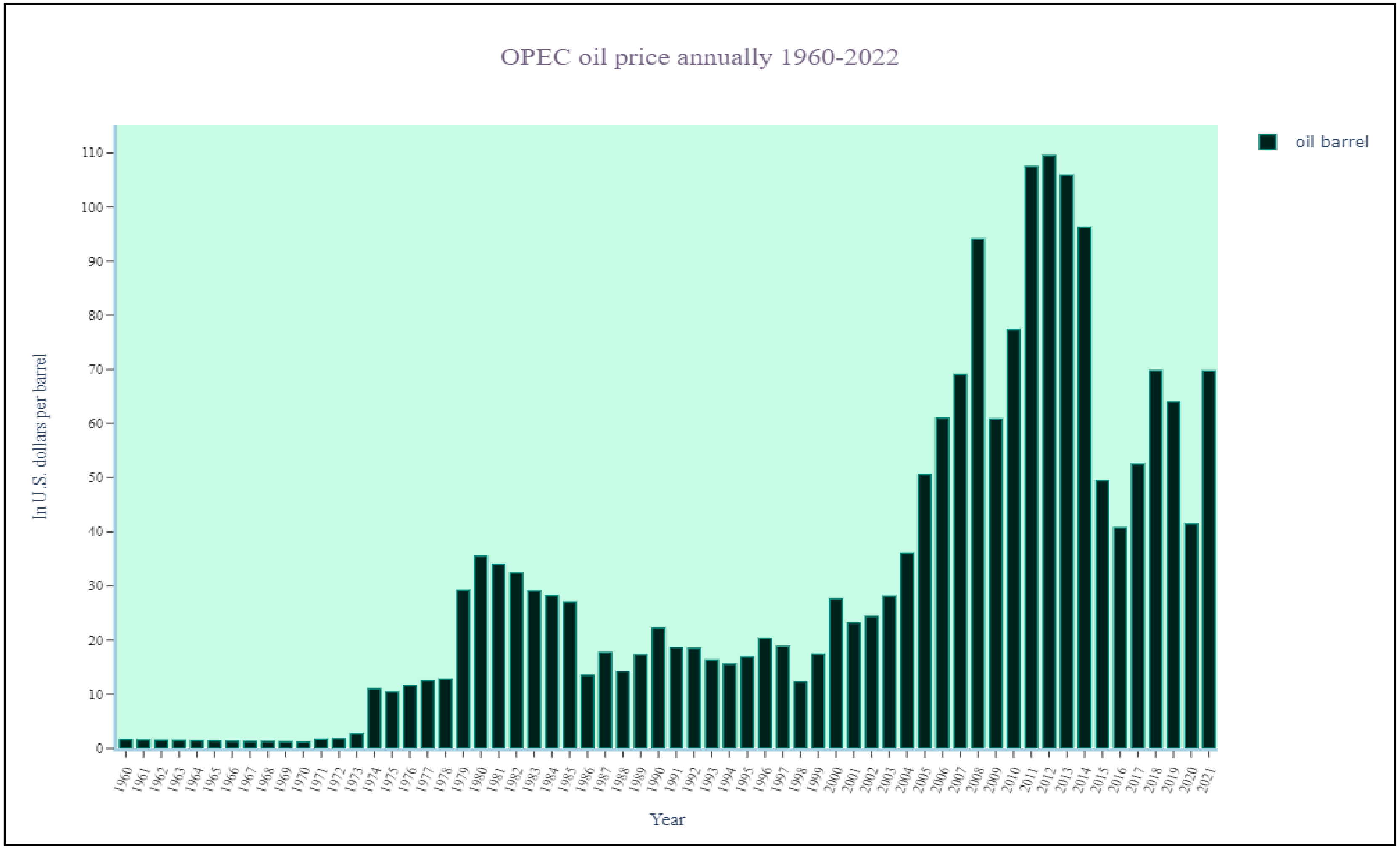


Figure 3.11: Bar chart with more than 10 bars

3.4.2 Horizontal Bar Chart

In figure 3.11 we have discussed the bar chart, there are diverse variations of bar chart, and we discuss the horizontal variation in figure 3.12. The chart illustrates the brand value of the 10 most valuable quick service restaurant brands worldwide in years 2020 and 2021 [68]. Two variables are discussed in the chart, numeric and categorical demonstrated on x and y axes, respectively. Typically a bar for each numeric or categorical value is expected in a bar chart, but when we have more than one bar for each value one or more of them can be missed depending on the total number of bars for each value. In our chart we have two horizontal bars for each categorical value shown on the y axis except the first value on the top of the axis. A question might arise, if there is any difference between a bar chart with vertical bars and one with horizontal bars? There is no distinction between them, depending on the visualization preference their direction can be chosen. As it can be observed there is an obvious gap between each pair of bars as well as between the bars themselves. Having spaces between bars help us to distinguish the dissimilarity between a bar chart and a histogram. A bar chart is one of the straightforward charts that users can effortlessly acquire the needed information and make a decent comparison. As it can be viewed a reader can quickly comprehend the difference between the years 2020 and 2021 by checking the bars. We may have more than two bars in each pair but, including many bars in a single pair make the chart sophisticated because the width of bars must be narrower.

As a result the final bar chart will not be pleasant for readers. The number of ticks on the x axis is also important to help the users approximate the value of the corresponding bar. The colors of bars in a bar chart with multiple bars for a value should match to build good-looking chart. The colors of bars and the background color can be modified by the Python code of the graph. Users can click on one of the years in the legend to make the year invisible temporarily to help them have a faster comparison. By clicking back on the year, the corresponding bars will be appeared again. The value and the name of the restaurant will be appeared once the user hovers over any bar. Zooming in on any specific region of the bar chart is a feature may be helpful if the readers prefer to have bigger and magnified bars.

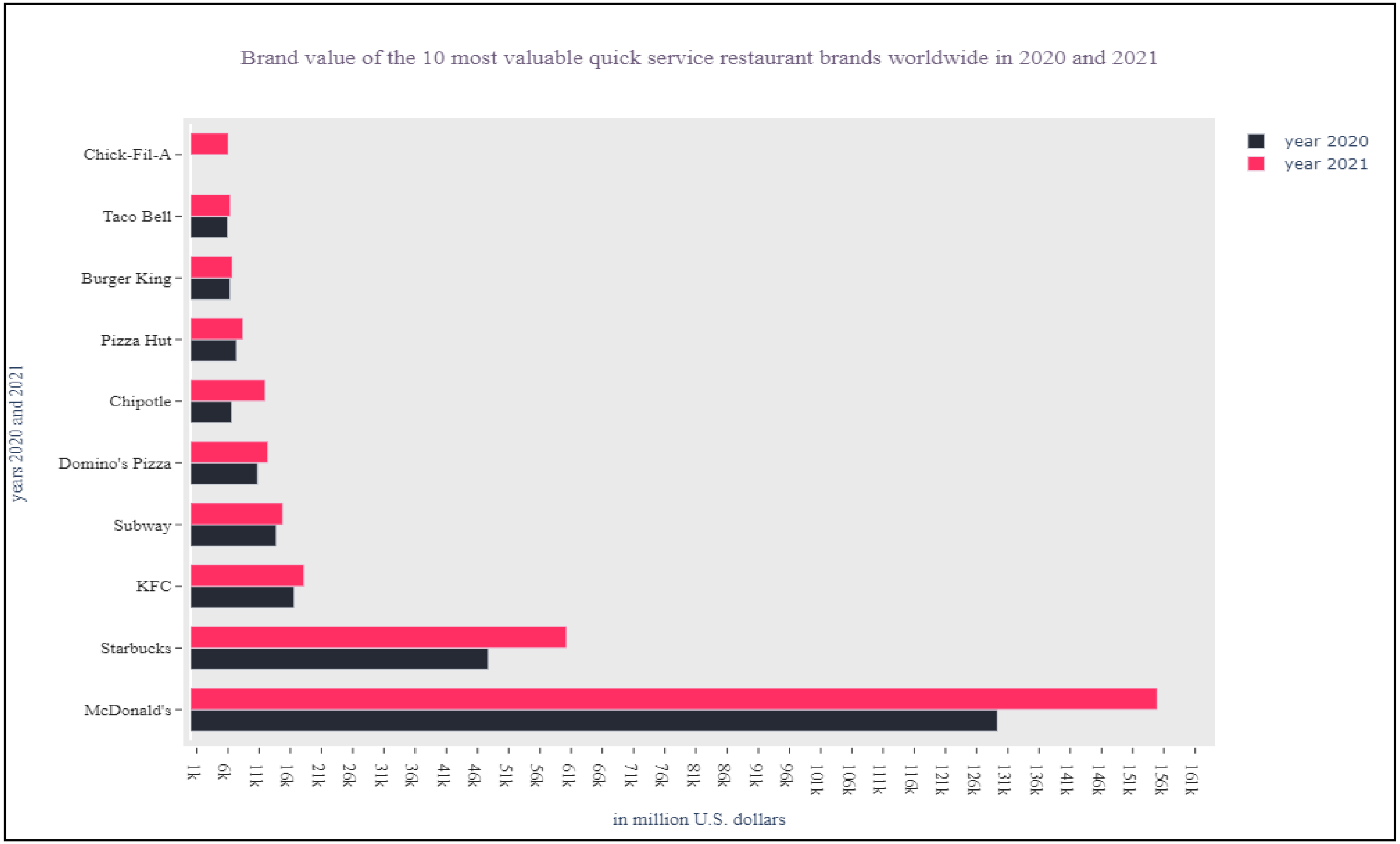


Figure 3.12: Horizontal bar chart

3.4.3 Stacked Bar Chart

Two variations of bar chart have been discussed so far in figures 3.11 and 3.12. Now we review another variation of bar chart called stacked. The stacked bar chart in figure 3.13 explains the brand value of the 10 most valuable quick service restaurant brands worldwide in years 2020 and 2021 [68]. There are two variables, one categorical and numeric which are represented by x and y axes, respectively. In this variation of bar chart the comparison between variables can be effortlessly done thanks to the segments of data stacked on top of each other. The bars can be either horizontal or vertical depending on our enthusiasm or preference. The brand value of McDonald’s is the highest among all other brands with an obvious difference and the least valuable brand is Chick-Fil-A. In some cases we might not have all the available data to visualize, in stacked bar chart it can cause a tremendous problem unless the data must be inserted to our chart. As it can be seen the data of year 2020 is missing for Chick-Fil-A restaurant, but we are still capable of acquiring fairly precise information. Number of stacked segments can be more than two there is no limit to the number of data segments, but having a bar with many stacked portions will make the readers confused and present a sophisticated chart. In this example the number of ticks on the x axis is fixed because we are working with the names of the restaurant unless we need to neglect a piece of data, but the number of ticks on the y axis can be increased or decreased based on the accuracy that we want to provide to the readers. As we have mentioned in figure 3.23 which was regarding a histogram, the evident difference between a histogram and a bar chart is the gap between the bars. The space between each bar can be either reduced or expanded based on the amount of data or choice. The colors of data segments, background color can be changed using the Python code. If the user zoom in on some bars, they can observe them with more details. To see the value of each portion and the name of the corresponding restaurant, the user requires to hover over any data chunk. Clicking on each year in the legend makes the related data portions invisible temporarily in order to concentrate on one or some specific segments.

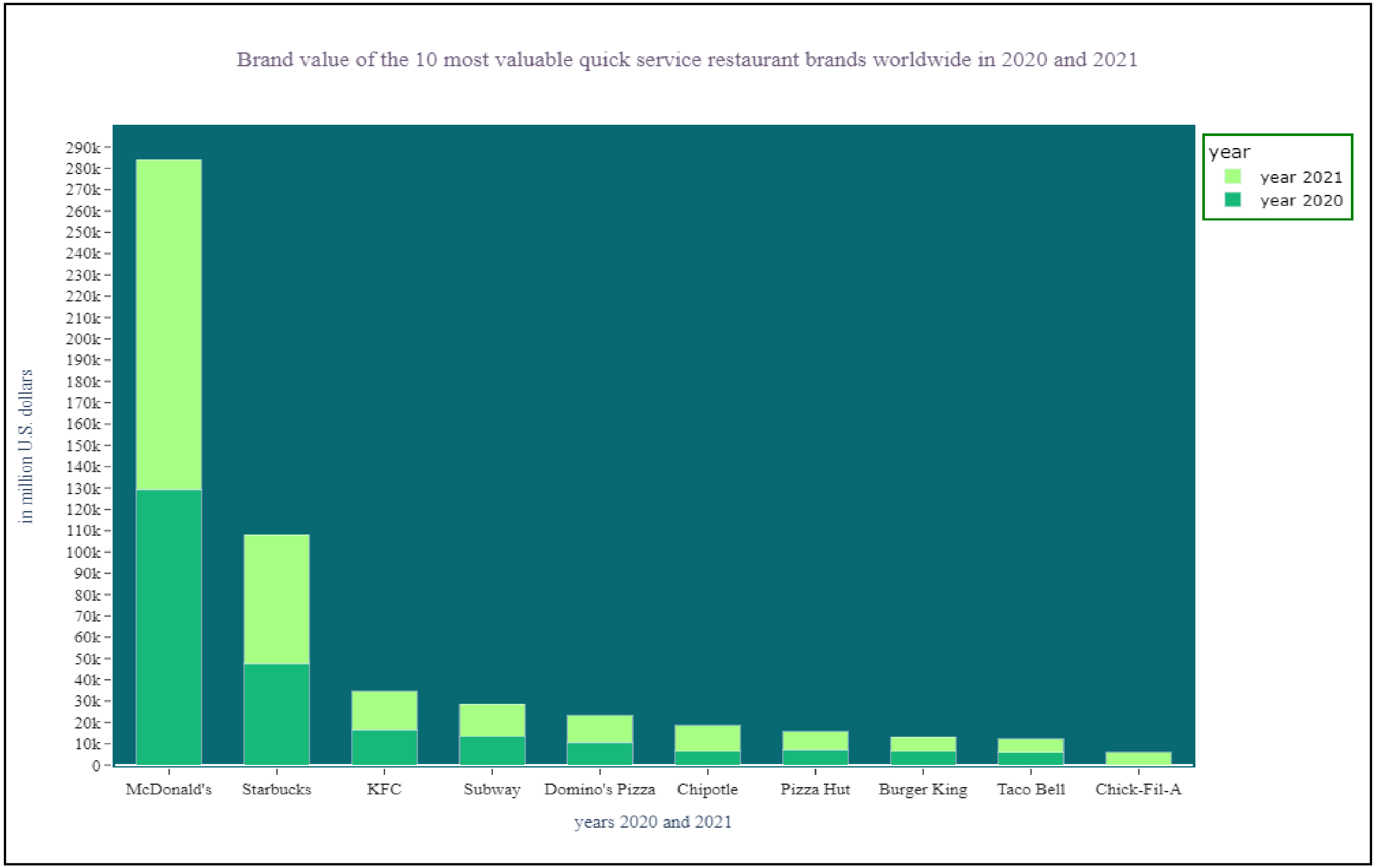


Figure 3.13: Stacked vertical bar chart

3.5 Box Plots

A box plot summarizes the distribution of values across measured groups by using boxes and whiskers. The placements of the box and whisker ends indicate the regions containing the majority of the data. When there are numerous groups to compare, box plots are most usually used; when there is only one group to plot, alternative charts with more information are favoured [8]. There are some relevant notes regarding which assist us to have a better understanding of box-and-whisker plots in the following [9]:

- Box and whiskers plot is another name for it.
- The median value is shown by the line in the centre of the box. This signifies that half of the data is above the median value and half of the data is below the median value.
- Medians are valuable because, unlike means, they are not influenced by outliers.
- There is **25%** of data higher than the median and **25%** of data below the median within the box, hence **50%** of the data is contained within the box.
- Using this type of graphic, we can quickly identify outliers and the plot's distribution.

Figure 3.14 [10] demonstrates all the above-mentioned notes.

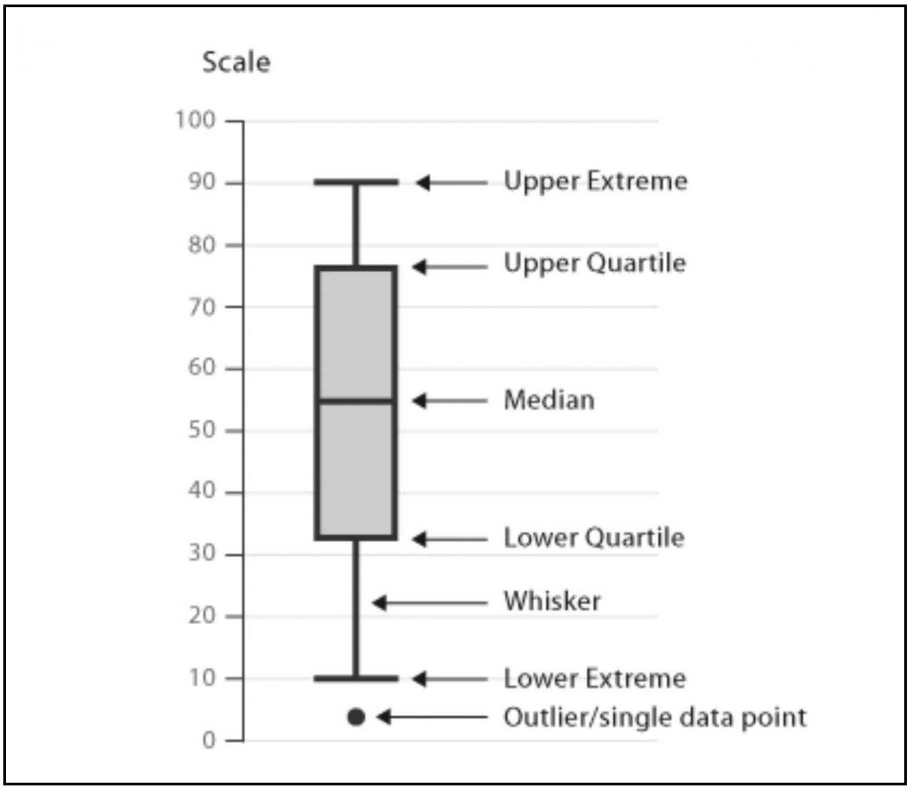


Figure 3.14: box-and-whisker plot

3.5.1 Box-and-Whisker

Box-and-whisker plot is well-known when the purpose is to demonstrate distribution. The box plot in figure 3.15 indicates beer and wine consumption according to the locations of countries in the hemispheres [69]. In the plot we do have two main variables, one is the beer and wine consumption per capita which is a numeric variable designated to y-axis and hemisphere which is a categorical variable that has three sub categories; “north”, “south” and “both” assigned to x-axis. The two diverse colors used for boxes give the reader the opportunity to easily spot the difference between boxes. We have six boxes in the plot which are grouped into three pairs to make the comparison straightforward. Each box consists of the minimum value, the maximum value and some points next to it. The points include both outliers and all the other values from our dataset. The horizontal lines at the end of each whisker demonstrates the minimum and maximum values located on the top and the bottom, respectively.

As it can be observed, in each box we have a solid horizontal line that represents the median value. Talking about median brings the other famous concepts, mean and standard deviation. In our boxes the dashed horizontal line defines the mean and the standard deviation.

At the end of each box we also have the first and the third quartiles. The ticks on the y-axis are assigned precisely in order to show the readers a detailed plot while the three ticks on the x-axis are stable. The legend next to the box plot can be beneficial when we need to figure out what is happening in the plot. The white-color grids can assist the readers to find the desired values simply. In the interactive version of our box plot the user can easily click on one of colors in the legend to make its corresponding boxes invisible to observe only the desired boxes. Once the reader hovers over one of the boxes they will be able to see all the mentioned statistical details. By a small alteration in the Python code we can have only the outliers of the dataset. One fairly uncommon feature concerning the box plot in figure 3.15 which produces a relatively unique plot is that all the boxes are vertical while most people are used to see the box-and-whisker plots horizontally.

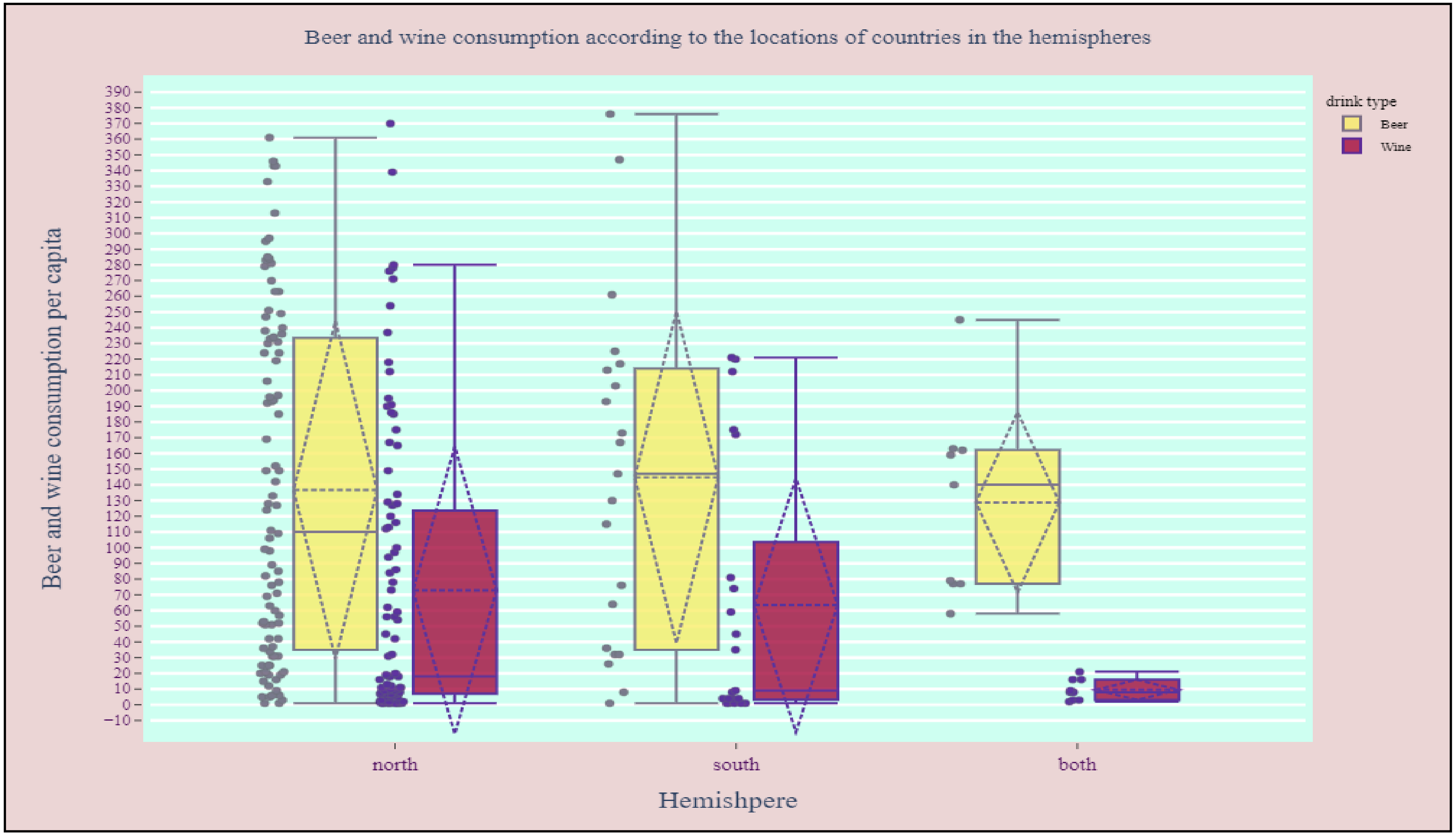


Figure 3.15: A chart with 6 box plots

3.6 Bubble Charts

A bubble chart is a sort of scatterplot with a third dimension: the size of the dots represents the value of an extra numeric variable. We will need three numerical variables as input: one for the X axis, one for the Y axis, and one for the dot size [11]. Circles indicate values in the basic bubble graph. These charts, like bar charts, are used to compare values between categories. Humans, unlike bar charts, are not very adept at comparing the diameters of circles. Circles, on the other hand, are more aesthetically appealing, may reinforce an image or metaphor, and are an excellent choice when determining exact proportions is not critical [1]. Avoidance of bubble charts is recommended when the dataset is small [12].

3.6.1 Bubble Chart Visualization

A bubble chart is seen as a scatter plot at first glimpse because of the dots spread all over the plot, but the sizes of dot bubbles are proportionally different based on our data. In some cases bubbles can have the same sizes. The bubble chart in figure 3.16 demonstrates the population by each country in year 2020 [70]. In the bubble chart we consider the relationship between percentage yearly change in population and total number of migrants for each country. The size of each bubble can be calculated by 3.1. Instead of number 2 in formula 3.1 [13] we can specify another positive integer in the numerator, but opting for a positive number can develop enormous bubbles in size which cover the whole plot. We construct a list comprising the populations of countries. We choose the maximum value from the list to have bubbles with proportional sizes. Number 90 in the formula can be any positive number greater than zero, but we need to take the sizes of bubbles into account because if we use a really small number the bubbles will not be visible on the plot and on the other size if we assign a very large number we might cover the whole plot and not be able to observe scattered bubbles. We utilized a continuous color scale in order to comprise both integers and decimals.

$$\frac{2 \times \max(size)}{90^2} \tag{3.1}$$

Where: size is the name of the list that we have constructed in our Python code, max is a built-in function in Python to find the maximum element in a sequence of values.

The x-axis represents the total number of migrants in million and y-axis represents the yearly change in population in percentage. One critical point that should be noted is the distribution of bubbles is directly related to the attributes that we select from our dataset. Some bubbles are overlapping in the chart, by changing the opacity value most of the bubbles can be seen. The horizontal color bar on top of the chart can give the audience an idea quickly on account of the continuous colors, as the color gets lighter we have negative change in the population but by going to the right of the color bar the color becomes darker which is an indication of positive change in the population. As it is shown in the chart each bubble has a pink line around which can be omitted, thicker or thinner based on our preference. For the purpose of accuracy we have many ticks on both x and y axes. The vertical grids are removed to have a more readable chart. Modification in the color scale, grid color, background color, and axes' colors can simply be made by changing the color names in the python code. The thick green is the zero line to separate positive and negative values for more readability. In the interactive version of our bubble chart when we hover over a bubble we see a tuple with two values, number of migrants and yearly change, respectively. The name of a country can be appeared when the mouse cursor is on the bubble. In the default mode of the chart we cannot observe all countries in particular the ones with small populations, but the user can see all countries by zooming in on a specific part of the chart.

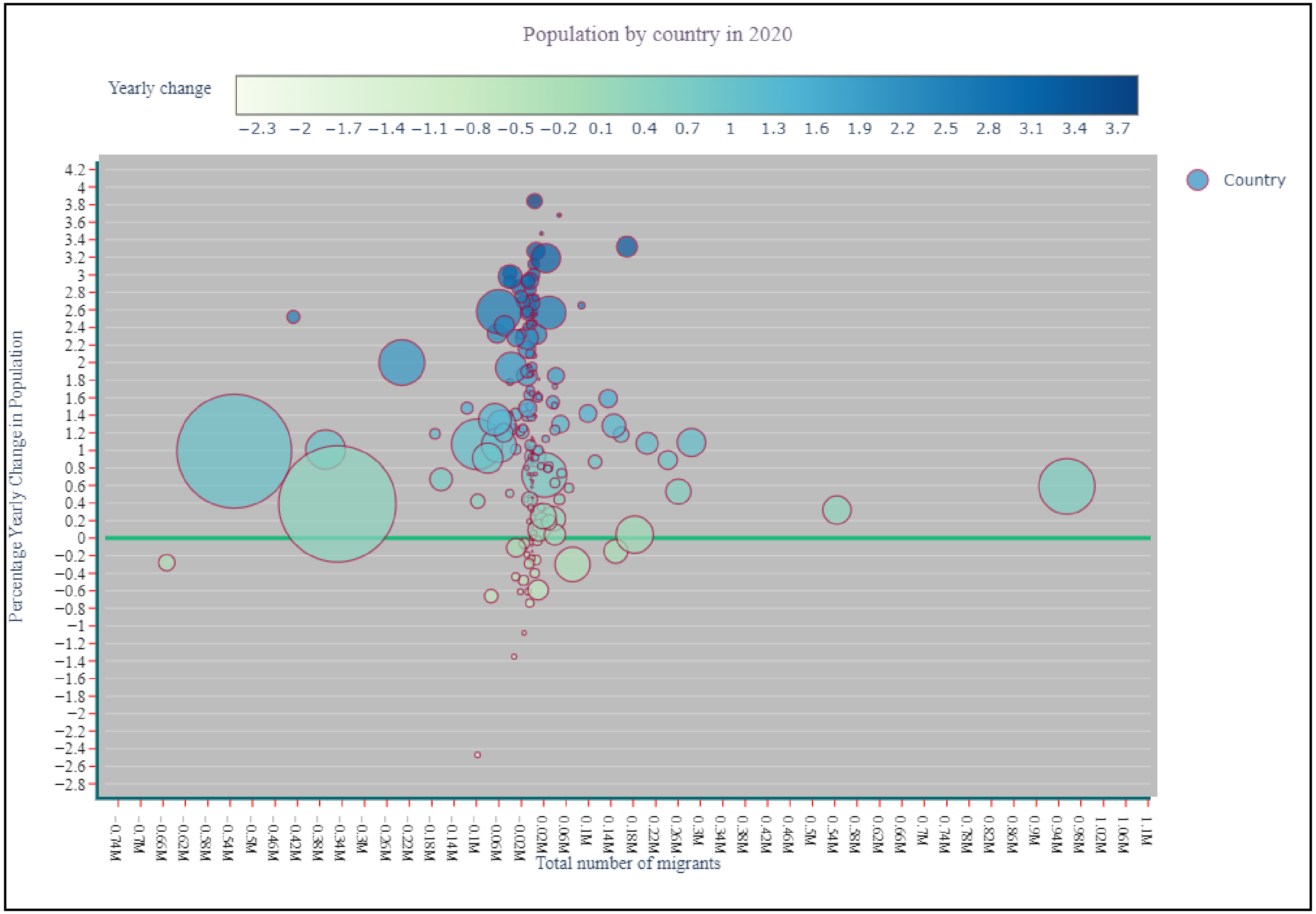


Figure 3.16: Bubble chart

3.7 Contour Plots

A Contour Plot (figure 3.17) [15] is a two-dimensional visual depiction of the connections between three numerical variables. The X and Y axes are represented by two variables, and the contour levels are represented by a third variable, Z [14]. The contour levels are shown as curves, with the region between curves colour coded to illustrate interpolated values.

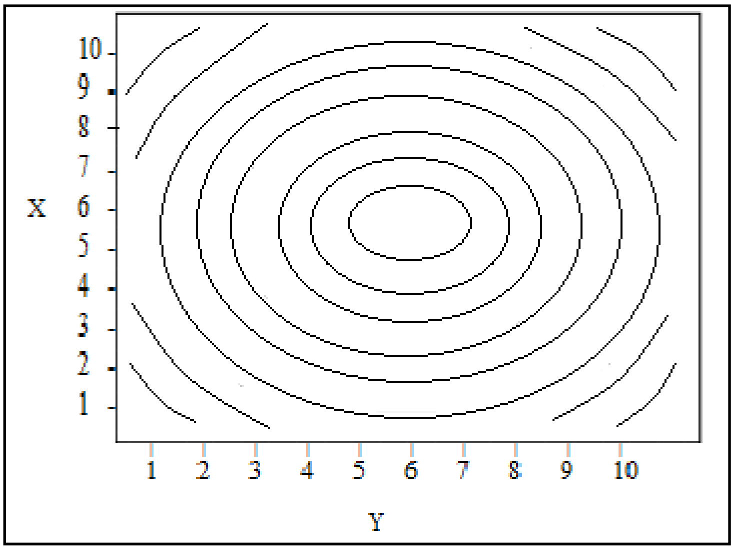


Figure 3.17: A simple contour plot

3.7.1 Contour Plot Visualization

The plot in figure 3.18 is called "contour plot" which is not commonly utilized compared to the other usual charts and graphs such as bar chart, line chart and histogram. The plot demonstrates the correlation among three numeric values, weight, height and body mass index of 500 people in a two dimensions [66]. The direction and the size of each shape of the plot can be diverse based on the data that we visualize. The axes x and y represent "weight in kilograms" and "height in centimeters". The third numeric variable is the BMI (Body Mass Index) which is the values in white at the border of each level of the contour plot. The number of contours is chosen automatically to be less than or equal to the value of “ncontours” which is an attribute in the python code. As it can be viewed there are 9 different levels with a unique color for each. The colors applied to the contours are continuous in order to include the decimals for the third numeric variable.

The size of each contour implies the number of values it contain, the more values the contour has the larger its area gets. The color bar assists us to detect the relationship between the color and the corresponding contour. We can also have unfilled contours only with borders visible or contours with mesh color which make them similar to a heat map. Any user can swiftly realize there are many people with BMI more than or equal to 4 by only looking at the plot. A valuable advantage of contour plots is that they are straightforward to be read and deliver the purpose of data to the readers in the simplest way. We should keep in mind that contour plots are not appropriate for two numeric variables. If we prefer to have filled contours with colors we will not be capable of seeing the borderlines, but by making the contours unfilled we will be able to alter the width, color and style of the borderline. In the interactive version of the plot the user can zoom in on a particular area to see more values on the contours. When the mouse cursor is on any part of the contour all three values of the numeric variables can be seen accordingly.

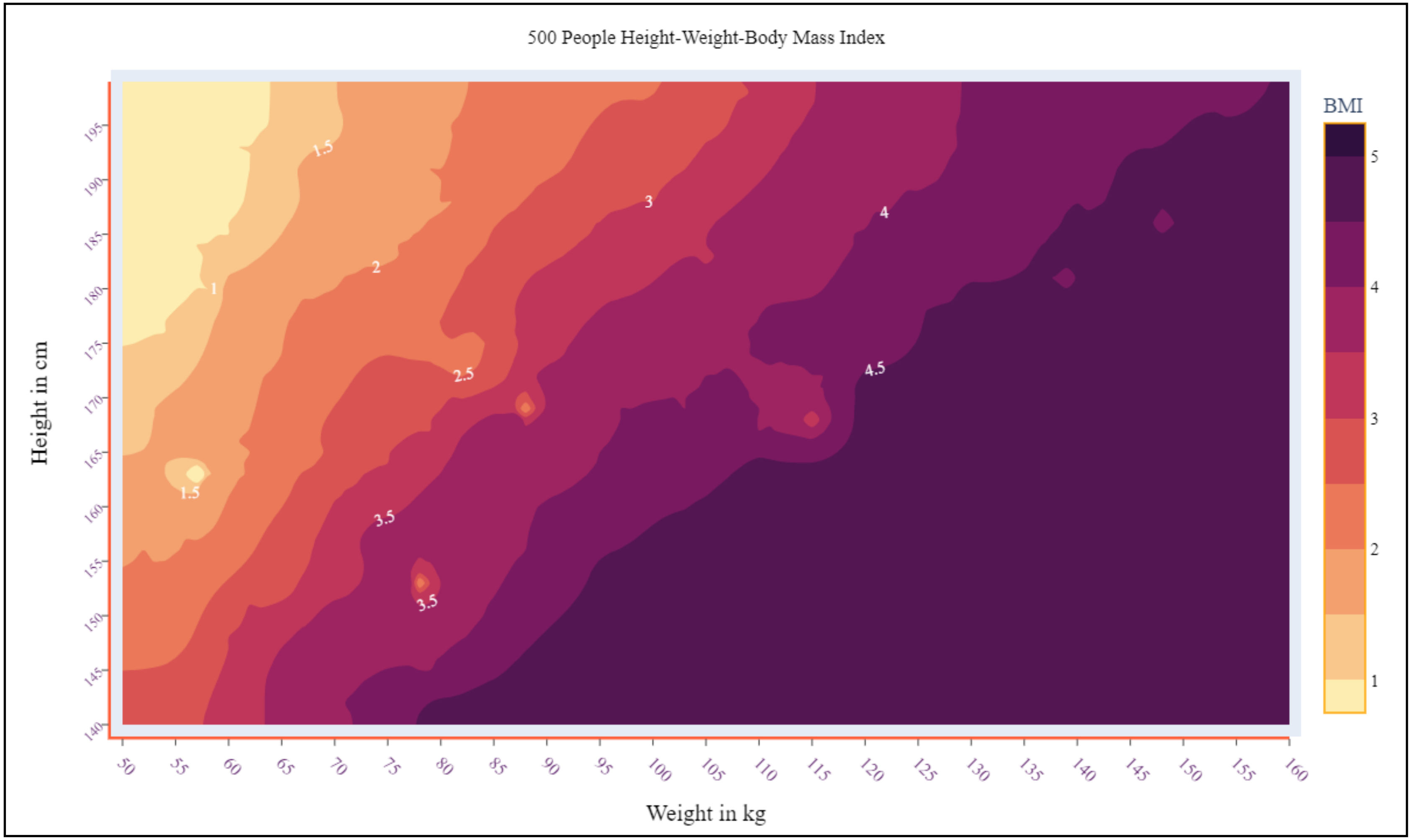


Figure 3.18: Contour plot on a 2D plane

3.8 Funnel Charts

A funnel chart (figure 3.19) [17] is a graphical depiction of how data flows through a process. In a funnel chart, the value of the dependent variable decreases as the process progresses. Funnel charts are commonly used to depict sales funnels, recruiting processes, and order fulfilment procedures [16].

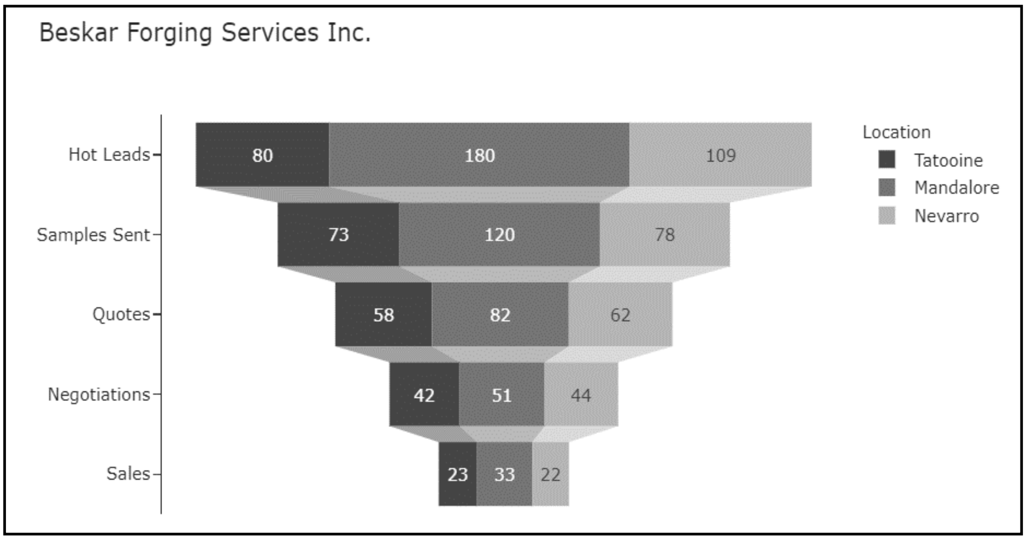


Figure 3.19: Funnel chart with 3 segments at each level

3.8.1 Funnel Chart with Two Numeric Variables

The plot in figure 3.20 is a funnel chart which illustrates the brand value of the 6 most valuable quick service restaurant brands worldwide in 2020 and 2021 [68]. In many circumstances there is a single series of data, but in the funnel chart shown in figure 3.20 there are two series of data, for years 2020 and 2021. The name “funnel” certainly suits the shape of the chart because the values of the dataset create a funnel-shaped chart. In most funnel chart there will be no use of x axis because a funnel chart displays some values based on a categorical variable as a result the existence of y-axis is mandatory. Our funnel chart demonstrates a comparison between two years, 2020 and 2021. Any type of user either novice or professional will quickly understand what the chart shows thanks to the values appeared on each segment and the sizes. The chunks of the funnel chart are stack on top of each other pair by pair, each pair depicts two years, and by changing a value of an attribute in the Python code of the chart we can have pairs grouped or overlaid. We should pay attention to the number of series that we need to visualize by a funnel chart to select the best funnel mode, in some cases we might hide the other series by choosing overlay mode.

A crucial important question arises is if we can have more than two series of data and the answer is yes, but the chunks might be very small where the values cannot be appeared properly. The comparison between years 2020 and 2021 has been made easy as a result of utilizing a funnel chart. All the segment on the right hand side of the chart occupy a bigger part from the whole, it is not necessary to always have larger right hand side. As we go down to the funnel the values are diminishing which an apparent characteristic of funnel charts unlike the pyramid charts. By looking at the chart we are able to observe that all the six brands have had an increase in their brand values compared to the previous year. In some conditions we might not have all the values in a series of data, but we will still be able to create a funnel chart with missing values. Creating a funnel chart with many missing values may result to a chart which will not look like a funnel that is an important note we should always consider in the process of visualization. The colors of segments, the colors of shades underneath each segment and the borderline color can be easily changed by our Python code. In the interactive version of the chart when the user hovers over a segment they can see the name and the value in percentage.

Figure 3.20: Funnel chart comparing two numeric variables

3.9 Heatmaps

Colours and colour saturations are used in heatmaps to depict data values. A heatmap (figure 3.21) [18] is just a table with cells which are color-coded. They are frequently used to show high-frequency data or where recognizing overall patterns rather than exact numbers is more essential [1]. Heatmaps may also assist in recognizing patterns, making them useful for studying trends that change fast, such as ad conversions [2]. They can also assist with:

- Competitor analysis
- Customer opinion
- Campaign influence
- Demographics of customers

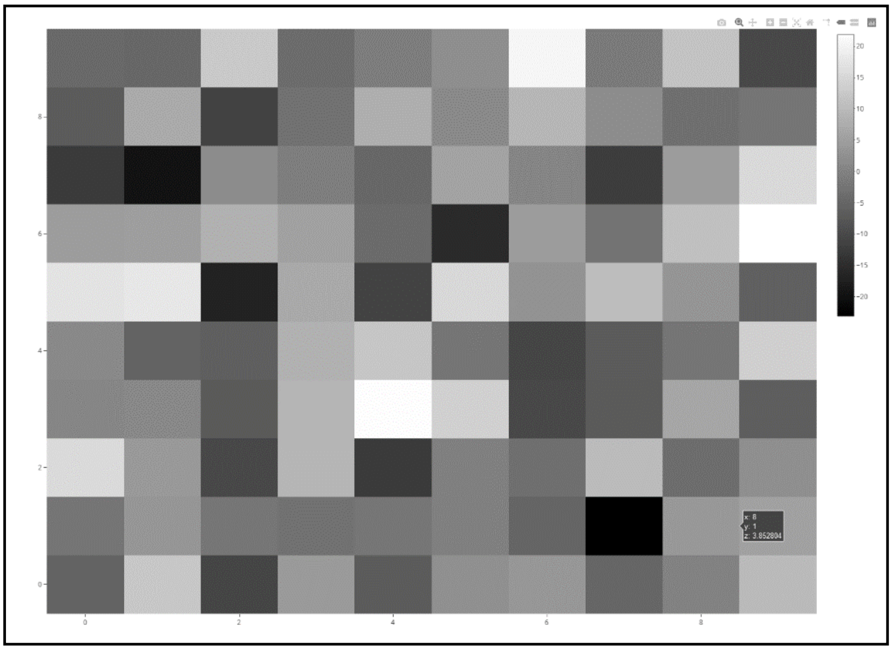


Figure 3.21: Heatmap with a color bar

3.9.1 Heatmap Visualization

The chart in figure 3.22 demonstrates the average rainfall in the United Kingdom from 2010 to 1019 [63]. Similar to other charts and plots in figures 3.7, 3.8, 3.16 and 3.18

we have applied different colors, in other words we utilized color saturation. A heatmap fairly similar to a table or a matrix because it contains many cells, one cell for each individual value of the main dataset. The color bar next to the heatmap helps the reader to navigate through the heatmap effortlessly and find the desired cell based on the corresponding color. The colors used in heatmaps are continuous to comprise all numbers not only integer numbers. As it is illustrated in the chart both x and y axes are existing, the x axis represents the years from 2010 to 2019 and the y axis show the months in a year. The borderline of each rectangular cell can be visible by changing a piece of the Python code, but normally heatmaps do not have borderlines for their cells in order to show the importance of colors in the visualization.

Heatmaps are utilized to attract the audience and display the data in smooth way, but we should not forget that visualizing a large dataset using a heatmap can generate many cells which will be tiny in size. In our chart we have three variables, one categorical and two numeric. The y and x axes depict the categorical and numeric variables, respectively and the other numeric variable belongs to the value of each cell which will be appeared on the each cell when the mouse cursor is on it in the interactive version of the chart. This sort of chart does not require any vertical or horizontal grids because of the cells cover the entire grid. Furthermore adding any gap between the cells either horizontally or vertically is uncommon in heatmap charts. Choosing another set of colors and changing the axes’ colors are based on the preference of the visualizer. If the simplicity and elegance are important, one of the best choices is to utilize a heatmap. In the interactive version of the chart the user can hover over any cell to observe its value which corresponds to the third numeric variable which the amount of the average rainfall in millimeters and the values of the corresponding x and y ticks, respectively. Zooming in on a specific part of the heatmap is also provided in the interactive version.

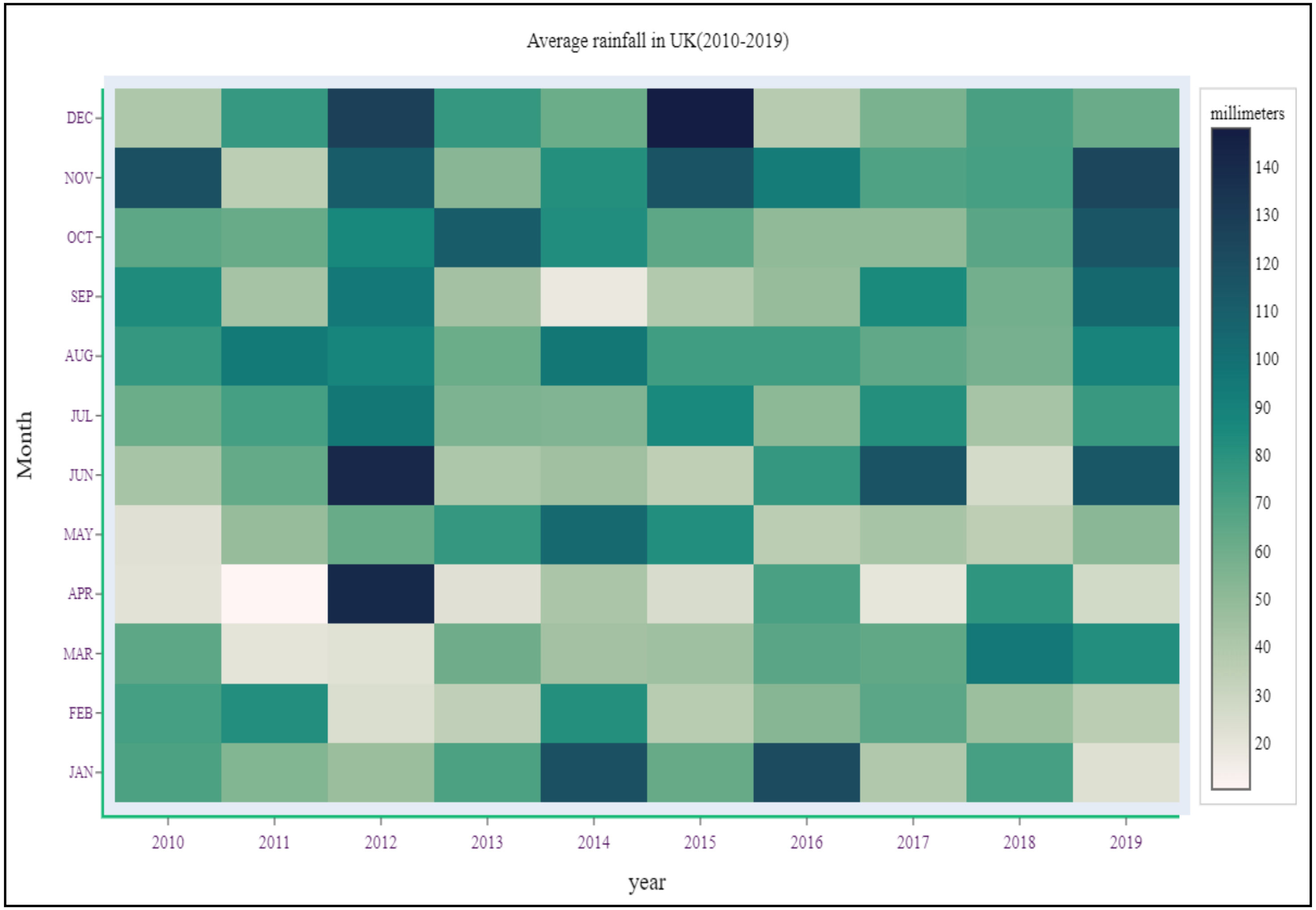


Figure 3.22: Heatmap visualization

3.10 Histograms

A histogram is a sort of vertical bar graph that displays numerical data and its frequency distribution. As the name implies, the distribution is frequently depicted across time, although the data might also be plotted on any historical scale, such as temperature, elevation, or monetary value. While histograms are most commonly associated with bar graphs, the notion may also be used to line graphs and other designs that rely on plotting two axes [19]. The primary step in creating a histogram is to bin the range of values and later calculate the number of values within each interval. A rectangle is drawn with a height proportionate to the count and a width equivalent to the bin size, such that the rectangles touch [20].

3.10.1 Histogram Visualization

One of the most popular charts in the field of statistics and data visualization is histogram. The histogram in figure 3.23 demonstrates total sulfur dioxide in 1600 red wine samples [71]. The main objective of a histogram is to depict how the data are distributed, it can simply be observed most values are within the interval 10 to 70 in our histogram. Regardless the knowledge of statistics an unprofessional user will be able to see the intervals which comprises the most and the least numbers of values. It might seem bizarre at first glance that there is a box-and-whisker plot on the top of the histogram, the reason is to have a diverse and more comprehensive histogram. The points distributed next to the right hand side whisker are the outliers. The maximum and minimum values of the box-and-whisker plot can be approximated thanks to the bins underneath them. If the reader checks carefully they can roughly obtain the maximum and the minimum values. The histogram includes 50 bins, but seventeen of them are shown properly and the reason is the other bins do not contain any values or have much less values compared to the other bins. The number of bins can be modified by the Python code depending on the accuracy of the histogram. One recognizable distinction between a bar chart and a histogram is that we do not have any gap between the bins in the histogram, but adding gaps in our histogram can be easily added using the Python code. In this chart we consider only one numeric variable which is represented by the x axis.

The interval for each bin can be modified based on our preference or how accurate we need the histogram to be. The ticks on the y axis make the users to acquire a precise value as much as possible, in histograms the number of tick on axes can play an important rule so we need to have sufficient number of ticks to give the reader at least a fairly exact value. In order not to have a chaotic histogram the vertical grid lines are omitted, but having the horizontal grid lines can assist the readers to find the corresponding value for each bin. The color and the opacity of the bins can be altered by the Python code. In the interactive version when the user hovers over the bins they can view the interval and the number of values in each bin. Furthermore, by hovering over the box plot on the top we will be able to observe quartiles, box's minimum and maximum values and all the outliers with their values.

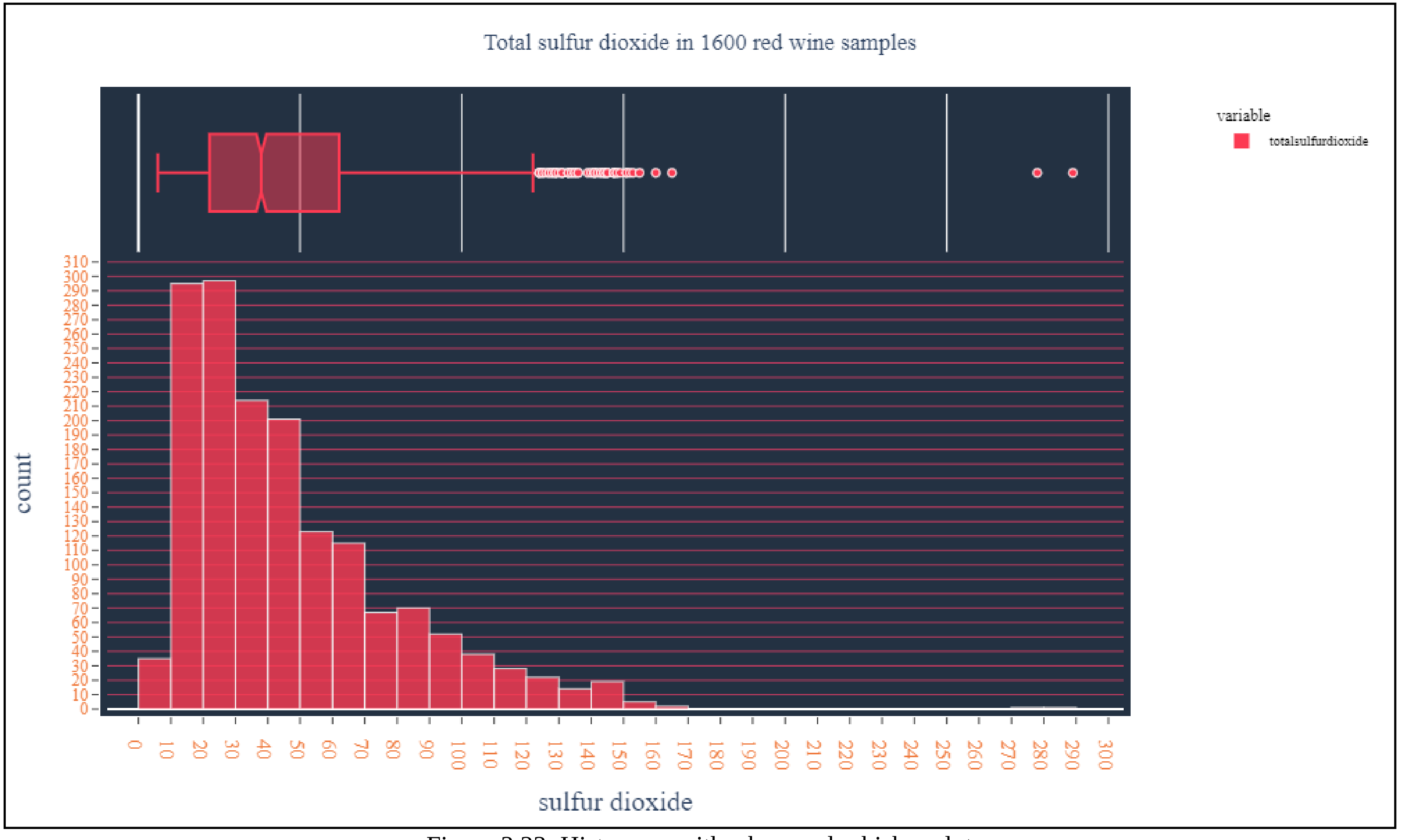


Figure 3.23: Histogram with a box-and-whisker plot

3.11 Pie Chart

A pie chart is occasionally used to demonstrate a part-to-whole connection or a composition by representing data in percentages. Pie charts are not intended to compare individual parts or to reflect exact numbers, instead we can use a bar chart for that [21]. While pie charts are not without criticism and complaints, they are also fairly familiar charts form for many individuals, and familiarity may be beneficial. Curves are also more appealing to humans than sharp points, according to research. In figure 3.24 we have a donut chart which is a variation of pie chart and used in many advertisements [23]. The following are some common disadvantages of pie charts [22]:

- It is impossible to compare two data sets
- The total represented by the chart is unknown
- And the chart is hard to interpret without labels (in particular with similarly sized slices)

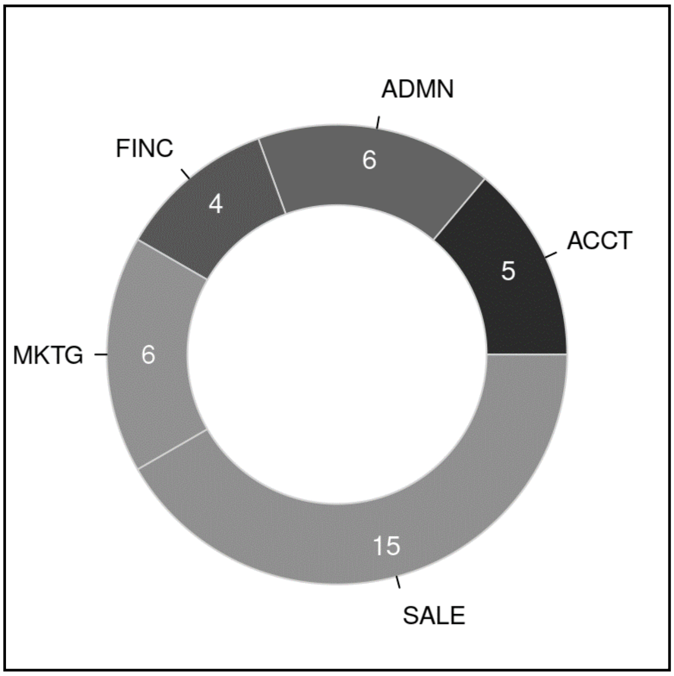


Figure 3.24: Donut chart with 5 segments

3.11.1 Pie Chart Visualization

In figure 3.25 we have one of the most well-known charts called pie. The pie chart expresses the number of deaths caused by selected communicable diseases annually worldwide as of 2019 [72]. There are 6 different diseases which mean our pie chart needs to have 6 slices and the sum of the slices must be 100. While working on pie charts each slice should have a value in percentage. At first sight either a professional or a novice can observe which sort of disease causes more deaths in the world. There is no use to have any x and y axes in a pie chart or even gridlines. Our concentration is the number of slices and their sum, one of the most common issues that might occur is the sum of slices, the problem can be easily solved by accurate calculations. The chart demonstrates that the top three contagious diseases are Tuberculosis, HIV/AIDS and Malaria. The slices belonged to Measles, Cholera and Hepatitis have much less proportion in the pie chart.

Visualization a huge amount of data by a pie chart will definitely result to an illegible and uninformative chart due to excessive number of slices. What we have visualized in the pie chart is a categorical variable. As it can be seen the biggest slice which represents Tuberculosis disease is split from the other slices, the reason is to quickly deliver the information to the reader. Corresponding colors give a decent appearance to the chart, the edge color around each slice is optional to have and can be removed or modified by the Python code. By hovering over each slice of the pie chart the user can see the name of the disease, the number of values and the percentage of the number of values. Another feature that the chart has is when we click on each square next to the disease we can remove its corresponding slice from the pie chart temporarily, but we need to be cautious because the values on each slice will be changed.

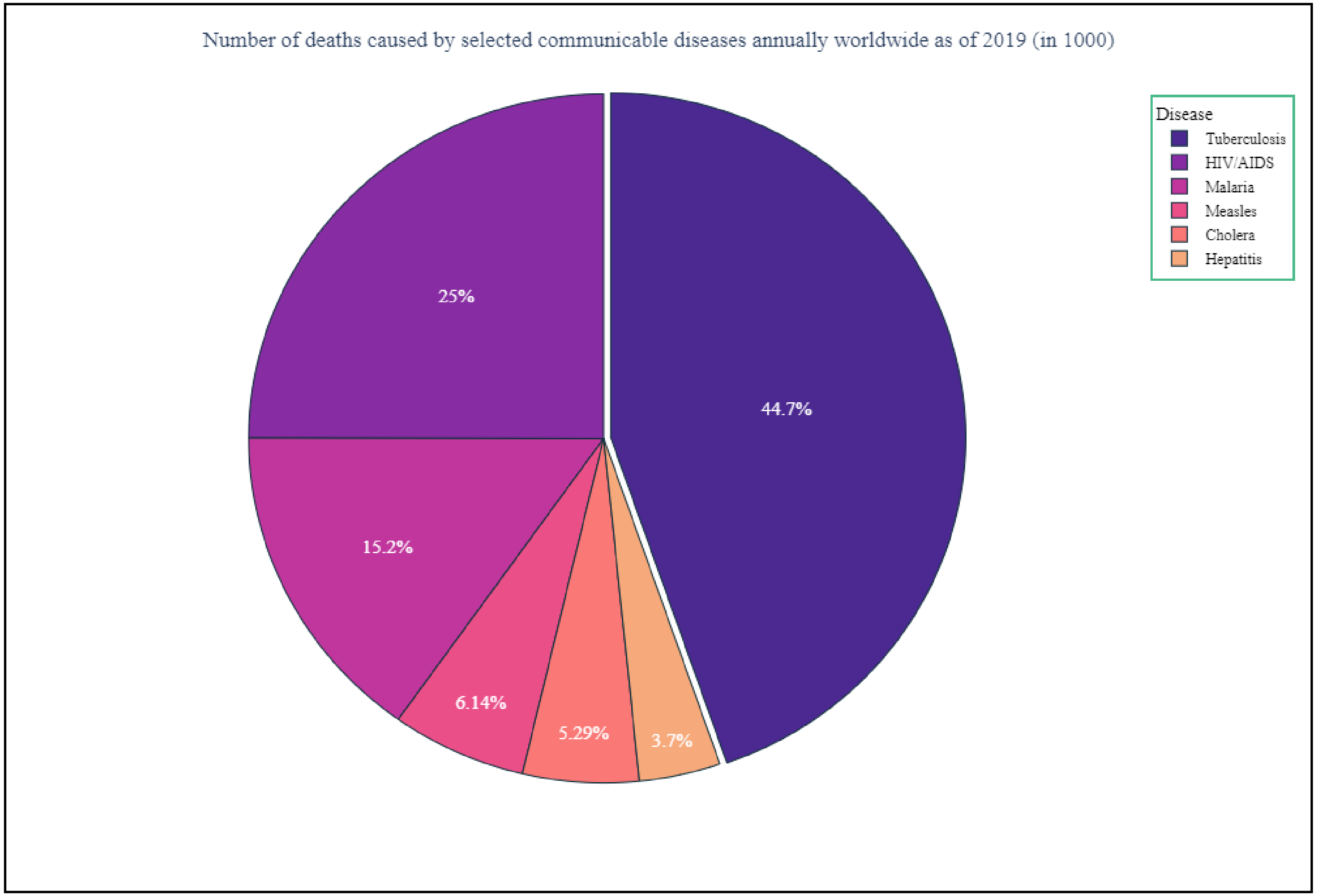


Figure 3.25: Pie chart with 6 slices

3.12 Radar Charts

A radar graph (figure 3.26) [25], also known as a web chart, a spider chart [24] or a star plot, is a two-dimensional depiction that presents data sets with three or more variables. The quantitative value of each variable is mirrored over an axis which normally begins in the middle of the chart. A line connects the points on each axis when each item’s variables are plotted, generating an irregular polygon which does not mirror a star or spider webs in some cases. On a single radar chart, we can compare many data sets by representing each of them with a various color, designated by labels, or go along with key [19]. Similar to other charts that have been discussed so far radar charts also have some flaws which are mentioned below [22]:

- They are not able to compare more than two or three plots at the same time.
- It can be difficult to distinguish which points correspond to which without coloring.
- Too many axes make it harder to read and less straightforward than other types of graphs.

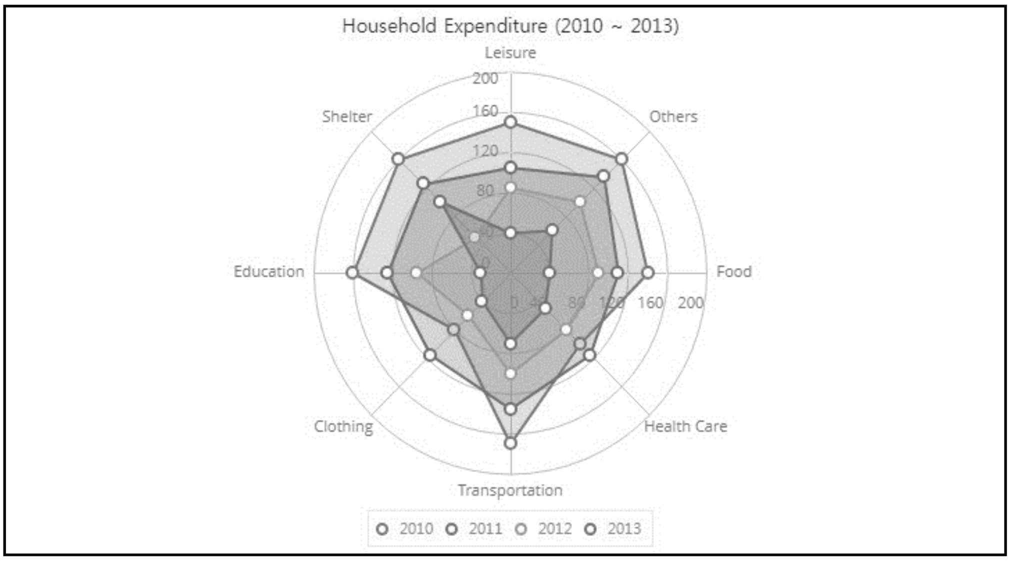


Figure 3.26: Radar chart with 4 polygons

3.12.1 Radar Chart with Two Polygons

The chart demonstrated in figure 3.27 is known as Radar chart, web chart, spider chart and polar chart. In this thesis we refer to it a radar chart. The chart depicts the most valuable commercial services providers in years 2020 and 2021 [73]. The primary application of a radar chart is to compare the correlation between two or more variables. In the chart we do not utilize vertical or horizontal axes, instead we have worked with radial axes. The radar chart has two orange and red webs for years 2020 and 2021, respectively. The ticks outside the circle indicates the names of the companies and the data points scattered on the polar coordinate carry the values of the companies in billion dollars. The horizontal line drawn on the right side of the coordinate has some numeric values to assist the user navigate through the radar chart with more precision. One of advantageous characteristics of radar charts is that they are highly customizable, for instance we can alter the directions of the webs by setting the value of "direction" attribute in the python code to either clockwise or counterclockwise. One noticeable item is the webs themselves, we might fill them with colors or just keep the borderlines around them. The radius of the chart can also be modified, we can change the color, the width and the number of ticks appeared on it. The thickness of the outer borderline of our main chart can be altered according to our preferences. By having a closer look the reader can realize that most data points have values greater than 18. The width and the color of radial gridlines can be changed based on our choices.

In some circumstances data points of each web can overlap with other data points of the other webs, one simple solution can fix the problem. We can reduce or enlarge the sizes of data points, add borderlines with different colors and widths. A unique feature concerning the radar charts is that we can choose how many sectors of our circle we require in degrees, the values can be selected in the form of a closed interval from 0 up to 360. By clicking on a year mentioned in the legend, the corresponding web will be disappeared temporarily. In order to observe data values of each point of the related web, hovering over the point suffices.

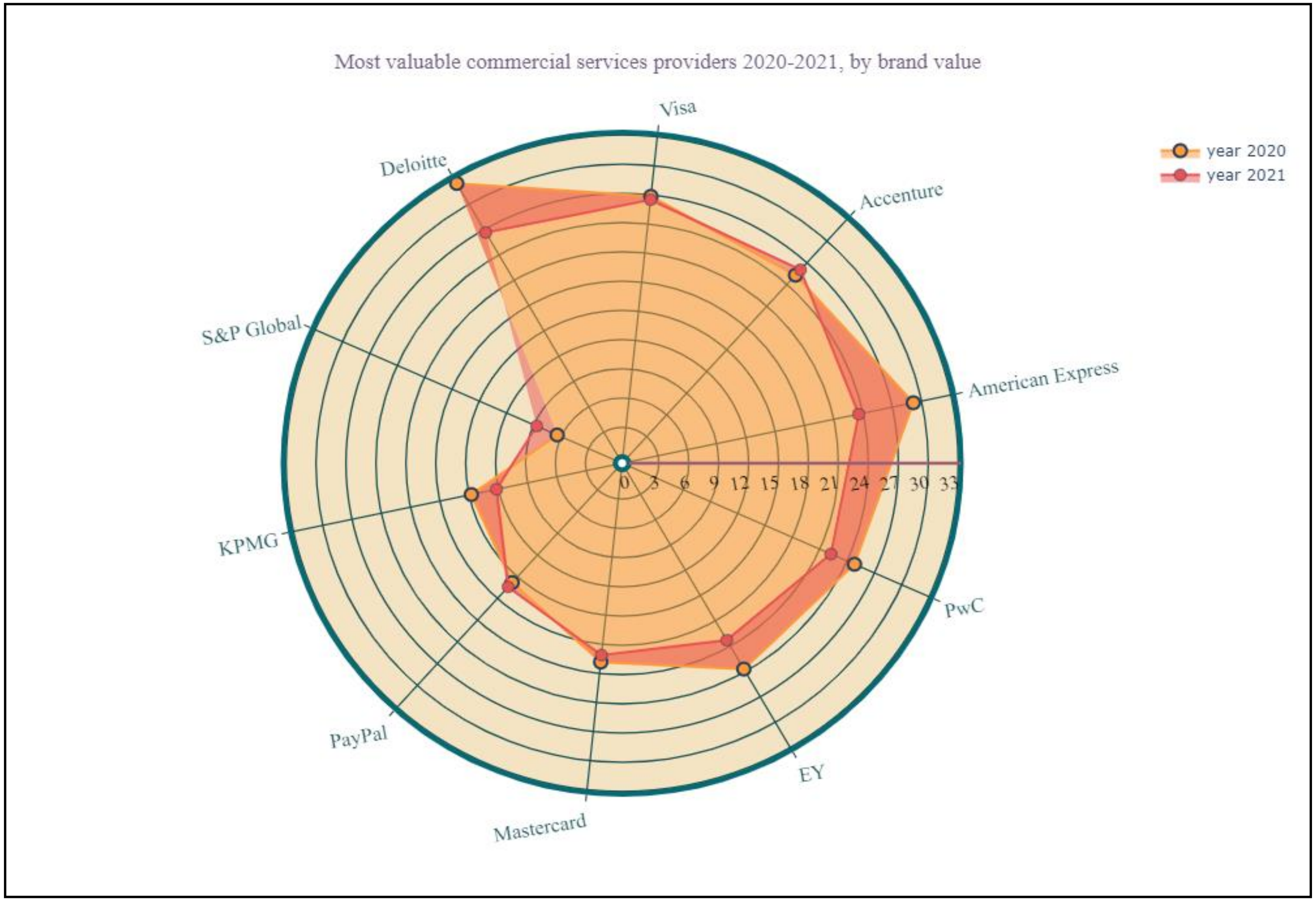


Figure 3.27: Circular radar chart with 2 polygons

3.13 Violin Plots

In contrast to the box-and-whisker storyline, select particular distribution points or the histogram in which values are arranged together into different intervals, the violin chart (figure 3.28) [27] depicts the overall shape of the distribution [1].

Each "violin" symbolizes a different group or variable. The shape shows the variable's density estimate: the more data points in a certain range, the bigger the violin for that range. It is quite similar to a boxplot, but it provides for a more in-depth knowledge of the distribution [26].

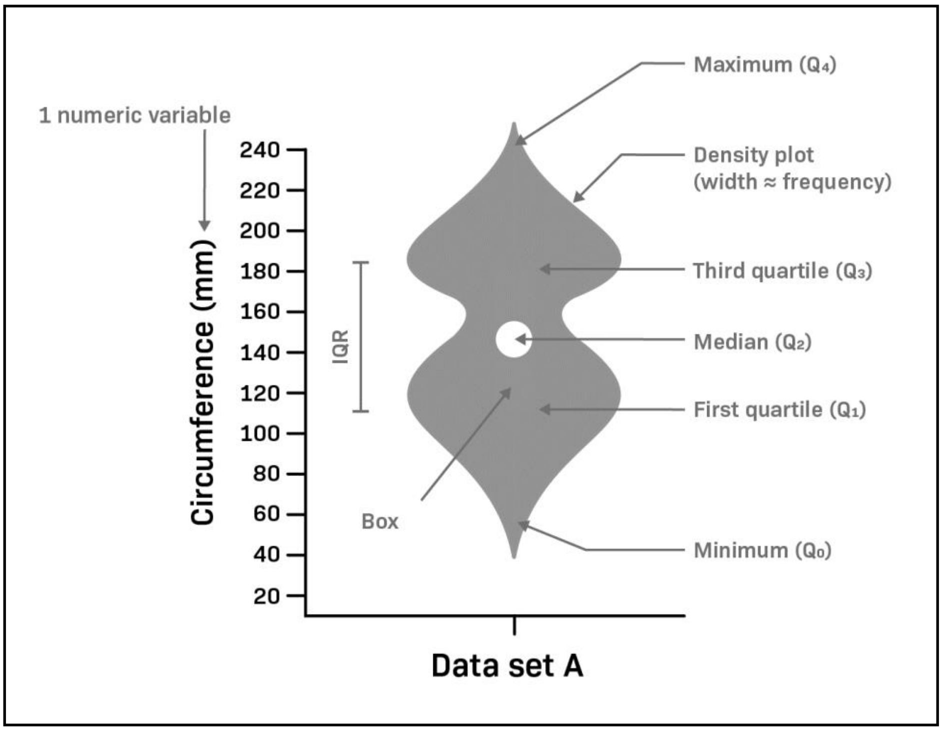


Figure 3.28: Violin plot

3.13.1 Violin Plot Visualization with Numeric Variables

In figure 3.29 we are going to analyze a chart called violin because of its shape. A violin chart is moderately similar to a box-and-whisker plot, but there are some major and trivial differences between them. In a violin chart we do not have neither whiskers nor lines indicating the maximum and the minimum values. Another dissimilarity is that a violin chart comprises kernel density estimation or KDE which is a smoothed out continuous version of a box plot [28]. Our violin chart illustrates the beer and wine consumption according to the locations of countries in the hemispheres [69]. We have a categorical variable which refers to the hemispheres and a numeric variable which expresses the consumption per capita. As it is observed we have 6 violin-shaped plots in total and two different colors, yellow and red to demonstrate the consumption of beer and wine accordingly. There three ticks on the x axis representing the locations of countries in the hemisphere, north, south and both. In each pair hemisphere we have violin charts to make the reader able to have a decent comparison between beer and wine consumption probability wise. The violins in our example are getting wider and narrower at some levels which is totally reasonable in a violin chart. When a violin gets wider, it refers to high probability and be getting narrower we will have lower probability. The points which are scattered next to each violin are the values from our dataset in order to give the reader a rough approximation about the distribution of people who consume either beer or wine in the corresponding hemisphere. At first glance the user can view that the yellow violins have boxes inside which assist the user to compare a box-and-whisker plot with a violin chart.

The existence of boxes within yellow violins are optionally, but our aim was to create a comprehensive visualization. The green dashed line drawn in each violin is the mean for the related violin. The number of ticks on the y axis can provide more precision for the readers especially when we have horizontal gridlines which makes the estimation much simpler. Almost all the features that we have mentioned previously can be utilized in the violin chart. For instance, by bringing the mouse cursor over any violin we will be able to see the maximum and the maximum values, quartiles, mean and median values, respectively. The colors of the background and the violin can modified based on our interest. The last point to be mentioned is that when the user clicks on the drink type in the legend, the corresponding violins will be omitted temporarily in order to concentrate on the other violins.

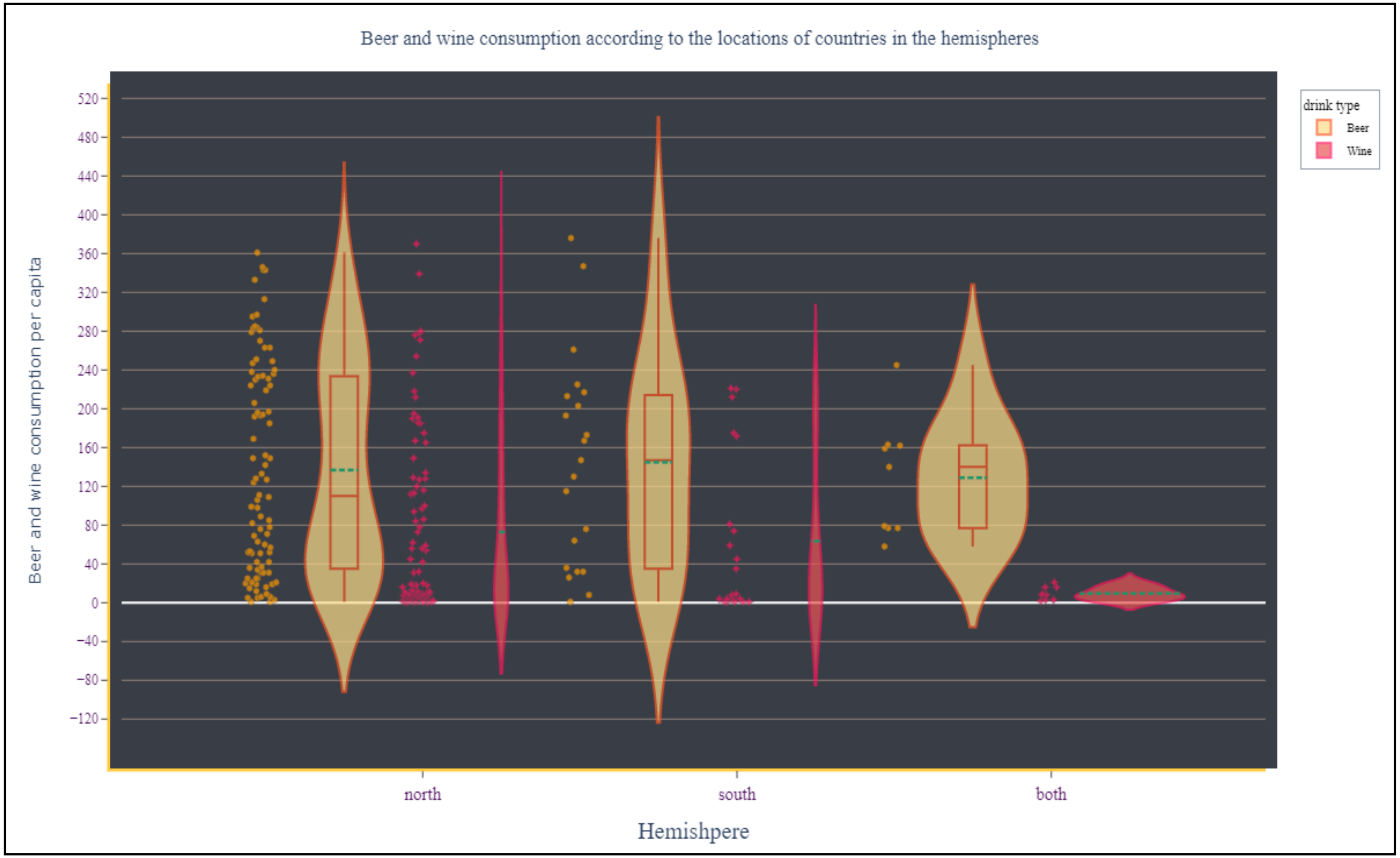


Figure 3.29: Violin plot with 6 violins

Chapter 4

THE FUTURE OF DATA VISUALIZATION

The realm of data visualization is always transforming. Previously, data visualization and advanced business intelligence were so difficult that they could only be utilized by data scientists and analytics professionals. Any company stakeholder can now effortlessly gather and analyze vast amounts of data, and utilize data visualization to acquire instant insights from complicated trends and patterns [74].

Data visualization will become increasingly strong and available to non-technical audiences in the next years. Data visualization is embracing new technology to make itself more accessible, engaging, and informative to any sort of audience, such as fascinating infographics or interactive images. Here's a deeper look at how we'll interact with data visualizations in the future [74].

REFERENCES

[1] Schwabish, J. (2021). Better Data Visualizations - A guide for scholars, researchers, and wonks. Columbia University Press.

[2] Oetting, J. (2022, June 3). 14 best types of charts and graphs for data visualization [+ guide]. HubSpot Blog. Retrieved from https://blog.hubspot.com/marketing/types-of-graphs-for-data-visualization

[3] The spaghetti plot - data to viz. (n.d.). Retrieved from https://www.data-to-viz.com/caveat/spaghetti.html

[4] Healy, Y. H. and C. (n.d.). Scatter plot. Scatter plot – from Data to Viz. Retrieved from https://www.data-to-viz.com/graph/scatter.html

[5] Healy, Y. H. and C. (n.d.). Barplot. Barplot – from Data to Viz. Retrieved from https://www.data-to-viz.com/graph/barplot.html

[6] Holtz, Y. (n.d.). Lollipop chart. The Python Graph Gallery. Retrieved from https://www.python-graph-gallery.com/lollipop-plot/

[7] Holtz, Y. (n.d.). Most basic circular barplot with python and matplotlib. The Python Graph Gallery. Retrieved from https://www.python-graph-gallery.com/circular-barplot-basic

[8] Essential chart types for data visualization. Chartio. (n.d.). Retrieved from https://chartio.com/learn/charts/essential-chart-types-for-data-visualization/

[9] How to choose the Right Chart for Data Visualization. Analytics Vidhya. (2021, September 1). Retrieved from https://www.analyticsvidhya.com/blog/2021/09/how-to-choose-the-right-chart-for-data-visualization/

[10] Box and whisker plot. Box and Whisker Plots - Learn about this chart and its tools. (n.d.). Retrieved from https://datavizcatalogue.com/methods/box_plot.html

[11] Healy, Y. H. and C. (n.d.). Bubble plot. Bubble plot – from Data to Viz. Retrieved from https://www.data-to-viz.com/graph/bubble.html

[12] Metwalli, S. A. (2021, July 15). Data Visualization 101: How to choose a chart type. Medium. Retrieved from https://towardsdatascience.com/data-visualization-101-how-to-choose-a-chart-type-9b8830e558d6

[13] Bubble. Bubble charts in Python. (n.d.). Retrieved from https://plotly.com/python/bubble-charts/#bubble-chart-with-plotlygraphobjects

[14] Contour plot. Data Viz Project. (2017, August 28). Retrieved from https://datavizproject.com/data-type/contour-plot/

[15] Stephanie. (2021, May 7). Contour plots: Definition, examples. Statistics How To. Retrieved from https://www.statisticshowto.com/contour-plots/

[16] What is a funnel chart? TIBCO Software. (n.d.). Retrieved from https://www.tibco.com/reference-center/what-is-a-funnel-chart#:~:text=A%20funnel%20chart%20is%20a,recruitment%2C%20and%20order%20fulfilment%20processes

[17] Funnel charts with python. Funnel Charts with Python. (n.d.). Retrieved from https://morioh.com/p/2b6ed205faa9

[18] Create heatmap in R (3 examples): Base R, ggplot2 & plotly package. Statistics Globe. (2022, March 17). Retrieved from https://statisticsglobe.com/heatmap-in-r

[19] 19 innovative ways to use information visualization across a variety of fields. Visme Blog. (2019, September 8). Retrieved from https://visme.co/blog/information-visualization/

[20] Histogram. Data Viz Project. (2017, September 29). Retrieved from https://datavizproject.com/data-type/histogram/

[21] Data Visualization – how to pick the Right Chart Type? eazyBI. (n.d.). Retrieved from https://eazybi.com/blog/data-visualization-and-chart-types

[22] Advantages and disadvantages of different types of graphs. Hebrew website. (n.d.). Retrieved from http://www.kmrom.com/Site-En/Articles/ViewArticle.aspx?ArticleID=416

[23] Donut chart in R with lessr. R CHARTS | A collection of charts and graphs made with the R programming language. (2021, December 5). Retrieved from https://r-charts.com/part-whole/donut-chart/

[24] Wikimedia Foundation. (2022, July 1). Radar Chart. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Radar_chart

[25] Koolchart User's Manual: Radar Chart. (n.d.). Retrieved from https://www.koolchart.com/demo/Docs/radar-chart.html

[26] Healy, Y. H. and C. (n.d.). Violin plot. Violin plot – from Data to Viz. Retrieved from https://www.data-to-viz.com/graph/violin.html

[27] LabXchange. (n.d.). Retrieved from https://www.labxchange.org/library/items/lb:LabXchange:46f64d7a:html:1

[28] Violin plots 101: Visualizing Distribution and probability density. Mode. (n.d.). Retrieved from https://mode.com/blog/violin-plot-examples/#how-to-read-violin-plot

[29] What is information visualization? TIBCO Software. (n.d.). Retrieved https://www.tibco.com/reference-center/what-is-information-visualization

[30] Brush, K., & Burns, E. (2020, February 20). What is data visualization and why is it important? SearchBusinessAnalytics. Retrieved from https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization

[31] Falkowitz, R. S. (2022, April 25). Information visualization or data visualization? This view of service management... Retrieved from https://www.3cs.ch/information-visualization-data-visualization/#:~:text=Information%20visualization%20and%20data%20visualization%20are%20related,creation%20of%20an%20information%20visualization

[32] What is information visualization? definition and faqs. What is Information Visualization? Definition and FAQs | HEAVY.AI. (n.d.). Retrieved from https://www.heavy.ai/technical-glossary/information-visualization

[33] Contributor, T. T. (2005, September 21). What is Digital Library? - definition from whatis.com. WhatIs.com. Retrieved from https://www.techtarget.com/whatis/definition/digital-library

[34] Team, T. B. I. G. L. (2022, February 15). Health care vs healthcare - what's the difference? BIG Language Website. Retrieved from https://biglanguage.com/blog/health-care-vs-healthcare/

[35] Twin, A. (2021, September 16). What market research tells companies about new products and services. Investopedia. Retrieved from https://www.investopedia.com/terms/m/market-research.asp#:~:text=Market%20research%20is%20the%20process,in%20the%20product%20or%20service

[36] Encyclopædia Britannica, inc. (n.d.). Manufacturing. Encyclopædia Britannica. Retrieved from https://www.britannica.com/technology/manufacturing

[37] A comprehensive guide on Data Visualization in python. Analytics Vidhya. (2021, October 14). Retrieved from https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-data-visualization-in-python/

[38] A brief history of data visualization. Premium Reporting & Data Analytics - Dundas Data Visualization. (n.d.). Retrieved from https://www.dundas.com/resources/blogs/introduction-to-business-intelligence/brief-history-data-visualization

[39] Sridharan, M., About the Author: Mithun Sridharan I am Mithun Sridharan, & Sridharan, I. am M. (2022, July 6). Data Visualization - history and origins. Think Insights. Retrieved from https://thinkinsights.net/data-literacy/data-visualization-history/

[40] Data visualization : What is this and why it matters. Analytics Vidhya. (2021, August 1). Retrieved from https://www.analyticsvidhya.com/blog/2021/08/data-visualization-what-is-this-and-why-it-matters/

[41] Peterman, M. (n.d.). 15 statistics that prove the power of Data Visualization. csg solutions. Retrieved from https://blog.csgsolutions.com/15-statistics-prove-power-data-visualization

[42] How we really make decisions. Association for Psychological Science - APS. (2013, November 8). Retrieved from

https://www.psychologicalscience.org/faces-minds/how-we-really-make-decisions.html

[43] Wikimedia Foundation. (2021, December 30). IBM Cognos Analytics. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/IBM_Cognos_Analytics

[44] Qlik sense: Modern cloud analytics. Qlik. (n.d.). Retrieved from https://www.qlik.com/us/products/qlik-sense

[45] Mihart. (n.d.). What is Power Bi? - power bi. Power BI | Microsoft Docs. Retrieved from https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview

[46] Fusion middleware user's guide for oracle data visualization. Moved. (2016, October 11). Retrieved from https://docs.oracle.com/middleware/12212/biee/BIVUG/GUID-0FBED0C2-2807-462A-8CD6-F01B3F068E5A.htm#BIVUG-GUID-0FBED0C2-2807-462A-8CD6-F01B3F068E5A

[47] Oracle. Exploring Data Using Oracle BI Cloud Service Visual Analyzer. (n.d.). Retrieved from https://www.oracle.com/webfolder/technetwork/tutorials/obe/cloud/bics/va/va.html#overview

[48] SAP Lumira - Overview. Tutorials Point. (n.d.). Retrieved from https://www.tutorialspoint.com/sap_lumira/sap_lumira_overview.htm

[49] Tibco Spotfire®. TIBCO Software Inc. (n.d.). Retrieved from https://www.tibco.com/products/tibco-spotfire

[50] Data Analytics & Business Intelligence Platform - Zoho Analytics. Zoho. (n.d.). Retrieved from https://www.zoho.com/analytics/

[51] Wikimedia Foundation. (2022, April 1). D3.js. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/D3.js

[52] Wikimedia Foundation. (2022, June 27). Project jupyter. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Project_Jupyter

[53] Wikimedia Foundation. (2022, June 15). Microstrategy. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/MicroStrategy

[54] Wikimedia Foundation. (2022, April 8). Google charts. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Google_Charts

[55] How to choose the Right Chart for Your Data. Infogram. (n.d.). Retrieved from https://infogram.com/page/choose-the-right-chart-data-visualization

[56] Nediger, M. (2021, November 26). How to choose the best types of charts for your data. Venngage. Retrieved from https://venngage.com/blog/how-to-choose-the-best-charts-for-your-infographic/

[57] 13 types of graphs and charts (plus when to use them). Indeed Career Guide. (n.d.). Retrieved from https://www.indeed.com/career-advice/career-development/types-of-graphs-and-charts

[58] 44 types of graphs & charts [& how to choose the best one]. Visme Blog. (2020, October 15). Retrieved from https://visme.co/blog/types-of-graphs/

[59] What is python? executive summary. Python.org. (n.d.). Retrieved July 14, 2022, from https://www.python.org/doc/essays/blurb/

[60] Getting started with plotly-python. GeeksforGeeks. (2021, July 15). Retrieved from https://www.geeksforgeeks.org/getting-started-with-plotly-python/

[61] Plotly. Plotly Python Graphing Library. (n.d.). Retrieved July 14, 2022, from https://plotly.com/python/

[62] Bharath, K. (2020, December 22). Everything you need to know about Jupyter notebooks! Medium. Retrieved from https://towardsdatascience.com/everything-you-need-to-know-about-jupyter-notebooks-10770719952b

[63] Krishnan, A. (2022, March 30). Average rainfall and temperature in UK (2010-2019). Kaggle. Retrieved from https://www.kaggle.com/datasets/aryakrishnanar/average-rainfall-and-temperature-in-uk20102019

[64] Best, R. de. (2022, July 14). Bitcoin price history 2013-2022. Statista. Retrieved from https://www.statista.com/statistics/326707/bitcoin-price-index/

[65] Joshi, A. (2021, March 28). Faang Historical Stock Price Data. Kaggle. Retrieved from https://www.kaggle.com/datasets/specter7/amazon-amzn-historical-stock-price-data?select=Facebook_Historical_StockPrice2.csv

[66] Ersever, Y. (2018, July 3). 500 person gender-height-weight-body mass index. Kaggle. Retrieved from https://www.kaggle.com/datasets/yersever/500-person-gender-height-weight-bodymassindex

[67] Sönnichsen, N. (2022, July 11). Oil prices 1960-2021. Statista. Retrieved from https://www.statista.com/statistics/262858/change-in-opec-crude-oil-prices-since-1960/

[68] Lock, S. (2021, September 14). Fast Food Industry Market Share Worldwide, by brand 2019. Statista. Retrieved from https://www.statista.com/statistics/273057/value-of-the-most-valuable-fast-food-brands-worldwide/

[69] Pessotto, M. (2018, October 28). Happiness and alcohol consumption. Kaggle. Retrieved from https://www.kaggle.com/datasets/marcospessotto/happiness-and-alcohol-consumption

[70] Anandhu, A. (2022, March 20). Countries in the world by population 2022. Kaggle. Retrieved from https://www.kaggle.com/datasets/anandhuh/countries-in-the-world-by-population-2022

[71] Learning, U. C. I. M. (2017, November 27). Red Wine Quality. Kaggle. Retrieved from https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

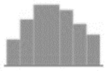
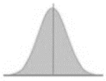
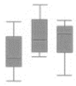
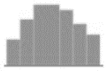
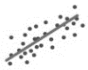
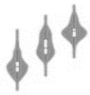
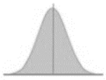
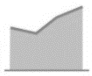
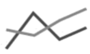
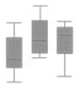
[72] Elflein, J. (2022, May 20). Deaths communicable diseases annually worldwide 2019. Statista. Retrieved from https://www.statista.com/statistics/282715/deaths-from-communicable-diseases-worldwide/

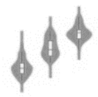
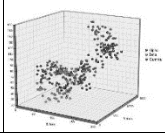
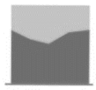
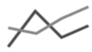
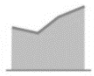
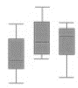
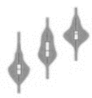
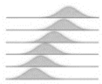
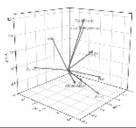
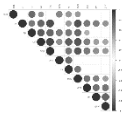
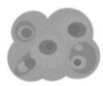
[73] Statista Research Department. (2022, July 6). Brand value of commercial services providers worldwide 2021. Statista. Retrieved from https://www.statista.com/statistics/809291/brand-value-of-commercial-services-providers/

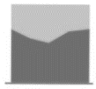
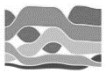
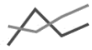
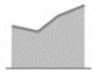
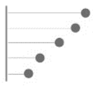
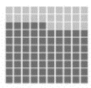
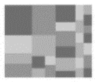
[74] Sunarto, N., Tomboc, K., & Sahni, H. (2022, February 2). 55 Facts & Statistics showing why data visualization is important. Piktochart. Retrieved July 17, 2022, from https://piktochart.com/blog/data-visualization-statistics

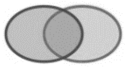
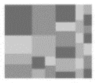
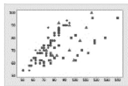
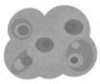
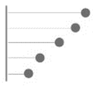
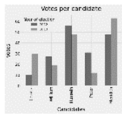
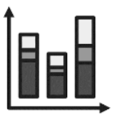
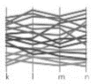
APPENDIX

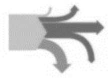
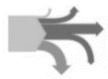
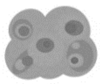
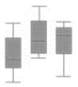
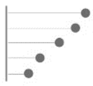
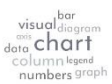
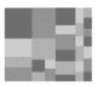
Charts’ Usages

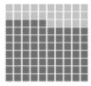
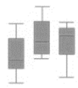
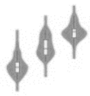
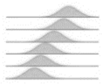
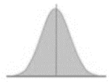
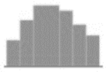
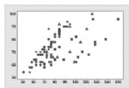
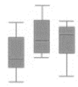
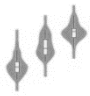
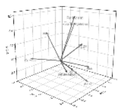
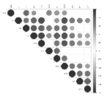
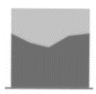
<table>
<tr><th>Type of the variable</th><th colspan="2">Number of variables</th><th colspan="2">Charts and Graphs</th><th>Presentation type</th></tr>
<tr><td rowspan="10">Numeric</td><td colspan="2" rowspan="2">one numeric variable</td><td>Histogram</td><td></td><td>Distribution</td></tr>
<tr><td>Density plot</td><td></td><td>Distribution</td></tr>
<tr><td rowspan="8">Two variables</td><td rowspan="6">Not ordered</td><td>Box plot</td><td></td><td>Distribution
Comparison</td></tr>
<tr><td>Histogram</td><td></td><td>Distribution</td></tr>
<tr><td>Scatter plot</td><td></td><td>Correlation
Distribution</td></tr>
<tr><td>Violin plot</td><td></td><td>Distribution
Comparison</td></tr>
<tr><td>Density plot</td><td></td><td>Distribution</td></tr>
<tr><td>2D density plot</td><td></td><td>Correlation</td></tr>
<tr><td rowspan="2">Ordered</td><td>Connected scatter plot</td><td></td><td>Correlation</td></tr>
<tr><td>Area plot</td><td></td><td>Evolution
Distribution</td></tr>
<tr><td rowspan="2">Numeric</td><td>Two numeric variables</td><td>Ordered</td><td>Line plot</td><td></td><td>Evolution
Distribution</td></tr>
<tr><td>Three variables</td><td>Not ordered</td><td>Box plot</td><td></td><td>Distribution
Comparison</td></tr>
</table>

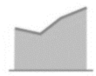
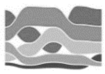
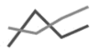
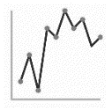
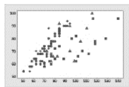
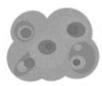
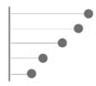
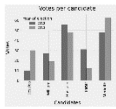
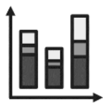
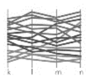
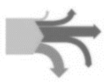
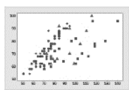
			Violin plot		Distribution Comparison
			Bubble plot		Comparison Distribution Evolution
			3D scatter		Correlation
		Ordered	Stacked area plot		Evolution Comparison
			Stream graph		Evolution Distribution Comparison
			Line plot		Evolution Distribution
			Area plot		Evolution Distribution
	Several variables (more than three)	Not ordered	Box plot		Distribution Comparison
			Violin plot		Distribution Comparison
Numeric	Several variables (more than three)	Not Ordered	Ridge plot		Distribution
			PCA		Correlation
			Correlogram		Correlation
			Heatmap		Correlation Comparison Distribution Evolution

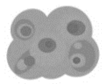
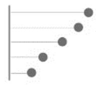
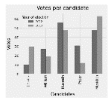
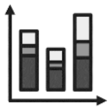
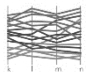
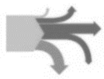
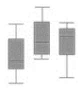
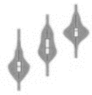
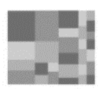
			Dendrogram		Correlation
		Ordered	Stacked area plot		Evolution Comparison
			Stream graph		Evolution Distribution Comparison
			Line plot		Evolution Distribution
			Area plot		Evolution Distribution
Categorical	One variable		Bar plot		Ranking Comparison
			Lollipop chart		Ranking Comparison Evolution
Categorical	One variable		Waffle chart		Part of a whole Comparison
			Word cloud		Ranking Comparison Distribution
			Doughnut chart		Part of a whole Comparison
			Pie chart		Part of a whole Comparison
			Treemap		Part of a whole Comparison Distribution
			Circular Packing		Part of a whole Comparison Distribution

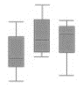
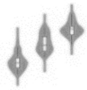
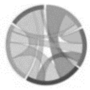
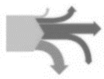
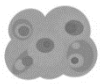
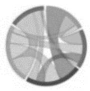
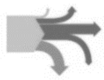
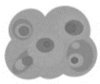
	Two or more variables	Two independent lists	Venn diagram		Part of a whole Correlation
		Nested	Treemap		Part of a whole Comparison Distribution
			Circular packing		Part of a whole Comparison Distribution
			Sunburst diagram		Comparison
			Bar plot		Ranking Comparison
Categorical	Two or more variables	Nested	Dendrogram		Correlation
		Subgroup	Grouped scatter		Correlation
			Heatmap		Correlation Comparison Distribution Evolution
			Lollipop chart		Ranking Comparison Evolution
			Grouped bar plot		Ranking Comparison
			Stacked bar plot		Part of a whole Comparison
			Parallel plot		Comparison
			Spider plot		Comparison Distribution

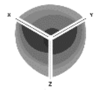
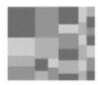
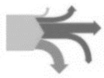
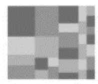
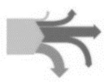
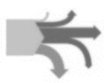
			Sankey diagram		Flow Comparison
		Adjacency	Network		Flow Correlation
			Chord chart		Flow Correlation Comparison
			Arc diagram		Flow Comparison Correlation
Categorical	Two or more variables	Adjacency	Sankey diagram		Flow Comparison
			Heatmap		Correlation Comparison Distribution Evolution
One numeric and one categorical	One observation per group		Box plot		Distribution Comparison
			Lollipop chart		Ranking Comparison Evolution
			Doughnut chart		Part of a whole Comparison
			Pie chart		Part of a whole Comparison
			Word cloud		Ranking Comparison Distribution
			Treemap		Part of a whole Comparison Distribution
			Circular packing		Part of a whole Comparison Distribution

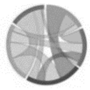
		Waffle chart		Part of a whole Comparison
	Several observati ons per group	Box plot		Distribution Comparison
One numeric and one categorical	Several observations per group	Violin plot		Distribution Comparison
		Ridge line		Correlation
		Density plot		Distribution
		Histogram		Distribution
One categorical and several numeric	No order	Grouped scatter plot		Correlation
		2D density plot		Correlation
		Box plot		Distribution Comparison
		Violin plot		Distribution Comparison
		PCA		Correlation
		Correlogram		Correlation
	A numeric variable is ordered	Stacked area plot		Evolution Comparison

			Area chart		Evolution Distribution
One categorical and several numeric	A numeric is ordered		Stream graph		Evolution Distribution Comparison
			Line plot		Evolution Distribution
			Connected scatter		Correlation
	One value per group		Grouped scatter plot		Correlation
			Heatmap		Correlation Comparison Distribution Evolution
			Lollipop chart		Ranking Comparison Evolution
			Grouped bar plot		Ranking Comparison
			Stacked bar plot		Part of a whole Comparison
			Parallel plot		Comparison
			Spider plot		Comparison Distribution
			Sankey diagram		Flow Comparison
Several categorical and one numeric	Subgroup	One observation per	Grouped scatter		Correlation

Several categorical and one numeric	Subgroup	One observation per group	Heatmap		Correlation Comparison Distribution Evolution
			Lollipop chart		Ranking Comparison Evolution
			Grouped bar plot		Ranking Comparison
			Stacked bar plot		Part of a whole Comparison
			Parallel plot		Comparison
			Spider plot		Comparison Distribution
			Sankey diagram		Flow Comparison
		Several observations per group	Box plot		Distribution Comparison
			Violin		Distribution Comparison
	Nested	One observation per group	Bar plot		Ranking Comparison
			Dendrogram		Correlation
			Sunburst diagram		Comparison
Several categorical and one numeric	Nested	One observation per	Treemap		Part of a whole Comparison Distribution

			Circular packing		Part of a whole Comparison Distribution
		Several observations per group	Box plot		Distribution Comparison
			Violin		Distribution Comparison
	Adjacency		Network		Flow Correlation
			Chord diagram		Flow Correlation Comparison
			Arc diagram		Flow Comparison Correlation
			Sankey diagram		Flow Comparison
			Heatmap		Correlation Comparison Distribution Evolution
Network	Network		Network diagram		Flow Correlation
			Chord diagram		Flow Correlation Comparison
			Arc diagram		Flow Comparison Correlation
Network	Network		Sankey diagram		Flow Comparison
			Heatmap		Correlation Comparison Distribution Evolution

			Hive plot		Comparison Correlation Distribution
	Nested or hierarchical	No value	Dendrogram		Correlation
			Tree map		Part of a whole Comparison Distribution
			Circular packing		Part of a whole Comparison Distribution
			Sunburst diagram		Comparison
			Sankey diagram		Flow Comparison
	Nested or hierarchical	Value for leaf	Dendogram		Correlation
			Tree map		Part of a whole Comparison Distribution
			Circular packing		Part of a whole Comparison Distribution
			Sunburst diagram		Comparison
Network	Nested or hierarchical	Value for leaf	Sankey diagram		Flow Comparison
		Value for edges	Dendrogram		Correlation
			Sankey diagram		Flow Comparison

			Chord diagram		Flow Correlation Comparison