COMP1800

Data Visualization Coursework

Name: Nazri

Student ID: 001278084

Course: MSc Big Data and Business Intelligence

Submission Date:06/04/2023

School of Computing and Mathematical Science

Table of Contents

Introduction to Data Visualisation 3
Visual Data Exploration on Chrisco 4
Data Exploration on Daily Customer Data Frame 6
Bar Chart- Total Customers Visited in each Outlet 6
Justification 6
Description 6
Pie Chart- Circular Representation for Total Distribution of Daily Customers 8
Justification 8
Description 9
Box Plot- Distribution of Outlets with High Volume of Daily Customers 10
Justification 10
Description 11
Monthly Trend on Outlets with High Volume Customer Data 12
Justification 12
Description 12
Interactive Line Plot - Distribution of Outlets with Extremely Lower Volume of Daily Customers 14
Justification 14
Description 14
Data Exploration on Summary Data Frame 16
Pair Plot-To find the Correlation between all Attributes of Summary Data Frame 16
Justification 16
Description 17
Heat Map- To find the Correlation between each Attribute in the Summary Data Frame 18
Justification 18
Description 18
Interactive Scatter Plot- To Study the Positively Correlated attributes in the Summary Data Frame 20
Justification 20
Description 20
Critical Review 22
Conclusion 23
References 24

Introduction to Data Visualisation

Data visualisation refers to the use of graphical or visual representations to communicate information or data to an audience in a way that is easy to understand and interpret. The process involves transforming complex sets of data into clear and concise visual forms such as charts, graphs, maps, or other visual aids that allow viewers to easily grasp patterns, trends, and correlations in data, that might not be apparent when looking at raw data alone.

The primary goal of data visualization is to make complex data understandable to people who are not data experts. The use of data visualization has become increasingly important in today's data-driven world, due to the increasing amount of data available, and the need to communicate insights effectively. Data visualization is commonly used in fields such as business, science, journalism, healthcare, education, and government to help decision-makers, researchers, and others to better understand and communicate complex information. It can also be used to highlight key insights and findings, and to present data in a compelling and engaging way.

Overall, data visualization is a powerful tool that allows people to better understand and analyze data, and to make informed decisions based on this information.

Visual Data Exploration on Chrisco

As part of the coursework, we need to conduct a visual data exploration for ChrisCo, a fictional but successful retail company managing several outlets across UK. The company collects large amount of data about its customers visiting each outlet through a loyalty card system, which is then aggregated and averaged to provide information on the company's 45 outlets, each identified by a unique three-letter code like ABC, XYZ, etc. We will be using a Python Notebook, either in Colab or Jupyter, to carry out this exploration.

To explore the data, ChrisCo's daily customer data visiting its outlet and the overall performance of each outlet, are provided as the data sources in CSV format. We need to use basic Pandas functions such as head(), info(), and describe() to explore the data to get an initial understanding of its structure, size, and content. Then, visualize the data using visualization libraries such as Seaborn, Matplotlib or other to create visualizations such as scatter plots, bar charts, and line charts to identify patterns, trends, correlations, and insights in the data.

Attributes of the Dataset

We need to explore 5 different CSV datasets that contain the daily customer data visiting each outlet and the overall performance of each outlet, such as outlet marketing, outlet overheads, outlet size, and outlet staff.

- **Daily Customer Data:** This dataset is a time series dataset containing information on the number of customers visiting each outlet each day in the year 2021 (01-01-2021 to 31-12-2021). The data is in integer format and have 365 data entries for each outlet. i.e., 365 rows and 45 columns, indicating a daily time series throughout the year. The dataset does not contain any null or negative values.
- **Outlet Marketing:** This dataset shows the amount spent by each outlet in marketing. Each row represents an outlet identified by a 3-letter word and its associated marketing expense in pounds (£). The data set contains 45 outlets with marketing expenditures varying from £1000 to £67000.
- **Outlet Overheads:** This dataset includes details on different costs. Each row depicts an outlet, with the first column holding the outlet's Id and the second column containing the associated overhead expense. This dataset contains 45 outlets with overhead expenses varying from £11000 to £99000.
- **Outlet Size:** This file includes information on the extent of a company's various outlets. Each row indicates an outlet, and the columns include the outlet's unique identifier (Id) and the outlet's area in square meters. There are 45 entries in the collection. The dataset contains numerical values.

- **Outlet Staff:** This dataset contains data on the number of employees working in each outlet. Each row indicates a distinct outlet, and the first column contains the outlet's identification code. The second column lists the number of employees working in each outlet. The number of employees working in each outlet ranges from 1 to 67.

The data needs to be combined into two data frames: one containing 'Daily Customer Data' and one row for each date, and the other containing 'Summary Data' and one row for each outlet, assembled from all of the CSV files, including the daily customer data.

Data Exploration on Daily Customer Data Frame

Bar Chart- Total Customers Visited in each Outlet

Justification

A bar chart is used to represent data using rectangular bars of equal width. Bar charts allow you to easily compare data across different categories or groups. The length or height of each bar represents the value of the data it represents, making it easy to see which data points are higher or lower than others. Bar charts are simple and easy to understand. They are a great way to present data in a clear and concise manner, making it easy for viewers to quickly grasp the information being presented. Bar charts can be used for large datasets, as long as the categories or groups are manageable. The rectangular bars are visually easy to compare, even when there are many data points being presented. These are the reasons why I used bar charts to visualize the 'Daily Customer Data' Dataframe.

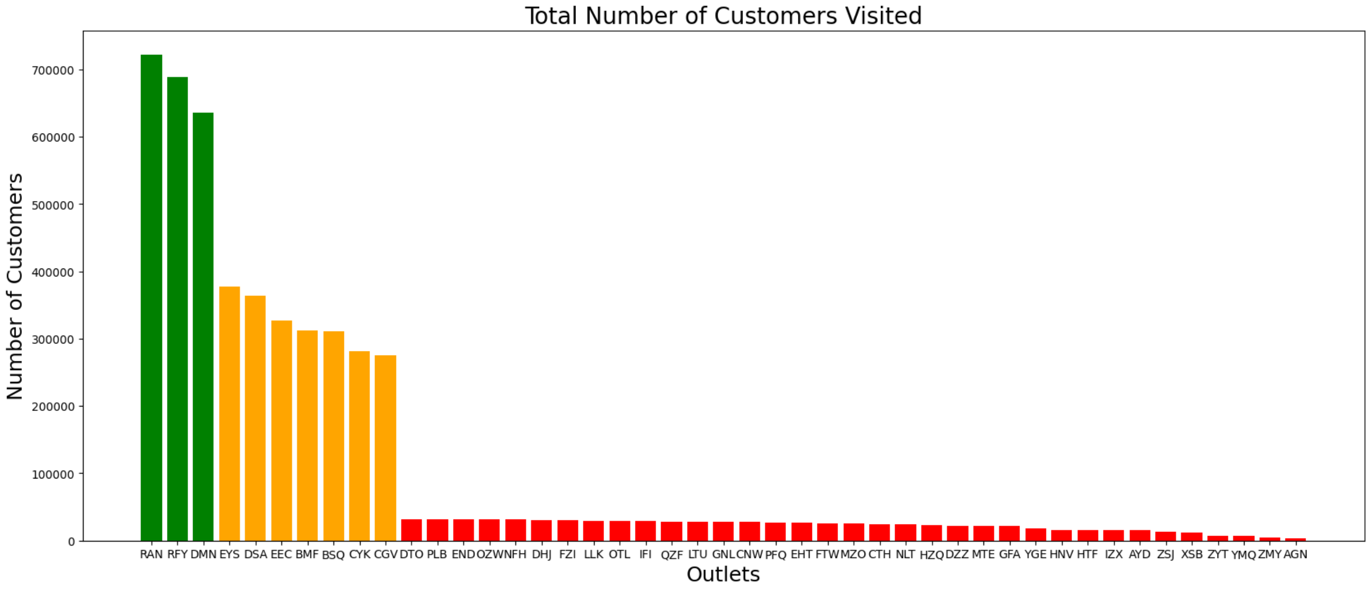

Figure 1 Bar Chart- Total customers visited in each outlet

Description

Figure 1 displays the bar chart which shows the total number of customers visited each outlet in the 'Daily Customer Data' dataframe. The x-axis displays different outlets whereas, the y-axis displays the number of customers. The colour of the bars is determined by the total number of customers visited each outlet. If the total number of customers visited exceeds 400000, the bar turns green. If the total number of customers exceeds 100000, the bar turns orange and if the total number of customers exceeds 0, the bar turns red. Otherwise, the bar is black.

This Bar chart allows us to compare the performance of each outlet. We can observe that the outlets 'RAN', 'RFY' and 'DMN', outlets have the highest number of customers.

Also, we can understand we can split the whole dataset into 3 segments that are ‘High Volume’, ‘Medium Volume’, and ‘Low Volume’. High volume outlets have customers count greater than 400000. The Medium volume outlets have customers count greater than 100000. And the rest of the outlet will be in the Low volume category.

Pie Chart- Circular Representation for Total Distribution of Daily Customers

Justification

A pie chart is a circular graph that is divided into slices to represent numerical proportions. The size of each slice is proportional to the value it represents. Pie charts are commonly used to display percentages or proportions of a whole. Pie charts are particularly useful for displaying data that is made up of several parts or categories. They can be used to represent the relative sizes of each category, and to show how the categories are related to each other. They are also useful for displaying data in a way that is easy to understand and interpret. They are visually appealing and can be used to communicate complex data in a simple and intuitive way. These are the reasons why I have used Pie Chart to visualize the 'Daily Customer Data' Dataframe.

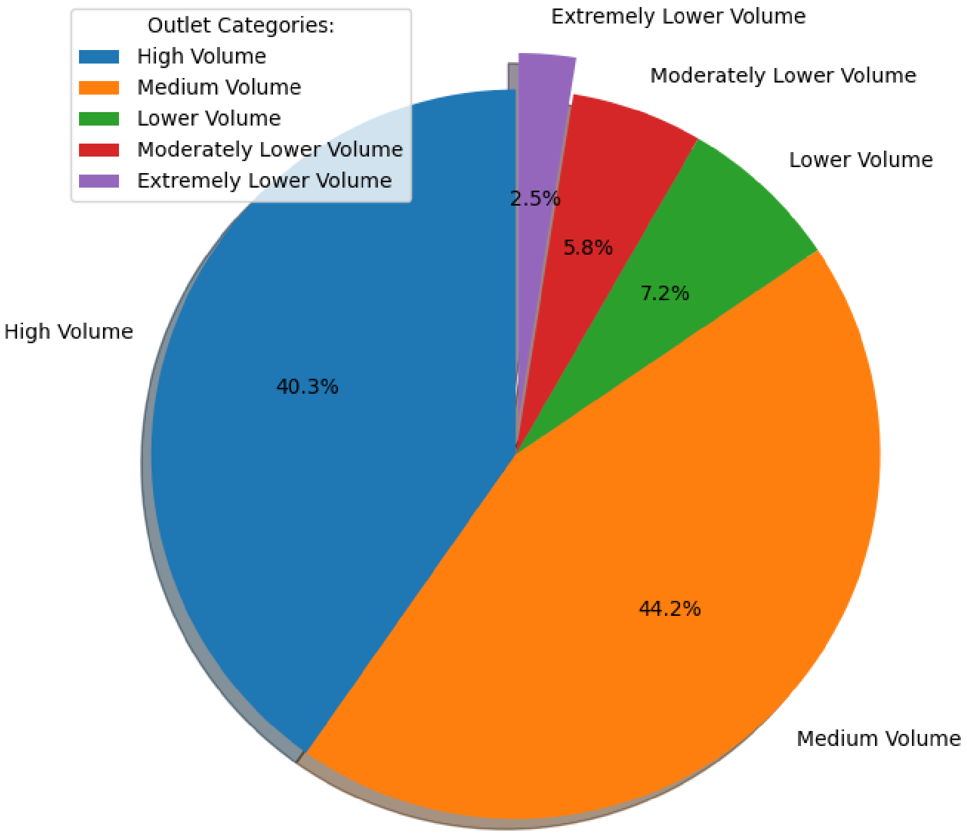

Figure 2 Pie chart representation for the total distribution of daily customers

Description

Figure 2 shows the Pie chart representation for the total distribution of the daily customers in each segment. As shown in Figure 2, the outlets classified as 'Medium volume' have highest number of daily customer count, resulting in 44.2% of the total customers. The outlets in the 'High Volume' group make up to 40.3% of all total customers who visited. The majority of customers fall into the 'High Volume' and Medium Volume' categories. The 'Low Volume' outlet group is further subdivided into three different sub-categories: 'Lower Volume', 'Moderately Lower Volume' and 'Extremely Lower Volume' and comprises of 7.2%, 5.8% and 2.5% respectively of the total customers who visited all outlets. The combined daily customer distribution across these three subcategories is 15.5%.

Box Plot- Distribution of Outlets with High Volume of Daily Customers

Justification

A box plot, also known as a box-and-whisker plot, is a graphical representation of data that shows the distribution of a dataset, particularly for large datasets. It can be used to compare the distribution of multiple datasets side-by-side. This allows you to see differences in the central tendency, variability, and skewness of the data, making it useful for comparing groups or sub-groups.Box plots are particularly effective for identifying outliers or extreme values in the data. Any values that fall outside of the whiskers are considered outliers and are plotted as individual points, making them easy to spot.Box plots can simplify complex data by summarizing the data into key statistical measures, such as the median, quartiles, and range. This makes it easier to understand the data and draw insights from it.

Overall, box plots are a powerful tool for visualizing and summarizing data. They can help to simplify complex data, identify outliers, and compare multiple datasets, making them a valuable tool for data analysis and interpretation. This is the reason why I have chosed this visualisation tool.

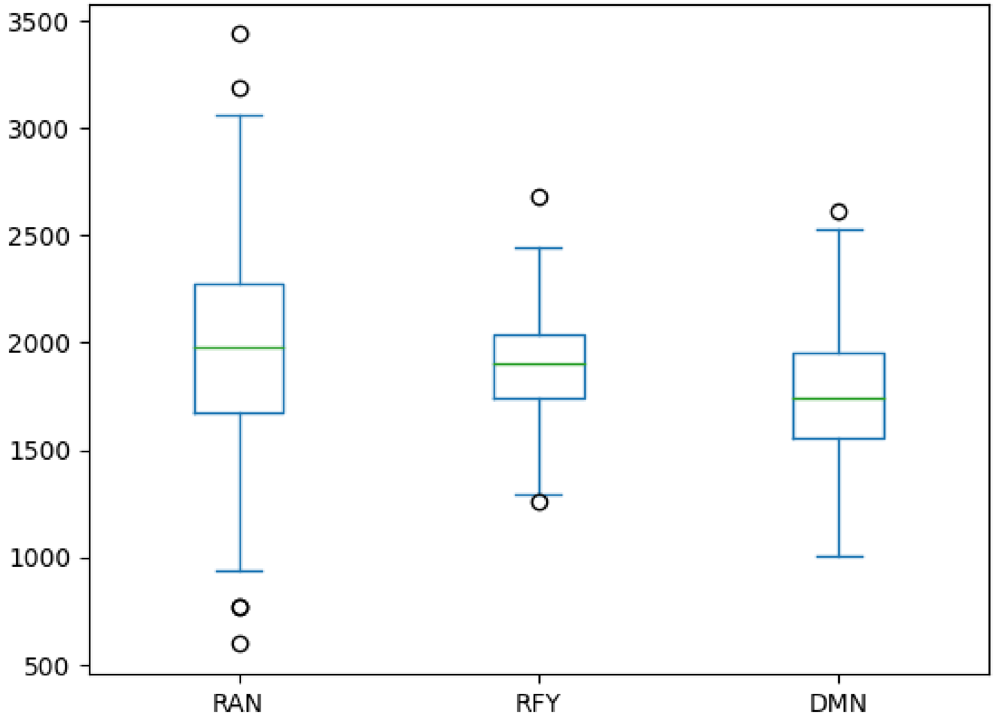

Figure 3 Box Plot- Distribution of Outlets with High Volume of Daily Customers

Description

A box plot consists of a rectangle (the box) and two whiskers that extend from the box. The box represents the interquartile range (IQR) of the data, which is the range of values between the 25th and 75th percentiles of the dataset. The line inside the box represents the median of the data.

The whiskers represent the range of the data outside of the IQR. Typically, the whiskers extend to 1.5 times the IQR or to the minimum and maximum values of the data, whichever is shorter. Any values outside of the whiskers are considered outliers and are plotted as individual points.

As shown in the figure 3, The outlet 'RFY' has a low variance i.e., most of its data is tightly clustered within a narrow range. This suggests that the number of customers visited the outlet 'RFY' remains consistent on a daily basis. Also, the outlet 'DMN' has a low variance, but high when compared with RFY. The box plot reveals that there is an outlier indicated by a circle in the lower whisker and upper whisker, matching the data point labelled as ‘RAN' and 'RFY'. This implies that 'RFY' and ‘RAN’ contains a major anomaly. Also, there is an outlier indicated by a circle in the upper whisker, matching the data point labelled as ‘DMN'. This implies that 'DMN' contains a minor anomaly. The box graph shows that the distribution of data is more spread out in the case of the outlet labelled 'RAN' than in the other outlets. This indicates that the number of customers visited the outlet 'RAN' is not stable, on the daily basis and that the statistics may be more variable.

Monthly Trend on Outlets with High Volume Customer Data

Justification

The importance of trends comes from the fact that they provide useful information about how a particular event or occurrence changes or evolves over time. This knowledge can help people and organizations better comprehend how to adapt to changes in the future. Businesses can keep their competitiveness by understanding trends and remaining up to current with shifting market conditions and customer preferences. The best practice to understand or visualize the trend we use line graphs. Line charts are especially useful for showing changes in one or more factors, and they can aid in the discovery of associations or correlations between different data points. I used a Line graph to represent the monthly trend to figure out which month receives the most or least visits.

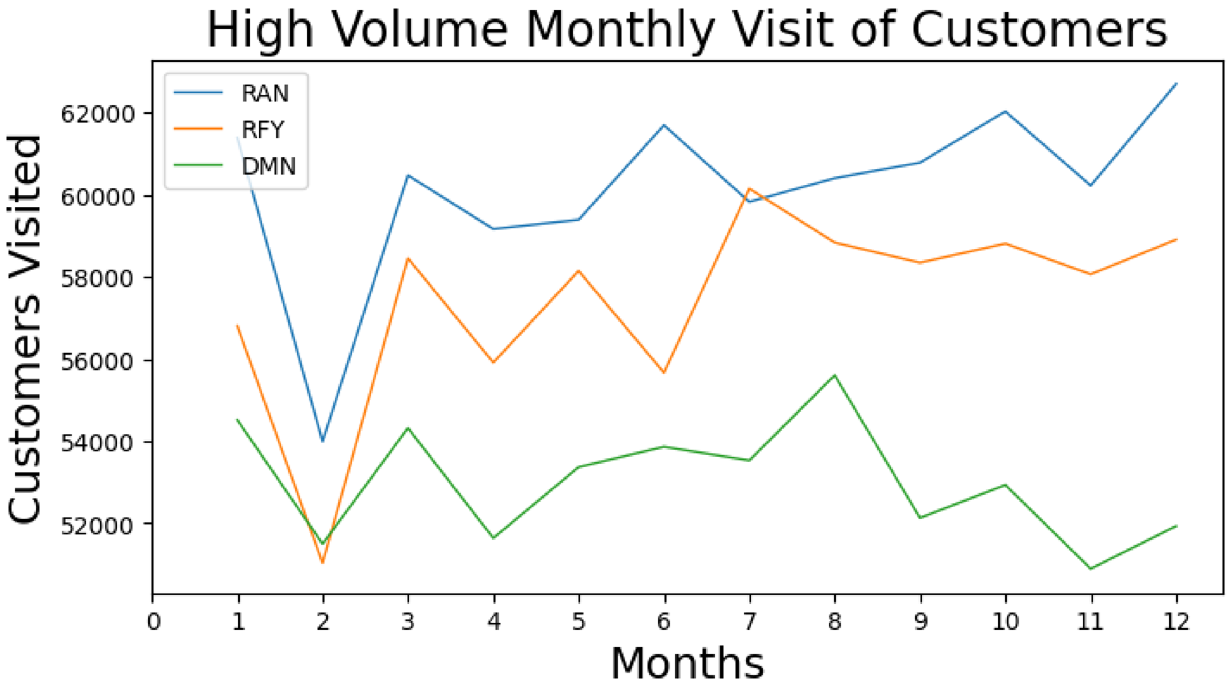

Figure 4 Monthly trend on High volume customer data

Description

According to Figure 4, there has been a significant decrease in customer visits during the month of February, which could indicate that customers are unwilling to visit the outlets during this time period. In addition, in March, there is a significant increase in customer visits for the outlets 'RFY' and 'RAN' whereas 'DMN' has not increased much. Furthermore, during the intermediate months, the number of customer visits fluctuates, with both increases and decreases observed. In this instance, the store 'RAN' received the highest number of customer visits in June and October, while the outlet 'RFY' received the highest number of customer visits in July. In contrast, the 'DMN' outlet had lower customer visits in November as compared with the other 2 outlets.

Interactive Line Plot - Distribution of Outlets with Extremely Lower Volume of Daily Customers

Justification

An interactive line plot is a type of data visualization that allows users to interact with the plot, manipulating or exploring the data in real-time. Users can zoom in and out of the plot to focus on specific areas of interest or to view the plot in greater detail. When the user hovers over a point on the line plot, a tooltip can appear that provides additional information about that data point. Interactive line plots can be particularly useful for exploring complex data, allowing users to quickly identify trends, patterns, and outliers. They can also be used to communicate insights to a wider audience, making the data more accessible and engaging.

Here, I have used Interactive Line Plot to explore the outlets in the Extremely Lower Volume in detail to identify whether new outlets have been opened during the year 2021 or any outlets have been closed by the company due to the low number of customer visit.

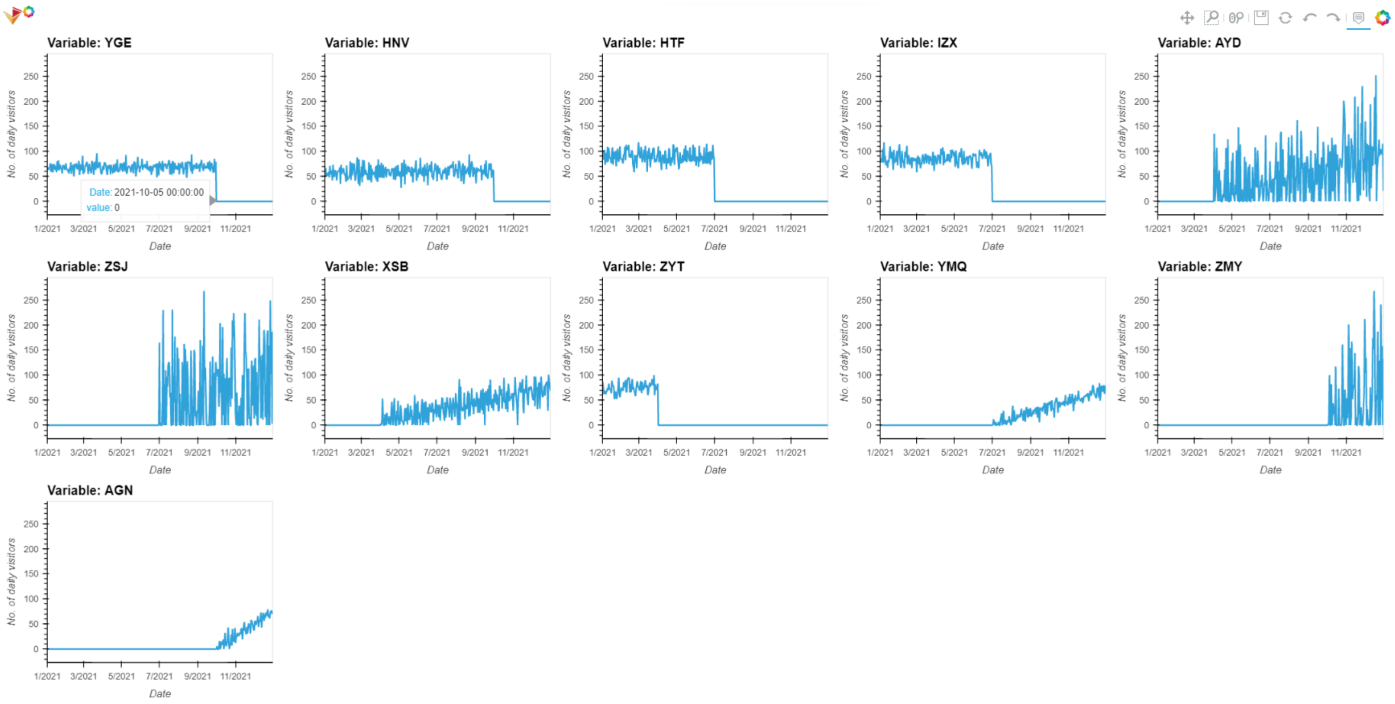

Figure 5 Interactive Line Plot- Distribution of Outlets with Extremely Lower Volume of Daily Customers

Description

From the figure 5 we can have the following findings:

Outlets closed during the year 2021: The outlets ‘YGE’,’HNV’,’HTF’,’IZX’,’ZYT’ were closed during the year 2021. These outlets were working in the beginning of the year with very few customer visits. There was no significant increase in the number of customers visiting these outlets. So, it was closed during the middle of the year 2021. Outlets ‘YGE’ and ’HNV’ were closed in the start of October month.

Outlets ’HTF’ and ’IZX’ were closed in the start of July and the outlet ZYT was closed in the start of April in the year 2021.

Outlets opened during the year 2021: The outlets ‘AYD’,’ZSJ’,’XSB’,’YMQ’, ‘ZMY’,’AGN’ were opened during the year 2021. The outlets ‘AYD’, and ’XSB’ were opened in the April and the outlet ’ZSJ’ and ’YMQ’ were opened in the July and the outlets ‘ZMY’,’AGN’ were opened in the October month of the year 2021

Data Exploration on Summary Data Frame

Pair Plot-To find the Correlation between all Attributes of Summary Data Frame

Justification

A pair plot is a type of data visualization that displays the pairwise relationships between different variables in a dataset. It is also known as a scatterplot matrix. In a pair plot, each variable is plotted against every other variable in the dataset, resulting in a grid of scatterplots. The diagonal of the grid shows a histogram of the distribution of each variable.

Here Pair plots are used to explore the relationships between multiple attributes in the 'Summary Data Frame', and to identify any patterns or trends that may exist. They can also be used to detect outliers, clusters, and correlations between attributes. A pair plot is a type of data visualization that displays the pairwise relationships between different variables in a dataset. It is also known as a scatterplot matrix.

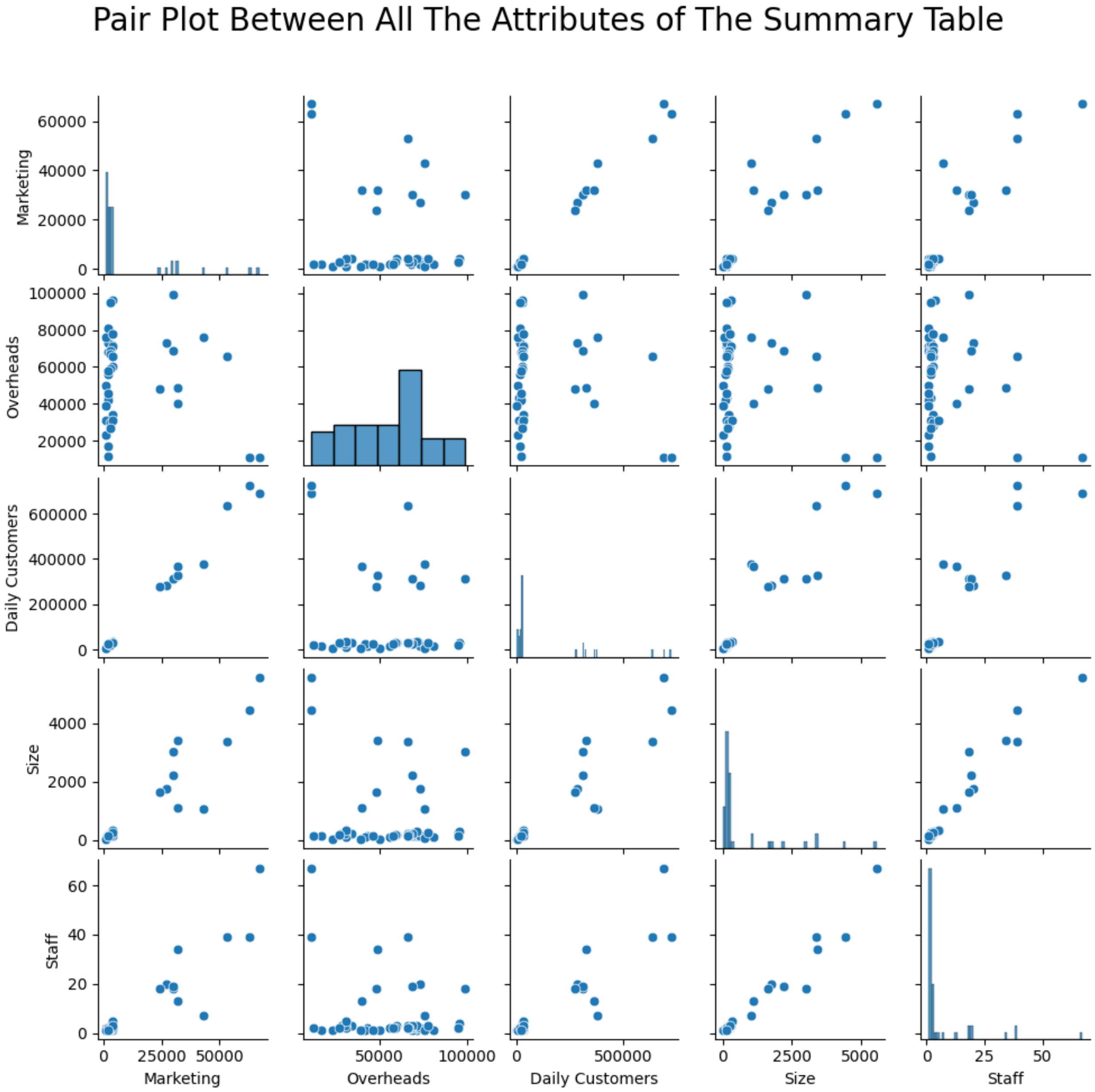

Figure 6 Pair Plot-To find the Correlation between all Attributes of Summary Data Frame

Description

In the figure 6 of Pair Plot each attribute ‘Outlet Marketing’, ‘Outlet Overheads’, ‘Outlet Size’, ‘Outlet Staff’ and the ‘Daily Customer Data’ of the ‘Summary Data Frame’ are plotted against every other attribute in the dataframe, resulting in a grid of scatterplots. The diagonal of the grid shows a histogram of the distribution of each attribute. From this grid, we can find the following observations

- ‘Outlet Marketing’ attribute has positive correlation with all the other attributes except ‘Outlet Overheads’ attribute. ‘Outlet Overheads’ have no relation with any of the attributes.
- ‘Outlet Marketing’ and ‘Daily Customers’ are positively correlated i.e. As the cost of marketing for each outlet increases, the number of daily customers visiting the outlets also increases.
- ‘Outlet Marketing’ and ‘Outlet Size’ are highly positively correlated i.e. As the size of the outlet increases, the cost of marketing for each outlet also increases.
- ‘Outlet Marketing’ and ‘Outlet Staff’ are positively correlated i.e., As the cost of marketing for each outlet increases, the number of staff in the outlets also increases.
- As the size of the outlet increases, the number of staff to manage the whole outlet also increases. i.e., The attributes ‘Outlet Size’ and ‘Outlet Staff’ have a high positive correlation with each other.
- ‘Outlet Size’ and ‘Daily Customers’ are positively correlated i.e. As the size of the outlet increases, the number of daily customers visiting the outlets also increases.
- ‘Outlet Staff’ and ‘Daily Customers’ are correlated with a coefficient, r = 0.92. As the number of customers visiting the outlets daily increases, the number of staff needed to attend as many customers should be increased as well.

Heat Map- To find the Correlation between each Attribute in the Summary Data Frame

Justification

Heat maps are often used to visualize large data sets and to identify patterns, trends, and relationships in the data. In the heat map, each data point is assigned a color based on its value. Typically, the highest values are represented by darker colors such as red, while lower values are represented by lighter colors such as green or blue. The result is a visual representation of the data that makes it easy to identify areas of high and low value. The correlation coefficient measures the strength of the linear relationship between two variables, and ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation.

Here I have used heat map to depict the relationship between each attribute of the 'Summary Data' like 'Outlet Marketing', 'Outlet Overheads', 'Outlet Size', 'Outlet Staff' and the 'Daily Customer Data'

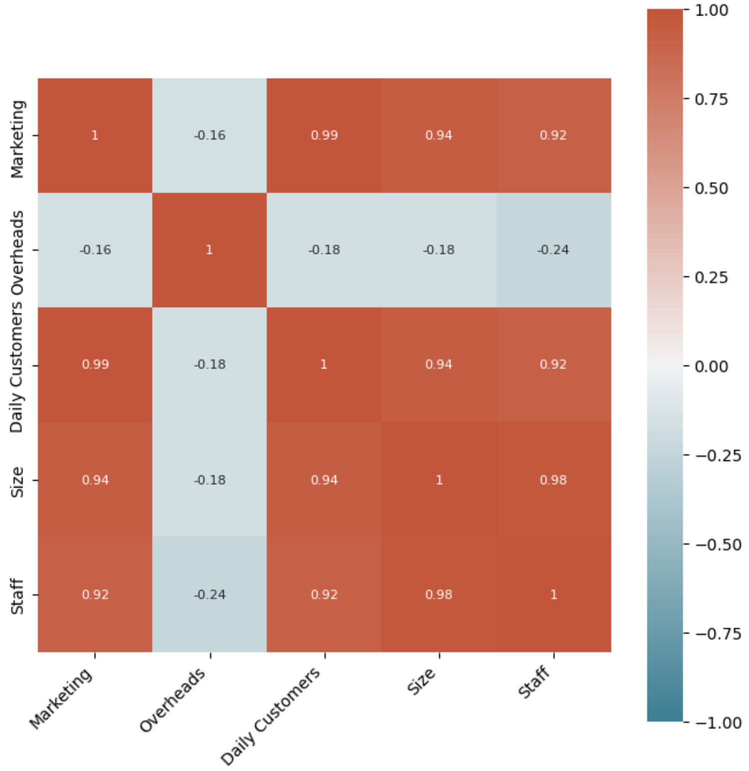

Figure 7 Heat Map- To find the Correlation between each attribute in the Summary Data Frame

Description

From the figure 7, we can understand the high positive correlation between the attributes 'Outlet Marketing', 'Outlet Overheads', 'Outlet Size', 'Outlet Staff' and the 'Daily Customer Data' of the 'Summary Data Frame' as follows:

- 'Outlet Size' and 'Outlet Staff' are correlated with a coefficient, r = 0.98. i.e. As the size of the outlet increases, the number of staffs to manage the whole outlet should be increased
- 'Outlet Marketing' attribute has greater correlation with all the other attributes except 'Outlet Overheads' attribute. 'Outlet Overheads' have no relation with any of the attributes.

- ‘Outlet Marketing’ and ‘Daily Customers’ are correlated with a coefficient, r = 0.99. i.e. As the cost of marketing for each outlet increases, the number of daily customers visiting the outlets also increases.
- ‘Outlet Marketing’ and ‘Outlet Size’ are correlated with a coefficient, r = 0.94. i.e. As the size of the outlet increases, the cost of marketing for each outlet also increases.
- ‘Outlet Marketing’ and ‘Outlet Staff’ are correlated with a coefficient, r = 0.92
- ‘Outlet Size’ and ‘Daily Customers’ are correlated with a coefficient, r = 0.94. i.e. As the size of the outlet increases the number of daily customers visiting the outlets also increases.
- ‘Outlet Staff’ and ‘Daily Customers’ are correlated with a coefficient, r = 0.92. As the number of daily customers visiting the outlets increases, the staff needed to attend as many customers should be increased as well.

Interactive Scatter Plot- To Study the Positively Correlated attributes in the Summary Data Frame

Justification

An interactive scatter plot is a type of data visualization that allows users to interact with the plot and explore the data in a more dynamic way. Interactive scatter plots allow users to zoom in and out on specific regions of the plot and extend to different areas of the plot, which can help to explore the data in more detail. Interactive scatter plots can display the exact values of the x and y variables about each point on the plot when the user hovers over it with their mouse, Interactive scatter plots are useful for exploring complex datasets and can provide a more dynamic and engaging way to visualize data compared to static plots. Here Interactive Scatter plot is used to study in detail about the positively correlated attributes of the 'Summary Data Frame'

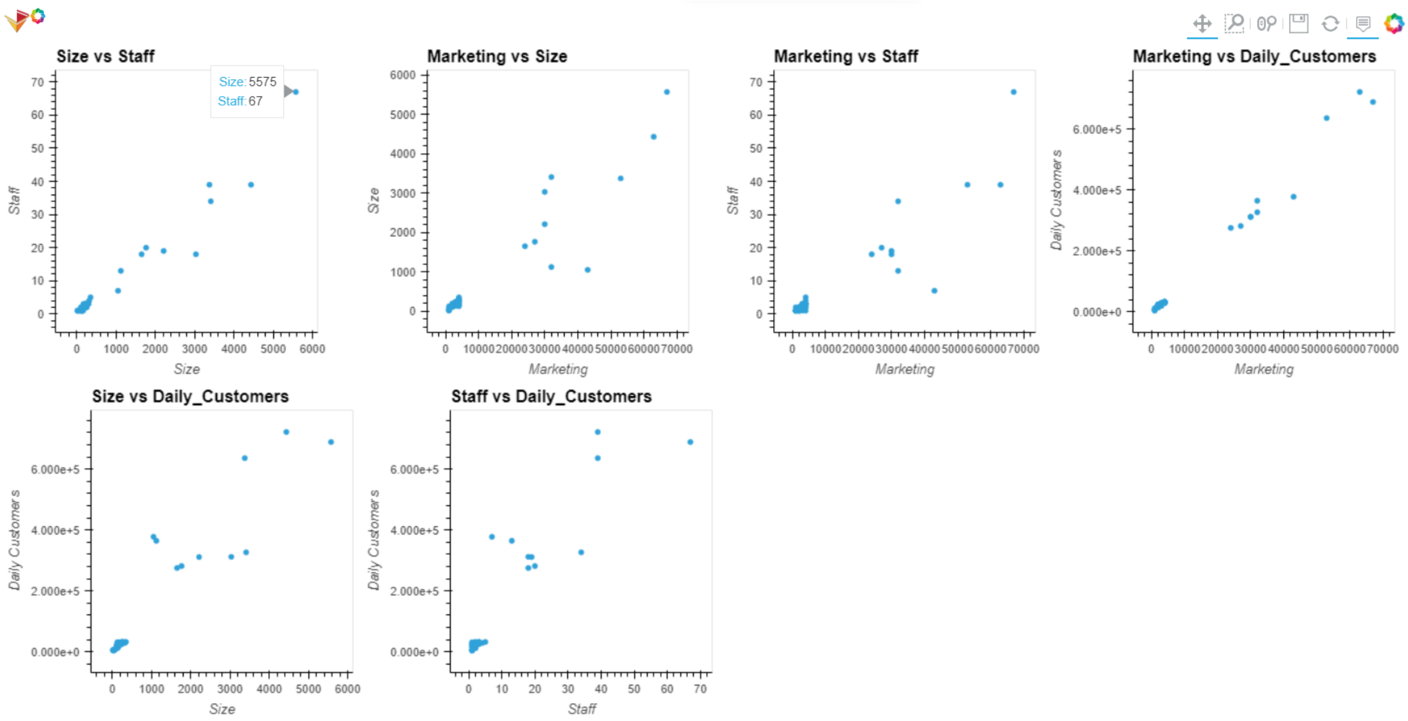

Figure 8 Interactive Scatter Plot- To find the Positive Correlation between each attribute in the Summary Data Frame

Description

From the figure 8 of Interactive Scatter Plot we can understand the high positive correlation between the attributes 'Outlet Marketing', 'Outlet Size', 'Outlet Staff' and the 'Daily Customer Data' of the 'Summary Data Frame' as follows:

- 'Outlet Size' and 'Outlet Staff' are correlated i.e. When the size of the outlet increases, the number of staffs to manage the whole outlet should be increased
- 'Outlet Marketing' and 'Daily Customers' are correlated. i.e. As the cost of marketing for each outlet increases, the number of daily customers visiting the outlets also increases.
- 'Outlet Marketing' and 'Outlet Size' are correlated. i.e. As the size of the outlet increases, the cost of marketing for each outlet also increases.

- ‘Outlet Marketing’ and ‘Outlet Staff’ are correlated i.e., As the cost of marketing for each outlet increases, the number of staff in the outlets also increases.
- ‘Outlet Size’ and ‘Daily Customers’ are correlated. i.e. As the size of the outlet increases the number of daily customers visiting the outlets also increases.
- ‘Outlet Staff’ and ‘Daily Customers’ are correlated. As the number of daily customers visiting the outlets increases, the staff needed to attend as many customers should be increased as well.

Critical Review

ChrisCo provided us with five datasets that contain daily customer visit counts and additional details for each of their 45 locations. The datasets include information such as annual customer visits, overhead expenses, local marketing costs, staff numbers, and outlet size. We converted the datasets into two separate data frames for research purposes.

We conducted exploratory data analysis on the first data frame, which contained the daily customer visit counts for each location. This allowed us to determine the dataset's shape and obtain figures such as the minimum, maximum, and average values. We also performed exploratory data analysis on the second data frame, which contained information about each outlet. This helped us understand the data structure and facilitated the creation of visualizations and additional analysis.

To analyze the first data frame, we used various types of charts such as bar, line, and pie charts to visualize and represent the data. These charts helped us identify trends and seasonal patterns in customer visits to outlets over time. Additionally, we used a box plot to understand the data distribution and identify any outliers.

The second data frame included yearly data for each of the 45 outlets, which allowed us to comprehend the relationships between different characteristics. We used a variety of graphs such as scatter plots, heat maps, and radar plots to analyze this data. These graphs helped us understand the connections between the different characteristics and identify any anomalies or outliers in the data. We used interactive plots with basic modification tools to allow for more in-depth analysis due to the large number of characteristics.

Conclusion

In conclusion, data visualization is an essential part of data analysis that allows us to present complex data in a simplified and understandable manner. Visualizations help us identify patterns and trends, detect outliers, and communicate insights in a clear and concise way.

The use of different types of charts and graphs, such as bar charts, pie charts, line charts, scatter plots, heat maps, and box plots, allows us to explore and analyze data from various angles. Interactive visualizations with basic modification tools also allow for more in-depth analysis of complex data sets.

Moreover, visualizations can be used to enhance decision-making processes by helping to identify areas of opportunity or concern, track progress, and forecast future trends. They are also useful in communicating data-driven insights to non-technical audiences, such as executives or policymakers.

Overall, data visualization is an important tool that enables us to transform raw data into meaningful insights and improve decision-making processes. As data continues to play an increasingly important role in various fields, the ability to visualize and communicate insights through data visualization will become even more critical.

References

García-Peñalvo, F. J., Seoane Pardo, A. M. & García-Holgado, A., 2022. The impact of visualization types on data analysis: A review of the state of the art. *International Journal of Information Management*, Volume 62, p. 102423.

Soltani, K. & Boehmke, B. C., 2022. The effect of data visualization on user comprehension and decision making: A systematic review. *Computers in Human Behavior*, Volume 127, p. 107192.