COURSEWORK

Module: COMP-1800-M01-2022-23

Data Visualisation

DIVYA PYNENI

001207695

Table of Contents

List of Figures

Introduction:

Data visualization is the process of representing data and information in a visual format, such as graphs, charts, and maps, to make it easier to understand and analyze. It is a valuable tool that enables users to identify patterns, trends, and relationships that may not be evident in raw data. With data visualization, users can explore and interpret large datasets and communicate insights effectively.

Data visualization has various applications in different fields, including business, science, journalism, and technology. In business, it can be used to monitor sales performance, track customer behavior, and identify market trends. In science, it can help researchers analyze experimental data and present research findings. In journalism, it is used to create visual representations of news stories, and in technology, it can be used to monitor network traffic and identify security threats.

Data visualization has become increasingly important with the rise of big data, which generates vast and complex datasets. It allows individuals and organizations to understand complex data sets and make informed decisions based on data-driven insights. As such, it is a crucial component of data-driven decision-making, empowering individuals and organizations to achieve their goals and objectives by using data to gain insights and make better decisions.

Implementation:

Bar Graph:

From the source of data, I have created two Data Frames, daily_customer_df and summary_data. In which daily_customer_df includes the data of daily Customers Visits and summary_data includes all the remaining data sets named daily Customers data, marketing, Overheads, Size and Staff.

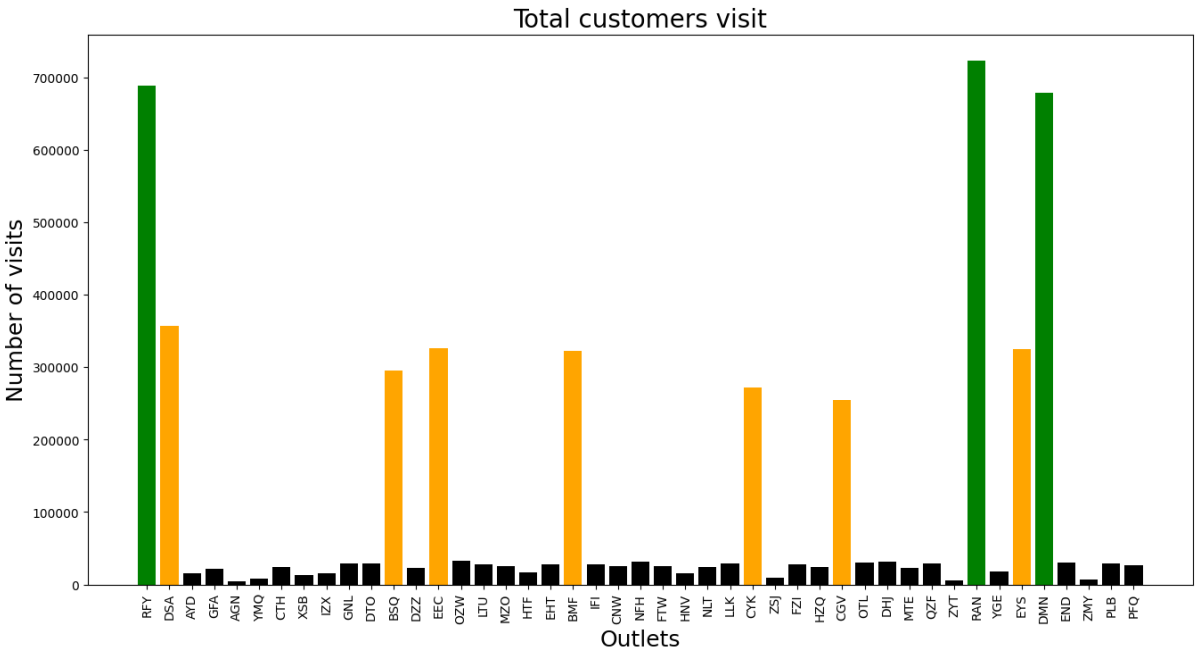

Fig 1: Bar graph

Justification:

Bar graphs are an effective way to present categorical data because they make it easy to compare values, it can be understood by both technical and non-technical audiences. These can help engage the viewer in the data analysis process. It can make the data more interesting and memorable, which can increase engagement and retention.

Description:

From the above figure, we have taken Outlets on the x-axis and the Total Number of Visits on the y-axis. I have segmented the daily customers data into three parts based on values, which are High, Medium and Low Customers Visit. To represent these in bar graph, we have used different colors. Green colored outlets represents the High Number of visits, Orange colored Outlets represents the Medium Number of visits and black colored outlets represents the Low Number of visits.

Line Plot

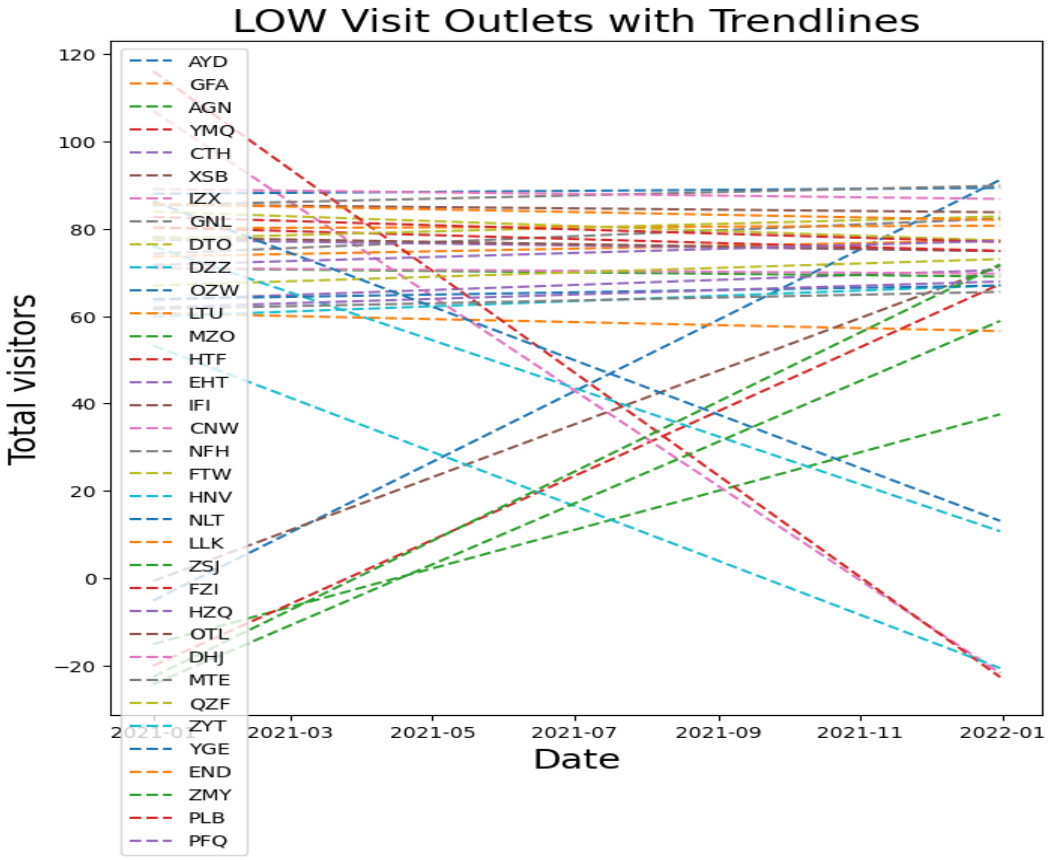

Fig 2 : Line plot for lower visits with trendlines

Justification:

Line plots are a useful tool for visualizing trends and patterns in data over time. They can show changes and fluctuations in the data with a single glance and are useful for identifying outliers, clusters, and trends. Compared to other types of plots, line plots are simple, easy to interpret, and don't require a lot of data preparation. Additionally, they can be used to plot multiple variables on the same plot for comparison purposes.

Description:

The above picture reveals the trends for low visited Outlets. Some outlets has the increasing trends which are ZSJ, ZMY, MZO, PLB, NLT AND IFI. And some outlets has the decreasing trends which are YMQ, IZX,OZW,DZZ and HNV and remaining outlets has no observed trends in low visited Outlets. And has more information on High and Medium visited outlets in the Notebook. If we observe them, RFY, DMN both has increasing trends, remaining doesn’t possess any trend. Every outlet has visits at the beginning of the year as well as at the end of the year. Therefore, no new Outlet has started during the year or closed during year.

Scatter Plot

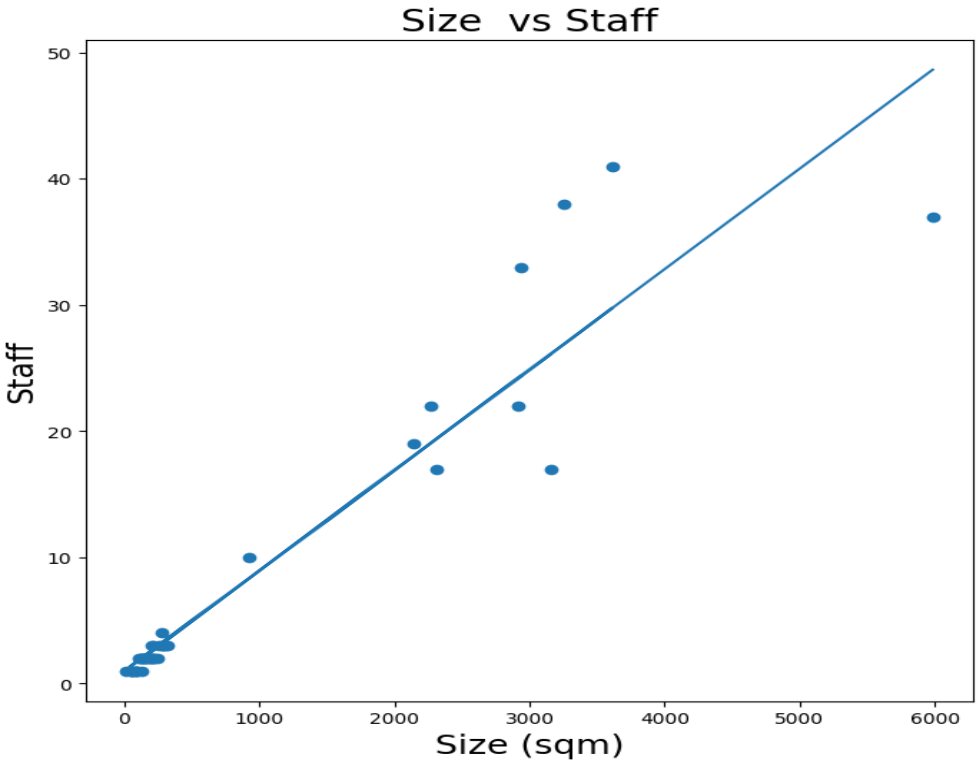

Fig 3 : Scatter plot

Justification:

Scatter plots are useful when we want to visualize the relationship between two variables. They are particularly helpful when we want to see if there is any correlation between the variables, or if there are any outliers in the data. The use of dots in the plot makes it easy to see individual data points and patterns in the data. Scatter plots can also be used to identify clusters of data points, which can be helpful in data clustering and classification tasks.

Description:

From the above scatter plot, it is plotted between Staff and Size and it is strongly correlated. Which means staff that have been allocated to each outlet is perfectly correlated with the area or size (square per meter) for each outlet. On the other hand, if we do scatter plot for both high and medium visits following outlets has the moderate strong positive correlation.

- High visits -- RFY vs DMN
- Medium visits -- EEC vs CGV, BMF vs CGV, CGV vs EYS

Auto Correlation Plot – (Distribution)

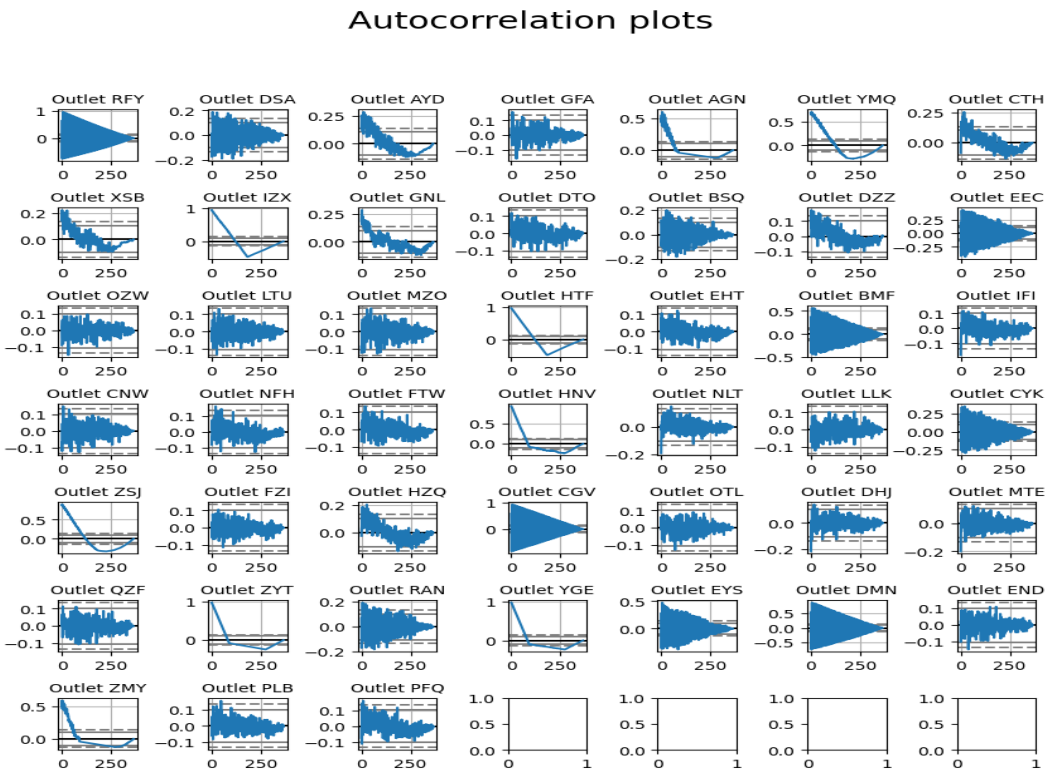

Fig 4: Auto-correlation for all Outlets

Justification:

Auto-correlation plots are used to visualize the correlation between a variable and its lagged values. It helps to identify the presence of any patterns or trends in the data.

Distribution:

From the above plot, we can observe that autocorrelation have been done all outlets and plotted as sublpots. x-axis indicates lag value and y-axis indicates the range of values from -1 to 1. So now, we can identify trends and seasonality. The following outlets has the seasonality.

- Outlets RFY, AYD, CTH, DTO, MZO, ETH, BMF, and FTW have higher correlation values at lags that are multiples of 7, suggesting weekly seasonality.
- Outlets IFI, NFH, CNW, ZSJ, CGV, DHJ, RAN, and YGE have higher correlation values at lags that are multiples of 12, indicating monthly seasonality.
- Outlets HZQ, ZYT, DMN, and END have higher correlation values at lags that are multiples of 4, suggesting quarterly seasonality.

Box plot

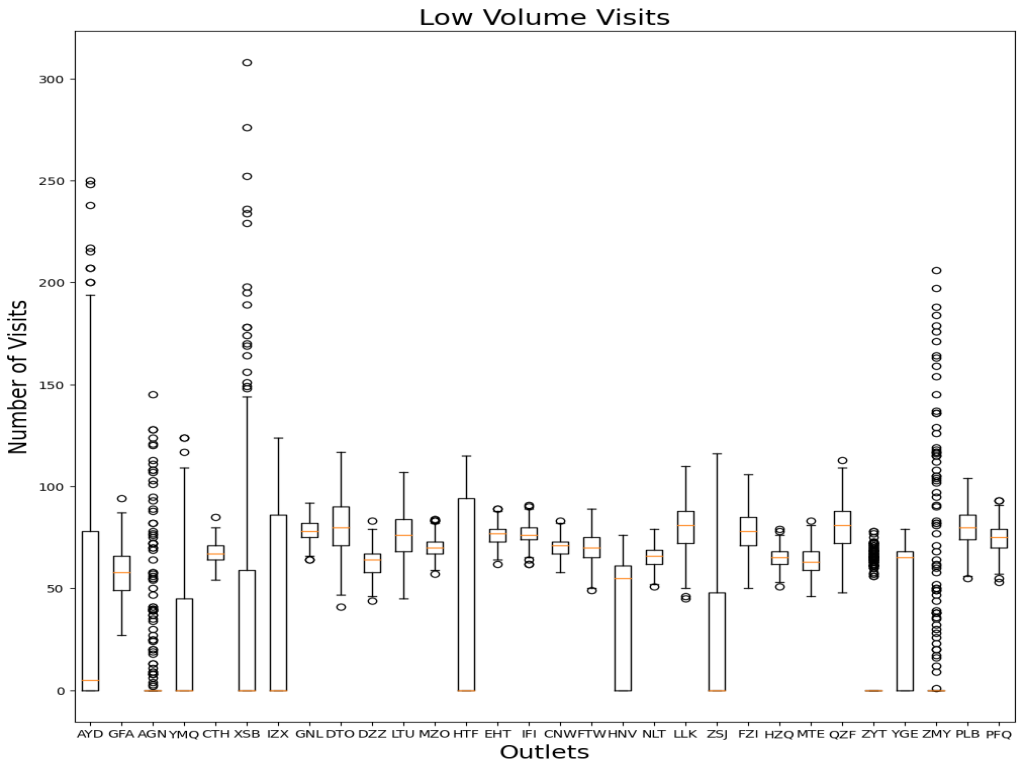

Fig 5 : Box plot for low visits

Justification:

Box plots provide a way to visualize the distribution of data, median, whisk as well as any potential outliers or extreme values, in a clear and concise manner.

Description:

From the above figure, we can observe that Box plot has done for Low Visits of Customers and outlets has taken on x-axis and taken number of visits on y-axis. Outliers are the points which are beyond the minimum and maximum values. Most of the outliers has been observed at AYD, AGN, XSB, ZYT and ZMY. Following Outliers has the less Outliers.

- High Visits – Outlier RAN
- Medium Visits – Outliers BSQ, EEC, CYK and EYS
- Low Visits –Outliers GFA,YMQ,CTH,GNL,DTO,DZZ,MZO,EHT,IFI,CNW,FTW,NLT,LLK,HZQ,MTE,QZF,PLB and PFQ.

Heat Map

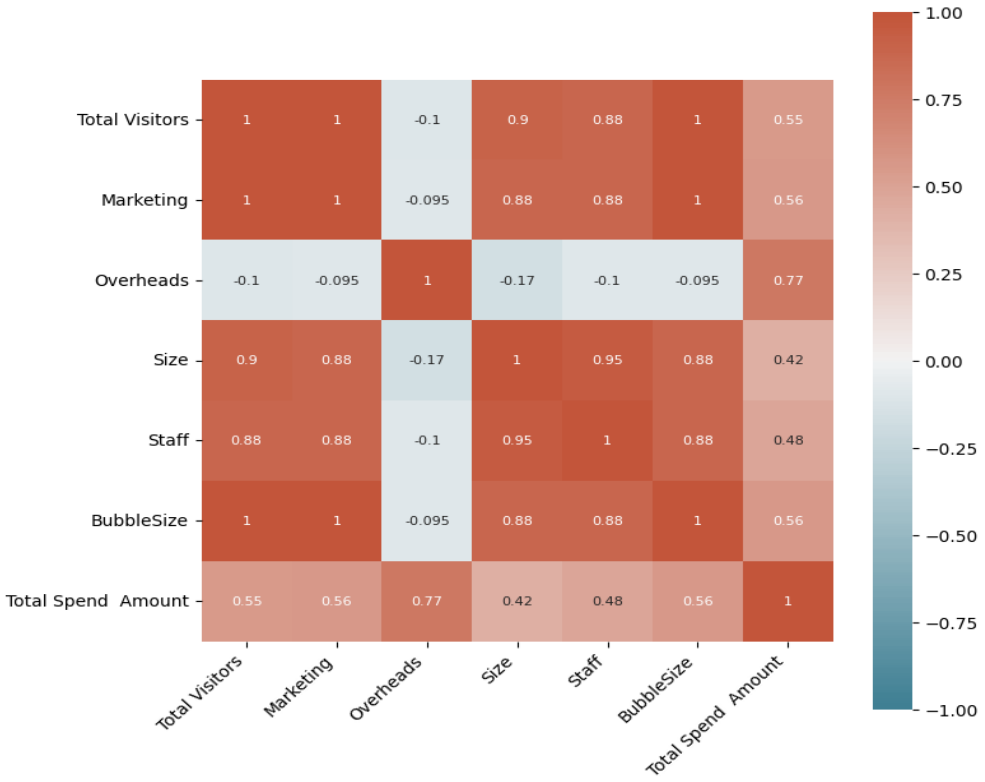

Fig 6 : Heat map for summary data

Justification:

Heat maps display correlation values using color-coded cells, with darker shades indicating higher correlation values and lighter shades indicating lower correlation values. These coefficients have values that range from -1 to +1, with higher absolute values indicating stronger correlations.

Description:

From the above heatmap, we have plotted for the summary data, Total Visitors, Marketing, Overheads, Size, Staff, BubbleSize and Total Spend Amount. The following columns has perfect correlations. Those are, Total visitors vs staff, Total visitors vs size, Total visitors vs marketing, Marketing vs staff, Marketing vs size, size vs staff, size vs Bubble size, staff vs Bubble size.

Interactive Bubble Plot

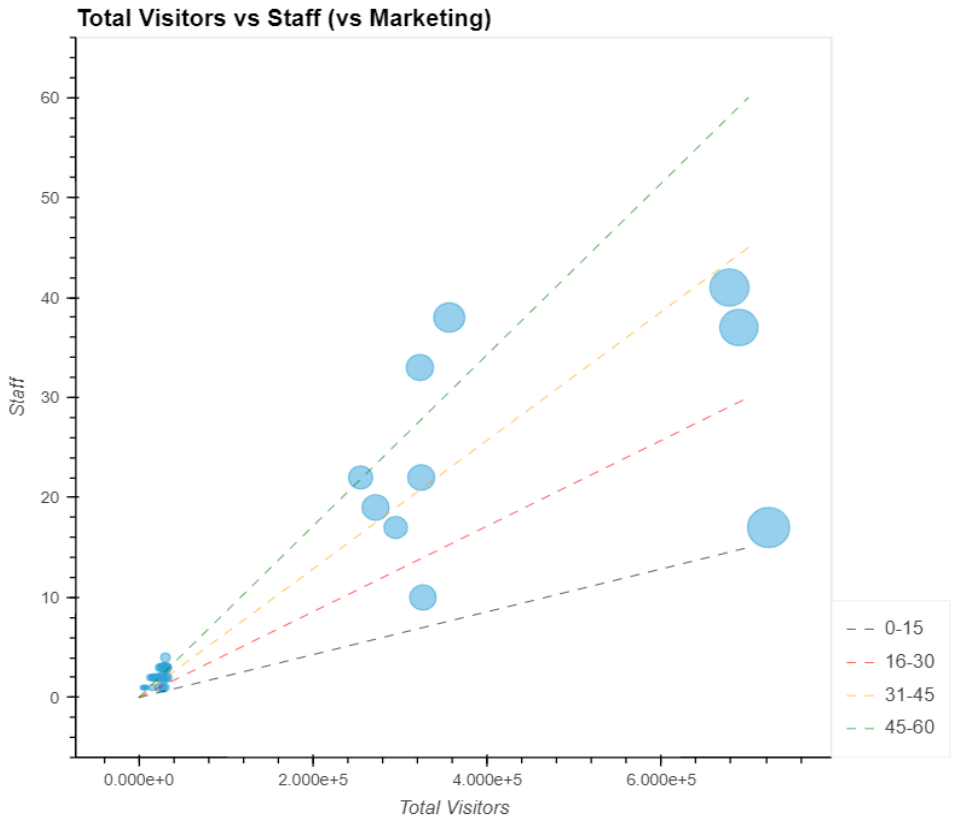

Fig 7: Interactive Bubble plot

Justification:

Interactive bubble plots are an effective way to visualize data that has three or more variables. They offer an interactive experience where users can modify the bubbles' size, color, and position based on various variables. This capability helps users detect patterns and relationships that may not be evident with other visualization tools. Interactive bubble plots can also aid in detecting outliers and clusters within the data.

Description:

The above Interactive Bubble plot is between Total visitors vs Staff (vs Marketing). On x-axis we have taken total visitors and on y-axis we have taken staff from summary data. In this bubble plot we have partitioned the staff into 4 categories which are (0 to 15, 16 to 30, 31 to 45 and 45 to 60). Those Outlets which are having the corresponding staff have been plotted. The use of this interactive bubble plot is we can see the details of each bubble if we place cursor on it. From the plot, we can see that some of the Outlets has highest number of visitors but comparatively having less number of Staff. For example, Outlet RAN has the highest number of visits but having less staff 17 comparing with the other high visited outlets.

Interactive Histogram

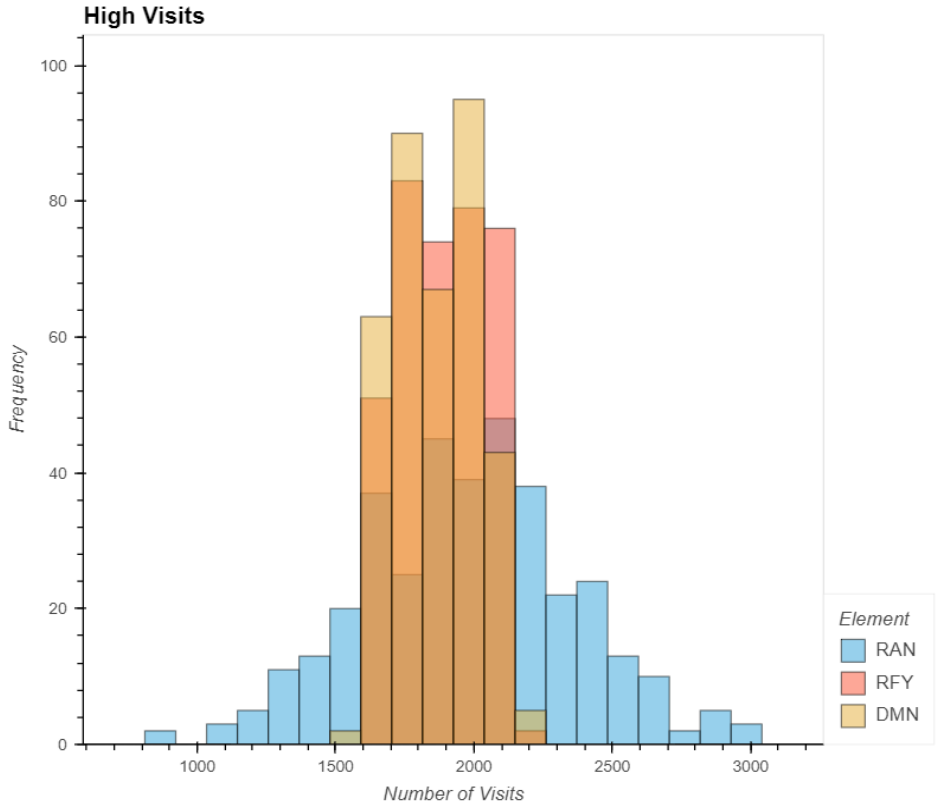

Fig 8 : Interactive Histogram

Justification:

Interactive histograms are effective in visualizing the distribution of a single variable or comparing the distribution of multiple variables. They allow users to interactively adjust the bin width, overlay multiple histograms, and filter the data based on different variables. This helps in identifying patterns and trends in the data that may not be easily observable using other visualization tools. Additionally, interactive histograms can help in identifying outliers and skewness within the data.

Description:

The above interactive histogram is for the High Number of visitors, Total number of visitors is taken on the x-axis and the Frequency has taken on the y-axis. In a histogram, frequency refers to the number of times a particular value or range of values occurs in a dataset. There are three histograms that are placed side by side. The Outlet DMN has the highest frequency, which means it has occurred many visits.

Critical Review:

Data visualization tools such as bar graph, pie chart, line plot, scatter plot, auto-correlation, Heat map, correlograph, radar chart or spider chart, bubble plot, and interactive plots and much more has been used to visualize the data that have been given. As per the requirement, company mostly focuses on High and medium Visits of Customers for each outlet. And they want a short information on Low Visits of Customers. Then we can go for bar graph or pie chart to represent the raw data. As the Daily Customers data includes one year information, which has a business to do with Time Series. Which plays a major role to identify trends and patterns. After working with correlations, I understood that, it gives the information which useful further.

In my work, I came to a point where if I want to find outliers, I would go for Box plot which makes information more understandable. Heatmaps are useful to find accurate correlations based on values which ranges from (-1 to 1). Interactive plots helps to notice the information more accurately it includes many tools like zoom, padding, etc., Interaction plots supports line plots, histogram, bar graphs, heat maps, scatter plots and bubble plots. However, while working with data, I preferably say that some datasets are not at all useful to work with except making summary data.

Summary:

Brief summary of each visualization tool:

● Bar graph: This graph is used to show the total number of visits to each outlet, with the outlets segmented into high, medium, and low visit categories. The colors represent the different visit categories. The graph shows trends for low-visited outlets, with some outlets having increasing trends and others having decreasing trends.

● Scatter plot: This graph shows the correlation between the number of staff allocated to each outlet and the size of the outlet in square meters. The plot shows a strong positive correlation between the two variables.

● Autocorrelation plot: This graph shows the correlation between a variable and itself over time, with lag on the x-axis and correlation on the y-axis. The graph shows that some outlets have weekly, monthly, or quarterly seasonality.

● Box plot: This graph is used to show the number of visits to each outlet, with outliers shown as points beyond the minimum and maximum values. The graph shows that some outlets have more outliers than others.

● Heatmap: This graph shows the correlation between different variables in the data, with darker colors indicating stronger correlations. The heatmap shows that some variables have perfect correlations with each other.

● Interactive bubble plot: This plot shows the number of visitors to each outlet plotted against the number of staff, with the size of the bubble indicating the amount spent on marketing. The plot allows the user to interactively explore each bubble for more information.

● Interactive histogram: This plot shows the frequency of high numbers of visitors to each outlet. The histogram shows that some outlets have higher frequencies than others.

Overall, the plots and graphs in the report provide a comprehensive and informative overview of the data. They allow the user to explore different aspects of the data and gain insights into patterns and trends.