Data Visualisation Coursework

Submitted by: Dorjada Halili

Student ID: 001293509

Subject: COMP1800 DATA VISUALISATION

SCHOOL OF COMPUTING AND MATHEMATICAL SCIENCE

Table of Contents

Table of Figures

Introduction to Data Visualisation

Data Visualisation can be defined as the presentation of data using common graphics such as charts, graphs, plots, and even animations. The purpose of Data visualisation is to communicate information clearly and give insights in a way that is easier to understand to a range of audiences, not just data experts. The main goal of data visualisation is to make pattern recognition, trend recognition and outlier recognition simpler in massive datasets. Data Visualisation helps to highlight the most useful insights from a dataset, making it easier to spot trends, patterns, outliers, and correlations. To create an effective data visualisation, it is important to choose the right type of visualization because each type of visualisation is designed to highlight different aspects of the data. In addition, it is highly important to consider the design elements of the visualization, such as colour, layout and labelling. A well-designed data visualisation can make data more accessible and understandable, whereas a poorly designed visualisation can be confusing and misleading. Understanding data is advantageous not only in the STEM field but also in various other sectors such as government, finance, marketing, history, consumer goods, service industries, education, sports, etc.

There are two types of data visualisation:

- Exploration (helps to figure out what's in your data, takes place while still analysing the data)
- Explanation (helps to communicate what is found and comes towards the end of the process)

Justifications and Descriptions of the 8 Visualisations:

1. Bar Graph

A bar graph, also known as a bar plot, is a type of visual representation of data and is used to display categorical data with numerical values. It uses a series of bars that displays data in two axes, x and y particularly. The height or length of each bar represents the frequency or proportion of each category. The values of each category are shown on the y-axis while the categories of data in the x-axis. When plotting a bar chart, categories are grouped in different colours. A bar graph is better suited to use in the case of our data because it is easier to interpret and ensures clear comparisons between the lengths of the bars, while a pie chart would be used for showing how a single variable is broken down into percentages.

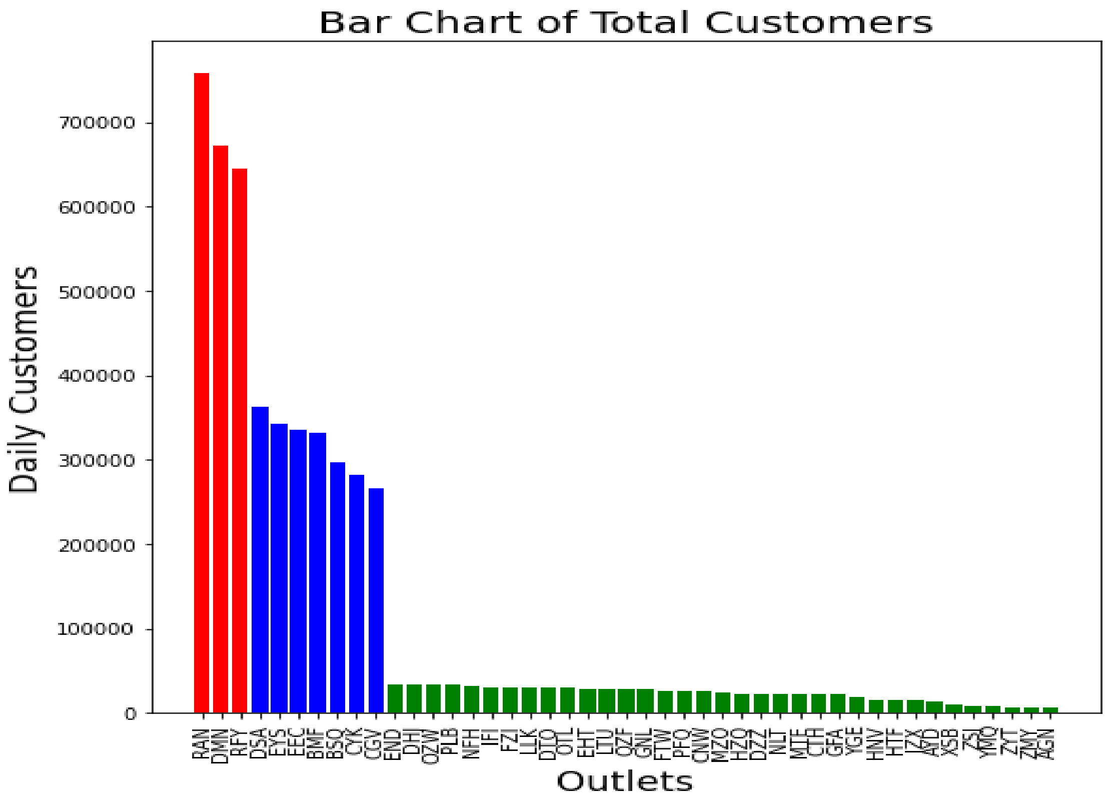


Figure 1. Bar Chart showing the customers of all outlets for the year 2021

The above bar chart displays the total number of customers for all outlets in the year 2021. Because it is a large dataset, to make the visualisation more distinct it is important to segment the outlets into 3 categories i.e. High, Medium and Low. From the graph, we can see that each outlet has got its bar and is being visualised in descending order from the highest to the lowest one. Each bar represents a different outlet category, with the height of the bar indicating the number of visitors to that category. Each category is shown in a particular colour, red for High volume, blue for medium volume and green for low volume outlets. From the bar chart, we can see that RAN, DMN and RFY have the highest number of visitors throughout the year while XSB, ZSJ, YMQ, ZYT, ZMY and AGN have the lowest number of visitors. This type of bar chart allows us to compare the number of visitors to each outlet category and identify which categories are most popular or least popular.

2. Heat Map

Using a heatmap allows us to understand how variables are related to each other and the strength of the relationship. Heatmaps are a powerful data visualisation tool that uses colour coding, so each cell of a grid is assigned a colour that represents the corresponding value of the data, in this way it is easier to compare the values of different variables. A heatmap could be used to show the correlation between different variables in the data, such as the correlation between daily customers and marketing spend, overheads and size of the outlets and so on. The strength of the correlation between two variables can also be represented by the intensity of the colour in the heatmap where warmer colours mean positive correlation while cells with low values are represented with cool colours. The value of the correlation can take any value between -1 to 1.

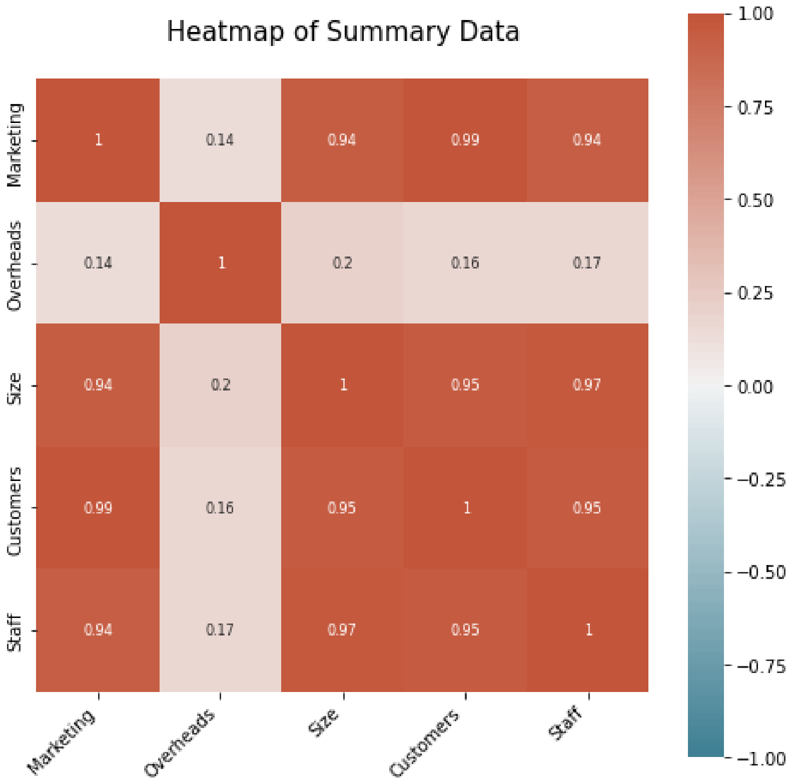


Figure 2. Heat Map showing correlation among the summarised categories

The above Heat Map shows the correlation between the summarised categories that are: marketing, overheads, size, customers and staff. The red colour that takes value ‘1’ indicates a perfectly positive correlation and this means that as one variable increases, the other variable also increases linearly, whereas the blue colour that takes the value ‘-1’ perfectly negative correlation means that one variable increases, the other variable decreases linearly. In this Heat Map we can see that marketing and customers have a value of 0.99 which means these two variables are perfectly positively correlated, so if the number of customers increases so does marketing expenses. Moreover, there is also a positive correlation between marketing and staff. Overheads and marketing are the least correlated to each other as shown in the above figure the value is 0.14. Marketing and staff, and marketing size share the same value that implies if marketing increases so do the two other sales.

3. Scatter Plot

A Scatter Plot is a type of visualisation that uses dots to display the relationships between two numerical variables. It consists of a set of points that represent a pair of values for the used variables. Each point in the scatter plot represents a data point and the value of each data point is indicated by the positioning of each dot in the x and y-axis. The strength and nature of the relationship can be revealed by the pattern of points in the scatter plot. Using a Scatter plot we can identify trends, outliers, clusters and correlations in the data. An outlier in a scatter plot is a data point that does not fit the pattern and deviates from the trend of data. To have a better understanding of the results in the scatter plot it is common to add a trend line that provides additional insights into how strong the relationship between the two variables is.

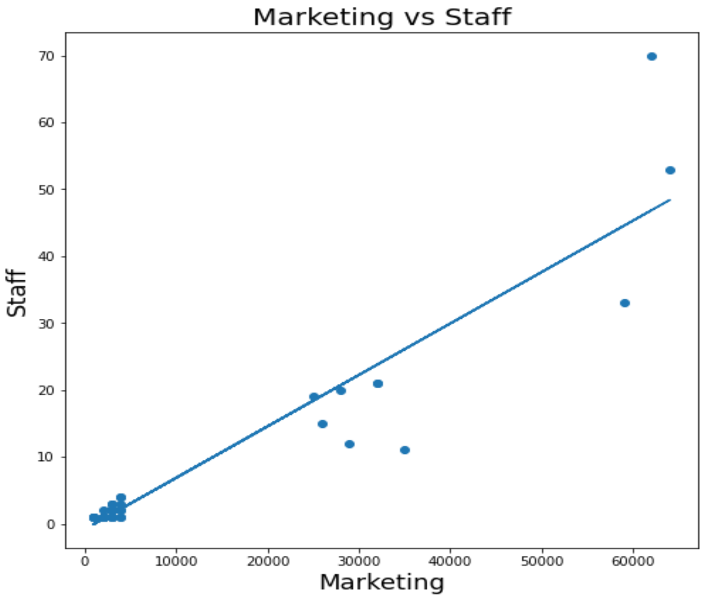
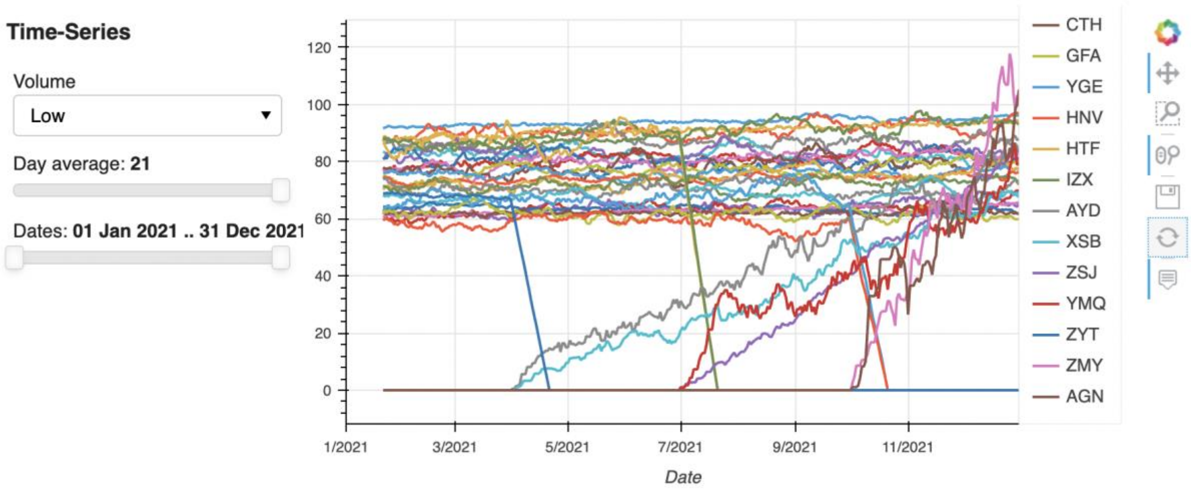


Figure 3. Scatter Plot for Staff on Marketing

The above visualisation is a scatter plot with a trend line for two variables Marketing and Staff respectively. Looking at the plot we can see that these two variables are positively correlated to each other which means if expenses on marketing increase so do the number of full-time staff of the outlets. Using Scatter Plot provides further analysis from the Heat Map that we used before, and it is obvious that we still got the same results. However, there is a data point that does not fit in the pattern and it is identified as an outlier, which explains that not always these two variables affect each other. The trend line is added to the graph to make the pattern even clearer. In our case, there is a positive trend line, which also indicated that there is a positive correlation between the two variables.

4. Line Graph

To get additional information about the customers who visited outlets for the year 2021, we used a Line Graph. A Line Graph or Line Chart is a way of displaying the data over some time. Interactive visualisation is needed because it enables users to manipulate and explore the data through various interactive features. Line Chart is very popular when depicting a pattern in data across a time series, so the lines are frequently drawn in chronological order. Line charts are commonly used to display trends over time or to compare the relationship between two variables. The line connecting the data points can reveal trends in the data, such as an increase or decrease over time or a correlation between two variables.



Time-Series

Volume: Low

Day average: 21

Dates: 01 Jan 2021 .. 31 Dec 2021

Figure 4. Line Graph showing outlets with low volume sales with weekly trends.

In Figure 4 we can see the Line Chart of the outlets having low-volume customers. The change in visitors during the year 2021 is displayed in the graph represented by a different colour as shown in the legend for each low-volume outlet. As this graph is interactive, there are some features we can use. Moreover, we can change the attributes value and the rolling average for each category and observe how the graph change according to the chosen setting. The rolling average in a Line Chart allows us to see the overall trend over time without being distracted by daily fluctuations. ZYT outlet has no activity after 5 months, and the IZX outlet has no activity after 7 months. XSB, AYD, ZMY, and YMQ are the newly opened outlets and have a high volume of visitors. Moreover, they also show seasonality because it increases with a fixed pattern on a particular period, while GFA, HNV, and YGE are the ones that have a low volume of customers during the year.

5. Box Plot

Box plots are a powerful tool in data visualisation that represents a dataset’s five-number summary. It consists of the minimum value, first quartile (Q1), median, third quartile (Q3), and maximum value. The distribution of data is shown by a box and whiskers. The box represents 50% of the data, which is the interquartile range (IQR). The interquartile range (IQR) is the difference between the first and third quartiles of the data. Q1 indicates the data's 25th percentile, and Q3 represents the data's 75th percentile. The whiskers in a box plot provide important information about the data's range and distribution, and they can be useful in identifying any observations that deviate significantly from the rest of the data. Any data point that extends beyond the whiskers is considered an outlier.

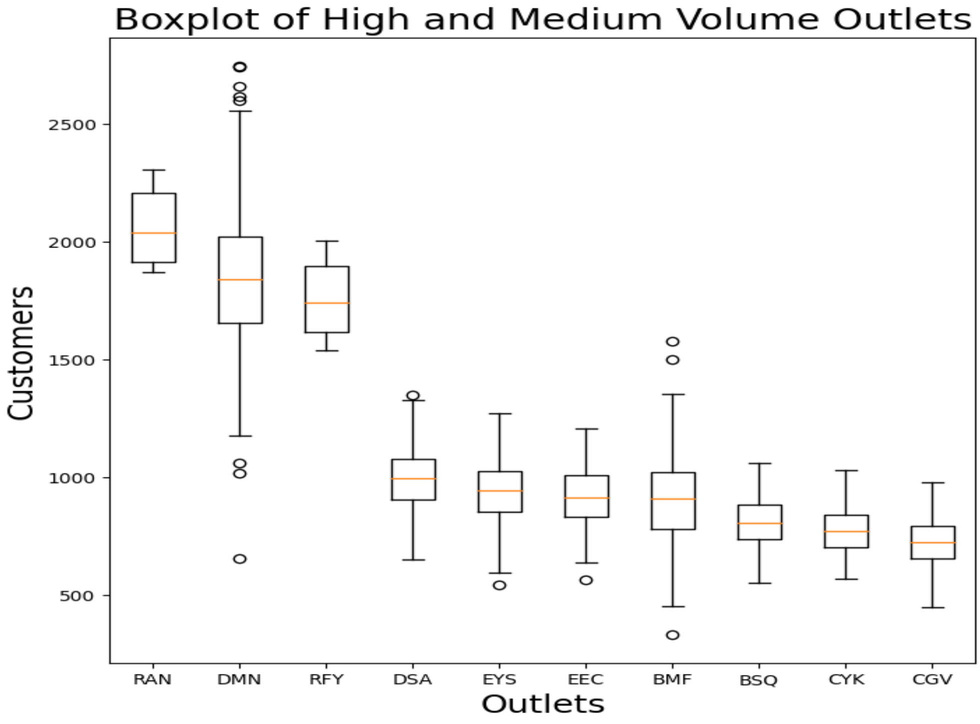


Figure 5. Box Plot showing outliers of High and Medium Volume Outlets

The above visualisation shows the distribution of high and medium-volume outlets.

The first three boxes show whiskers in big size among the others and share nearly the same median approximately 2000. DMN outlet has some outliers, dots at the upper limit of whiskers that indicate high values of the number of customers that visited that outlet, and low values to the third one. The remaining 7 outlets, which correspond to medium-volume outlets display different sizes of boxes and a median value of 600 – 1000 units, which is less than the range of high-volume outlets. There are outliers in almost all of them expect BSQ, CYK and CGV. BMF has outliers at the top and bottom.

6. Radar Chart

A Radar Chart, also known as a spider chart, can be used to display multivariate data in the form of a two-dimensional chart. This is a type of graph that illustrates comparisons of data groups and entities using distinct features and colour coding. In the Radar Chart, each axis represents a variable, and there are five variables in this case: marketing, customers, staff, size, and overheads. Each variable is represented by a line or curve that connects a series of points and radiates from a central point. This type of visualisation enables users to compare outlets and identify relationships and trends. A Radar Chart is included in this report because it is a powerful tool for communicating complex data in a clear and concise way, especially when dealing with more than three variables.

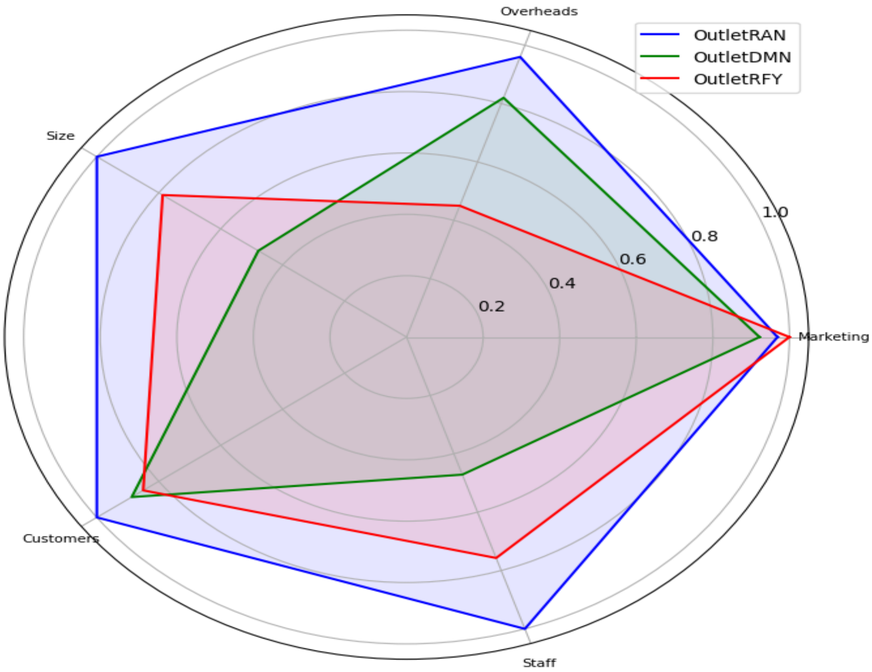
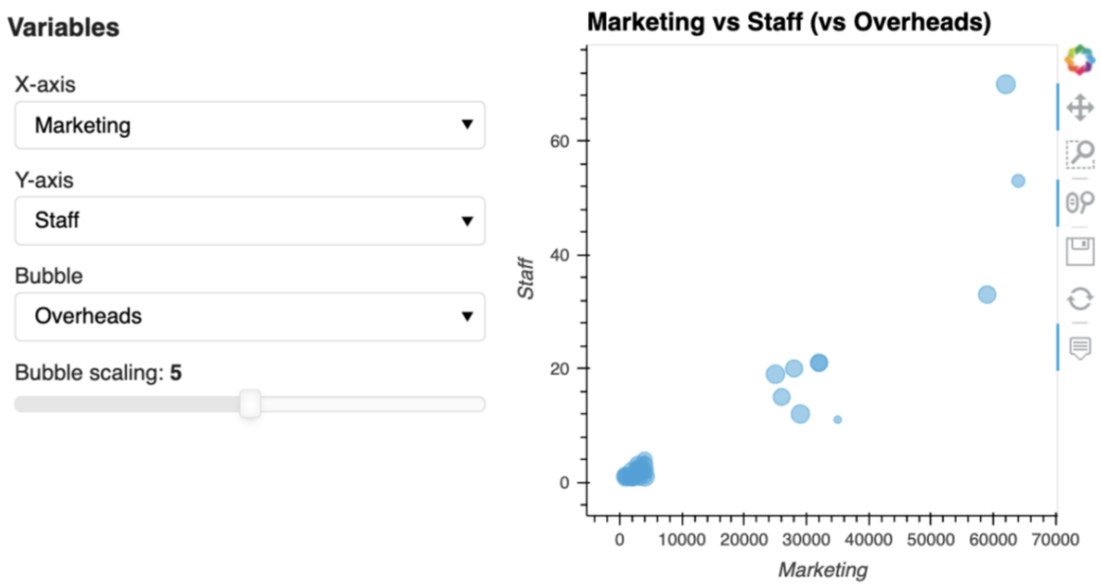


Figure 6. Radar Plot showing Outlets with High Volume of Customers.

In Figure 5, the Radar Chart shows the outlets with a high volume of customers throughout the year 2021. We have 5 variables used and 3 outlets and which of them has been defined a different colour for a better understanding. RAN is blue, DMN green and RFY red as seen in the legend to the right. In the graph, we can see that marketing and staff are positively correlated to each other as shown in the scatter plot and the heatmap in the above visualisations. We can also see a positive relationship between marketing and customers, as one increases so does the other. For the RFY outlet, we have less customers compared to other high-volume outlets and higher marketing costs. Overheads are less than other outlets, but the size is bigger than DMN and less than RAN.

7. Bubble Chart

A bubble chart is included to further analyse and demonstrate the relationship between the three variables. The first two variables correspond to the x and y axes, while the third determines the size of the bubble. We used scatter plots to identify relationships, but using a bubble chart adds another dimension of information to the chart, making it more appealing. A bubble is used to represent each individual data point. Overall, bubble charts are an effective way of displaying multiple variables and identifying relationships between them.



Variables

X-axis

Marketing

Y-axis

Staff

Bubble

Overheads

Bubble scaling: **5**

Figure 7. Bubble Plot showing Marketing vs Staff (Overheads)

Figure 7 demonstrates a bubble chart with three variables. Marketing is on the x-axis, and Staff is on the y-axis, with Overheads representing the size of the bubble. Once again, the Radar Graph demonstrates that there is a positive correlation between the variables, however, there are also some data points that deviate from the pattern of the data, known as outliers. As marketing increases, so does the bubble size. This is an interactive graph, so we can use other features or change the attributes to see what happens to the graph when we change the variables. Because of the coordination of the size of the bubbles, we have a better understanding of the relationships between variables and a very informative graph.

8. Histogram

A histogram is a useful data visualisation tool as it demonstrates the distribution of a dataset as a series of bars. Each bar's height corresponds to the number of data points in that bin, and the width corresponds to the range of values within that bin. We can adjust the bins' number and size, the bars' colour and style, and the axis label and title to customize a histogram. A histogram indicates whether the values in a dataset are concentrated around a limited range or are more widely dispersed. The reason why we are using a histogram to analyse the distribution of all outlets with low-volume customers is that it can provide a clear and concise summary of the data. We use a histogram to analyse the distribution of all outlets with low-volume consumers since it provides a clear and concise summary of the data. From a histogram, we can see where the peaks of the distributions are, whether the distribution is skewed or symmetric and if there are any outliers.

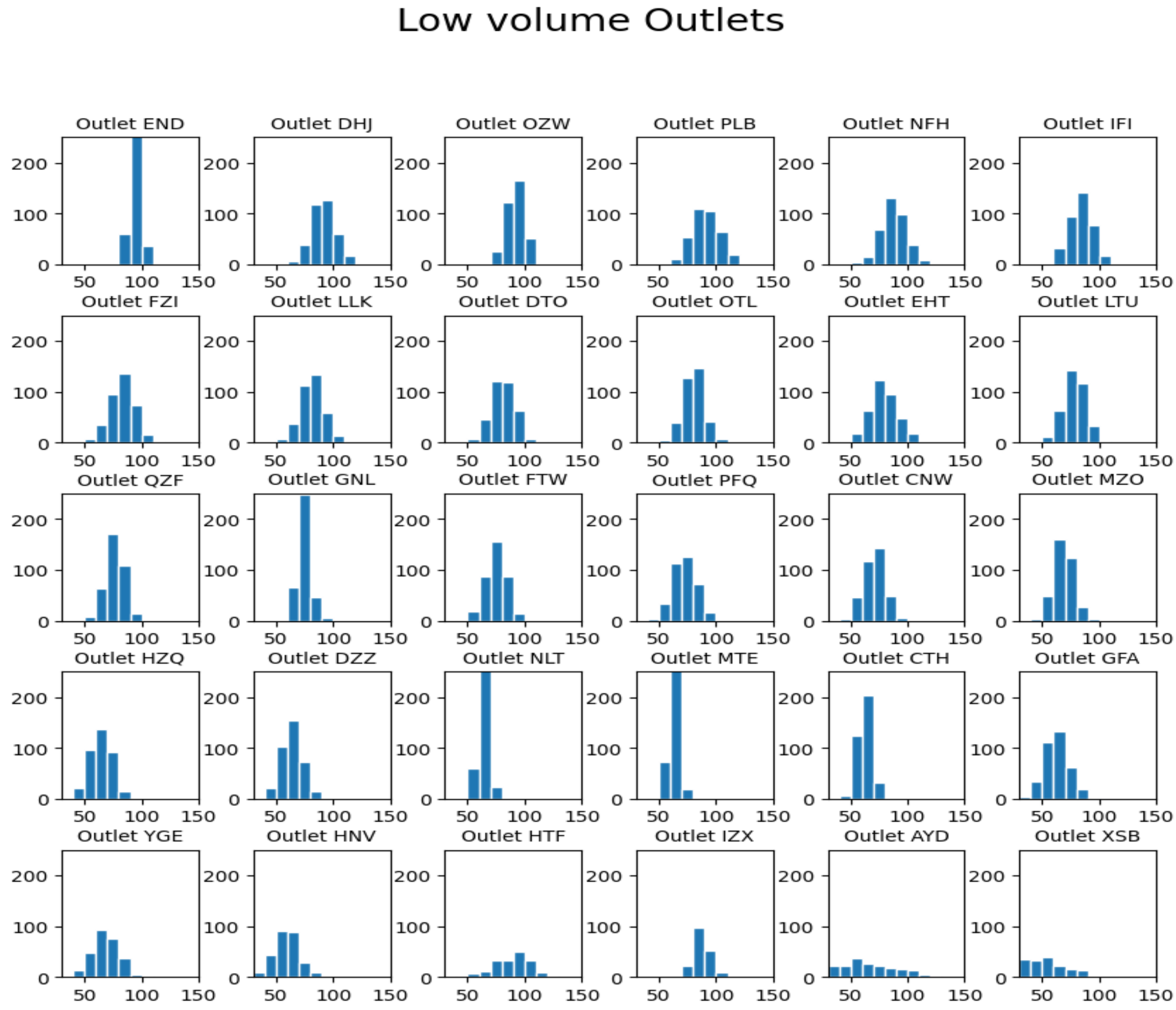


Figure 8. Histogram showing distribution of low volume customers for outlets.

Figure 8 Histogram depicts the customer distribution for low-volume outlets, and practically all of them have empty columns, as indicated in the graphs. The number of consumers increases until the middle of the year for the majority of the stores, then decreases. The distribution of most outlets is skewed and symmetric as well.

Outlets END, GNL, NLT, and MTE have all hit a peak at some point, which suggests that numerous customers have visited these outlets during that time period.

Critical Analysis

The COMP 1800 Data Visualisation module has provided an excellent introduction to the principles and best practices of data visualisation. This module has taught me the importance of communicating information visually and how to choose the appropriate graphs. I've also learned how to create various forms of visualisations that are clear, concise, and appealing, but what's more essential is that they provide useful information. I applied my knowledge in the coursework by creating visualisations aimed at demonstrating trends and patterns in data.

One of the best practices demonstrated in the coursework is clarity in data visualisation. The visualisation should be understandable and easy for the user to get the right insights and information from data. This can be achieved by the use of labelling and colour choices. Based on what I have learned in this module, I can now identify and define the basic concepts of visualisation, as well as evaluate which methods of visualisation are appropriate for different types of data. I have created a dashboard with interactive visualisations that allow users to adjust the attributes and receive real-time use cases. In conclusion, the data visualisations curriculum has given me a great set of skills and knowledge that I can apply to coursework and other applications. I can build successful data visualisations that help communicate insights and patterns from data by following best practices and focusing on clarity and simplicity. After successfully finishing the module, I am confident enough to pursue a career in this industry to improve my skills.

Conclusion

The dataset was analysed and several observations were made.

- The outlets were divided into high, medium, and low-volume categories.
- The data was found to be noisy, making it difficult to detect patterns and trends by using simple visualisations like line plots or bar charts.
- The analysis showed that some outlets were closed during the year, while others were newly opened, highlighting the need for a more detailed analysis.
- The data showed seasonality for low-volume outlets.
- The study found that the total number of customers, staff, outlet size, and marketing were strongly correlated. This means that higher marketing costs are linked to more customers per outlet, and larger outlets require more staff.
- No significant correlation was observed between overheads and the other four variables.
- Some outliers were identified in the data, mostly among low-volume outlets, but also in high and medium-volume outlets.

The coursework includes eight visualisations, two of which are interactive that allow the user to explore and analyse the data in a more detailed and personalized way. Interactive visualizations can provide a more comprehensive understanding of the dataset and enable more informed decision-making. In conclusion, the analysis shows the importance of comprehensive data analysis and appropriate data visualization techniques in gaining valuable insights into datasets.