Data Visualisation

Name: Raheel Shaikh

Contents

Table of figures

Abstract

In this report, we'll use graphics and data visualisation techniques to highlight the findings from data provided by a fictional firm named ChrisCo, which has a vast quantity of information about the number

of customer visits along with information about other attributes such as the overhead cost, the amount spent on marketing, the size of the store, and the number of staff who work for an outlet.

Introduction

The graphical depiction of data and information using graphs, charts, and other visual components is known as data visualisation. It is a crucial tool for communicating complex information to a variety of groups, including academics, executives from corporate organisations, decision-makers, and the general public.

We may identify correlations, patterns, and trends in data that may not be immediately clear from tables or text by using data visualisation. By visualising data, we may compare diverse datasets, find trends and outliers more quickly, and come to more precise and well-informed judgements.

Data visualisations come in a variety of formats, such as bar charts, line charts, scatterplots, histograms, and heatmaps. Each form of visualisation is appropriate for various sorts of data and inquiries.In many disciplines, including science, business, media, and public policy, data visualisation is becoming more and more significant. It is employed to disseminate research results, spot trends and business prospects, and assist in decision-making.

In general, data visualisation is a vital tool for making sense of complicated data and conveying ideas and conclusions to others. Data visualisation aids in decision-making and promotes change in the world by making data more approachable and interesting (Fayyad et al.,2001).

Visualisation:1 – Segmentation

The below bar graph in figure 1 shows us the daily customer visits for all the outlets of the company. The bar chart is an ideal choice in this senario since we can compare the number of customer visits in each outlet and segment the data in four different catergories.

a. ‘High Volume’- customer visits greater than 600,000.

b. ‘Medium Volume’ – customer visits greater than 200,00 but less than 600,000.

c. ‘Low Volume’ – customer visits greater than 20,000 but less than 200,000.

d. ‘Very Low Volume’ – customer visits less than 20,000.

We have three bars in green that represent the outlets with high volumes of customer visits. While the 7 orange bars make up for the outlets having a medium volume customer influx. The 24 orange bars in the middle represent the ‘Low Volume’ outlets. Whereas, the 11 black bars towards the end are marked for the outlets with "very low volume customer visits.

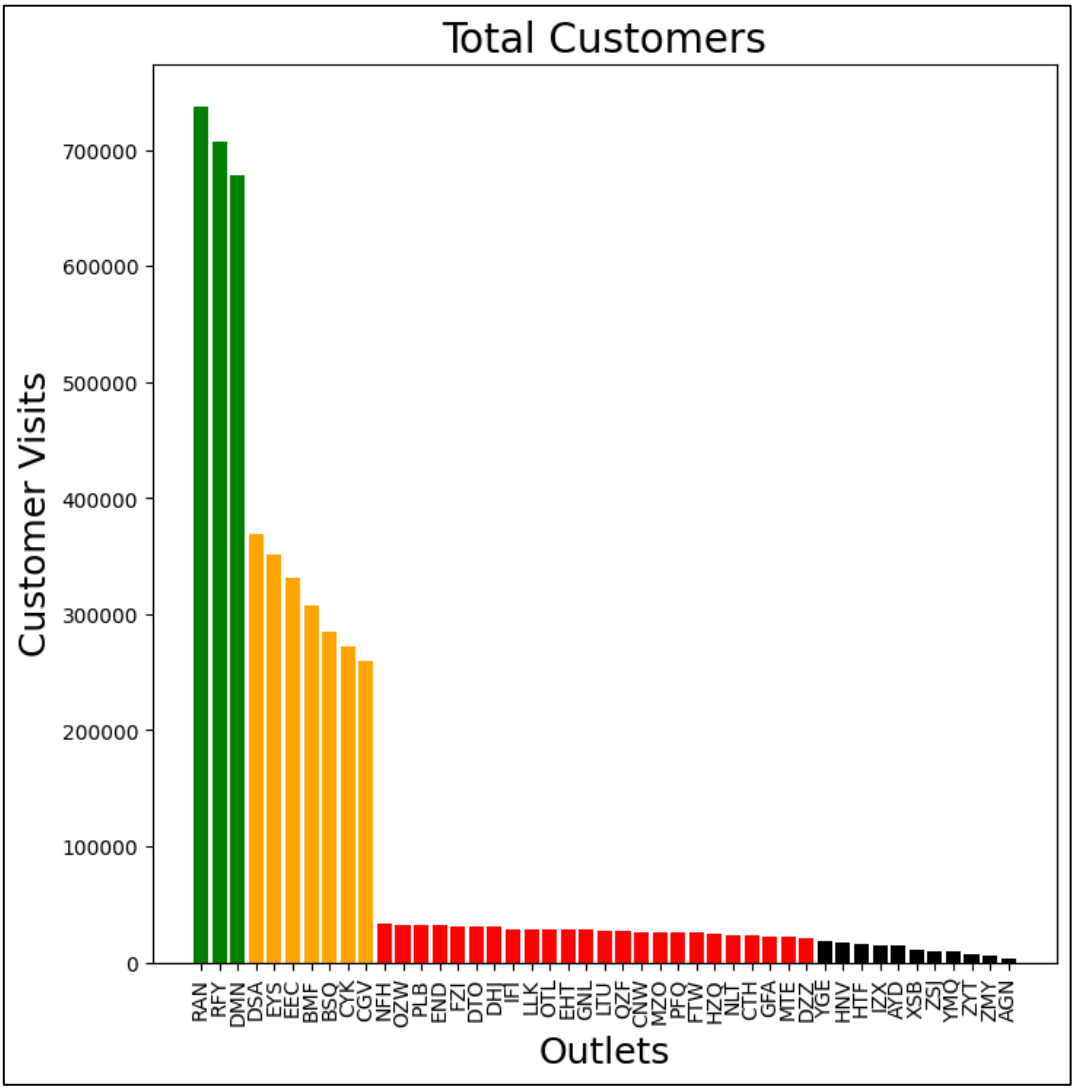

Figure 1: Bar chart showing customer segmentation.

Visualisation:2 – Store opened or closed

We must examine the quantity of customer visits for each date in order to locate the recently opened and closed stores. Line plots are the most effective visual representation for this time-series data. Since we have a catergory of outlets with very low volume, we may infer that there is a possibility that some of these outlets have just started or recently closed during the year, which would explain their extremely low number of customer visits.

From the figure 2 below we can see that there are 5 outlet for which the customer visits have dropped to zero and continue to remain the same for the rest of the year, these are the stores that have closed down during the year. We have stores YGE and HNV close down in the October 2021. While HTF and IZX close down in July 2021 and ZYT close down in April 2021.

There are 6 outlets that have a trend line starting for zero, this indicated that these store have opened up recently during the year. Outlet AYD and XSB have opened up in March 2021. Whereas, ZSJ and YQM were opened in June 2021. While ZMY and AGN have opened between October-November 2021.

Recently opened :- AYD,XSB,ZSJ,YMQ,ZMY and AGN

Recently closed: YGE,HNV,HTF,IZX and ZYT

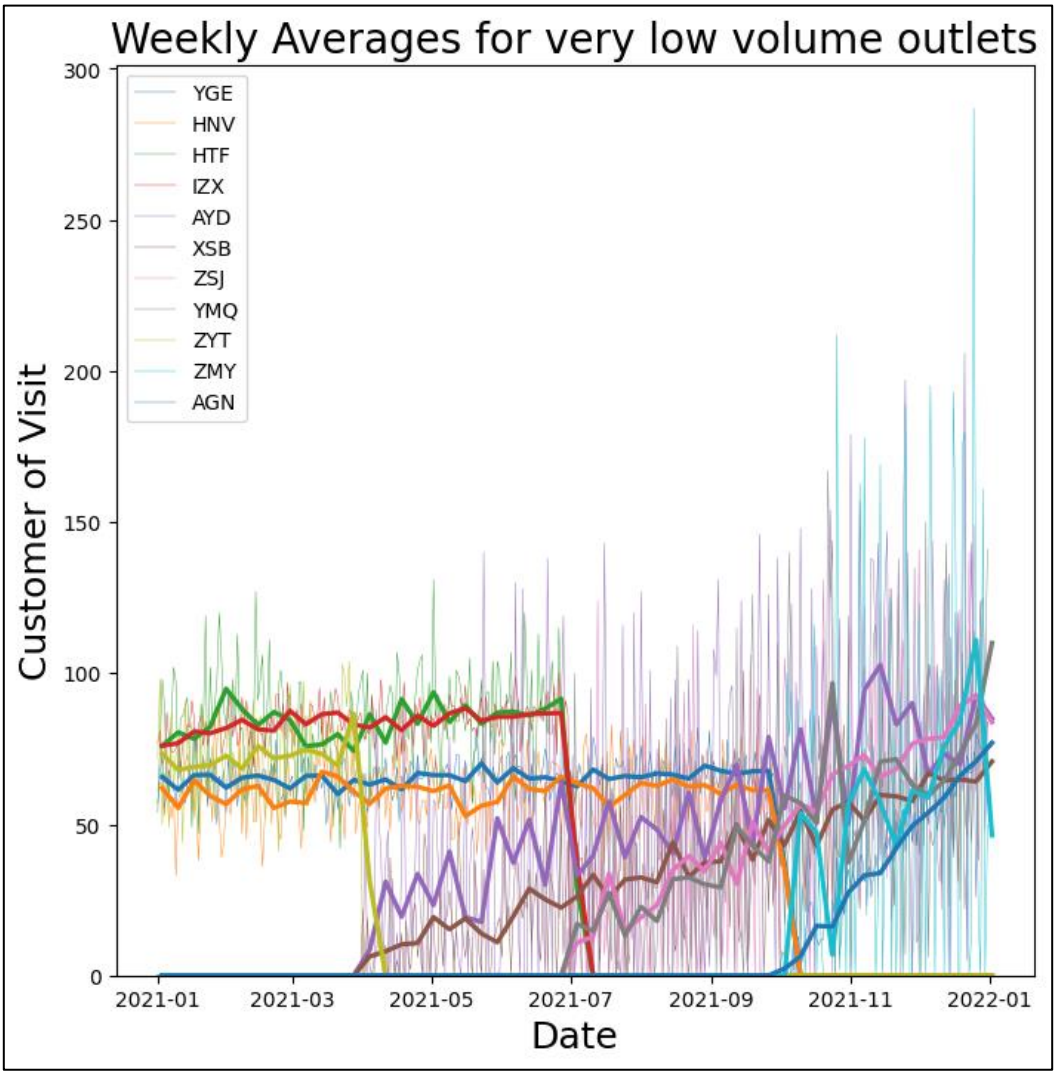

Figure 2: Weekly average for very low volume outlets.

Visualisation:3 – Dirstribution of customer visits over the year for ‘Medium Volume’ outlets.

Box plots are an effective tool for figuring out how a dataset is distributed and for contrasting different datasets. They offer a simple and consistent approach to view a dataset's most important elements and may be used to spot any peculiar or unexpected aspects.

The box plot in figure 3 shows us the distirbution of the customer visits for the medium volume outlets. The top line tell us the maximum number of cutomer visits a store had on a particular day whereas, the lower line tell us the minimum number of visits a store had. The line in the middle of the box represents the median value of the data. The dots show the outliers, or the unusual fluctuations in the number of visits for a specific days. This could be values that are either too low or high as compared to the rest of the data. We can see that stores DSA, EEC and BMF had outliers that were lower than the usual minmium i.e., DSA had two outlier of 708 and 699. While, EEC and BMF had values that decreased below 600 and 500 respectively. On the other hand BSQ had an increase in customer visits i.e., more than 1100 visits. Since it’s a interactive plot we can hover over it to get the value displayed in the graph.

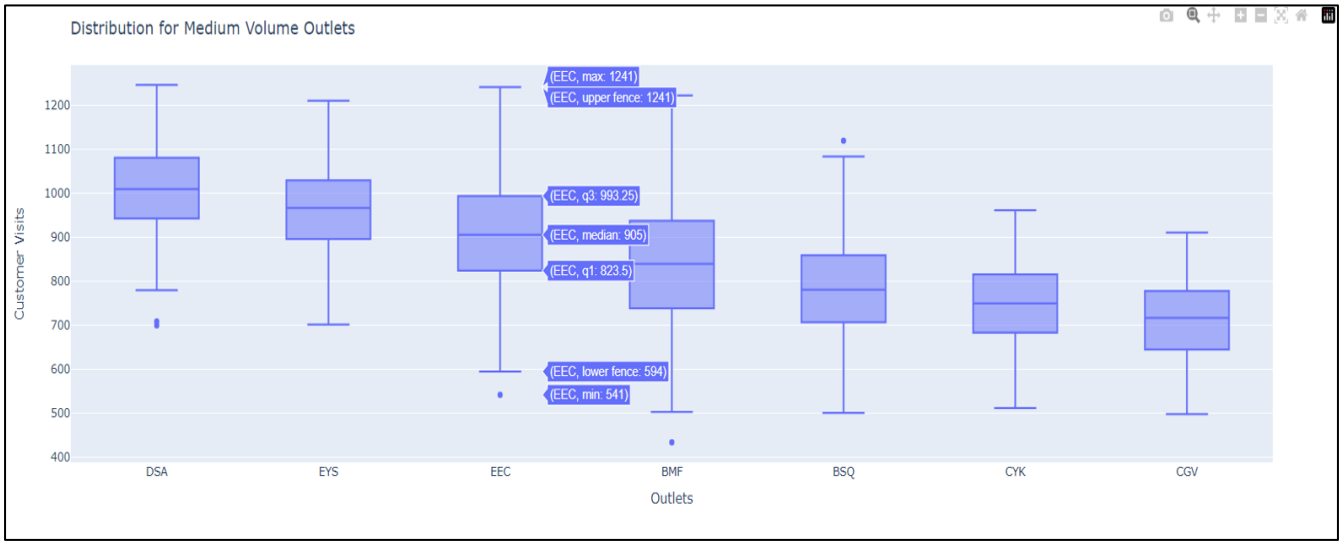

Figure 3: Interactive box plot for medium volume outlets.

Visualisation:4 – Heat Map

The relationship heatmap shows the correlation between the attributes in the summary data frame. The values on the heatmap range from -1 to 1. The index provides a colour's shade value. The dark red colour denotes a strong direct link and a +1. When the red shade fades, the value falls and reaches white, which represents a significant inverse proportion, at which point the value changes to -1.

In the below heatmap in figure 4, we can see a significant positive correlation between most of the attributes of the summary data. We have a very high correlation between marketing and the number of customer visits. Another positive correlation that is well highlighted in the heat map is the correlation between the number of staff and the size of the outlet.

Another intriguing phenomenon that can be seen is that 'Overheads, which represents the overhead cost of the store, has a negative correlation with all the other items attributed. This indicates that overhead is independent of all the other attributes.

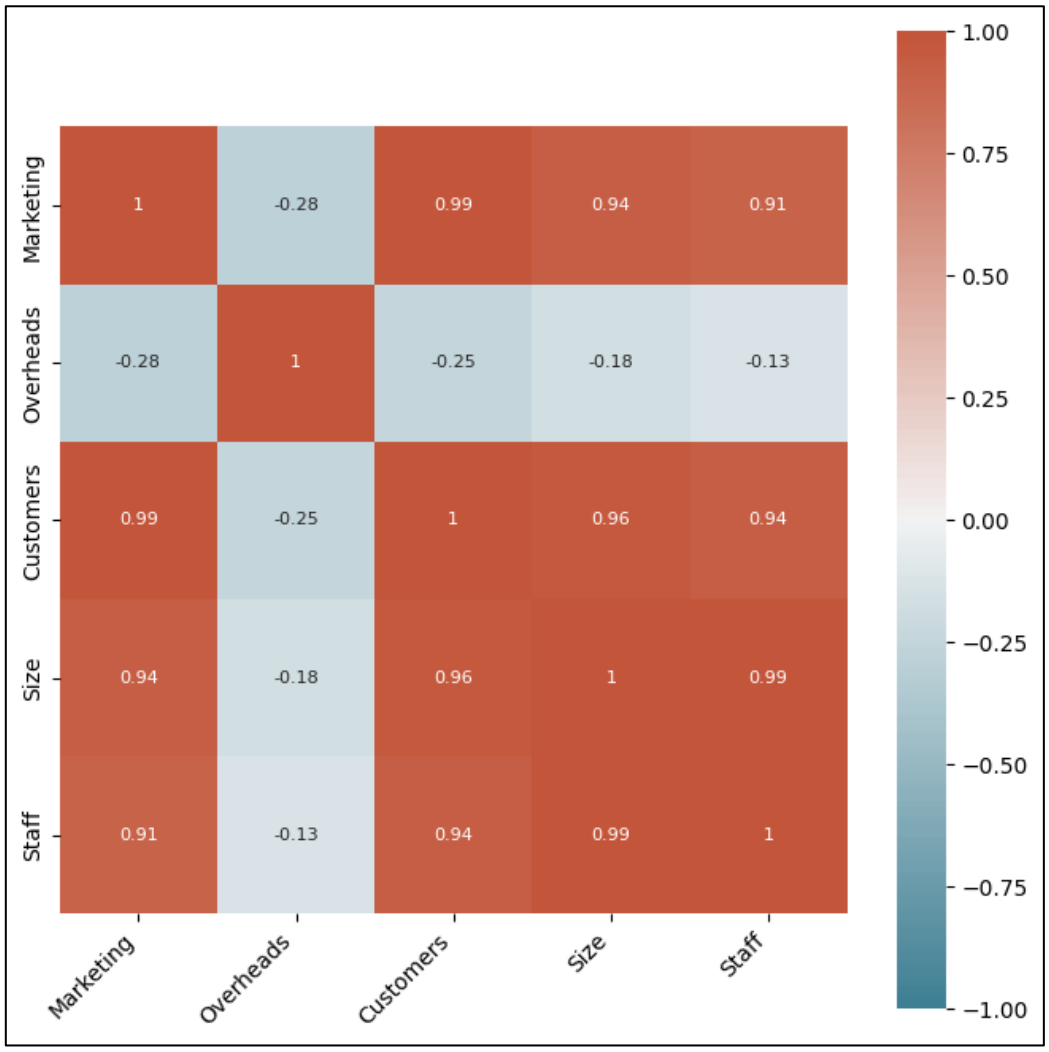

Figure 4: Heat map for summary data

Visualisation:5 -Investigating the correlation between the marketing and the customer visits.

We use the scatter plot with a regression line to check the relationship between customer visits and marketing costs. The regression line in the plot tells the direction and strength of the relationship between the two attributes. An upward line from left to right implies a positive correlation. While a downward line from left to right indicated a negative correlation between the variables,

The plot in figure 5 has four different categories plotted with different colours for each category. For example, blue is used to plot the high volume, orange for the medium volume, green for the low volume, and red for the very low volume outlet. We can see clearly from the graph that the regression line is an upward line that runs from left to right, and the scatter plots follow this line, so we can say that the two attributes are correlated to each other for all 4 categories. This indicated that as the amount spent on marketing increased, so did the volume of customer visits.

Figure 5: Scatter plot for customer vs marketing with regression line.

Visualisation:6 - Looking for anomalies/outliers.

Here we want to see if there are any insights in the customer visits along with overheads whilst at the same time looking at the marketing cost for the outlets. Since the bubble plot allows us to visualise three variables on a two-dimensional plane, it is an ideal choice in this case.

The plot in Figure 6 has overheads, i.e., the overhead cost for the outlets, on the X-axis, while the number of customer visits is on the Y-axis, and we use the bubble size to represent the amount spent on marketing. In our plot, we have four lines representing the overhead cost per customer visit. In our plot, we can see that RFY, DMN, and RAN have large bubble sizes. This can be due to the fact that they are high volume outlets, and we know that marketing is correlated to the volume of customer visits. It's important to note that RAN has a very high overhead cost as compared to the other two high volume store but spent almost the same amount on marketing as the other two. Almost all of the high- and medium-volume outlets have an overhead cost of less than £ 0.25 per visit. Besides, there is only one medium-volume BSQ that spends more than £ 0.25 on overhead.

The majority of the outlets with low and very low customer visits have high overhead costs of more than £ 2.00 per customer visit, but at the same time spend very little on marketing as well. This indicates a possible anomaly in this case. For instance, AGN has very high overhead costs but very low customer visits and spends very little on marketing; this can be considered an outlier.

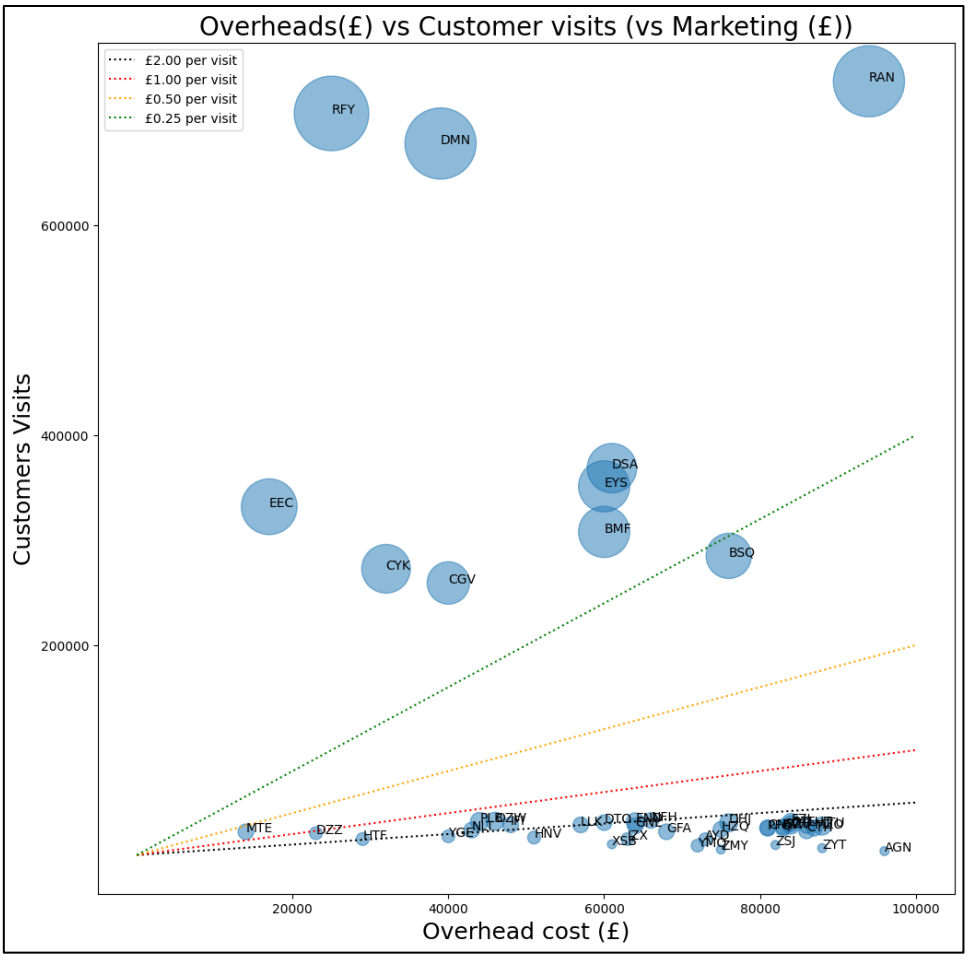

Figure 6: 3-dimension bubble plot (Overheads vs Customer Visits vs Marketing).

Visualisation:7 – Analysing the high volume outlets.

Here we make use of a radar plot, also known as spider charts, as we want to compare multiple attributes of the summary data; hence, a radar plot is one of the ideal choices since it provides a visual overview of multivariate data, enabling us to see trends, contrast variables, and acquire understanding of the connections between them.

The radar plot below in Figure 7 is for the outlets with high volumes of customer visits. We can see that the outlet RAN has the highest customer visits amongst the three plots. It is also the outlet with the highest overhead costs to run the store. Whereas, RFY is a little under RAN in terms of customer visits and is the outlet with the second highest number of customer visits. It is smaller in store size than the other two, while having less staff than the others and the lowest overhead costs. The outlet DMN is the store with the third highest customer visits but has the most staff compared to the other two. It spends about the same on marketing as RAN, which is the second-largest store among the three after RAN.

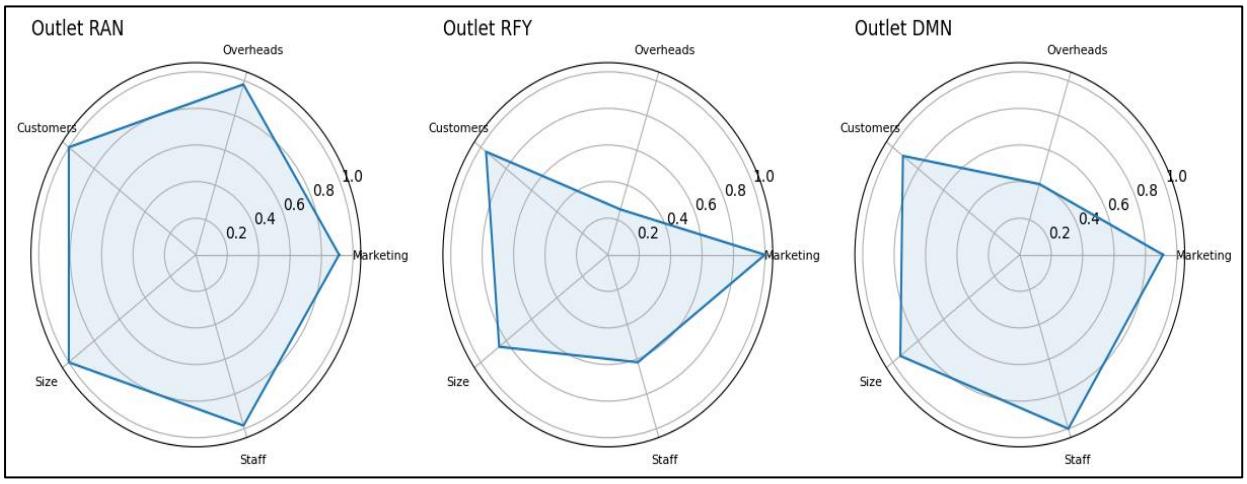

Figure 7: Radar plot for high volume outlets.

Visualisation:8 – Looking for any possible seasonality in the data.

In this case a bar graphs had been used to show a time series of visitor volumes. Since we have a large number of data point to be accommodated on the x-axis, we make use of bar plot rather than a line graph to depict the data.

This interactive graph in figure 8 displays the total number of customer visits across all the store on various days. The graph exhibits a regular trend, showing that during the middle of the week i.e., on Tuesday, Wednesday, and Thursday the stores receive high volume of customer visits, which then steadily declines over the weekend i.e., Friday, Saturday, and Sunday and then starts to rise once again on Mondays. This indicates a weekly seasonality in the customer visit pattern.

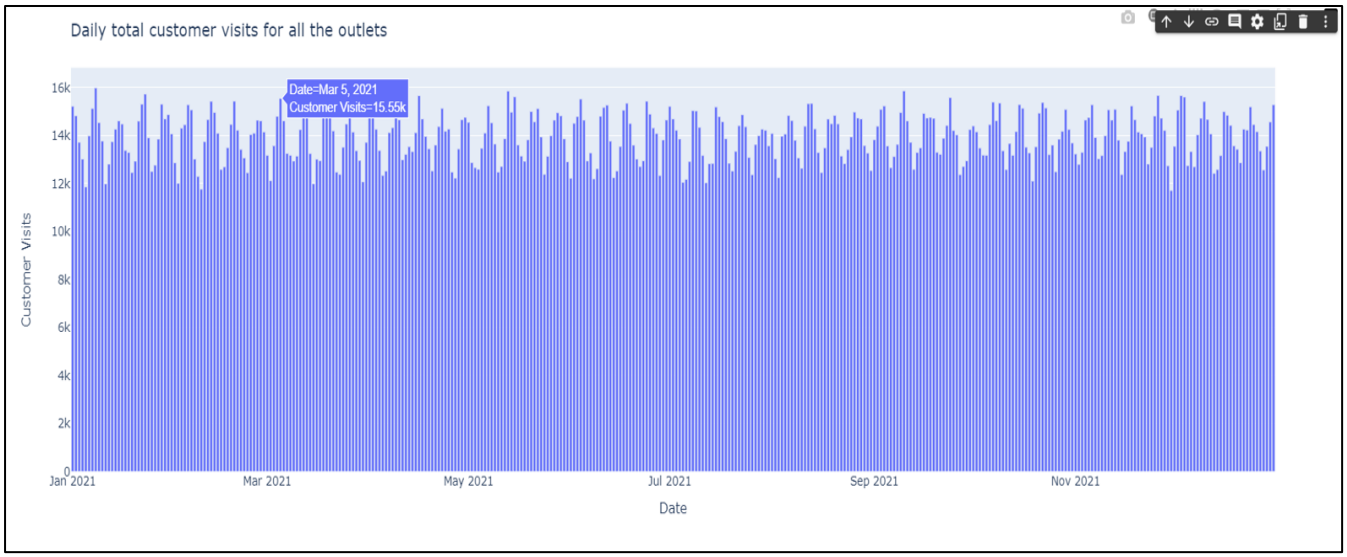

Figure 8: Interactive bar chart to depict seasonality.

Critical review

The Data Visualisation course has provided me with the knowledge of different methods for manipulating and analysing data through visually appealing graphs that can be readily comprehended by a wide variety of audiences. All of the data visualisations in the course were made using two well-known Python tools called Matplotlib and Seaborn, the code was written in an efficient manner by avoiding any form of redundancy in the program.

Also, I thought Plotly's interactive visualisations were really well-done. Plotly provides a variety of interactive elements for various plot styles, improving the usability and engagement of the visualisation process.

I also gained knowledge of the limitations of various visualisations, such as the inadequacy of line plots for showing a high number of data points. The necessity of choosing the appropriate plot style depending on the research topic and the data being examined was emphasised in the course. For instance, the heatmap used the varying colour shades to depict the correlations between numerous characteristics.

Overall, I thought the course on data visualisation was interesting and helpful, and it has really improved my knowledge of how to properly represent and evaluate data using visualisation methods.

Summary

After analysing the data provided by the Chrisco company and using different visualisation techniques we get some useful insight from the data.

- The data provided can be divide into 4 categories and the number of outlets in category is as follow:
  - 3 in ‘High Volume’ outlets.
  - 7 in ‘Medium Volume’ outlets.
  - 24 in ‘Low Volume’ outlets.
  - 11 in ‘Very Low Volume’ outlets.
- The 3 high volume outlets are ‘RAN’,’RFY’ and ’DMN’.
- The outlet RAN has the highest number of customer visits amongst all the outlets.
- The company had a number of outlets that opened during the year and a few outlets that closed down. There was a total of 11 outlets and these outlets belonged to the ‘Very Low Volume’ category and hence this justifies them being the outlets with very low customer visits for the year.

  Recently opened :- AYD,XSB,ZSJ,YMQ,ZMY and AGN

  Recently closed: YGE,HNV,HTF,IZX and ZYT

- The customer visit data consists of outliers. For instance in the medium volume outlets 4 of the store (DSA, EEC, BMF, and BSQ) have outliers in their data.
- The data set displays a significant correlation between most of the attributes of the dataset. We got to see a very high possitive correlation between the the money spent on marketing and the customer visits. Thus, implying that higher the cost spent on marketing the more customers visit the store.

- It was also noted that the overhead cost had a negative correlation with most of the other attributes. This indicates that the overhead amount spent by an outlet is independent of the other attributes.
- Most of the low and very low outlets have a very high overhead cost inspite of having less staff and spending less on marketing.
- The customer visit data displays a weekly seasonality. With the number of visits highest during the middle of the week and dropping over the weekend.

References

Fayyad, U. M., Grinstein, G. G., & Wierse, A. (2001). *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann.