COMP 1800

DATA VISUALISATION

Coursework

Name: Fariha Tasnim Khan

Student ID: 001249887

Introduction to Data Visualization

The representation of data using graphs, maps, and other visual tools is known as data visualization. The amount of data being generated globally from a variety of sources is rising dramatically. Every minute, data is being generated from just a click online. Large amounts of data being produced every day is increasing the importance of data visualization. With data visualization large amounts of information can be seen in one glimpse. This makes it easier for the viewer to spot patterns and trends in data which can be challenging to do when large amounts of data are presented as rows and columns. With the found trends and patterns, better future decisions can be taken and also with the help of visualizations performance can be monitored regularly. For example- Line chart show trends and the changes that took place over time. From such a plot, viewers can discover that there was a fall is sales at a particular month and an increase in sales at a particular month. With this information, they can research more and find the causes of such an increase or decrease. On the other hand, in a dataset, there are hundreds of columns and rows, and it is very hard to identify such patterns by just looking at it. In addition, from visualizations such as of box plot, viewers can find outliers which can be used to identify why a value is outside the normal range of values. Therefore, visualization can be used to monitoring, identification and decision making.

There are many ways of visualizing a dataset. But before visualizing a dataset it is first important to find the purpose which means what insights is expected to be found from a visualization. This is necessary because different plots show different information. For example- If the purpose is to find the shop outlets with the highest number of customer visits, then implementing a heat map is not going to fulfill the purpose. Hence in order to find the correct information, the correct visualization needs to be implemented.

VISUALIZATIONS:

Visualization 1: Bar Chart

The first visualization that I implemented is a Bar Chart. Bar chart shows data in the form of rectangular bars whose lengths are proportional to the values they represent. I have used the dataset of customer visits to ChrisCo's outlets for implementation of the bar chart. The purpose of this visualization is to find out which outlets have a high, medium, and low volume of customer visits.

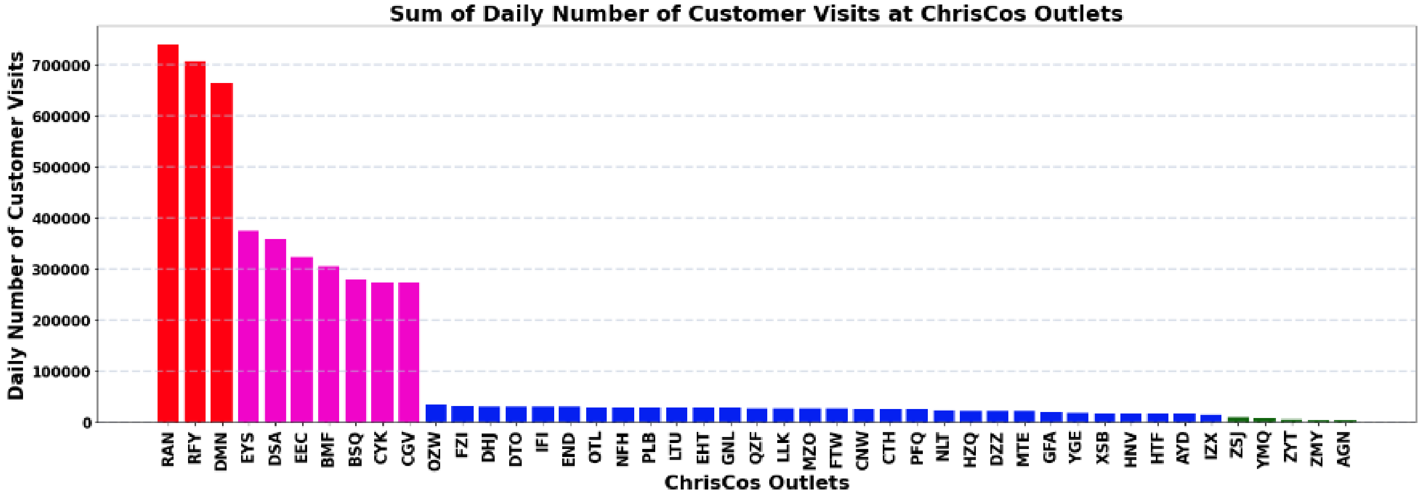

Figure 1: Bar Char showing sum of daily number of customer visits at ChrisCo's outlets.

Justification:

I have used a bar chart because it is ideal to show and compare each category in a dataset. In this case the outlets are the categories, and the height of the bar represents the number of daily customer visits to ChrisCo's outlets. Each category has different heights based on the number of customer visits to the outlet and in order to further categorize the outlets in terms on high volume, medium volume, low volume and very low volume of customer visits I have used different colors. I have used the red color for showing high volume, the pink color for medium volume, the blue color for low volume and the green color for very low volume. I have also sorted the dataset, to display all the bars in a descending order. Therefore, using bar chart I can visually show a comparison between all the outlets number of customer visits and also differentiating it in different categories of volume using colors. So, the purpose of using a bar charts is to do a comparison between the outlets of ChrisCo highlighting the different volumes of customer visits to outlets using a variety of colors.

Description:

From the above bar chart in figure 1, outlets with different bar heights have been categorized with different colors. The bars greater than 400000 are categorized as the outlets with a high number of customer visits and is colored in red. The bars greater than 200000 are categorized as the outlets with a medium number of customer visits and is colored in pink. The bars greater than 8914 are categorized as the outlets with a low number of customer visits and is colored in blue and finally the bars who are even lower are categorized as the outlets with a very low number of customer visits and is colored in green.

It can be seen in figure 1 that there are 3 bars with red color. These 3 bars represent the outlets with a high number of customer visits, and they are: “RAN”, “RFY” and “DMN”. Similarly, there are 7 bars with pink color. These 7 bars represent the outlets with a medium number of customer visits, and they are: “EYS”, “DSA”, “EEC”, “BMF”, “BSQ”, “CYK” and ‘CGV”. The rest of the outlets colored in blue, and green are the outlets with a low volume of customer visit.

Visualization 2: Line Plot with Monthly Averages

The second visualization that I implemented is a line plot with monthly averages. The plot displays the average value of the high and medium number of customer visits to ChrisCo's outlets over a year from 2021-01 to 2022-01. The average values are shown in thick lines in different colors which represent different outlets. It is found in figure 1 bar chart, that there 3 outlets with high volume of customer visits to outlets and 7 outlets with a medium volume of customer visits. I have selected these 10 outlets and plotted a line plot to find in detail about its trend and patterns.

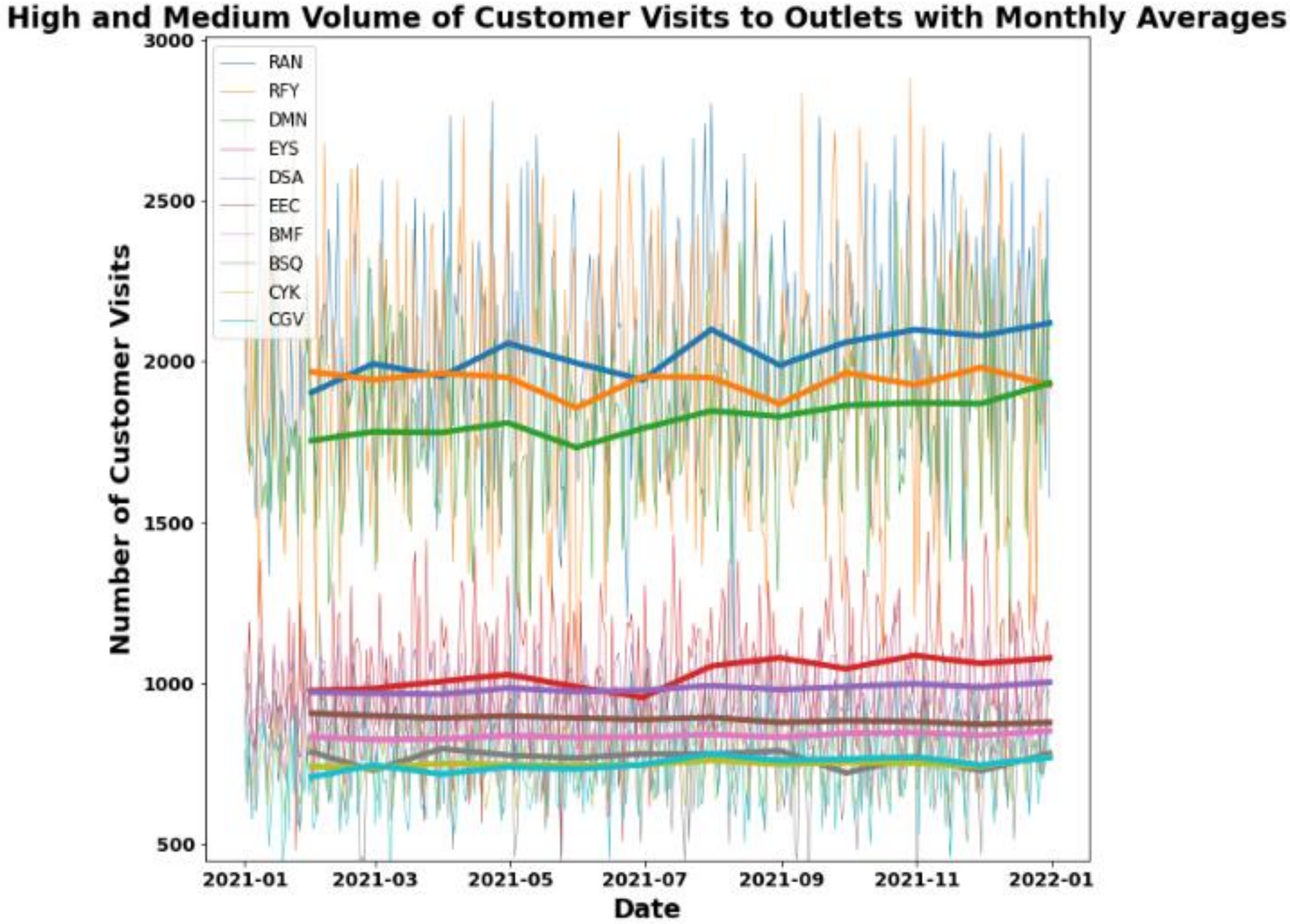

Figure 2: Line plot of high and medium volume of customer visits to outlets with monthly averages.

Justification:

A line plot is a type of plot where individual data points are connected by a line. A line graph shows quantitative data over a specified time frame. (PETERS, 2022) The reason why I chose a line plot to show the outlets with high and medium volume visits is because with a line plot, the changes in data over time can be seen. Here, the changes in visits can be seen from 2021 to 2022. Finding changes in data over time is important as with this, we can find trends and patterns, from which it is possible to identify why at a particular month the visit is low, which may be due to seasonality or a change in customer behavior. Thus, it will help in decision making.

The line plot I have implemented also includes monthly averages, which helps to identify trends and patterns on the monthly averages of the selected outlets. The reason why I chose to implement monthly averages in the line plot is because it allows to see the overall trend of a month, which gives an overall idea of the number of visits every day. This is much simpler to find trends and patterns on, than a plot which displays data points of every day for one year. In addition, monthly averages also eliminate some of the noise that may exist in the dataset.

Description:

The top 3 thick lines that have data points greater than 1500 are the outlets with a high volume of customer visit and these thick lines represent the average number of visits. Starting from the top, the blue line represents the outlet “RAN”, the orange line represents the outlet “RFY”, and the green line represents the outlet “DMN”. The outlet “RAN” has some up and downs till 2021-09 after which it has an increasing trend and increased over 2000 visits in 2022-01. The outlet “RFY”, has a constant rate till 2021-05 after which it has some up and downs till 2021-12 and then has a gradual decreasing trend. The outlet “DMN” has a slight fall in 2021-06 but has an overall increasing trend.

The 7 thick lines with data points greater than 500 and less than 1500 are the outlets with a medium volume of customer visits and these lines represent the average number of visits. The medium volume outlets are “EYS”,” DSA”,” EEC”,” BMF”,” BSQ”,” CYK” and” CGV”. All the medium volume outlets have an overall constant rate. Amongst which the outlet “EYS” have a slight increase to over 1000 visits in 2022-01.

Visualization 3: Interactive Line Plot

Figure 2 is an interactive line plot which shows changes in data points over time with additional features of interactivity such as zooming, hovering, etc. The purpose of implementing an interactive line plot for outlets with low volume of visits is to identify if any outlets closed or any new outlets opened.

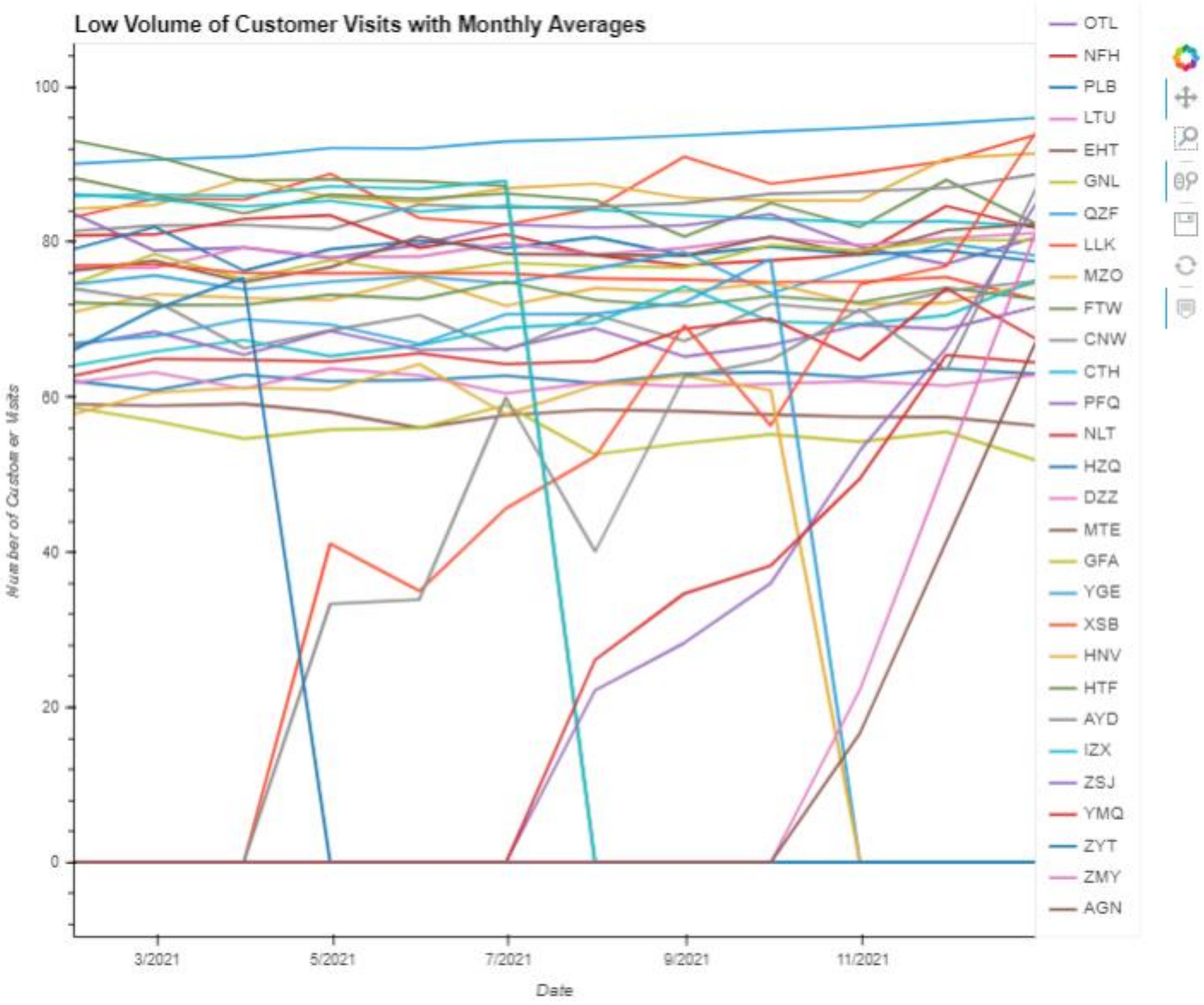

Figure 3: Interactive line plot for low volume outlet visits with monthly averages.

Justification:

As mentioned above, line plot is ideal to show changes over time. Figure 3 is an interactive line plot and interactive line plots enables the user to take several actions, such as zooming, hovering, selecting a variable to display, and more. This enables the viewers to explore the data more effectively to identify trends and patters which may not be possible to spot in static plots. For example, figure 2 is not an interactive plot. If the mouse is hovered on top of the plot no details such as outlet name or exact date of the data point can be found. Instead, the outlet names have to be found by identifying the colors which is hard to identify when there are too many lines in a single plot. But with interactive plots, even if there are many lines in a plot, each of the line's information can be identified by just hovering the mouse on top the lines. In addition, for a clearer view, the zoom feature can be used which will help zoom in and see all the trends and patterns in details. These features save time when finding trends and patterns from plots. One of the reasons of using the interactive line plot is because there are many outlets with low volume of visits and it is hard to find the patterns of each of the outlet with a static line plot, as identification of so many different colored lines is difficult, time consuming and may lead to mistakes.

Description:

The line chart in figure 3 shows all the low volume visits outlets, amongst which there are outlets which have been identified as closed and newly opened. All the outlets have different colors and can be identified by hovering on top of the lines to know about the outlet name and the exact date. There are 5 outlets which have been closed and they are: "YGE", "HNV", "HTF", "IZX" and "ZYT". If the mouse is kept on the dark blue line situated on the very left side of the plot, it can be seen that this line represents the outlet "ZYT", which had outlet visits of 75 on 2021-03-31. However, the outlet "ZYT" had 0 outlet visits on 2021-04-30 and it continued. Thus, it can be said that the outlet "ZYT" closed. Similarly, the outlet "IZX" represents the sky-blue line and had 87 outlet visits on 2021-06-30 but then it closed on 2021-07-31 which can be said because it had 0 outlet visits from this date. Behind the line of the outlet "IZX" is the line of the outlet "HTF" and this outlet closed on the same date on 2021-07-31. If the name of "IZX" is clicked on the legend, then it is removed from the plot which enables to see the line of "HTF" clearly. The outlet "HNV" is represented by the yellow line and this outlet also closed on 2021-10-31. The last outlet that closed is "YGE" and this outlet also closed on the same date as "HNV" which is on 2021-10-31.

There are also 6 outlets which have opened. If the interactive plot is viewed from the value 0 in y axis, it can be seen that there are 6 lines which rose to high number of visits from 0. These are the outlets which can be identified as newly opened, and they are: "ZSJ", "YMQ", "ZMY", "AGN", "XSB" and "AYD". From the very left side of the plot, the orange line represents the outlet "XSB" and it can be seen that on 2021-03-31, the outlet opened, and the visits rose to 69 on 2021-08-31. On the same day as "XSB", the "AYD" outlet also opened which is shown by a grey line. The line red and purple represents the outlets "YMQ" and "ZSJ" and these two outlets opened on the same day on 2021-06-30. The last two outlets "ZMY" and "AGN" also opened on the same day on 2021-09-30.

Therefore, from this interactive line chart, 5 outlets that closed and 6 outlets that opened during the year has been identified.

Visualization 4: Interactive Heat Map showing Correlations of Outlet Visits

Figure 4 is an interactive heat map which shows correlation between the different outlets number of visits by customers. Correlation shows how much two variables vary in respect to one another and is measured on a scale from + 1 to -1. This is an interactive heat map which has additional features of interactivity such as zooming, hovering, etc.

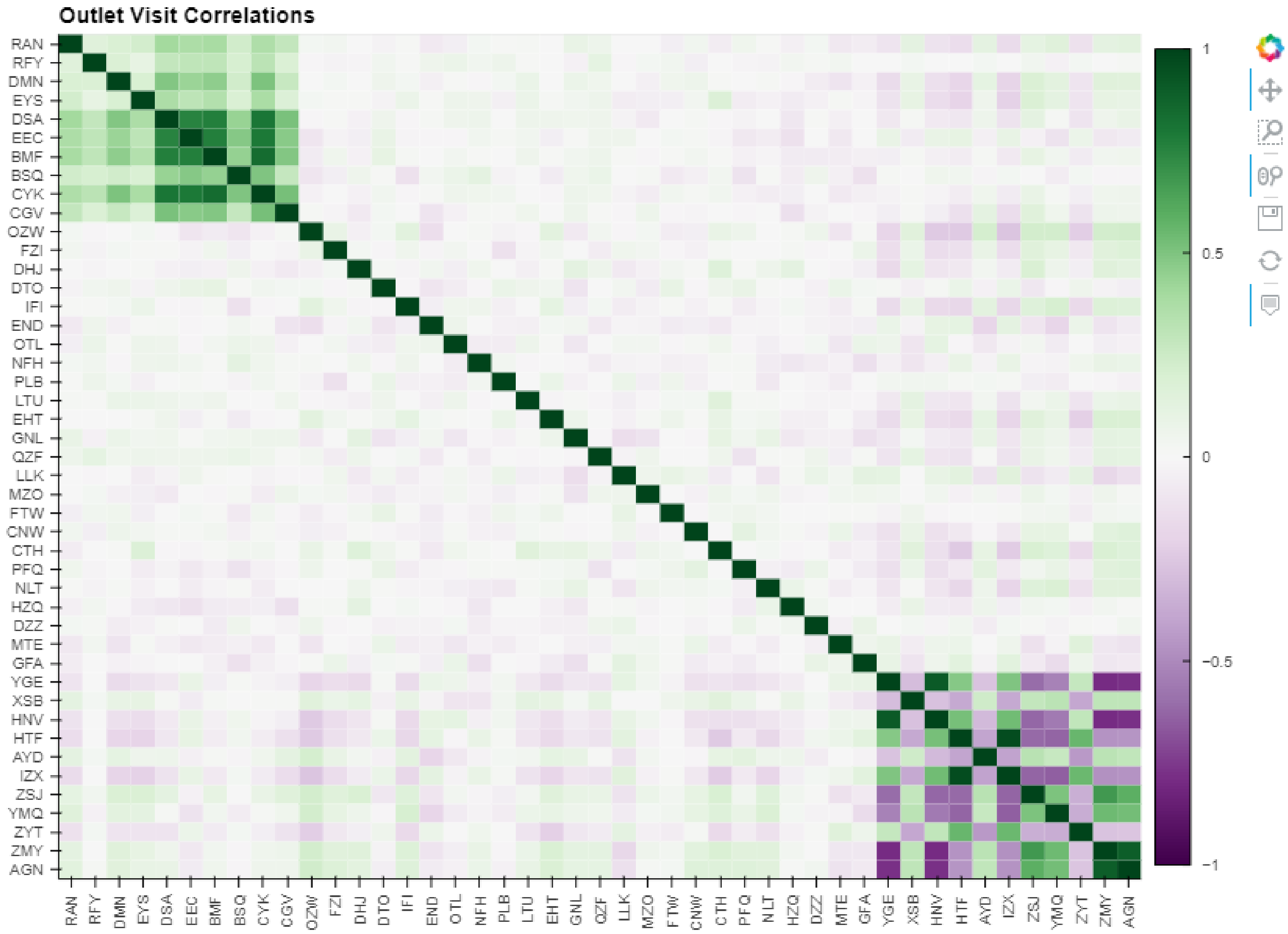

Figure 4: Interactive heat map showing correlations of outlet visits by customers.

Justification:

For implementing the heat map, I have used the dataset which contains information regarding customer visits of the different outlets. There are 45 outlets in this dataset which shows that the dataset is very big and is not ideal to just use a static heat map to represent the correlation values. The interactive heat map that I implemented in figure 4 contains many cells, but correlation information can still be extracted from the heat map by hovering a mouse, by which the name of the outlet and the value of correlation can be seen for each cell. Viewers can also zoom in and out for a clearer view of the cells. This is not possible in a static heat map. If this dataset is used in a static heat map, it will result into too many cells and too many correlation values. It is not possible to view each of cells and its values even if the figure size is increased as there is no feature of zooming in and hovering to find value of each cell. In addition, in such a big plot, it is difficult to figure out the column and row name. Hence, this is why I have chosen an interactive heat map for this dataset as it offers more options for exploration and engagement.

Description:

The heap map consists of values on a scale from + 1 to -1 and each value is represented using different colors. The color scale is given on the right side of figure 4, where the dark green color represents the correlation value greater than 0.5, and the dark purple color represents the correlation value less than 0.5. A correlation is considered as a strong positive correlation when the value is greater than 0.75. It can be seen that there is a total of 19 dark green colored cells with a positive correlation value of greater than 0.7. (ZACH, 2020) This means that the sets of variables of these 19 cells move together in the same direction. The outlets with the highest positive correlation value are “IZX” and “HTF” of 0.96654. Therefore, it can be said that when the number of visits of the outlet “IZX” rises, the visits of “HTF” also increases.

Similarly, the values that are less than 0 have a negative correlation. There are 8 dark purple cells which represents the sets of variables with a strong negative correlation of value greater than -0.7. This indicates that if one variable rises, the other will decrease, and vice versa. The outlets with the highest negative correlation value are “HNV” and “ZMY” of -0.79107. Therefore, it can be said, that if the number of visits of the outlet “HNV” increases, the visits of “ZMY” falls.

It is found earlier in figure 3, that there are some outlets which closed and some outlets which opened. Amongst these outlets were “HNV” and “ZMY”, where the outlet “HNV” closed and the outlet “ZMY” opened. Therefore, it can be said, that because one outlet closed and the visits dropped to 0, the other outlet which opened faced an increase in visits. This proves the negative correlation between the outlets.

Visualization 5: Heat Map for Summary Data

A statistical measure called correlation shows how much two variables fluctuate in relation to one another. A scale from + 1 to -1 is used to calculate the correlation coefficient. In order to find the correlation between each of the two variables in ChrisCo's summary data, I have used a Heat Map. The summary data consists of the marketing and overhead costs for each outlet, the outlet sizes and the outlet staff employed. In this heat map in figure 5, the darkest shade of color represents a positive correlation of +1 and the lightest shade of color represent a negative correlation of -1.

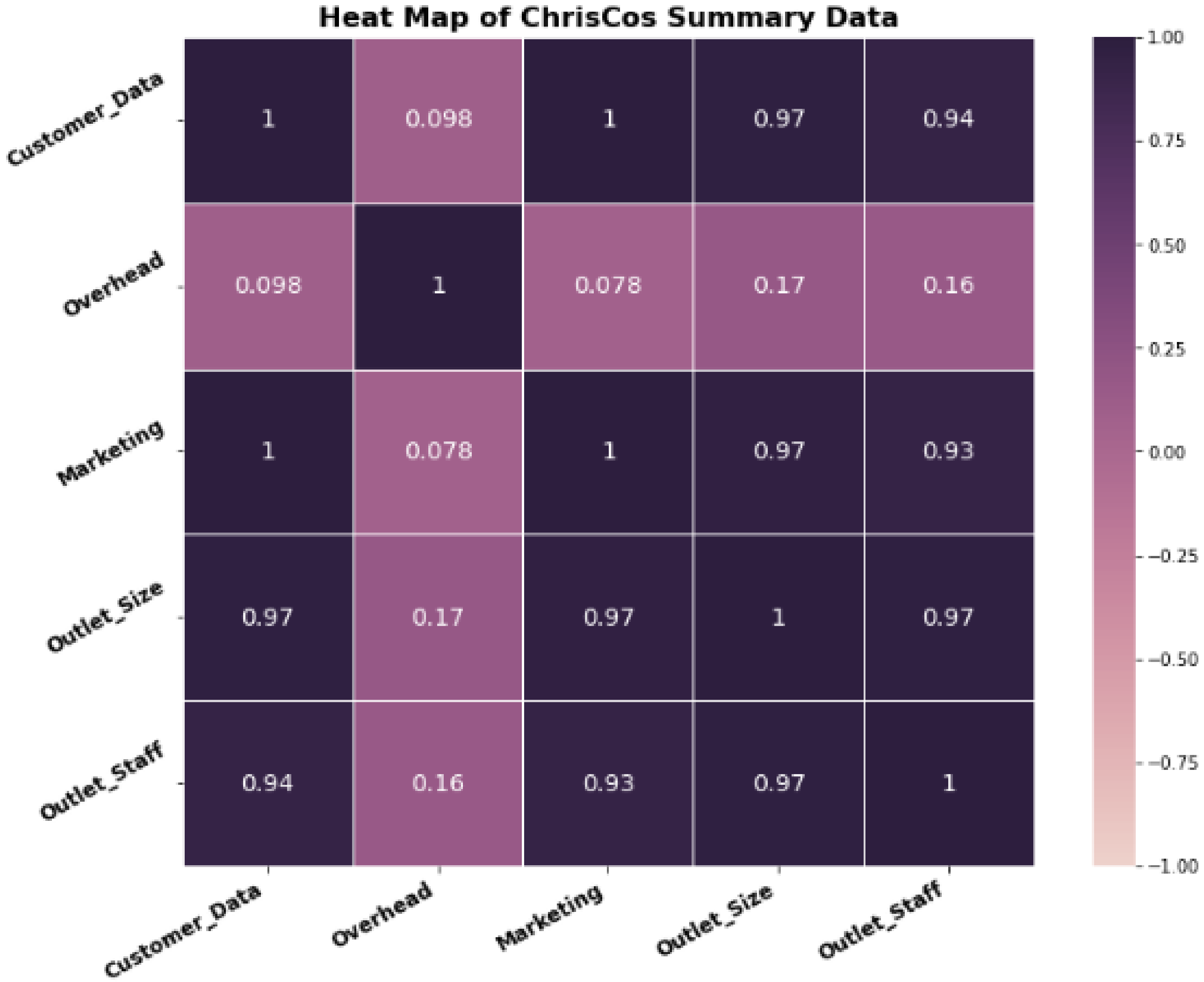

Figure 5: Heat Map of Summary Data.

Justification:

I have used a Heat Map as it is ideal to show large amounts of data in the form of a diagram where data values are represented by colors. Heat maps use different shades of colors to represent relationships between two variables where the darkest shade indicates a positive correlation, and the lightest shade indicates a negative correlation. The variables are mentioned in x and y axis. Due to the usage of different colors in heat map, identifying patterns and strong and weak correlation is easy to find with just a quick glance. (Kumar, 2022)

Scatter plots also serve the same purpose of finding correlation between two variables. However, the reason why I did not choose a scatter plot is because, in a scatter plot viewers need to visually inspect the points and look for patterns to find the correlation between two variables, which is time-consuming. However, a heat map shows the value of correlation for each of the two variables in one single plot with different colors, which makes it easier to identify which variables of a dataset is positively or negatively correlated. This makes the identification process faster without having to interpret the plotted points. Thus, this is why I have chosen a heat map for the purpose of finding correlation.

Description:

In the above heat map in figure 5, all the correlation values have been shown with different intensity of colors. The dark colors represent a strong positive correlation, and a medium dark color represents a weak positive correlation. Just by looking at the heat map, it can be said that most of the cells have a dark color which indicates that most of the variables have a strong positive correlation with each other.

It can be seen in figure 5, that the variable overhead has a weak positive correlation with all other variables. A weak positive correlation is considered if the correlation value is between the range of 0.25 to 0.5. As all the values of the variable overhead lie within this range, it can be said that the variable overhead has a weak positive correlation with all other variables. This means that if the other variables increase, the overhead variable increase as well but in a relatively weak manner. (ZACH, 2021)

A strong positive correlation is considered when the correlation value is greater than 0.75 and it means that both the variable moves in the same direction. (ZACH, 2020) In figure 5, it can be seen that the variable customer data has a strong positive correlation with marketing, outlet staff and outlet size of a correlation value greater than 0.9. Amongst which, customer data has a perfect positive correlation of 1 with the variable marketing, which indicates that both the variables marketing, and customer data move in the same direction at all times. Thus, it can be said that when that cost of marketing increases, the number of outlets visits increases as well and at all times. The variable marketing has a strong positive correlation with the variable’s outlet size and outlet staff of correlation value greater than 0.9. The variables outlet staff and outlet staff also have a perfect positive correlation of 1 which means that these two variables always move in the same direction.

Visualization 6: Box Subplots for ChrisCo’s Summary Data

Figure 6 is a box sub plot for ChrisCo’s summary data showing information regarding marketing, overhead, outlet staff and the size of the outlets.

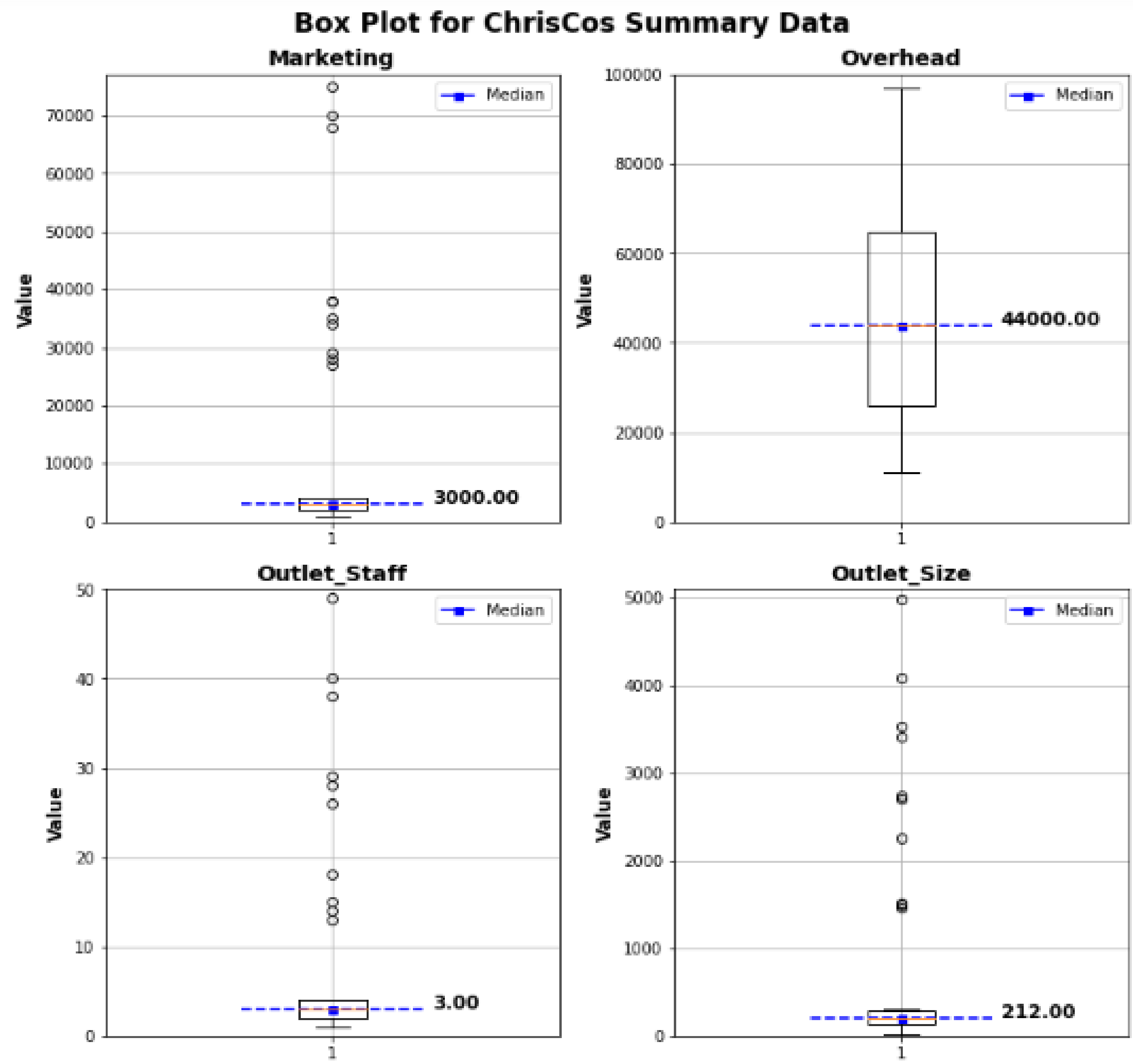

Figure 6: Box plot for ChrisCo’s Summary Data

Justification:

A box plot is a visual representation of a group of numerical data through its quartiles. The box plot shows the median, the maximum and the minimum, and high and low quartiles. The outliers can also be found from a boxplot. The distances between the different parts of the box show outliers and illustrate how spread and skewed the data are. The purpose of using a box plot is to find the distribution of data, outliers, the maximum, median and the minimum values of each variable in ChrisCo's summary data. This information is important in decision making, as having outliers may mean that something may have happened, which caused the value to be significantly different from the rest of the values and it is required to know the cause to be able to consider this in future. For example- it can be seen in marketing box subplot in figure 6, that marketing has many outliers. This may be due to seasonality, unusual conditions or a drop in visits which caused an increase in marketing cost. In addition, outliers may also mean having an error in the data.

Description:

The box plot above in figure 6, shows the maximum, minimum and median for each of the variables. It can be seen that the median value of marketing is 3000. This means that half of the outlets have an annual marketing cost greater than 3000 and the other half of the outlets have an annual marketing cost less than 3000. The points that are far away are considered as outliers. It can be seen that the subplot marketing has many outliers, and the last outlier is considered as the maximum value of marketing which is above 72000 and the minimum value is less than 2000. For marketing subplot, the end of whisker represents the minimum value. It can also be said from the box plot, that marketing has a narrow spread in data.

Similarly, the median value for overheads is 44000 which means that half of the outlets have annual overheads costs greater than 44000 and the rest half has annual overhead cost less than 44000. There are no outliers for overheads and hence the two ends of whiskers represent the maximum and the minimum value. It can be said by viewing the plot that the maximum value is above 95000 and the minimum value is around 10000. Moreover, overhead has a wide spread in data.

The third subplot is a subplot of number of employees employed at outlets. It can be seen that the median value is 3 which means half of the outlets have employees more than 3 and other half has employees less than 3. The minimum value of employees is around 1. There are many outliers, where the highest outlier value is around 49. These outliers may be caused by seasonality which may have caused an increase in outlet visits causing ChrisCo to hire more employees.

The last subplot is a subplot of outlet sizes where the median of outlet size is 212. Hence if can be said that half of the outlets have a size greater than 212 and the other half has outlet sizes less than 212. There are many outliers and highest value of outlet size is around 4993. Both the spread of data for outlet size and outlet staff is narrow.

Visualization 7: Bar Subplots for ChrisCo’s Summary Data

Figure 7 shows bar plots of ChrisCo’s summary data into subplots. Each of the subplot contains bar plots based on each of the columns from the summary data which are marketing, overhead, outlet staff and outlet size. I have dropped one column which contains information regarding customer visits to outlet, as a bar chart regarding this has have already been shown in figure 1.

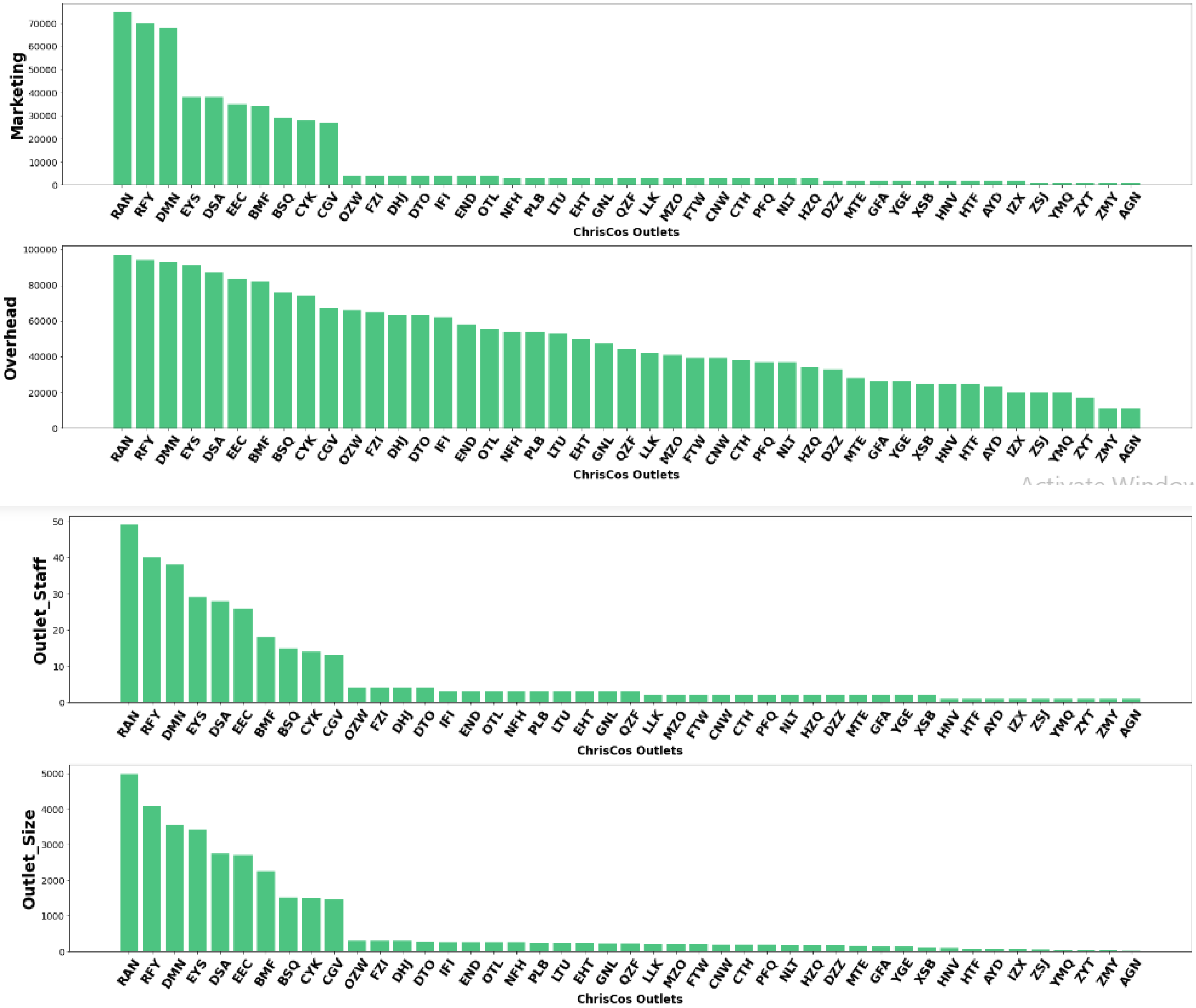

Figure 7: Bar Subplots for ChrisCo’s Summary Data.

Justification:

The purpose of using a bar chart as subplots to present this information, is because it is difficult to show it all together in one bar plot. When combined in one single plot, the bars represent each of the columns of the summary data and its value but does not consider the rows. The rows represent the outlets, and this is why I have implemented it as subplots so that I can present the rows of the dataset as well. The reason behind this is because it is important to find out which outlets size is bigger, which outlets have higher cost, which outlets have the highest number of employed staff, etc. With such a plot, these questions can be answered as bar subplots also show the rows. Hence this is why I have chosen this plot as with this I can compare each outlet individually. This purpose could not be fulfilled with a box plot as well which is implemented in figure 6. It only showed the maximum, minimum, and median values of each of the columns, but did not display the outlet names.

Description:

It is previously found in figure 6, regarding the maximum and minimum values of each of the columns in ChrisCo's summary data, but which outlet had the highest or lowest was not found. Figure 7 consists of 4 bar subplots where the first bar subplot is of Marketing, which shows the annual spending on local marketing for each outlet. It can be seen that the highest spending on marketing is on the outlets "RAN", "RFY" and "DMN" of over 7000. A medium amount of spending is on the outlets, "EYS", "DSA", "EEC", "BMF", "BSQ", "CYK" and "CGV". It is found earlier in figure 5, that there a high correlation between marketing and customer visits. It is also found in figure 1 that "RAN", "RFY" and "DMN" are the outlets with a high volume of visits. Hence it can be said that the correlation between these two variables is proved that an increase in marketing cost is what caused an increase in number of visits. Similarly, the outlets with a medium volume of visits found in figure 1 have a medium amount of cost on marketing.

The second sub plot shows information regarding ChrisCo's annual spending on overheads. Both the high and medium volume of outlets found in figure 1, have an overall similar amount of overhead cost. The outlets with a low volume of visits have a lower cost comparatively, which may be due to smaller outlet size.

The third subplot shows information regarding number of full-time staff employed at ChrisCo's outlet. The outlets "RAN", "RFY" and "DMN" have number of outlets staff greater than 40. Also, the medium volume outlets "EYS", "DSA", "EEC", "BMF", "BSQ", "CYK" and "CGV" have number of outlet staff greater than 10. As found in figure 5, the outlet staff has a strong correlation with outlet size and so the outlets "RAN", "RFY" and "DMN" have the highest outlet size shown in figure 7 sub plot number 4.

Even without consideration to figure 5, it can be said from figure 7 that the variables marketing, outlet staff and outlet size are correlated as the height of the outlets in each of the variable's subplots is similar to each other which shows that they have a relation with each other.

Visualization 8: Autocorrelation Plot of Outlets with High Volume of Customer Visits.

The last plot I have implemented is an autocorrelation plot with the 3 outlets that have the highest volume of customer visits.

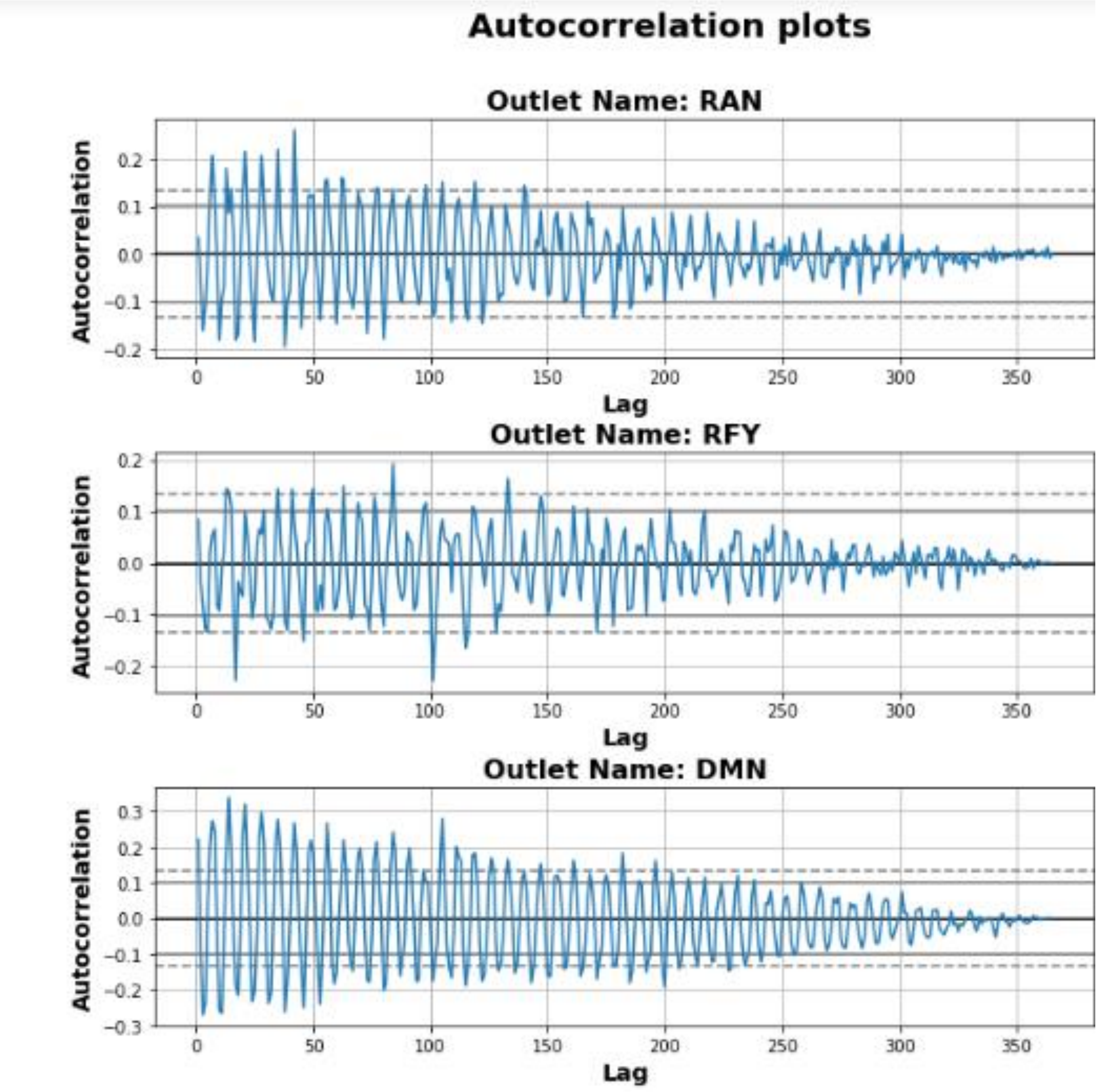

Figure 8: Auto correlation plot of outlets with the highest volume of customer visits.

Justification:

Autocorrelation shows the relationship between a variable’s present value and its previous values. The plot shows the correlation of the outlets at different lags and how correlation changes as the lags increase. The dotted blue line in the plots represent the significance threshold, where the data points below the dotted line is considered as not significant and data points above the dotted line is considered as significant. (Kumar & Ben-Assuli, 2017) When data are seasonal, the autocorrelations for the seasonal lags will be greater than for other lags. (Hyndman & Athanasopoulos, 2018) In addition, the data points above this line also means it is correlated significantly with its past values.

It has been found earlier in figure 1 that the outlets “RAN”, “RFY” and “DMN” are the outlets with the highest customer visits. The purpose of the autocorrelation plot in figure 8 is to explore the seasonality of the 3 selected outlets.

Description:

For the outlet “RAN”, it can be seen in the first subplot within the lag range 0 to 50, that there are 5 high data points which exceeded the value of 0.2. This indicates that there is seasonality. These data points are above the dotted line which also means they are significant. Other points inside the two dotted line indicates it is not significant.

For the outlet “RFY”, it can be seen in the second subplot there is a peak within the lag range 50 to 100, which indicates seasonality. In addition, there is one point in the lag range 0 to 50 and another data point within the lag range 50 to 100, where the value fell below -0.2. This also indicates seasonality. It can also be seen that most of the points are in between the dotted lines, which are considered as not significant.

Lastly, the third subplot represents the outlet “DMN”, where it can be seen that the lag range 0 to 50 has many high points which shows seasonality. In addition, it can also be said that after every 50 lags till lag 120, there is a rise in the value which also shows seasonality. Moreover, it can be seen that most of the points till lag 120 are above the dotted line which means they are significant.

Critical Review:

In this coursework, I have implemented 8 visualizations with which I tried to find information that the dataset contains and is hard to find just in a table format. The identified information includes the outlets with a high, medium and low volume of visits, the monthly averages of visits in the outlets, the outlets which have closed and newly opened, the seasonality for which outlets may have faced a change in its visit and the correlation. In addition, I have identified the maximum, minimum and median value of annual spending on marketing and overheads, the size of ChrisCo's outlets and the staff employed. I have also found the correlation between these and discovered which outlets have the highest cost, the biggest outlet size, and the highest number of employed staff.

In order to finish the coursework, I have applied all the knowledge that I have gained from the module COMP 1800 Data Visualisation. In this module, I have learned what visualization is and why it is necessary. When it comes to exploring large datasets, it is hard to understand and find information from it by just looking at a table. However, if the same dataset is visualized, a lot of the information can be extracted from the plots. For example- to find out the highest value of a outlet, thousands of rows in table has to be viewed one by one which is time consuming but with the help of a bar chart, it can be found within minutes. Hence visualization is important. In this module I have learned various techniques of visualization which includes bar charts, line plots, histogram, pie chart, area plot, bubble plot and a lot more. The best techniques of visualization and finding insights from the plots has been taught. I have learned to do visualization in python. I have also learned to do interactive plots which is very useful and effective as it includes additional features such as zooming in and out and hovering to find the details of the plot. This is very effective in terms of finding information on plots which contains a lot of information all together. For example- in an interactive line chart there may be many lines in one single plot but due to the option of hovering, it is easy to find details of each of the lines. I have utilized all this knowledge to do the coursework. I have applied the best techniques of visualization in my coursework in order to find the highest number of insights from the dataset. While choosing the best techniques, I have implemented a few plots and then chose the one which looks visually appealing and is useful in terms of finding information.

Summary:

- ChrisCo has 45 outlets, among which there are 3 outlets with a high volume of customer visits and 7 outlets with a medium volume of customer visits. The 3 high volume outlets are “RAN”, “RFY” and “DMN” and the 7 medium volume outlets are “EYS”, “DSA”, “EEC”, “BMF”, “BSQ”, “CYK” and ‘CGV”. The rest of the outlets have a low volume of customer visit. This information is shown in figure 1 with the help of a bar chart.

- The identified outlets with a medium volume of visits have an average monthly visit trend that is constant. Amongst which the outlet. “EYS” have a slightly increasing trend in 2022-01. Also, among the outlets with a high volume of outlet visits, the outlet “DMN” have an overall increasing trend. The other two outlets “RAN” and “RFY” have an overall fluctuating trend. This information regarding monthly visit trends of outlets with high and medium volume of visits is shown in figure 2 with a line plot.

- There are 5 ChrisCo’s outlets that closed during the year, and they are: “YGE”, “HNV”, “HTF”, “IZX” and “ZYT”. In addition, there are 6 outlets that opened, and they are: “ZSJ”, “YMQ”, “ZMY”, “AGN”, “XSB” and “AYD”. This information has been identified in figure 3, with the help of an interactive line chart which has features of zooming in and out, hovering, etc.

- In figure 4, an interactive heat map has been implemented based on the number of visits of different outlets. The heat map has additional features of zooming in and out, hovering etc. Amongst which a total of 19 group of outlets have been found with a positive correlation value of greater than 0.7. In each group there are two outlets. The highest positive correlation value is of the outlets “IZX” and “HTF” of 0.96654. In addition, 8 group of outlets have been identified with a negative correlation of value greater than -0.7. The highest negative correlation value is of the outlets “HNV” and “ZMY” of -0.79107.

- Figure 5 is heat map of ChrisCo’s summary data, which includes data regarding marketing, overhead, outlet staff, outlet size and customer data regarding number of outlet visits. The heat map shows there is a strong positive correlation between every variable other than with the variable overhead. The variable overhead has a weak positive correlation with all other variables.

- Figure 6 is a box plot for ChrisCo’s summary data, where the maximum, median, minimum values, and outliers of marketing, overhead, outlet size and outlet staff have been found. It has been found that the variable overhead has no outliers and has a median value of 44000. The median value of marketing is 3000 and has a number of outliers amongst which the highest value of outlier is above 72000. The median number of outlet staff is 3 and has several outliers. Lastly, the median of outlet size is 212 and also has a number of outliers. These outliers may have caused due to errors or seasonality.

- Figure 7 shows subplots of bar plot of ChrisCo's summary data, where the outlets with the highest or lowest costs, size or staff have been identified. The highest amount spent on marketing is on the outlets "RAN", "RFY" and "DMN" of over 7000. Also, the outlets "RAN", "RFY" and "DMN" have number of outlets staff greater than 40 which is the highest. Similarly, these 3 outlets have the highest outlet size as well. Lastly, the overhead cost for majority of the outlet is over 2000.

- Figure 8 shows autocorrelation of the 3 high volume outlets "RAN", "RFY" and "DMN", where seasonality has been identified among these outlets.

References

Hyndman, R. J. & Athanasopoulos, G., 2018. *Forecasting: Principles and Practice.* 2 ed. s.l.:s.n.

Kumar, A., 2022. *Data Analytics.* [Online]
Available at: https://vitalflux.com/correlation-heatmap-with-seaborn-pandas/#:~:text=with%20each%20other.-,What%20is%20Correlation%20Heatmap%3F,the%20strength%20of%20this%20relationship.
[Accessed 31 March 2023].

Kumar, S. A. & Ben-Assuli, O., 2017. Predicting Obesity Rate and Obesity-Related Healthcare Costs using Data Analytics. *Health Policy and Technology,* February.

PETERS, K., 2022. *Investopedia.* [Online]
Available at: https://www.investopedia.com/terms/l/line-graph.asp
[Accessed 31 March 2023].

ZACH, 2020. *Statology.* [Online]
Available at: https://www.statology.org/what-is-a-strong-correlation/
[Accessed 31 March 2023].

ZACH, 2021. *Statology.* [Online]
Available at: https://www.statology.org/what-is-a-weak-correlation/#:~:text=The%20correlation%20between%20two%20variables,is%20between%200.25%20and%200.5.
[Accessed 31 March 2023].