# COMP 1800

# Data Visualization

# Coursework


# Adesegun Emmanuel Sunmola

# 001250122

# M.Sc. Big Data & Business Intelligence

# Table of Contents

# Table of Figures

# Chapter One – Introduction to Data Visualization

Data visualization is a method of presenting data in graphical or visual form, which facilitates the comprehension and interpretation of complex information. This technique plays a crucial role in a variety of fields, including business, science, and research, as it enables analysts and decision-makers to extract valuable insights from data in a timely and efficient manner (Cairo, 2019). Given the rising volume and complexity of data generated on a daily basis, the significance of effective data visualization has never been greater.

Visualizing data entails choosing the most appropriate type of visual representation, such as charts, graphs, or maps, that accurately reflects the data and conveys the intended message (Few, 2012). It also entails adhering to design principles and best practices to produce visualizations that are both effective and visually appealing (Tufte, 2001). When done correctly, data visualization can uncover patterns, trends, and relationships in data that may not be evident through numerical analysis alone.

In summary, data visualization is an indispensable tool for comprehending data and communicating insights to others. As the significance of data continues to grow, so too does the demand for skilled data visualization professionals who can convey complex information in visual form.

# Chapter Two – Exploratory Data Analysis

## 1.1 EDA – Summary

Exploratory data analysis (EDA) is a crucial step in the analysis of any dataset. It involves summarizing the main characteristics of the data and identifying patterns, trends, and relationships between variables. The purpose of EDA is to gain a better understanding of the data and to identify any issues that need to be addressed before proceeding with further analysis.

## 1.2 Import Dataset into Python using Pandas Library

To load the dataset into Python for the project, I will be using the Pandas library in Python. The first step is to import the Pandas library and upload the file.

```
#Import Dataset into Python using the Pandas Library
daily_customers_df = pd.read_csv('https://tinyurl.com/ChrisCoDV/001250122/OutletDailyCustomers.csv', index_col=0)
marketing = pd.read_csv('https://tinyurl.com/ChrisCoDV/001250122/OutletMarketing.csv', index_col=0)
overheads = pd.read_csv('https://tinyurl.com/ChrisCoDV/001250122/OutletOverheads.csv', index_col=0)
size = pd.read_csv('https://tinyurl.com/ChrisCoDV/001250122/OutletSize.csv', index_col=0)
staff = pd.read_csv('https://tinyurl.com/ChrisCoDV/001250122/OutletStaff.csv', index_col=0)
pd.plotting.register_matplotlib_converters()
daily_customers_df.index = pd.to_datetime(daily_customers_df.index)
```

*Figure 1-Import Data*

The code uses the pandas library in Python to import data from five distinct CSV files. The data comprises information about daily customers, marketing expenses, overhead costs, outlet sizes, and staff counts. The files are located on the internet and accessed using URLs, and the 'index_col=0' parameter specifies that the first column of each file will serve as the index for the resulting pandas DataFrame. The code also registers matplotlib converters for pandas plotting. Additionally, it converts the daily_customers_df DataFrame's index to a datetime format with the pd.to_datetime() function. This is to get the data ready for analysis or plotting that requires a time series format.

## 1.3 Create DataFrame and Merge the Datasets

This code uses the pandas library to create a new DataFrame named 'summary_data.' The DataFrame is set up with its index being the same as the columns from the 'daily_customers_df' DataFrame.

```
#Merge and Read the Dataset
summary_data = pd.DataFrame(index = daily_customers_df.columns)
summary_data['Daily Customers'] = daily_customers_df.sum().values
summary_data['Marketing'] = marketing.values
summary_data['Overheads'] = overheads.values
summary_data['Outlet Size'] = size.values
summary_data['Staff'] = staff.values
```

*Figure 2-Dataframe*

Next, the code adds five columns to the 'summary_data' DataFrame: 'Daily Customers', 'Marketing', 'Overheads', 'Outlet Size', and 'Staff.' The 'Daily Customers' column gets its values from the sum of the values in each column of the 'daily_customers_df' DataFrame. The other four columns are populated with the corresponding values from the 'marketing,' 'overheads,' 'size,' and 'staff' DataFrames.

All in all, the purpose of this code is to consolidate information from various sources into one summary DataFrame that displays crucial data on daily customers, marketing expenses, overhead costs, outlet sizes, and staff counts.

### 1.3.1 Read the Outlet Daily Customers Datasets

The code is calling the "head()" method on the "daily_customers_df" DataFrame. The "head()" method is a pandas function that returns the first 5 rows of the DataFrame (or the number of rows specified within parentheses).

```
# Read the Outlet Daily Customers - the columns (outlets) show the number of customers who visit the outlets and the rows show the specific dates.
daily_customers_df.head()
```

| Date | IZX | YGE | ZYT | DMN | OZW | OTL | GNL | QZF | AGN | BSQ | ... | PFQ | CYK | DTO | FZI | HTF | DSA | PLB | MZO | YMQ | EHT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2021-01-01 | 92 | 70 | 78 | 2196 | 81 | 75 | 81 | 79 | 0 | 847 | ... | 69 | 840 | 99 | 72 | 82 | 1284 | 82 | 69 | 0 | 62 |
| 2021-01-02 | 79 | 57 | 63 | 1941 | 103 | 86 | 68 | 68 | 0 | 654 | ... | 66 | 881 | 72 | 85 | 101 | 1030 | 84 | 55 | 0 | 73 |
| 2021-01-03 | 80 | 61 | 83 | 1482 | 88 | 88 | 81 | 88 | 0 | 635 | ... | 75 | 698 | 84 | 84 | 90 | 1055 | 89 | 68 | 0 | 92 |
| 2021-01-04 | 89 | 66 | 60 | 1577 | 88 | 92 | 70 | 72 | 0 | 436 | ... | 66 | 769 | 89 | 86 | 79 | 891 | 86 | 70 | 0 | 74 |
| 2021-01-05 | 87 | 55 | 61 | 1598 | 75 | 83 | 83 | 78 | 0 | 805 | ... | 72 | 599 | 85 | 76 | 86 | 878 | 80 | 78 | 0 | 64 |

5 rows × 45 columns

*Figure 3- Read Dataset (Daily Customers)*

By calling this method, the code is displaying the first five rows of the "daily_customers_df" DataFrame, which can be useful to quickly check that the data has been imported correctly and to get a sense of the data's structure and content.

### 1.3.2 Read the “Summary Data” Datasets

The code is calling the "head()" method on the "summary_data" DataFrame. The "summary_data" is the newly created DataFrame consolidating the customers, marketing, overheads, size and staff data.

```
print(" ")
print("This is the summary data")
summary_data.head(10)
```

```
This is the summary data
```

| | Daily Customers | Marketing | Overheads | Outlet Size | Staff |
|---|---|---|---|---|---|
| IZX | 15112 | 2000 | 64000 | 123 | 2 |
| YGE | 17338 | 2000 | 34000 | 63 | 1 |
| ZYT | 6646 | 1000 | 71000 | 23 | 1 |
| DMN | 658979 | 64000 | 57000 | 6979 | 37 |
| OZW | 33769 | 3000 | 97000 | 147 | 1 |
| OTL | 30904 | 3000 | 79000 | 272 | 3 |
| GNL | 28639 | 4000 | 92000 | 218 | 3 |
| QZF | 27504 | 3000 | 27000 | 94 | 1 |
| AGN | 4399 | 1000 | 57000 | 23 | 1 |
| BSQ | 281864 | 25000 | 42000 | 2499 | 17 |

*Figure 4- Read Dataset (Summary_data)*

## 1.4 Inspect the Data

Inspecting the data is an important step in the data analysis process because it helps to understand the structure and content of the data, as well as to identify any potential issues or problems with the data. The purpose of inspecting the data is to gain a basic understanding of the data and to ensure that the data was loaded correctly. This includes checking for missing values, understanding the type of data in each column, and getting a sense of the overall distribution of the data.

By inspecting the data, we can make informed decisions about how to proceed with the data analysis, including what kinds of models or techniques are appropriate for the data, what data cleaning or preprocessing steps are necessary, and questions to answer with the data.

### 1.4.1 Shape and Size

To get a sense of the size of the data, we can use the shape method of the pandas DataFrame. The shape method returns the number of rows and columns in the data.

```
#Shape and Size of Data - summary_data
print("Number of rows: ", summary_data.shape[0])
print("Number of columns: ", summary_data.shape[1])
```

```
Number of rows:  45
Number of columns:  5
```

```
[29] #Shape and Size of Data - daily_customers_df
     print("Number of rows: ", daily_customers_df.shape[0])
     print("Number of columns: ", daily_customers_df.shape[1])
```

```
Number of rows:  365
Number of columns:  45
```

*Figure 5 - Shape & Size*

From the above, it shows that the “summary_data” contains 45 rows and 5 columns. The rows contain the 45 retail outlets for ChrisCo. The columns contain the daily visitors, marketing cost, overhead cost, outlet size and outlet staffs.

The above diagram also shows provides information on the daily visitors over the period of 365 days, from 1 January 2021 to 31 December 2021. It contains 365 rows and 45 columns. The 365 rows contain daily customer visit information for the 45 retail outlets (columns) over the period of 365 days.

### 1.4.2 Data Types

To understand the data types of each column in the data, we can use the ‘dtypes’ method of the pandas DataFrame. In addition, knowing the data types of the columns can be useful for tasks such as data cleaning, manipulation, and analysis.

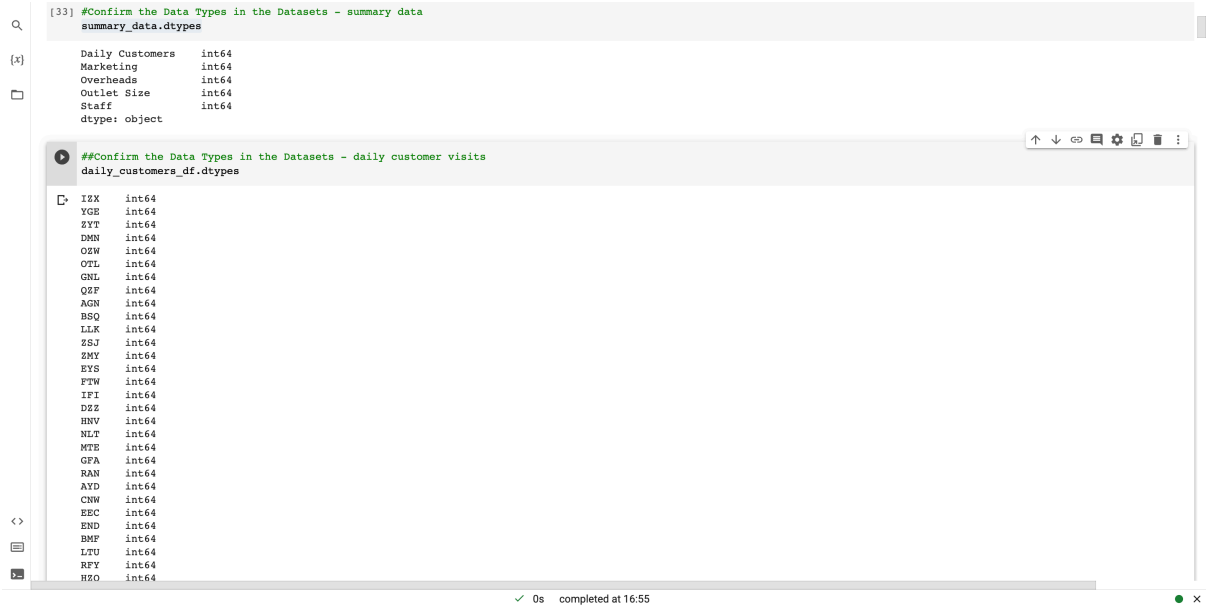


*Figure 6 - Data Types*

In this case, all columns are shown to be of integer data type, represented by "int64". This means that the values in each column are whole numbers with no decimal places.

### 1.4.3 Missing Values

To check for missing values in the data, we can use the ‘isnull’ method of the pandas DataFrame and the ‘sum’ method to get the total number of missing values in each column.

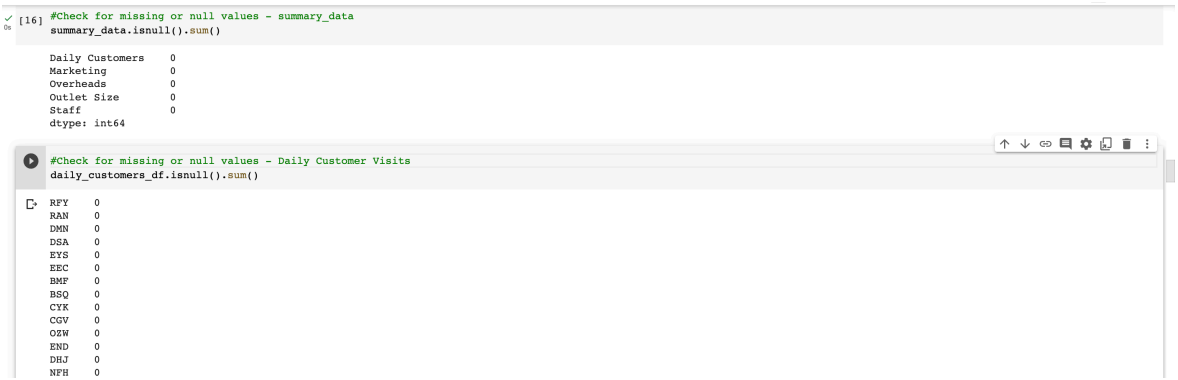


*Figure 7 - Missing Values*

The result shows that there are no null values.

# Chapter Three – Data Visualizations

## 3.1 Correlation Matrix Heatmap

### 3.1.1 Justification

A correlation matrix is a square matrix that shows the correlation coefficients between a set of variables. The matrix is symmetric, meaning that the correlation between variable A and variable B is the same as the correlation between variable B and variable A. The correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. The values of the correlation coefficient range from -1 to +1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and +1 indicates a perfect positive correlation.

Correlation matrices are often visualized using heatmaps, with warmer colors representing higher positive correlations and cooler colors representing higher negative correlations. The Correlation matrix can help ChrisCo to understand how different factors are related to daily customers and identify any potential drivers of sales.

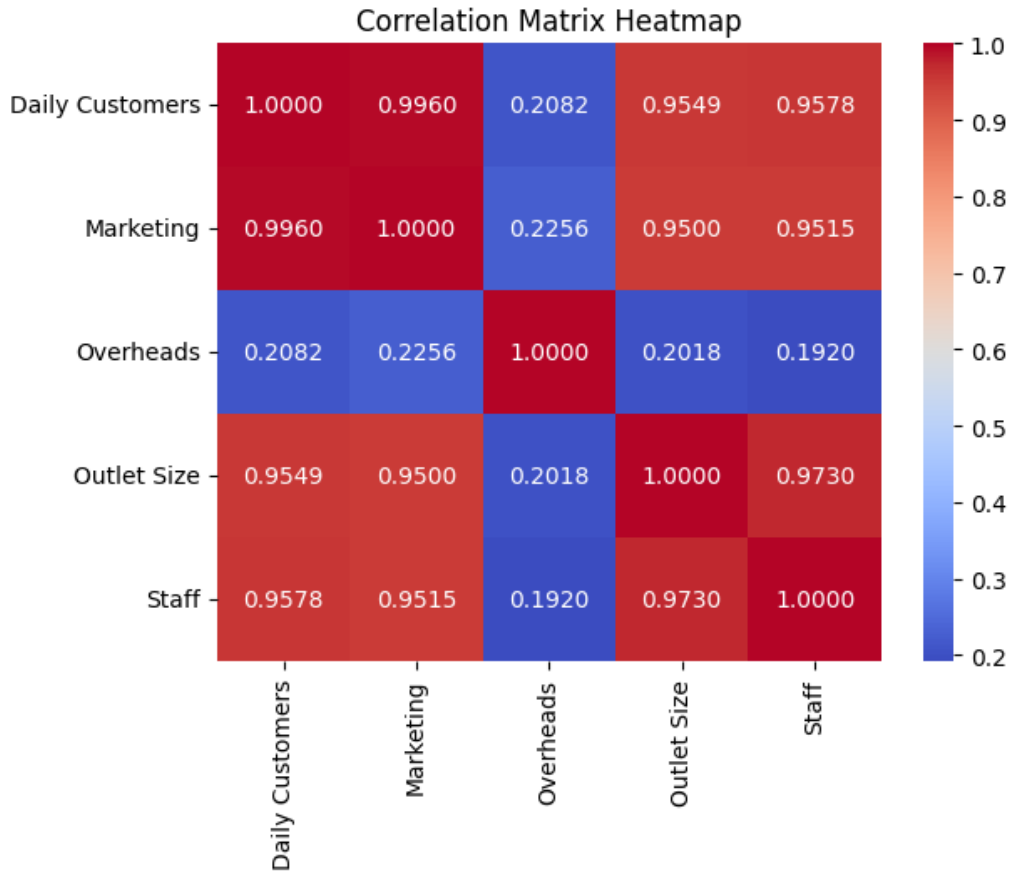


*Figure 8 - Correlation Matrix Heatmap*

### 3.1.2 Description

The correlation matrix reveals that there is a strong positive correlation (0.9960) between daily customers and marketing, indicating that outlets that invest more in marketing tend to have higher daily foot traffic of customers. There is also a strong positive correlation (0.9549) between daily customers and outlet size, suggesting that larger outlets will attract more customers. There also appears to be a positive correlation between daily customers and the number of staff, although this relationship is less clear than the others. Finally, overheads are relatively consistent across outlets, regardless of daily customer volume.

Overall, the correlation matrix suggests that marketing and outlet size may be important factors in driving sales, while overheads and staff may not have a strong impact on daily customers. This information can help ChrisCo. to make strategic decisions about where to allocate resources and how to improve performance.

## 3.2 Time Series Plot

### 3.2.1 Justification

A time series plot is a type of data visualization that displays how a variable changes over time. It is a graph with time plotted on the x-axis and the variable of interest plotted on the y-axis. The plot typically shows how the variable changes over days, weeks, months, or years.

A time series plot is a useful visualization to include because it helps to identify any patterns or trends in the data over time. By looking at the plot, we can see if there are any seasonal patterns, trends, or outliers that may be affecting the daily customers data. This can help ChrisCo to understand how sales are changing over time and identify any factors that may be driving these changes.

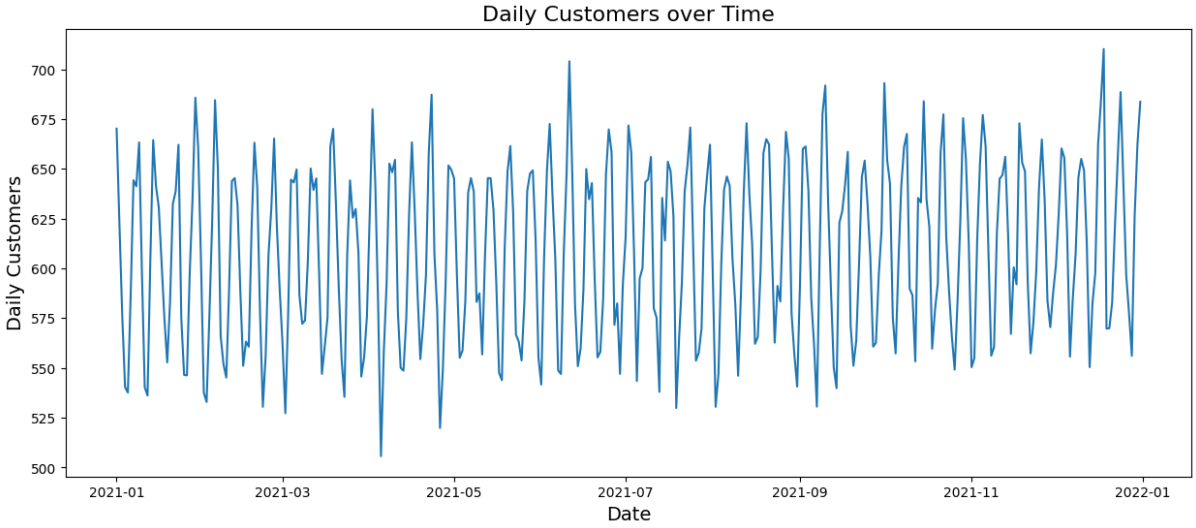


*Figure 9 - Time Series Plot*

### 3.2.2 Description

The time series plot provides a clear picture of the seasonality in the daily customers data. The dip in customers in April may be due to the start of the school year, while the peak in June and December may be related to summer holidays and tourism. This information can be used by the company to plan staffing levels, inventory, and marketing activities to match the seasonal trends in customer demand. Additionally, the plot suggests that the company may want to investigate any factors that may be driving the seasonal pattern and determine if there are any opportunities to increase sales during the low season. Overall, the time series plot provides valuable insights into the daily customers data that the company can use to make informed decisions about how to manage its outlets.

## 3.3 Bar Plot

### 3.3.1 Justification

A bar plot is a type of chart used to represent categorical data with rectangular bars. The length of each bar is proportional to the value it represents. Bar plots are commonly used to compare the values of different categories or groups, as well as to show changes in data over time. They can also be used to highlight patterns, trends, or anomalies in the data.

The justification for including this visualization is that it allows us to easily compare the total daily customers across different outlets. By visualizing this information in a bar chart, we can quickly identify which outlets have the highest and lowest volumes of customers, as well as any potential outliers.

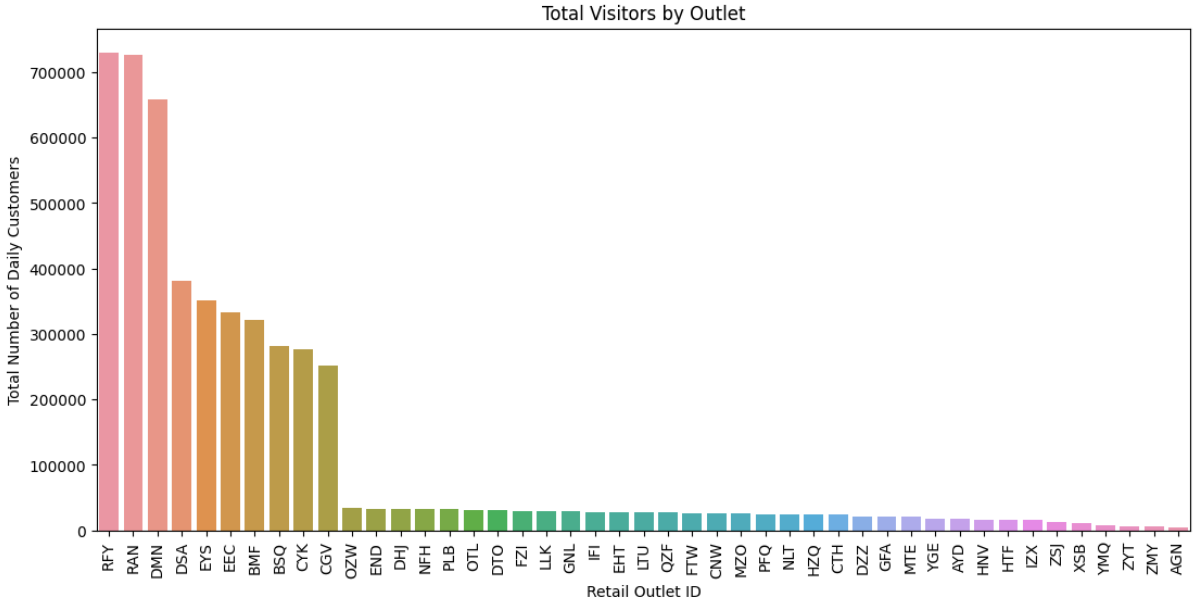


*Figure 10 - Bar Plot*

### 3.3.2 Description

From this visualization, we can draw several conclusions about the data. First, Outlet RFY has consistently high daily customer visits, suggesting that it may be a successful and popular location. Second, Outlet AGN has consistently low daily customer visits, which could indicate a potential issue that needs to be addressed. Finally, Outlets BSQ and CYK all have similar total daily customer counts, which could suggest that these outlets are comparable in terms of their performance. Overall, the bar chart is a straightforward way to present this information and can help ChrisCo. identify which outlets are performing well and which may need additional attention.

## 3.4 Box Plot

### 3.4.1 Justification

A box plot is a graphical representation of the distribution of a set of numerical data through their quartiles. It is also known as a box-and-whisker plot. A box plot consists of a box that spans the interquartile range (IQR), which is the range between the first and third quartiles of the data. The median of the data is marked by a line inside the box. The whiskers extend from the box to the minimum and maximum values that are not considered outliers, which are represented as points outside of the whiskers.

Box plots are useful for visually summarizing the distribution of data and identifying any outliers or skewness in the data. They can be used to compare the distribution of different groups of data, such as comparing the distribution of data between different categories or segments. Box plots are commonly used in statistical analysis and data visualization to quickly and easily assess the distribution of a dataset.

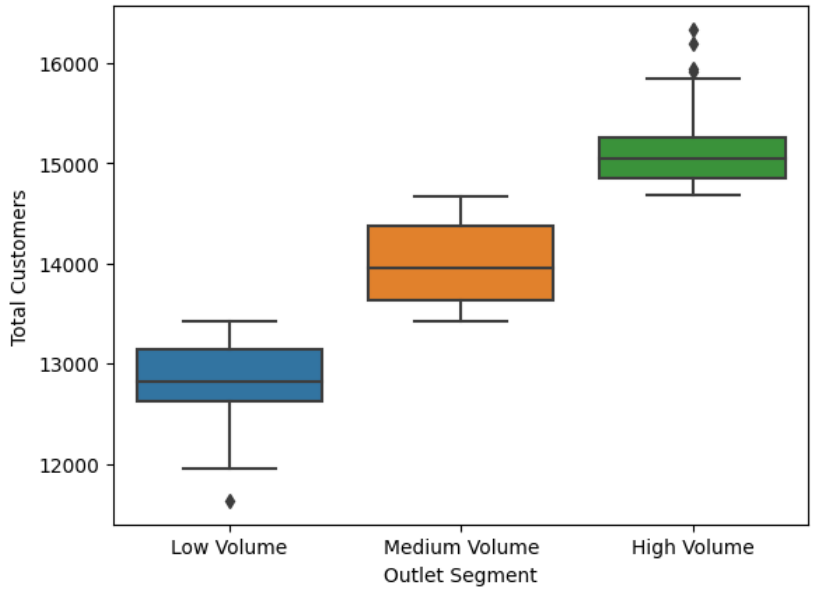


*Figure 11 - Box Plot*

### 3.4.2 Description

The box plot reveals that the median number of customers for the high volume segment is significantly higher than the median number of customers for the medium and low volume segments. Additionally, the high volume segment has a wider distribution of total customers, indicating that there is more variation in the number of customers between these outlets. The

box plot also shows that there are a few outliers in the high volume segment, indicating that there may be some outlets that are performing exceptionally well in terms of the number of customers. On the other hand, the low volume segment has a narrower distribution of total customers, indicating that there is less variation in the number of customers between these outlets.

In conclusion, the box plot is a useful visualization to compare the total number of customers between outlet segments and to identify any outliers or skewness in the data. It reveals that the high volume segment has a significantly higher median number of customers and a wider distribution of total customers compared to the medium and low volume segments. The low volume segment has a narrower distribution of total customers, indicating that there is less variation in the number of customers between these outlets.

## 3.5 Radar Chart (High Volume Retail Outlets)

### 3.5.1 Justification

A radar chart, also known as a spider chart or a star chart, is a graphical representation of multivariate data that displays data points in a two-dimensional plane with multiple variables represented on axes starting from the same point. The variables are plotted using a line or shape that connects the data points to show the overall shape of the data.

The radar chart is a suitable visualization for comparing multiple variables across different outlets in ChrisCo. It allows for a quick comparison of the performance of different outlets by plotting the values of each outlet on a set of axes, with each axis representing a different variable. The radar chart is used to visualize the normalized values of different variables, such as the number of customers, sales, and revenue, for each outlet in the "High_Traffic" list. The chart allows for an easy comparison of the values of these variables for each outlet and helps identify which outlets are performing well and which ones need improvement.

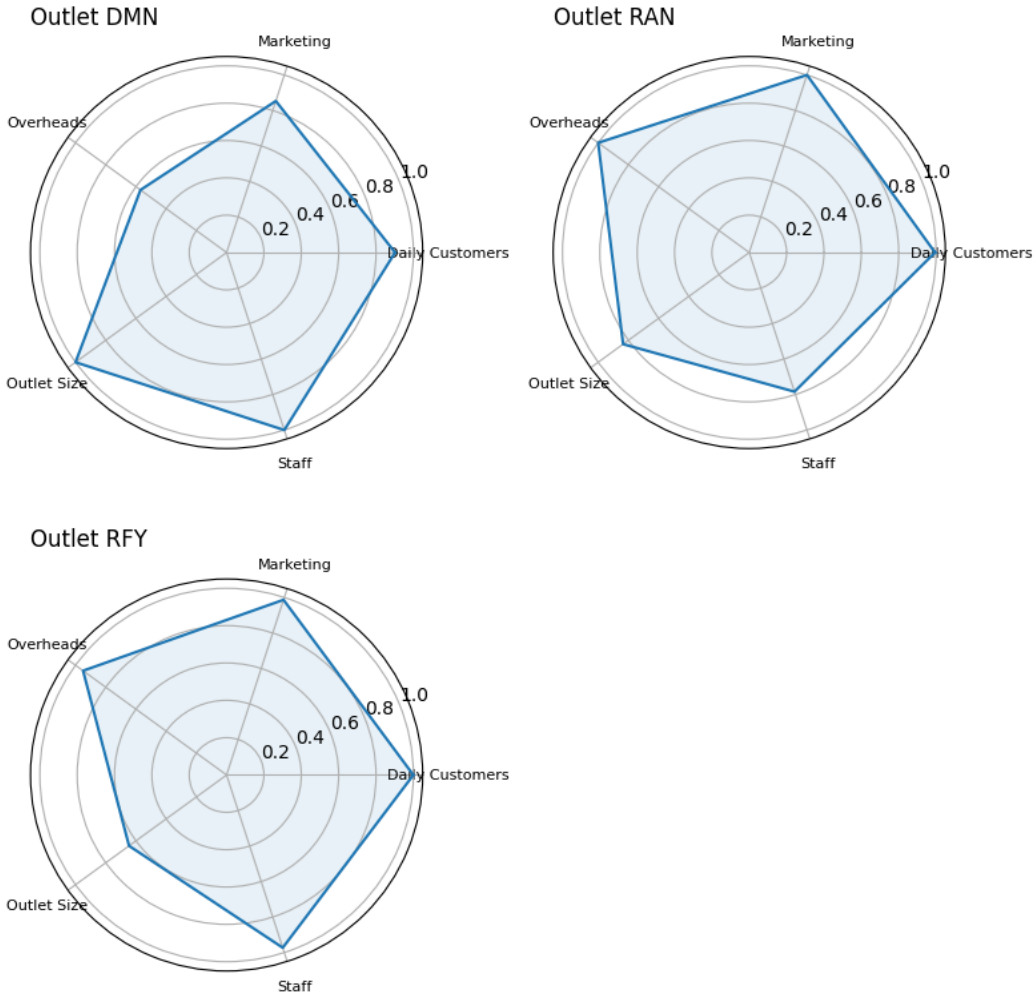


*Figure 12 - Radar Plot*

### 3.5.2 Description

The radar chart is used to compare the high-traffic retail outlets (represented by the different subplots) based on multiple attributes (represented by the different axes). Each axis represents an attribute such as daily customers, marketing, overheads, outlet size, and staff. The value of each attribute for each outlet is shown as a line that connects the values on the respective axes.

From the chart, it can be observed that some outlets perform better than others in certain attributes. For example, outlet RFY has high daily customer visits and marketing, while DMN has the smallest marketing cost and daily customer visits. This information can be useful for ChrisCo. in identifying the strengths and weaknesses of their outlets and developing strategies to improve their performance.

## 3.6 Autocorrelation Plots (Medium Traffic Retail Outlets)

### 3.6.1 Justification

An autocorrelation chart, also known as an autocorrelation plot, is a graphical representation of the correlation between a time series and a lagged version of itself. The x-axis represents the lag time, and the y-axis represents the correlation coefficient between the time series and the lagged version of itself.

ChrisCo could use the autocorrelation chart to identify patterns and trends in customer visits over time. Autocorrelation plots can help identify the presence of seasonality and trends, as well as the optimal lag time for forecasting future values. By analyzing the autocorrelation chart for each outlet, ChrisCo can determine if customer visits are influenced by previous visits at the same time of day, week, or month. This information can be used to improve staffing levels and inventory management, as well as inform marketing and promotional strategies. Additionally, autocorrelation charts can help identify potential outliers or anomalies in the time series data.

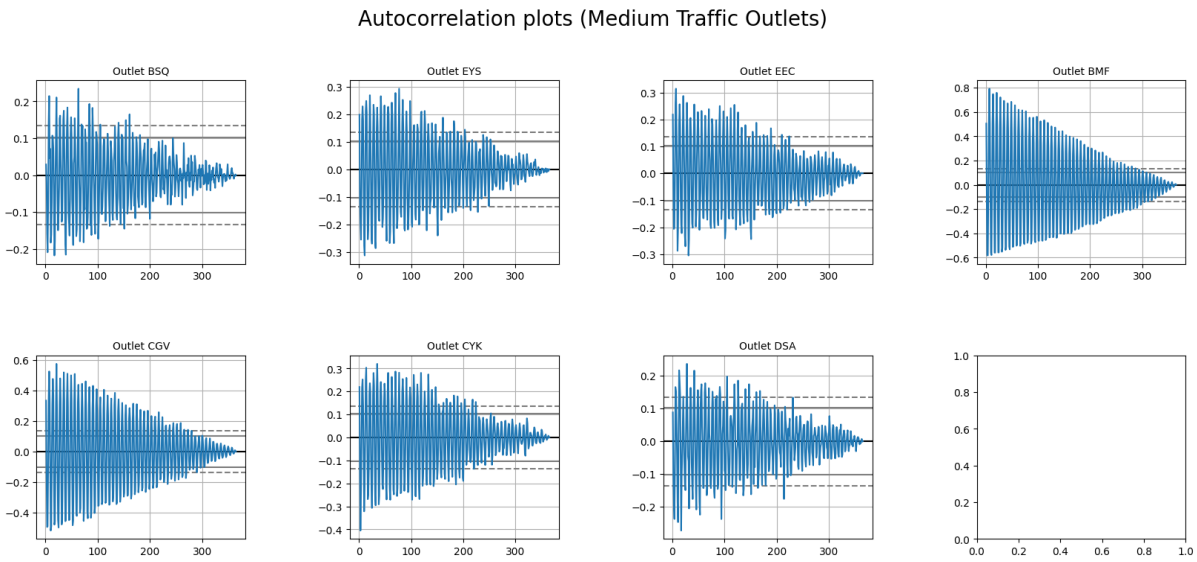


*Figure 13 - Autocorrelation Plot*

### 3.6.2 Description

The autocorrelation chart is showing the autocorrelation plot for each medium traffic outlet. The autocorrelation plot measures the correlation between a time series and a lagged version of itself at different lags. The horizontal axis represents the lag and the vertical axis represents the correlation coefficient. The dashed blue lines in the plot indicate the confidence interval for the correlation coefficients.

In the plot, we can see that for some outlets like DSA, BSQ and EYS, the correlation coefficients are high for the first few lags, indicating a strong correlation between the values at these lags. As the lag increases, the correlation coefficients decrease and eventually become statistically insignificant. This suggests that these outlets have a seasonality pattern, where the values are correlated with their previous values at certain lags.

For other outlets, the correlation coefficients are mostly insignificant, indicating a lack of strong correlation between the values at different lags. This suggests that these outlets do not have a strong seasonality pattern.

## 3.7 Interactive Line Plot (High Volume Retail Outlets)

### 3.7.1 Justification

An interactive line plot is a data visualization tool that allows users to interact with the plotted data by hovering over or clicking on the plotted lines to reveal more detailed information. The plot can be customized in various ways, such as changing the colors, markers, or line styles of the plotted lines.

For ChrisCo, an interactive line plot could be useful in analyzing the daily customer visits to their retail outlets over time. By providing an interactive visualization, ChrisCo can allow their stakeholders to explore the data in more detail and gain a better understanding of the trends and patterns in customer visits over time. This can help them make more informed decisions about marketing strategies, store operations, and resource allocation. Additionally, ChrisCo can easily compare and contrast the performance of their different outlets over time.

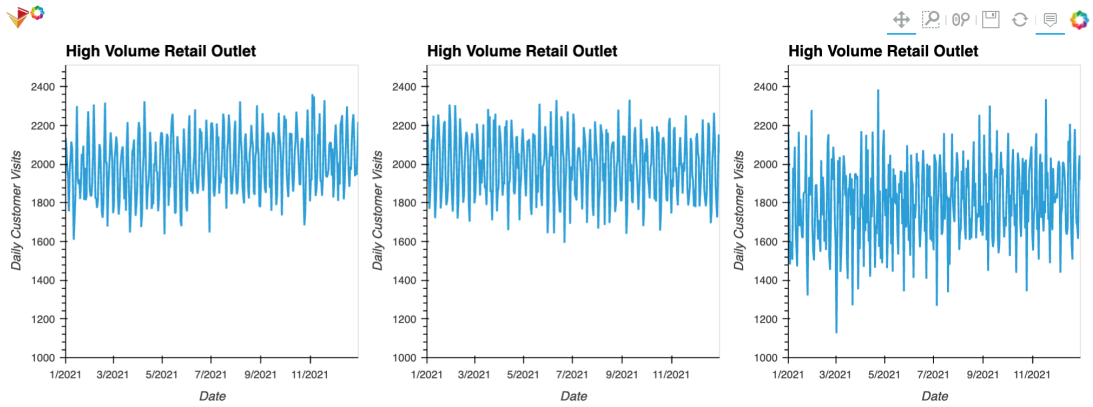


*Figure 14 - Interactive Line Plot 1*

Hovering on the RFY line chart displays additional information which helps to understand the information better. As shown below, hovering over the data point show the customers traffic for 24th October 2021.

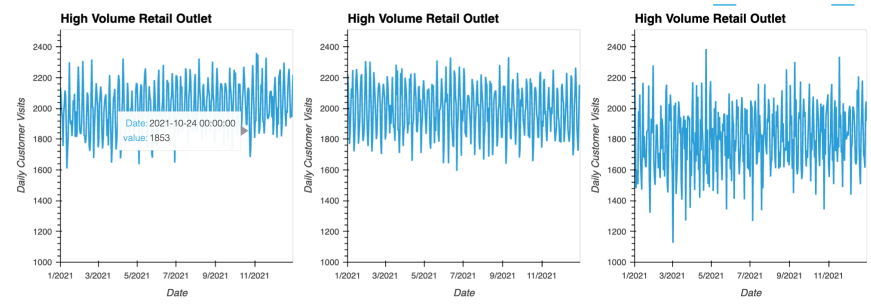


*Figure 15 - Interactive Line Plot 2*

Zooming in on the RAN interactive line plot displays a more detailed view of the selected data, as shown below:

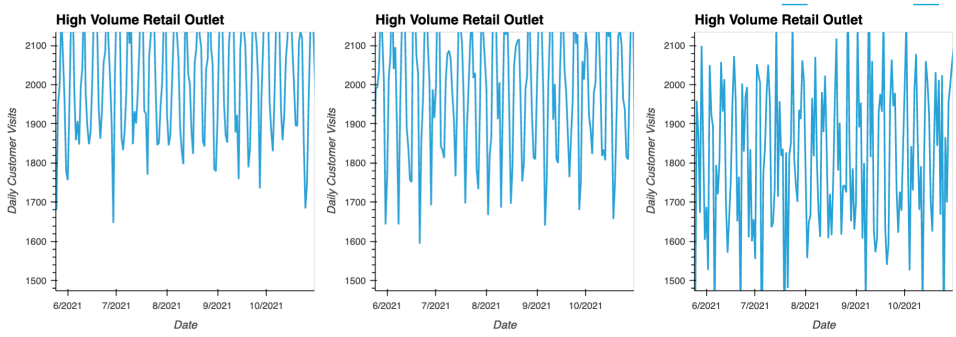


*Figure 16 - Interactive Line Plot 3*

### 3.7.2 Description

The chart is an interactive line plot of the daily customer visits over time for three high volume retail outlets (RFY, RAN, and DMN) in the dataset. The x-axis represents time, and the y-axis represents the number of daily customer visits.

From the chart, we can observe that RFY has the highest number of daily customer visits, followed by RAN, and then DMN. There are some fluctuations in the number of daily customer visits for each outlet over time, but overall, the trend seems to be relatively stable. We can also see that there is a slight increase in the number of daily customer visits for all three outlets towards the end of the time period represented in the chart.

The use of subplots allows for easy comparison of the daily customer visits between the three outlets, and the visualization provides a clear and concise way of presenting this information.

## 3.8 Interactive Scatter Plot – Overheads Vs Daily Customers

### 3.8.1 Justification

An interactive scatter plot is a graphical representation of data points in a two-dimensional space, where each point represents the values of two variables. In an interactive scatter plot, users can interact with the graph by zooming, panning, and hovering over individual data points to see more information.

ChrisCo can use an interactive scatter plot to visualize the relationship between two variables, such as the relationship between daily customers and overhead costs. The interactivity of the plot allows users to explore the data and identify any patterns or outliers that may not be immediately visible in a static plot. In addition, an interactive scatter plot can be useful for data exploration and hypothesis testing. By selecting different variables to plot and adjusting the size of the data points, users can quickly identify any correlations or trends in the data. Overall, an interactive scatter plot is a powerful tool for data visualization and analysis.

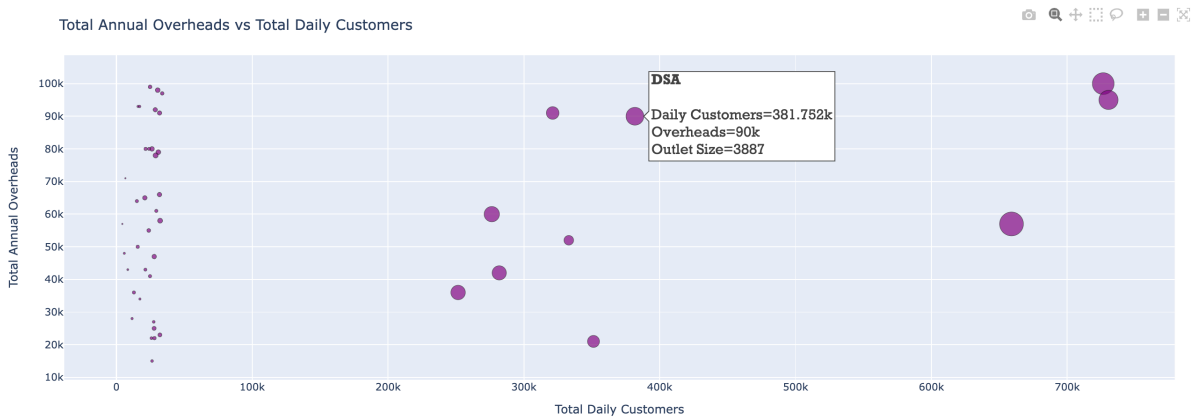


*Figure 17 - Interactive Scatterplot*

### 3.8.2 Description

The chart is an interactive scatter plot showing the relationship between total annual overheads and total daily customers for ChrisCo's outlets. The size of the markers represents the outlet size, and hovering over each marker shows the name of the corresponding outlet. The x-axis represents the total daily customers while the y-axis represents the total annual overheads. The scatter plot indicates that there is no clear correlation between total annual overheads and total daily customers. The markers are distributed randomly throughout the chart, indicating that there is no obvious pattern to the relationship between these two variables.

# Chapter Four – Critical Review

The objective of the visualization project is to explore the characteristics of different outlets of ChisCo company based on their relevant characteristics, including correlations, seasonal behaviour, outliers, etc. to help ChrisCo. segment the data and make informed decisions.

The report provides a clear methodology for data exploration, normalization, and transformation to ensure that the visualizations accurately reflect the underlying data. Once the exploratory data analysis is completed, and the data is confirmed to be clean, various visualizations were developed on the Google Colab environment.

The correlation matrix map was visualized to understand the relationships between the different variables, the time series plot was visualized to display how the "daily customer visit" variable changes over time. The bar plot shows the number of customers that visit the individual retail outlets making it easy to identify the high performing outlets and the low performing outlets. The box plot reveals that the median number of customers for the high volume segment. The radar chart compares multiple variables across different outlets in ChrisCo, and was visualized after the normalization conducted. The autocorrelation chart to identify patterns and trends in customer visits over time.

The visualizations would help ChrisCo. to identify patterns and trends, segment customers and outlets and identify anomalies or low volume areas, and provide insights for making great decisions.

# Chapter Five – Conclusions

In conclusion, the analysis and visualization revealed several key insights about the retail outlets and customer behavious at ChrisCo. It was revealed found that there is a significant positive correlation between outlet size and the number of daily customers, indicating that larger outlets tend to attract more customers. We also identified seasonal patterns in customer visits, with higher volumes during the summer months and when students are on break, and lower volumes during the winter months and at the end of summer.

Overall, the visualizations provided a clear and concise representation of the data, making it easy to identify trends and anomalies. These insights can help inform future business decisions and strategies for ChrisCo.

## References:


Cairo, A. (2019). How charts lie: Getting smarter about visual information. W. W. Norton & Company.

Few, S. (2012). Show me the numbers: Designing tables and graphs to enlighten. Analytics Press.

Fry, B. (2008). Visualizing Data. O'Reilly Media, Inc.

Healy, K. (2018). Data Visualization: A Practical Introduction. Princeton University Press.

Kelleher, C., & Tierney, B. (2018). Data Science: An Introduction. CRC Press.

Kirk, A. (2016). Data Visualisation: A Handbook for Data Driven Design. SAGE Publications.

McCandless, D. (2010, February 26). The beauty of data visualization. TED. https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization

Tufte, E. R. (2001). The Visual Display of Quantitative Information. Graphics press.

Ware, C. (2012). Information visualization: Perception for design. Elsevier.

Wilke, C. O. (2019). Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures. O'Reilly Media, Inc.