Data Visualization

(COMP-1800)

Coursework submitted in partial fulfillment

of the requirements for the degree of

Master of Science

In

Data Science

By

Rabin Chhetri

(ID: 001185145)

Visual Data Exploration for ChrisCo

Submitted to: Prof. Chris Walshaw

Module Leader (COMP-1800)

School of Computing and Mathematical Sciences

UNIVERSITY of GREENWICH

Table of Contents

List of Figures

1. Introduction

In general words, Data Visualisation (Tableu, 2022) can be defined as the graphical representation of any information or data to provide an easy access to understand different aspects of the data. Being one of the important steps of data analysis and data science, the main motive of data visualization is to pass information in a very clear and efficient way to the users using various graphical means (Vitaly Friedman, 2008).

According to (Stephen Few, 2004), data are of two types which are combinely used to support meaningful information: Categorical and Quantitative data. Categorical data includes items with a particular features, it can be either nominal or ordinal whereas Quantitative data includes items with a proper measurement, and it can be either discrete or continuous. It is most important to properly distinguish between these two types of data as they require different types of visualization techniques. There are many types of data visualization techniques and some of them include bar chart, line plot, heat map, scatter plot, pie chart, histogram, correlogram, box plot, radar plot, bubble plot and many more. Visualisation helps in generating out the proper information from the data after removing all the noise from it. Before visualizing the data, the data needs to be cleaned and shaped correctly so that it can be visualized clearly.

2. Visualizations

This section contains all the eight visualization based on the sales and website data of ChrisCo. Visualizations are done including all the six datasets provided breaking down into two data frames. The main dataset contains total of 40 venues and daily records for each venue during the entire year so these venues are categorized into three sections as High, Medium and low Volume venues depending on the number of visitors on each venue to effectively explore the data.

2.1. Bar Chart

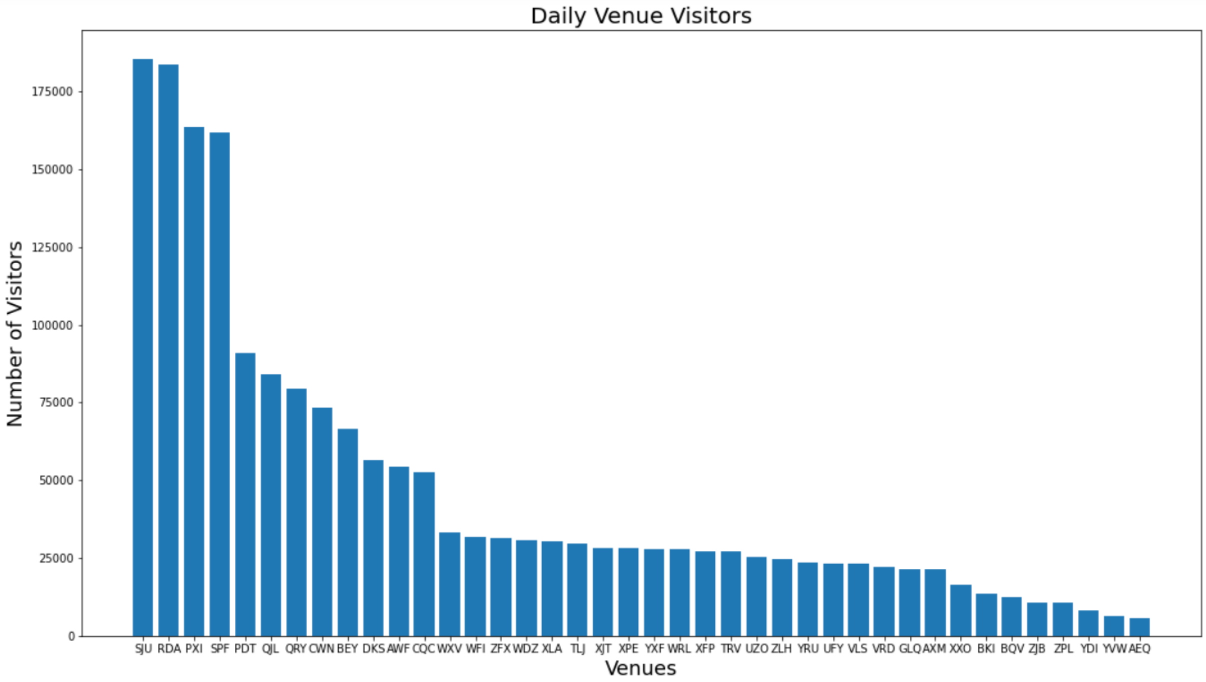

Figure 1: Bar Chart of Daily Venue Visitors

The above figure (Figure 1) represents the total number of visitors in all the 40 venues in the form of Bar Chart. X-axis represents each venue whereas y-axis represents the number of visitors. The data is sorted in descending order before visualising so that it would be easier to explore the findings.

From Figure 1, as the data is sorted we can easily segment the data into different categories. Considering venues with more than 100000 visitors as High Volume Venues, the first four venues can be categorized as the High Volume venues, similarly in the range of 50000 to 100000 visitors venues from PDT up to CQC can be categorized as Medium Volume Venues. Likewise below that up to AXM can be categorized as Low Volume Venues and all others can be categorized as Very Low Volume venues. It is clear from the above Figure 1 that Maximum number of visitors prefer to visit venue SJU and only very few visitors rarely visit venue AEQ. From this bar chart, we can easily identify what proportion of the total data belong to high, medium, low and very low categories.

2.2. Comparative Bar Chart

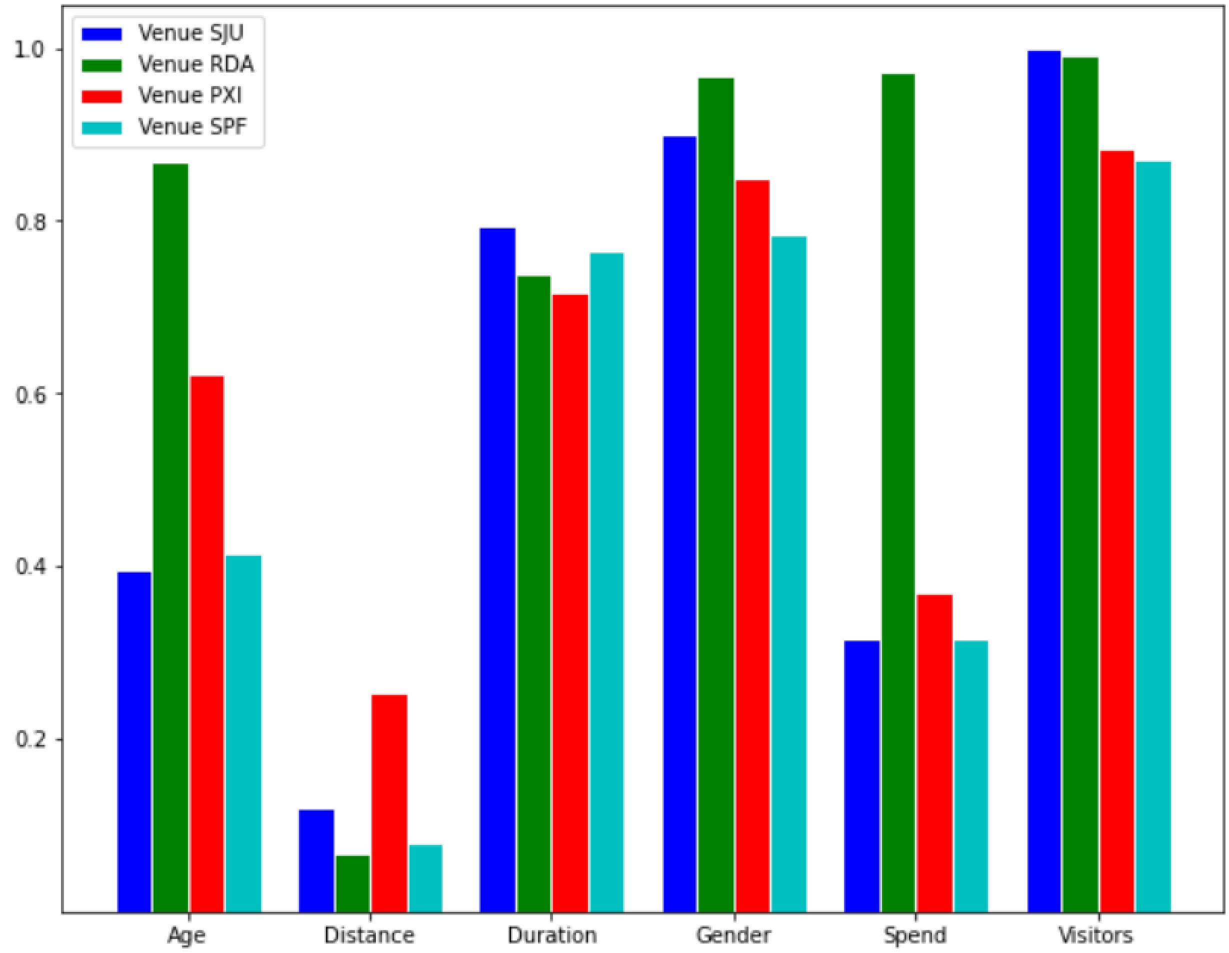

Figure 2: Comparative Bar Chart for High Volume Venues

The above figure (Figure 2) shows the comparative bar chart for all the high volume venues with different color for each bar. This bar chart contains the normalized value of number of visitors in each high volume venues in order to bring every venues in the same scale.as we have five dimensions in the summary data including age, distance, duration, gender and amount spend.

From above Figure 2, it can be noticed that Venue RDA has the maximum number of female visitors, old age visitors and all the visitors travel very less distance from the venue. Also, the visitors visiting venue RDA spent comparatively more amount during their stay. It is also clear from the above figure that Venue SJU attracts maximum number of visitors and they stay there for longer duration of time as the average age of visitor is comparatively less than for other venues.

2.3. Correlogram

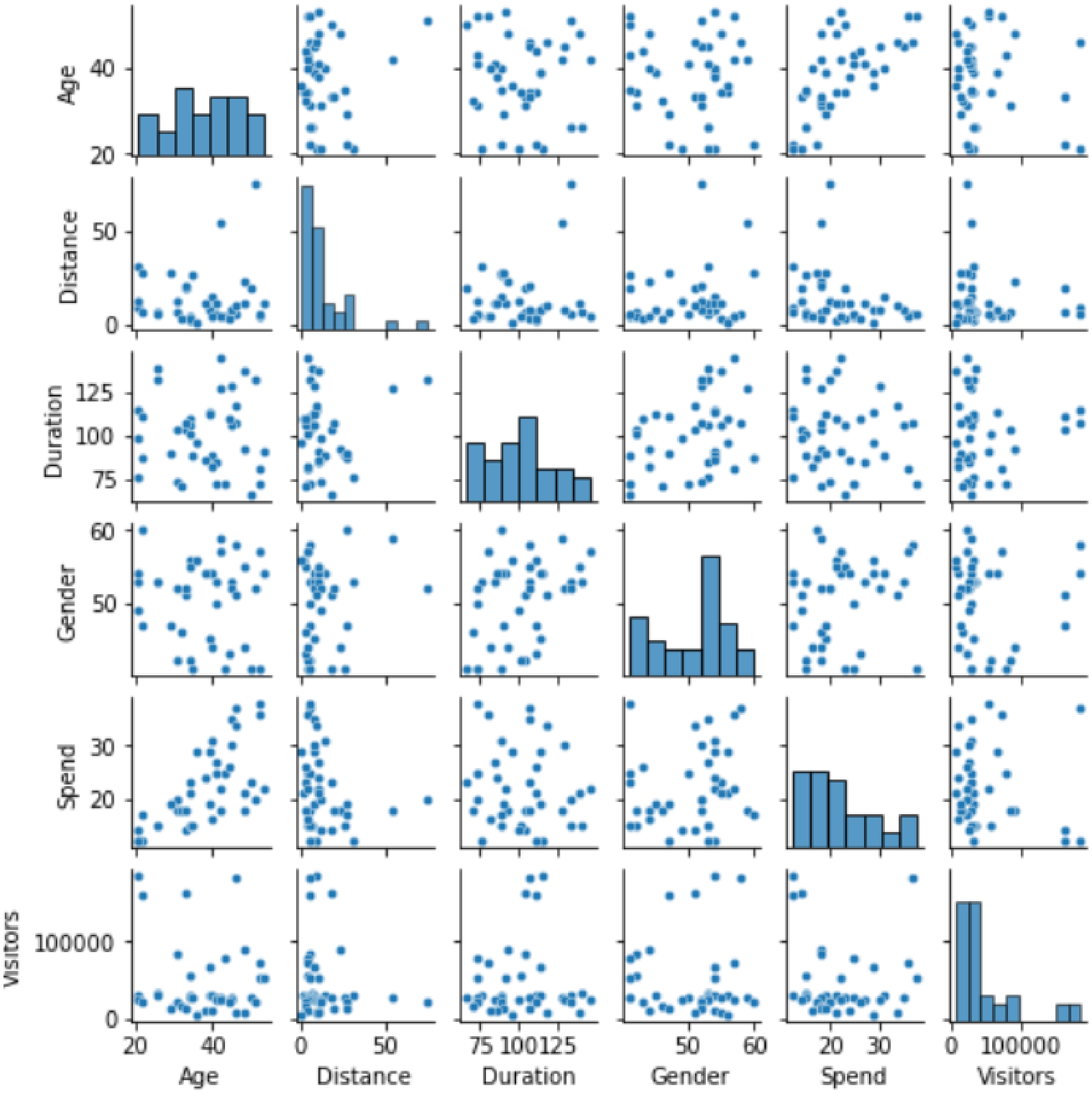

Figure 3: Correlogram of Summary Data

The above figure (Figure 3) shows the correlogram plot of summary data with all the dimensions and number of visitors in each venue. This plot is helpful in identifying the correlation between each dimensions of the dataset. Correlogram plots are really good in visualising datasets with higher dimensional values.

As there is no point in showing the correlation of a variable with itself, all the diagonal portion is filled with the histogram. It is noticed that there is a positive correlation between gender and duration. Similarly, we can also correlate the age of visitors and the amount they spend in each venue. This plot helps us in finding interdependence between the dimensions of the dataset. Most of the female visitors spend more amount of time while visiting the venue. Also, it is noticeable that younger visitors spend comparatively less amount of money than the older aged visitors and Maximum number of old aged visitors travel less distance from the venue.

2.4. Heat Map

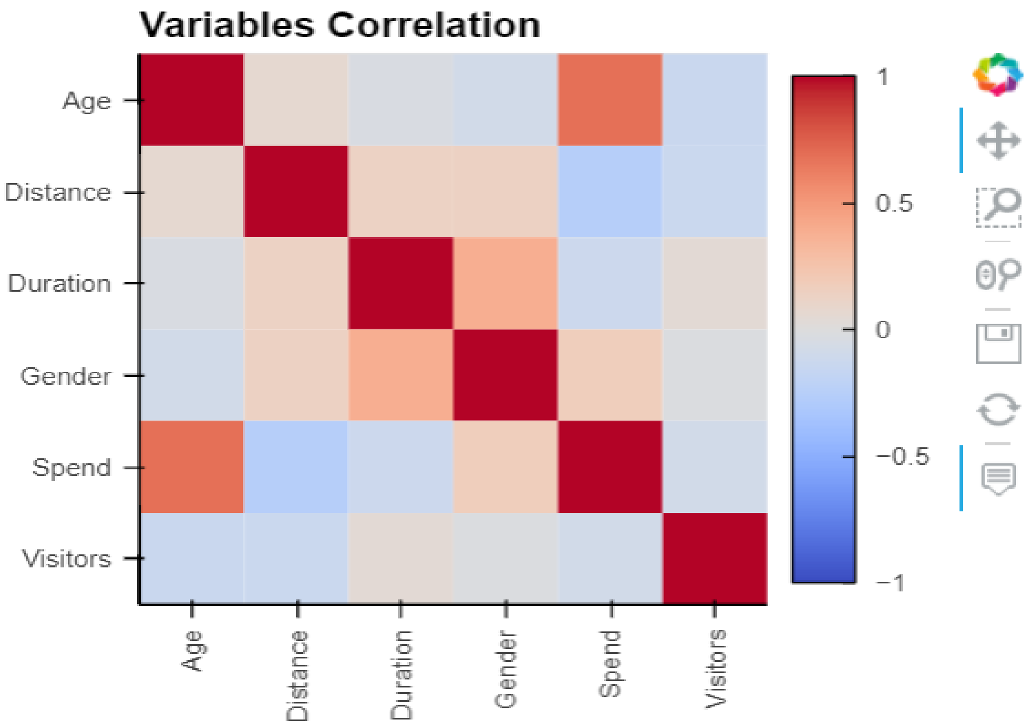

Figure 4: Interactive Heat Map of Summary Variables

The figure (Figure 4) shows the interdependence of all the dimension in the summary data using Heat Map. Here, blue color (-1) signifies the cold region or the negative correlation whereas red color (1) signifies the hot region or the positive correlation. Since a variable is perfectly correlated with itself so all the diagonal element are perfect red in color.

The above heat map also shows the correlation between dimensions. Here, we can see age and amount spend are more correlated than other dimensions. It indicates visitors with higher average age spend more amount of money than lower aged visitors. Similarly, dimensions distance and spend are not related to each other.

2.5. Scatter Plot

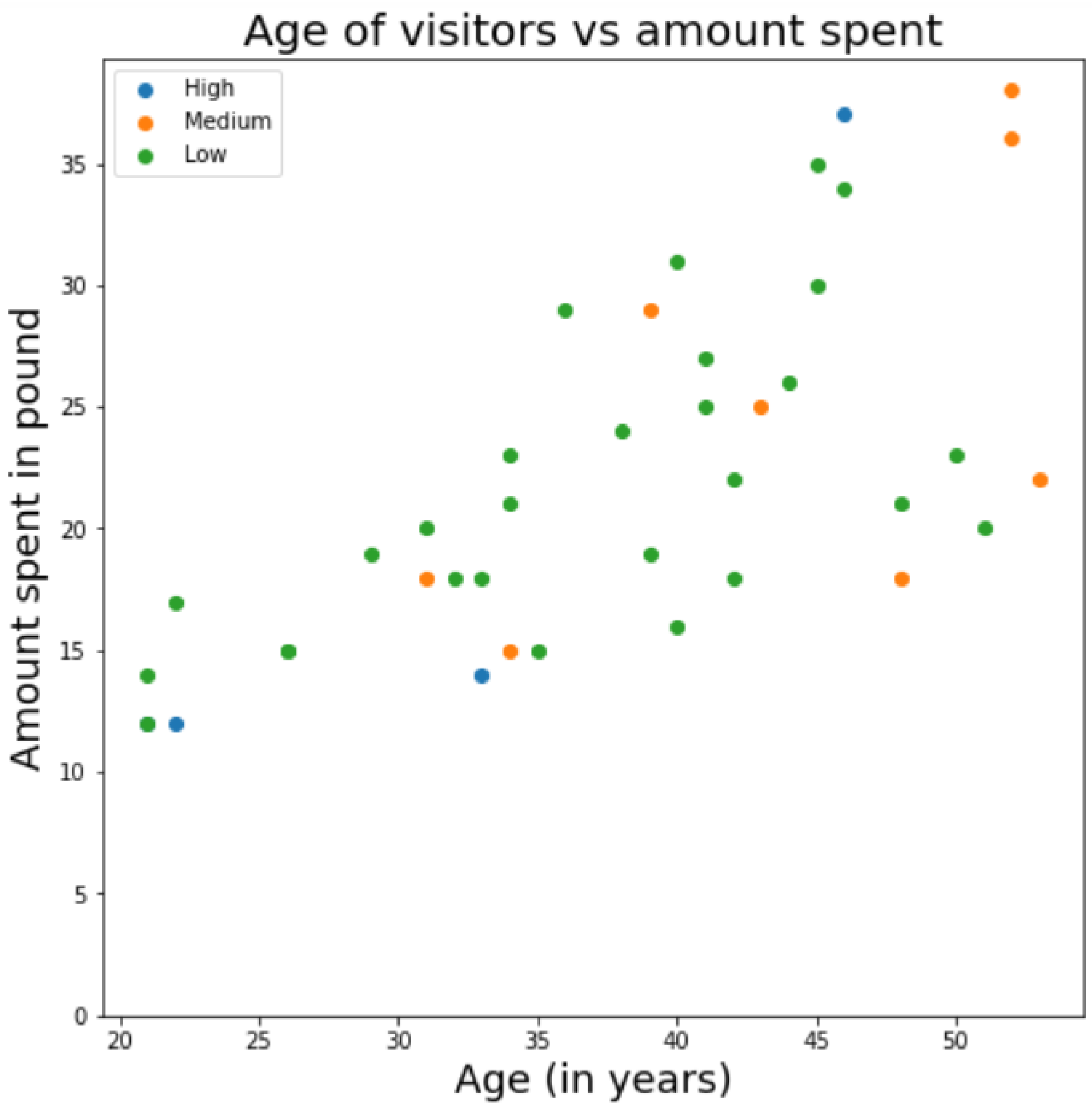

Figure 5: Scatter Plot showing relation between Age and Amount spent in a venue

The above figure (Figure 5) shows relationship between age of visitors and the corresponding amount they spent during their visit in the form of Scatter plot as scatter plots are easier in analyzing the correlation between the variables. Here, three different colors are used to categorize the data into 3 different types. Blue represents High volume venues, orange represents medium volume venues and green represents the low volume venues.

It shows that for the High volume venues, age of visitors and the amount they spend are interdependent to each other that is as the age of visitor increases the amount they spend also increases and vice- versa whereas for the medium and low volume venues it is not applicable. This plot also gives information that the visitors spend maximum amount of money in medium volume venues.

2.6. Histogram

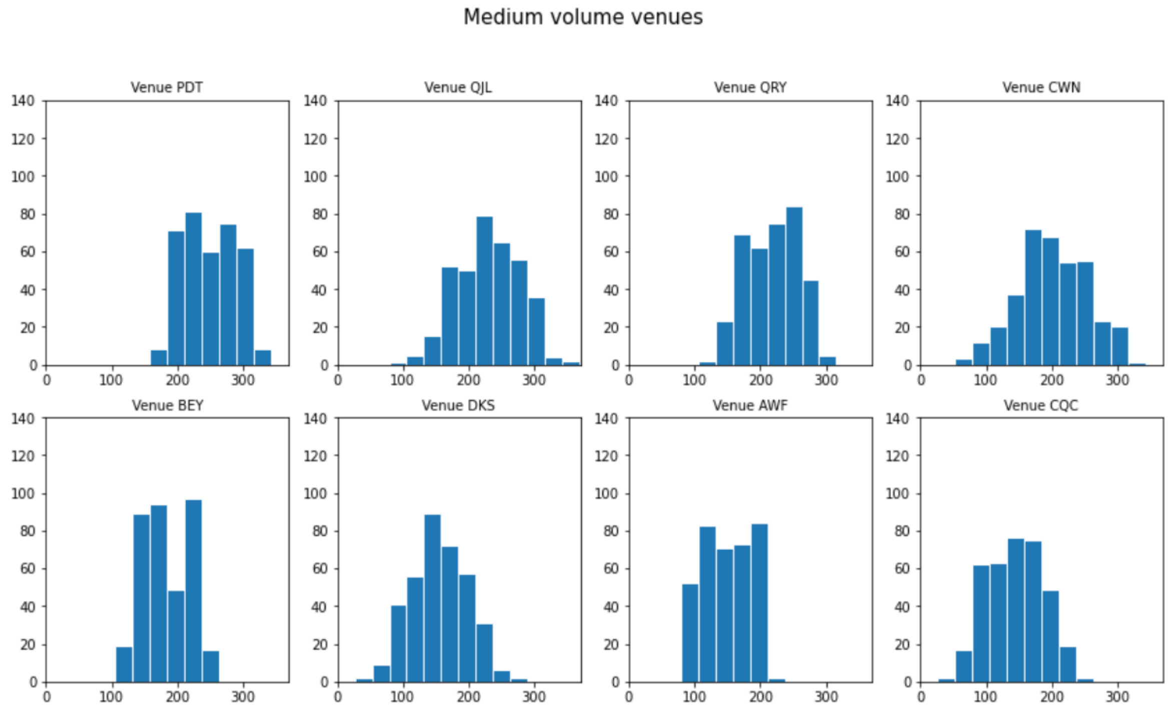

Figure 6: Histogram for Medium Volume Venues

The above figure (Figure 6) shows the statistics of daily visitors in the medium volume venues in the form of Histogram. This plot was chosen as it gives simple and versatile representation of the data. This Histogram shows the frequency of the number of visitors in each medium volume venues.

Here, the number of days and number of daily visitors are scaled down to lower value representing them in y-axis and x-axis respectively. Maintaining bin width as 25 there are total of 15 bins used in creating this Histogram. Subplots are created for each individual venue and each subplot shows the normal distribution of data. From the above Histogram, it can be concluded that in case of medium volume venues where the total number of visitors lies in the range if 50000 to 100000, the distribution of visitors among the venues is clearly normal.

2.7. Box Plot

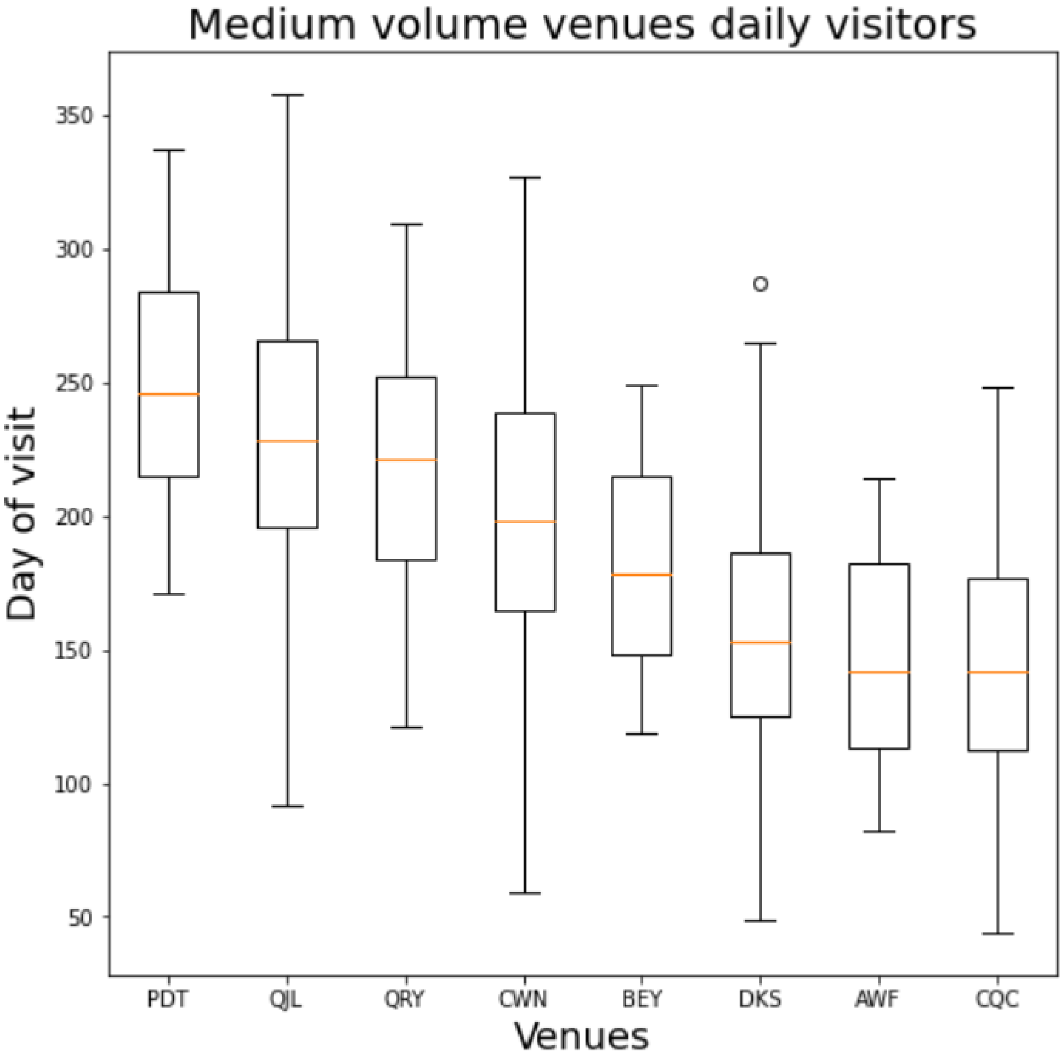

Figure 7: Box Plot for Medium Volume Venues

The above figure (Figure 7) shows the distribution of number of daily visitors in the medium volume venues in the form of Box plot. This plot also gives similar information as in Histogram that the distribution of visitors is normal in all the medium volume venues.

Here, in case of venues BEY and AWF the whiskers are too short in length which represents that maximum data is within the box around the median and these two venues receive visitor in a minimum number of days during the entire year. Venue CWN receives the visitors in maximum number of days throughout the year. Venue CQC starts receiving visitors in the beginning of the year whereas Venue QJL attracts visitors till the end of the year. Also, Venue DKS also shows outliers in its visitors distribution which indicates it has some visitors beyond its maximum limit. Overall, this plot shows normal distribution of visitors in the medium volume venues.

2.8. Line Plot

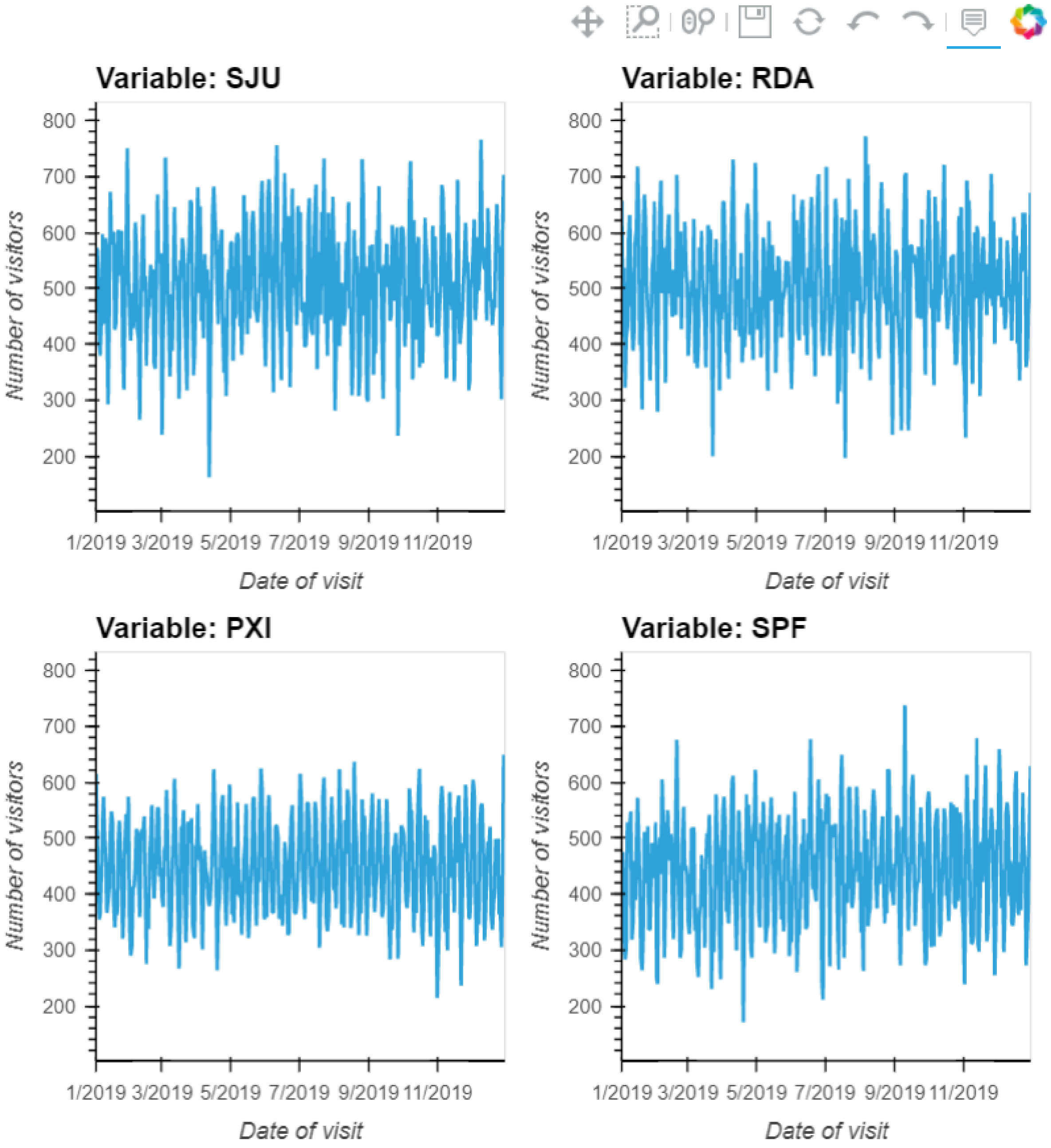

Figure 8: Interactive Line Plot for High Volume Venues

This figure (Figure 8) shows the frequency of visitors in all the high volume venues throughout the year in the form of line plot. This line plot is interactive in nature where the reader can get the depth information from the plot using the interactive tools available within it. This plot is chosen as line plot are really helpful in representing the time series data and it is easier to identify the seasonal factors affecting the distribution of data.

This line plot gives us the information that among the high volume venues SJU receives maximum number of visitors throughout the year whereas SPF receives comparatively lower visitors. As a whole, the distribution of visitors is normal as there is no such extreme change in the plot because of any seasonal effects or anomalies in the data.

3. Critical Review

In order to explore the data of the ChrisCo and generate some insight from it, I have used eight different techniques of data visualization which includes bar chart, comparative bar chart, heat maps, scatter plot, histogram, box plot, correlogram and line plot.

Beginning with the bar chart it showed the proportions of total data distributed among all the venues from where it was easier to segment the venues as high, medium, low and very low categories. The summary data is normalized to bring down the scale to same level for all the dimension. After that Comparative Bar chart was used which helped in identifying the interdependence between all the dimensions of the dataset (age, distance, duration, gender, and amount). Correlogram, Heat maps and Scatter plots were used to explore the correlations between each dimensions of the data and found some relation between age of the visitors and the amount of money spent during their visit for all categories of the data though it was extremely correlated incase if high volume venues. Following the exploration of correlation, Histogram and Box plot were implemented over the data of medium volume venues to look over the distribution of the visitors. The distribution is found to be normal but one outlier is detected in case of DKS venue. Finally Line plot was used with added features of interactivity to see if there are any seasonal behaviors and anomalies but the data is found to be free from anomalies and seasonal attacks throughout the year.

4. Conclusion

After visualizing the given data of the ChrisCo using different techniques of Data Visualization, following points can be observed.

- There is no any missing information in the dataset provided.
- These data points gives information about the count of visitors in 40 different venues throughout the year with five dimensions (age, gender, distance, duration and amount spend) affecting the number of visitors.
- Among 40 venues, Venue SJU and Venue AEQ are the most visited and the rarely visited venues respectively.
- More than half of the venues receive lesser amount of visitors below 50000 whereas only four venues (SJU, RDA, PXI, and SPF) receive more than 10000 visitors.
- Most of the older aged visitors prefer to visit venue RDA as they stay very nearby to the venue.
- Most of the female visitors visit venue BEY while they visit venues ZPL, VLS, PXI and SPF rarely.
- People travel maximum distance to visit venue SJU whereas they travel very less distance to visit venue AEQ.

- Visitors spend maximum amount of time in venue BQV and stay for shorter duration in venue PXI.
- People spend maximum amount (38 pounds) in venue VLS whereas they spend very less (12 pounds) in venues VRD, XXO and QRY.
- For High volume venues, Age of the visitor and amount they spend are positively correlated.
- All the given data points can be best segmented into different categories using sorting.
- Venue DKS possess some outliers
- Overall, there is no such seasonal behavior or anomalies in the data points and the distribution is normal in nature.

5. References

Stephen Few, 2004. *Selecting the right graph for your message.* [Online]
Available at: http://www.perceptualedge.com/articles/ie/the_right_graph.pdf
[Accessed 7 4 2022].

Tableu, 2022. *tableu.* [Online]
Available at: https://www.tableau.com/en-gb/learn/articles/data-visualization
[Accessed 7 4 2022].

Vitaly Friedman, 2008. *Data Visualization and Infographics.* [Online]
Available at: https://www.smashingmagazine.com/2008/01/monday-inspiration-data-visualization-and-infographics/
[Accessed 7 4 2022].