UNIVERSITY of GREENWICH

Coursework for COMP-1800-Data Visualization (Term 2)

Author: Neharika Joshi

Student ID: 001200608

Date: 8th April, 2022

Course: MSc. Data Science

Module Leader: Prof. Chris Walshaw

Deadline: 8th April, 2022

Acknowledgement

To begin with, I would like to express my deepest appreciation to those who provided us with the opportunity to attend this module.

Many thanks go to the course leader, Prof. Chris Walshaw who has invested his full effort in guiding the students in achieving the goal.

Lastly, I would like to thank my beloved husband, Mr. Deep Khadka and my fellow classmates for providing me with insightful remarks and improving my work for this module.

Table of Contents

1. Introduction 5

2. Exploring Data 5

3. Visualizations 6

3.1 Segmentation with Bar Charts 6

3.2 Venues Opening & Closing 7

3.3 Distribution of High and Medium Volume Visitors with Box Plot 8

3.4 Weekly Seasonality with Auto Correlation Plot 9

3.5 Heat Map of High & Medium Volume Venues 10

3.6 Pair Plot of Summary Data 11

3.7 Radar Plot of High-Volume Summary data 12

3.8 Comparative Bar Chart for High Volume 13

3.9 Summary of Low Volume & Very Low Volume Venues 14

4. Critical Review 16

Conclusion 17

References 18

Table of Figures

Figure 1 : Segmentation of Venues 6

Figure 2 : Area plot for opened and closed venues 7

Figure 3: Box plot for High & Medium Volume Venues 8

Figure 4: Autocorrelation plots for High & Medium Volume Venues 9

Figure 5: Heat Map of High & Medium Volume Venues 10

Figure 6: Pair Plot for Summary Data 11

Figure 7: Radar Plot of High-Volume Venues 12

Figure 8: Comparative Bar Chart for High Volume 13

Figure 9: Heat map of low volume venues 14

Figure 10: Heat map of low volume venues 15

1. Introduction

Data visualization is the visual representation of data or information so that it may be easily understood. When you look at raw data, it just looks like a bunch of numbers that don't make sense. However, once you begin editing and processing the data, all of the hidden information becomes apparent. Because everyone cannot grasp the concept of data science, it is easier for everyone to understand what is going on by looking at a graphical representation. It may be used to find trends, outliers, seasonal behavior, and patterns in large datasets. The advent of Big Data necessitates its representation in a readable format, and people prefer to absorb data visually since it is easier to comprehend. Computers are used to construct visualizations these days, and the most frequent techniques include line plots, bar charts, and scatter plots. Interactive visualizations, which allow users to describe how data is displayed, are also popular these days. Patterns, trendlines, and possible correlations between data components are displayed using computer graphics (Sadiku, et al., 2016).

2. Exploring Data

ChrisCo is a company that manages 40 venues around the United Kingdom and collects a massive quantity of data by using a loyalty card scheme to track every visit. In addition, six separate datasets from various categories must be evaluated. The data sets are:

- Daily Visitors in the venues for the year 2019
- Average age of the customer
- Maximum Distance travelled to reach each venue (in miles)
- Average time spent in each venue (in minutes)
- Rate of female visitors at each venue
- Average Money spent at each venue (in GBP)

The dataset is explored in two different data frames, 'data' for the daily visitors and 'sum_data' which is the complied data for all the rest of the datasets including sum of daily visitors.

3. Visualizations

3.1 Segmentation with Bar Charts

A bar chart is a graph in which each bar shows a relationship between two variables: the category and the value. The height of the bar is proportional to the category's value. A bar chart can effectively depict how the venues can be labelled when segmenting the data into distinct groups.

Figure 1 depicts the total number of visitors to all 40 locations in the year 2019. There are four venues (RDA, SJU, PXI, and SPF) that have a large number of visitors and are classified as **High-Volume Venues**. Similarly, **Medium-Volume Venues** are defined as eight venues (PDT, AWF, BEY, QJL, QRY, CWN, CQC, DKS) with average visits. In addition, as shown in Figure 1, there are 20 venues (UFY, XLA, XFP, WFI, WRL, WXV, VRD, WDZ, GLQ, XPE, YRU, TRV, TLJ, YXF, XJT, AXM, ZLH, ZFX, VLS, UZO) that have nearly the same number of visitors. These are classified as a **Low-Volume Venues**. Finally, the remaining eight venues with the fewest visitors (XXO, BKI, AEQ, BQV, YVW, ZPL, YDI, ZJB) are classified as **Very Low-Volume Venues**.

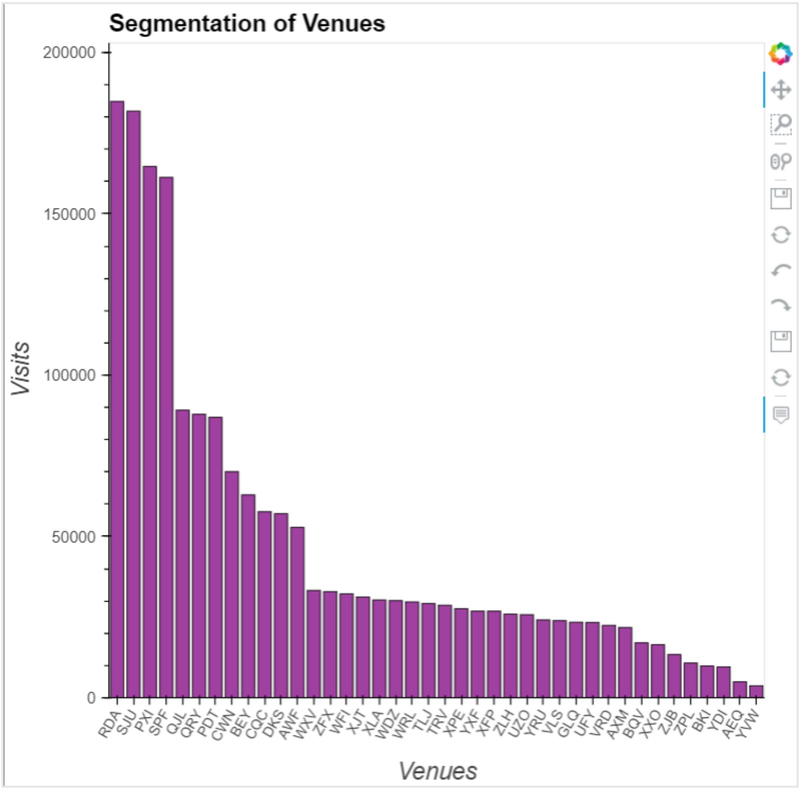

Figure 1 : Segmentation of Venues

3.2 Venues Opening & Closing

Area plots are based on line charts and are used to compare two or more similar items. It can also be used to discover any anomalies in data over a period of time. Due to the fact that several venues opened and closed during the year 2019, an area plot was utilized to depict those anomalies.

Figure 2 illustrates that XXO and ZPL had a significant boost in the first half of 2019, but were forced to close in July. BKI and YDI, on the other hand, both opened in the same month and made significant growth in only five months. Figure 2 further shows that ZJB and BQV were only established in April of this year. After opening in October 2019, the monthly visits to the venues YVW and AEQ nearly doubled.

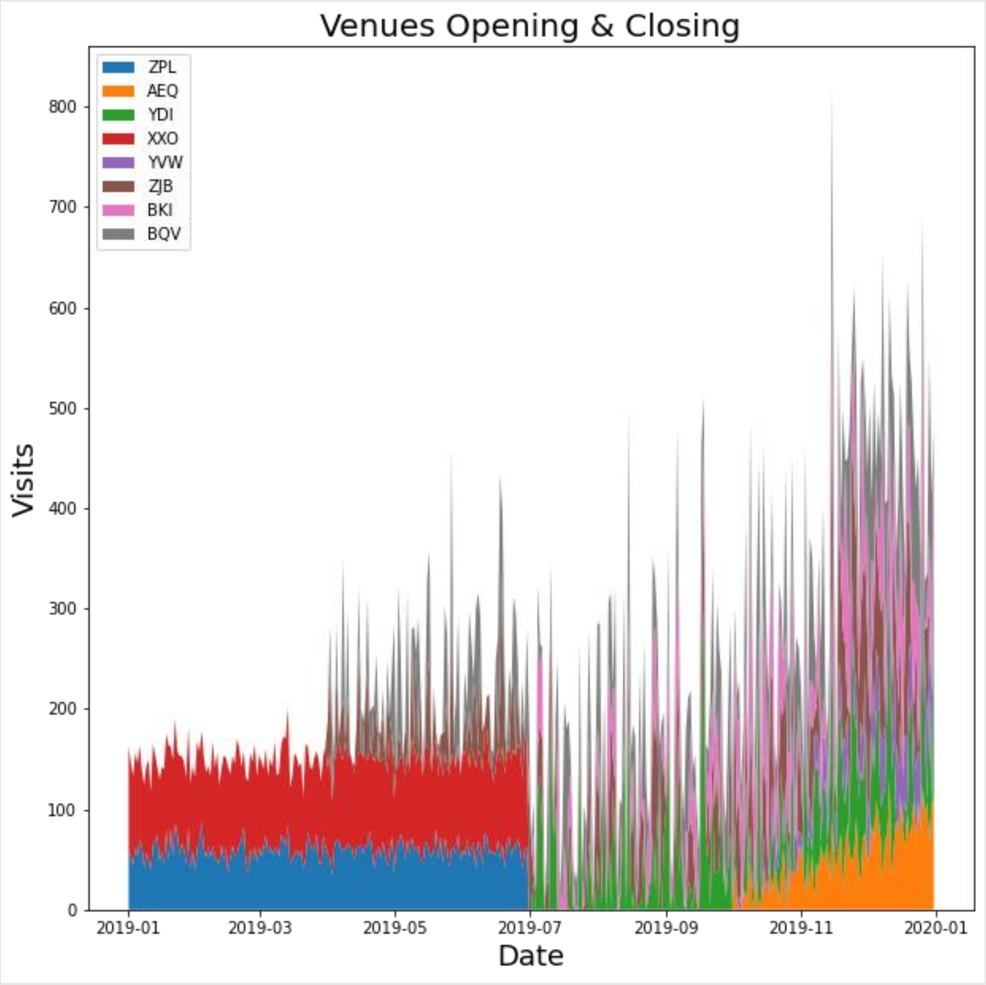

Figure 2 : Area plot for opened and closed venues

3.3 Distribution of High and Medium Volume Visitors with Box Plot

A box plot is a visual depiction of data distribution in five dimensions: minimum, first quartile, median, third quartile, and maximum. It also detects outliers, which are points that are outside of the lowest and maximum extrema.

Figure 3 depicts the year-round distribution of High and Medium Volume visitors. Outliers were found in the three venues RDA, PDT, and QRY.

The following details can be seen in the plot:

- ✓ SJU has the widest distribution among all the venues
- ✓ AWF and CQC are tightly distributed compared to other venues

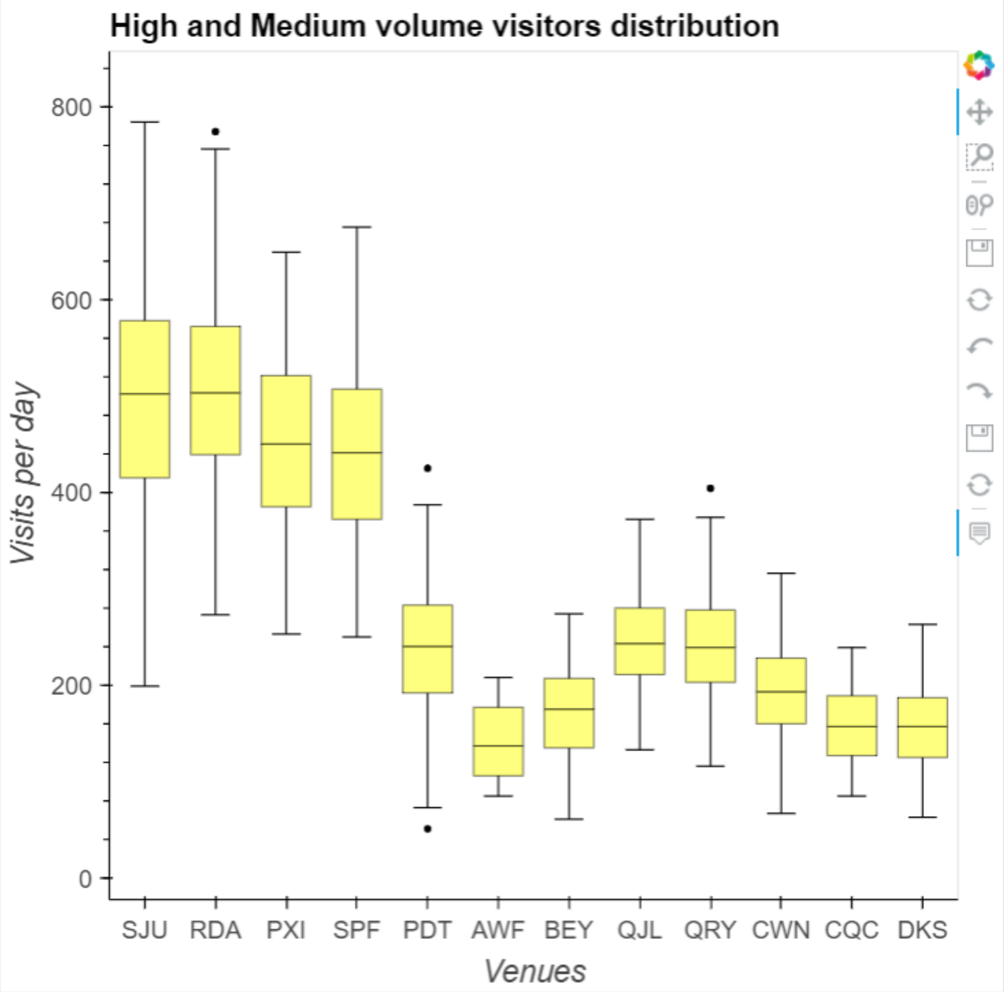

Figure 3: Box plot for High & Medium Volume Venues

3.4 Weekly Seasonality with Auto Correlation Plot

Each spike, whether decreasing or increasing from the dashed lines, indicates that it has a value other than zero; this phenomenon is known as Autocorrelation. Seasonal trends in retail stores will almost always be present. AutoCorrelation plots, which compute the randomness in data across a time series, may quickly discover these trends. As a result, this plot was chosen to demonstrate the possibility of seasonality.

The Autocorrelation plots for High and Medium Volume Venues are shown in Figure 4, with the lag limit set to 0 to 20 days. The figures show that all of the spikes are statistically significant, implying that the venues are significantly connected with itself in a 7-day lag. All of the high and medium volume products show signs of seasonality.

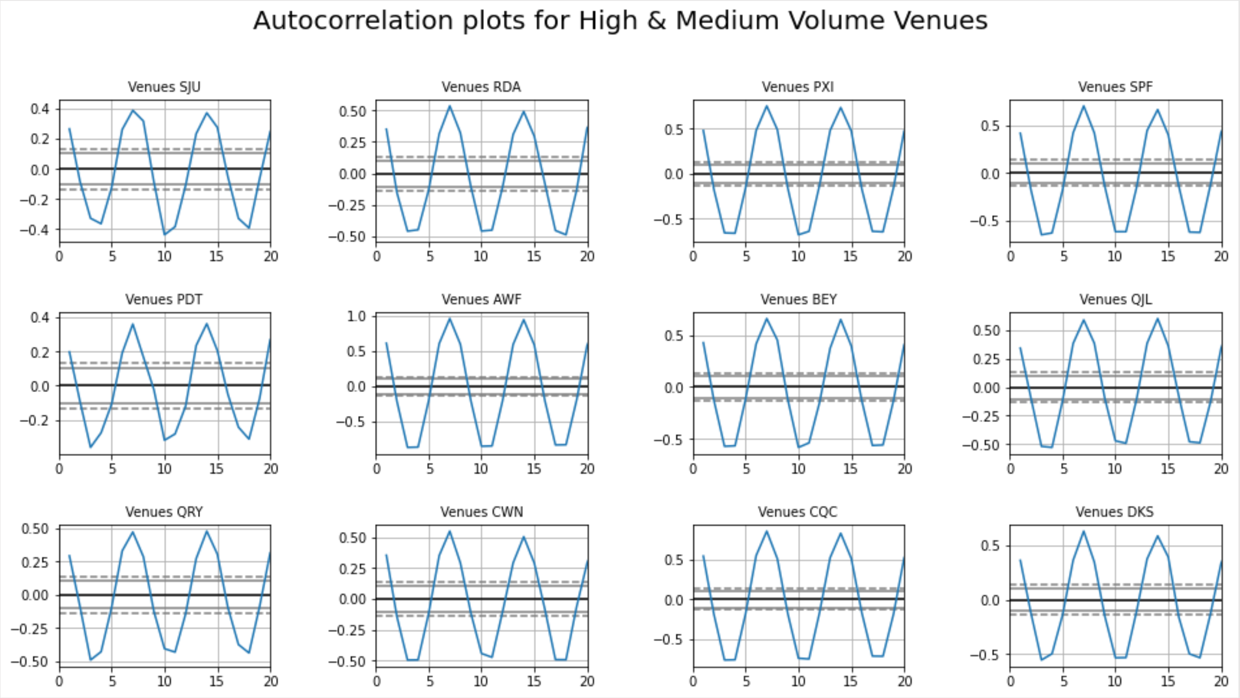

Figure 4: Autocorrelation plots for High & Medium Volume Venues

3.5 Heat Map of High & Medium Volume Venues

Heat Maps depict the effective correlation between one or more components and include the correlation coefficient. As a result, it displays all positive and negative correlations in a color-coded, easier-to-read format.

Figure 5 shows that the strongest positive correlations with correlation coefficients greater than 0.8 are AWF, PXI, SPF, and CQC. AWF and CQC are the most strongly positively correlated venues, with a correlation coefficient of 0.93.

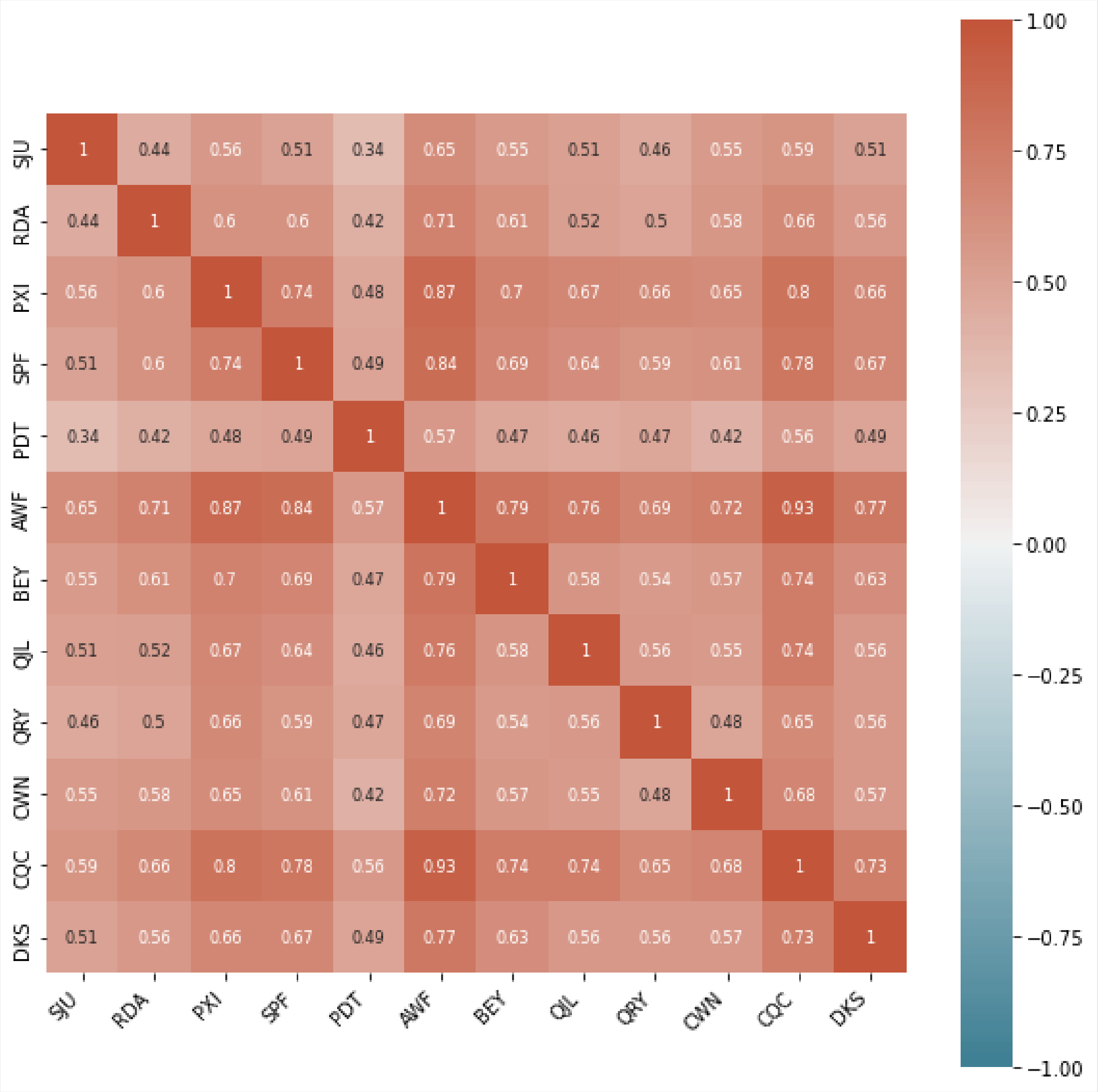

Figure 5: Heat Map of High & Medium Volume Venues

3.6 Pair Plot of Summary Data

Pair Plot or Scatterplot Matrix is a scatter plot matrix that compares distinct variables to one other. It's similar to a heatmap, however instead of color coding and presenting correlation coefficients, this graphic uses visual scatter plots. With a histogram drawn diagonally in the center, it can be used to detect correlation between variables.

Figure 6 shows that the total number of visits has a strong positive correlation with the maximum distance covered by visitors (Distance) (Visits). Furthermore, the average age of visitors (Age) and the average amount of money spent on venues (Spent) are highly correlated. Furthermore, there is a mild correlation between the average amount of time visitors spend in venues (Duration) and the percentage of female visitors (Gender).

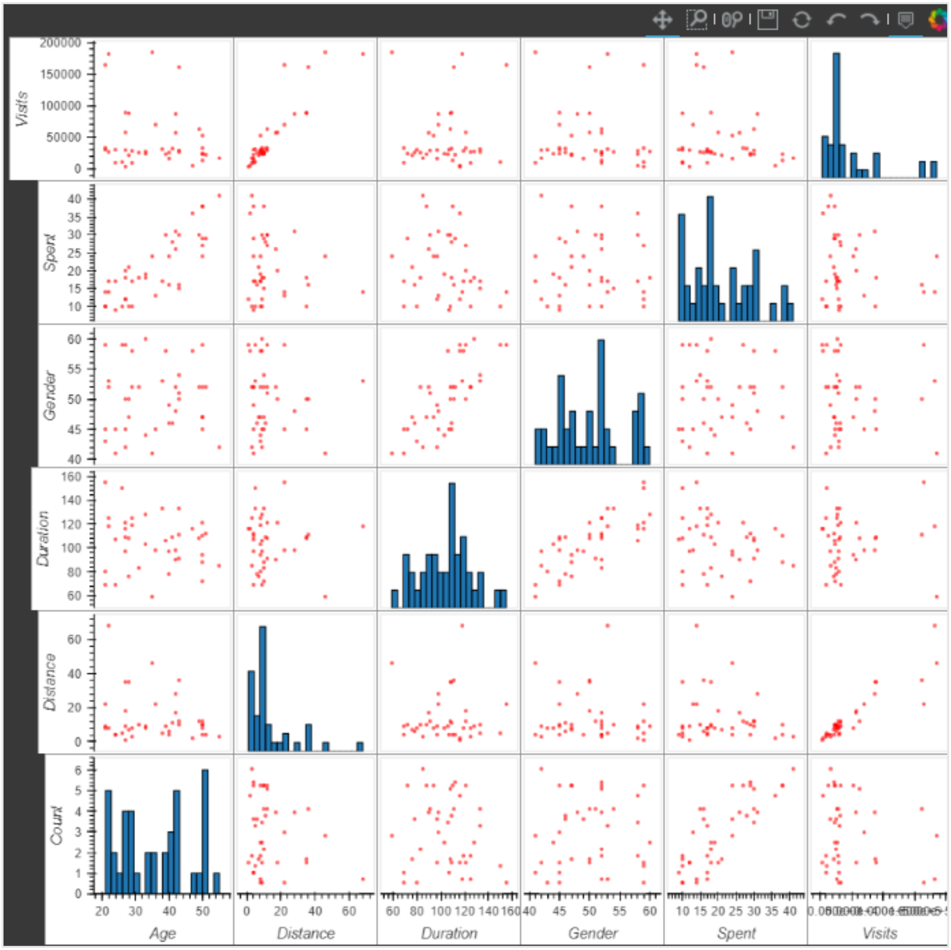

Figure 6: Pair Plot for Summary Data

3.7 Radar Plot of High-Volume Summary data

Multiple variables are compared inside their own axis over a group of values forming a polygon using a radar or spider plot. As a result, this plot was chosen for recognizing some form of pattern within high-volume venues. Figure 3 depicts significant similarities between Distance – Visits and Duration – Gender, as well as Time spent – Money spent.

Figure 7 shows that regardless of the distance that visitors must travel to reach the venue, the visitors are nearly identical in every location. Time spent at the venue where there are more female visitors also shows a tendency. Despite the fact that consumers spend less time at the venues, it appears that they spend more money. Furthermore, in Venues SJU and PXI, persons of the same age spent about the same amount of money.

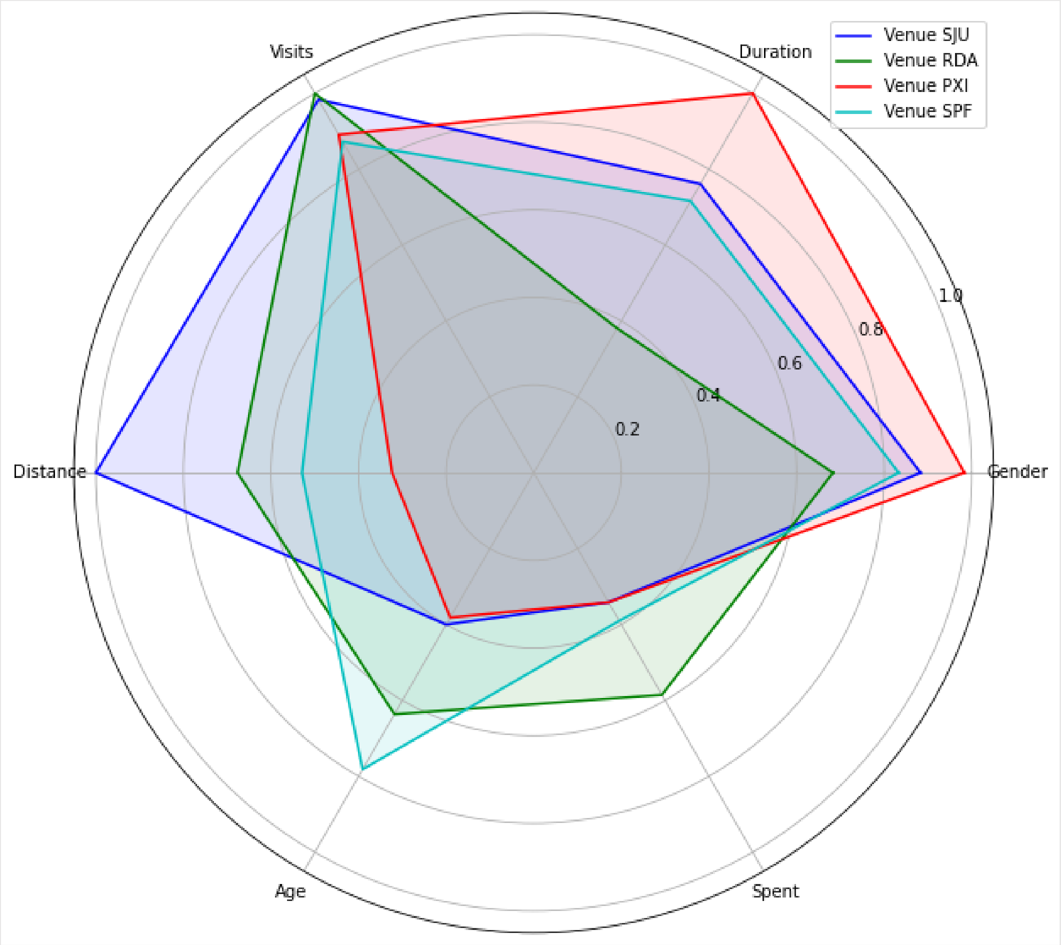

Figure 7: Radar Plot of High-Volume Venues

3.8 Comparative Bar Chart for High Volume

A comparative bar chart is used to compare items that have the same components but are distinct. This graphic was included because it depicts some key findings from the summary data.

For the various datasets from the company, Figure 8 reveals a lot of diversity among the high-volume venues. Figure 8 shows the following information:

- ✓ People spending time in Venue PXI is much higher than in other venues, but it is much lower in Venue RDA.
- ✓ In comparison to the other two venues, it appears that Venue SPF and Venue RDA attract a larger number of senior visitors.
- ✓ When compared to other venues, the money spent at Venue RDA is higher.
- ✓ The figure clearly shows that customers had to travel the greatest distance to Venue SJU and the shortest distance to Venue PXI.

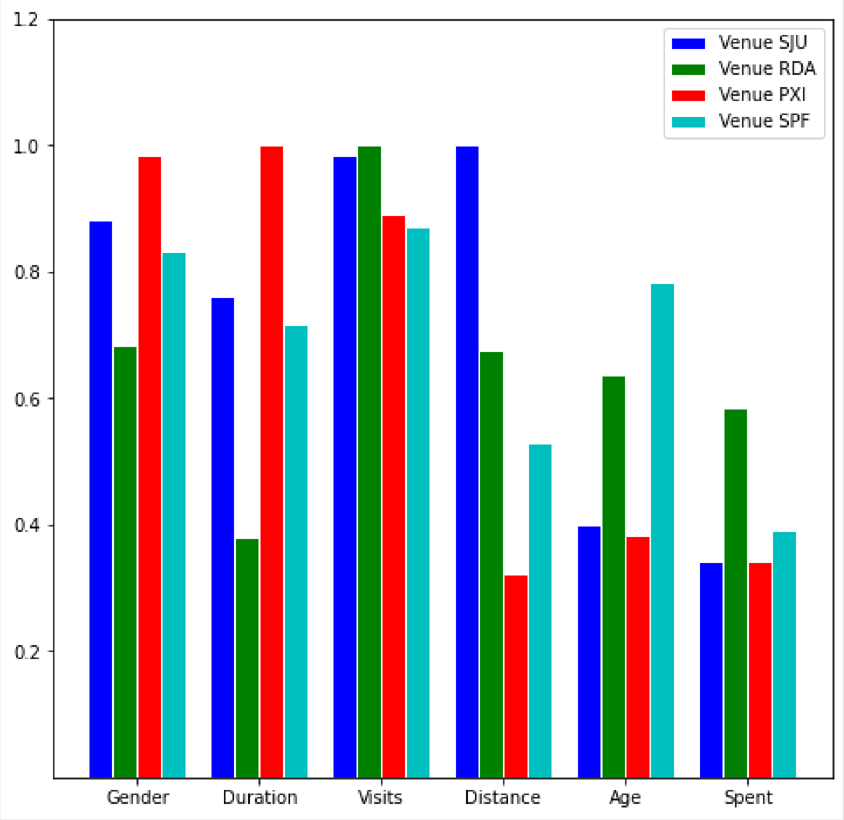

Figure 8: Comparative Bar Chart for High Volume

3.9 Summary of Low Volume & Very Low Volume Venues

In comparison to other categories, low volume data is very noisy and lacks significant information. However, all of the venues in the very low volume group opened or closed at different times throughout the year. More information about the anomalies can be found in the item **'3.2 Venues Opening & Closing'** above.

Figure 9 shows a heat map of low-volume sites with a lot of residual data. Figure 10 reveals that XXO, ZPL, YVQ, and AEQ are all highly connected venues. Furthermore, there is a slight negative correlation between ZPL and XXO against BKI.

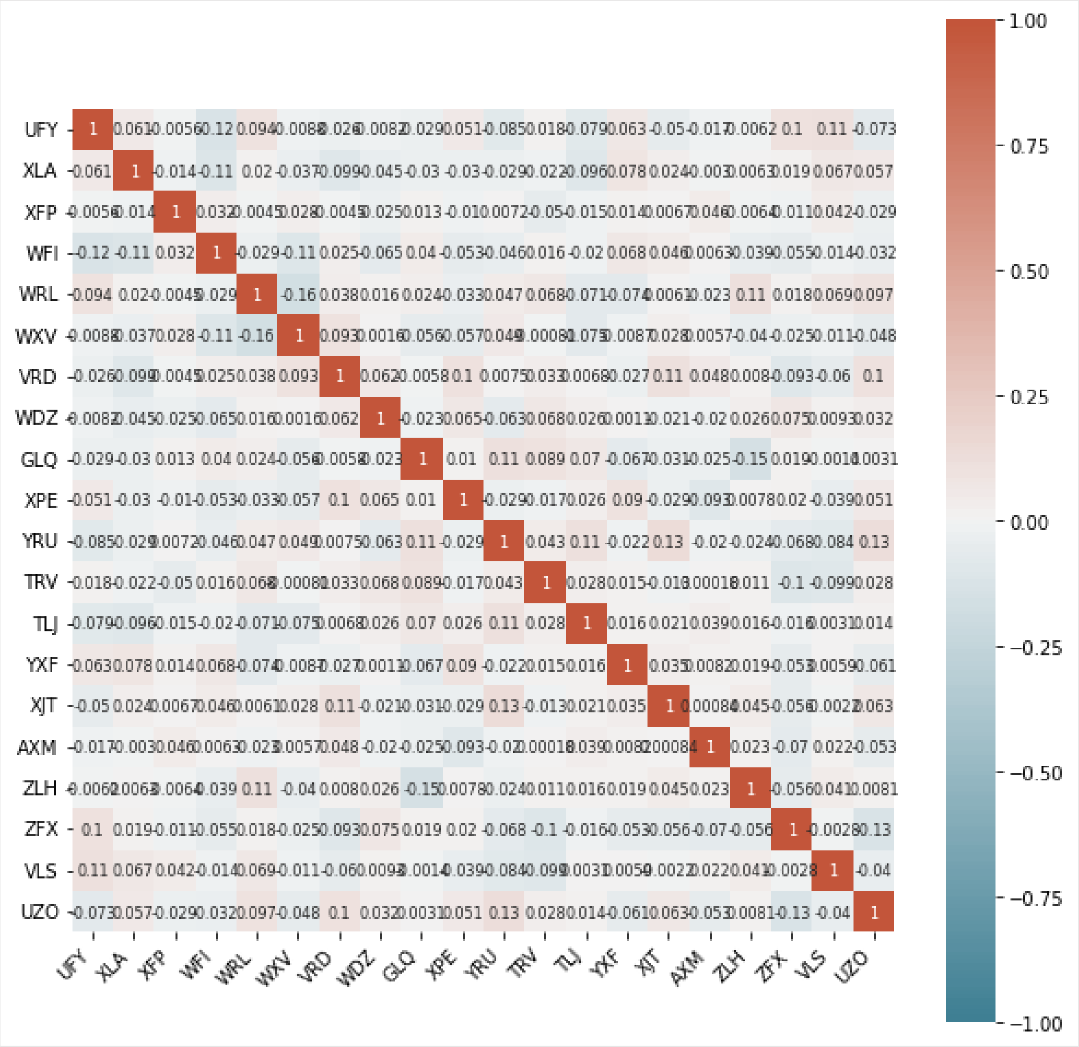

Figure 9: Heat map of low volume venues

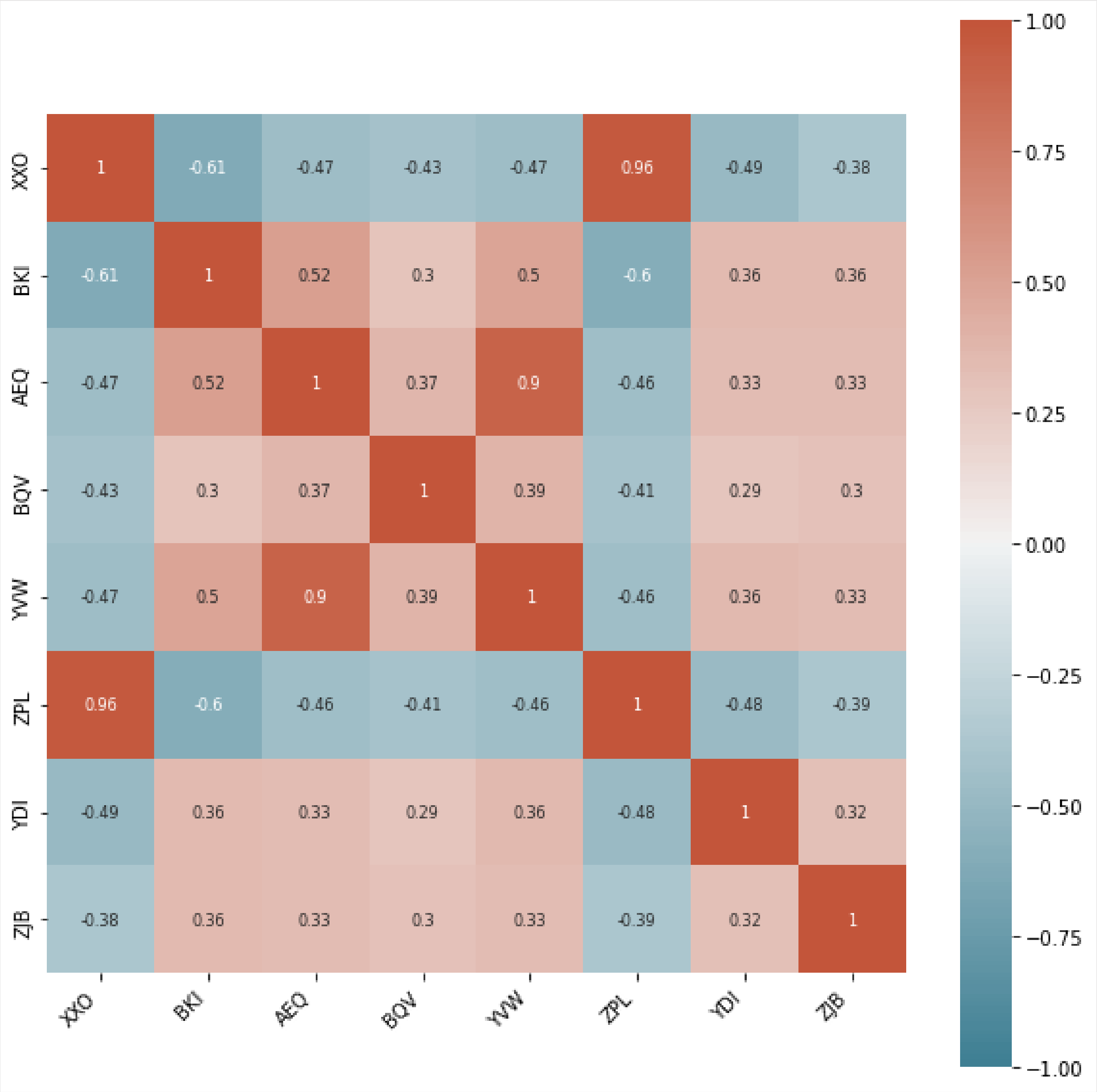

Figure 10: Heat map of low volume venues

4. Critical Review

To begin, the data was segmented using a python loop to divide the total number of visits into multiple groups, which is significantly more efficient than selecting locations manually. Second, monthly visitor data was reviewed to determine when venues closed and opened. Because the area plot was deemed to be the most effective for spotting anomalies, it was utilized to illustrate the exact month when the venues' plans were altered. To see if the dataset had any seasonality, the autocorrelation plot was utilized. The box plot was also displayed as an interactive graphic so we can zoom in on outliers and explore other components. Also, to display the positively strongly correlated venues, an interactive scatter-plot matrix was used, with hover and zoom in possibilities to study each data point, demonstrating the power of interactive visualizations. A radar plot and a comparative bar chart are also included, which show a variety of patterns and trends among the high-volume locations. Finally, rather than using heatmaps and scatter plots to find correlation, a pair plot was used, which does a good job of finding correlation just by looking at the matrix.

I would have looked at more visualizations not featured in this module if I had more time. I only examined visualizations that I was already familiar with from tutorials and lectures. In the same way, I would have spent more time creating dashboards, which could be more valuable.

Conclusion

To summarize, the closing and opening of ChrisCo venues across the United Kingdom had many ups and downs. Some outliers were observed in the high and medium volume venues. Furthermore, highly correlated locations were found and further investigated to see if any conclusions could be drawn. In all of the high and medium volume areas, weekly seasonality was also detected. Various trends and patterns were also observed among the summary data attributes. Finally, data from low-volume venues was too noisy to be analyzed, while very low-volume data revealed some substantial positive associations.

References

Sadiku, M. et al., 2016. Data Visualization. *International Journal of Engineering Research And Advanced Technology (IJERAT),* 2(12), pp. 11-16.