Table of Contents

Data Visualisation

Data visualization is the graphical representation of information and data. Data visualization tools that use visual elements such as charts and maps allow to see trends and outliers in the data.

The main purpose of data visualization; to make it easy to identify patterns, trends, and outliers in large data sets. The term is often used interchangeably, including infographics, information visualization, and statistical charts. Data visualization is also an element of the broader discipline of data presentation architecture that aims to describe, find, format and deliver data in the most efficient way possible. As a result, they can help with comprehension, accessibility, and memory. The process of transforming data into graphics such as maps, graphs, and charts is known as visualisation. These are then utilised to explore or explain data discoveries. The purpose is to support audiences in decision-making and analytical sense.

Data visualisation involves two different categories: exploration and explanation. The two serve different purposes, and so there are tools and approaches that may be appropriate only for one and not the other. The explanatory approach allows the author to guide the user, define the take-home message, steer the conversation while the exploratory approach invites users to ask questions, inspect the evidence, find their own answers.

Explanatory visualizations Explanatory visualisations depend more heavily on message and are author-driven with a specified route across the visualisation. When telling a story or communicating effectively, this method works best. Explanatory visualisations can be used to explain the major findings of an investigation, in which the data to be displayed is decided in advance and the presentation is based on the needs of the end user. The story it is trying to tell is known to the author in the first place, and therefore it can be specifically designed to accommodate and highlight that story. In other words, some editorial decisions have to be made about what information stays inside, what is distracting or irrelevant and should come out. This is the process of selecting focused data to support the story it is trying to tell.

Exploratory visualizations User-driven exploratory visualisations have no set sequence, little messaging, and a high level of interaction. These visualisations are frequently interactive, with a graphical interface that allows the reader to pick and limit particular parameters, allowing to uncover for itself whatever insights the dataset may have. By giving alternative viewpoints into the data to be evaluated, exploratory visualisations can help users to independently uncover patterns and untangle relationships. These might be new ideas that the visualization's creator hasn't yet considered.

Figure 1

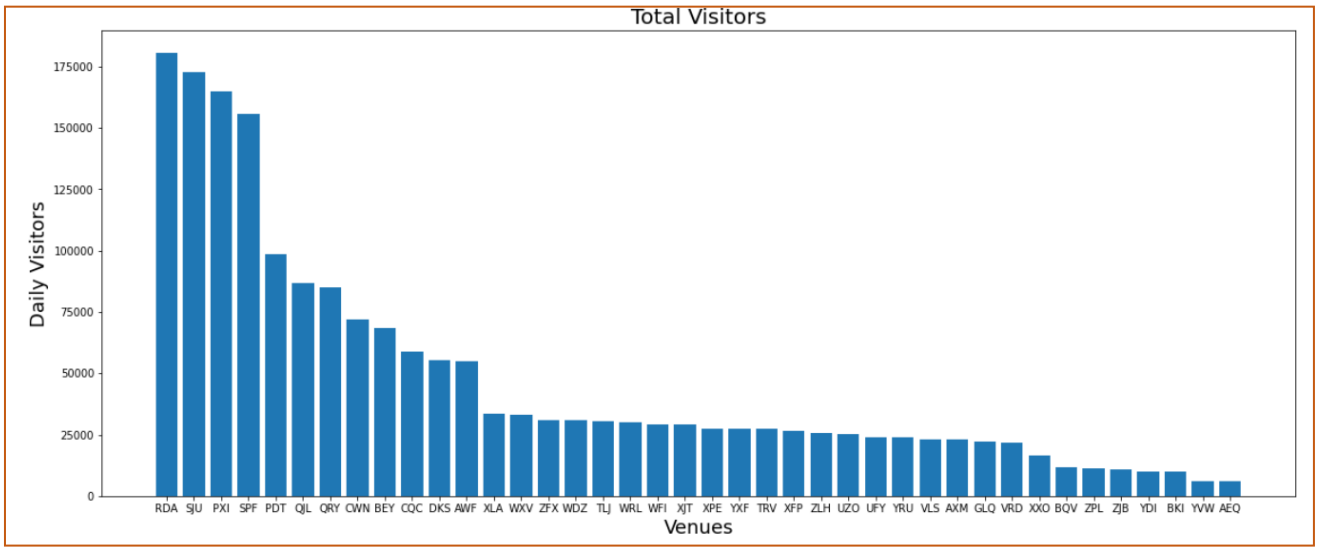


Figure 1 Total Visitors Throughout 2019

A Bar Chart, see Figure 1, is included in this report because bar charts are straightforward to interpret and have a clear relationship between size and value that allows for easy comparison. In addition, it helps in presenting very large or very small values more emphatically.

Figure 1 is a Bar Chart showing the number of visitors at 40 venues in 2019. The x-axis shows the venues, while the y axis shows the number of visitors these venues have received throughout the year. The data is "High Volume " venues with 150000 or more visitors in 2019, venues have between 54000 and 150000 visitors are "Medium Volume ", " Low Volume " venues with 25500 to 54000 visitors and all values below this value are segmented as " Very Low ". Only an overview of the data can be obtained from this figure. Apart from that, why high-volume venues get more visitors or what causes the big visitor difference between high volume venues and medium volume venues will be detailed in the rest of the report with other visual aids.

Figure 2

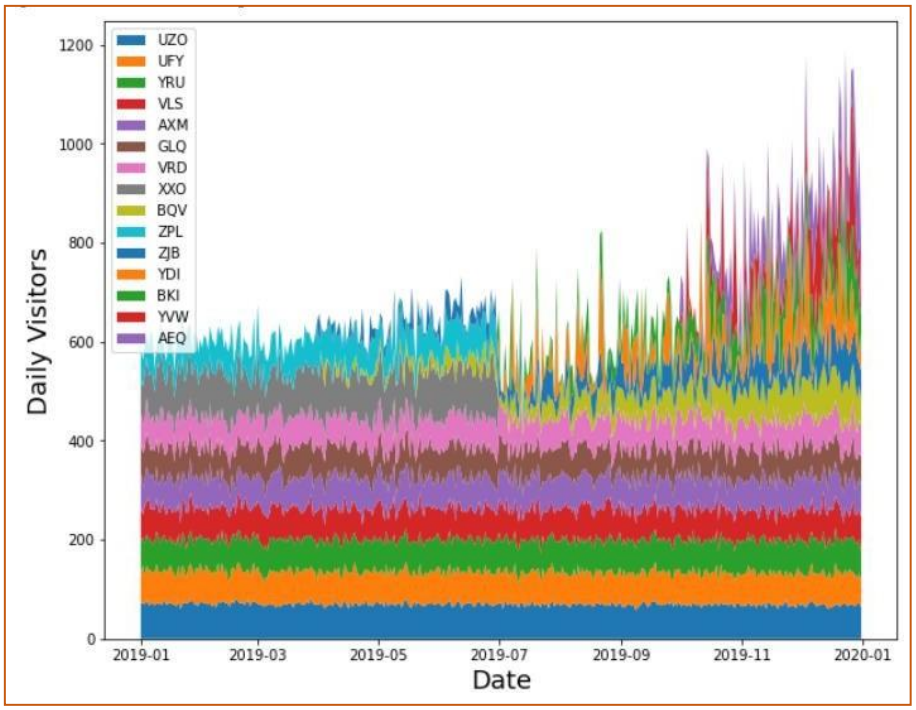


Figure 2 An Area Plot Showing Total Visitors of The Low Volume Venues

Area Plot, is quite revealing like a line plot, but rather than putting all lines on top of each other and superimposed on top of each other, it can be stacked one on top of each other and be filled in the area under each line. With Area Plot, we have a much clearer picture of how individual venues receive visitors. However, it is not known how the venues are in terms of visitors as a whole, Area Plot was used to get answers to more detailed questions such as at which times of the year more visitors are received or to determine the places that are not active during certain periods of the year.

With Figure 2, the visitors received by low volume venues throughout 2019 are shown as above. As seen in the graph, the number of visitors received by some venues has increased overall. It is seen that while there are around 500 visitors per day in the beginning, there are venues that receive close to 1000 visitors per day at the end of the year.

In addition, some anomalies can be detected in the graph. Some venues appear to be unavailable at certain times of the year. Venues at the bottom of the graph are venues that receive more visitors than other venues. For example, Venues ZPL and XXO receive approximately 500 visitors per day in the begining, and it appears to have stopped serving as of 2019-07. BQV and ZJB started to serve after 2019-03, and there was a noticeable decrease in these venues in 2019-07. While YDI-BKI was an unavailable venue in the first seven months of the year, it started to serve after this date, and there was a slight increase in the number of visitors towards the end of the year. Finally, AEQ-YVW venues started to serve after the ninth month of the year.

Figure 3

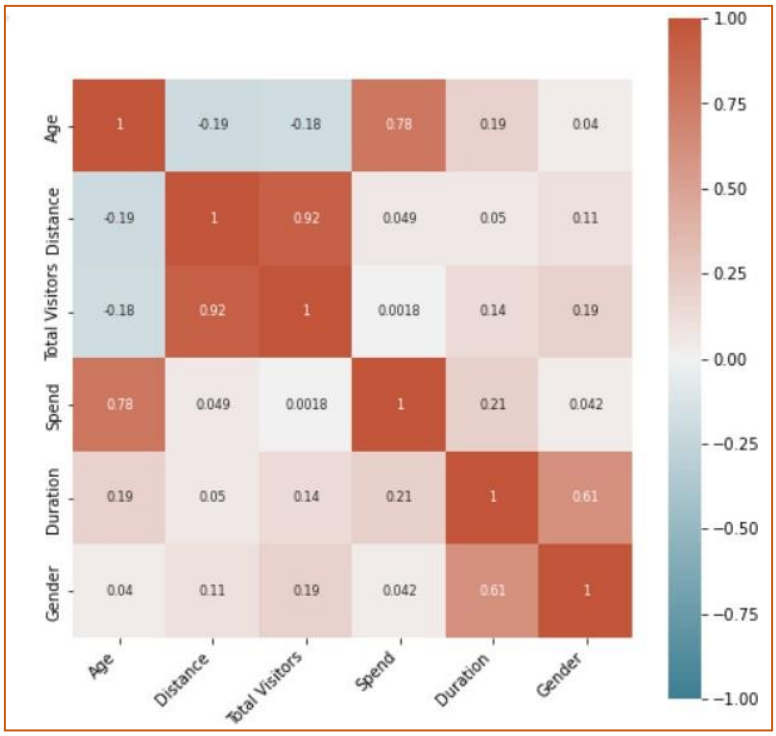


Figure 3 A Heat Map for All Categories

Correlation heatmap created to understand the linear relationship between different variables in the data. The above heatmap aims to check the relationship between “Age”, “Distance”, “Total Visitors”, “Spend”, “Duration”, and “Gender” categories how they move concerning to each other. Correlation heatmap contain the same information in a visually appealing way. It shows in glance which variables are correlated, to what degree, in which direction, and alerts us to potential multicollinearity problems. Discovered, the results of the correlations can then be examined to find out what caused this correlation.

Figure 3 shows that there is only one strong correlation between “Distance” and “Total Visitors” with a value of 0.92. The heatmap shows that as the distance length increases, the total number of visitors increases at the right rate. Due to the popularity of the venues, long distances may be travelled more by visitors. The other high correlation after 0.92 positive correlation value is between “Spend” and “Age”. This shows that older visitors spend more on their visits than younger visitors. The blue boxes in the heatmap indicate a non-strong inverse correlation between the values of 0.18 and 0.19 and the “Total Visitors” - “Age” and “Distance” - “Age” categories. The fields with negative correlation between six variables are only these fields. Although at a low rate, this shows that older people visit less frequently and visit distant venues less than younger people.

Figure 4

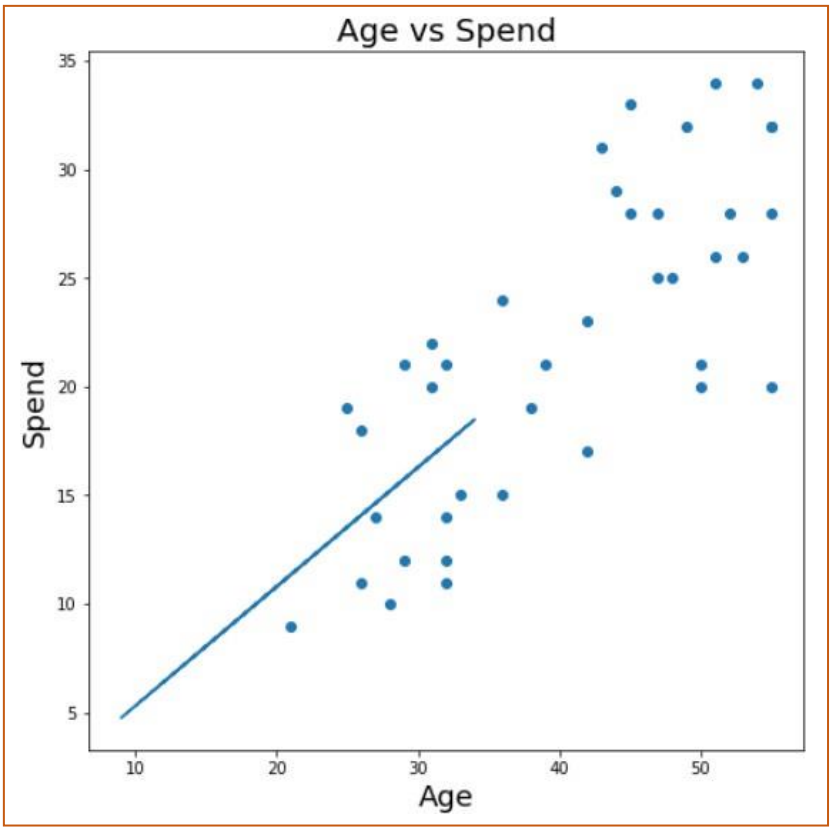


Figure 4 A Scatter Plot of the Age against the Spend

A Scatter graph is used to show the potential correlation between two different variables. A Scatter graph can be used when the data from both variables under investigation is continuous. With Scatter Plot it is aimed to examine the connection between "Age" and "Spend" categories in detail.

The data points or dots, which appear in Figure 4, represent each of the venues' individual values and allow pattern identification when looking at the data holistically. The Scatter Graph shows that when looking at the locations of the dots, it is seen that there is a positive correlation between Spend and Age, supporting the Heatmap. They get a very close together straight line that means that they correlated. As can be seen from the graph, older people spend more in the places they visit.

Figure 5

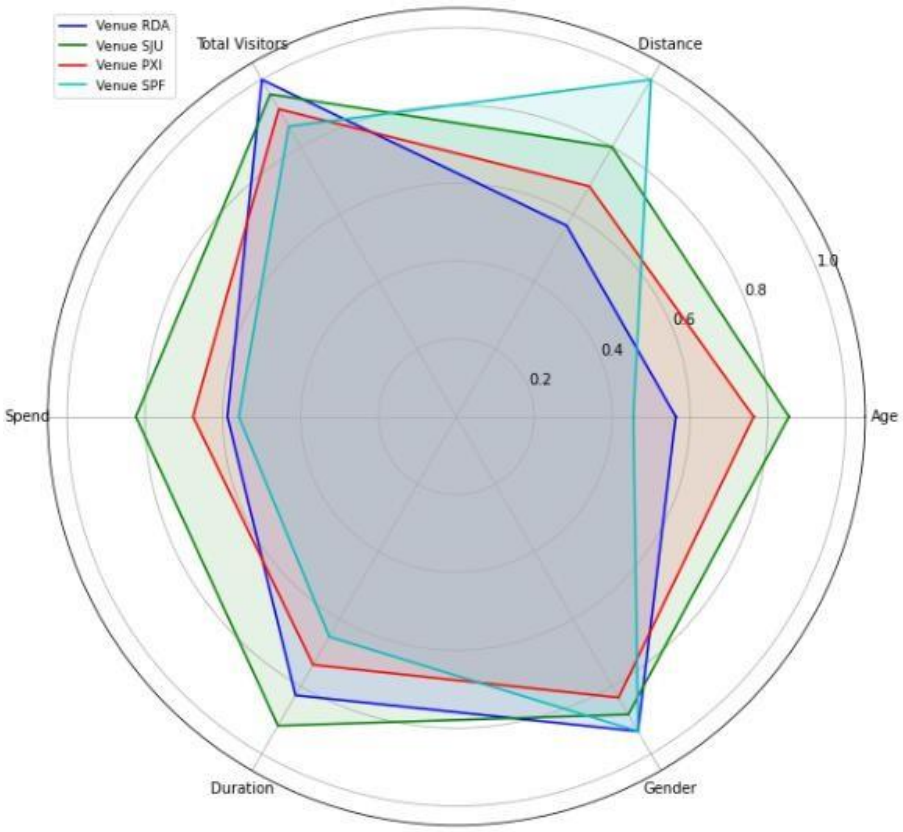


Figure 5 A Radar Plot of the High Visited Venues

Radar Chart allows us to compare all categories of venues 'Age', 'Gender', 'Total Visitors', 'Spend', 'Duration', 'Distance', rather than just comparing one category to another. It more clearly identifies other relationships between categories that may have been missed in other charts.

Figure 5 is a chart of venues with a high number of visitors compared in six categories during the year. The graph shows that even though RDA is the venue with the most visitors, less spend is spent at this venue than SJU, and less time is spent. This may be because that SJU may contain a section similar to a shop where visitors can spend more. In addition, although SPF receives the least number of visitors compared to others, it is seen that it mostly receives younger visitors. This may also be that visitors are more distant from the venue than others.

Also, Figure 5 supports the positive correlation that was shown in Figure 3 between the “Distance” and “Total Visitors”, as can be seen in the RDA and SJU venues. The correlation between the Gender and Duration categories, which was detected as a small correlation of 0.61 in Figure 3, is seen more clearly with Radar Plot. The correlation would not be displayed as strongly within the Heatmap. However, it can be seen that female visitors stay more in these venues because the venues with high visitors shown in Radar Plot appeal to female visitors more.

Figure 6

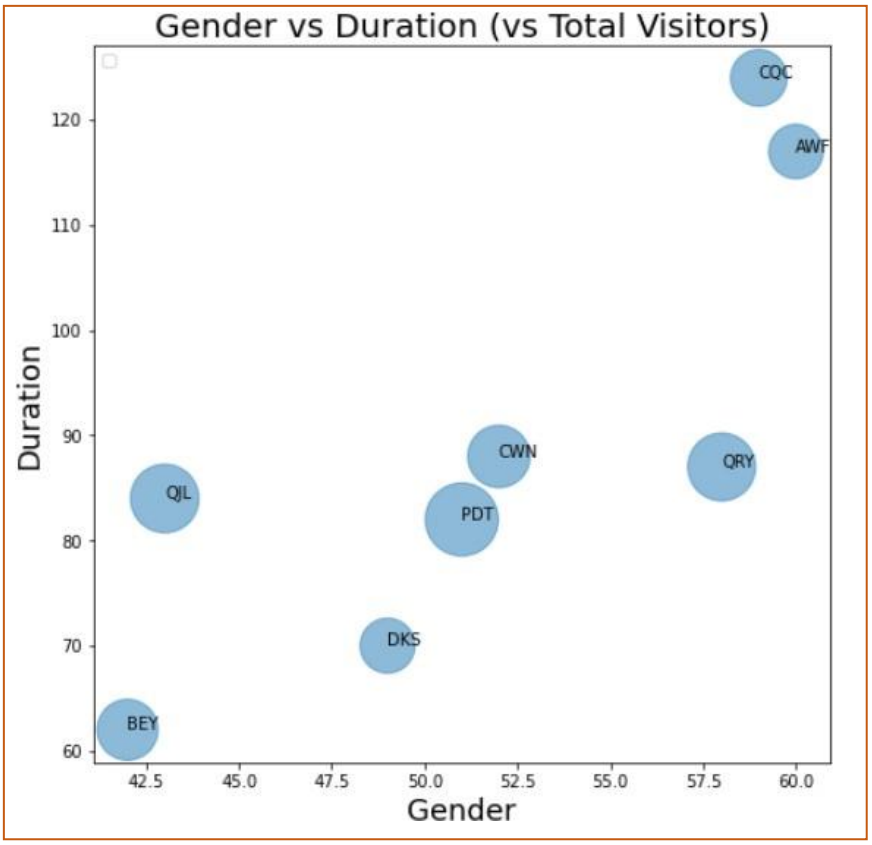


Figure 6 A Bubble Plot of Duration and Gender for Medium Volume Venues

It is intended to show the relationship between the bubble plot dimensions "Duration," "Total Visitors," and "Gender". The addition of bubble size as a dimension allows comparisons to be made between three variables instead of simply two. Three separates pairwise comparisons (Duration vs. Total Visitors, Total Visitors vs. Gender, Duration vs. Gender) as well as an overall three-way comparison may be made in a single bubble chart. To obtain the same number of insights, numerous two-variable scatter plots would be required; even then, inferring a three-way relationship between data points would not be as straightforward as in a bubble chart.

Figure 6 shows that when we look at the locations of the bubbles, it is seen that there is a positive correlation between “Gender” and “Duration”, supporting Figure 3 and Figure 5. The correlation between these 3 categories can be clearly seen from the size and location of the bubbles. As the proximity of the bubbles and the number of female visitors increase, the proportional increase in duration in the venue once again supports the positive correlation. (Gender: the percentage of visitors identifying as female at each venue). This indicates that women visit medium volume venues more and keep their visits longer than male visitors.

Figure 7

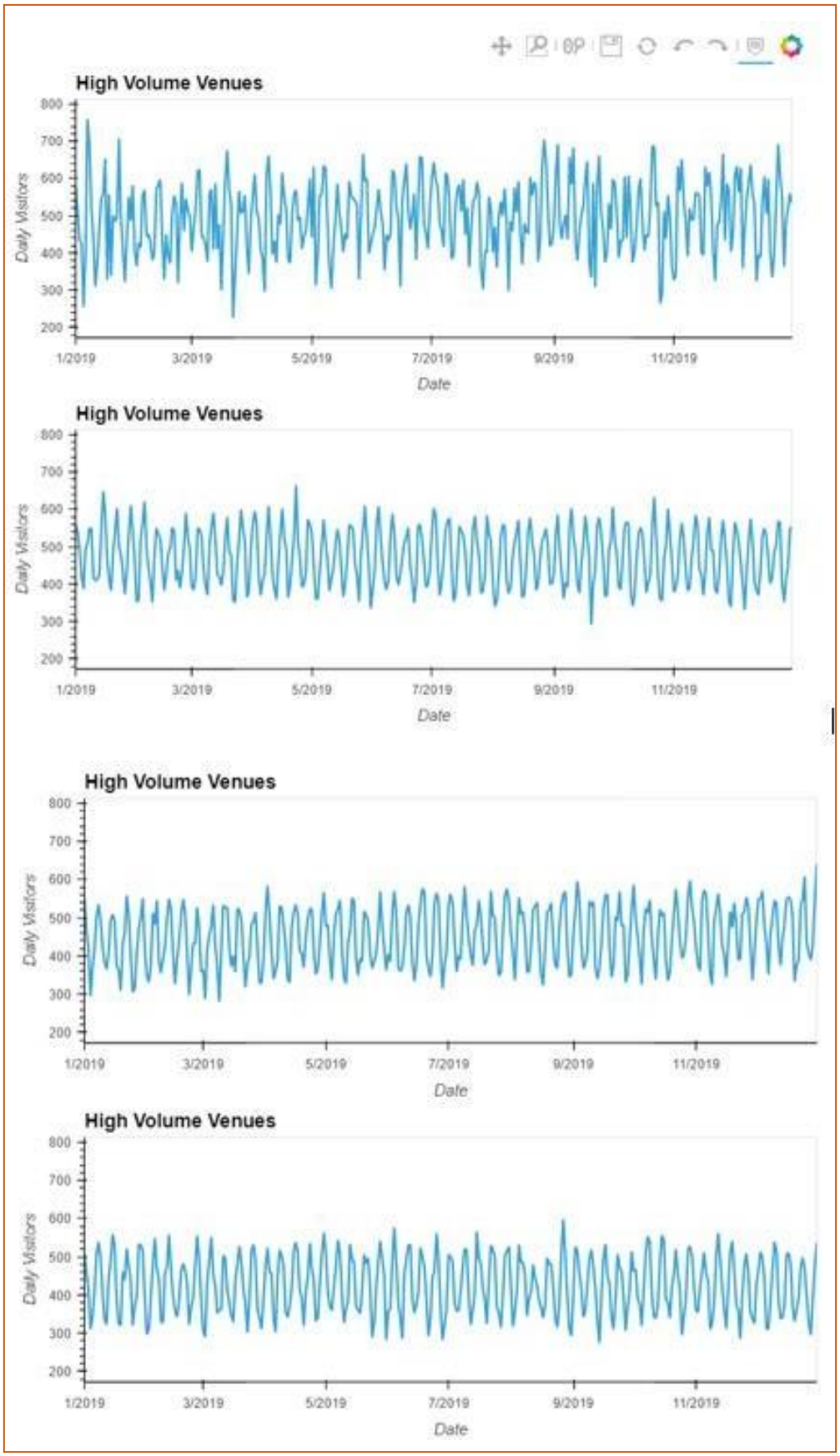
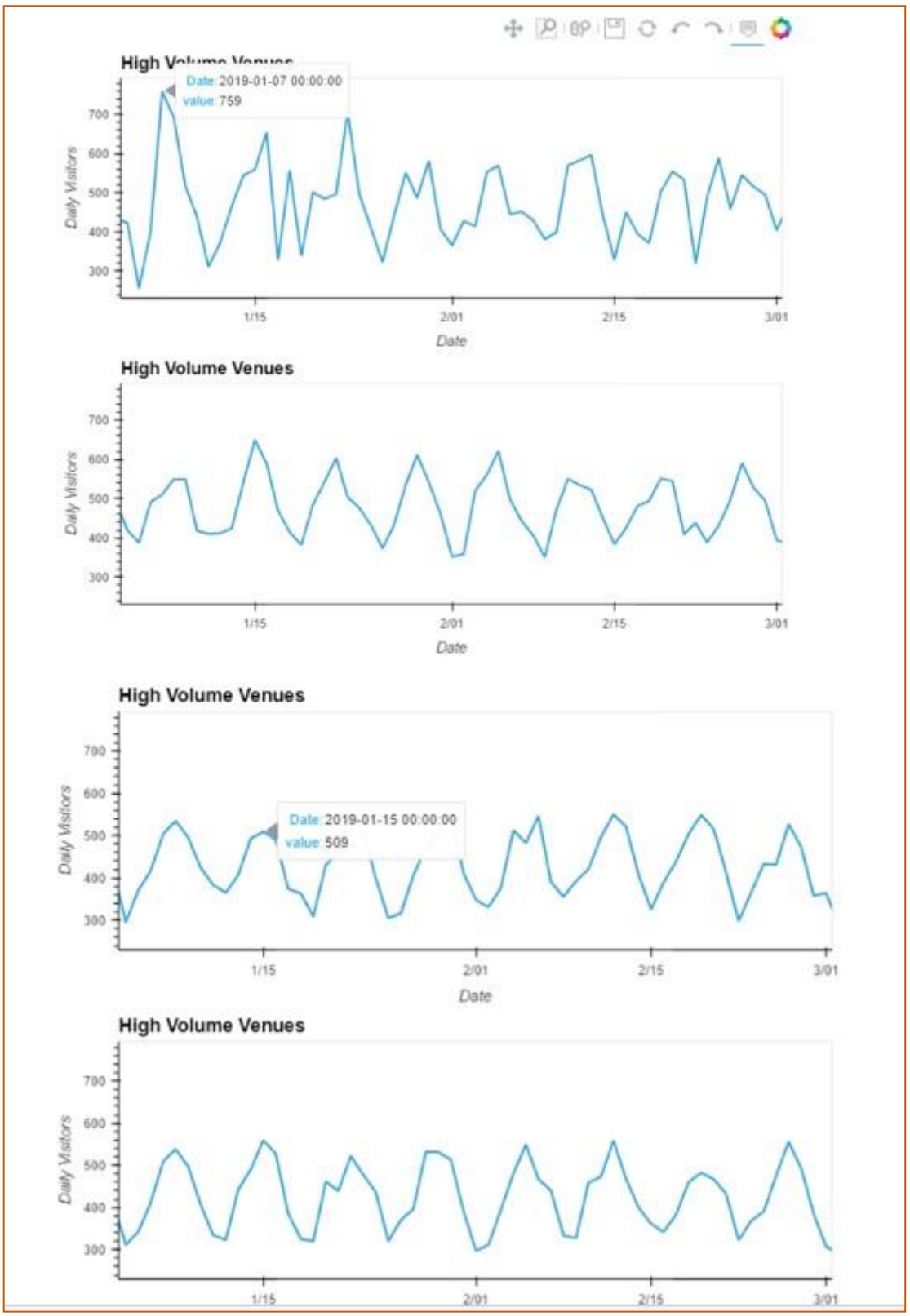


Figure 7 A Line Plot for High Volume Venues

The interactive Line plot above allows the freedom to fully explore analysed data. Interactive visualization can take the concept a step further by interactively changing the data seen and how it is processed to examine charts and graphs for more detail. The tools (zoom, save etc.) in the upper right part of Figure 7 allow us to analyse the Graph by months or even weeks.

Figure 7 shows the number of daily visitors the four high-visiting venues receive throughout the year. As can be seen in Figure 7, the Graph shows that four venues have a regular behaviour throughout the year, but with minor changes. As can be seen, there have been many sudden ups and downs in the number of visitors throughout the year. When we zoom between 1/2019 and 3/2019 with the zoom feature of the interactive Graph, the 2-month line graph is obtained. It is seen that the number of visitors to venues regularly peaks every week.


High Volume Venues
Date: 2019-01-07 00:00:00
value: 759
Daily Visitors
700
600
500
400
300
1/15
2/01
2/15
3/01
Date
High Volume Venues
Daily Visitors
700
600
500
400
300
1/15
2/01
2/15
3/01
Date
High Volume Venues
Date: 2019-01-15 00:00:00
value: 509
Daily Visitors
700
600
500
400
300
1/15
2/01
2/15
3/01
Date
High Volume Venues
Daily Visitors
700
600
500
400
300
1/15
2/01
2/15
3/01

They all got the peak in the same place. The venues have troughs and peaks fairly consistent intervals throughout the year. There may be a case of more visits on weekends rather than at the beginning of the week. So, it shows that there is some weekly seasonality. It can be detected easily by using the tools that are shown in the righthand side of the Figure 7.1.

Figure 8

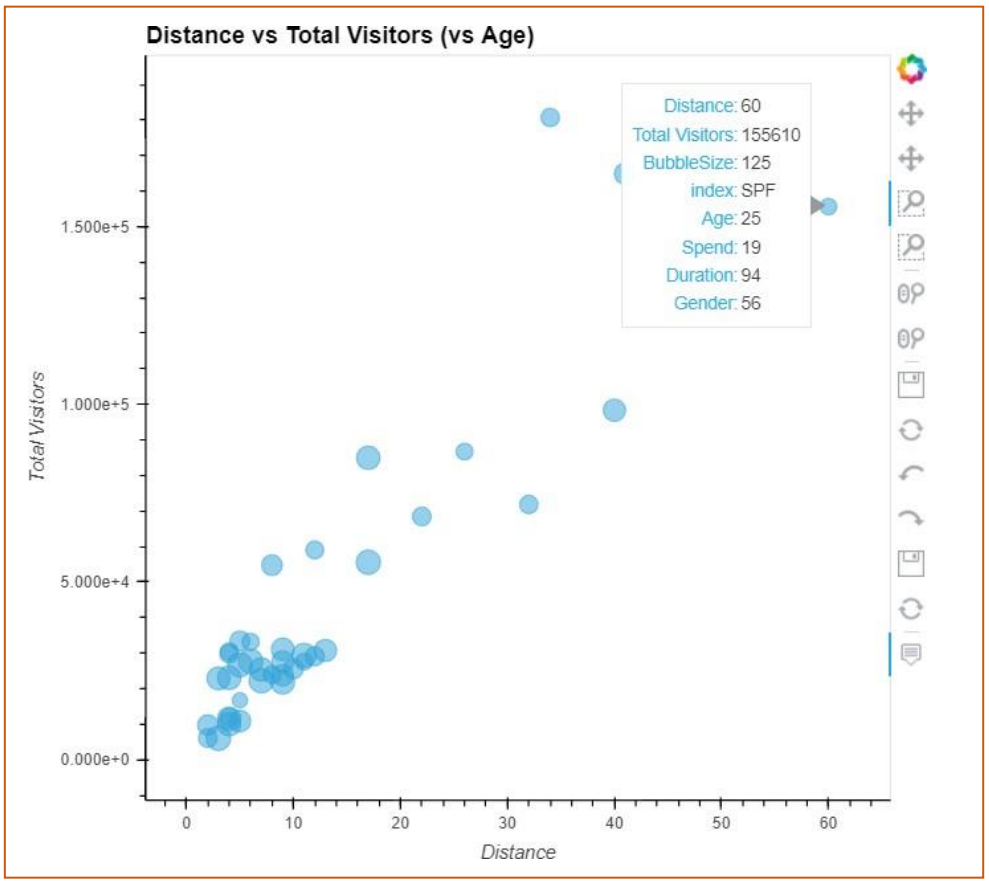


Figure 8 A Bubble Plot of Total Visitors and Distance

It is intended to show the relationship between the bubble plot dimensions "Total Visitors," "Distance" and "Age". The addition of marker size as a dimension allows comparisons to be made between three variables instead of simply two. Three separates pairwise comparisons (Total Visitor vs. Distance, Total Visitors vs. Age, Distance vs. Age) as well as an overall three-way comparison may be made in a single bubble chart.

The Bubble Graph shows that there is a positive correlation between "Total Visitors" and "Distance" when looking at the locations of the bubbles, supporting the Heatmap. However, looking at the density of the bubbles, it can be seen that the venues with high distance appeal to the young audience. This may be due to the fact that older people find it challenging to travel long distances. The venues with longer distances attract more visitors and appeal more to the younger crowd.

Critical Review

The change in the number of visitors during 2019 under specific categories of the 40 venues specified in the study was examined. Since the data presented is too large to be analysed as a whole, it is divided into sections; the venues are divided into four segments high, medium, low and very low, according to the total number of visitors they receive in a year. While examining the data visually, many visualization methods have been tried in order to detect correlations, outliers and seasonal

behaviour appropriately, and the visualization methods that give the most apparent results and are interconnected have been selected for the report. A total of 8 visualizations were used in this study: "Bar Chart", "Area Plot", "Heat Map", "Scatter Plot"," Radar Plot"," Bubble Plot", "Interactive Line Plot", and "Interactive Bubble Plot".

The Bar Chart is included in this report because bar charts are very easy to interpret and have an obvious relationship between size and value that allows for easy comparison. It helps in presenting very large or very small values more emphatically. By sorting the data, the total number of visitors received by the venues in a year is clearly shown for each segment. Figure 2 Area Plot was used to get answers to more detailed questions such as at which times of the year more visitors are received or to determine the places that are not active during specific periods of the year. With Area Plot, some anomalies were detected in very-low volume venues. Some venues have been found to be unavailable at certain times of the year. For example, the ZPL-XXO venues seems to be inactive after the seventh month of the year.

In order to obtain more satisfactory results, the connection between other categories and how these categories affect each other were examined. The correlation between the categories "Age", "Distance"," Total Visitors"," Spend"," Duration", and "Gender" were visualized with a Heatmap. The results of the discovered correlations were examined to find out what caused this correlation. After the correlation between categories was refined, the Scatter Graph was used to examine the connection between the "Age" and "Spend" categories, supporting the Heat Map. A radar chart was used in the next step in order to more clearly identify the relationships that might have been missed in other charts. For this reason, Bubble Plot was used in the next step, which provides the opportunity to compare three different categories with a single graphic. This shows that women visit medium volume venues prominently and keep their visits longer than male visitors.

In addition, Figure 6 An interactive Line Plot was used. The tools (zoom, save etc.) graphic on the upper right of Figure 7 has enabled it to be analysed according to months or even weeks. When the graph is analysed for two months with the zoom feature, the weekly seasonality is determined. Finally, the relationship between the Interactive bubble plot (Figure 8) and "Total Visitors", "Age" and "Distance" has been tried to be detailed. While venues with longer distances received more visitors, it was seen that they appealed more to young people.

The visualisations used helped to examine the data in more detail. Anomalies in the data, seasonal behaviour, and correlations between categories are properly detected.

Conclusion

Below are the results obtained with eight visualisations used to analyse the number of visitors received by high volume, medium volume, low volume and very low volume venues throughout the year and the factors affecting this:

- Some anomalies have been detected. Some of the venues with very low visitors are operating during some times of the year and inactive at other times of the year. Venues ZPL and XXO appears to have stopped serving as of 2019-07. BQV and ZJB started to serve after 2019-03, and there is a noticeable decrease in these venues in 2019-07. While YDI and BKI are an unavailable venue in the first seven months of the year, it started to serve after this date, and there was a slight increase in the number of visitors towards the end of the year. Finally, AEQ-YVW venues started to serve after the ninth month of the year
- Strong correlation has been identified between “Distance” and “Total Visitors”. Due to the popularity of the venues, long distances may be travelled more by visitors.

- Positive correlation has identified between “Spend” and “Age”. Older people might spend more in the places they visit.
- BKA is the most visited venue, where less was spent than SJU and at the same time visitors spent less time. This may be because SJU may include a section similar to a gift shop where visitors can spend more. SPF receives the least number of visitors compared to others, it is seen that it mostly receives younger visitors. This may also be due to the fact that visitors are more distant from the venue than others.
- Positive correlation identified between “Gender” and “Duration”. This indicates that women visit medium volume venues prominently and keep their visits longer than male visitors.
- Weekly seasonality has spotted. It is observed that the number of visitors to high volume venues peaks every week. There may be a case of more visits on weekends rather than at the beginning of the week.
- The relationship between "Total Visitors", "Distance", "Age" was compared. It shows that the venues with longer distances might attract more visitors and appeal more to the younger crowd.