COMP- 1800 Data Visualisation

Name: Yogi Bhoot

ID: 001189309

University of Greenwich

8th April 2022

Module Leader – Prof. Chris Walshaw

School of Computing and Mathematical Science

Contents

- ## Introduction

Data visualisation is a representation of data and information. There are two types of visualisation. One is Data representation which means data exploration and data explanation. Another one is infographic representation which means a combination of information and image presentation. Data visualisation is an easy way to understand data characteristics and make data-driven decisions. By using visual elements like pie, line, bar, and map. There are many ways that have provided different solutions to represent one data. Each graph uses for different purposes like a bar graph used for normalizing data, a line graph used for overall trends such as upward and downward, and seasonality which gives patterns in time series. Also, there are many plots used for outliers, correlation, etc.

Data Visualisation is very important because the human brain is not enough to understand and recognize raw data and noisy patterns. So, data visualisation tries to arrange data and make understandable and usable information. Data visualisation takes raw data from various sources and makes a model that helps the company to make a conclusion and decision. Moreover, we identified which areas need to be improved and what types of measurements need to be applied to make a good lifestyle and profit for the company.

- ## Data Visualisation 1

➢ Visual 1 :

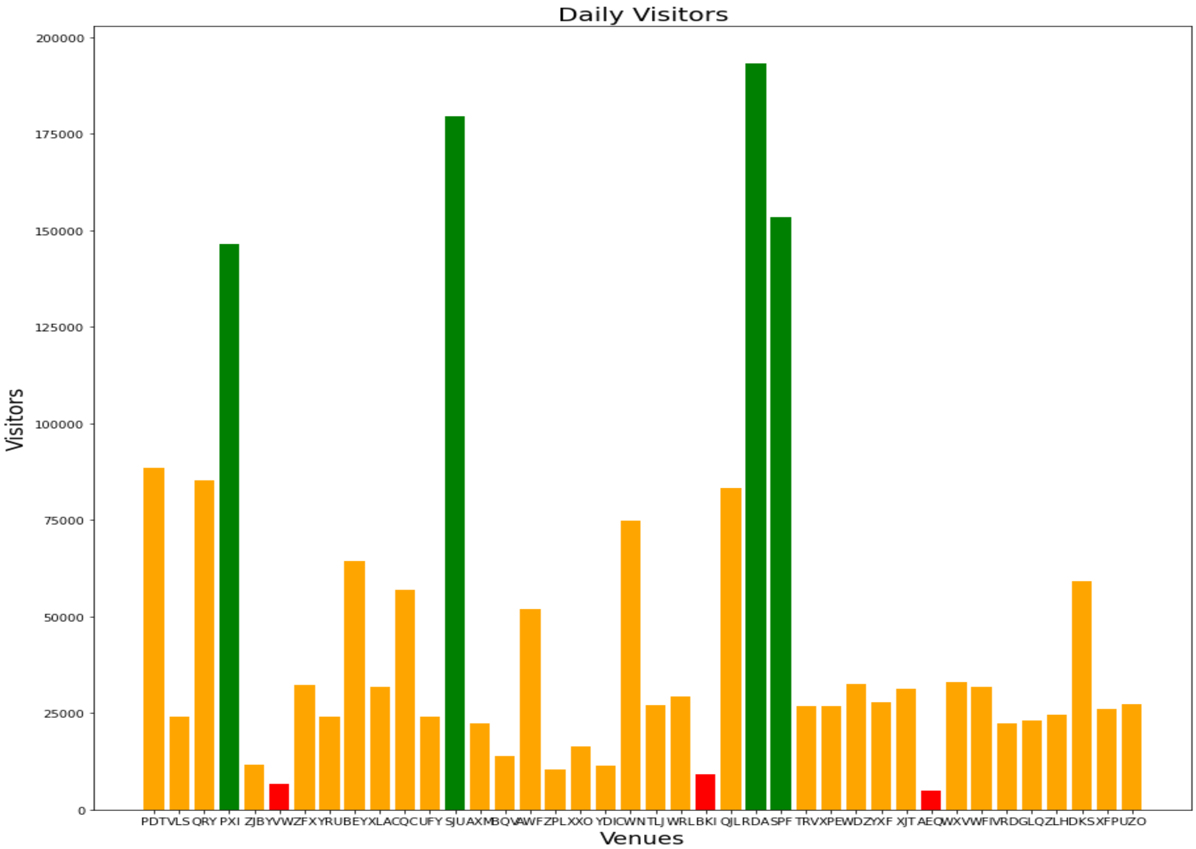


Figure 1 category wise bar graph

➢ Justification :

A bar graph helps us to compare venues given period. Here, we need to know which venues are higher, average, and lower. So, we implement a bar graph using the total number of visitors that has the best way to find venues. In addition, we plotted graphs using colors. So, we clearly understand the graph.

➢ Description :

We visualised the bar graph daily visitor VS date and gave specific visitor numbers to divide into categories. Also, we provided colors to categories which are green for high visitor venues, orange for average visitor venues, and red for low visitor venues. On the x-axes, we can see the venue, and on the y-axes, we can see the number of visitors. According to the bar graph, we can decide on venues that have higher, average, and low visitors. According to our bar graph, we have 4 green venues, 3 red venues, and the rest orange venues that show high, medium, and low venues.

- ## Data Visualisation 2

➢ Visual 2:

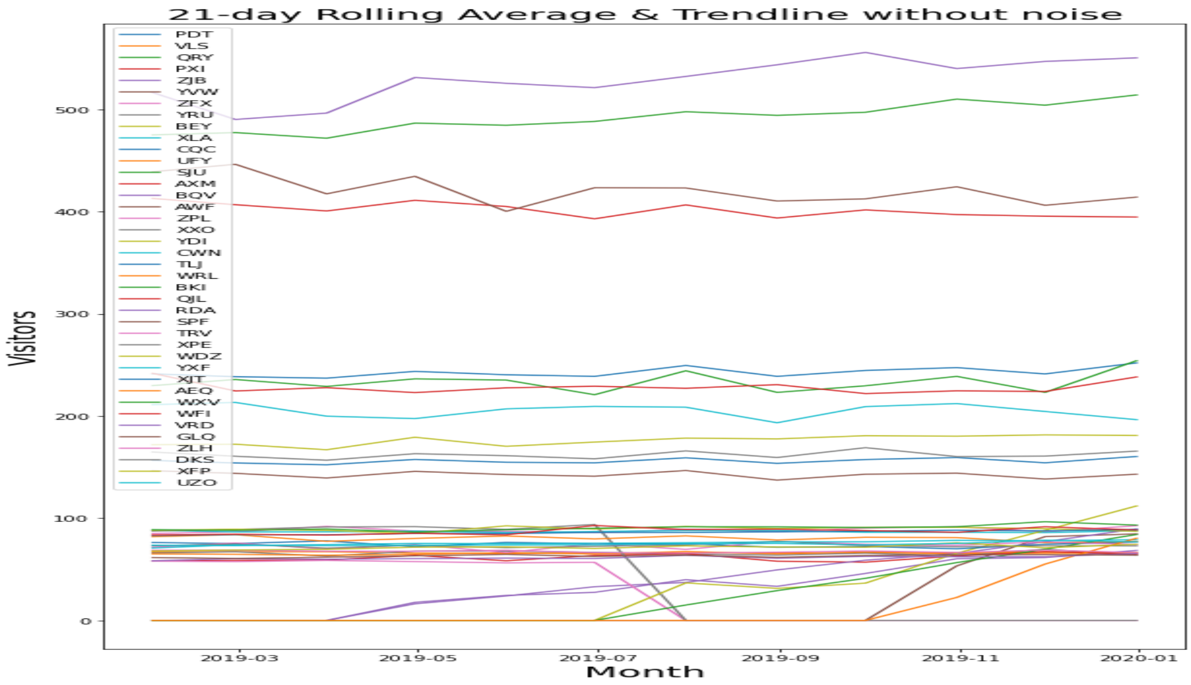


Figure 2Line graph date Vs venues

➢ Justification :

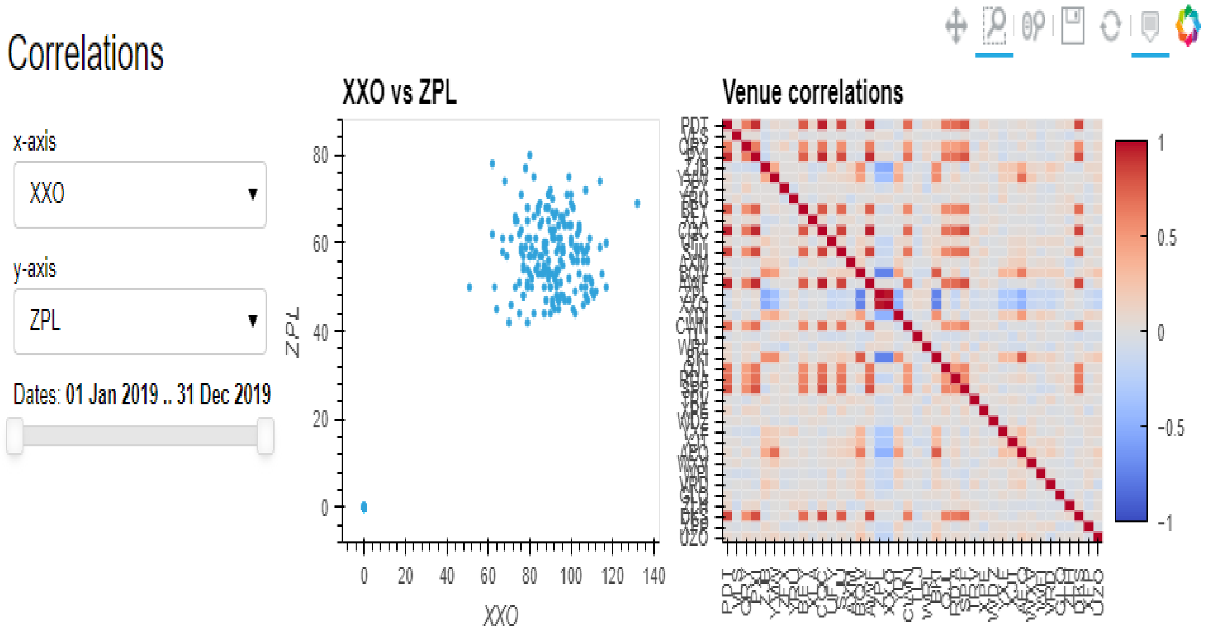
The line graph is an analysis of data throughout the period. The line graph is connected by several points in a straight line. The line graph has provided many functionalities to show different kinds of line plots including rolling average, subplot, and many more. In this section, we implement a rolling method to find which venues are famous over the period and which ones are closed due to lack of visitors and newly open venues.

➢ Description :

Above visualisation, we find venues that have higher, medium, and lower visitors. We found 4 higher venues, 3 lower venues, and 33 medium venues according to our given condition. On this part, we display only trend monthly time series for venues that are closed and open during periods on the x-axes we can display venue names, and on the y-axes, we can display the number of visitor average value because we can easily see clear trend using the 21 days rolling function. Also, we will find which venues are going to make more profit according to their trend line. We are showing the total of 8 venues closed and open in which, only 2 venues are closed and 6 are opened over the period.

- **Data Visualisation 3**

➢ **Visual 3 :**



Correlations

x-axis

XXO ▾

y-axis

ZPL ▾

Dates: 01 Jan 2019 .. 31 Dec 2019

Figure 3 correlation dashboard venue vs venues

➢ **Justification :**
Correlation is the most important function for computing average data that gives the relationship between two variables. The correlation result should be positive means both variables go in the same direction, negative means opposite direction, and neutral means no changes in variable values. Here, we created a dashboard system but, a little difference between time series and correlation. I want to know both close venues XXO and ZPL are related to each other because only two venues are close during a given period. Also, the heat map provides all venues vs all venues correlation. There are dark red values that show positive and dark blue negative.

➢ **Description :**
In this section, in one frame we provide specific two venues and time limits for analysis of how they are related and affect their visitors and another one provides all possible correlation results. According to the second frame, a few venues are showing a negative correlation. We gave two required venues, on the x-axes, and the y-axes, and I want to know the relationship between them. We don’t change in time because we are analyzing the whole period.

- ## Data Visualisation 4

➢ Visual 4 :

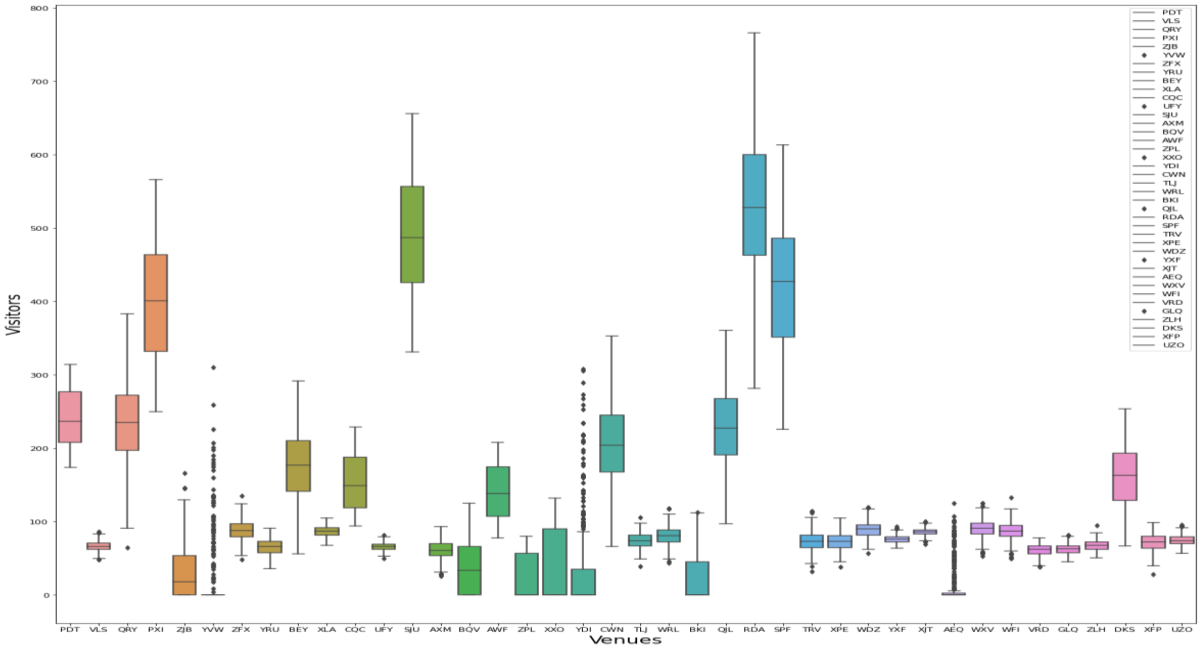


Figure 4 Box plot for outlier

➢ Justification :

The box plot is used for normal distribution when you compare multiple variables (groups) at the same time. It gives minimum, Q1 (the first quartile means 25%), Q2 (median means 50%), Q3 (the third quartile means 75%), IQR (Interquartile range means Q3-Q1), and maximum values. Also, the box plot gives skewness which has a positive and a negative. For further implementation, we need to know which venues are affected by wrong visitor values.

➢ Description :

We created a box plot for finding an outlier for each venue. Our main focus is on outliers because it makes a huge impact on the results of the analysis. The higher visitor box plots are more starch and have no outliers to impact on result compared to medium visitors. Lower visitor venues have more error values and minimum and maximum values are near look like no box plot which gives fault result. Apart from YDI being the only venue that has the highest outliers, all medium visitor venues are fewer false values.

- ## Data Visualisation 5

- **Visual 5 :**

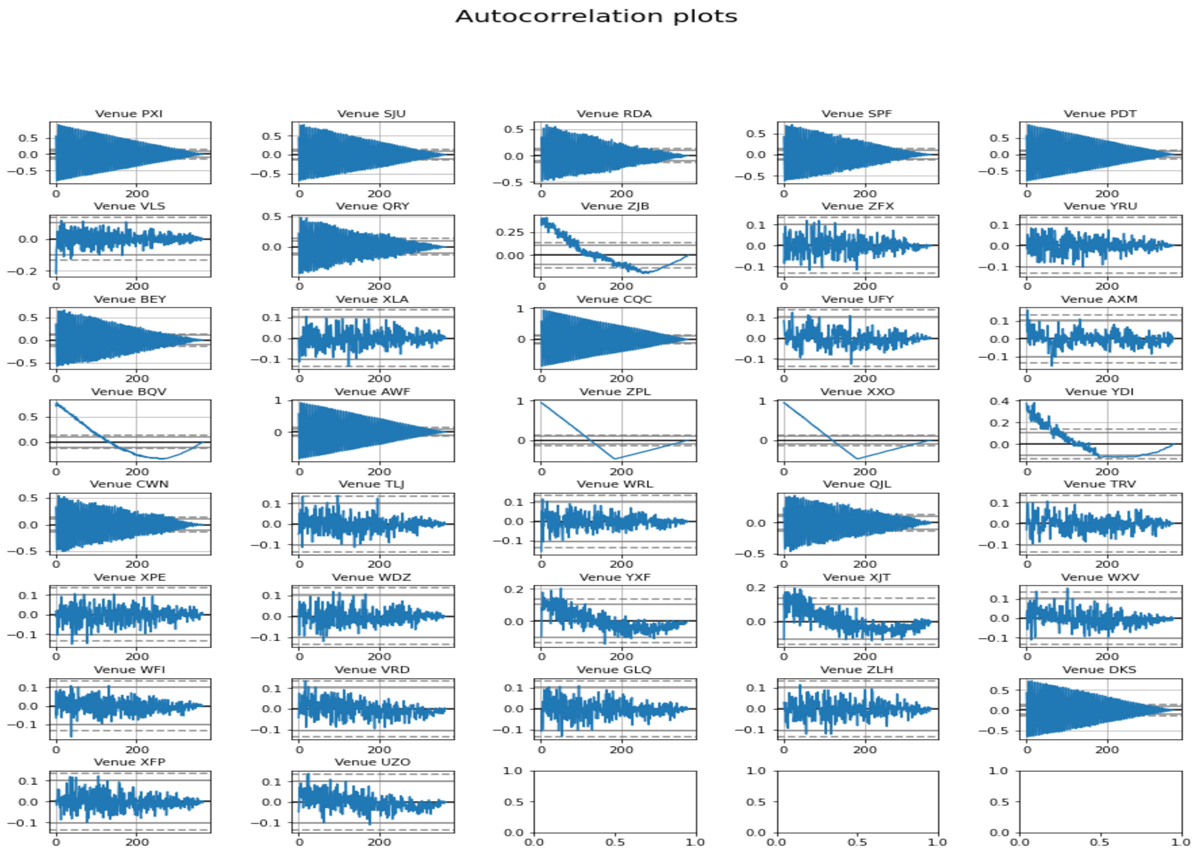


Figure 5 Auto-correlation plot of high and medium venues

- **Justification :**

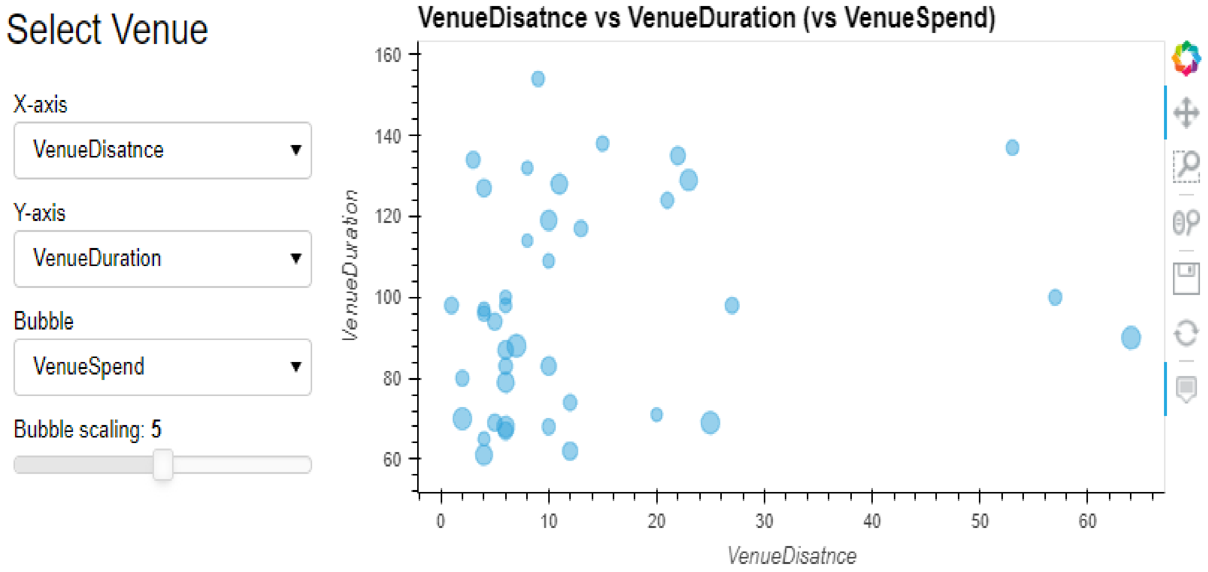
Autocorrelation plots show the similarity between the real-time value and the previous value of the variable. The output will be in three categories such as negative, positive, and independent of each other. Autocorrelation has a default limit between -1 to 1. It will be useful to make a decision for more people visiting the place and earning profit.

- **Description :**

Here, we visualize an autocorrelation graph for high and medium venues and give a specific range for correlation values between 0.5 to -0.5 means -0.5 to 0 is a negative trend and 0 to 0.5 is a positive trend. The higher visitor venues trend is positive and the medium visitor venues trend is mixed negative and positive. Some of the venues are between our given limit means it’s neutral. But, ZJB, XXO, BQV, ZPL, and YDI venues are very famous during their establishment after that, popularity goes down, and then it increases in popularity. Now, those venues are average visitors.

- ## Data Visualisation 6

➢ Visual 6 :



Select Venue

X-axis

VenueDisatnce ▾

Y-axis

VenueDuration ▾

Bubble

VenueSpend ▾

Bubble scaling: 5

Figure 6 Bubble plot Comparing 3 variables

➢ Justification :

The bubble graph is more interactive and makes more than two variables relationship compared to a scatter plot and the dot point size shows the third variable relationship. Also, make a fixed variable bubble plot. But, we created the bubble plot using the dashboard system which gives more accessibility to all variables and bubble size. We need to explore venue-spend time compare to venue distance and venue duration.

➢ Description :

We visualize the bubble graph using three variables. One is venue distance, the second is venue duration, and the last one is venue spend. Along with this, we give a bubble size around 5. The biggest bubble size is 148 which shows the highest venue spent relationship between the longest distance Vs medium duration. Also, there are two bubble sizes nearly the highest bubble size which shows the minimum distance Vs medium duration.

- ## Data Visualisation 7

➢ Visual 7 :

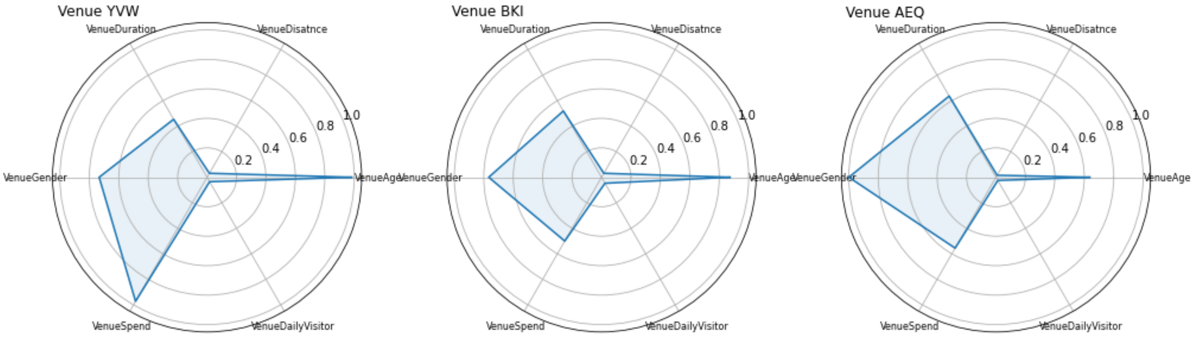


Figure 7 Scatter plot for low visitors venues

➢ Justification :

Radar graphs visualize multivariable data into one graph. Also, it displays more groups according to common multivariable data without giving separate axes means all variable has own axes are connected by a central point. A comparison between our multivariable data shows the performance of variables that are high and low within our data. So, we can make a decision on factors that are needed to improve according to their related variable performance.

➢ Description :

In the implementation, we visualize a radar plot for our lower visitor venues, because, we need to maintain all venues which are going famous and attract more visitors. According to our radar graphs, there are very less visitors to this venue. Venue YUW is a very old venue, famous among middle-aged people, and open availability is low even people spend more time visiting the place. Both BKI and AEQ venues have the same performance for the visitor. AEQ is slightly more visitors compared to BKI venues.

- ## Data Visualisation 8

➢ Visual :

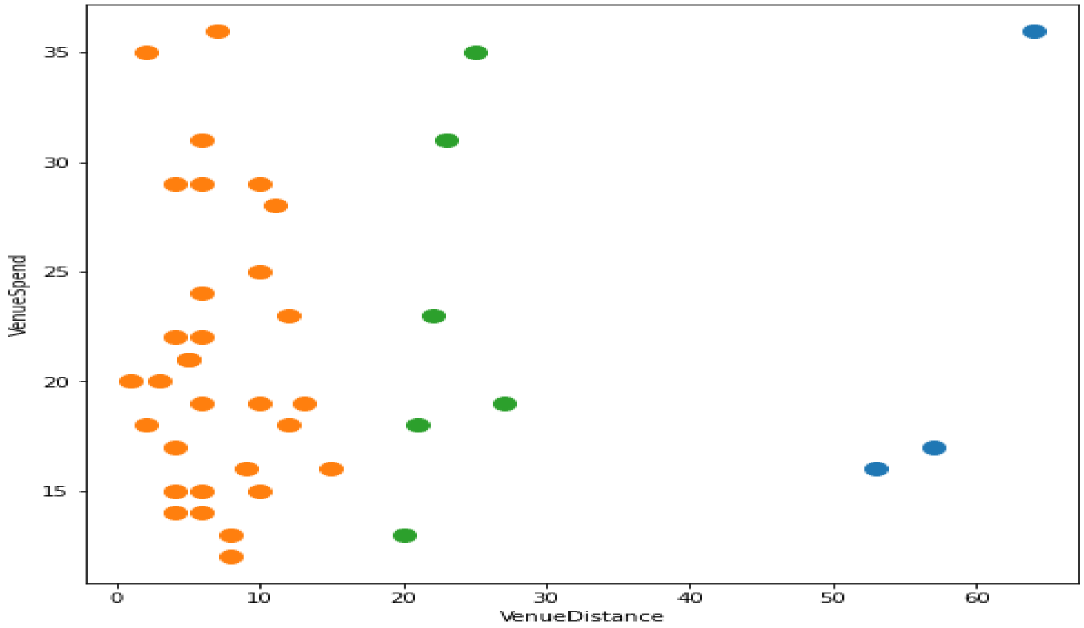


Figure 8 cluster plot venue spend Vs venue distance

➢ Justification :

The clustering algorithm is centroid based which means every cluster has connected to a centroid. In the following, every value is connected to the centroid but, the condition is the minimum distance allow to connect. K-means algorithm performs in a two-step. The first step is to find the K value which provides how many groups will be created and the second step is computing the minimum distance and assigning a near centroid. Here, we use K-means to visualise for anomaly using venue distance and venue spend

➢ Description :

Here, we use the silhouette Coefficient clustering method. Also, we give a limit of -1to 1 for the clustering method. We find out the K value is 3 using step 1according to our data. Then, we create the main graph using two variables one is people who spend money on visiting venues and the second is the Distance of Venus means people travel from their home to the venue. According to our graph, we can say that our anomaly value (false value) in blue which shows the lower visitor venue and distance is more than 60 miles and time is more than 35.

- ## Critical Review

The report is created and analysed by performing on the student dataset which has provided by the professor. We have 6 types of datasets including the Daily time series which has 40 venues and their characteristics. First, we worked on daily basis time series. Time series provide more interesting and important information about all venues including their historical values. We find something interesting points such as trends, relations between venues, and false values which impact their visitors using the bar, line, and auto-correlation. Then, we done analysis on the second data frame. In the summary data, provide how their variables impact on visitors and the relationship between them. The overall evolution and exploration show that a lower percentage of visitor venues are more distance which takes more time to travel for visitors and the duration of availability is short. If we increase availability, it will be more interesting for analysis. It might be more critical to explore data and make decision for venues profit.

Conclusion

- Finally, we can conclude from our analysis of the daily data frame and summary data frame. There are many things revealed to using visualisation for higher, medium, and lower visitor venues.
- The highest number of visitors visiting RDA venues is around 193188, while the lowest number of visitors visits around 4843.
- Apart from that, ZPL and XXO venues are closed over the period. However, there are 6 venues such as YDI, BQV, ZJB, YVW, BKI, and AEQ newly open during the same period. In addition, we can show that newly open venues are more famous compared to others where there are open.
- We can also observe that in the outlier plot which shows only one medium visitor's venue YDI has more false values compared to other medium venues.
- Moreover, there is a good relationship between venue-daily-visitor VS venue-distance and venue-age VS Venue-spend that can higher impact on them. If some changes appear on one variable, it can also impact another related variable.
- Dashboards were provided many functionalities which makes the plot a very interesting way to visualize and we utilize more in deep study.