**INTRODUCTION**

Data visualization is the graphical representation of information and data. It is a particularly efficient way of communicating when the data is numerous, by using visual elements like charts, graphs, and maps. Data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in data sets.

Data visualization provides a quick and effective way to communicate information in a universal manner using visual information.

**DISCUSSION OF FINDINGS.**

The data we are working on comes from the 40 venues of ChrisCo company. By going through their website , we were able to get some information, we also got some information which we are analyzing from their sales department. Below are the 40 venues about which we will bring out valuable information on some of them.

**'ZJB', 'UZO', 'UFY', 'XPE', 'WXV', 'BKI', 'QJL', 'DKS', 'ZLH', 'QRY', 'TLJ', 'WFI', 'VRD', 'CWN', 'YRU', 'VLS', 'ZPL', 'XJT', 'YDI', 'WRL', 'YXF', 'TRV', 'SJU', 'PXI', 'XFP', 'RDA', 'ZFX', 'GLQ', 'BEY', 'AWF',**

**'XXO', 'XLA', 'BQV', 'WDZ', 'AEQ', 'PDT', 'AXM', 'CQC', 'SPF', 'YVW'.**

Figure 1: Plot of daily visit of each venue to date

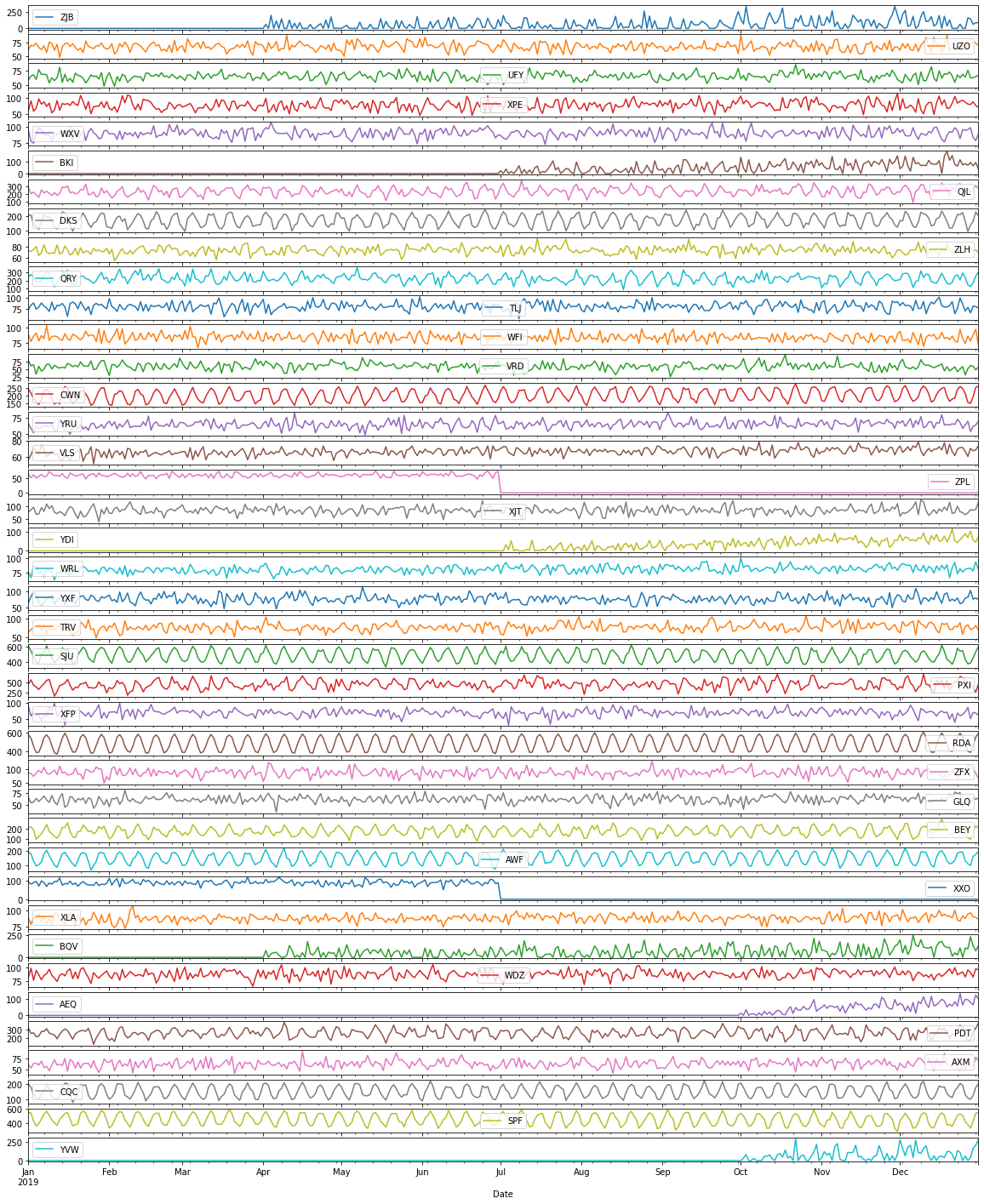


Figure 1 shows how the data changes over time in each venue. Though the plot above shows a very few venues that have some spikes, the ups and downs of most venues are too perfectly alike, with almost no trend. We can also see the new venues ZJB , BKI , YDI , BQV , AEQ and YVW that were opened in the year and the two branches ,ZPL and XXQ that closed down before the end of the year. For the new branches , ZJB opened 171 days out of 365 , BKI opened 156 days, YDI 164 days , BQV for 207 days , AEQ opened for 88 days while YVW opened for just 69 days. For the branches that closed down . The total number of all visits for all branches is 1888139 visits ,out of which only four branches , RDA , SJU , SPF , PXI(the outliers) have 681212 visits ,36% of the total visits, while the remaining 36 branches share the remaining 64%. These branches ZPL , BKI , YDI , YVW and AEQ have the lowest total visits for the year 2019.

Figure 2: Relationship plot of variables

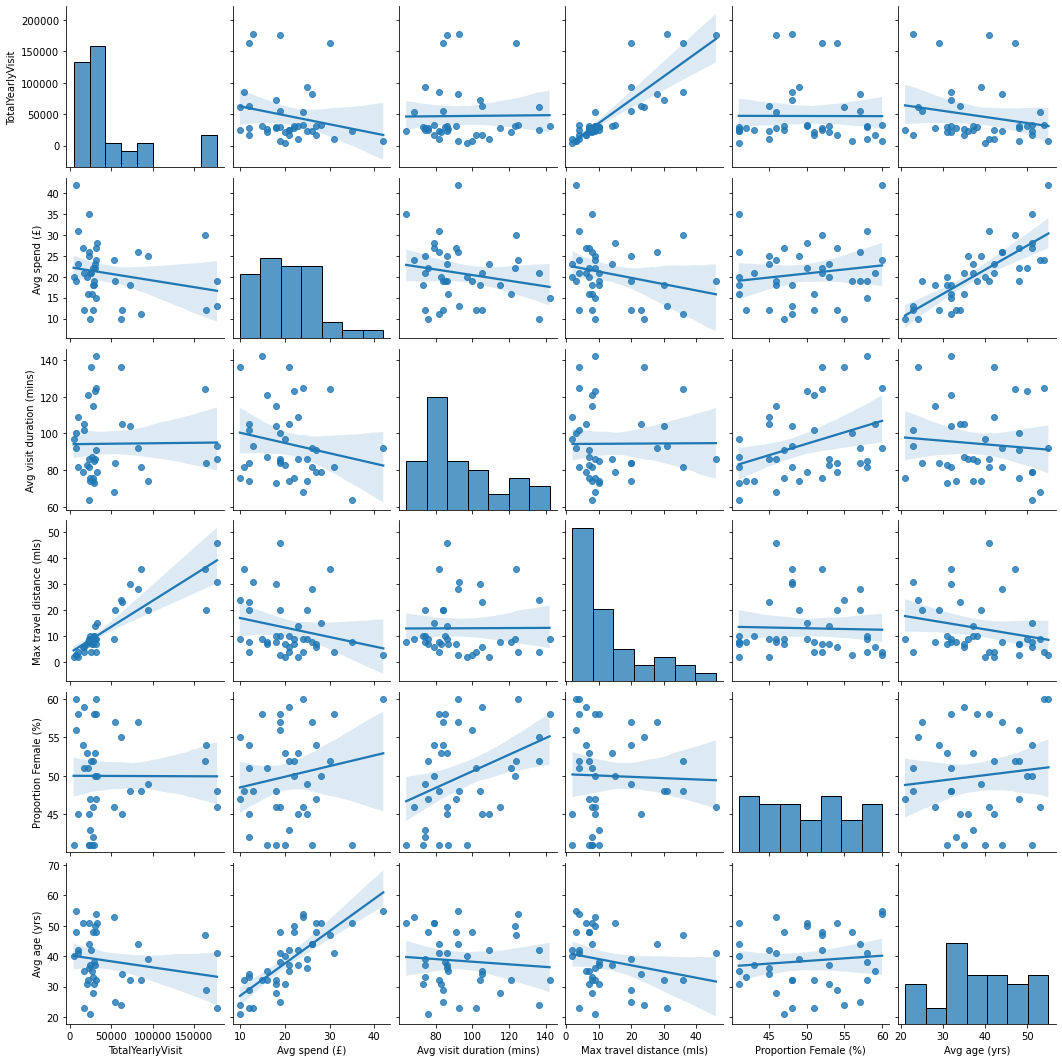


Figure 2 shows that the maximum travel distance has a high positive correlation with Total visit for the year ,i.e the higher the maximum Travel Distance , the higher the number of visits. Maximum travel distance also has a low negative correlation with **Avg Spend (£)**. The **Average spend (£)** has a high positive correlation with the **Average age(Yrs)”**. The Proportion of females(%) is having a positive relationship with the **Average Visit Duration.** The scatterplot of TotalYearlyvisits shows that it has has some outliers, the histogram of TotalYearlyVisit is right skewed, this means that most of the data are concentrated on the lower values while only a few have high values. The “Max travel distance(mls)” and “Avg Visits Duration(mins) also have more of smaller values than the higher values just like the TotalYearlyVisit.

The main reason for using Figure 2 is to show the degree of association that exist among variables, and then see what can be altered to maximize the revenue of the venues.

Figure 3:

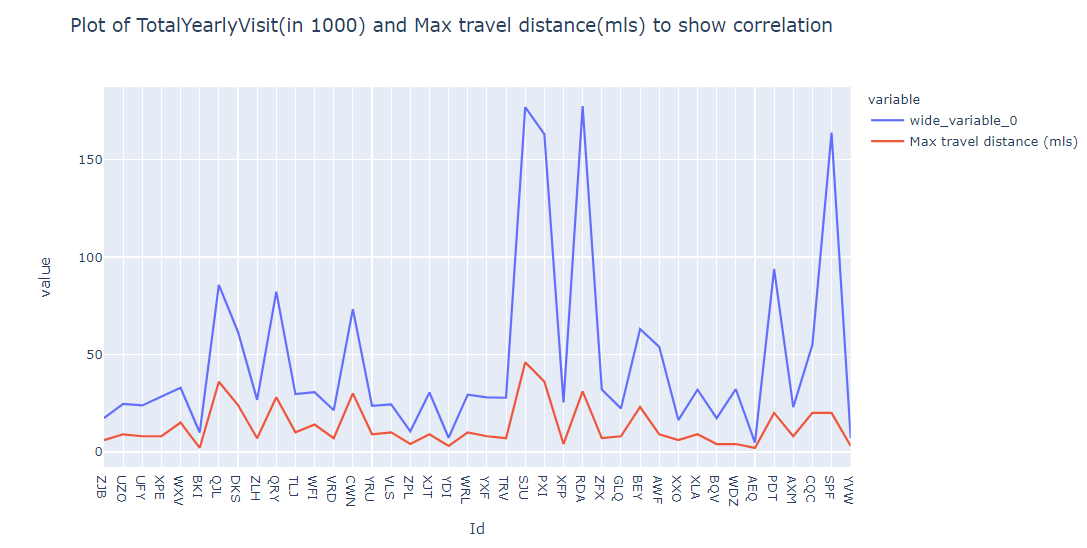


Figure 3 shows that that the degree of correlation of TotalYearlyVisit and Maximum travel distance is very high, the ups and downs of both plots are so much alike. There is need for a more comprehensive study of both variables in order to increase the number of visits in all branches.

Figure 4: Distribution of ages across all venues

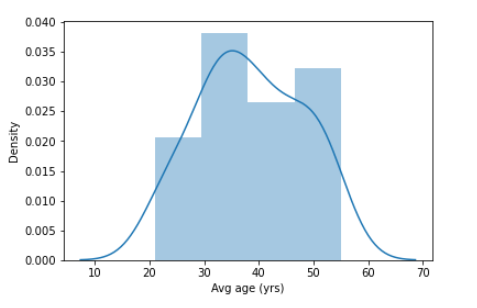


Figure 4 shows that the age range 20+ to 50+ as almost normally distributed across all the venues. Knowing this will make conscious efforts to be taken in order to increase visits by these age group. The peak age is between 30-35.

Figure 5: BARPLOT OF TOTALYEARLYVISIT ARRANGED FROM HIGHEST TO LOWEST

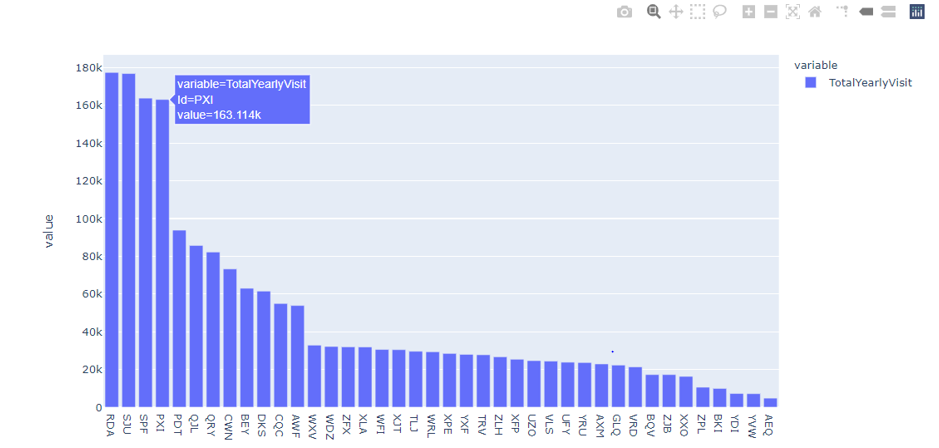


Figure 5 shows the the barplot of venues arranged based on their TotalYearlyVisit in descending order, from highest to lowest. The first four venues RDA , SJU , PXI and PDT are outliers , the difference between PDT , which is the fourth and QJL is over 60,000 visits.

The venues with lowest TotalYearlyVisit are XXO , ZPL , BKI , YDI , YVW , AEQ.

Figure 6: BARPLOT OF TOTALYEARLYVISIT ARRANGED FROM HIGHEST TO LOWEST AFTER ESTIMATING THE EXPECTED TOTALYEARLYVISIT OF VENUES THAT DID NOT OPEN AT THE START OF THE YEAR AND THOSE THAT CLOSED DOWN BEFORE THE YEAR END

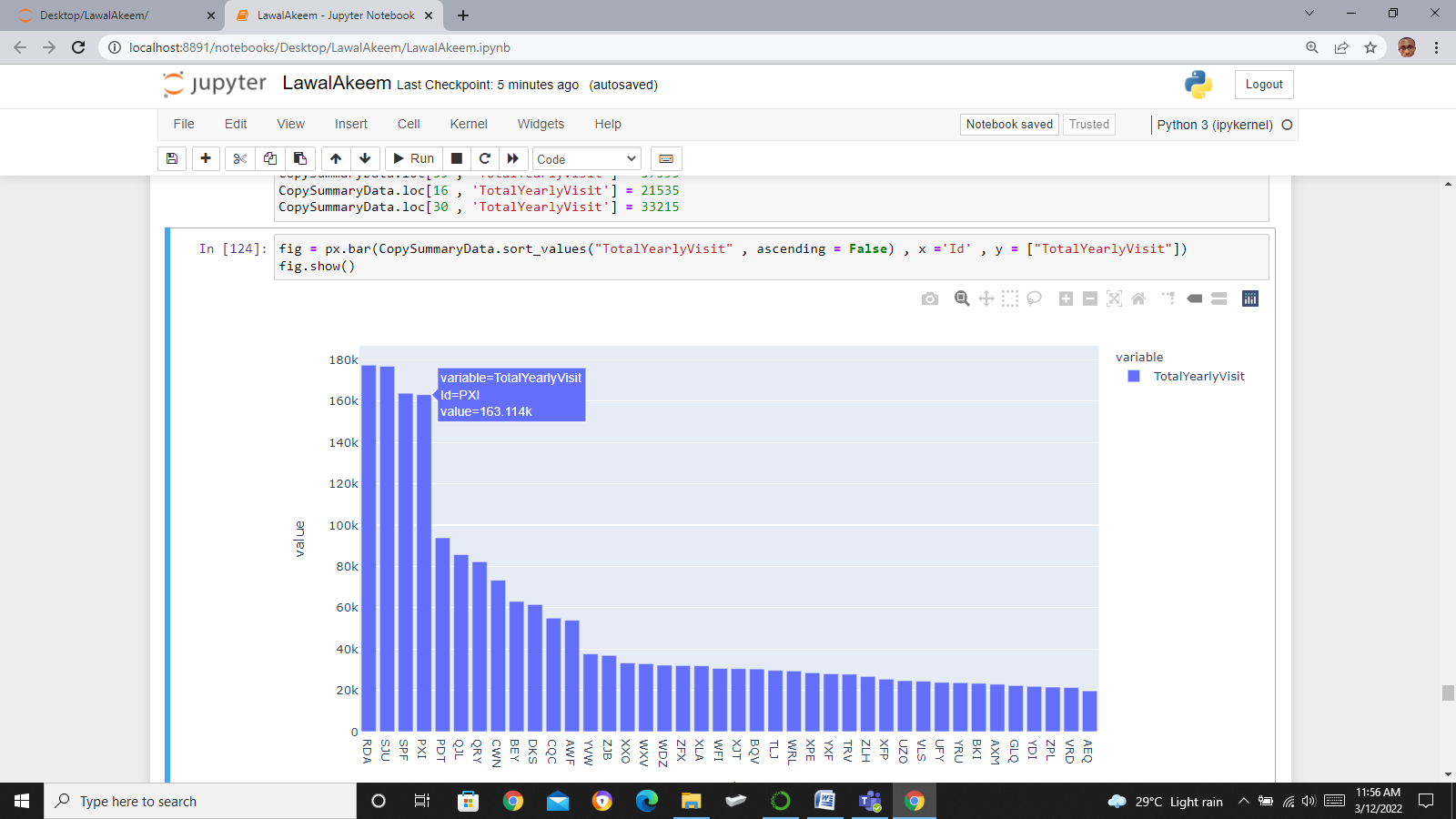


Figure 6 shows the barplot of the TotalYearlyVisit in descending order after factoring the expected TotalYearlyVisit of venues that did not open at the start of the year and those that closed down before the year end. Comparing Figure 5 and Figure 6 , we can see that the first twelve venues from RDA to AWF maintain their lead , while the venues with lowest TotalYearlyVisit changed ,with AXM , GLQ , YDI , ZPL , VRD , AEQ as the new group. i.e XXO,BKI,YVW leaving the group of lowest venues with TotalYearlyVisit The first four venues RDA , SJU , PXI and PDT are outliers , the difference between PDT , which is the fourth and QJL is over 60,000 visits.

Factoring the expected TotalYearlyVisit affects and corrects the distribution of medium volume venues and the low volume venues. If we remove the outliers , its likely we have an almost perfectly distributed values of TotalYearlyVisit across all venues.

Figure 7: Pie Chart showing estimated revenue of each venue

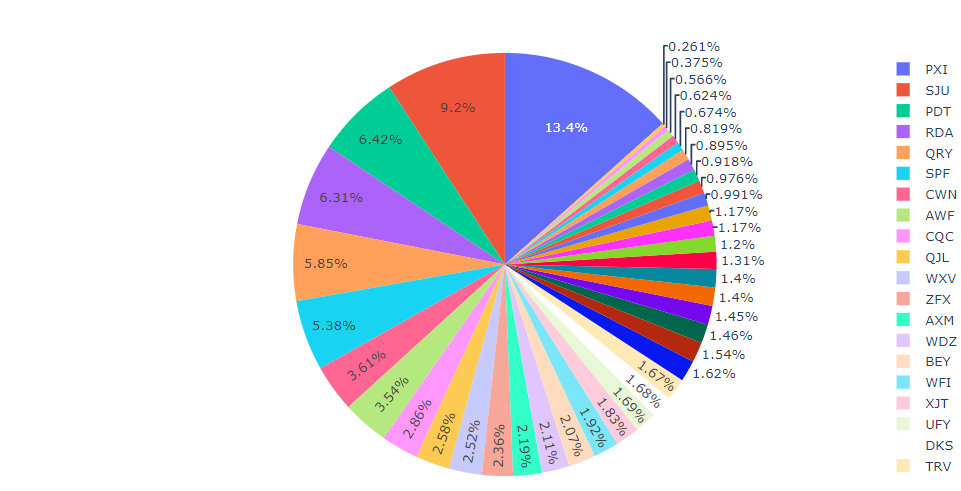


Figure 7 shows the pieplot of estimated revenue of each venue, which was calculated by multiplying the total visit for the year in each venue by Average spend(£). This reveals that seven branches which include PXI, SJU , PDT , RDA , QRY ,SPF and CWN accrued 50% of the revenue for the year 2019 while the remaining 33 venues.

With this pie chart , we can see the percentage of estimated revenue of each venues to the overall Revenue of all venues.

**THE ESTIMATED REVENUES OF THREE CATEGORIES OF VENUES.**

Low volume venues are:

ZJB GLQ BKI XXO VRD BQV ZPL AEQ YDI YVW

count =10 , estimated revenue = 2879155

Medium Volume Venues are:

UZO UFY XPE WXV ZLH TLJ WFI YRU VLS XJT

WRL YXF TRV XFP ZFX AWF XLA WDZ AXM CQC

count =20 , estimated revenue = 13022799

High volume venues are:

QJL PXI DKS RDA QRY BEY CWN PDT SJU SPF

count =10 , estimated revenue =20643770

Figure 8: Pie Chart showing the Revenue contributions of the three categories of venues

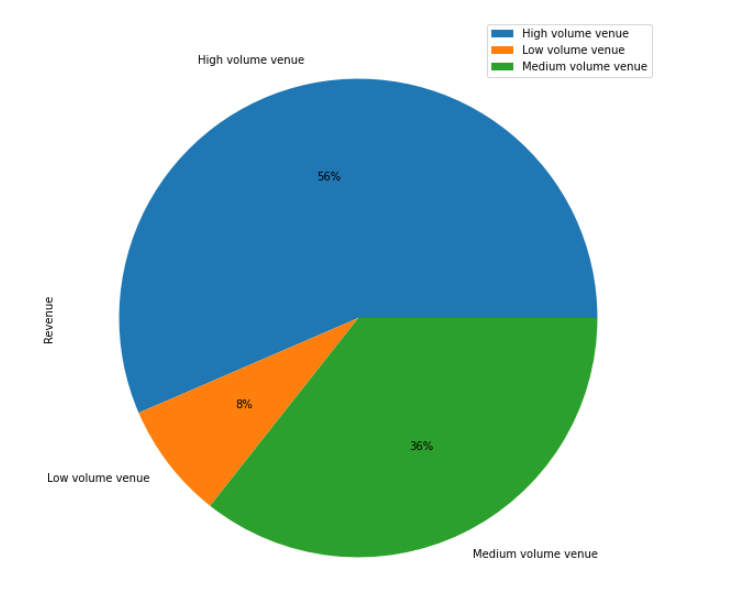


Figure 8 shows the estimated revenue of the three categories of venues. High volume venues consisting of 10 venues(25% of the total number of venues) has an estimated contribution of 56% of total revenue , medium volume venues consisting of 20 venues(50% of total number of venues) has an estimated contribution of 36% of total revenue, while the low volume venues consisting of 10 venues(25% of total number of venues) has an estimated contribution of 8%.

If we insert the expected total yearly visit of the venues that did not start at the start of the year and the venues that closed down before the end of the year

ZJB = 36865 ,BKI = 23360 ,YDI = 21900 ,BQV = 30295 ,AEQ = 19710 , YVW = 37595 , ZPL = 21535

XXO = 33215

We have a new group of low volume venues , medium volume venues and high volume venues

NEW LOW VOLUME VENUE

UFY ZPL BKI YDI VRD GLQ YRU AEQ VLS AXM

NEW MEDIUM VOLUME VENUES

ZJB TLJ TRV XLA UZO WFI XFP BQV XPE XJT

ZFX WDZ WXV WRL AWF CQC ZLH YXF XXO YVW

NEW HIGH VOLUME VENUES

QJL PXI DKS RDA QRY BEY CWN PDT SJU SPF

**JUSTIFICATION OF VISUALISATION USED**

Figure 1 was used to show the behavior of daily visits of each venue through the year. As a seasonal data time plot is best suited to show seasonal behavior.

Figure 2, Relationship plot of variables as the name implies , was used to visualize the degree of association between the variables. We were also able to see the distribution of values to each variables,

Figure 3: The reason for using Figure 3 is to show the almost perfect association that exists between TotalYearlyVisit and Maximum travel distance.

Figure 4 : The reason for using Figure 4 is to show how age groups are distributed across all venues, Conscious efforts can then be made to study what motivates certain age group and then be used to motivate other age groups

Figure 5 and Figure 6:Bar charts are good for comparison of variables. With this we can clearly see the branches with highest TotalYearlyVisit till we reach the branches with the lowest visit. As the data were arranged in ascending order , it makes visualization and insights easier.

Figure 7 and Figure 8: The two primary use case for a pie chart are (1) to make the audience have a general sense of the part-to-whole relationship in our data and (2) to convey that one segment of the total is relatively large or small. We have used pie chart in Figure 7 and Figure 8 for exactly these two reasons, Figure 7 showing the part to whole relationship of each venue to the total of TotalYearlyVisit and Figure 8 showing the Aggregate to total relationship of three categories of venues

**CRITICAL REVIEW**

The plot of daily visits to their dates were made. We were able to bring out the venues that did not open at the start of the year and the two venues that closed down before the year ends. The total yearly visit of each venue were calculated by adding all the daily visits for each venue. The given data (Venue age , venue distance , venue duration venue gender and venue spend) were merged together with the TotalYearlyVisit into a dataframe called SummaryData. The count of daily visit with zero values were gotten for venues that did not open at the start of the year and the two venues that closed down before the year ends. It was noted that some of these venues opened for as low as 69 days out of 365. Comparing these venues with venues that have full year daily visits will not give us the true picture of these venues. The SummaryData was grouped into three categories , low-volume-venues , medium-volume-venues and high-volume-venues according to the 0 - 25th percentile , 26th -75th percentile and 76th -100th percentile.

In order not to change our SummaryData dataframe , a deep copy of Summary data was done and named CopySummaryData. Then the cell value for TotalYearlyVisit were replaced with calculated expected TotalYearlyVisit, then grouped into three categories , low-volume-venues , medium-volume-venues and high-volume-venues according to the 0 - 25th percentile , 26th -75th percentile and 76th -100th percentile. The venues previously in high-volume-venue maintained their position , while the remaining venues scattered across medium-volume-venues and low-volume-venues.

**SUMMARY**

The TotalYearlyVisit of the company as we get it from the original data is not enough to make generalizations and comparisons about the whole venues. Four venues RDA ,SJU ,SPF and PXI are considered to be outliers according to their totalyearlyvisit.

The TotalYearlyVisit is so much positively associated to the Maximum Travel distance, the higher the Maximum Travel distance , the higher the TotalYearlyVisit. This looks absurd as both variables are expected to be negatively correlated.

The **Average spend (£)** has a high positive correlation with the **Average age(Yrs)”.** The distribution of ages across all venues is almost normally distributed.

The **Proportion of females(%)** is having a positive relationship with the **Average Visit Duration.**

In the original data given, High volume venues has an estimated contribution of 56% of total revenue , medium volume venues has an estimated contribution of 36% of total revenue, while the low volume has an estimated contribution of 8%.

In the modified data , i.e CopySummaryData, High volume venues has an estimated contribution of 52% of total revenue , medium volume venues has an estimated contribution of 35% of total revenue, while the low volume has an estimated contribution of 13%.

References

What Is Data Visualization? Definition, Examples, And Learning Resources.[Online] Available at <https://www.tableau.com/learn/articles/data-visualization> (Accessed: 12th March 12, 2022)

# Set Pandas Conditional Column Based on Values of Another Column[Online] Available at

<https://datagy.io/pandas-conditional-column/> (Accessed: 12th March 12, 2022)

Kate B , Ed B (2020) . Data visualization [Online] Available at <https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization>

[The Data Visualisation Catalogue](https://datavizcatalogue.com/index.html) , Bar chart. [Online] Available at <https://datavizcatalogue.com/methods/bar_chart.html> (Acessed: 14th March, 2022)