Analyzing Crisco company data

1. Introduction

In today's world, when 2.5 quintillion bytes of data are created every day, data and patterns must be examined before data-driven choices can be made. Data visualization makes trends, anomalies, and patterns in data visible and understandable. Data visualization is the graphical display of data in a visual context (charts, maps, or graphs) to make trends, outliers, and patterns in data simpler to interpret (Goyal, 2021). Data visualization is a set of tools that may be used to study a dataset and information, allowing you to see trends, corrupt data, outliers, and more.

Some benefits of data visualization are as follows:-

- It helps in the detection of correlations between independent variable connections.
  - It helps to find the trend over time.
- It helps in the data cleaning process by finding incorrect data and corrupted or missing values.
- It helps in data reduction by playing a crucial role while combining the categories.

2. Exploring data

ChrisCo is a fictional company that manages a number of venues around the United Kingdom. ChrisCo gathers a lot of information of the people that come to its venues. ChrisCo is made up of 40 venues, identified by its own three-letter code. Visualization is used to study the data of ChrisCo in accordance with the requirements of the academic module.

The main purpose of the assignment is to use the visualization approach to check for correlations, seasonality, and trends in the dataset, as well as to discover anomalies.

We need a way to display all of the data acquired by the ChrisCo so that we can understand it. By presenting data in a visual context, such as maps or graphs, data visualization which can helps us to understand what it means. This makes the data more understandable to the human mind, making it simpler to see trends, patterns, and outliers in vast data sets.

3. Visualization

3.1. Bar charts

A bar chart is a graphical representation of a data which has vertical axis with numbers on it, and a horizontal axis showing values. Bar's height (or widths) are proportional to the numbers they indicate. Bar charts is used to display the graphical representation of the venue daily visitor.

Figure 1, shows the bar chart representation of the total daily visitor data where the x-axis represents the list of the venue whereas the y-axis represents the volume of daily visitors in the venue.

Basically, the pictorial diagram shows that there is 40 venue with the width range of venues with the maximum range around 181164. Moreover considering the minimum range is about 5184.

After representing the data into bar chart, we can analyze that RDA, SJU, SPF, and PXI venue has the high volume of visitors, whereas, PDT, QRY, QJL, CWN, BEY, DKS, CQC, and AWF are the venue has the medium volume of visitor, so on, WXV, ZFX, WFI, XLA, WDZ, WRL, XJT, YXF, TRV, XPE, TLJ, XFP, ZLH, VLS, UZO, UFY, YRU, GLQ, AXM, VRD, XXO are the venue that has the low volume of visitor. Similarly, BQV, ZPL, BKI, ZJB, YDI, AEQ, and YVW are the venue which has the very low volume of visitor.

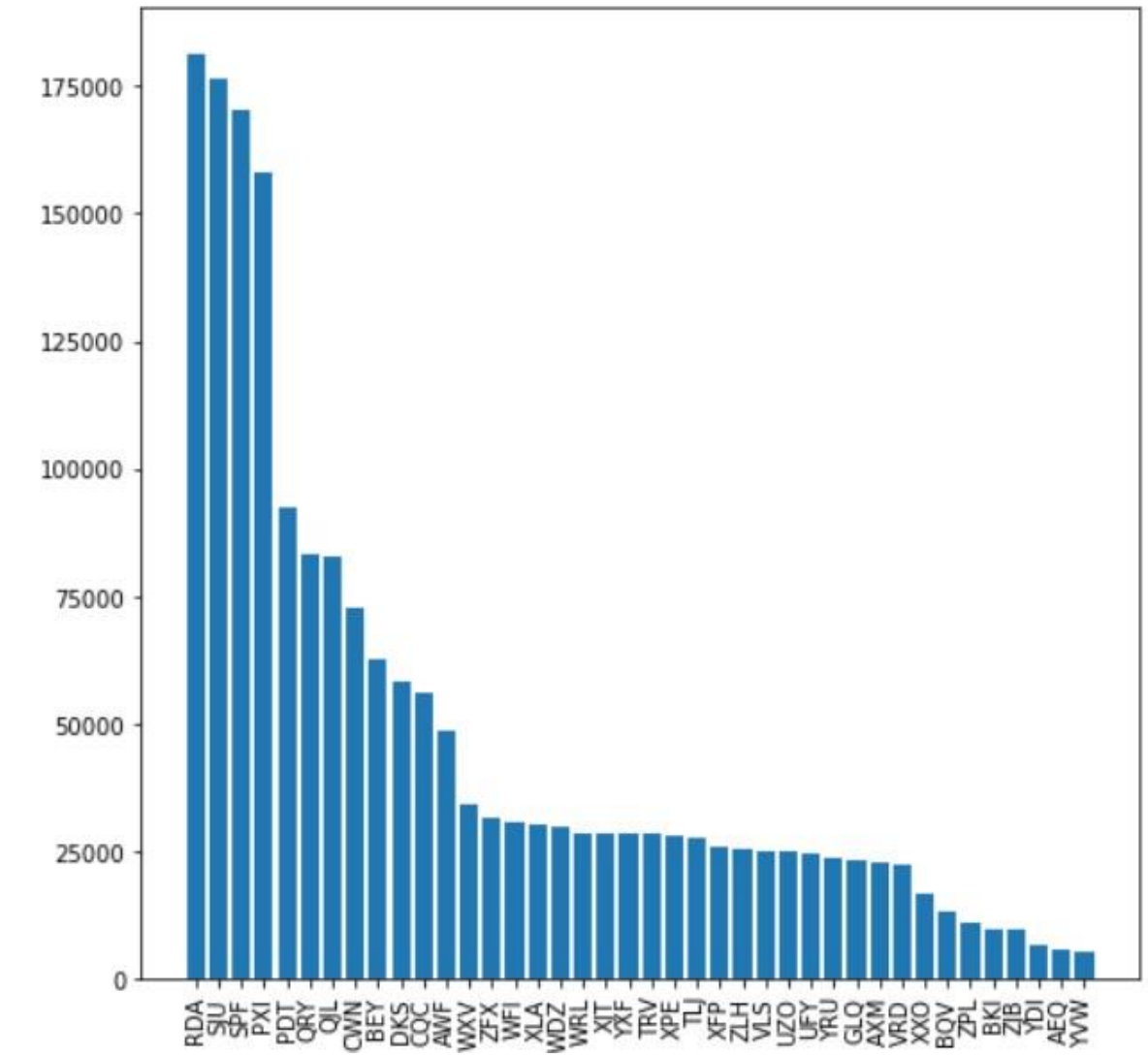

Figure 1 bar plots of daily visitors

3.2. Box plot

A box plot is a standardized way of representing distribution of data based on a summary. Box plot is implemented inorder to study the outliers, symmetrical and how data is grouped or the data is skewed.

Figure 2 shows, the box plot represents of high and low venue visitors where the x-axis represents the list of the venues whereas the y-axis represents the volume of daily visitors to the venue. From the box plot we can identified the outliers and if a statistical data set is normally distributed or skewed. Four outliers can be seen at the point of SJU, PDT, and QRY. We can identified that the middle line of the PDT box lies outside the PXI which is likely to be a difference between the two groups.

We can identified from the box plot that RDA, SJU, SPF, QRY data are symmetric.

Similarly, there are numbers of left-Skewed and right-Skewed among them PXI is an example of right-skewed and XXO is the example of left-skewed. Also, PXI, BEY, and AWF have the longer box which means it has the more dispersed data.

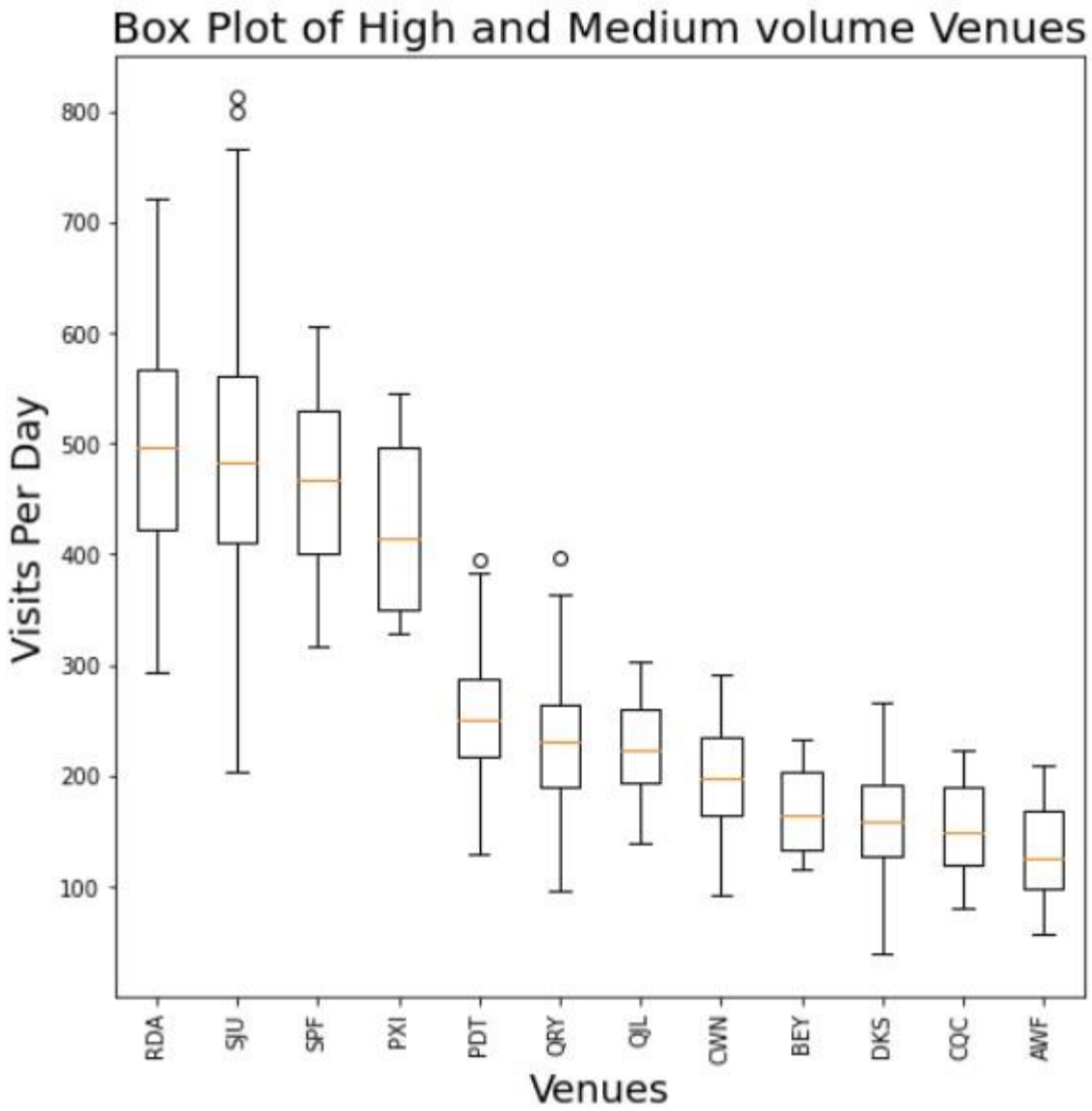

Figure 2 Box plot of high and medium venues

3.3. Heat map

Heat map is a two-dimensional representation of data in which values are represented by colors. Heat map is used to find the correlation of the time series based on their coefficient value.

Figure 3, represents the correlation between the medium volumes of a daily visitor. All the diagonal squares are red as each time series and the co-efficient value is almost greater than 0.75 which means strong correlation is formed. Whereas mild correlation is seen between the QRY and PDT.

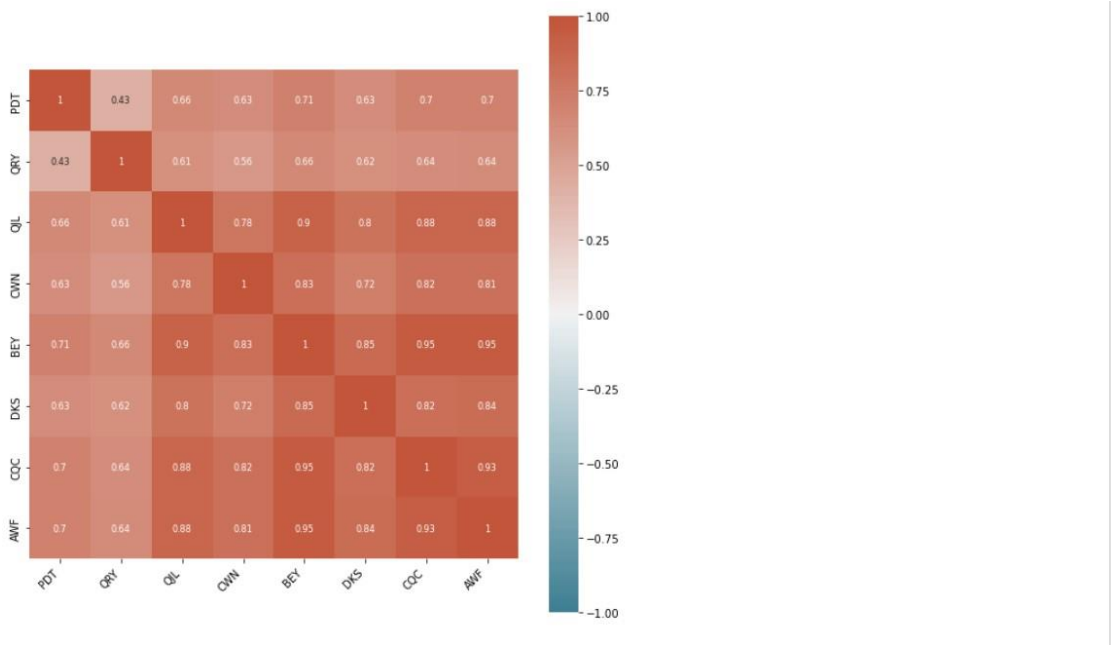

Figure 3 Heat map of a medium volume

3.4. Scatter plot

Dots are used to indicate values for two separate numeric variables in a scatter plot. The values for each data point are shown by the position of each dot on the horizontal and vertical axes. Scatter plots are used to see how variables relate to one another.

Figure 4, shows the correlation exist between the variable of high volumes. We can identify from the figure that all the variable of high volumes are positively correlated to each other. Similarly, outlier is formed between the variable RDA VS PXI, SJU VS PXI and SPF vs PXI.

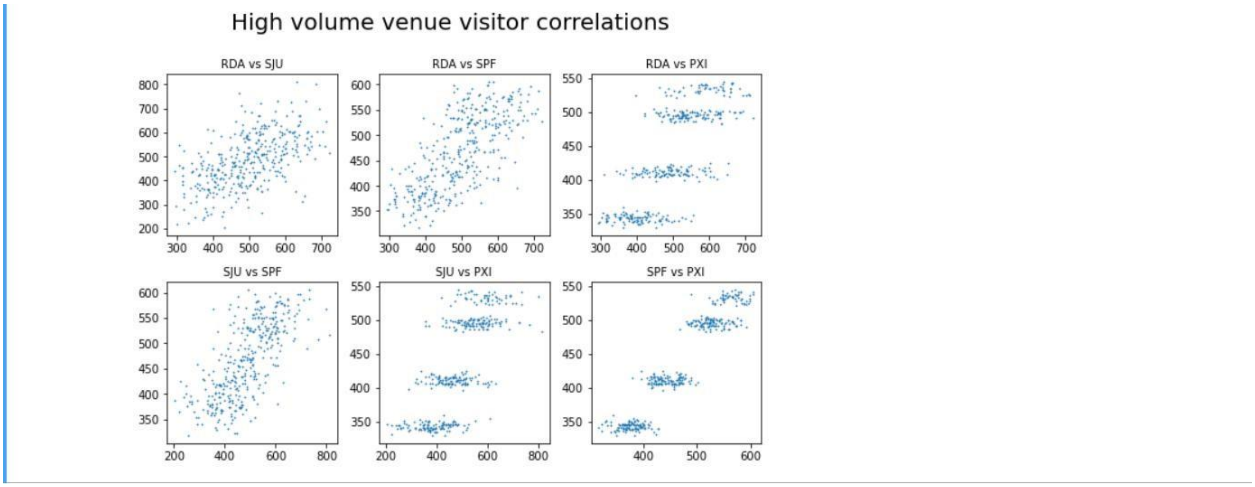

Figure 4 High volume venue visitor correlations

3.5. Stack plot

Stack plot is a plot that shows the whole dataset with easy visualization of how each part makes up the whole. Each constitute of the stack plot is stacked on top of each other.

Figure 5, shows the stack plot to determine the venues opening and closing. x- Axis contains the date of the venue whereas, y- axis contains the volume of visit. From the figure we can determine that XXO, ZPL, BK2, ZJB, YDI, AEQ, and YVW are the venue taken into consideration for the stack plot. From the figure we can find that XXO and ZPL venues are closed on July, BKI and YDI venues are open on July, ZJB venue is open on April, and AEQ and YVW venues are open on OCT.

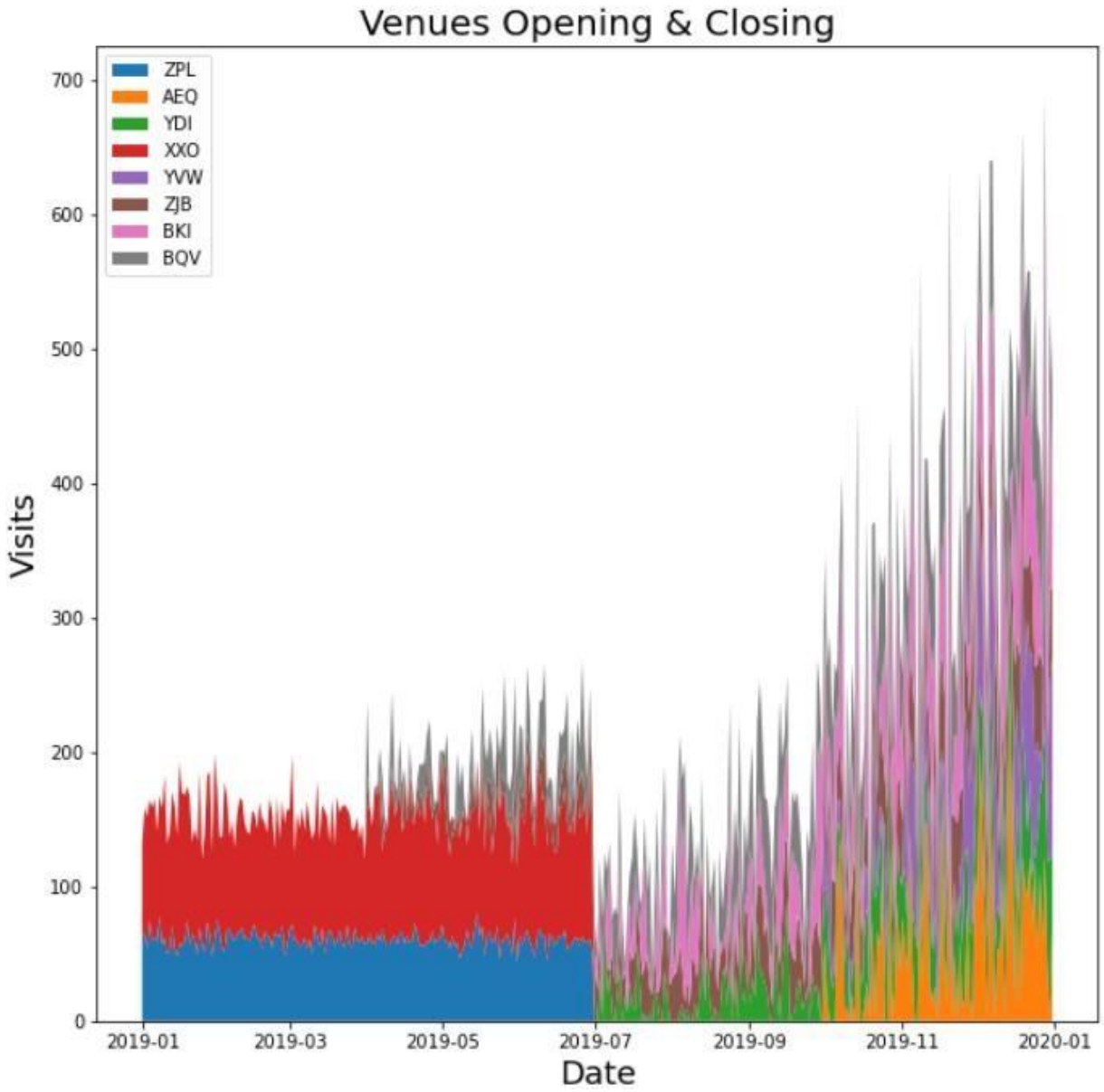

Figure 5 Stack plot to determine the anomalies

3.6. Interactive plot of the segmented dataset (Hvplot)

Hvplot is a high-level charting API based on HoloViews that provides a consistent and comprehensive API for plotting data in all of the following formats.

Figure 6, shows the Hvplot of the segmented data which illustrates the volume of visitors of the different venues. The X-axis represents the date as well as y-axis represents the volume.

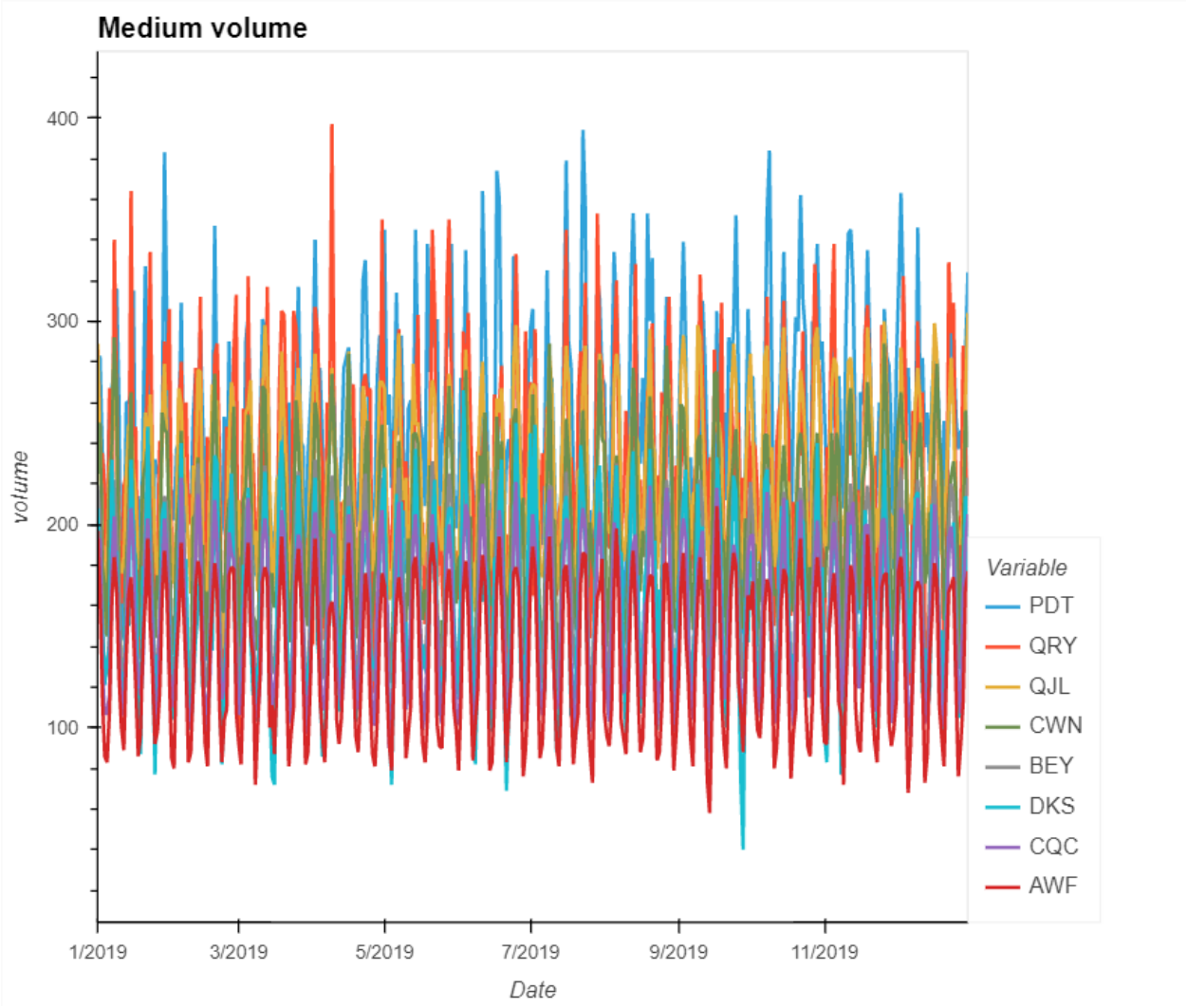

Figure 6 Interactive plot of the medium volume

3.7. Heat map interactive

Figure 7 is an interactive heat map that shows the correlation between the summary variables. Positive correlation and high correlation are not established, according to the magnitude of the coefficient in the figure. However, because the coefficient value is

0.62431, there is the establishment of a weak co-relation between age and spend. Similarly, a moderate connection exists between gender and duration, age and spend, as their coefficient variable is about 0.62.

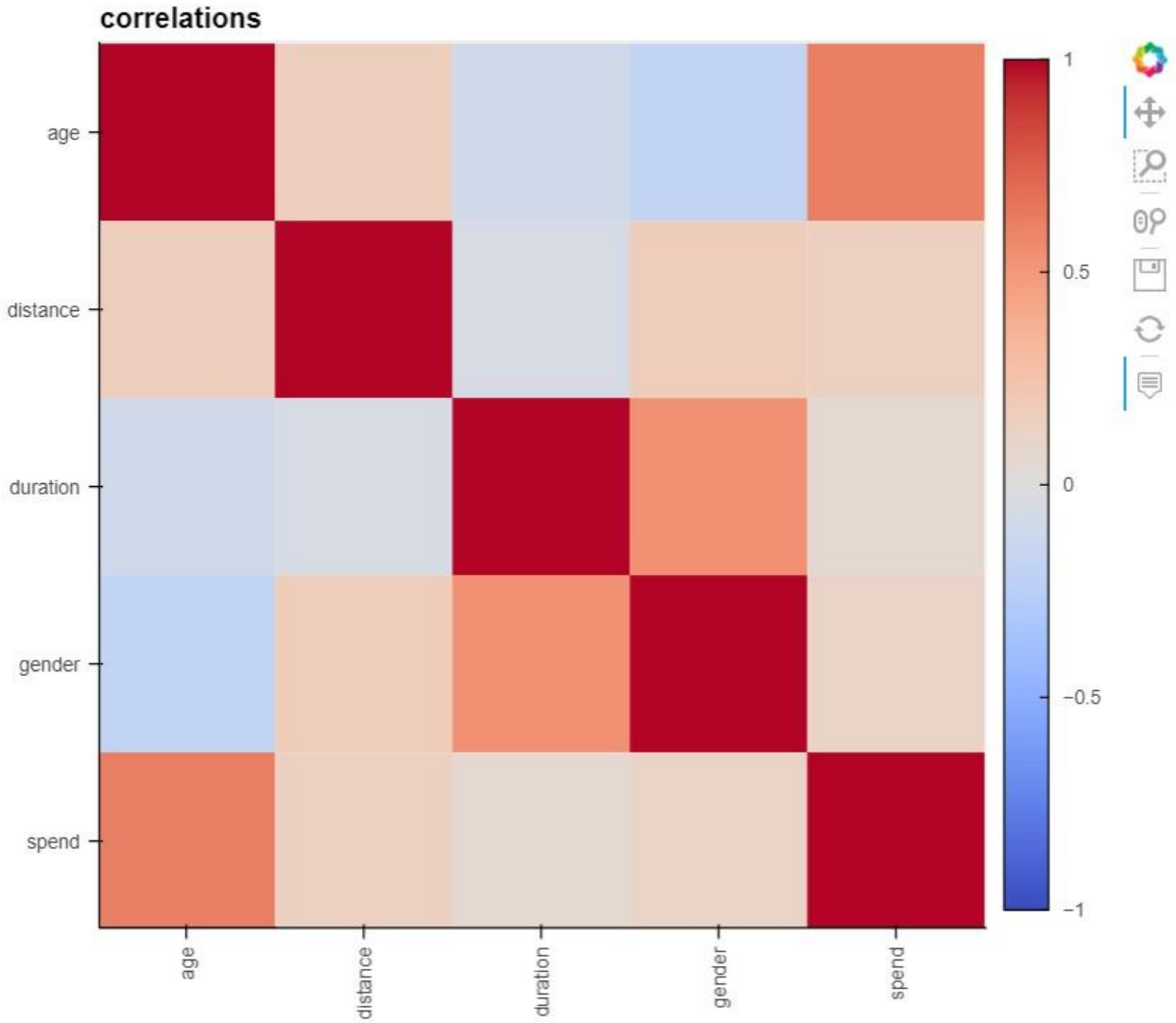

Figure 7 Interactive heat map of the summary

3.8. Histogram of all subplots

A histogram is a graphical representation that divides a set of data points into a specified ranges.

Figure 8, illustrate the group of bar chart containing the information about the distribution of the variable of the summary dataset.

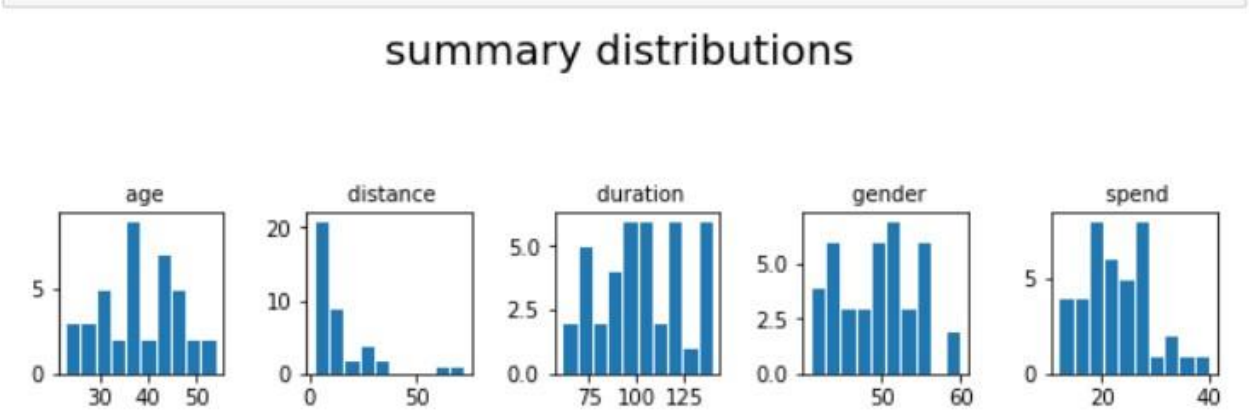

Figure 8 Summary distribution

3. Critical review

The dataset of the ChrisCo Company is taken into a consideration for visualizing. Six dataset (VenueDailyVisitors.csv, VenueAge.csv, VenueDistance.csv, venueDuration.csv, venueGender.csv, VenueSpend.csv) are used for visualizing. Two data frames are created, one containing daily visitors data and other containing summary data including all the CSV files. Segmentation is done in venueDailyVisitors.csv and the dataset is divided into four segment. i.e. High_volume, medium_volume, low_volume, and very_low_volume.

Different visualization approaches are used to investigate these datasets; a total of 8 visualization techniques are used, and various information such as data correlation, seasonality, outliers, and anomalies are discovered. The graphical view of the venueDailyVisitors.csv is displayed using bar charts. The data distribution of high and medium volume is displayed using a box plot. The correlation between the variables of the medium volume is displayed using a heat map. Scatter plots are used to show the association between large volume variables as dots. Stack charts are used to detect high volume opening and closing days. Hvplot is a tool for interactive medium volume visualization. Histogram is used for a graphical representation that divides a set of data points into a specified ranges.

Similarly, by merging the VenueAge.csv, VenueDistance.csv, venueDuration.csv, venueGender.csv, and VenueSpend.csv, the data frame is created to display the summary.

However, several of the data visualization approaches used did not operate as planned. The academic research includes all of the methodologies used to explore the chrisCo data. There are still several efficient visualization approaches left, such as tree maps, correlation matrices, and so on, that might be more efficient.

4. Conclusion

In conclusion, there are several approaches for displaying datasets. The effectiveness of visualization techniques is determined by the nature of the dataset, hence while displaying, it is critical to select the appropriate approaches based on the dataset's nature. Crisco’s dataset is explored as much as possible using various visualization approaches. Visualization extracts many types of data that might be useful to ChrisCo in improving its business. The cycle of data is studied using technical analysis such as correlation, seasonality, outliers, and anomalies.