stratics
engaging your customer

REPORT

Tickets Business Case

Ricardo Castañeda

5/July/2023

Contents

1.0 INTRODUCTION

This report presents the analysis done in the tickets and clients datasets; for this, we covered three main areas: First, an exploratory data analysis (EDA) that aims to reveal relevant and hidden aspects from both datasets (irregularities and trends). Second, the data preprocessing and modeling part, where we aim to modify data features, and in this case perform a customer segmentation analysis using the K-means clustering algorithm. Finally, the data visualization part where we use Power BI to develop interactive dashboards.

2.0 EXPLORATORY DATA ANALYSIS (EDA)

Note: The results associated with this chapter can be seen in the "EDA" notebook.

To start, the tickets dataset was inspected, but no relevant issues were found as there were no duplicates, null values, or evident data inconsistencies. However, some extreme values were identified like customers that paid $\$1971$ and purchased 191 tickets or customers that returned -61 tickets and needed reimbursements of around $\$-1168$, while the average trend was to buy nearly 3 tickets with a related cost of $\$26$. We assumed these extreme values to be real values and not registration errors, but this can be double-checked with management and fixed if necessary.

Following this, the clients' dataset was also inspected, but this case presented more evident inconsistencies. For instance, the variable "age" ranged from -6990 to 523 years, which is a clear error. Moreover, there were null values registered in the form of wrong strings ("(blank)","*", etc) in the columns ("age", "postalCode", "countryCode", and "favoriteStore"). Nonetheless, no duplicated rows were found.

3.0 PREPROCESSING

For this part of the project, we used object-oriented programming (OOP), to create modules with classes and objects (python files with functions) that will modify the data for different parts of the analysis.

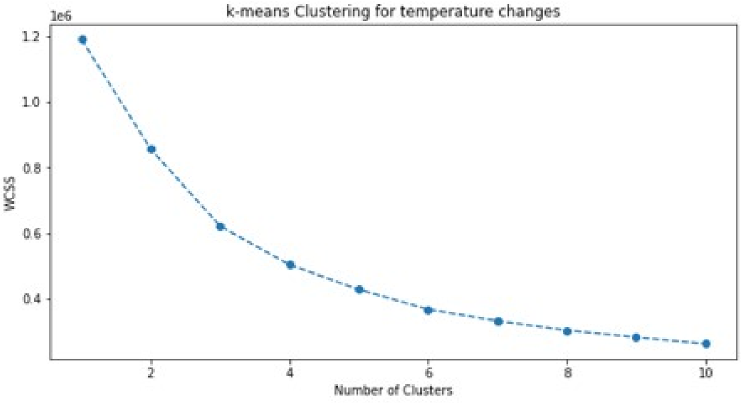
The first module we created aims to clean the client's data. Here we imputed the age values and replace "(blank)" with the median age, and also we replaced ages under 10 and over 95 years old with the mean value as well. We used the median value as this is a robust estimator of the mean and will not be influenced by the wrong values (e.g. -6900). The wrong values in "contryCode" attribute were replaced in Power BI while doing the dashboards.

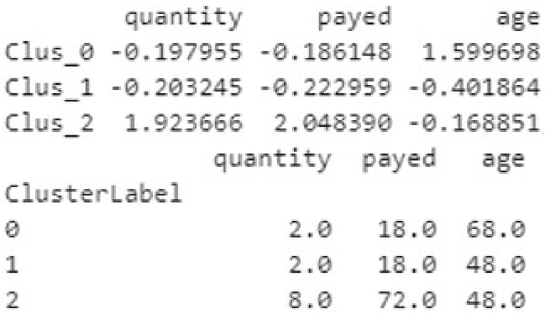
For the tickets dataset, we only created a function that included an extra column to classify positive payments as "sells", and negative payments as "returns". This was mainly done to visualize the impact of positive and negative values in the dashboards.

Finally, there is a module called process that merges the $clients$ and $tickets$ datasets using the "clientID" attribute as the key, and also it adds "s_" to the "storeID" variable, so the computer will recognize it as a category and not as an integer.

4.0 CUSTOMER SEGMENTATION

The goal of this part is to find hidden trends in the data that allow us to classify the different customers, into small groups. Hence, further marketing strategies could be performed based on the specific patterns that these groups reflect. To this end, we implemented the K-means clustering algorithm which is an unsupervised machine-learning method that simply splits N observations into K numbers of clusters. In summary, the algorithm uses the mean (centroids) of each group to assign a classification label to those observations close to those centroids. To remark, the number of clusters was decided based on the elbow method, which comes from trying the model with a different number of clusters (1 to 11 in this case) and plotting the error for all of them, and the point where the graph looks like an elbow indicates the right number of clusters (3 for us). Also, the model only accepts numerical variables, so we used ”age”, ”paid” and ”quantity” for this classification.





```
           quantity     payed       age
Clus_0 -0.197955 -0.186148  1.599698
Clus_1 -0.203245 -0.222959 -0.401864
Clus_2  1.923666  2.048390 -0.168851
              quantity  payed   age
ClusterLabel
0                  2.0   18.0  68.0
1                  2.0   18.0  48.0
2                  8.0   72.0  48.0
```

Figure 4.1: K-means Error Plot and Results

As mentioned before the elbow plot suggested three groups that classify the customers as follows:

- cluster_0: Represents clients that buy few units and present constant returns (-0.19 quantity), that also pay small amounts of money and request constant reimbursements (-0.18 payed), and that in general tend to have an advanced age (1.59 age).
- cluster_1: This group also represents clients with the same trends mentioned in cluster_0, but in this case, they tend to be young people (-0.40 age).
- cluster_2: This group represents the ideal population for the company, as these customers buy tickets constantly or multiple units at a time, which generates high payments (1.92 quantity, 2.04 payed), and they tend to be around 48 years old which is the average age (-0.16 age)

5.0 DASHBOARDS AND EXTRAS

In this part, we create visualizations from the merged dataset "clients_tickets", for top 10 clients, stores performance, and sales per country. it is relevant to mention that individual pages for filters were created in some of the dashboards and with (crl +click) the filters can be accessed. This means that there is a button in the dashboard's windows that takes you to the filters. The purpose of this was to optimize the space and increase the number of visuals per output.

Finally, as an extra effort, we try to use Random fores to make a prediction for future sales, but the performance of the model was low. Therefore, these are not reliable predictions, and that is why no details will be provided. Nevertheless, we keep a notebook with the process as further research can be done regarding this topic and different models can be tried.