Solution Approach for Crayon Data Analytics Hackathon

Date: 22-Dec-19 |  Chennai

Team Ravaangi

Vijendra kumar saini SHA005417 & Vibhaas Vinay Kotwal SHA005341

2019

Contents

[**Introduction** 2](#_Toc27940355)

[Problem definition 2](#_Toc27940356)

[Goal 2](#_Toc27940357)

[Customer Behavior Analysis 3](#_Toc27940358)

[**Literature review** 4](#_Toc27940359)

[Binary classification 4](#_Toc27940360)

[**Methodology** 5](#_Toc27940361)

[Framework 5](#_Toc27940362)

[Tool selection 6](#_Toc27940363)

[**Data** 7](#_Toc27940364)

[Useful observations or visualizations made on the data 7](#_Toc27940365)

[Data understanding 9](#_Toc27940366)

[Features 10](#_Toc27940367)

[Transaction Aggregation by customer 11](#_Toc27940368)

[**Implementation** 13](#_Toc27940369)

[Classification 13](#_Toc27940370)

[**Appendix A** 16](#_Toc27940371)

[Algorithm Implementations 16](#_Toc27940372)

# **Introduction**

## Problem definition

The theme of the challenge is to predict the future behavior of users based on what they have done in the past.

## Goal

Generate transaction for the customers based on the historical data available. Please refer to the data dictionary for the columns available for you to model.

The sample output is expected to have either  
 Result type 1: Customer Id, Merchant Id (Mcc\_code), Frequency of transaction  
  OR  
 Result type 2: Customer Id, Date, Merchant Id

## Customer Behavior Analysis

To predict the customer behavior first we will have to analyze the data provided for transactions done. Based on the transactions Customer first needs to be segmented into buyer personas based on their common characteristics.

Once you have customer segmented, we can predict the future transactions based on the behaviors and spending trends of a Customer segment.

1. **Segment your audience**

Its important to understand Customer characteristics. Demographic traits such as gender, age, and location as well as engagement tendencies like product category.

1. **Identify the key benefit for each group**

Each customer persona will have its own unique reason for purchasing items, and it's extremely to identify it. Looking at historical data we need to get insights into these reasons for each segment.

1. **Machine learning models**

To perform such analysis, we will have to make use of statistics, and machine learning models comes in handy for doing this type of analysis. We will see in coming sections what all models we can utilize and which one will work best in our case.

# **Literature review**

We are listing down some of the commonly available models but we are not explaining in length as these are very commonly available and pretty standard models.

## Binary classification

Based on historical data, we can predict whether the purchase will or will not be done in future. If the outcome need to be in True/False, we can use Binary classification.

The most common type of machine learning algorithms for a binary classification task are vector-based methods. Belonging to this category are Decision Tree, Random Forest, SVMs and Logistic Regression.

* 1. **Decision Trees**

DTs consist of a set of split conditions which divide a heterogeneous population into smaller, more homogeneous subgroups regarding a certain variable. The aim is to create the most homogeneous subgroups.

* 1. **Random Forest**

Bagging is another example for ensemble trees, where many large trees are fit to the bootstrap re-sampled versions of the data and are classified by majority vote. RF improves on Bagging by de-correlating the trees. After each tree split a random sample of features is chosen and only these are considered for the next split. The results are again based on the majority vote of the single trees. By using a large number of classifiers, Bagging and RFs, improve on the weaknesses of non-ensemble DTs, such as robustness and over-fitting.

* 1. **Support Vector Machines**

An SVM separates two classes by fitting a hyperplane between them. Doing this, only one hyperplane is used, which differs from DTs where a hyperplane is added after each split.

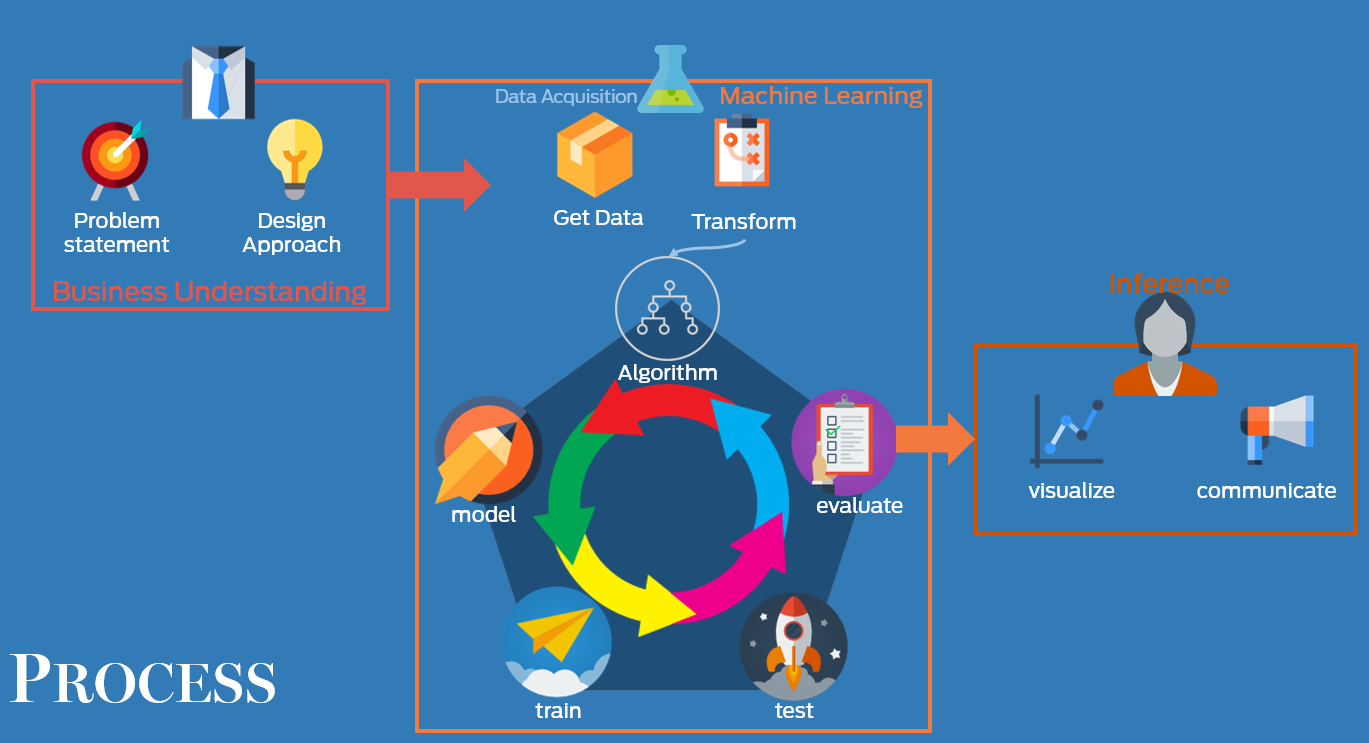
* 1. **Logistic Regression**

LR, also called Logit regression, belongs to the class of generalized linear models and is used to predict categorical target variables. This is achieved through a logistic function, which has the shape of a sigmoid curve, taking values between 0 and 1.

# **Methodology**

This chapter contains an explanation of the methodology framework that was used to structure this study, followed by the explanation of evaluation metrics on which the model comparison will be focused.

## Framework

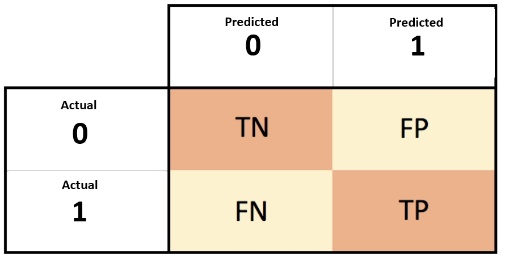


Steps are as follows:

1. **Business understanding**
2. **Data understanding**
3. **Data preparation**
4. **Modeling**
5. **Evaluation of the model**
6. **Deployment of the model**
7. **Evaluation metrics**

This section describes the evaluation metrics used to compare the different algorithms. Confusion matrix, showing true positives, true negatives, false positives.

and false negatives.



## Tool selection

Python will be used for implementing the different machine learning algorithms. Python is a general-purpose high-level programming language. It is widely used today by machine learning community due to its rich libraries that contain various predictive analytics algorithms. One of the most known libraries is Scikit-learn, which will also be used in this thesis (Sk-learn, 2017). It provides state-of-the-art implementations of many machine learning algorithms Libraries:

1. scikit-learn
2. seaborn
3. numpy
4. pandas
5. matplotlib

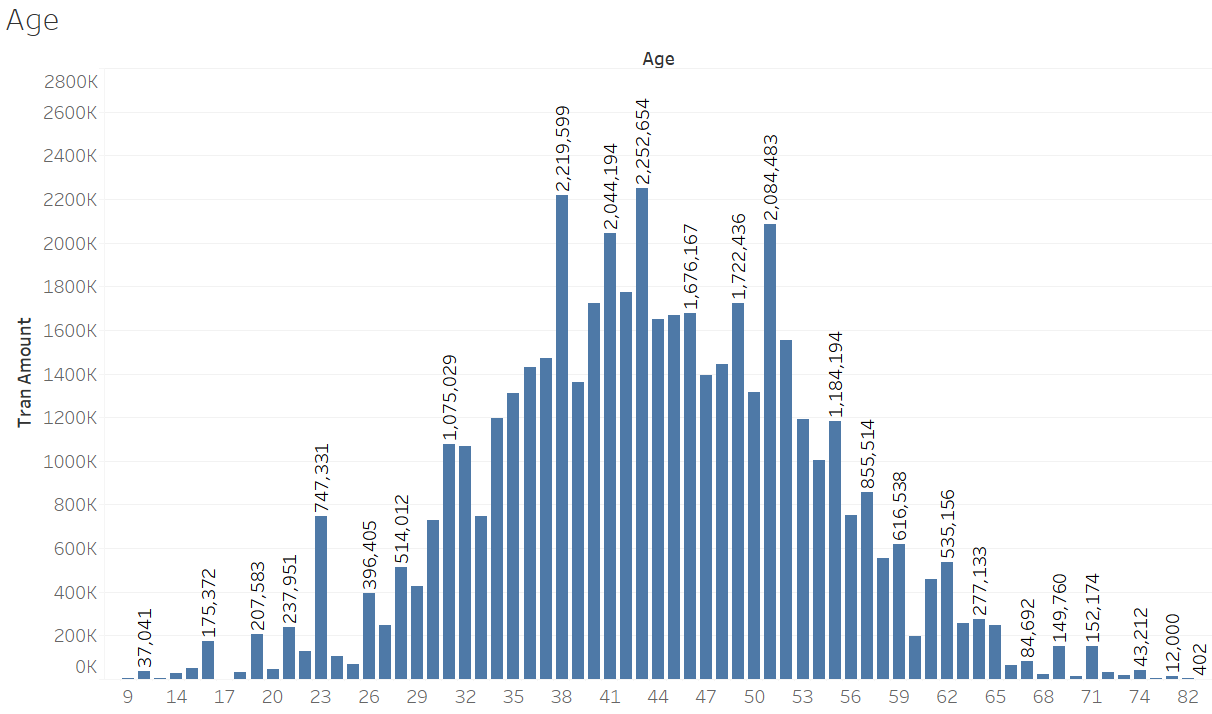
Alteryx & Tableau has ben used for data manipulation and visualization.

# **Data**

In this part, the data being used for the training and testing of the models is described.

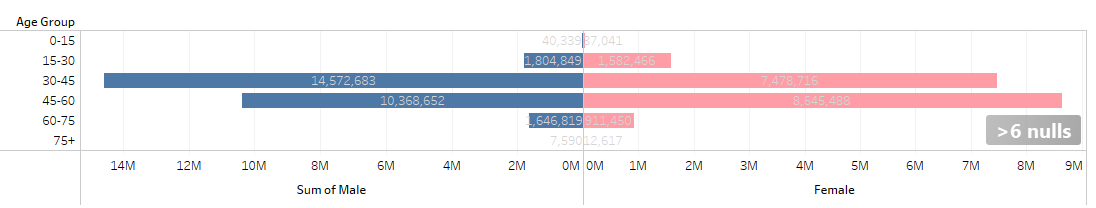
The utilized data is transaction & customer demographics dataset.

## Useful observations or visualizations made on the data

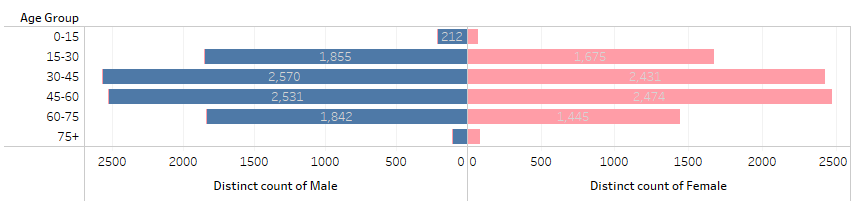


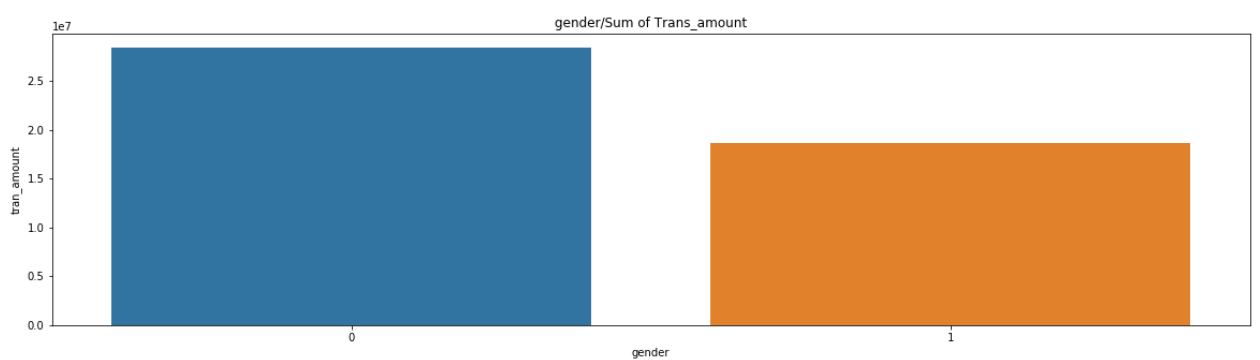
The trend of the age on the transaction amount.

This clearly is a Normal curve and clearly customers from around 35 to 55 are big spenders.

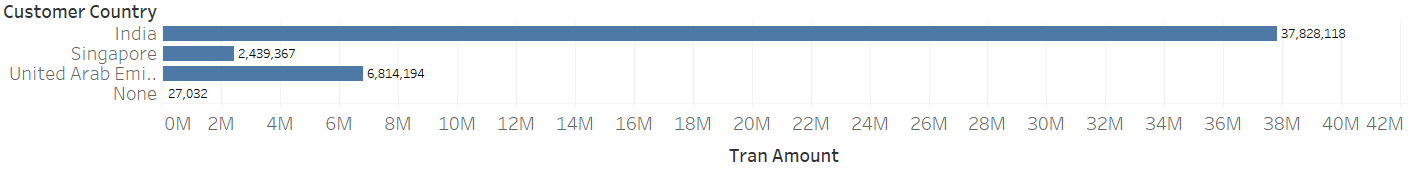


The trend of the age group and gender on the transaction amount. Above shows us that male spent more than females in the age group 30-45, but this trend changes drastically for age group 45-60 where total spent by females is very close to males in same age group.

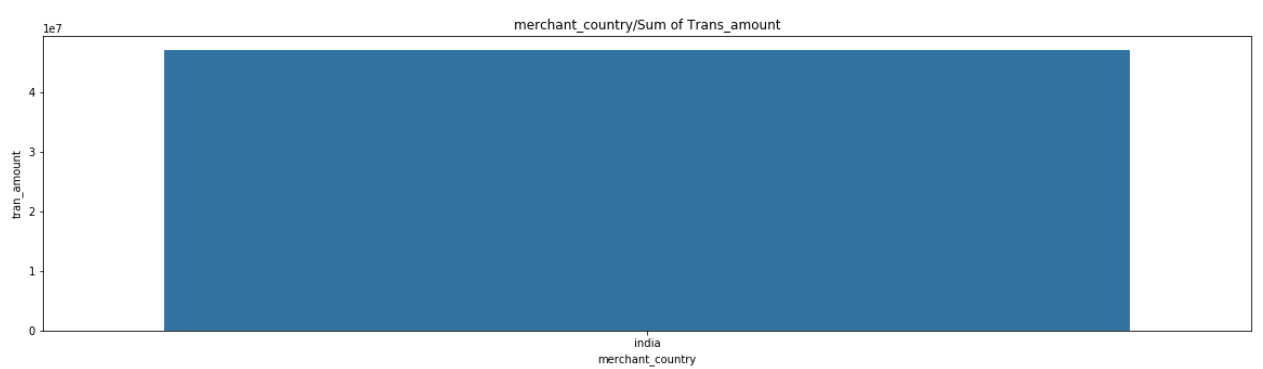
 Count of male/female in both groups.

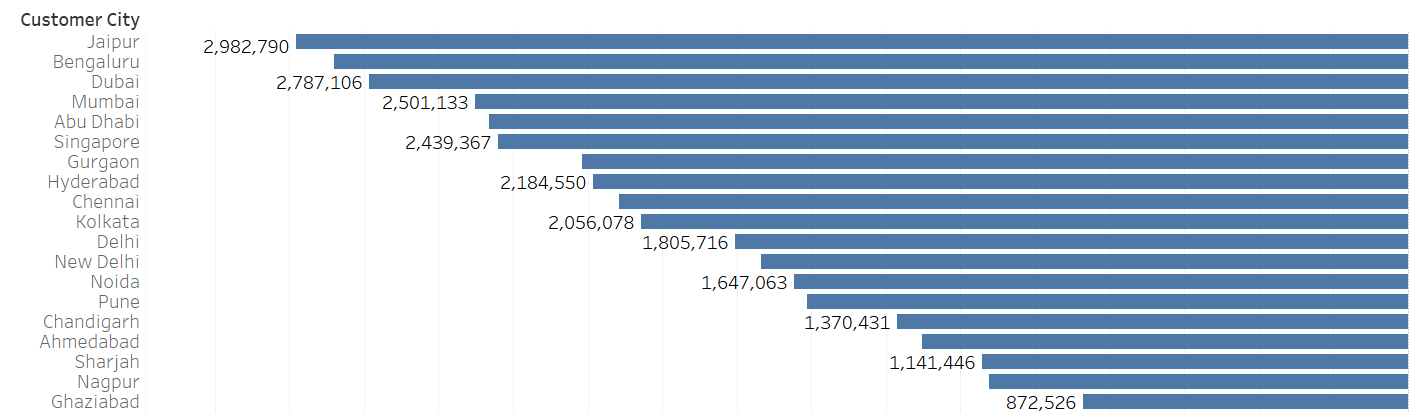
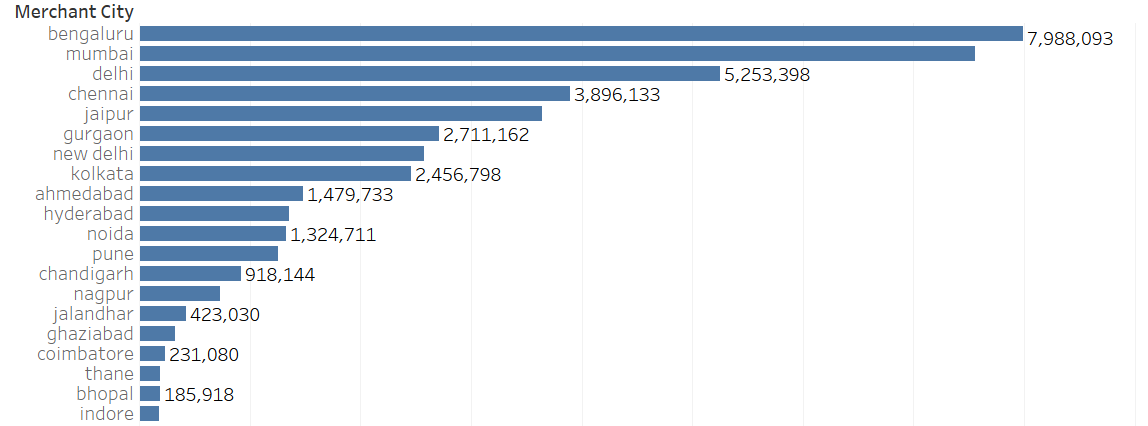


Above bar graph is self-explanatory. Females(0) overall spend more than males( taken as 1)

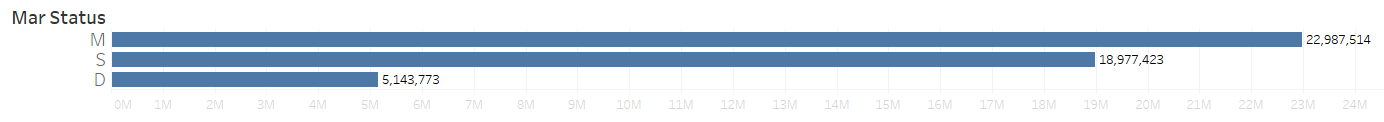


The trend of the customer country on the transaction amount.

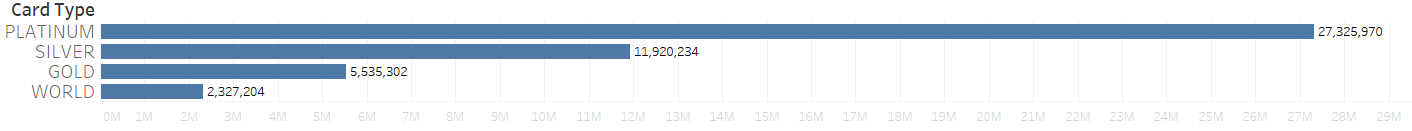


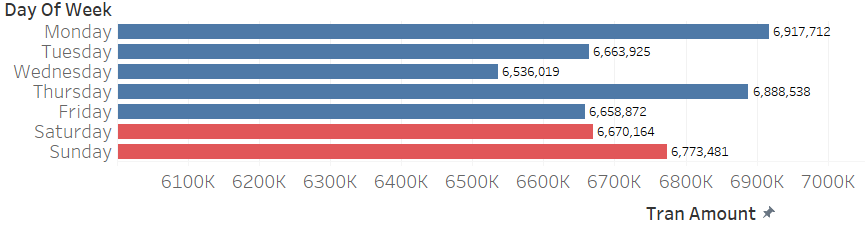
The trend of the customer & merchant city on the transaction amount.



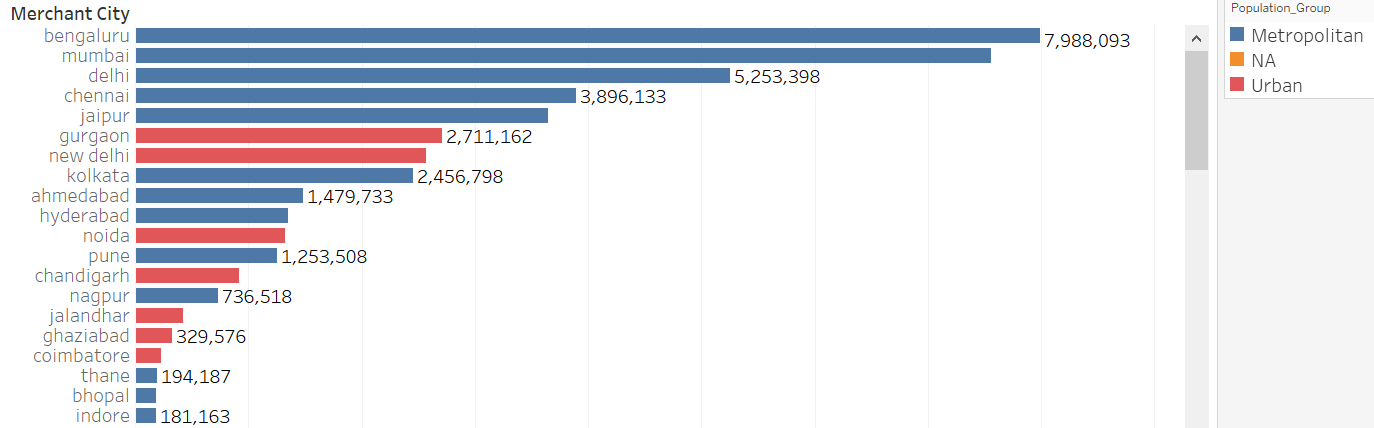
The trend of the marital status on the transaction amount.



The trend of the card type on the transaction amount.



The trend of the weekday on the transaction amount.



The trend of the merchant city on the transaction amount.

## Data understanding

To generate a first understanding of both the sequence as well as the customer data

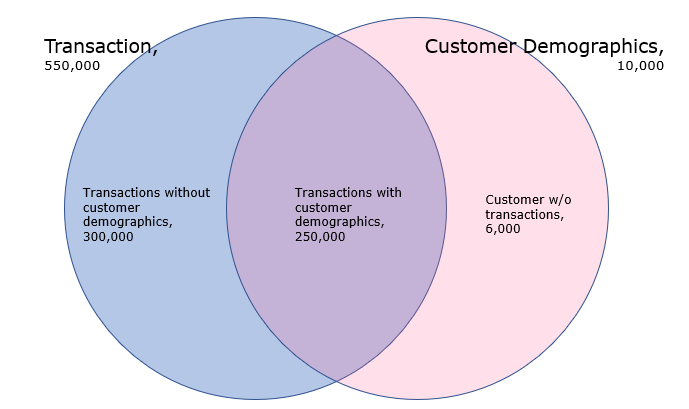
a first exploratory data analysis based on Excel was conducted.

* 1. **Transaction data set**

|  |  |  |
| --- | --- | --- |
| transaction |  |  |
| customer\_id | Unique customer ID | |
| tran\_id | Unique transaction ID | |
| tran\_date | Transaction Date |  |
| tran\_amount | Transaction Amount |  |
| merchant\_name | Merchant Name |  |
| merchant\_country | Merchant Country |  |
| merchant\_city | Merchant City |  |
| mcc\_code | Merchant category code | |
| card\_id | Unique card ID |  |

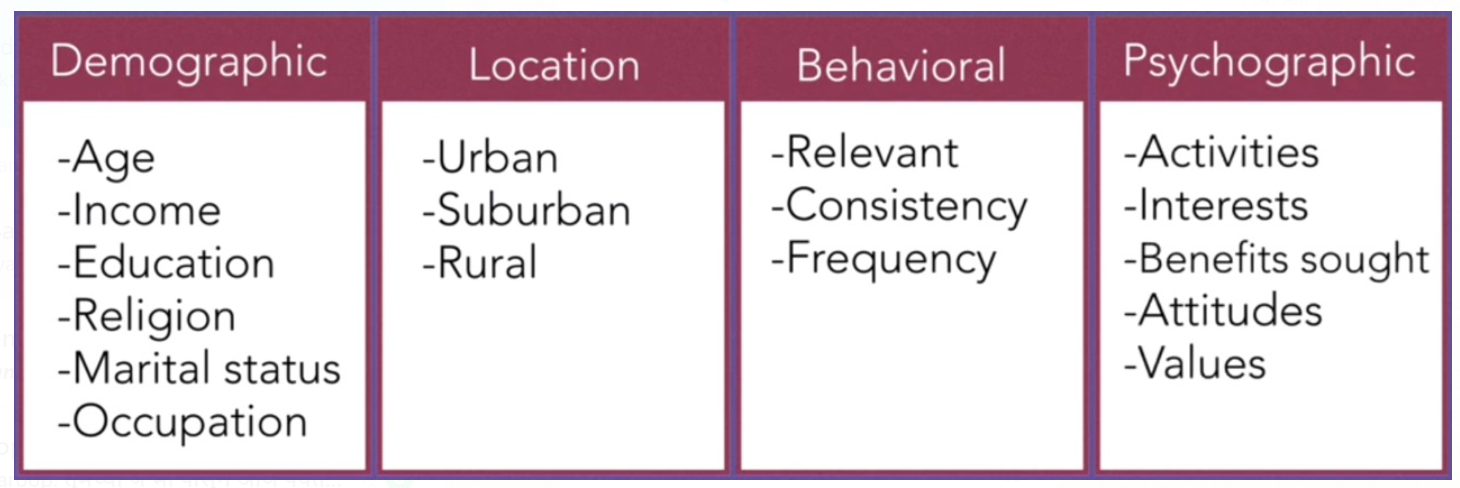
* 1. **Customer data set**

|  |  |  |
| --- | --- | --- |
| Customer Demographics | |  |
| customer\_id | Unique customer ID | |
| card\_type | CardType | SILVER, GOLD, PLATINUM, WORLD |
| card\_id | Unique card ID |  |
| mar\_status | Marital Status | S, M, D |
| age | Customer Age |  |
| gender | Gender | M, F |
| customer\_country | Customer Country |  |
| cr\_lim\_group | Credit limit |  |
| customer\_city | Customer Ciy |  |
| customer\_id\_new | Unique customer ID | |
| customer\_uid | Unique customer ID | |



## Features

Attributes for customer profiling.



* 1. **Feature engineering**
* **MCC codes:** A Merchant Category Code (MCC) is a four-digit number listed in ISO 18245 for retail financial services. An MCC is used to classify a business by the types of goods or services it provides. ["List of MCC codes in CSV, ODS, XLS formats"](https://github.com/greggles/mcc-codes). *github*.
* City Classification: Cities classified by population (0, 1, 2) <https://en.wikipedia.org/wiki/Classification_of_Indian_cities>
* Age Groups: group of 10 years interval created
  1. **Categorical variables**

Most algorithms require for categorical variables to be encoded.

Dummy variable created for following features :

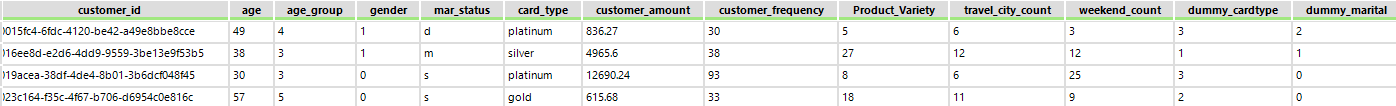
* dummy\_india
* dummy\_marital
* dummy\_gender
* dummy\_cardtype
* dummy\_crlimit
  1. **Missing values**

Lastly, missing values had to be imputed. The most common techniques to solve the problem of missing values is to either remove, predict or impute them.

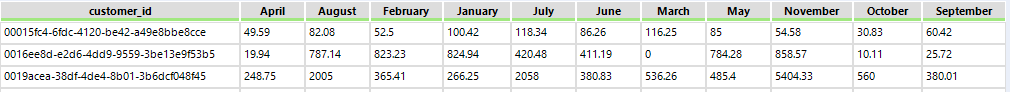
* merchant\_city
* age\_group
* customer\_city
* Product Category – NA
* Mer\_city\_grp
* Cust\_city\_grp
* Mer\_state
* Cust\_state

## Transaction Aggregation by customer id

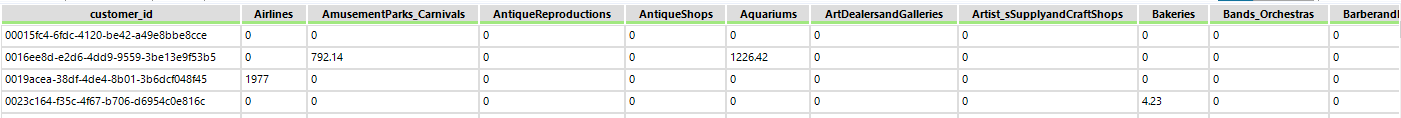
* 1. **Transaction Summary**



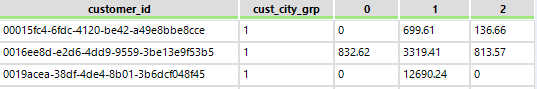
* 1. **Transaction – Month trend**



* 1. **Transaction – Product Interest**



* 1. **Transaction – City group**



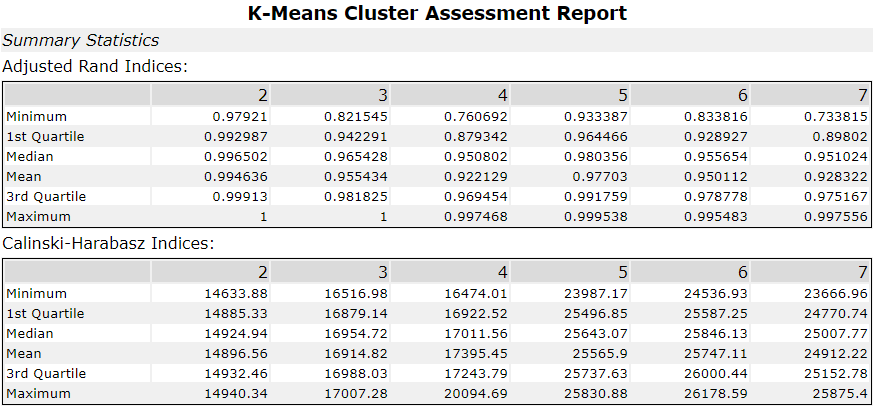
# **Implementation**

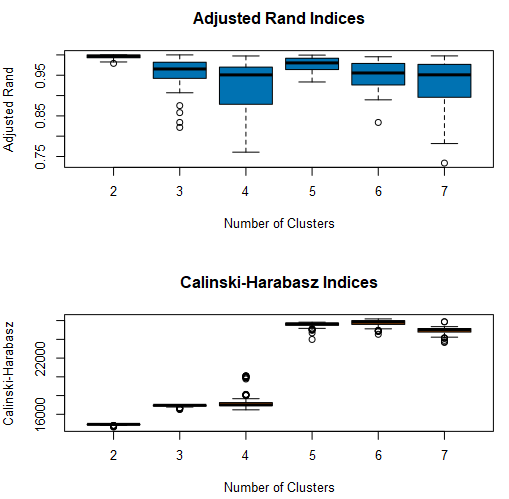
## Classification

* 1. **K means**

K means is used for clustering.

Selecting number of clusters: The report itself includes two representations of each of these metrics, a table and a box-plot. For each target cluster number, the index is calculated for all bootstrap replicates. The tables and plots report the indices of each of these trials as overall distributions.





Below is a standard text from :

<https://community.alteryx.com/t5/Alteryx-Designer-Knowledge-Base/Tool-Mastery-K-Centroids-Diagnostics/ta-p/302311>

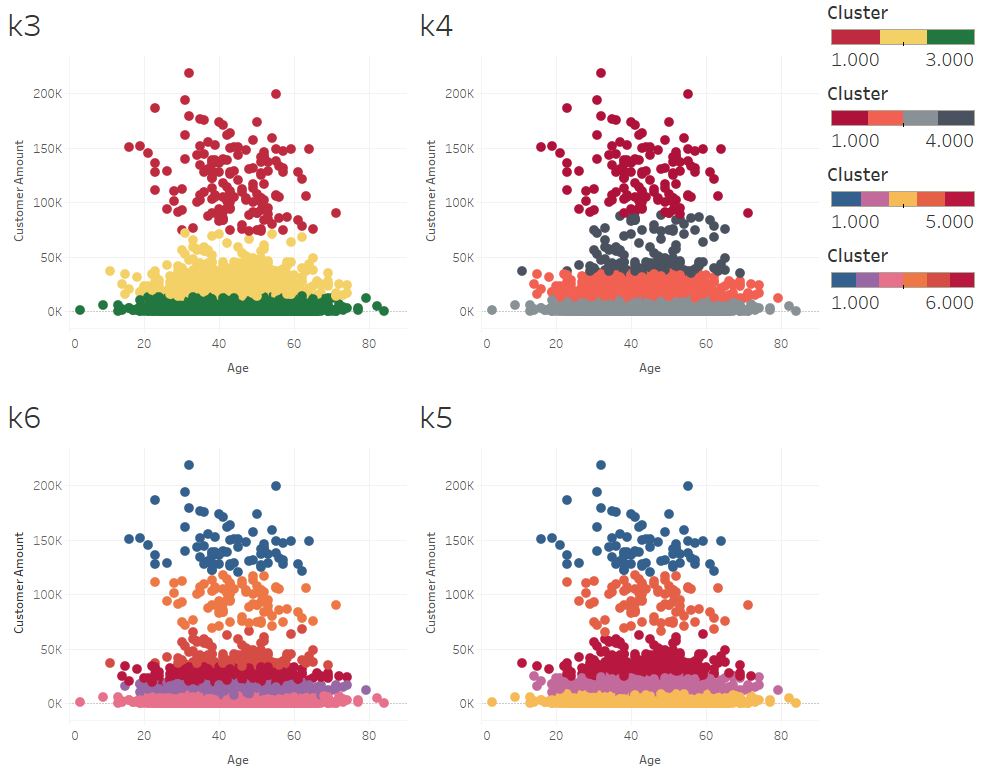
*The table is reported in quartiles, which highlight the overall distribution of the calculated values. The minimum Adjusted Rand and Calinski-Harabasz Indices are reported as the first row in the table, the middle value between the minimum and median (Q1), the median (Q2), and so on.*

*The box-plots are a visualization of the distributions of index values across the bootstrap replicates (and clustering solutions). The second and third quartiles are depicted by the shaded box, the dashed lines display the spread of the first and fourth quartiles, and points indicate any outliers.*

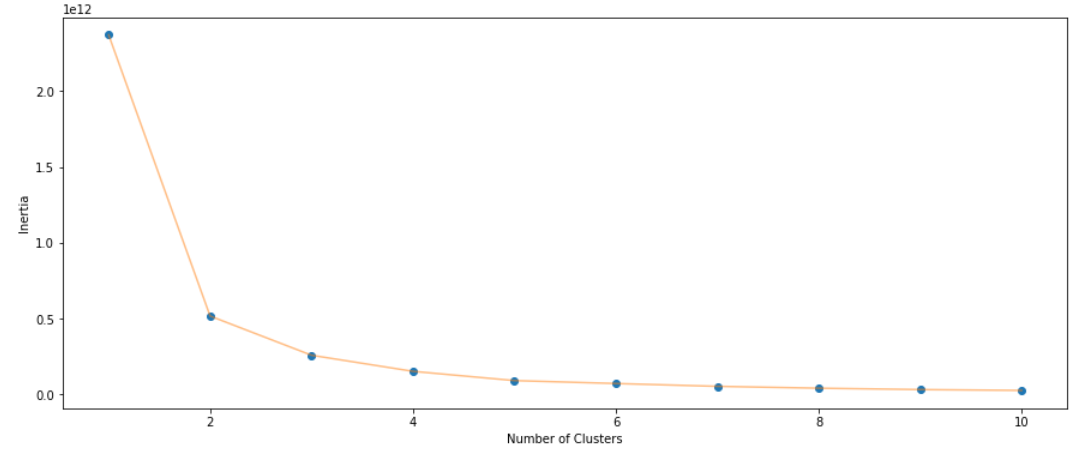
*The Adjusted Rand Index is on a scale of 0 to 1, where 0 indicates totally random clusters. It is a measure of agreement between clustering partitions. Higher values indicate higher agreement.*

*For the Calinski-Harabasz Indices, higher values also indicate a better solution. The Calinski-Harabasz index tends to be most accurate when the clusters and approximately spherical in shape, and compact in the middle (i.e., normally distributed). This index all tends to prefer cluster consisting roughly of the same number of records.*

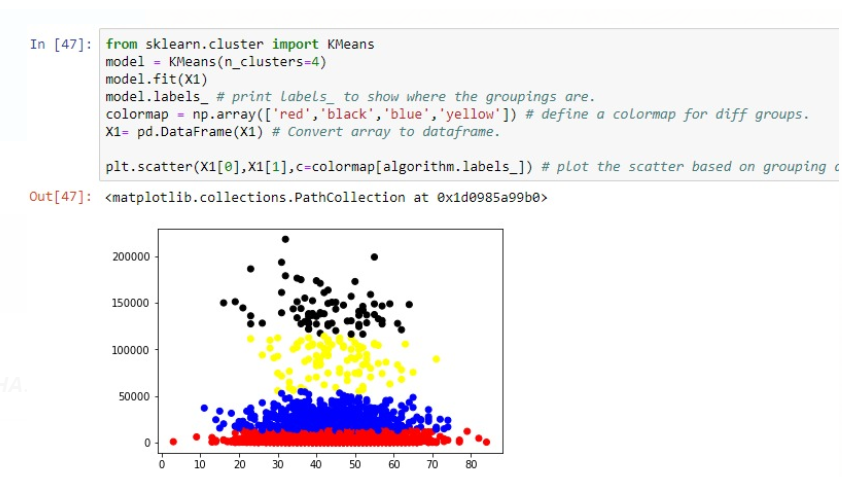
*As a rule, you are looking for the number of clusters where the index values are maximized, and the spread of the quartiles are minimized. In this example, two clusters seems to result in consistently high values for the Adjusted Rand Index, and three clusters has a high median, but at least one very low index value. The Calinski-Harabasz index shows that the 3-cluster solutions seem to have the higher values, although there is an low-value outlier.*



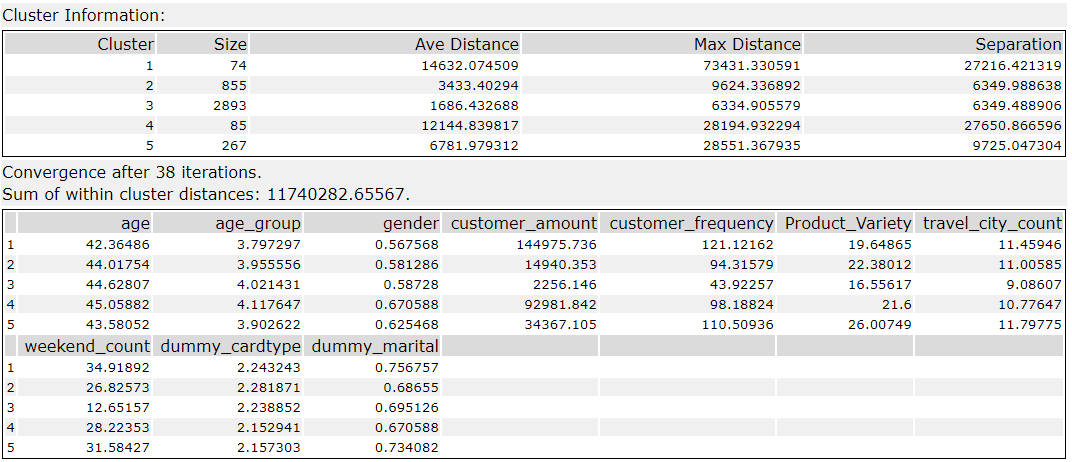
Selecting N Clusters based in Inertia (Squared Distance between Centroids and data points, should be less). Also called as Elbow method.



**4 selected**



**5 selected**



Average of features by cluster.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster | Avg\_age | Avg\_age\_group | Avg\_gender | Avg\_dummy\_marital | Avg\_customer\_amount | Avg\_customer\_frequency |
| 1 | 42.36486 | 3.797297 | 0.567568 | 0.756757 | 144975.7 | 121.1216 |
| 2 | 44.01754 | 3.955556 | 0.581287 | 0.68655 | 14940.35 | 94.31579 |
| 3 | 44.62807 | 4.021431 | 0.58728 | 0.695126 | 2256.146 | 43.92257 |
| 4 | 45.05882 | 4.117647 | 0.670588 | 0.670588 | 92981.84 | 98.18824 |
| 5 | 43.58052 | 3.902622 | 0.625468 | 0.734082 | 34367.1 | 110.5094 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Avg\_Product\_Variety | Avg\_travel\_city\_count | Avg\_weekend\_count | Avg\_dummy\_cardtype |
| 1 | 19.64865 | 11.45946 | 34.91892 | 2.243243 |
| 2 | 22.38012 | 11.00585 | 26.82573 | 2.281871 |
| 3 | 16.55617 | 9.08607 | 12.65157 | 2.238852 |
| 4 | 21.6 | 10.77647 | 28.22353 | 2.152941 |
| 5 | 26.00749 | 11.79775 | 31.58427 | 2.157303 |

# **Appendix A**

## Algorithm Implementations

K Means for clustering

