Market Segmentation Analysis on OLIST E-Commerce Data

MIS 64038 Project Report, Spring 2020

Sumanth Avunuri, savunuri@kent.edu
Harika Penjerla, E-mail: spenjerl@kent.edu
Mallikarjun Sasnur, E-mail: msasnur@kent.edu
Chujun Huang, E-mail: chuang23@kent.edu
Susmita Singh, E-mail: ssingh35@kent.edu

Contents

1.Business Requirements

Background

Market Analysis is a process of uncovering insights into your marketplace by surveying a representative sample of its participants. Consumer segmentation, product segmentation and sales forecasting are few strategies that primarily helps business to improve their marketplace. understanding the preferences, attitudes, and behaviors of consumers in a market-based economy, and it aims to understand the effects and comparative success of marketing campaigns. The importance of it is widely recognized among business and this strategies and dashboards could potentially help stakeholders take better decision that significantly impacts businesses.

Business Problem

The data might contribute in determining sales improvements, customer churn and delivery performance and could also help sellers in product recommendations. This project aims to improve customer retention, delivery performance and sales prediction with respect to geo codes.

2.Overview of data

Data Sources

Most of the datasets that concerns with orders, products, customers and geo-locations have been downloaded from here. The datasets however have many duplicates and redundant values. We can observe dataset model below.

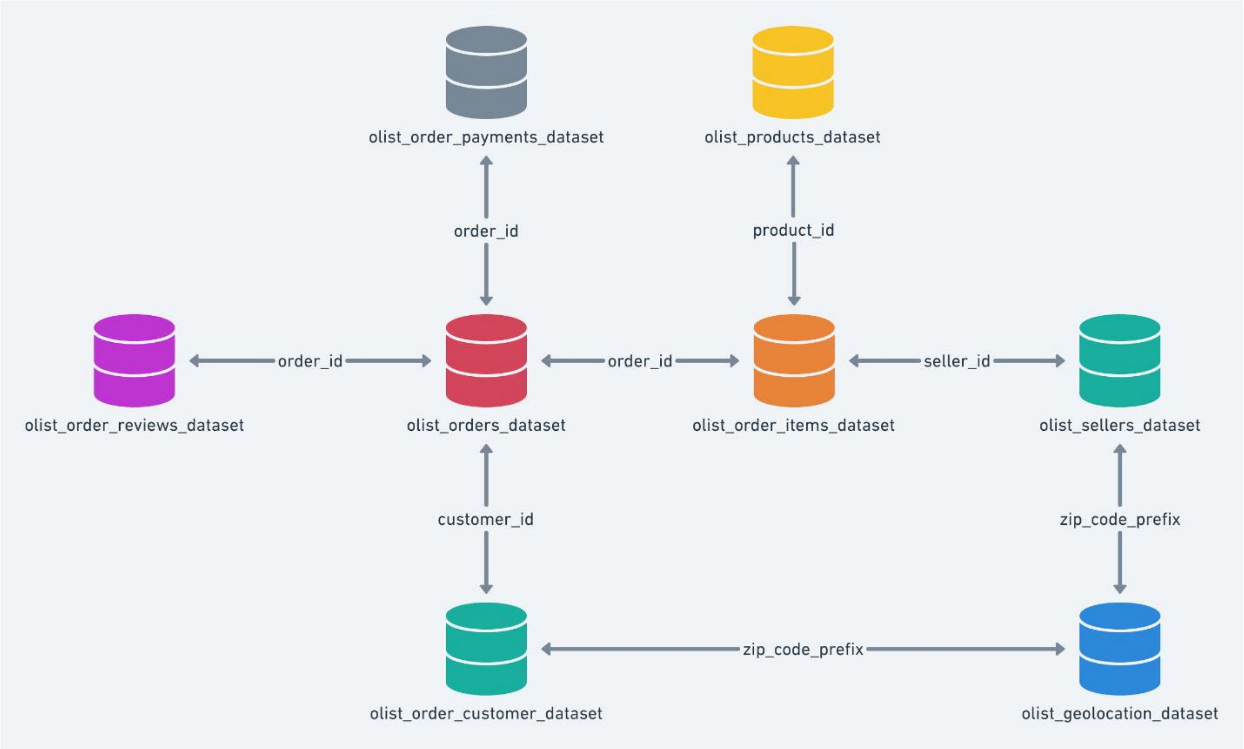
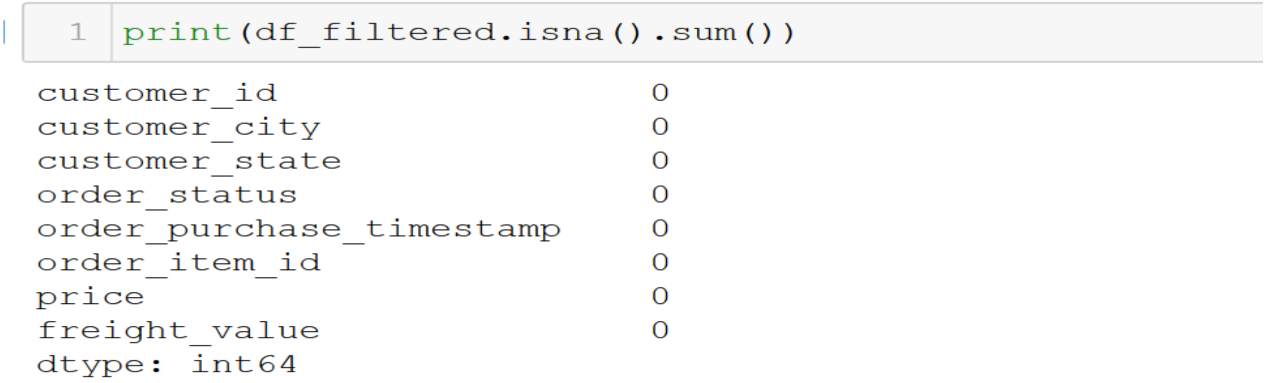


Figure 1 Data Scheme

Data Cleaning

The Team has performed data cleaning using python and R for two types of analysis. We removed all the missing values, removed duplicates, and merged the data using unique columns.

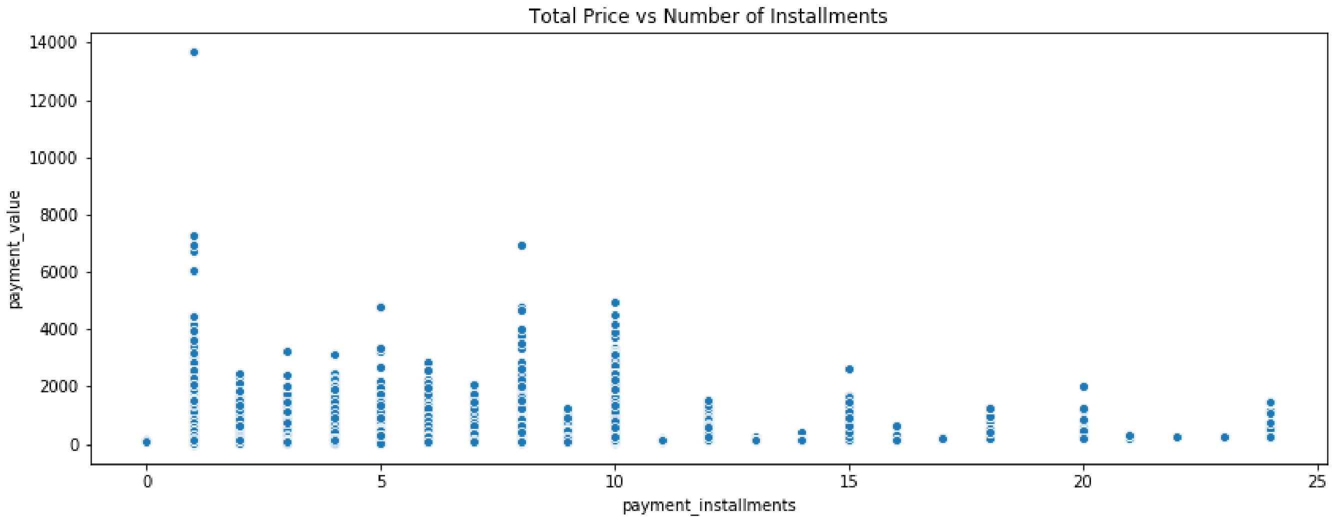


```
print(df_filtered.isna().sum())
```

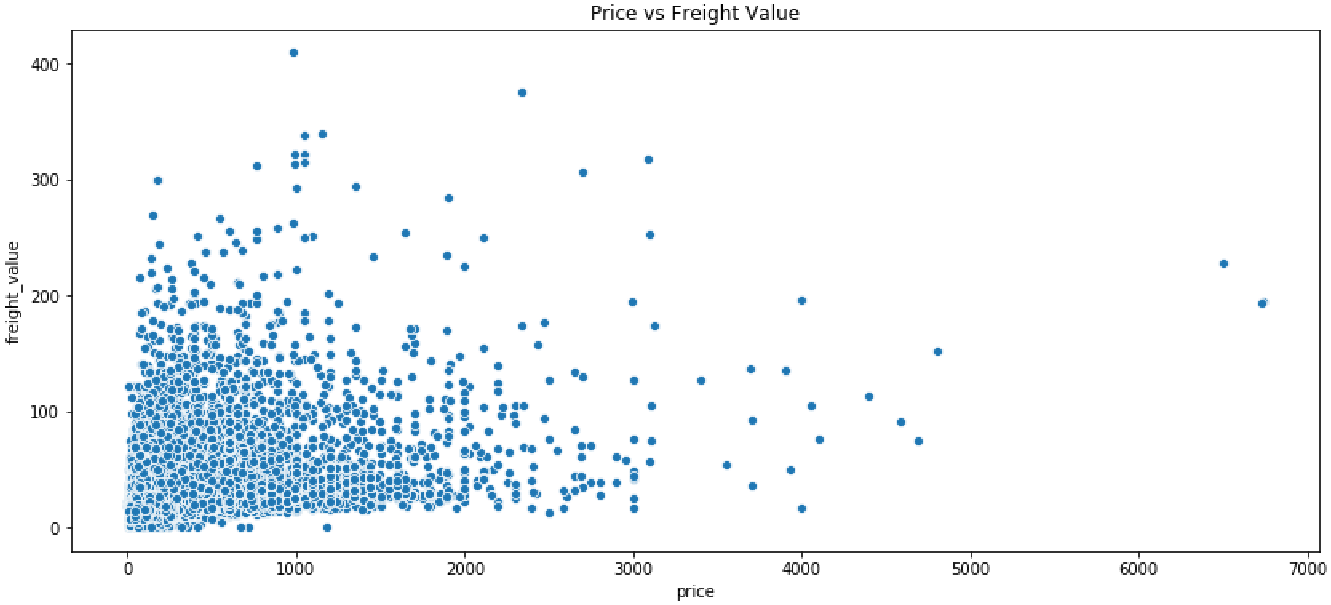
```
customer_id                 0
customer_city               0
customer_state              0
order_status                0
order_purchase_timestamp    0
order_item_id               0
price                       0
freight_value               0
dtype: int64
```

3.Data Exploration Analysis

By observing the relationship between payment value and payment installments, we can infer that most of the installment of the customers ranges from 2 to 10 with a payment value ranges from 0 to 5000 dollars.



There is steep correlation between the price of the order and freight value. We must also understand that there is no causal relationship between them.



4.Details of Modeling Strategy

The Team has chosen two types of modelling strategies to support our objective i.e., Time series forecasting using Auto regression and K means clustering for segmentations by RFM analysis.

Time series forecasting

An autoregressive (AR) model **predicts future behavior based on past behavior**. It is used for forecasting when there is some correlation between values in a time series and the values that precede and succeed them. You *only* use past data to model the behavior, hence the name *auto*regressive. The process is basically linear regression of the data in the current series against one or more past values in the same series.

An AR(p) model is an autoregressive model where specific lagged values of y_t are used as predictor variables. Lags are where results from one time period affect following periods.

The value for "p" is called the *order*. For example, an AR would be a "first order autoregressive process." The outcome variable in a first order AR process at some point in time *t* is related only to time periods that are one period apart (i.e. the value of the variable at t – 1). A second or third order AR process would be related to data two or three periods apart.

The AR(p) model is defined by the equation:

$$y_t = \delta + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-1} + A_t$$

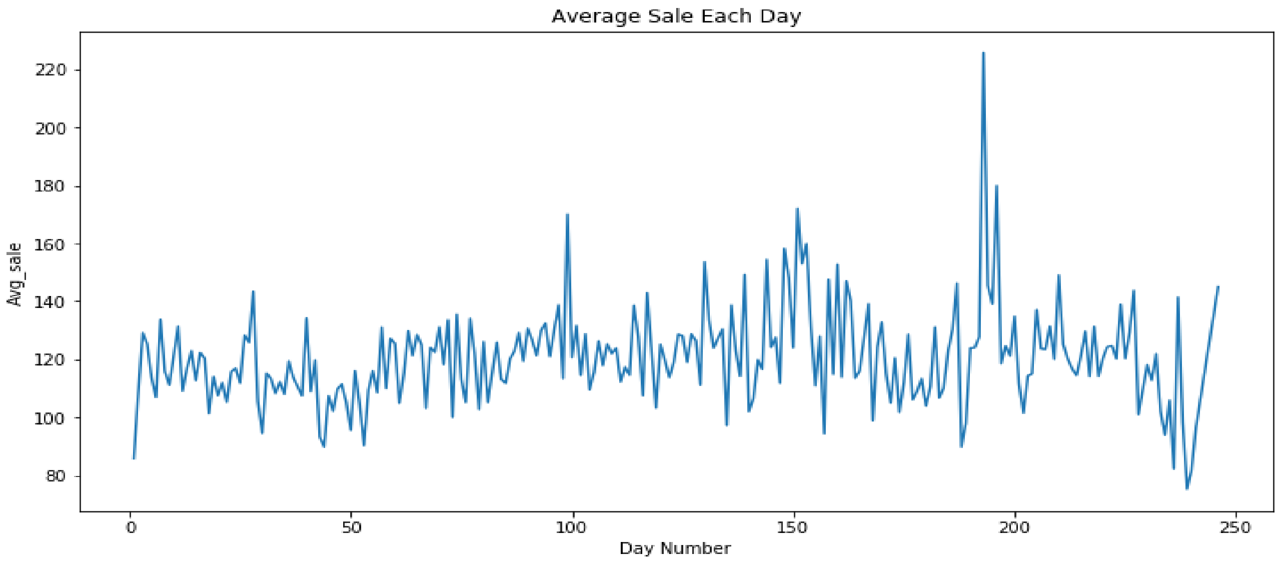
Where:

- y_{t-1}, $y_{t-2} \ldots y_{t-p}$ are the past series values (lags),
- A_t is white noise (i.e. randomness),
- and δ is defined by the following equation:

$$\delta = \left(1 - \sum_{i=1}^{p} \phi_i\right) \mu \text{ ,}$$

Where μ is mean.

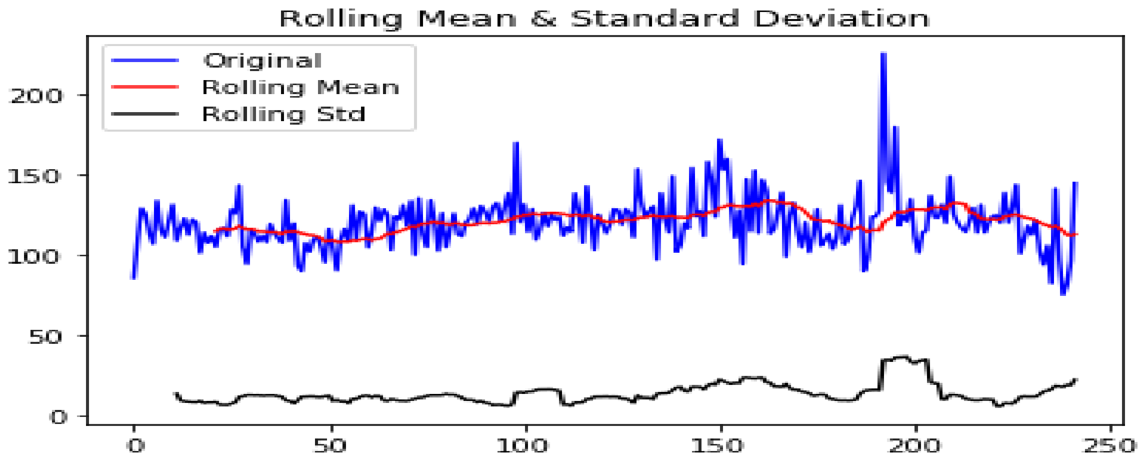
The nature of the Average sales to days trend is stationary. The way we concluded the series whether its stationary or non-stationary is by using dickey fuller test. There is no seasonality observed in the graph.



The rolling mean and standard deviation would show mean line to the window of that time series. This would help us in determining the fluctuations in the series.

1. Plotting rolling statistics
2. Dickey-Fuller test

Here the null hypothesis is that the TS is non-stationary. The test results comprise of a Test Statistic and some Critical Values for difference confidence levels. If the 'Test Statistic' is less than the 'Critical Value', we can reject the null hypothesis and say that the series is stationary.

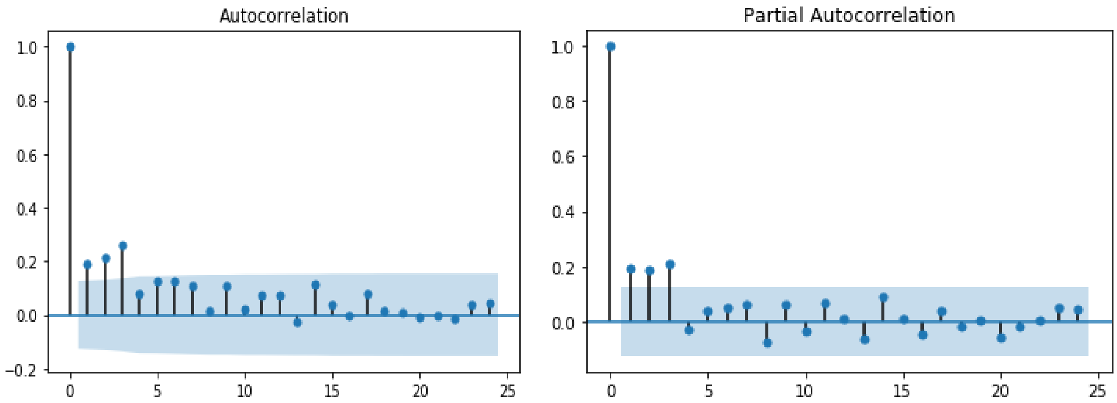


```
ADF Stastistic: -5.704077
p-value: 0.000001
```

```
The graph is stationery
Critical values:
        1%: -3.458
        5%: -2.874
        10%: -2.573
```

From the above dickey fuller test, we can confirm that the series is stationary. Now, let’s check the autocorrelation and lag plot.

```
price     price
price  1.000000  0.193328
price  0.193328  1.000000
```



The stats model’s library provides an autoregression model where you must specify an appropriate lag value and trains a linear regression model. It is provided in the AR() class.

We can use this model by first creating the model AR() and then calling fit() to train it on our dataset. This returns an model_fit() object. Once fit, we can use the model to make a prediction by calling the predict() function for a number of observations in the future. This creates 40-day forecast.

Below are the model results on the train data. The model chooses lag value as 14.

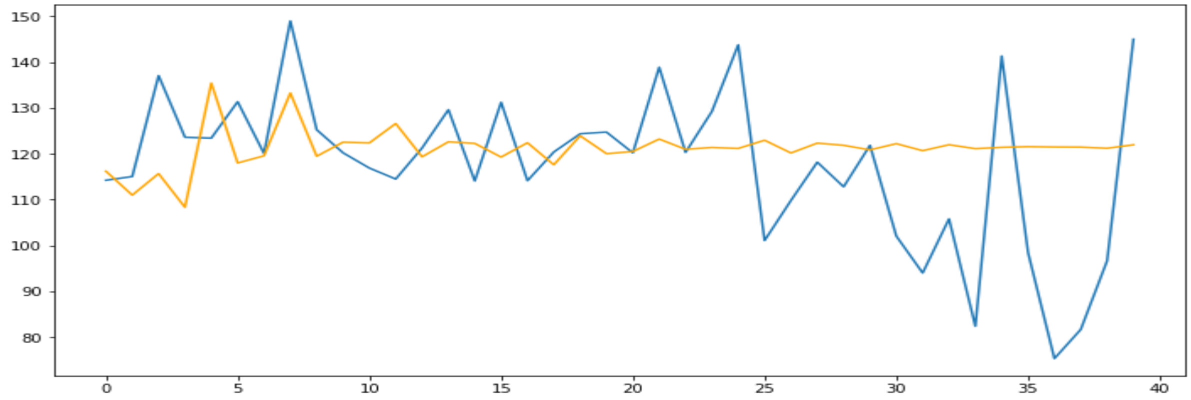
```
  AR Model Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  201
Model:                         AR(14)   Log Likelihood                -777.714
Method:                          cmle   S.D. of innovations             15.486
Date:                Thu, 30 Apr 2020   AIC                              5.651
Time:                        19:48:00   BIC                              5.927
Sample:                             0   HQIC                             5.763

==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         49.7904     19.987      2.491      0.013      10.616      88.965
L1.y           0.1168      0.077      1.519      0.129      -0.034       0.268
L2.y           0.2047      0.077      2.657      0.008       0.054       0.356
L3.y           0.2452      0.079      3.116      0.002       0.091       0.399
L4.y          -0.0669      0.081     -0.827      0.408      -0.226       0.092
L5.y          -0.0962      0.081     -1.192      0.233      -0.254       0.062
L6.y           0.0510      0.081      0.634      0.526      -0.107       0.209
```

```
L7.y            0.1197      0.082       1.464      0.143      -0.041       0.280
L8.y           -0.0684      0.082      -0.837      0.403      -0.229       0.092
L9.y            0.0776      0.081       0.956      0.339      -0.082       0.237
L10.y          -0.0728      0.091      -0.801      0.423      -0.251       0.105
L11.y           0.0191      0.091       0.211      0.833      -0.159       0.197
L12.y          -0.0207      0.089      -0.232      0.816      -0.196       0.154
L13.y          -0.0777      0.088      -0.883      0.377      -0.250       0.095
L14.y           0.1585      0.089       1.773      0.076      -0.017       0.334
```

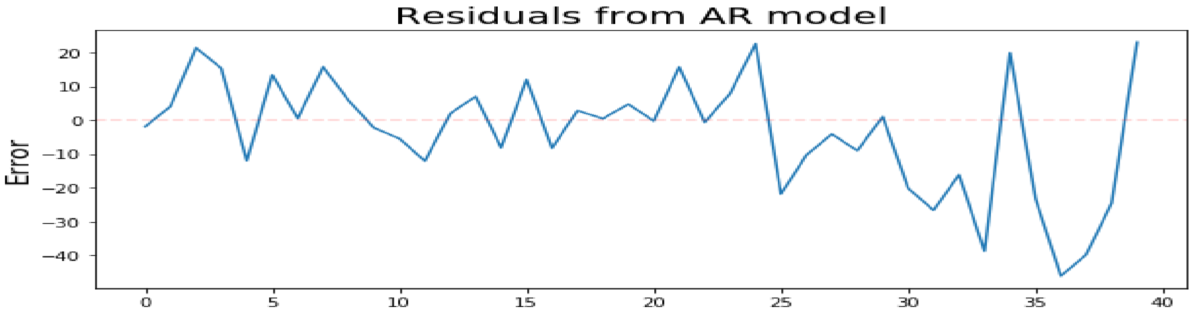
```
The lag value chose is: 14
Coefficients:
 [ 4.97903649e+01  1.16843049e-01  2.04666701e-01  2.45182997e-01
 -6.69013491e-02 -9.62345596e-02  5.10130457e-02  1.19711790e-01
 -6.83890160e-02  7.76463727e-02 -7.27649655e-02  1.91299821e-02
 -2.07247659e-02 -7.76613088e-02  1.58505803e-01]
```

The model predicted values are plotted against the real values as we can see though the series has dynamic fluctuations, we managed to fit the optimal series.



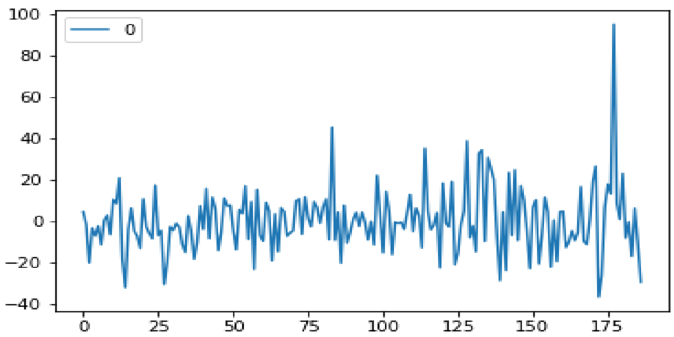
Estimation of the model’s performance

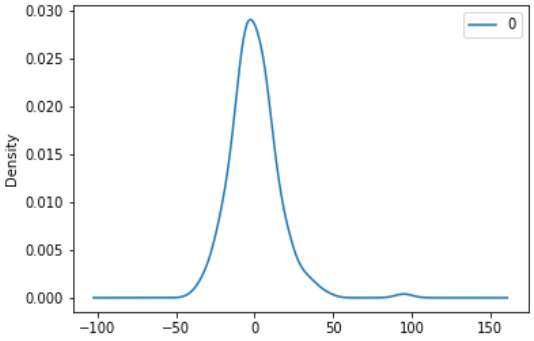
The difference between what was expected and what was predicted is called the residual error plot. We managed to get decent generalized plot restricting error to min.



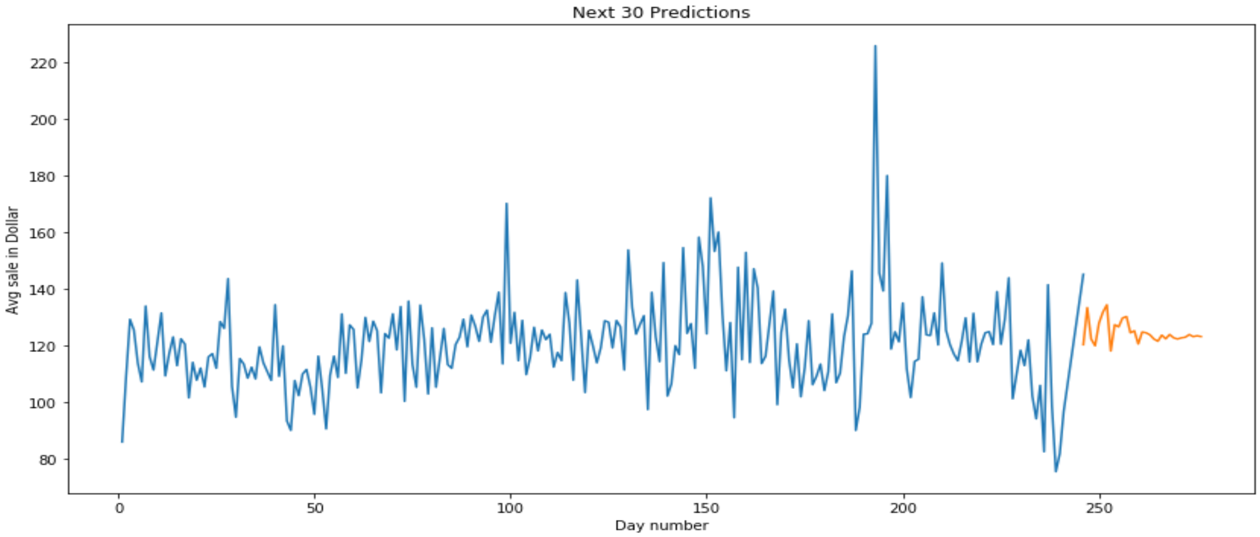
[1]

We can also the residual plot and density plot of the total sales of the order data.





The final model forecasted plot with the prediction values of next 30 days. We restricted our predictions to 30 days since the sales are highly dynamic and we should be very cautions while predicting for longer time periods.



[1] Code File: TimeSeries Forecasting.ipynb

Customer Segmentation Using K-means

Customer segmentation is used in marketing to better understand customer behavior of the business. The most common types of customer segmentation are

Demographic Segmentation

Geographic Segmentation

Behavioral Segmentation

Lifecycle Based Segmentation

Segmentation used for this analysis was based on the purchase behavior of the customers based on features like recency, frequency, monetary values which is commonly known as RFM.

Recency: Number of days the user has been inactive, from the moment of last purchase to the latest time in t[2]he dataset.

Frequency: Total number of transactions completed by a customer.

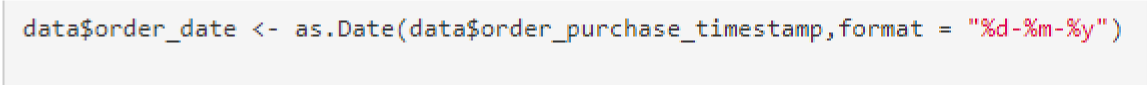
Monetary: Total revenue generated by the customer.

Import Data:

Initially we need to join the data tables and import them to the R studio.

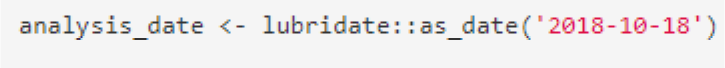
Data Cleaning:

Changing the order timestamp to date format using as.Date



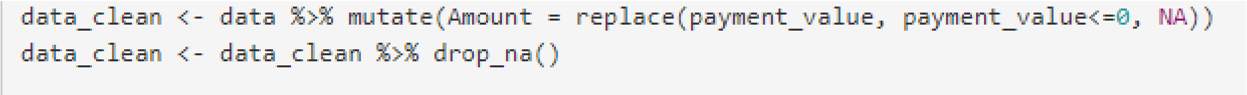
```
data$order_date <- as.Date(data$order_purchase_timestamp,format = "%d-%m-%y")
```

Consider analysis date for calculating recency value



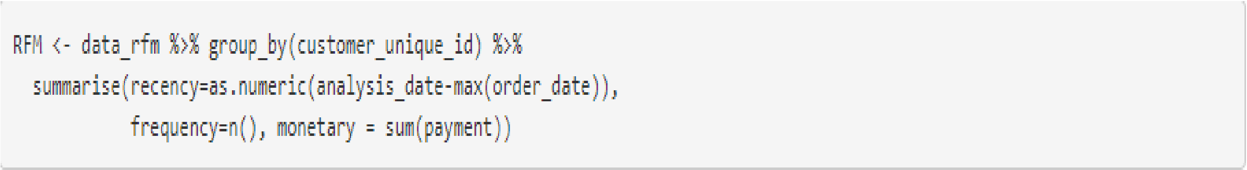
```
analysis_date <- lubridate::as_date('2018-10-18')
```

Removing any negative values from payment



```
data_clean <- data %>% mutate(Amount = replace(payment_value, payment_value<=0, NA))
data_clean <- data_clean %>% drop_na()
```

Calculating RFM values

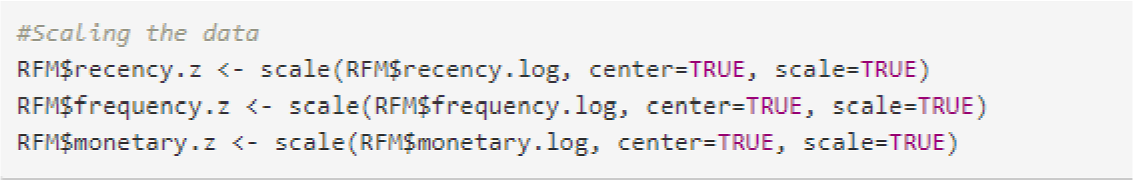


```
RFM <- data_rfm %>% group_by(customer_unique_id) %>%
  summarise(recency=as.numeric(analysis_date-max(order_date)),
            frequency=n(), monetary = sum(payment))
```

[2] Code File: Clustering.Rmd

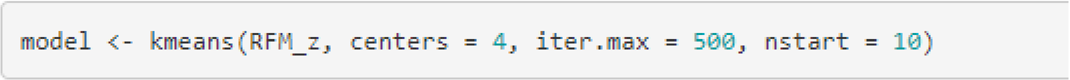
The segmentation is performed using K-means clustering. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. A cluster in this algorithm refers to collection of data points aggregated together because of certain similarities.

Important step when K-means clustering is to be scaling the data because in RFM data we have different units for the different variables. So scaling represent the true distance among the variables.



```
#Scaling the data
RFM$recency.z <- scale(RFM$recency.log, center=TRUE, scale=TRUE)
RFM$frequency.z <- scale(RFM$frequency.log, center=TRUE, scale=TRUE)
RFM$monetary.z <- scale(RFM$monetary.log, center=TRUE, scale=TRUE)
```

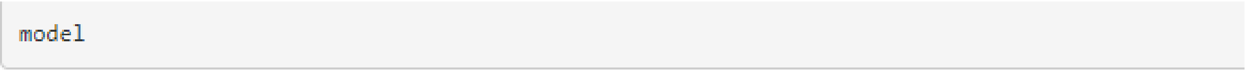
Now applying the K-means algorithm in r using kmeans function and k =4 (clusters)



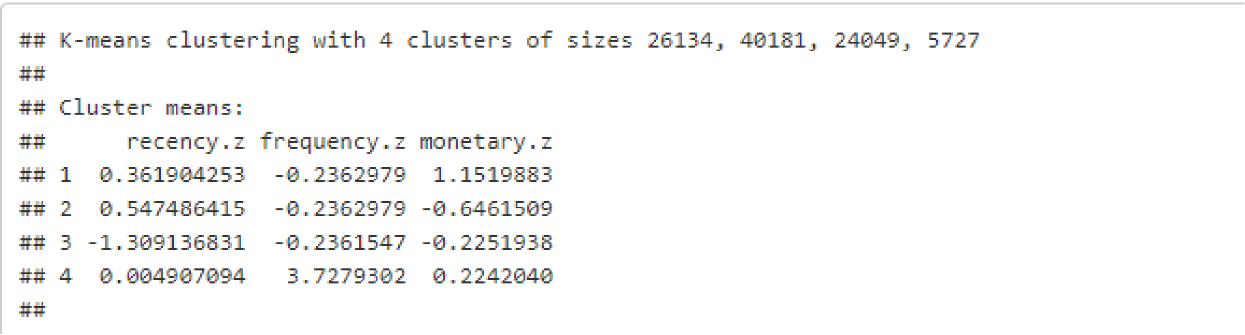
```
model <- kmeans(RFM_z, centers = 4, iter.max = 500, nstart = 10)
```

Estimation of the model's performance

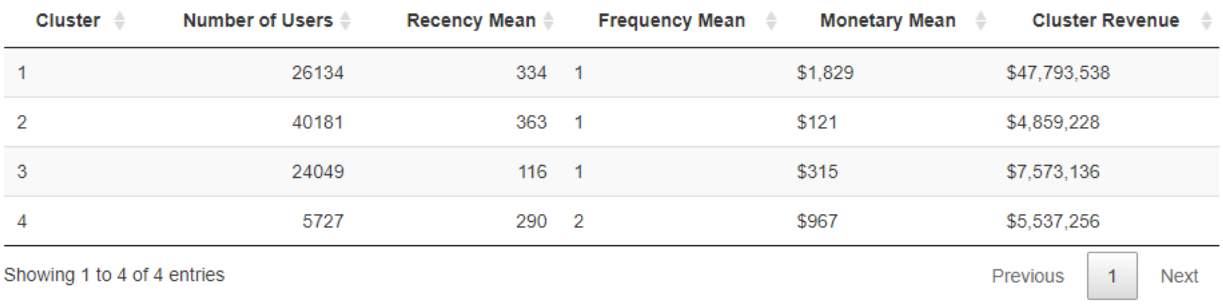
K-means model gives 4clusters with following characteristics.



```
model
```



```
## K-means clustering with 4 clusters of sizes 26134, 40181, 24049, 5727
##
## Cluster means:
##      recency.z frequency.z monetary.z
## 1  0.361904253  -0.2362979  1.1519883
## 2  0.547486415  -0.2362979 -0.6461509
## 3 -1.309136831  -0.2361547 -0.2251938
## 4  0.004907094   3.7279302  0.2242040
##
```



Cluster	Number of Users	Recency Mean	Frequency Mean	Monetary Mean	Cluster Revenue
1	26134	334	1	$1,829	$47,793,538
2	40181	363	1	$121	$4,859,228
3	24049	116	1	$315	$7,573,136
4	5727	290	2	$967	$5,537,256

Showing 1 to 4 of 4 entries

Previous
1
Next

From above k-means clustering analysis with 4 clusters.

Cluster1: Customers who spend more “Big Spenders” Cluster.

Cluster3: Customers who spend recently.” Most Recent” Cluster.

Cluster4: Customers who spend frequently. “Loyal Customers” Cluster.

Cluster2: Customers who are Inactive, lost etc. “Others” Cluster.

5.Insights and Conclusions

The main contribution of this two-stage modelling plus Dashboard building to current scholarly discourse is that it provides a more data-driven approach in generating the business strategies for stakeholders. This methodology can be readily applied to other customer segmentation and other Quantitative Marketing Analysis. Because the clustering process consists only of unsupervised learning techniques, some clusters may have no economic explanations, hence resulting in less significance when regressed against the data.

- By using customer segmentation analysis, we can mainly focus on the special offers/promotions to increase loyalty among customers. Also select best communication channel for each group of customers and improve marketing strategies. For example, the product promotion can prioritize based on the recency, other customers segments who is likely to make purchases. Loyal customers need more attention in this business so by providing more promotions/offers can be a useful strategy.
- Credit cards is the most used Payment type by the customers so to improve customer satisfaction by offering credit card rewards points. Sales forecasting for the next 30 days seems to be around 130 million dollars which is less than last two months. So, it necessary improve customer online experience and product categories.
- Time series prediction results shows that the next 15 days, the sales could potentially hit low. This could assist organization officials in taking a better decision.
- The dashboard assists with the visual interpretation on various aspects of the data like customer segmentations, product segmentations and geo location analysis.

References

http://en.wikipedia.org/wiki/RFM_(market_research)

https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/

https://www.tableau.com/learn/tutorials/on-demand/dashboard-interactivity-using-actions

https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/

https://www.business2community.com/customer-experience/4-types-of-customer-segmentation-all-marketers-should-know-02120397

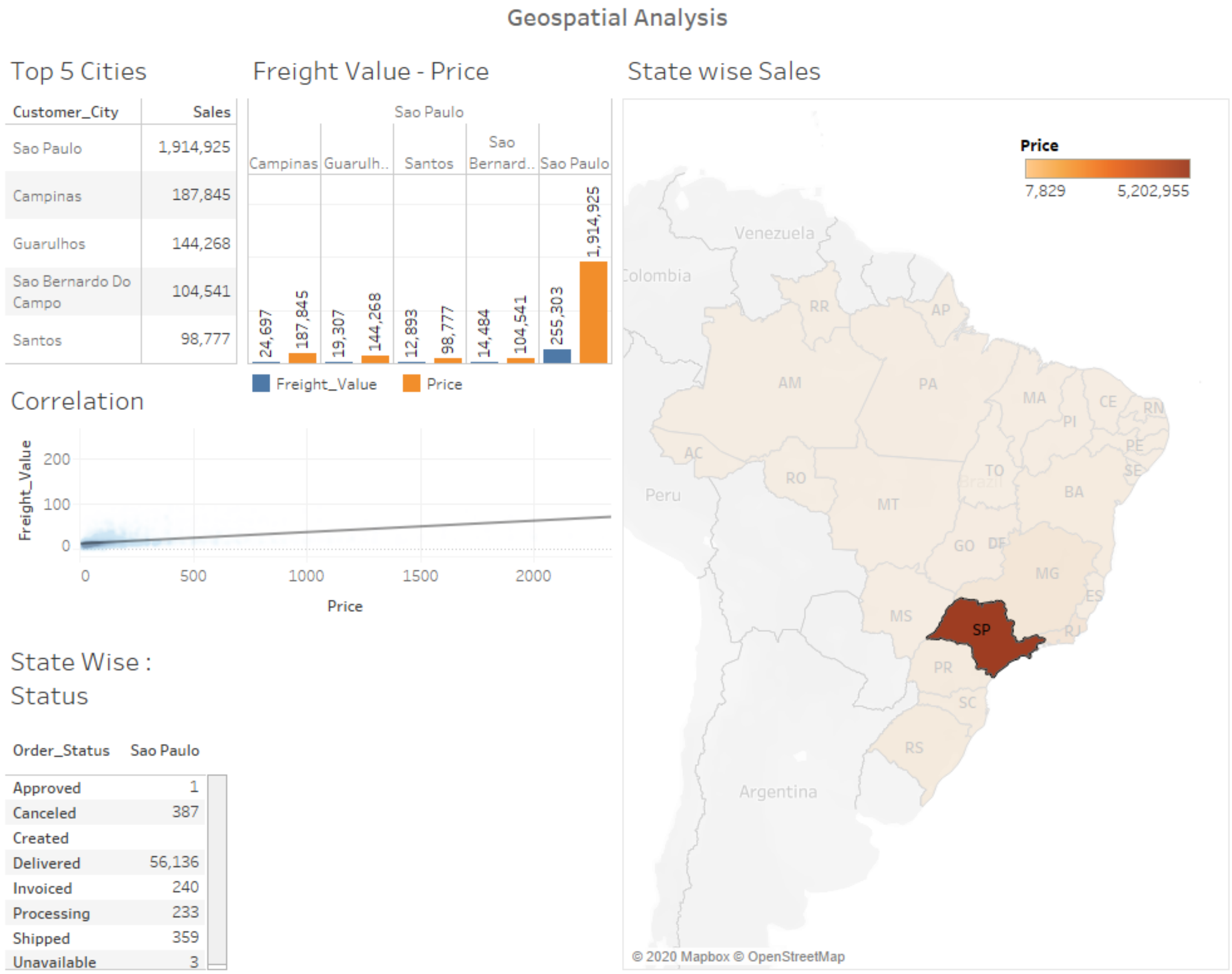
https://help.tableau.com/current/guides/get-started-tutorial/en-us/get-started-tutorial-build.html

https://www.tableau.com/solutions/marketing-analytics

Appendix

1.Geolocation Dashboard[3]

The dashboard gives information about Top5 cities, Correlation between Freight value and price, State wise Sales.



Geospatial Analysis

Top 5 Cities

Customer_City	Sales
Sao Paulo	1,914,925
Campinas	187,845
Guarulhos	144,268
Sao Bernardo Do Campo	104,541
Santos	98,777

State Wise : Status

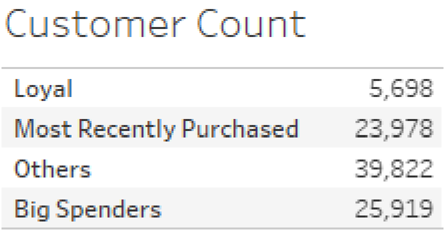
Order_Status	Sao Paulo
Approved	1
Canceled	387
Created	
Delivered	56,136
Invoiced	240
Processing	233
Shipped	359
Unavailable	3

Figure 2 Geospatial Dashboard

[3] Refer File: Geospatial Analysis.twbx

2.Dashboard for Customer Segmentation

The data used for the dashboard is the joining of clusters dataset, order dataset, product dataset with English product categories. The dashboard consists of four sections which can filtered according to year and segmentation. Customer count gives the number of customers in each segment for specific year.



Customer Count

Loyal	5,698
Most Recently Purchased	23,978
Others	39,822
Big Spenders	25,919

Figure 3 Customer Count Worksheet

Product category gives the number of customers for each category. The top five product categories are bed_bath_table, health_beauty, sports_leisure, computers_accessories and furniture_decor.

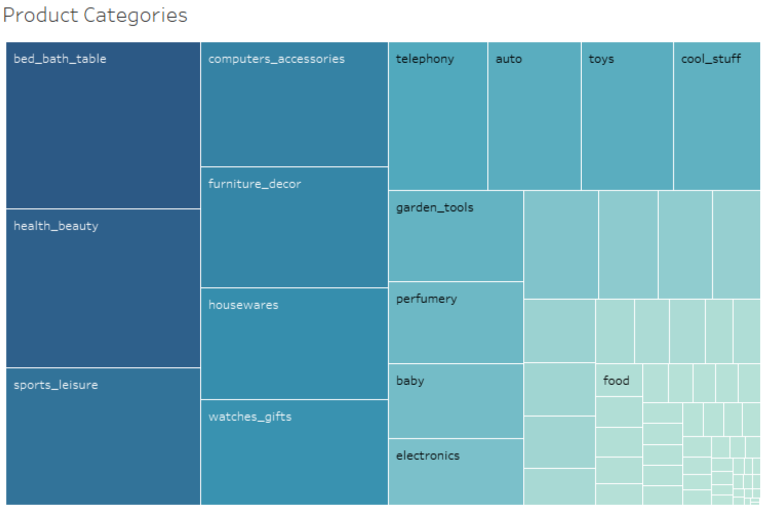


Figure 4 Product Categories Worksheet

Revenue sections consist of total revenue for each month from different customer segmentation.

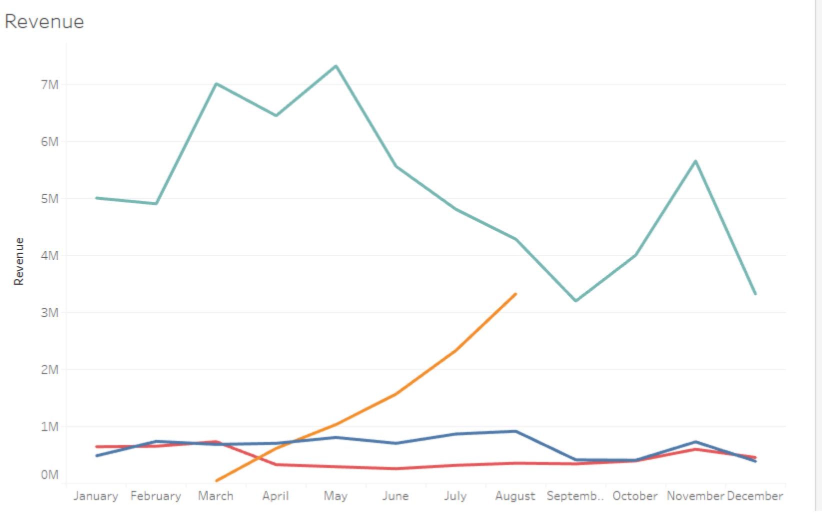


Figure 5 Revenue Worksheet

Payment Type

In this ecommerce business we have four payment type as debit card, boleto (payment method in Brazil which is regulated by Brazilian Federation of Banks), voucher, credit cards according to year. Overall, most of the customers used credit card payment type.

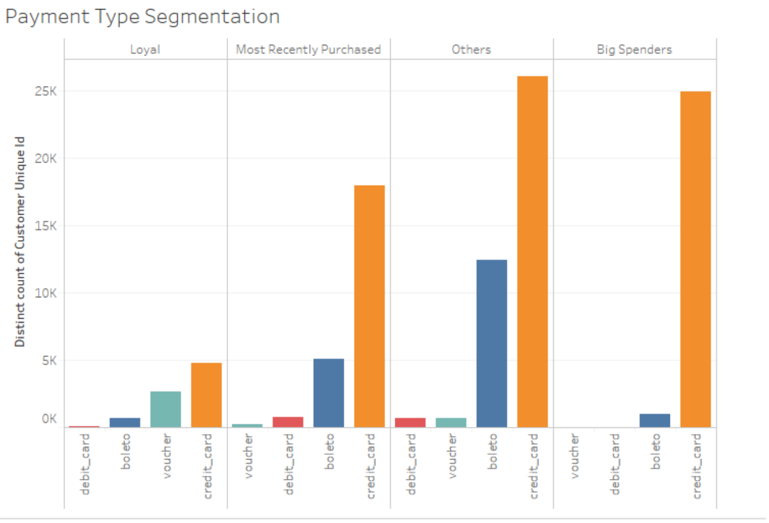
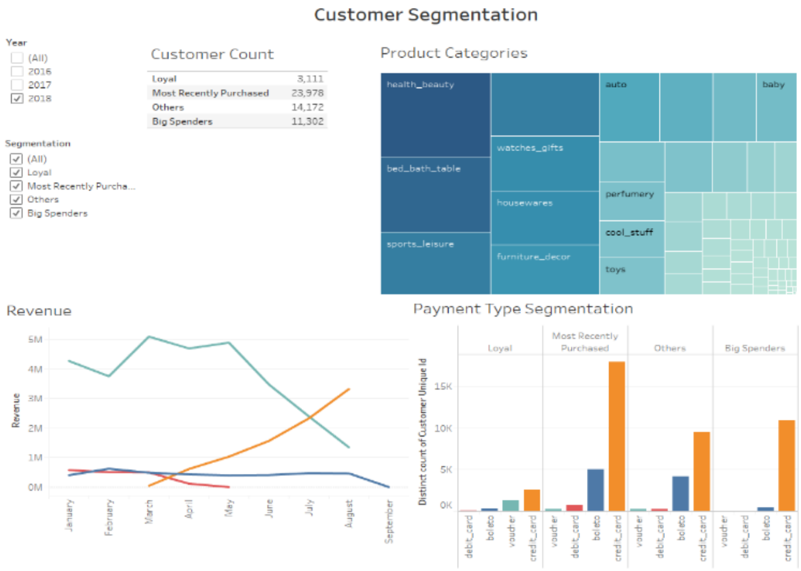


Figure 6 Payment Type Worksheet

The overall dashboard looks as in figure5.



Customer Segmentation

Year

- □ (All)
- □ 2016
- □ 2017
- ☑ 2018

Segmentation

- ☑ (All)
- ☑ Loyal
- ☑ Most Recently Purcha...
- ☑ Others
- ☑ Big Spenders

Customer Count

Loyal	3,111
Most Recently Purchased	23,978
Others	14,172
Big Spenders	11,302

[4]Figure 7 Customer Segmentation Dashboard

[4] Refer File: Clustering.twb