

**Corporación Favorita Grocery Sales Forecasting**

Raksha Kaverappa

Emily Strong

Index

Abstract 4

Exploratory Data Analysis 6

Feature Selection……………………………………………………………………………………………......11

Clustering……………..………...……………………………………………………………………………11

4.1 Unit Sales Clustering………………………………………………………………………………….11

4.2 Transactions Clustering………………………………………………………………………………..11

Unit sales Prediction……………………………………………………………………………………………11

Time Series Forecast 11

5.1 Estimation and Forecasting using ARIMA 12

Azure Experiment-Unit Sales 14

4.4 Predict the Unit Sales 25

4.5 Create a Web Service for the above model 26

Web Application 30

Reference 32

# **Abstract**

The dataset chosen is a Kaggle competition dataset hosted by Corporación Favorita. This is an Ecuadorian grocery chain with over 100 stored carrying over 200,000 products. Currently we are predicting the sales of just grocery, but this can be further extended to other classes of groceries and beauty products which you might find a super market.

The link to our dataset is <link>

The dataset has the following files and properties:

* Train.csv: Consists of train data with unit sales per iter per day.
* Stores.csv: Consists of all the stores, their location and their individual store numbers
* Items.csv: Consists all the items, their family, classes and the item number
* Holidays.csv: Consists of the holidays and events metadata.
* Oils.csv: Consists of Daily oil prices.

We are predicting the Unit Sales for the grocery items by clustering them based on the item classes. We have used Neural Networks to predict the Unit sales of the items.

We are also forecasting the future transactions of each store and studying the effect of oil prices on the transactions since Ecuador is an oil dependent country.

# **Exploratory Data Analysis**

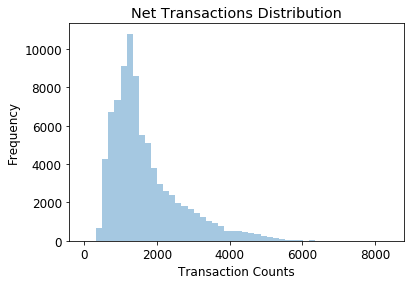
The Stores are distributed across Ecuador and we got the following plot for the same:

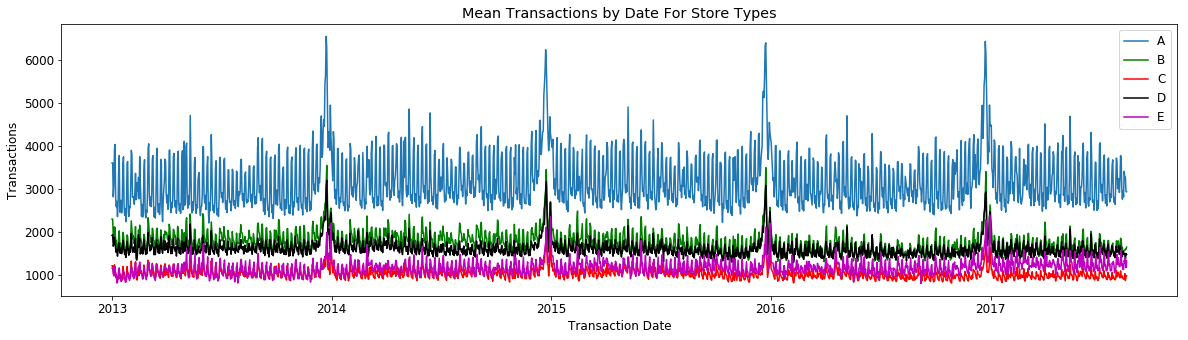


Figure 1: Density of Stores Per City. Legend: Black = 1 store, Blue = 2 stores, Purple = 3 stores,

Red = >5 stores, Orange = >10 stores

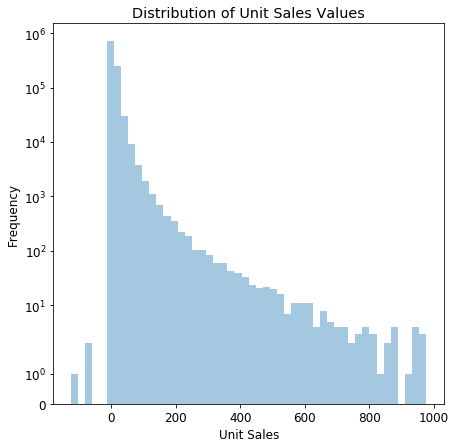
The frequency distribution of transactions and the mean transactions by date for each store type is as shown below:

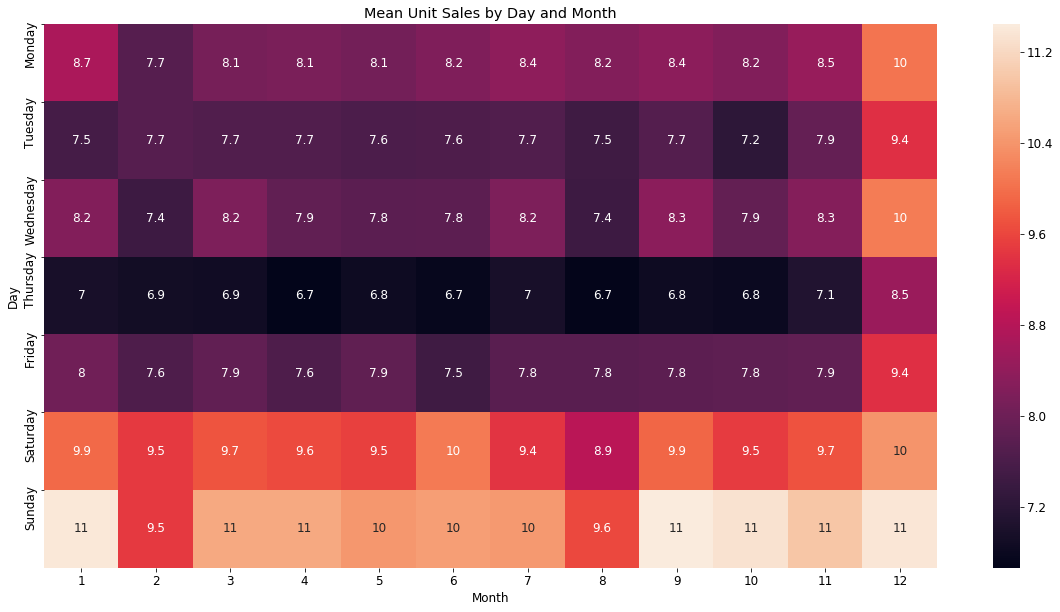




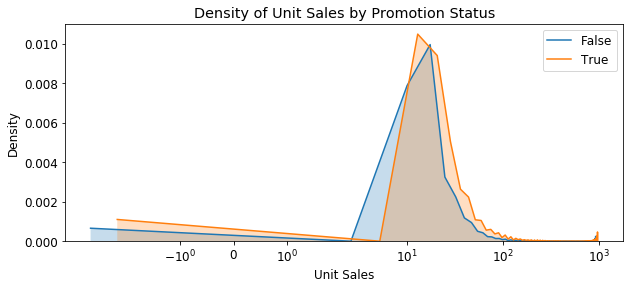
We can see that the transactions based on stores can we divided into 5 clusters. We have used this analysis for accurately clustering out Transactions data for time series forecasting.

The distribution of unit sales and the mean unit sales by day and month is as shown below:

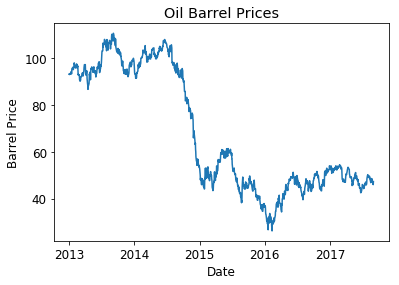




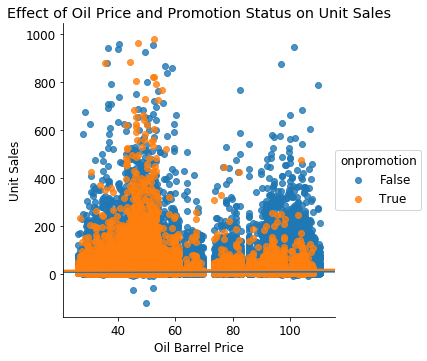
We can clearly see that the unit sales for every store was maximum on the weekends.



We also observed that there wasn’t much variation in unit sales when item were on promotion.



We also plotted the trend of oil prices by date. Further, we inspected the variation of unit sales based on oil prices and grouped them on the basis of promotion.



# **Clustering**

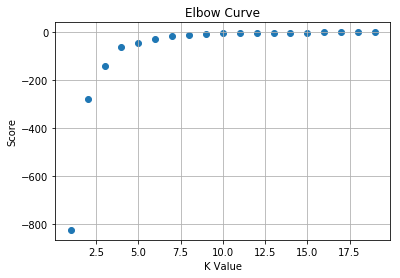
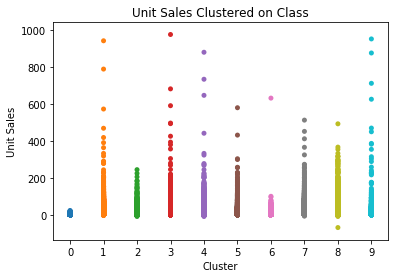
**Unit Sales:**

Based on our EDA, we selected the following features for our unit sales prediction:

* onpromotion
* store transaction counts
* oil barrel price
* item class
* state
* day of week
* month
* local and regional holidays (flags)
* national holidays and events based on feature reduction below

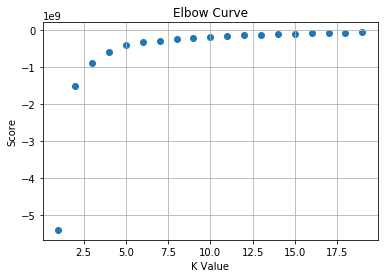
For the unit sales dataset, we were able to select features for just 20% of the total dataset due to the large size of the dataset.

We manually clustered the dataset by finding the optimal value for the number of clusters using the elbow curve shown below. We have chosen 10 clusters for our dataset. Since we are dealing with a large amount of categorical data, we decided that Kmeans is not a good approach to use. Hence, we manually clustered the dataset based on item class.

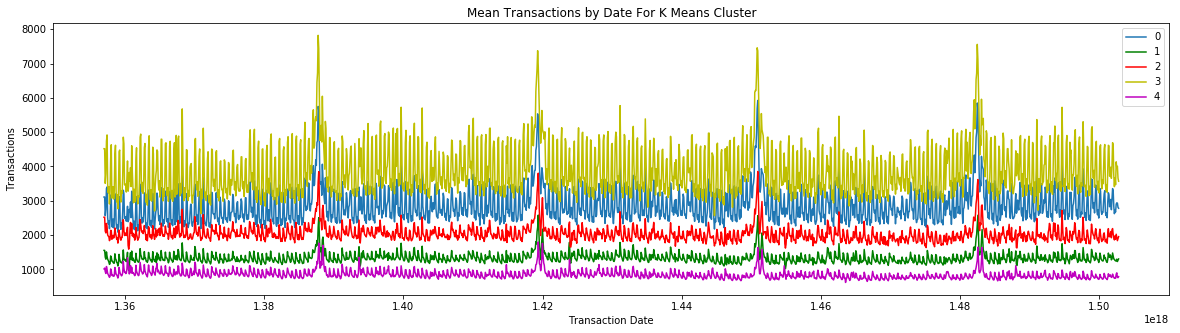
 

**Transactions:**

The transactions dataset was transformed to a time series dataset where each column was shows the transactions by date for a particular store number. We Clustered the transactions dataset with a K value of 5(as a result of the elbow curve shown below).



The clustered data is as shown below



# **Unit Sales Prediction**

We used linear regression to predict the unit sales per store and per item. We selected the features using stepwise linear regression in R after which we trained the model using linear regression. The MAE and MAPE for each of the clusters is as shown below:

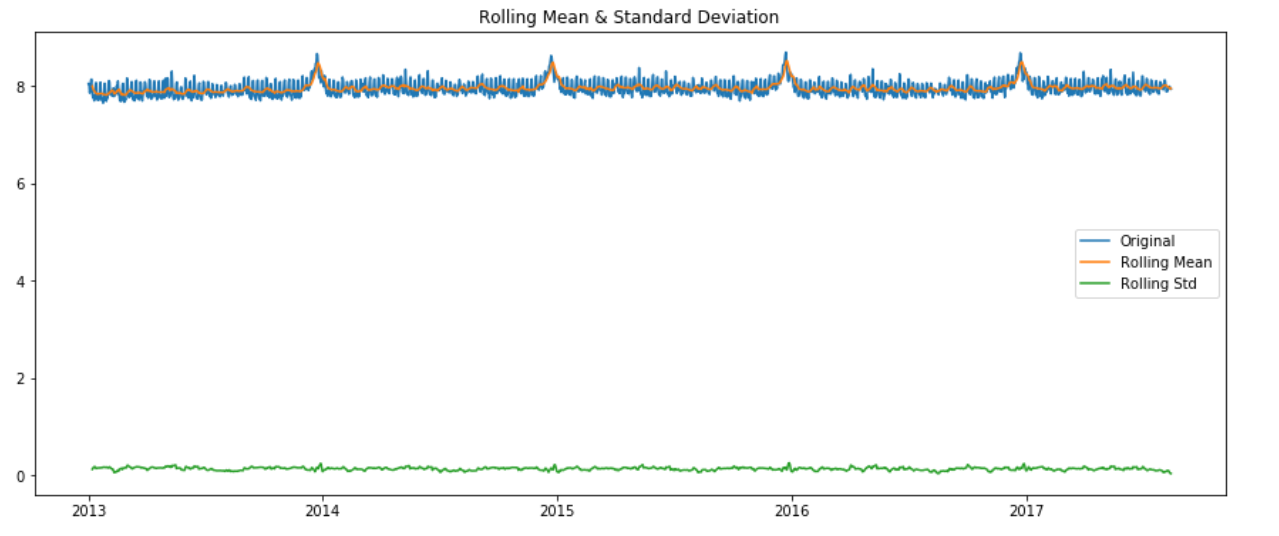
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **MAE-Train** | **MAE-Test** | **MAPE-Train** | **MAPE-Test** |
| Cluster 0 | 1.09045943211 | 1.09007295475 | 60.472444309 | 60.5225811096 |
| Cluster 1 | 2.1046738779 | 2.1046738779 | 87.8291263265 | 87.8361760686 |
| Cluster 2 | 4.94627656152 | 5.00750308326 | 115.838648041 | 117.783078074 |
| Cluster 3 | 3.69877857453 | 3.69564205273 | 118.355929414 | 118.107319756 |
| Cluster 4 | 1.60452993812 | 1.60753634475 | 75.6795277545 | 75.8338208977 |
| Cluster 5 | 0.522646305249 | 0.534347661827 | 36.9559092456 | 37.3638001872 |
| Cluster 6 | 2.6238135117 | 2.61366682971 | 99.1629727511 | 98.8520829999 |
| Cluster 7 | 6.73416341233 | 6.73520174937 | 153.053919285 | 154.79421774 |
| Cluster 8 | 3.28522056073 | 3.28369014909 | 109.630326372 | 109.850182934 |
| Cluster 9 | 4.81458060488 | 4.82504755436 | 132.063044691 | 132.02729994 |

# **Time series forecast**

***Transactions***

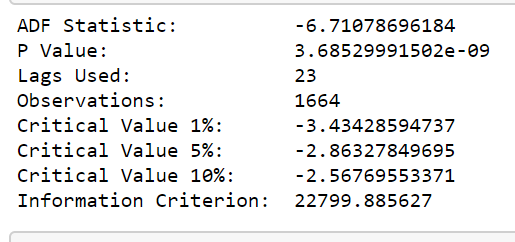
After clustering the transactions data, we forecasted the transactions of each store in that cluster. We used ARIMA models for forecasting the future transactions.

We removed the seasonality and trend and converted it into a stationary model. We test if the model is stationary using rolling mean, rolling variance and Dicker Fuller method. The results obtained are as shown below:

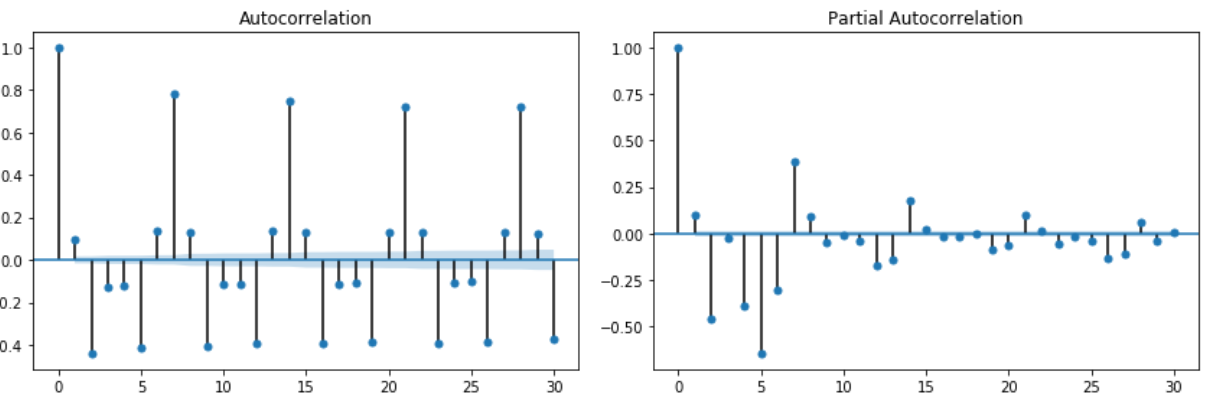


As we can see, the model has been made stationary by decomposing it and removing trend and seasonality.

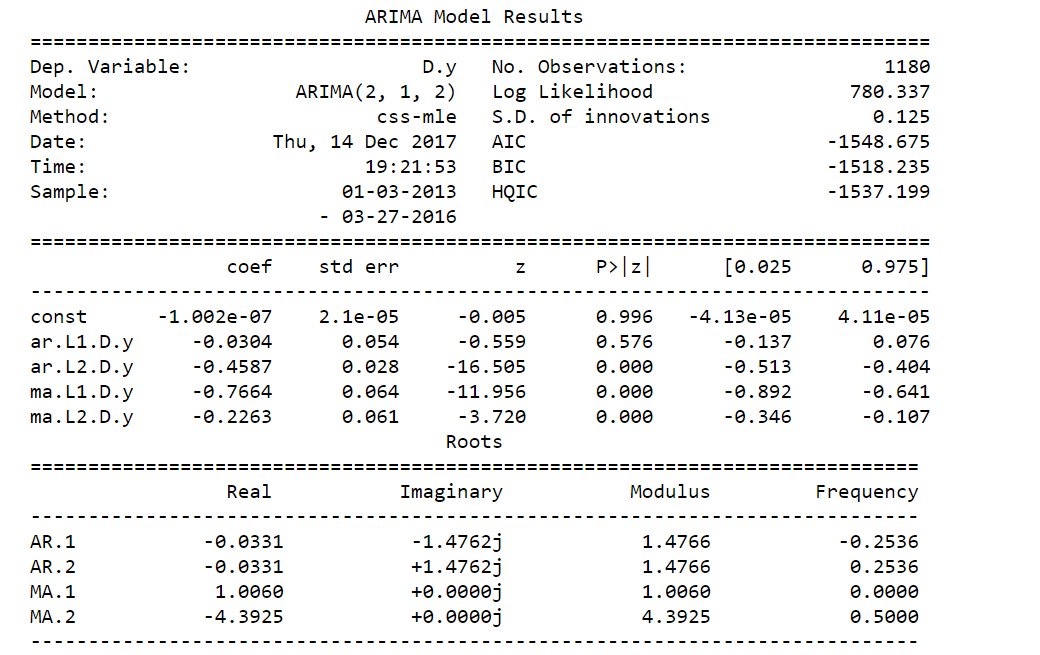
We also used Dicker Fuller method to check if the model is stationary. We can see that the ADF statistic is less than 5% of the critical value.



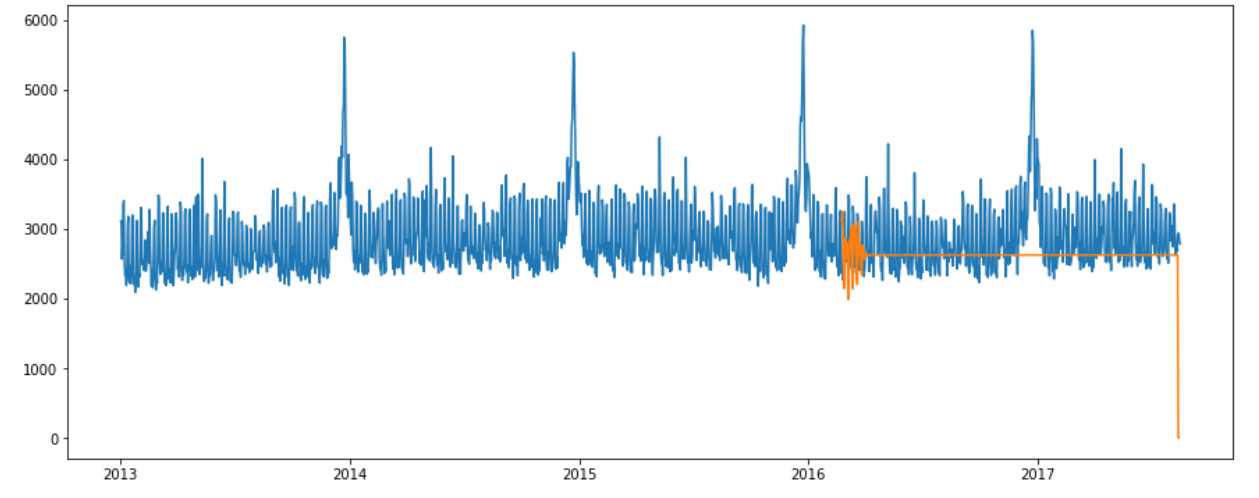
We also used ACF and PACF to find the optimal parameters for the ARIMA model



The results of the ARIMA model is as shown below.



The prediction for the future time is as shown



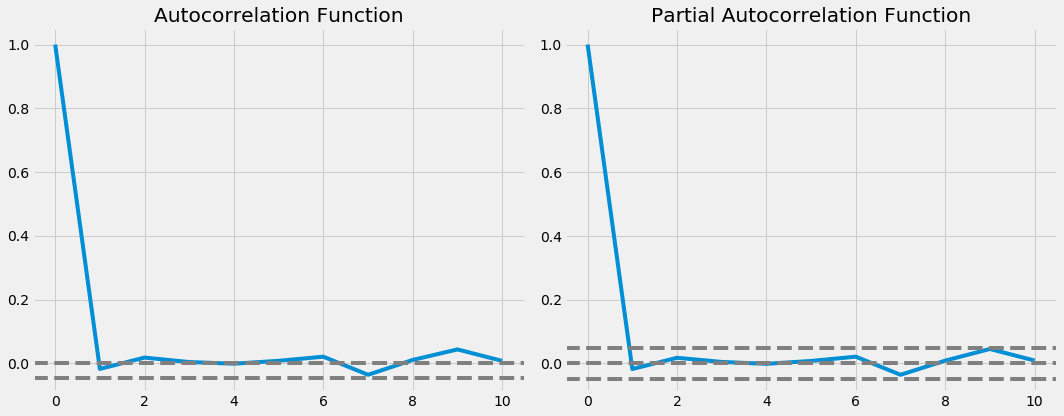
The MAE, RMSE and BIAS of each cluster is as shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Arima Model** | **Cluster 0** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **cluster 4** |
| Mean Forcast error | 2917.716282 | 1331.051354 | 1998.506271 | 3925.654602 | 790.403602 |
| Mean Absolute error | 2919.97953 | 1331.701283 | 1997.410354 | 3927.966568 | 790.815995 |
| RMSE | 1157.3588 | 493.0613 | 1570.2317 | 1433.6683 | 186.9574 |

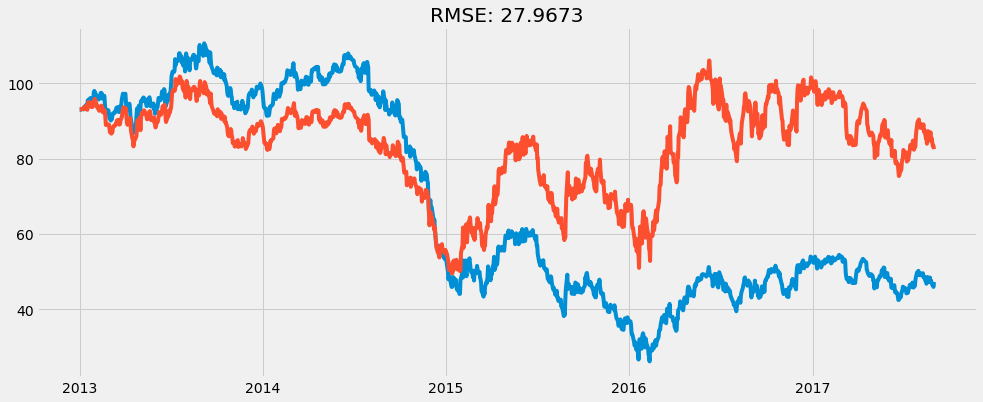
***Oil Sales:***

Oil sales dataset would possibly have an impact on transactions and Unit sales, hence we forecasted the future oil prices as well. We removed Seasonality and trend and predicted the future oil prices.

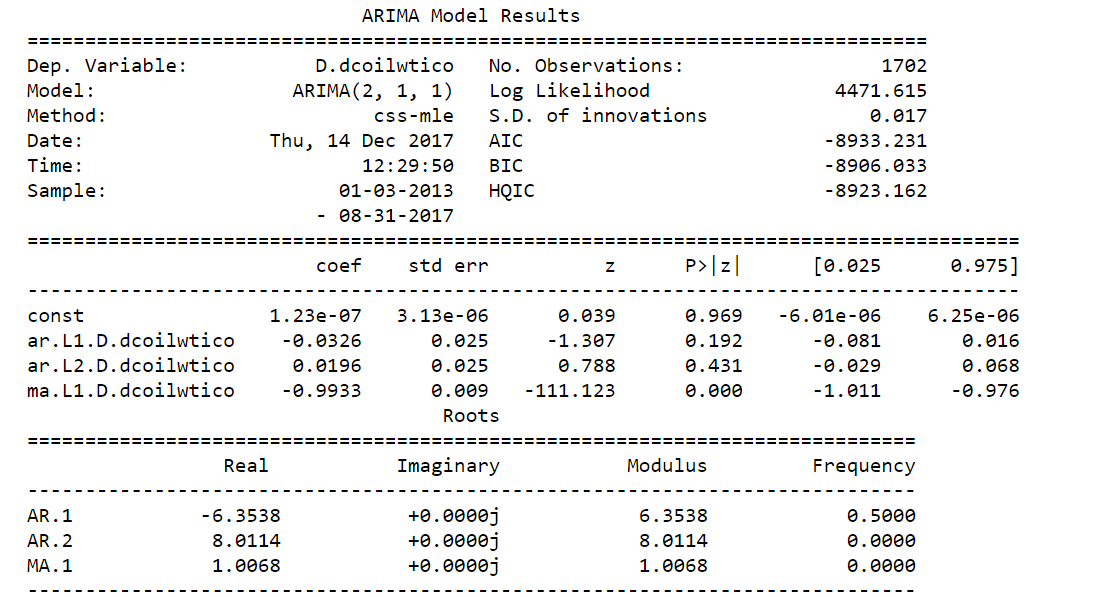
The ACF and PACF are as shown below. We determined the optimal parameters using this plot.



The oil Sales prediction is as shown below:



The results of the ARIMA model for oils dataset is as shown below:



# **Azure-ML studio**

The linear regression for unit sales was deployed on Azure and the web service was deployed. The model is as shown below

