Corporación Favorita Grocery Sales Forecasting

Raksha Kaverappa
Emily Strong

Final Draft: 12/15/2017

Presentation: https://www.slideshare.net/EmilyStrong/team-4-presentation-grocery-sales-forecasting
Github: https://github.com/erstrong/FavoritaGrocery/

Index

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 I items, 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 https://www.kaggle.com/c/favorita-grocery-sales-forecasting

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_Events.csv: Consists of the holidays and events metadata.
- Oils.csv: Consists of Daily oil prices.

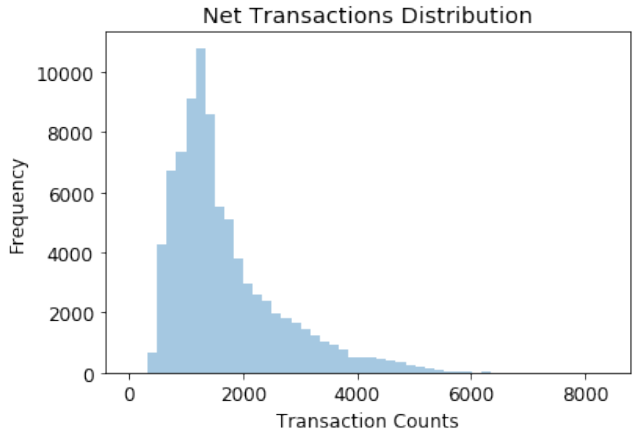
We are predicting the Unit Sales for the grocery items by clustering them based on the item classes. We have used linear regression to predict the unit sales of the items.
We are also forecasting the future transactions of each store and studying the effect of oil prices on the transactions since Ecuador is an oil dependent country.

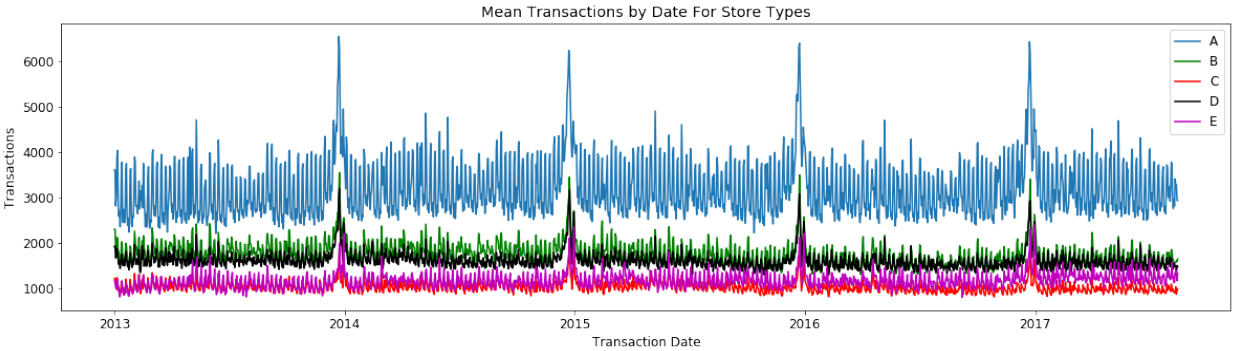
Exploratory Data Analysis

The Stores are distributed across Ecuador, with the largest concentration in Quito where the company was founded:

Figure 1: Density of Stores Per City. Legend: Black = 1 store, Blue = 2 stores, Purple = 3 stores, Red = >5 stores, Orange = >10 stores

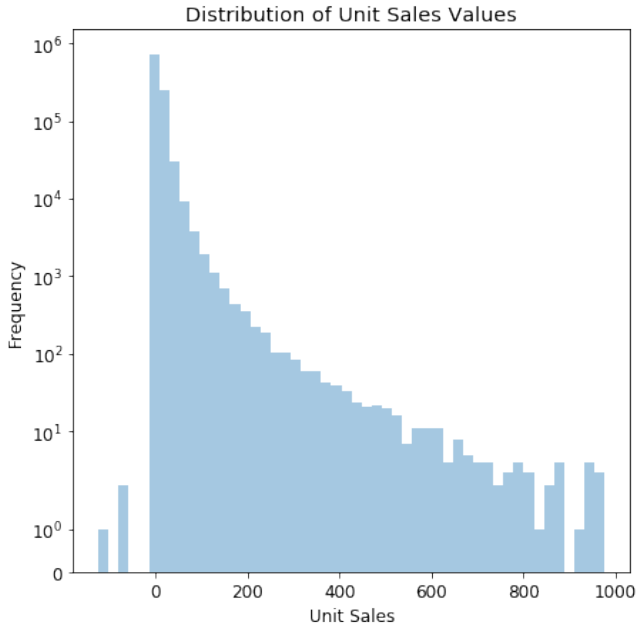
The frequency distribution of transactions and the mean transactions by date for each store type is as shown below:

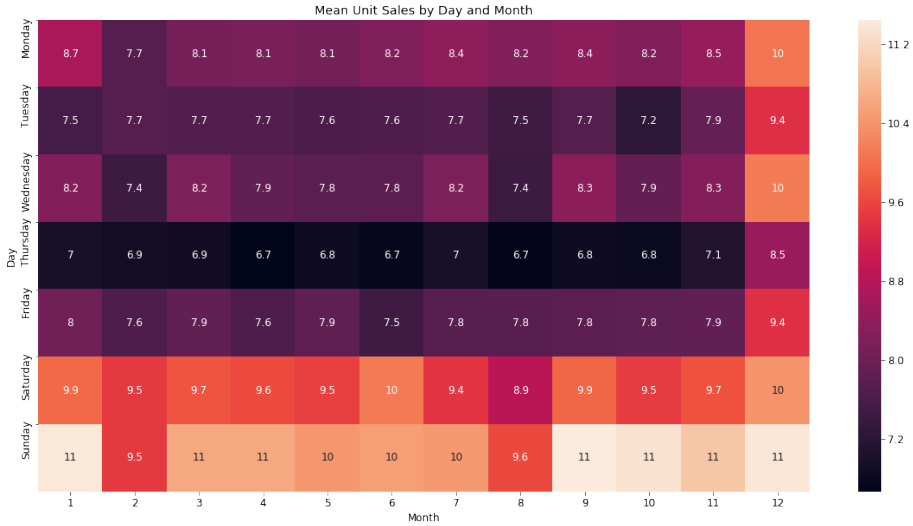




We can see that the transactions based on stores can be divided into 5 clusters. We have used this analysis for accurately clustering out Transactions data for time series forecasting.

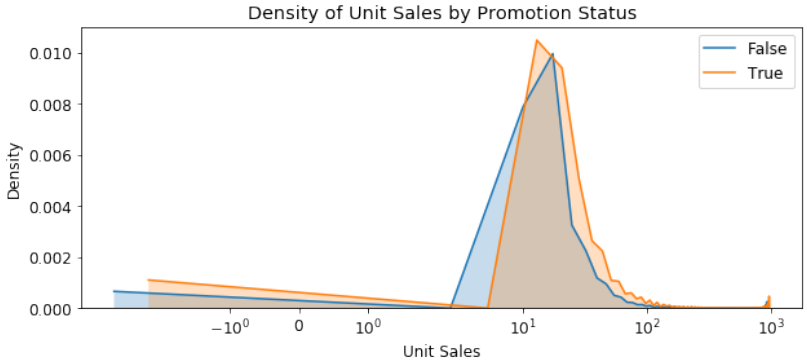
The distribution of unit sales and the mean unit sales by day and month is as shown below:



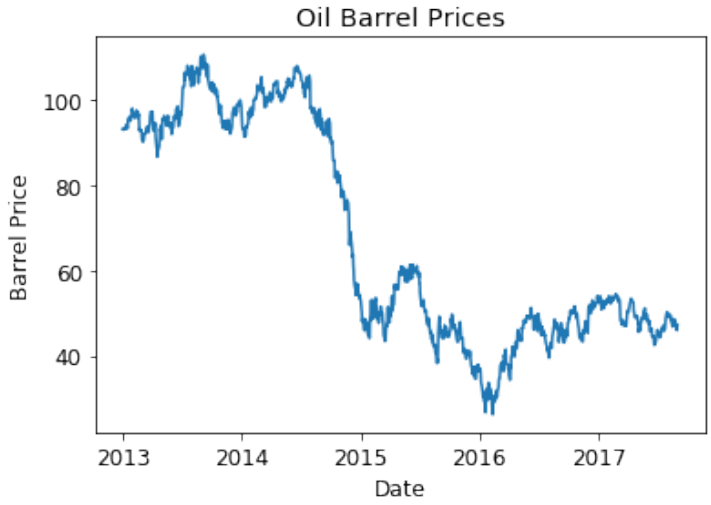


We can clearly see that unit sales are time-dependent, with higher sale volumes on weekends and during December.

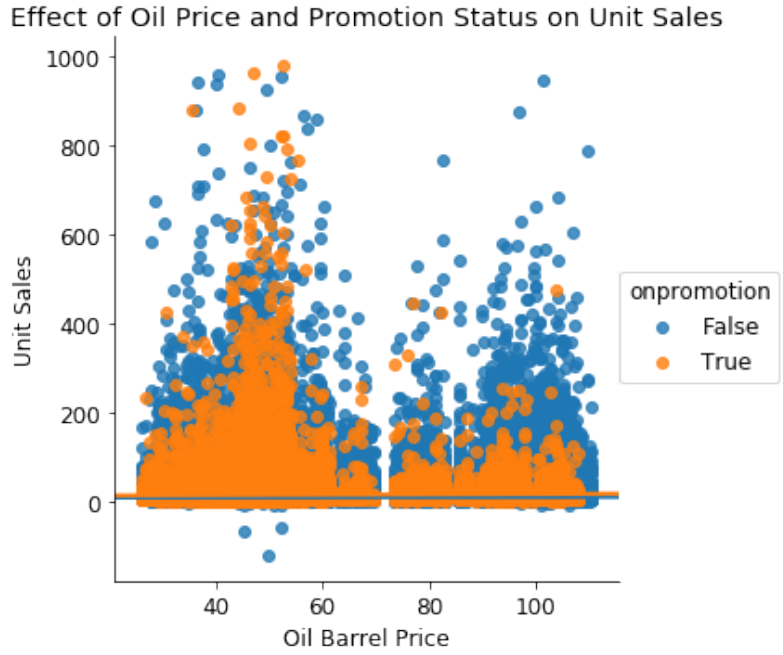
We also observed that being on promotion does correspond to a small increase in unit sales:



Since the Ecuadorian economy is dependent on oil, oil barrel price is likely a good indicator of the strength of the economy. There was a permanent break in the oil price in 2014:



We also examined the relationship between oil price and promotion status. When oil prices are lower, promotion status does appear to be more important to unit sales:



Clustering

Unit Sales:

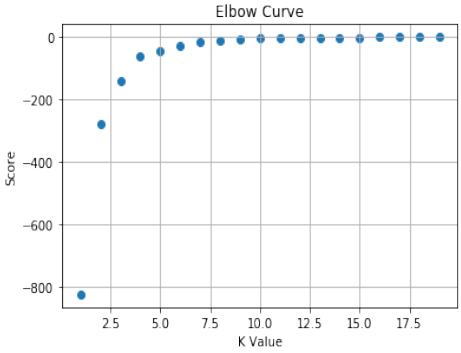
Based on our EDA, we selected the following features for our unit sales prediction:

- on promotion
- store transaction counts
- oil barrel price
- item class
- state in which the store is located

- day of week
- month
- local and regional holidays (flags)
- national holidays and events binned based on whether they correspond to an increase or decrease in unit sales

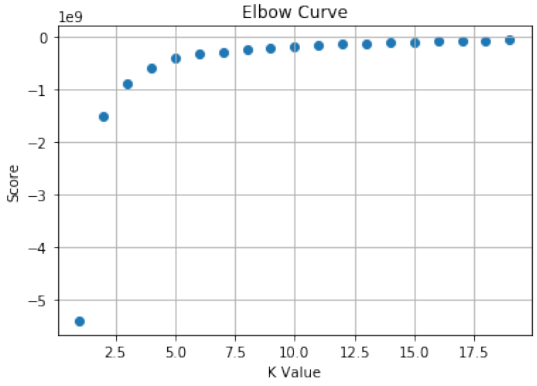
For the unit sales dataset, we were able to select features for just 20% of the total dataset due to the large size of the dataset.

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.

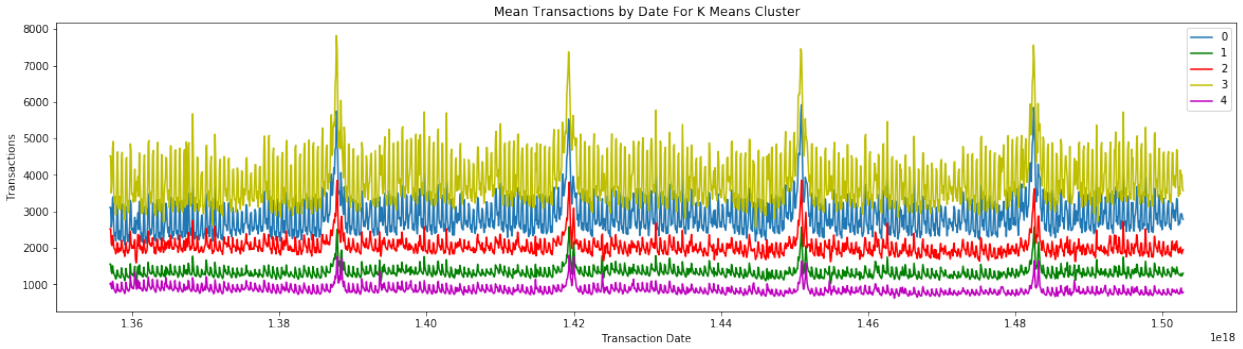


Transactions:

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, KNN, random forest and neural networks to predict the unit sales per store and per item, ultimately choosing linear regression as the best predictor for the data. We selected the features for each cluster using stepwise linear regression in R after which we trained the model again 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

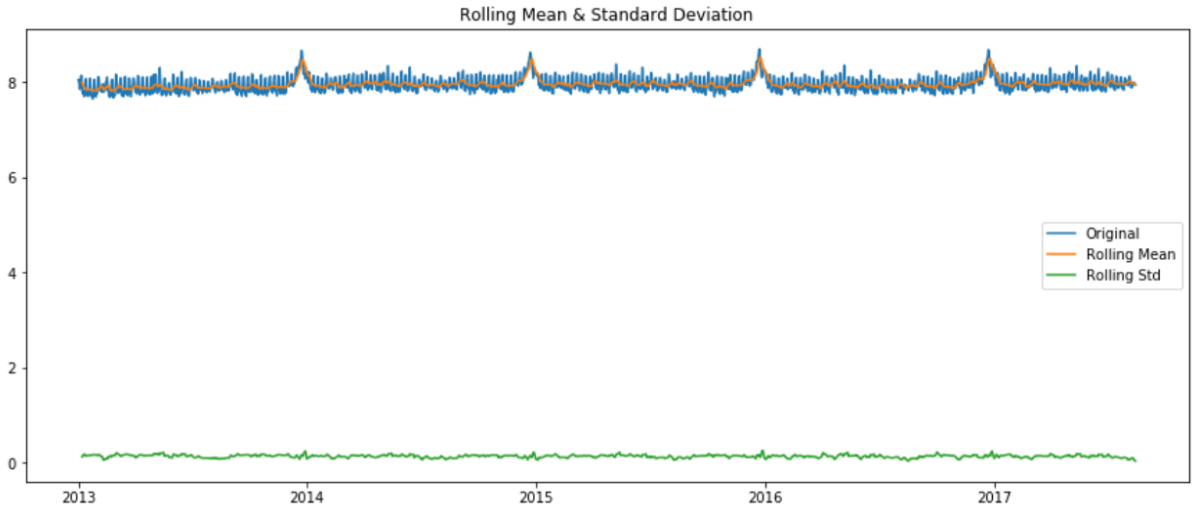
The MAE for each API on Azure is equivalent to the MAE-Test values we obtained with Python.

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:

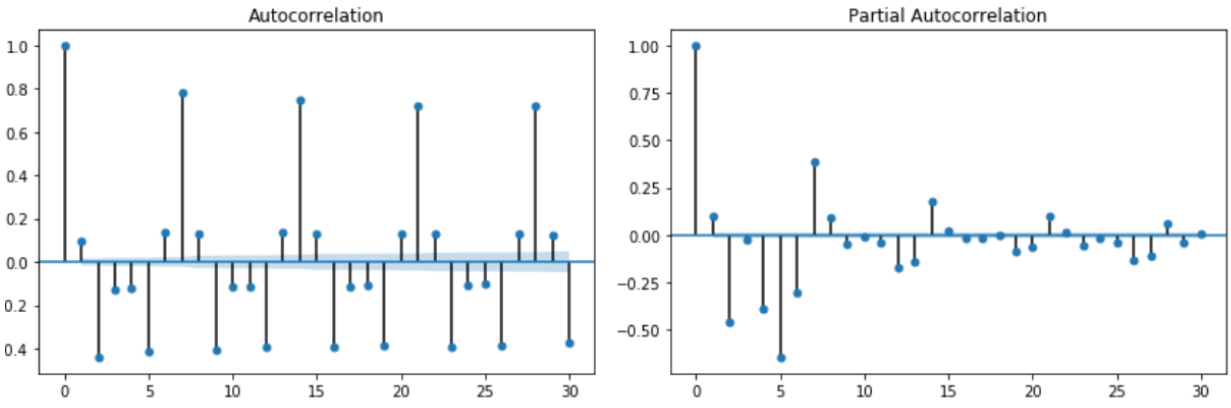


As we can see, the model has been made stationary by decomposing it and removing trend and seasonality.

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.

```
ADF Statistic:          -6.71078696184
P Value:                3.68529991502e-09
Lags Used:              23
Observations:           1664
Critical Value 1%:      -3.43428594737
Critical Value 5%:      -2.86327849695
Critical Value 10%:     -2.56769553371
Information Criterion:  22799.885627
```

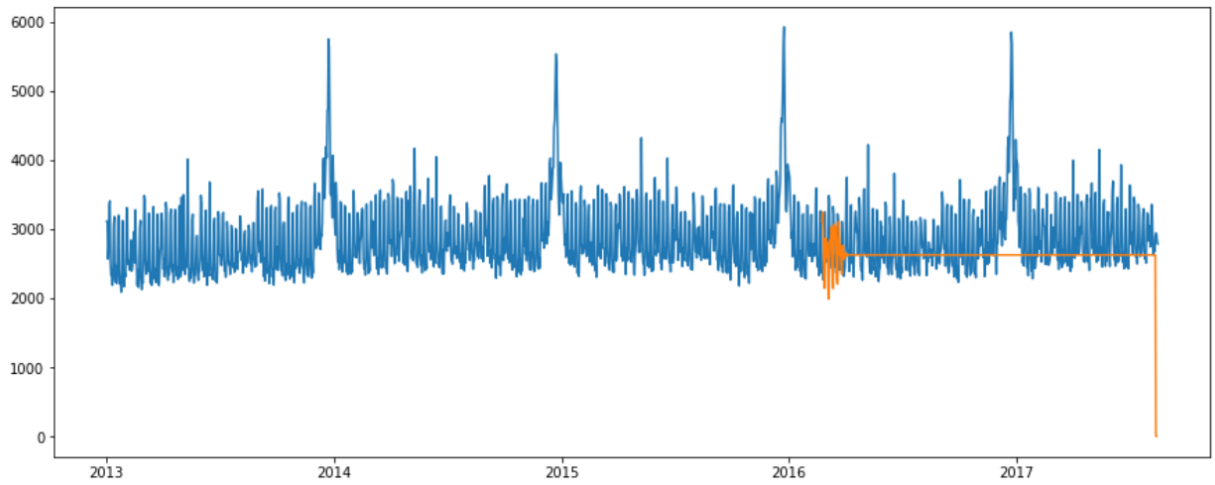
We also used ACF and PACF to find the optimal parameters for the ARIMA model



The results of the ARIMA model is as shown below.

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                    D.y   No. Observations:                 1180
Model:                 ARIMA(2, 1, 2)   Log Likelihood                 780.337
Method:                       css-mle   S.D. of innovations              0.125
Date:                Thu, 14 Dec 2017   AIC                          -1548.675
Time:                        19:21:53   BIC                          -1518.235
Sample:                    01-03-2013   HQIC                         -1537.199
                         - 03-27-2016
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.002e-07    2.1e-05     -0.005      0.996   -4.13e-05    4.11e-05
ar.L1.D.y     -0.0304      0.054     -0.559      0.576      -0.137       0.076
ar.L2.D.y     -0.4587      0.028    -16.505      0.000      -0.513      -0.404
ma.L1.D.y     -0.7664      0.064    -11.956      0.000      -0.892      -0.641
ma.L2.D.y     -0.2263      0.061     -3.720      0.000      -0.346      -0.107
                                    Roots
=============================================================================
                 Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -0.0331           -1.4762j            1.4766           -0.2536
AR.2           -0.0331           +1.4762j            1.4766            0.2536
MA.1            1.0060           +0.0000j            1.0060            0.0000
MA.2           -4.3925           +0.0000j            4.3925            0.5000
-----------------------------------------------------------------------------
```

The prediction for the future time is as shown

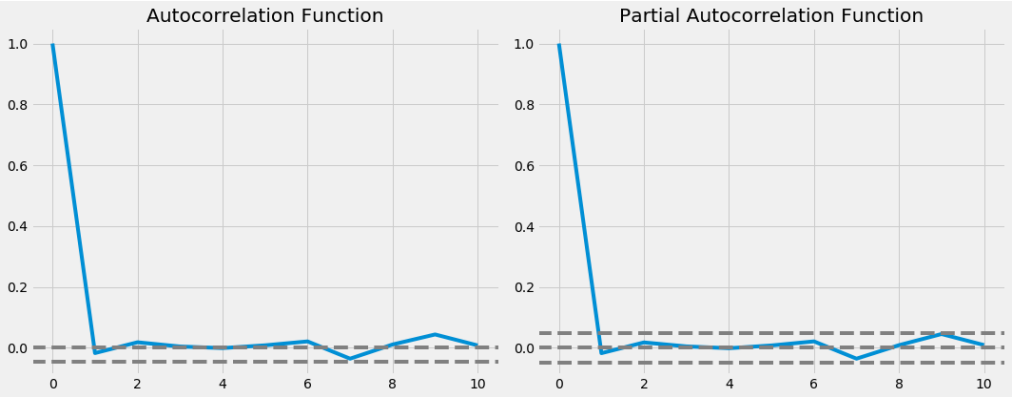


The MAE, RMSE and BIAS of each cluster is as shown below:

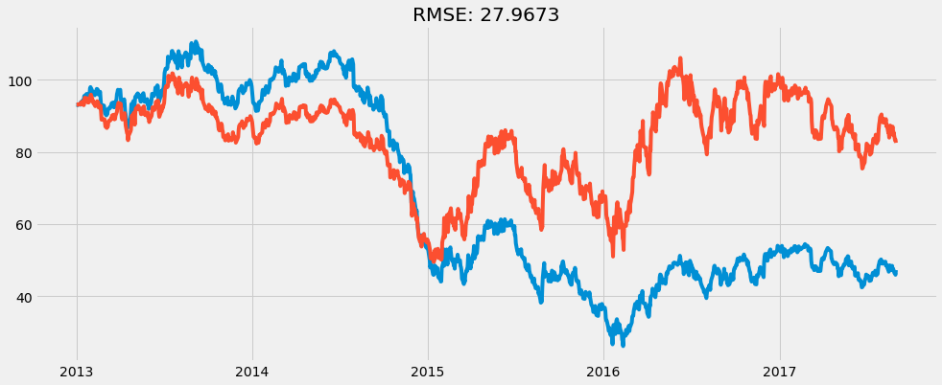
Arima Model	Cluster 0	Cluster 1	Cluster 2	Cluster 3	cluster 4
Mean Forcast error	2917.716282	1331.051354	1998.506271	3925.654602	790.403602
Mean Absolute error	2919.97953	1331.701283	1997.410354	3927.966568	790.815995
RMSE	1157.3588	493.0613	1570.2317	1433.6683	186.9574

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:

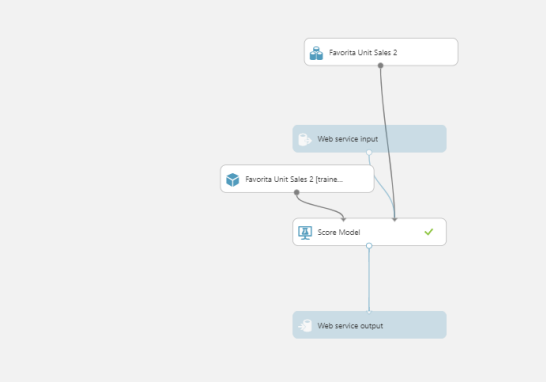


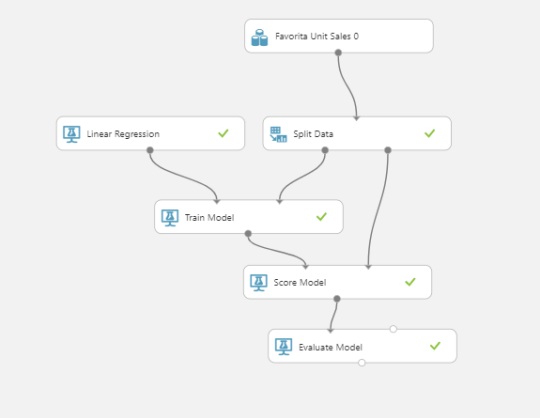
The results of the ARIMA model for oils dataset is as shown below:

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:           D.dcoilwtico   No. Observations:                 1702
Model:                 ARIMA(2, 1, 1)   Log Likelihood                4471.615
Method:                       css-mle   S.D. of innovations              0.017
Date:                Thu, 14 Dec 2017   AIC                          -8933.231
Time:                        12:29:50   BIC                          -8906.033
Sample:                    01-03-2013   HQIC                         -8923.162
                         - 08-31-2017
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                  1.23e-07   3.13e-06      0.039      0.969   -6.01e-06    6.25e-06
ar.L1.D.dcoilwtico     -0.0326      0.025     -1.307      0.192      -0.081       0.016
ar.L2.D.dcoilwtico      0.0196      0.025      0.788      0.431      -0.029       0.068
ma.L1.D.dcoilwtico     -0.9933      0.009   -111.123      0.000      -1.011      -0.976
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -6.3538           +0.0000j            6.3538            0.5000
AR.2            8.0114           +0.0000j            8.0114            0.0000
MA.1            1.0068           +0.0000j            1.0068            0.0000
-----------------------------------------------------------------------------
```

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





We were unsuccessful in deploying our time series models to Azure ML Studio due to its limited support for time series. We thus chose to use these models as pickle files in our web application.

Web Application

Live application: http://groceryforecasting.herokuapp.com/

We attempted to build our application using Flask however we encountered several problems:

- Our time series models would not accept dates outside the ranges they were trained on.
- API calls that were successfully able to parse predicted values in Jupyter Notebook threw an error about converting Series data to JSON in Flask. We were unable to resolve this issue.

Conclusion

For most of our time series predictions and unit sales predictions we had high error rates. It is possible that using external data such as weather forecasts would have provided better predictors. As for the predictors we did use, when we performed feature selection on the separate clusters, each one selected transaction count, oil price, at least one day, and at least one month. Of the remaining features, each was selected by at least one cluster.