**Time Series Project**

*Augusto Perez, James Ghosn, German Baltazar*

This week’s project consisted of a series of Wal-Mart sales information to predict future profits. The dataset had 421,570 observations divided by the number of stores and departments and the weekly date of each observation. An additional dataset with extra features was given, as shown in Table I. The first step was to get the total sum of the weekly sales of each store, combining the individual sales of each of their departments, resulting in a total of 6,435 observations.

A quick EDA shows that the sales of every single store may follow a stationary trend, as shown in Figure 1. Following this thought, it may be assumed that the sales predictions can be directly obtained from a time-series model predictor.

Table I. Dataset features

| Date | Store | Dept | Weekly\_Sales | Temperature |
| --- | --- | --- | --- | --- |
| Fuel\_Price | MarkDown1 | MarkDown2 | MarkDown3 | MarkDown4 |
| MarkDown5 | CPI | Unemployment | IsHoliday |  |

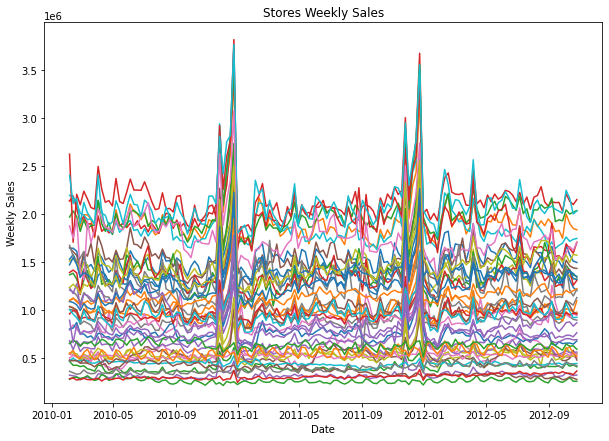


Figure 1. Stores' weekly sales behavior

The second step was dealing with the NaN values of the MarkDowns. Usually, they would’ve been changed for a zero-value, but since the problem revolves around sales values, it is best to substitute with the median value. However, it is noticeable in Figure 1 that working individually with every single store will end with too many models, making the solution non-practical. A different approach to reducing the number of predicting models is by implementing an unsupervised method, like K-Means.

The optimal number of clusters k was determined by implementing the Elbow Method, evaluating the Within-Cluster Sum of Square (WCSS) value, and selecting a total number of 5 groups for this particular problem. Figure 2 shows the behavior of this selection model, while Figure 3 shows an example of how the dataset was grouped into its clusters compared in a 2D space. Finally, Figure 4 shows how clustering can group the data based on the number of sales, making a group for high, regular, medium, and low sales.

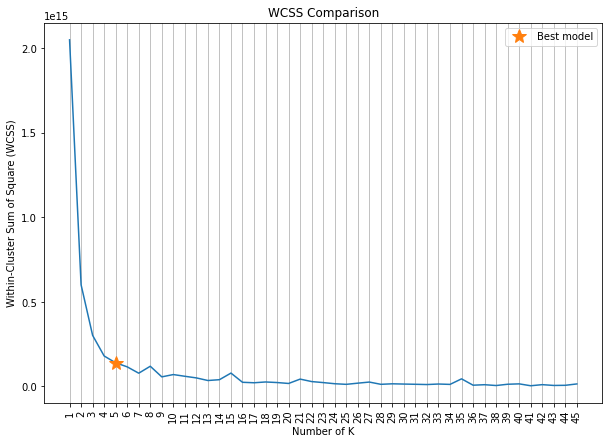


Figure 2. WCSS analysis for K-Means

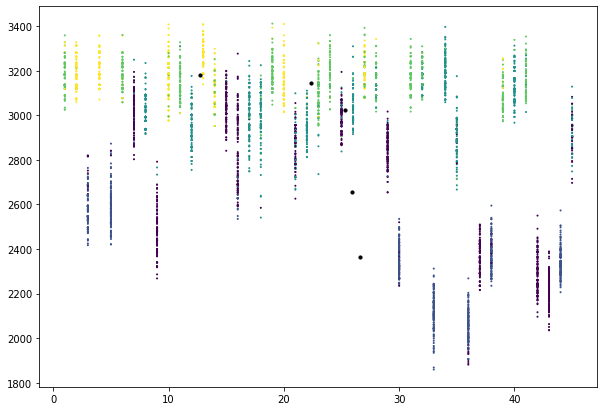


Figure 3. K-Means clustering

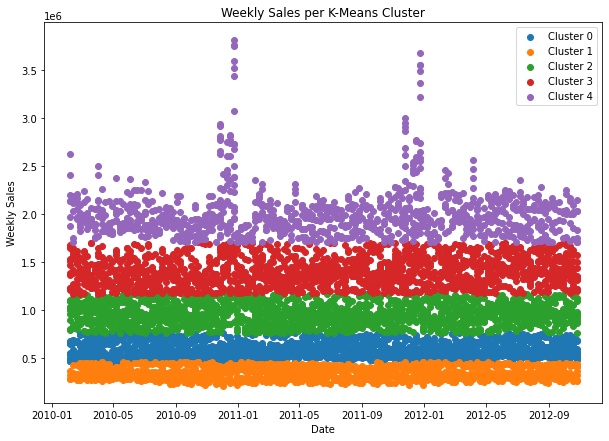
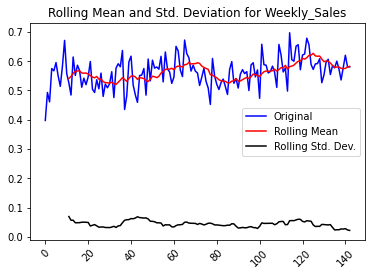
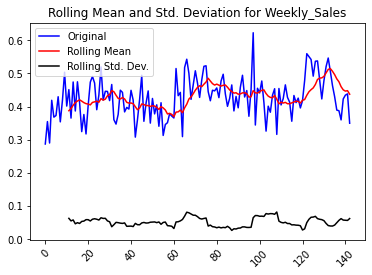


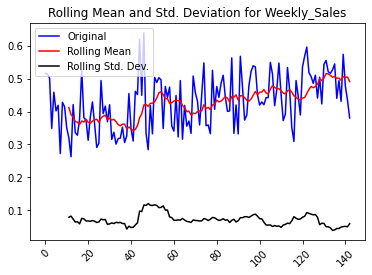
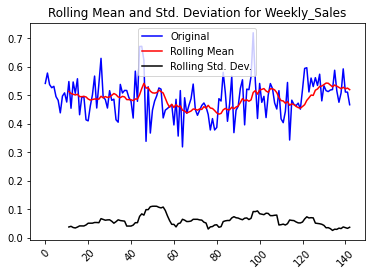
Figure 4. Weekly sales per cluster

The data were normalized with the min/max method, based on each group to reduce variation. Additionally, since the clusters were no longer evaluated based on the number of stores, the Store and Dept features were dropped; the IsHoliday feature was also dropped since the predictor model should work properly with just the datetime of Date. Finally, each cluster had its dataset based on the mean value of each feature, based on every single day available on the original dataset. This process generated five different datasets of 143 observations and 11 attributes each.

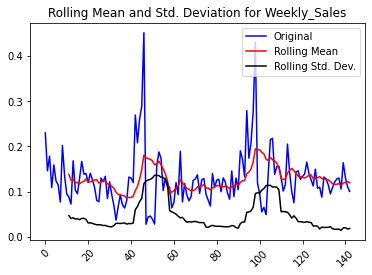
The predictions for each cluster's sales were made using ARIMA and Facebook’s Prophet model. Every dataset was evaluated for stationarity before implementing any of the mentioned methods. Figure 5 shows the original behavior of each cluster, noticing in this case that groups 1 and 3 were the only ones with a p-value above 0.05. For these two cases, it was necessary to subtract the difference to make the models stational. After that, the autocorrelation of every single cluster was obtained to get the optimal p and q values for the ARIMA model. Figure 6 shows the autocorrelations for getting the p values, while Figure 7 shows the ones for the q parameter. Finally, the best model for each cluster was assigned with the Auto ARIMA method. Table II shows the general architecture of the ARIMA model of every cluster.



a) b)

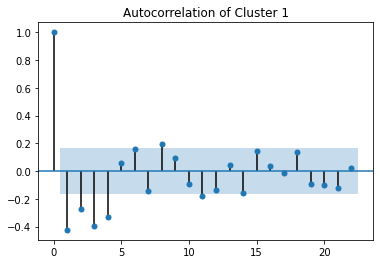
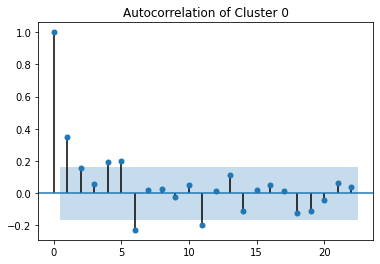


c) d)

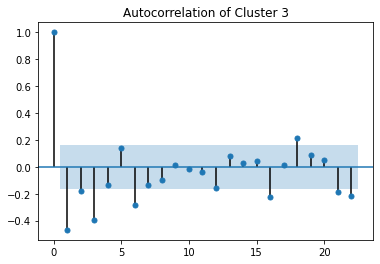
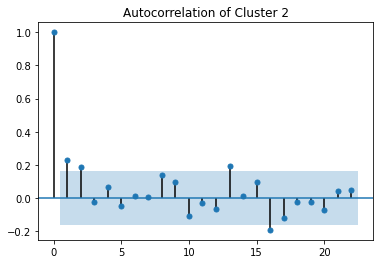


e)

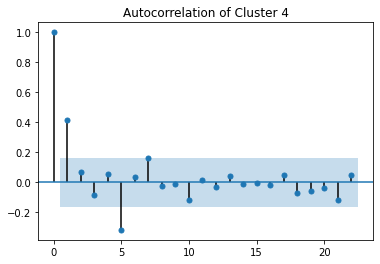
Figure 5. Stationarity test for a) cluster 0, b) cluster 1, c) cluster 2, d) cluster 3, and e) cluster 4



a) b)

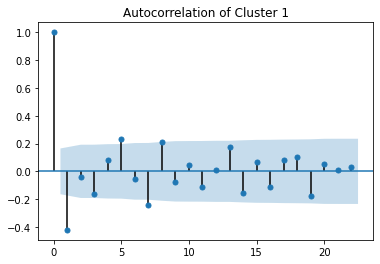
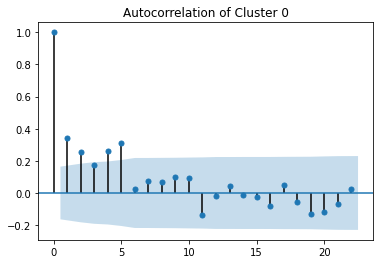


c) d)

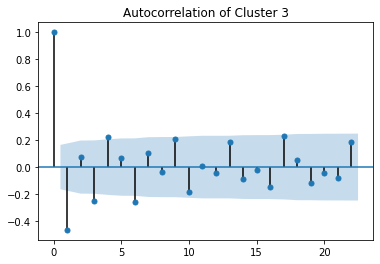
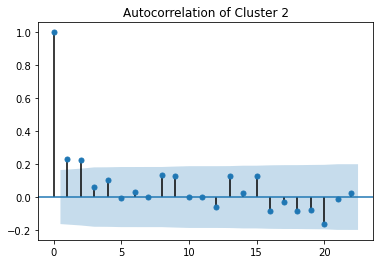


e)

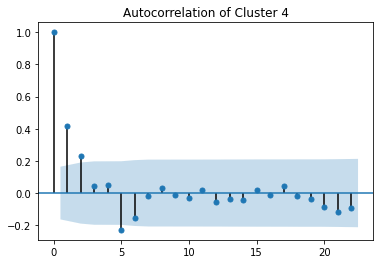
Figure 6. Autocorrelation for p parameter in a) cluster 0, b) cluster 1, c) cluster 2, d) cluster 3, and e) cluster 4



a) b)



c) d)



e)

Figure 7. Autocorrelation for q parameter in a) cluster 0, b) cluster 1, c) cluster 2, d) cluster 3, and e) cluster 4

Table II. ARIMA parameters for cluster predictions

|  | **p** | **d** | **q** |
| --- | --- | --- | --- |
| *Cluster 0* | 0 | 1 | 1 |
| *Cluster 1* | 3 | 0 | 4 |
| *Cluster 2* | 0 | 1 | 1 |
| *Cluster 3* | 1 | 0 | 1 |
| *Cluster 4* | 1 | 0 | 5 |

In order to use the Facebook’s Prophet, the data had to be prepared by performing EDA which was stated above, but unlike the ARIMA method the dataset of each cluster had to be changed similar to that of Figure 8 with two columns, one titled ‘ds’ which is the date and the other titled ‘y’ which is the weekly sales. Also in order to be able to work with this model the stationarity conditions stated above should be proven valid in the same way as in the previous model. Now the data is ready to feed to the Facebook Prophet model, like any model first create the model Prophet with the changepoint prior scale of 1 and fit the modified data to this model. Then using the make future dataframe feature with the model create a dataframe with the required periods needed to be predicted in this case it was 90 days. Now the model is ready to perform a forecast of the first cluster of the dataframe that was just created using the predict to predict the values. In order to best understand the results check the Figures below for clarification. Looking at Figure 9 we notice that the forecast begins after the black dots (changepoint) end. In order to check the predicted weekly sales we have to denormalize the values of yhat (which are the predicted values in the forecast dataframe), to do that the values should be multiplied by the difference of the max and min values than added to the min value, these values are shown in Figure 11.After that was completed the same steps were followed for the rest of the clusters and each cluster was predicted at three different periods which are 5 days, 30 days and 90 days.

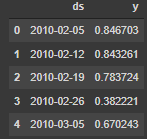


Figure 8. New Data Frame for Prophet

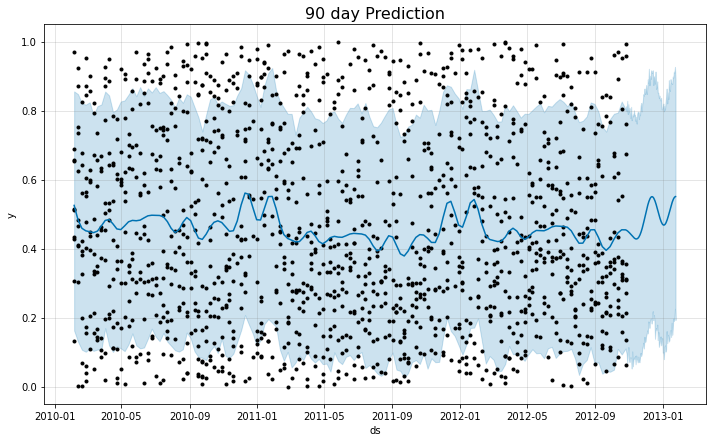


Figure 9. A prediction of the sales after 90 days

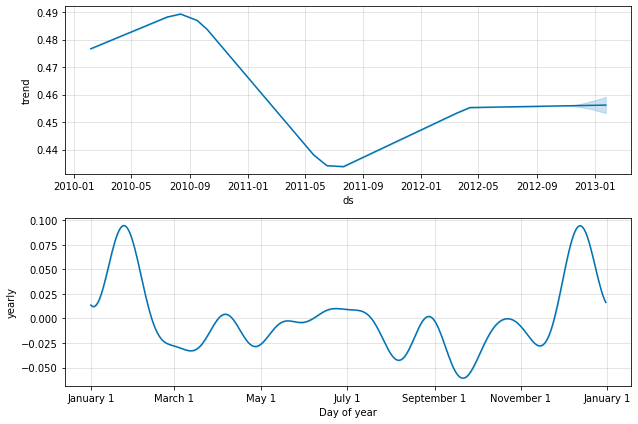


Figure 10. Forecast Components

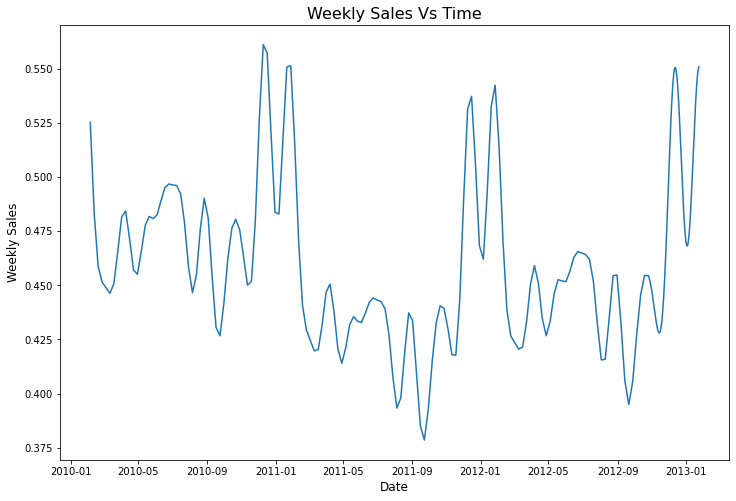
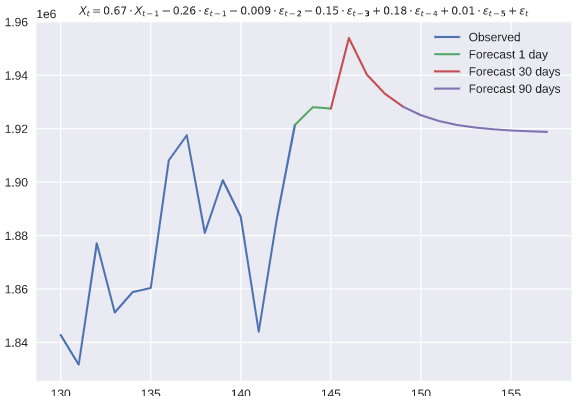


Figure 11. Weekly Sales with respect to the Date

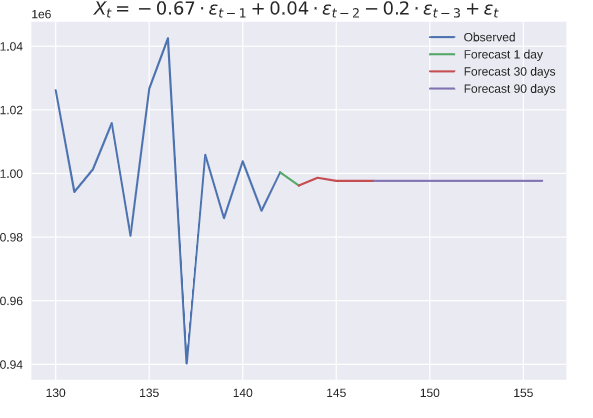
# Plots Predictions and models

## Cluster 0Cluster 1

## Cluster 2



## Cluster 3



## Cluster