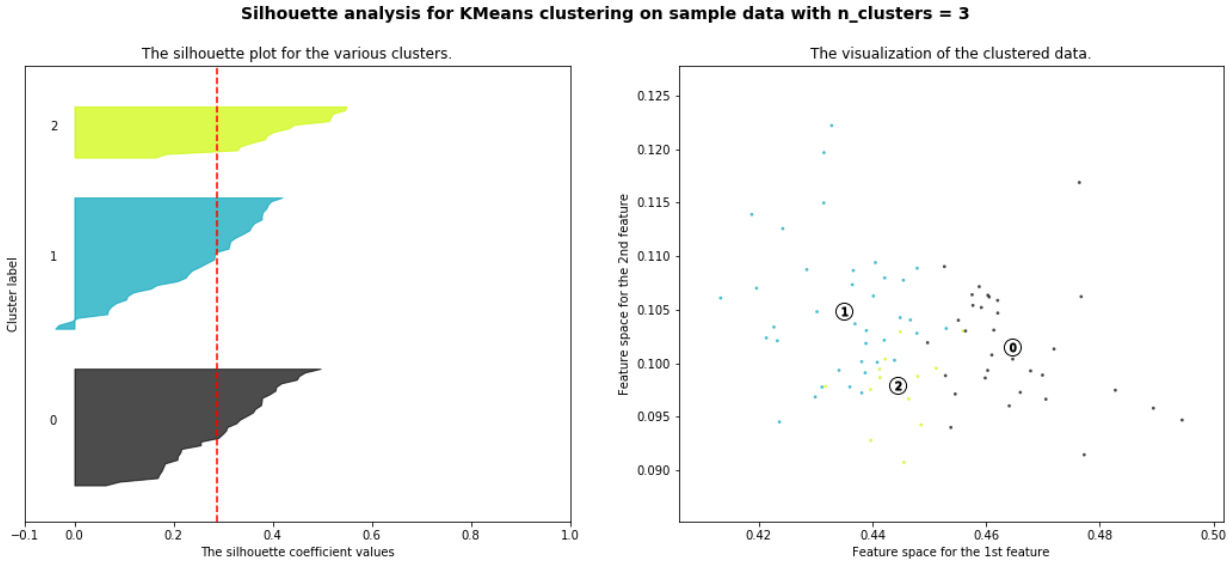
Determining Store Formats for Existing Stores

- Defining the optimal number of store formats.



Number of Clusters	Silhouette Score
2	0,247
3	0,289
4	0,253
5	0,213
6	0,229
7	0,244
8	0,239
9	0,249

Using the data from 2015 of the existing stores' sales, I clustered the stores with the k-means algorithm comparing the efficiency of the number of cluster varying from two to nine. According to the Silhouette Plot and Scores the optimal number of cluster is 3, since in the plot it is possible to see that the clusters have about the same size and in the table has the highest score.

- Checking how many stores fall into each format

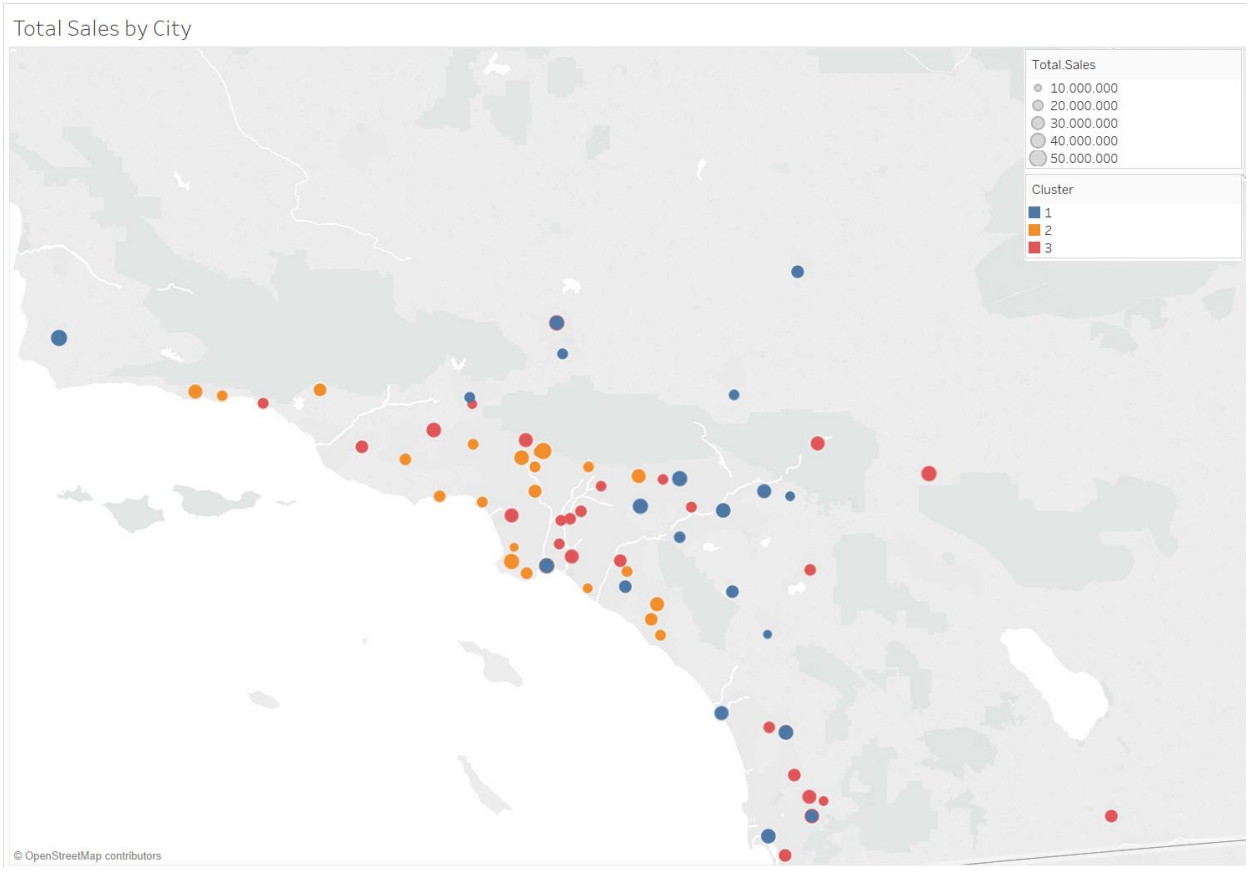
Cluster	Size
0	37
1	33
2	15

- Verifying one way that the clusters differ from one another.

Cluster	General Merchandise	Total Sales	General_Merchandise_Perc
0	6,7685,106.61	741,838,363.72	0.09124
1	55,824,735.43	796,715,969.04	0.070069
2	60,569,632.57	935,779,513.54	0.064726

One way the clusters vary is the total of the General Merchandise Sales with a difference of almost seven million, even though Cluster 0 has more stores across the three and the Cluster 2 has less across the three.

- In the map below is the location of the stores, with colors indicating the cluster and the size, the total sales.



Formats for New Stores

- Joining the data from stores and its respective clusters with its demographics data.

  I joined the data of stores' sales with its clusters and the demographics data from the city that each store is located, therefore now I have a dataset with every store ID, demographics and the segment the existing store falls into, with this data I will use to train 3 classification models to later predict each new store' segment. The classification models chosen were Decision Tree, Random Forest and AdaBoost(Decision Tree), with 20% as the validation data, the models were fitted and compared along in terms of accuracy and f1-score.

- Comparing classification models.

Model	**Accuracy**	**F1**
Random Forest	0.8235	0.8197
Decision Tree	0.7647	0.6857
Boosted Model	0.5294	0.4519

The chosen model was the Random Forest Model, the model highly overcome all other both in Accuracy as in F1 Score.

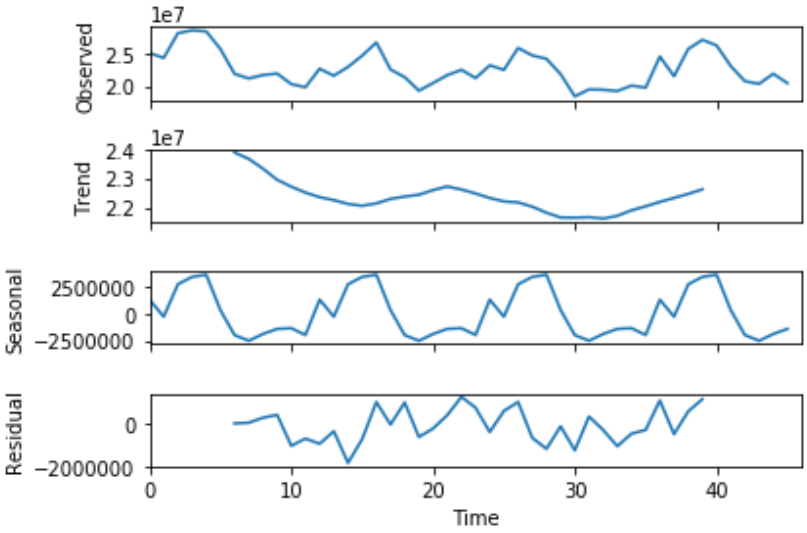
- Predicting the 10 new stores' segment with the Random Forest

Store Number	**Segment**
S0086	3
S0087	1
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Predicting Produce Sales for 2016

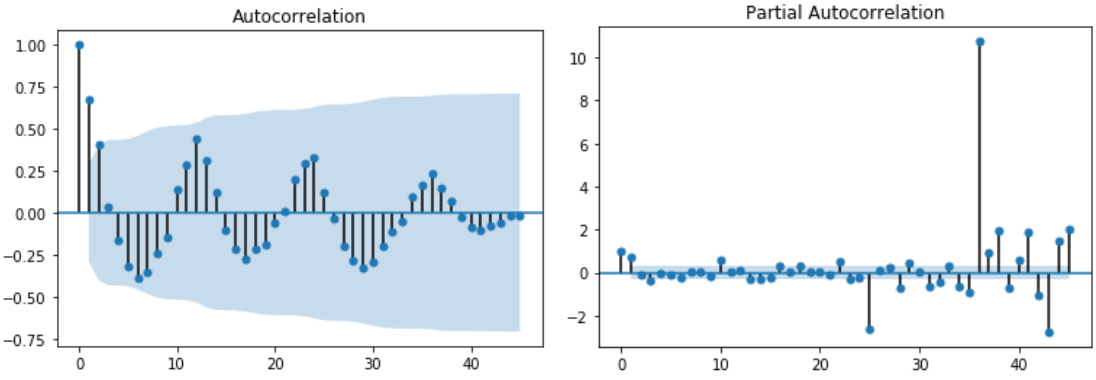
In the stores' sales data, there is data from March 2012 to December 2015 for sales in every type of product the stores sell. Then, they were used to forecast the stores' sales for each type of product for the year of 2016. For this, I used two Times Series algorithms called Exponential Smoothing (ETS) and Autoregressive Integrated Moving Average (ARIMA).

- Defining the terms for ETS

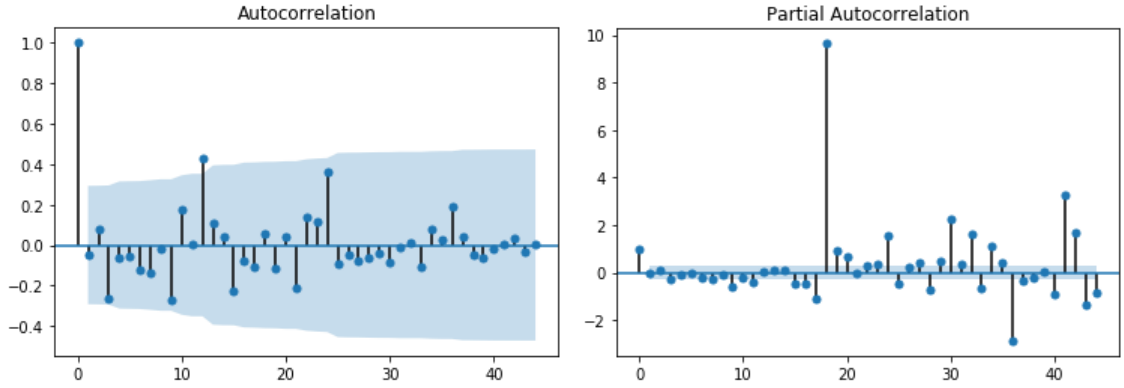


The ETS model was built from the analysis of the Decomposition Plot above and it shows that the Error is Multiplicative because the remainder plot is fluctuating between large and small errors over time, the Trend is None because there is no clear trend and the Seasonality is Multiplicative because the seasonal fluctuations tend to increase and decrease with the level of the time series. Therefore, the configuration for the model is ETS(M,N,M).

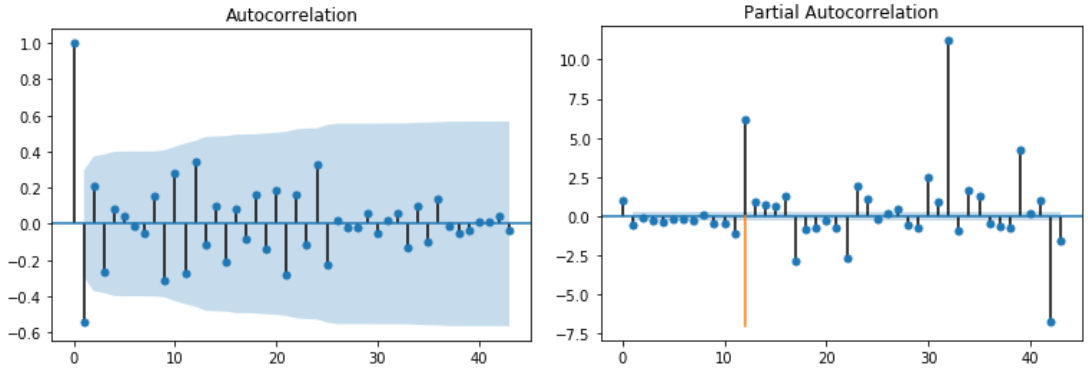
- Defining the terms for ARIMA



Analyzing the ACF/PACF graph it is possible to see that the series is not stationary, we can verify in the ACF graph that neither the mean nor the variance is constant over time, so it is needed to take the Seasonal Differencing.



After the Seasonal Difference is taken we can see in the ACF/PACF graphs that the series is still not stationary but the seasonal component is gone, therefore we need to take the First Differencing.



After the First Difference it is possible to see that the series is now stationary except for Lag 12.

So I took the a seasonal difference and a first difference and this suggest that the lower case d and the upper case D of the ARIMA model have the value of 1. The M term is 12 because there are 12 months or periods in each season. Therefore the configuration for the model is ARIMA(0,1,0)(0,1,0)[12].

Model	**RMSE**	**MAE**	**MASE**
ETS	798,987.79	726,330.51	0.4274
ARIMA	1,538,537.78	1,349,195.97	0.7939

I then compared the ETS(M,N,M) and the ARIMA(0,1,0)(0,1,0)[12] and the ETS model obtained the smallest error measures for RMSE and MASE in the validation sample, so I used this configuration to forecast the produce sales for 2016.

- Forecasting Every Store's Produce Sales for 2016

Date	Forecast Existing Stores	Forecast New Stores
2016-01	21,539,936.01	2,584,383.53
2016-02	20,413,770.60	2,470,873.92
2016-03	24,325,953.10	2,906,307.87
2016-04	22,993,466.35	2,771,532.13
2016-05	26,691,951.42	3,145,848.57
2016-06	26,989,964.01	3,183,909.28
2016-07	26,948,630.76	3,213,977.72
2016-08	24,091,579.35	2,858,247.21
2016-09	20,523,492.41	2,538,173.64
2016-10	20,011,748.67	2,483,550.17
2016-11	21,177,435.49	2,593,089.19
2016-12	20,855,799.11	2,570,200.44