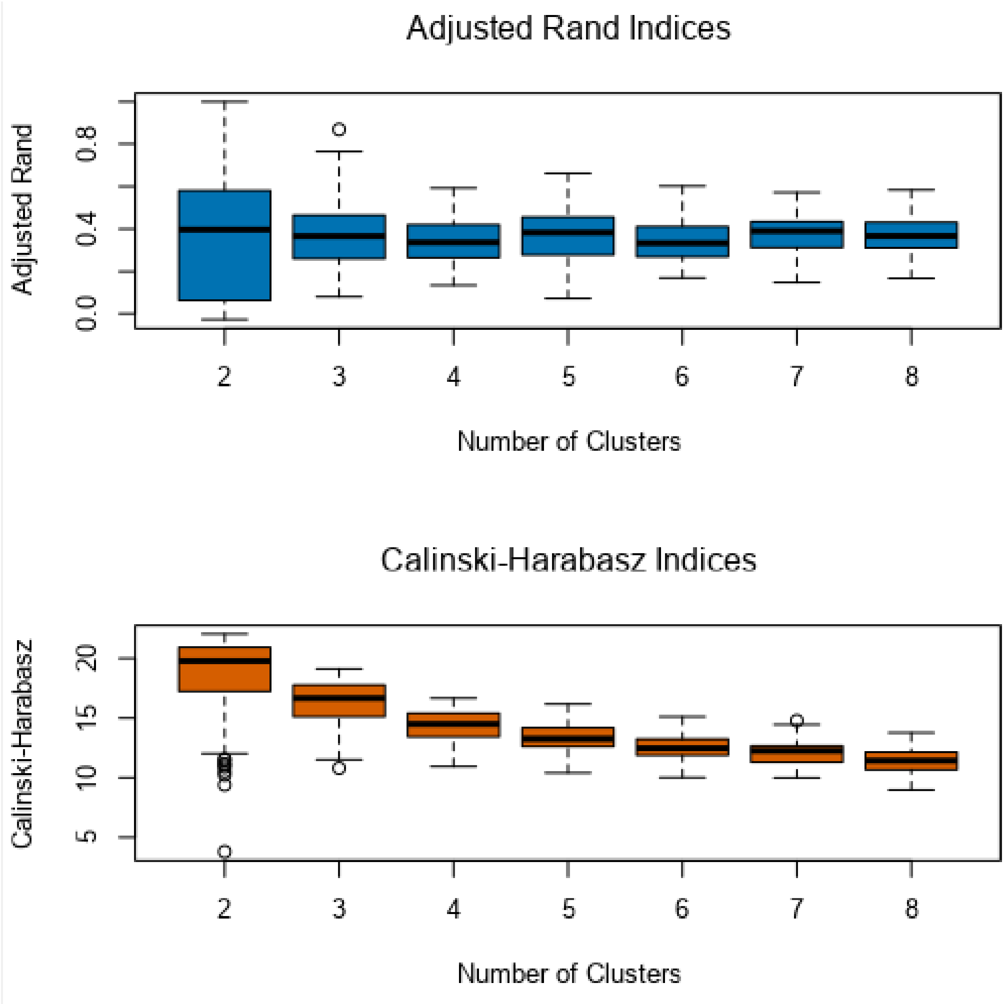
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

The optimal number of store formats is 3. I arrived at this number by looking at the K-Means Diagnostic.



Adjusted Rand Index tells us how similar data points are within clusters; the higher the index, the more similar they are. CH (Calinski-Harabasz) Index tells us how dense and well-separated clusters are; the higher the index, the more they are. Although 2 clusters model has higher median and 3rd quartile on both plots, the datapoints are also more spread out. We can see that 2 clusters have much higher variance than 3 clusters. Furthermore, when looking at the CH plot, we see 2 clusters have many outliers.

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.02775	0.08019	0.134532	0.072217	0.169763	0.147906	0.167458
1st Quartile	0.070414	0.259363	0.263959	0.27706	0.269708	0.3148	0.31042
Median	0.397046	0.366173	0.336574	0.383341	0.332336	0.390168	0.366566
Mean	0.378383	0.391248	0.349083	0.372545	0.342064	0.384792	0.369649
3rd Quartile	0.580007	0.463864	0.420191	0.45296	0.40902	0.433926	0.42987
Maximum	1	0.868402	0.591965	0.662271	0.601961	0.571377	0.583551

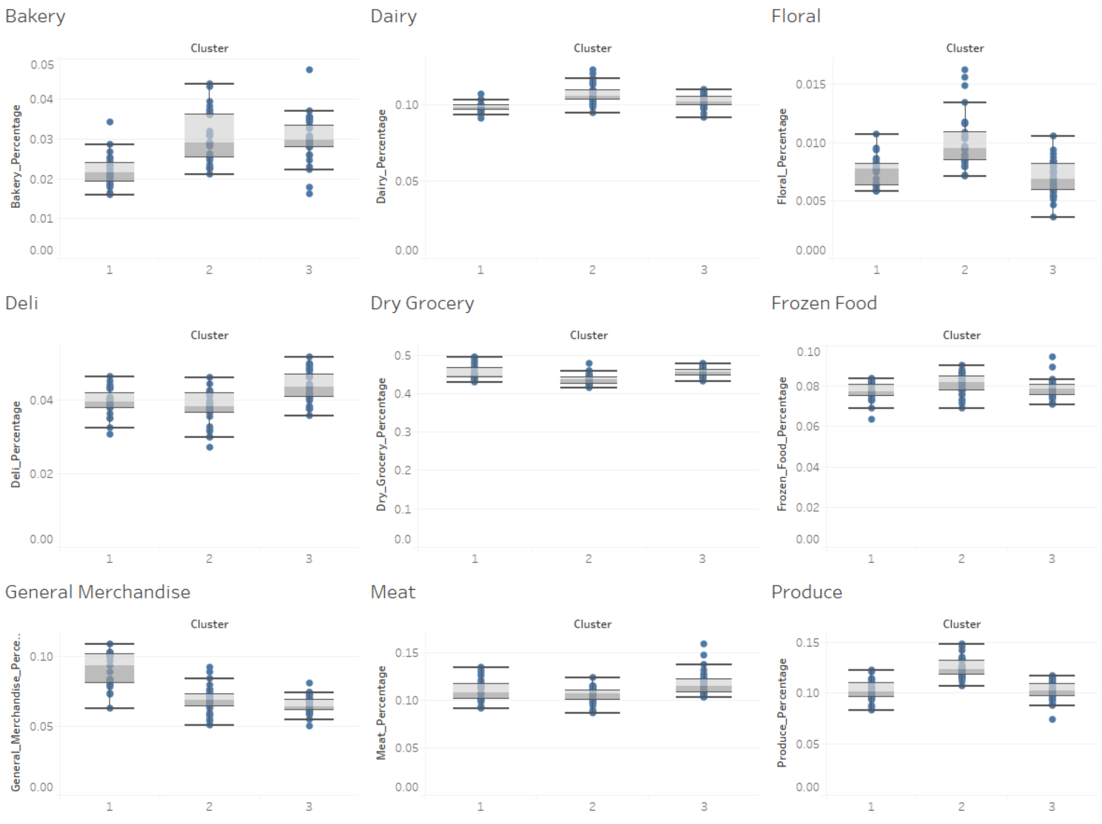
Calinski-Harabasz Indices:

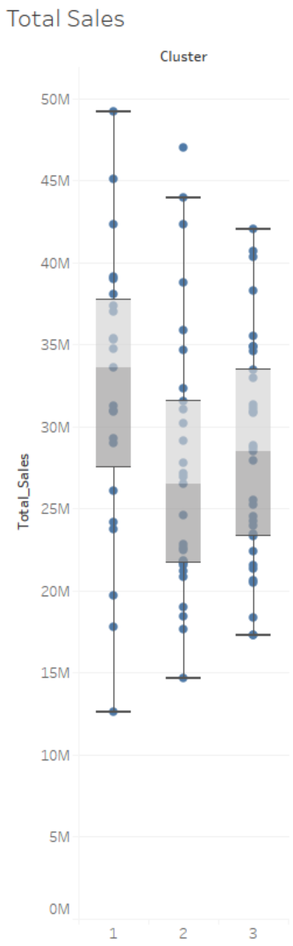
	2	3	4	5	6	7	8
Minimum	3.787506	10.80678	10.94687	10.41103	10.00938	9.984881	8.967392
1st Quartile	17.357005	15.1359	13.42154	12.64046	11.90145	11.318974	10.675962
Median	19.779239	16.6847	14.49294	13.27144	12.49155	12.271615	11.446404
Mean	18.386203	16.25191	14.35029	13.27766	12.59697	12.08478	11.460153
3rd Quartile	20.911233	17.78993	15.40083	14.17785	13.23228	12.689791	12.157365
Maximum	22.061691	19.11366	16.68051	16.18035	15.11493	14.780739	13.759128

Based on the above data I concluded that 3 is the optimal number of clusters.

Stores in Clusters

Cluster Number	Number of Stores
1	23
2	29
3	33



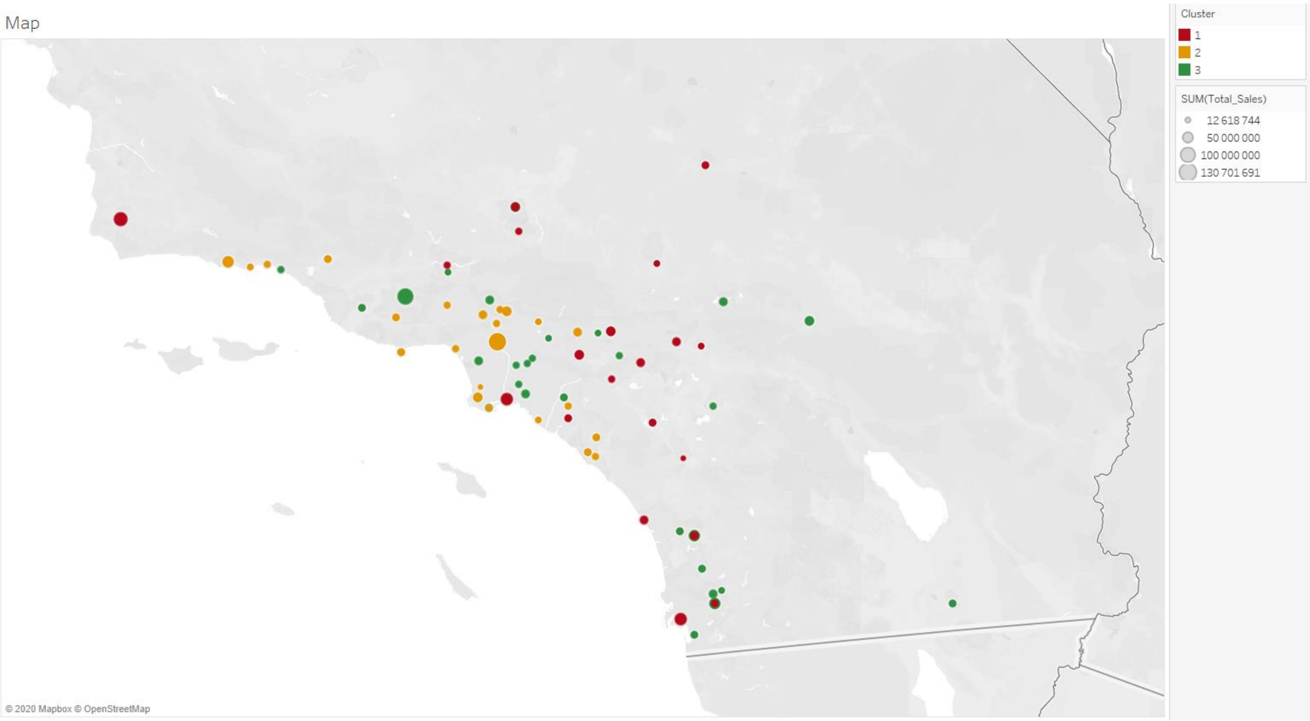


Looking at the plot on the left, we can see that clusters differ a lot in Total Sales. The 1st cluster has the highest median Total Sale but is also more spread out whereas the 3rd cluster has significantly lower median but is also more compact.
Also, we can examine whisker plots for each of the category variable (shown as a percentage of the total sale). Interestingly, some variables show little variance between different clusters (like Dry Grocery) while others experience major differences (like General Merchandise). This might be explained partially by the fact that Dry Grocery category contains essential products and Merchandise category does not thus does not have the same amount of variance in sales which the plot above illustrates. This might indicate that stores in segment 1 are in more popular locations where are presumably more tourists. This translates to lower sales in products like bread and higher sales of merchandise.

The stores' map below shows how clusters are distributed spatially.

Link to the map:

https://public.tableau.com/profile/szymon.trochimiak#!/vizhome/Task1NanodegreeFinalProject/Map?publish=yes



Task 2: Determine the Store Format for New Stores

Since store format is a categorical variable, I used a non-binary classification model, specifically a Boosted Model.

The table below shows accuracies and F-values of all classification models I used.

Model	Accuracy	F1
Forest	0.8235	0.8426
Decision Tree	0.7059	0.7685
Boosted	0.8235	0.8889

Although Forest and Boosted models have the same accuracy, Boosted Model has higher F1 value (weighted average of the recall (true positive rate) and precision).

Below are confusion matrices for each model.

Confusion Matrix of Decision Tree Model

	Actual 1	Actual 2	Actual 3
Predicted 1	3	0	2
Predicted 2	0	4	2
Predicted 3	1	0	5

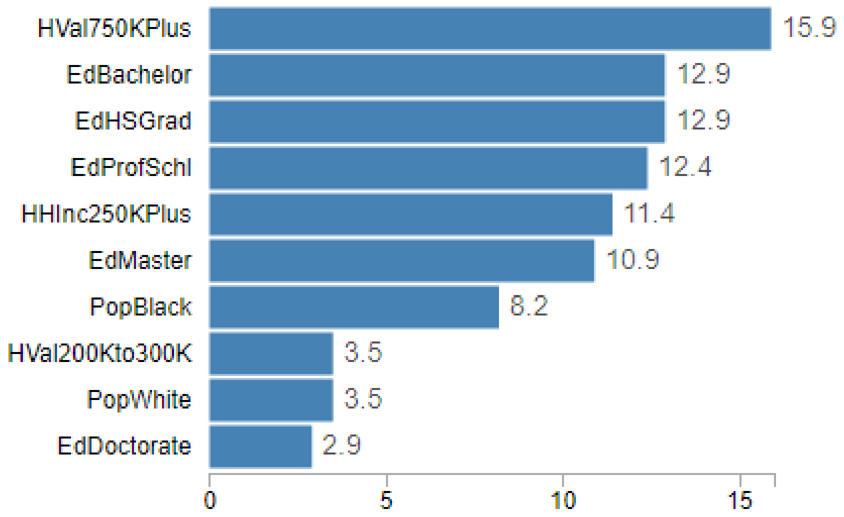
Confusion Matrix of Forest Model

	Actual 1	Actual 2	Actual 3
Predicted 1	3	0	1
Predicted 2	0	4	1
Predicted 3	1	0	7

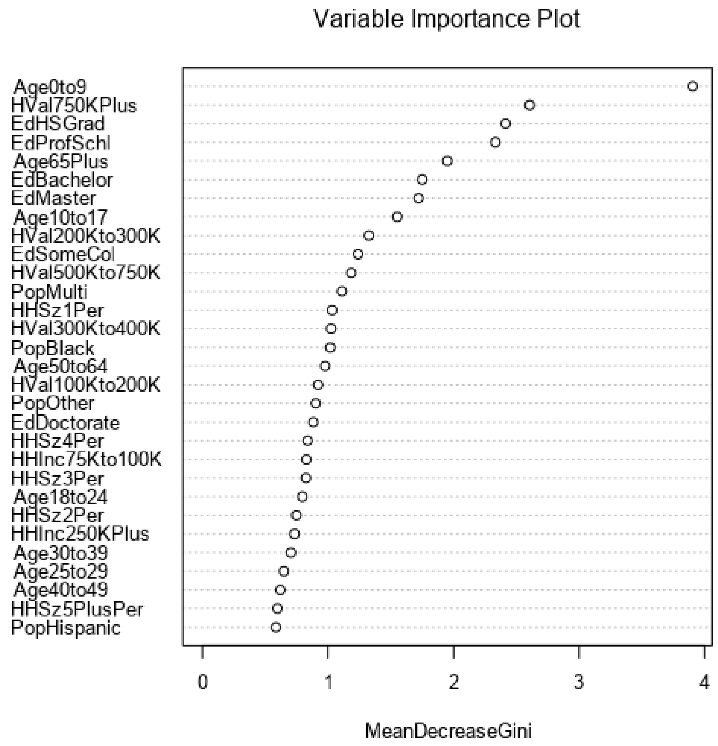
Confusion Matrix of Boosted Model

	Actual 1	Actual 2	Actual 3
Predicted 1	4	0	1
Predicted 2	0	4	2
Predicted 3	0	0	6

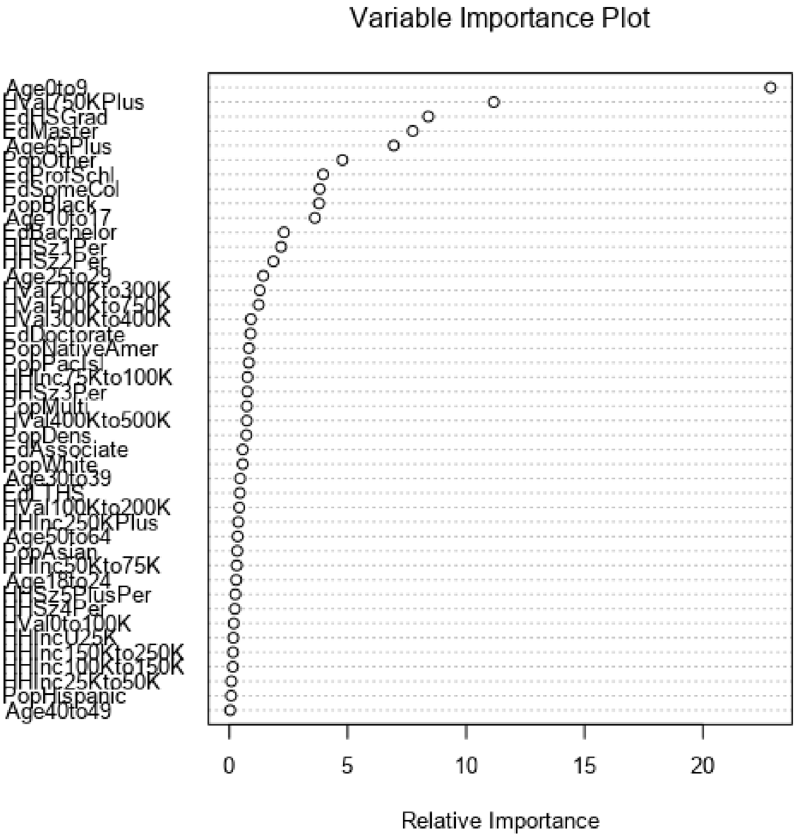
Decision Tree Model Variable Importance



Forest Model Variable Importance



Boosted Model Variable Importance



By looking at the plots above we see that **HVal750KPlus**, **EdHSGrad** and **Age0to9** are the most important variables overall. Even though **Age0to9** didn't show up on the Decision Tree's Variable Importance plot, it is the most important variable for Forest and Boosted model thus it clearly must be a very good predictor variable.

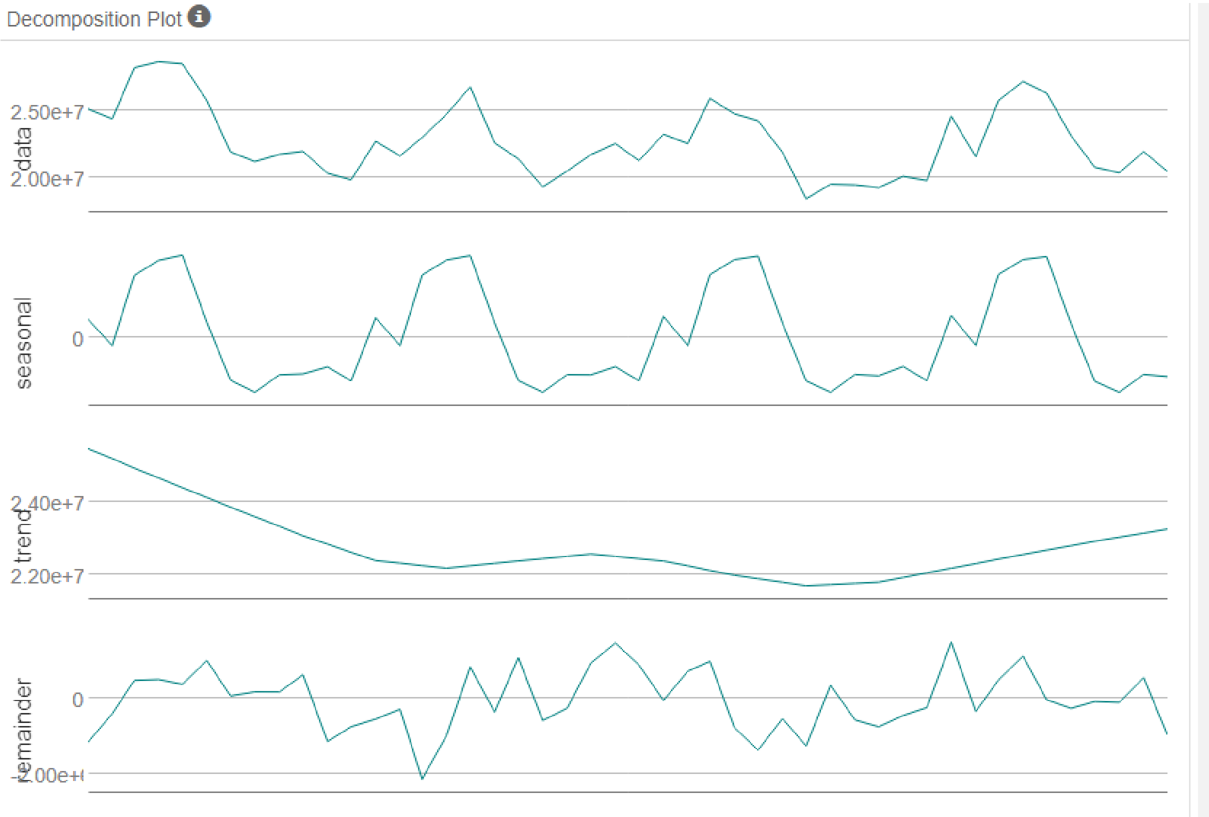
Below is the table describing to which segment each new store should belong.

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

Task 3: Predicting Produce Sales

I used ETS(M,N,M) for my forecast.

First, I examined the decomposition plot of the data.



I noticed that there is no trend, remainder has irregular pattern – it is not constant then I should use multiplicative method. Lastly I noticed that there are seasonal patterns. At first glance they look constant. However, when I ran two different ETS models (one used additive method for seasons and other multiplicative) the multiplicative one heavily outperformed the additive one.

I also tried ARIMA model (I set it to full AUTO) but it didn't perform nearly as well as ETS.

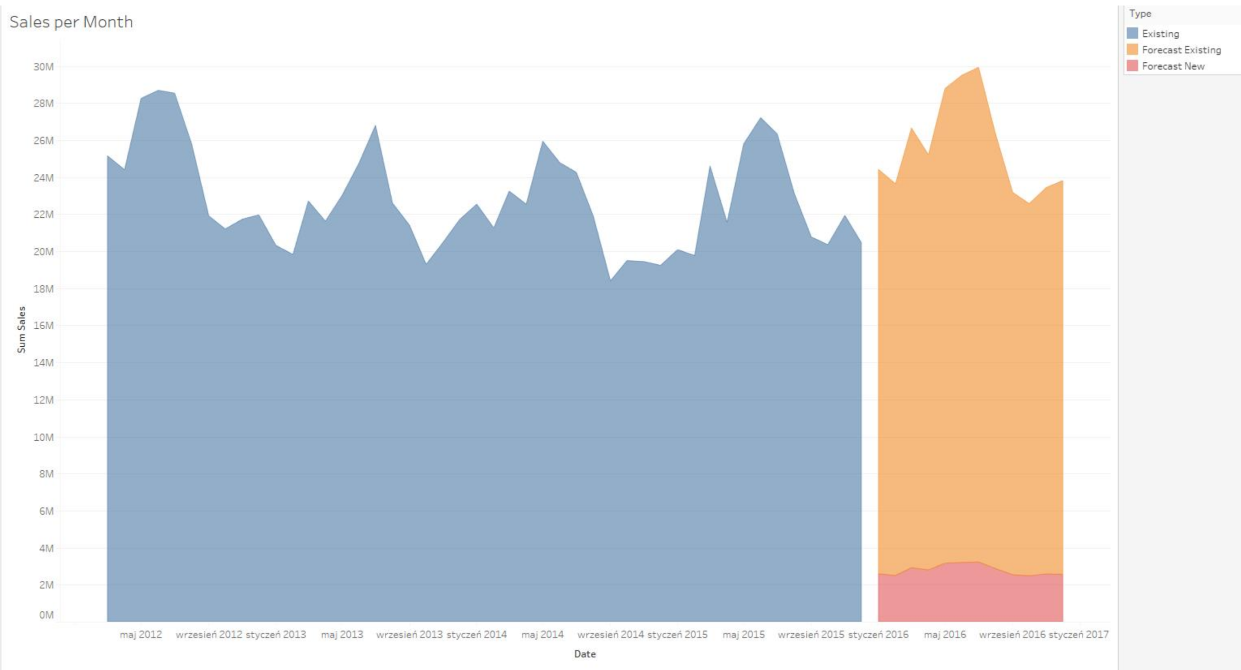
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_AUTO	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463
ETS_M_N_A	-939996.84	1096383.6	965002.2	-4.3116	4.4256	0.5678
ETS_M_N_M	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257

We can see that MASE (Mean Absolute Scaled Error) is almost twice as big in ARIMA and ETS(M,N,A) as it is in ETS(M,N,M). Other errors also are several times bigger than ETS(M,N,M) errors. This clearly indicates that this model is superior.

Forecasted Values

Month	New Stores	Existing Stores
Jan-16	2588356.56	21829060.03
Feb-16	2498567.17	21146329.63
Mar-16	2919067.02	23735686.94
Apr-16	2797280.08	22409515.28
May-16	3163764.86	25621828.73
Jun-16	3202813.29	26307858.04
Jul-16	3228212.24	26705092.56
Aug-16	2868914.81	23440761.33
Sep-16	2538372.27	20640047.32
Oct-16	2485732.28	20086270.46
Nov-16	2583447.59	20858119.96
Dec-16	2562181.70	21255190.24



Link to the above chart

https://public.tableau.com/profile/szymon.trochimiak#!/vizhome/Forecast_15896521829990/SalesperMonth?publish=yes