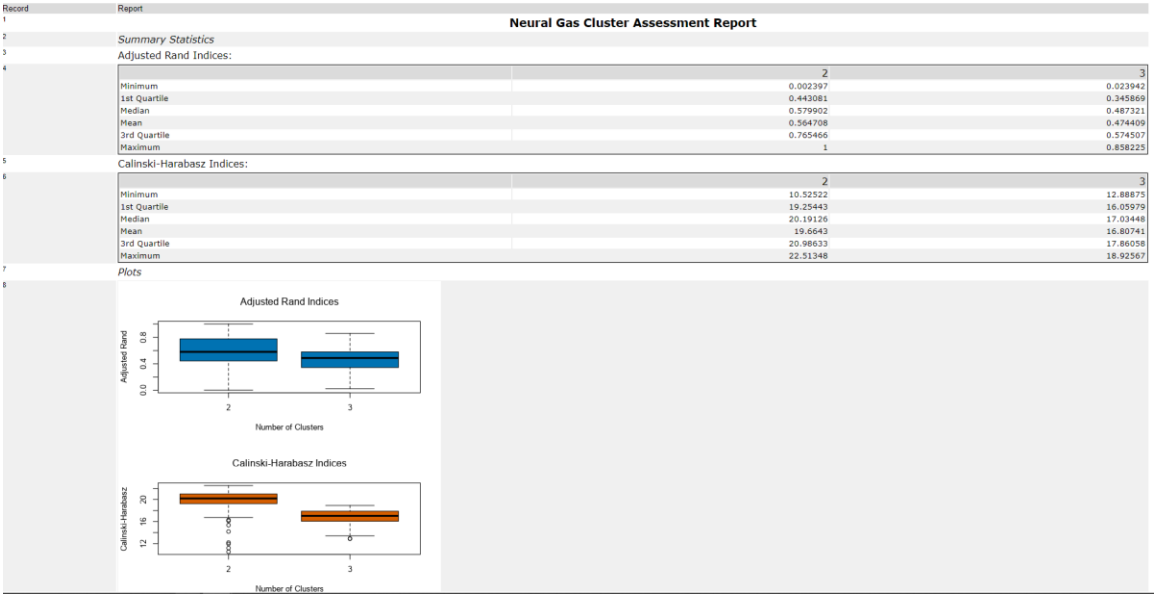
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?

   I used the K-Centroids Diagnostic tool to check for optimum number of clusters, Neural Gas, K-Medians Cluster & K-Mean Cluster all had 2 and 3 clusters having the highest values for adjusted Rand and Calinski-Harabasz indices.

   In my opinion, the number of optimal store formats is 3.



Neural Gas Cluster Assessment Report

Summary Statistics

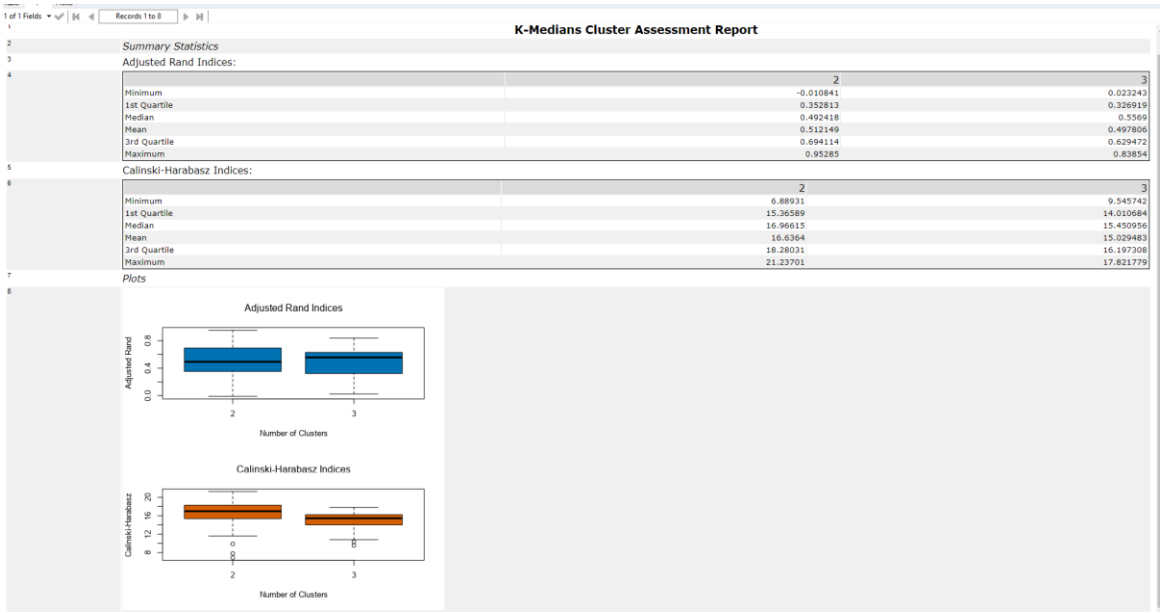
Adjusted Rand Indices:

	2	3
Minimum	0.002397	0.023942
1st Quartile	0.443081	0.345869
Median	0.579902	0.487321
Mean	0.564708	0.474409
3rd Quartile	0.765466	0.574507
Maximum	1	0.858225

Calinski-Harabasz Indices:

	2	3
Minimum	10.52522	12.88875
1st Quartile	19.25443	16.05979
Median	20.19126	17.03448
Mean	19.6643	16.80741
3rd Quartile	20.98633	17.86058
Maximum	22.51348	18.92567

Plots



K-Medians Cluster Assessment Report

Summary Statistics

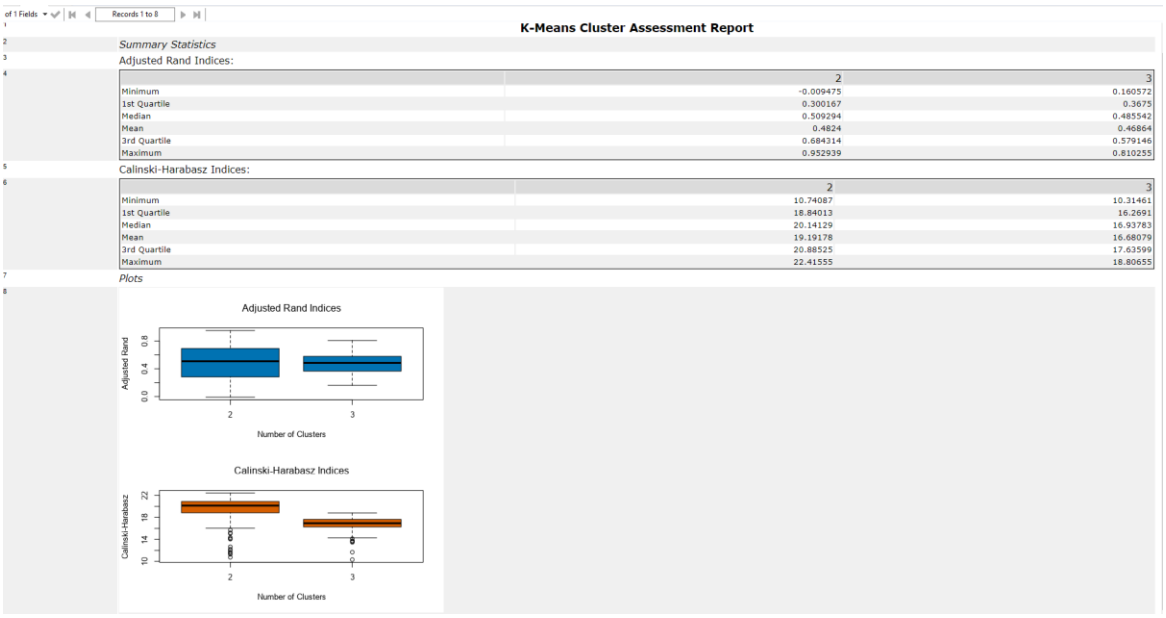
Adjusted Rand Indices:

	2	3
Minimum	-0.010841	0.023243
1st Quartile	0.352813	0.326919
Median	0.492418	0.5569
Mean	0.512149	0.497806
3rd Quartile	0.694114	0.629472
Maximum	0.95285	0.83854

Calinski-Harabasz Indices:

	2	3
Minimum	6.88931	9.545742
1st Quartile	15.36589	14.010684
Median	16.96615	15.450956
Mean	16.6364	15.029483
3rd Quartile	18.28031	16.197308
Maximum	21.23701	17.821779

Plots



K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3
Minimum	-0.009475	0.160572
1st Quartile	0.300167	0.3675
Median	0.509294	0.485542
Mean	0.4824	0.46864
3rd Quartile	0.684314	0.579146
Maximum	0.952939	0.810255

Calinski-Harabasz Indices:

	2	3
Minimum	10.74087	10.31461
1st Quartile	18.84013	16.2691
Median	20.14129	16.93783
Mean	19.19178	16.68079
3rd Quartile	20.88525	17.63599
Maximum	22.41555	18.80655

Plots

2. How many stores fall into each store format?

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Store Format/**Cluster**	Number of Stores/**Size**
1	23
2	29
3	33

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

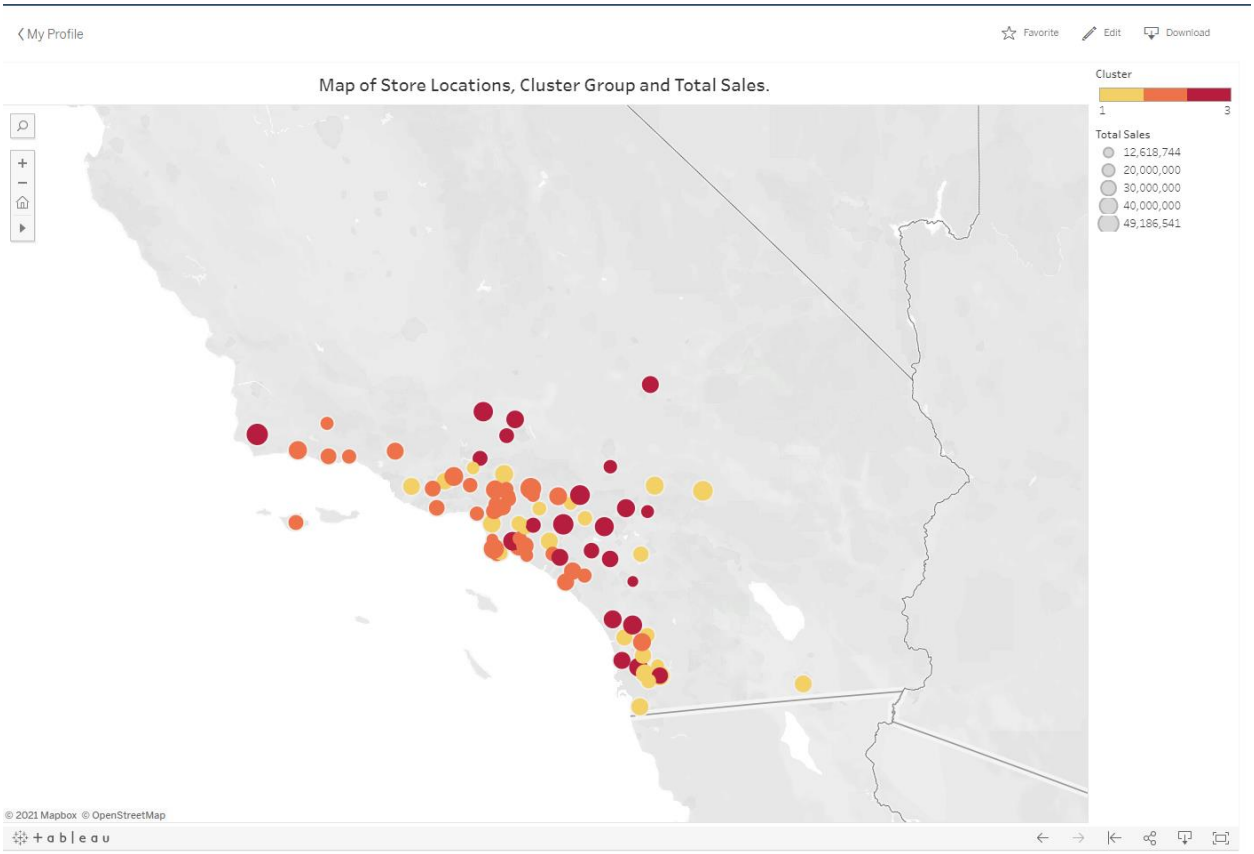
   The third cluster has the smallest Average distance 2.11, being the most compact and stable among the other & more separated from the other two clusters.

   While the second cluster has the highest Maximum distance 4.47 from the centroid.

4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.

   Link to the Viz on Tableau Public:

   https://public.tableau.com/profile/joshua.okafor#!/vizhome/MapofStoreLocationsandClusters_16209101900290/Map?publish=yes



Task 2: Formats for New Stores

1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

   To predict the store formats for the new stores, demographic data from StoreDemographicData.csv was used. All the variables were kept as predictor variables and run through a boosted, decision tree and random forest model. An 80/20 split of the data was used for training and validating the models.

   I trained the training data using Decision Tree, Forest, and Boosted Models on the stores with cluster values and using the accuracy, classification reports, ROC Graphs and accuracy in cluster categories I chose the **Boosted Model** as it performed better with the highest F1 Score.

Model Comparison Report

Fit and error measures

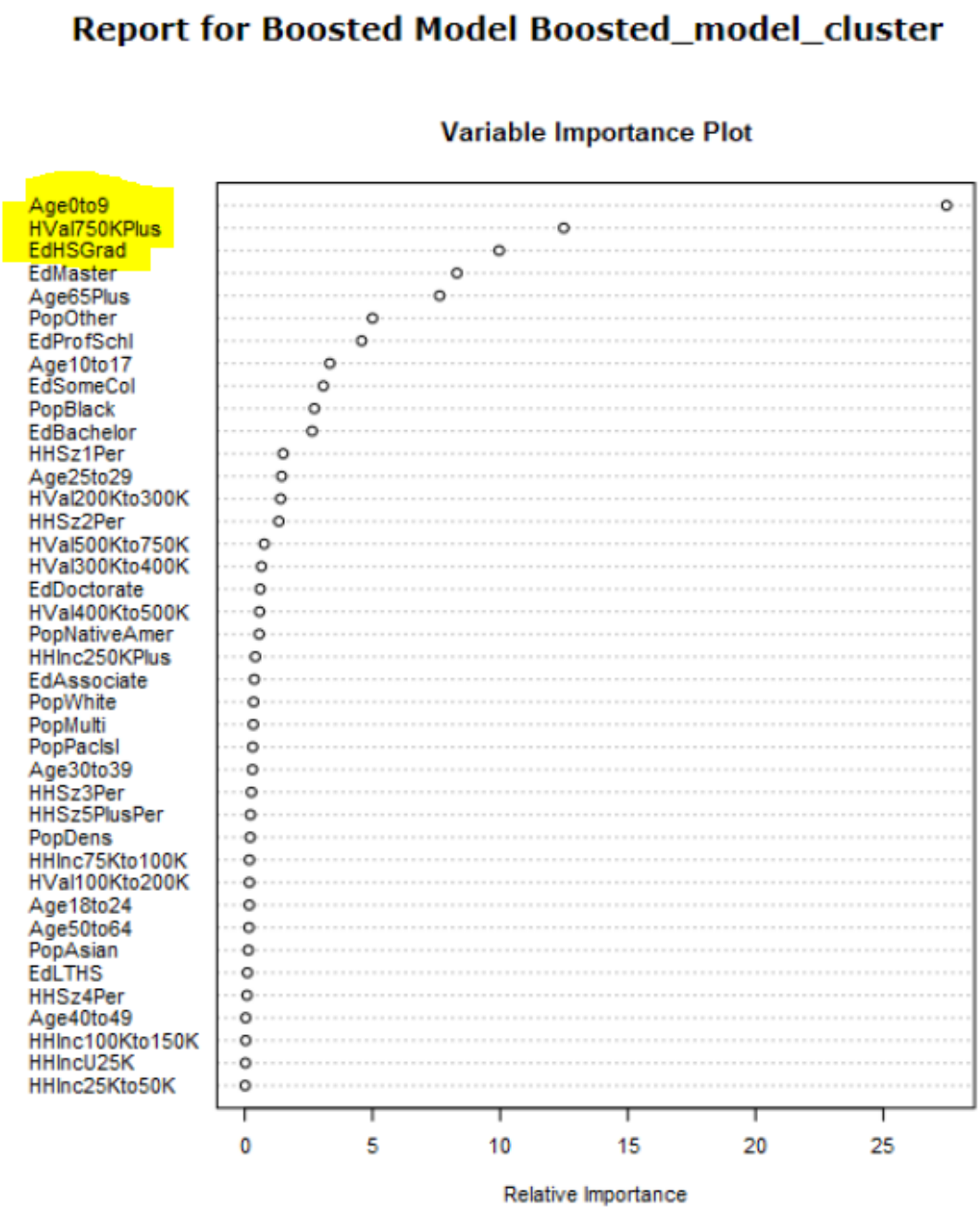
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_cluster	0.8235	0.8426	0.7500	1.0000	0.7778
Fores_model_cluster	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_model_cluster	0.8235	0.8889	1.0000	1.0000	0.6667

The confusion matrix of Boosted Model shows the highest True positives values and high sensitivity, it means rarely fail diagnosis.

Confusion matrix of Boosted_model_cluster

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Below is the variable importance plot for the Boosted model chosen for the final prediction.



Based on the Above plot, the 3 most important variables for the boosted Model are **Age0to9, HVal750KPlus, EdHSGrad**.

2. What format do each of the 10 new stores fall into? Please fill in the table below.

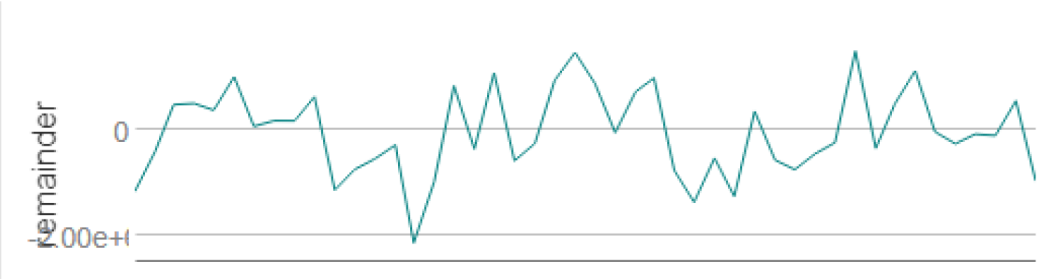
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Task 3: Predicting Produce Sales

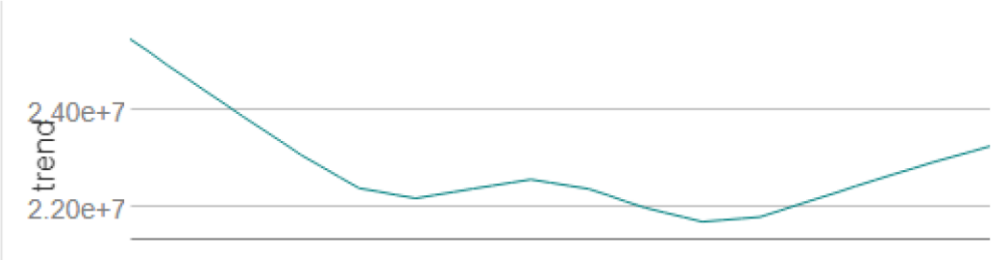
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

I Compared the performance of ETS and Arima model to select which to be applied for the stores forecast.

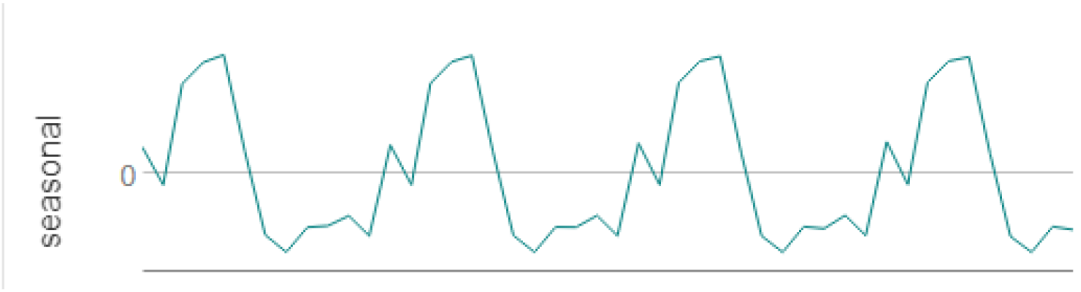
ETS: From the Decomposition Plot we observed the M, N, M pattern.
The Error plot shows variance along the years, it is fluctuating with different sizes, this means we used the error multiplicatively (M.)



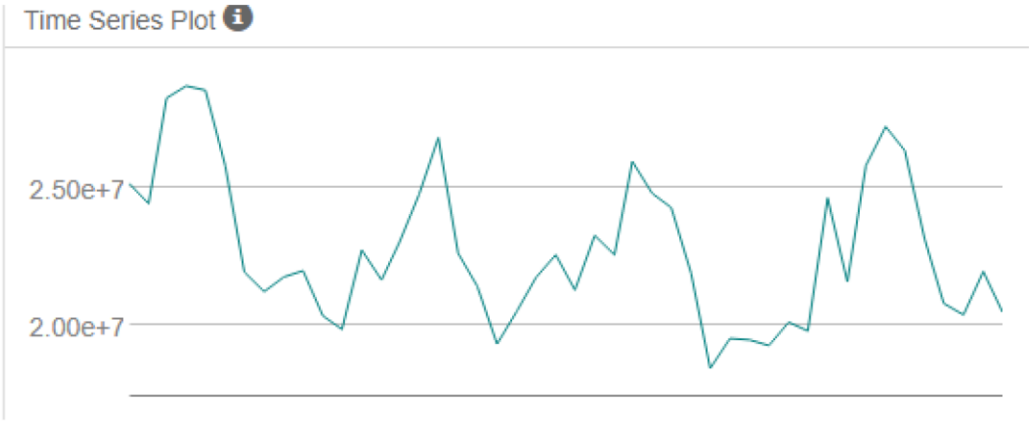
The Trend Plot, we observed that the trend moves uptrend and downtrend.



The Seasonal plot shows peaks and valleys in similar periods of time, this suggests applying seasonality in a multiplicative method (M.)



ARIMA: Using the TS Plot tool, the dataset showed a time series plot non-stationary because it had too many peaks and valleys, as shown in the graph below.



Autocorrelation Function Plot (ACF) and Partial Autocorrelation Function Plot (PACF) graphs showed the positive correlation, being suggested the AR 1 and MA 0 terms. ARIMA (1,1,1) (0,1,0) [12] depicts that there was a positive autocorrelation at period 1.

Comparison Between **ETS & ARIMA**:

After applying the TS Compare tool to Original Data, we obtained the comparison of the models.

ETS model in overall has the best accuracy measures (report below); especially the lowest MASE, MAPE, and RMSE. Therefore, we choose the ETS model to the forecast produce sales for the new and existing stores.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_MNM	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
ARIMA	-604232.3	1050239.2	928412	-2.6156	4.0942	0.5463

2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Month	Year	Existing Store Sales Forecast	New Store Sales Forecast
Jan	2016	21,539,930	2,587,451
Feb	2016	20,413,771	2,477,353
Mar	2016	24,325,953	2,913,185
Apr	2016	22,993,466	2,775,746
May	2016	26,691,951	3,150,867
Jun	2016	26,989,964	3,188,922
Jul	2016	26,948,631	3,214,746
Aug	2016	24,091,579	2,866,349
Sept	2016	20,523,492	2,538,727
Oct	2016	20,011,749	2,488,148
Nov	2016	21,177,435	2,595,270
Dec	2016	20,855,799	2,573,397

Link to Viz on Tableau Public:
https://public.tableau.com/profile/joshua.okafor#!/vizhome/TableauSalesForecasts_16209158348970/Task3?publish=yes