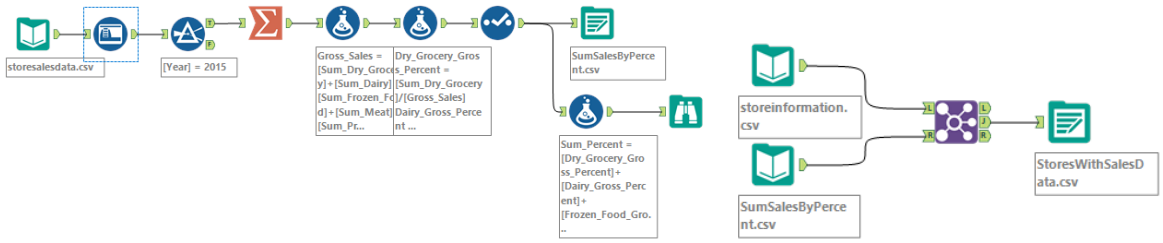
Predictive Analytics

Brian Wenger

The business problem for this project is to identify clusters for current stores, analyze demographic data to assign new stores to those clusters, and then forecast the average produce sales for existing and new stores.

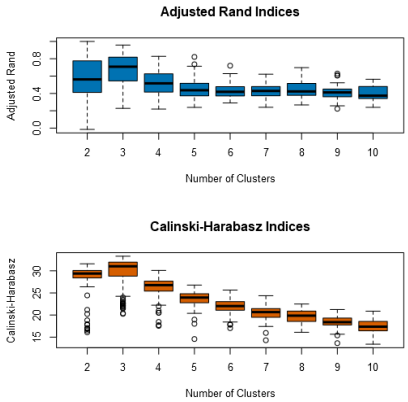
Task 1: Determine Store Clusters for Existing Stores

Data Prep



In order to accurately cluster the existing stores, first the data must be cleaned and prepared. The data needs to be filtered by year (“2015”) and by type (“Existing”). The sales data is summed by StoreID and category. A new field is created for the gross sum of all product categories for each store. New fields are created for each product which calculates the percentage of total store sales (category sales divided by total sales).

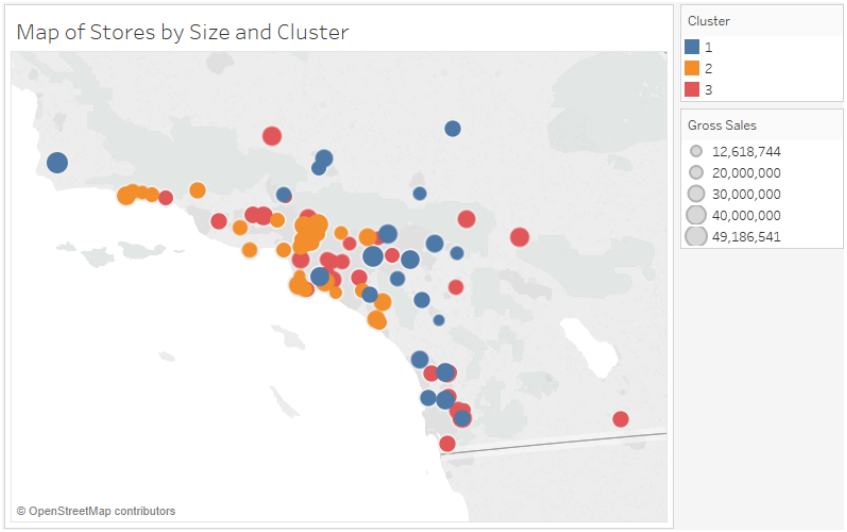
Clustering and Analysis



To best determine the number of clusters we need to run a K-Centroid diagnostic, for this project a range of 2 to 10 clusters were used for evaluation. As we are going with a K-Means approach we choose that model type and assess the results with the Adjusted Rand and Calinski-Harabasz charts. The Adjusted Rand chart shows which cluster number has higher stability, in this case the highest cluster is number 3 with a median index of 0.7083. The Calinski-Harabasz chart shows which cluster number is more distinct and compact, in this case the highest cluster is number 3 with a median index of 31.02. In both chart types, the result with the highest index shows the better performer. As both results show a cluster size of three being the best performer, three will be the number of clusters used.

Record #	Cluster	Count
1	1	23
2	2	29
3	3	33

After confirming the number of clusters to use (3), a K-Centroid Cluster Analysis tool and an Append Cluster tool are used to assign the clusters to individual stores. We apply the same settings as we performed in the diagnostics, however we use a higher starting seed for better accuracy. The store counts for the three clusters are as follows: Cluster 1 has 23 stores, cluster 2 has 29 stores, and cluster 3 has 33 stores.

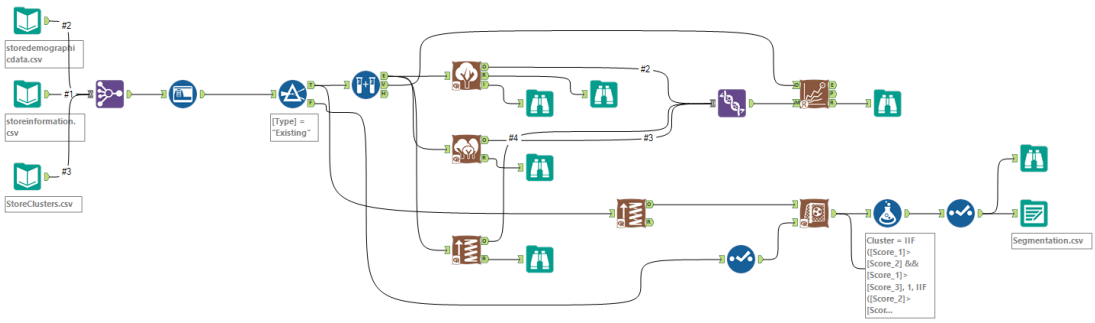


	Dry_Grocery_Gross_Percent	Dairy_Gross_Percent	Frozen_Food_Gross_Percent	Meat_Gross_Percent	Produce_Gross_Percent	Floral_Gross_Percent	Deli_Gross_Percent
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Bakery_Gross_Precent	General_Merchandise_Gross_Percent
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Based on the results of the clustering model, one way that the clusters differ from one another is where they typically reside on the map. Cluster 1 tends to be located in the mainland, spreadout from other stores. Cluster 2 resides along the northern coastline. And Cluster 3 tends to reside along the souther coast, but not as compact as Cluster 2.
Another way they differ is in what percentage of sales in specific categories are bought at their stores. Dairy for example, Cluster 1 spends far less on dairy than Cluster 2, which spends the most. Cluster 3 is in the middle between the two.

Task 2: Clusters for New Stores



To find the segment each new store should be contained in, first the demographic data must be prepared and organized. Data is combined from two of the included data sets (StoreDemographicData and StoreInformation) as well as the results of Task 1 (StoreClusters). The data is then filtered to separate the existing stores from the new stores.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DecTree	0.7059	0.7327	0.6000	0.6667	0.8333
Forest	0.8235	0.8251	0.7500	0.8000	0.8750
Boosted	0.8235	0.8543	0.8000	0.6667	1.0000

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Boosted

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

Confusion matrix of DecTree

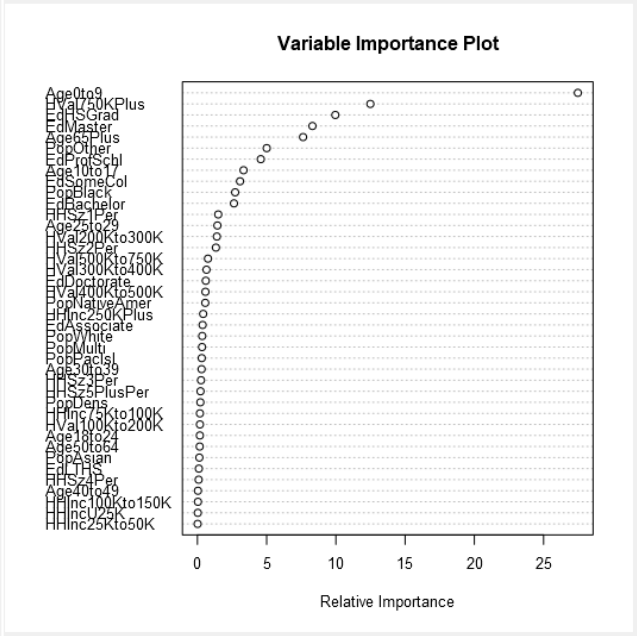
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of Forest

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

To conduct the analysis, there is a 20% hold-out sample to compare and verify the performance of each classification model. The remaining 80% is used for estimation. The only variables that are used for training the model are the demographic data variables. Three classification models are implemented: Decision Tree, Forest, and Boosted. Using the Model Comparison tool, by testing each model against the hold-out sample, the Boosted had the highest accuracy at 82.35% with the highest F1 of 85.43%. With this being the highest accuracy, it will be used to assign clusters to the new stores. The full dataset is then used to perform the final analysis and score the new stores. Using a Model Score tool followed by nested IF statements, a new field is created with assigned cluster. The following table shows the assigned cluster and the percentage of confidence the Forest model had for each cluster for each store.

Record #	Store	Cluster	Score_1	Score_2	Score_3
1	S0086	1	0.494734	0.012388	0.492878
2	S0087	2	0.063566	0.824404	0.11203
3	S0088	3	0.457695	0.060476	0.481829
4	S0089	2	0.02465	0.926534	0.048815
5	S0090	2	0.023626	0.918913	0.057461
6	S0091	1	0.927411	0.003404	0.069185
7	S0092	2	0.039425	0.886528	0.074047
8	S0093	1	0.890308	0.005433	0.104259
9	S0094	2	0.006694	0.965396	0.02791
10	S0095	2	0.109779	0.508989	0.381232

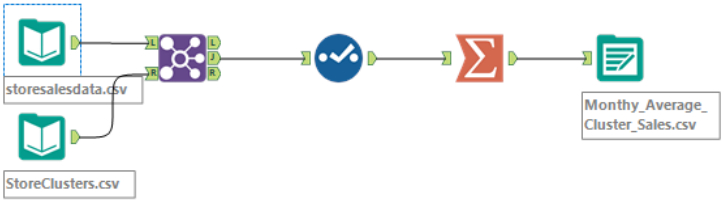


The three most important variables are “Age0to9”, “HVal750KPlus”, and “EdHSGrad”. The new stores fall into the following clusters:

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

Task 3: Predicting Produce Sales

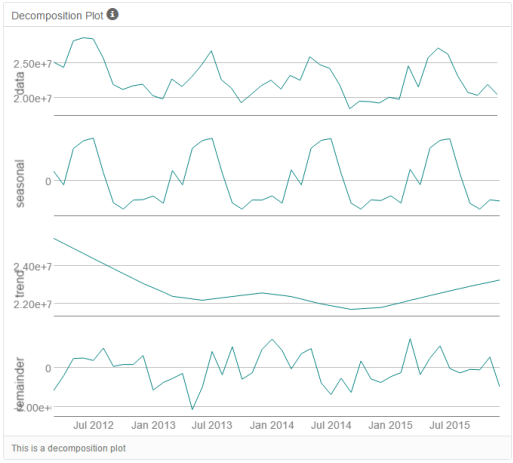
Data Cleanup



StoreSalesData and StoreClusters are combined to pull in the clusters into the data set. The

summarize tool is used to group by year and month, and sum produce.

Trend, Seasonal, and Error

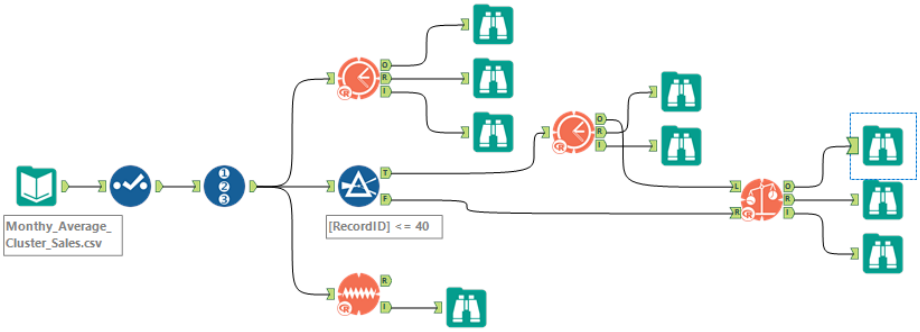


Trend – The trend in this plot shows no clear trend as can be seen from the trend plot on the decomposition plot. The plot appears to being going down and then back up, with one peak in the middle, therefore no trend component is included (N).

Seasonality – This time-series has a defined seasonality with the peaking occurring at its peak in July of each year and it valleys in October of each year. Over time the amount of variance for the seasonality decreases slightly, therefore this is multiplicative (M).

Error – The error or “remainder” chart is fluctuating between large and small errors over time, the error is multiplicative (M).

ETS

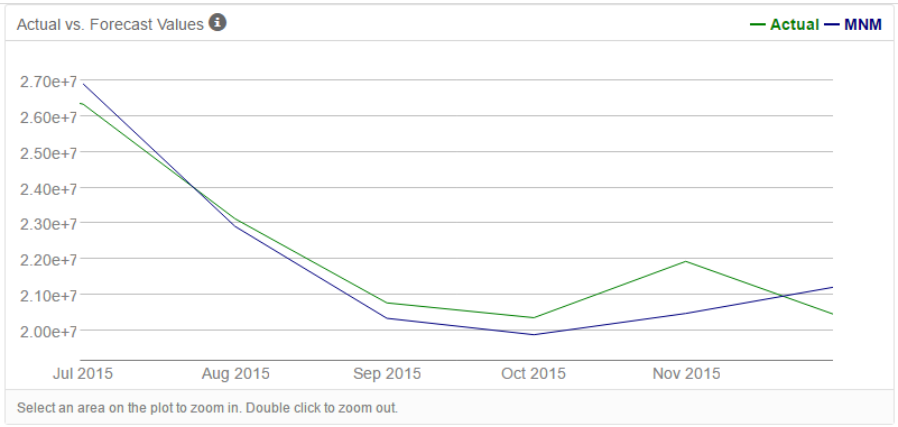


Within the model a holdout group of 6 is taken. We set the target field to “Sum_Produce” (which is grouped monthly from above) and ensure the frequency is set to ‘Monthly.’ Based on the

analysis from above we set the error to be multiplicative (M), the trend to No Trend (N), and the seasonality to multiplicative (M).

Based on the above analysis we will be testing the following model:
ETS (m, n, m)

After comparing the model to the actual data, the following was discovered:



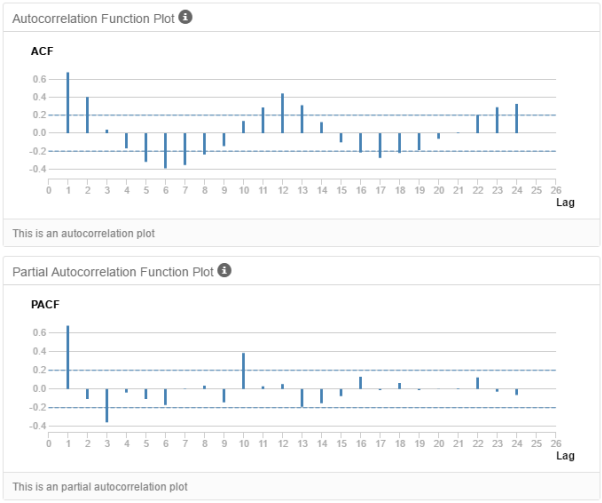
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
MNM	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA

The RMSE (Root Mean Squared Error) of this model is 760267.3, which is the standard deviation from the mean in the model. The model with a lower number here indicates a more consistent and more accurate model for prediction.

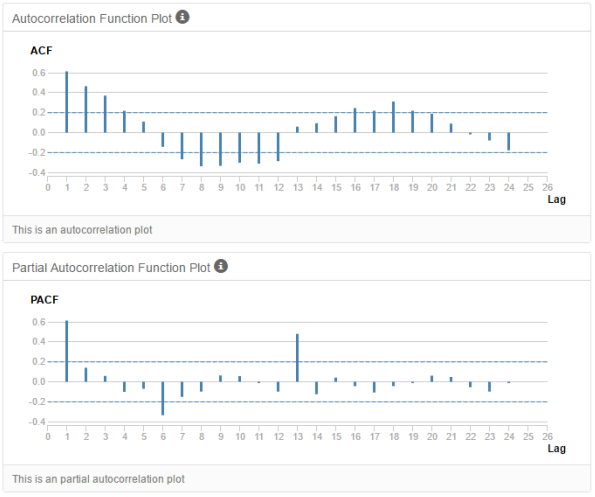
The MASE (Mean Absolute Scaled Error) here is 0.3822. The MASE ideally should be less than 1 to be considered effective.

ARIMA

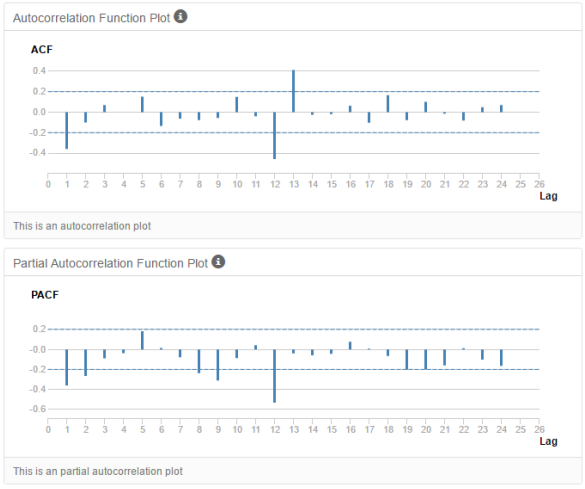
In the ARIMA model, it needs to be decided what values to use for the AR, I, and MA values. In order to do this, first we need to differentiate the values to check for correlation and partial correlation among the values in the monthly sales variable. The first image below shows the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) charts, these were used to decide on how to implement the ARIMA model.



Before differentiation

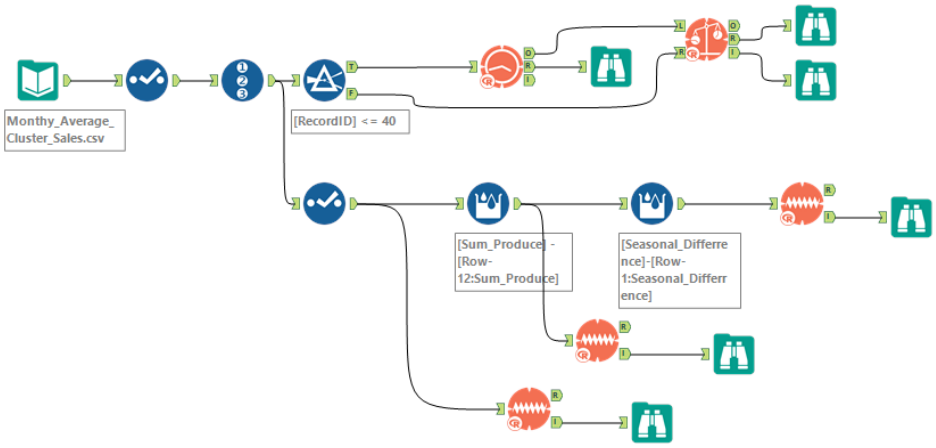


After differentiation



After seasonal differentiation

ARIMA(p,d,q)(P,D,Q)



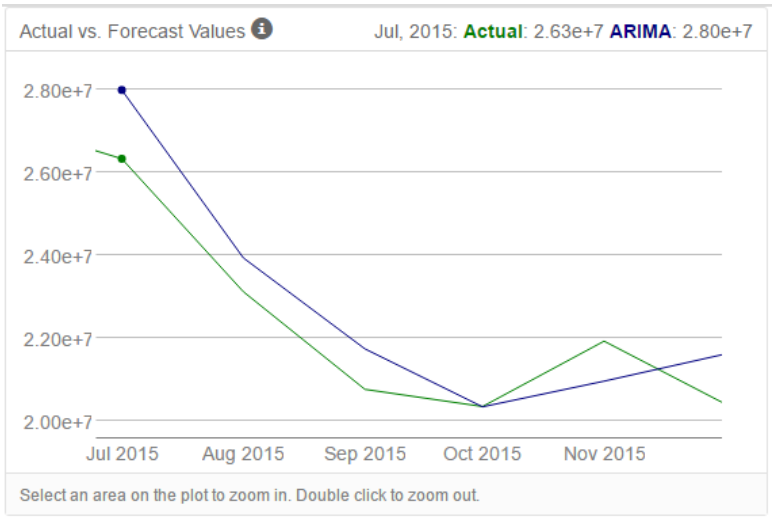
Autoregressive Component (p, P)
The reason to have AR components include positive correlation in the first lag in the PACF, where the ACF gradually decreases to 0. The first lag in the PACF is positive, the ACF gradually decreases to 0. We will set the autoregressive component the AR seasonality will be set to 1.

Integrated Component Term (d, D)
Because the data was non-stationary, we need to use differing to make the time-series stationary. This was done once to make the trend stationary and once to make the seasonality stationary. There was no need to do a first differencing therefor the integrated component is set to 0, and since there was differentiating done on the seasonality we will set the seasonal integrated component to 1 as well.

Moving Average Component (q, Q)
The reason to have MA components include negative correlation in the first lag in the ACF, where the PACF gradually decreases to 0. This model shows the positive correlation in the ACF and undefined in the PACF. We will be setting the MA component to 0. This model shows the positive correlation in the ACF and undefined in the PACF. We will be setting the seasonal MA component to 0.

Based on the above analysis we will be testing the following model:
ARIMA(1, 0, 0) (1, 1, 0)

After comparing the model to the actual data, the following was discovered:



Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	-604232.3	1050239	928412	-2.6156	4.0942	0.5463	NA

The RMSE (Root Mean Squared Error) of this model is 1050239, which is the standard deviation from the mean in the model. The model with a lower number here indicates a more consistent and more accurate model for prediction.

The MASE (Mean Absolute Scaled Error) here is 0.5463. The MASE ideally should be less than 1 to be considered effective.

Forecast

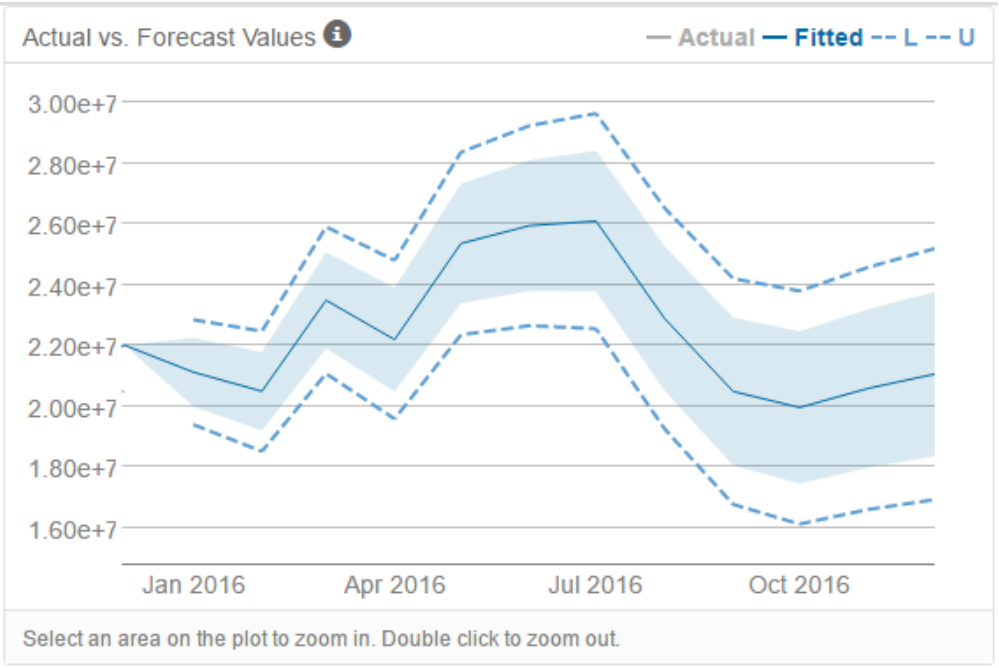
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	-604232.3	1050239	928412	-2.6156	4.0942	0.5463	NA

ARIMA

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
MNM	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA

ETS

Based off the lower (indicating higher accuracy) RMSE and the MASE the ETS (m,n,m) model was chosen to forecast the values for January 2016 through December 2016. The model was then run on the whole set of 46 cases (the 6 holdouts were returned) with the same parameters used for accuracy testing.



Record #	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	2016	1	21136208.135109	22863751.647268	22265788.122301	20006628.147918	19408664.622951
2	2016	2	20506604.689889	22485979.825084	21800848.524632	19212360.855146	18527229.554694
3	2016	3	23506131.457397	25923604.543644	25086832.145154	21925430.769639	21088658.371149
4	2016	4	22207971.238436	24819551.269971	23915591.635728	20500350.841144	19596391.206902
5	2016	5	25376698.322185	28385663.710055	27344155.037671	23409241.6067	22367732.934316
6	2016	6	25963559.446576	29258459.785154	28117978.976999	23809139.916154	22668659.107998
7	2016	7	26113357.20163	29660962.648063	28433011.720628	23793702.682632	22565751.755197
8	2016	8	22904671.917667	26542287.656104	25283181.003148	20526162.832187	19267056.179231
9	2016	9	20499151.00121	24219766.868399	22931930.9538	18066371.048621	16778535.134021
10	2016	10	19970808.947309	23811395.340529	22482033.410444	17459584.484174	16130222.554089
11	2016	11	20602232.29737	24592117.996351	23211078.329341	17993386.2654	16612346.59839
12	2016	12	21072786.922156	25209448.553796	23777604.577977	18367969.266335	16936125.290516

Year	Month	New Stores	Existing Stores
2016	1	2,486,612.72	21,136,208.14
2016	2	2,412,541.73	20,506,604.69
2016	3	2,765,427.23	23,506,131.46
2016	4	2,612,702.50	22,207,971.24
2016	5	2,985,493.92	25,376,698.32
2016	6	3,054,536.41	25,963,559.45
2016	7	3,072,159.67	26,113,357.20
2016	8	2,694,667.28	22,904,671.92
2016	9	2,411,664.82	20,499,151.00
2016	10	2,349,506.93	19,970,808.95

2016	11	2,423,792.03	20,602,232.30
2016	12	2,479,151.40	21,072,786.92