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?

   The optimal number of store formats that was derived was three using the k-means clustering tool. The 85 existing stores for 2015 were used for the analysis. To determine this number the following was done :

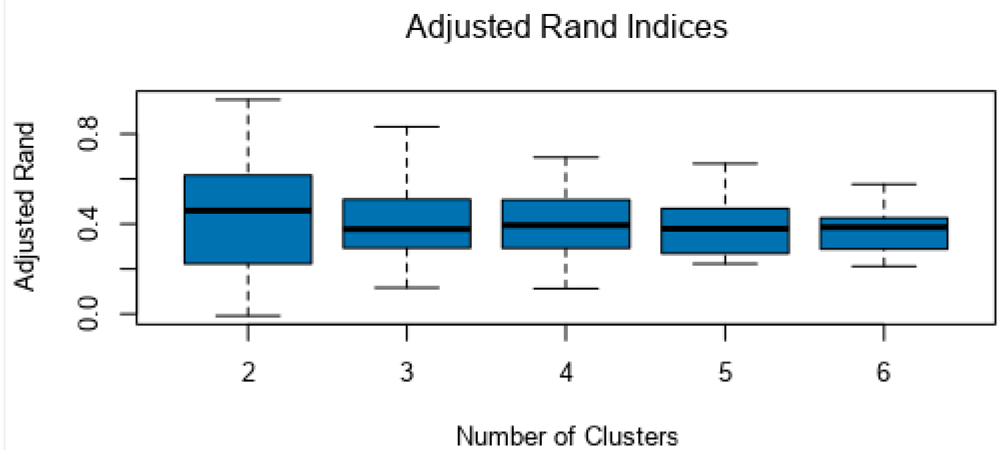
   - The sum of the sales data was calculated by store_id and year.
   - The percentage sales per category per store was used for clustering.

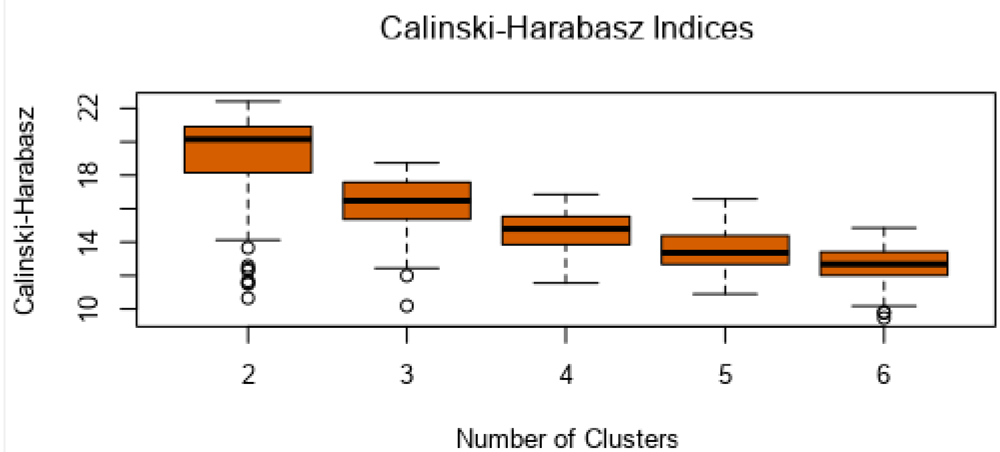
Results - Auto Field (3) - Output

11 of 11 Fields | 85 records displayed | Search | Data | Metadata

Record	Store	Year	Percent_dry_grocery	Percent_diary	Percent_frozen	Percent_meat	Percent_produce	Percent_floral	Percent_deli	Percent_bakery	Percent_general
1	S0001	2015	0.4613	0.1031	0.0772	0.1077	0.0972	0.0068	0.0435	0.0355	0.0677
2	S0002	2015	0.4575	0.1064	0.0788	0.1149	0.1013	0.0074	0.0398	0.0297	0.0641
3	S0003	2015	0.4213	0.1024	0.069	0.1147	0.1254	0.0096	0.0418	0.0361	0.0797

The clustering diagnostics tool was used on all of the 9 predictor variables shown above.





The result is Adjust Rand and CH indices.
The box plots above indicate that 3 numbers of clusters is the most suitable. Number 2 has too many outliers and the box plot for Adjust Rand is very wide. Number 2 also has many outliers. The CH indices boxplots show Number 3 has the largest median from the rest. Also, the Adjusted Rand shows a narrow boxplot for 3.

2. How many stores fall into each store format?

The summary report is show below with the number of stores in each cluster.

Cluster Information:

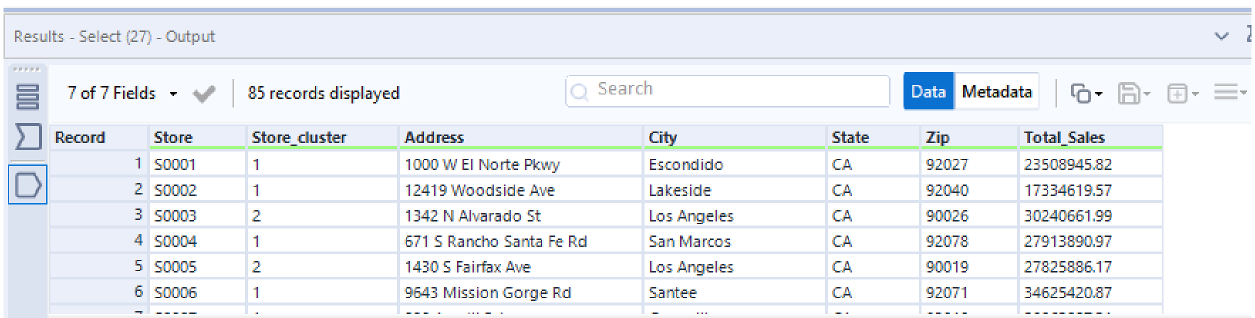
Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.100598	4.823985	2.193986
2	35	2.475232	4.410756	1.9441
3	25	2.287649	3.582763	1.723182

Convergence after 8 iterations.

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

Based on the results of the clustering model, there are more stores in cluster 2 than cluster 1 or cluster 3. The average distance of the datapoints representing the stores is also higher for cluster 2. The assumption would be that cluster 2 will also have the largest proportion of the new stores.

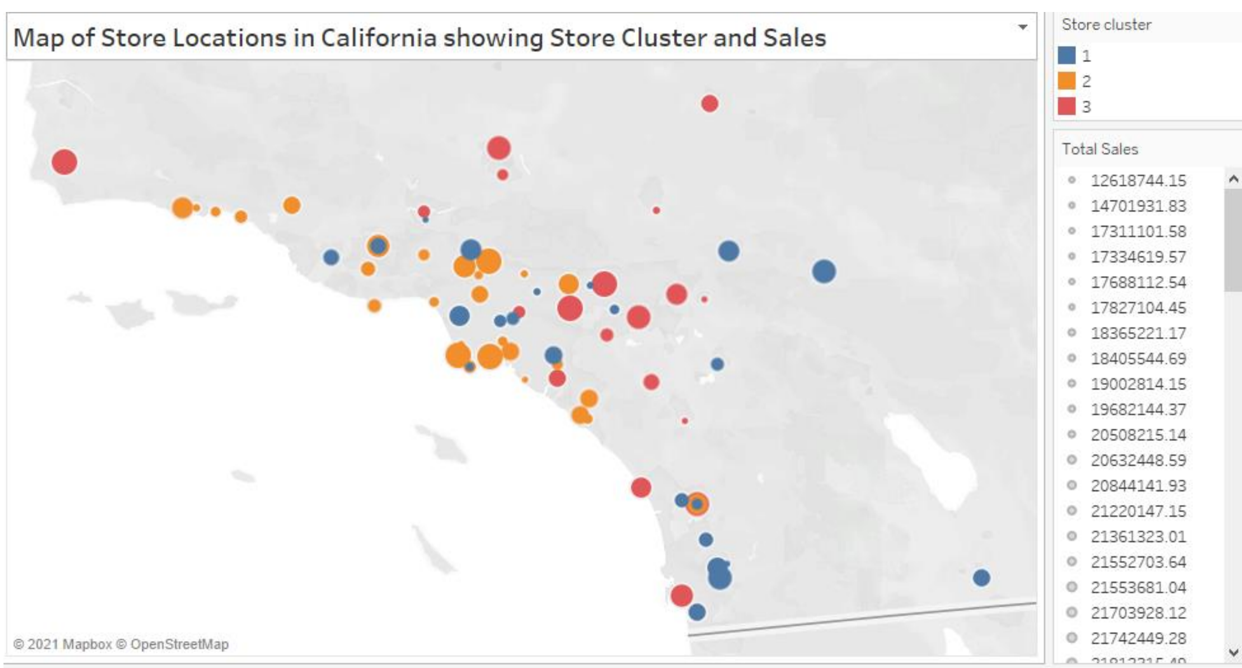
4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.


Results - Select (27) - Output

7 of 7 Fields | 85 records displayed | Search | Data | Metadata

Record	Store	Store_cluster	Address	City	State	Zip	Total_Sales
1	S0001	1	1000 W El Norte Pkwy	Escondido	CA	92027	23508945.82
2	S0002	1	12419 Woodside Ave	Lakeside	CA	92040	17334619.57
3	S0003	2	1342 N Alvarado St	Los Angeles	CA	90026	30240661.99
4	S0004	1	671 S Rancho Santa Fe Rd	San Marcos	CA	92078	27913890.97
5	S0005	2	1430 S Fairfax Ave	Los Angeles	CA	90019	27825886.17
6	S0006	1	9643 Mission Gorge Rd	Santee	CA	92071	34625420.87

The table shows the cluster assignment to each store and the total sales which was imported into tableau for visualization.



The map of store location in California indicate the store cluster 2 is along the shore more than cluster 1 and cluster 3.

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.)

Model Comparison Report

Fit and error measures

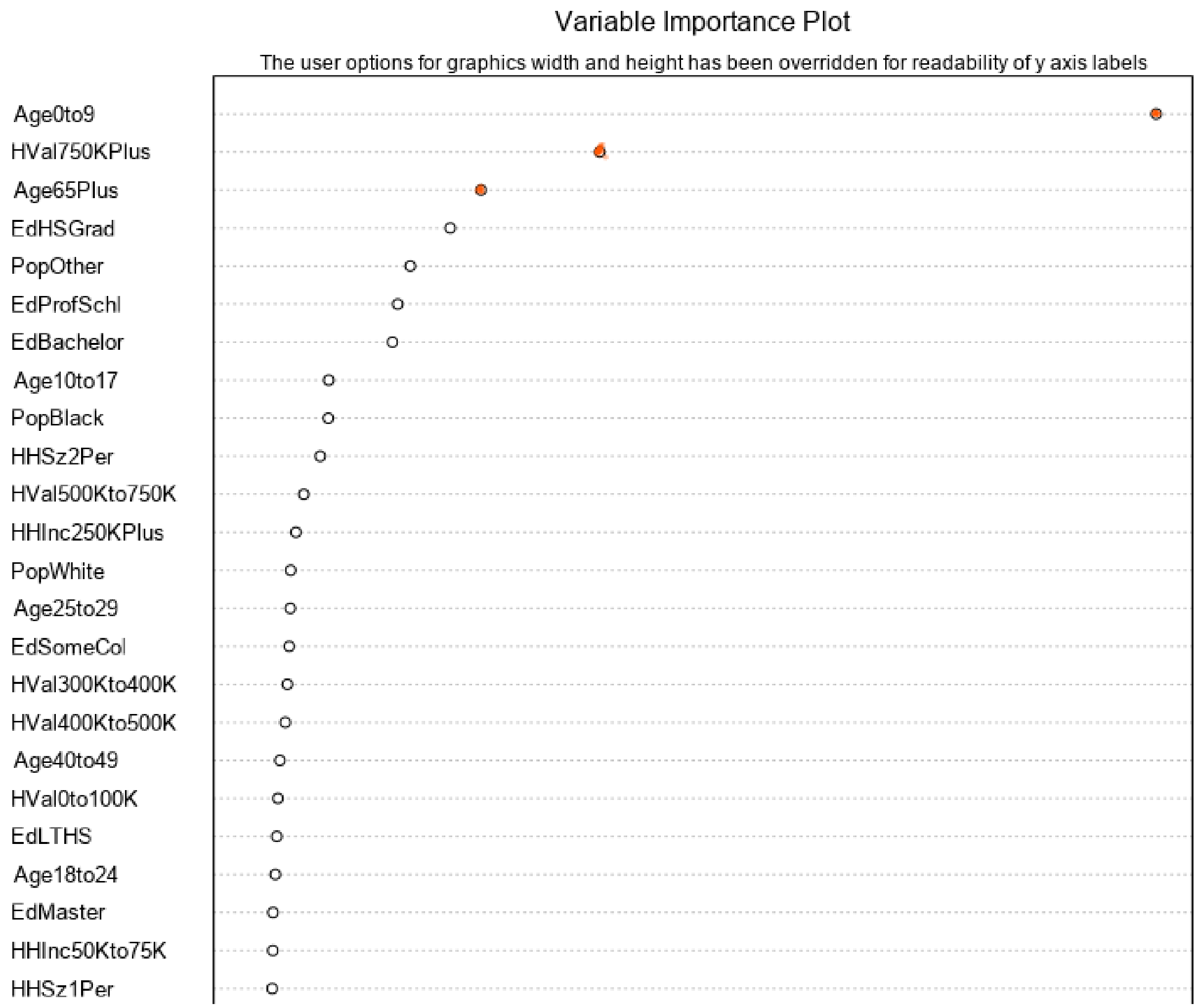
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_cluster	0.7059	0.7083	0.6250	1.0000	0.5000
FM_store_cluster	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted	0.7647	0.8333	0.5000	1.0000	1.0000

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Decision Tree , Forest Model, and Boosted model were used to predict the store format for the new stores. The model best suitable for the prediction is the Boosted mode. It has the highest accuracy of 0.7647 and F1 score of 0.8333.

2. What are the three most important variables that help explain the relationship between demographic indicators and store formats?



The variable importance plot from the boosted model shows the three variables Age0to9, HVal750KPlus, and Age65Plus (points in red) as the most important variables to explain the demographic indicators and store formats. It seems that age and wealth are important variables for predicting store formats.

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

Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2

S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

Task 3: Predicting Produce Sales

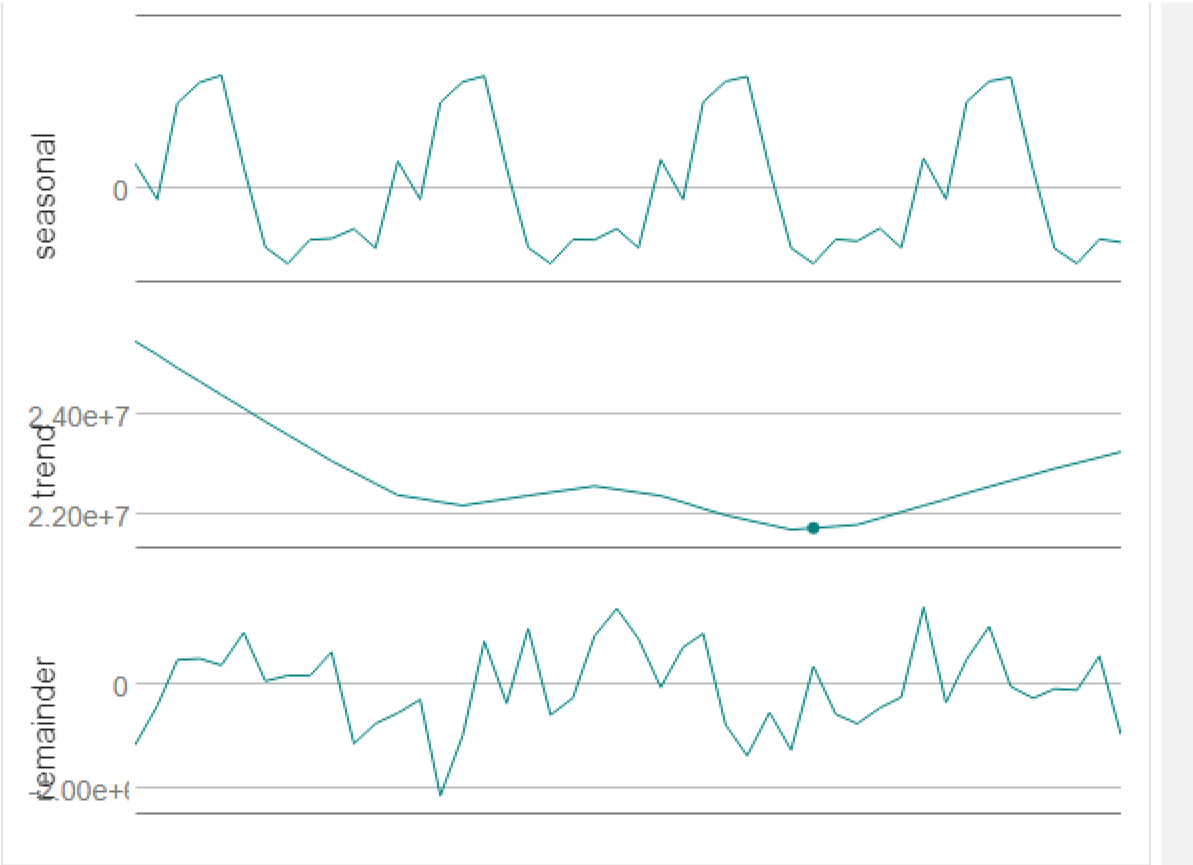
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

STEP 1: FORECAST PRODUCE SALES OF THE 85 EXISTING STORES

To forecast the full year of 2016 for produce sales the ETS(MNM) was used.

ETS(MNM)

To get MNM , the decomposition graphs of error, time series, and seasonal series were plotted.

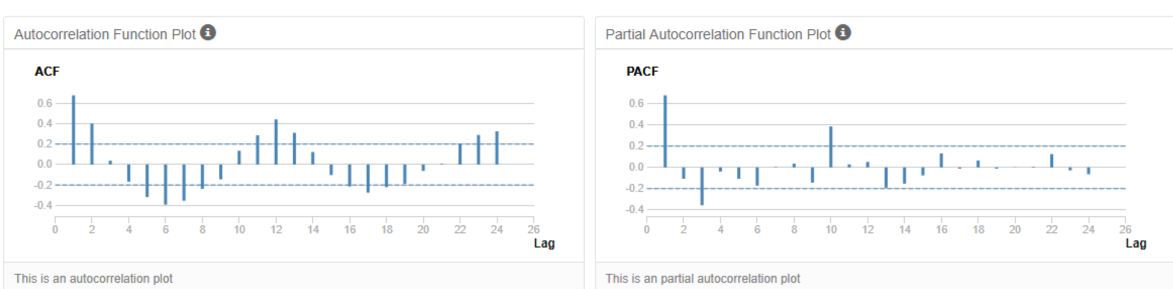


This is a decomposition plot

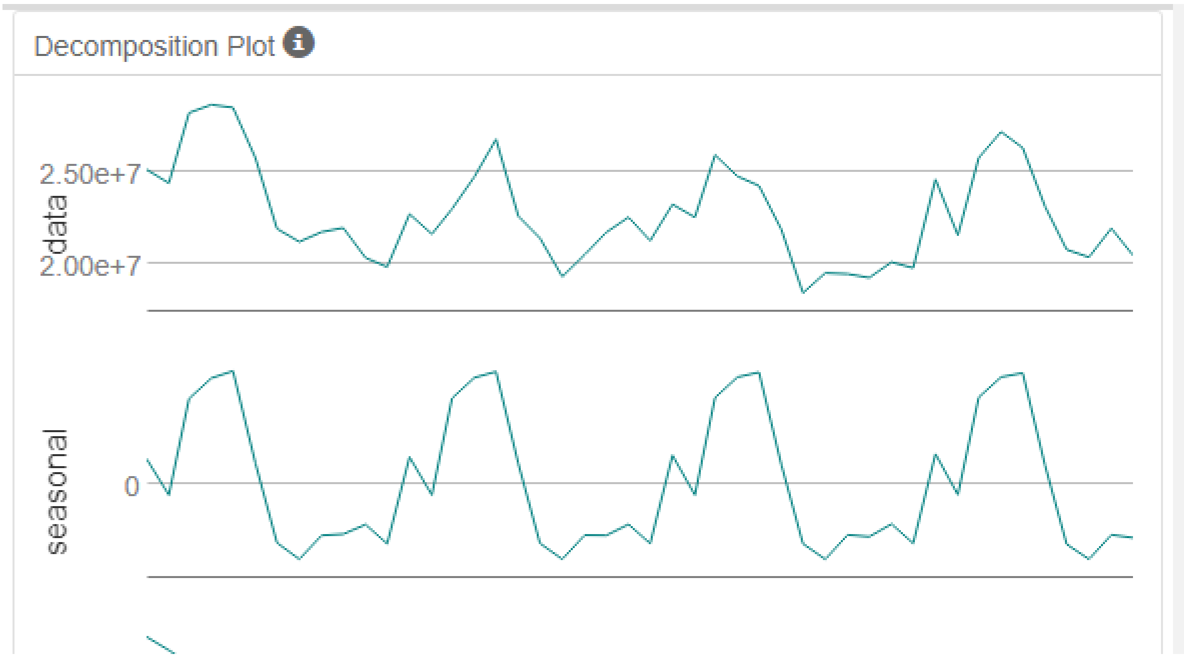
Since the remainder plot shows strong up and down spikes the multiplicative (M) was used for error. There is no trend suggesting None (N). The seasonal graph shows a pattern and a slow decline suggesting multiplicative (M)

ARIMA(1,0,0)(1,1,0)12

To calculate p,d,q,P,D,Q for ARIMA the following graphs were used.



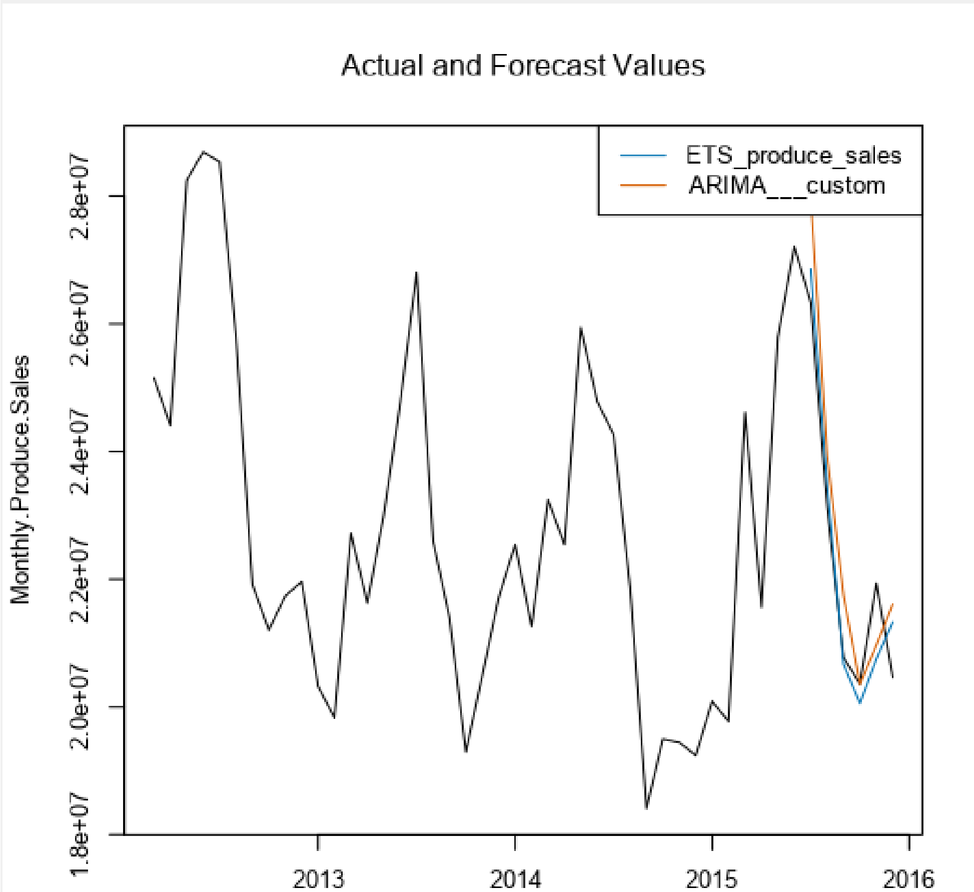
The strong positive correlation at lag-1 for both ACF and PACF graphs suggest AR.



The time plot suggests that differencing is required to stabilize around 0 and the seasonal plot suggests that ARIMA has a seasonal component.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_produce_sales	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA___custom	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

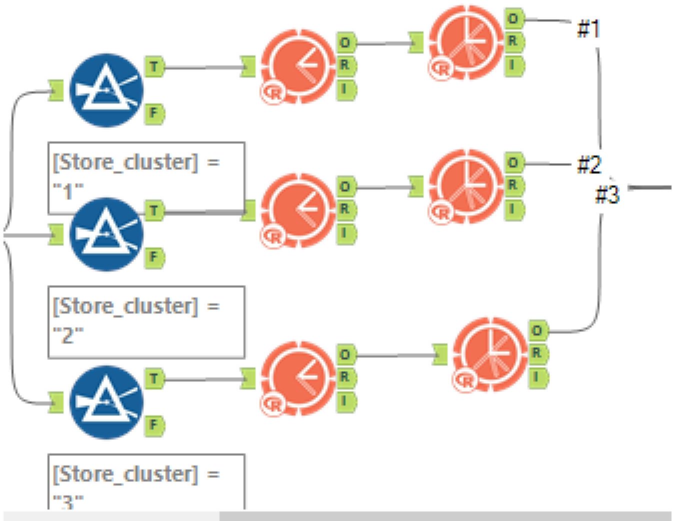


ETS (MNM) was used since it has overall MASE and RMSE errors compared to ARIMA.

STEP 2: FORECAST SALES OF THE 10 NEW STORES

To forecast produce sales for new stores the following was done:

The average produce sales for the existing stores for each segment was forecasted by month. The ETS(MNM) model was used on each segment.



To get the forecast for the new stores the average store produce sales forecast was multiplied by the number of new stores in that segment.

Output Column	Data Preview
ForecastAvgSumProduce	218778.52

```
IF [store_cluster] = '1'  THEN [ForecastAvgSumProduce] * 1
ELSEIF [store_cluster] = '2' THEN [ForecastAvgSumProduce] * 6
ELSE [ForecastAvgSumProduce] * 3 ENDIF
```

Then sum of forecasts were summed for each segment to get overall forecast for new stores.

2 of 2 Fields 12 records displayed

Record	Month	Sales_all_stores
1	2016-01	2563357.93
2	2016-02	2483924.76
3	2016-03	2910944.20
4	2016-04	2764881.89
5	2016-05	3141305.92
6	2016-06	3195054.23
7	2016-07	3212390.98
8	2016-08	2852385.83
9	2016-09	2521697.18
10	2016-10	2466750.92
11	2016-11	2557744.62
12	2016-12	2530510.81

STEP 3: SUM FORECASTS OF THE EXISTING AND NEW STORES TOGETHER

Results - Formula (30) - Output

4 of 4 Fields | 12 records displayed | Search | Data | Metadata

Record	Month	sales_forecast_existing	sales_forecast_new	Total_sales_forecast
1	2016-01	21829060.03	2563357.93	24392417.96
2	2016-02	21146329.63	2483924.76	23630254.39
3	2016-03	23735686.94	2910944.2	26646631.14
4	2016-04	22409515.28	2764881.89	25174397.17
5	2016-05	25621828.73	3141305.92	28763134.65
6	2016-06	26307858.04	3195054.23	29502912.27
7	2016-07	26705092.56	3212390.98	29917483.54
8	2016-08	23440761.33	2852385.83	26293147.16
9	2016-09	20640047.32	2521697.18	23161744.5
10	2016-10	20086270.46	2466750.92	22553021.38
11	2016-11	20858119.96	2557744.62	23415864.58
12	2016-12	21255190.24	2530510.81	23785701.05

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.

TABLE WITH FORECASTED PRODUCE SALES FOR NEW AND EXISTING STORES

Month	New Stores	Existing Stores
2016-01	2563357.93	21829060.03
2016-02	2483924.76	21146329.63
2016-03	2910944.2	23735686.94
2016-04	2764881.89	22409515.28
2016-05	3141305.92	25621828.73
2016-06	3195054.23	26307858.04
2016-07	3212390.98	26705092.56
2016-08	2852385.83	23440761.33
2016-09	2521697.18	20640047.32
2016-10	2466750.92	20086270.46
2016-11	2557744.62	20858119.96
2016-12	2530510.81	21255190.24