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?**

   There are three optimal number of store formats. To determine that number we have used k-centroid diagnostic method. First, I have prepared the data by using existing store data and the given conditions. Then we used the data as an input of k-centroid diagnostic for k-means, which is given in project detail. We get following result :

Report

K-Means Cluster Assessment Report

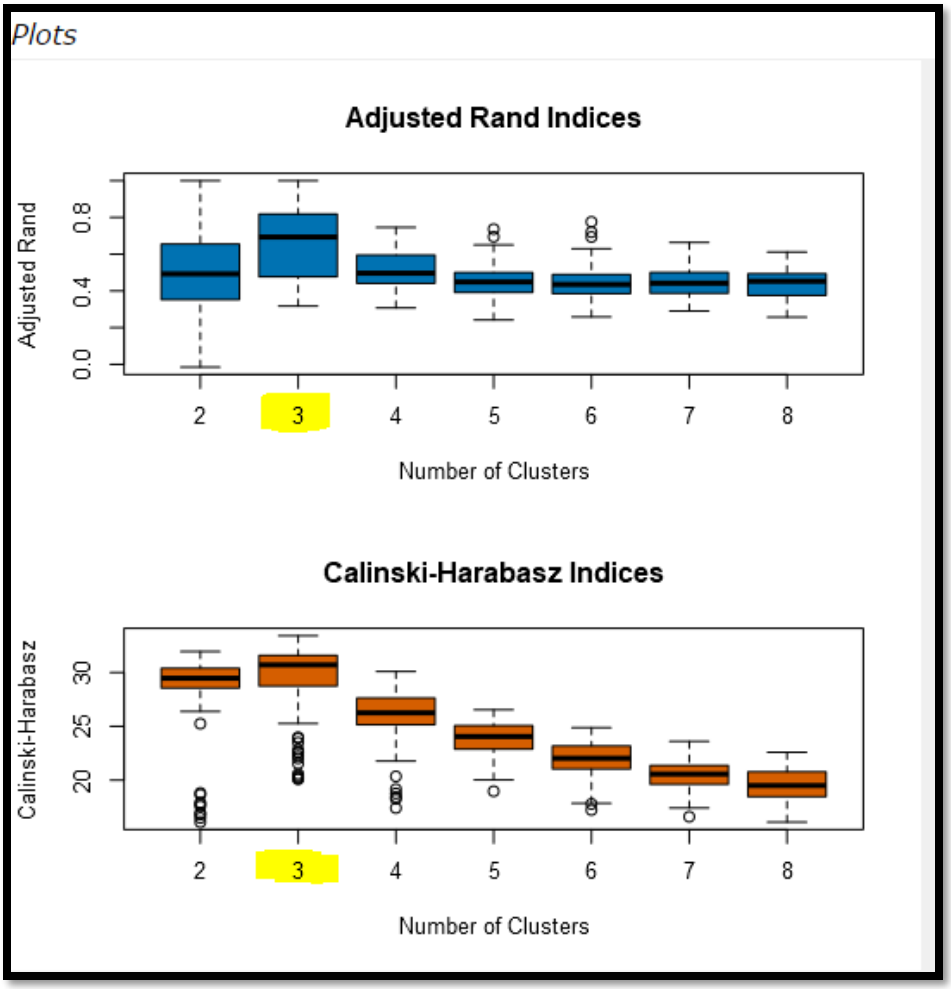
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.0152	0.3171	0.3072	0.2412	0.2586	0.2903	0.2568
1st Quartile	0.352	0.4819	0.4431	0.3943	0.3896	0.3877	0.377
Median	0.4926	0.6936	0.4964	0.4487	0.4348	0.4417	0.4526
Mean	0.484	0.6575	0.5125	0.4623	0.4532	0.4498	0.4411
3rd Quartile	0.655	0.816	0.5913	0.4982	0.489	0.4997	0.491
Maximum	1	1	0.7458	0.7366	0.7762	0.6637	0.6118

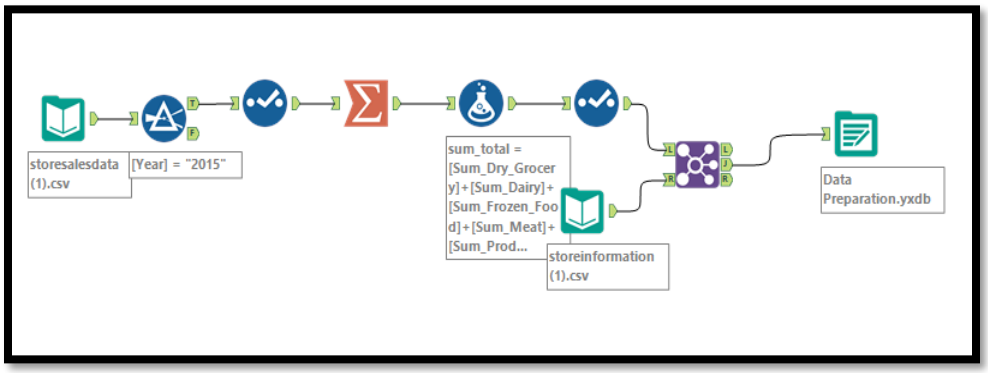
Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.1	20.09	17.41	18.98	17.24	16.61	16.11
1st Quartile	28.61	28.76	25.16	22.91	21.05	19.61	18.46
Median	29.47	30.7	26.25	24.05	22.02	20.56	19.5
Mean	28.41	29.47	25.99	23.88	21.96	20.48	19.62
3rd Quartile	30.39	31.58	27.62	25.06	23.14	21.35	20.77
Maximum	31.95	33.41	30.09	26.53	24.87	23.6	22.59

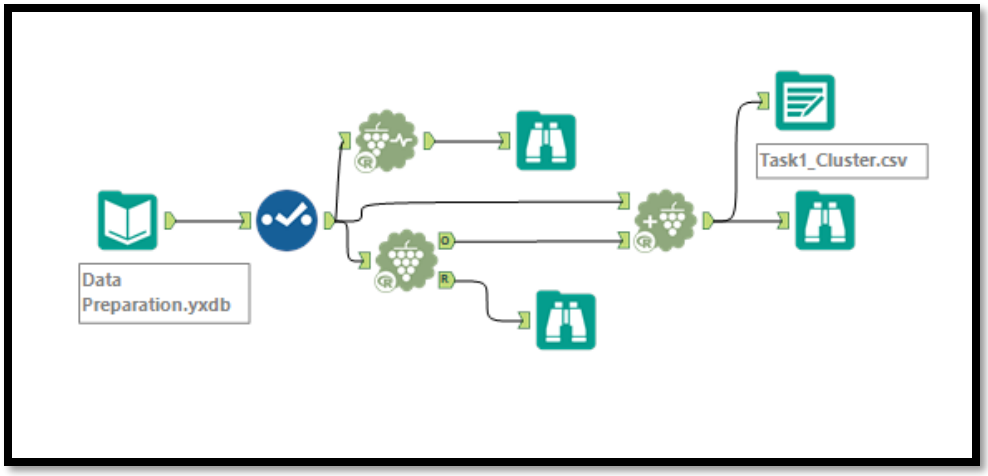


AR and CH indices are used to determine the optimal method and number of clusters. The higher the median and smaller the variation the better. As per the above box and whisker plot, three number of clusters has the highest median. (AR-0.6936, CH-30.7) Therefore, we will choose three number of clusters for further analysis.

Data preparation workflow:



Cluster Analysis



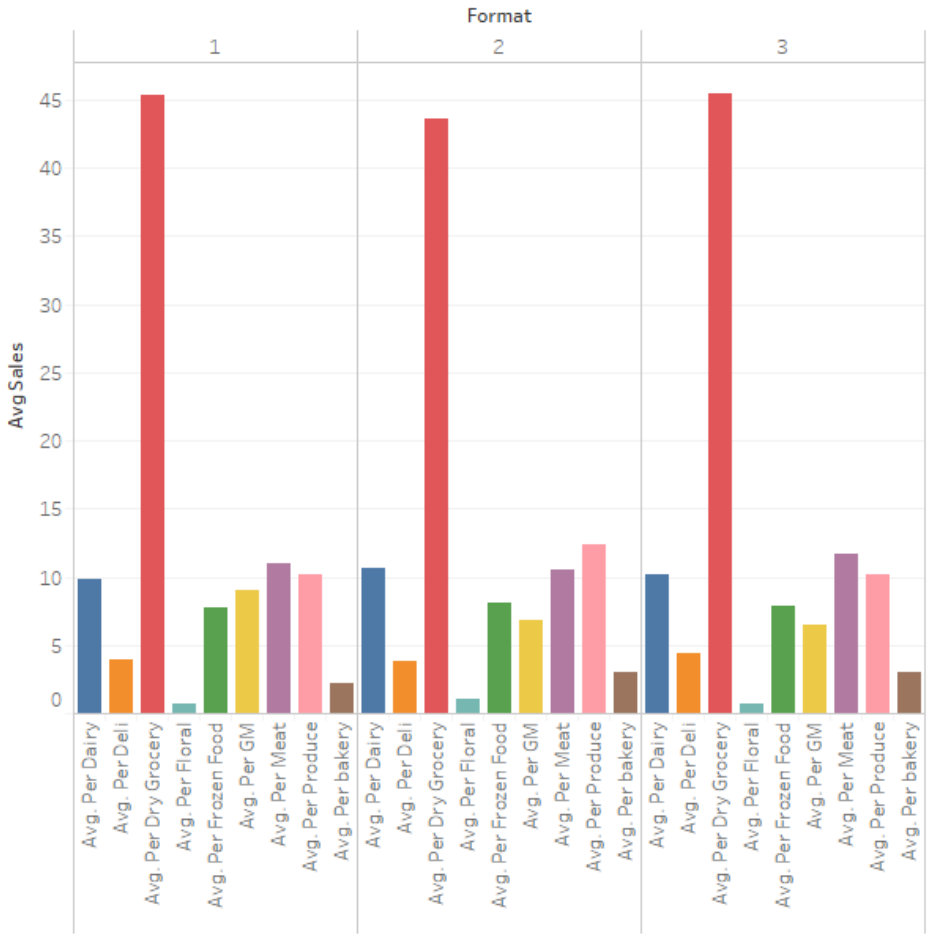
2. **How many stores fall into each store format?**

   After using K-Centroid cluster analysis tool of Alteryx, we get following number of stores(size) in each store format(Cluster)

Cluster	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?**

   Suppose, If we compare Average sales category wise for each category then as per result we can say that 33 stores will sell more than 29 and 23 stores. However, if we take look at cluster result, then category dairy, frozen food, produce, floral and bakery perform well in format 2.

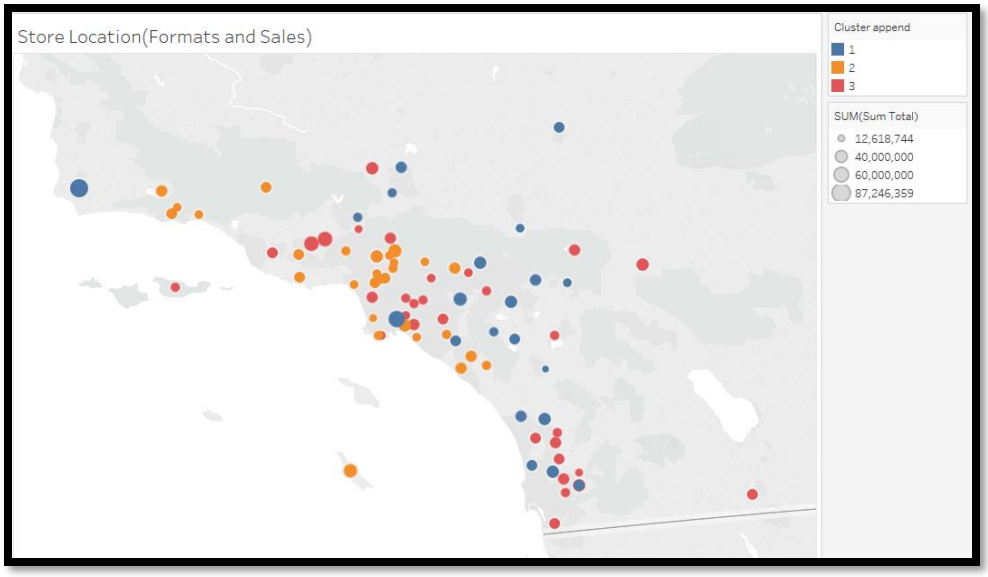


	Per_Dry_Grocery	Per_Dairy	Per_Frozen_Food	Per_Meat	Per_Produce	Per_Floral	Per_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Per_bakery	Per_GM
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

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.

[Task 1]



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.)
   The goal is to determine store formats, which is non-binary classification problem. Therefore, we will test three different models (Forest Model, Decision Tree model, Boosted Model) and choose best of them. After model comparison, we get below output. According to result, all the model has same accuracy. However, F1 is slightly higher for boosted model. So, we will take it for predicting the outcome.

Model Comparison Report

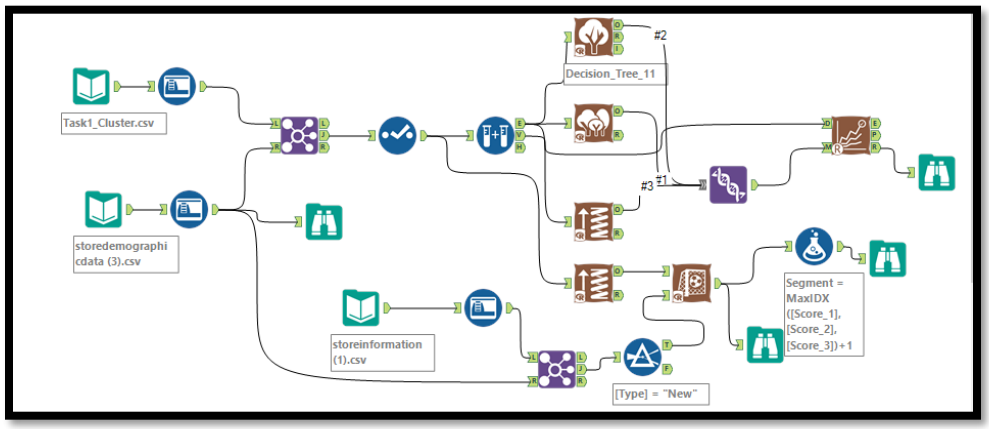
Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_	0.8235	0.8251	0.7500	0.8000	0.8750
Decision_Tree_	0.8235	0.8251	0.7500	0.8000	0.8750
Boosted	0.8235	0.8543	0.8000	0.6667	1.0000

2. 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	1
S0092	2
S0093	1
S0094	2
S0095	2

Alteryx Workflow:



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?

Existing Store Forecast:

ETS is the best model for existing store.

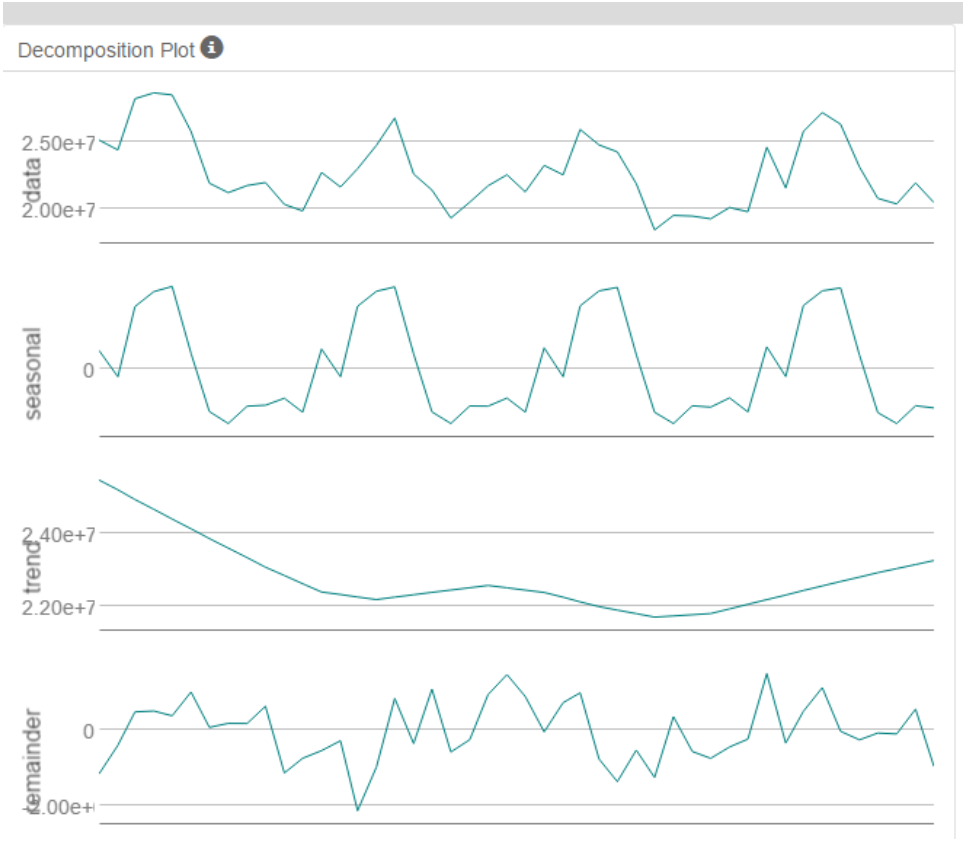
Building and Choosing Model:

Below is the Decomposition plot. As shown in graph seasonal, trend and remainder represent seasonality, trend and error respectively. The graph represents time(X-Axis) V/S monthly sales(Y-Axis).

Trend: From the trend, we can say that it is not linear nor exponential.

Seasonality: The seasonal portion shows that the regularly occurring spike in sales each year changes in magnitude, even so slightly rather than being constant. In Alteryx, we will need to hover our mouse over the seasonal graph in Interface mode to be able to see that the seasonal numbers are slightly increasing. This is important because having seasonality suggests that any ARIMA models used for analysis will need seasonal differencing. The change in magnitude suggests that any ETS models will use a multiplicative method in the seasonal component.

Error: The remainder plot is fluctuating between large and small errors over time.



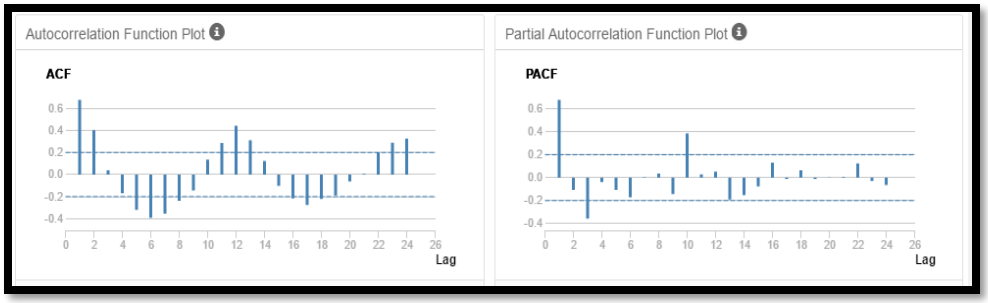
As per above observation our ETS term would be (M,N,M).

For selecting ARIMA term, we will observe ACF and PACF as below. According to plots, we can say that seasonal first difference of the series has removed most of the significant lags from the ACF and PACF. Therefore, the differencing terms will be d(1) and D(1).

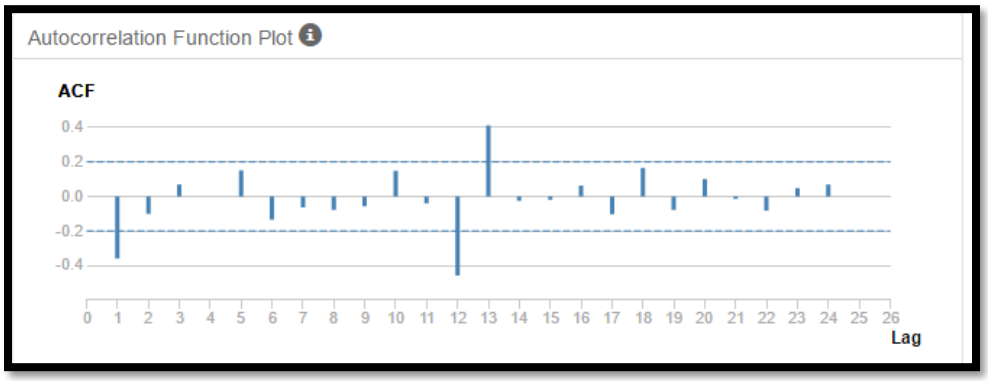
For, non-seasonal AR and MA term, the ACF plot shows a strong negative correlation at lag 1 which is confirmed in the PACF. This suggests an MA(1) model or q(1) since there is only 1 significant lag and AR(0) model or p(0).

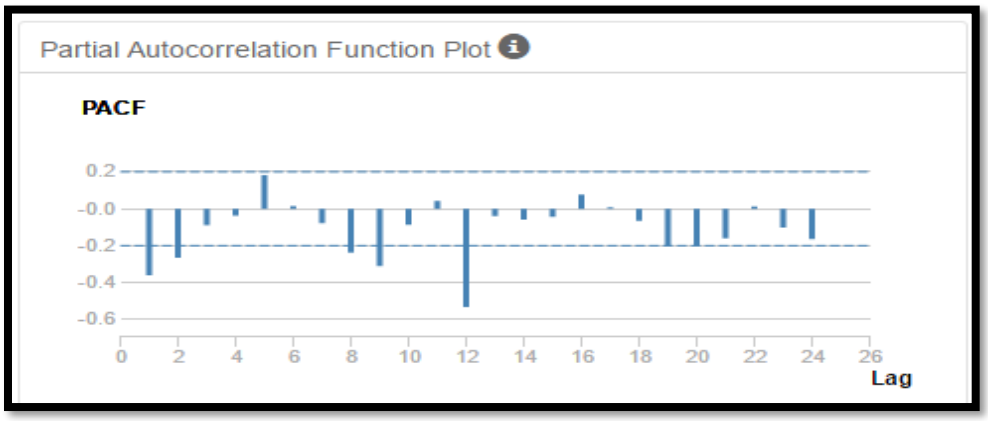
For seasonal component, AR(0), MA(1) as ACF is negative at lag 1, 12 , which is confirmed in PACF. M =12, which indicate seasonal periods which is 12 months in this case. So ARIMA model will be (0,1,0) (0,1,0) (12).

<u>ACF PACF plots before any differencing:</u>



ACF PACF plots after first differencing:





Model Comparison:

After model comparison for holdout sample, we get following result.

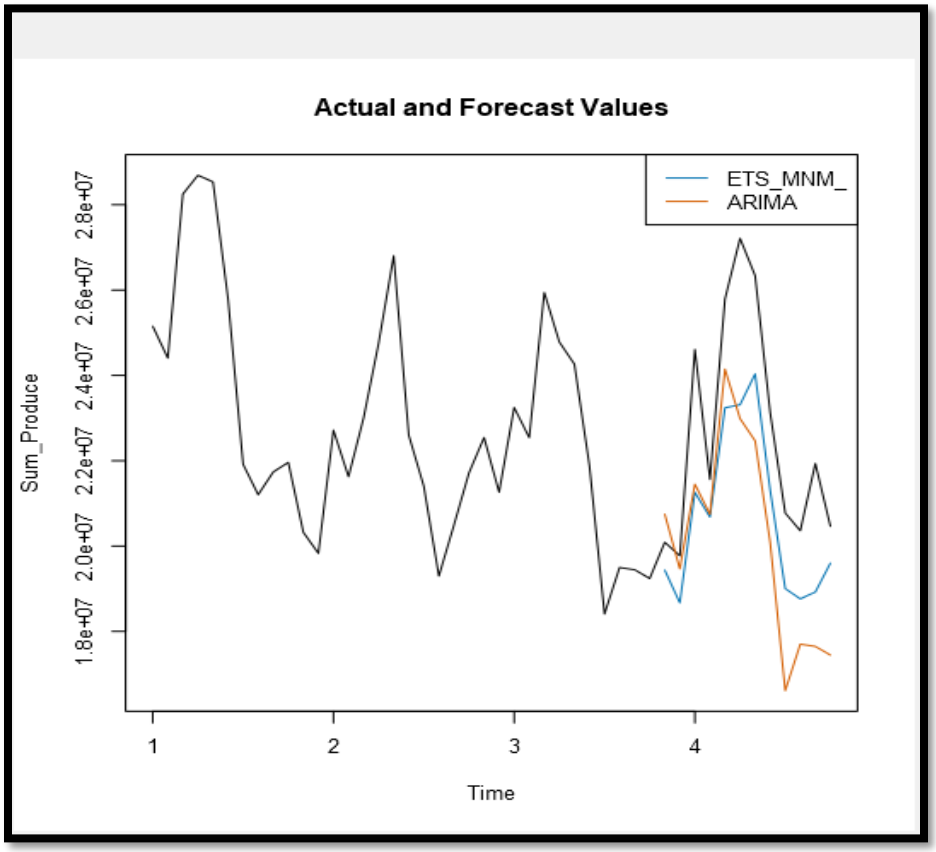
According to result, we can say that MASE(1.2691) of ETS model is lower than MASE(1.6988) of ARIMA. And MASE of ETS is closer to zero than MASE of ARIMA.

Apart from that RMSE of ETS (2226513) is also lower than RMSE of ARIMA(2999244), which indicates it has a smaller standard deviation from the mean.

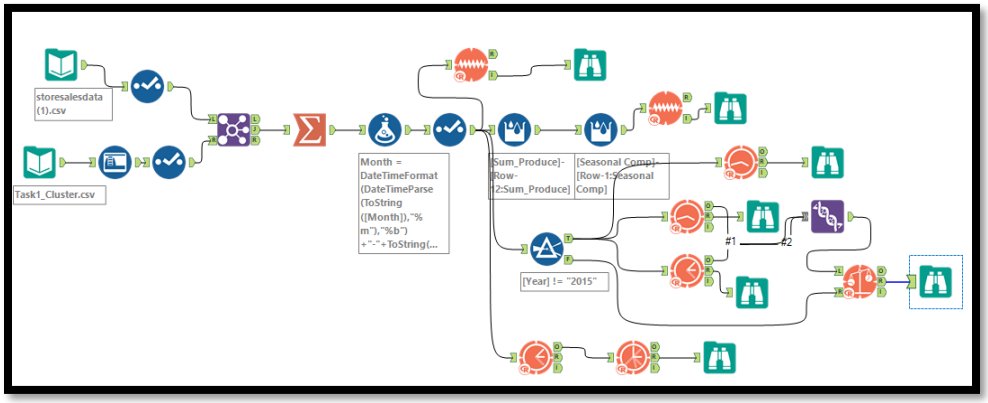
In conclusion, smaller the value of MASE and RMSE, better the model is. Thus, ETS is best model for further forecasting.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_MNM_	1983593	2226513	1983593	8.4729	8.4729	1.2691	NA
ARIMA	2545369	2999244	2655219	11.0071	11.5539	1.6988	NA

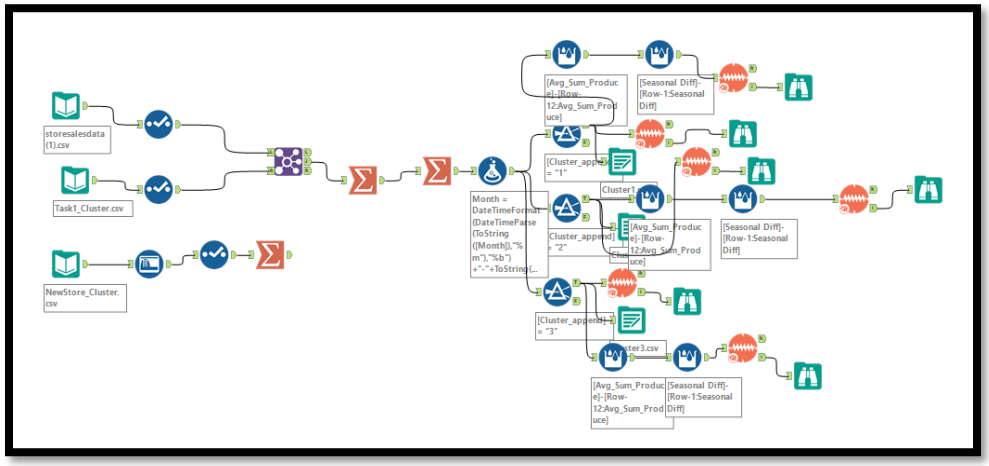


Alteryx Workflow:

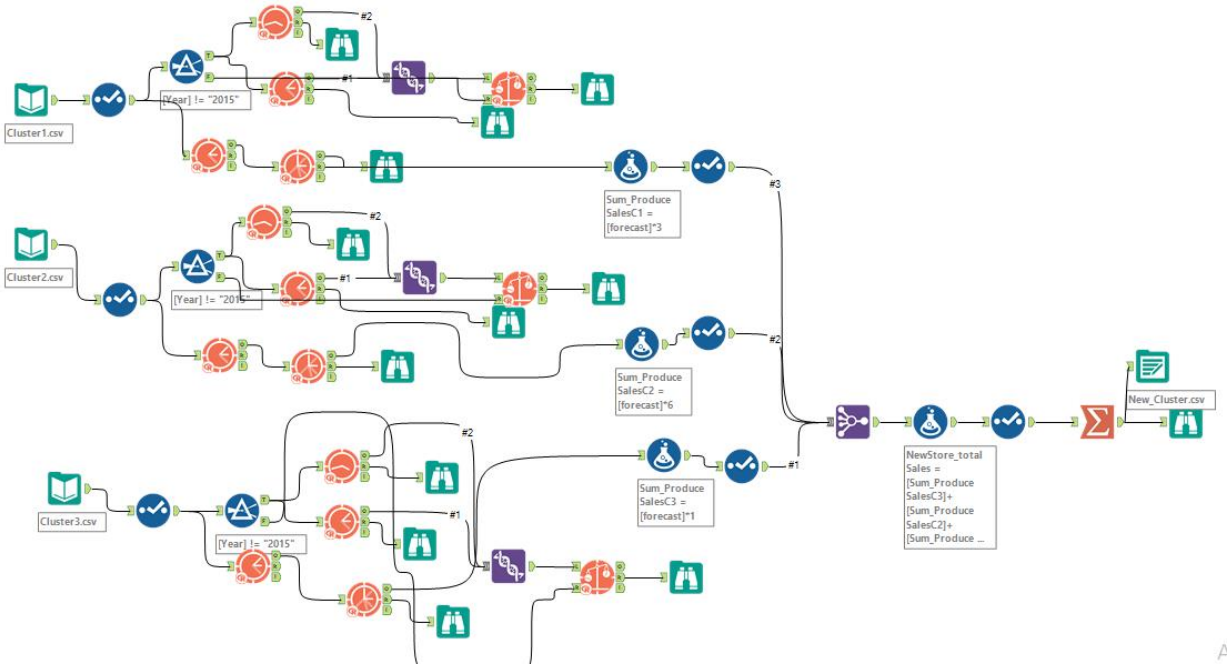


New Store Forecast

First, we prepared dataset in Alteryx. Below is the workflow :



After preparing dataset, we did model comparison on cluster level. Below is the model comparison for each cluster.

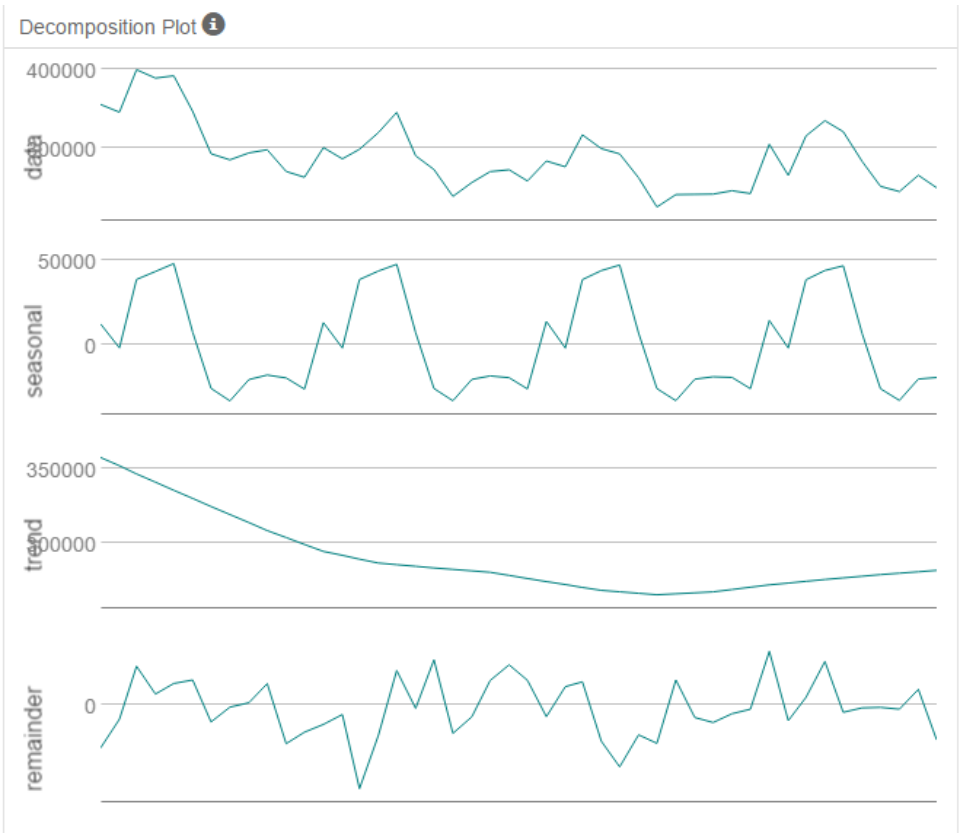


Cluster 1:

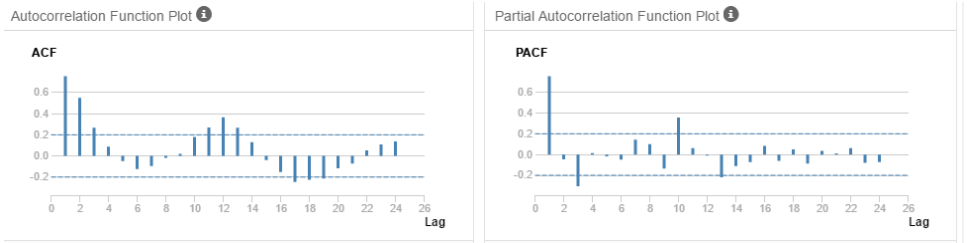
According to decomposition plot below we can say that there is no trend present.

If we hover pointer on seasonal plot, then we can say that there is slight increase or decrease.

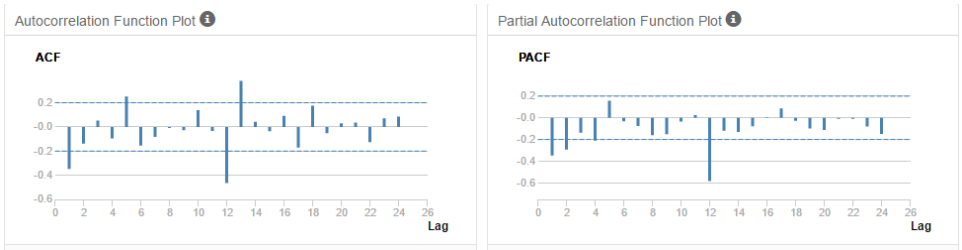
Error is growing and shrinking.



ACF, PACF before any differencing:



ACF, PACF plot after first differencing:



After above observation below are our Model Term :

ETS Model: Error is not constant = Multiplicative, Trend = None, Seasonality is not constant= Multiplicative. So, ETS(M,N,M) will be used.

ARIMA Model: In ACF and PACF plot the seasonal first difference of the series has removed most of the significant lags from the ACF and PACF so there is no need for further differencing. The remaining correlation can be accounted for using autoregressive and moving average terms and the differencing terms will be d(1) and D(1).

For, non-seasonal AR and MA term, the ACF plot shows a strong negative correlation at lag 1 which is confirmed in the PACF. This suggests an MA(1) model or q(1) since there is only 1 significant lag and AR(0) model or p(0).

For seasonal component, AR(0), MA(1) as ACF is negative at lag 1, 12 , which is confirmed in PACF. M =12, which indicate seasonal periods which is 12 months in this case. So ARIMA model will be (0,1,0) (0,1,0) (12).

Model Comparission

After model comparison for holdout sample, we get following result.

According to result, we can say that MASE (0.8895) of ETS model is lower than MASE(1.5886) of ARIMA. And MASE of ETS is closer to zero than MASE of ARIMA.

Apart from that RMSE of ETS (21747.98) is also lower than RMSE of ARIMA(36945.74), which indicates it has a smaller standard deviation from the mean.

In conclusion, smaller the value of MASE and RMSE, better the model is. Thus, ETS is best model for further forecasting.

Accuracy Measures:

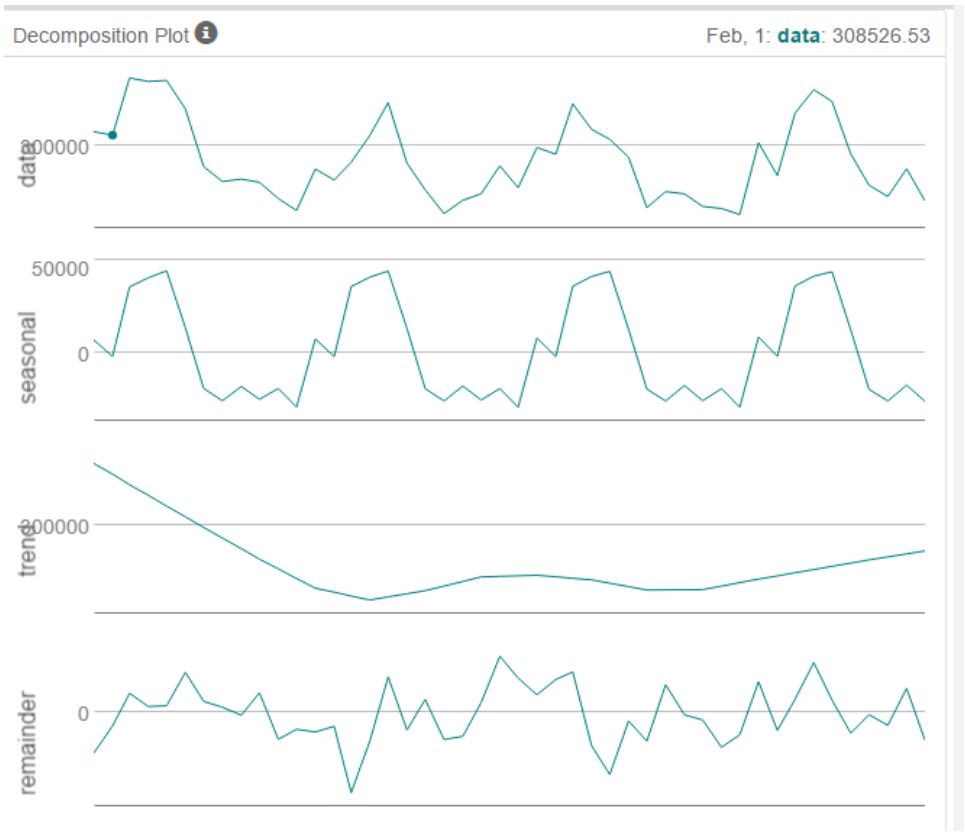
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_MNM_	18175.99	21747.98	18175.99	6.3296	6.3296	0.8895	NA
ARIMA	31784.36	36945.74	32460.52	11.1886	11.4635	1.5886	NA

Cluster 2:

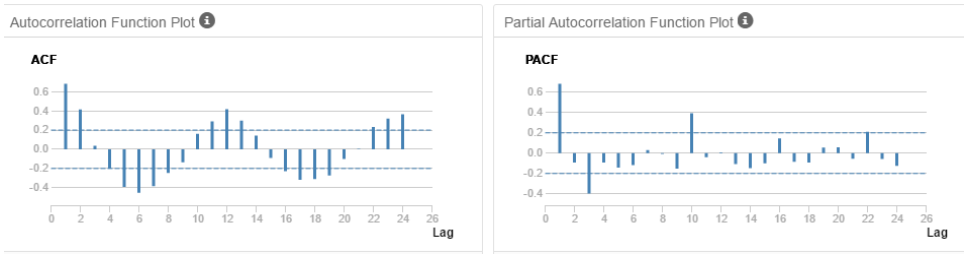
According to decomposition plot below we can say that there is no trend present.

If we hover pointer on seasonal plot, then we can say that there is slight increase or decrease.

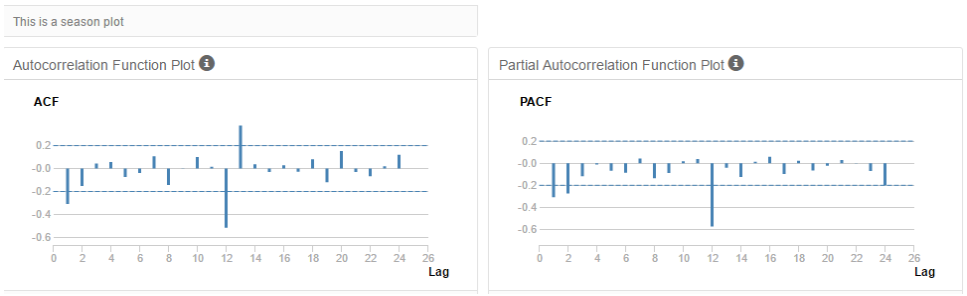
Error is growing and shrinking.



ACF, PACF plot after any differencing:



ACF, PACF plot after first differencing:



After above observation below are our Model Term :

ETS Model: Error is not constant = Multiplicative, Trend = None, Seasonality is not constant= Multiplicative. So, ETS(M,N,M) will be used.

ARIMA Model: In ACF and PACF plot the seasonal first difference of the series has removed most of the significant lags from the ACF and PACF so there is no need for further differencing. The remaining correlation can be accounted for using autoregressive and moving average terms and the differencing terms will be d(1) and D(1).

For, non-seasonal AR and MA term, the ACF plot shows a strong negative correlation at lag 1 which is confirmed in the PACF. This suggests an MA(1) model or q(1) since there is only 1 significant lag and AR(0) model or p(0).

For seasonal component, AR(0), MA(1) as ACF is negative at lag 1, 12 , which is confirmed in PACF. M =12, which indicate seasonal periods which is 12 months in this case. So ARIMA model will be (0,1,0) (0,1,0) (12).

Model Comparission

After model comparison for holdout sample, we get following result.

According to result, we can say that MASE (0.6017) of ETS model is lower than MASE (0.9322) of ARIMA. And MASE of ETS is closer to zero than MASE of ARIMA.

Apart from that RMSE of ETS (14117.92) is also lower than RMSE of ARIMA(20113.41), which indicates it has a smaller standard deviation from the mean.

In conclusion, smaller the value of MASE and RMSE, better the model is. Thus, ETS is best model for further forecasting.

Accuracy Measures:

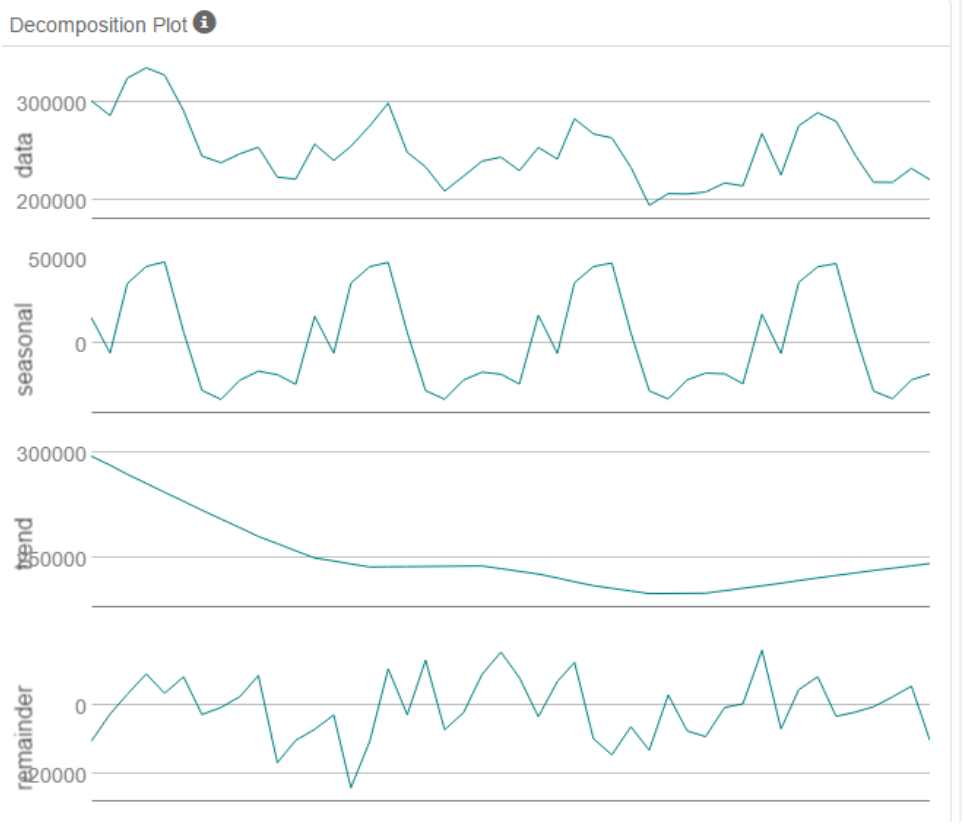
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_MNM_	8413.646	14117.92	10698.37	2.6434	3.5428	0.6017	NA
ARIMA	5360.661	20113.41	16574.92	1.4082	5.7532	0.9322	NA

Cluster 3:

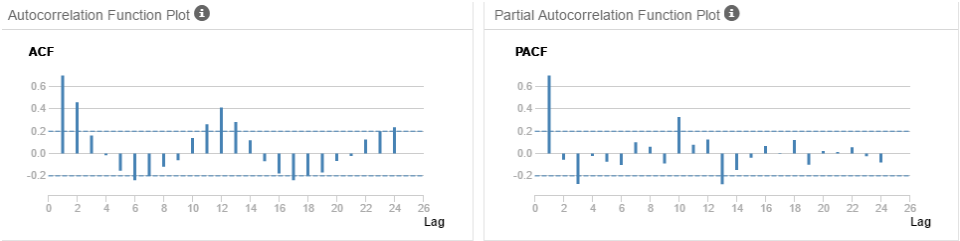
According to decomposition plot below we can say that there is no trend present.

If we hover pointer on seasonal plot, then we can say that there is slight increase or decrease.

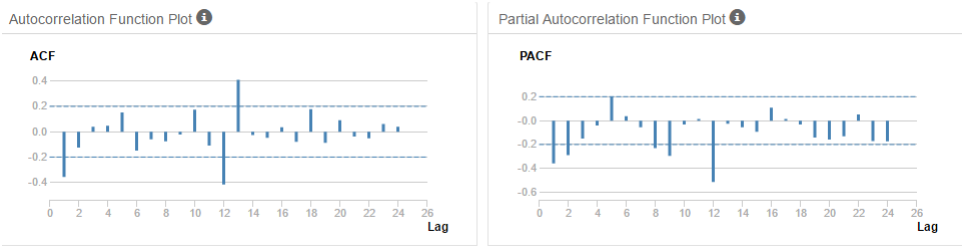
Error is growing and shrinking.



ACF, PACF plot after any differencing:



ACF, PACF plot after first differencing:



After above observation below are our Model Term :

ETS Model: Error is not constant = Multiplicative, Trend = None, Seasonality is not constant= Multiplicative. So, ETS(M,N,M) will be used.

ARIMA Model: In ACF and PACF plot the seasonal first difference of the series has removed most of the significant lags from the ACF and PACF so there is no need for further differencing. The remaining correlation can be accounted for using autoregressive and moving average terms and the differencing terms will be d(1) and D(1).

For, non-seasonal AR and MA term, the ACF plot shows a strong negative correlation at lag 1 which is confirmed in the PACF. This suggests an MA(1) model or q(1) since there is only 1 significant lag and AR(0) model or p(0).

For seasonal component, AR(0), MA(1) as ACF is negative at lag 1, 12 , which is confirmed in PACF. M =12, which indicate seasonal periods which is 12 months in this case. So ARIMA model will be (0,1,0) (0,1,0) (12).

Model Comparission

After model comparison for holdout sample, we get following result.

According to result, we can say that MASE (1.204) of ETS model is lower than MASE (1.6146) of ARIMA. And MASE of ETS is closer to zero than MASE of ARIMA.

Apart from that RMSE of ETS (25139.78) is also lower than RMSE of ARIMA(35296.92), which indicates it has a smaller standard deviation from the mean.

In conclusion, smaller the value of MASE and RMSE, better the model is. Thus, ETS is best model for further forecasting.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_MNM_	23290.6	25139.78	23290.6	9.4038	9.4038	1.204	NA
ARIMA	30980.96	35296.92	31233.89	12.7083	12.8248	1.6146	NA

After choosing ETS(M,N,M) as best model, we get following forecast result.

Month	New Store	Existing Store	Total Store
Jan-16	2661149.648	21539936.01	24201085.66
Feb-16	2587168.908	20413770.6	23000939.51
Mar-16	2840241.719	24325953.1	27166194.82
Apr-16	2778280.21	22993466.35	25771746.56
May-16	2999553.997	26691951.42	29691505.42
Jun-16	3023860.398	26989964.01	30013824.41
Jul-16	3045103.291	26948630.76	29993734.06
Aug-16	2850318.494	24091579.35	26941897.84
Sep-16	2649579.294	20523492.41	23173071.7
Oct-16	2618296.682	20011748.67	22630045.35
Nov-16	2679615.951	21177435.49	23857051.44
Dec-16	2636991.297	20855799.11	23492790.41

2. Please provide a Tableau Dashboard (saved as a Tableau Public file) that includes a table and a plot of the three monthly forecasts; one for existing, one for new, and one for all stores. Please name the tab in the Tableau file "Task 3".

Task 3