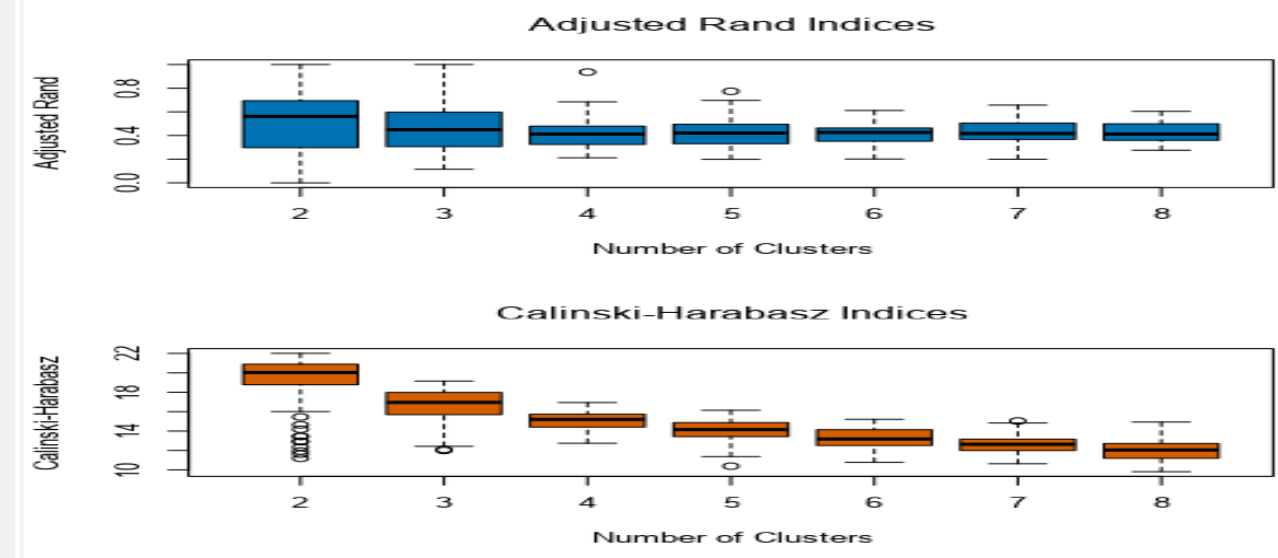
Project: Predictive Analytics Capstone

By Bipin Kumar Sultania

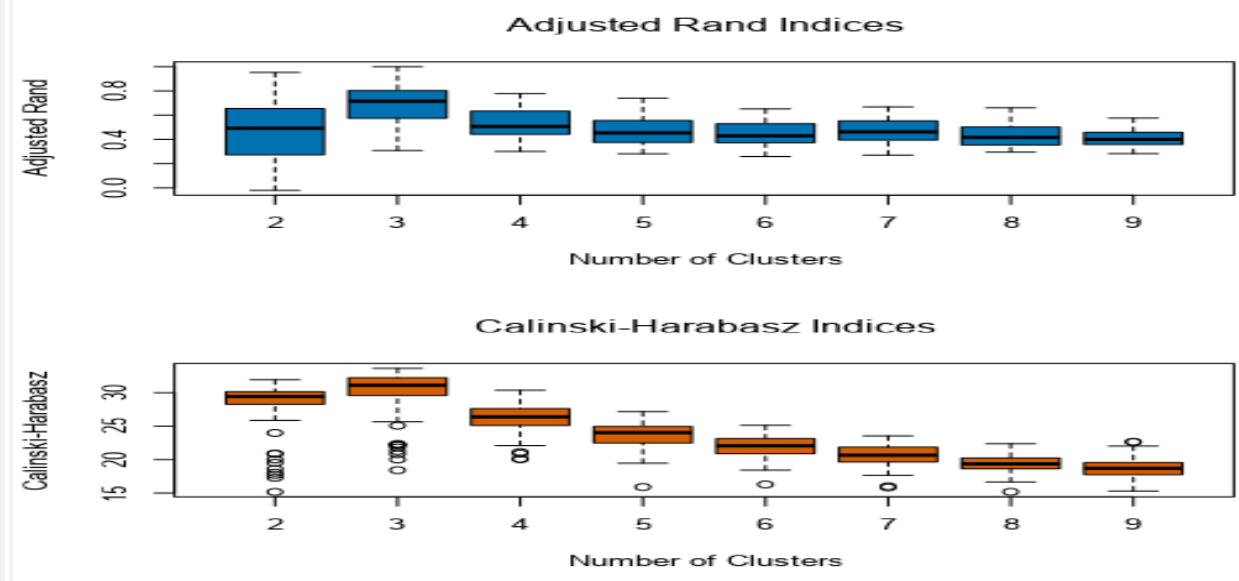
Determine Store Formats for Existing Stores

Ques 1: What is the optimal number of store formats? How did you arrive at that number?
Answers: We determine the ideal number of store formats based on sales data 2015. As mentioned, I used the percentage of sales by category per store to perform the grouping (sales of each category as a percentage of the total store sales). I have also standardized the variables using the Z-Score. The KCentroids tool Diagnostic allows us to make an assessment of the appropriate number of clusters.
The clustering algorithm is K-Means. Two measures examined are the adjusted Rand index and the Calinski-Harabasz index. This shows that 3 clusters are the ideal process, since the box-whisker plots in the Adjusted Rand Indices show how close the indices for each data point are in each other. Even though cluster 2 looks like the optimum number of clusters, it is actually 3 because the variance is too large for 2 clusters, whereas we see more compactness and even higher median values when we have 3 clusters.

With Standardizing the field with Z-Score



Without Standardizing the field with Z-Score



Ques 2: How many stores fall into each store format?
Answer: Following is the cluster information generated by the Cluster Analysis Tool K-Centroids.

Cluster Information:

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

Clusters	Stores
1	23
2	29
3	33

Ques 3: Based on the results of the clustering model, what is one way that the clusters differ from one another?
Answer: Following is the cluster information generated by the Cluster Analysis Tool K-Centroids.

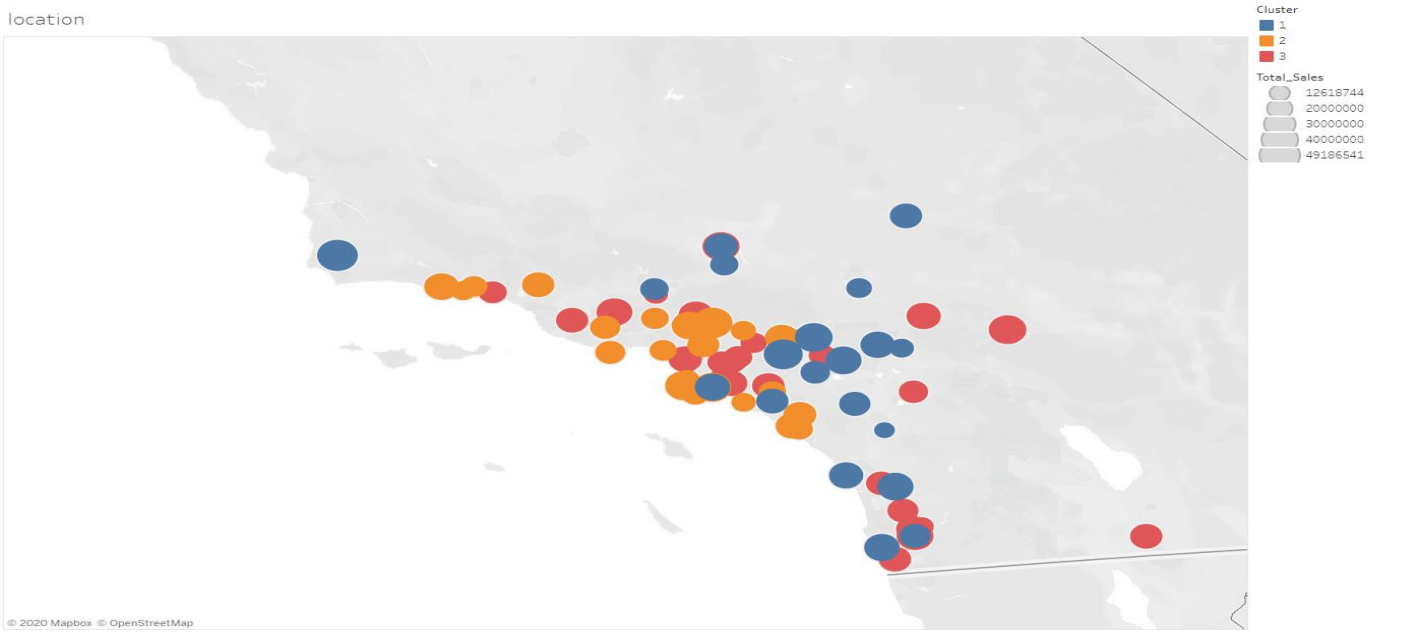
Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_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

	Percent_Bakery	Percent_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

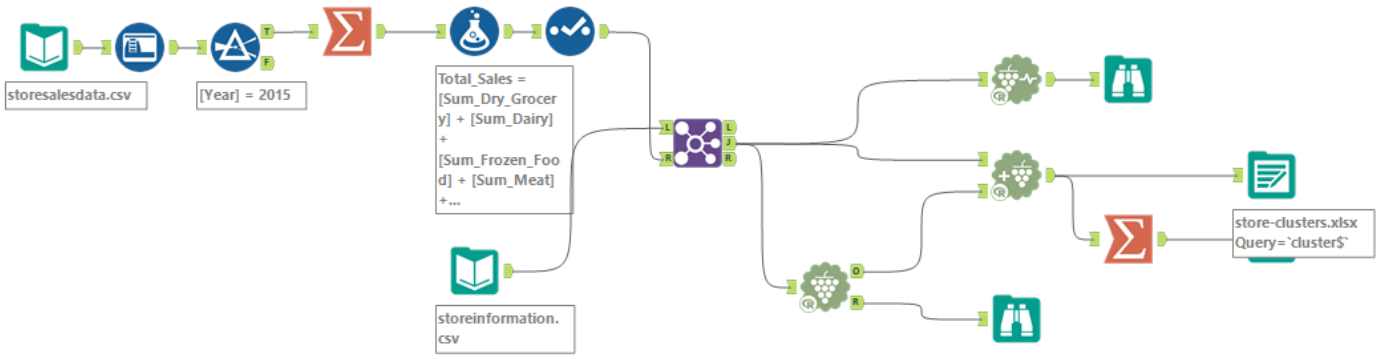
It was calculated as the average for each variable within each final cluster. The centers of final clusters reflect the characteristics of the typical case of each cluster. In particular, I deduce that
Cluster 1 stores are characterized by high percentage of merchandise sales in general.
Cluster 2 stores are characterized by high produce and floral sales in general.
Cluster 3 stores are characterized by high deli sales in general.

Ques 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.
Answer: The pic show store locations and use colors to show clusters and sizes to show total sales.



Link: https://public.tableau.com/profile/bipin7440#!/vizhome/Store-Location/location

Alteryx Workflow



Formats for New Stores

Ques 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.)

Answer: To predict which segment each store fits into, based on the characteristic demographic and socioeconomic status of the population living in the area around for each new store, I used a decision tree, forest and "Boosted" model. The model comparison tool, compares the performance of the different predictive models, follows the results:

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted	0.8235	0.8889	1.0000	1.0000	0.6667
Random_Forest	0.8235	0.8426	0.7500	1.0000	0.7778

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 Decision_Tree

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

Confusion matrix of Random_Forest

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

I choose the “Boosted” model since, despite having the same accuracy (0.8235) of the “Decision Tree”, the F1 value (0.8889) is slightly higher.

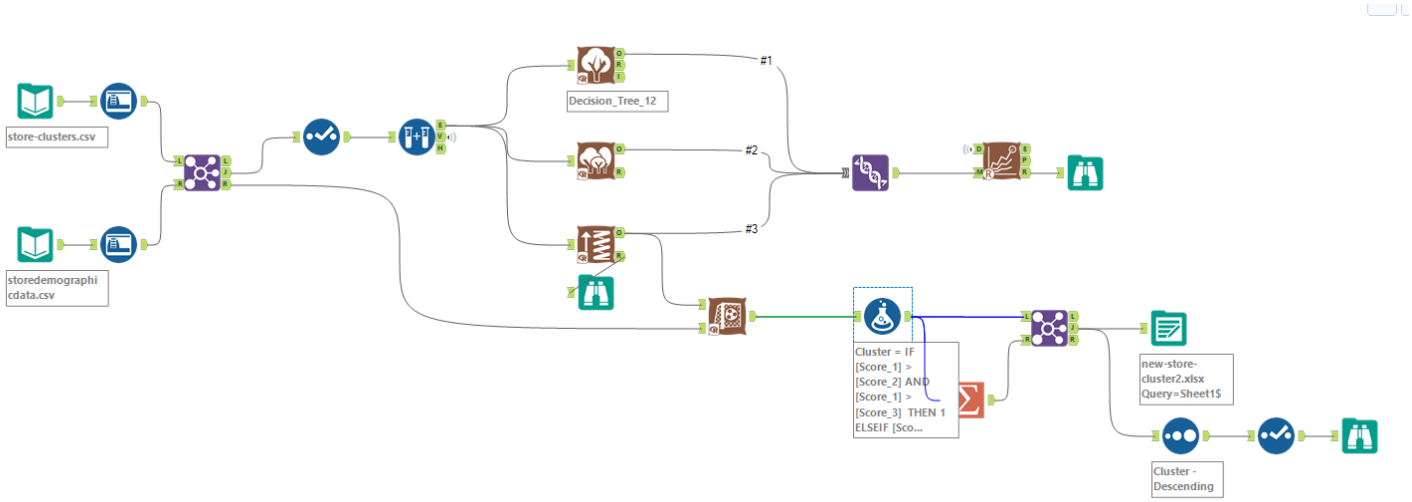
Ques 2: What format do each of the 10 new stores fall into? Please fill in the table below.

Store Number	Segment
S0086	3

S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2

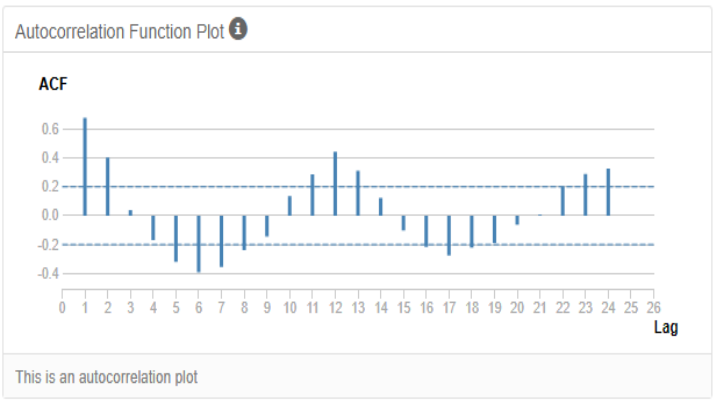
I used the following Formula to come up with the predicted cluster for each store:
If [Score_1] > [Score_2] AND [Score_1] > [Score_3] THEN 1 ELSEIF [Score_2] > [Score_1] AND [Score_2] > [Score_3] THEN 2 ELSE 3 ENDIF

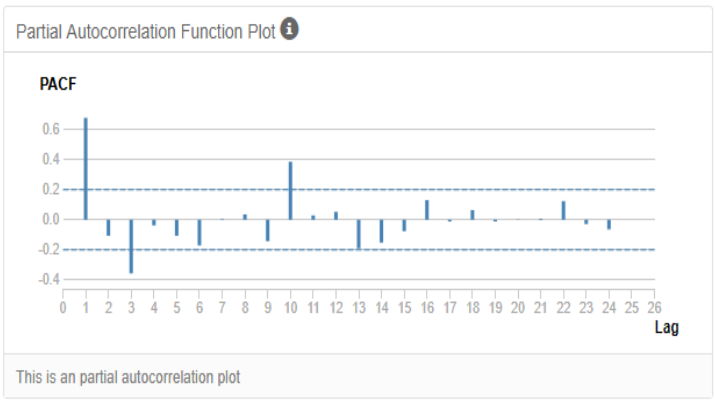
Alteryx Workflow



Predicting Produce Sales

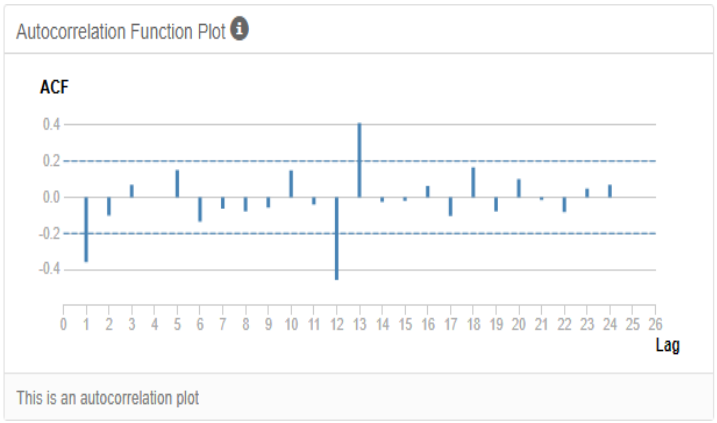
Ques 1 What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?
Answer: I prepared a forecast with monthly granularity for product sales for the year 2016 for existing and new stores. To forecast sales for stores existing, I aggregated sales in all stores per month and produce a forecast. The time series is broken down into three time series which is the seasonal component, the trend component and the rest. Below, we report the decomposition graph:

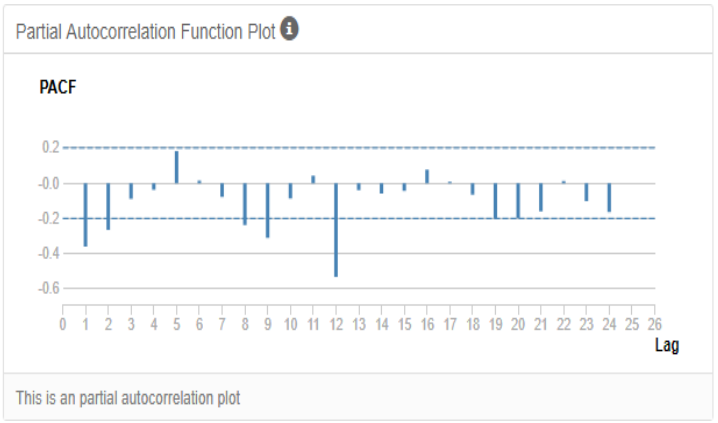




I then built the ETS model, examining the seasonal component, component of trend and rest component on the time series decomposition graph. The seasonality is growing slightly over time (peaks are increasing very slowly), so we apply this multiplicatively, the series there is no trend, the error is increasing or decreasing over time, then we apply the error multiplicatively, so we choose the ETS (M, N, M). The assembly of an ARIMA model requires that the series be stationary. The graphics autocorrelation (ACF) or partial autocorrelation (PACF) help us to determine the existence of autocorrelation:

I noted that the ACF shows an oscillation, indicating a seasonal series, in the "lags", it can be observed in several seasonal periods. In the monthly data, I noted that in lags 12, 24, peaks occur at 12-month intervals and 24 months furthermore we observed that a peak in delay 1 on an ACF chart indicates a strong correlation between each value in the series and the previous value. Then I adjusted the series with the seasonal model ARIMA. Non-stationary series can be corrected by a transformation like differentiation. Applying the first seasonal difference, we can see in the graph below that the time series was parked. Observing the ACF and PACF autocorrelation graphs from the first seasonal difference, we can identify the numbers of ARIMA terms needed.





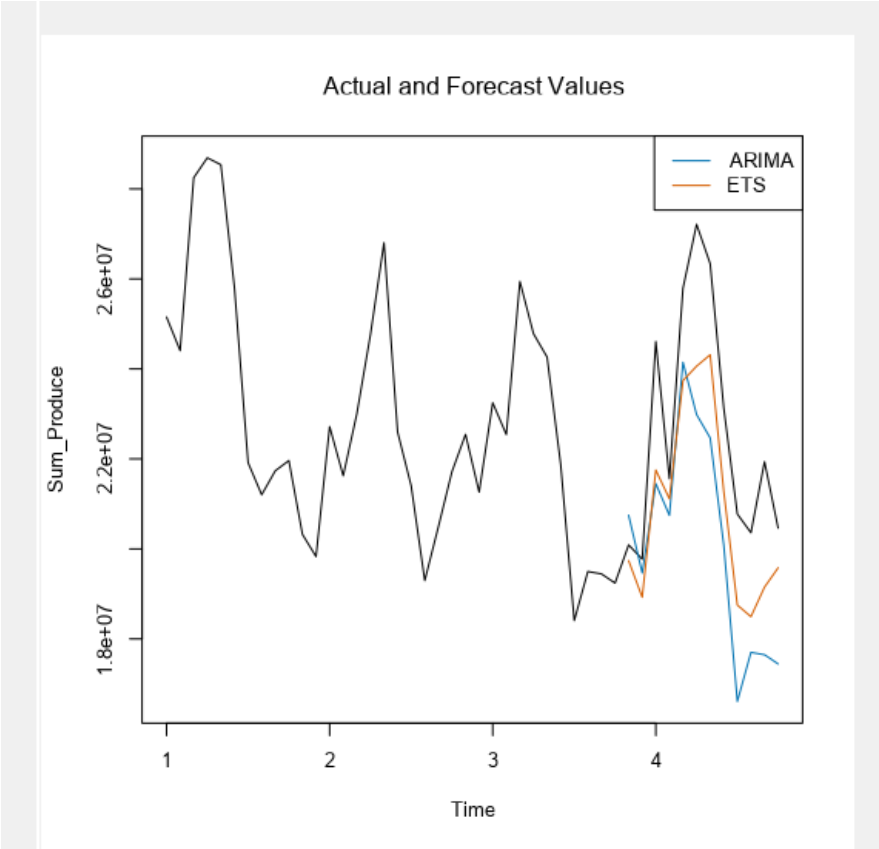
Observing the two negative peaks in the FAC in lag 1, which indicates non-seasonal BF terms. For seasonal terms, I noted that there is a peak negative at 12-month intervals. This indicates seasonal MA terms. Then the model that fits is ARIMA (0, 1, 1) (0, 1, 1) 12.

When choosing models, I used some of the data available for testing, such as validation sample and we use the rest of the data to verify the model. The holdout sample size must be the number of periods I want to predict and given that, the goal is to provide a forecast for the next 12 months of sales. Data points from 2015-01 to 2015-12 have been removed from the series of data, we can see the table of precision statistics for each model

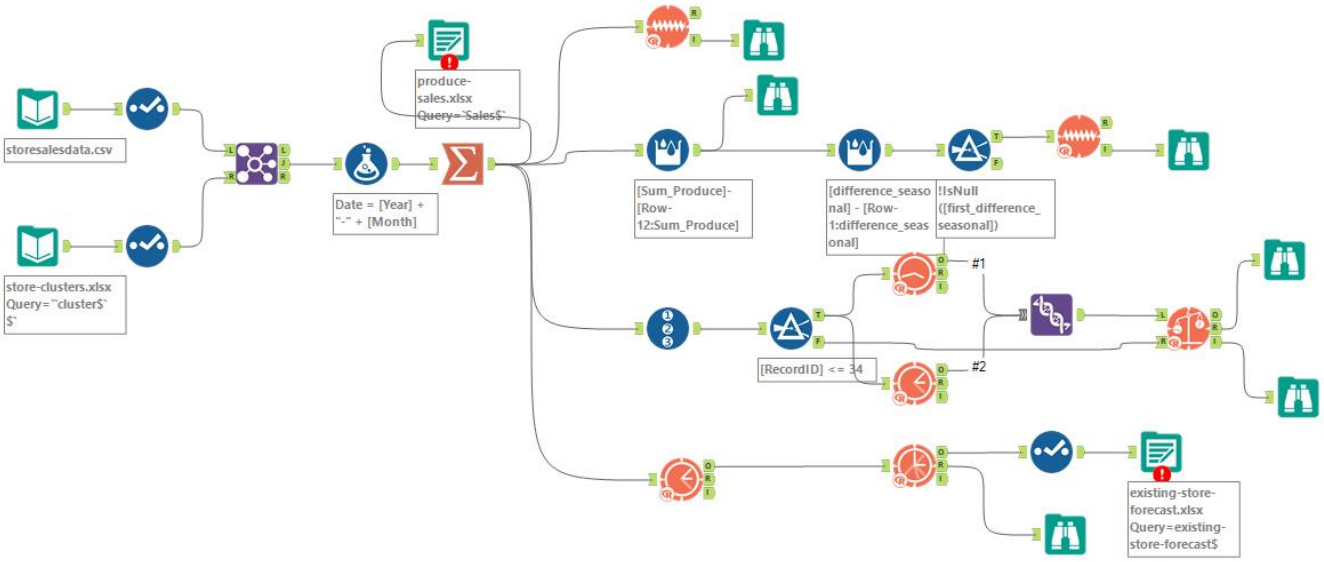
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	2545369	2999244	2655219	11.0071	11.5539	1.6988
ETS	1761302	1978476	1761302	7.5704	7.5704	1.1269

From the values in the table, I concluded that the ETS model is better than the ARIMA model for this problem, since EMS model RMSE and MASE are lower than the ARIMA model. For the ETS model, RMSE is 1978476 and MASE 1.1269 and for the ARIMA model, RMSE is 2999244 and MASE is 1.6988.
Following is the graph that shows all-time series values and forecast values for all compared models.
In the test we can see how the ETS model behaves more accurately than the ARIMA model for this data set, reaffirming the use of ETS for our problem.



Alteryx Workflow

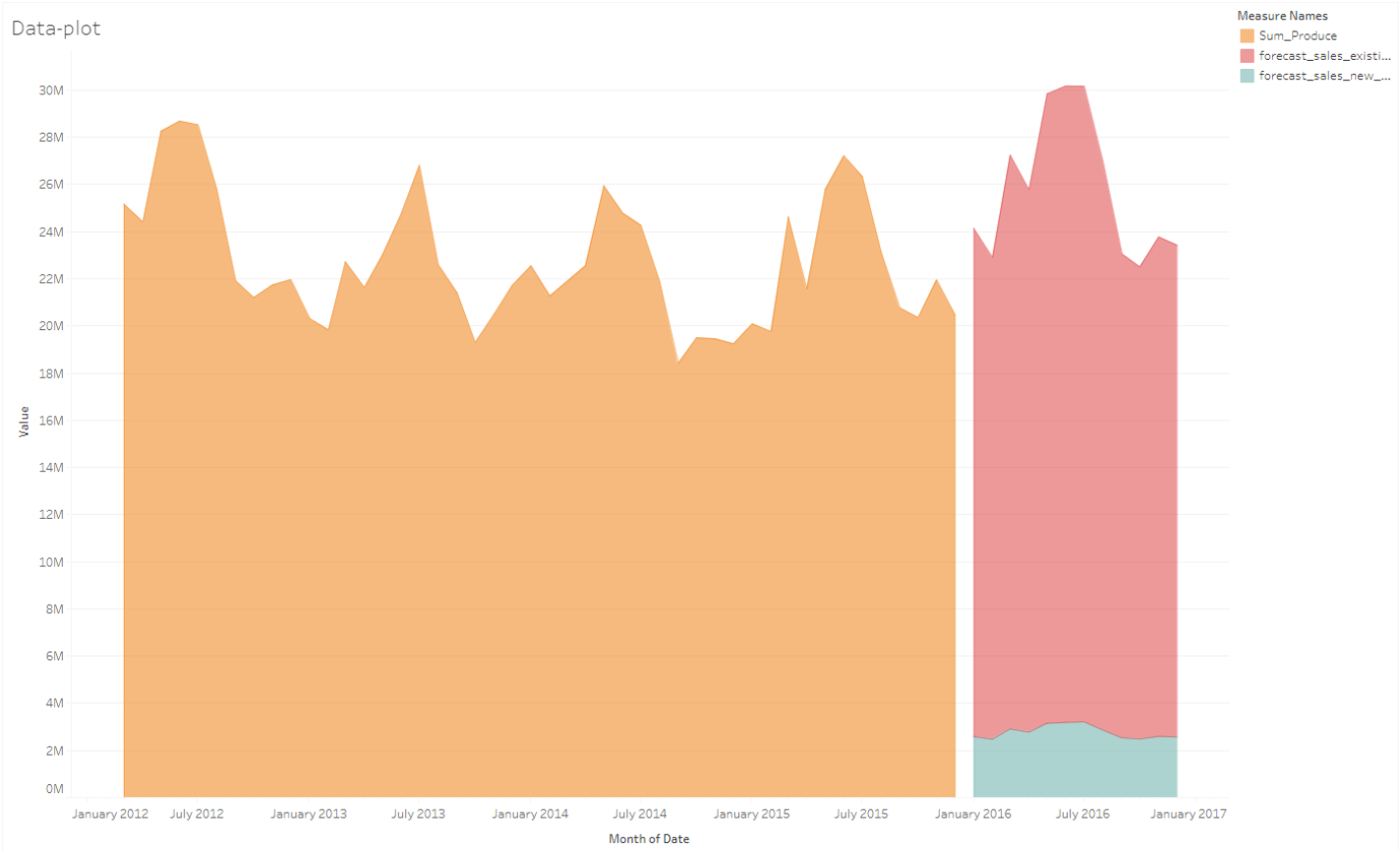


Ques 2: Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Answer:

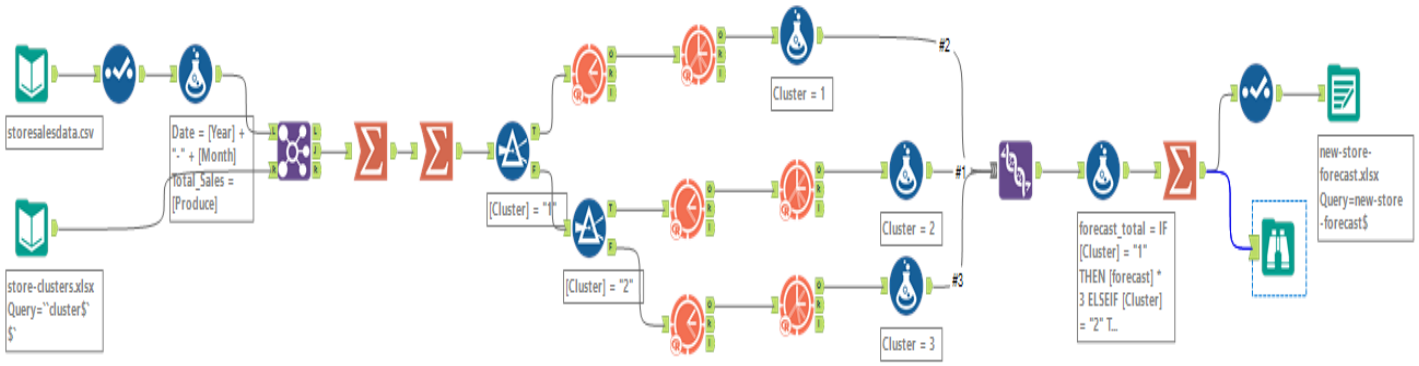
Year	Month	forecast_sales_existing_stores	forecast_sales_new_stores
2016	1	21829060	2588357
2016	2	21146330	2498567
2016	3	23735687	2919067

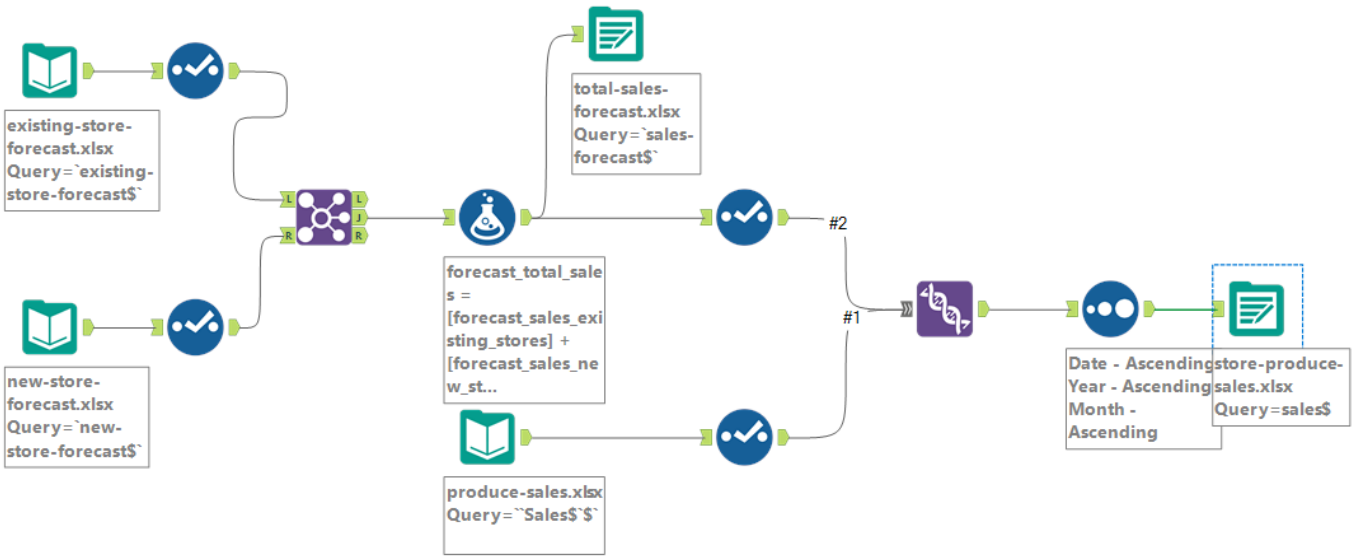
2016	4	22409515	2797280
2016	5	25621829	3163765
2016	6	26307858	3202813
2016	7	26705093	3228212
2016	8	23440761	2868915
2016	9	20640047	2538372
2016	10	20086270	2485732
2016	11	20858120	2583448
2016	12	21255190	2562182



Link : https://public.tableau.com/profile/bipin7440#!/vizhome/ProduleSales/Data-plot

Alteryx Workflow




existing-store-forecast.xlsx Query=`existing-store-forecast$`
new-store-forecast.xlsx Query=`new-store-forecast$`
total-sales-forecast.xlsx Query=`sales-forecast$`
forecast_total_sales = [forecast_sales_existing_stores] + [forecast_sales_new_st...
produce-sales.xlsx Query=``Sales$`$`
#2
#1
Date - Ascending
Year - Ascending
Month - Ascending
store-produce-sales.xlsx Query=sales$