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?

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4
Minimum	-0.010301	0.105996	0.14205
1st Quartile	0.110724	0.290955	0.297785
Median	0.428735	0.411022	0.409202
Mean	0.409553	0.440623	0.410116
3rd Quartile	0.714527	0.580392	0.51712
Maximum	1	0.85143	0.7173

Calinski-Harabasz Indices:

	2	3	4
Minimum	9.056197	8.594103	11.10884
1st Quartile	17.485413	15.481045	14.09839
Median	19.901347	17.173811	14.87037
Mean	18.543358	16.554277	14.87413
3rd Quartile	20.917592	18.032112	15.87772
Maximum	21.992647	19.089004	16.77123

Fig: K-Means Cluster Assessment Report

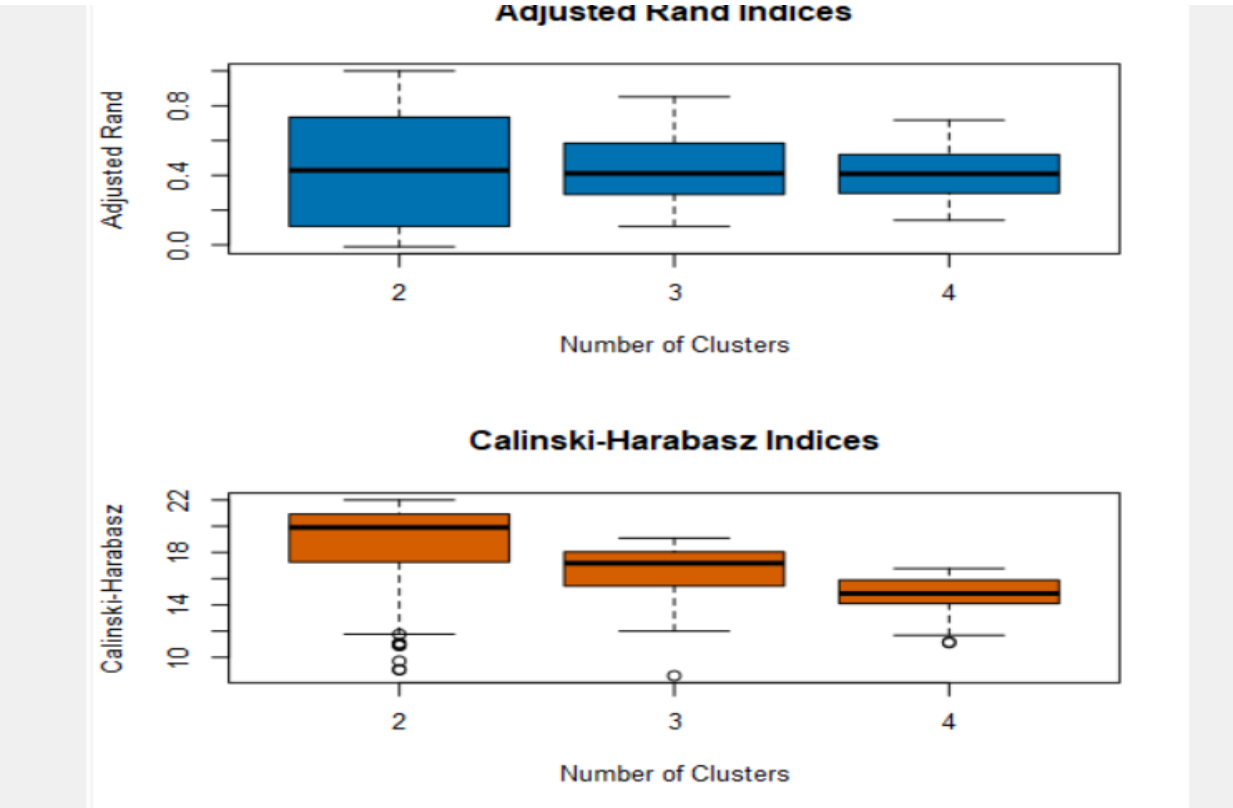


Fig: Adjusted Rand Indices and Calinski-Harabasz Indices

Based on the K-means report, Adjusted Rand and Calinski-Harabasz indices below, the optimal number of store formats is **3** because for 3 clusters CH and AR indices have high median value and are compact.

2. How many stores fall into each store format?

Cluster 1 has 23 stores, cluster 2 has 29 stores while cluster 3 has 33 stores.

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

Fig: Cluster Information

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

Cluster 1 stores have the highest total sales. Cluster 2 stores sold more Produce. Cluster 1 stores sold more General Merchandise.

Cluster 1 stores have the highest median total sales when compared to the other 2. Cluster 3 stores are the most similar in terms of sales due to their compact range.

Summary Report of the K-Means Clustering Solution ClusterByStore

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Grocery_PC_Sale + Diary_PC_sale + FrozenFood_PC_Sale + Meat_PC_Sale + Produce_PC_Sale + Floral_PC_Sale + Deli_PC_Sale + Bakery_PC_Sale + GM_PC_Sale, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

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

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

	Grocery_PC_Sale	Diary_PC_sale	FrozenFood_PC_Sale	Meat_PC_Sale	Produce_PC_Sale	Floral_PC_Sale	Deli_PC_Sale
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_PC_Sale	GM_PC_Sale
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Fig: K-Means Clustering Summary

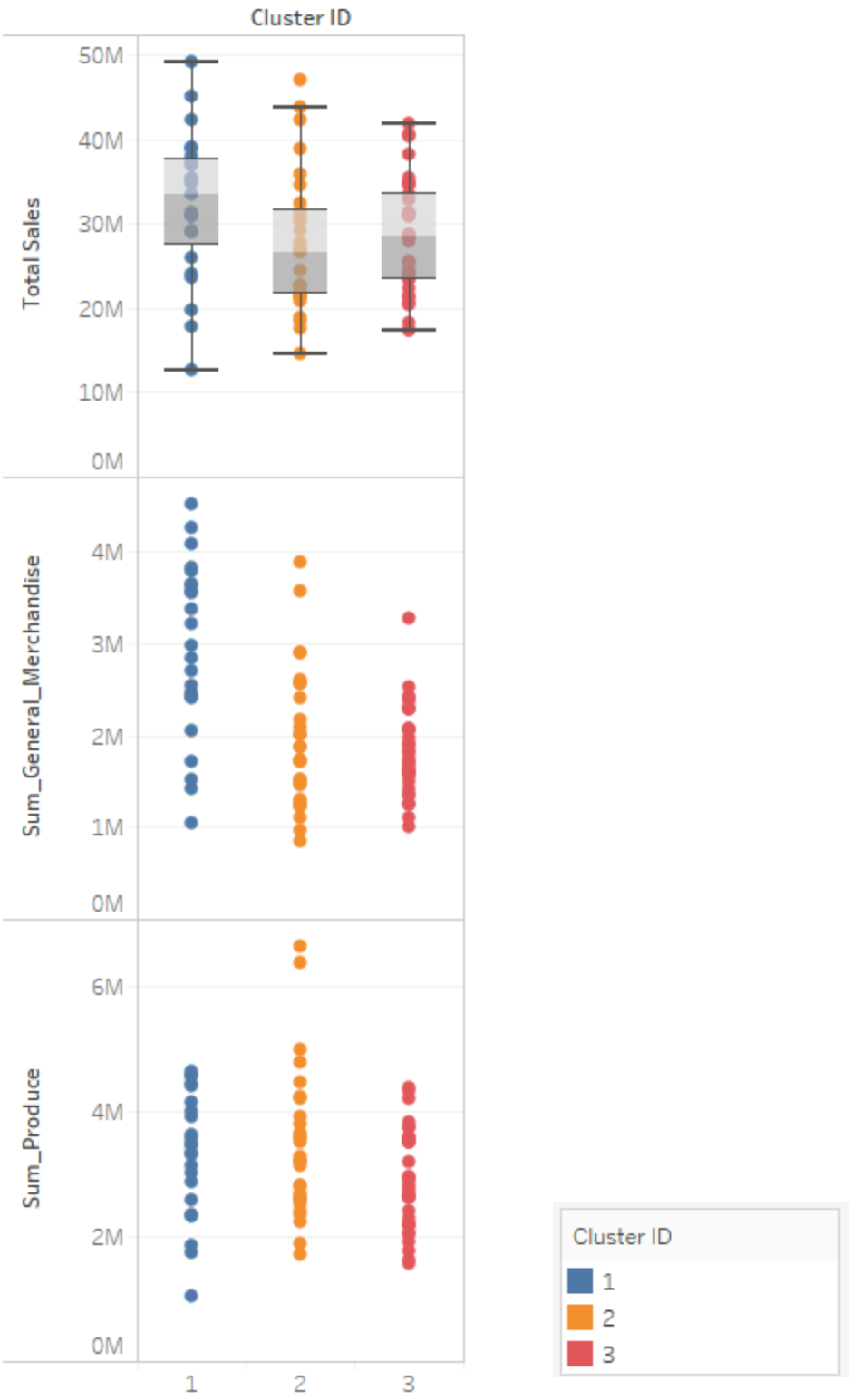
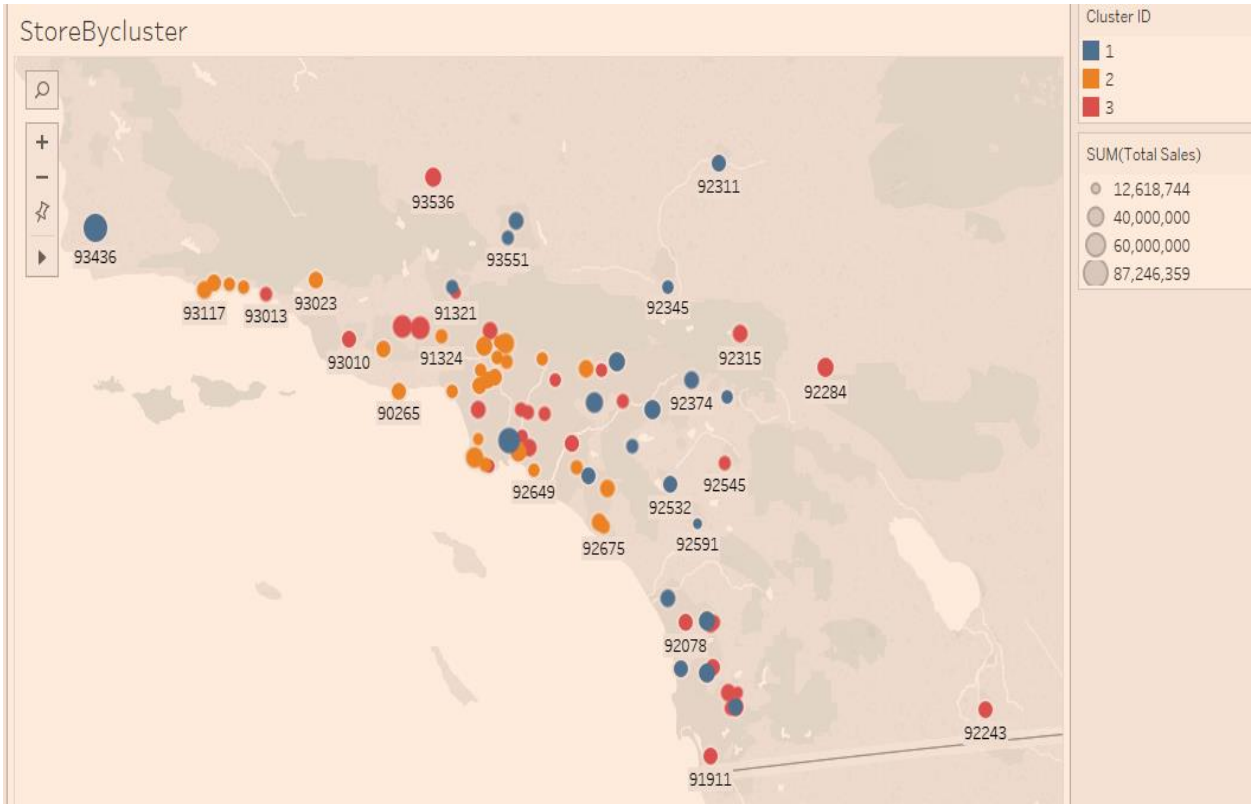


Figure 4: Tableau Visualization

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.

https://public.tableau.com/profile/doibajit.medhi#!/vizhome/StoreByClustersSales/StoreBycluster?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.)

The model comparison report below shows comparison matrix of Decision Tree, Forest Model, and Boosted Model.

Boosted Model is chosen despite having the same accuracy as Forest Model due to higher F1 value.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT	0.7059	0.7685	0.7500	1.0000	0.5556
FM	0.8235	0.8426	0.7500	1.0000	0.7778
BM	0.8235	0.8889	1.0000	1.0000	0.6667

Fig 6: Model Comparison Report

Confusion matrix of BM

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

Confusion matrix of DT

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

Confusion matrix of FM

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

Fig: Confusion Matrices for all models.

Report

Report for Boosted Model BM

Basic Summary:

Loss function distribution: Multinomial
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1612

Plots:

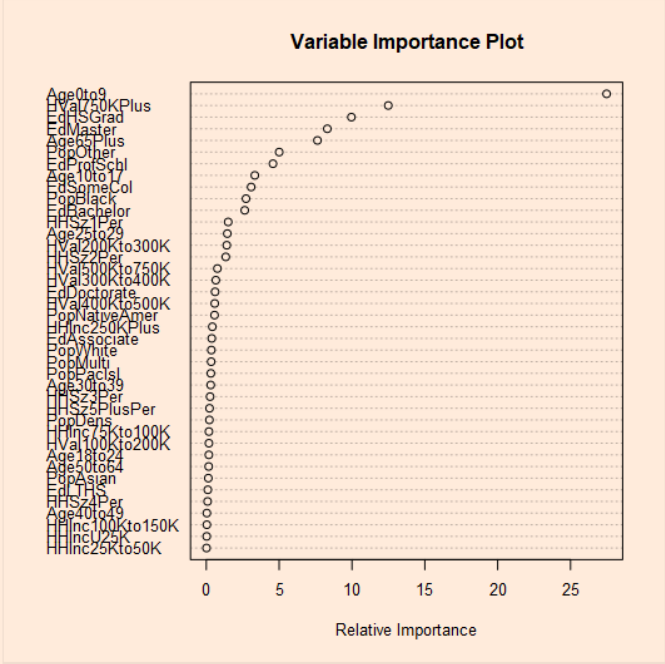


Fig: Variance Importance Plot

Ave0to9, *HVal750KPlus*, and *EdHSGrad* are the three most important variables.

3. What format does 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

Fig: Store Number and Segment

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?

ETS(M, N, M) with **no dampening** is used for ETS model.

The seasonality shows increasing peaks and valley trend and should be applied multiplicatively. The trend is not clear therefore nothing should be applied. The error is irregular and should be applied multiplicatively.

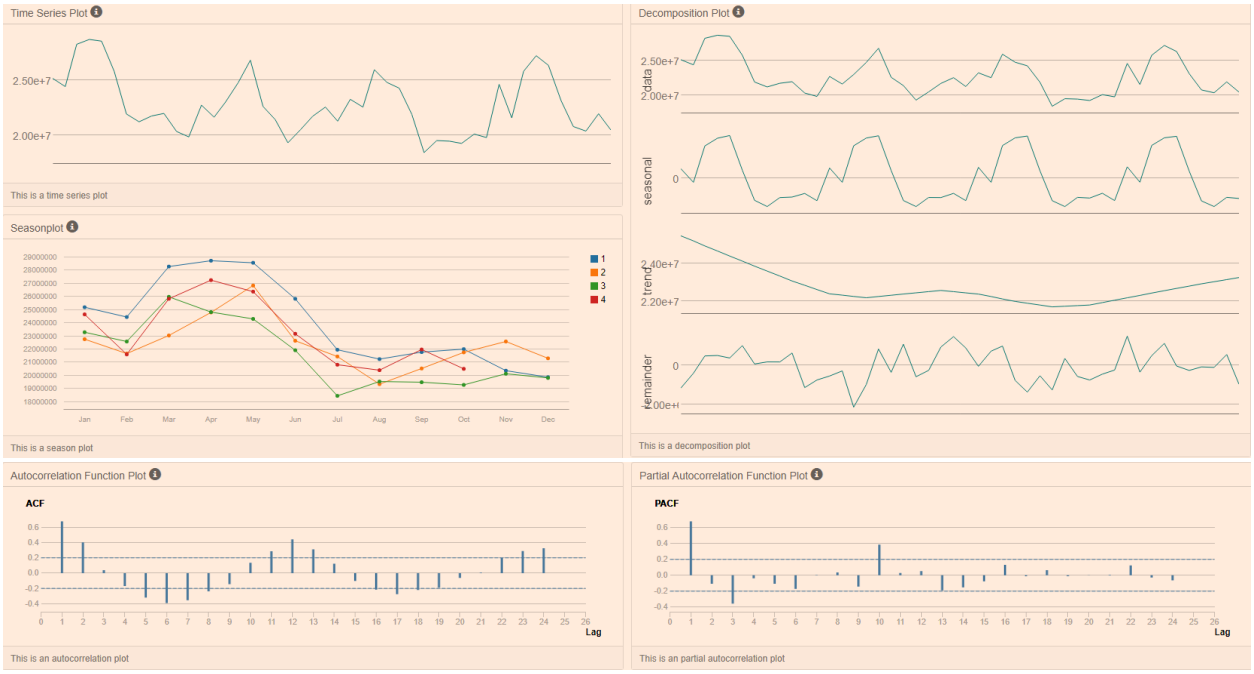


Fig: TS Plot

Summary of Time Series Exponential Smoothing Model MNM

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788

Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

Smoothing parameters:

Parameter	Value
alpha	0.539196
gamma	0.000128

Fig: ETS Model summary

For the ARIMA model, the model is set to calculate options automatically. ARIMA (1,0,0)(1,1,0)[12] is selected.

Summary of ARIMA Model ARIMA

Method: ARIMA(1,0,0)(1,1,0)[12]

Call:
auto.arima(Sum_Produce)

Coefficients:

	ar1	sar1
Value	0.79852	-0.700441
Std Err	0.126448	0.140181

sigma^2 estimated as 1671079042075.49: log likelihood = -437.22224

Information Criteria:

AIC	AICc	BIC
880.4445	881.4445	884.4411

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

Fig: ARIMA Model Summary

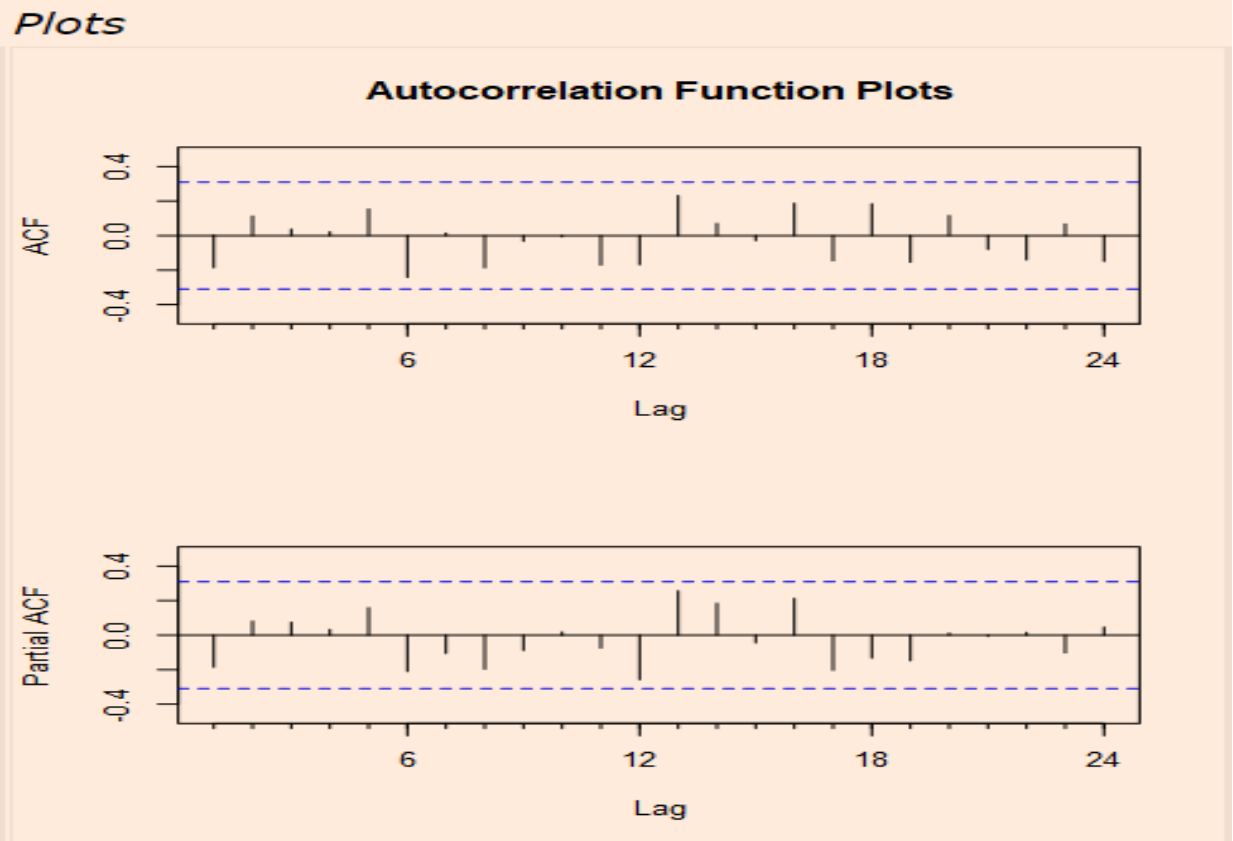


Fig: ARIMA model ACF and PCF plots

ETS model's accuracy is higher when compared to the ARIMA model. A holdout sample of 6 months of data is used. Its RMSE of 1020596 is lower than ARIMA's 1042209. ETS model also has a higher AIC in 1283 while ARIMA's AIC is 880.

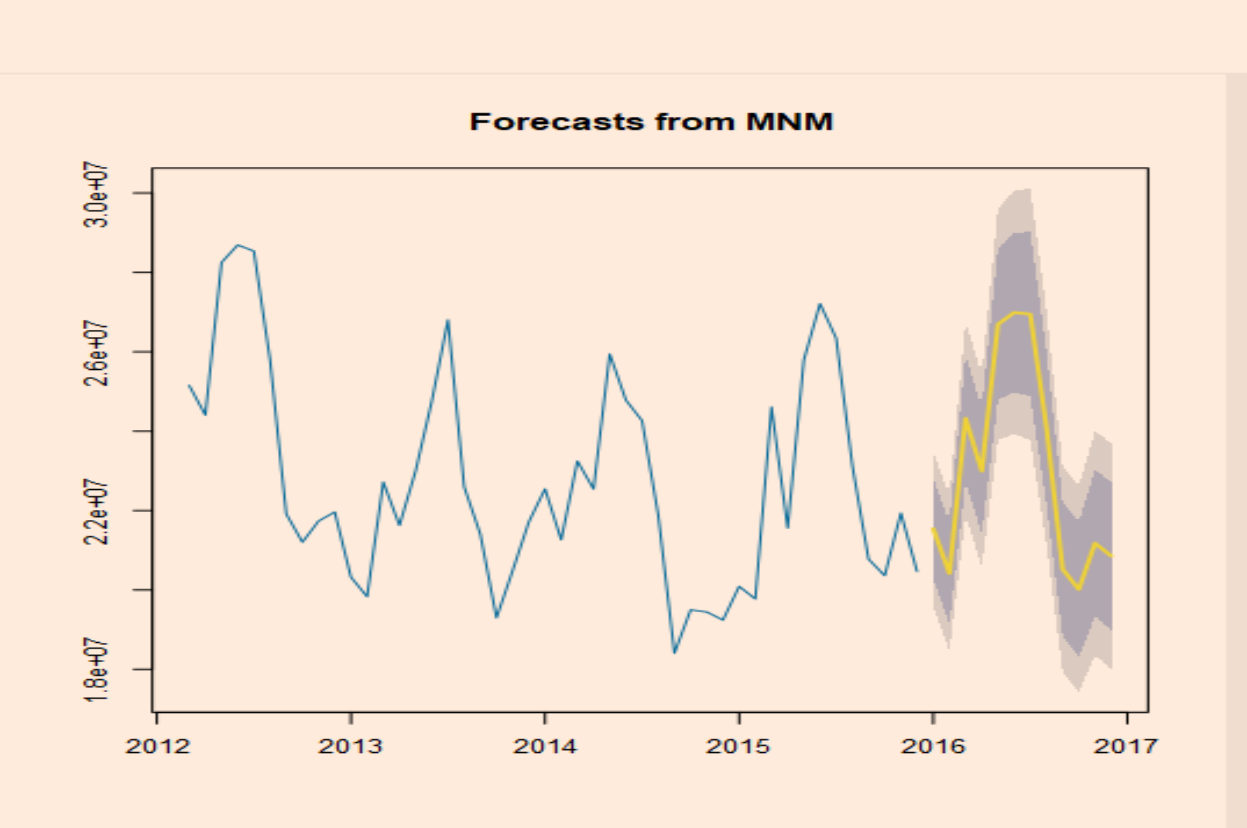


Fig: 12 Period Forecast from ETS MNM

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
2016	2	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
2016	3	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981697
2016	4	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
2016	5	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
2016	6	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
2016	7	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
2016	8	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
2016	9	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
2016	10	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
2016	11	21177435.485839	23994279.191514	23019270.585553	19335600.386124	18360591.780163
2016	12	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

Fig: Forecast with confidence intervals

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.

https://public.tableau.com/profile/doibajit.medhi#!/vizhome/TableauForecastforProduceSales/Task3-Forecast?publish=yes

Period	Month	Existing stores Sales	New stores Sales	Total Produce Sales
2016	1	21539936.01	2534110.119	24074046.13
2016	2	20413770.6	2401620.071	22815390.67
2016	3	24325953.1	2861876.835	27187829.93
2016	4	22993466.35	2705113.688	25698580.04
2016	5	26691951.42	3140229.579	29832181
2016	6	26989964.01	3175289.884	30165253.89
2016	7	26948630.76	3170427.149	30119057.91
2016	8	24091579.35	2834303.453	26925882.8
2016	9	20523492.41	2414528.519	22938020.93
2016	10	20011748.67	2354323.373	22366072.04
2016	11	21177435.49	2491462.998	23668898.48
2016	12	20855799.11	2453623.425	23309422.53

Fig: Produce Forecast table for 2016

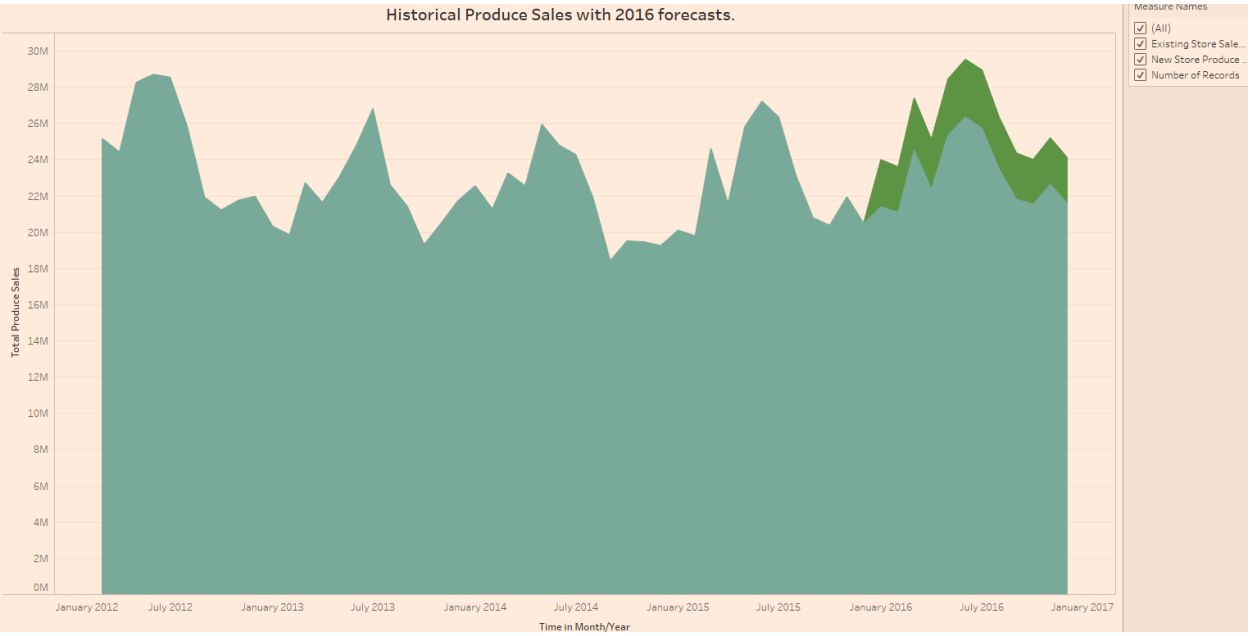
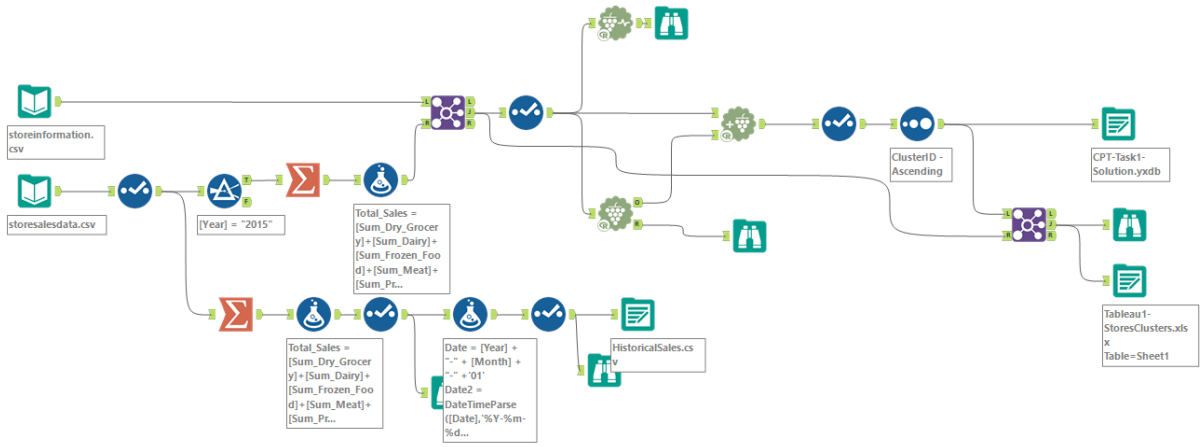
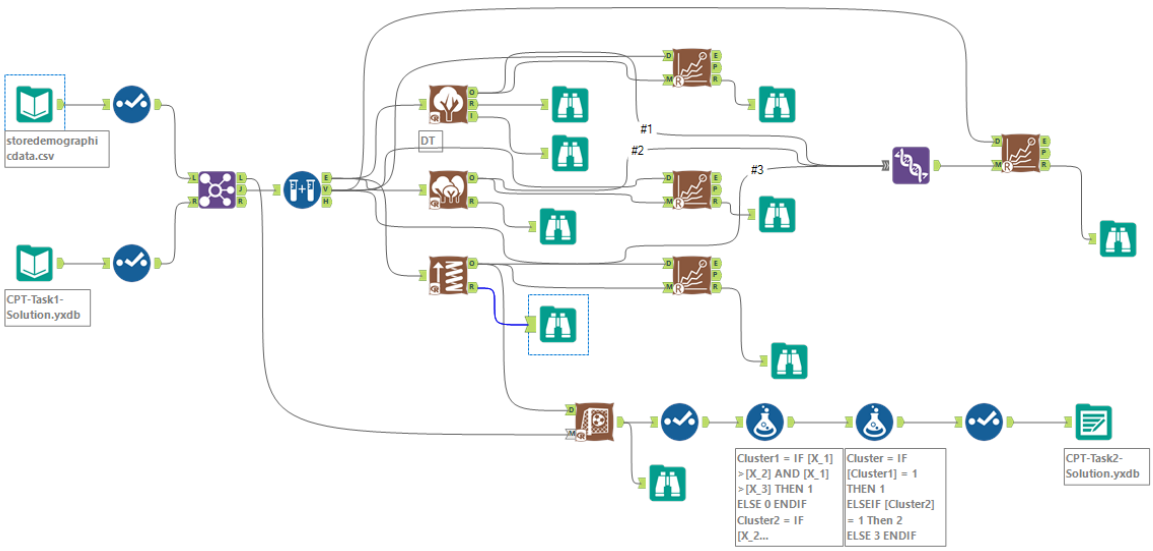
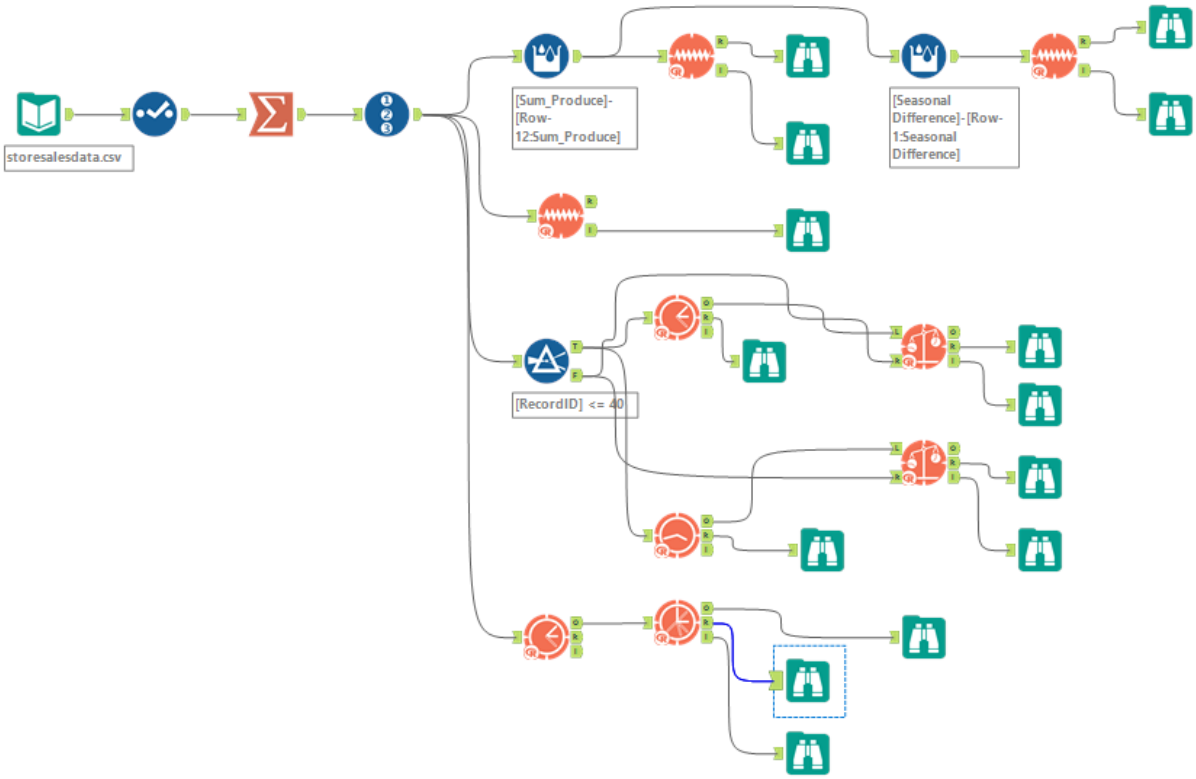


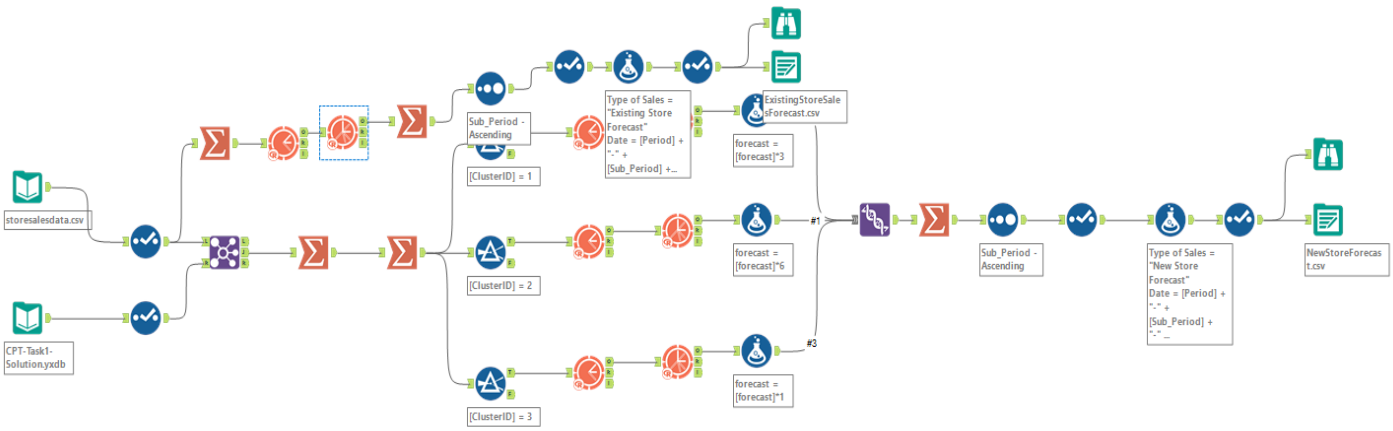
Fig: Tableau visualization for historical and forecast sales for existing stores and new stores

Alteryx Workflows:






storesalesdata.csv
[Sum_Produce]-[Row-12:Sum_Produce]
[Seasonal Difference]-[Row-1:Seasonal Difference]
[RecordID] <= 40


storesalesdata.csv
CPT-Task1-Solution.yxdb
Sub_Period - Ascending
[ClusterID] = 1
[ClusterID] = 2
[ClusterID] = 3
Type of Sales = "Existing Store Forecast" Date = [Period] + "-" + [Sub_Period] +...
ExistingStoreSaleForecast.csv
forecast = [forecast]*3
forecast = [forecast]*6
forecast = [forecast]*1
#1
#3
Sub_Period - Ascending
Type of Sales = "New Store Forecast" Date = [Period] + "-" + [Sub_Period] + "-" ...
NewStoreForecast.csv