Project: Predictive Analytics Capstone

(Abdulwasiu Tiamiyu)

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 is 3. Although 2 has the highest median values within both the Adjusted Rand Indices and Calinski-Harabasz Indices, it has a lot of outliers and higher spread. 3 has smaller spread, showing compactness.

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4
Minimum	-0.008598	0.047321	0.190877
1st Quartile	0.21411	0.311458	0.260379
Median	0.427746	0.425431	0.393611
Mean	0.426051	0.438655	0.37657
3rd Quartile	0.60704	0.577371	0.443479
Maximum	0.862177	0.806806	0.728735

Calinski-Harabasz Indices:

	2	3	4
Minimum	10.84432	10.18405	10.90095
1st Quartile	18.29771	15.23665	13.71761
Median	20.0721	16.6871	14.68046
Mean	19.04128	16.26252	14.49592
3rd Quartile	20.98638	17.42509	15.44396
Maximum	22.44228	18.75042	16.86351

Figure 1.1: K-Means Cluster Assessment Report for Adjusted Rand and Calinski-Harabasz Indices

Plots

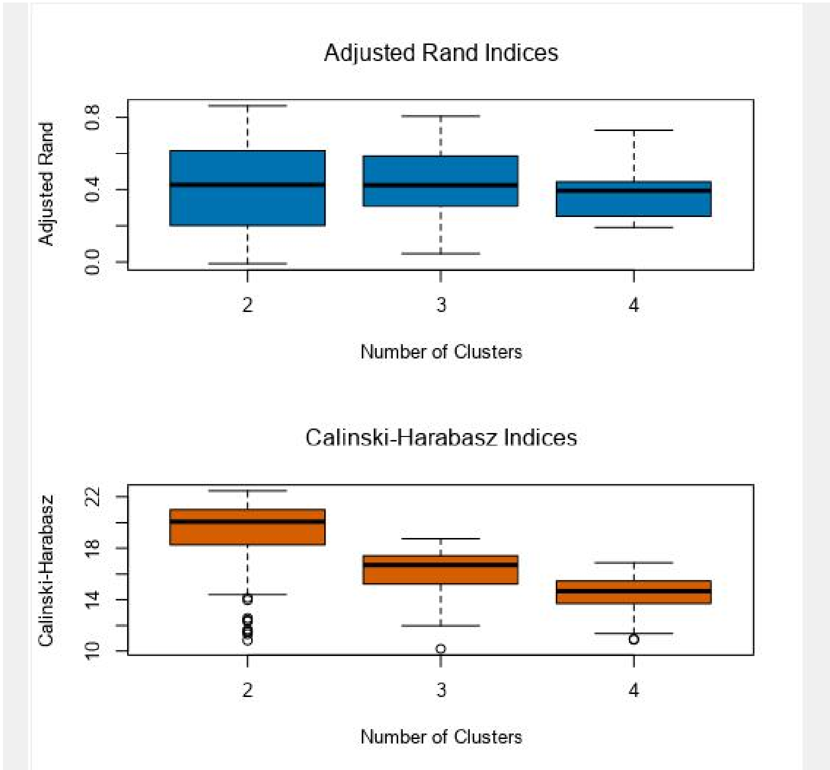


Figure 1.2: Plots for Adjusted Rand and Calinski-Harabasz Indices

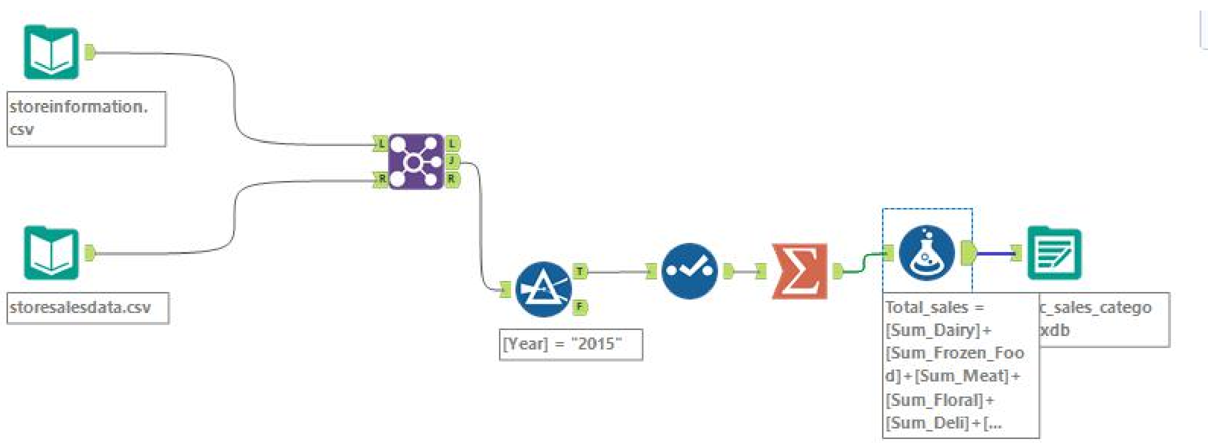


Figure 1.3: Percentage category workflow

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

Cluster 1 has 25 stores, Cluster 2 has 35 stores, and Cluster 3 has 25 stores

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

Figure 1.4: Cluster Distribution

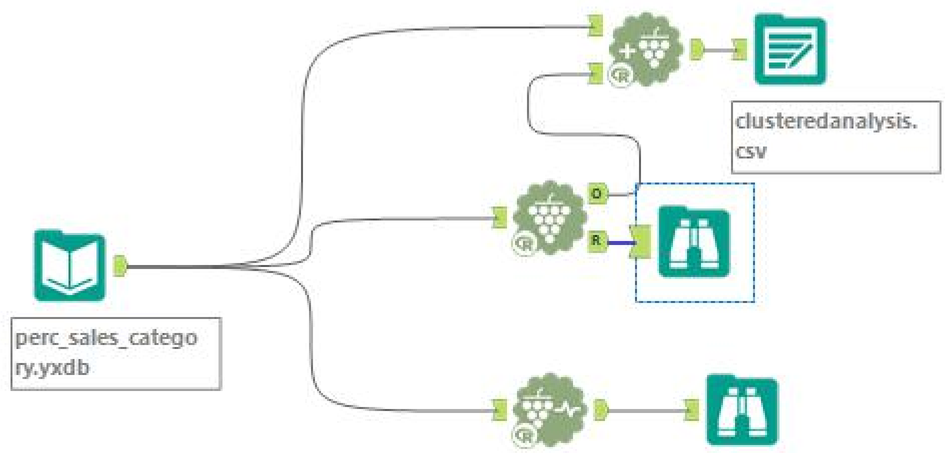


Figure 1.5: Clustered Analysis workflow

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

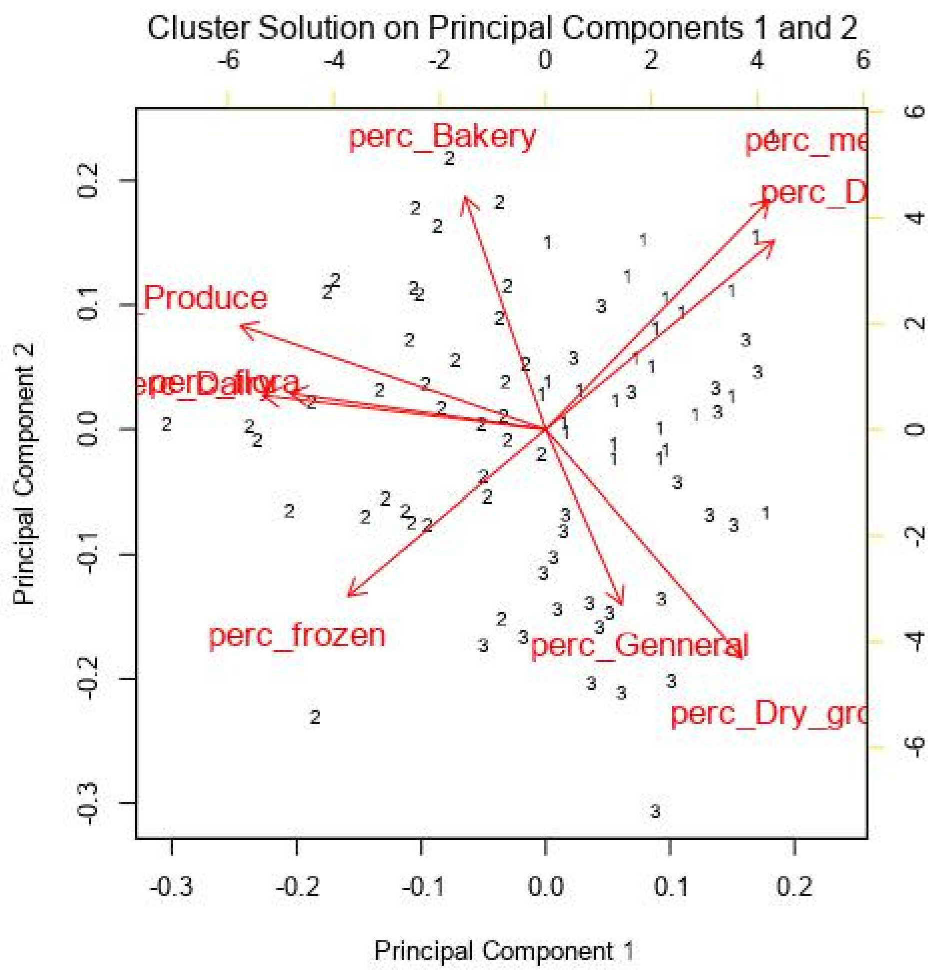
Stores that fall under cluster 3 for example would want an increase in General Merchandise as compared to stores that fall under cluster 1 & 2. Also, stores that fall under cluster 2 sold more Floral and Produce. Cluster 1 sold more Deli.

Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

	Perc_Dairy	perc_frozen	perc_meat	perc_flora	perc_Deli	perc_Bakery	perc_Genneral
1	-0.215879	-0.261597	0.614147	-0.663872	0.824834	0.428226	-0.674769
2	0.655893	0.435129	-0.384631	0.71741	-0.46168	0.312878	-0.329045
3	-0.702372	-0.347583	-0.075664	-0.340502	-0.178481	-0.866255	1.135432

	perc_Produce	perc_Dry_groce
1	-0.655027	0.528249
2	0.812883	-0.594802
3	-0.483009	0.304474

Figure 1.6: Summarized report of K-Means clustering



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.**

Tableau Visualization Link here:

https://public.tableau.com/app/profile/abdulwasiu.tiamiyu/viz/Clusterlocation/cluster?publish=yes

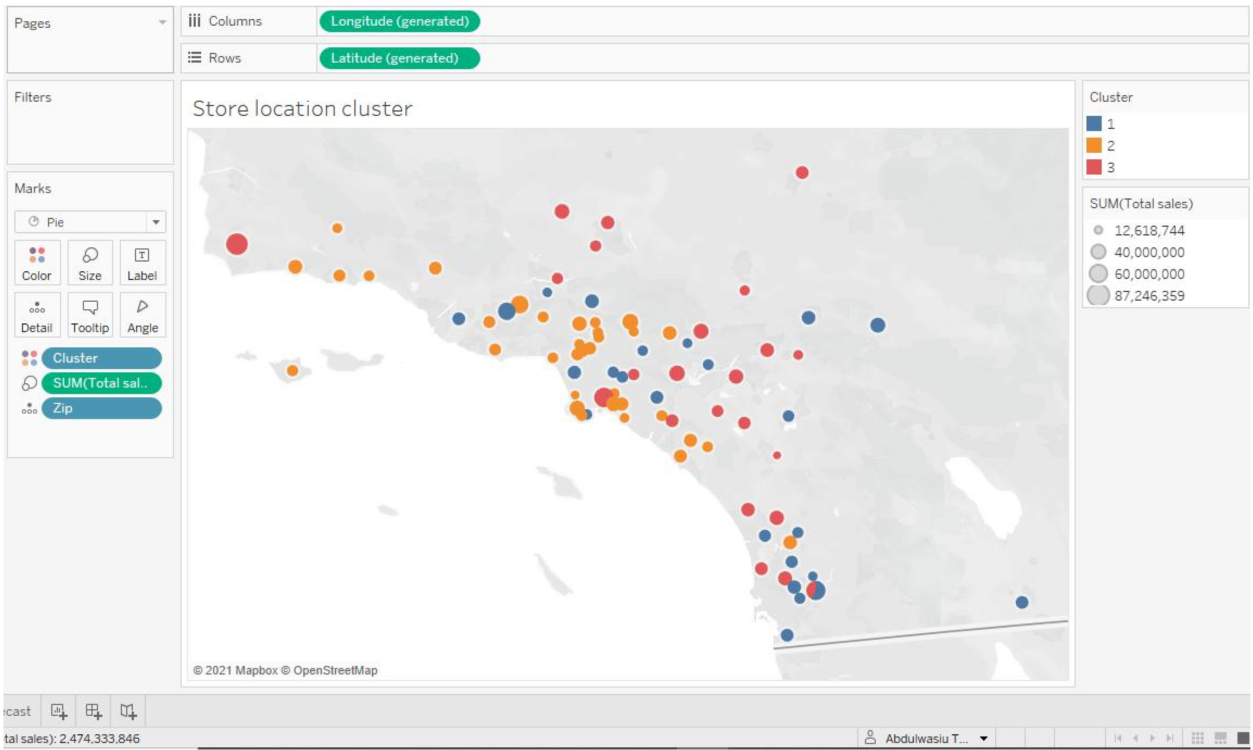


Figure 1.7: Location of stores

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.)**

I did not run a logistic regression model because this a non-binary classification problem. A decision tree, forest, and boosted model were created to predict the store formats for the new stores.

We can see in the figure below that the Forest Model and the Boosted Model have highest F1 score and highest average accuracy rate across Accuracy_1, 2 and 3 (similar), so I can use any of the two to score our 10 new scores and assign it to either cluster 1,2 or 3.

I used the Boosted Model for my prediction because of its proven record over the forest model. I created an output on the boosted model tool (BoostedModel.yxdb).

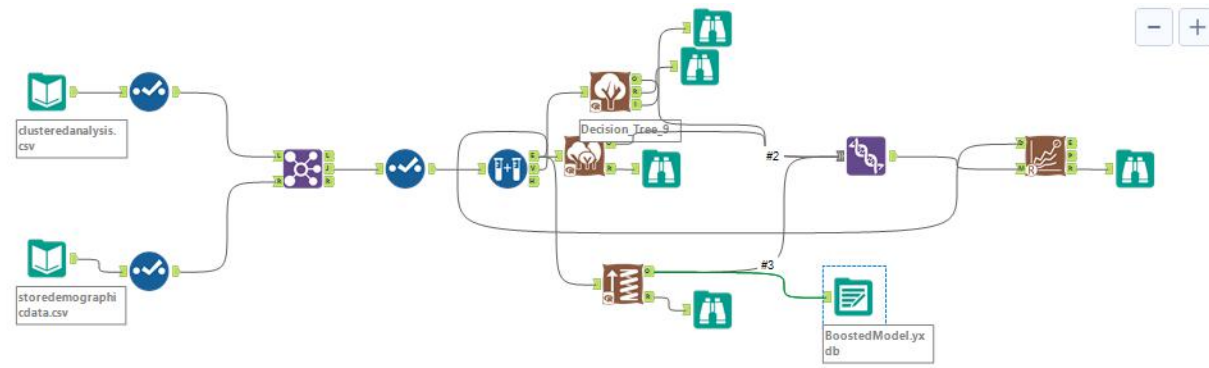


Figure 2.1: Models comparison workflow

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_9	0.6471	0.6667	0.5000	1.0000	0.5000
Forest_model	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted_model	0.7059	0.7500	0.5000	1.0000	0.7500

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boosted_model

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

Confusion matrix of Decision_Tree_9

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

Confusion matrix of Forest_model

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

Figure 2.2: Comparison report & Confusion matrix for the three models

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

	A	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW
1	Store	PopPacIsl	PopWhite	HVal0to100	HVal100Ktc	HVal200Ktc	HVal300Ktc	HVal400Ktc	HVal500Ktc	HVal750KPl	PopDens	Cluster_1	Cluster_2	Cluster_3	Cluster
2	S0086	0.000756	0.1796193	0.1303828	0.1375598	0.0885	0.1130383	0.1214115	0.3253589	0.0837	2094.40702	0.637404194	4.92E-02	0.313368473	1
3	S0087	0.00312	0.5063675	0.0178	0.0184	0.0921	0.1133235	0.097	0.3454497	0.3159161	6256.72792	9.78E-02	0.793082535	0.10915621	2
4	S0088	0.00483	0.0431	0.0471	0.0282	0.0949	0.1556588	0.1330899	0.3757053	0.1652838	8043.56289	0.264103739	0.241280574	0.494615687	3
5	S0089	0.00531	0.4530063	0.0357	0.06	0.061	0.0803	0.0688	0.4156907	0.2784904	7547.02571	1.52E-02	0.965359323	1.94E-02	2
6	S0090	0.00659	0.5270993	0.0223	0.0196	0.0176	0.0702	0.0543	0.2911979	0.524818	7621.04393	1.83E-02	0.96536773	1.63E-02	2
7	S0091	0.00629	0.4057886	0.1010909	0.2109091	0.3687273	0.1672727	0.0444	0.0749	0.0327	1054.5224	3.77E-02	2.77E-03	0.959523187	3
8	S0092	0.00162	0.471739	0.027	0.0489	0.1379258	0.1368566	0.1510616	0.3571101	0.1411333	8639.43653	1.11E-02	0.971299349	1.76E-02	2
9	S0093	0.00438	0.4697714	0.0137	0.1928491	0.3464557	0.1974586	0.108135	0.1127445	0.0287	3207.43809	2.53E-02	5.98E-03	0.968680302	3
10	S0094	0.00189	0.7136448	0.00848	0.0193	0.1210253	0.1620736	0.0998	0.1896319	0.3996917	4435.82352	8.05E-03	0.985449104	6.50E-03	2
11	S0095	0.00216	0.5671289	0.1960163	0.0534	0.1837936	0.184699	0.1896786	0.1810774	0.0113	2663.8341	9.77E-02	0.844647175	5.77E-02	2

Figure 2.3: Formats for the 10 new stores

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

Table 2.1: Segment clusters for new stores

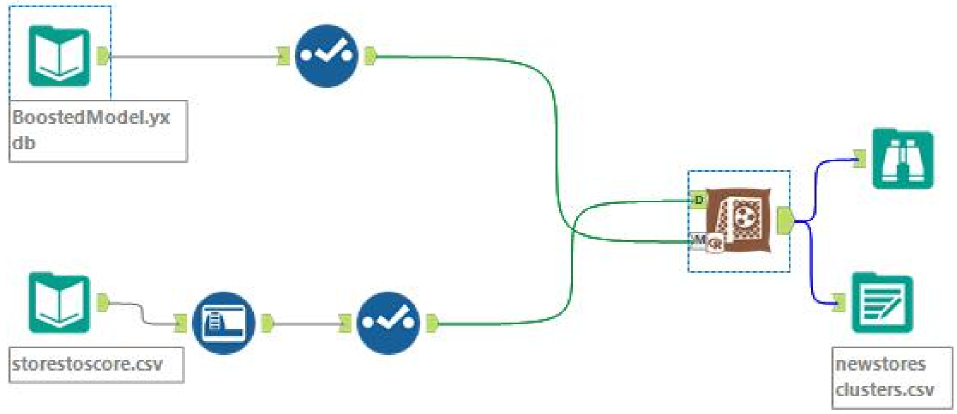


Figure 2.4: Workflow for the new formats

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?

I have chosen to use ETS (M,N,M) model (with no dampening) for each of my forecast as it gave the best result. We can see in the decomposition plot below, seasonal is multiplicative and error is multiplicative (trend not applied).

Figure 3.1: Time Series Plot/Decomposition Plot of historical monthly sales

After comparing the results against the holdout sample, the ETS performs better against the ARIMA model in term of accuracy. The RMSE and MASE is also lower than that of the ARIMA

Actual and Forecast Values

Actual	ETS	ARIMA
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	-604232.3	1050239	928412	-2.6156	4.0942	0.5463

Figure 3.2: ETS and ARIMA Model comparison

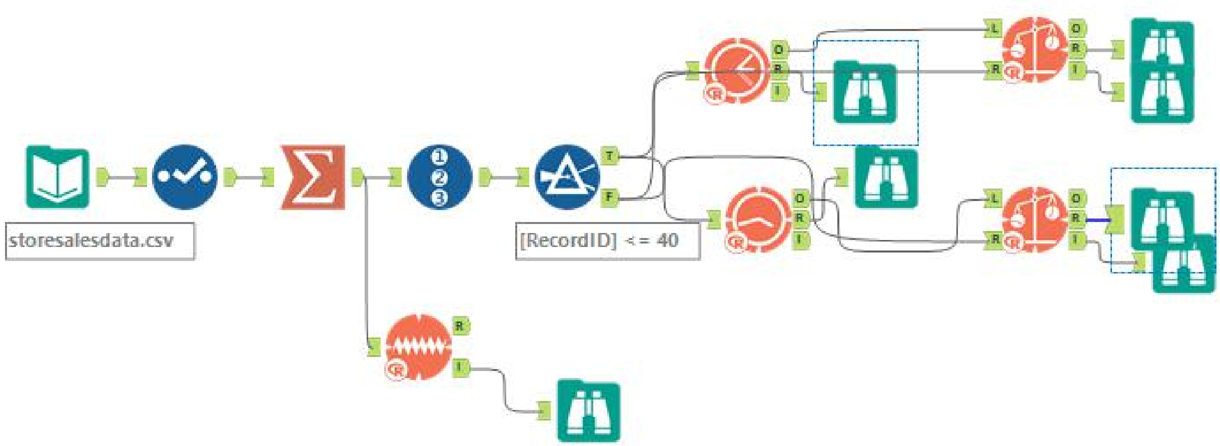


Figure 3.2: ETS and ARIMA Model comparison workflow

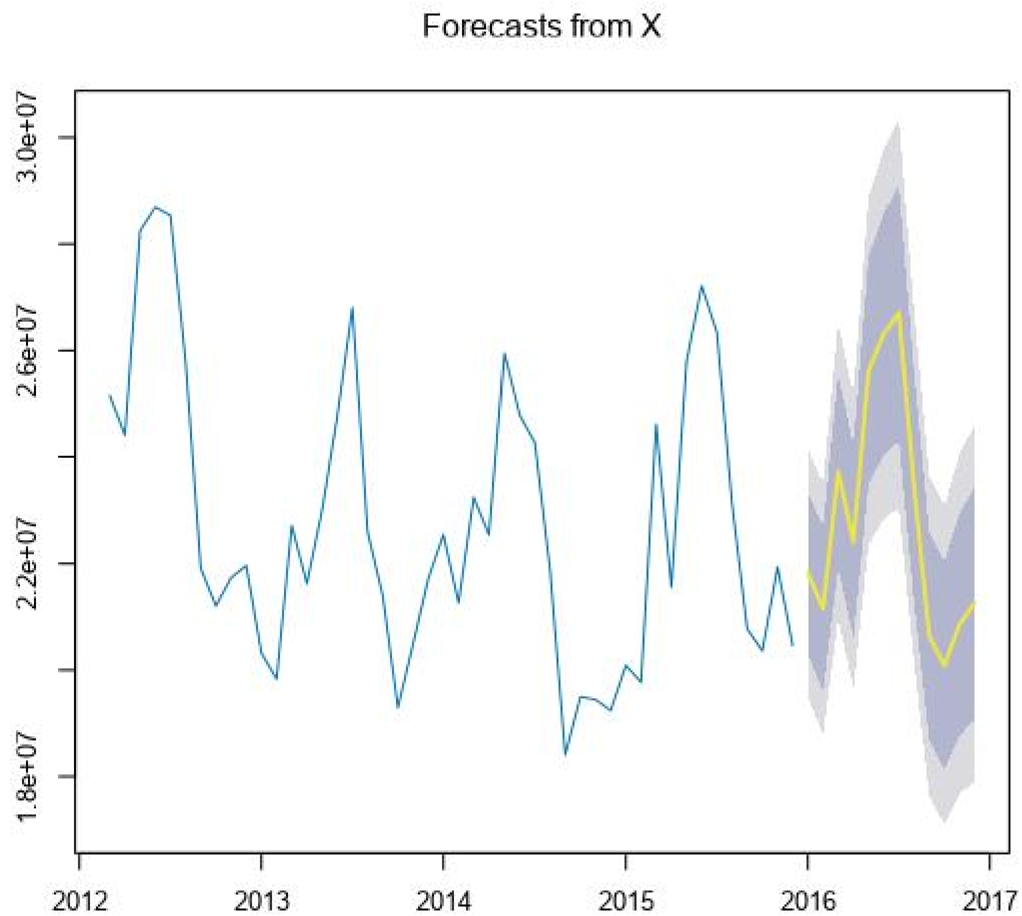


Figure 3.4: ETS forecast plot

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21829060.031666	24149899.115321	23346575.14138	20311544.921952	19508220.948011
2016	2	21146329.631982	23512577.365832	22693535.862148	19599123.401815	18780081.898131
2016	3	23735686.93879	26517865.796798	25554855.912929	21916517.964651	20953508.080782
2016	4	22409515.284474	25150243.401256	24201581.075733	20617449.493214	19668787.167691
2016	5	25621828.725097	28880596.484529	27752622.431914	23491035.018279	22363060.965665
2016	6	26307858.040046	29777680.067343	28576652.715009	24039063.365084	22838036.01275
2016	7	26705092.556349	30348682.320364	29087507.847195	24322677.265503	23061502.792334
2016	8	23440761.329527	26742106.733295	25599395.061562	21282127.597491	20139415.925758
2016	9	20640047.319971	23635033.372194	22598363.439189	18681731.200753	17645061.267747
2016	10	20086270.462075	23084199.797487	22046511.090727	18126029.833423	17088341.126662
2016	11	20858119.95754	24055437.105831	22948733.269445	18767506.645635	17660802.809249
2016	12	21255190.244976	24596988.126893	23440274.43075	19070106.059202	17913392.363058

Figure 3.5: ETS forecast table for existing data

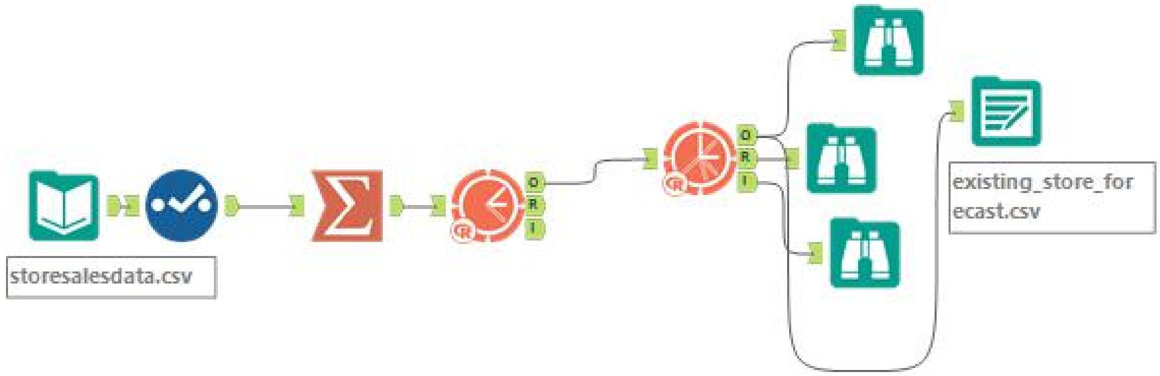


Figure 3.6: ETS forecast workflow for existing data

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.

Month	New Stores	Existing Stores
Jan 16	2,491,319	21,829,060
Feb 16	2,408,385	21,146,330
Mar 16	2,833,157	23,735,687
Apr 16	2,679,433	22,409,515
May 16	3,054,886	25,621,829
Jun 16	3,106,152	26,307,858
July 16	3,132,699	26,705,093
Aug 16	2,776,154	23,440,761
Sep 16	2,451,566	20,640,047
Oct 16	2,401,772	20,086,270
Nov 16	2,477,302	20,858,120
Dec 16	2,452,170	21,255,190
Total Annual Sales	**$32,264,995**	**$274,035,760**

Table 3.1: Sales forecast for Existing and New Stores for the next 12 months

Tableau Visualization Link here:

https://public.tableau.com/app/profile/abdulwasiu.tiamiyu/viz/Clusterlocation/Forecast?publish=yes

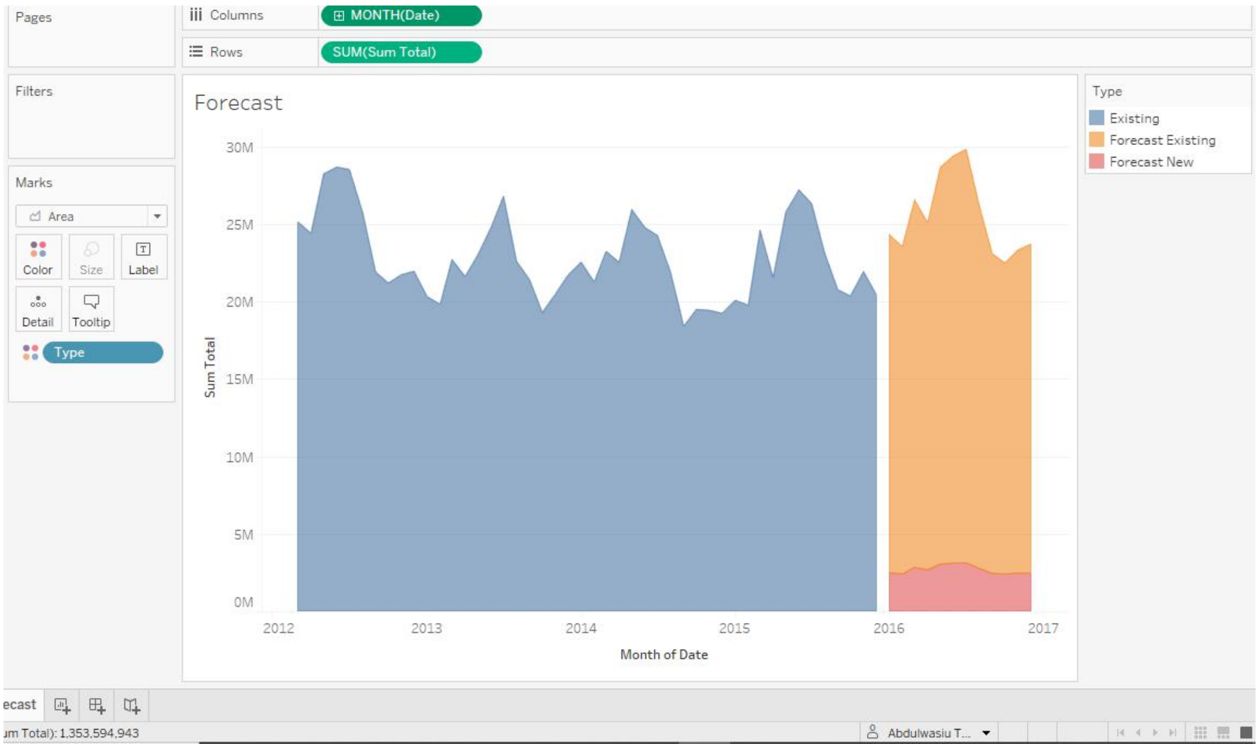


Figure 3.7: Historical and forecast sales for existing and new stores from Mar-12 to Dec-16

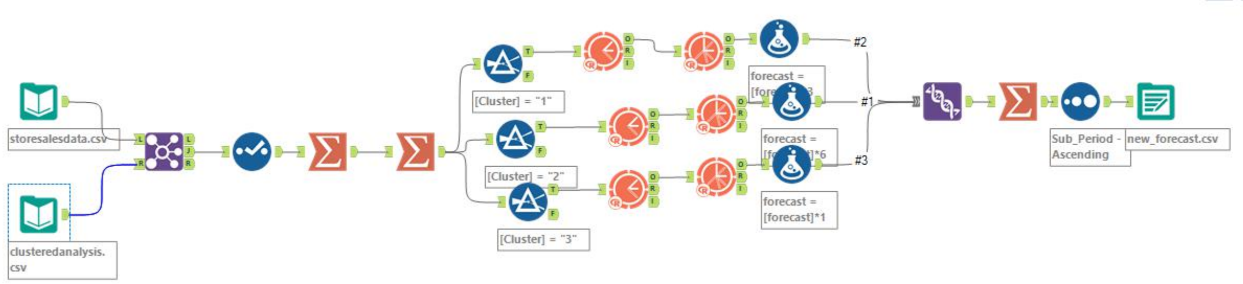


Figure 3.8: ETS forecast workflow for new data