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 K-Centroid Diagnostic Tool and clustering method by k-means was used to analyze the data. Based on this the report from the tool suggest that the optimal number of clusters is 3.

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.007972	0.356712	0.23729	0.318515	0.258423
1st Quartile	0.331034	0.639482	0.457252	0.40282	0.373217
Median	0.544552	0.727654	0.549672	0.490682	0.434725
Mean	0.528373	0.734917	0.552725	0.499096	0.451346
3rd Quartile	0.775917	0.895298	0.637192	0.575423	0.487863
Maximum	0.952941	1	0.958456	0.83378	0.833241

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	16.61829	17.73061	20.68272	20.28254	17.51664
1st Quartile	28.23418	30.15424	25.51023	22.94745	21.32559
Median	29.42988	31.06788	27.07099	24.1643	22.25157
Mean	28.30798	30.52361	26.68859	24.07233	22.1438
3rd Quartile	30.11131	32.24634	27.91717	25.24158	23.30232
Maximum	31.71569	33.63781	30.24214	26.96229	24.56733

Figure 1: K-Means Cluster Assesment Report

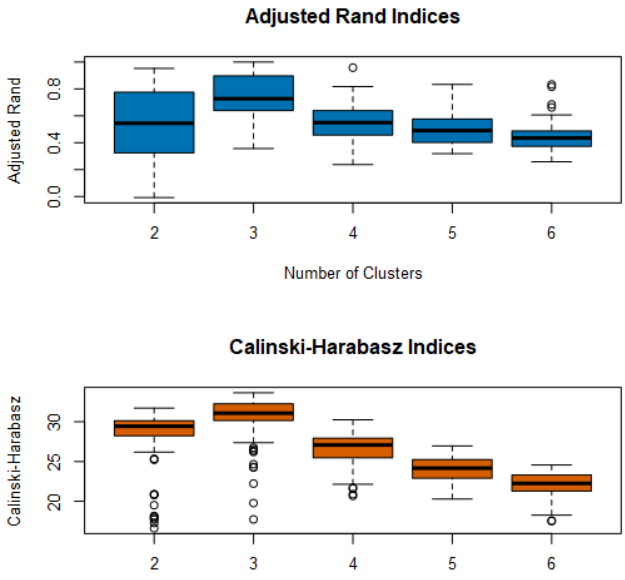


Figure 2: Adjusted Rand Indices and Calinski-Harabasz Indices

This is mainly derived by looking at the adjusted Rand indices and the Calinski-Harabasz Indices. Both show the highest mean at 3 Clusters.

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	Size	Ave Distance	Max Distance
1	23	2.320539	3.55145
2	29	2.540086	4.475132
3	33	2.115045	4.9262

Figure 3: 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 sold more General Merchandise in terms of the percentage while Cluster 2 stores sold more of Produce.

Cluster 1 stores have the highest medial total sales when compared to the other clusters. Its range of the total sales and most of other sales are also the largest. Cluster 3 stores are the most similar in terms of sales. This can be seen due to the compact state of the data.

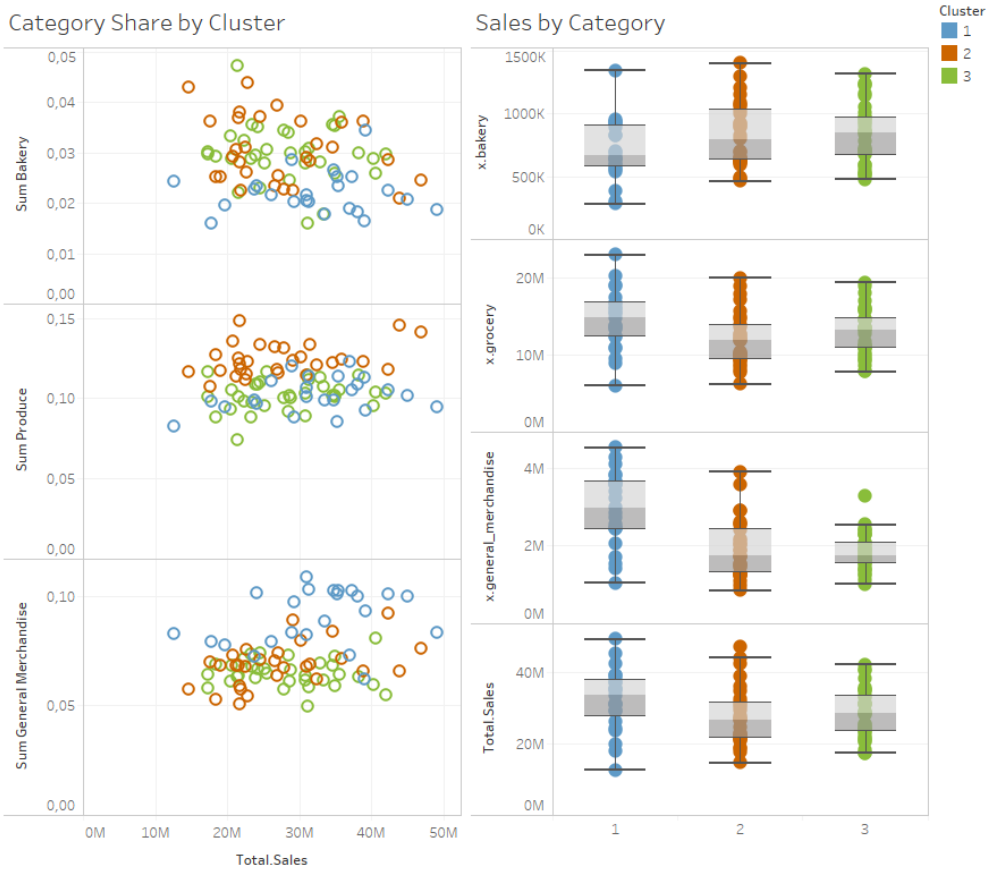


Figure 4: Tableau Visualization

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.**

Tableau Profile:
https://public.tableau.com/profile/aljoscha.grunwald#!/vizhome/Task1_4_2/Task4_1

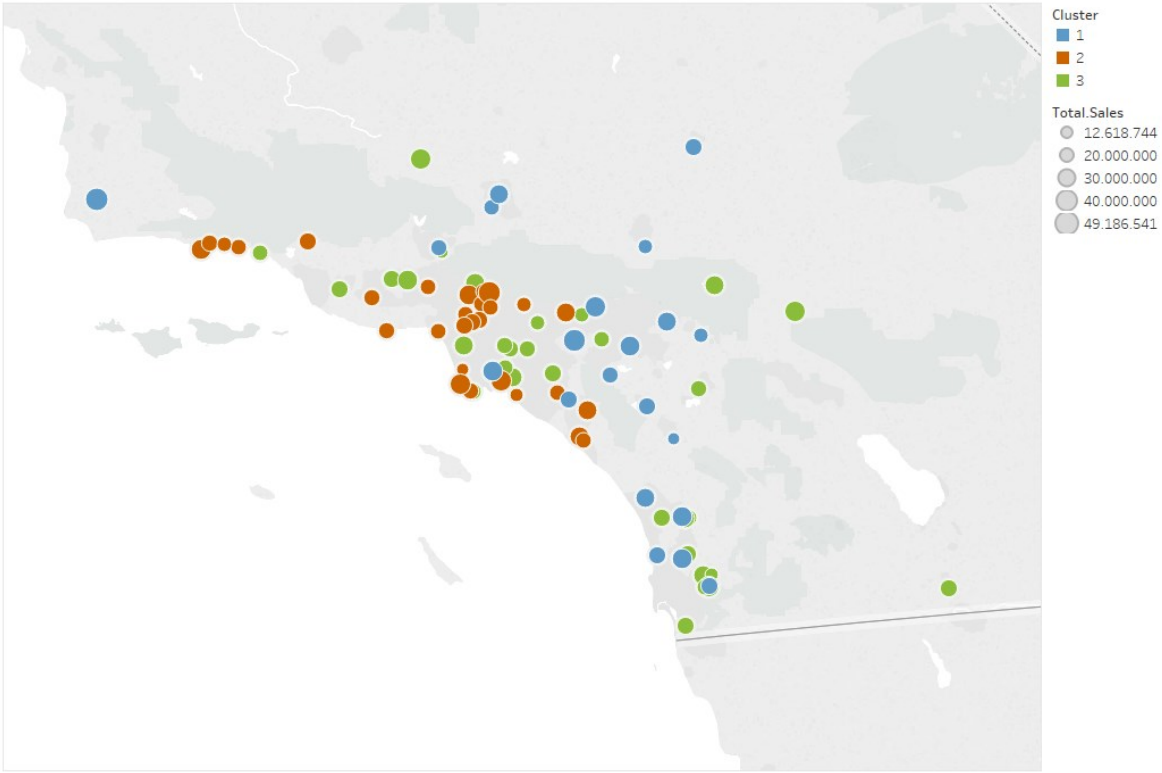


Figure 5: Location of the Stores

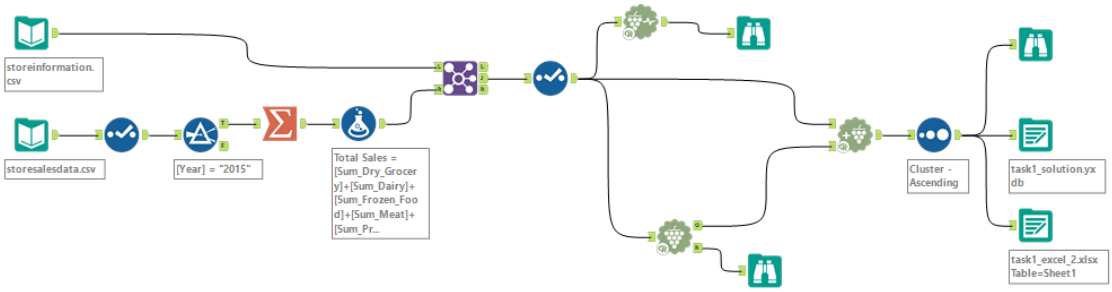


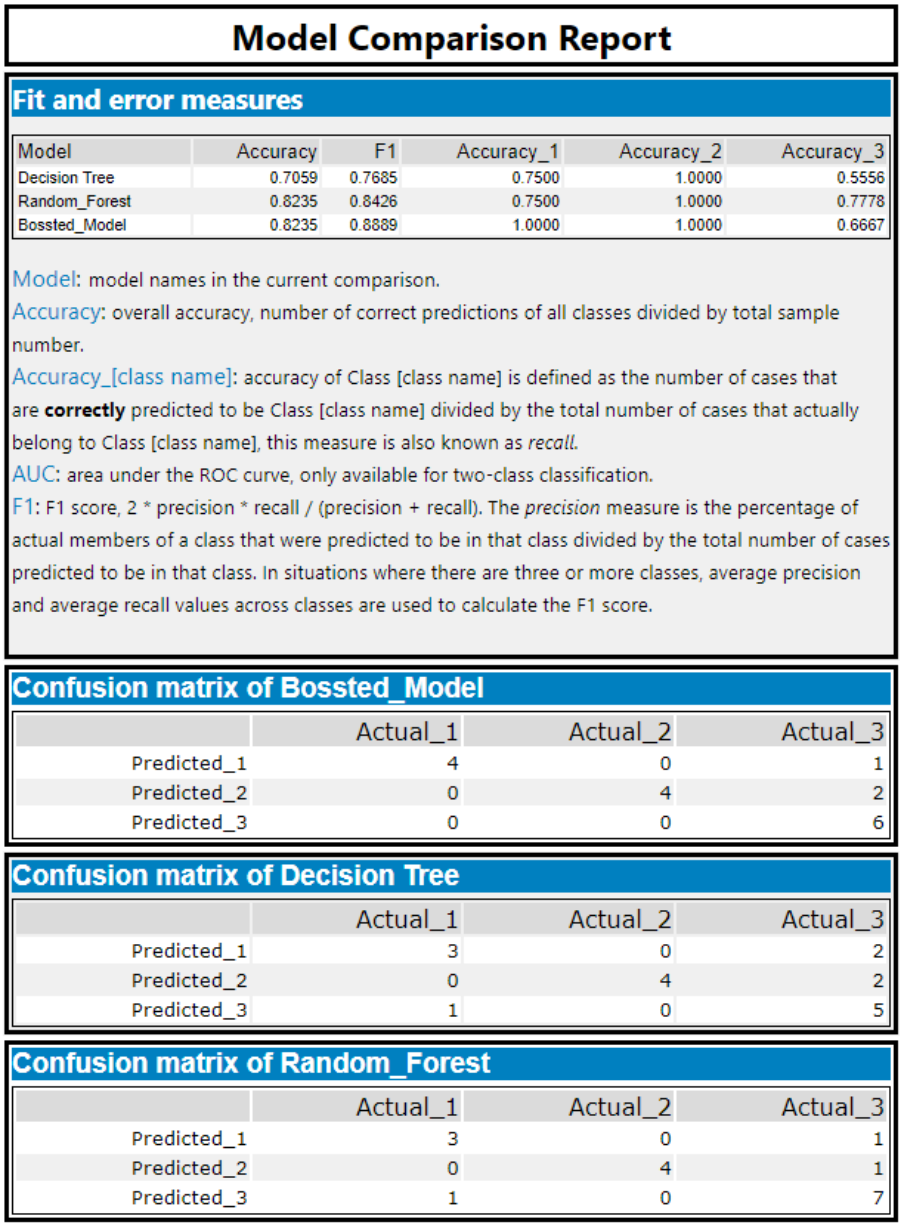
Figure 6: Alteryx Workflow Task 1

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

Despite the identical accuracy of the forest model and boosted model. The choice should be the boosted model because of the higher F1 score.



Model Comparison Report

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
Random_Forest	0.8235	0.8426	0.7500	1.0000	0.7778
Bossted_Model	0.8235	0.8889	1.0000	1.0000	0.6667

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 Bossted_Model

	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

Figure 7: Model Comparison Report

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

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

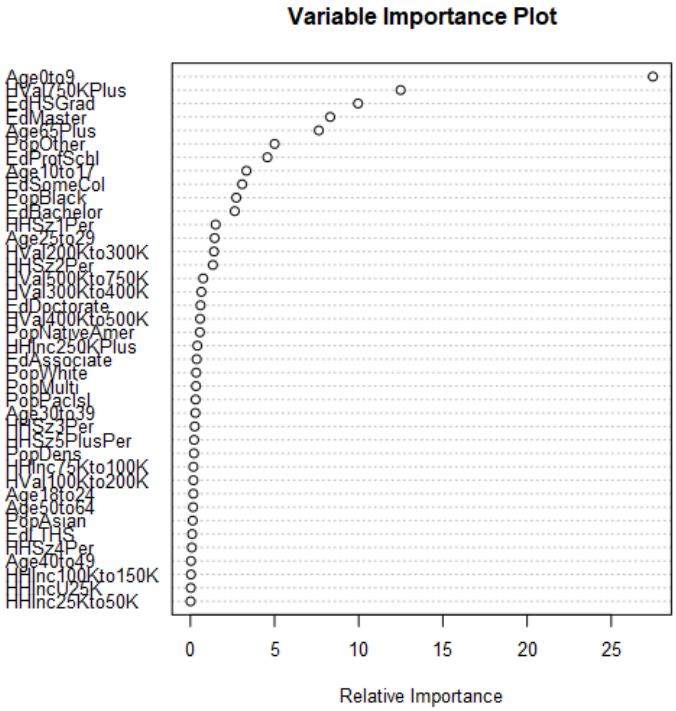


Figure 8: Variable Importance Plot

3. **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
S0095	2

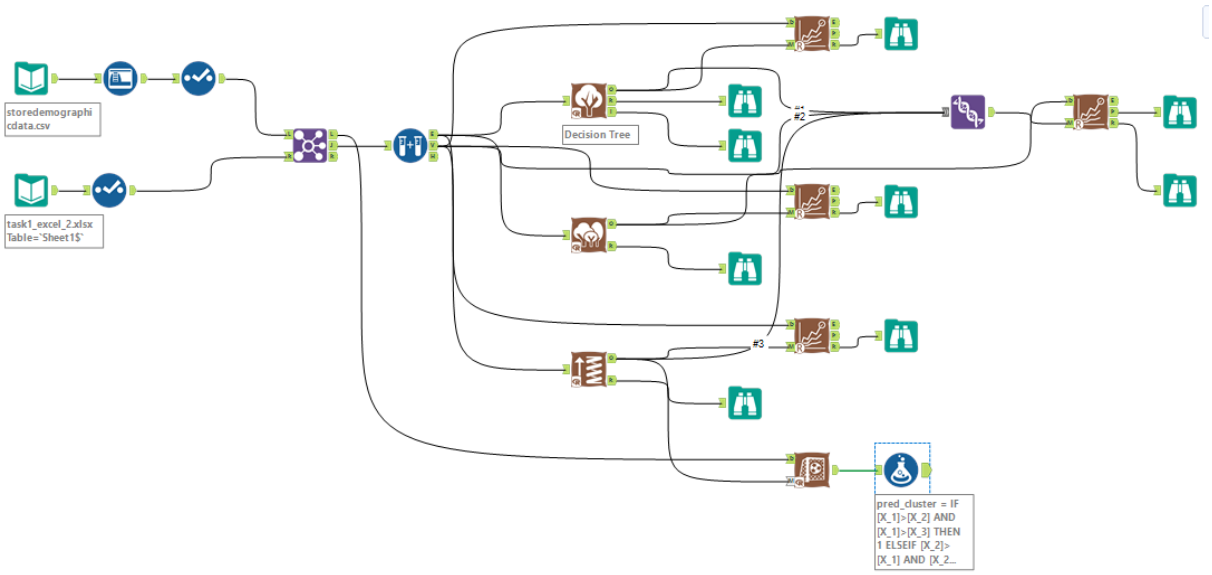


Figure 9: Alteryx Workflow Task 2

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?**

ETS(M,N,M) with no dampening should be used for the ETS model.

The seasonality shows an increasing trend and should be applied multiplicatively. The trend can't be derived, and nothing should be applied regarding this. The error of the plot is irregular and should therefore be applied multiplicatively.

Figure 10: Decomposition Plot Part 1

ARIMA(0,1,2)(0,1,0) is used because, the seasonal difference and seasonal first difference shows that there is a lag at -2.

Figure 11: Decomposition Plot Part 2

The ETS model's accuracy is higher when compared to the ARIMA model. A holdout sample of 6 months data was used. the RMSE of 1,020,597 is smaller than the ARIMA's 1,429,296 while its MASE is 0.45 and therefor lower compared to the ARIMA's 0.53. in addition, the ETS also has a higher AIC at 1,283 while ARIMA's AIC is 859.

ETS:

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

Figure 12: ETS Model Measures

ARIMA:

Information Criteria:

AIC	AICc	BIC
858.7774	859.8209	862.665

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.054315	1429296.2983494	951432.2560696	0.6151859	4.2022854	0.531117	-0.0260961

Figure 13: ARIMA Model Measures

After using the ETS for the forecasts the following solution was calculated:

Period	Sub_Period	Prognose	Prognose_high_95	Prognose_high_80	Prognose_low_80	Prognose_low_95
4	11	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
4	12	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
5	1	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981697
5	2	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
5	3	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
5	4	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
5	5	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
5	6	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
5	7	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
5	8	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
5	9	21177435.485839	23994279.191514	23019270.585553	19335600.386124	18360591.780163
5	10	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

Figure 14: Table of forecasted values

The Graph shows the actual and forecast value with a confidence interval of 80& and 95%

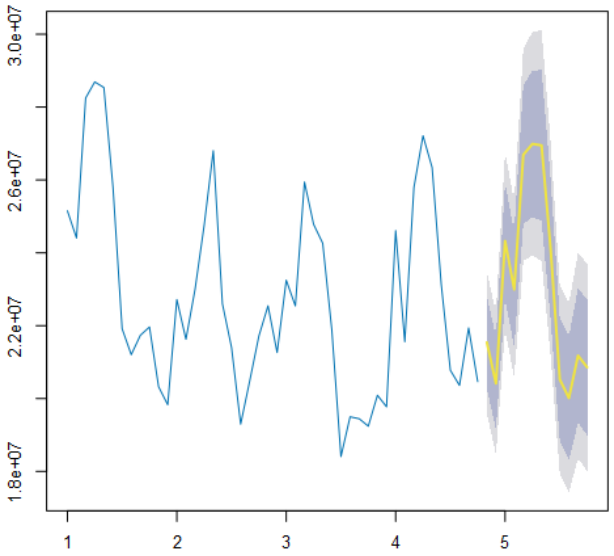
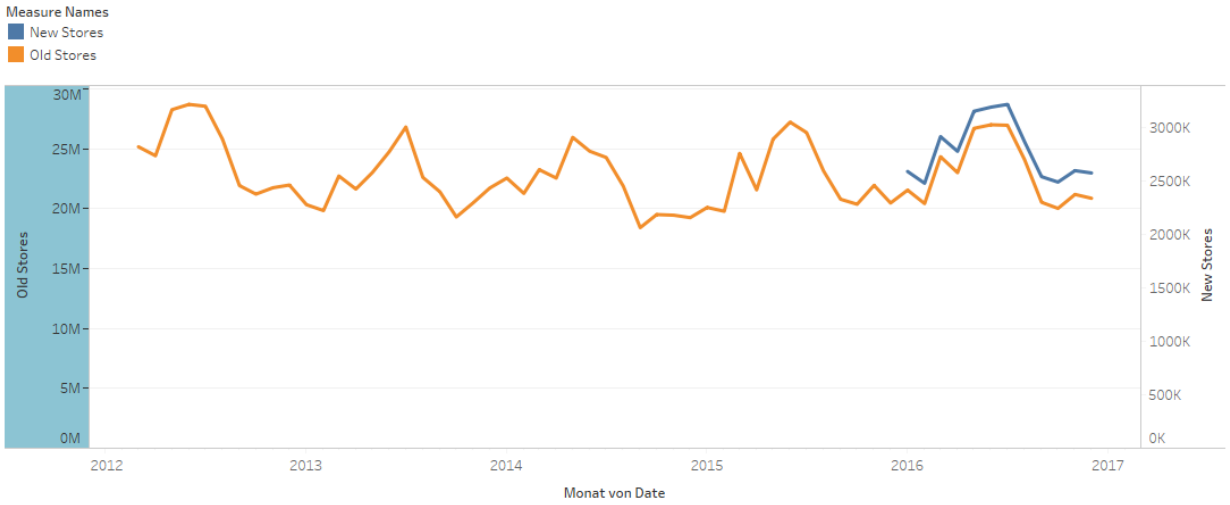


Figure 15: Graph of forecasted values

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.**

The table and graph below show the forecasted sales for existing stores and new stores. New store sales is obtained by using the ETS(M,N,M) analysis with all individual clusters to obtain the average sales per store.

The average sales value (3x cluster 1, 6x cluster 2, 1x cluster 3) are accumulated to calculate the new store sales.



Year	Month	New Stores	Old Stores
2016	1	2.587.451	21.539.936
	2	2.477.353	20.413.771
	3	2.913.185	24.325.953
	4	2.775.746	22.993.466
	5	3.150.867	26.691.951
	6	3.188.922	26.989.964
	7	3.214.746	26.948.631
	8	2.866.349	24.091.579
	9	2.538.727	20.523.492
	10	2.488.148	20.011.749
	11	2.595.270	21.177.435
	12	2.573.397	20.855.799

Figure 16: Tableau Extract

Tableau Link:
https://public.tableau.com/profile/aljoscha.grunwald#!/vizhome/Task3_283/Task3

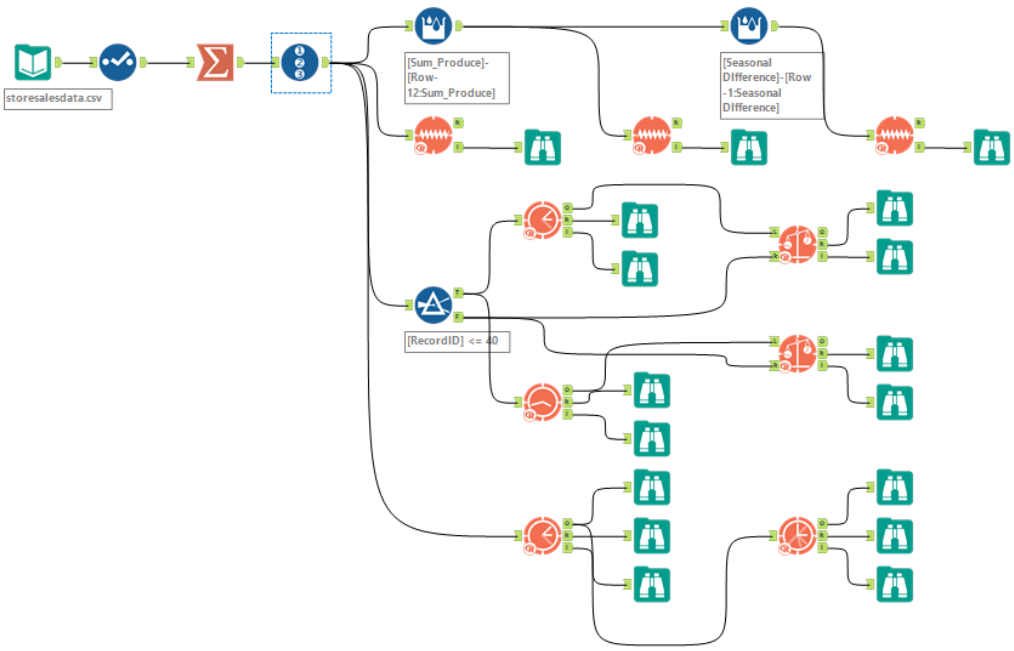


Figure 17: Alteryx Workflow Task 3