Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.016485	0.27351	0.31976	0.274316	0.235718
1st Quartile	0.35943	0.594017	0.46406	0.39294	0.377774
Median	0.544023	0.705326	0.53195	0.456588	0.421798
Mean	0.524263	0.69161	0.548167	0.470346	0.435429
3rd Quartile	0.694147	0.800179	0.635682	0.520656	0.493589
Maximum	0.952939	0.969034	0.942222	0.841981	0.677532

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	17.281	17.38103	18.89398	16.69676	15.71092
1st Quartile	28.22121	29.21236	25.03471	22.86498	21.10249
Median	29.4157	31.14178	26.33467	24.22188	21.96958
Mean	28.56936	30.07118	26.18037	23.72205	21.92474
3rd Quartile	30.21867	32.17467	27.4999	25.09459	22.95561
Maximum	31.71569	33.63781	30.1583	26.63063	24.72038

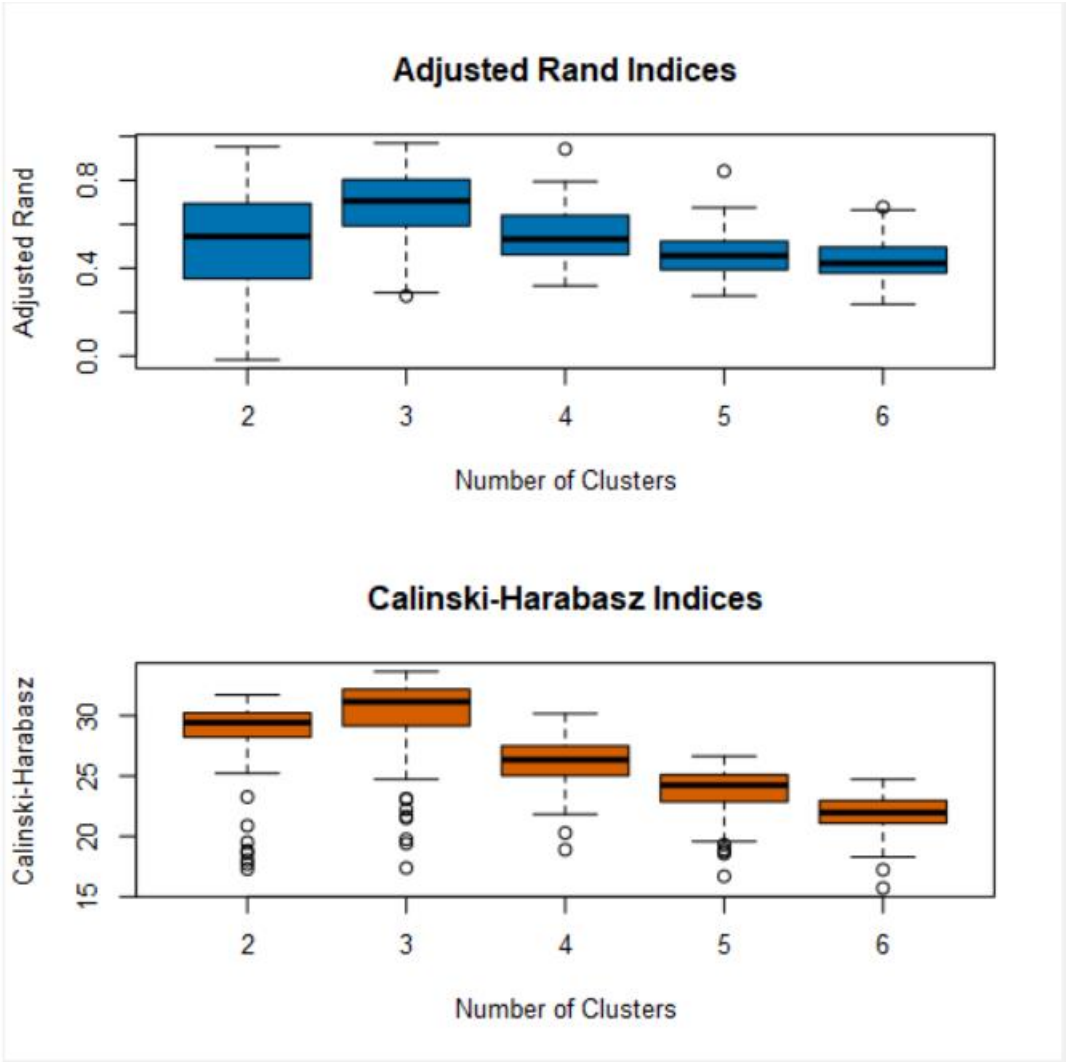


Figure 1. K-Means Cluster Assessment Report

According to the K-Means analysis, both Adjusted Rand Indices and Calinski-Harabasz Indices shows highest median value at 3, indicating that the optimal number of store formats is 3.

2. How many stores fall into each store format?

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

According to the Cluster Information, Cluster 1 has 23 stores, cluster 2 has 29 stores and cluster 3 has 33 stores.

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

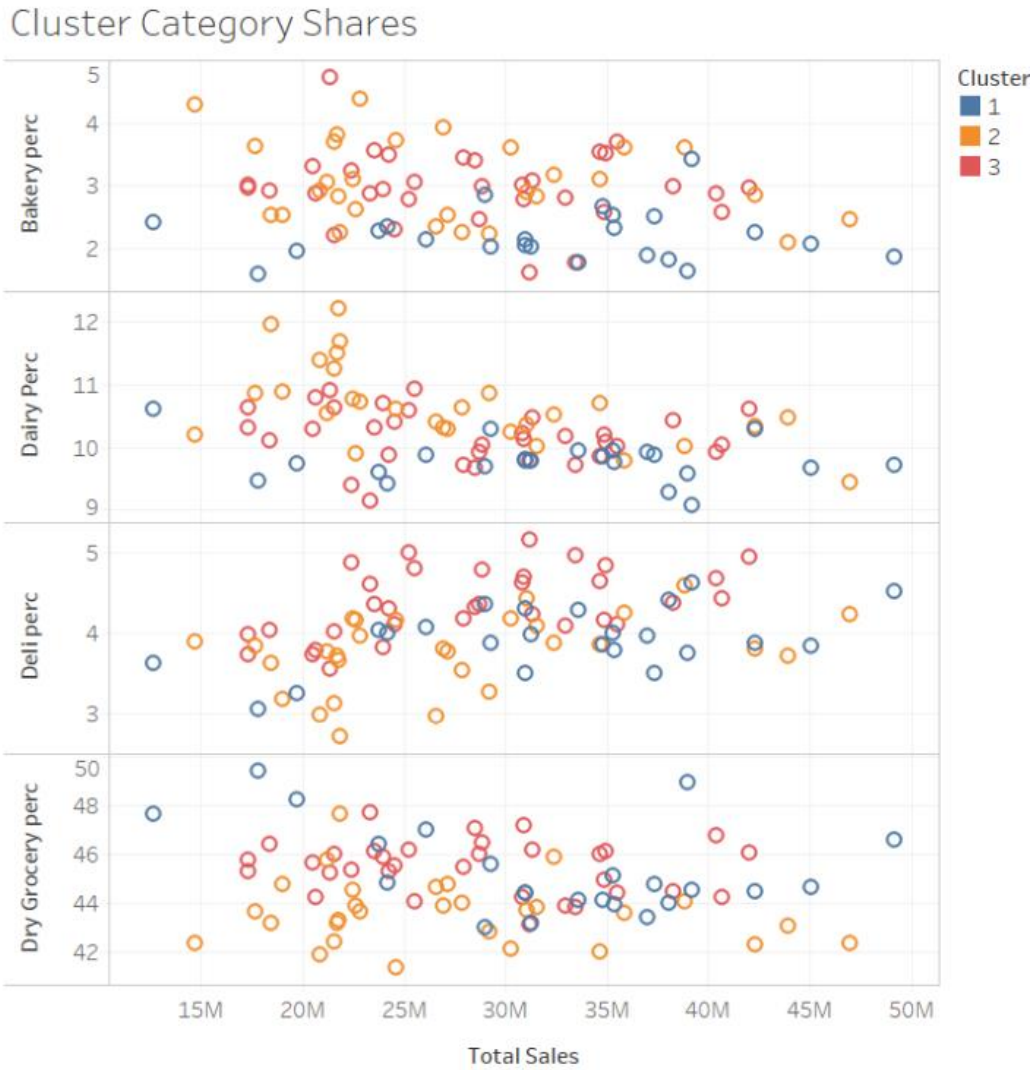


Figure 2. Cluster Category Shares

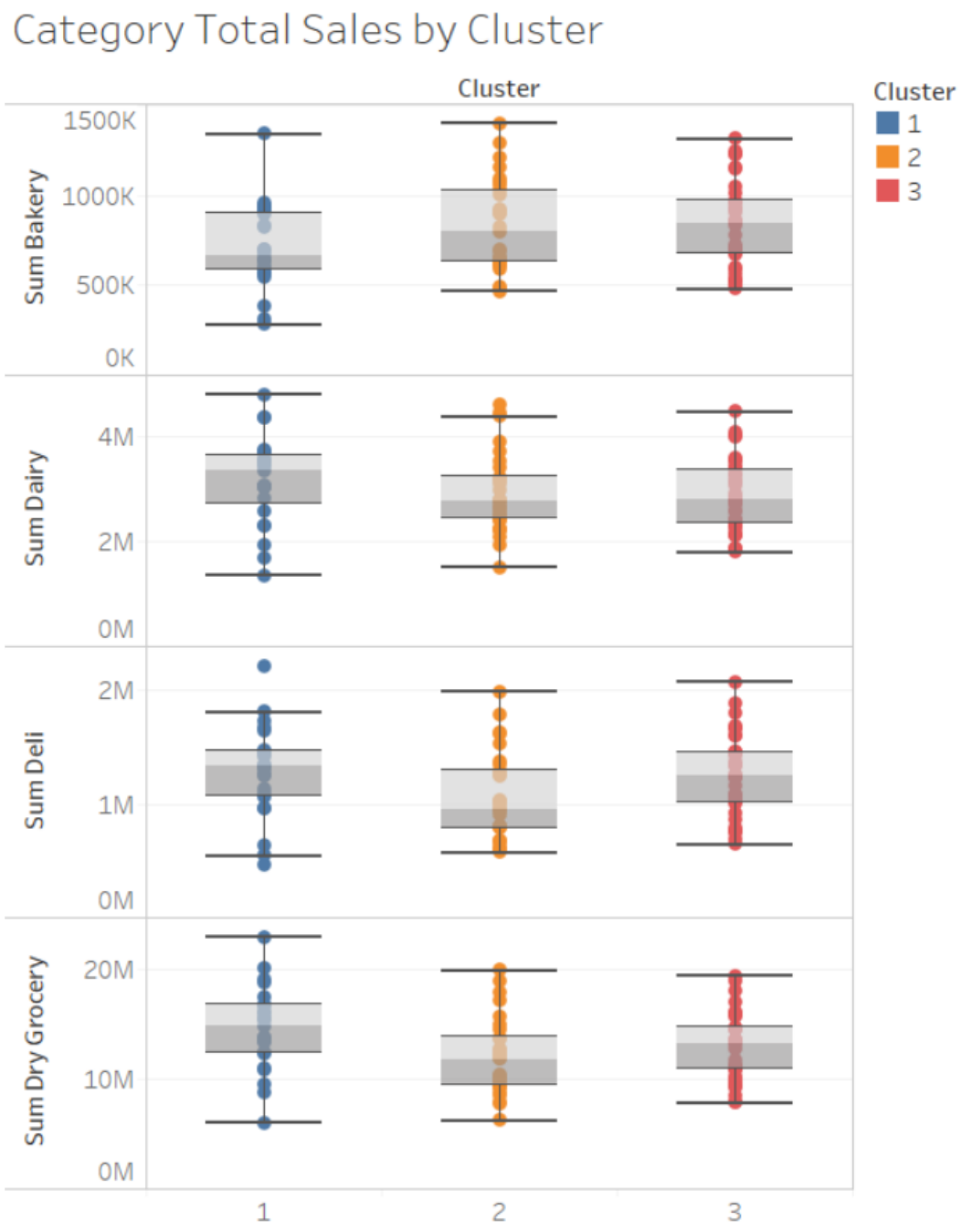


Figure 3. Category Total Sales by Cluster

Based on the Category Share by cluster and category total sales by cluster plots, Cluster 1 has the widest range of the sales and its sales are largest for most categories. Cluster 2 has the highest percentage sales in Diary, while Cluster 3 has highest percentage sales in Deli.

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.

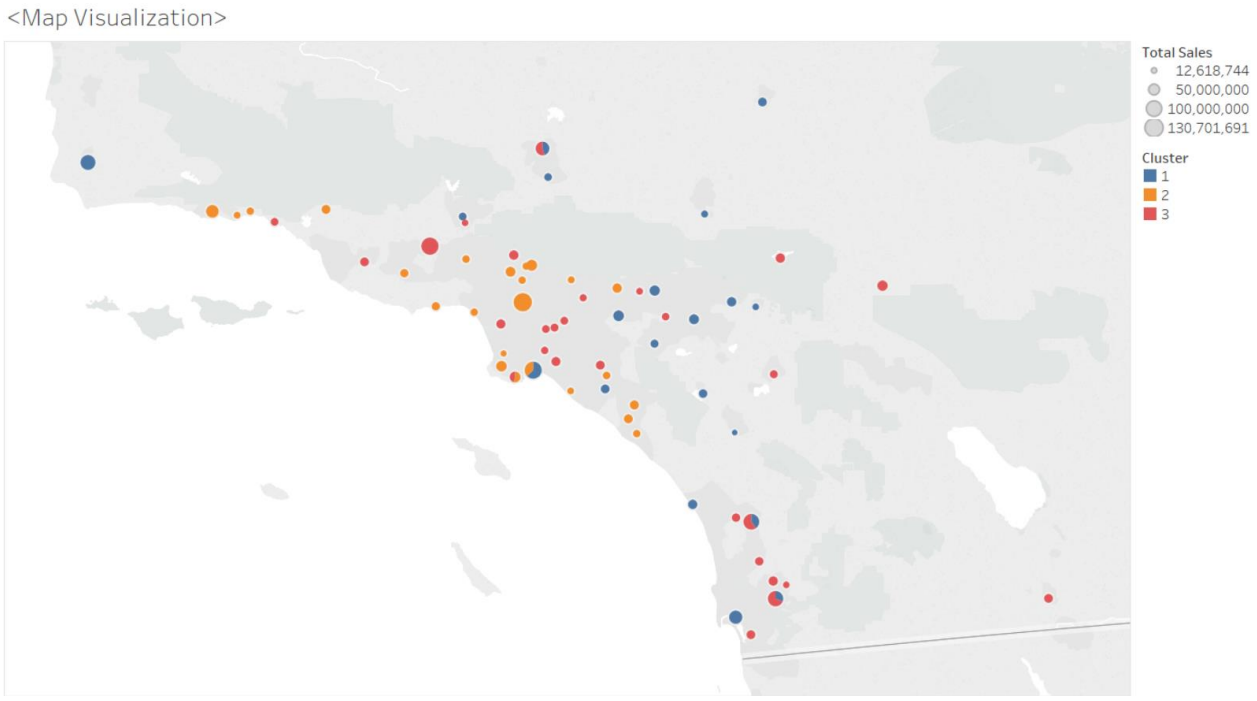


Figure 4. Store Location and Total Sales (Tableau)

Tableau Profile: https://public.tableau.com/shared/SWZNTNK79?:display_count=yes

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.)

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest	0.8235	0.8426	0.7500	1.0000	0.7778
Decision	0.7647	0.7593	0.5000	1.0000	0.7778
Boosted	0.8235	0.8889	1.0000	1.0000	0.6667

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

	Actual_1	Actual_2	Actual_3
Predicted_1	2	0	0
Predicted_2	0	4	2
Predicted_3	2	0	7

Confusion matrix of Forest

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

Figure 5. Model Comparison Report

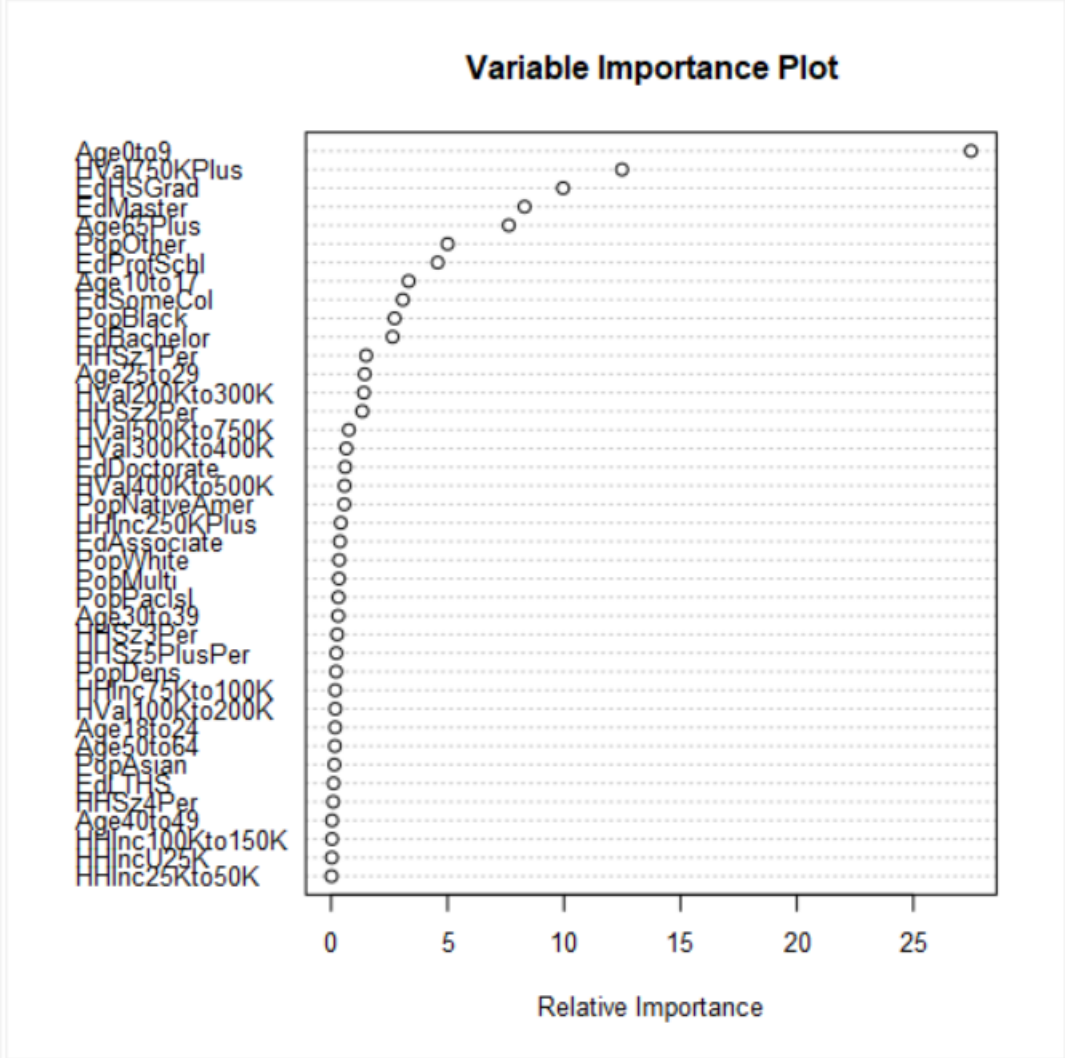


Figure 6. Variable Importance Plot

Based on the results from the Model Comparison Report, Boosted Model is appropriate to use for classifying format of the new stores. Forest Model had the same overall accuracy of 0.8235, but Boosted Model was chosen because it has higher F1 value.

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
S0095	2

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?

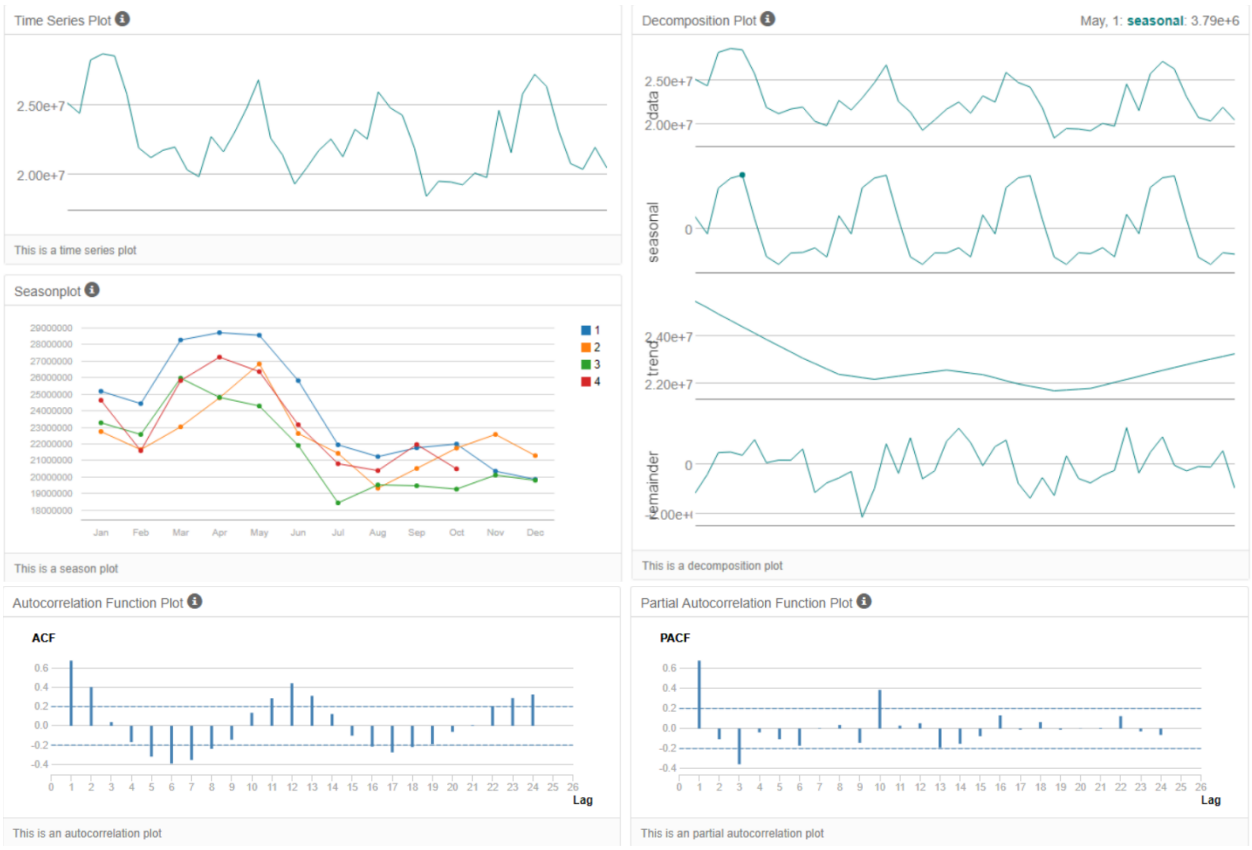


Figure 7. TS Plot

The Error (Remainder) shows irregular pattern so it should be applied

multiplicatively. Trend does not show any linear or exponential behavior, so nothing should be applied for Trend. The Seasonality shows increasing trend; therefore, it should be applied multiplicatively. Thus, ETS(M,N,M) with no dampening should be used for ETS model.

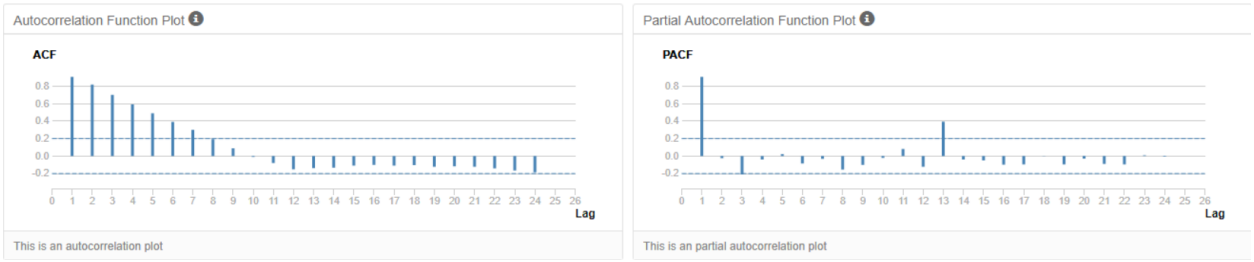


Figure 8. ACF and PCF after adjusting Seasonal Difference

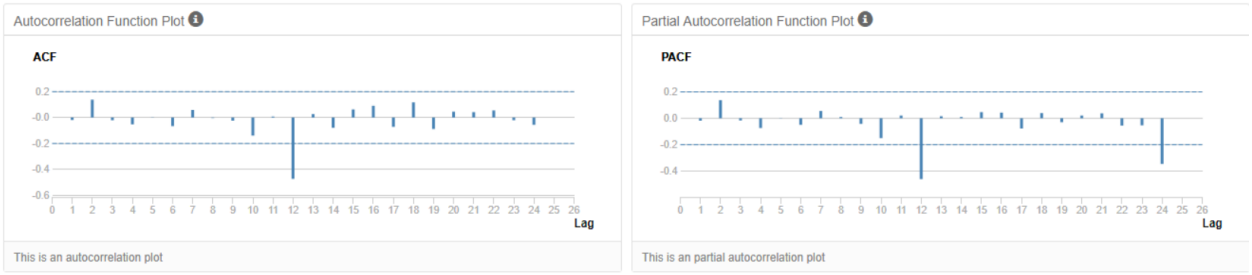
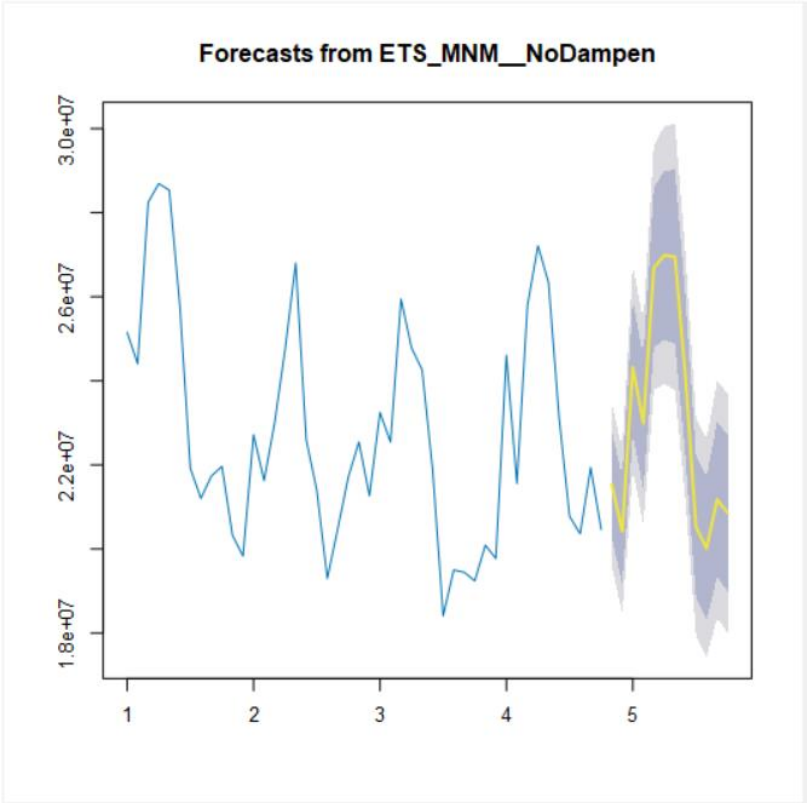


Figure 9. ACF and PCF after adjusting First Seasonal Difference



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_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 10. 12 Period Forecast from ETS(MNM) Model

Actual and Forecast Values:

Actual	ETS_MNM_NoDampen	ARIMA_012__010_
26338477.15	26907095.61191	25515002.53492
23130626.6	22916903.07434	22982398.33693
20774415.93	20342618.32222	19509673.05693
20359980.58	19883092.31778	20599981.42693
21936906.81	20479210.4317	20547162.64693
20462899.3	21211420.14022	20342794.22693

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_MNM_NoDampen	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
ARIMA_012__010_	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909

Figure 11. Model Comparison

Based on the ACF and PCF changes after adjusting seasonal and first seasonal difference, it can be concluded that ARIMA(0,1,2)(0,1,0) model is used.

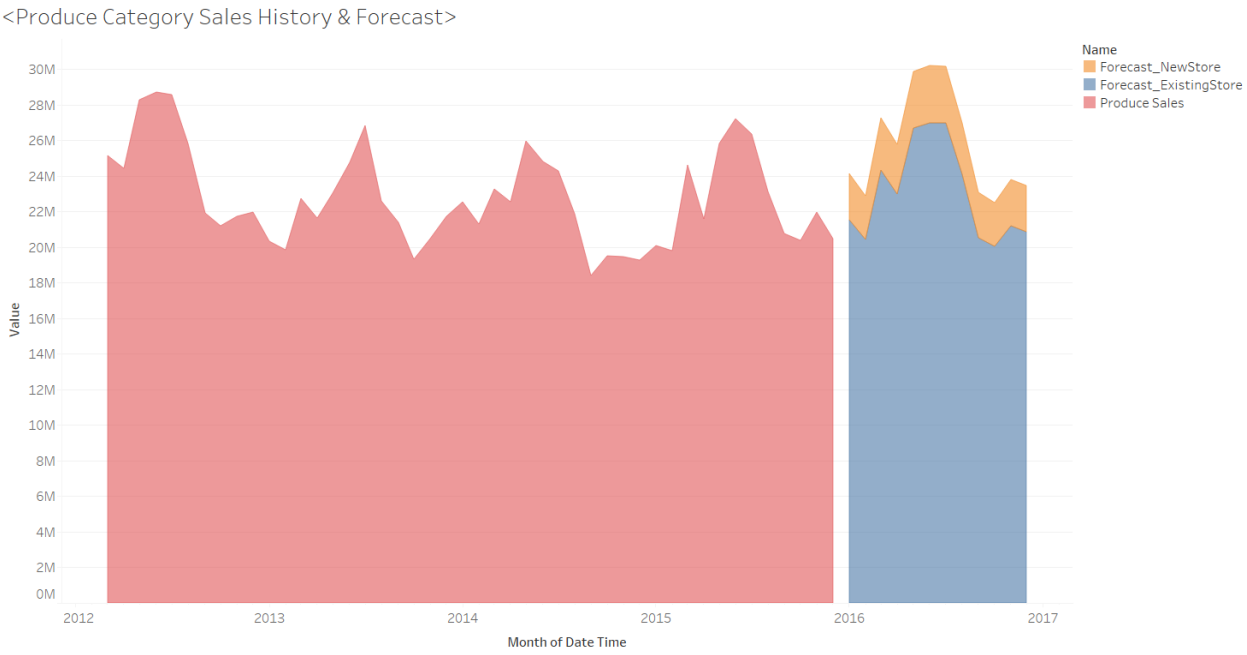
After comparing all the possible ETS and ARIMA model for forecasting the produce sales, we see that RMSE(760267.3) and MASE(0.3822) value for the ETS model is lower than that of ARIMA Model: RMSE (584382.4) MASE (0.3909).
We conclude that ETS (M,N,M) – No Dampening is the most appropriate forecast model.

Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

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.

Year	Month	Forecast_ExistingStore	Forecast_NewStore	Total_Forecast
2016	1	$ 21,539,936	$ 2,587,451	$ 24,127,387
2016	2	$ 20,413,771	$ 2,477,353	$ 22,891,123
2016	3	$ 24,325,953	$ 2,913,185	$ 27,239,138
2016	4	$ 22,993,466	$ 2,775,746	$ 25,769,212
2016	5	$ 26,691,951	$ 3,150,867	$ 29,842,818
2016	6	$ 26,989,964	$ 3,188,922	$ 30,178,886
2016	7	$ 26,948,631	$ 3,214,746	$ 30,163,376
2016	8	$ 24,091,579	$ 2,866,349	$ 26,957,928
2016	9	$ 20,523,492	$ 2,538,727	$ 23,062,219
2016	10	$ 20,011,749	$ 2,488,148	$ 22,499,897
2016	11	$ 21,177,435	$ 2,595,270	$ 23,772,706
2016	12	$ 20,855,799	$ 2,573,397	$ 23,429,196

Table 1. Sales forecast fir existing and new stores

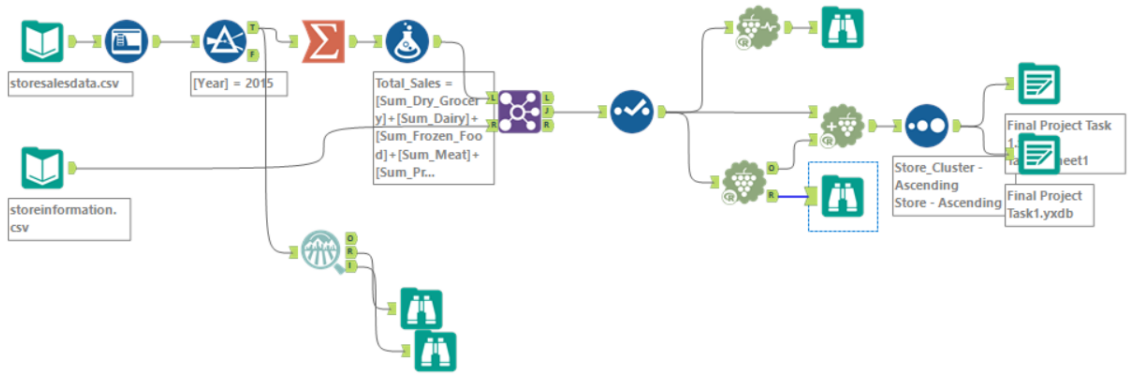


The chart shows the historical and forecast sales of Produce categoty for existing stores and new stores from 2013 to 2016.

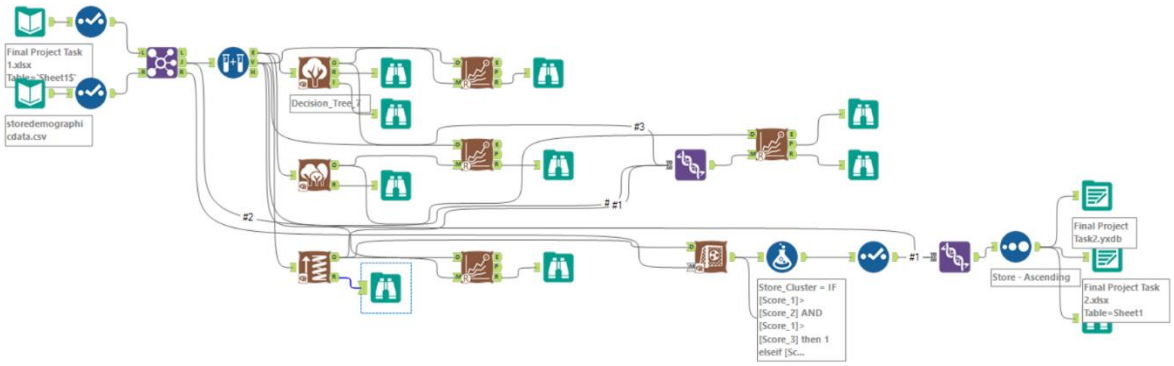
Tableau Profile
https://public.tableau.com/views/Task3_301/Sheet1?:embed=y&:display_count=yes

Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

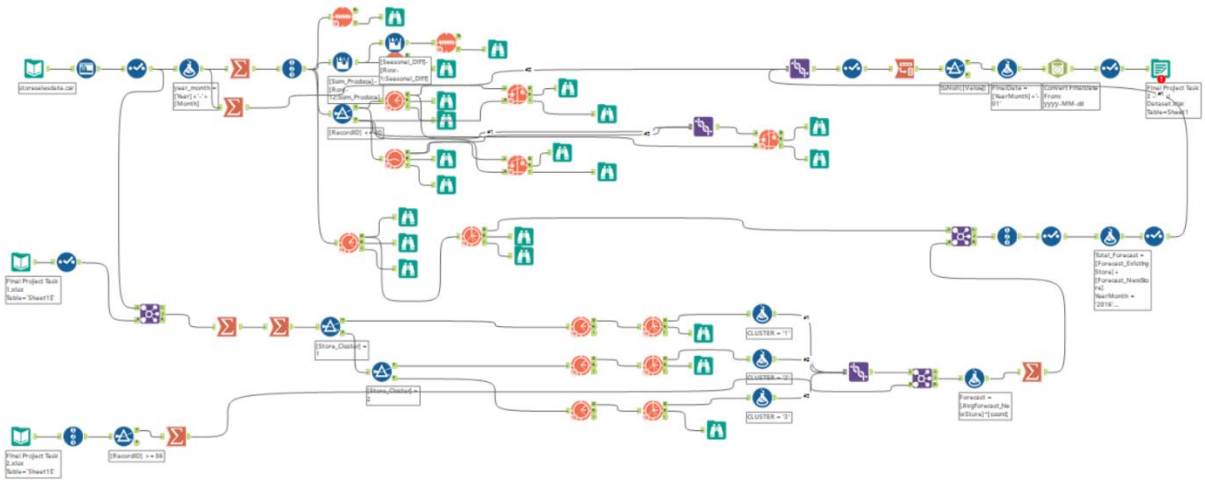
Alteryx Workflow



Workflow 1: Alteryx Workflow for Task 1



Workflow 2: Alteryx Workflow for Task 2



Workflow3: Alteryx Workflow for Task 3