<u>Project: Predictive Analytics Capstone</u>

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

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 the store formats is 3.

   As shown in the following chart, cluster 3 has higher median and mean in both AR and CH methods compared to all other cluster number except 2. Based upon my version of alteryx, cluster 2 is also very competitive. For AR, cluster 3 has better mean. For CH, cluster 2 is better than cluster 3 in both mean and median performance.

Report

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

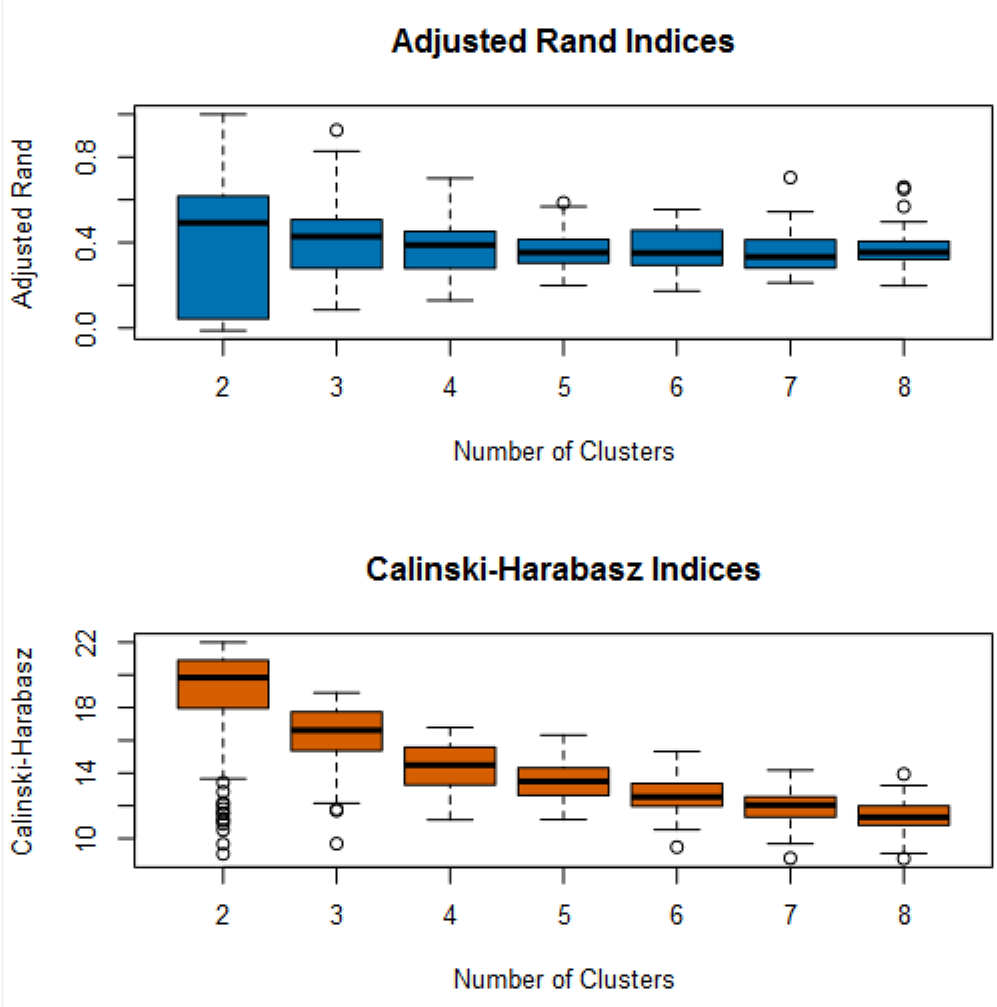
	2	3	4	5	6	7	8
Minimum	-0.012332	0.085005	0.129167	0.198479	0.172868	0.211424	0.197457
1st Quartile	0.055047	0.28273	0.279896	0.303745	0.294079	0.281472	0.321616
Median	0.492542	0.428163	0.388131	0.353296	0.351385	0.333331	0.353529
Mean	0.406457	0.411914	0.372189	0.366041	0.367644	0.354859	0.369188
3rd Quartile	0.61678	0.50506	0.450843	0.41474	0.453322	0.409187	0.404819
Maximum	1	0.925732	0.70085	0.586379	0.5548	0.703966	0.660004

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	9.056197	9.683921	11.14097	11.15269	9.474469	8.797239	8.769803
1st Quartile	17.976426	15.402516	13.27496	12.65426	11.988572	11.311079	10.838622
Median	19.836525	16.618434	14.49044	13.49543	12.537825	12.043325	11.303199
Mean	18.604945	16.309418	14.37112	13.46494	12.624375	11.910413	11.376818
3rd Quartile	20.889876	17.734502	15.56523	14.30924	13.365637	12.535052	11.963996
Maximum	21.992647	18.908142	16.79342	16.32568	15.329887	14.179165	13.936724

However, based upon the graphs below, we could see cluster 3 has apparently lower variance than cluster 2. Cluster 3 is much more compact than cluster 2 in AR method. In CH method, Cluster 3 is still more compact. And cluster 2 has more outliers.

Therefore, I will use 3 as the number of optimal clusters.



2. How many stores fall into each store format?

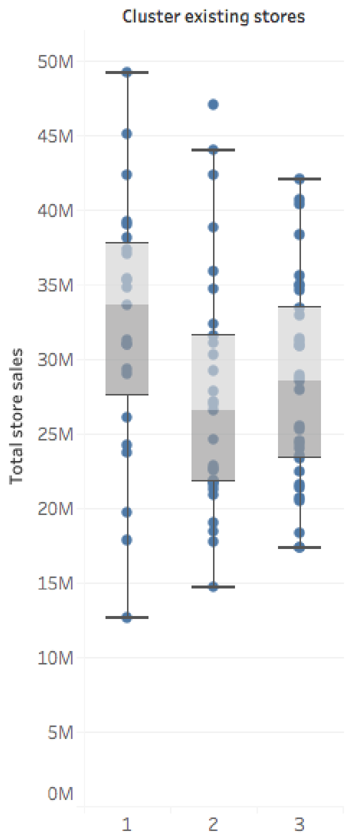
   According to the summary chart below, there are 23 stores in store format 1, 29 stores in format 2, and 33 stores in format 3.

Record #	Cluster_existing_stores	Count
1	1	23
2	2	29
3	3	33

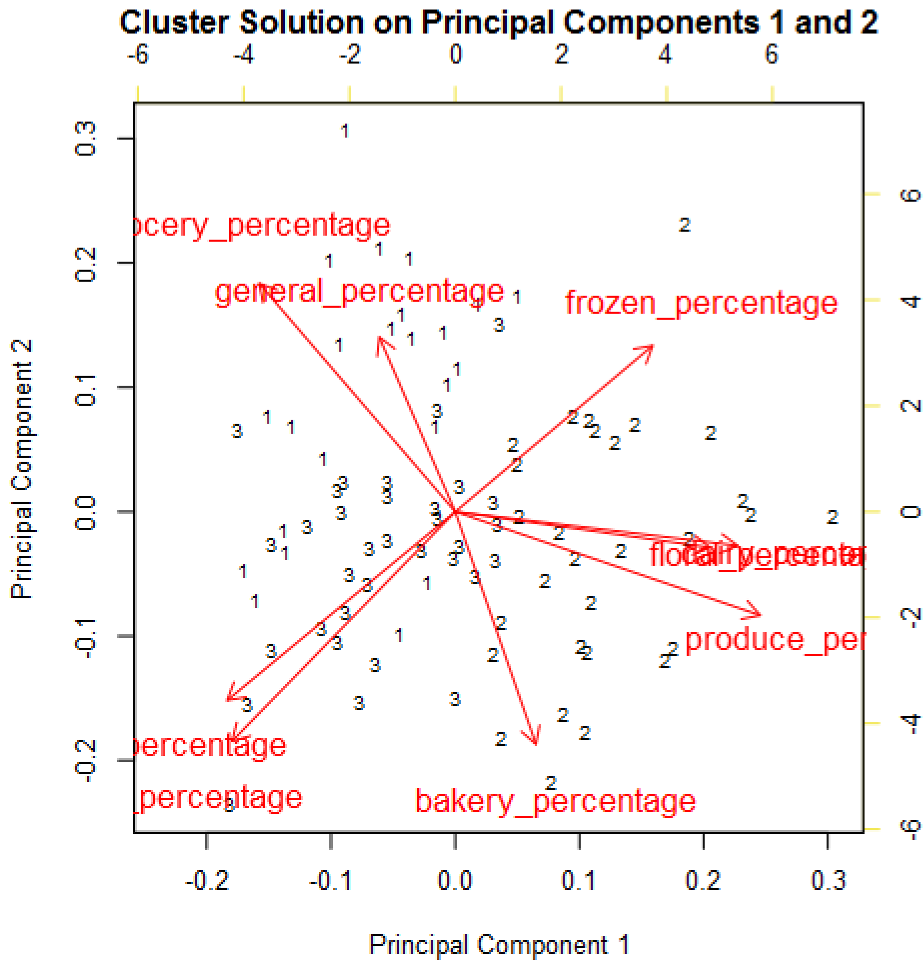
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

   Based upon the below box-whisker graph, we could see the median of total store sales of the 3 clusters are different.

box-whister

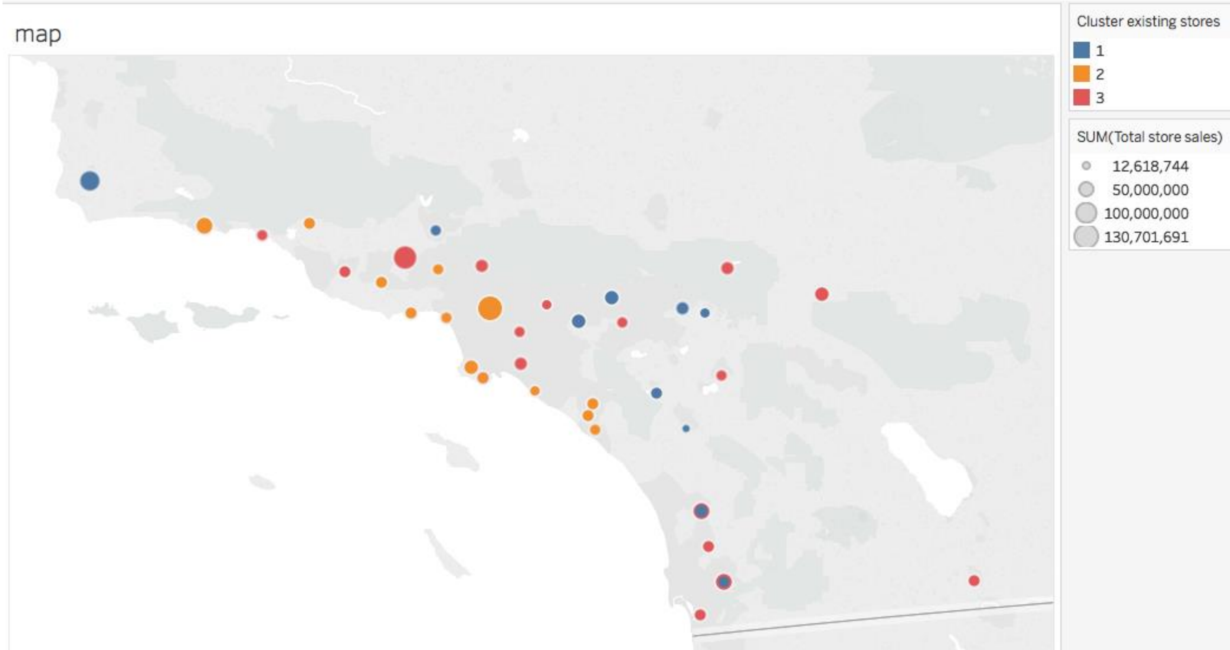


Based upon the cluster information from the report of K-Centroids tool, we could visualized the comparison as below. It indicates the difference in product mixes of different store formats.



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.

See as below.



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.)

For the decision method, the Root node error is 0.63235. The confusion matrix below shows 66.7% positive prediction is correct and 78.9% negative prediction is correct.

Actual	Actual Positive	Actual Negative
Predicted Positive	24 (66.7%)	12 (33.3%)
Predicted Negative	4 (21.1%)	15 (78.9%)

To review the below summary chart, the decision tree method has a 76.5% total accuracy.

Accuracy **76.5 %**

Proportion of correct predictions in the data

F1_Score **73.2 %**

Harmonic mean of Recall and Precision

Precision **68.2 %**

Proportion of values predicted positive, that were actually positive

Recall **78.9 %**

Proportion of values actually positive, that were predicted positive

By model comparison report, we could see the overall accuracy is 70.59%.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_task2	0.7059	0.7685	0.7500	1.0000	0.5556

We could also see for decision tree, both cluster 1 and cluster 2 are relatively difficult to predict.

Confusion matrix of Decision_Tree_task2

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

For forest model, the OOB estimate of the error rate is 25%. The forest model did better in prediction cluster 2 than cluster 1 or 3.

OOB estimate of the error rate: 25%

Confusion Matrix:

	Classification Error	1	2	3
1	0.368	12	1	6
2	0.08	0	23	2
3	0.333	5	3	16

By model comparison report, we could see the overall accuracy is 82.35%.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
forest_task2	0.8235	0.8426	0.7500	1.0000	0.7778

Confusion matrix of forest_task2

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

For the boosted model, we could see from the model comparison report, the overall accuracy is 82.35%. The model did very good in clustering 1 and 2.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
boosted_task2	0.8235	0.8889	1.0000	1.0000	0.6667

Confusion matrix of boosted_task2

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

Overall, we could see, for the three models, Forest and the Boosted have the same overall accuracy, which is better than the Decision Tree Model. The Boosted has the same good performance of prediction cluster 2 as Forest. But the Boosted is better in predicting cluster 1 than Forest, but worse in predicting cluster 3 than Forest.

So I put all the three models in model comparison tools together to compare. As showed in the below charts, besides what we observed above regarding the accuracy, the Boosted Model has less bias showed in the confusion matrix. Therefore, I will use Boosted Model.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
forest_task2	0.8235	0.8426	0.7500	1.0000	0.7778
boosted_task2	0.8235	0.8889	1.0000	1.0000	0.6667
Decision_Tree_task2	0.7059	0.7685	0.7500	1.0000	0.5556

Confusion matrix of Decision_Tree_task2			
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of boosted_task2			
	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Confusion matrix of forest_task2			
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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

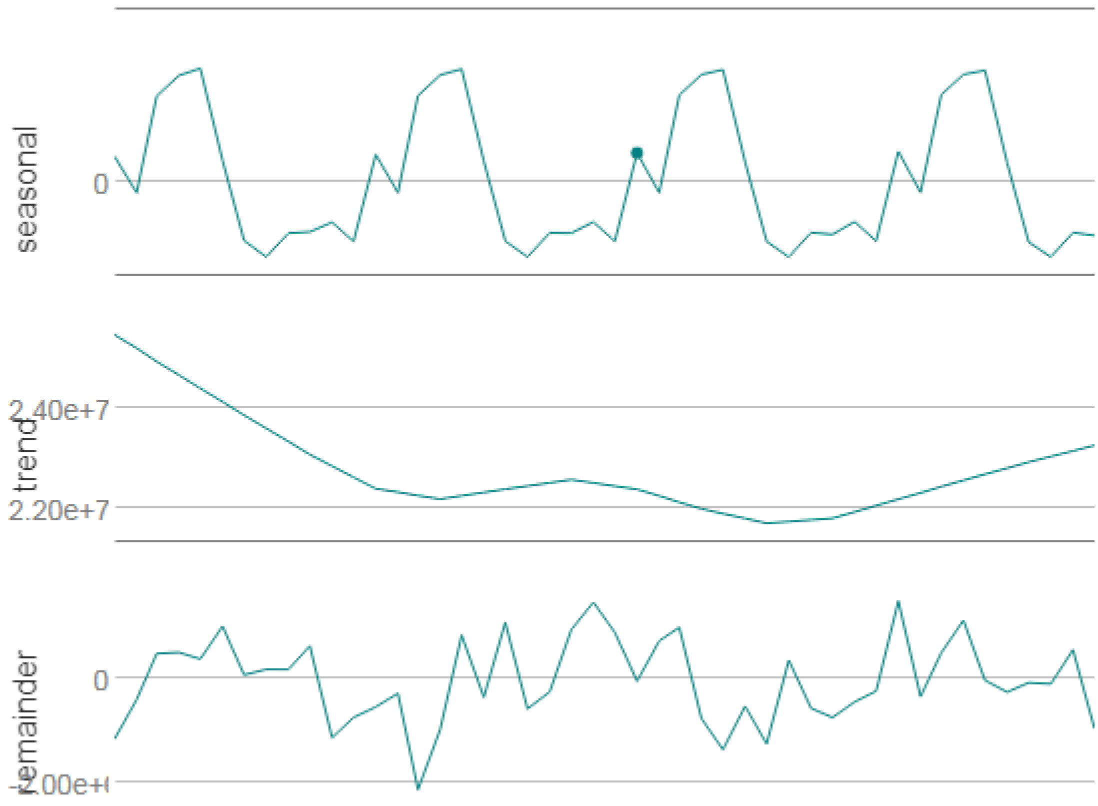
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

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?
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.

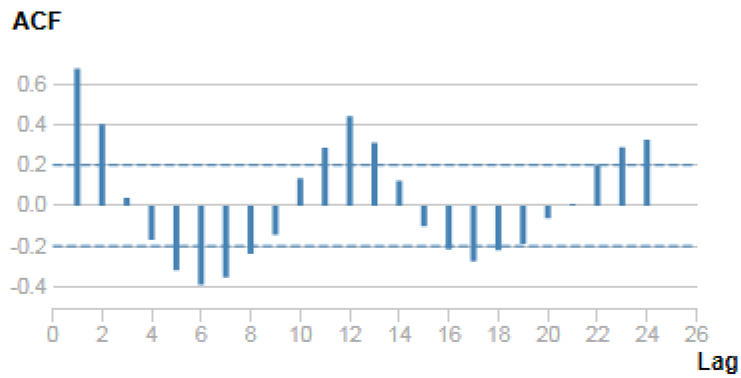
1.
For the forecast of existing stores, here goes my choice:

By applying TS plot tool, we could see from the decomposition report that the trend does move in a linear fashion or an exponential trend. Since it moves not like a trend line, I will use none. The seasonality looks constant. But when showing the exact number, we could see a slightly decrease. Therefore, I will use multiplicative; the error shows changing variance as the time series moves along, suggesting applying multiplicative. I will not try damped since the trend is none. An auto setting model was tried to test, and further confirmed my settings. Therefore, the best fit ETS model is ETS (M,N,M).

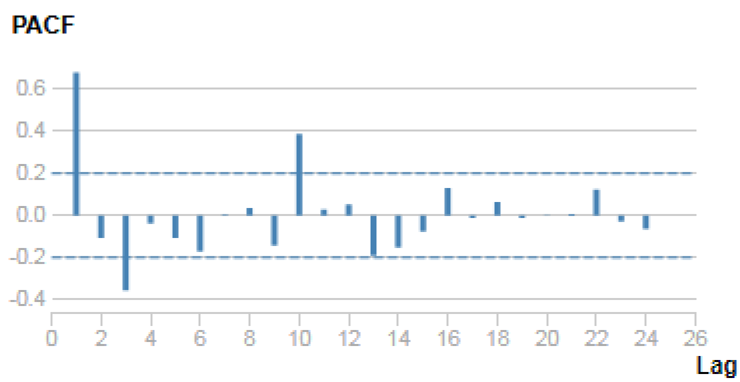


To configure ARIMA, we need to see ACF and PACF plots. By the TS plot function, we could see the plots as below. The stationarized series has a positive correlation at lag-1 in ACF, indicating AR rather than MA. The PACF shows only one spike in lag-1, indicating AR as well. There is no other significant lags in either ACF or PACF, indicating an AR(1) model.

Autocorrelation Function Plot



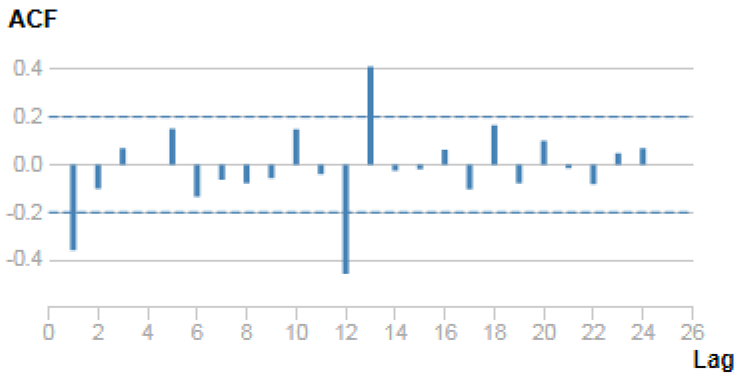
Partial Autocorrelation Function Plot



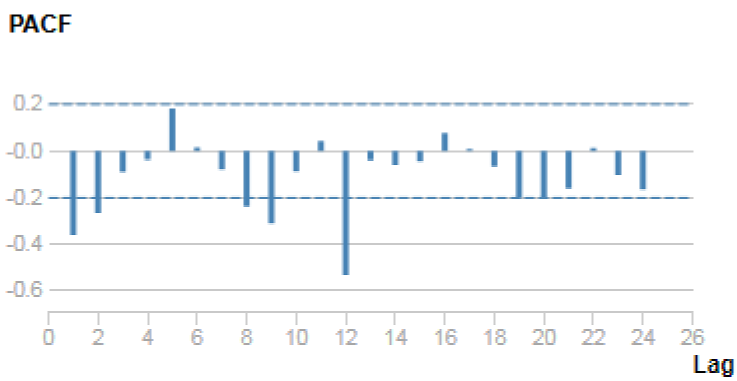
Further, by checking TS plot for seasonal differencing, I used first seasonal differencing and then see no serial correlation in ACF and PACF plots. It also means d and D will be 1. The seasonal autocorrelation is negative, indicating MA. The ACF cuts off to zero, while the PACF has spikes decaying towards zero, both indicating MA. The ACF plot shows a negative autocorrelation at lag 1 and it is confirmed in the PACF, indicating q=1. In the ACF and PACF plots of the seasonal first difference of produce data, I see a significant negative lag at 12 and cuts off to 0 at lag 24, suggesting Q=1. And since our period is 1 year, m=12.

To compare an additional model, I copied all other parameters for another ARIMA except setting Q to 2 in this one to ensure the seasonal component is accounted for. I also set up an auto model to compare.

Autocorrelation Function Plot



Partial Autocorrelation Function Plot



Compare the three ARIMA models, we could see ARIMA(0,1,1)(0,1,1)[12] has RMSE as 935292.1712234; ARIMA(0,1,1)(0,1,2)[12] has RMSE as 763923.4295347; the auto set ARIMA(1,0,0)(1,1,0)[12] has an RMSE as 1042209.8528363. Therefore the Q=2 model has a much smaller RMSE and a slightly higher AIC than the Q=1 model.

By using TS compare tools, we compare ETS (M,N,M) and ARIMA(0,1,1)(0,1,2)[12] models.

For the ETS (M,N,M), the forecast error measurements against the holdout sample are:

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_MNM_task3_cluster1	-1725.3	7187.581	5539.896	-0.6531	2.09	0.2534

For ARIMA (0,1,1)(0,1,2)[12], the forecast error measurements against the holdout sample are:

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_TASK3_Q2	-990040.7	1310865	1155297	-4.3955	5.1488	0.6798

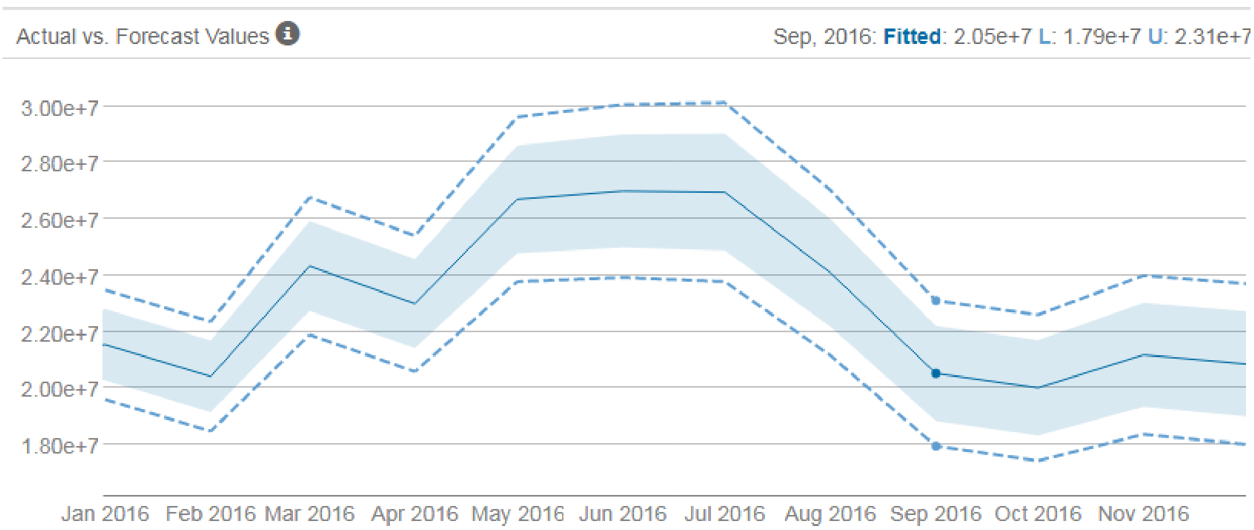
Comparing the MAPE and ME of the two models upon the forecast error measurements against the holdout sample, we could see ETS is doing better. I will use ETS(M, N, M) to forecast.

3. For forecast tables, here goes the analysis.

The forecast for 2016 of existing stores are:

Period	Sub_Period	forecast_task3_existing
2016	1	21539936.007499
2016	2	20413770.60136
2016	3	24325953.097628
2016	4	22993466.348585
2016	5	26691951.419156
2016	6	26989964.010552
2016	7	26948630.764764
2016	8	24091579.349106
2016	9	20523492.408643
2016	10	20011748.6686
2016	11	21177435.485839
2016	12	20855799.10961

The graph using 95% and 80% confidence intervals is:



The forecast for 2016 average produce in cluster 1 is:

Period	Sub_Period	forecast_cluster1
2016	1	256056.032949
2016	2	244548.923224
2016	3	293254.587434
2016	4	275841.952548
2016	5	314668.287235
2016	6	316655.428983
2016	7	318463.410907
2016	8	278092.991554
2016	9	247574.917662
2016	10	241544.741016
2016	11	254424.713942
2016	12	257905.506922

Since there are three new stores in cluster 1, the forecast for the new stores in cluster 1 will be:

Month	New Stores_cluster1_average	New Stores_cluster1_sum
16-Jan	256056.0329	768168.0988
16-Feb	244548.9232	733646.7697
16-Mar	293254.5874	879763.7623
16-Apr	275841.9525	827525.8576
16-May	314668.2872	944004.8617
16-Jun	316655.429	949966.2869
16-Jul	318463.4109	955390.2327
16-Aug	278092.9916	834278.9747
16-Sep	247574.9177	742724.753
16-Oct	241544.741	724634.223
16-Nov	254424.7139	763274.1418
16-Dec	257905.5069	773716.5208

The forecast for 2016 average produce in cluster 2 is

Period	Sub_Period	forecast_task3_cluster2
2016	1	265594.847766
2016	2	253264.72445
2016	3	295443.526216
2016	4	285116.608029
2016	5	321995.572552
2016	6	326046.639417
2016	7	329587.121571
2016	8	297122.98882
2016	9	263666.455329
2016	10	258452.686811
2016	11	268672.564962
2016	12	261568.455979

Since there are six new stores in cluster 2, the forecast for the new stores in cluster 2 will be:

Month	New Stores_cluster2_average	New Stores_cluster2_sum
16-Jan	265594.8478	1593569.087
16-Feb	253264.7245	1519588.347
16-Mar	295443.5262	1772661.157
16-Apr	285116.608	1710699.648
16-May	321995.5726	1931973.435
16-Jun	326046.6394	1956279.837
16-Jul	329587.1216	1977522.729
16-Aug	297122.9888	1782737.933
16-Sep	263666.4553	1581998.732
16-Oct	258452.6868	1550716.121
16-Nov	268672.565	1612035.39
16-Dec	261568.456	1569410.736

The forecast for 2016 average produce in cluster 3 is

Period	Sub_Period	forecast_task3_cluster3
2016	1	225713.666052
2016	2	224117.77602
2016	3	260760.316649
2016	4	237520.103949
2016	5	274888.538311
2016	6	282675.879908
2016	7	281832.684105
2016	8	249331.755812
2016	9	214003.363899
2016	10	212797.943546
2016	11	219960.854853
2016	12	230269.372411

Since there is one new stores in cluster 3, the forecast for the new stores in cluster 3 will be:

Month	New Stores_cluster3_average	New Stores_cluster3_sum
16-Jan	225713.6661	225713.6661
16-Feb	224117.776	224117.776
16-Mar	260760.3166	260760.3166
16-Apr	237520.1039	237520.1039
16-May	274888.5383	274888.5383
16-Jun	282675.8799	282675.8799
16-Jul	281832.6841	281832.6841
16-Aug	249331.7558	249331.7558
16-Sep	214003.3639	214003.3639
16-Oct	212797.9435	212797.9435
16-Nov	219960.8549	219960.8549
16-Dec	230269.3724	230269.3724

To sum all the ten new stores produce sales forecasts as:

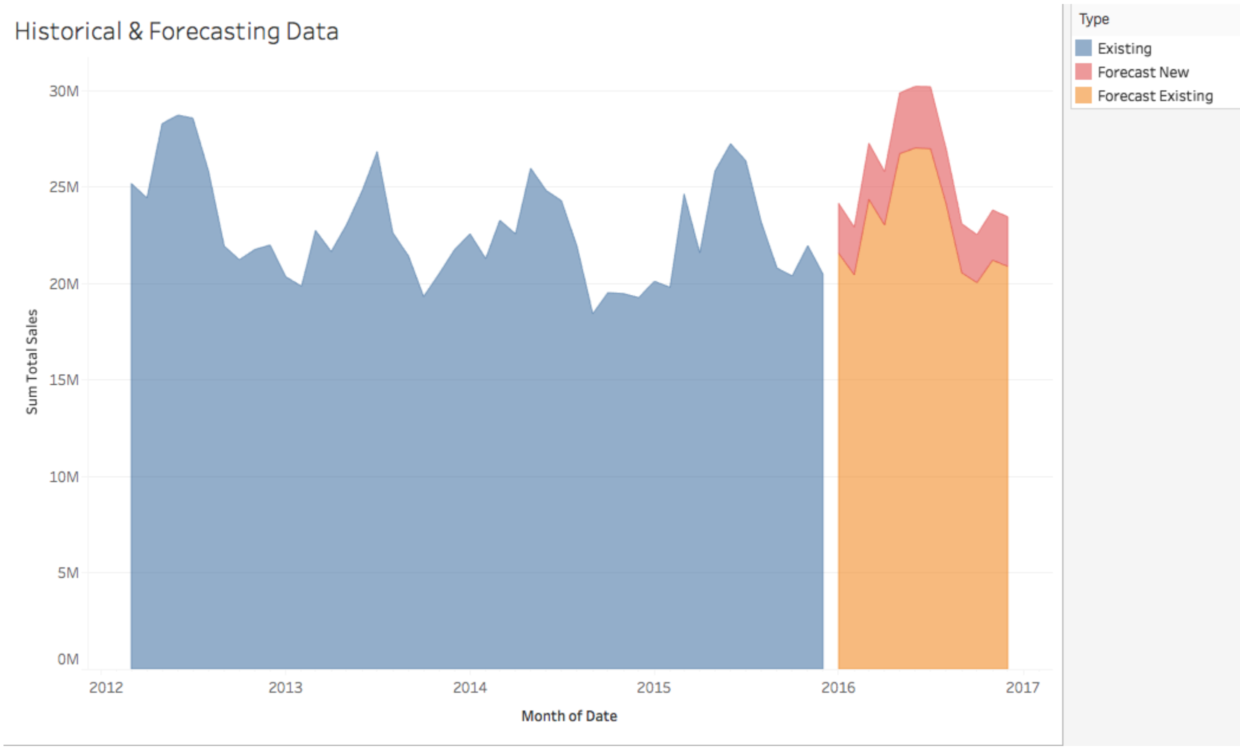
Month	New Stores_cluster1_sum	New Stores_cluster2_sum	New Stores_cluster3_sum	New Stores Sum
16-Jan	768168.0988	1593569.087	225713.6661	2587450.851
16-Feb	733646.7697	1519588.347	224117.776	2477352.892
16-Mar	879763.7623	1772661.157	260760.3166	2913185.236
16-Apr	827525.8576	1710699.648	237520.1039	2775745.61
16-May	944004.8617	1931973.435	274888.5383	3150866.835
16-Jun	949966.2869	1956279.837	282675.8799	3188922.003
16-Jul	955390.2327	1977522.729	281832.6841	3214745.646
16-Aug	834278.9747	1782737.933	249331.7558	2866348.663

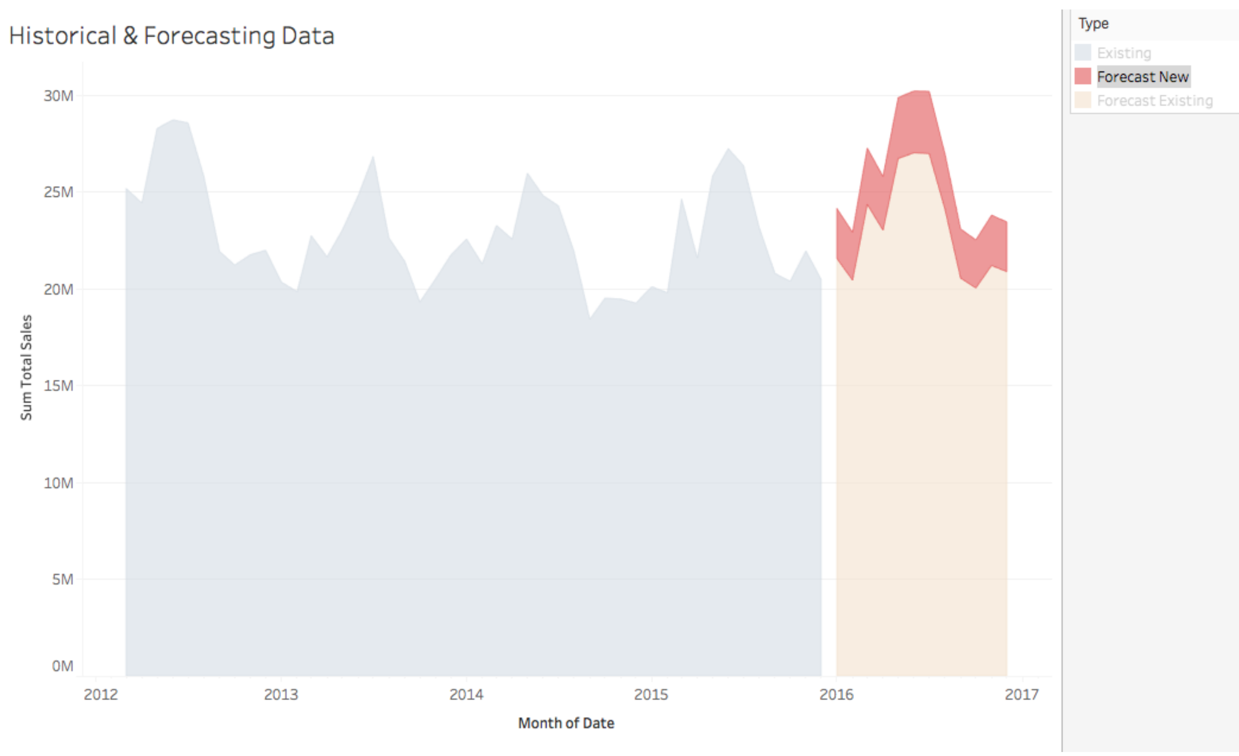
16-Sep	742724.753	1581998.732	214003.3639	2538726.849
16-Oct	724634.223	1550716.121	212797.9435	2488148.287
16-Nov	763274.1418	1612035.39	219960.8549	2595270.386
16-Dec	773716.5208	1569410.736	230269.3724	2573396.629

The forecast table for existing and new stores will be:

Month	New Stores	Existing Stores
1-Jan	2587450.851	21539936.01
1-Feb	2477352.892	20413770.6
1-Mar	2913185.236	24325953.1
1-Apr	2775745.61	22993466.35
1-May	3150866.835	26691951.42
1-Jun	3188922.003	26989964.01
1-Jul	3214745.646	26948630.76
1-Aug	2866348.663	24091579.35
1-Sep	2538726.849	20523492.41
1-Oct	2488148.287	20011748.67
1-Nov	2595270.386	21177435.49
1-Dec	2573396.629	20855799.11

A visualization of the forecast that includes historical data, existing stores forecasts, and new stores forecasts goes as follows:


Historical & Forecasting Data
Type
Existing
Forecast New
Forecast Existing
30M
25M
20M
15M
10M
5M
0M
Sum Total Sales
2012
2013
2014
2015
2016
2017
Month of Date


Historical & Forecasting Data
Type
Existing
Forecast New
Forecast Existing
30M
25M
20M
15M
10M
5M
0M
Sum Total Sales
2012
2013
2014
2015
2016
2017
Month of Date