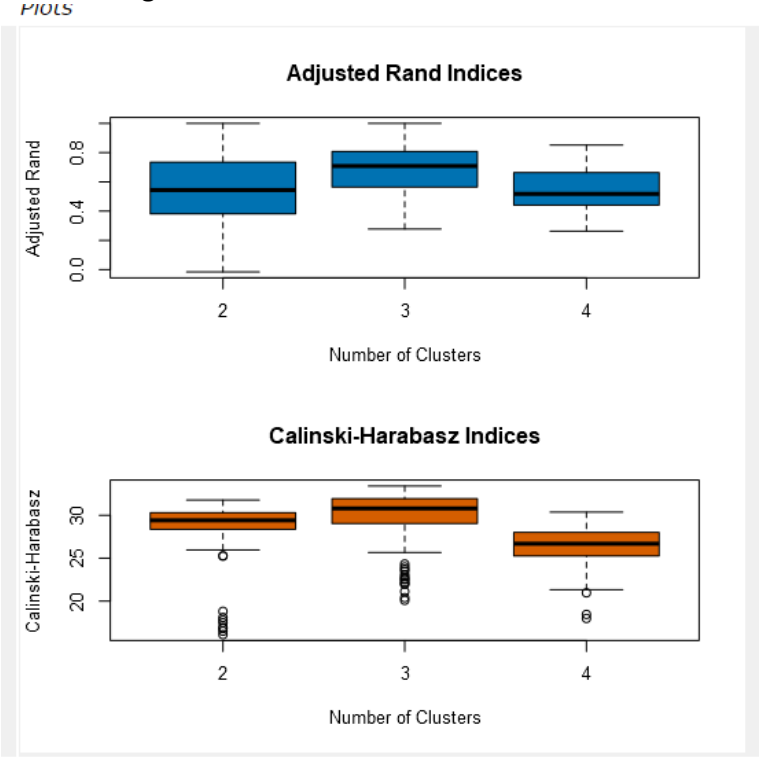
Project: Predictive Analytics Capstone

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 optimal number of store formats is 3. Based on the K-mean cluster diagnostics below are the RAND and CH indices:

   K-mean Diagnostics:



   Based on the above plots, we can see that cluster 3 has the highest mean and median values for the indices in both RAND and CH plots. Hence concluding that the optimal number of store formats (segments) would be 3.

2. How many stores fall into each store format?

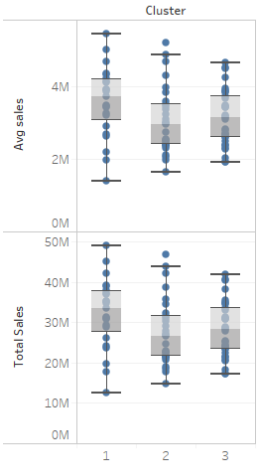
   Below is a list of optimal number of stores per cluster:

Cluster	Count
1	23
2	29
3	33

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

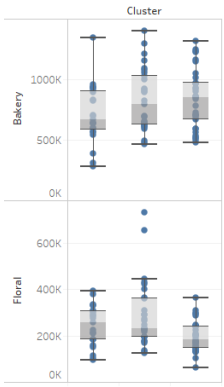
One way to test the clusters is by comparing Total Sales of each store because this was not used as a variable in cluster analysis. I also took average sales at each store in each cluster and based on the below image we can say that the clusters behave differently from each other.

Avg and Tot sales box



Avg sales and Total Sales for each Cluster.

Floral and Bakery box

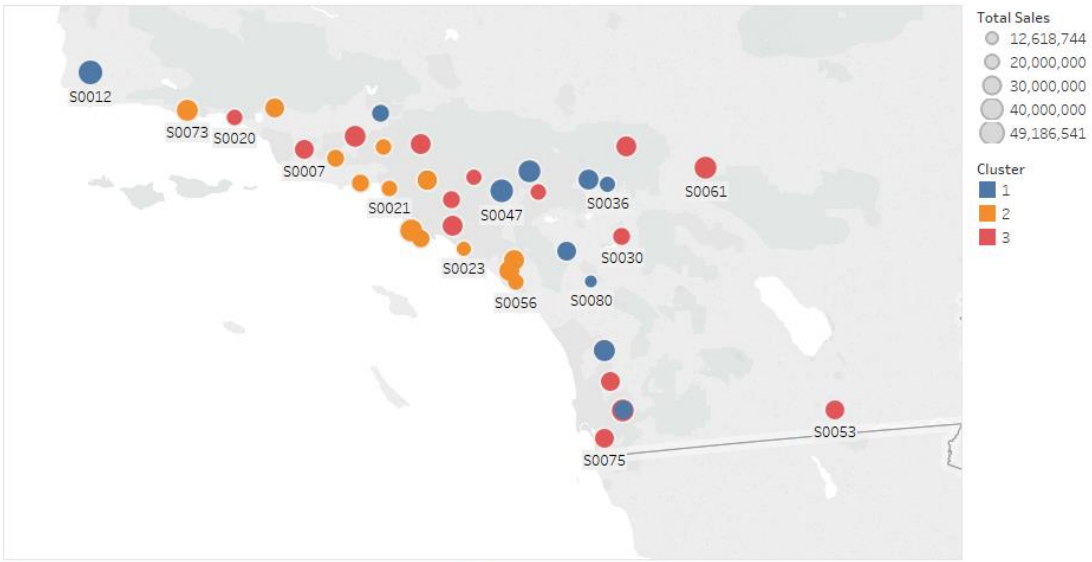


Bakery and Floral for each Cluster.

I also verified the individual category sales for Bakery and Floral to verify the clusters and the above image on the right shows that they all are varied.

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.

cluster map

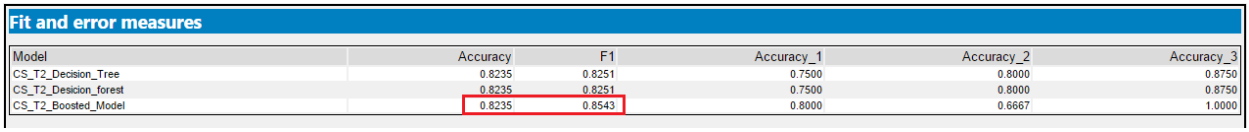


Map based on Longitude (generated) and Latitude (generated). Color shows details about Cluster. Size shows Total Sales. The marks are labeled by Store. Details are shown for City.

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.)

   Based on the report below, even though the accuracy measures are same for all three models, I choose Boosted model because it has slightly higher F1 score.
   In a situation, where the classes are unbalanced and all the classes are important we would choose a model with higher F-measure. We can consider F-score in this scenario because this is non-binary target variable and is related to segmentation rather than having a weighted average.



Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
CS_T2_Decision_Tree	0.8235	0.8251	0.7500	0.8000	0.8750
CS_T2_Desicion_forest	0.8235	0.8251	0.7500	0.8000	0.8750
CS_T2_Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000

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

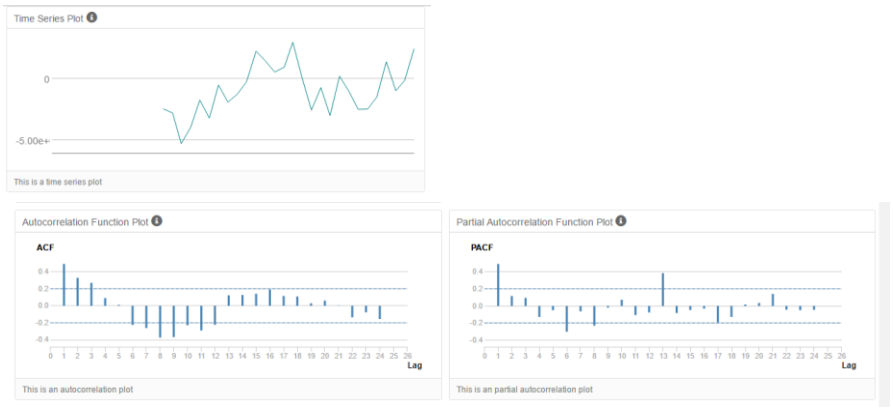
Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Task 3: Predicting Produce Sales

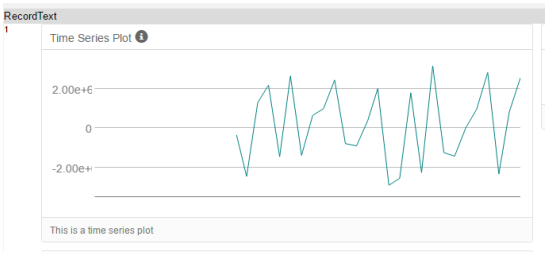
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS (a, m, n) or ARIMA (ar, i, ma) notation. How did you come to that decision?

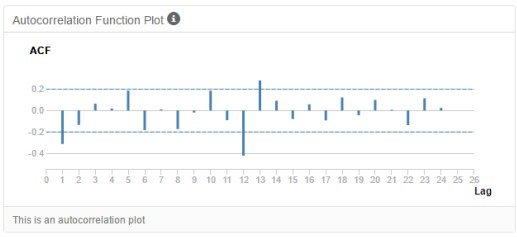
   I used ETS (M, N, M) because the data suggests that we have no trend, the seasonality increases in magnitude over time and error component also increases in magnitude hence adding Seasonality and Trend multiplicatively and not adding any trend terms.

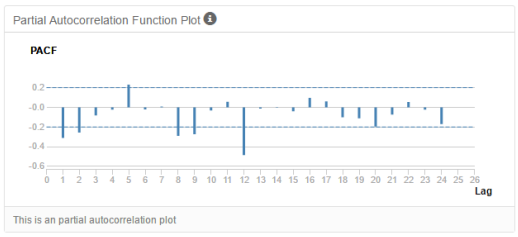
For seasonal ARIMA model I have used ARIMA (0, 1, 1) (0, 1, 1) [12] considering it is monthly data for 12 months. I have taken seasonal differencing and below is the plot.



Looking at the first plot above we can see that the mean and variance are not close to zero. Also examining the ACF and PACF plots we can that significant autocorrelation exists between lags. So, I am taking the seasonal first difference and below are the plots.







After examining the above Time series plot and ACF and PACF plots from the seasonal first differencing we can say that the mean and variance are close to 0. The ACF and PACF plots do not show autocorrelation. Hence we can now determine our ARIMA terms since we have standardized the data.

Now, below are the in-sample errors for ETS and ARIMA model. I have also used TS compare tool to compare the ETS and ARIMA models and below are the results.

ETS model:

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

ARIMA model:

Information Criteria:

AIC	AICc	BIC
849.8292	850.8727	853.7167

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
150815.8641194	935292.1712234	628801.7029024	0.6312352	2.7761535	0.3510153	-0.0469226

Forecast error measures:

Actual and Forecast Values:

Actual	CS_T3_ETS	CS_T3_ARIMA
26338477.15	26907095.61191	27182961.16627
23130626.6	22916903.07434	24073582.27177
20774415.93	20342618.32222	21223756.4441
20359980.58	19883092.31778	20648299.23319
21936906.81	20479210.4317	21205988.81004
20462899.3	21211420.14022	21622151.40814

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
CS_T3_ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA
CS_T3_ARIMA	-492238.8	792197.3	735878.2	-2.1992	3.3098	0.433	NA

Looking at the results, ETS is a better model because it does a good job at forecasting the values than the in-sample.

Looking at the RMSE values of both models we see that ETS model has lower RMSE which means that the forecasts will have narrow possible range of values.

Also the MASE for ETS is 0.38 (which is less than 0.43 of ARIMA) which also suggests that this is a better model.

Hence, I can conclude that ETS (M, N, M) is the best model for forecasting the produce for next

12 months.

3. 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.

Below are the forecast values for the existing and new stores for the next 12 months:

Year	Month	Existing_stores_forecast	New_Stores_forecast
2016	1	21539936	2587451
2016	2	20413771	2477353
2016	3	24325953	2913185
2016	4	22993466	2775746
2016	5	26691951	3150867
2016	6	26989964	3188922
2016	7	26948631	3214746
2016	8	24091579	2866349
2016	9	20523492	2538727
2016	10	20011749	2488148
2016	11	21177435	2595270
2016	12	20855799	2573397

Below is the visualization with existing and forecast data: