Project: Predictive Analytics Capstone

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 store formats is 3. I have compared number of clusters to choose the optimal with min number of clusters 2, max number of clusters 6 and number of starting seed 10. The following plot shows that.

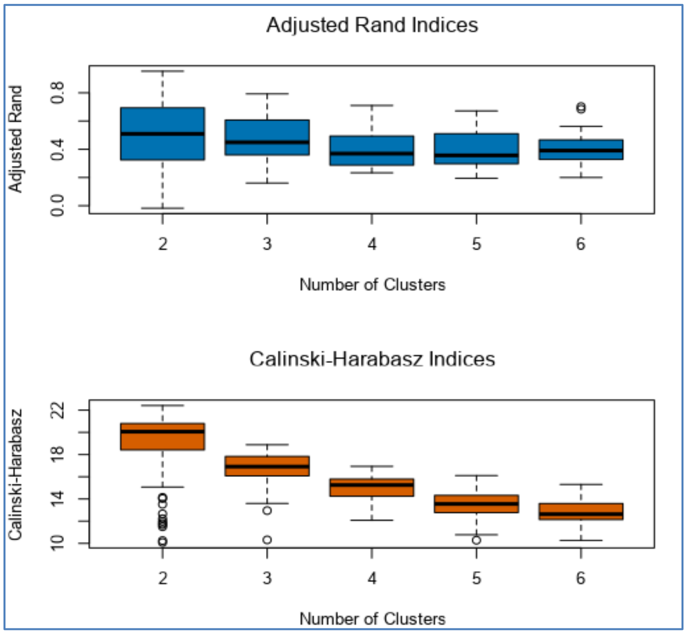


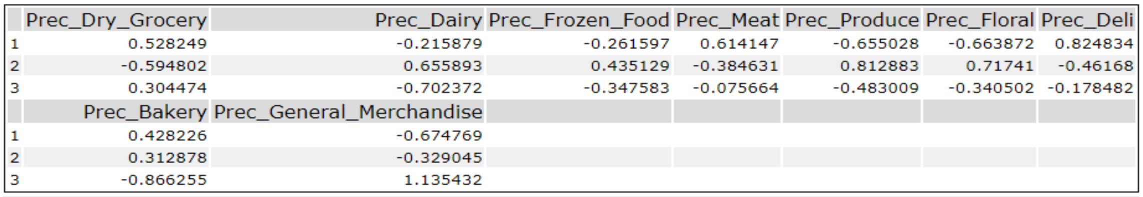
Figure 1: K-Means Plots of Cluster Assessment

   According to the plot, the increase of cluster number leads to compact data but the 3 clusters in K-means have the optimal median compared with other clusters. Because of the difference between Q1 and Q3 is small which means the data closer than 2 clusters which has the highest median, but the data is not compacted like 3 clusters.

2. How many stores fall into each store format?

   The cluster 1 has **25** stores.
   The cluster 2 has **35** stores.
   The cluster 3 has **25** stores.

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

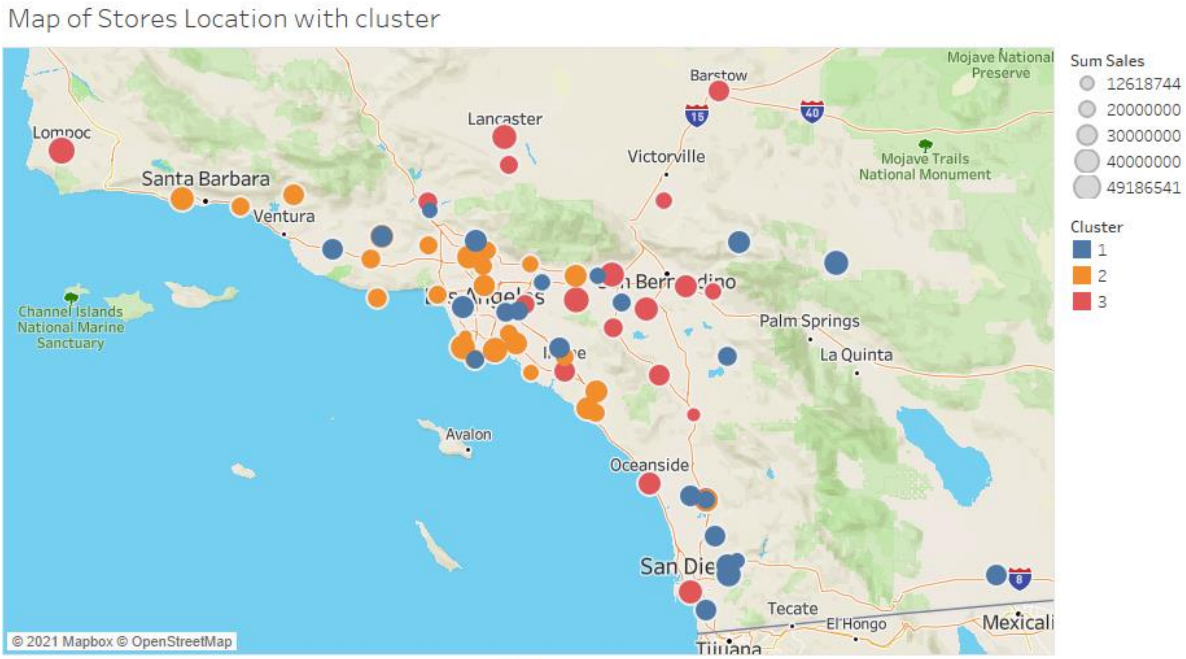


	Prec_Dry_Grocery	Prec_Dairy	Prec_Frozen_Food	Prec_Meat	Prec_Produce	Prec_Floral	Prec_Deli	Prec_Bakery	Prec_General_Merchandise
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834	0.428226	-0.674769
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168	0.312878	-0.329045
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482	-0.866255	1.135432

Figure 2: Cluster CentroidsTable

As the cluster centroids table shown, the cluster 1 has more attention on sales of deli while the sales of floral and general merchandise are the least important. In the cluster 2, the sales of produce and floral are the most important which is the opposite of cluster 1. The general merchandise is the most important sales in cluster 3 while the bakery sales is the least important.

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.



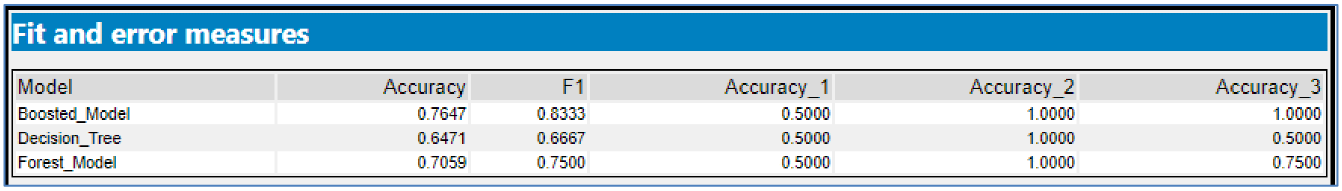
Map based on Longitude (generated) and Latitude (generated). Color shows details about Cluster. Size shows details about Sum Sales. Details are shown for City, State and Zip.

Figure 3: Map of Stores Location with Clusters

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 boosted model is the methodology used to predict the format for new stores.



Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Boosted_Model	0.7647	0.8333	0.5000	1.0000	1.0000
Decision_Tree	0.6471	0.6667	0.5000	1.0000	0.5000
Forest_Model	0.7059	0.7500	0.5000	1.0000	0.7500

Figure 4: Fit and Error Measures

According to the above figure, the boosted model has the best accuracy and least error. Since the accuracy is 0.76 and F1 is 0.83. As a result, the boosted model has been chosen to predict the 10 new stores.

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	3
S0092	2
S0093	3
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?

In the forecast, we have used ETS model with (M,N,M) components. The decomposition plot on 40-month sales is going to show how we make that decision. While the rest 6-month sale for validation.

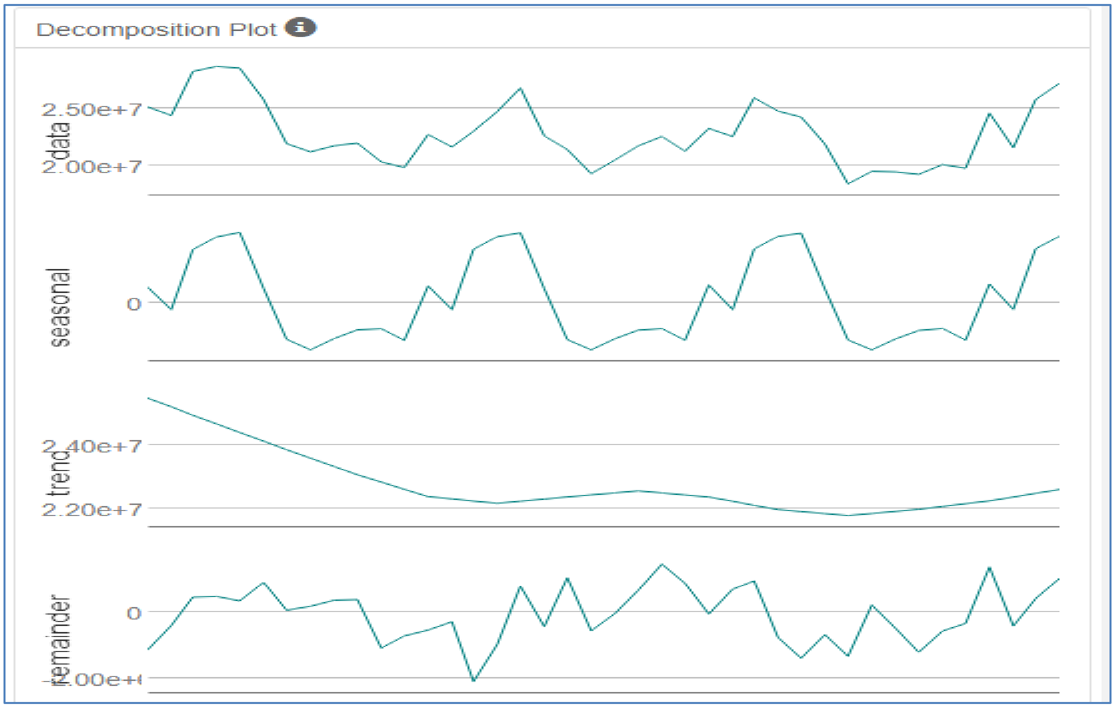
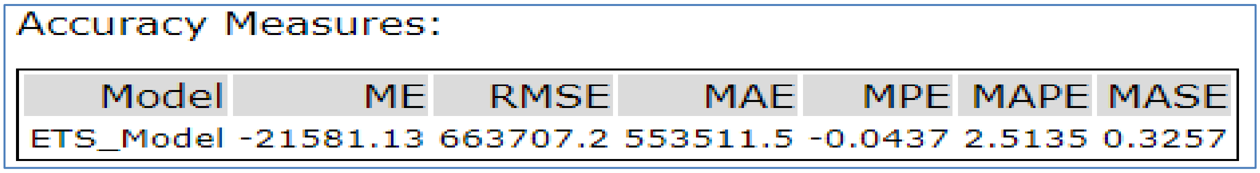


Figure 5: Decomposition Plot

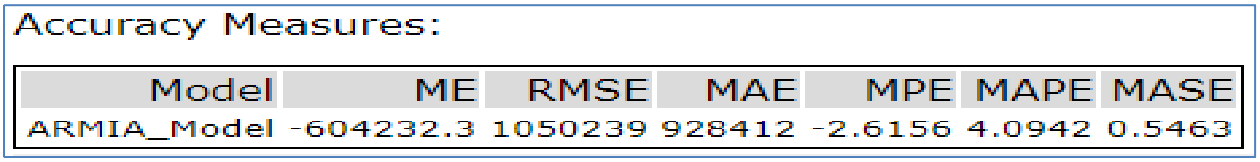
According to the decomposition plot, the error has huge variance over time, so it multiplicative. The trend has been decreased and slow growth rate, so it is none. The seasonal is multiplicative.



Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_Model	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257

Figure 6: Accuracy Measures of ETS



Accuracy Measures:

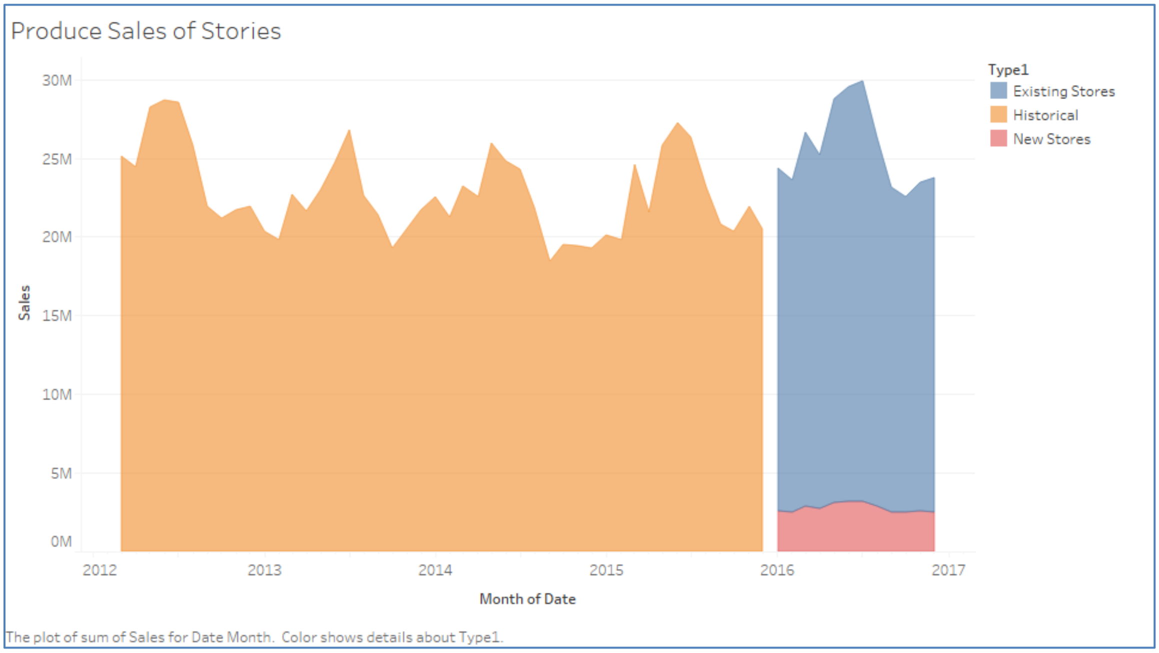
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARMIA_Model	-604232.3	1050239	928412	-2.6156	4.0942	0.5463

Figure 7: Accuracy Measures of ARIMA

The ETS has been chosen based-on the accuracy measures which is closer to true values than the ARIMA.

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.

Month	New Stores	Existing Stores
Jan-16	2,563,357.91004118	21,829,060.031666
Feb-16	2,483,924.72756208	21,146,329.6319817
Mar-16	2,910,944.1456874	23,735,686.9387899
Apr-16	2,764,881.86969732	22,409,515.2844737
May-16	3,141,305.86730493	25,621,828.7250966
Jun-16	3,195,054.20380398	26,307,858.0400465
Jul-16	3,212,390.95408986	26,705,092.5563487
Aug-16	2,852,385.7691978	23,440,761.3295266
Sep-16	2,521,697.18679037	20,640,047.3199708
Oct-16	2,466,750.89369629	20,086,270.4620746
Nov-16	2,557,744.58771366	20,858,119.95754
Dec-16	2,530,510.80513342	21,255,190.2449756


Produce Sales of Stories

The plot of sum of Sales for Date Month. Color shows details about Type1.

Figure 8: Area charts of Produce Sales

<u>Before you submit</u>

Please check your answers against the requirements of the project dictated by the rubric. Reviewers will use this rubric to grade your project.