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?

I first filtered and prepared the data for year 2015 sales data. I then, created the percentage sales per category per store. Finally, I used K-Centroid diagnostic tool to determine the optimal number of clusters. The setting suggested that having 3 categories is the optimal number of clusters. The plots below show the results. We know that the higher the median and the smaller the variation the better.

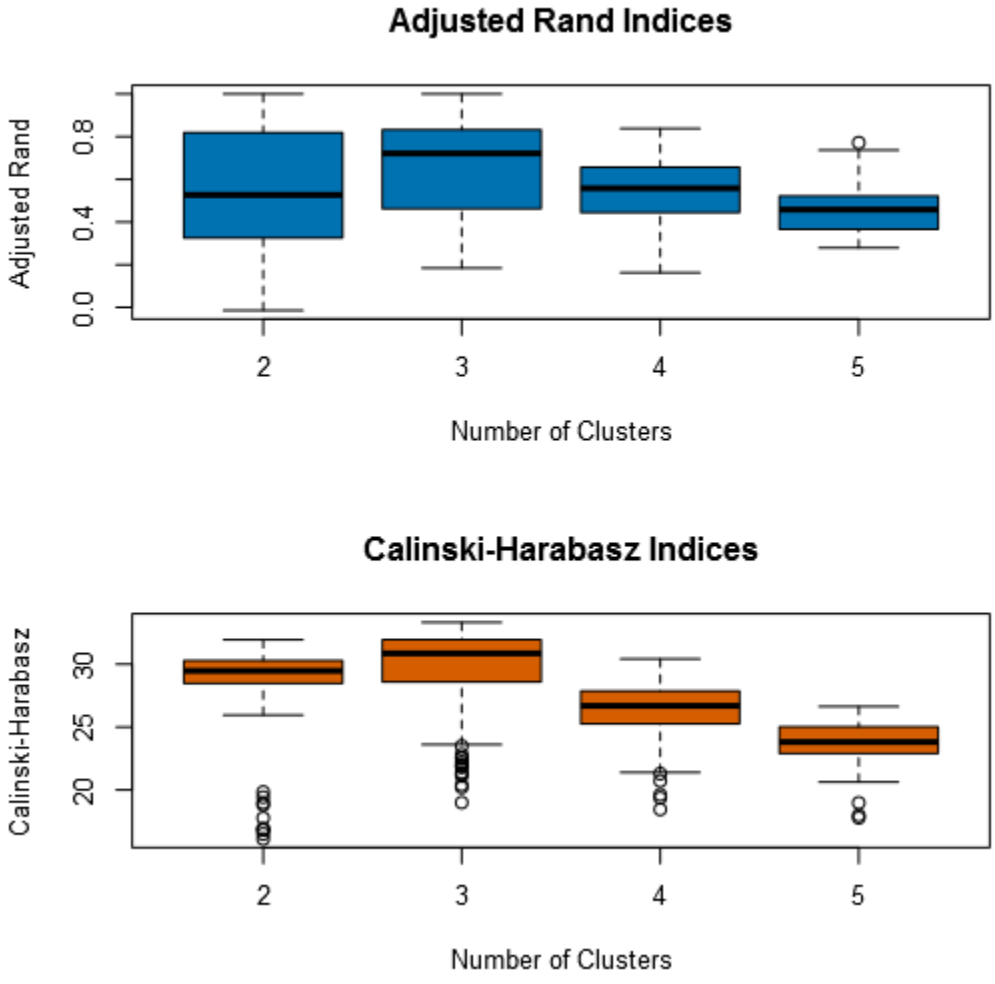


Figure 1 - K-Means Cluster Assessment Report - without PCAs

The adjusted rand plot for the setting shows that 3-cluster has the highest median among all the other settings. In terms of CH index, the plot shows that CH for 3 clusters is the highest among

the other numbers but does not necessarily have the narrowest variation. Therefore, I conclude that 3 categories provide the best outcome.

2. How many stores fall into each store format?

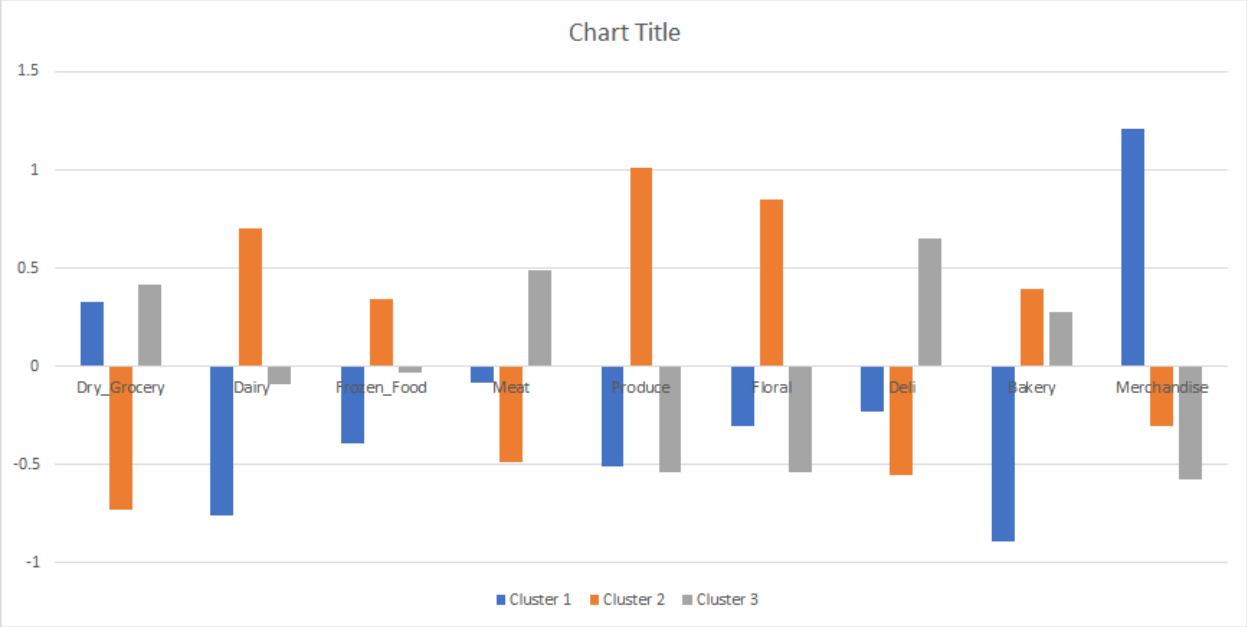
Cluster 1 has 23 stores
Cluster 2 has 29 stores
Cluster 1 has 33 stores

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

I have brought the Summary Report of the K-Means Clustering Solution below. Looking at Cluster Centroids (Table below), this gives standardized value of each cluster member for each of the measures. I have brough another plot using the table values to better understand what distinguishes the clusters from one another.

Percentage Sum_Sum_ Dry_ Grocery	Percentage Sum_Sum_ Dairy	Percentage Sum_Sum_ Frozen_ Food	Percentage Sum_Sum_ Meat	Percentage Sum_Sum_ Produce	Percentage Sum_Sum_ Floral	Percentage Sum_Sum_ Deli
0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

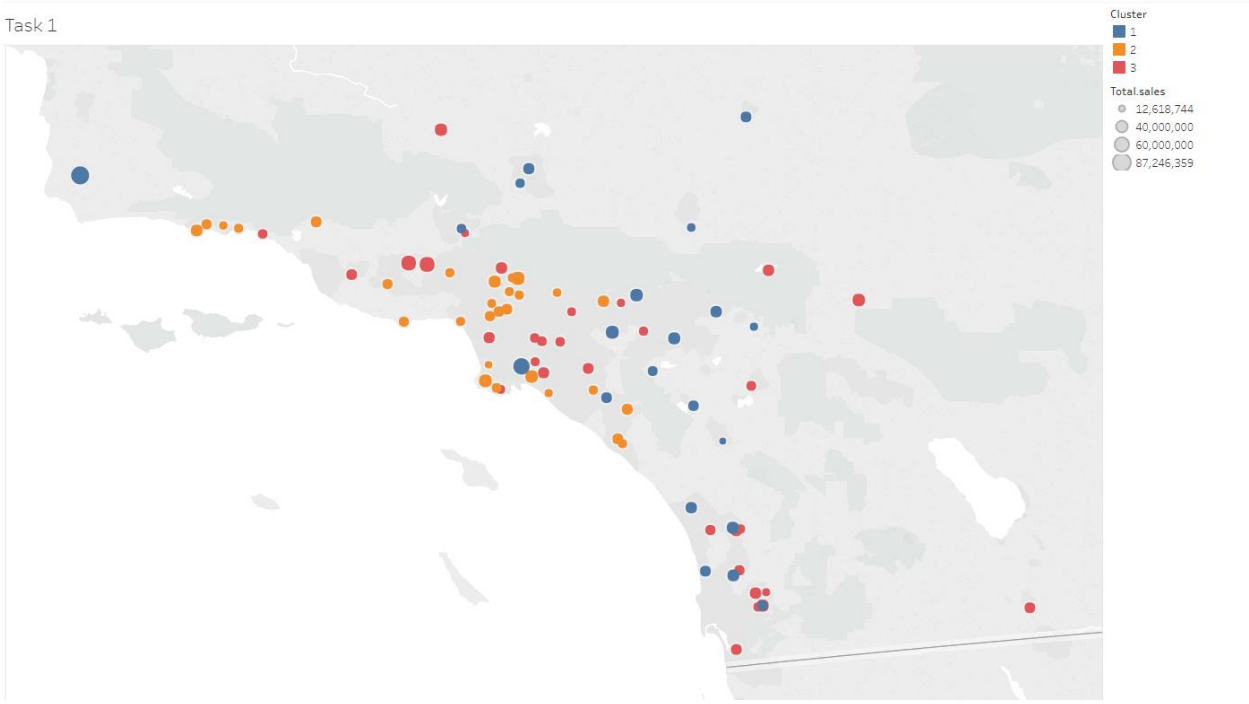
Percentage Sum_Sum_ Bakery	Percentage Sum_Sum_ General_ Merchandis
-0.894261	1.208516
0.396923	-0.304862
0.274462	-0.574389



For instance, as we can see from the plot above, Cluster 1 has the highest percentage of Merchandise items, Cluster 2 has the highest percentage of Produce and lastly, Cluster 3 has the highest percentage of Deli items sold.

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.

Here is the (link).



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.)

In sales prediction phase, I ran 3 different models (Decision Tree, Forest Model, and Boosted Model simultaneously and compared them against each other using Model Comparison tool. The results of the comparison test is presented below:

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_Default	0.7059	0.7327	0.6000	0.6667	0.8333
FM_Default	0.8235	0.8251	0.7500	0.8000	0.8750
BM_Default	0.8235	0.8543	0.8000	0.6667	1.0000

Figure 1 - Model Comparison Report

Considering Accuracy measure of Forest and Boosted models, both have shown a similar overall accuracy (0.8235). Howveer, Boosted model has a higher F1 value (closest to 1 is the best). As we can see from the figure above, the Boosted Model has produced the most accurate results compared to the other 2 models. Therefore, I chose Boosted Model to predict the best store format for the new stores.

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?

First the results for the existing stores is presented. The composition plot for the time series is presented below. As you can see the time series does not have any trend and the seasonal portion does not show any growth. However, the remainder (Error) shows a widening range suggesting a multiplicative term. Hence, in terms of ETS model I conclude that ETS(M,N,M) would be my choice.

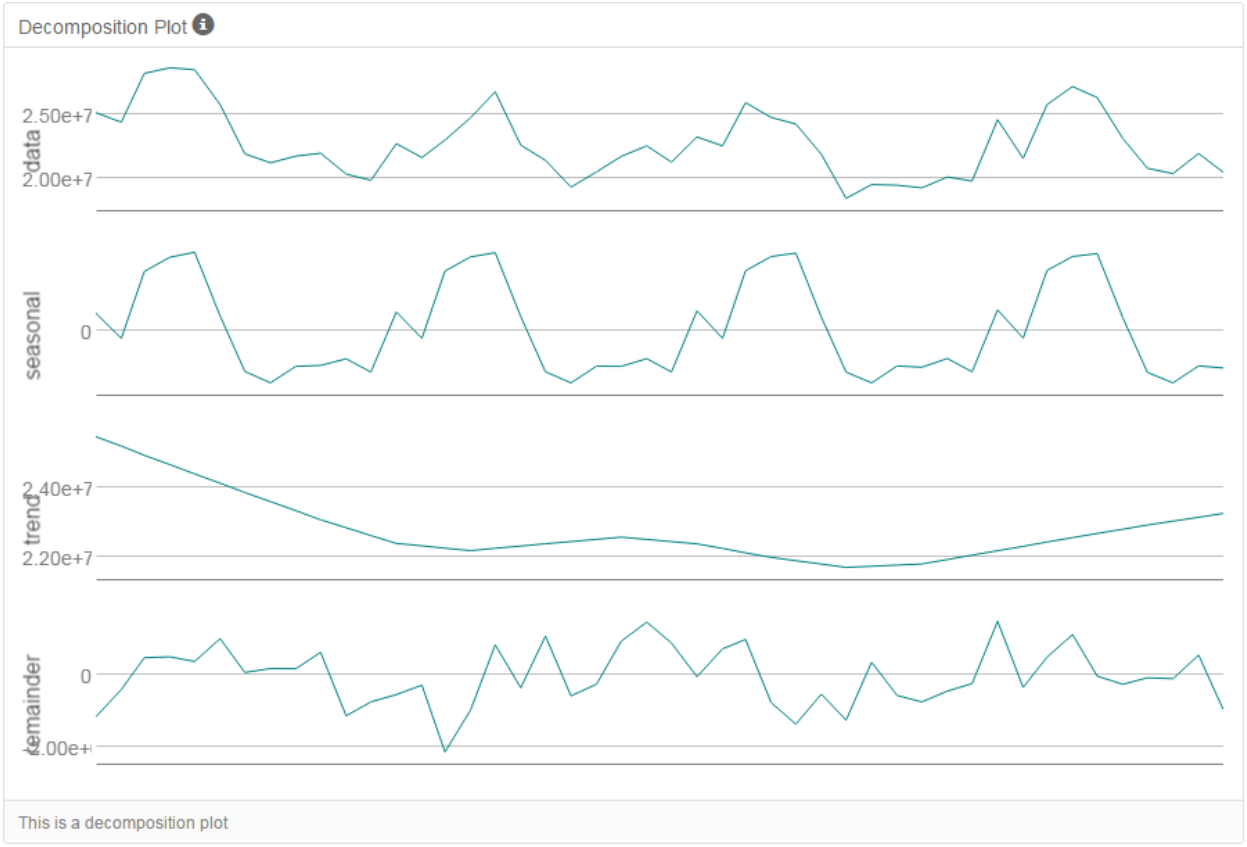
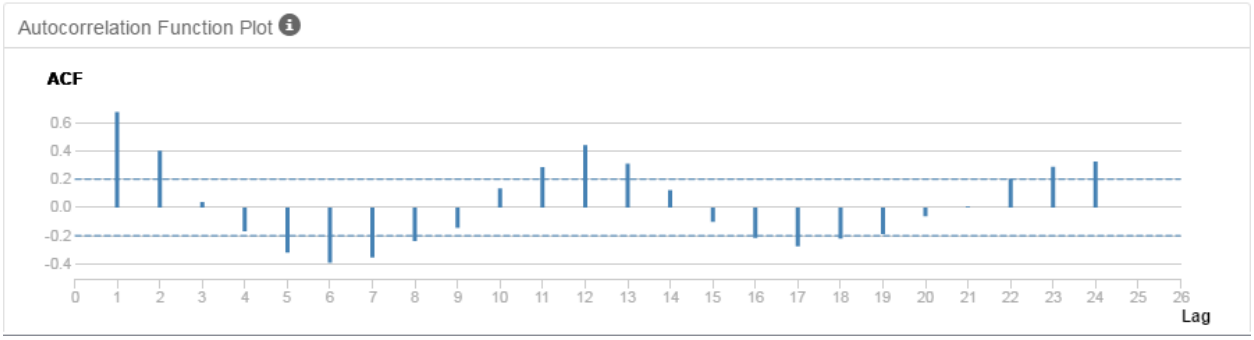


Figure 2 - Decomposition Plot



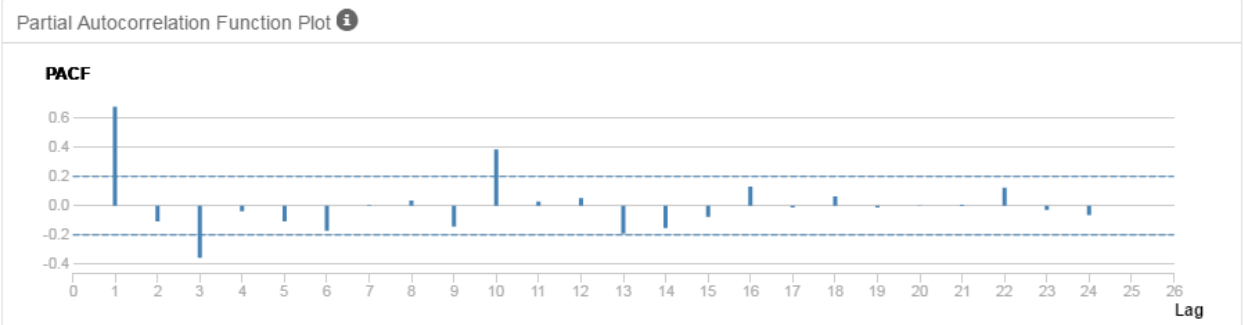


Figure 3 - ACF and PACF before any differencing

As we can see from the ACF and PACF, there is only one significant spike at lag 1, so I conclude that AR term for the non-seasonal portion is 1. So far, we have ARIMA (1,0,0)

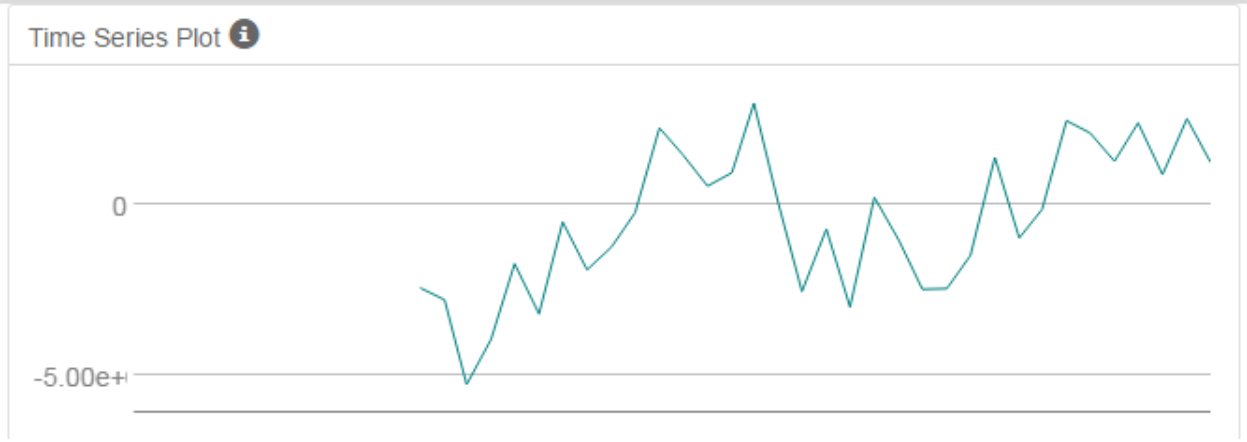
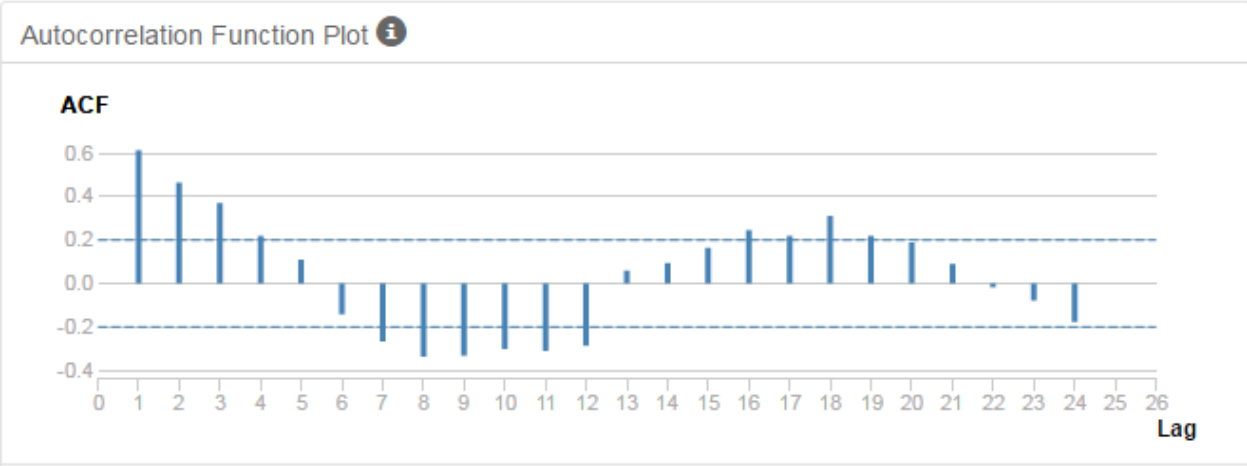


Figure 4 - Time Series Plot after seasonal differencing



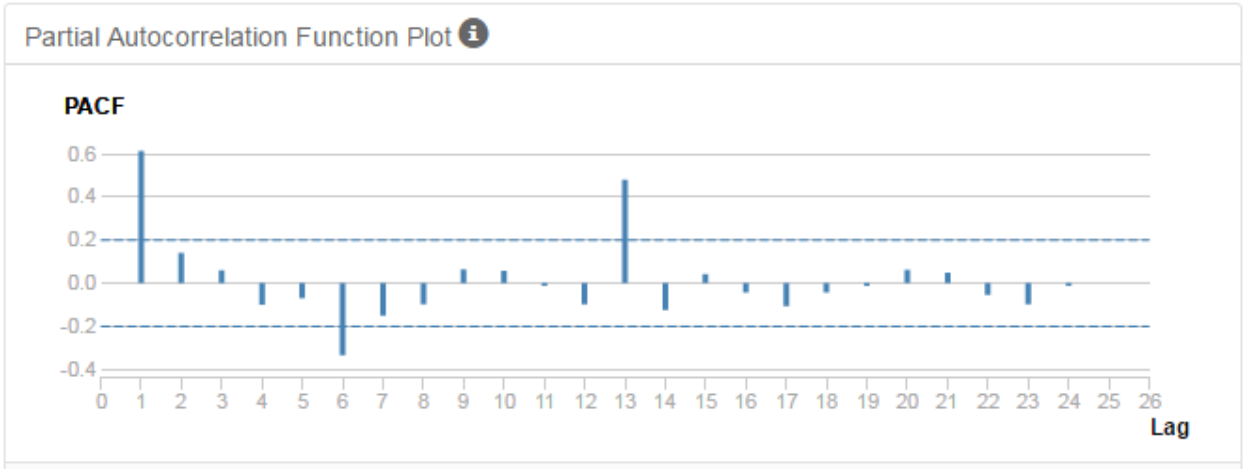


Figure 5 - ACF and PACF after seasonal differencing

After seasonal differencing the plot shows a median of 0 and obviously the data set is stationary. So, the D term is 1. From PACF and ACF after seasonal differencing we can see that there is no significant correlation left in the time series. Therefore, we have (0,1,0) for the seasonal component.
The optimum setup for ARIMA would be ARIMA(1,0,0)(0,1,0)12.

To compare the 2 different models (ARIMA and ETS) I have brough the in-sample results for both of them below. As we can see, the ETS model has shown better in-sample error measures. We also use holdout sample of 6 months to compare the 2 models. Below, we can see the table showing the results of the comparison between the ETS and ARIMA model:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	-604232.3	1050239.2	928412	-2.6156	4.0942	0.5463	NA
ETS	-324792.3	680122.7	596442.4	-1.4619	2.7002	0.351	NA

Figure 6 - Comparison of time series models

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-266968.7825838	1385800.2923691	961223.1598628	-1.2966978	4.3808852	0.5121821	-0.1664469

Figure 7 - In-sample error measures for ARIMA Model

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-241658.3191268	886787.7565482	699047.4732299	-1.1576764	3.1317204	0.3724833	0.069077

Figure 8 - In-samole error measurements for ETS model

Considering the results from the comparison test, we can easily conclude that ETS model has produced more accurate predictions. Smaller ME, RMSE, MAE, MASE all show that ETS model is a better fit. I would definitely use ETS model to predict the sales for existing stores.

Now, I present the results of the comparison for the new stores.

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA_Task3	40053.93	46036.27	40518.37	12.8607	13.0316	1.8686	NA
ETS_Task3	29207.48	32368.36	29207.48	9.2574	9.2574	1.347	NA

Figure 9 - Results for Cluster 1

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA_Task3	34044.01	38475.36	34185.83	12.4381	12.4971	1.6654	NA
ETS_Task3	24280.54	26668.9	24280.54	8.8024	8.8024	1.1829	NA

Figure 10- Results for Cluster 2

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA_Task3	22362.313	34979.79	25373.095	8.2041	9.6429	1.5935	NA
ETS_Task3	4617.547	10666.91	8951.217	1.6167	3.631	0.5622	NA

Figure 11 - Results for Cluster 3

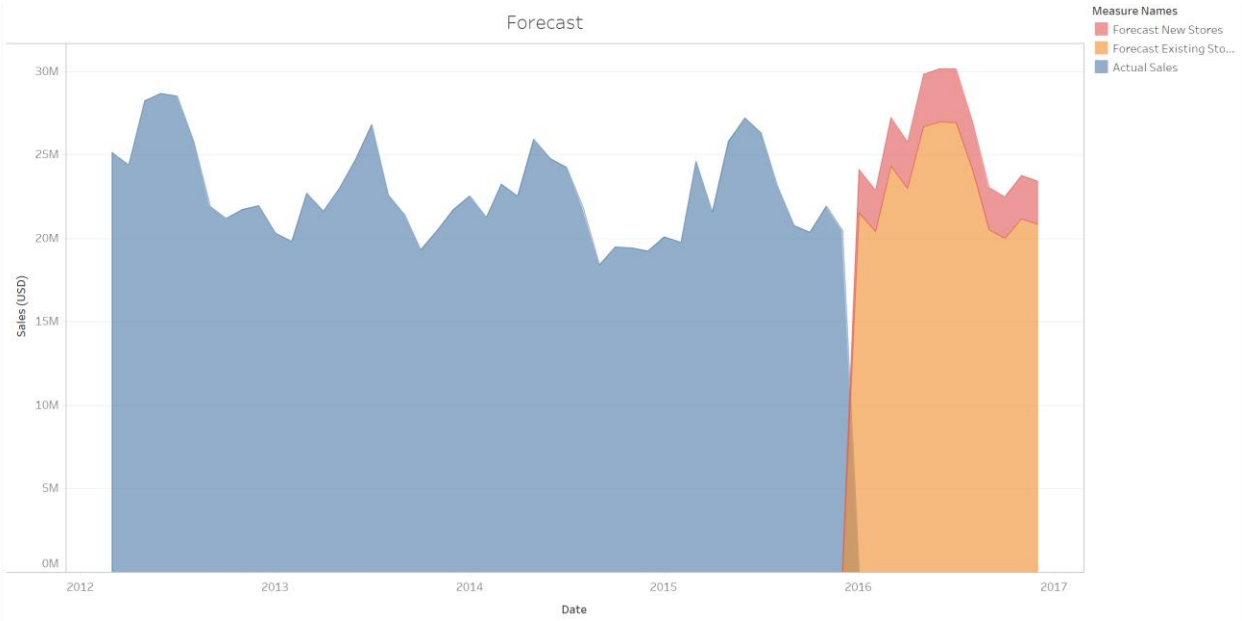
As we can see, in all the three different clusters, the ETS model has produced a more accurate prediction. Therefore, I will use ETS for the sales prediction of the new stores too.

2. Please provide a Tableau Dashboard (saved as a Tableau Public file) that includes a table and a plot of the three monthly forecasts; one for existing, one for new, and one for all stores. Please name the tab in the Tableau file "Task 3".

Month	New Stores	Existing Stores
Jan-16	2,584,383.53	21,539,936.00
Feb-16	2,470,873.92	20,413,771.00
Mar-16	2,906,307.87	24,325,953.00
Apr-16	2,771,532.13	22,993,466.00
May-16	3,145,848.57	26,691,951.00
Jun-16	3,183,909.28	26,989,964.00
Jul-16	3,213,977.72	26,948,631.00
Aug-16	2,858,247.21	24,091,579.00
Sep-16	2,538,173.64	20,523,492.00
Oct-16	2,483,550.17	20,011,749.00
Nov-16	2,593,089.19	21,177,435.00
Dec-16	2,570,200.44	20,855,799.00

The table above shows the sales figures for existing stores (aggregate sum for 85 stores) and sales for New Stores (10 stores).

The plot below shows the actual sales existing stores from March 2012 to December 2015 followed by the forcasted sales figures for existing store and new stores per month over the year 2016. The link to the Tableau Public file is also (here).



Before you submit

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