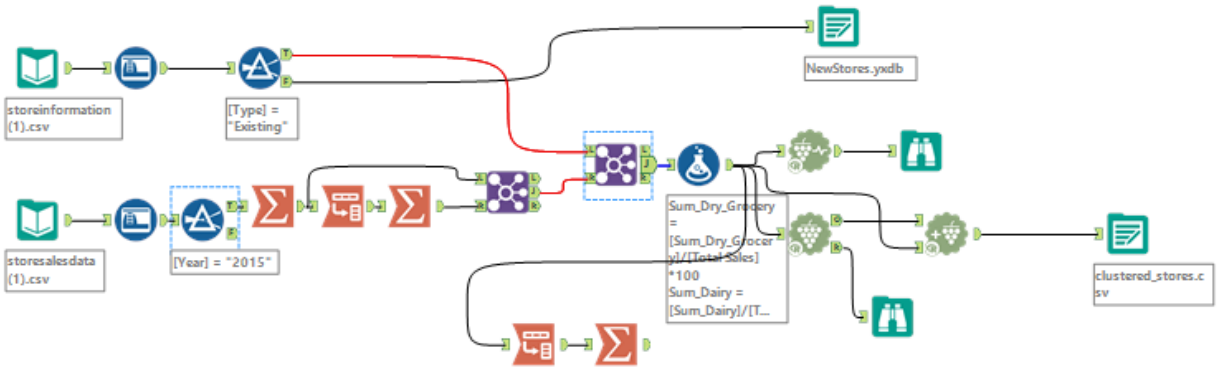
Project: Predictive Analytics Capstone

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

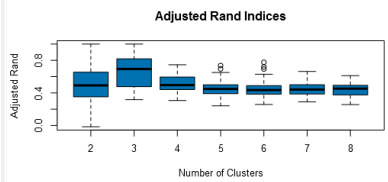
Task 1: Determine Store Formats for Existing Stores

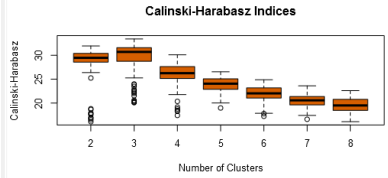
1. What is the optimal number of store formats? How did you arrive at that number?

The optimal number of store formats is 3. The first thing I had to do is filter the store sales data to only include 2015 and then add the various categories to make total sales field. I then joined the sales data with the store information. Once that was done, I placed a formula tool to have the categories show the contribution(%) it had on total store sales. I did not make use of the principal components tool/PCA process because we don't have a large number of variables that would make good use of it.



Once the data was ready, I input each category's % in the k-centroids diagnostic tool with k-means method as required) in order to identify the best number of clusters. The result showed cluster 3 as the ideal number as it has the highest median in both the adjusted rand indices and the Calinski-Harabasz Indices. It may not however have the lowest 1st and 3rd quartile but the median suggests that out of all clusters, cluster 3 is the most significant. I then moved on and added a K-Centroids Cluster Analysis tool to cluster the stores. Lastly, I appended the clusters to its corresponding stores.





Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.0152	0.3171	0.3072	0.2412	0.2586	0.2903	0.2568
1st Quartile	0.352	0.4819	0.4431	0.3943	0.3896	0.3877	0.377
Median	0.4926	0.6936	0.4964	0.4487	0.4348	0.4417	0.4526
Mean	0.484	0.6575	0.5125	0.4623	0.4532	0.4498	0.4411
3rd Quartile	0.655	0.816	0.5913	0.4982	0.489	0.4997	0.491
Maximum	1	1	0.7458	0.7366	0.7762	0.6637	0.6118

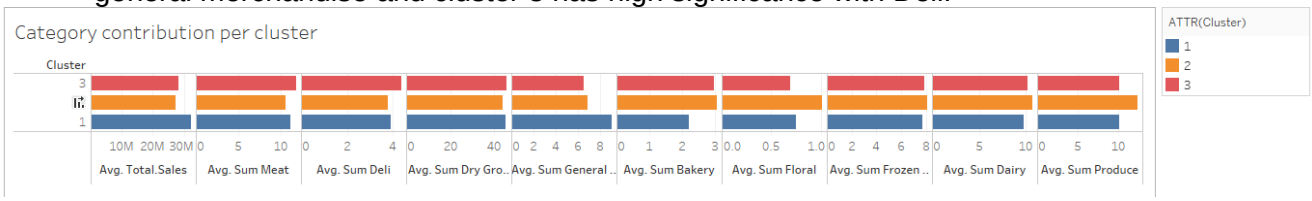
Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.1	20.09	17.41	18.98	17.24	16.61	16.11
1st Quartile	28.61	28.76	25.16	22.91	21.05	19.61	18.46
Median	29.47	30.7	26.25	24.05	22.02	20.56	19.5
Mean	28.41	29.47	25.99	23.88	21.96	20.48	19.62
3rd Quartile	30.39	31.58	27.62	25.06	23.14	21.35	20.77
Maximum	31.95	33.41	30.09	26.53	24.87	23.6	22.59

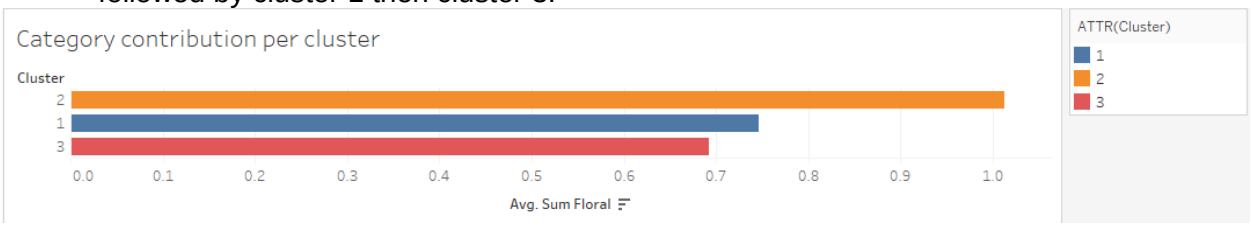
2. How many stores fall into each store format?

Cluster	Size
1	23
2	29
3	33

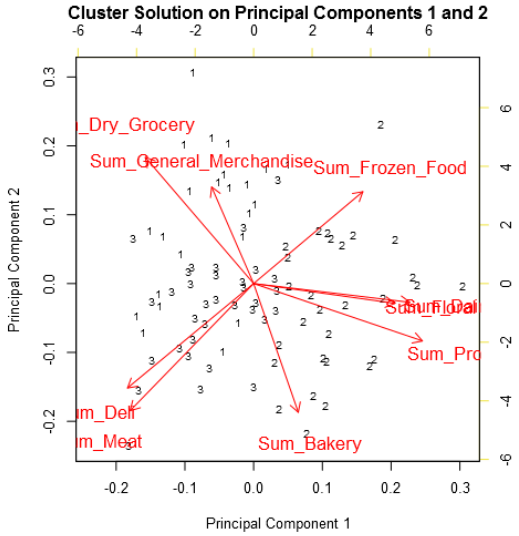
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Each cluster has its significant categories. Just like cluster 1 has high significance with general merchandise and cluster 3 has high significance with Deli.



However one notable example that can easily distinguish from the 3 clusters can be seen in the floral category. Stores in cluster 2 have the highest average contribution followed by cluster 1 then cluster 3.



The scatter plot shown by the K-centroids Cluster analysis has most of the 2's(cluster 2) leaning towards the floral category.

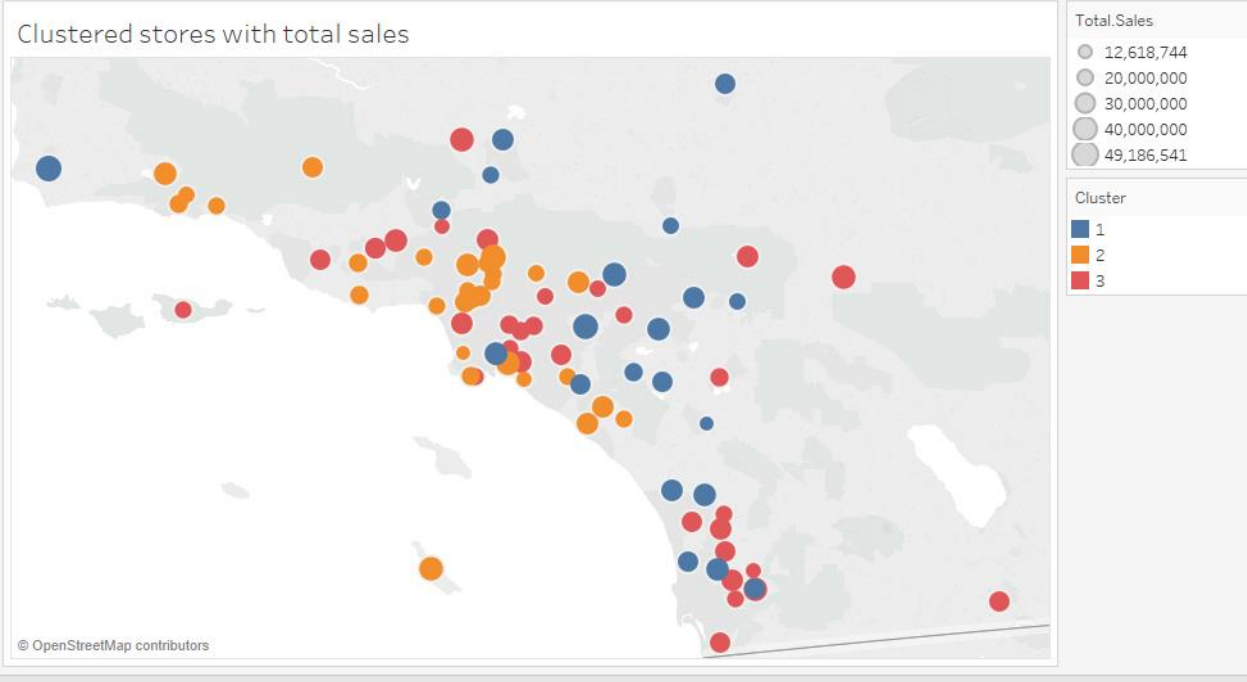


The summary also shows cluster 2 have a .85 significance in floral while the rest have it negatively but cluster 1 still has a better significance than 3.

	Sum_Floral
1	-0.301524
2	0.851718
3	-0.538327

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.

   https://public.tableau.com/shared/ZKCYMYMPX?:display_count=yes



Task 2: Formats for New Stores

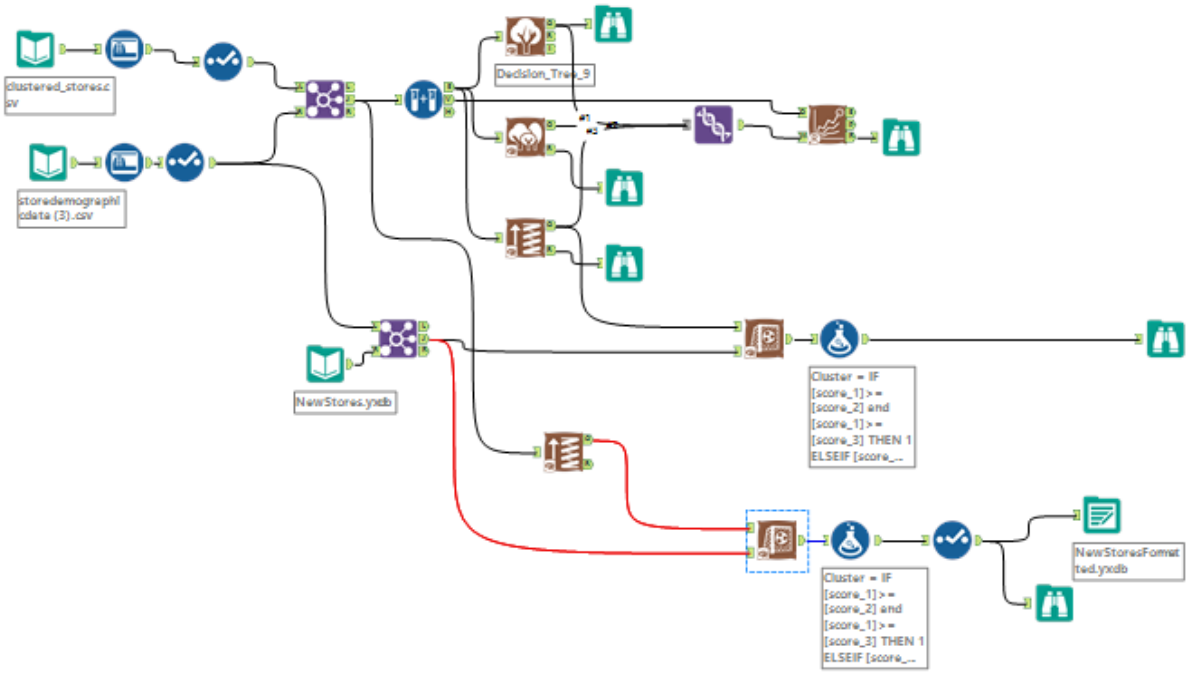
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

   I made use of the boosted model to determine the segments for the new stores. I compared the results of 3 models; decision tree, forest and the boosted model by using a validation sample with 20% of the data. The results of model comparison tool show decision tree as having the lowest accuracy and has forest tree and boosted model with the same accuracy. Then I take look at the F1 score, boosted model has a better weighted average of precision and recall as its closer to 1 than the forest tree.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7647	0.7810	0.7500	0.6667	0.8571
forest_tree	0.8235	0.8251	0.7500	0.8000	0.8750
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000

   Now that we have the boosted model as the selected model, I built another boosted model but this time using 100% of the data rather than 80% just so we could improve the model even just slightly. I then made use of the model results in the score tool to get each store's chance of being part of that cluster. And finally for each store, I selected the cluster with the highest chance.



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

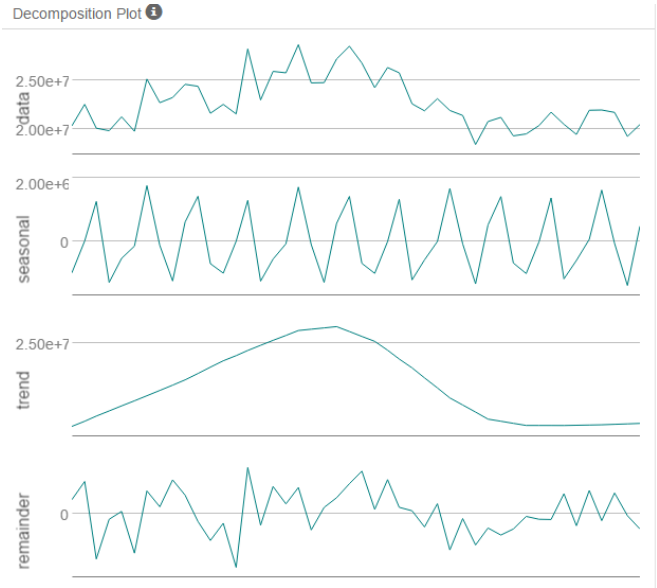
Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

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?
    a. **Existing stores forecast:**

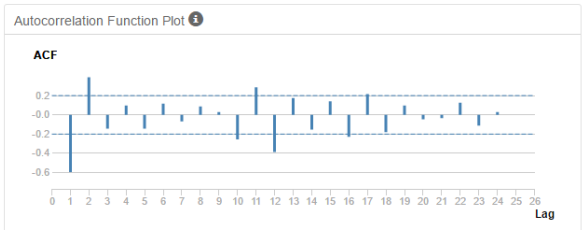
    I selected ETS(M,N,M) model for the existing stores. I used the decomposition chart to select the ETS terms. From the remainder plot, we can see error fluctuate between large and small therefore, we would use a multiplicative term for error. The trend plot shows no clear pattern as its neither growing nor shrinking exponentially.

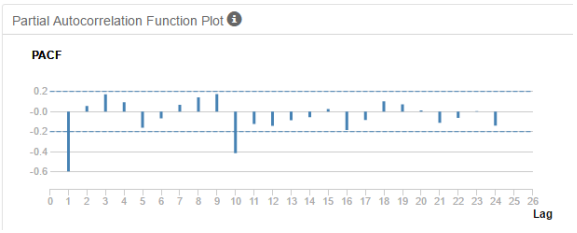
It’s also not linear as it grows then shrinks and so we should add an N term for trend. For the seasonality plot I will apply an Additive term. This results to a ETS(M,N,M) model.



The ARIMA model I selected has the terms ARIMA(0,1,1)(0,1,0)[12]. Since the data was not stationary at first, I made a first difference. After the first difference, the data was now stationary and there were only few significant lags on the ACF and PACF plots meaning there is no need for further differencing. So I added I(1) term for differencing. Both the ACF and PACF show a strong negative correlation at lag1 therefore we should add 1 term for MA. The seasonal lags(12,24,etc.) of the ACF show a strong negative correlation at lag12 but it’s not the same for PACF so we don’t need to add a term for seasonal MA.

First Difference:





Now that we know the best models for the ETS and ARIMA, we can now compare their in-sample results. We can see that the ETS has a better ME because its value is closer to the actual but overall neither are particularly close. From the RMSE, the ETS has less residue. MPE of ETS is closer to 0 than the ARIMA. From the MASE, we can see the ETS still have a better value. Lastly, the MAPE shows ETS having a slightly lower percentage error than the ARIMA. Now when comparing the information criteria, the AIC of the ARIMA is lower than the ETS. Because of this we can't really say which model is better and so we will now compare the results using validation data.

ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-195624.5309981	1890111.5396223	1364835.0267607	-1.377762	6.0682509	0.3246142	0.0604254

Information criteria:

AIC	AICc	BIC
1329.5961	1349.5961	1354.9293

ARIMA(0,1,1)(0,1,0)[12]

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-230932.0542547	2644112.2554134	1564912.4701852	-1.3941023	6.8658229	0.3722009	-0.1457271

Information Criteria:

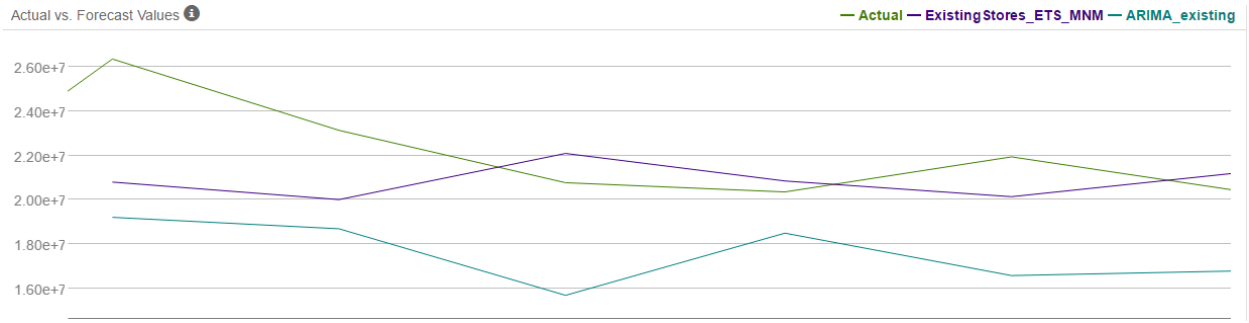
AIC	AICc	BIC
890.0515	890.5515	892.6432

The accuracy measures show the ME of the ETS forecast as being closer to the actual than the ARIMA. The RMSE value of the ETS is lower than the ARIMA meaning it has less residue. The MPE of the ETS is better as its value is definitely much closer to 0 than ARIMA. Both of the models have a MASE greater than 1 meaning they are quite significant but the ARIMA has a better value. The MAPE shows the percentage error of the models and the ETS has only 9.15% error while ARIMA has it higher with a 20.32% error.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ExistingStores_ETS_MNM	1318456	2770347	2161985	5.0608	9.1549	1.109	NA
ARIMA_existing	4577903	4852865	4577903	20.3157	20.3157	2.3482	NA

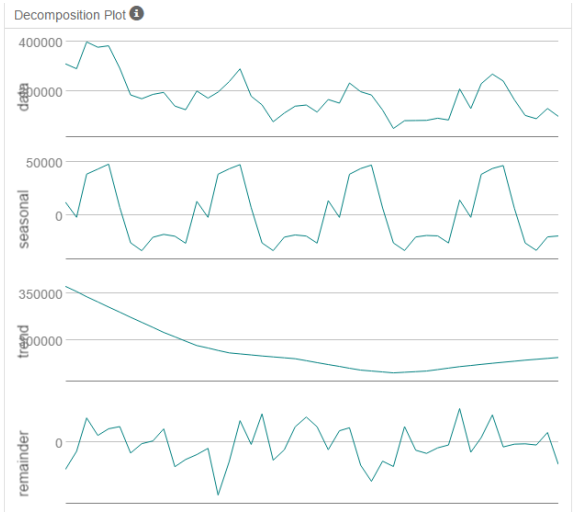
The actual vs. forecast plot makes it pretty clear that the ETS(M,N,M) is the best model for the existing stores because we can see that the ETS(M,N,M) has a forecast much closer to the actual values than the ARIMA(0,1,1)(0,1,0)[12].



b. New stores forecast:

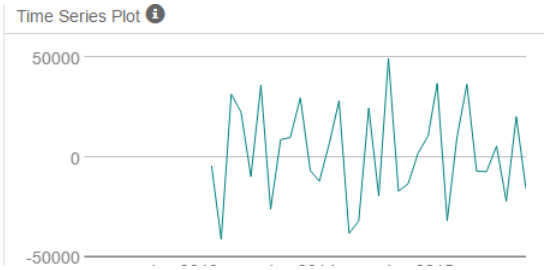
i. Cluster 1

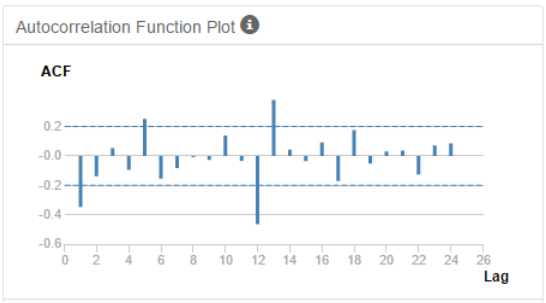
I selected ETS(M,N,M) as the better model for cluster 1. From the decomposition chart, we can see error fluctuate between large and small and that would suggest adding a Multiplicative term. The trend plot does not show any clear trend as its neither growing nor shrinking exponentially therefore we should have N for none as the term. The seasonal plot appears constant at first but when taking a closer look, each time it peaks there is a slight decrease. Therefore a multiplicative term should be added for seasonality. Giving us the ETS(M,N,M) model.

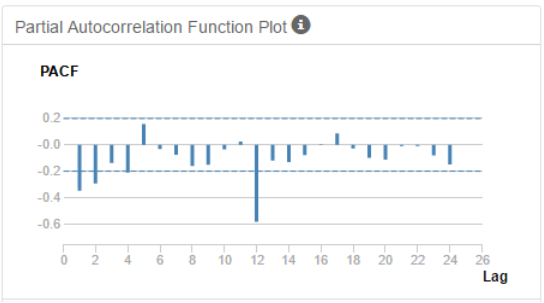


I compared this model with ARIMA(0,1,1)(0,1,1)[12]. Before I could add terms for the ARIMA, I had to make sure the data is stationary. Since it was not stationary, I made a first difference. After the first difference, we can see the data is stationary and we can also see that there are not too many significant lags from the ACF and PACF so there is no need for further differencing. So I add I(1) term for differencing. The ACF and PACF shows a strong negative correlation at lag 1 which is confirmed in the PACF therefore an MA(1) term should be added. The seasonal lags(12,24) in the ACF and PACF has a strong negative correlation at lag12 so one MA(1) term should be added to seasonal moving average.

Cluster 1 First Difference Results:







Now that we know the optimal models for both the ETS and ARIMA, we will compare their results. From the in-sample error measures, we can see that ETS has a better ME because its value is closer to the actual (on average). From the RMSE, ARIMA has a lower value. MPE shows a ETS have its value closer to zero. MAPE shows the ARIMA having a lower percentage error and MASE also has ARIMA with less error. Now we can compare the information criteria of both models, the ARIMA stands out because it has a lower AIC than ETS. So far it appears ARIMA is the better selection.

ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-1130.385839	14880.7246523	11521.8261611	-0.6204795	3.9101181	0.3700371	0.2741272

Information criteria:

AIC	AICc	BIC
943.0089	963.0089	968.3421

ARIMA(0,1,1)(0,1,1)[12]

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
2811.3574337	12835.973076	8400.7449835	0.9924202	2.9815979	0.2697999	-0.0421265

Information Criteria:

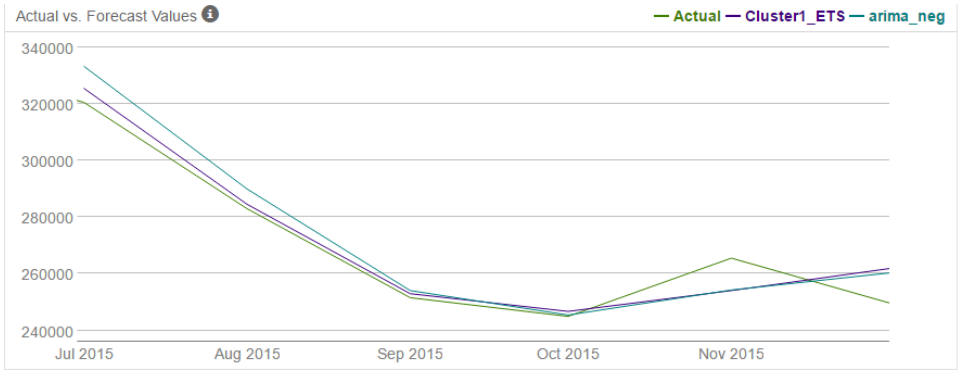
AIC	AICc	BIC
618.2325	619.2759	622.12

Now we will make a comparison based on the results using a validation(TS Compare tool). The accuracy measures show ETS as the overall better model despite the in-sample error measures. ETS has a better ME,RMSE,MAE,MPE,MAPE and MASE.

Accuracy Measures:

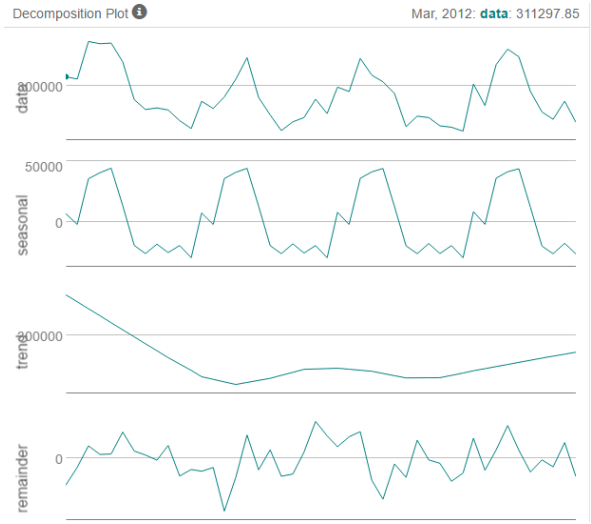
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
Cluster1_ETS	-1725.3	7187.581	5539.896	-0.6531	2.09	0.2534	NA
arima_neg	-3670.341	8708.559	7415.627	-1.2718	2.6825	0.3392	NA

The chart below makes it clear that ETS(M,N,M) is the better model as its estimates are closer to the actual value than ARIMA(0,1,1)(0,1,1)[12].



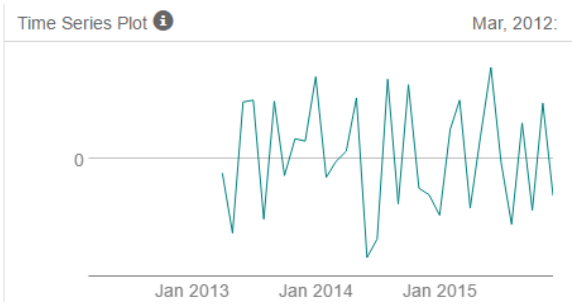
ii. Cluster 2

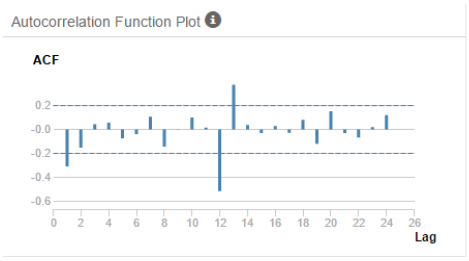
For cluster 2, I also ended up selecting ETS(M,N,M). From the decomposition chart, we can see error have no constant variance over time suggesting Multiplicative term. The trend plot also here does not show any clear trend therefore we should use N term for error. The seasonal plot has its peaks decrease slightly and so we should use a multiplicative term for seasonality and so we end up with ETS(M,N,M).

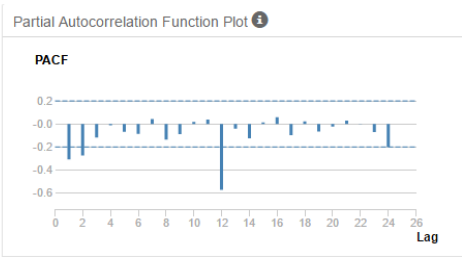


The ARIMA model I ended up with is ARIMA(0,1,1)(0,1,1)[12] which is exactly the same as Cluster 1. The data was also not stationary and so I add a first difference. After the first difference, the data was stationary and there were not too many significant lags from the ACF and PACF due to this I only added (1) I term for differencing. The ACF and PACF here also show a strong negative correlation at lag1 meaning 1 MA term should be added. Seasonal lags in the ACF and PACF show a strong negative correlation at lag 12 so one seasonal MA(1) term should be added. Leaving us with ARIMA(0,1,1)(0,1,1)[12].

Cluster 2 First Difference Results:







Now we compare both models in-sample error measures. We can see that ETS has a lower ME. The RMSE and MAE of the ARIMA has less error than the ETS. MPE shows a negative bias on ETS but its still closer to 0 than the ARIMA. MAPE shows the ARIMA having a lower %error. Neither have the MASE higher than 1. The information criteria of both models suggest the ARIMA as the better model as it has a lower AIC but just like in cluster 1 we will compare both models using the validation data.

ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
366.8623712	11644.6934199	9778.4883652	-0.0009094	3.3387125	0.4341274	0.0443467

Information criteria:

AIC	AICc	BIC
924.5555	944.5555	949.8887

ARIMA(0,1,1)(0,1,1)[12]

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
1822.0781914	11416.3586162	7622.7123717	0.5880446	2.6141063	0.3384192	-0.0096304

Information Criteria:

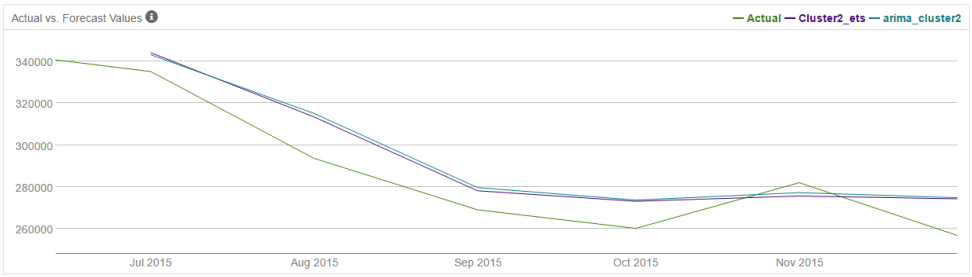
AIC	AICc	BIC
611.8923	612.9358	615.7799

The accuracy measures show ETS as the better model despite the in-sample errors just like cluster 1. ETS has a ME value lower than the ARIMA. RMSE also shows the ETS having less residue. The MPE shows a negative bias on both models but the ETS is still better because its closer to 0. The MASE shows neither of the models go beyond the value of 1 but ARIMA still has a slightly lower value. The MAPE confirms the difference in quality by showing the gap percent wise as ETS has a less percentage error than the ARIMA.

Accuracy Measures:

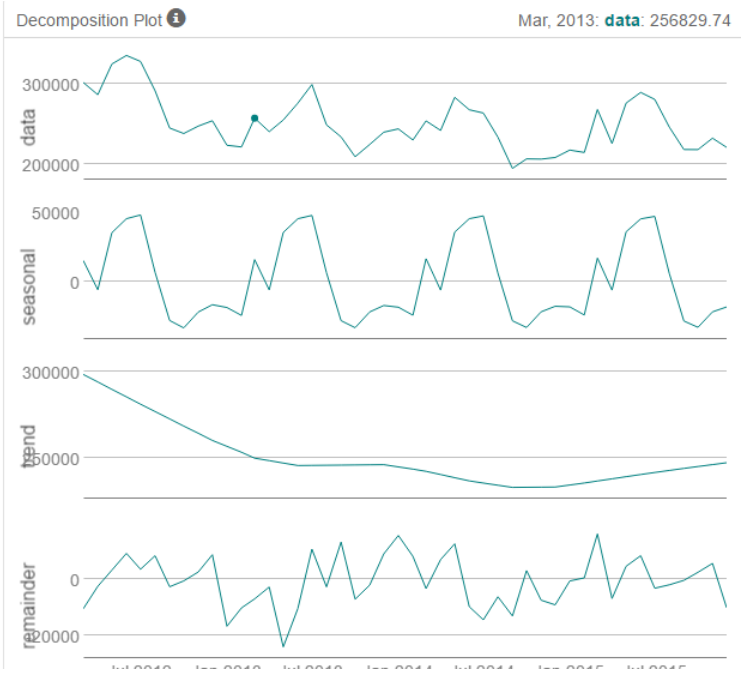
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
Cluster2_ets	-10357.11	13390.4	12486.06	-3.7348	4.4898	0.6549	NA
arima_cluster2	-11187.6	13999.39	12775.63	-4.045	4.6082	0.6701	NA

The actual vs. forecast chart makes it clear that ETS(M,N,M) is the better model for cluster 2 with estimates closer to the actual than ARIMA(0,1,1)(0,1,1)[12]. However, the ARIMA is just right behind in terms of accuracy.



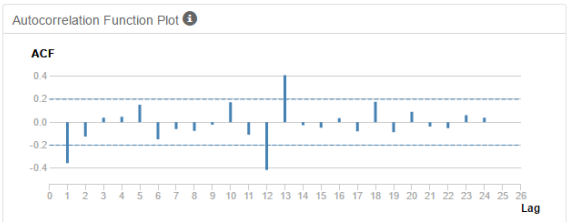
iii. Cluster 3

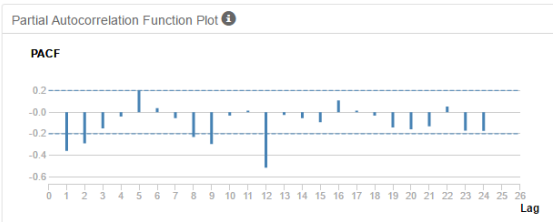
For cluster 3, I also selected ETS(M,N,M). The decomposition chart shows the residue plot fluctuate between small and large therefore we should make use of a Multiplicative term for error. The trend also has no clear exponential increase or decrease and so N term should be used for trend. Seasonality has every peak decrease slightly and so a Multiplicative term should be used. We end up with ETS(M,N,M) just like the other two clusters.



The ARIMA model I used for this cluster is the same as the other clusters ARIMA(0,1,1)(0,1,1)[12]. The reasoning is also the same. I add I(1) difference term because it was not stationary and had too many significant lags in both ACF and PACF. After the first difference, the data was good without many significant lags. The strong negative correlation at Lag1 of the ACF and PACF suggest 1(MA) term. 1 seasonal MA(1) term was added because lag12 had a significant negative correlation. Leaving us with ARIMA(0,1,1)(0,1,1)[12].

Cluster 3 First Difference Results:





The comparison of both models' in-sample error measures show the ARIMA as the better model just like in previous clusters. ETS has a better ME and MPE but aside from that the ARIMA has better measures. The RMSE and MAE of the ARIMA has less error. The MAPE has ARIMA with less % error and a lower value for MASE. The information criteria of models favor the ARIMA model.

ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-904.5824512	13281.7150443	11440.1260418	-0.7031141	4.56156	0.4704034	0.2823645

Information criteria:

AIC	AICc	BIC
935.7645	955.7645	961.0977

ARIMA(0,1,1)(0,1,1)[12]

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
2300.0238862	11068.3340838	7329.8933186	0.9100956	2.9923464	0.3013959	-0.0651335

Information Criteria:

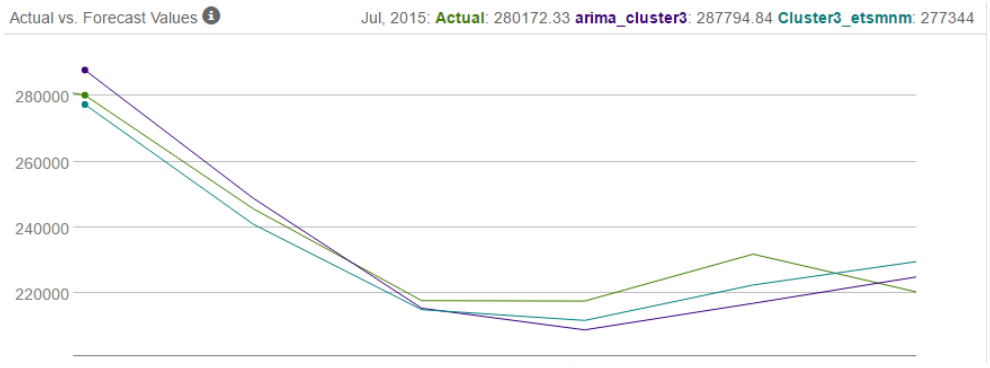
AIC	AICc	BIC
610.2603	611.3038	614.1479

The accuracy measures when comparing both models show ETS only slightly better than the ARIMA. The ETS has a higher ME than the ARIMA but the RMSE of the ETS has less residue. For MPE, the ARIMA has a better value as its closer to 0. The MASE has neither of the models with value higher than 1 but the ARIMA still has more significance. Lastly, the MAPE shows the ETS have a lower %error.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
arima_cluster3	1773.796	8088.549	6879.927	0.9068	2.9277	0.3314	NA
Cluster3_etsmnm	2719.045	6372.346	5776.787	1.1246	2.5113	0.2782	NA

The actual vs. forecast chart has the ETS(M,N,M) forecast a bit closer to the actual. The forecast appears to go a bit off towards the end of both models but the ETS still has a better output. And because of this I will select the ETS(M,N,M) over the ARIMA(0,1,1)(0,1,1)[12] as the model for cluster 3.



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".

   https://public.tableau.com/views/CapstoneTask3v2/Task3?:embed=y&:display_count=yes&publish=yes

Month	Existing Stores	New Stores
January 2016	21,504,678	2,587,451
February 2016	19,544,107	2,477,353
March 2016	18,363,161	2,913,185
April 2016	19,788,317	2,775,746
May 2016	20,801,881	3,150,867
June 2016	19,349,167	3,188,922
July 2016	19,841,356	3,214,746
August 2016	20,145,195	2,866,349
September 2016	22,044,999	2,538,727
October 2016	20,359,245	2,488,148
November 2016	19,922,751	2,595,270
December 2016	21,465,768	2,573,397