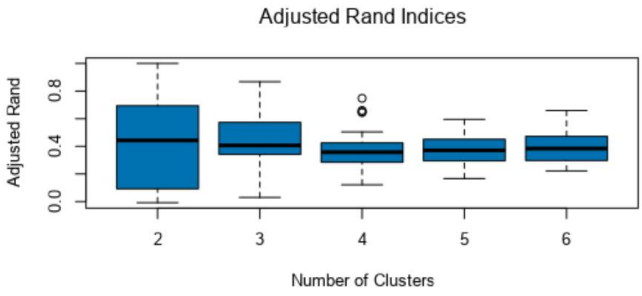
Project: Predictive Analytics Capstone

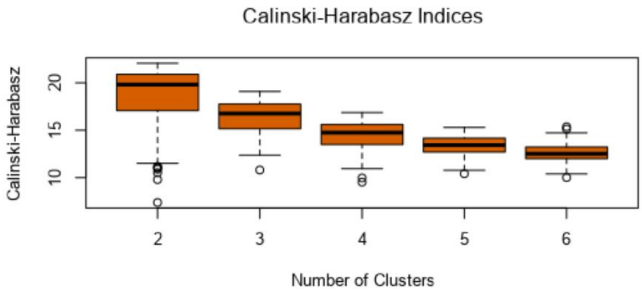
Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

Task 1: Determine Store Formats for Existing Stores

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

Using a.) % of Total Sales of different product categories as variables and (b.) K-Means Clustering method, the optimal number of store format is **3**. While AR and CH median indices are higher for k=2, the deviation for CH is high. The mean AR for k=3 at 0.44 is also higher compared to k=2 at 0.405. Hence, we choose k=3 with AR and CH means of 0.44 and 16.4, respectively.



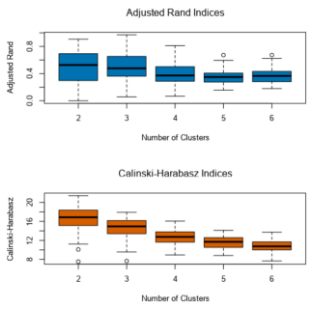
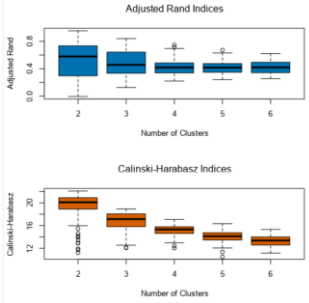


Adjusted Rand Indices:

	2	3
Minimum	-0.007639	0.029695
1st Quartile	0.094172	0.343478
Median	0.443213	0.406361
Mean	0.405201	0.443015
3rd Quartile	0.684276	0.56807
Maximum	1	0.868183

Calinski-Harabasz Indices:

	2	3
Minimum	7.376319	10.80678
1st Quartile	17.163364	15.15871
Median	19.816152	16.75762
Mean	18.520371	16.39173
3rd Quartile	20.893269	17.74967
Maximum	22.061691	19.089

Comparisons	
K-Median Method	**Neural Gas Method**
Result is also not as clear with this method since AR and CH are higher at k=2 but with many CH outliers. For k=3, AR is higher, but CH is lower vs. K-means Method. 	Result is also not as clear with this method since AR and CH are higher at k=2 but with many CH outliers. For k=3, AR and CH are higher vs. K-means. However, the project instructs the use of a K-means model. 

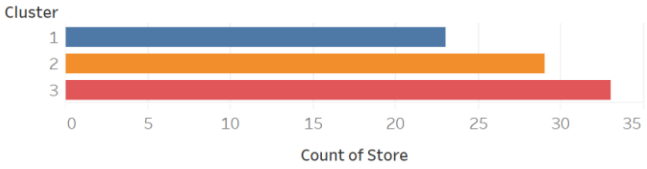
2. How many stores fall into each store format?

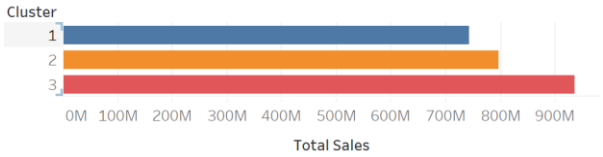
The stores are well-distributed. There are no segments with store count over 40 or less than 20.

Cluster Information:

Cluster	Size
1	23
2	29
3	33

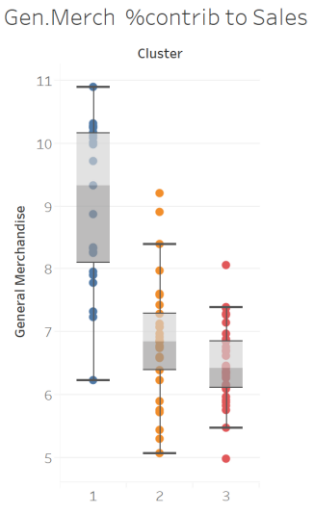
The total sales are highest for cluster 3 followed by cluster 2 the cluster 1, consistent with the store count within those clusters.



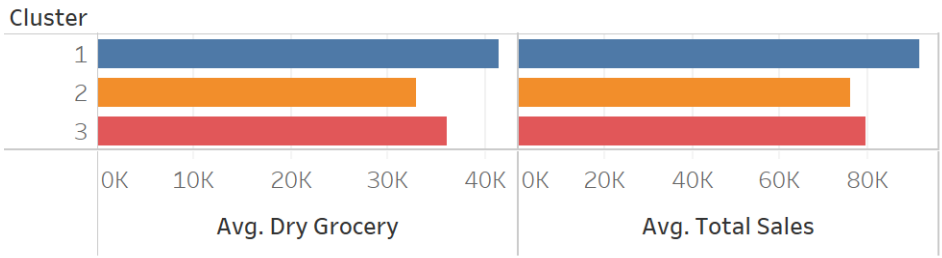


3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

One way they differ is in the %contribution of General Merchandise to Sales. As shown in the box and whiskers plot below, stores in segment 1 have generally higher values followed by segment 2, and lastly segment 3.



Because the highest contributor to overall sales are dry goods, the cluster with the highest average historical sales in dry goods also have the highest average historical total sales.



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.

Apart from the variables on sales distribution used, stores in Cluster 2 appear to be geographically linked as well. However, clusters 1 and 3 are spread out.

Link to Tableau public file:
https://public.tableau.com/profile/bianca6727#!/vizhome/Visualizations_CapstoneProject/Map

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.)

Using the validation data set to compare all the models, the Random Forest and Boosted models offer the same overall accuracy at 0.8235 but F1 is higher for the Boosted Model. Hence, we prefer the **Boosted Model**.

Below also shows that the accuracy for cluster #3 is lowest.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_1	0.7059	0.7685	0.7500	1.0000	0.5556
Forest_1	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted	0.8235	0.8889	1.0000	1.0000	0.6667

While our data set is quite small leaving only (85 x 20%) ~17 validation datapoints for our confusion matrix, we still investigate the breakdown for each below as well as the significant predictors. Recall that: $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$

Model: Decision Tree

Significant predictor is only: HVal750KPlus PopBlack

```
Model Summary
Variables actually used in tree construction:
[1] HVal750KPlus PopBlack
Root node error: 43/68 = 0.63235
n= 68
```

Confusion matrix of Decision_Tree_1

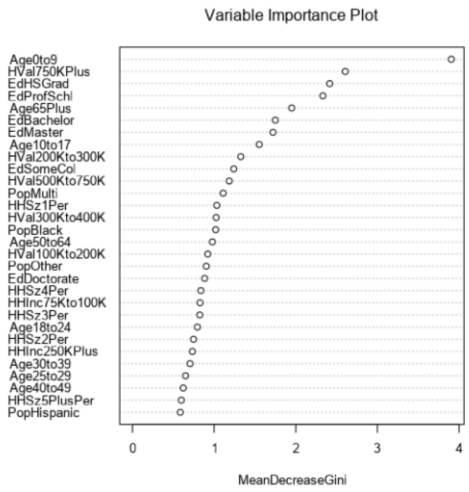
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Using the above confusion matrix, the average precision and recall are 0.70 and 0.77, respectively:

Cluster	TP	FP	FN	Precision	Recall
1	3	2	1	0.60	0.75
2	4	2	0	0.67	1.00
3	5	1	4	0.83	0.56
			Average:	0.70	0.77

Model: Random Forest

Most important variables are: Age0-9, HVal750KPlus, EdHSGrad



Confusion matrix of Forest

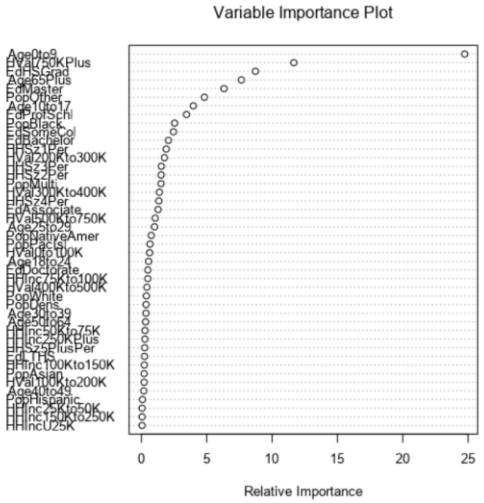
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

Using the above confusion matrix, the average precision and recall are 0.81 and 0.84, respectively. Both higher than those under Decision Tree:

Cluster	TP	FP	FN	Precision	Recall
1	3	1	1	0.75	0.75
2	4	1	0	0.80	1.00
3	7	1	2	0.88	0.78
			Average:	0.81	0.84

Model: Boosted Model

Most important variables are similar to those in RF: Age0-9, HVal750KPlus, EdHSGrad



Confusion matrix of Boosted

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

Using the above confusion matrix, the average precision and recall are 0.82 and 0.89, respectively. Both higher than those under Decision Tree and Random Forest models. Hence, we choose the Boosted model as our final model.

Cluster	TP	FP	FN	Precision	Recall
1	4	1	0	0.80	1.00
2	4	2	0	0.67	1.00
3	6	0	3	1.00	0.67
			Average:	0.82	0.89

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

Using the Boosted model, the following are the final segments for the 10 new stores:

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?

For existing stores, the model used for the forecast is **ARIMA (1,0,0)(1,1,0)[12]**. I arrived to this decision by first identifying the best ETS and ARIMA models and finally comparing these two models.

For ETS, from the charts below:

- E: error term appears have nonconstant variance, hence we apply **multiplicatively**
- T: no clear trend, hence we indicate **none**
- S: seasonality sizes (peaks and valleys) appear constant over time in the decomposition plot, but the "Seasonplot" shows otherwise, hence we test this out **additively** and **multiplicatively** and retain the better result.

The ETS (M,N,M) AIC is lower at 1279 vs. 1285 for ETS(M,N,A). As such, we choose **ETS(M,N,M)** to later compare to our ARIMA model, another time series and forecasting method.

Method:
ETS(M,N,A)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
19712.957305	1039083.8832672	865754.0619341	-0.1244787	3.7798396	0.4832889	-0.0130603

Information criteria:

AIC	AICc	BIC
1285.0321	1305.0321	1310.3653

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3502.9443415	969051.6076376	787577.7006835	-0.1381187	3.4677635	0.4396486	0.0077488

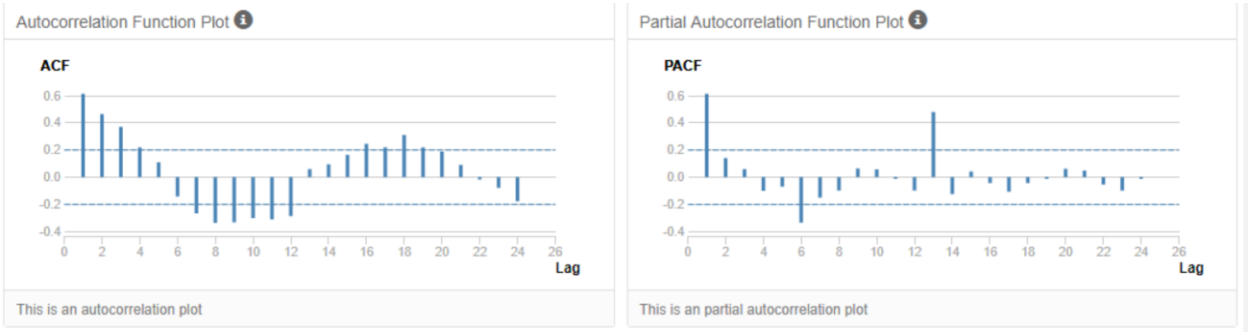
Information criteria:

AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

For ARIMA, the time series decomposition plot and Seasonplot (above) indicates the need for seasonal differencing.

The ACF and PACF of the **seasonally differenced** series tells me it is a SAR model:
-ACF decays to 0
-ACF at lag 1 is positive
-PACF cuts off quickly

From the PACF, there is significant lag at 1, hence p=1, and at lag=12, hence P=1.[1]



We test out ARIMA (1,0,0)(1,1,0)[12]. The Ljung-Box test p-value is greater than 0.05 which indicates residuals are independent, and the ACF and PACF plots now show no significant lags indicating there is no relationship left to model.

The Mean Error (ME), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) may look nominally large, but that is because these are scale dependent errors and our time series values are in the 25 million area.

[1] I initially differenced the data once more because of the high autocorrelations observed. However, the result would end up taking in a model with MA unit root.

Report

Summary of ARIMA Model SARIMA

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

Call:
Arima(Sum_Produce, order = c(1, 0, 0), seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:

	ar1	sar1
Value	0.79852	-0.700441
Std Err	0.126448	0.140181

sigma^2 estimated as 1671079042075.49: log likelihood = -437.22224

Information Criteria:

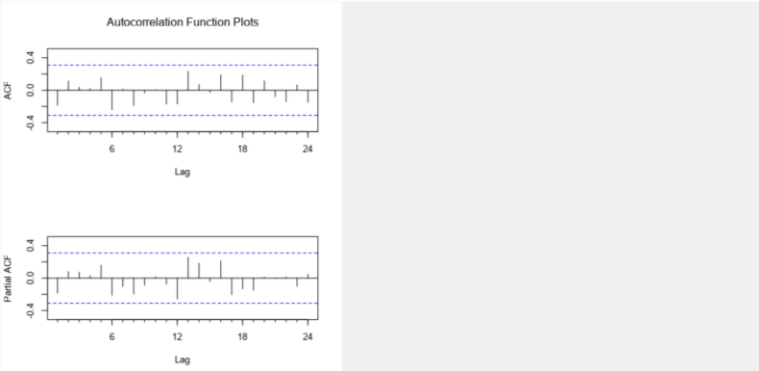
AIC	AICc	BIC
880.4445	881.4445	884.4411

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

Ljung-Box test of the model residuals:
Chi-squared = 15.0973, df = 12, p-value = 0.23616

Plots



Comparing the ETS model and the ARIMA model, the AIC of the ARIMA is lower at 880 vs ETS at 1279. The accuracy measures given by the TS compare tool also favor the ARIMA model given lower errors. Hence, we use the ARIMA model as our final model.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-141894.6	4374270	3675934	-2.5042	15.9008	2.1629
SARIMA	112812.6	4215678	3509555	-1.267	15.0978	2.065

We check the comparison again on the whole data set and it still favors the ARIMA model:

Accuracy Measures:

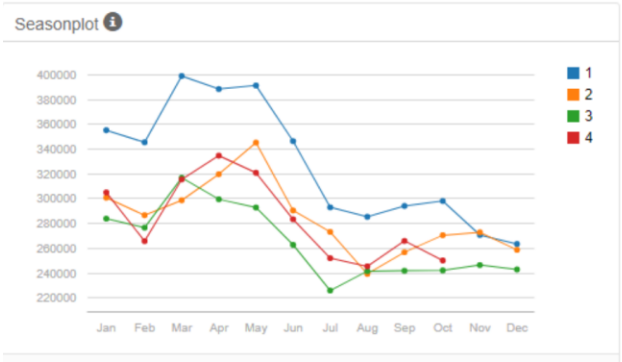
Model	ME	RMSE	MAE	MPE	MAPE	MASE
SARIMA_all	367200.7	2902559	2350338	0.7122	10.256	1.3883
ETS_all	-146477.6	3006574	2377632	-1.7025	10.526	1.4044

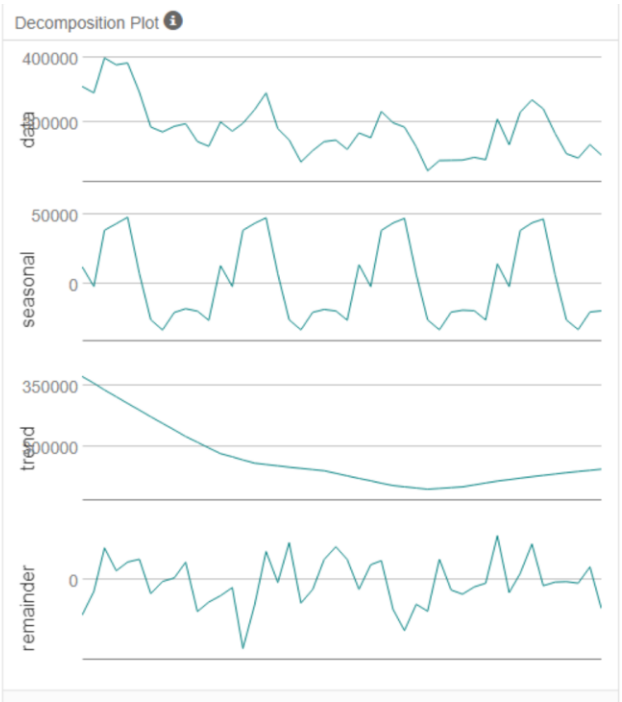
For new stores in Cluster 1, the model used for the forecast is **ETS(m,n,m)** I arrived to this decision by first identifying the best ETS and ARIMA models and finally comparing these two models.

For ETS, from the charts below:

- E: error term appears have nonconstant variance, hence we apply **multiplicatively**

- T: trend appears to be downward and plateaus at the end, hence we indicate **additive with trend dampening**
- S: seasonality sizes (peaks and valleys) appear constant over time in the decomposition plot, but the “Seasonplot” shows otherwise, hence we test this out **additively** and **multiplicatively** and retain the better result.





The ETS (M,Ad,M) AIC is lower at 938 vs. 944 for ETS(M,Ad,A). As such, we choose **ETS(M,Ad,M)** to later compare to our ARIMA model, another time series and forecasting method.

Method:
ETS(M,Ad,A)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-151.6583053	13485.3476492	10313.2567373	-0.2210662	3.5545702	0.3312225	0.016635

Information criteria:

AIC	AICc	BIC
944.2513	976.8228	974.6512

Method:
ETS(M,Ad,M)

In-sample error measures:

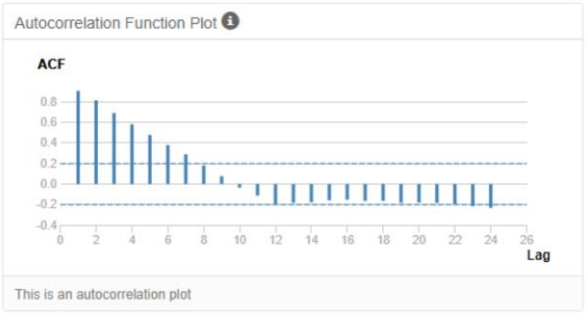
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
322.9709728	12465.3608163	10008.166743	-0.0055362	3.428519	0.3214242	-0.0409425

Information criteria:

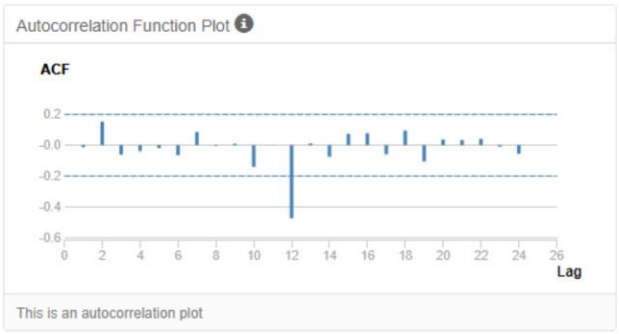
AIC	AICc	BIC
937.9875	970.5589	968.3873

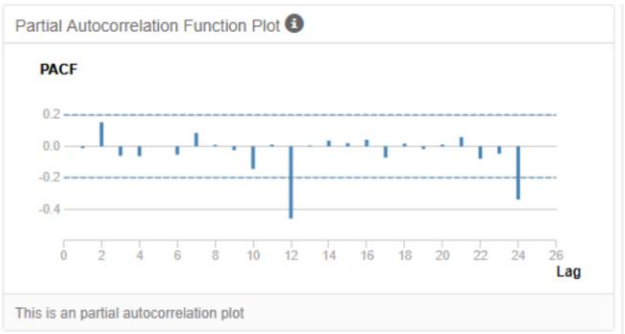
For ARIMA, the time series decomposition plot and Seasonplot (above) indicates the need for seasonal differencing.

The ACF and PACF of the **seasonally differenced** series still indicates non-stationary data with high autocorrelations in high lag levels.

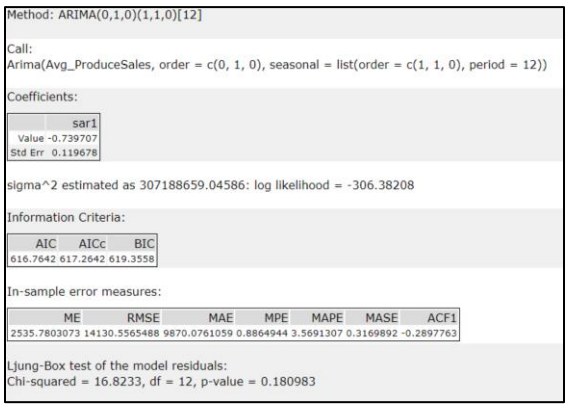


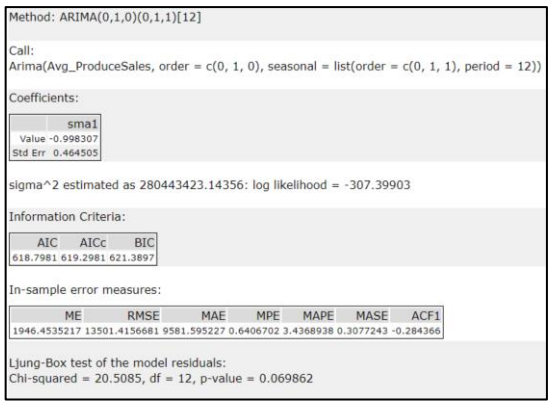
As such, we difference this data again at lag=1. The ACF and PACF indicate both p and q are zero. However, there is a spike at lag=12 for both ACF and PACF.





We test out both ARIMA (0,1,0)(1,1,0)[12] and ARIMA (0,1,0)(0,1,1)[12] and compare the results. The residuals are still close to significant level for the Seasonal MA(Q) model and the model with the lower AIC is ARIMA (0,1,0)(1,1,0)[12] at 616.76. Hence, we choose the **ARIMA (0,1,0)(1,1,0)[12] model for Cluster1.**





Comparing the ETS model and the ARIMA model using the TS compare tool favors the ETS model given lower error measures. Hence, we use the **ETS (M,Ad,M) model as our final model for Cluster 1.**

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
SARIMA_P	-12056.41	62801.29	51356.79	-6.781	17.9072	2.3492
ETS_MAM_	16787.07	57728.54	45013.47	3.5906	14.6112	2.059

We check the comparison again on the whole data set and it still favors the ETS model:

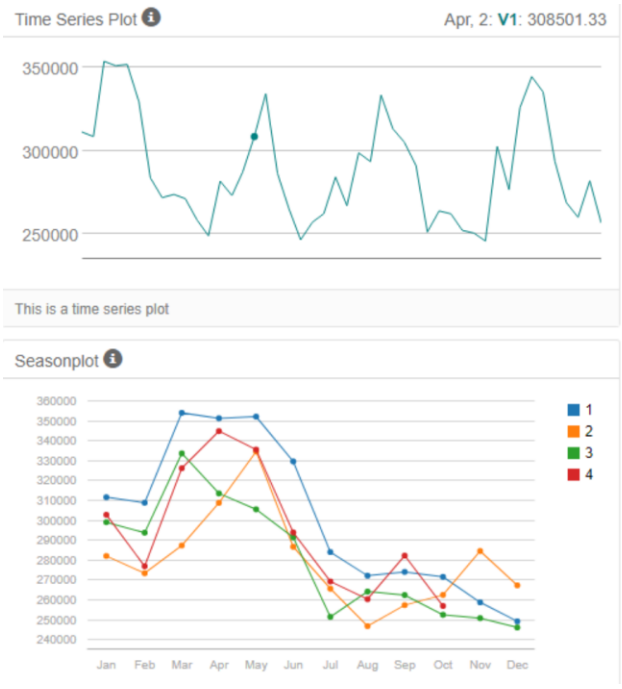
Accuracy Measures:

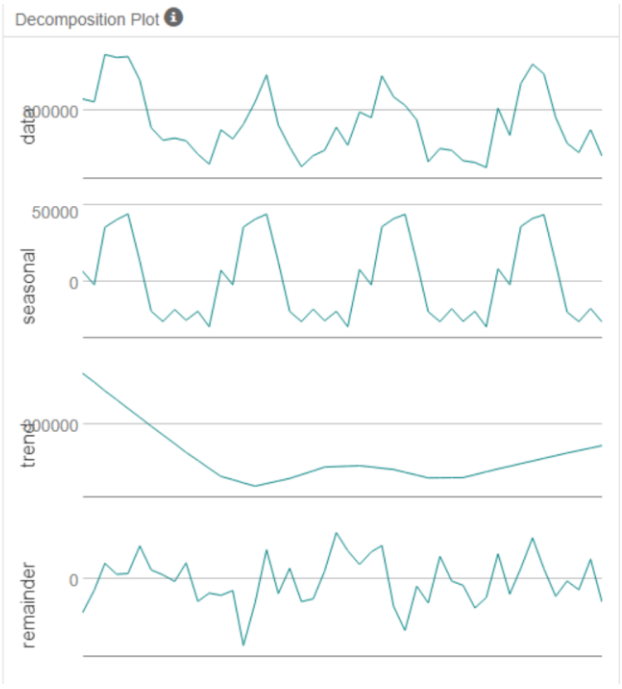
Model	ME	RMSE	MAE	MPE	MAPE	MASE
SARIMA_P	-15904.21	61816.79	51146.9	-8.0595	18.0679	2.3396
ETS_MAM_	15830.66	55123.86	42124.7	3.4627	13.7266	1.9269

For new stores in Cluster 2, the model used for the forecast is **ETS(M,N,M)** I arrived to this decision by first identifying the best ETS and ARIMA models and finally comparing these two models.

For ETS, from the charts below:

- E: error term appears have nonconstant variance, hence we apply **multiplicatively**
- T: no clear trend, hence we indicate **none**
- S: seasonality sizes (peaks and valleys) appear constant over time in the decomposition plot, but the "Seasonplot" shows otherwise, hence we test this out **additively** and **multiplicatively** and retain the better result.





The ETS (M,N,M) AIC is lower at 924 vs. 932 for ETS(M,N,A). As such, we choose **ETS(M,N,M)** to later compare to our ARIMA model, another time series and forecasting method.

Method:
ETS(M,N,A)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
464.099255	12784.3239436	10510.6283492	-0.0402222	3.5963842	0.4666316	0.0235294

Information criteria:

AIC	AICc	BIC
932.5075	952.5075	957.8406

Method:
ETS(M,N,M)

In-sample error measures:

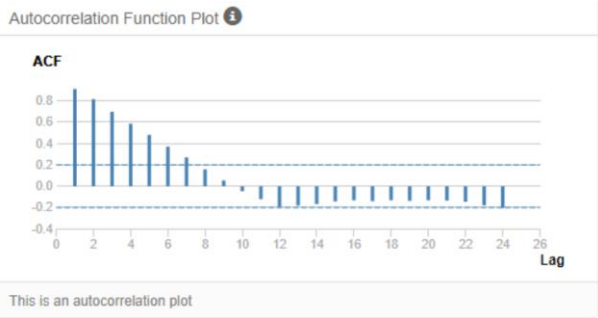
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
348.7384344	11620.1624901	9652.8680831	0.0041204	3.2984172	0.4285503	0.0305734

Information criteria:

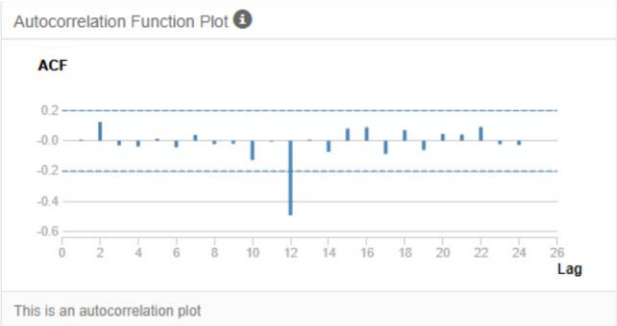
AIC	AICc	BIC
924.5778	944.5778	949.911

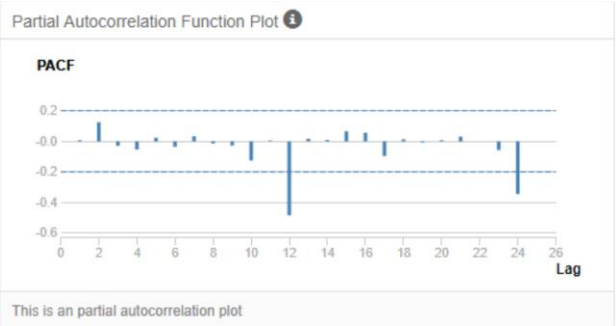
For ARIMA, the time series decomposition plot and Seasonplot (above) indicates the need for seasonal differencing.

The ACF and PACF of the **seasonally differenced** series still indicates non-stationary data with high autocorrelations in high lag levels.

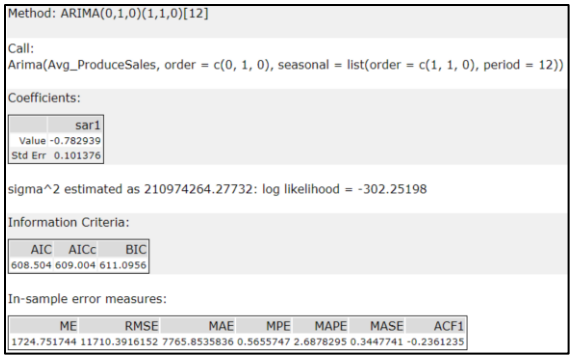


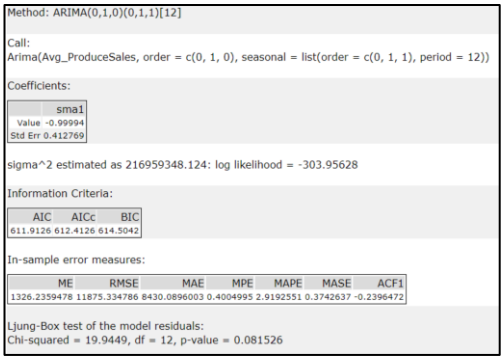
As such, we difference this data again at lag=1. Similar to Cluster1, the ACF and PACF indicate both p and q are zero. However, there is a spike at lag=12 for both ACF and PACF.





We test out both ARIMA (0,1,0)(1,1,0)[12] and ARIMA (0,1,0)(0,1,1)[12] and compare the results. The residuals are still close to significant level for the Seasonal MA(Q) model and the model with the lower AIC is ARIMA (0,1,0)(1,1,0)[12] at 608.5. Hence, we choose the **ARIMA (0,1,0)(1,1,0)[12] model for Cluster2.**





Comparing the ETS model and the ARIMA model using the TS compare tool favors the ETS model given lower error measures. Hence, we use the **ETS (M,N,M) model as our final model for Cluster 2.**

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-9655.898	50546.18	43809.12	-4.9068	15.2102	2.2977
SARIMA_P	-47317.25	71983.77	59976.78	-18.1771	21.8409	3.1457

We check the comparison again on the whole data set and it still favors the ETS model:

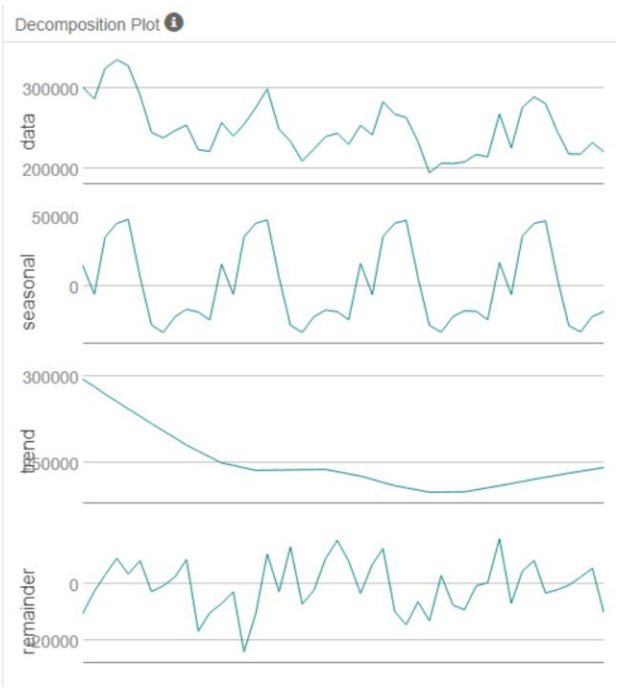
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-8486.591	48630.74	41681.71	-4.4172	14.4579	2.1861
SARIMA_P	-50201.525	72699.26	61209.82	-19.1518	22.3378	3.2103

For new stores in Cluster 3, the model used for the forecast is **ETS(M,N,A)** I arrived to this decision by first identifying the best ETS and ARIMA models and finally comparing these two models.

For ETS, from the charts below:

- E: error term appears have nonconstant variance, hence we apply **multiplicatively**
- T: trend appears to be downward and plateaus at the end, hence we indicate **additive with trend dampening**
- S: seasonality sizes (peaks and valleys) appear constant over time in the decomposition plot, but the “Seasonplot” shows otherwise, hence we test this out **additively** and **multiplicatively** and retain the better result.



The ETS (M,Ad,M) AIC is lower at 928 vs. 932 for ETS(M,Ad,A). As such, we choose **ETS(M,Ad,M)** to later compare to our ARIMA model, another time series and forecasting method.

Method:
ETS(M,Ad,A)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-91.7789664	11640.9076458	9524.0278323	-0.2314535	3.8225671	0.3916159	-0.0137693

Information criteria:

AIC	AICc	BIC
932.5154	965.0869	962.9153

Method:
ETS(M,Ad,M)

In-sample error measures:

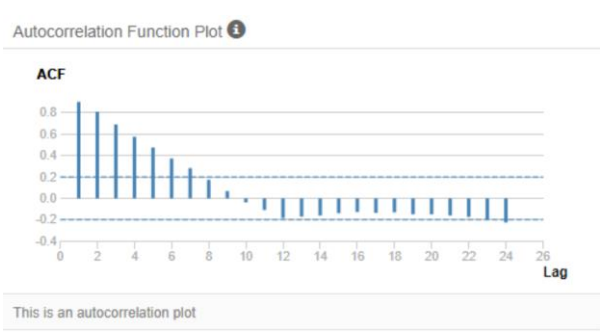
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
422.0953389	11036.9177776	8765.6225662	0.0328792	3.5363322	0.3604312	-0.0334552

Information criteria:

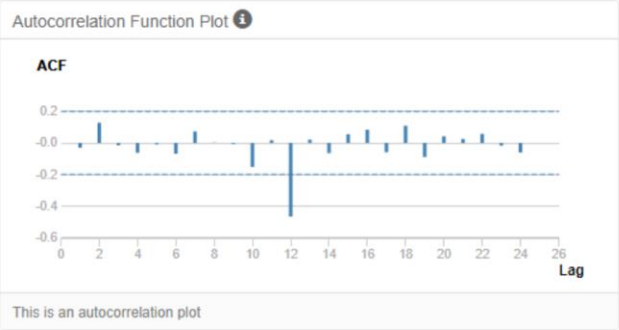
AIC	AICc	BIC
928.4196	960.991	958.8194

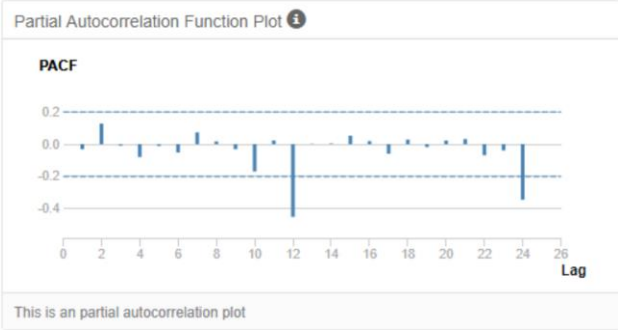
For ARIMA, the time series decomposition plot and Seasonplot (above) indicates the need for seasonal differencing.

The ACF and PACF of the **seasonally differenced** series still indicates non-stationary data with high autocorrelations in high lag levels.



As such, we difference this data again at lag=1. Like Clusters 1 and 2, the ACF and PACF indicate both p and q are zero. However, there is a spike at lag=12 for both ACF and PACF.





We test out both ARIMA (0,1,0)(1,1,0)[12] and ARIMA (0,1,0)(0,1,1)[12] and compare the results. The residuals are still close to significant level for the Seasonal MA(Q) model and the model with the lower AIC is ARIMA (0,1,0)(1,1,0)[12] at 611.989. Hence, we choose the **ARIMA (0,1,0)(1,1,0)[12] model for Cluster3.**

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

Call:
Arima(Avg_ProduceSales, order = c(0, 1, 0), seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:

	sar1
Value	-0.643765
Std Err	0.150801

sigma^2 estimated as 288527001.99018: log likelihood = -303.99451

Information Criteria:

AIC	AICc	BIC
611.989	612.489	614.5807

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
1963.061924	13694.617603	9502.9456748	0.7321188	3.9273029	0.390749	-0.3635958

Ljung-Box test of the model residuals:
Chi-squared = 17.901, df = 12, p-value = 0.139061

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

Call:
Arima(Avg_ProduceSales, order = c(0, 1, 0), seasonal = list(order = c(0, 1, 1), period = 12))

Coefficients:

	sma1
Value	-0.999792
Std Err	0.610273

sigma^2 estimated as 218970880.13457: log likelihood = -304.07911

Information Criteria:

AIC	AICc	BIC
612.1582	612.6582	614.7499

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
1457.3524002	11930.2586765	8554.5078977	0.5201664	3.538881	0.3517504	-0.3554955

Ljung-Box test of the model residuals:
Chi-squared = 19.8829, df = 12, p-value = 0.082908

Comparing the ETS model and the ARIMA model using the TS compare tool favors the ETS model given lower error measures. Hence, we use the **ETS (M,Ad,M) model as our final model for Cluster 3.**

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	1702.931	47850.28	39606.18	-1.4874	15.4743	1.9075
SARIMA_P	-12941.995	54672.59	45875.41	-7.6678	18.6461	2.2095

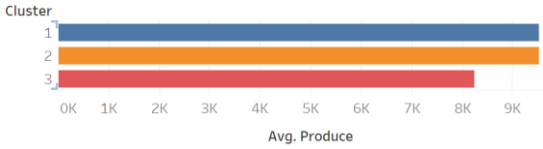
We check the comparison again on the whole data set and it still favors the ETS model:

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	735.2048	45740.57	37567.03	-1.7299	14.7543	1.8093
SARIMA_P	-15844.6707	53814	45594.55	-8.7637	18.707	2.1959

An additional observation is that while cluster 1 has the highest average produce sales

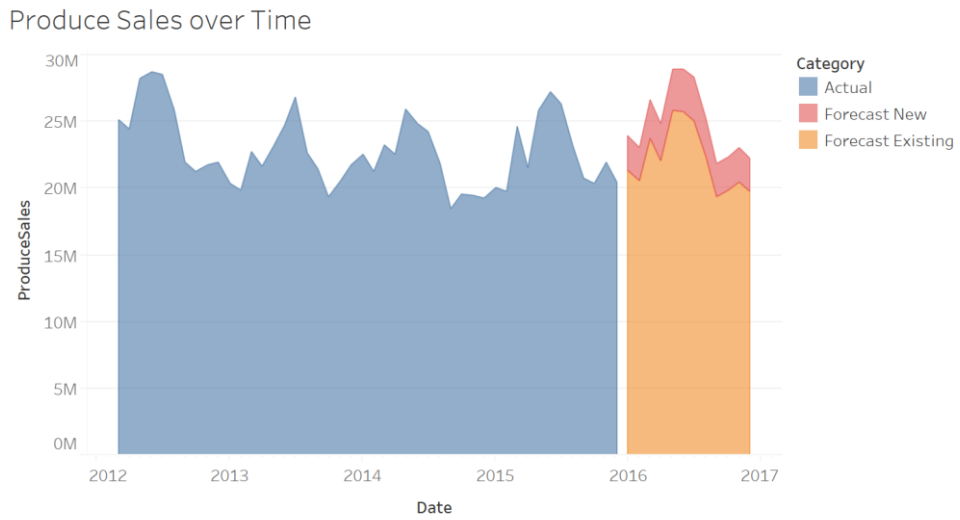
historically, the time series indicated a decline in overall sales, i.e. downward trend.



As such, the cluster with the highest (average) sales forecast is cluster 2.

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.

Month	New Stores	Existing Stores
Jan-16	2,592,072	21,370,818
Feb-16	2,492,866	20,525,731
Mar-16	2,900,298	23,684,288
Apr-16	2,774,619	22,073,944
May-16	3,135,161	25,826,610
Jun-16	3,187,182	25,708,731
Jul-16	3,220,571	25,059,365
Aug-16	2,858,047	22,355,893
Sep-16	2,525,367	19,333,714
Oct-16	2,476,852	19,829,131
Nov-16	2,565,588	20,428,496
Dec-16	2,536,407	19,720,851



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.