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

Q1- 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 result of Adjusted Rand indices and Calinski-Harabasz Indices, the cluster with the highest medium value is cluster 3

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.01155	0.3083	0.213	0.2837	0.2762
1st Quartile	0.3814	0.5258	0.4169	0.374	0.3965
Median	0.5619	0.6653	0.5107	0.4406	0.4256
Mean	0.5084	0.6594	0.5471	0.4704	0.4502
3rd Quartile	0.6942	0.7865	0.6427	0.5199	0.5067
Maximum	1	1	0.8902	0.8207	0.6626

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	16.1	18.94	18.45	17.02	17.37
1st Quartile	28.42	28.68	25.16	22.91	21.28
Median	29.47	30.83	26.61	23.98	22.17
Mean	28.24	29.58	26.34	23.7	21.95
3rd Quartile	30.31	31.97	27.85	24.9	22.84
Maximum	31.44	33.26	30.37	26.53	24.87

Figure 1: K-Means Cluster Assessment Report

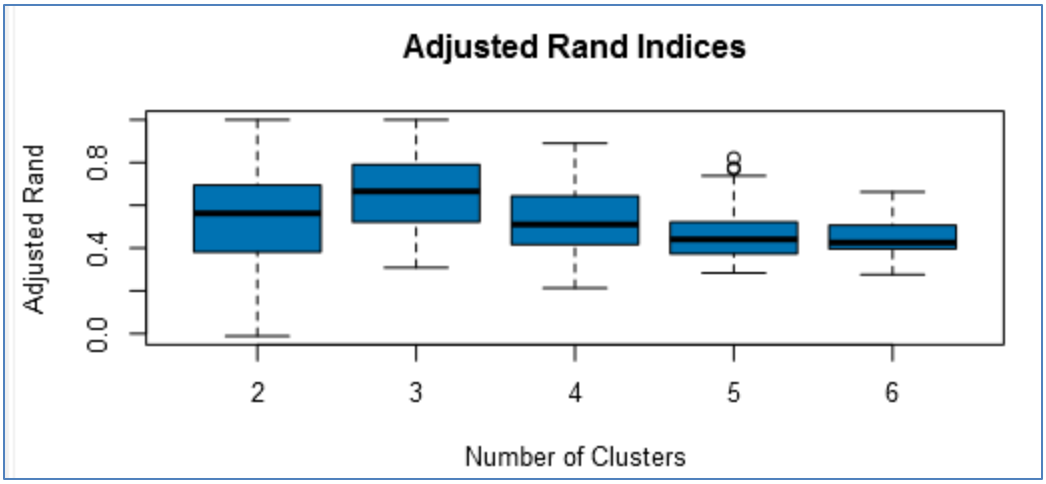

Figure 2: Adjusted Rand Indices Plot

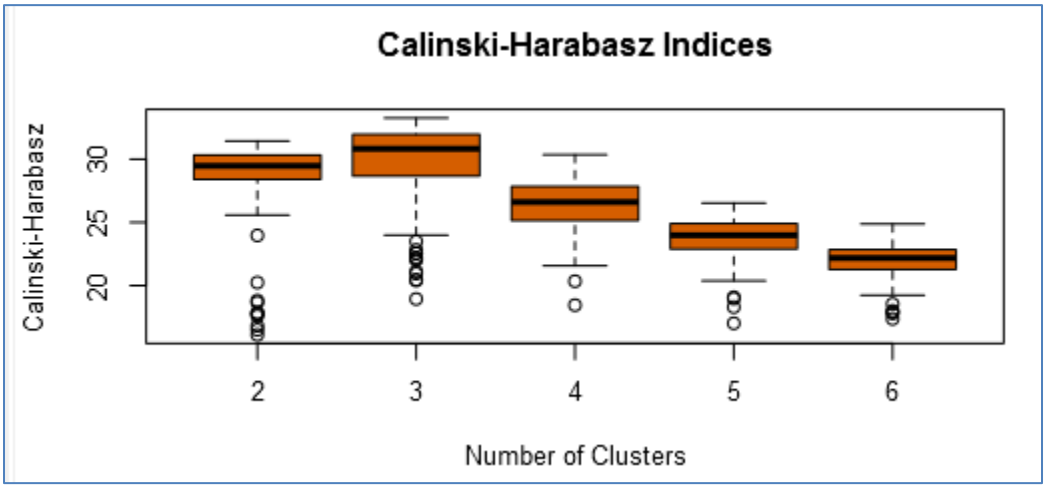

Figure 3: Calinski-Harabasz Indices Plot

Q2- How many stores fall into each store format?

- After applying the K-Centroids Cluster Analysis:
  - Cluster 1 has 23 stores
  - Cluster 2 has 29 stores
  - Cluster 3 has 33 stores

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Figure 4: Cluster Information

Q3- Based on the results of the clustering model, what is one way that the clusters differ from one another?

- Based on the box and whisker plot visualization
    - The total sales of Cluster 1 stores is the greatest
    - The total sales of Cluster 3 stores is the fewest

https://public.tableau.com/profile/manar3259#!/vizhome/Thetotalsalesofstoresforeachcategories/Thetotalsalesofstoresforeachcategories?publish=yes

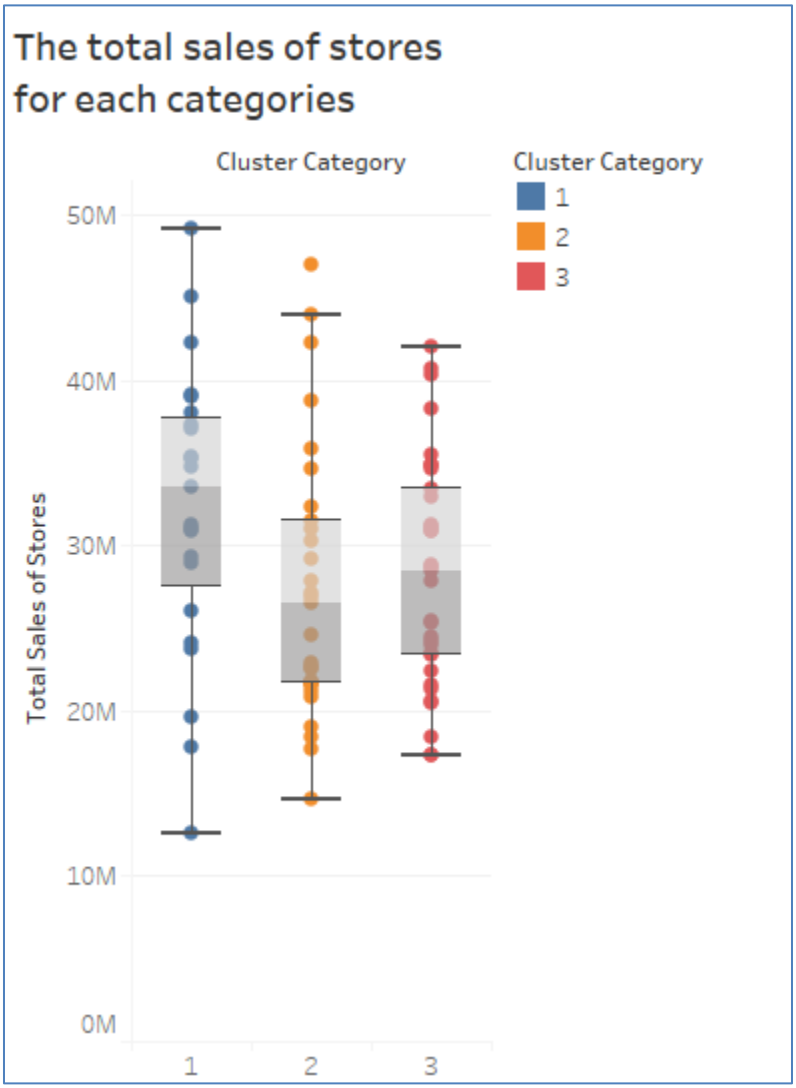

Figure 5: The total sales of stores for each categories

Q4- 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/profile/manar3259#!/vizhome/Task1-Q4-Thelocationofstores/Thelocationofstores?publish=yes

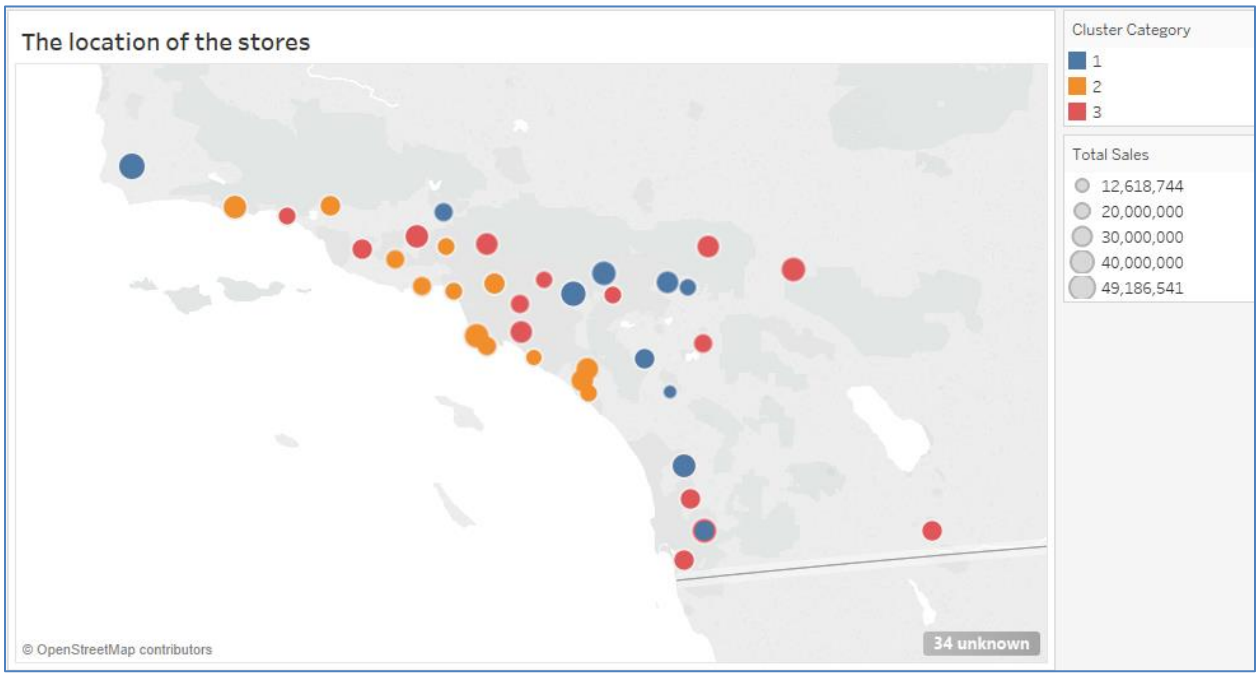

Figure 6: the location of stores

Task 2: Formats for New Stores

Q1- 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.)

- After applying the model comparison on the decision tree, forest, and boosted models
  - The accuracy values of forest and boosted models are the same
  - The F1 value of boosted model is the highest
  - So, the boosted model is chosen

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
S_DT	0.7059	0.7327	0.6000	0.6667	0.8333
S_FM	0.8235	0.8251	0.7500	0.8000	0.8750
S_BM	0.8235	0.8543	0.8000	0.6667	1.0000

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Figure 7: Model Comparison Report

Q2- 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

Q1- 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) ETS Model

Step 1 - Select ETS model terms:

- The Decomposition Plot shows how each of the trend, seasonal and error components should be applied.
    - The trend is not clear, nothing will be applied.
    - The Seasonality increases over time, it will be applied multiplicatively.
    - The error increases over time, it will be applied multiplicatively.
- So, the non-damped ETS(M,N,M) model is chosen

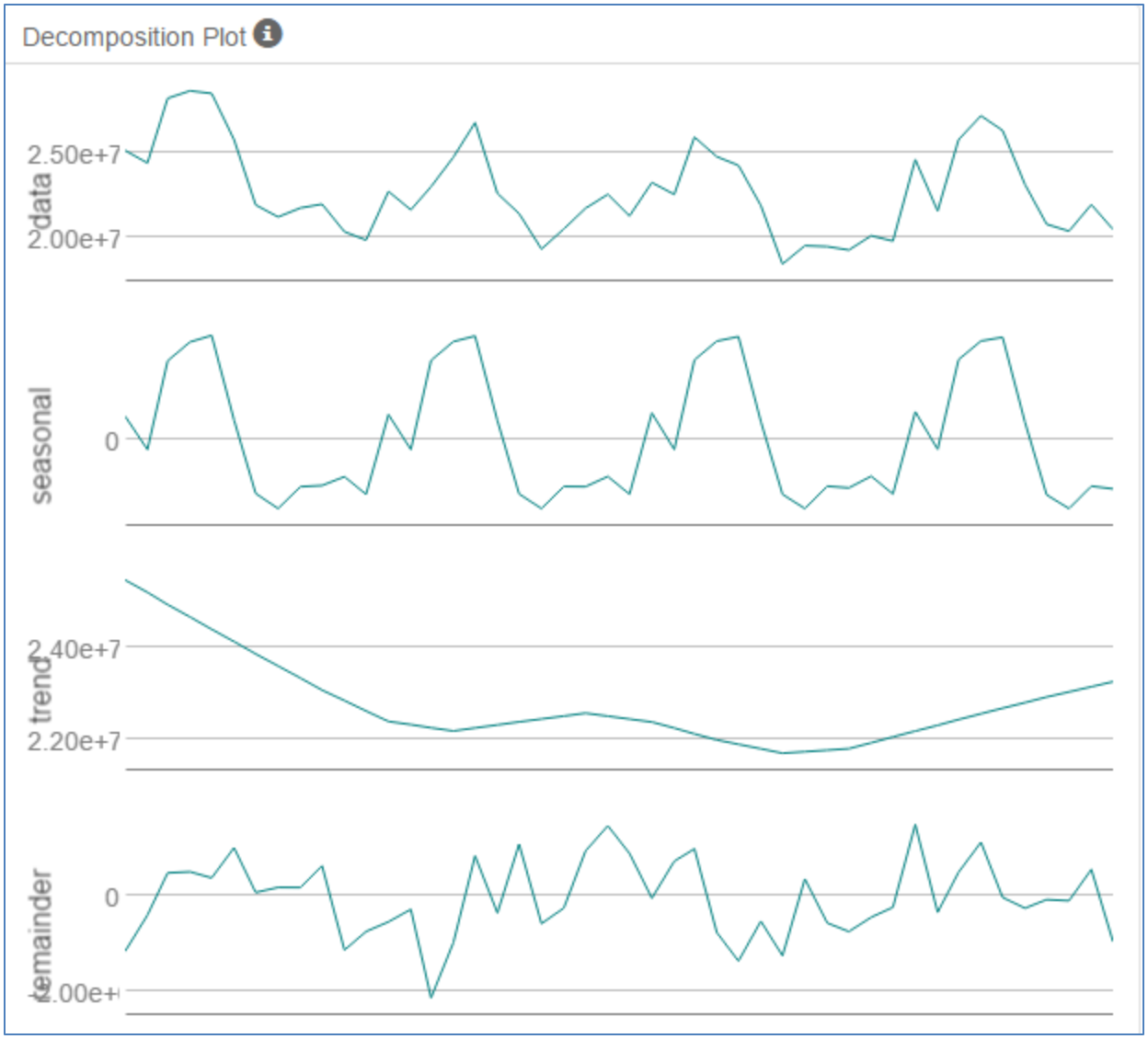

Figure 8: Decomposition Plot

Step 2 - Build the model:

- The result of in-sample errors of non-damped ETS(M,N,M) model
  - RMSE = 1020596.9042405
  - MASE= 0.4506721
  - AIC= 1283.1197

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

Figure 9: In-sample error measures and Information criteria of ETS(M,N,M)

Step 3 - Validate model:

- The result of Forecast error measurements of non-damped ETS(M,A,M) model
  - RMSE = 760267.3
  - MASE= 0.3822

Comparison of Time Series Models

Actual and Forecast Values:

Actual	ETS_M_N_M_
26338477.15	26907095.61191
23130626.6	22916903.07434
20774415.93	20342618.32222
20359980.58	19883092.31778
21936906.81	20479210.4317
20462899.3	21211420.14022

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_M_N_M_	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA

Figure 10: Actual and Forecast Values and Accuracy Measures of ETS(M,N,M)

B) ARIMA Model

Step 1 - Check stationarity:

- Time Series is non-stationary

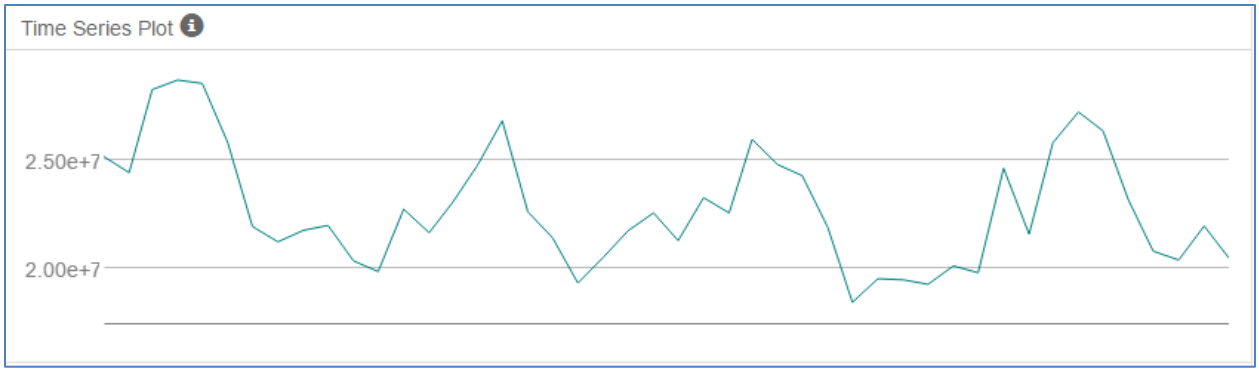

Figure 11: Time Series Plot of ARIMA

- The Auto-Correlation Function (ACF) indicates a high correlation between points.
- The Partial Autocorrelation Function Plots (PACF) displays a significant lag at point 11, because of seasonality.

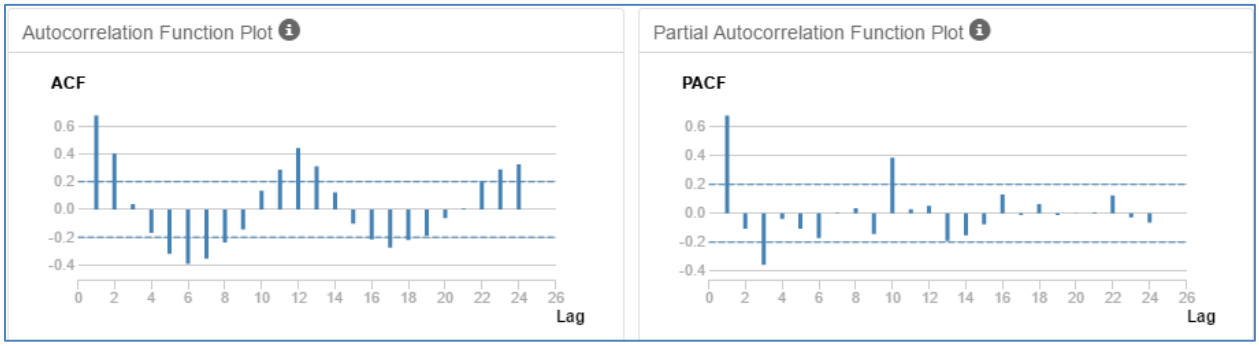

Figure 12: ACF and PACF of ARIMA

Step 2 - Difference:

Because the time series is non-stationary, we need to take a seasonal difference.

First seasonal difference

- After applying the first seasonal difference, the time series is still non-stationary.

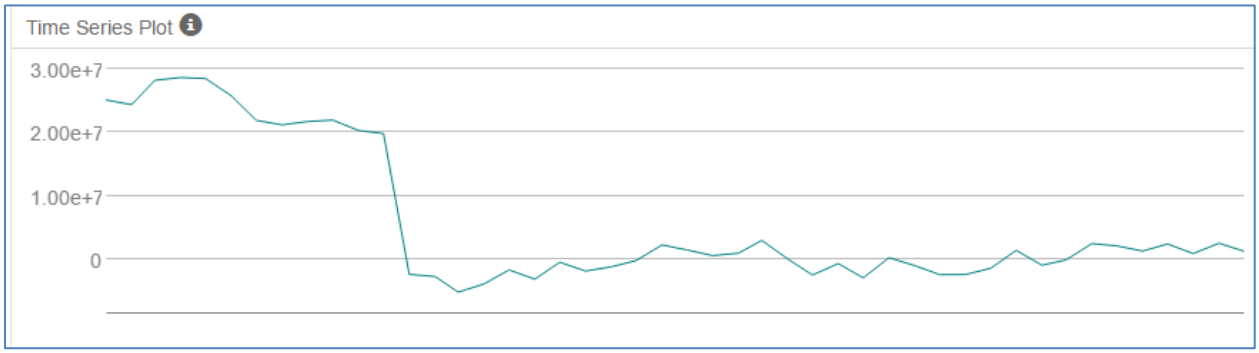

Figure 13: The first difference of a time series

- The correlation between points is high as shown in ACF
- There's no strong correlation between the points as shown in PACF

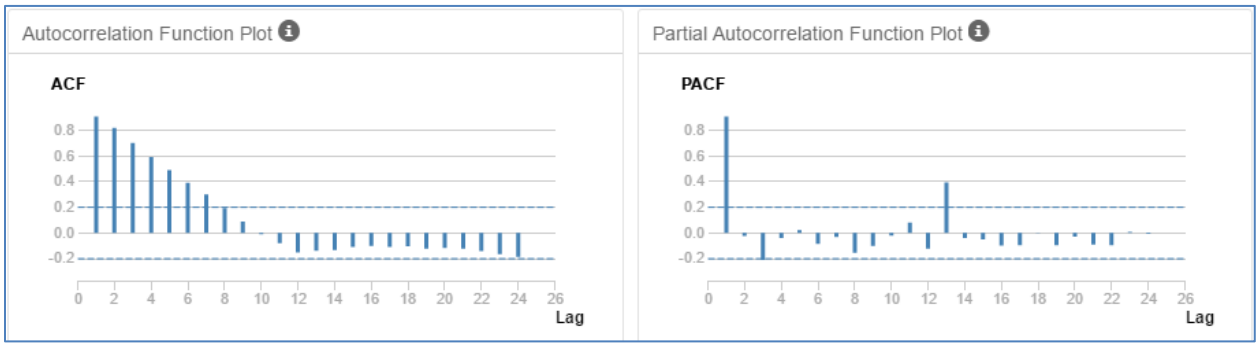

Figure 14: The first difference of ACF and PACF

Second seasonal difference

- After applying the second seasonal difference, the time series is now a stationary.

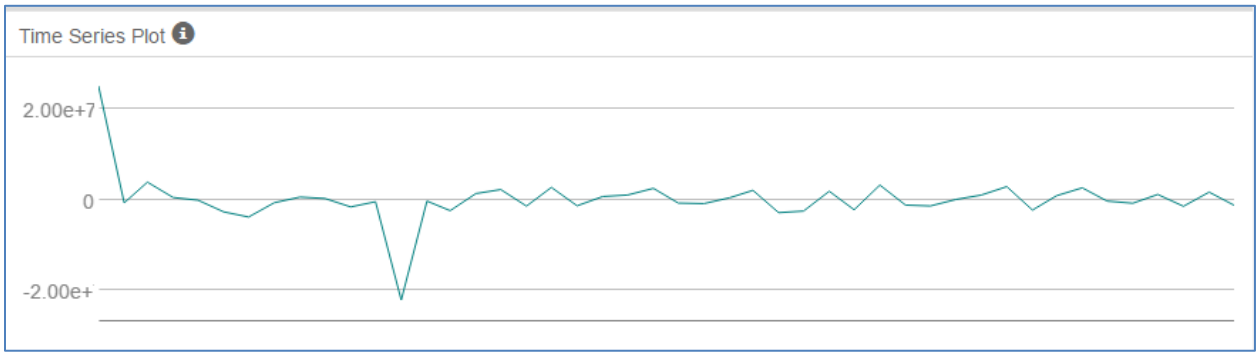

Figure 15: The second difference of a time series

- The correlation between points is not high anymore as shown in ACF
- There's no strong correlation between the points as shown in PACF

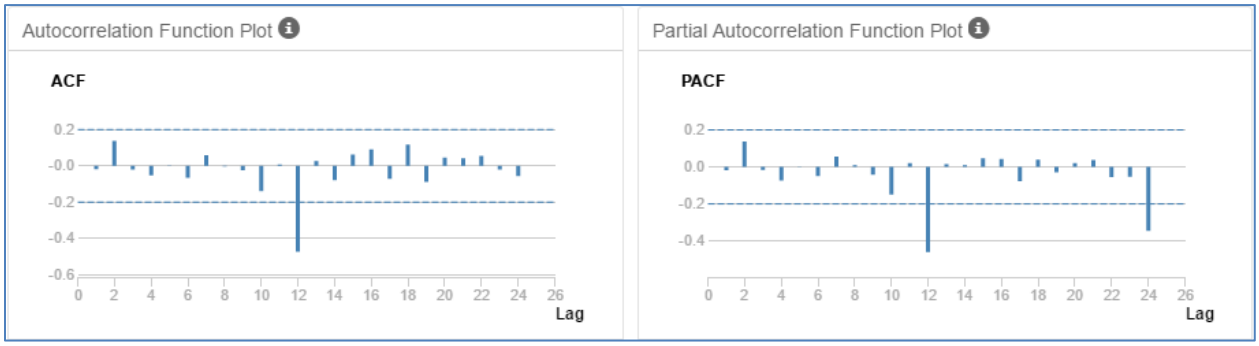

Figure 16: The second difference of ACF and PACF

Step 3 - Select AR and MA terms:

- The auto option of ARIMA model terms is ARIMA(1,0,0)(1,1,0)[12]
- The model includes a non-seasonal AR(1) term (p), a seasonal AR(1) term (P), Seasonal differencing (D), no regular difference (d), no MA terms (q,Q).

Step 4 - Build the model:

- The result of in-sample errors of ARIMA(1,0,0)(1,1,0)[12] model
  - RMSE = 1042209.8528363
  - MASE= 0.4120218
  - AIC= 880.4445

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

Figure 17: Information criteria and In-sample error measures of ARIMA(1,0,0)(1,1,0)[12]

Step 5 - Validate model:

- The result of Forecast error measurements of ARIMA(1,0,0)(1,1,0)[12] model
  - RMSE = 1050239
  - MASE= 0.5463

Comparison of Time Series Models

Actual and Forecast Values:

Actual	ARIMA_Auto
26338477.15	27997835.63764
23130626.6	23946058.0173
20774415.93	21751347.87069
20359980.58	20352513.09377
21936906.81	20971835.10573
20462899.3	21609110.41054

Accuracy Measures:

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

Figure 18: Actual and Forecast Values and Accuracy Measures of ARIMA(1,0,0)(1,1,0)[12]

C) Choose the best model

- ETS Model is better than ARIMA Model, because of the following:
  - The values of RMSE and MASE are smaller
    - RMAS value of ETS = 760267.3 and ARIMA= 1050239
    - MASE value of ETS = 0.3822 and ARIMA= 0.5463
  - It tends to predict values more accurately

Q2- 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.

A table with the 12 month forecasts for existing and new stores

ETS Model was used for forecasting sales of the existing and new stores in 2016

Month	New Stores	Existing stores
Jan-16	2567021.14154645	21136208.1351094
Feb-16	2457620.10648943	20506604.6898891
Mar-16	2891987.02636857	23506131.4573967
Apr-16	2751975.85507785	22207971.2384362
May-16	3125313.92103624	25376698.3221854
Jun-16	3184236.9573107	25963559.4465763
Jul-16	3221864.94574349	26113357.20163
Aug-16	2852755.80480619	22904671.9176674
Sep-16	2552504.13848453	20499151.0012101
Oct-16	2494823.97582237	19970808.9473091
Nov-16	2588702.83451654	20602232.2973705
Dec-16	2568322.19494756	21072786.9221559

A visualization of stores sales forecasts that includes historical data, existing stores forecasts, and new stores forecasts

https://public.tableau.com/profile/manar3259#!/vizhome/Thesalesdataforhistoricaldataexistingstoresforecastsandnewstoresforecasts/Sheet1?publish=yes

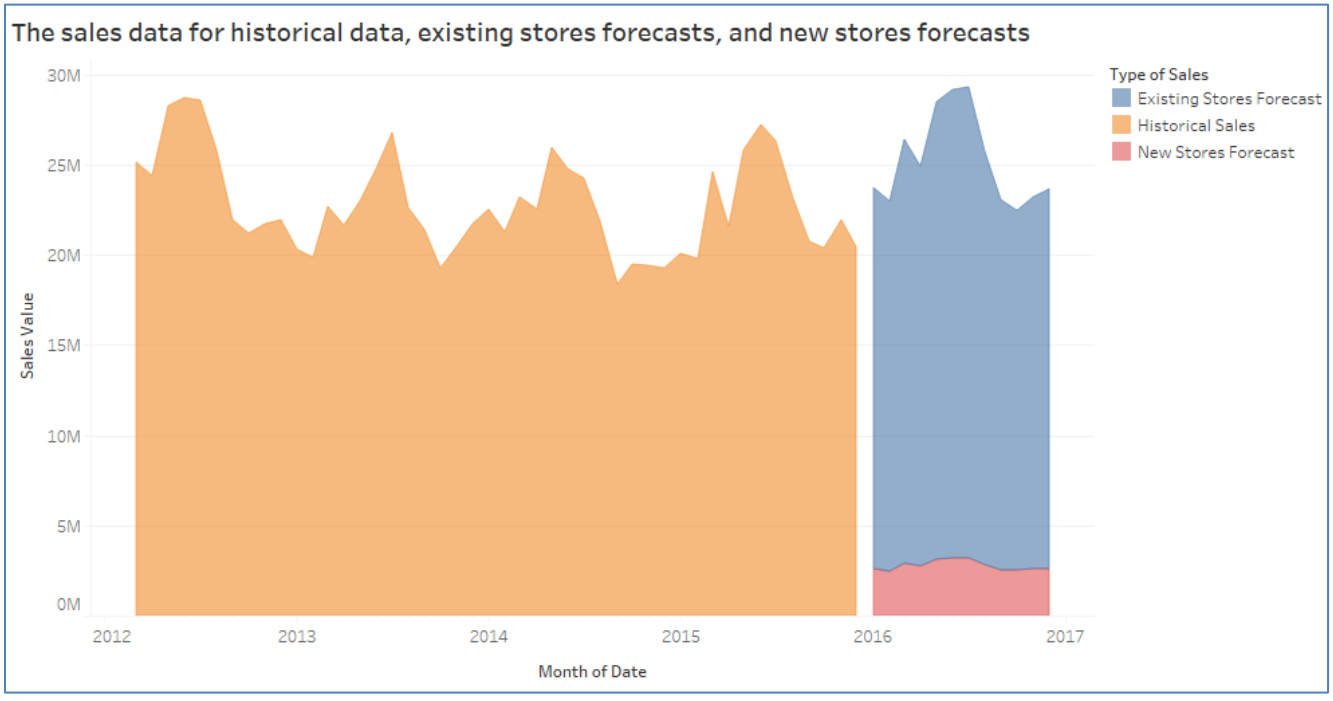

Figure 19: The sales data for historical data, existing stores forecasts, and new stores forecasts

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.