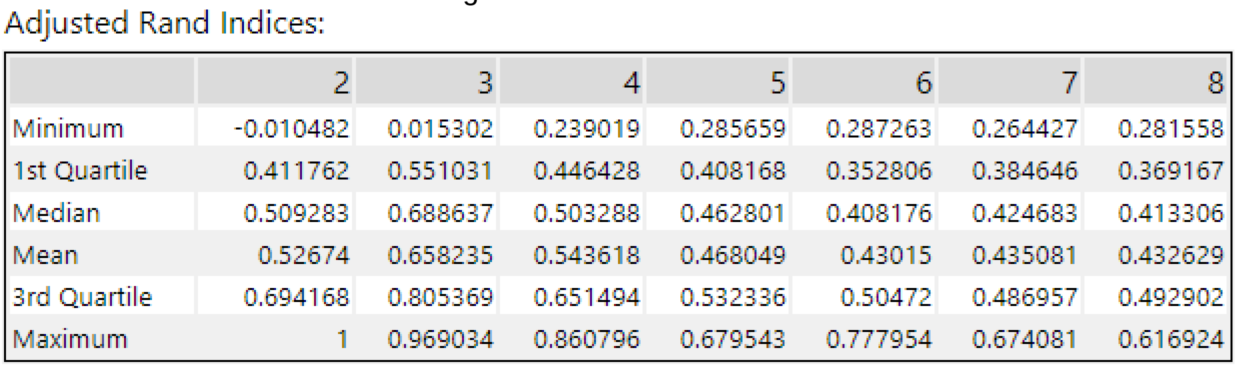
Project: Predictive Analytics Capstone

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

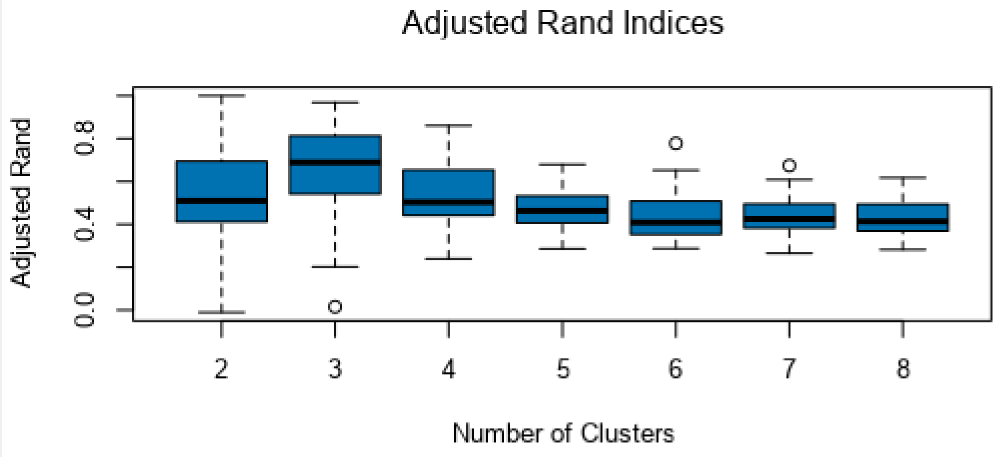
Task 1: Determine Store Formats for Existing Stores

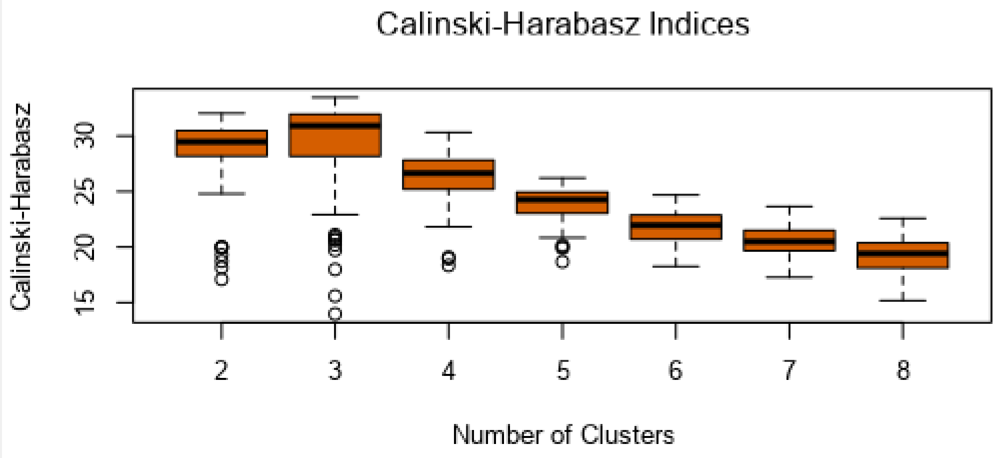
1. What is the optimal number of store formats? How did you arrive at that number?
   The optimal number of store formats is 3. I arrived this number by using K-Centroids Cluster Analysis and K-Centroids Diagnostics Tools with K-Means Clustering Method. As cluster 3 median in ARI is the largest



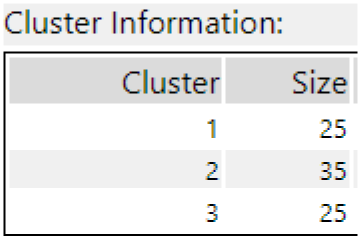
Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.010482	0.015302	0.239019	0.285659	0.287263	0.264427	0.281558
1st Quartile	0.411762	0.551031	0.446428	0.408168	0.352806	0.384646	0.369167
Median	0.509283	0.688637	0.503288	0.462801	0.408176	0.424683	0.413306
Mean	0.52674	0.658235	0.543618	0.468049	0.43015	0.435081	0.432629
3rd Quartile	0.694168	0.805369	0.651494	0.532336	0.50472	0.486957	0.492902
Maximum	1	0.969034	0.860796	0.679543	0.777954	0.674081	0.616924





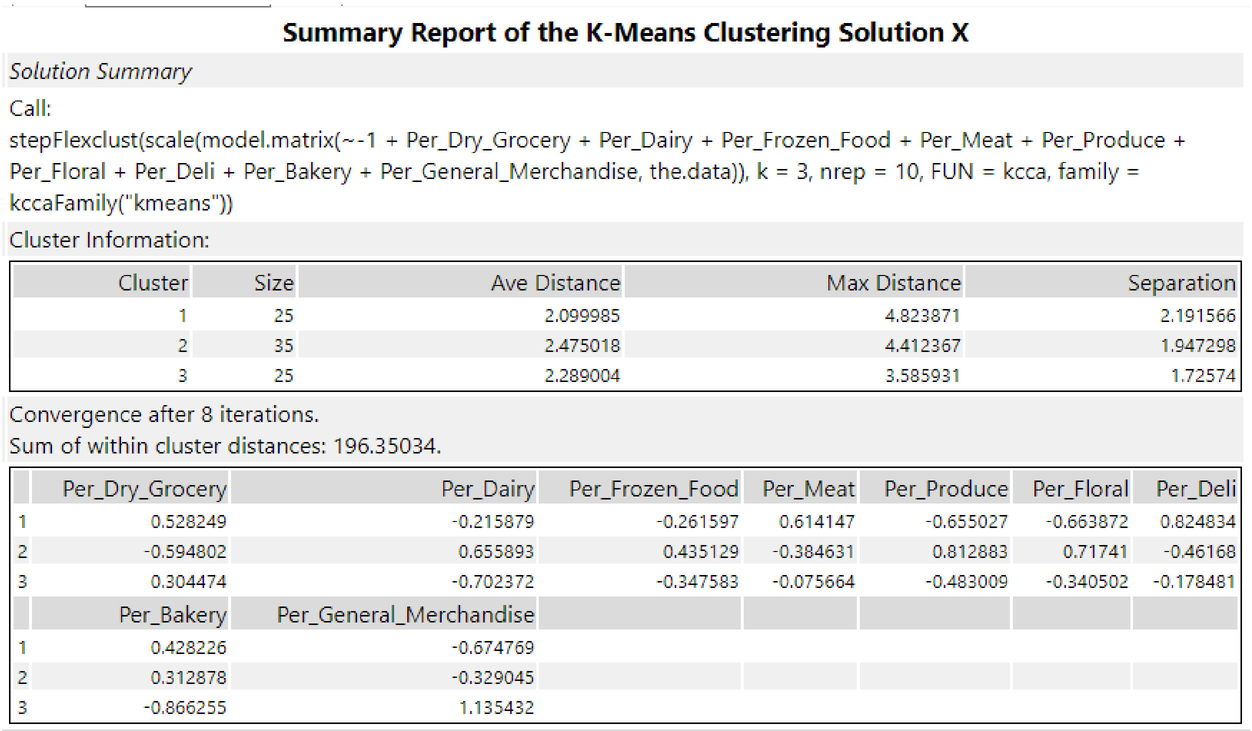
2. How many stores fall into each store format?



Cluster Information:

Cluster	Size
1	25
2	35
3	25

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
Cluster 1: lowest in average total sale and largest in (Dry_grocery, Meat,deli,Bakery)
Cluster 2: largest in (Dairy-frozen_food-produce-floral-)
Cluster 3: Largest in average total sale, largest in (General Merchandise)



Summary Report of the K-Means Clustering Solution X

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Per_Dry_Grocery + Per_Dairy + Per_Frozen_Food + Per_Meat + Per_Produce + Per_Floral + Per_Deli + Per_Bakery + Per_General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

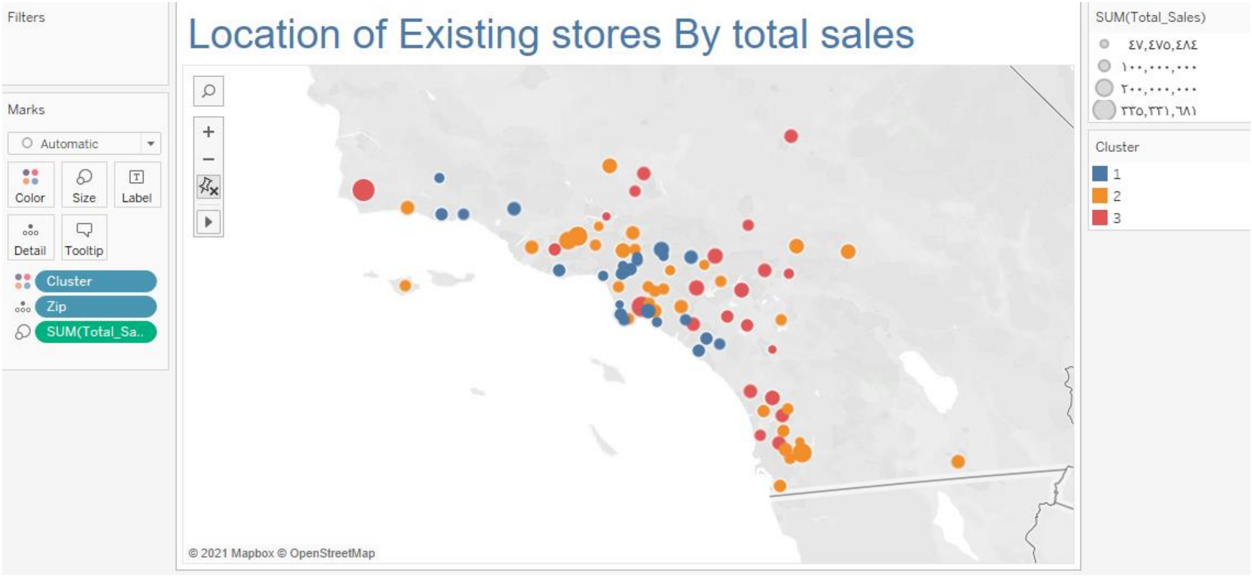
Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

	Per_Dry_Grocery	Per_Dairy	Per_Frozen_Food	Per_Meat	Per_Produce	Per_Floral	Per_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655027	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178481

	Per_Bakery	Per_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

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.

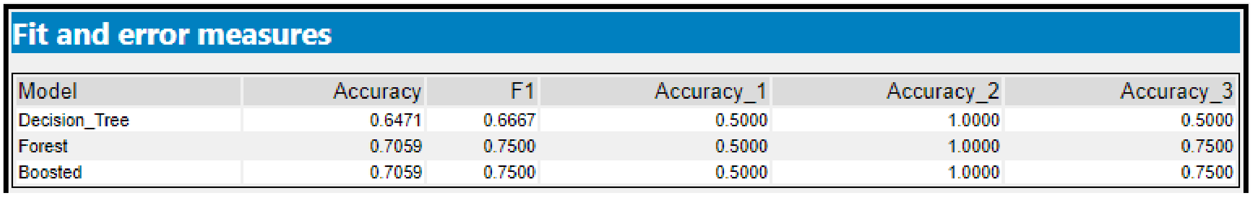


Location of Existing stores By total sales

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 the following table, any of the three methodologies can be used, However, I will eliminate the decision tree as it is the worst accurate.



Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.6471	0.6667	0.5000	1.0000	0.5000
Forest	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted	0.7059	0.7500	0.5000	1.0000	0.7500

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

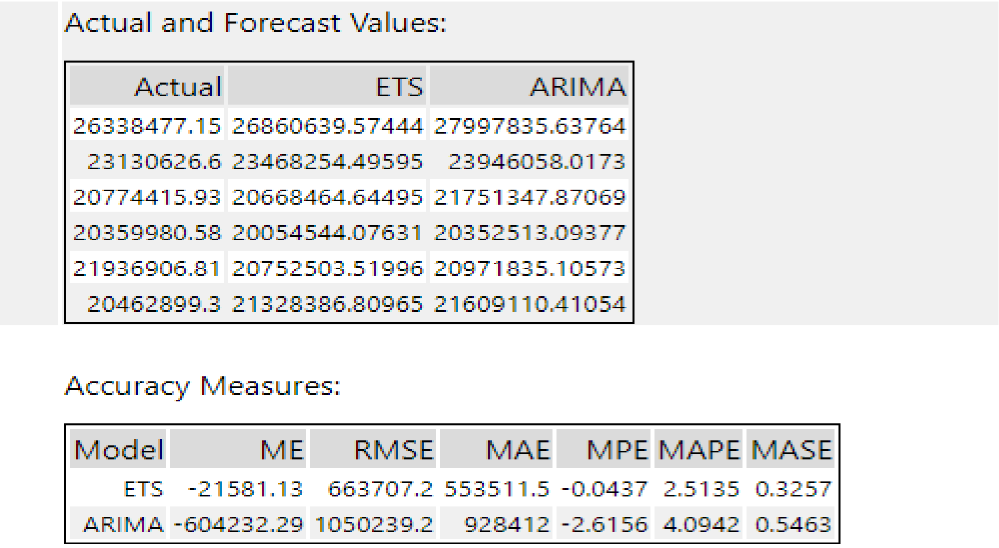
Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

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?

I used ETS model for forecast. I came to this decision after comparing between ETS and ARIMA and using TS Plot tool. Therefore, the forecast should be ETS(M, N, M)

Comparison of Time Series Models

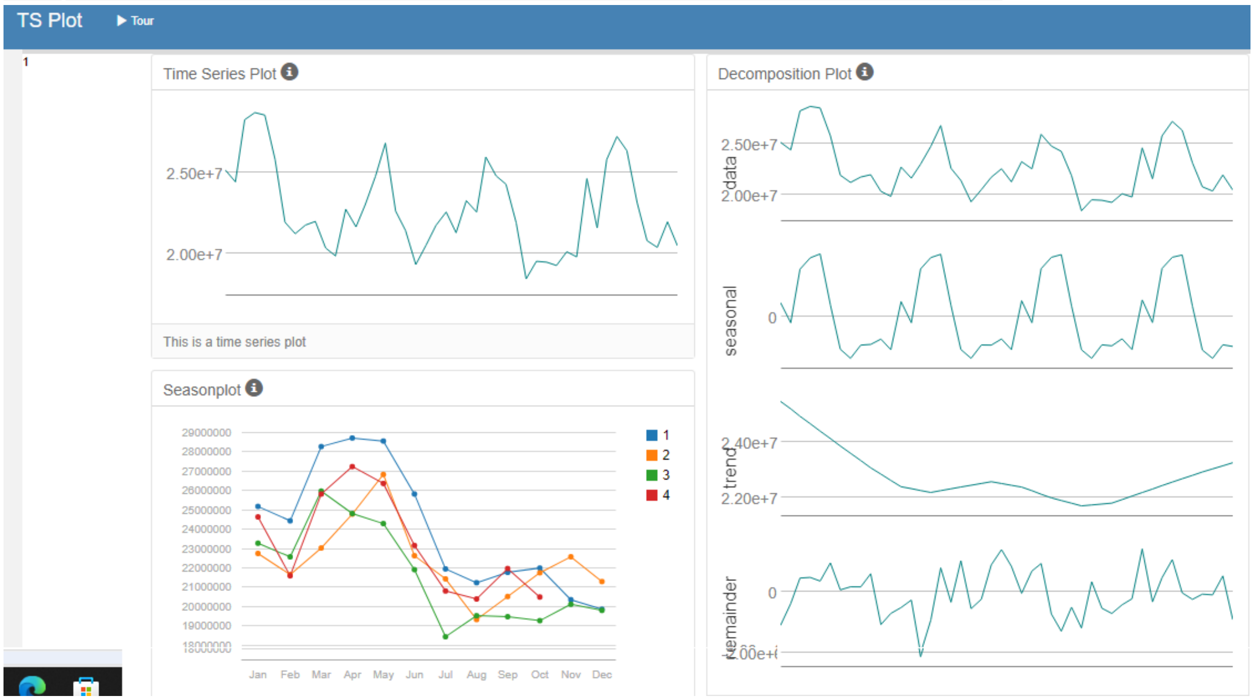


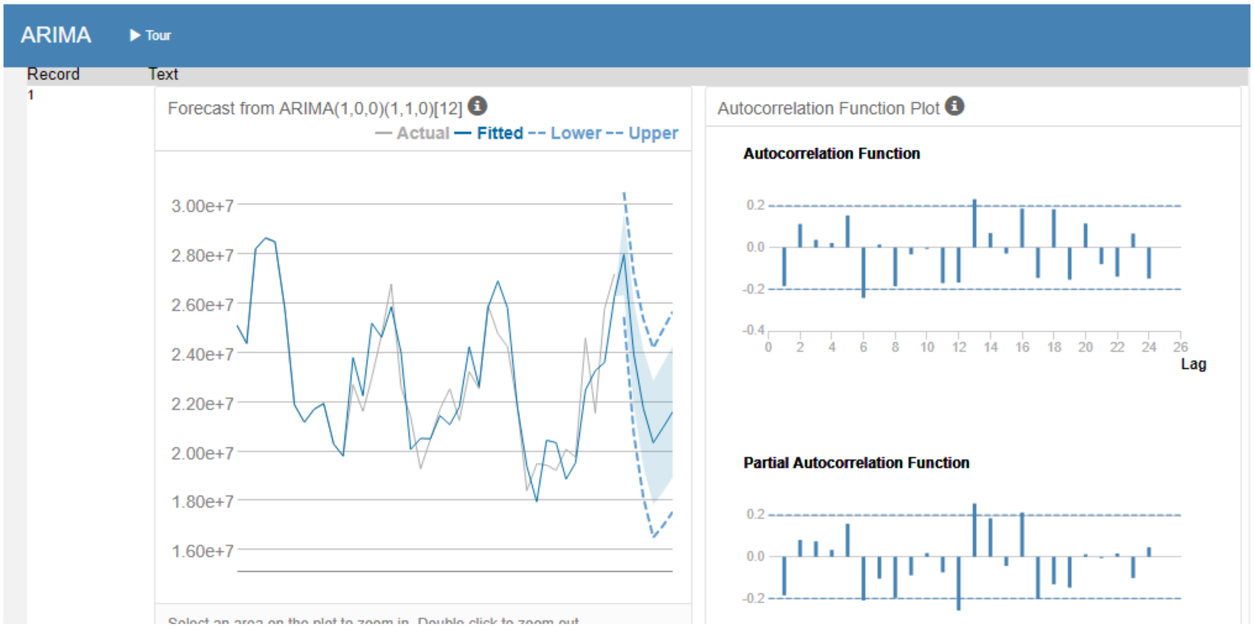
Actual and Forecast Values:

Actual	ETS	ARIMA
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

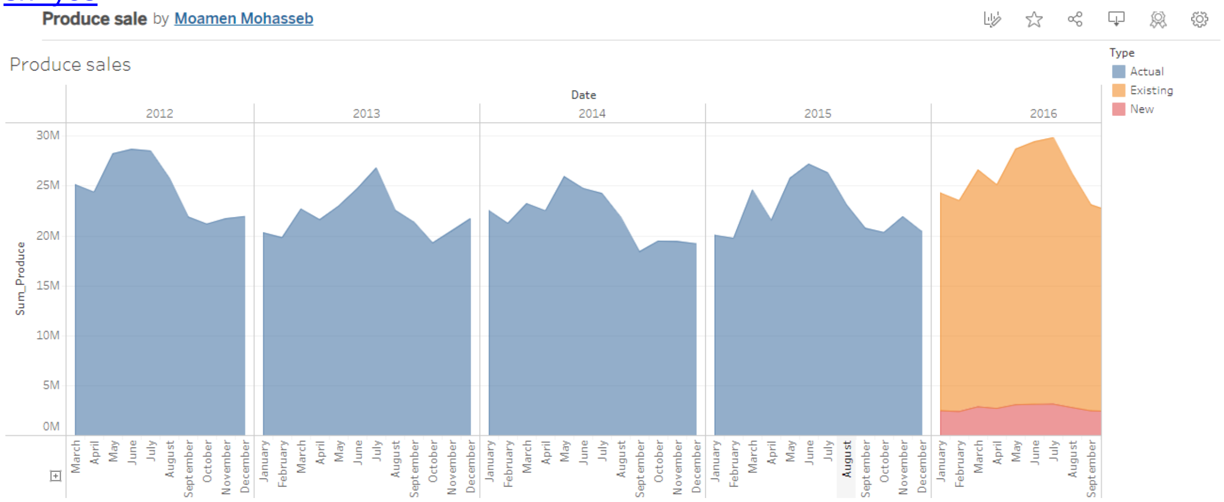



ARIMA
▶ Tour
Record
Text
1
Forecast from ARIMA(1,0,0)(1,1,0)[12]
— Actual — Fitted -- Lower -- Upper
3.00e+7
2.80e+7
2.60e+7
2.40e+7
2.20e+7
2.00e+7
1.80e+7
1.60e+7
Select an area on the plot to zoom in. Double click to zoom out
Autocorrelation Function Plot
Autocorrelation Function
0.2
0.0
-0.2
-0.4
0
2
4
6
8
10
12
14
16
18
20
22
24
26
Lag
Partial Autocorrelation Function
0.2
0.0
-0.2

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.

Year	Month	Forecast_Integer	New_Stores_Sales
2016	1	21829060	2493697
2016	2	21146330	2405584
2016	3	23735687	2879417
2016	4	22409515	2720393
2016	5	25621829	3089903
2016	6	26307858	3139497
2016	7	26705093	3155160
2016	8	23440761	2807733
2016	9	20640047	2482456
2016	10	20086270	2420097
2016	11	20858120	2510816
2016	12	21255190	2480120

https://public.tableau.com/app/profile/moamen.mohasseb/viz/Producesale/Sheet1?publish=yes



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.