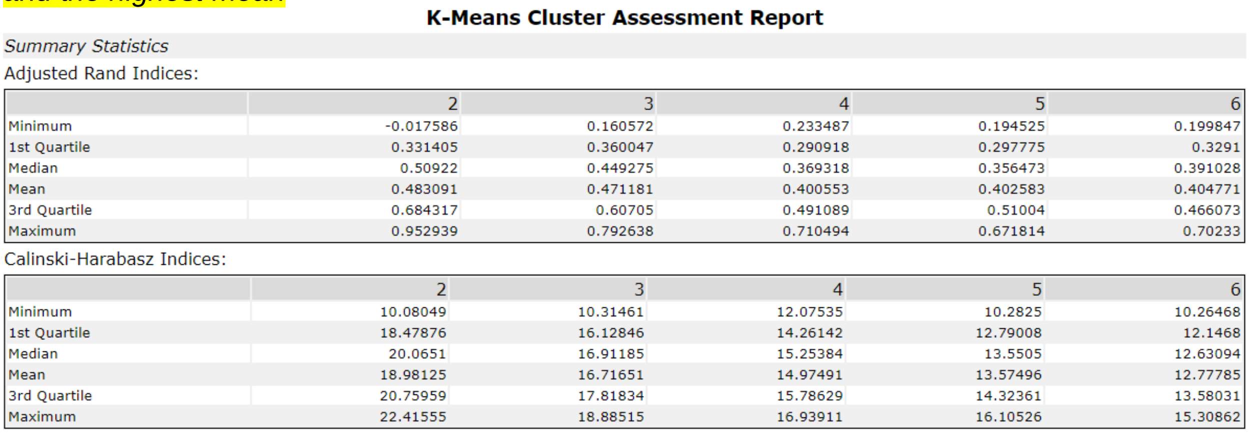
Project: Predictive Analytics Capstone

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?

To determine the best number of store formats that we are going to use, we carry out the analysis of k-centroids with the grouping of k-means up to k = 6. According to the diagnoses, k = 3 gave better results, I will use k = 3 since it has the most adjusted range and the highest mean



K-Means Cluster Assessment Report

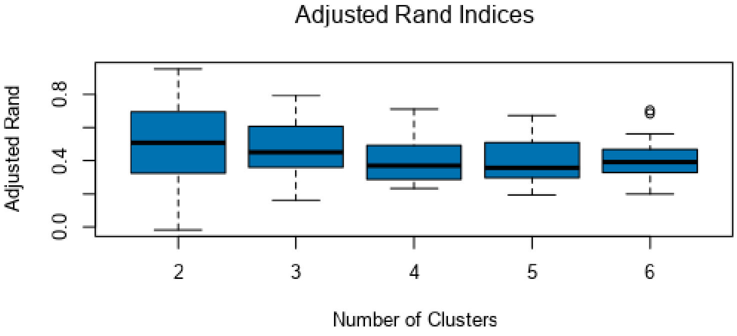
Summary Statistics

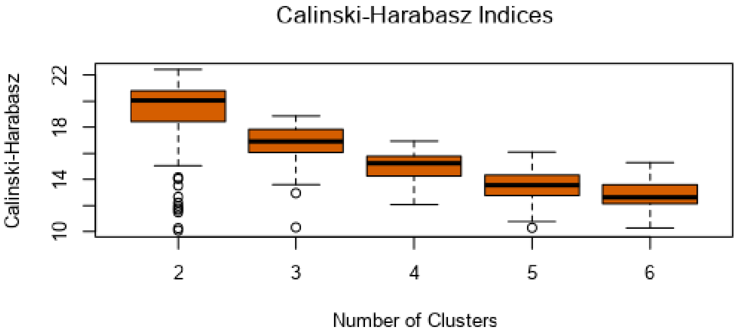
Adjusted Rand Indices:

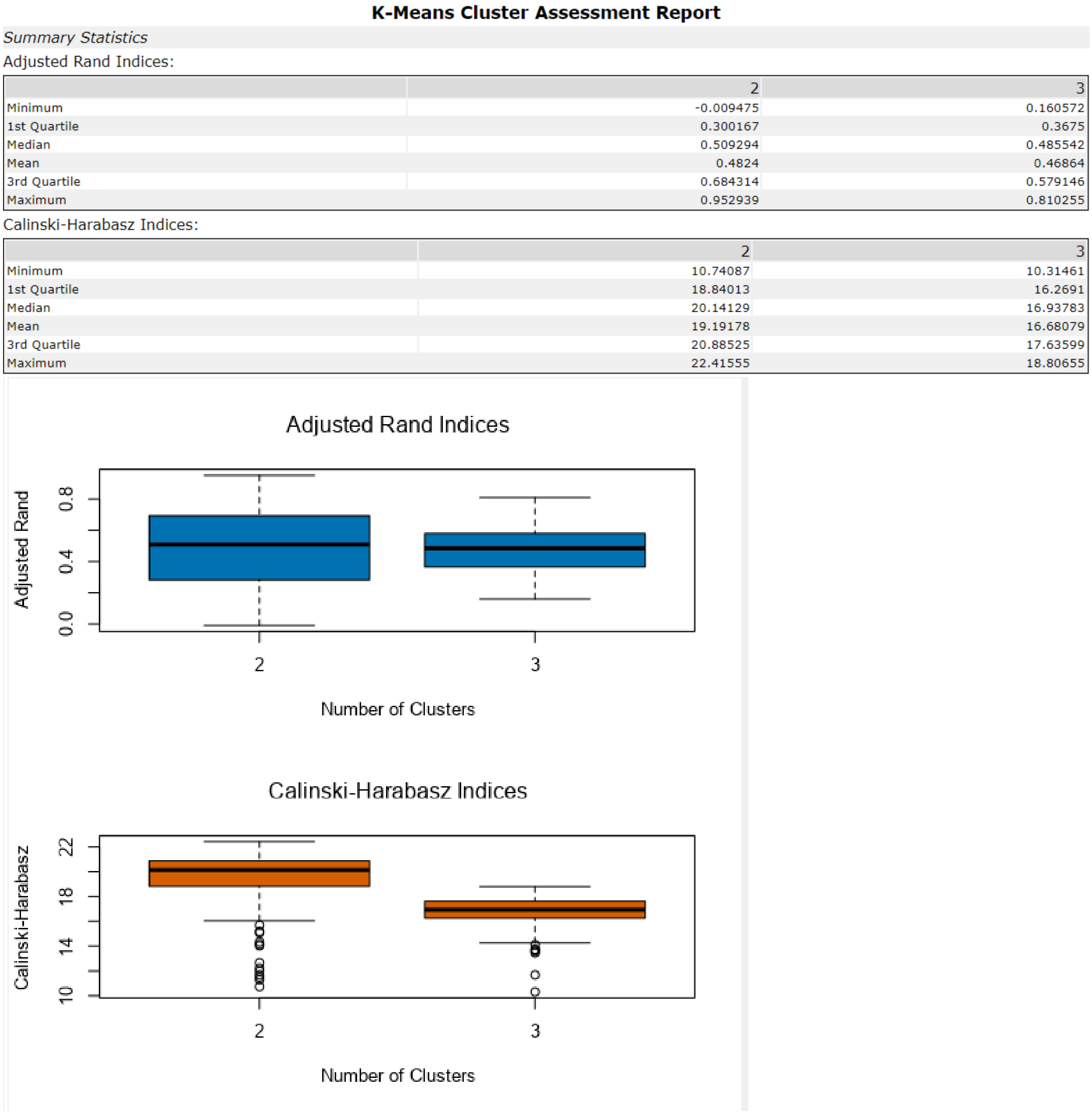
	2	3	4	5	6
Minimum	-0.017586	0.160572	0.233487	0.194525	0.199847
1st Quartile	0.331405	0.360047	0.290918	0.297775	0.3291
Median	0.50922	0.449275	0.369318	0.356473	0.391028
Mean	0.483091	0.471181	0.400553	0.402583	0.404771
3rd Quartile	0.684317	0.60705	0.491089	0.51004	0.466073
Maximum	0.952939	0.792638	0.710494	0.671814	0.70233

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	10.08049	10.31461	12.07535	10.2825	10.26468
1st Quartile	18.47876	16.12846	14.26142	12.79008	12.1468
Median	20.0651	16.91185	15.25384	13.5505	12.63094
Mean	18.98125	16.71651	14.97491	13.57496	12.77785
3rd Quartile	20.75959	17.81834	15.78629	14.32361	13.58031
Maximum	22.41555	18.88515	16.93911	16.10526	15.30862







K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

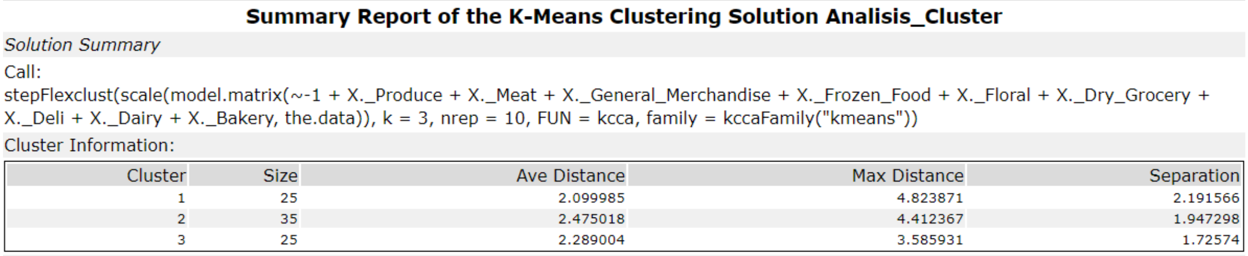
	2	3
Minimum	-0.009475	0.160572
1st Quartile	0.300167	0.3675
Median	0.509294	0.485542
Mean	0.4824	0.46864
3rd Quartile	0.684314	0.579146
Maximum	0.952939	0.810255

Calinski-Harabasz Indices:

	2	3
Minimum	10.74087	10.31461
1st Quartile	18.84013	16.2691
Median	20.14129	16.93783
Mean	19.19178	16.68079
3rd Quartile	20.88525	17.63599
Maximum	22.41555	18.80655

2. How many stores fall into each store format?

Cluster 1 25 stores
Cluster 2 35 stores
Cluster 3 25 stores



Summary Report of the K-Means Clustering Solution Analisis_Cluster

Solution Summary

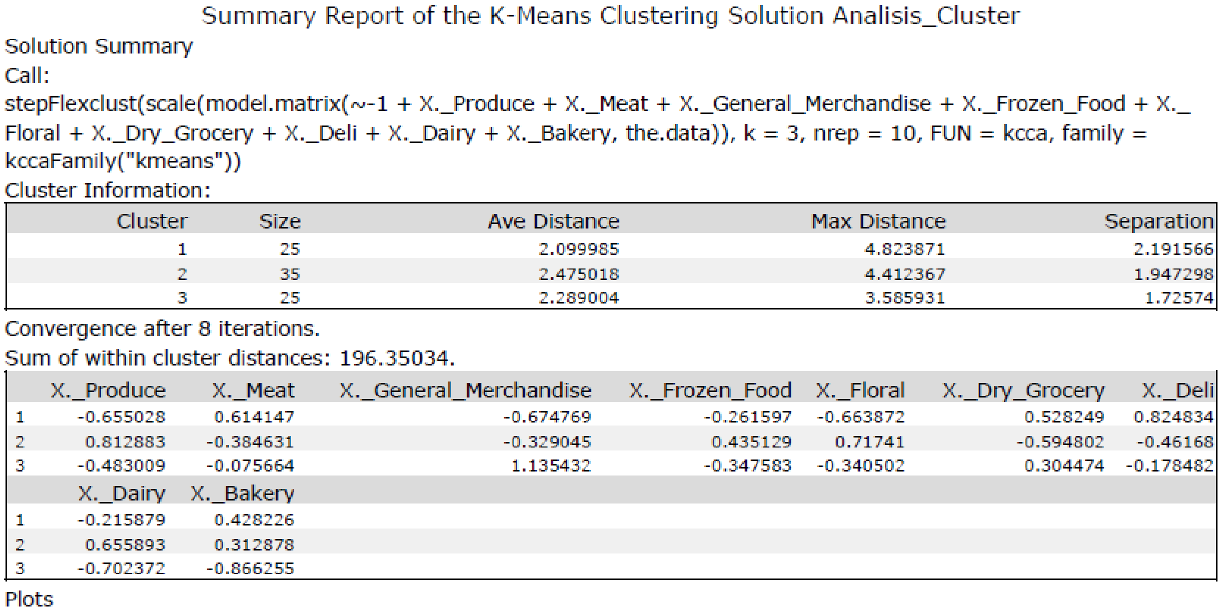
Call:
stepFlexclust(scale(model.matrix(~-1 + X._Produce + X._Meat + X._General_Merchandise + X._Frozen_Food + X._Floral + X._Dry_Grocery + X._Deli + X._Dairy + X._Bakery, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

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

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

We can see that cluster 3 is the most positive in percentage of sales of general merchandise against group 1, which is the most negative in percentages, we can deduce that these two clusters are the most different in percentage of sales of general merchandise, the group 2 has a great magnitude in edibles, fruits and vegetables, dairy and frozen food



Summary Report of the K-Means Clustering Solution Analisis_Cluster

Solution Summary

Call:

stepFlexclust(scale(model.matrix(~-1 + X._Produce + X._Meat + X._General_Merchandise + X._Frozen_Food + X._Floral + X._Dry_Grocery + X._Deli + X._Dairy + X._Bakery, 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.

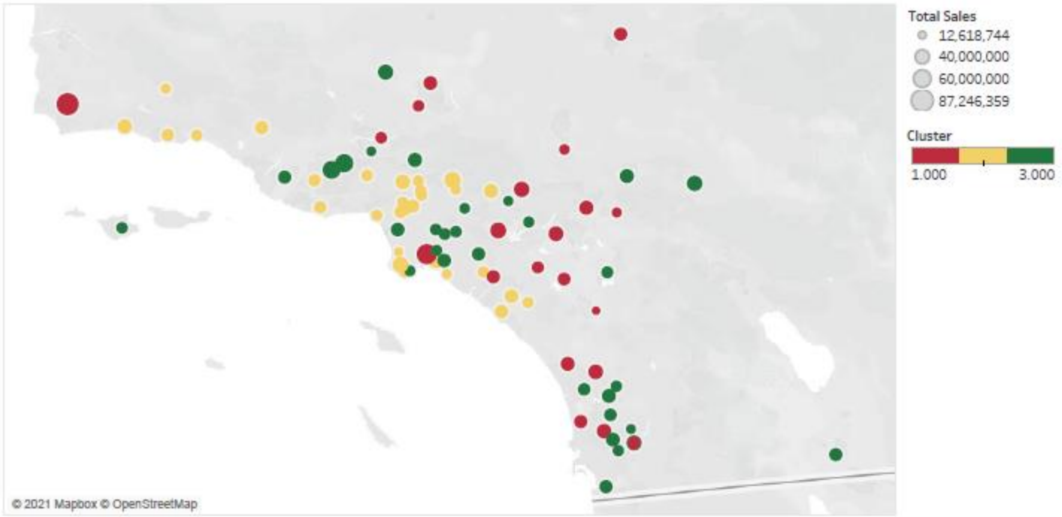
	X._Produce	X._Meat	X._General_Merchandise	X._Frozen_Food	X._Floral	X._Dry_Grocery	X._Deli
1	-0.655028	0.614147	-0.674769	-0.261597	-0.663872	0.528249	0.824834
2	0.812883	-0.384631	-0.329045	0.435129	0.71741	-0.594802	-0.46168
3	-0.483009	-0.075664	1.135432	-0.347583	-0.340502	0.304474	-0.178482

	X._Dairy	X._Bakery
1	-0.215879	0.428226
2	0.655893	0.312878
3	-0.702372	-0.866255

Plots

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.

Map_of_Store_Locations, Cluster_Group and Total_Sales



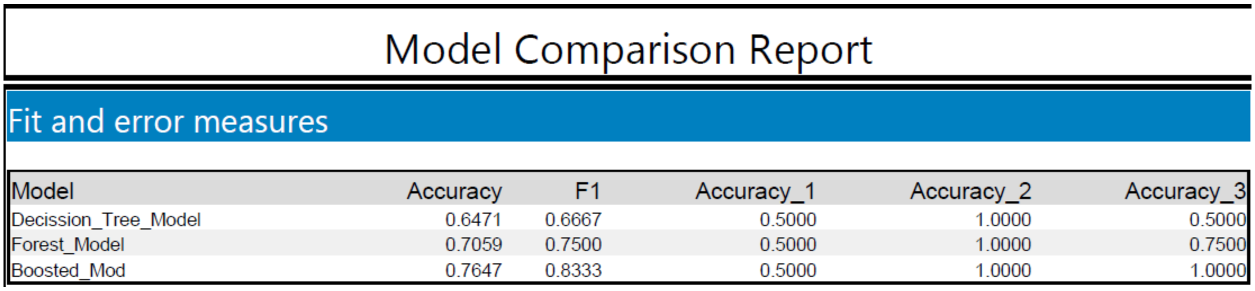
Map based on Longitude (generated) and Latitude (generated). Color shows details about Cluster. Size shows sum of Total Sales. Details are shown for Zip.

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.)

We use the StoreDemographicData.csv file, we run an improved model, decision trees and random forests, the respective accuracies are



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decission_Tree_Model	0.6471	0.6667	0.5000	1.0000	0.5000
Forest_Model	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted_Mod	0.7647	0.8333	0.5000	1.0000	1.0000

Observing the previous results, we will use the boosted model since it has the highest precision and the highest F1 score.

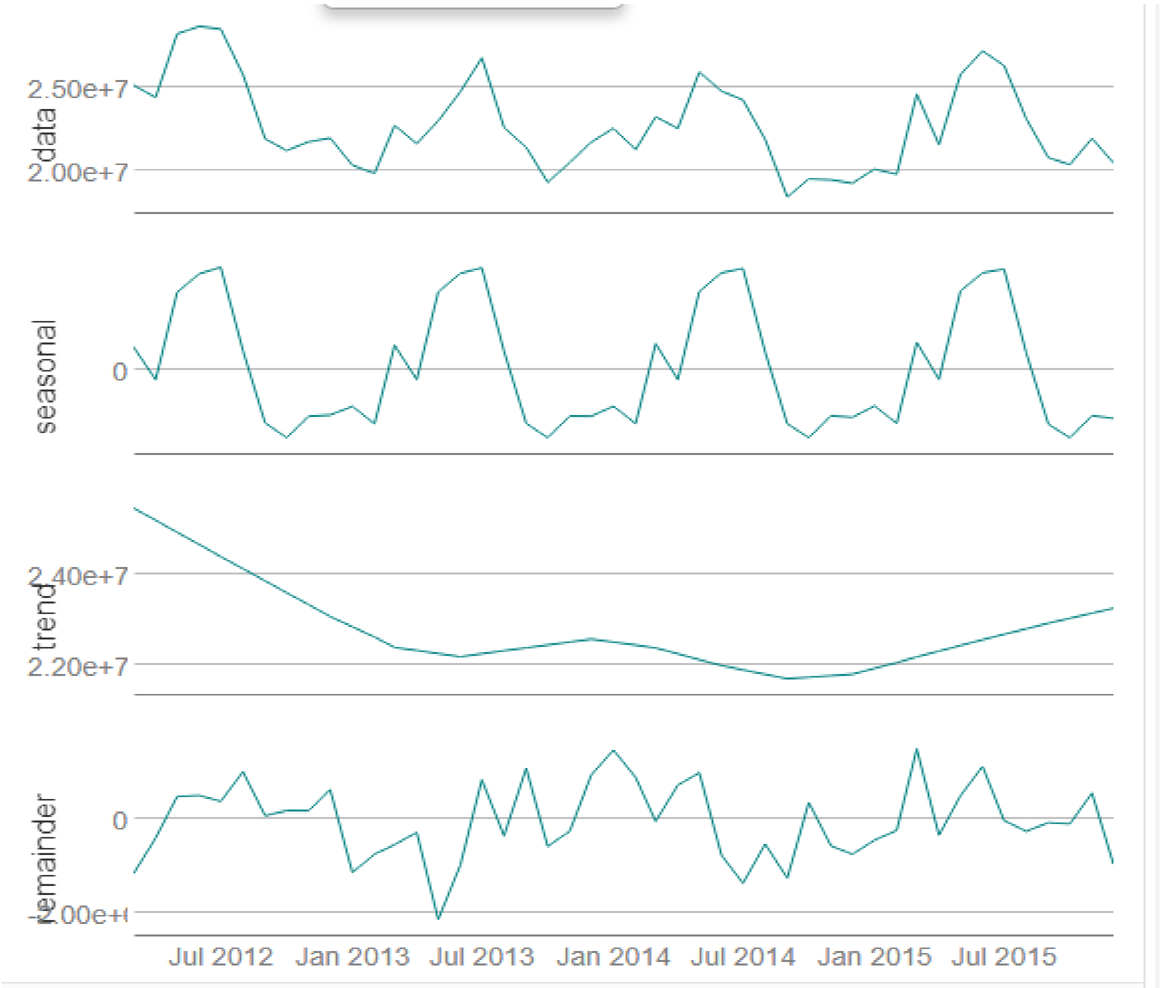
2. What format do each of the 10 new stores fall into? Please fill in the table below.

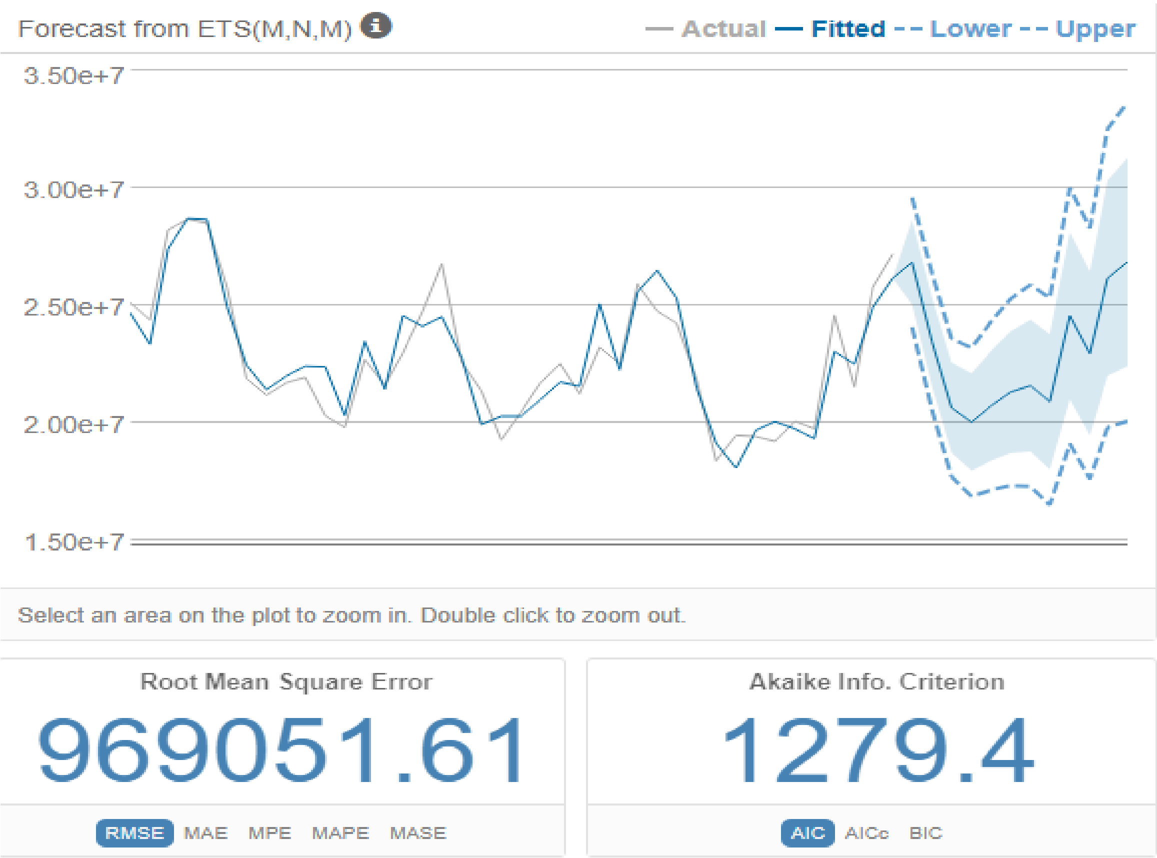
Store Number	Segment
S0086	3
S0087	2
S0088	3
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

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?

We train both ETS and ARIMA models to find the best prediction of total production from existing stores. The analysis of the initial times series decomposition plots gives us the opportunity to apply an analysis of the model's own parameters. We use the storesdata.csv, we create the graphs to understand the trend, seasonality and error, it can be seen that there is seasonality and the error tends to decrease over time. we have seasonality multiplicatively, the trend does not exist and the error multiplicatively.





Select an area on the plot to zoom in. Double click to zoom out.

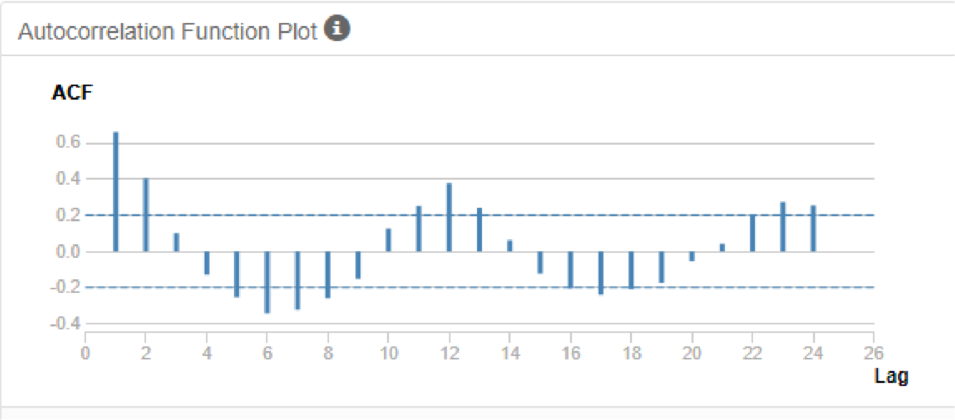
Root Mean Square Error

969051.61

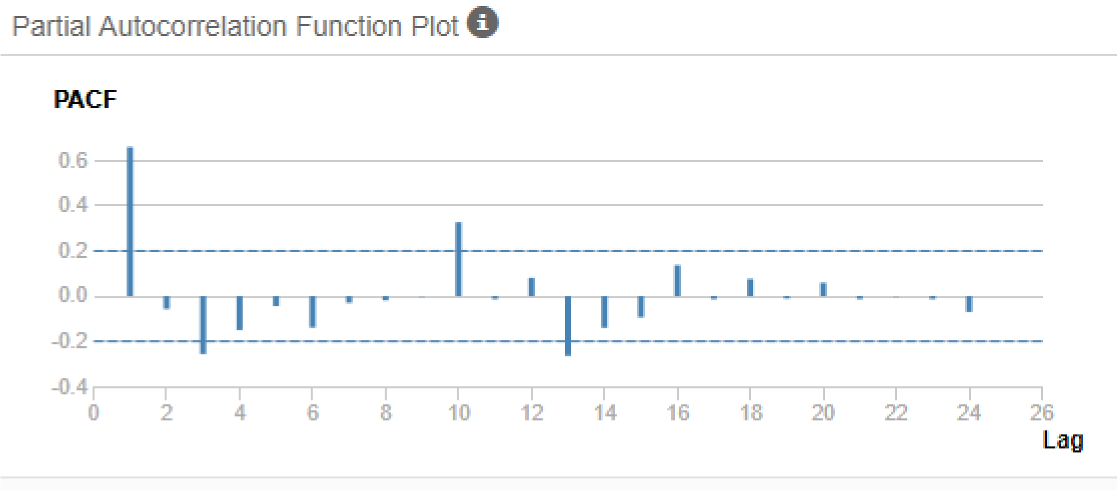
Akaike Info. Criterion

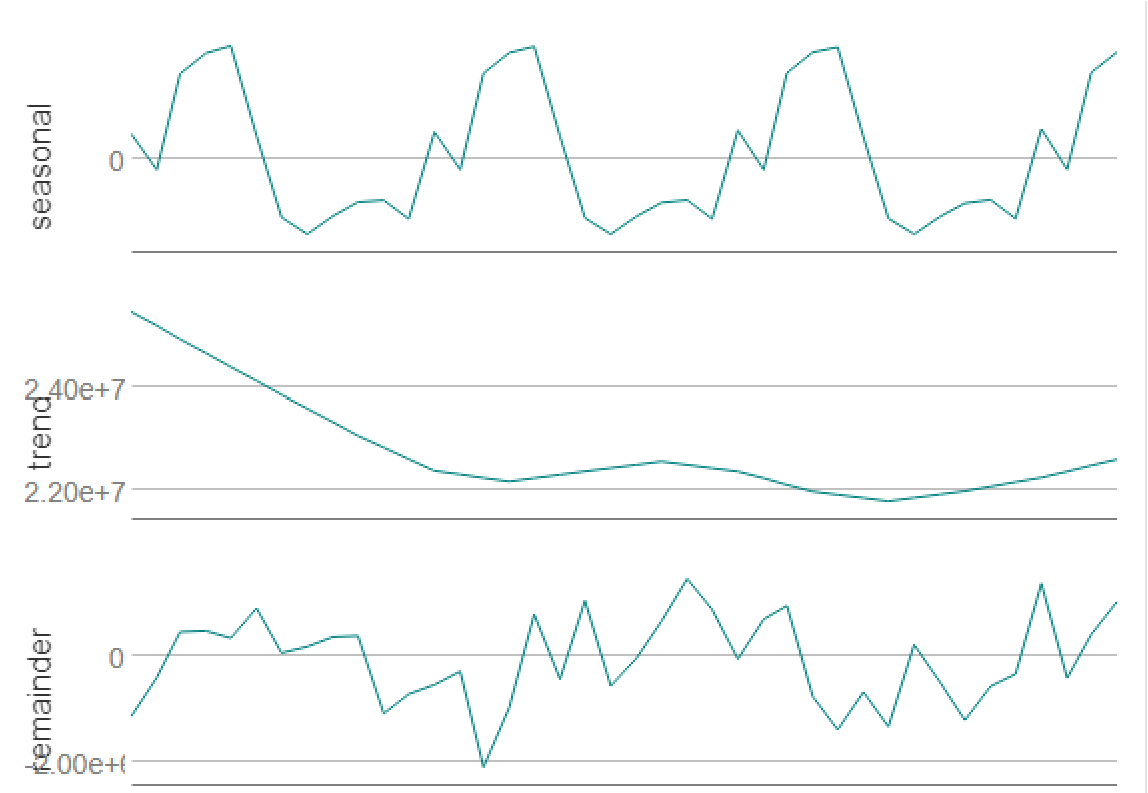
1279.4

We can see in the graphs that the correlation is seasonal and positive, that is why the term AR must be used

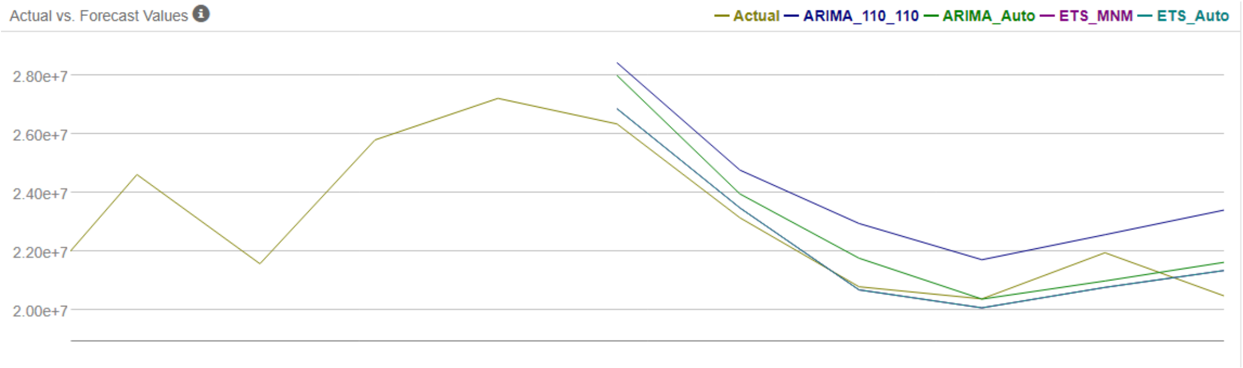


The following graph shows the fall of the correlation, thus suggesting the end AR (1)

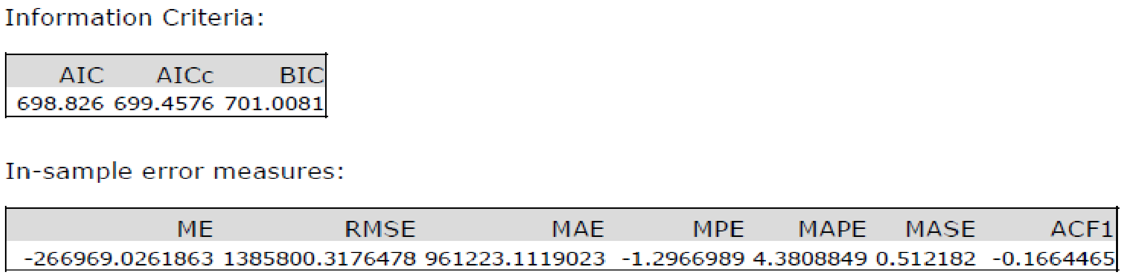




An increasing seasonality is also observed, use the term I (1) in the seasonal model as MA (0), the final model will be ARIMA (1,1,0) (1,1,0) 12 and comparing the 4 models , we observe the results in the following plot.



ARIMA(1,0,0)(0,1,0) 12 model summary



Information Criteria:

AIC	AICc	BIC
698.826	699.4576	701.0081

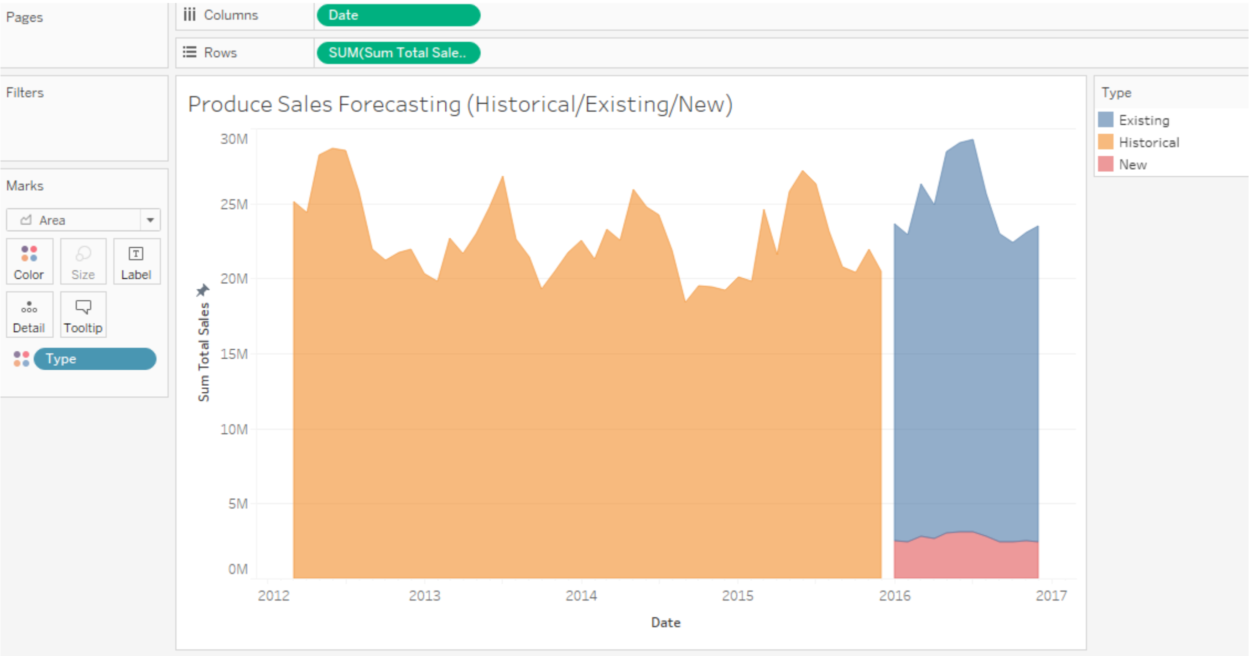
In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-266969.0261863	1385800.3176478	961223.1119023	-1.2966989	4.3808849	0.512182	-0.1664465

The highest AIC (1279.4) is from the ETS (M, N, M), also a better fit is appreciated, so I decide to use ETS (M, N, M)

3. 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.

Date	Period	Existing Store	New Stores
2,016	1	21,136,641.781775	2,491,319.093207
2,016	2	20,507,039.12384	2,408,384.783604
2,016	3	23,506,565.982355	2,833,157.321387
2,016	4	22,208,405.755153	2,679,433.371626
2,016	5	25,380,147.771963	3,054,885.876482
2,016	6	25,966,799.465113	3,106,151.779247
2,016	7	26,113,792.565116	3,132,699.144598
2,016	8	22,899,285.769116	2,776,154.195458
2,016	9	20,499,583.908226	2,451,565.941438
2,016	10	19,971,242.820704	2,401,771.574835
2,016	11	20,602,665.916965	2,477,301.916348
2,016	12	21,073,222.081854	2,452,170.069396



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.