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?
   On running the k-centroids diagnostic for the k-means model we get the report below -

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

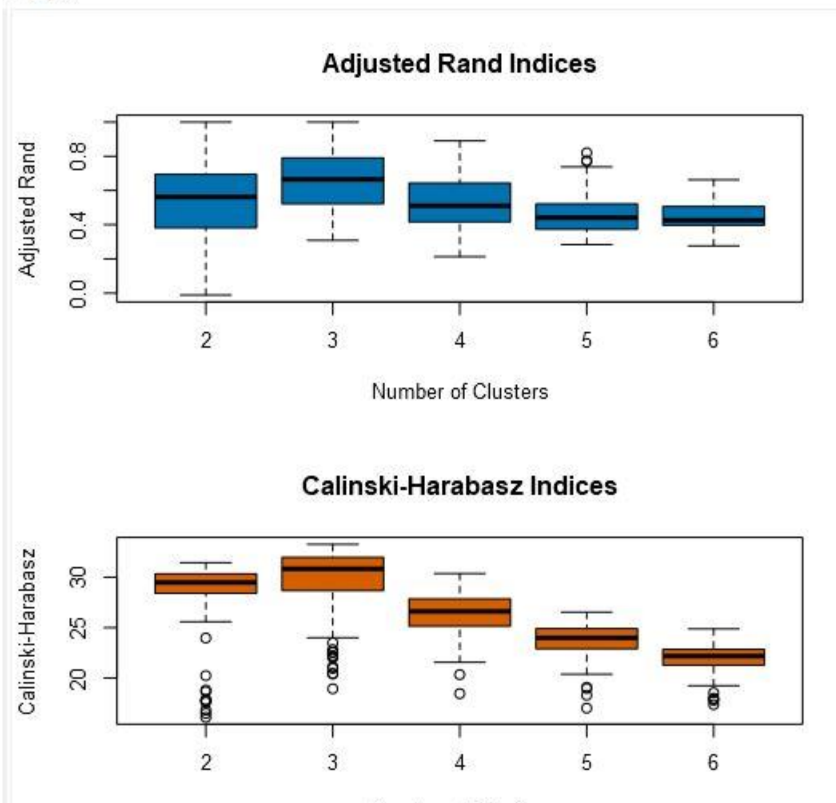
	2	3	4	5	6
Minimum	-0.01155	0.3083	0.213	0.2837	0.2762
1st Quartile	0.3814	0.5258	0.4169	0.374	0.3965
Median	0.5619	0.6653	0.5107	0.4406	0.4256
Mean	0.5084	0.6594	0.5471	0.4704	0.4502
3rd Quartile	0.6942	0.7865	0.6427	0.5199	0.5067
Maximum	1	1	0.8902	0.8207	0.6626

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

And the plots for the diagnostic are –

Plots



From the above two plots from the Adjust Rand indices and CH indices plot we find that the optimal number for store formats is 3. On comparing the box plots for the other cluster numbers , only the 3 cluster format has a high median value in both the indices.

2. How many stores fall into each store format?

Report

Summary Report of the K-Means Clustering Solution clustering_model

Solution Summary

Call:

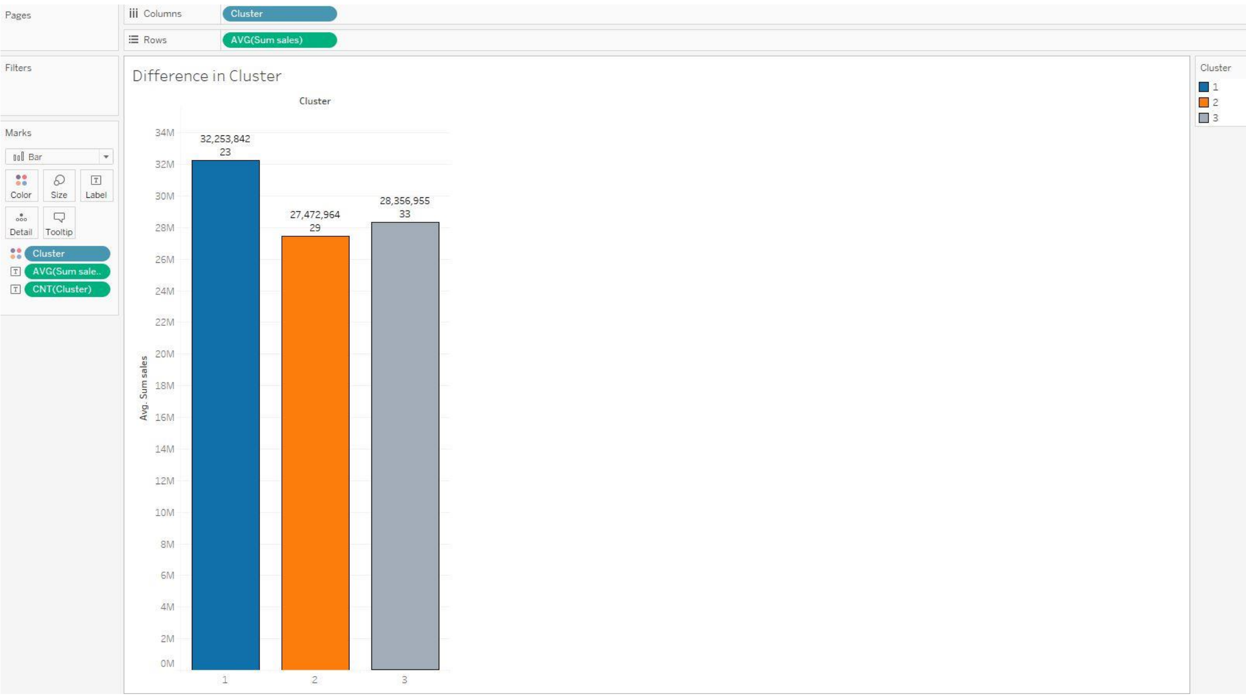
stepFlexclust(scale(model.matrix(~-1 + percent_dryGrocery + percent_dairy + percent_frozenfood + percent_meat + percent_produce + percent_floral + percent_deli + perecnt_bakery + percent_generalmerchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

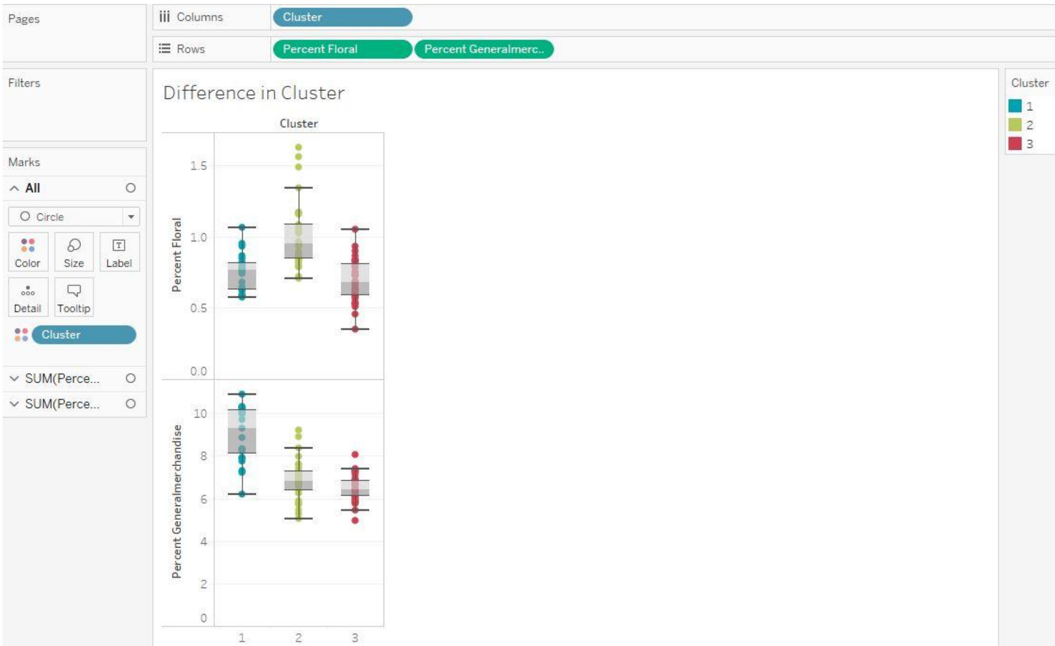
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

23 stores fall under cluster 1
29 stores fall under cluster 2
33 stores fall under cluster 3

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



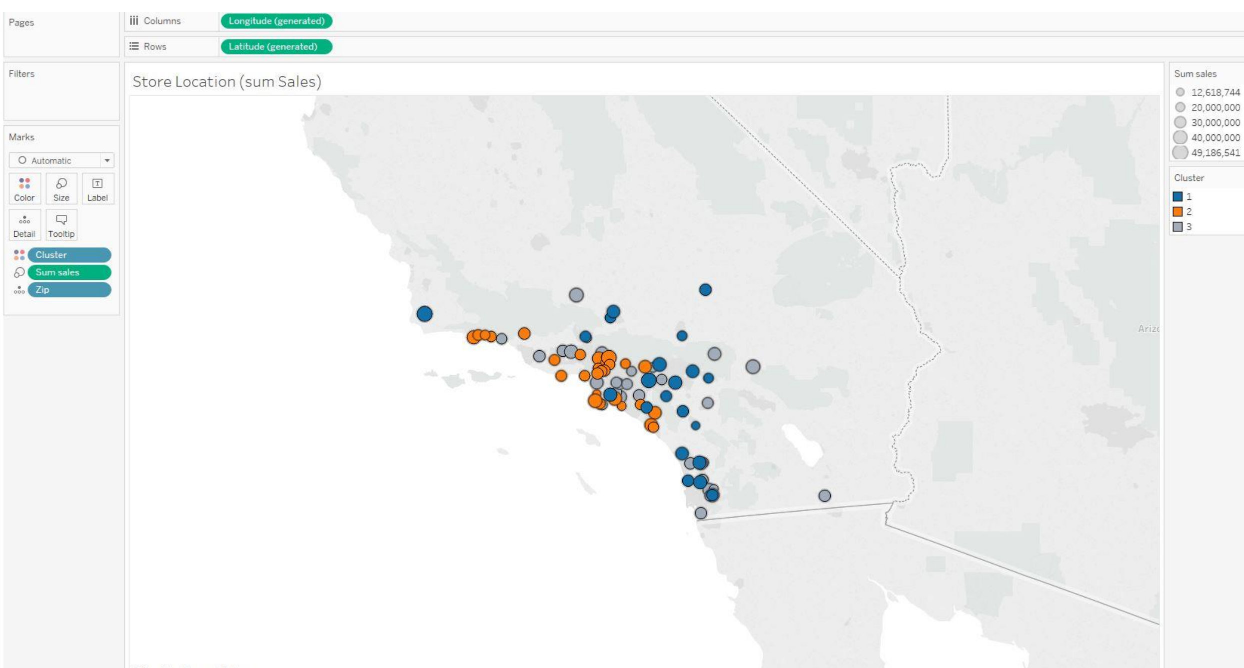
The one way that the clusters are different from each other is the average sum of sales. The average sum of sales is the highest for stores in cluster 1 and comparatively lowest in cluster 2.



The other ways of differentiating the clusters is box plots between different categories. The above box plots shows the clusters vs percent of floral sales for which the Cluster 2 is the highest, safe to assume that stores in cluster 2 are the ones that sell floral type items.

Respectively, we can see that stores that fall under cluster 1 have more general merchandise stock that they sell.

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.



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.)

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_model	0.8235	0.8251	0.7500	0.8000	0.8750
Decision_Tree_7	0.8235	0.8251	0.7500	0.8000	0.8750
boosted_model	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)

Confusion matrix of Decision_Tree_7

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

Confusion matrix of Forest_model

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

Confusion matrix of boosted_model

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

Even though the Accuracies for the 3 models are the same , we choose the boosted model as our model for classification as It has the highest F-value of 0.8543.

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	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

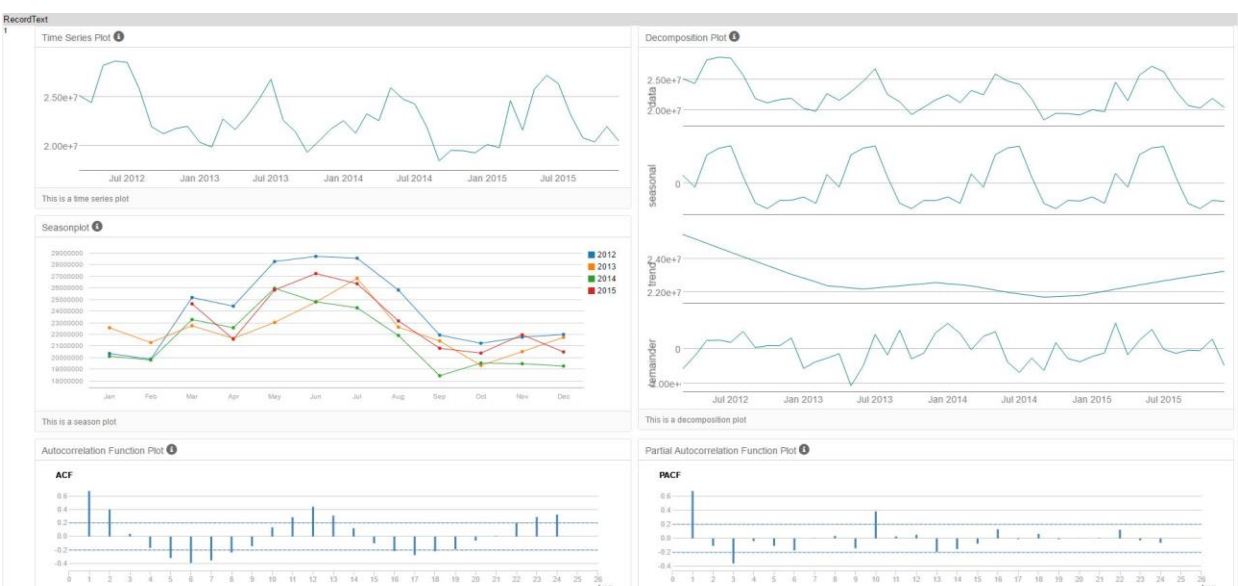
Task 3: Predicting Produce Sales

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?

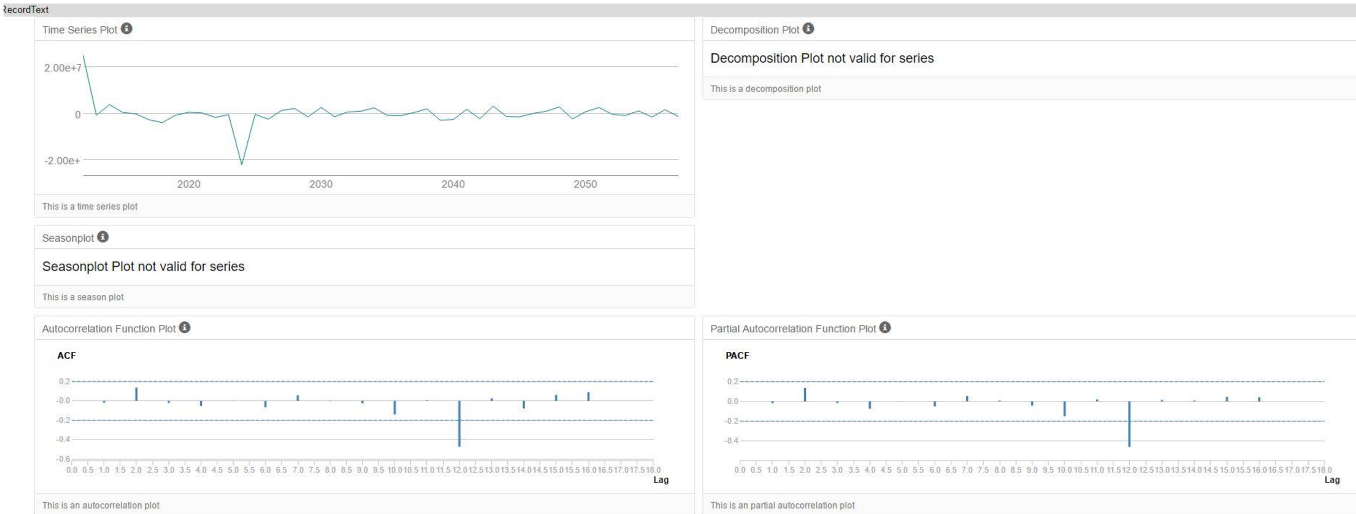
This Is the Time series plot for the ETS model and we can see that the seasonality and error/remainder can be applied multiplicatively since the graph for seasonality shows slight increase in its peaks with time and is not constant throughout the time frame we will apply seasonality multiplicatively.

Similarly for the error/remainder graph, we apply it as multiplicative as the plot is not constant.

And trend does not show any increase or decrease so we can apply nothing to it.



For our ARIMA model we use the seasonal first difference as the time series plot was not stationary.



As we can see there is a lag-2 and we can perform the ARIMA (0,1,2),(0,1,0)(12)

Method:
ETS(M,N,M)

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

Method: ARIMA(0,1,2)(0,1,0)[12]

Call:
Arima(Sum_Produce, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 0), period = 12))

Coefficients:

	ma1	ma2
Value	-0.415471	-0.054116
Std Err	0.219958	0.234438

sigma^2 estimated as 3268620653560.66: log likelihood = -426.38872

Information Criteria:

AIC	AICc	BIC
858.7774	859.8209	862.665

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.054315	1429296.2983494	951432.2560696	0.6151859	4.2022854	0.531117	-0.0260961

On using the TS compare tool from Alteryx for both the ETS and ARIMA model against the holdout sample we can see that the ETS model has a better forecast against the ARIMA as its values are closer than the ARIMA model's forecast.

Therefore, we will go ahead with the ETS model for our forecasting.

The table below shows the forecast for sum produce sales for the year 2016.

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
2016	2	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
2016	3	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981697
2016	4	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
2016	5	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
2016	6	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
2016	7	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
2016	8	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
2016	9	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
2016	10	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
2016	11	21177435.485839	23994279.191514	23019270.585553	19335600.386124	18360591.780163
2016	12	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

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.

We are looking for the average store per cluster, so wefirst sum the produce by cluster, store, year, month. And then take the average of that sum produce by cluster, year, month.

For the new stores we know that there are only –

3 stores in cluster 1

6 stores in cluster 2

1 store in cluster 3

We get the following table in the end –

Year	Period	New Store Sales
2016	1	2,626,198
2016	2	2,529,186
2016	3	2,940,264
2016	4	2,774,135
2016	5	3,165,320
2016	6	3,203,286
2016	7	3,244,464
2016	8	2,871,488
2016	9	2,552,418
2016	10	2,482,837
2016	11	2,597,780
2016	12	2,591,815

From our above forecasted sum of produce sales for the ETS model we have –

Period	Sub_Period	forecast
2016	1	21539936.007499
2016	2	20413770.60136
2016	3	24325953.097628
2016	4	22993466.348585
2016	5	26691951.419156
2016	6	26989964.010552
2016	7	26948630.764764
2016	8	24091579.349106
2016	9	20523492.408643
2016	10	20011748.6686
2016	11	21177435.485839
2016	12	20855799.10961

So adding these two tables we get –

Year	Period	New Store Sales	Existing Store Sales
2016	1	2,626,198	21,539,936
2016	2	2,529,186	20,413,771
2016	3	2,940,264	24,325,953
2016	4	2,774,135	22,993,466
2016	5	3,165,320	26,691,951
2016	6	3,203,286	26,989,964
2016	7	3,244,464	26,948,631
2016	8	2,871,488	24,091,579
2016	9	2,552,418	20,523,492
2016	10	2,482,837	20,011,749
2016	11	2,597,780	21,177,435
2016	12	2,591,815	20,855,799

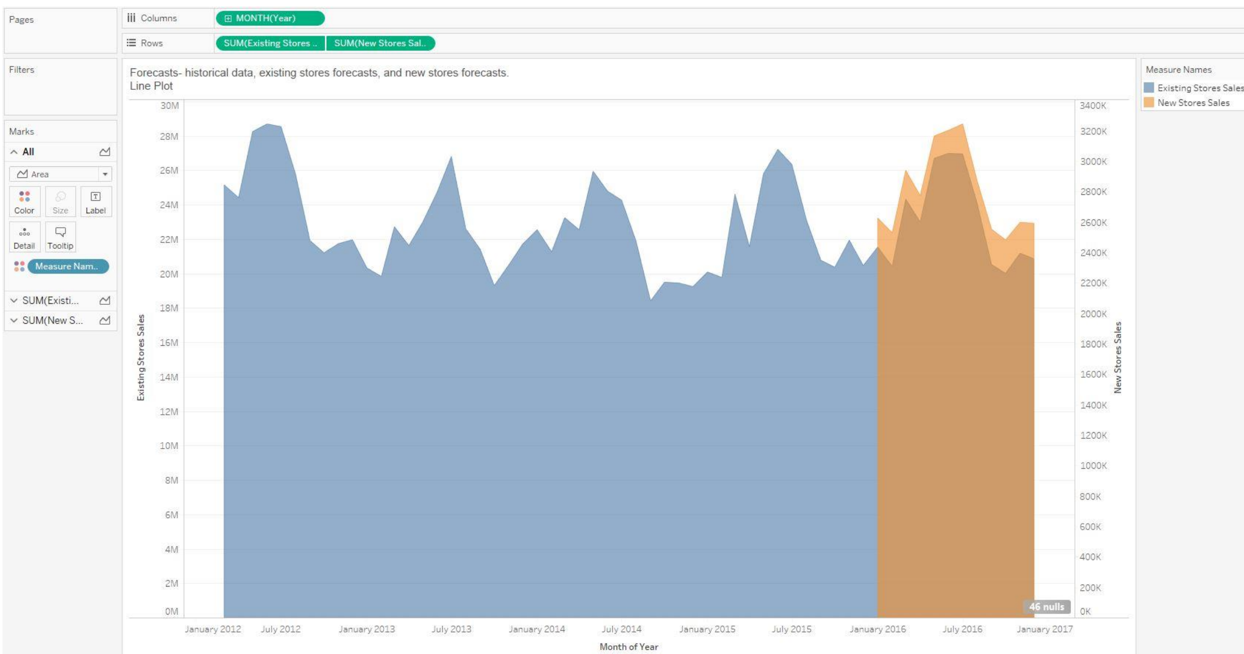


Tableau Public profile - https://public.tableau.com/profile/mohammadnadeem#!/