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 format is 3. This result was founded by using K-means centroid diagnostics tool. As it is shown below in the diagrams, number 3 has the highest adjusted rand index value. In CH index indicates that number 3 has the highest median and fairly compact spread.

Summary Statistics

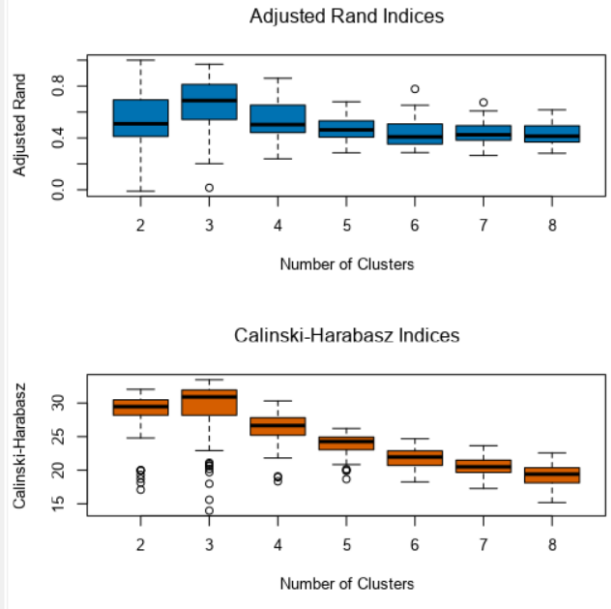
Adjusted Rand Indices:

	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

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	17.09341	13.97288	18.34333	18.6761	18.26096	17.27695	15.18428
1st Quartile	28.21314	28.18867	25.22107	23.05743	20.72628	19.67809	18.14144
Median	29.47306	30.92094	26.64745	24.26385	21.95763	20.49391	19.41545
Mean	28.69301	29.28225	26.3012	23.87402	21.72469	20.52895	19.28853
3rd Quartile	30.48102	31.95364	27.82114	24.94462	22.87944	21.41419	20.35482
Maximum	32.04793	33.47176	30.32206	26.22178	24.68592	23.65996	22.59031

Plots



2. How many stores fall into each store format?

As it is shown below, the following numbers fall into each cluster respectively 25,35 and 25.

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?
   To understand the difference between clusters, average distance, max distance and separation are the best indicators.

   As it shown below, cluster 1 has the least average distance. This result proves that cluster 1 is the most compact/precise cluster. In the same time, the highest value of max distance from the centroid appears in cluster 1. This indicates that there might be outliers. Finally, separation value tells us that cluster 1 has the most distance from the closest cluster.

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

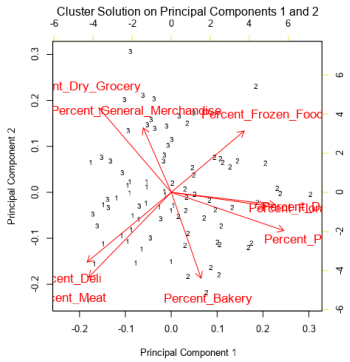
   Based on the results, the high and low values for each cluster are listed below. The highest value for each cluster is:

   - Cluster 1 – Deli
   - Cluster 2 – Production
   - Cluster 3 – General Merchandise

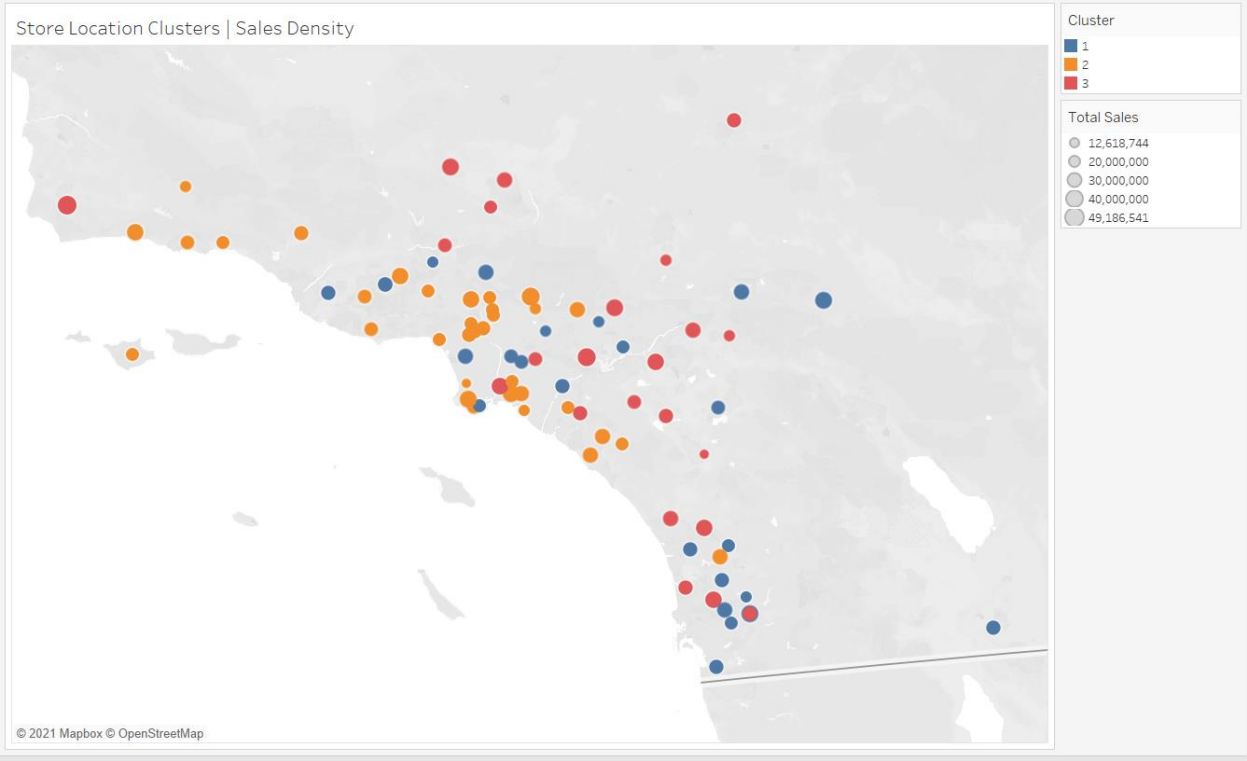
	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482

	Percent_Bakery	Percent_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

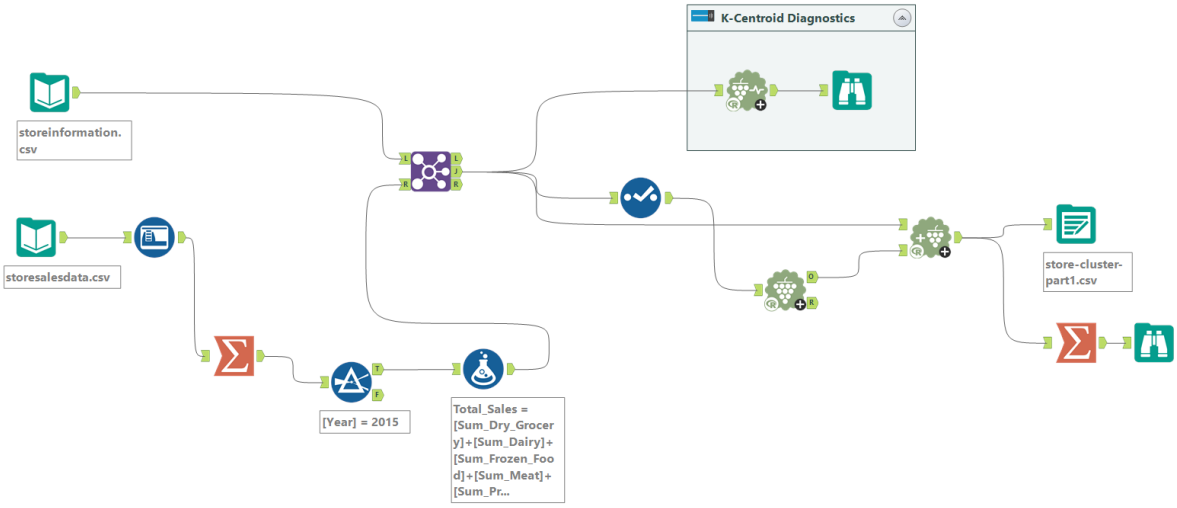
Plots



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.



https://public.tableau.com/views/task1_16320943627000/Sheet2?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link



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 this assignment, the prediction should be carried with cluster identifier. Thus, cluster field is selected as target variable and all the demographic data were used as predictor values. Since this is not a binary problem, non-binary predictive models were tested and compared. Based on accuracy, overall accuracy and F1 Score, the best prediction was calculated with boosted model. The results are listed below.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
RF_Model	0.6471	0.7083	0.3750	1.0000	0.7500
B_Model	0.7647	0.8333	0.5000	1.0000	1.0000
DT_Model	0.7059	0.7083	0.6250	1.0000	0.5000

Confusion matrix of B_Model

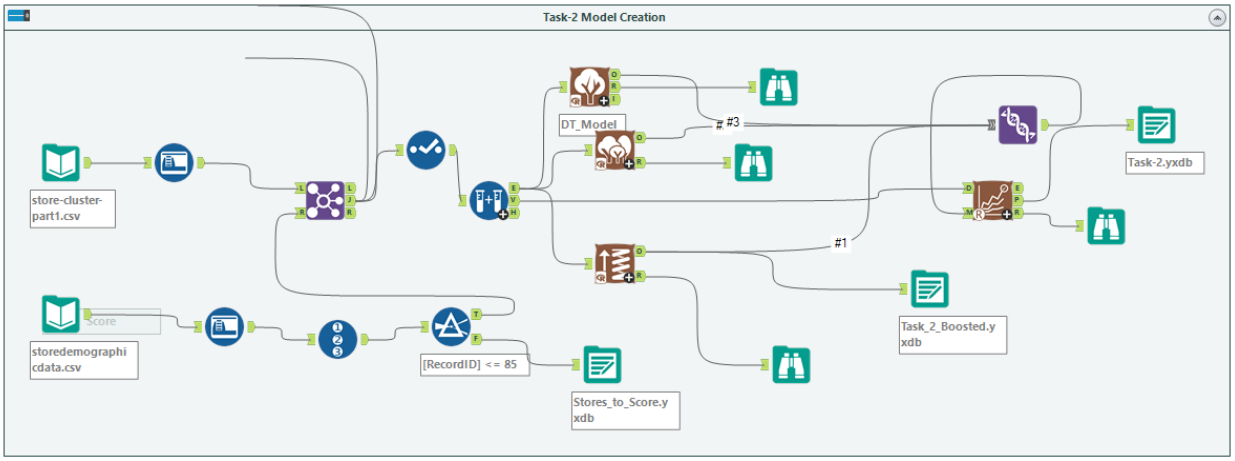
	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	0
Predicted_2	2	5	0
Predicted_3	2	0	4

Confusion matrix of DT_Model

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

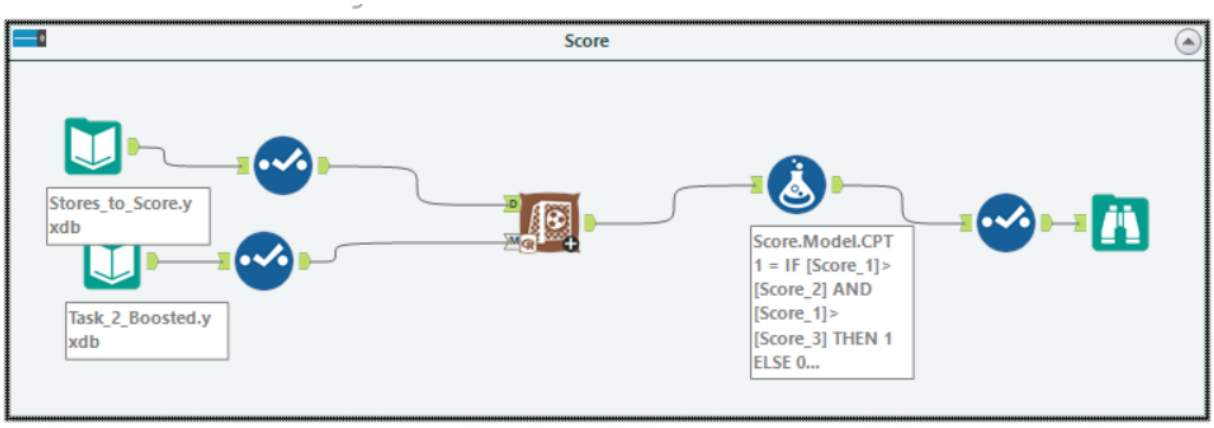
Confusion matrix of RF_Model

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



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



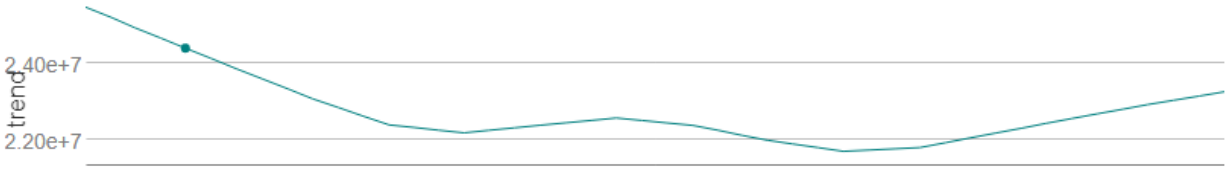
Task 3: Predicting Produce Sales

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?

Based on the comparison and results that listed below, ETS model was selected for this problem.
Firstly, I checked the time series decomposition plot. To be able establish this, I grouped the data based on “year”, “month” and calculated total and average of *produce* data. Then TS plot tool was used to identify the trends, seasonal and error components within a report. The target field was selected as “sum_produce” and data was plot in monthly format. As it shown in the plots below, the error and the seasonal appear **multiplicative** and the trend is **nonexisting**.

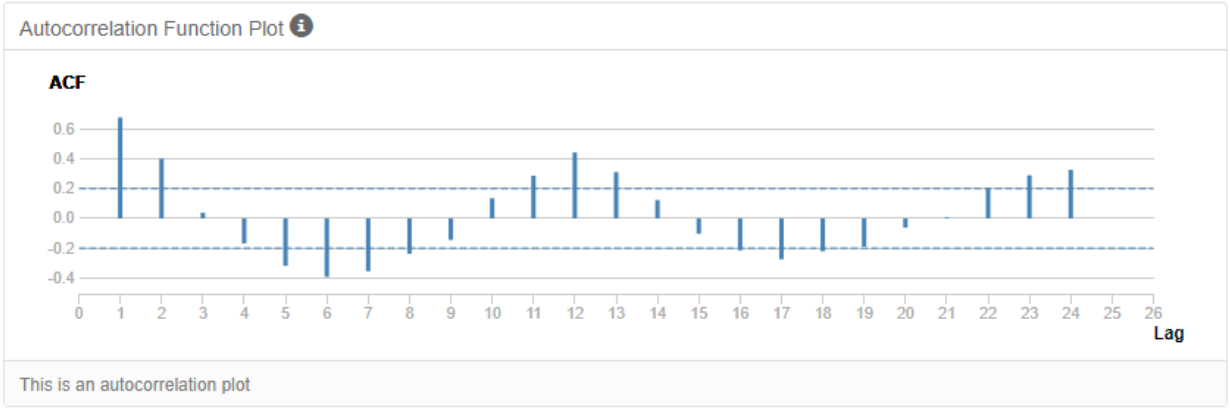
There is no regular pattern in error component.

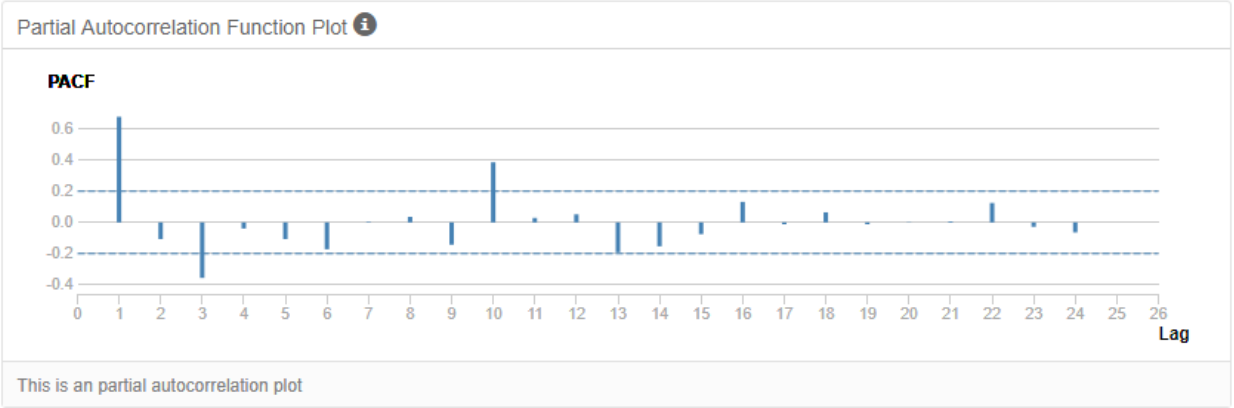
The trend doesn’t really show as a consistent pattern. First it goes down and continues as upward.

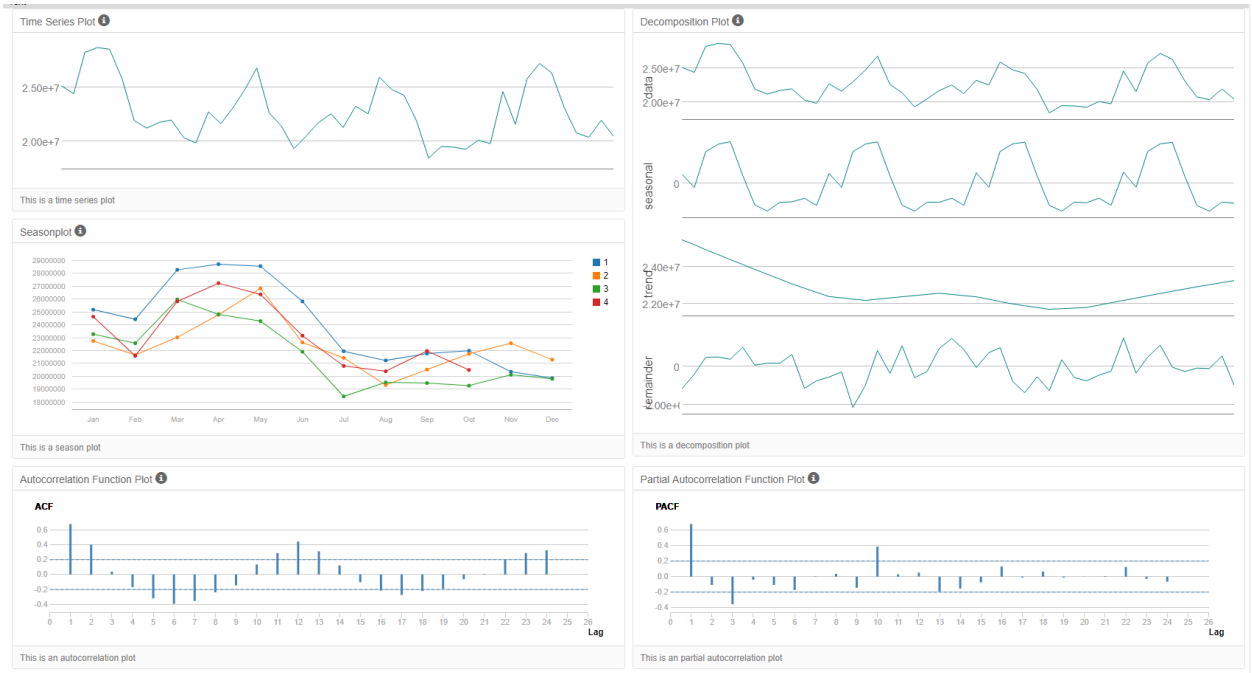


As it is shown in the figure below, there is very correlated seasonality. There is a repetition in the pattern. The sales gradually increase every year in July to April. Once it reaches to the highest point in April, it starts to decrease until July.

The next step is ACF and PACF graphs. There is decline in the overall view of ACF graph. It gradually declines and follows the same pattern. Thus, I identified that a seasonal difference should be applied in the time series to be able to get a stationary dataset.

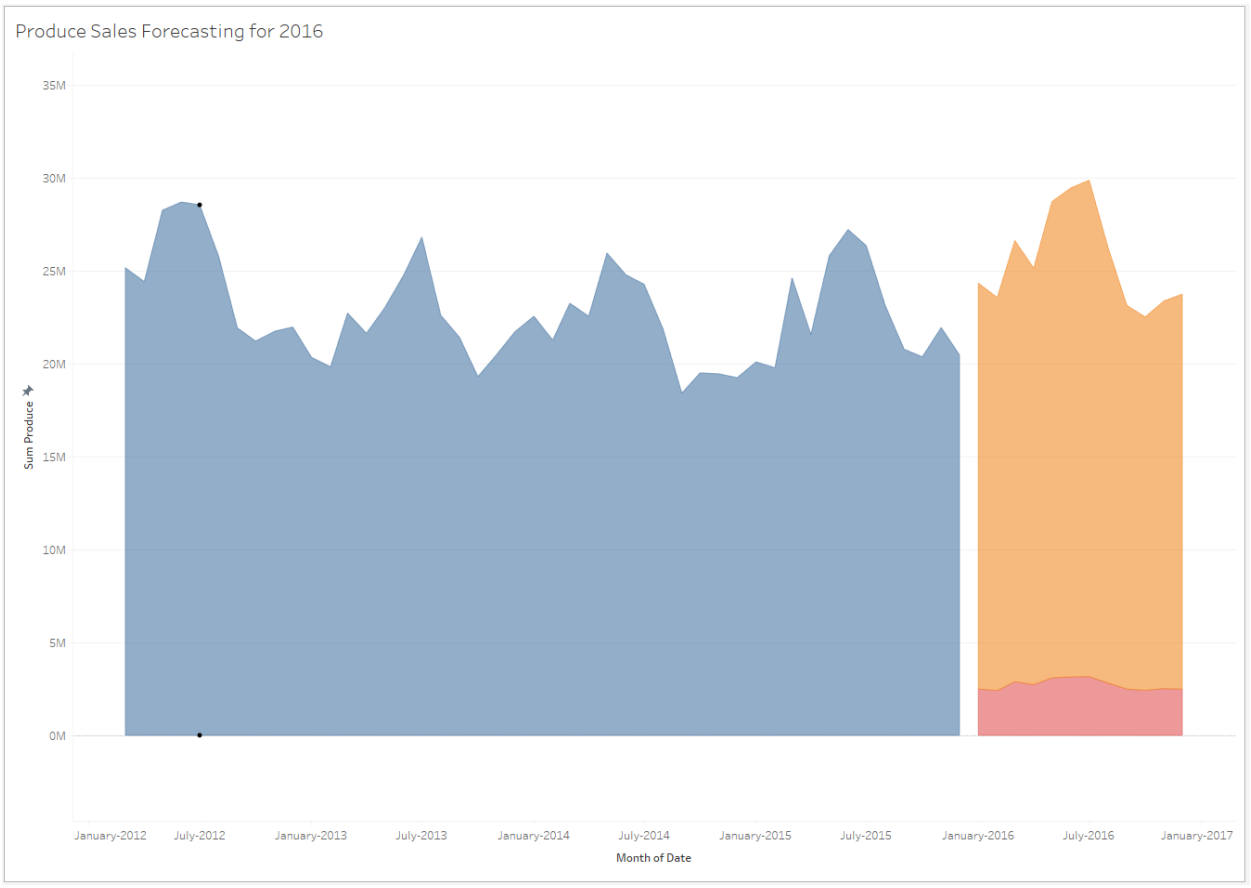




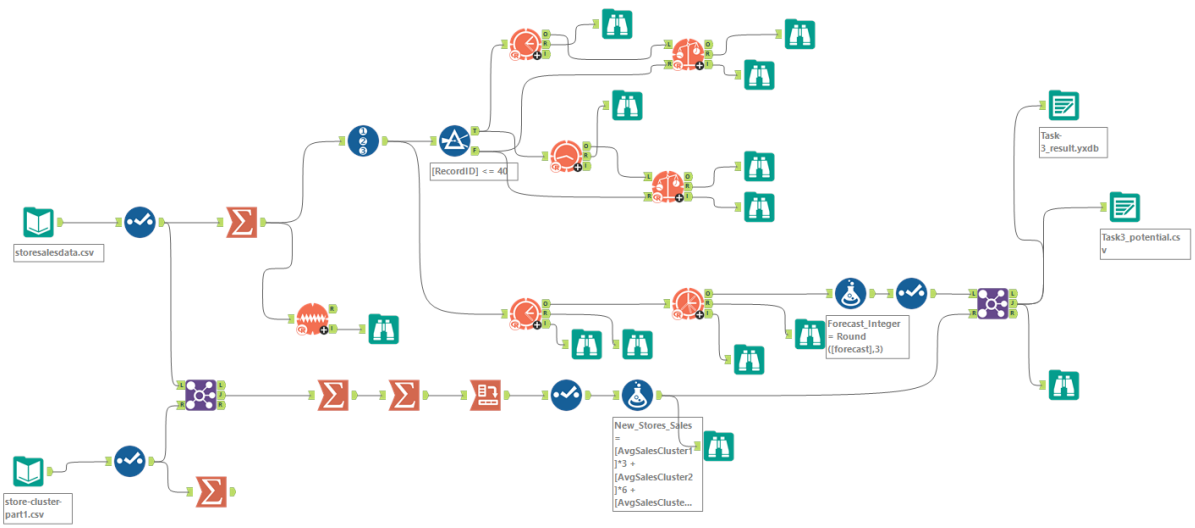


2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

2016	Forecast_Integer	New_Stores_Sales
January	21829059	2493697
February	21146331	2405584
March	23735688	2879417
April	22409514	2720393
May	25621830	3089903
June	26307858	3139497
July	26705094	3155160
August	23440761	2807733
September	20640048	2482456
October	20086269	2420097
November	20858121	2510816
December	21255189	2480120



https://public.tableau.com/app/profile/caner5694/viz/Produce_Sales_Forecasting_2016/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.