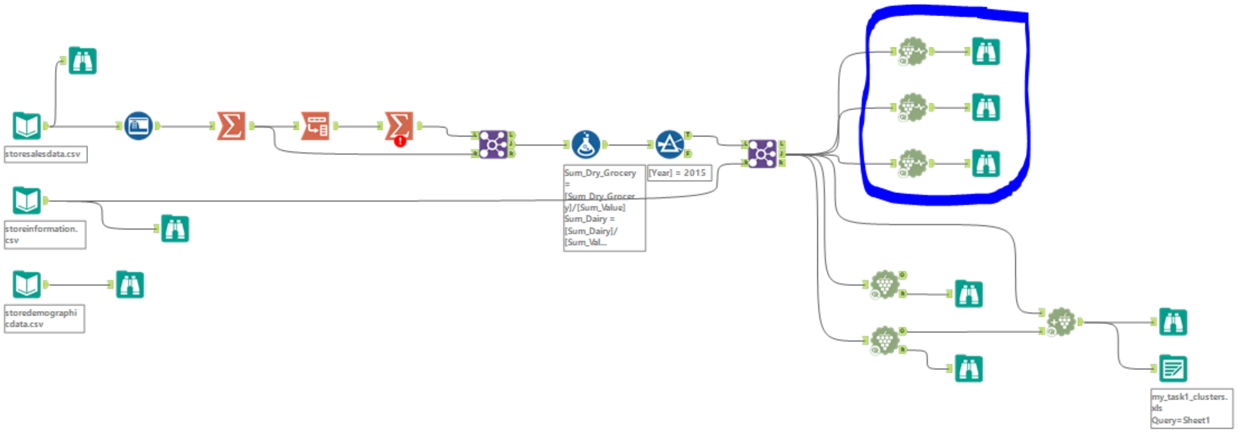
Project: Predictive Analytics Capstone

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?



storesalesdata.csv
storeinformation.csv
storedemographicdata.csv
Sum_Dry_Grocery = [Sum_Dry_Grocery]/[Sum_Value] Sum_Dairy = [Sum_Dairy]/ [Sum_Val...
[Year] = 2015
my_task1_clusters.xls Query=Sheet1

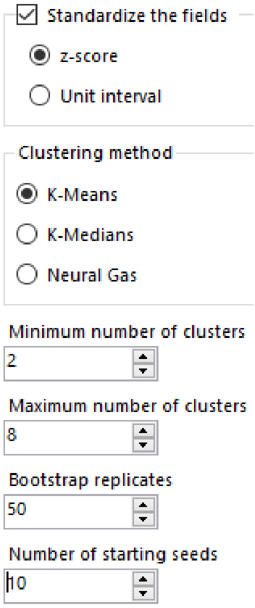
I summed up every category for each store per year

Every category was summed up also to determine the total sales for each store per year. To help us in the clustering algorithm, I calculated the percentage of each category out of the total sales and filtered for 2015.

I ran 3 different tests with the K-Centroids Diagnostics (example on the right) to compare the 3 clustering methods: K-Means, K-Medians and Neural Gas.

Even though it may seem that 2 clusters score higher than 3 clusters when evaluating the indices, I chose 3 as it would be much more efficient. We would combine the standardization and localization techniques to better customize the product according to the customer's needs and increase our sales.

Choosing 3 clusters also includes outliers from the '2' clusters variant. Additionally, the median from 3 and 2 clusters from Adjusted Rand Indices share almost the same value.



☑ Standardize the fields

● z-score

○ Unit interval

Clustering method

● K-Means

○ K-Medians

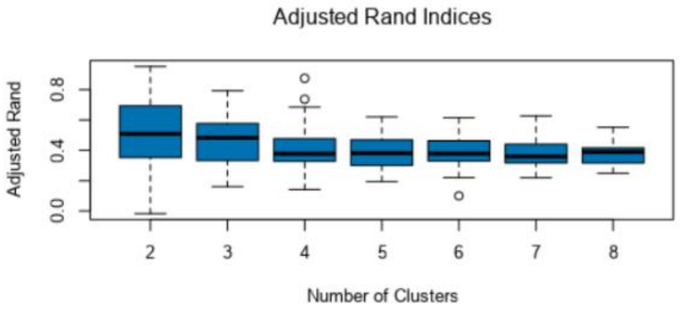
○ Neural Gas

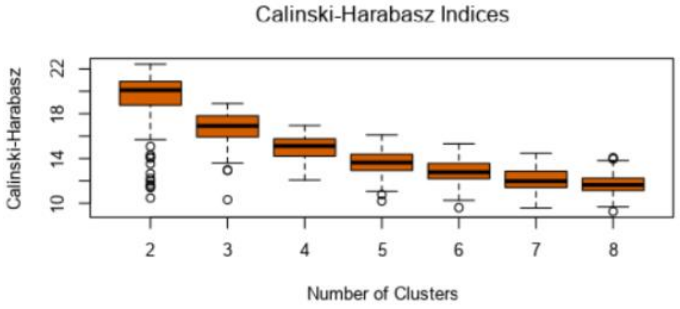
Minimum number of clusters: 2

Maximum number of clusters: 8

Bootstrap replicates: 50

Number of starting seeds: 10

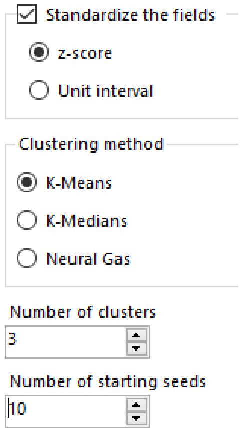




2. How many stores fall into each store format?

Selecting K-Means clustering method on the K-Centroids Cluster Analysis tool, the stores were distributed as follows:

Cluster 1 – 25
Cluster 2 – 35
Cluster 3 – 25



☑ Standardize the fields

● z-score

○ Unit interval

Clustering method

● K-Means

○ K-Medians

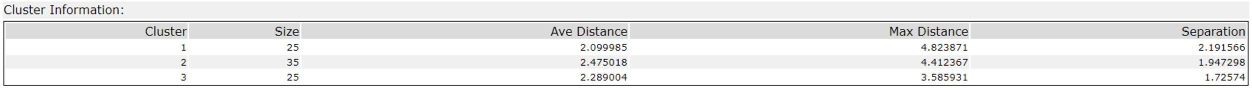
○ Neural Gas

Number of clusters

3

Number of starting seeds

10

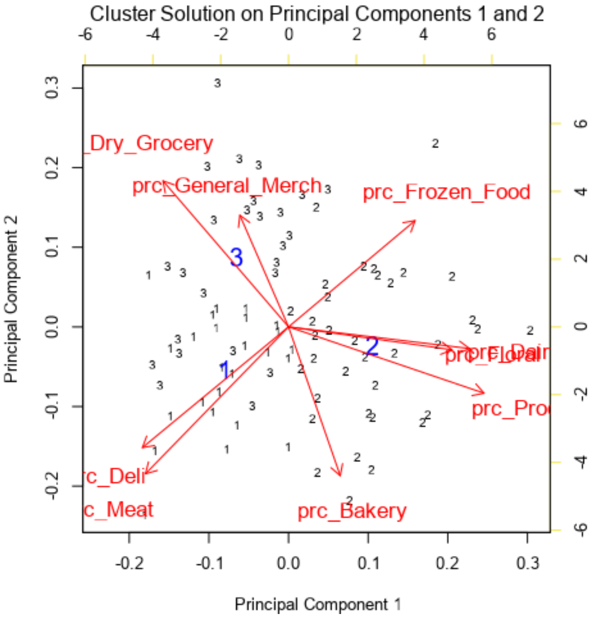


Cluster Information:

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?

The clusters are separated and differ based on the array of products they offer and the percentage of sales for each category.

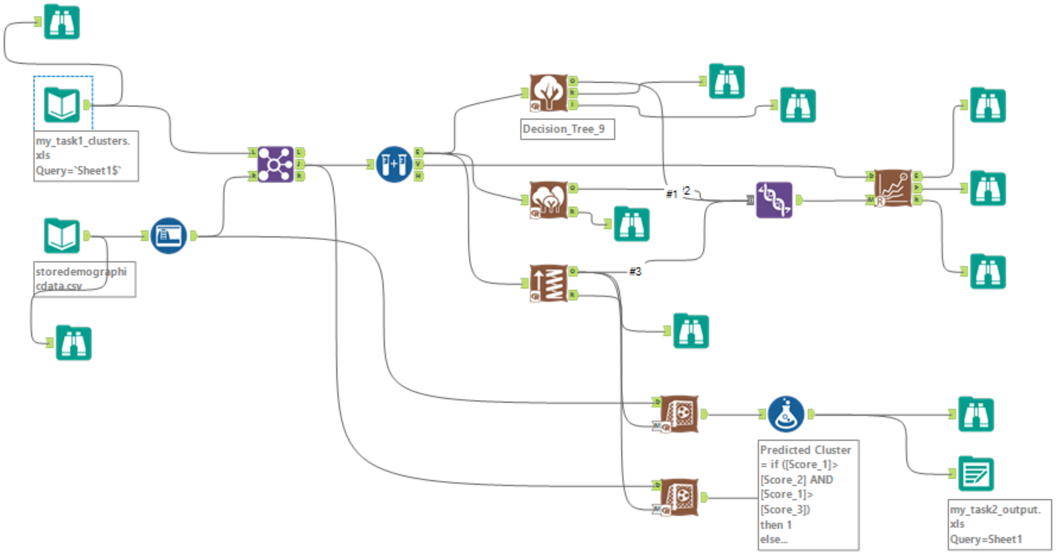


4. **Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.**

https://public.tableau.com/profile/felix.lupu#!/vizhome/task1_16117815879730/Sheet1?publish=yes

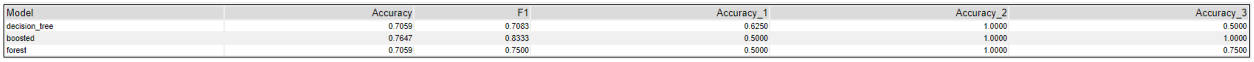
Task 2: Formats for New Stores

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.)**

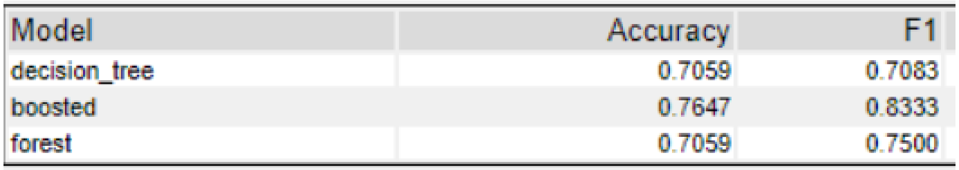


Using the data from Task 1 and the demographic data for the new stores, I ran the 3 different methods (Decision Tree, Forest Model, Boosted Model) to determine which one is the most accurate in choosing the appropriate cluster for each new store.

I used a 20% validation sample with Random Seed = 3 to compare the 3 models and the results can be seen in the pictures below:



Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
decision_tree	0.7059	0.7083	0.6250	1.0000	0.5000
boosted	0.7647	0.8333	0.5000	1.0000	1.0000
forest	0.7059	0.7500	0.5000	1.0000	0.7500



Model	Accuracy	F1
decision_tree	0.7059	0.7083
boosted	0.7647	0.8333
forest	0.7059	0.7500

The most accurate model for predicting the cluster each new store should be assigned to, is the boosted model with an accuracy of 76%, which is an appealing percentage. A very important metric is F1 score as it entails also the balance between precision and recall.

Therefore, Boosted Model has 83% F1 score. It scores the highest and we will use this one.

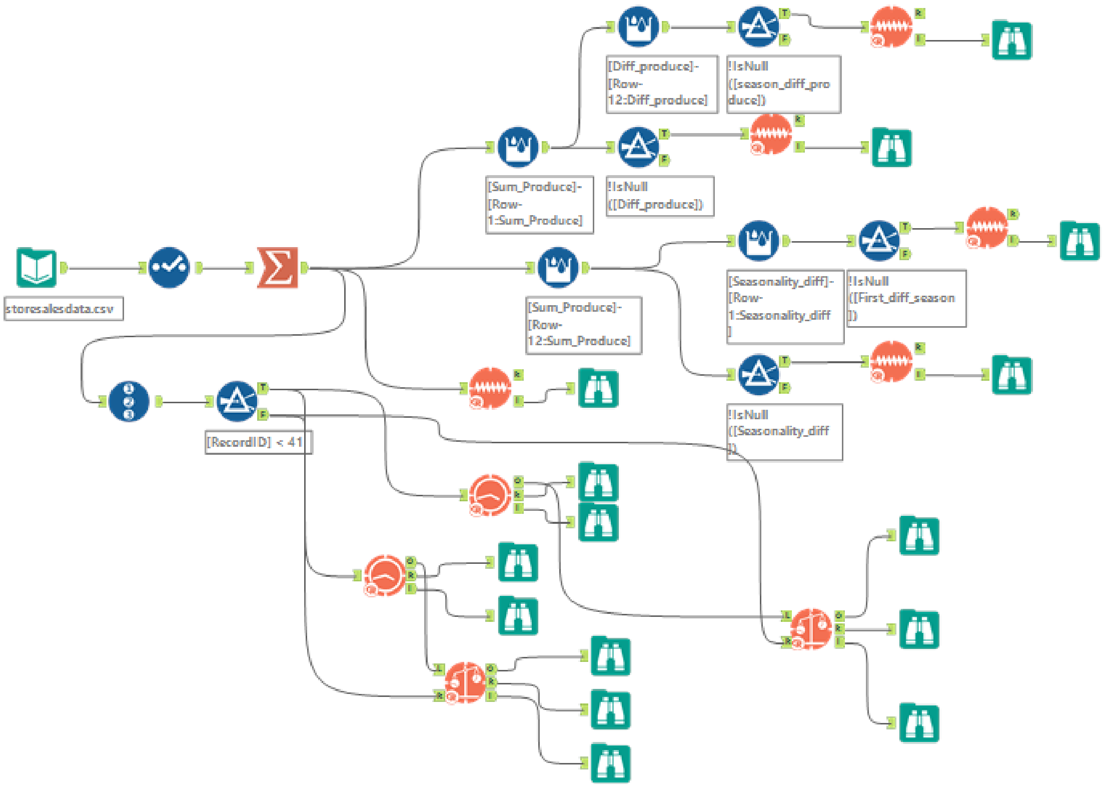
2. What format do each of the 10 new stores fall into? Please fill in the table below.

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

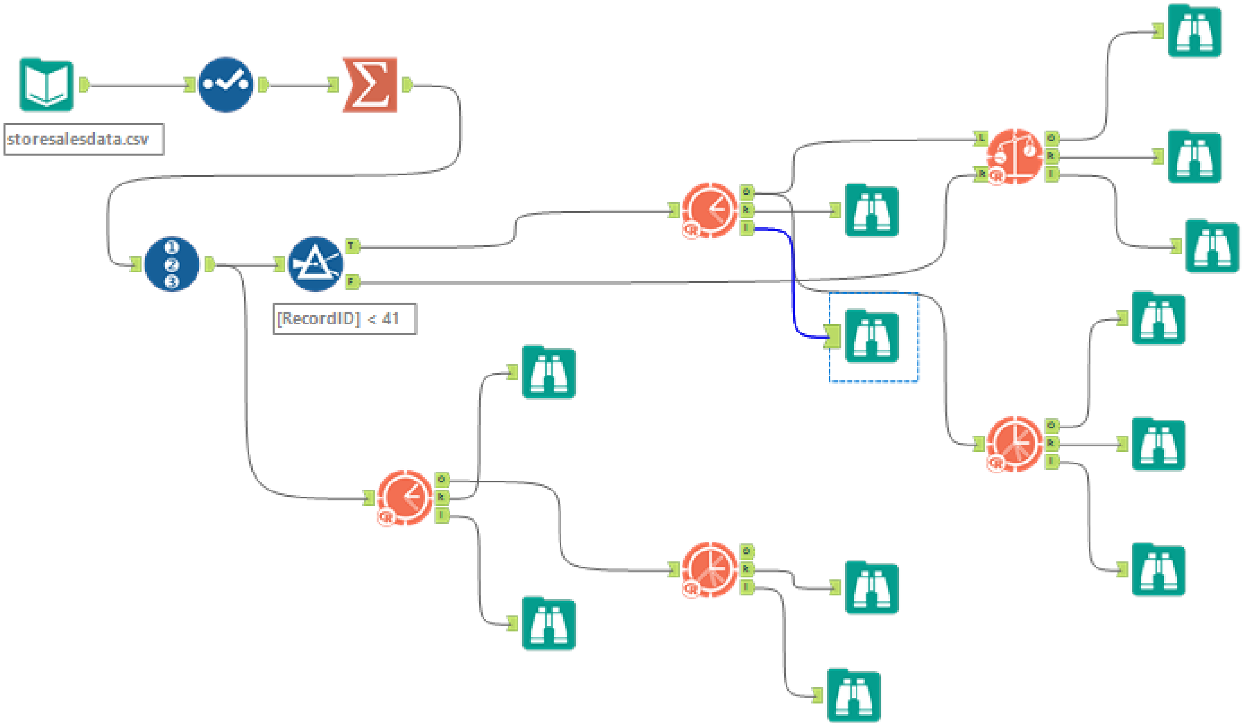
Task 3: Predicting Produce Sales

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?

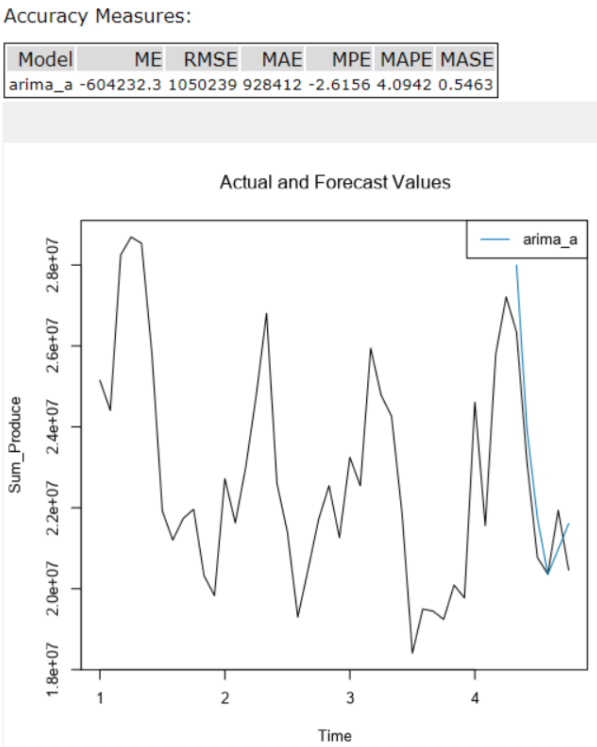
After running numerous possibilities and iterations for ARIMA and ETS, I came to the final comparison between ARIMA (1,0,0)(1,1,0)[12] and ETS (M,N,M).


storesalesdata.csv
[Diff_produce]-[Row-12:Diff_produce]
!IsNull([season_diff_produce])
[Sum_Produce]-[Row-1:Sum_Produce]
!IsNull([Diff_produce])
[Sum_Produce]-[Row-12:Sum_Produce]
[Seasonality_diff]-[Row-1:Seasonality_diff]
!IsNull([First_diff_season])
!IsNull([Seasonality_diff])
[RecordID] < 41

ARIMA workflow

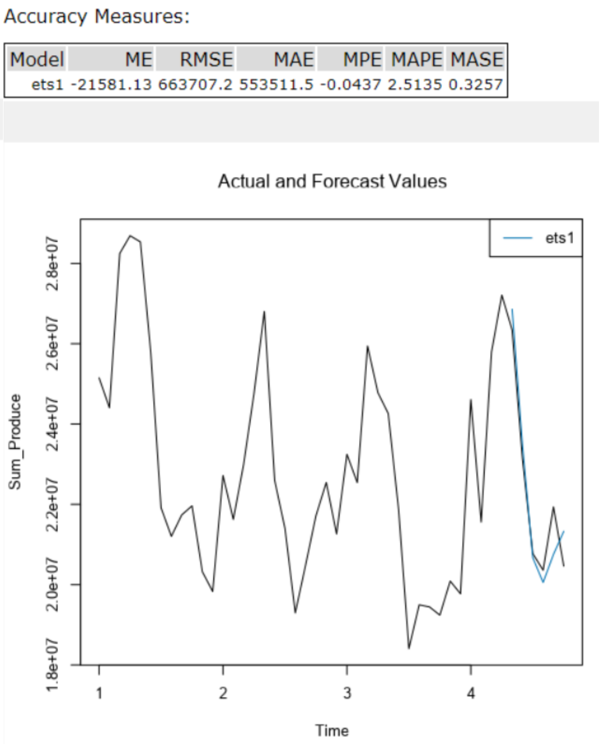


ETS workflow



Accuracy Measures:

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



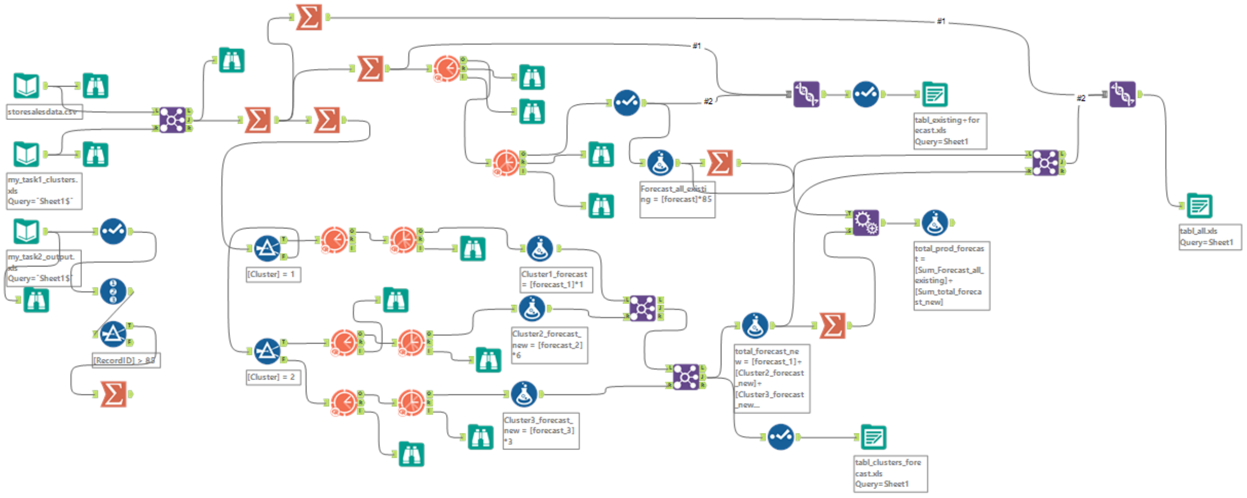
Accuracy Measures:

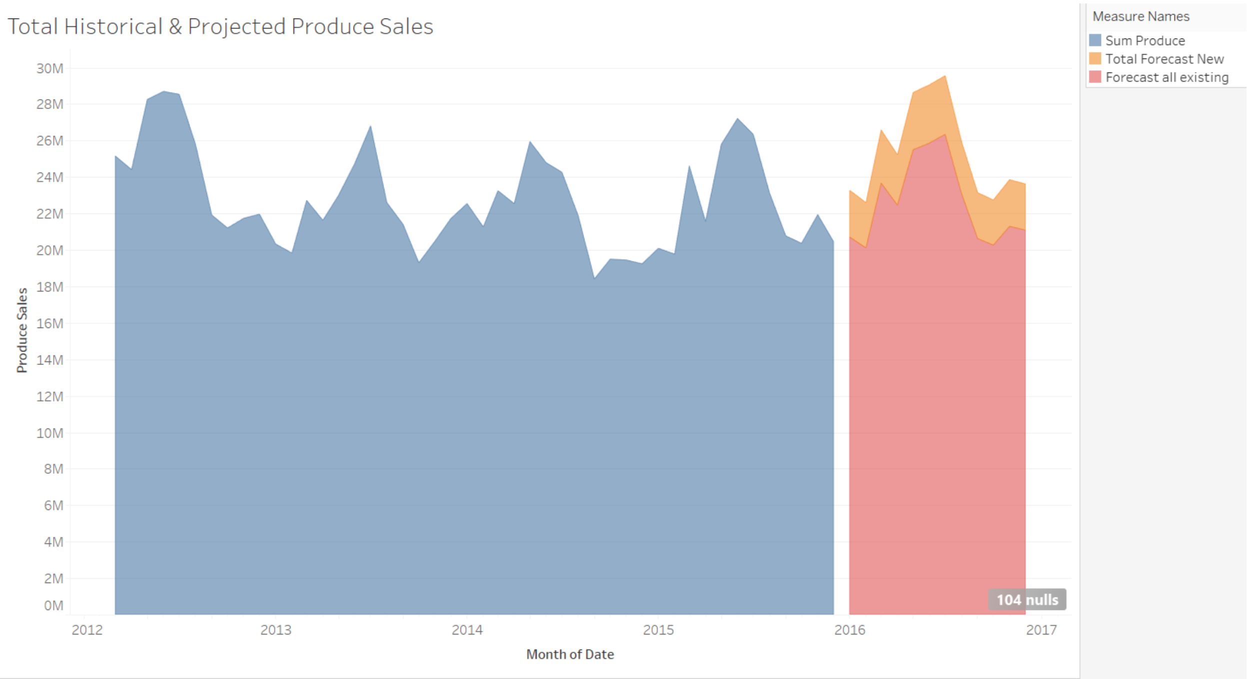
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ets1	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257

Placing the graphics side by side and comparing the errors provided by each model, we can clearly see that ETS performs mostly better than ARIMA, the errors are smaller and in the 6 month holdout sample, ETS closely follows the actual data.

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	New stores	Existing Stores
2,016	1	2,563,357.91	20,705,352.06
2,016	2	2,483,924.73	20,110,037.34
2,016	3	2,910,944.15	23,667,529.44
2,016	4	2,764,881.87	22,453,552.67
2,016	5	3,141,305.87	25,493,363.16
2,016	6	3,195,054.20	25,858,644.49
2,016	7	3,212,390.95	26,340,298.85
2,016	8	2,852,385.77	23,118,849.50
2,016	9	2,521,697.19	20,624,065.23
2,016	10	2,466,750.89	20,267,315.97
2,016	11	2,557,744.59	21,293,661.31
2,016	12	2,530,510.81	21,089,480.41


storesalesdata.csv
my_task1_clusters.xls Query=`Sheet1$`
my_task2_output.xls Query=`Sheet1$`
[RecordID] > 88
[Cluster] = 1
[Cluster] = 2
Cluster1_forecast = [forecast_1]*1
Cluster2_forecast_new = [forecast_2]*6
Cluster3_forecast_new = [forecast_3]*3
Forecast_all_existing = [forecast]*85
#1
#2
tabl_existing+forecast.xls Query=Sheet1
total_prod_forecast = [Sum_Forecast_all_existing]+[Sum_total_forecast_new]
total_forecast_new = [forecast_1]+[Cluster2_forecast_new]+[Cluster3_forecast_new...
tabl_clusters_forecast.xls Query=Sheet1
tabl_all.xls Query=Sheet1



https://public.tableau.com/profile/felix.lupu#!/vizhome/task3_16117886092750/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.