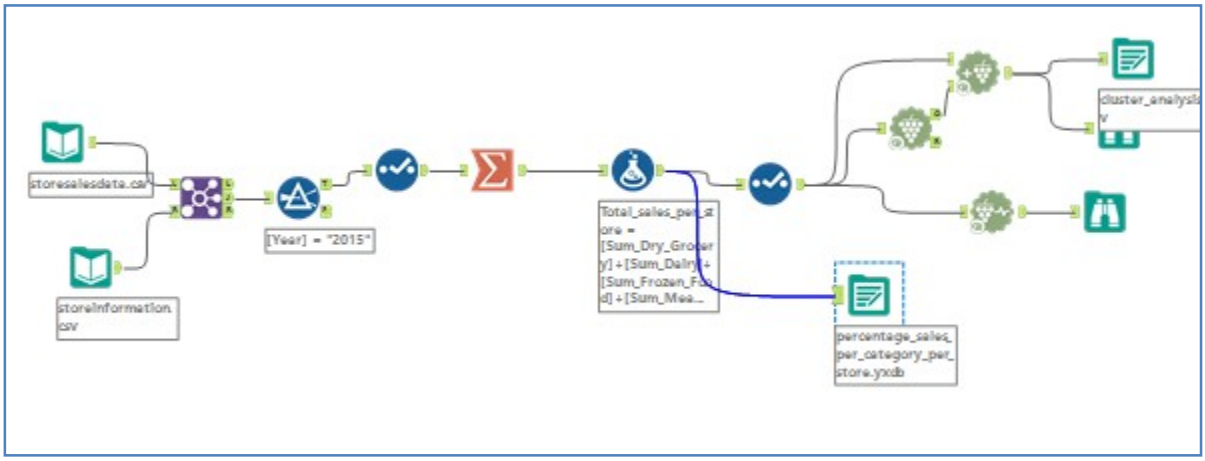
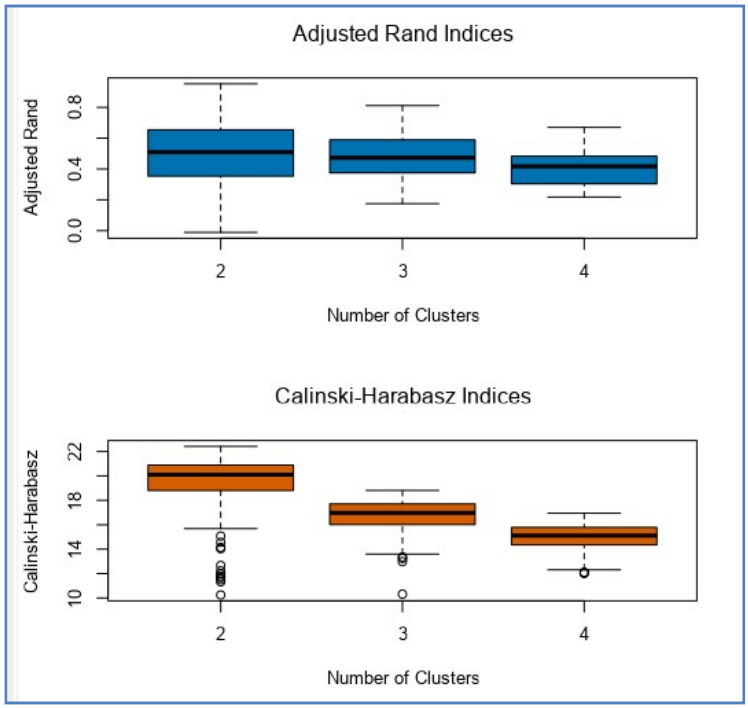
# Project: Predictive Analytics Capstone

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?
   3 is the optimal number of clusters because for 3 clusters AR and CH indices have high median values and also the variance is lower and are more compact.

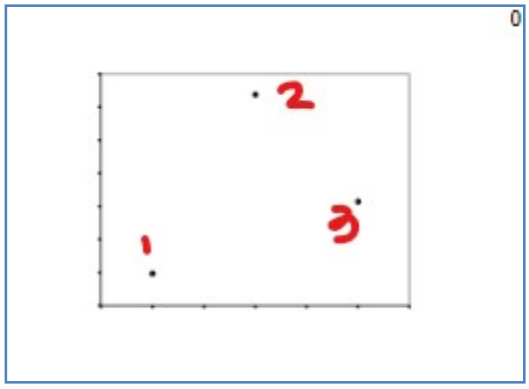


2. How many stores fall into each store format?
   Please find below number of store as per cluster

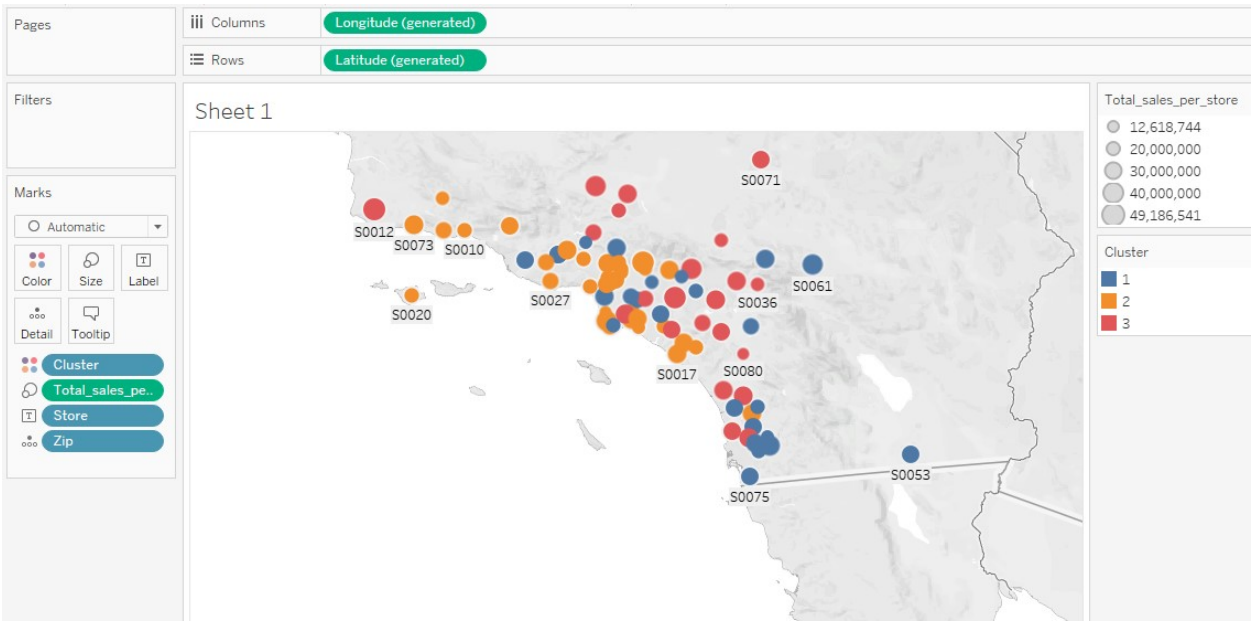
| Cluster | |
|---|---|
| 2 | 35 |
| 1 | 25 |
| 3 | 25 |

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Looking into sales of data of cluster, It's been observed that 2 Cluster is large sale cluster where as 3 is medium size cluster and 1 small size sales cluster

| Record | Cluster | Sum_Total_sales_per_store |
|---|---|---|
| 1 | 1 | 698302817.97 |
| 2 | 2 | 968990542.87 |
| 3 | 3 | 807040485.46 |

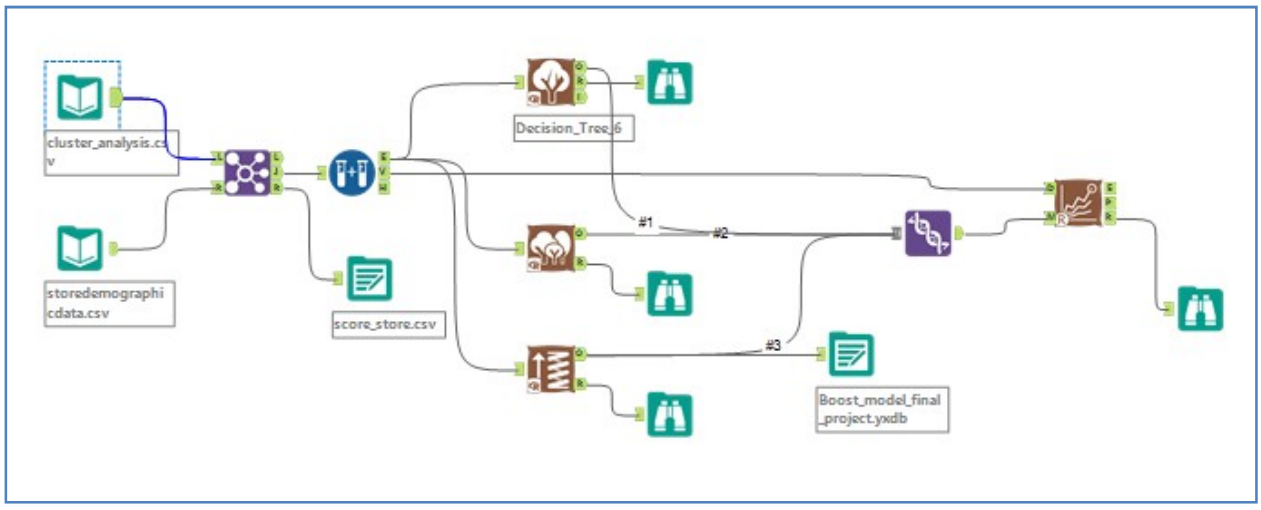



4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.




# 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.)
   Classification (Non-binary) is good methodology to predict the store format for new stores. Here we have compared decision tree, forest, and boosted model to select best model.




**Model Accuracy comparision**

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| Decision_Tree_6 | 0.7059 | 0.7083 | 0.6250 | 1.0000 | 0.5000 |
| forest_model | 0.7059 | 0.7500 | 0.5000 | 1.0000 | 0.7500 |
| boost_model | 0.7647 | 0.8333 | 0.5000 | 1.0000 | 1.0000 |

## Confusion matrix

**Confusion matrix of Decision_Tree_6**

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

**Confusion matrix of boost_model**

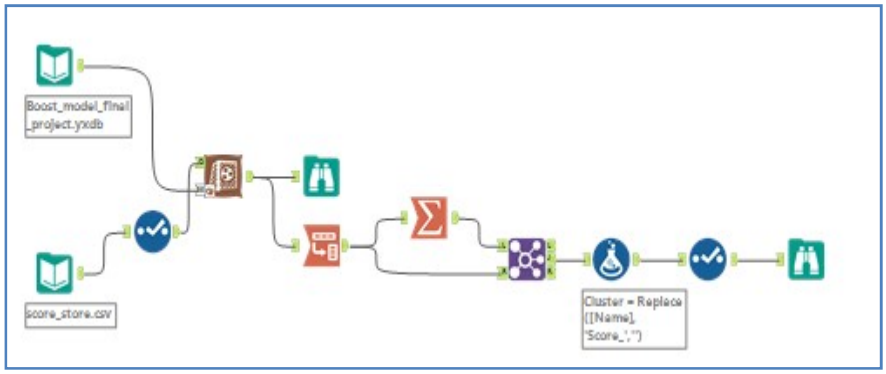
| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 4 | 0 | 0 |
| Predicted_2 | 2 | 5 | 0 |
| Predicted_3 | 2 | 0 | 4 |

**Confusion matrix of forest_model**

| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 4 | 0 | 1 |
| Predicted_2 | 2 | 5 | 0 |
| Predicted_3 | 2 | 0 | 3 |

We have selected **boost model** as its accuracy is high as compared to others.

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

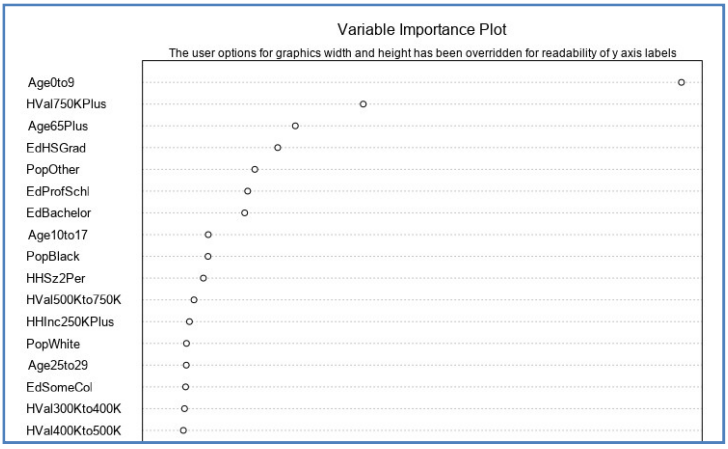



| R... | Store | Cluster |
|---|---|---|
| 1 | S0086 | 1 |
| 2 | S0087 | 2 |
| 3 | S0088 | 3 |
| 4 | S0089 | 2 |
| 5 | S0090 | 2 |
| 6 | S0091 | 3 |
| 7 | S0092 | 2 |
| 8 | S0093 | 3 |
| 9 | S0094 | 2 |
| 10 | S0095 | 2 |

| Store Number | Segment |
|---|---|
| S0086 | 1 |
| S0087 | 2 |
| S0088 | 3 |
| S0089 | 2 |
| S0090 | 2 |
| S0091 | 3 |

| | |
|---|---|
| S0092 | 2 |
| S0093 | 3 |
| S0094 | 2 |
| S0095 | 2 |

As per below, it is been observed that age0to9, HVal750Kplus, Age65Plus and EdHSGrad are variables helps to decide store formats



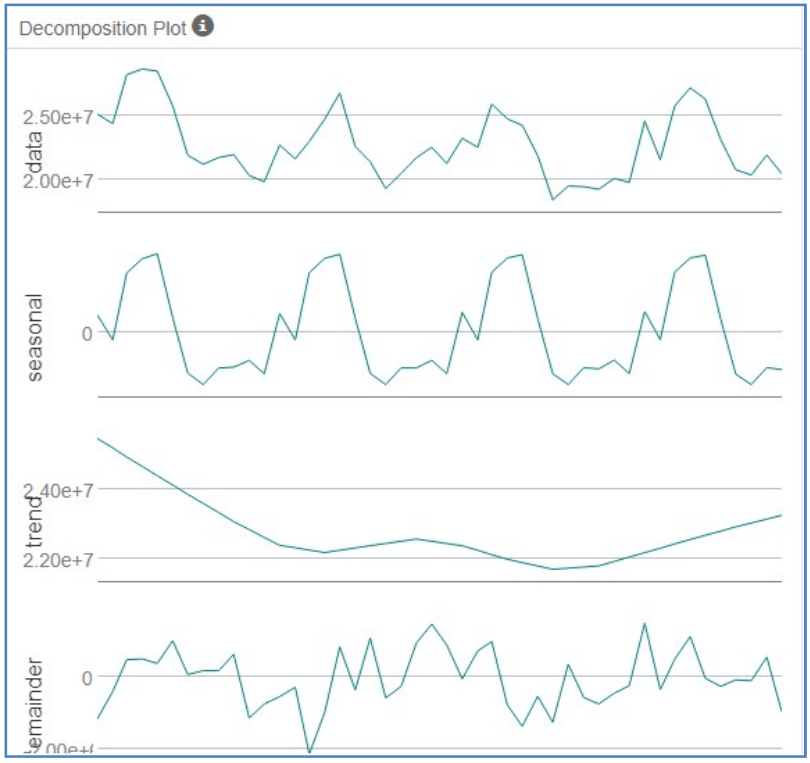

# Task 3: Predicting Produce Sales

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?

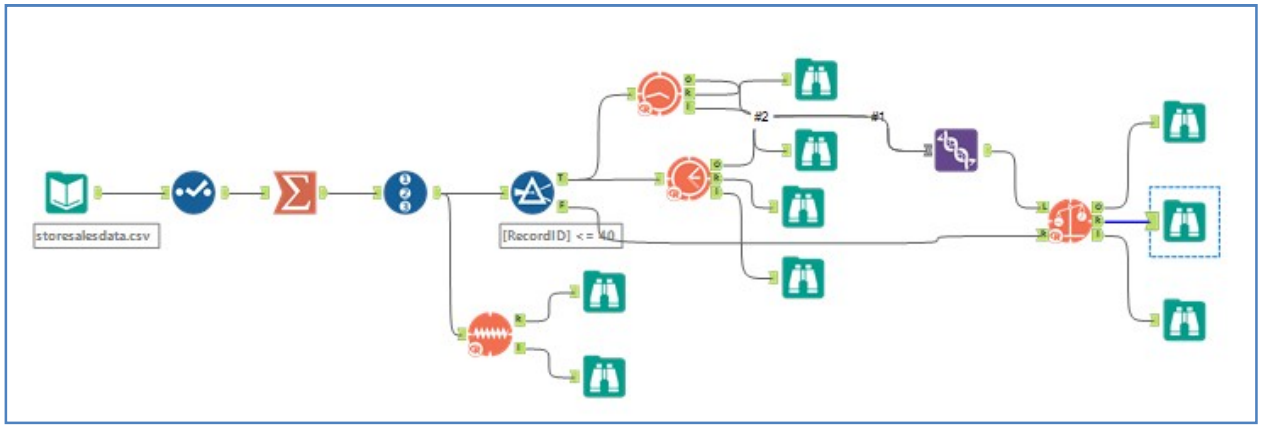
From decomposition plot, its been observed that there is no trend, seasonal is multiplicative and error is multiplicative. After comparing the results against the holdout sample, the ETS performs better against the ARIMA model. So I select ETS model to forecast.

Accuracy Measures:

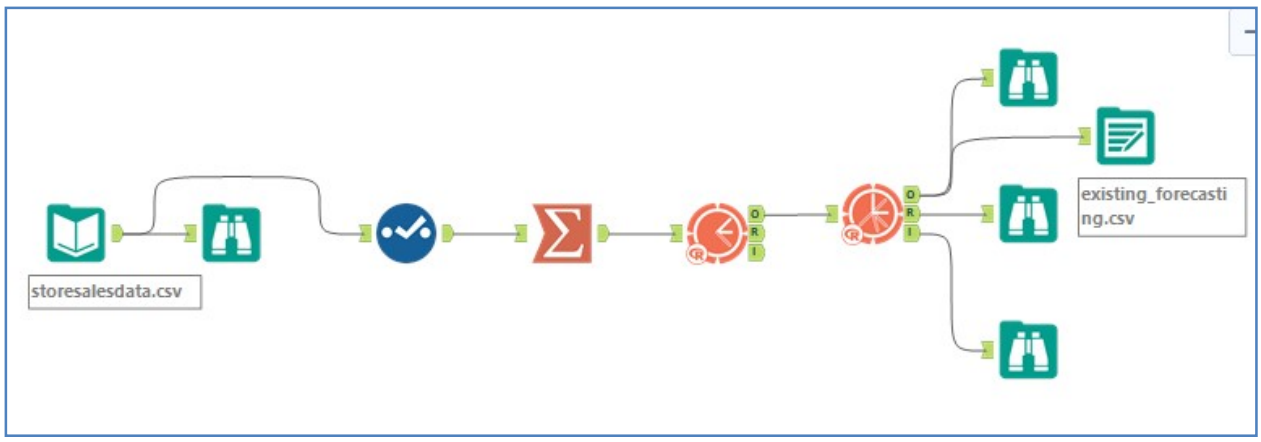
| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| ARIMA | -604232.29 | 1050239.2 | 928412 | -2.6156 | 4.0942 | 0.5463 |
| ETS | -21581.13 | 663707.2 | 553511.5 | -0.0437 | 2.5135 | 0.3257 |



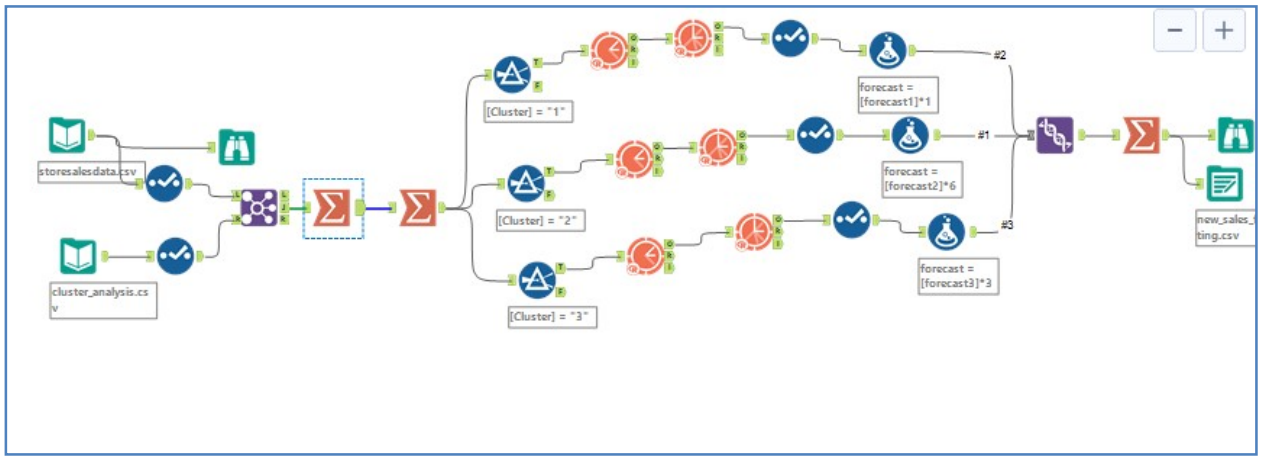

## Model Analysis:



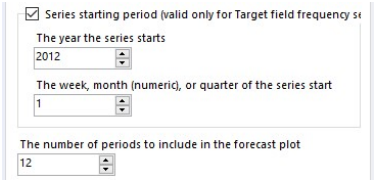

## Forecasting existing store sales:




**Forecast new store sales:**

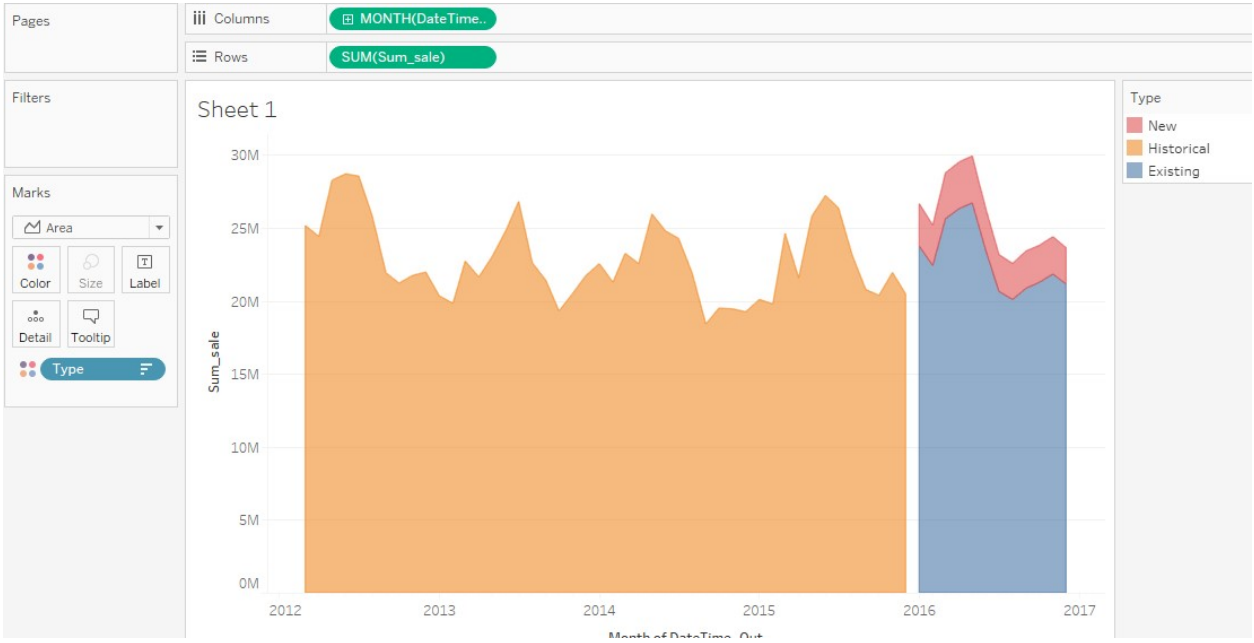



3. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.
   Note: as mentioned in review, I have mentioned starting year, but t shows no change in data




| DateTime_Out | Existing Store | New Store |
|---|---|---|
| 01-01-2016 | 23735686.94 | 2910944.146 |
| 01-02-2016 | 22409515.28 | 2764881.87 |
| 01-03-2016 | 25621828.73 | 3141305.867 |

| | | |
|---|---|---|
| 01-04-2016 | 26307858.04 | 3195054.204 |
| 01-05-2016 | 26705092.56 | 3212390.954 |
| 01-06-2016 | 23440761.33 | 2852385.769 |
| 01-07-2016 | 20640047.32 | 2521697.187 |
| 01-08-2016 | 20086270.46 | 2466750.894 |
| 01-09-2016 | 20858119.96 | 2557744.588 |
| 01-10-2016 | 21255190.24 | 2530510.805 |
| 01-11-2016 | 21829060.03 | 2563357.91 |
| 01-12-2016 | 21146329.63 | 2483924.728 |




## 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.