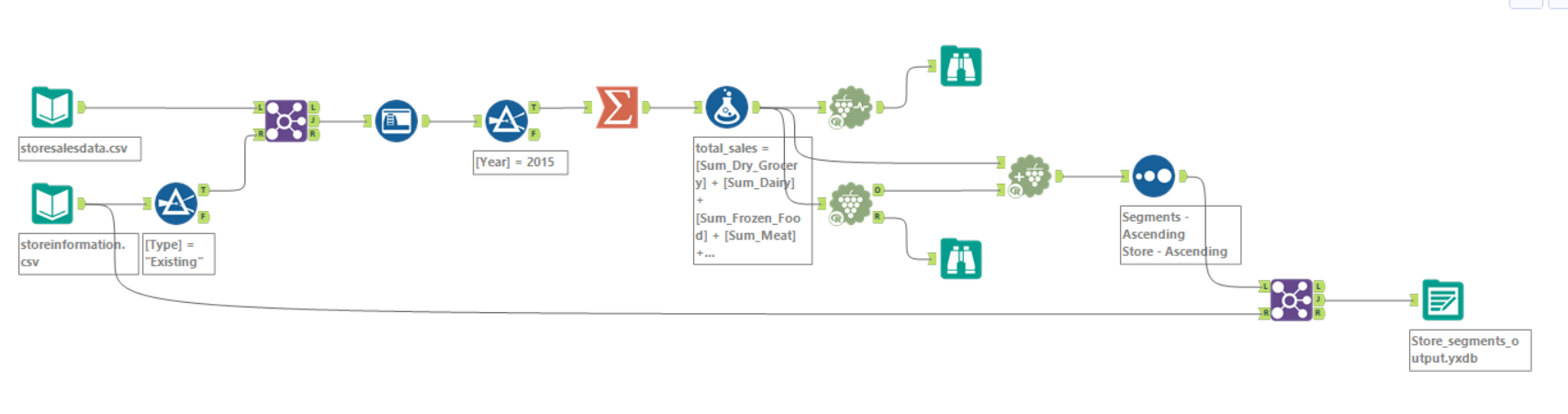
Project: Predictive Analytics Capstone

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>

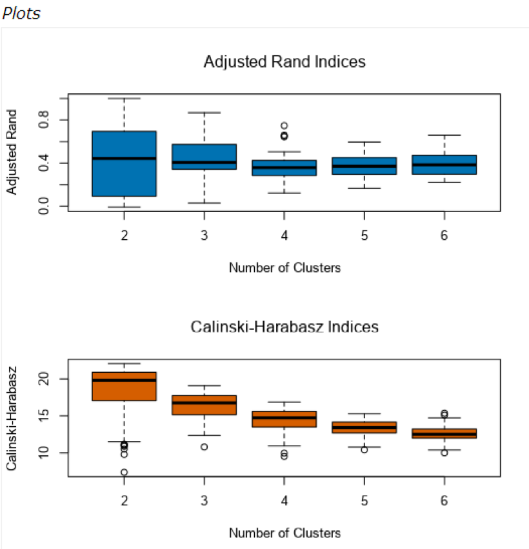
## Task 1: Determine Store Formats for Existing Stores

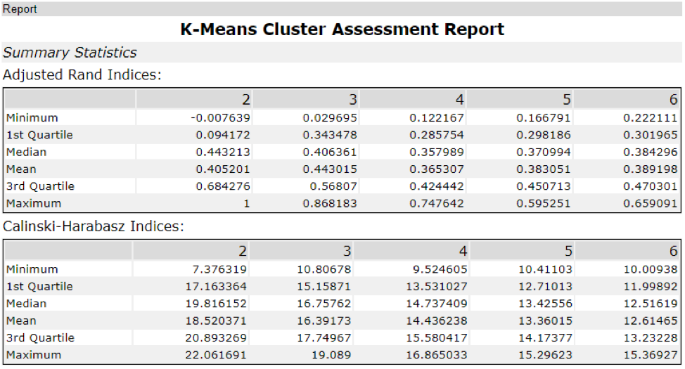


1. What is the optimal number of store formats? How did you arrive at that number?

The optimal number store formats are 3.

The optimal number of store formats is 3. This is because it has high median values within both the AR and CH index and smaller spread, showing compactness. We can get this through using the K-Centroids diagnosis tool on Alteryx.





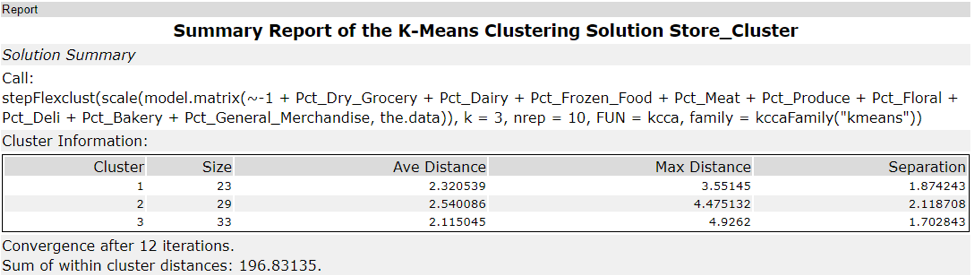
1. How many stores fall into each store format?

Format 1: 23 Stores

Format 2: 29 Stores

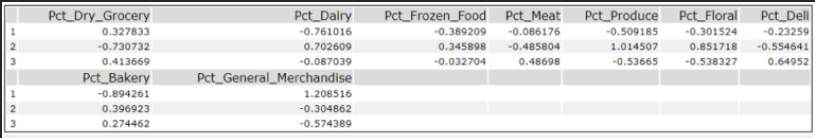
Format 3: 33 Stores

Using the K-Centroids Cluster Analysis tool using the same configuration as we used in K-Centroids Diagnostics tool:

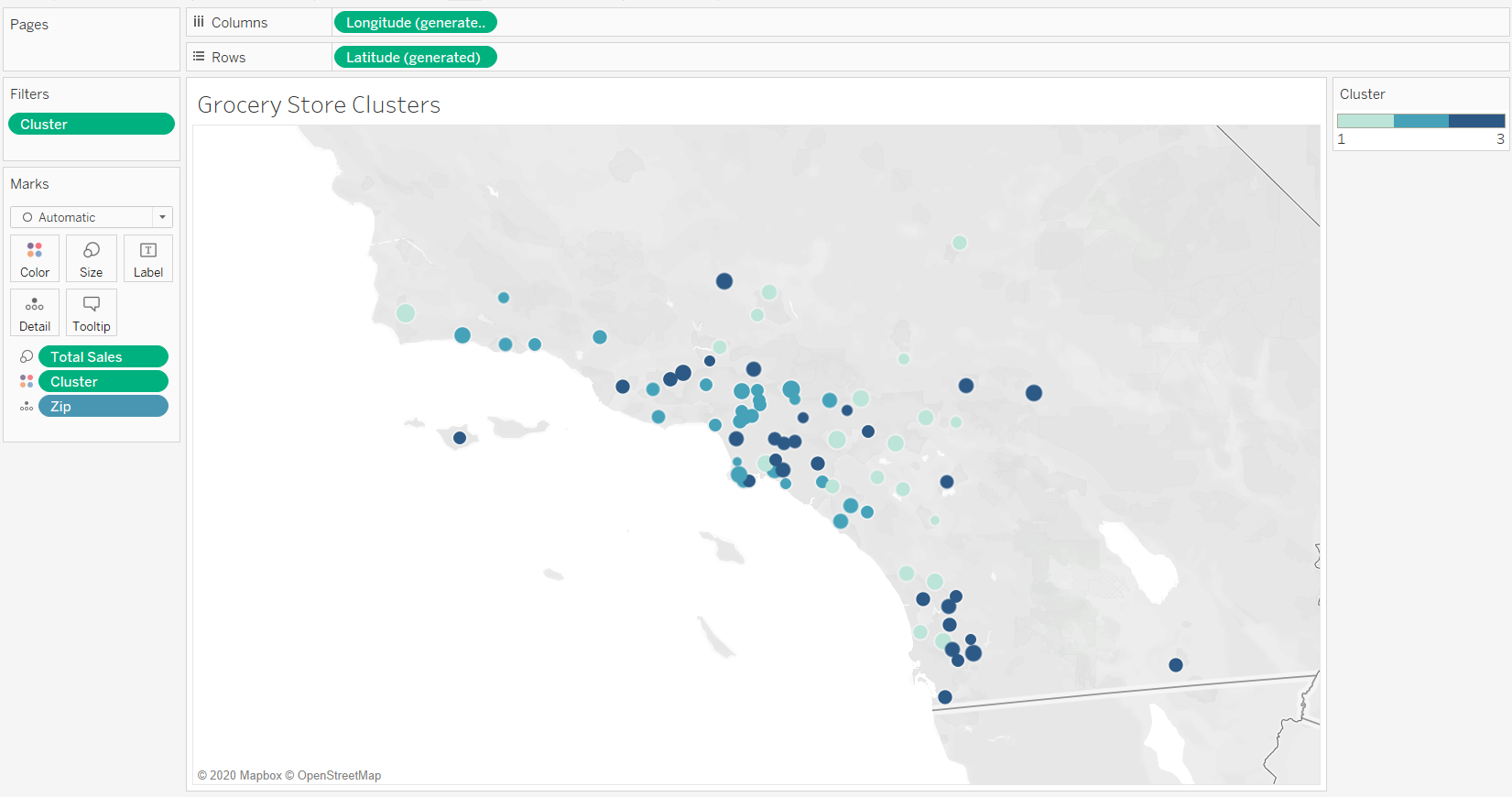


1. Based on the results of the clustering model, what is one way that the clusters differ from one another?

Based on the summary report of the K-Means Clustering solution, considering the percentage of sales by category of each store, cluster 1 sells more in general merchandise; cluster 2 sells more in produce and floral; and cluster 3 sells more in deli and meat; etc.

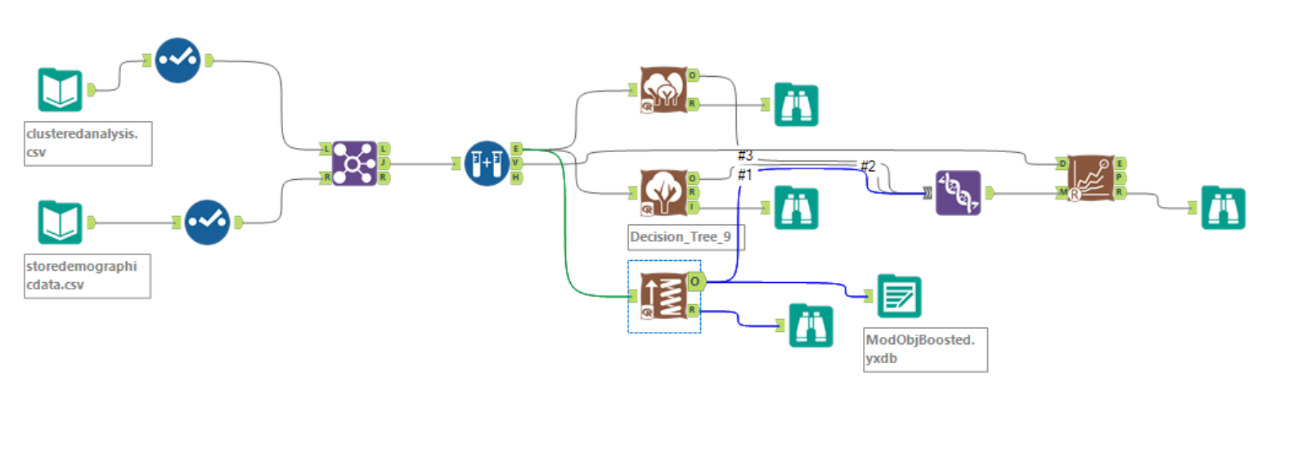


1. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.

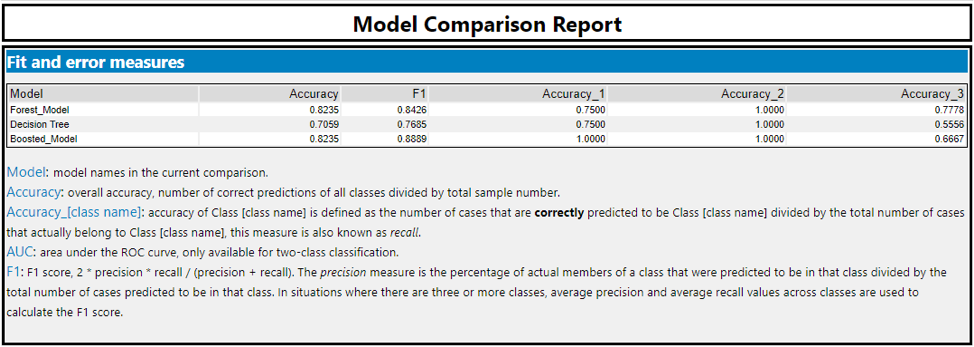


## Task 2: Formats for New Stores

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.)

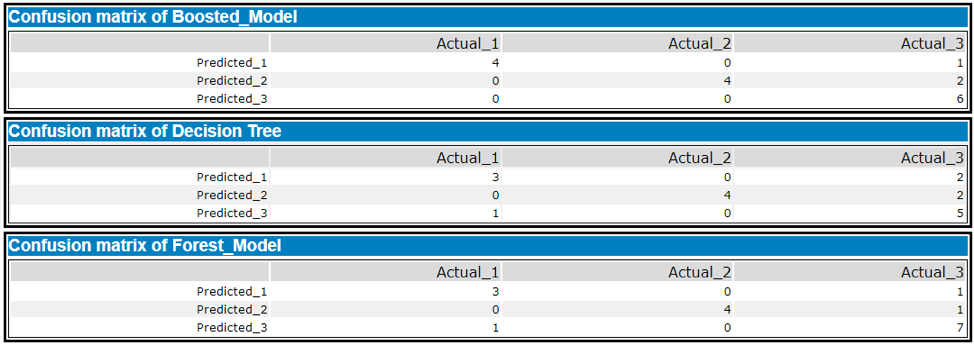


Since we must predict the clusters for the stores, so it is a classification problem and we have more than 2 classes to predict. So, we will use Decision Tree, Random Forest and Boosting. Using the workflow above, we get the statistics:

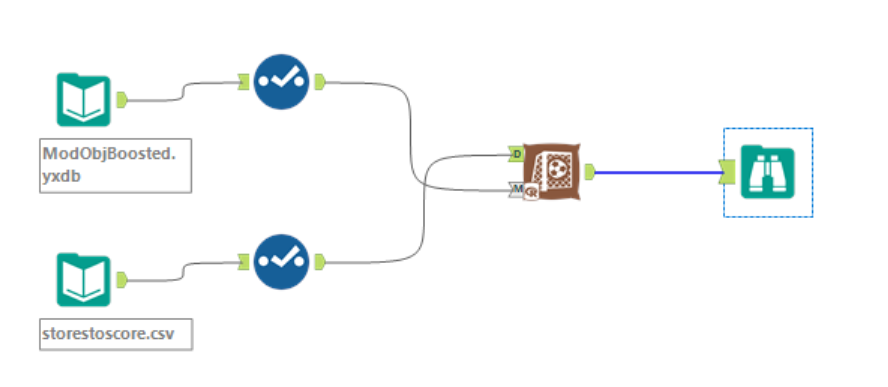


Based on the results of the model comparison, we can clearly decide that the **Boosting algorithm** is producing the best model as it had high accuracy of 82.35% and max F1 score of 0.8889. Hence, we will use the Boosting algorithm to make the prediction.

Use the above created model to predict the clusters for the new stores:



1. 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 |

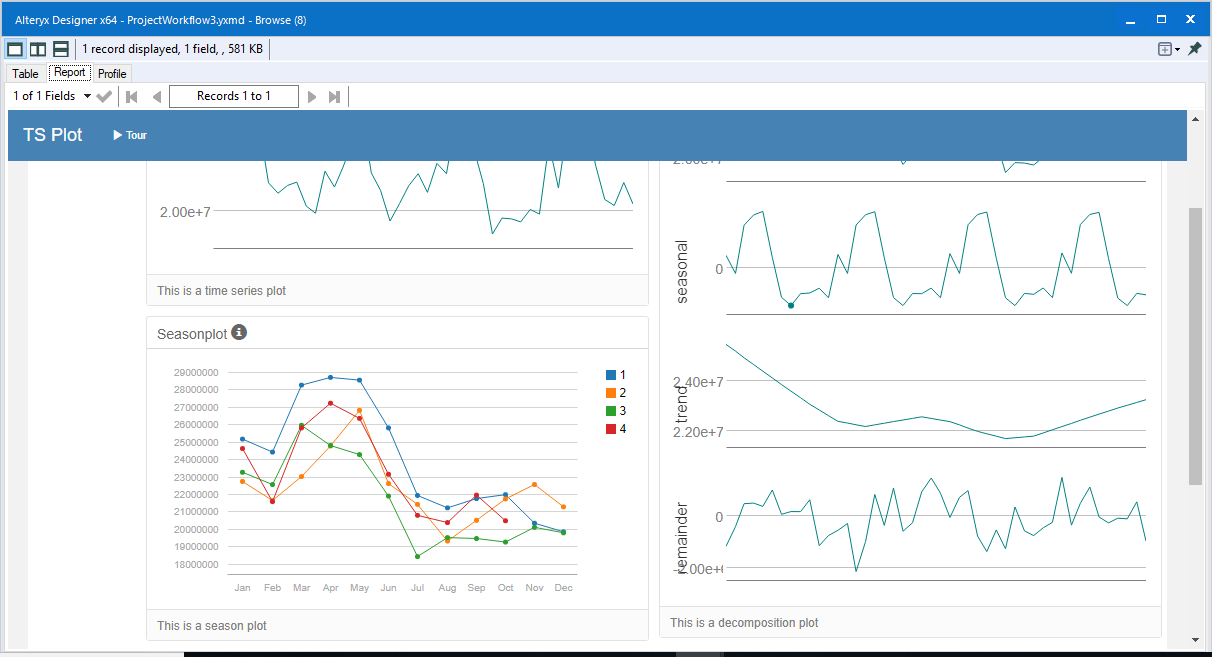
So, there are 3 stores in cluster 1, 6 stores in cluster 2 and 1 store in cluster 3.

## Task 3: Predicting Produce Sales

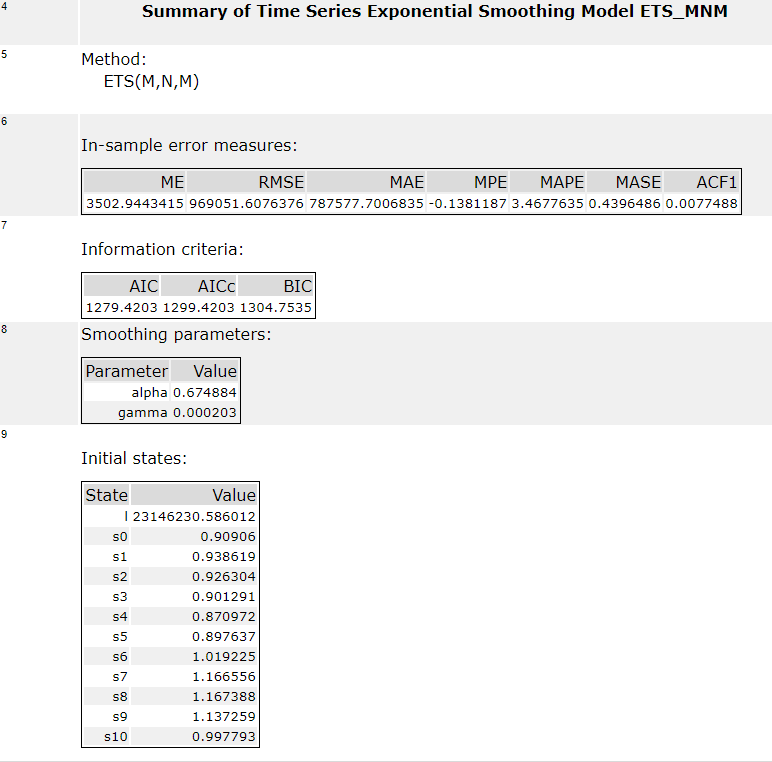
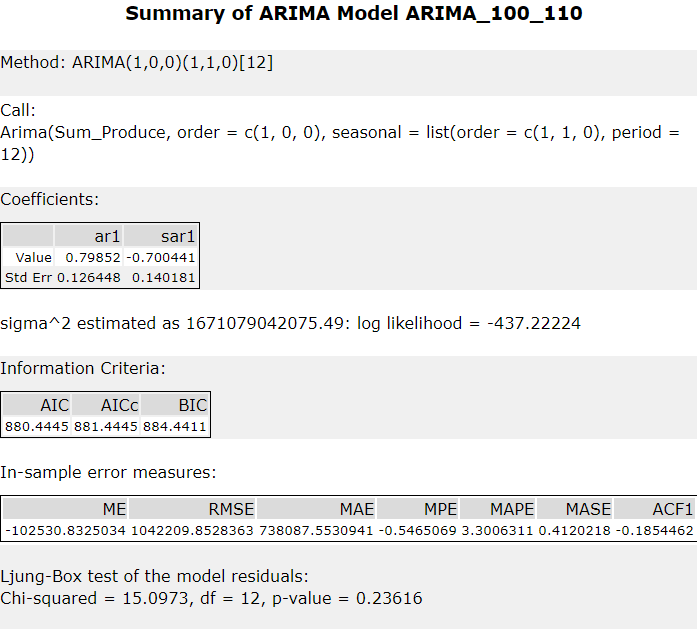
## 

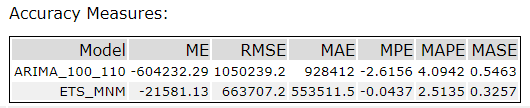
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

For existing stores, I used both the ETS and ARIMA models to find the best solution. For predicting the aggregate produce for the existing stores, I plotted the Decomposition plots to understand the trend, seasonality, and error. Looking at the three plots below, it is apparent that there exists seasonality and the error appears to decrease over time. Since the trend curve slopes upward after a period, I will not use that. So, I will have seasonality multiplicatively, trend as none, and remainder multiplicatively giving an ETS (M, N, M).



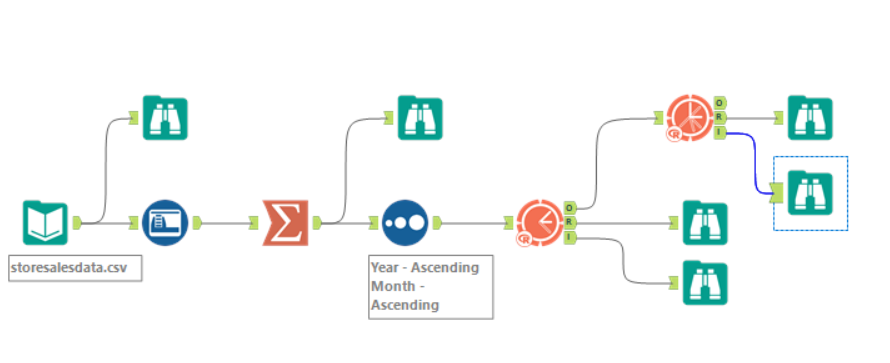
We can see in the decomposition plot above there is no trend, seasonal is multiplicative and error is multiplicative. After comparing the results against the holdout sample, the ETS(M,N,M) performs better against the ARIMA(1,0,0) (1,1,0) model.



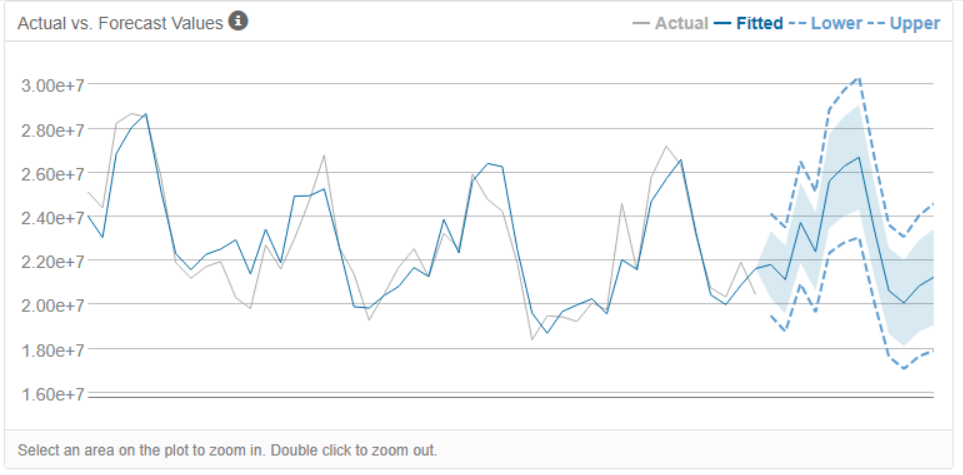


2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

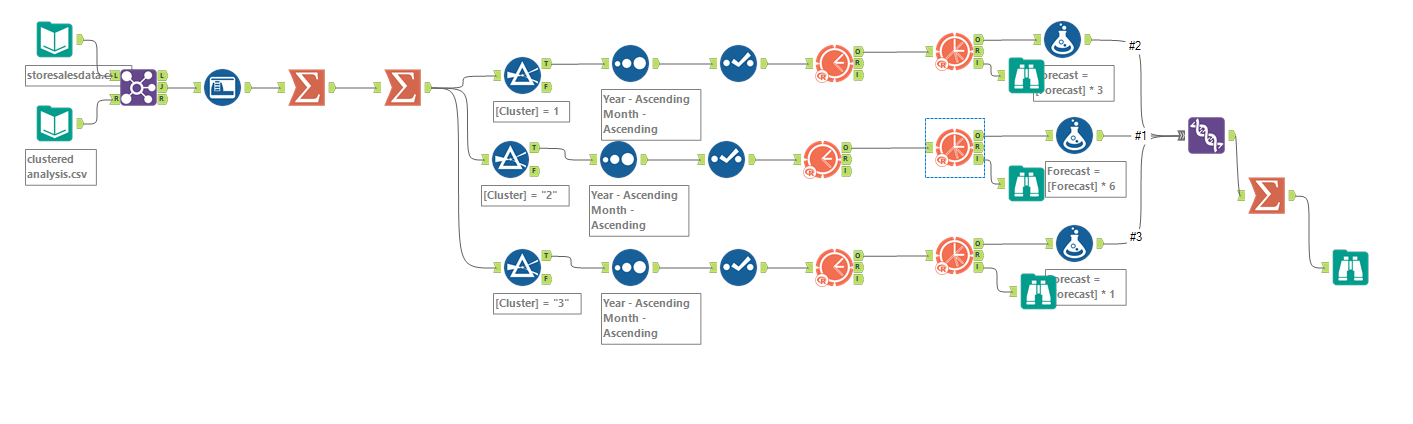
For Existing Stores - Workflow:



For Existing Stores - Forecast plot:

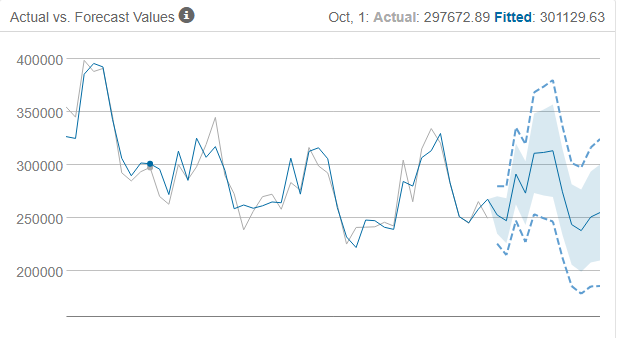


For New Stores - Workflow:

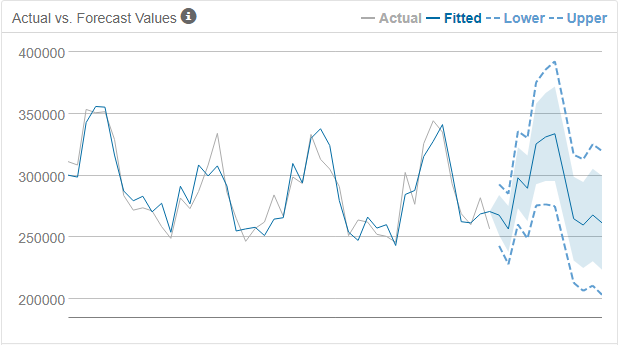


For New Stores – Individual Forecast plot:

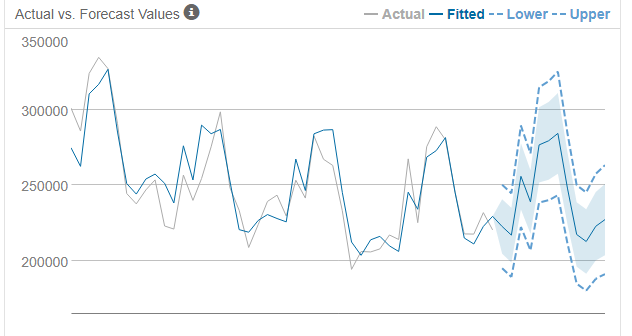
Cluster 1:



Cluster 2:



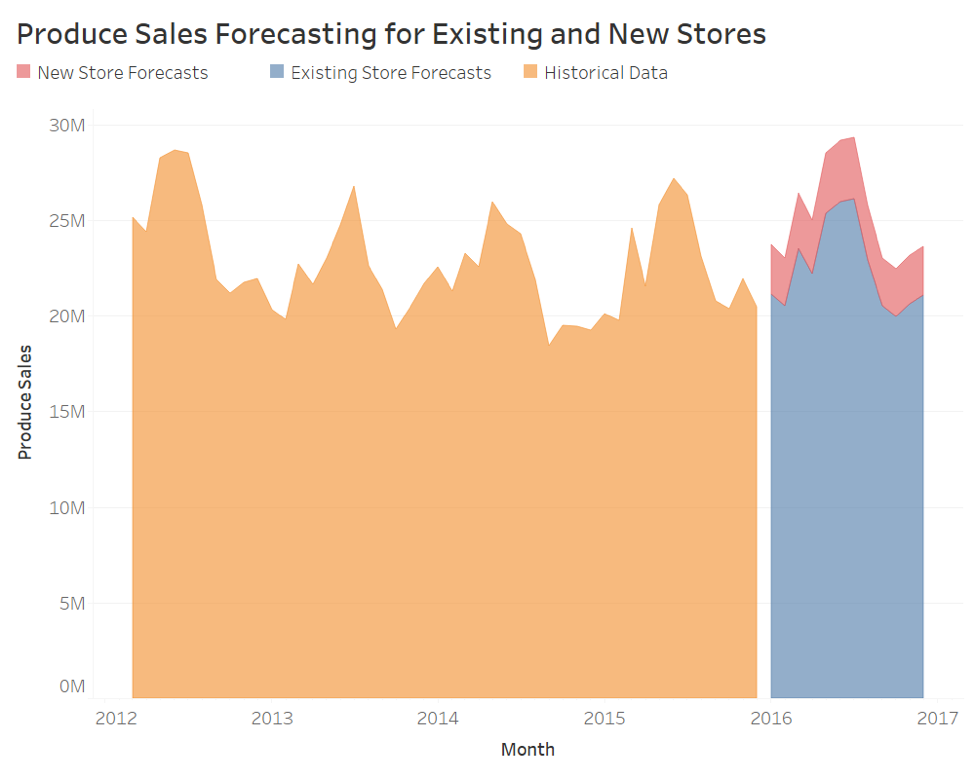
Cluster 3:



Forecast Table:

|  |  |  |
| --- | --- | --- |
| **Period** | **New Stores** | **Existing Stores** |
| Jan 16 | 2,587,451 | 21,539,936 |
| Feb 16 | 2,477,353 | 20,413,771 |
| Mar 16 | 2,913,185 | 24,325,953 |
| Apr 16 | 2,775,746 | 22,993,466 |
| May 16 | 3,150,867 | 26,691,951 |
| Jun 16 | 3,188,922 | 26,989,964 |
| July 16 | 3,214,746 | 26,948,631 |
| Aug 16 | 2,866,349 | 24,091,579 |
| Sep 16 | 2,538,727 | 20,523,492 |
| Oct 16 | 2,488,148 | 20,011,749 |
| Nov 16 | 2,595,270 | 21,177,435 |
| Dec 16 | 2,573,397 | 20,855,799 |

Tableau Visualisation:



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.