**Project: Predictive Analytics Capstone**

**(Abdulwasiu Tiamiyu)**

## **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 formats is 3. Although 2 has the highest median values within both the Adjusted Rand Indices and Calinski-Harabasz Indices, it has a lot of outliers and higher spread. 3 has smaller spread, showing compactness.

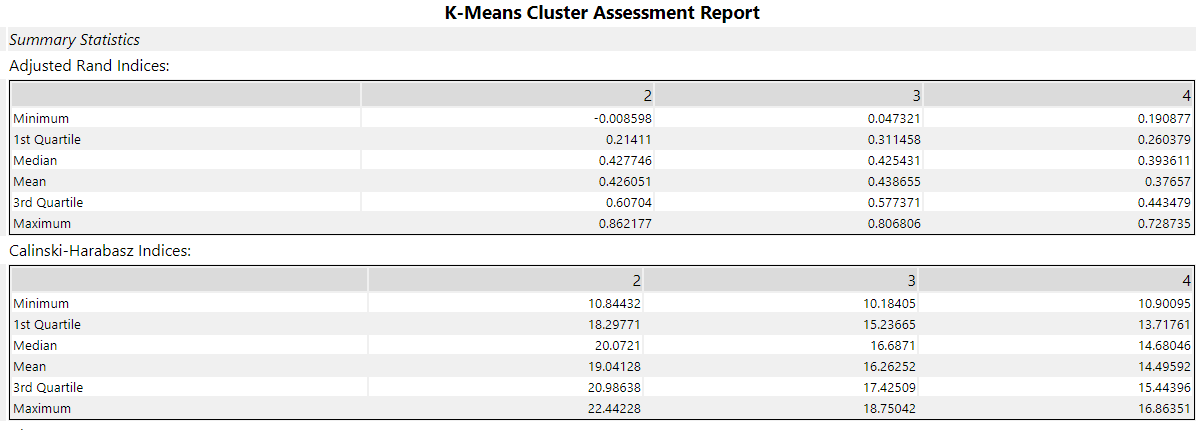


Figure 1.1: K-Means Cluster Assessment Report for Adjusted Rand and Calinski-Harabasz Indices

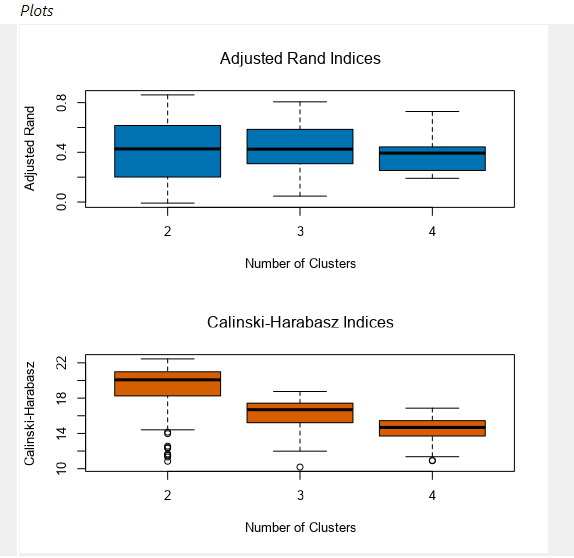


Figure 1.2: Plots for Adjusted Rand and Calinski-Harabasz Indices

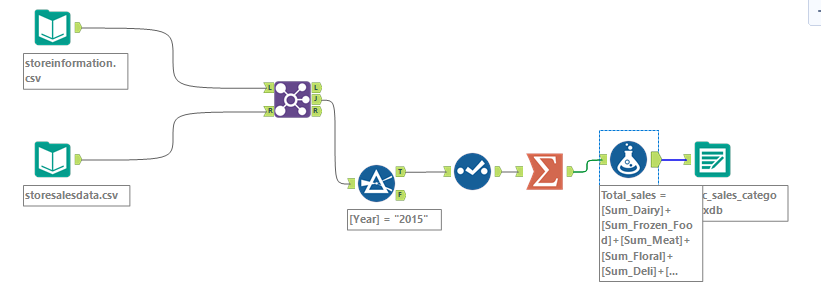


Figure 1.3: Percentage category workflow

1. **How many stores fall into each store format?**

Cluster 1 has 25 stores, Cluster 2 has 35 stores, and Cluster 3 has 25 stores

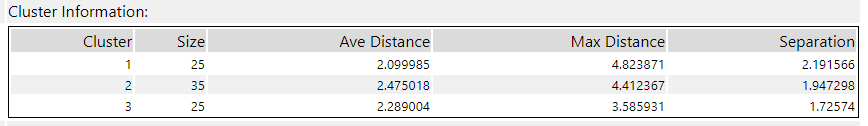


Figure 1.4: Cluster Distribution

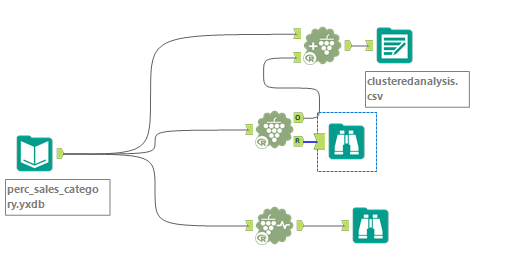


Figure 1.5: Clustered Analysis workflow

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

Stores that fall under cluster 3 for example would want an increase in General Merchandise as compared to stores that fall under cluster 1 & 2. Also, stores that fall under cluster 2 sold more Floral and Produce. Cluster 1 sold more Deli.

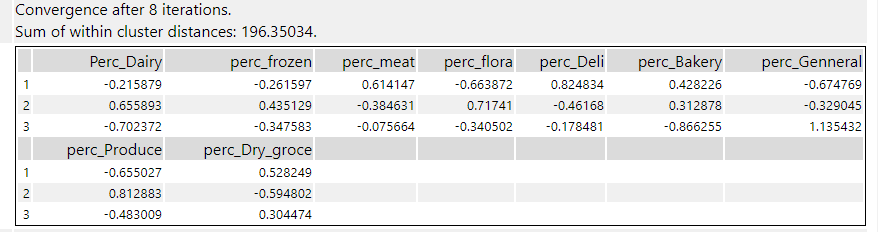
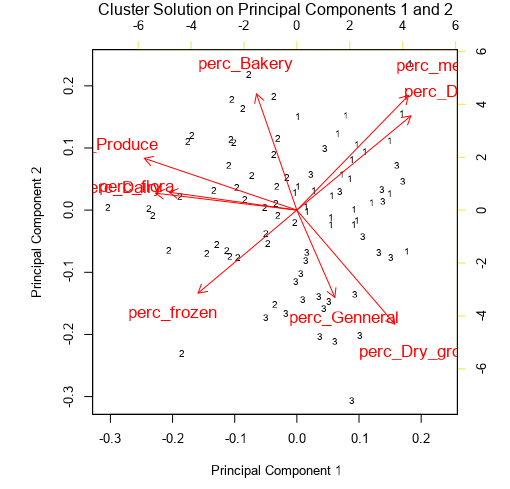


Figure 1.6: Summarized report of K-Means clustering



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.**

Tableau Visualization Link here:

https://public.tableau.com/app/profile/abdulwasiu.tiamiyu/viz/Clusterlocation/cluster?publish=yes

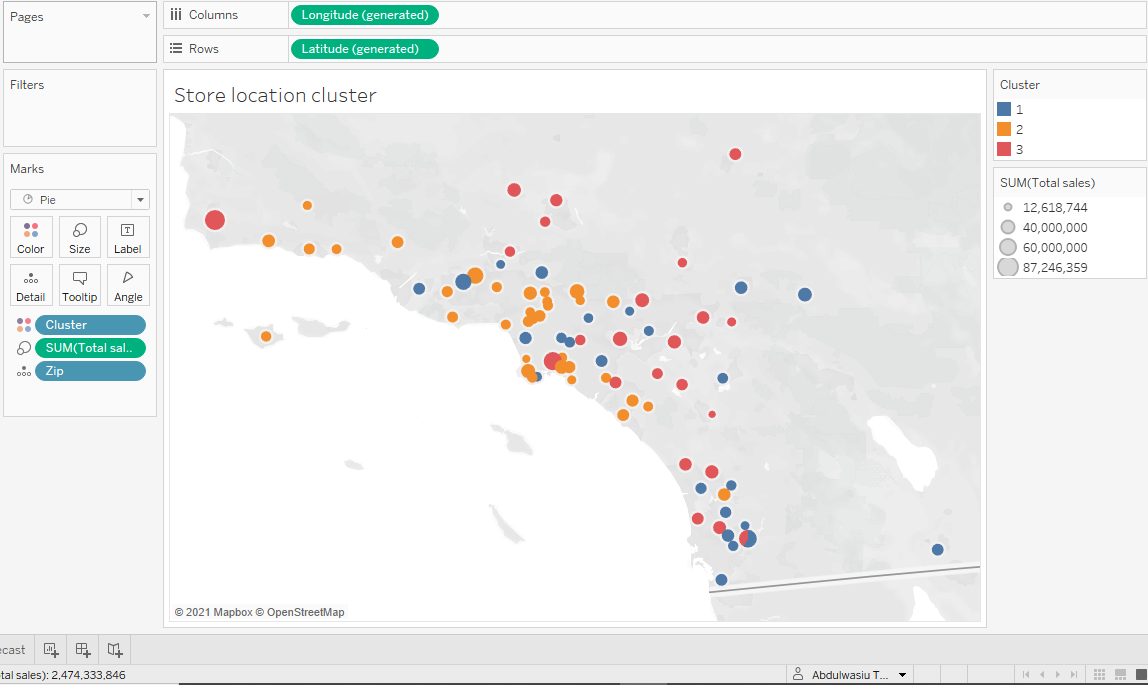


Figure 1.7: Location of stores

## **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.)**

I did not run a logistic regression model because this a non-binary classification problem. A decision tree, forest, and boosted model were created to predict the store formats for the new stores.

We can see in the figure below that the Forest Model and the Boosted Model have highest F1 score and highest average accuracy rate across Accuracy\_1, 2 and 3 (similar), so I can use any of the two to score our 10 new scores and assign it to either cluster 1,2 or 3.

I used the Boosted Model for my prediction because of its proven record over the forest model. I created an output on the boosted model tool (BoostedModel.yxdb).

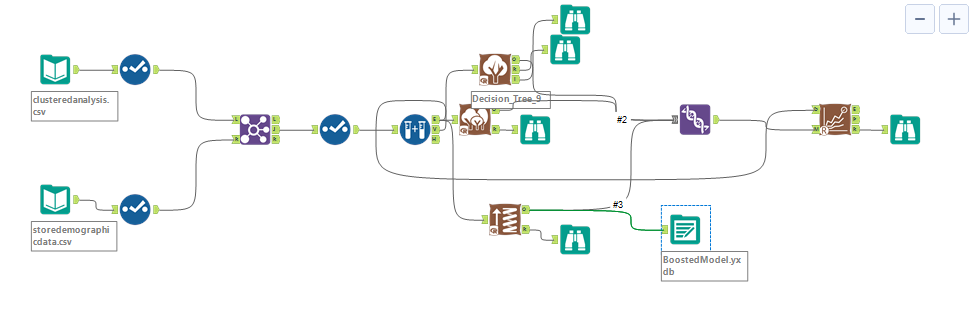


Figure 2.1: Models comparison workflow

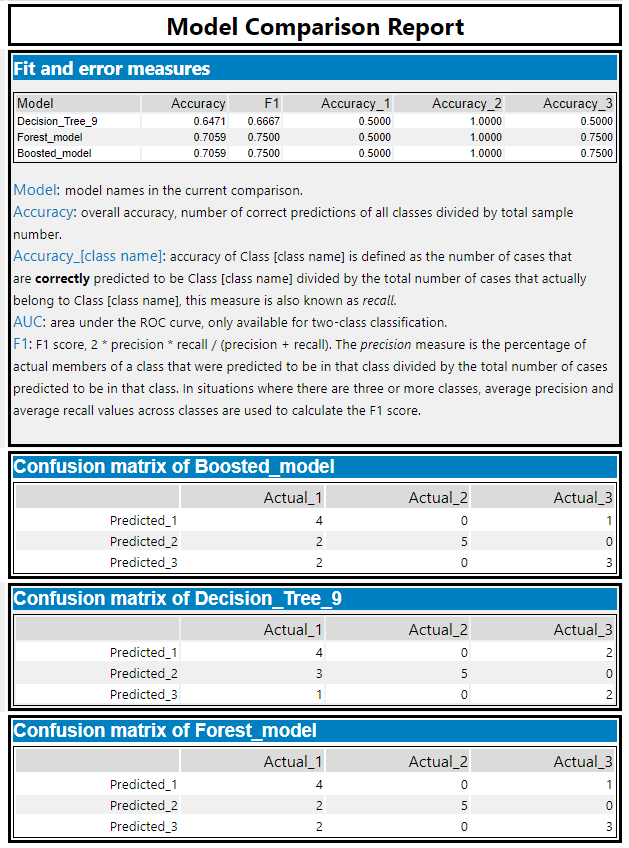


Figure 2.2: Comparison report & Confusion matrix for the three models

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

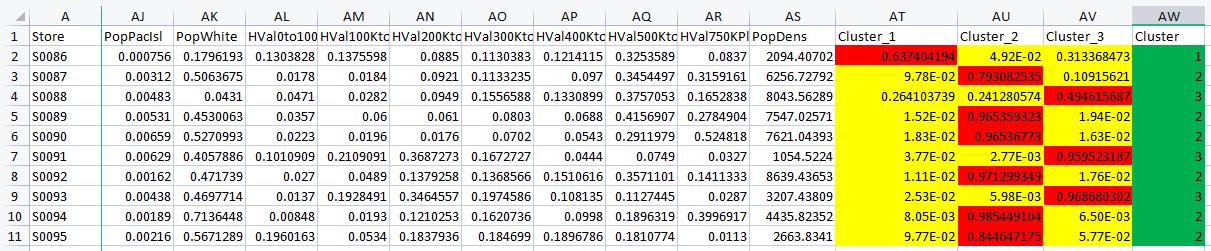


Figure 2.3: Formats for the 10 new stores

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

## Table 2.1: Segment clusters for new stores

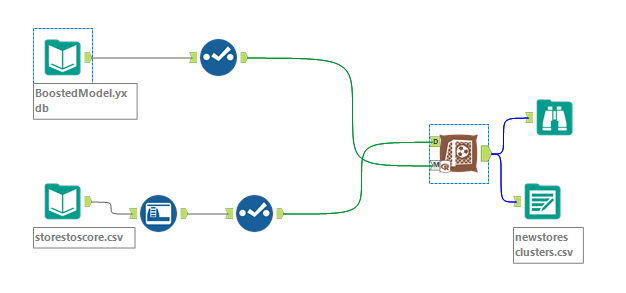


Figure 2.4: Workflow for the new formats

## **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?**

I have chosen to use ETS (M,N,M) model (with no dampening) for each of my forecast as it gave the best result. We can see in the decomposition plot below, seasonal is multiplicative and error is multiplicative (trend not applied).

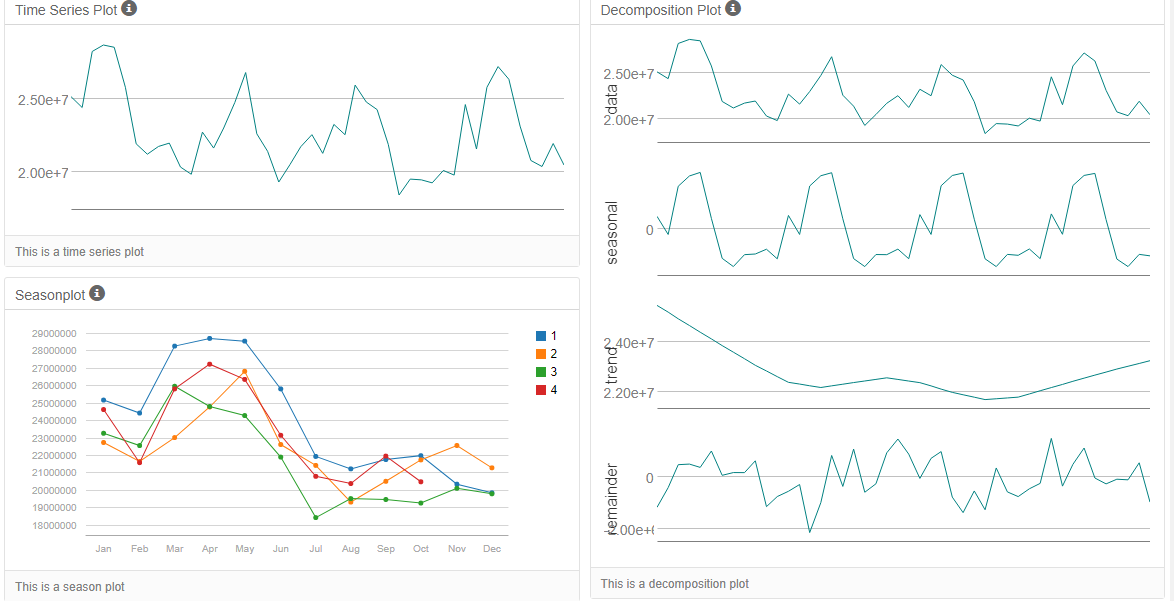
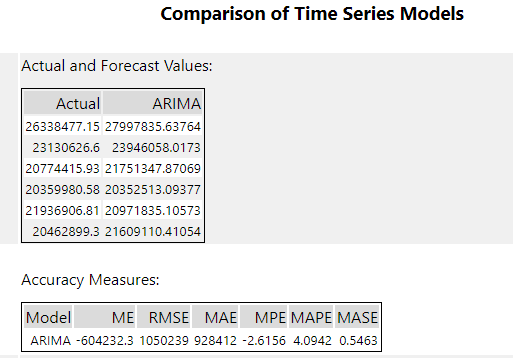
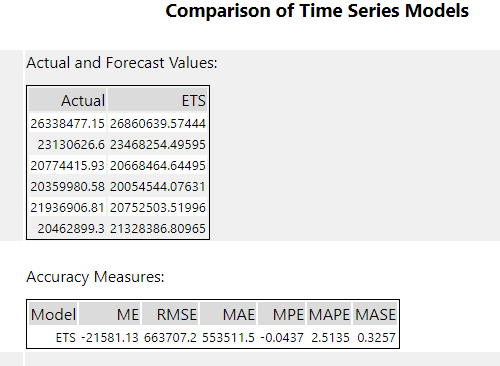


Figure 3.1: Time Series Plot/Decomposition Plot of historical monthly sales

After comparing the results against the holdout sample, the ETS performs better against the ARIMA model in term of accuracy. The RMSE and MASE is also lower than that of the ARIMA



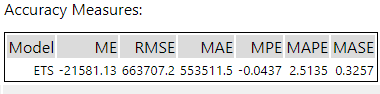




Figure 3.2: ETS and ARIMA Model comparison

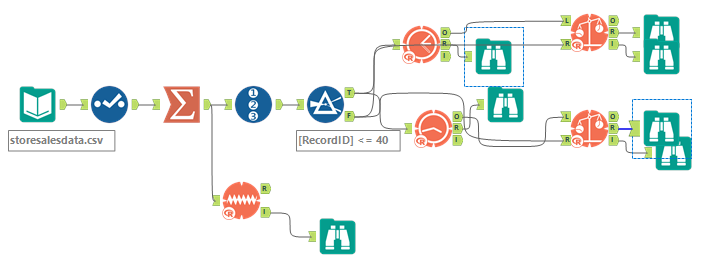


Figure 3.2: ETS and ARIMA Model comparison workflow

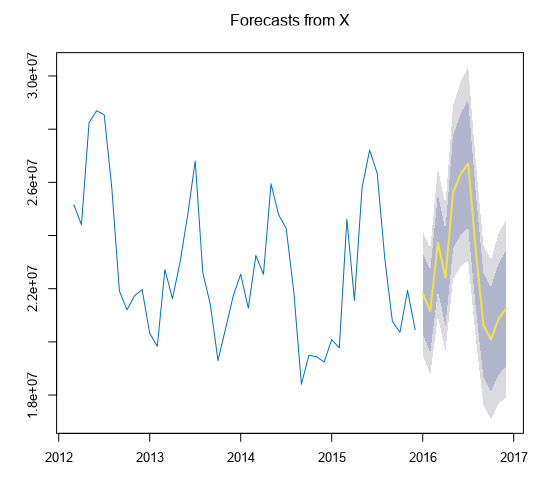


Figure 3.4: ETS forecast plot

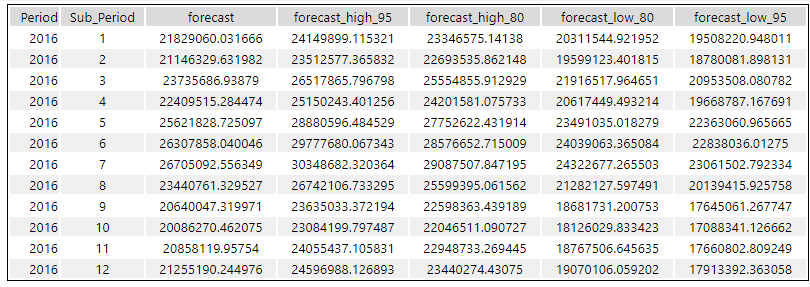


Figure 3.5: ETS forecast table for existing data

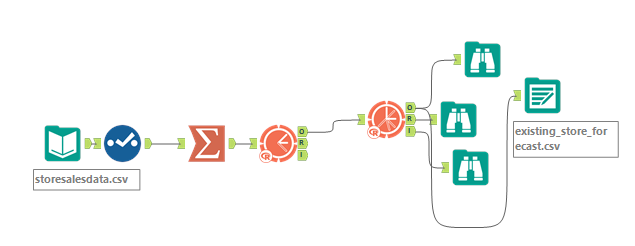


Figure 3.6: ETS forecast workflow for existing data

1. **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.**

|  |  |  |
| --- | --- | --- |
| **Month** | **New Stores** | **Existing Stores** |
| Jan 16 | 2,491,319 | 21,829,060 |
| Feb 16 | 2,408,385 | 21,146,330 |
| Mar 16 | 2,833,157 | 23,735,687 |
| Apr 16 | 2,679,433 | 22,409,515 |
| May 16 | 3,054,886 | 25,621,829 |
| Jun 16 | 3,106,152 | 26,307,858 |
| July 16 | 3,132,699 | 26,705,093 |
| Aug 16 | 2,776,154 | 23,440,761 |
| Sep 16 | 2,451,566 | 20,640,047 |
| Oct 16 | 2,401,772 | 20,086,270 |
| Nov 16 | 2,477,302 | 20,858,120 |
| Dec 16 | 2,452,170 | 21,255,190 |
| **Total Annual Sales** | **$32,264,995** | **$274,035,760** |

## Table 3.1: Sales forecast for Existing and New Stores for the next 12 months

Tableau Visualization Link here:

https://public.tableau.com/app/profile/abdulwasiu.tiamiyu/viz/Clusterlocation/Forecast?publish=yes

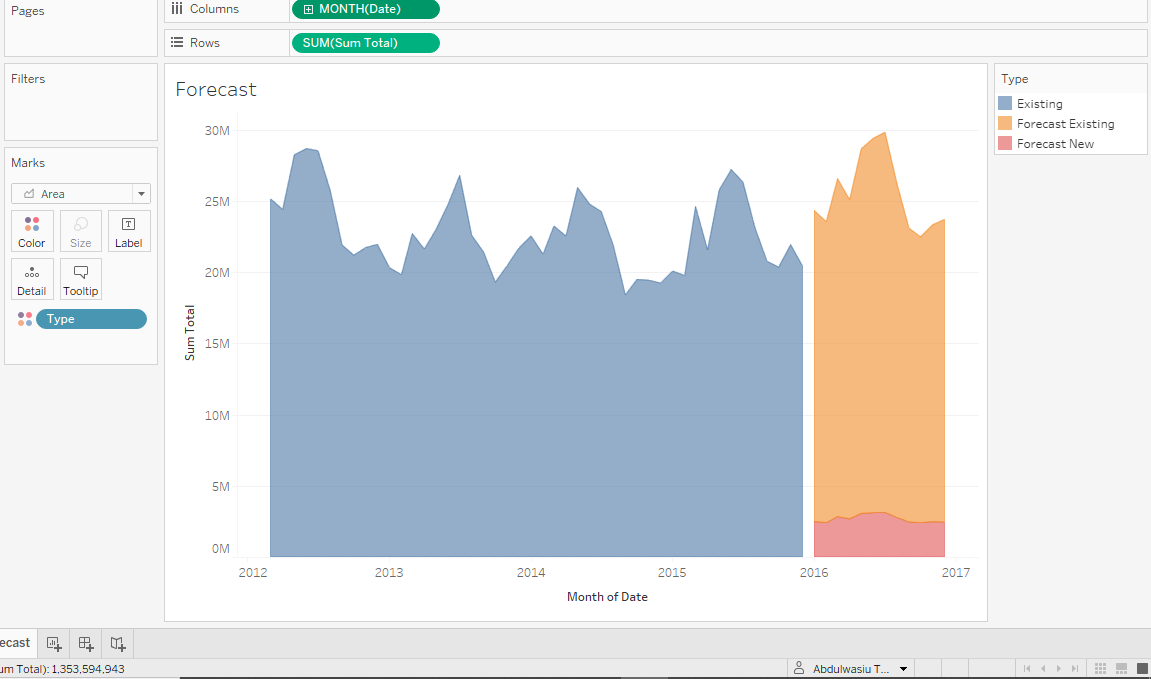


Figure 3.7: Historical and forecast sales for existing and new stores from Mar-12 to Dec-16

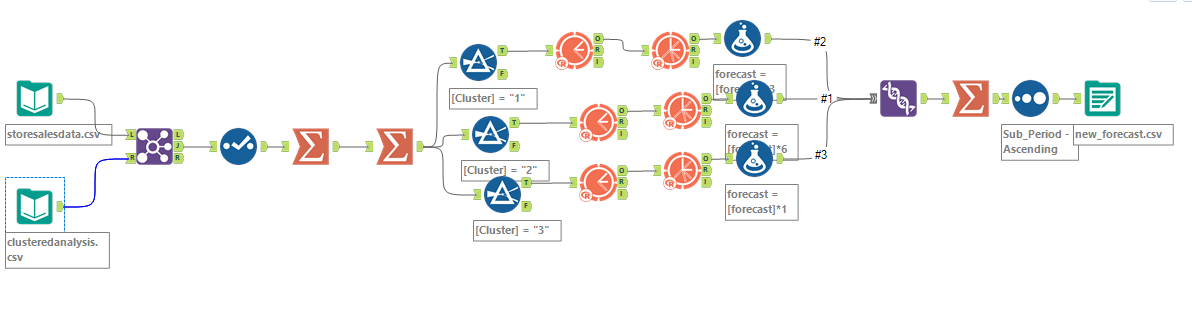


Figure 3.8: ETS forecast workflow for new data