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?

**Answer:**

The optimal number of clusters is 3. This was obtained by using Adjusted Rand indices and Calinski - Haragasz indices. The visualization below indicates that 3 is the best cluster number as it has the highest Adjusted Rand indices and Calinski - Haragasz indices



1. How many stores fall into each store format?

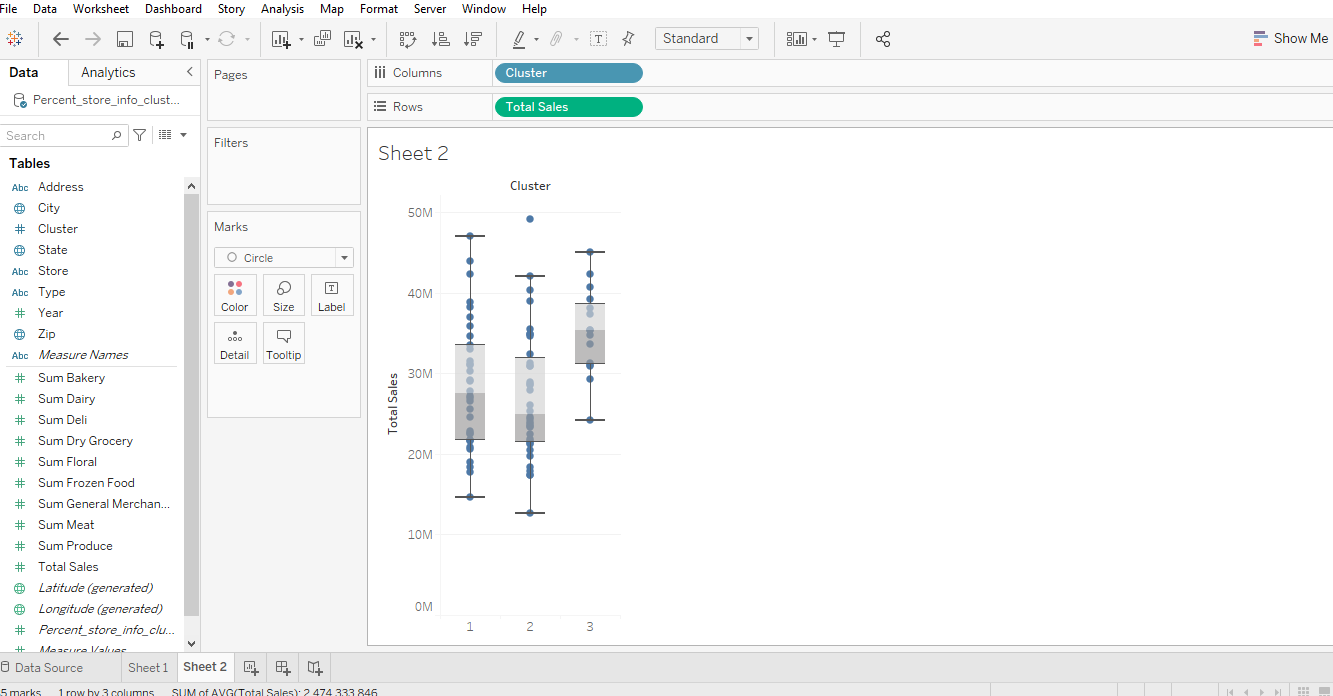
**Answer**

|  |  |
| --- | --- |
| Cluster (Format) | Number of stores |
| 1 | 25 |
| 2 | 35 |
| 3 | 25 |

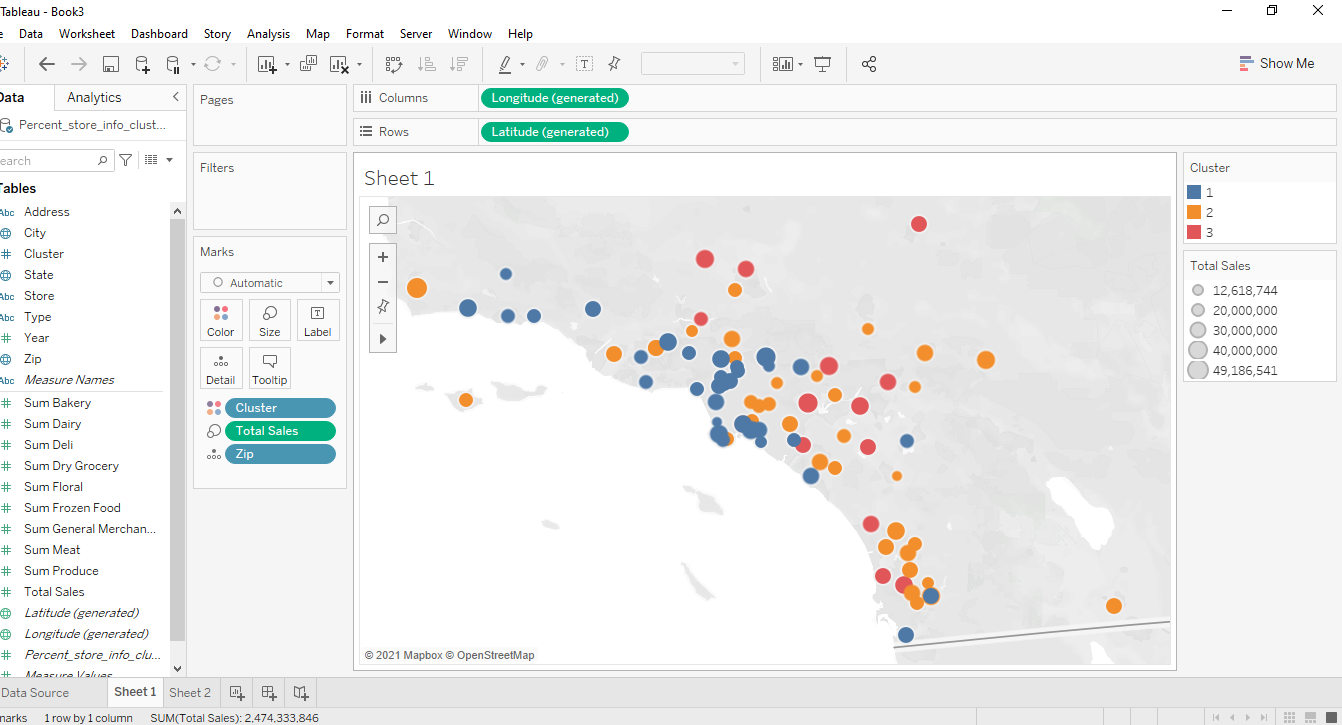
1. Based on the results of the clustering model, what is one way that the clusters differ from one another?

**Answer**:

The clusters differ from each other with respect to total sales. Stores in cluster 3 have bigger sales followed by stores in cluster 1. Comparatively, stores in cluster 1 have smaller total sales as depicted in the visualization below.



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.)

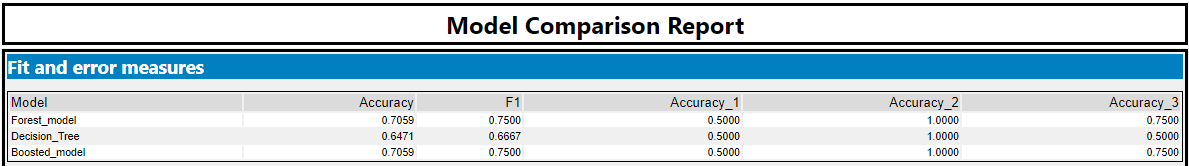
**Answer**

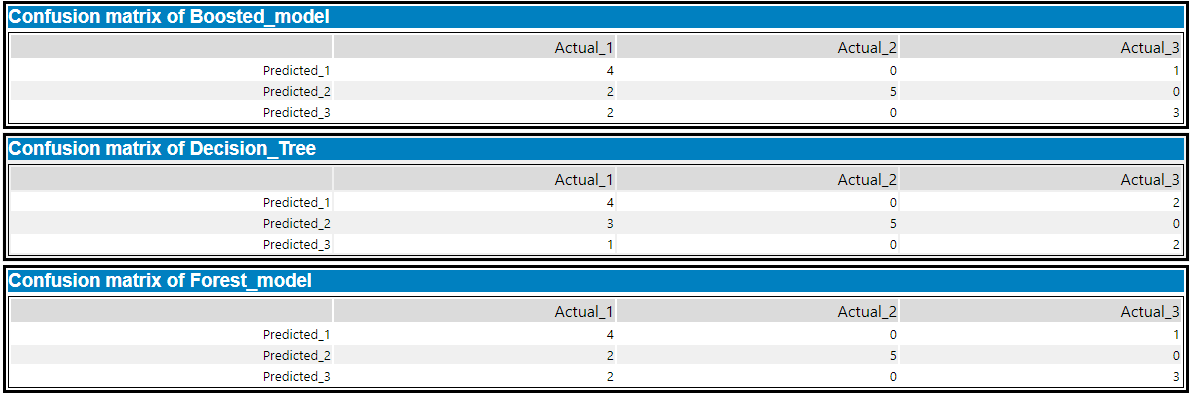
Multi-classification methodology was used with Boosted model to predict the best store format.

Multi-classification method was used because the target variable (store format) has more than two classes(categories) being 1,2,3 and there was also enough historical data to allow prediction.

Boosted model was chosen the multi-classification because it has high accuracy and F1 score. Since the Boosted model has the same performance as the Random Forest model due to the alteryx version used, the Boosted was chosen out personal preference over the Random Forest model.

The visualization below supports my opinion.





1. 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?

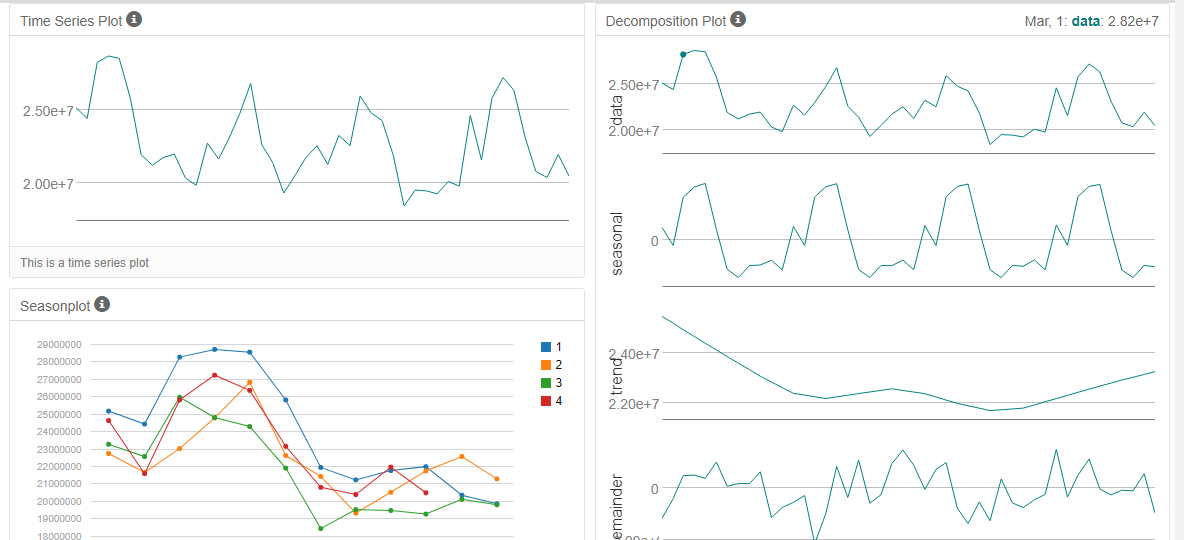
**Answer**

ETS (M, N, M) was used for each forecast.

I first run different ETS model ARIMA model taking into consideration the different time series components. The best ETS model was ETS (M, N, M) and the best ARIMA model was ARIMA (1,0,0) (1,1,0) [12].

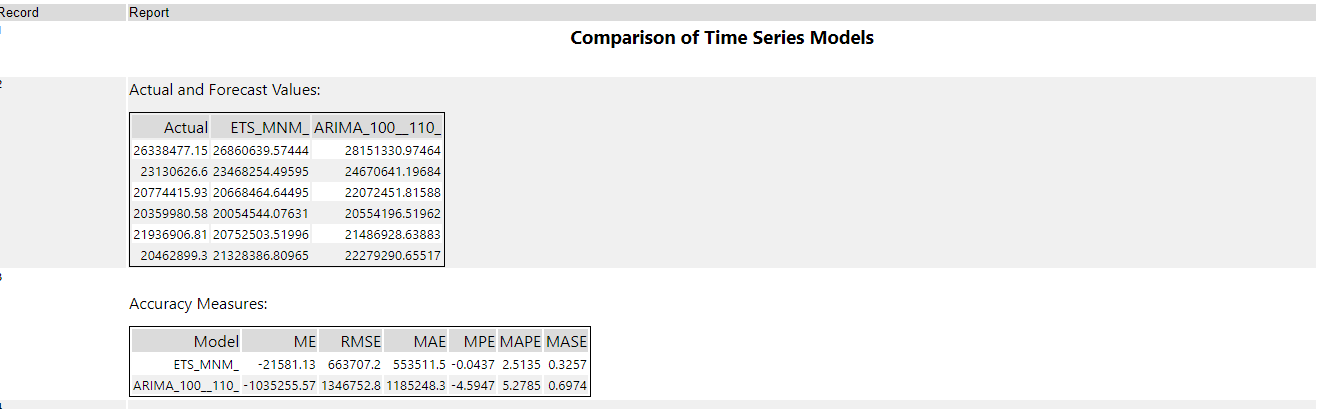
ETS (M, N, M) was chosen over any other ETS model based on the decomposition plot.

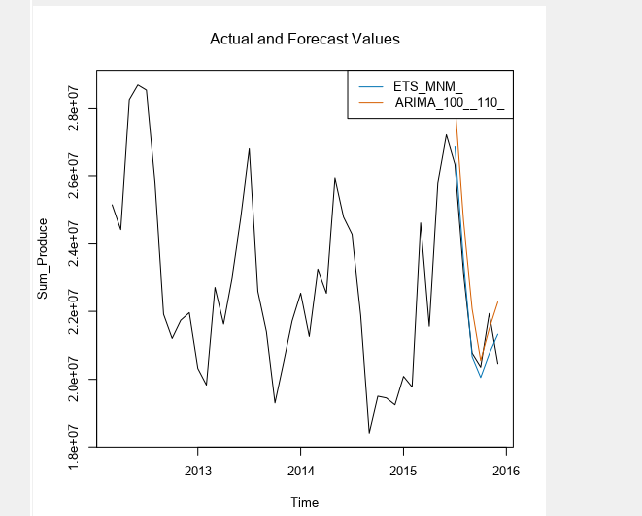
Considering the errors plot and the seasonal plot, their variations are not constant over time hence M was selected for both. N is selected for trend because the trend plot goes down then up indicating that there is no trend.



**Choosing ETS (M, N, M) over ARIMA (1,0,0) (1,1,0) [12**].

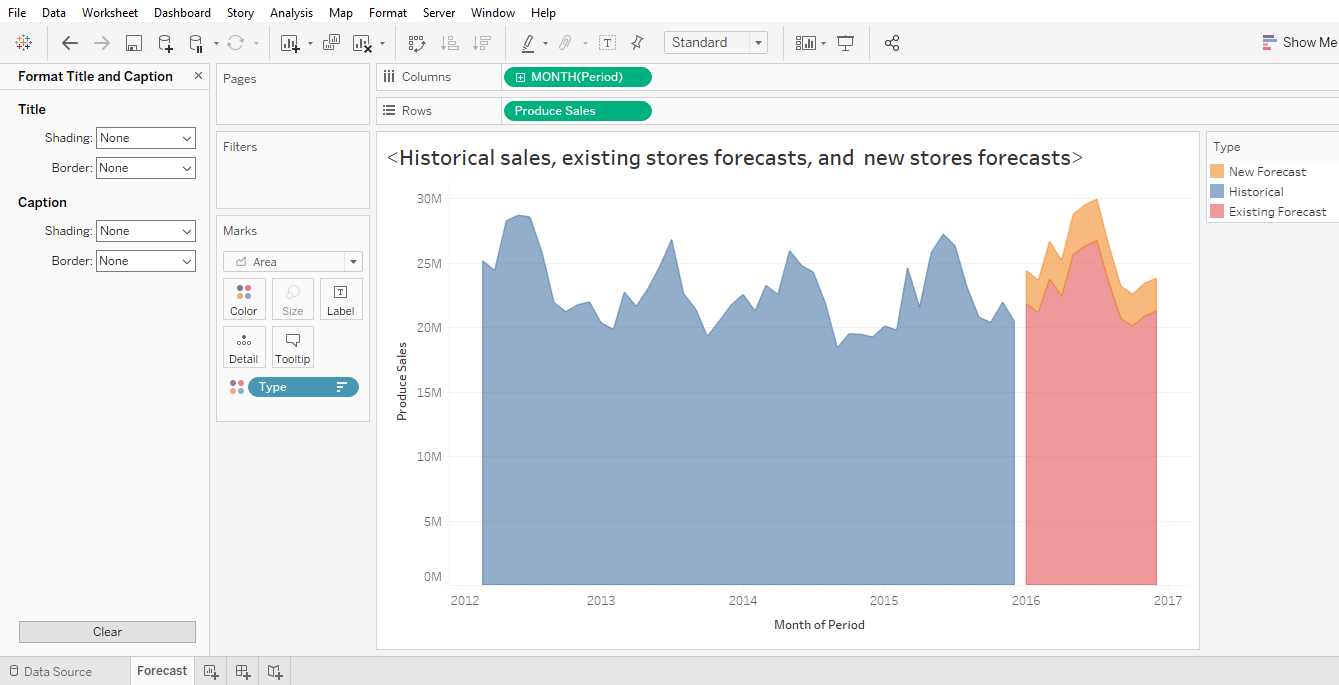
Comparatively, ETS (M, N, M) model out-performed the ARIMA (1,0,0) (1,1,0) [12] as it has the lowers RMSE and MASE and low forecast errors as shown in the visualizations below.





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.

|  |  |  |  |
| --- | --- | --- | --- |
| Period | Month | Existing Stores Forecast | New Stores Forecast |
| 2016 | January | 21829060.03 | 2563357.91 |
| 2016 | February | 21146329.63 | 2483924.728 |
| 2016 | March | 23735686.94 | 2910944.146 |
| 2016 | April | 22409515.28 | 2764881.87 |
| 2016 | May | 25621828.73 | 3141305.867 |
| 2016 | June | 26307858.04 | 3195054.204 |
| 2016 | July | 26705092.56 | 3212390.954 |
| 2016 | August | 23440761.33 | 2852385.769 |
| 2016 | September | 20640047.32 | 2521697.187 |
| 2016 | October | 20086270.46 | 2466750.894 |
| 2016 | November | 20858119.96 | 2557744.588 |
| 2016 | December | 21255190.24 | 2530510.805 |

S