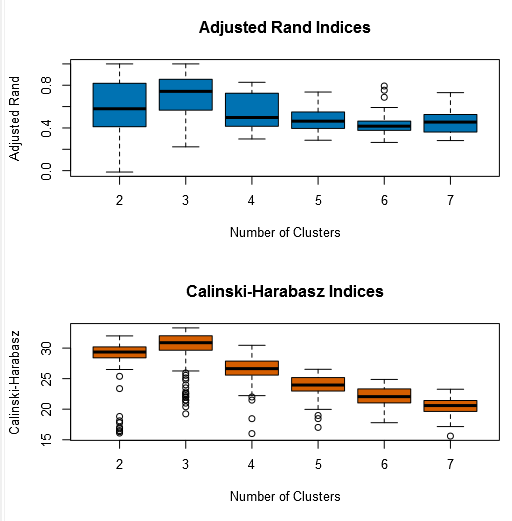
Project: Predictive Analytics Capstone

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

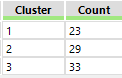
1. What is the optimal number of store formats? How did you arrive at that number?

The optimal number of store formats is three. I used a K-Means Clustering method. The results, depicted below, show that three is the ideal number of clusters.



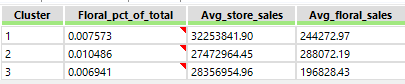
1. How many stores fall into each store format?

The number of stores per format are displayed below.

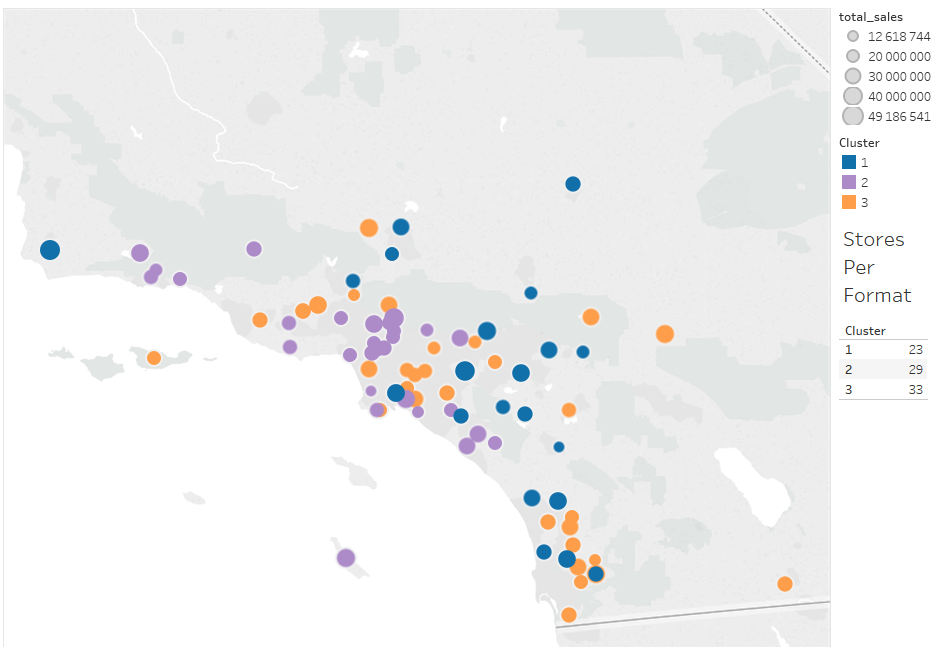


1. Based on the results of the clustering model, what is one way that the clusters differ from one another?

As depicted below, the clusters differ in Floral sales percent of total. After multiplying by the average store sales, the Floral dollar sales differ widely among the clusters.



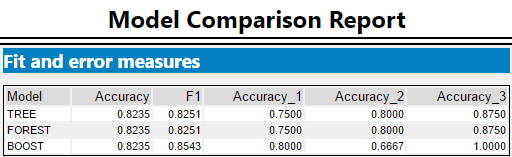
1. Please provide a map created in Tableau that shows the location of the existing stores, uses color to show cluster, and size to show total sales. Make sure to include a legend! Feel free to simply copy and paste the map into the submission template.



## 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 used the Boosted Model. As depicted in the screenshot below, while the accuracy of the three models is equivalent, the F1 measure of the Boosted Model is superior.



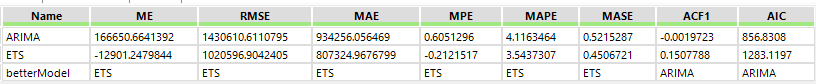
1. What format do each of the 10 new stores fall into? Please fill in the table below.

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

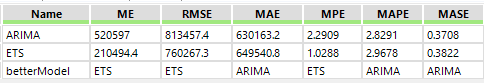
## Task 3: Predicting Produce Sales

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?

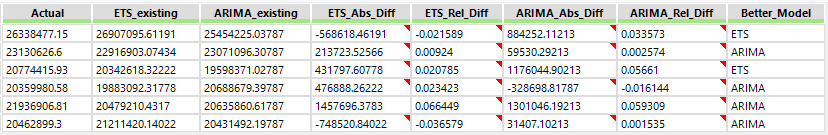
I chose the ETS(M,N,M) model because most of the in-sample error measurements, depicted below, are superior to the ARIMA model. The more important measures, RMSE and MASE, are smaller and therefore better in the ETS model.

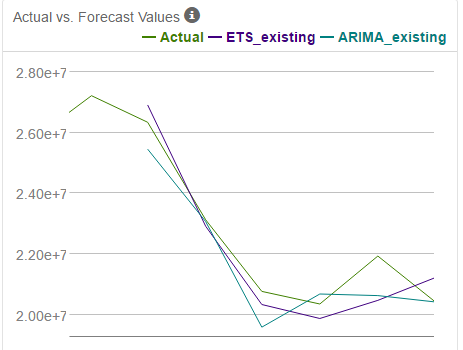


The accuracy measures back up the decision. Where the ETS model is better, it is much better, especially at the RMSE measurement. Where ARIMA is better, it is only slightly better, especially at the MASE measurement.



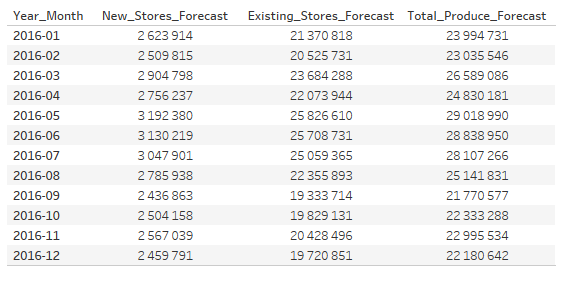
Finally, the forecasts show, against the holdout sample of six months, the ETS model makes extremely accurate predictions. This is shown in both the table and the chart below.

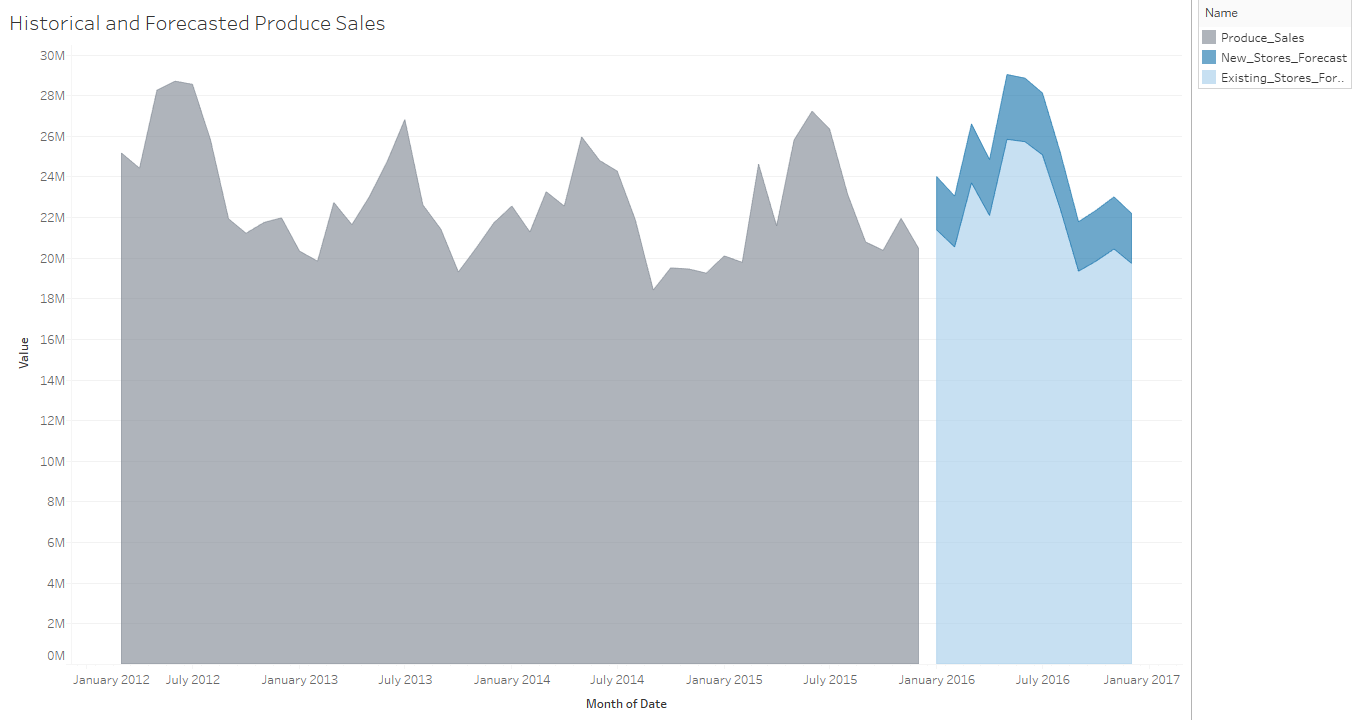




1. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

I have depicted below a table of produce forecasts for new stores, existing stores, and the total for each month in 2016. I have produced a stacked area chart below which shows historical produce sales with forecasts for existing and new store produce sales.





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.