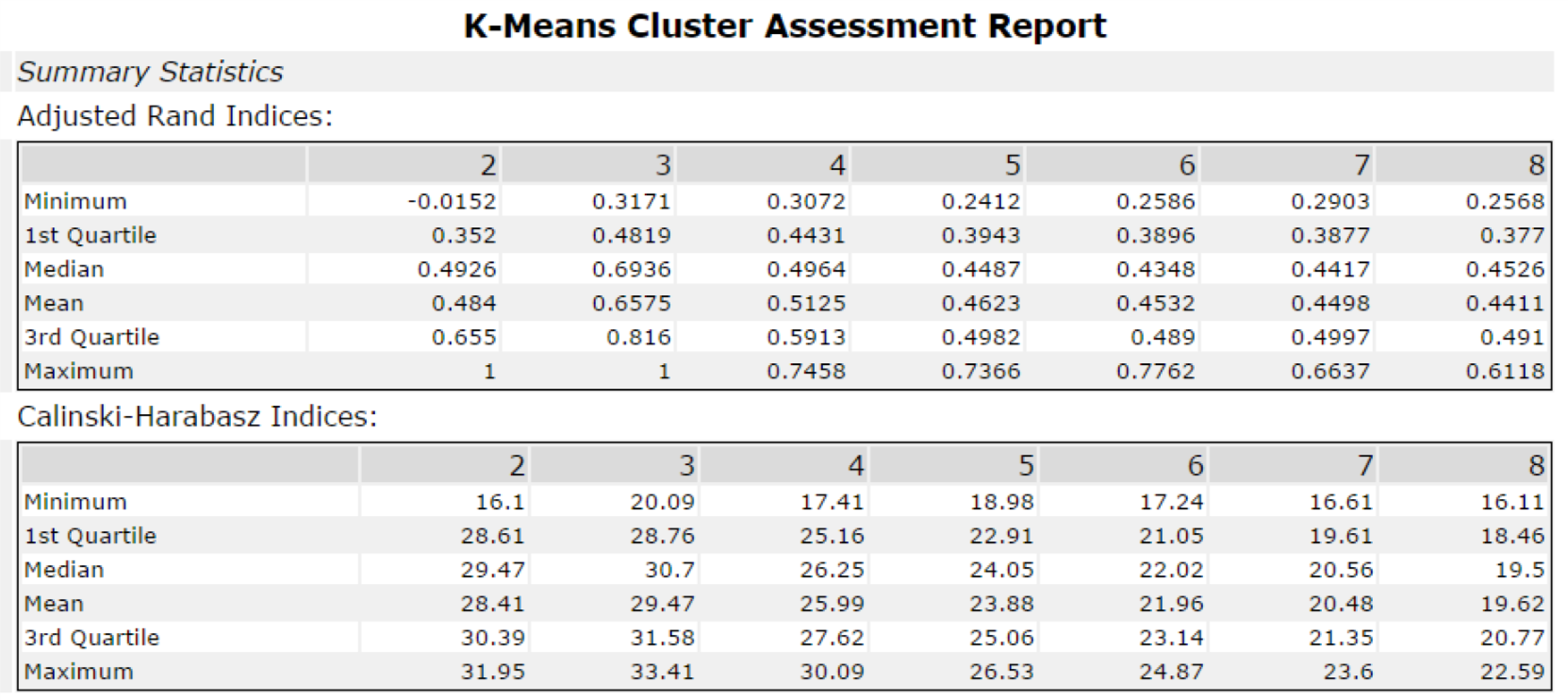
Project: Predictive Analytics Capstone

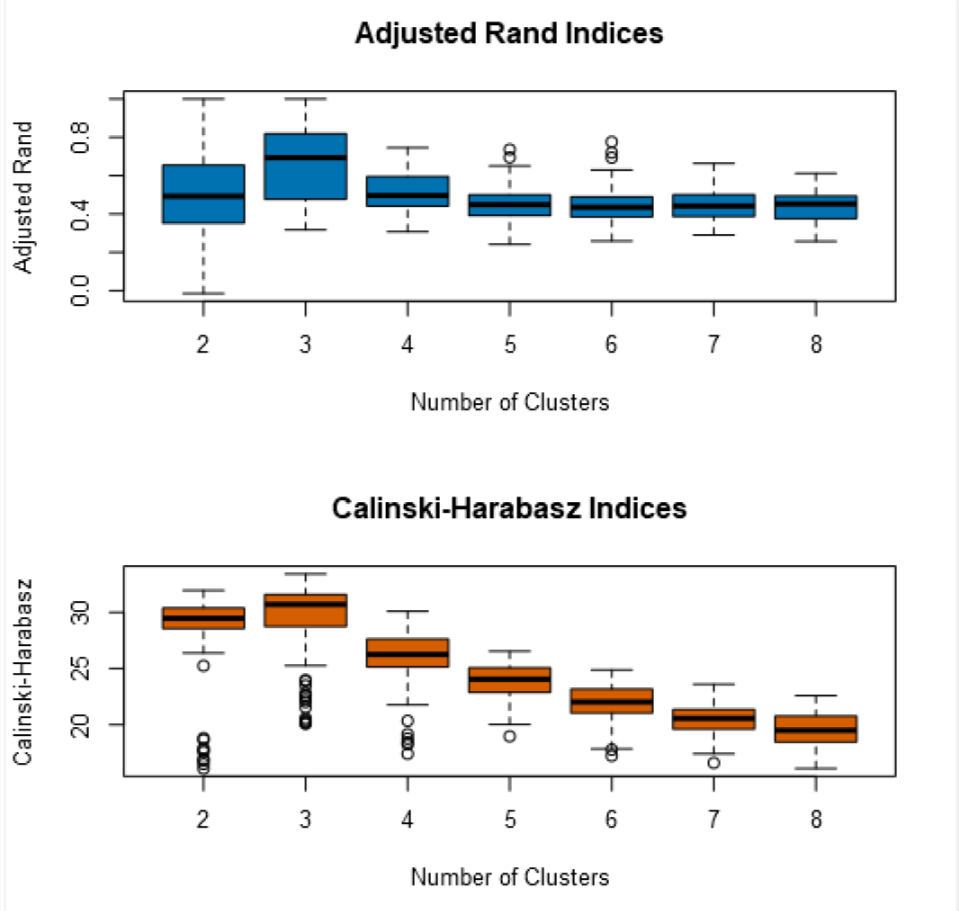
Complete each section. When you are ready, save your file as a PDF document and submit it here: <https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project>

## Task 1: Determine Store Formats for Existing Stores

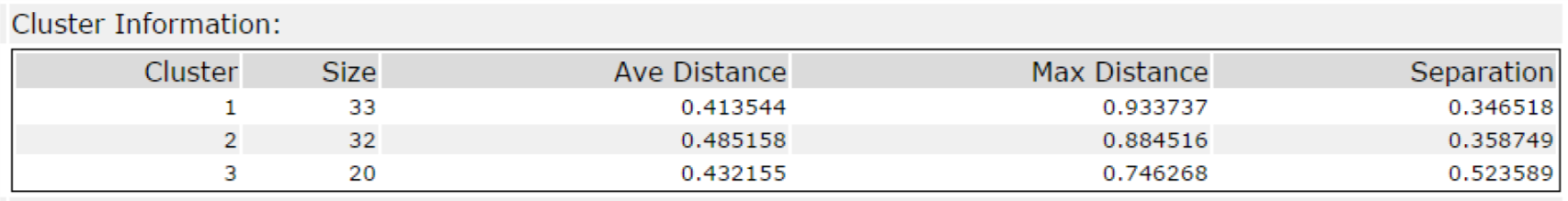
1. What is the optimal number of store formats? How did you arrive at that number?

The optimal number of store formats is 3. See the following analysis result from K-Centroids Diagnostics:



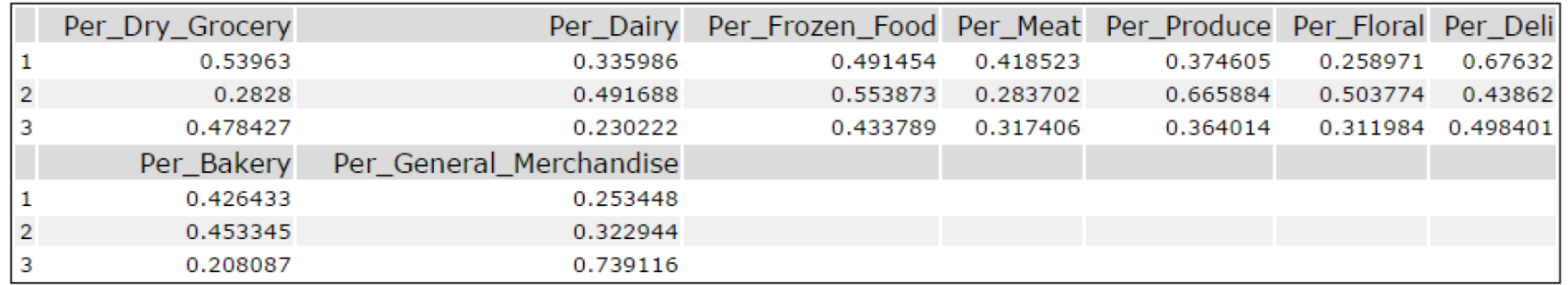


1. How many stores fall into each store format?



So sizes of the three clusters are 33, 32, and 20.

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



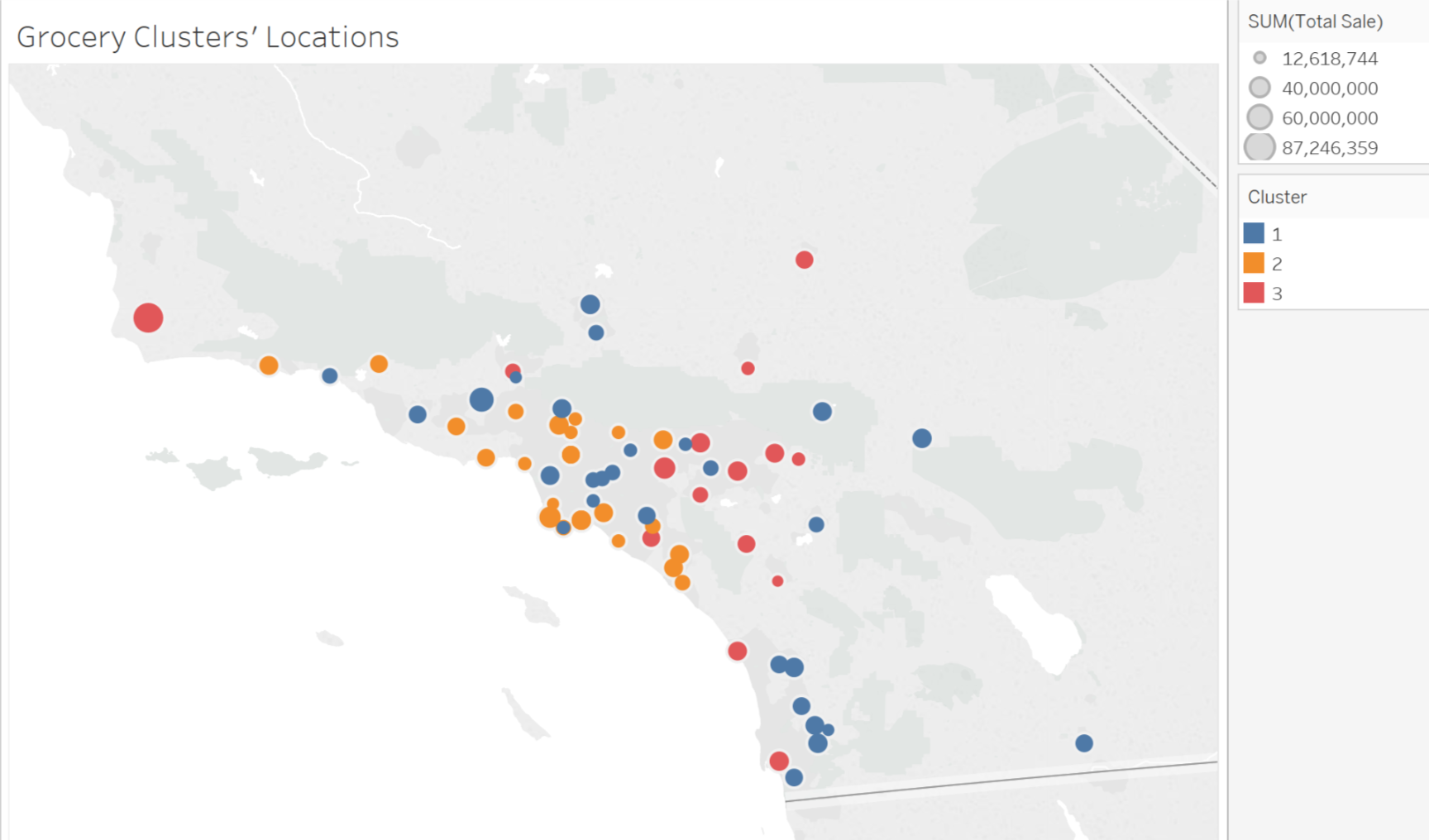
Cluster 2 differs from clusters 1 and 3 by focusing more on Dairy, Produce, and Floral.

Clusters 1 and 3 differ in Bakery and General\_Merchanidse.

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 Public file link:

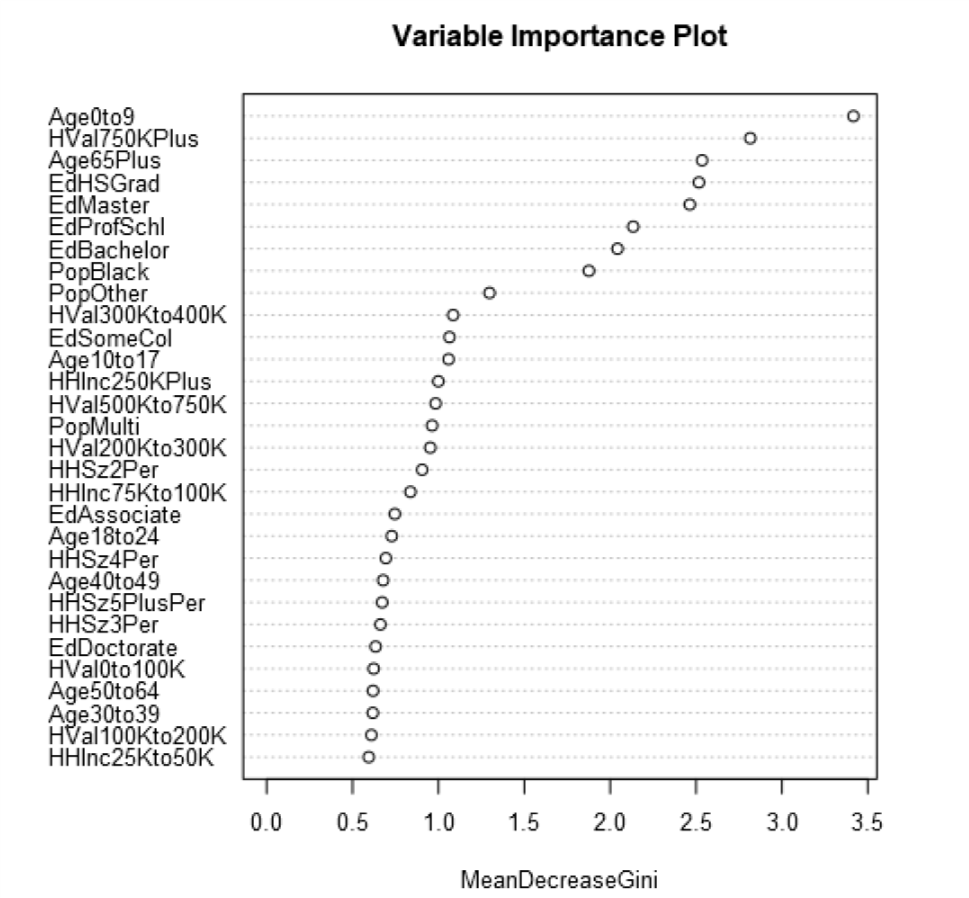
<https://public.tableau.com/profile/charlio#!/vizhome/grocery_geo/Sheet1?publish=yes>



## 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.)

We need to predict the cluster field for the new stores. Cluster includes 3 values. So this is a non-binary classification problem. Thus we shall use decision tree, forest or boosted models. After running all three models, it turns out forest and boosted have the same result on accuracy and confusion matrix, they both perform better than decision tree. So I used the forest model. The three most important variables are Age0to9, HVal750KPlus and Age65Plus.



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

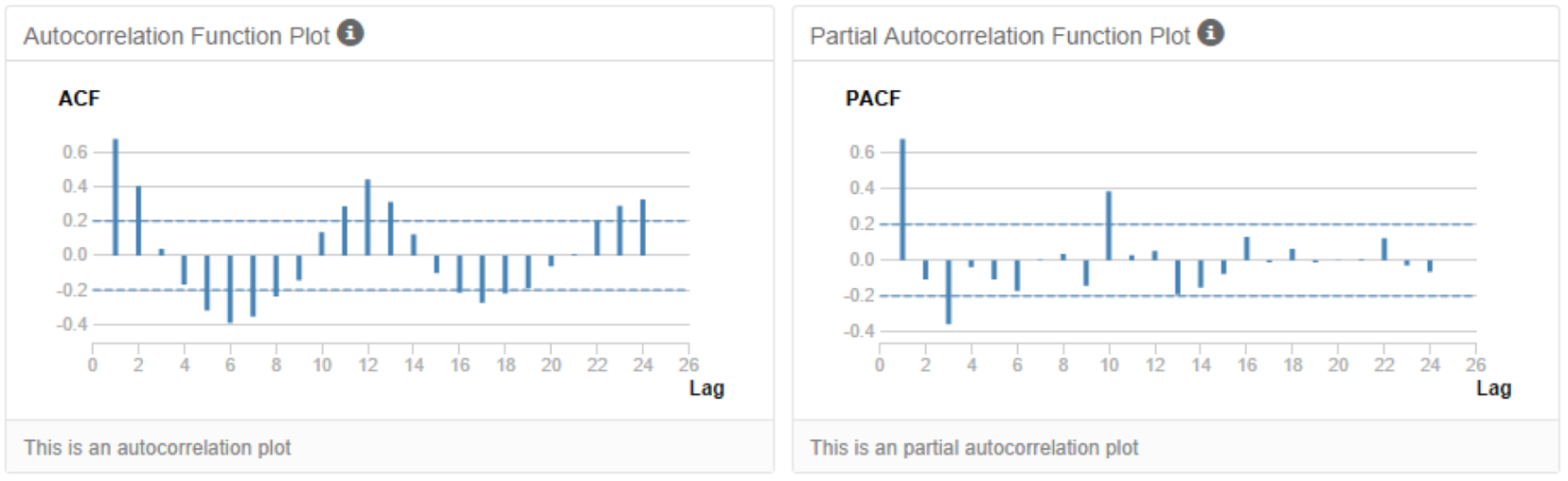
|  |  |
| --- | --- |
| Store Number | Segment |
| S0086 | 1 |
| S0087 | 2 |
| S0088 | 1 |
| S0089 | 2 |
| S0090 | 2 |
| S0091 | 3 |
| S0092 | 2 |
| S0093 | 3 |
| S0094 | 2 |
| S0095 | 2 |

## Task 3: Predicting Produce Sales

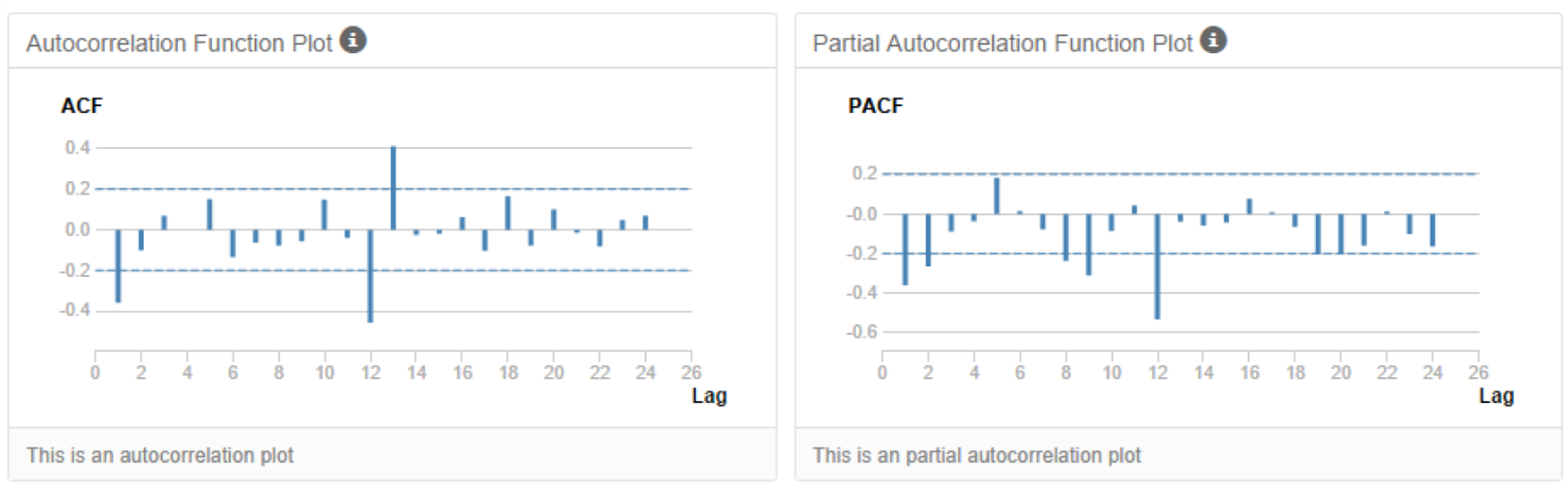
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

I used ARIMA(ar=0, i=1, ma=2)(P=0, D=1, Q=2).

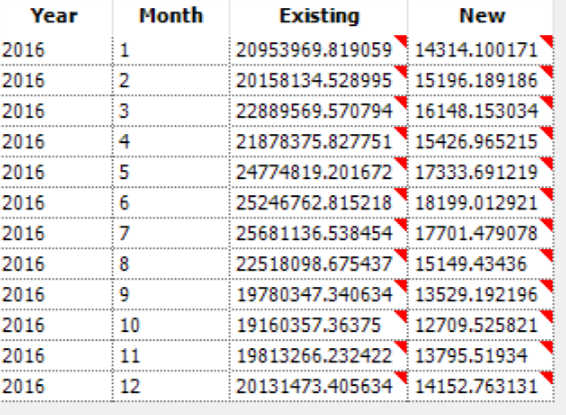
Below is the ACF and PACF for the raw total sales for all existing stores

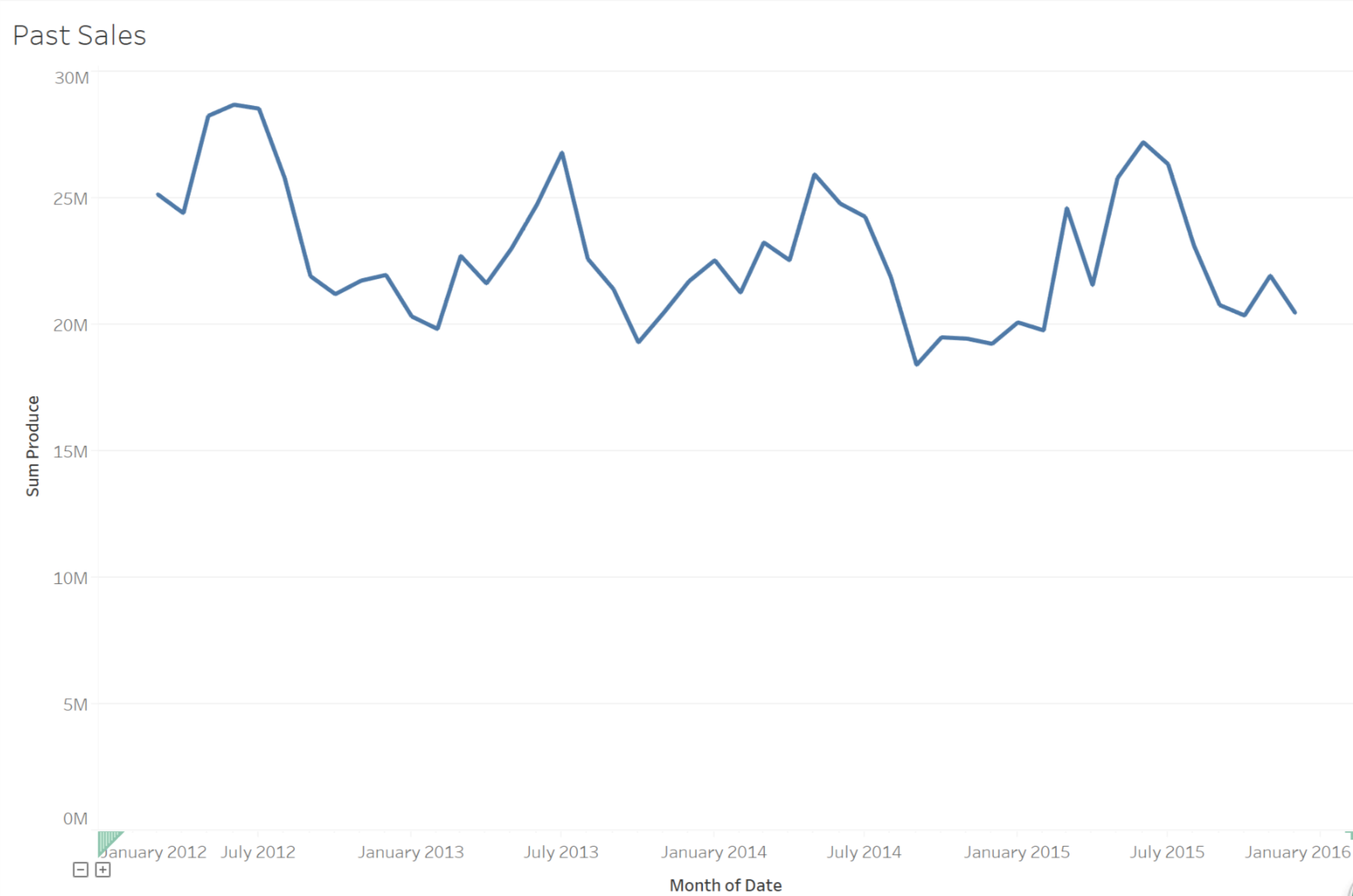


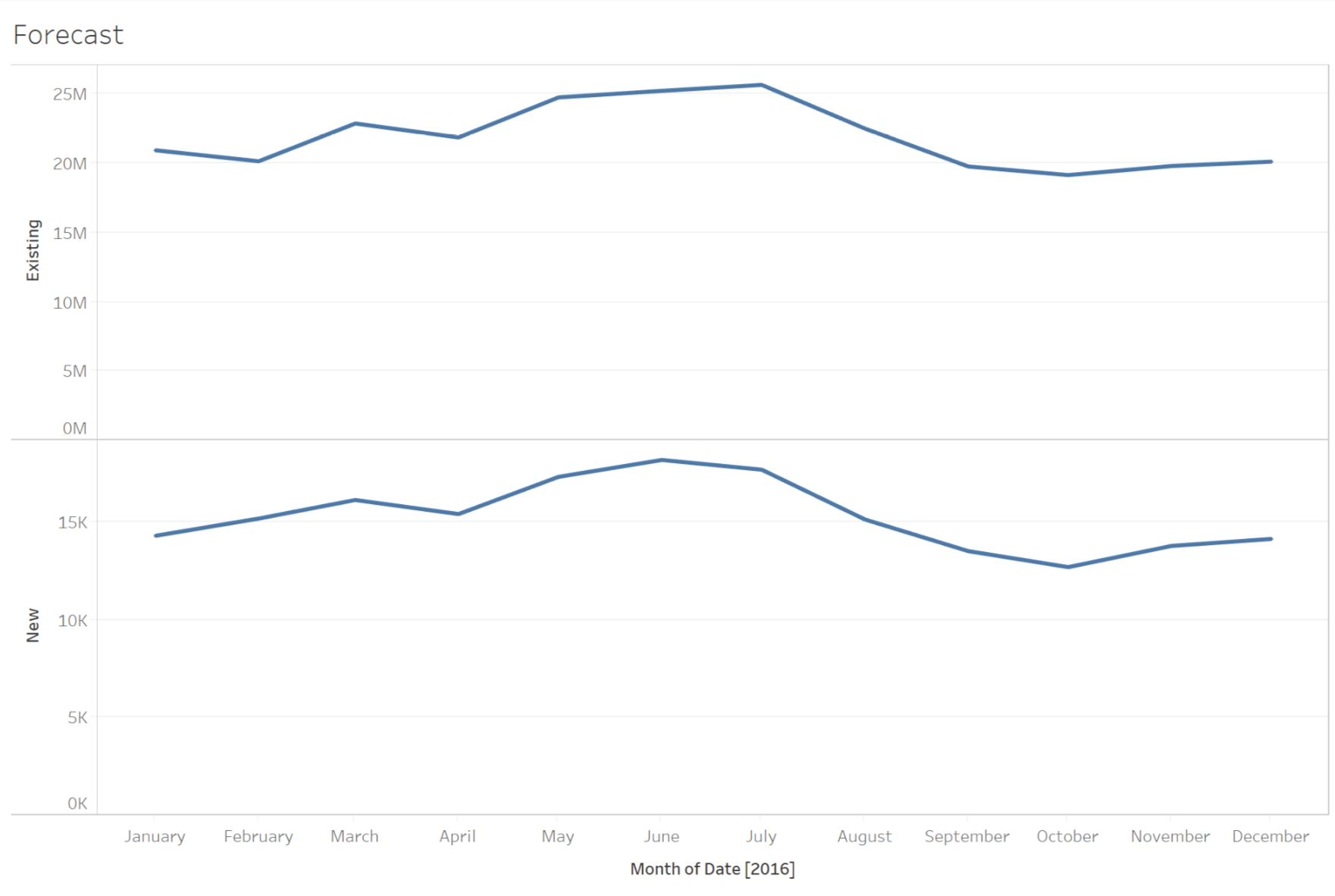
Below is the corresponding seasonal second difference plot:



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.







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.