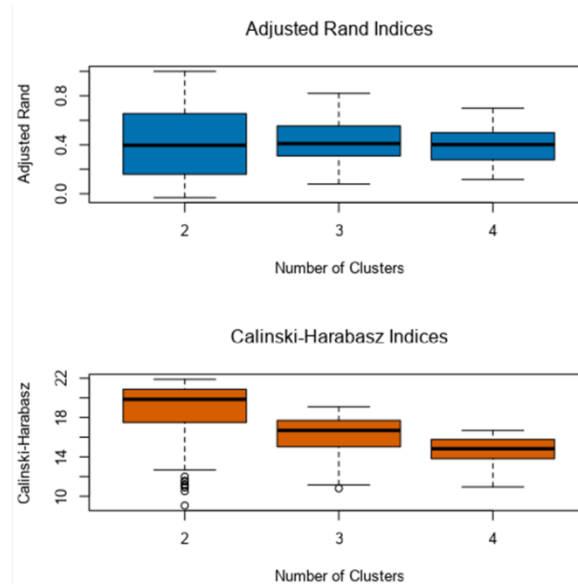# Project: Predictive Analytics Capstone

## Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?
   3 store formats, used clustering diagnostic tool to predict the optimal number of clusters with the following results of the indices:



2. How many stores fall into each store format?
   Store Format 01: 23 stores.
   Store Format 02: 29 stores.
   Store Format 03: 33 stores.

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Total sales percent is the one way to differentiate between each cluster. For example: Cluster 3 has the highest total sales for most categories, then cluster 2 and 1. See the following chart for more details about the categories for each cluster.
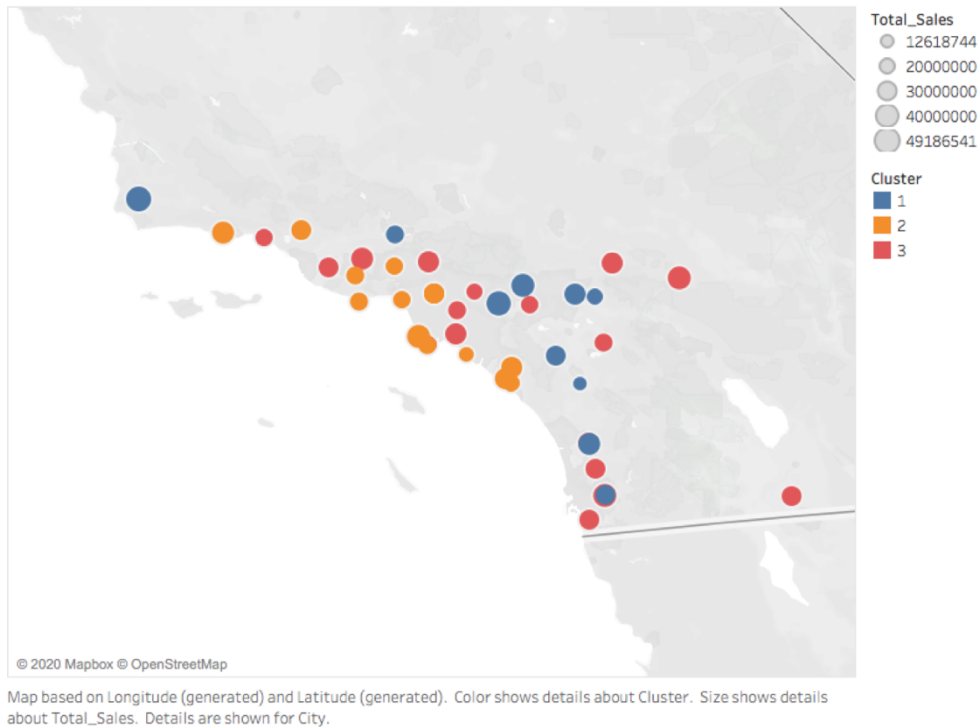
Differences between store formats
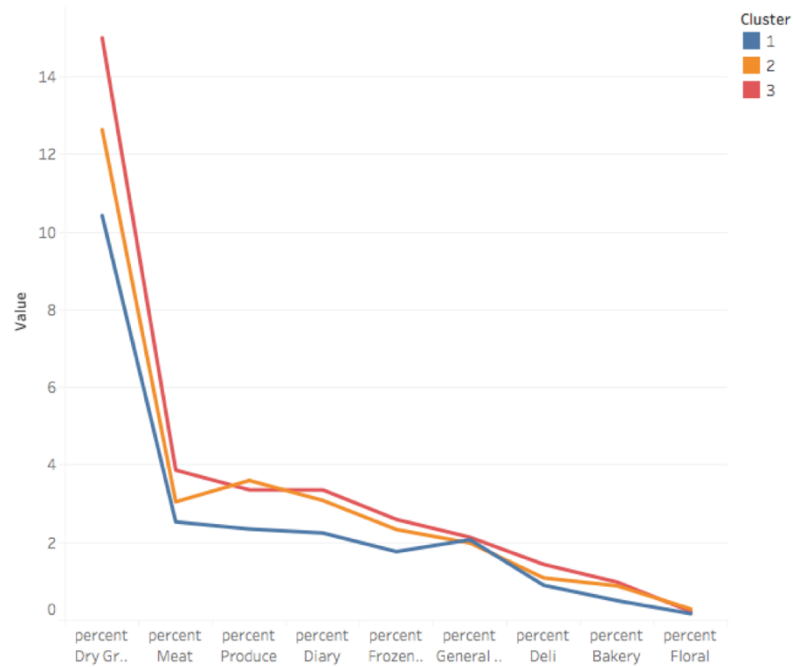


Percent Dry Grocery, percent Meat, percent Produce, percent Diary, percent Frozen Food, percent General Merchandise, percent Deli, percent Bakery and percent Floral. Color shows details about Cluster.

4. 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.

Grocery Stores Clusters



Map based on Longitude (generated) and Latitude (generated). Color shows details about Cluster. Size shows details about Total_Sales. Details are shown for City.

## 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?
   Used Boosted Model to predict the store format, as it has highest F1 value (0.8889) compared to other models.

2. 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 |

# 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?

Used ETS(M,N,M) for forecasting as it has better RMSE (663707. 2) than ARIMA (figure 01).

And used (M,N,M) for the model as the TS has seasonality that multiplicative and has no trend (figure 02).

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| ARIMA | -604232.29 | 1050239.2 | 928412 | -2.6156 | 4.0942 | 0.5463 |
| ETS | -21581.13 | 663707.2 | 553511.5 | -0.0437 | 2.5135 | 0.3257 |



Actual and Forecast Values



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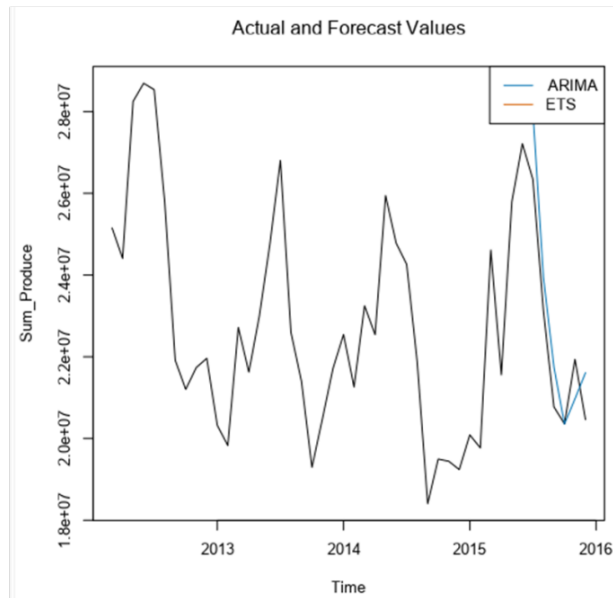
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

| Month | New Stores | Existing Stores |
|---|---|---|
| Jan-16 | 2588357 | 21146330 |
| Feb-16 | 2498567 | 23735687 |
| Mar-16 | 2919067 | 22409515 |
| Apr-16 | 2797280 | 25621829 |
| May-16 | 3163765 | 26307858 |
| Jun-16 | 3202813 | 26705093 |
| Jul-16 | 3228212 | 23440761 |
| Aug-16 | 2868915 | 20640047 |
| Sep-16 | 2538372 | 20086270 |
| Oct-16 | 2485732 | 20858120 |
| Nov-16 | 2583448 | 21255190 |
| Dec-16 | 2562182 | 21146330 |



Produce Sales Forecast
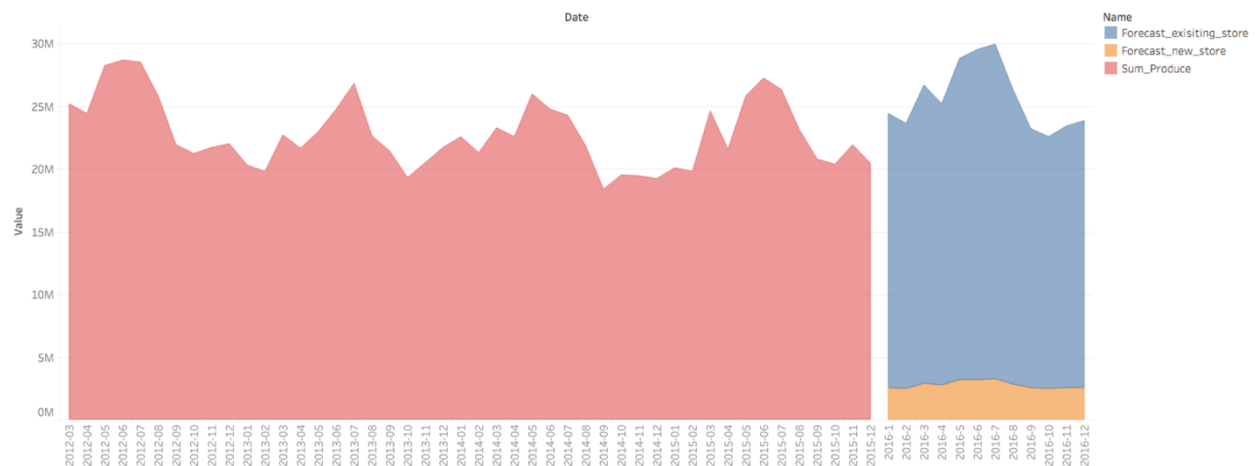
Sum of Value for each Date. Color shows details about Name.