**Project 7: Predictive Analytics Capstone**

**By: Huong Ivy Nguyen**

## **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. This value is chosen based on the results of the adjusted rand indices and the C-H indices obtained from running the Alteryx workflow 1 (Figures 1, 2).

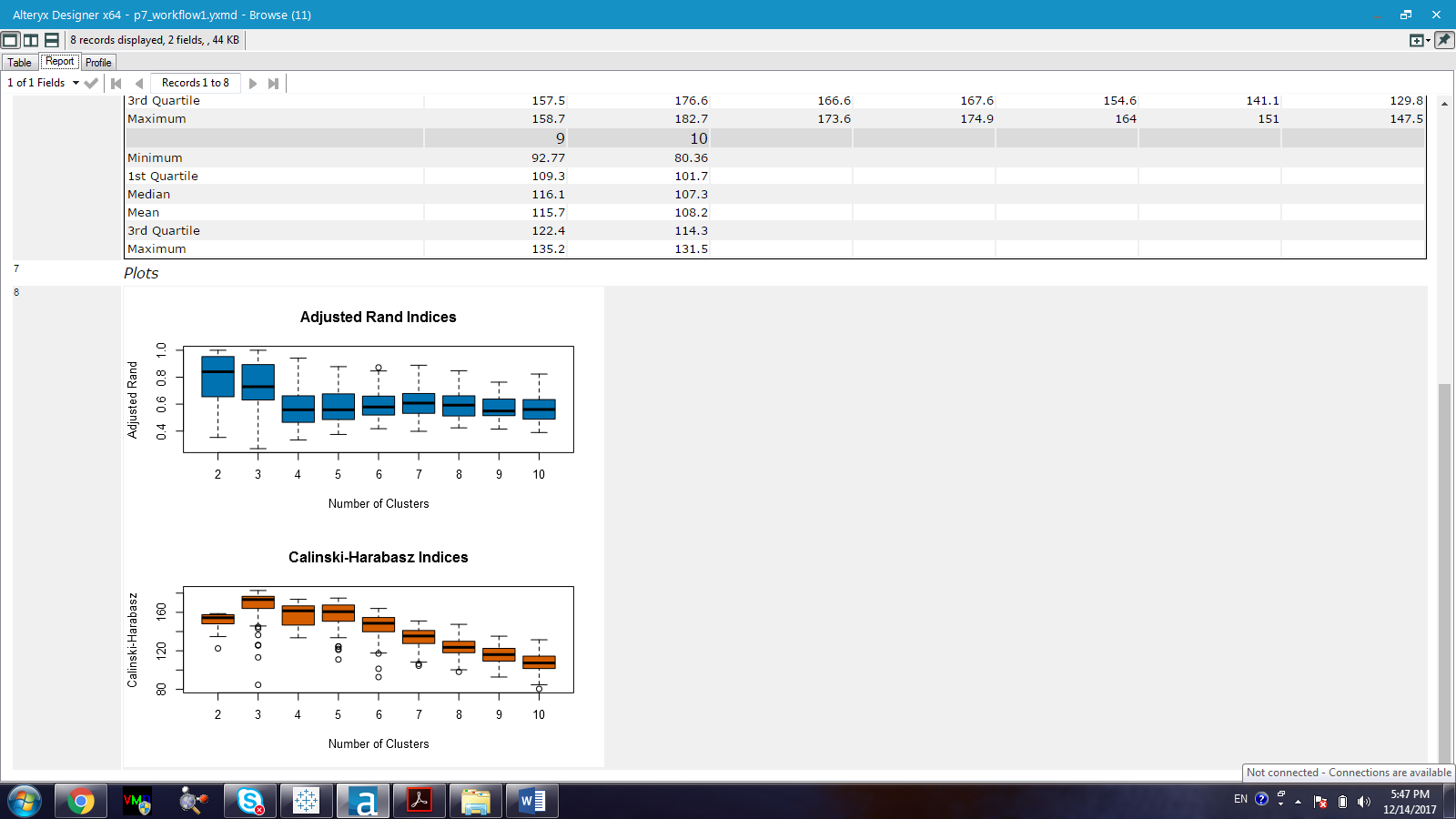


Figure 1. Adjusted rand indices and Calinski-Harabasz indices for the optimal number of clusters which should be used for the K-mean algorithm.

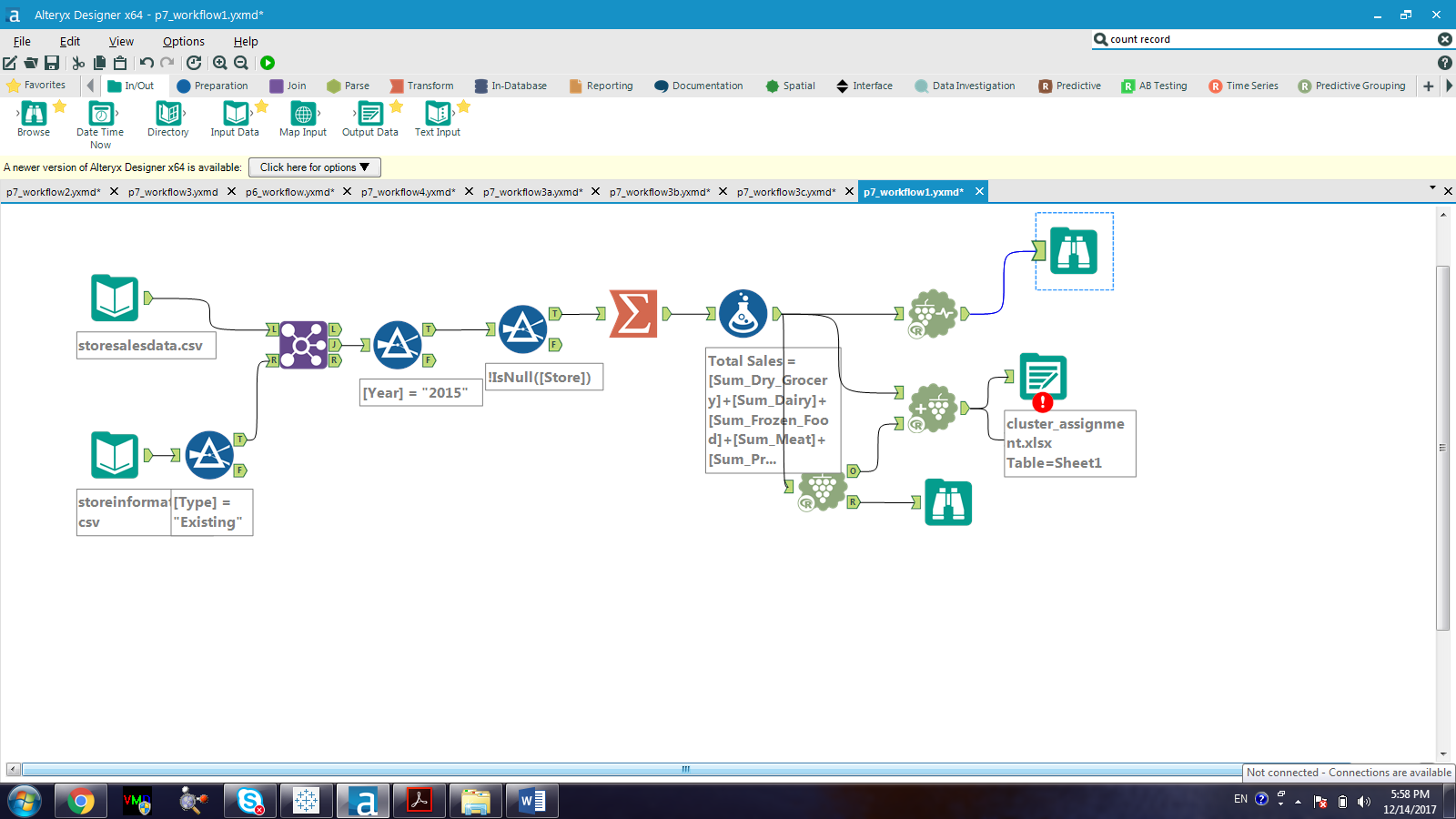


Figure 2. Evaluating the number of clusters that should be used in the K-mean clustering

Based on Figure 1, the following features can be pointed out:

* According to the adjusted rand indices, cluster 2 and 3 have a higher median value than the other clusters do, they also have a wider range of interquartile range (IQR).
* According to the C-H indices, cluster 3 has the highest median value with a narrow IQR range.

By combining the features obtained from both the adjusted rand indices and the C-H indices, cluster 3 is chosen due to its median value from both indices and its narrow IQR from the C-H indices ‘report.

1. How many stores fall into each store format?

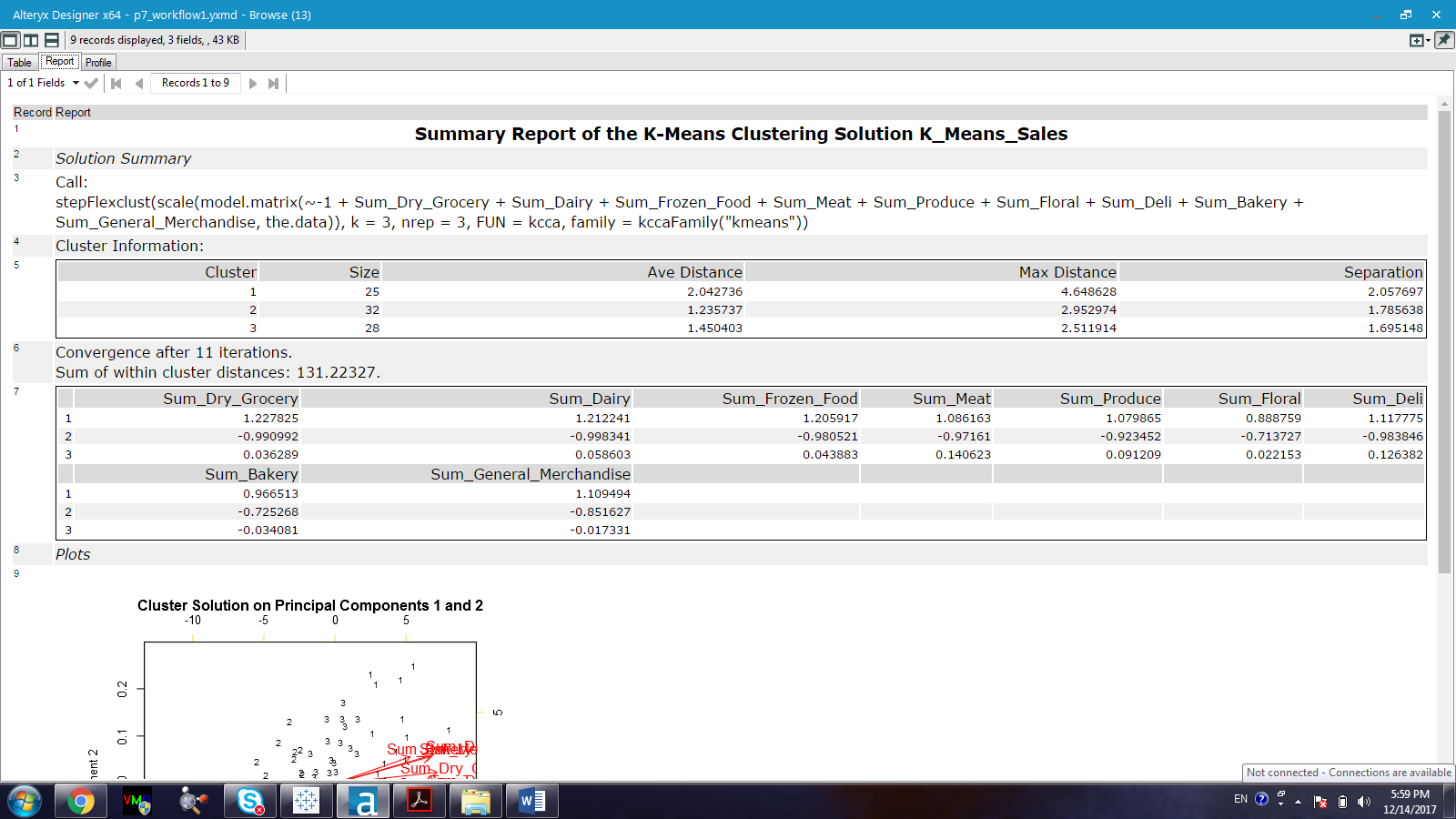


Figure 3. Report for the number of stores fall into each cluster

There are 25 stores in cluster 1, 32 stores in cluster 2, and 28 stores in cluster 3.

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

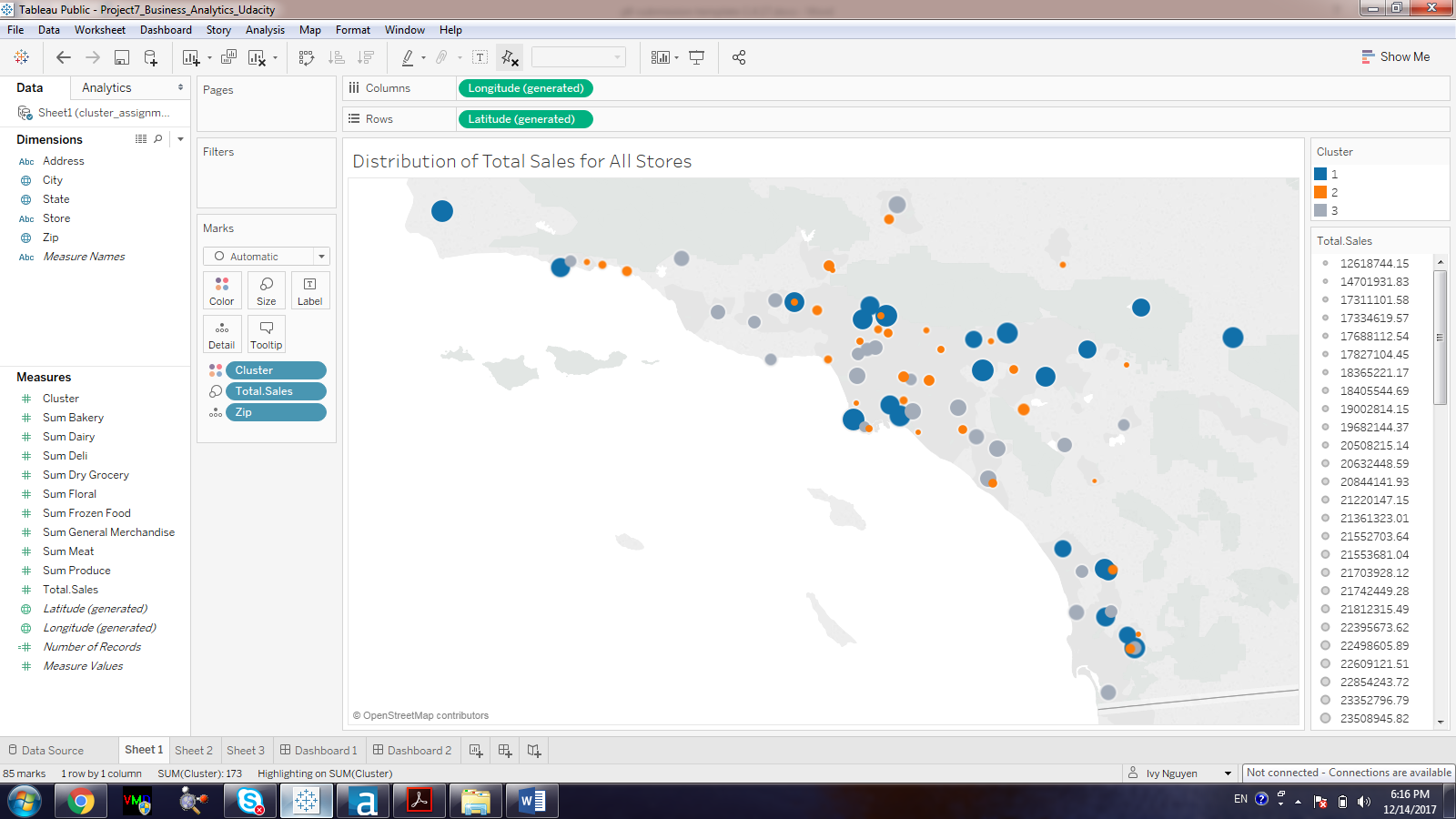


Figure 4. Total sales distribution of all the stores in all clusters where blue is cluster 1, orange is cluster 2, and gray is cluster 3. The size of the circle represents the amount of total sale for each store.

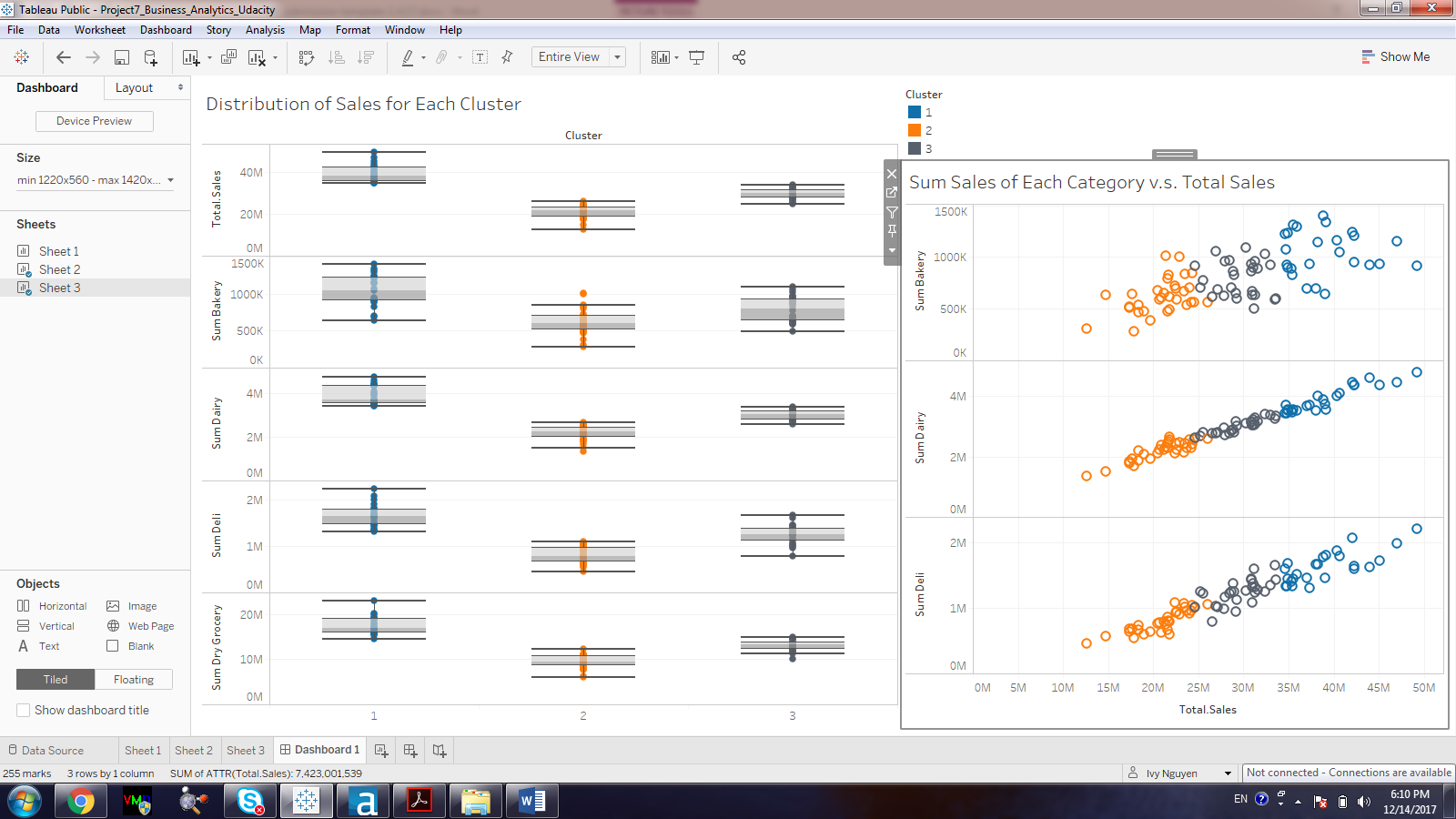


Figure 5. Left figure: whiskey-boxplot of the total sales, sum of bakery, sum of dairy, sum of deli, and sum of dry grocery for each cluster. Right figure: the relationship between total sales and sum sales for each category for each cluster. (Blue: cluster 1, Orange: cluster 2, Gray: cluster 3).

Based on Figure 4 and Figure 5, it appears that the total sales and the sum of sales for each category are lowest for cluster 2 in comparison to the other clusters. The sales for cluster 3 is lower than that of cluster 1. Therefore, the clusters differ from each other based on the number of sales including total sales and the sale for each category. Specifically, cluster 1 has a the highest number of sales followed by cluster 3 and then cluster 2.

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

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

## **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?

2. Please provide a Tableau Dashboard (saved as a Tableau Public file) that includes a table and a plot of the three monthly forecasts; one for existing, one for new, and one for all stores. Please name the tab in the Tableau file "Task 3".