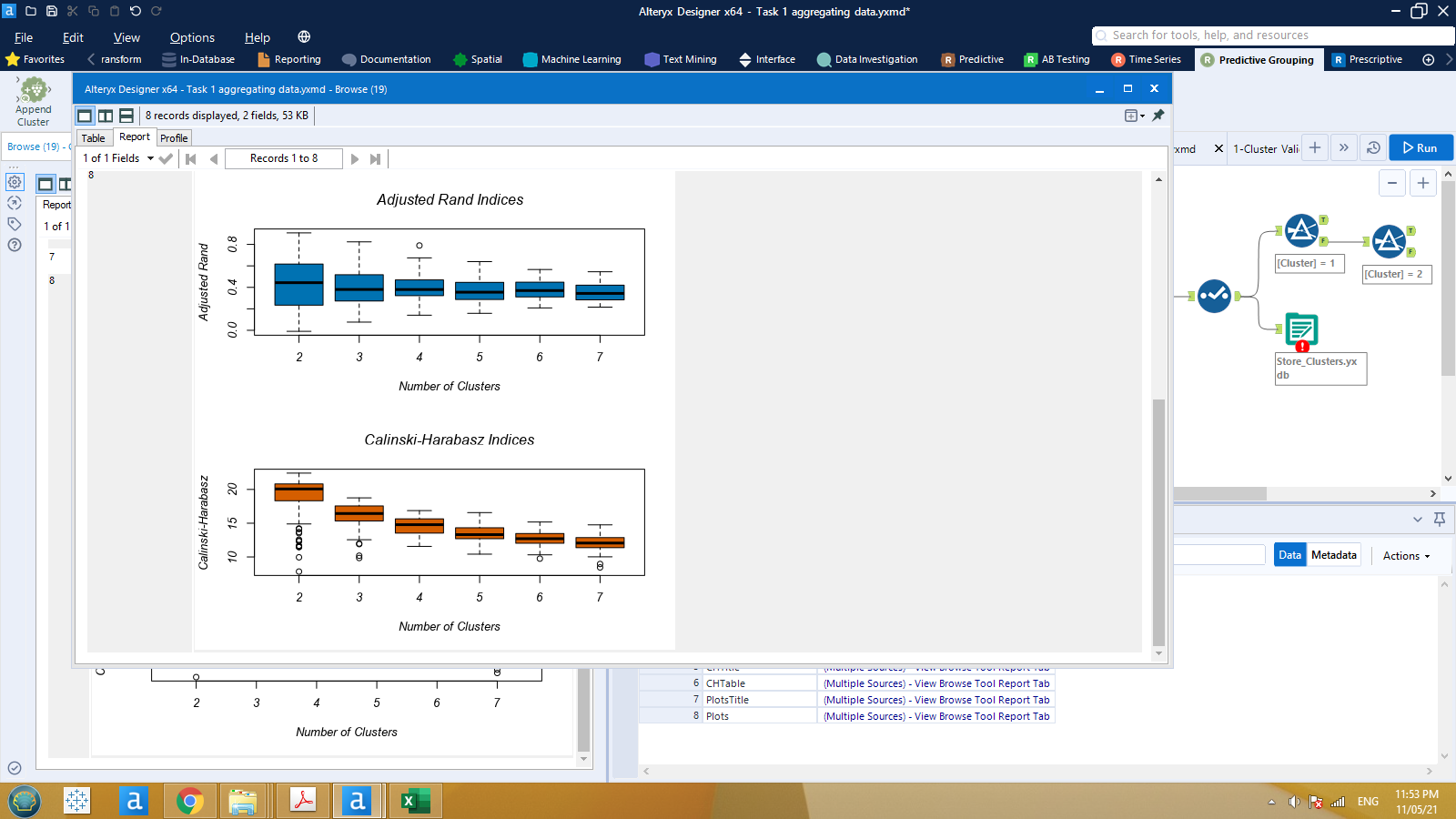
Project: Predictive Analytics Capstone

## Task 1: Determine Store Formats for Existing Stores

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

After aggregating the data, I’ve used the K-Centroids Diagnostics tool with K-Means method to choose the idle number of clusters, I’ve found that 2 and 3 Clusters having the highest values for Adjusted Rand and CH indices. But only 3 clusters have a compact data variability. So, I opted for 3 clusters.

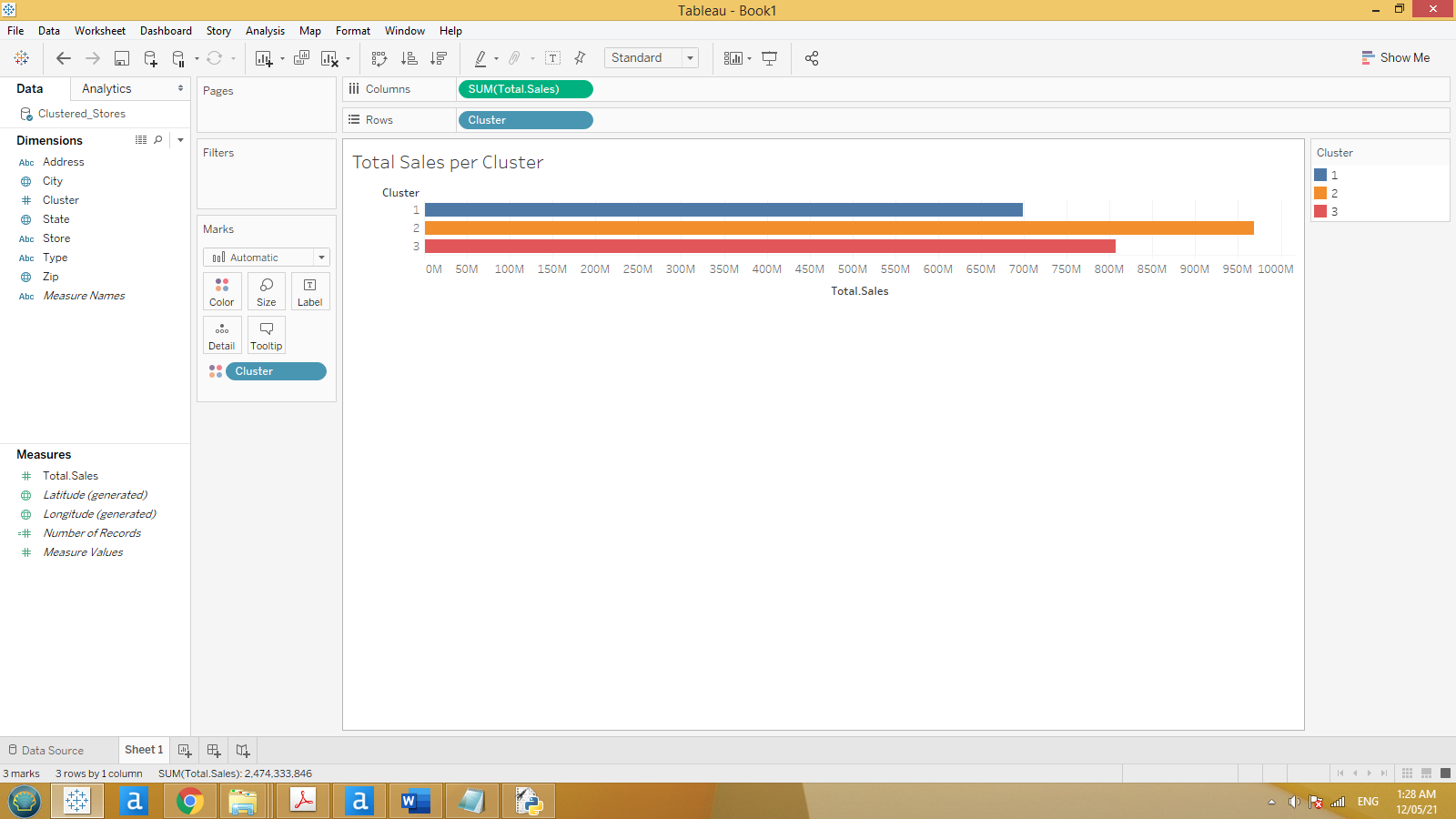


1. How many stores fall into each store format?

Cluster1 has 25 stores, cluster2 has 35 stores, and cluster3 has 25 stores.

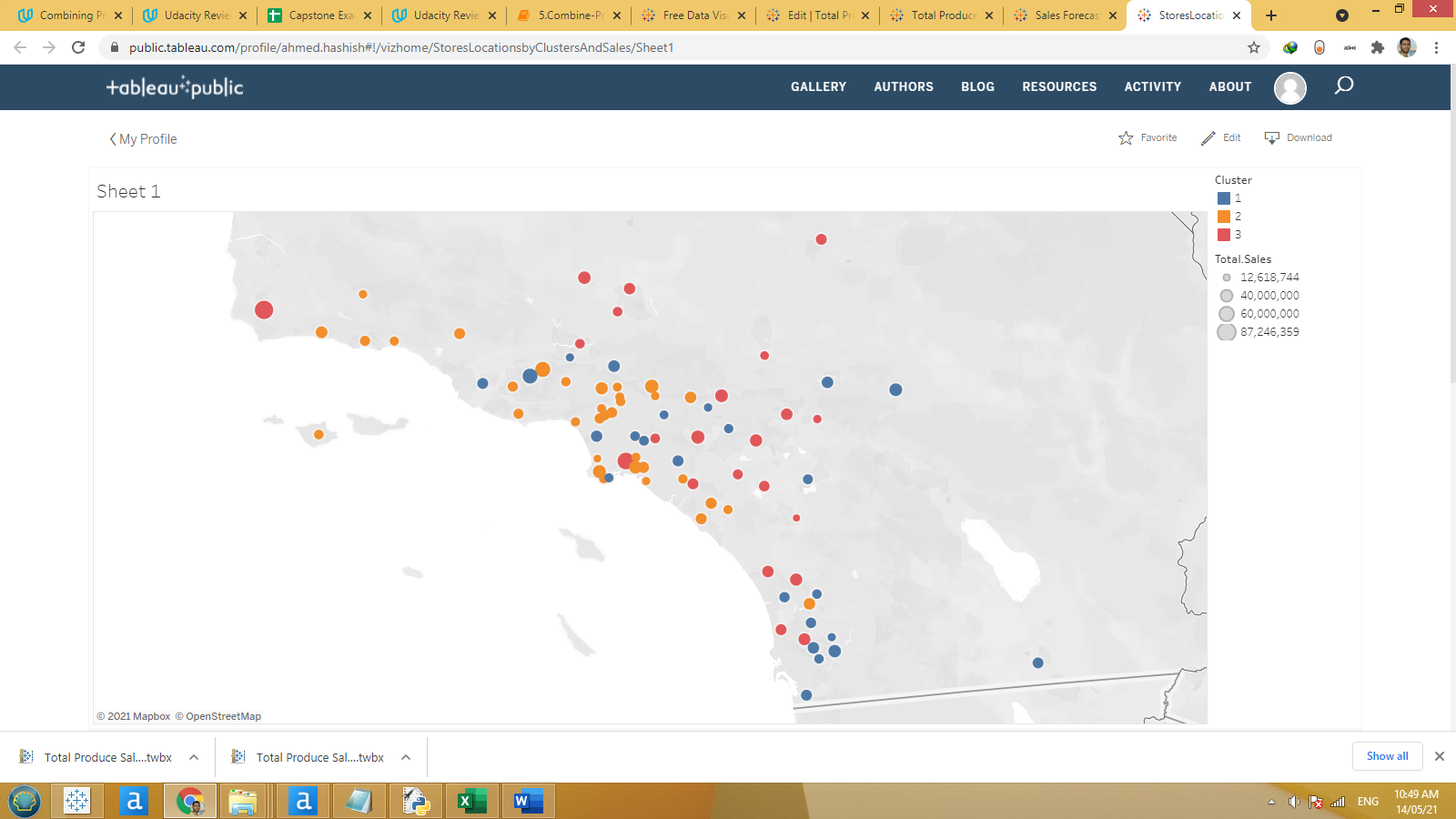
1. Based on the results of the clustering model, what is one way that the clusters differ from one another?

Clusters differ from another with 2015 total sales by their stores where cluster2 refers to the largest category of store in sales volume, cluster3 refers to the medium sales volume and, cluster1 refers to the least sales volume.



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.

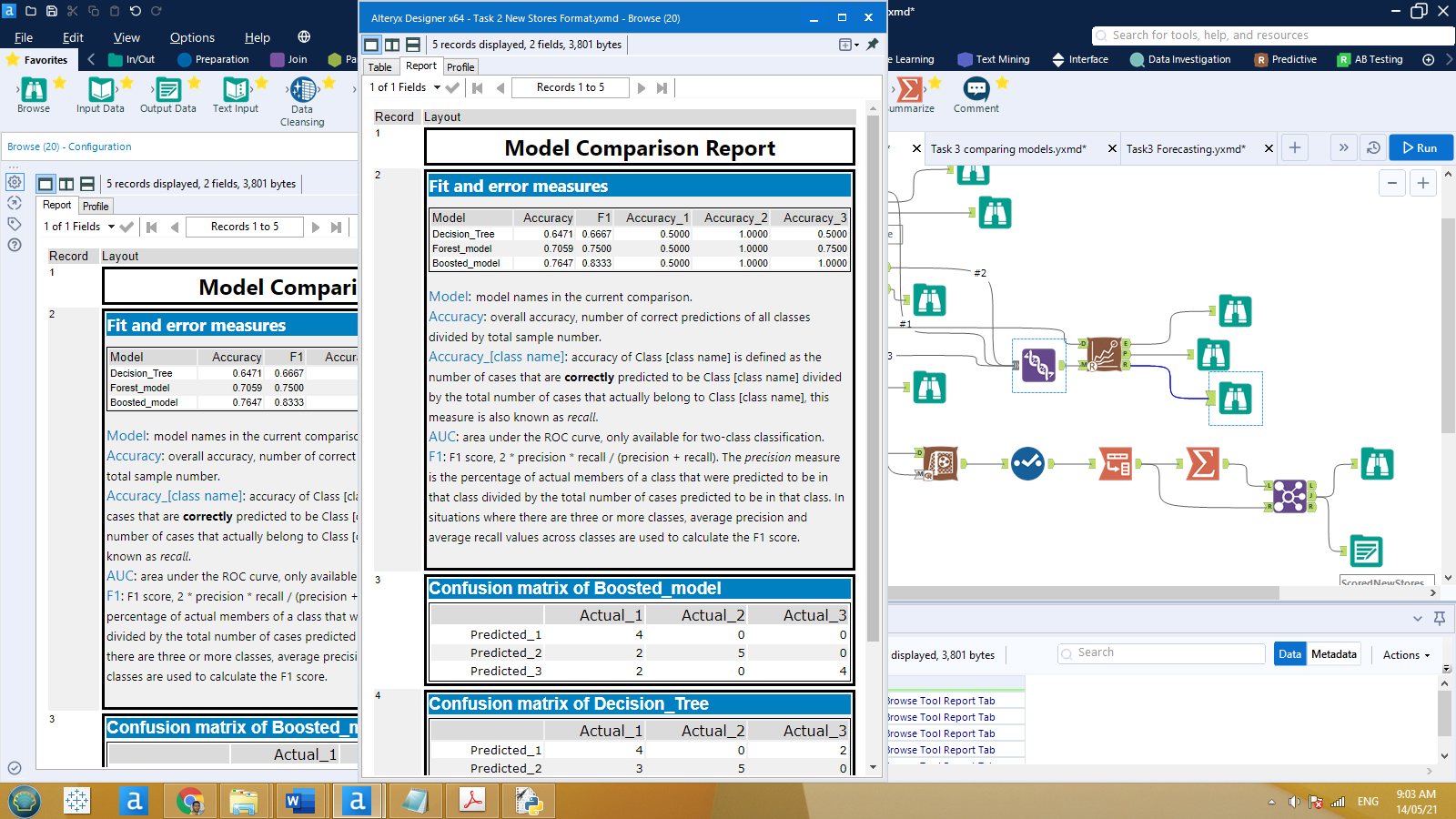
<https://public.tableau.com/profile/ahmed.hashish#!/vizhome/StoresLocationsbyClustersAndSales/Sheet1>

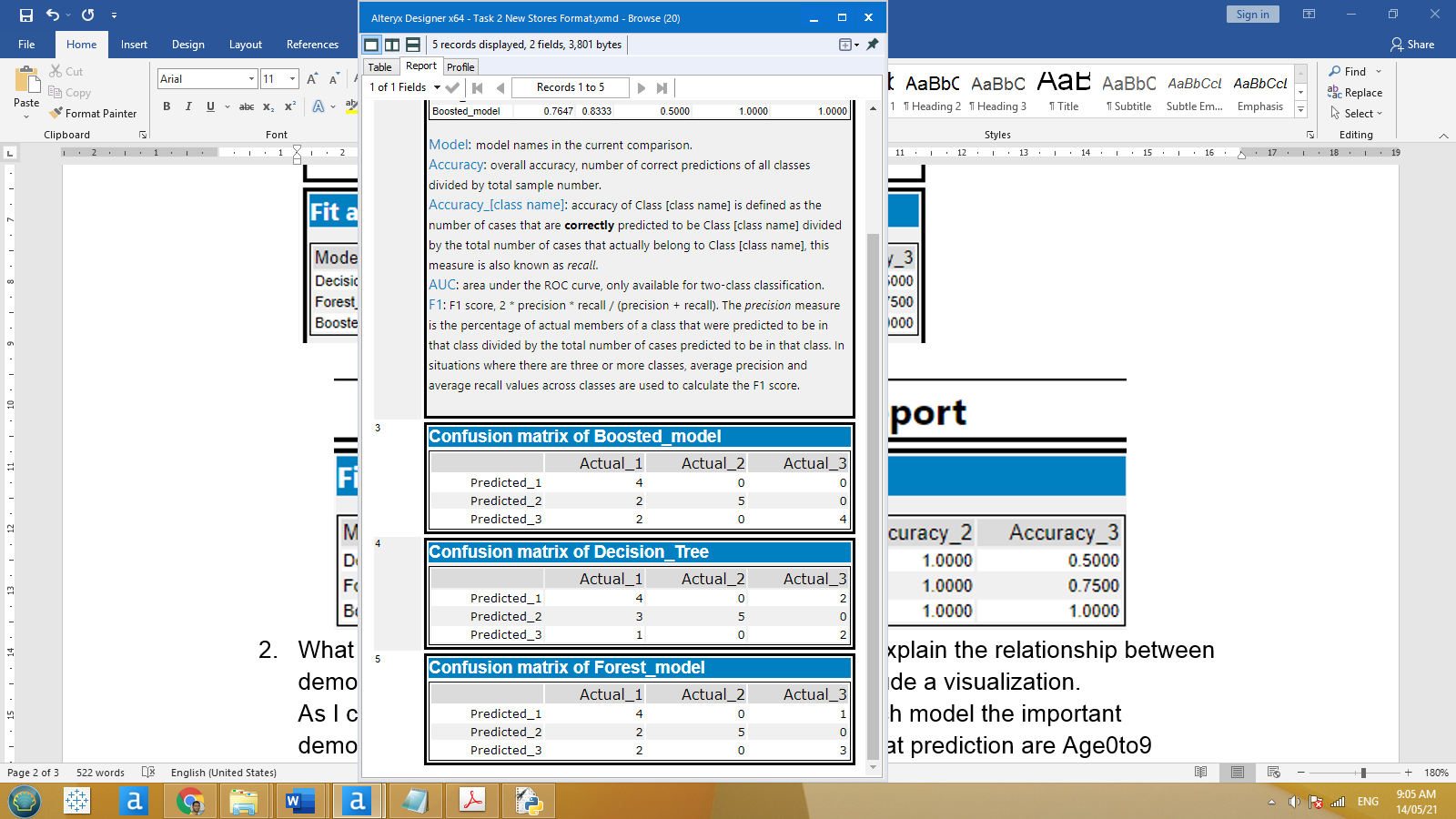


## Task 2: Formats for New Stores

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.)

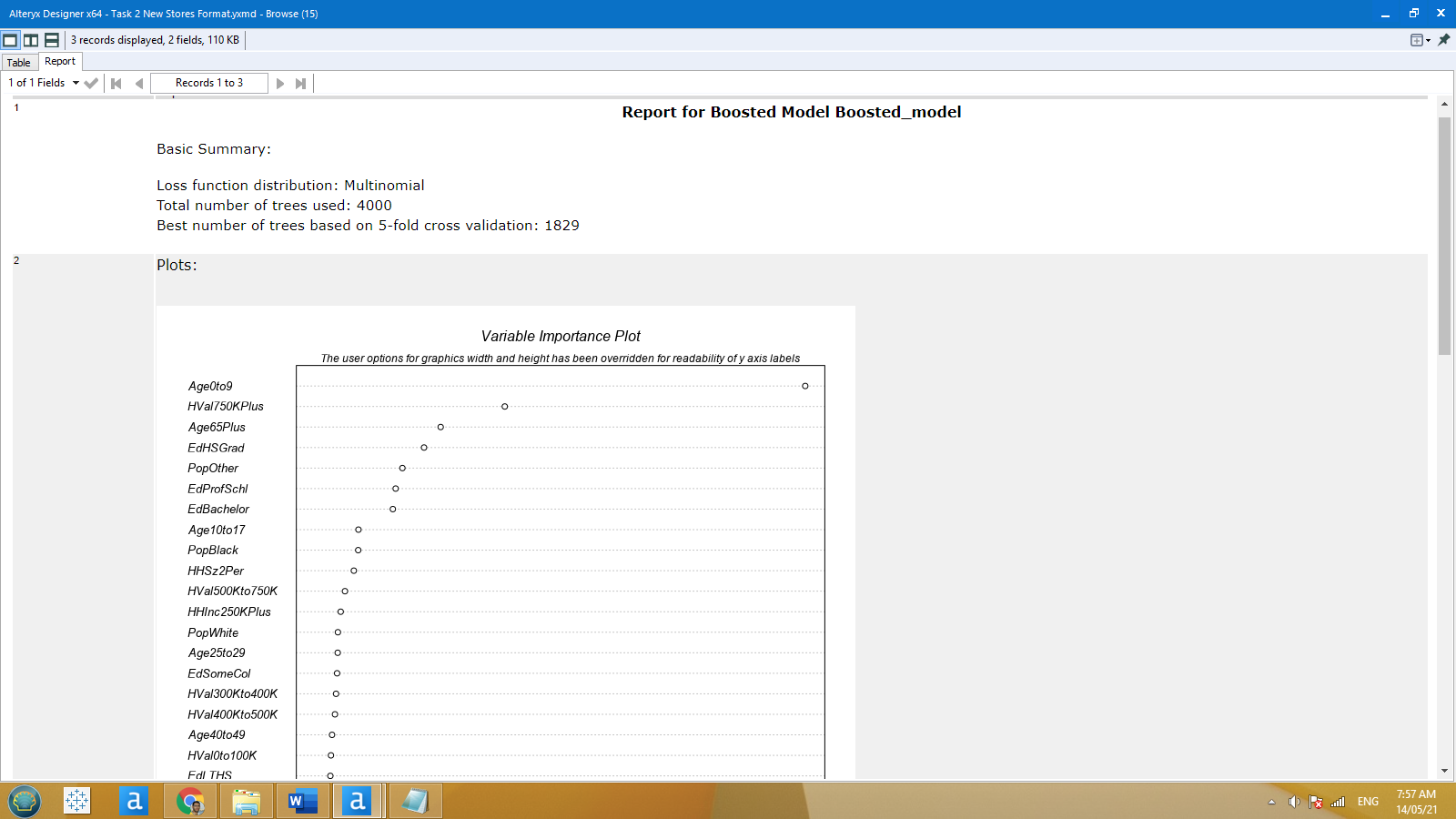
Since we are trying to predict the store format for 85 stores (which is a classification problem) using demographic data, I have trained the data using Decision Tree, Forest and Boosted Models on 68 stores (as 80% estimation sample) then I have validated and compared the three models on the residuals 17 stores (as 20% validation sample). The comparison led us to choose the Boosted Model as it performed better and it will be used to predict the new 10 stores.





1. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.

As I chose the best model was the boosted one, for such model the important demographic variables that mostly affect the store format prediction are Age0to9 followed by HVal750KPlus followed by Age65Plus.



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

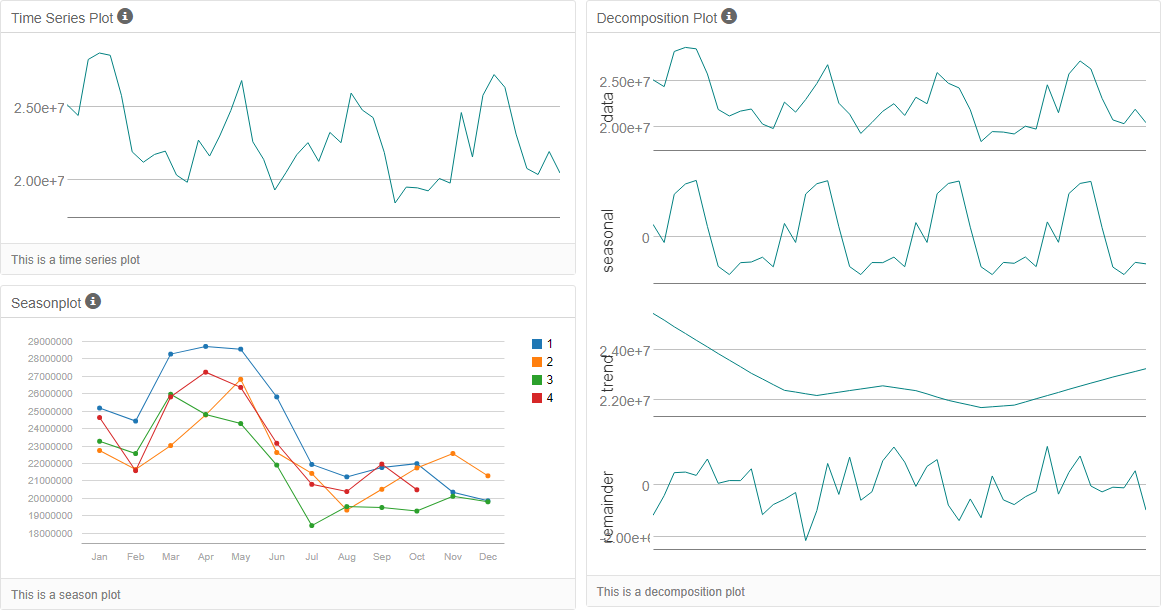
|  |  |
| --- | --- |
| Store Number | Segment |
| S0086 | 1 |
| S0087 | 2 |
| S0088 | 3 |
| S0089 | 2 |
| S0090 | 2 |
| S0091 | 3 |
| S0092 | 2 |
| S0093 | 3 |
| S0094 | 2 |
| S0095 | 2 |

## Task 3: Predicting Produce Sales

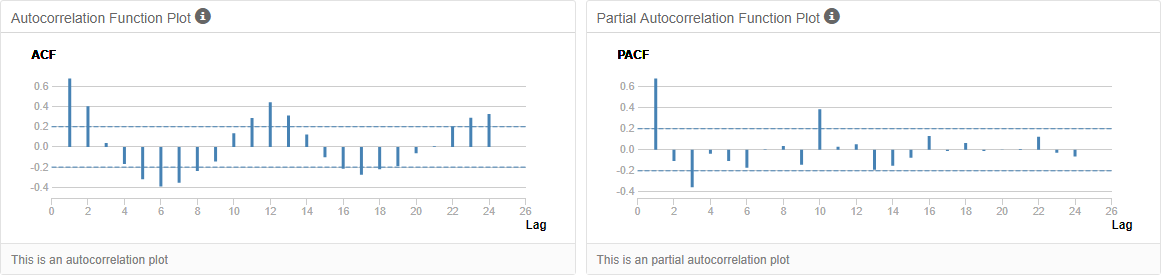
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

To decide whether we’ll use ETS or ARIMA model, first we will check how the Time Series behaves: Looking for seasonality, we could see that it shows increasing trend and should be multiplicatively. The trend plot doesn’t show any trending, and nothing should be applied. Its error is irregular and should be applied multiplicatively.

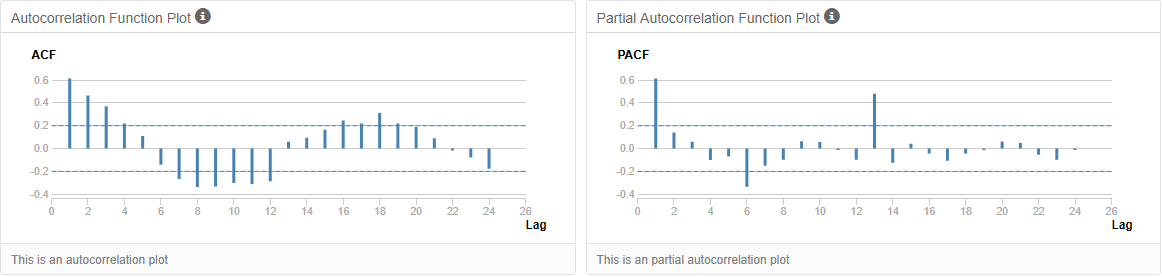
**ETS(M,N,M) with no dampening** should be used for ETS model.



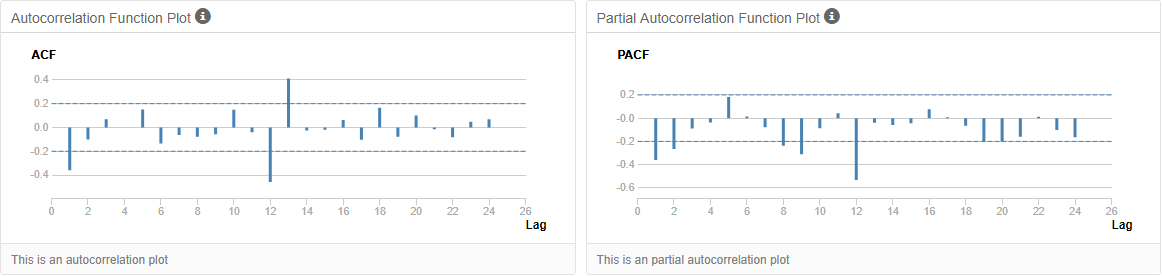
Because of the seasonality on these series, we need to differentiate our Time Series in order to Stationarize the series, as following.



First, we need to look at the seasonal differencing component, to allow us to account for the value as observed in the same season one year earlier, as figure below.

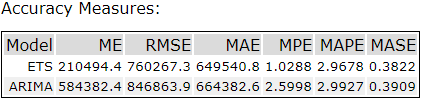


So, looks like we have to take the first seasonal difference to correct for seasonality before the dataset stationary



After plotting the first seasonal difference, we can see that the series has stationarized. We can see this through our ACF and PACF plots, the serial correlational has now disappeared.

For the ARIMA model, the set **ARIMA(0,1,2)(0,1,0)** was chosen, seasonal difference and seasonal first difference were performed. There is a lag-2. The parameters determined for the ARIMA are based on ACF and PACF plots (above)



Based on above Table results, which was obtained from running the two time-series models against the holdout sample of 6 months data, the **ETS model’s accuracy is higher when compared to ARIMA model**. ETS model has lower RMSE value and lower MASE value.

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.

Table of your forecasts for existing and new stores

|  |  |  |
| --- | --- | --- |
| Month | New Stores | Existing Stores |
| Jan-16 | 2,587,450.85 | 21,539,936.01 |
| Feb-16 | 2,477,352.89 | 20,413,770.60 |
| Mar-16 | 2,913,185.24 | 24,325,953.10 |
| Apr-16 | 2,775,745.61 | 22,993,466.35 |
| May-16 | 3,150,866.84 | 26,691,951.42 |
| Jun-16 | 3,188,922.00 | 26,989,964.01 |
| Jul-16 | 3,214,745.65 | 26,948,630.76 |
| Aug-16 | 2,866,348.66 | 24,091,579.35 |
| Sep-16 | 2,538,726.85 | 20,523,492.41 |
| Oct-16 | 2,488,148.29 | 20,011,748.67 |
| Nov-16 | 2,595,270.39 | 21,177,435.49 |
| Dec-16 | 2.573396.63 | 20,855,799.11 |

Visualization of my forecasts that includes historical data, existing stores forecasts, and new stores forecasts is published below

<https://public.tableau.com/profile/ahmed.hashish#!/vizhome/SalesForecast_16209771467160/Sheet1>

