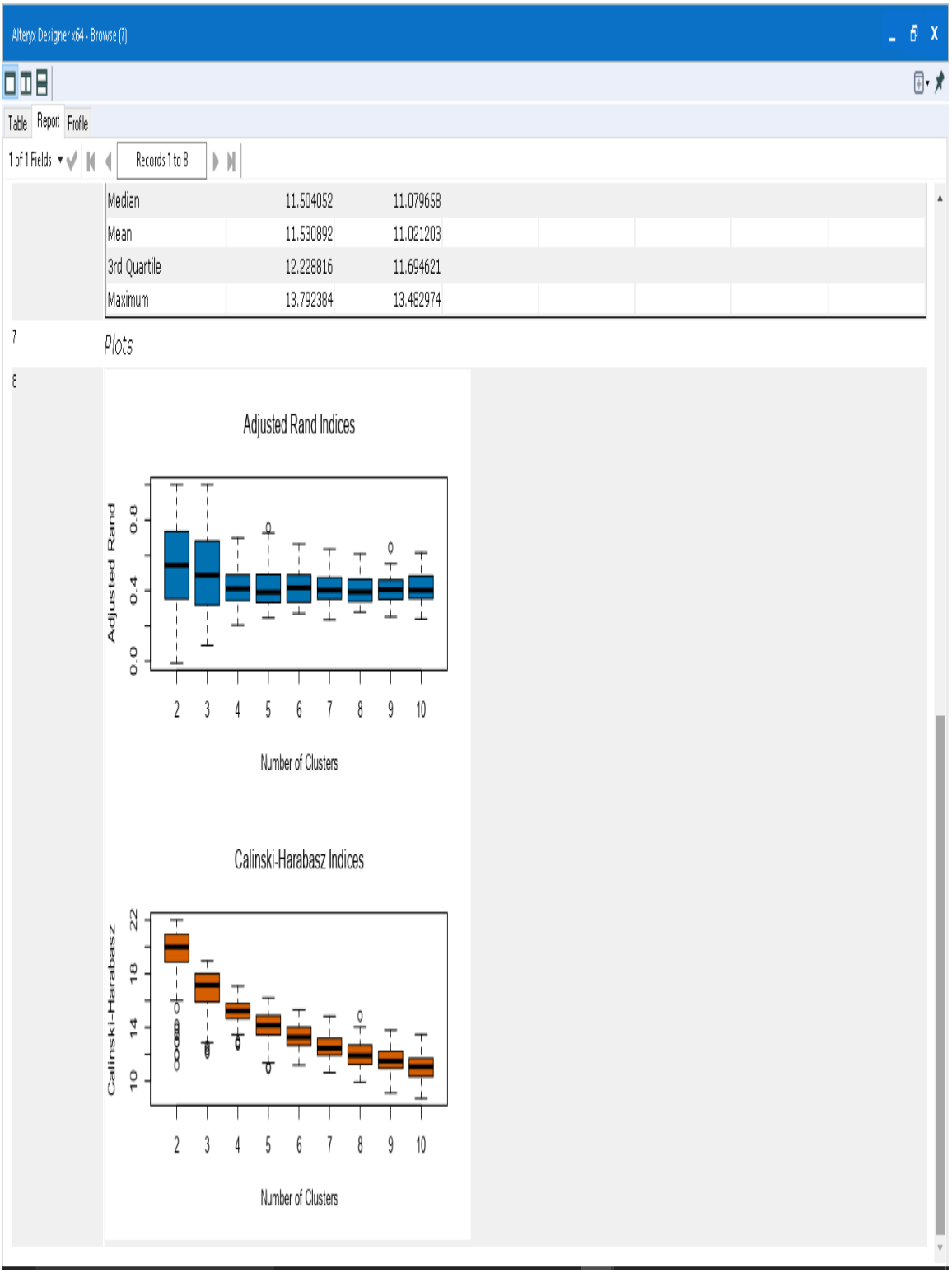
Project: Predictive Analytics Capstone

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

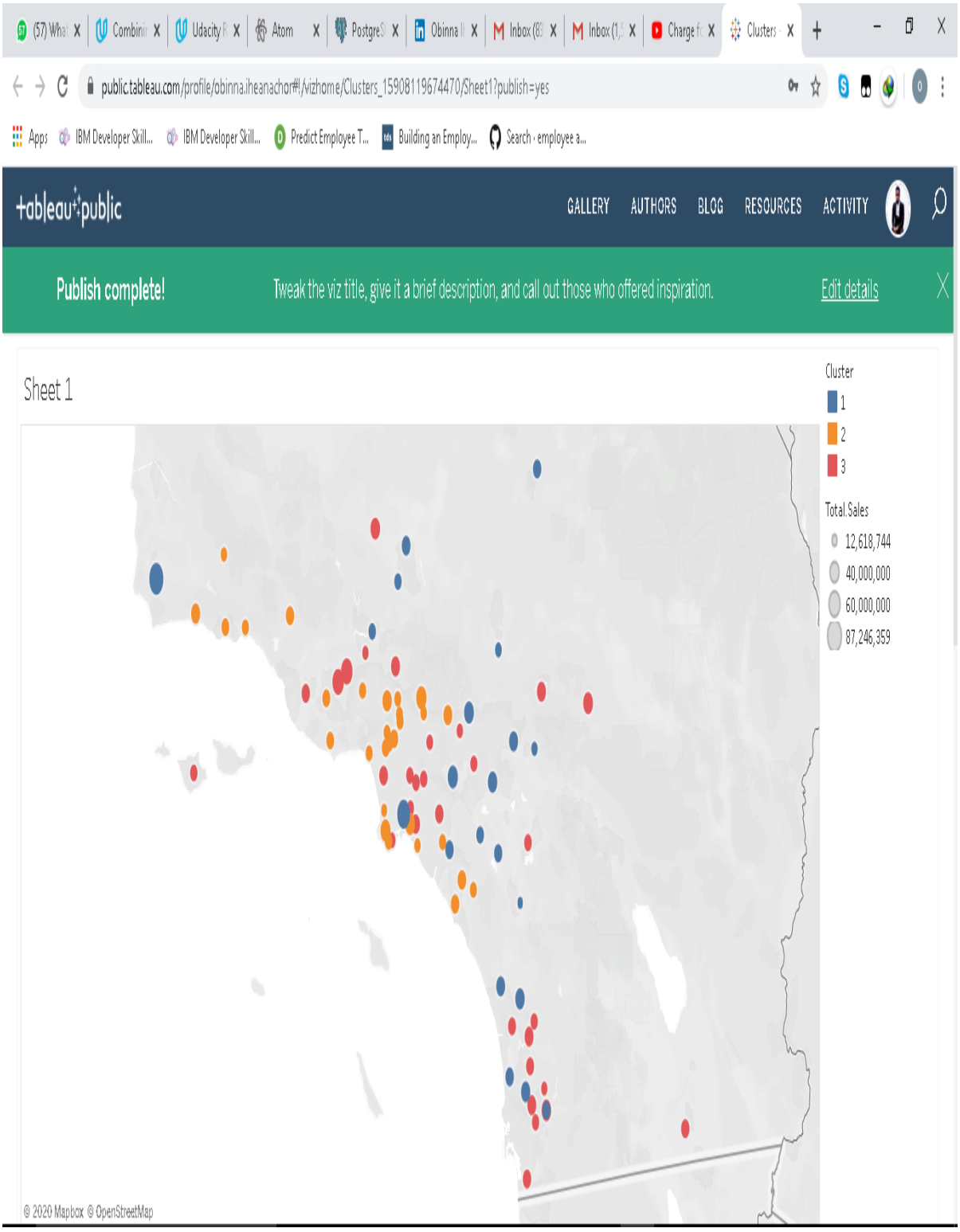
Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?
   Ans: Three
   After aggregating the data, I used the K-Centroids Analysis tool to check for optimum number of clusters. Using K-Means, K-median and Neural Gas all had 2 and 3 Clusters having the highest values for Adjusted Rand and Calinski-Harabasz indices.
   I opted for 3 clusters.



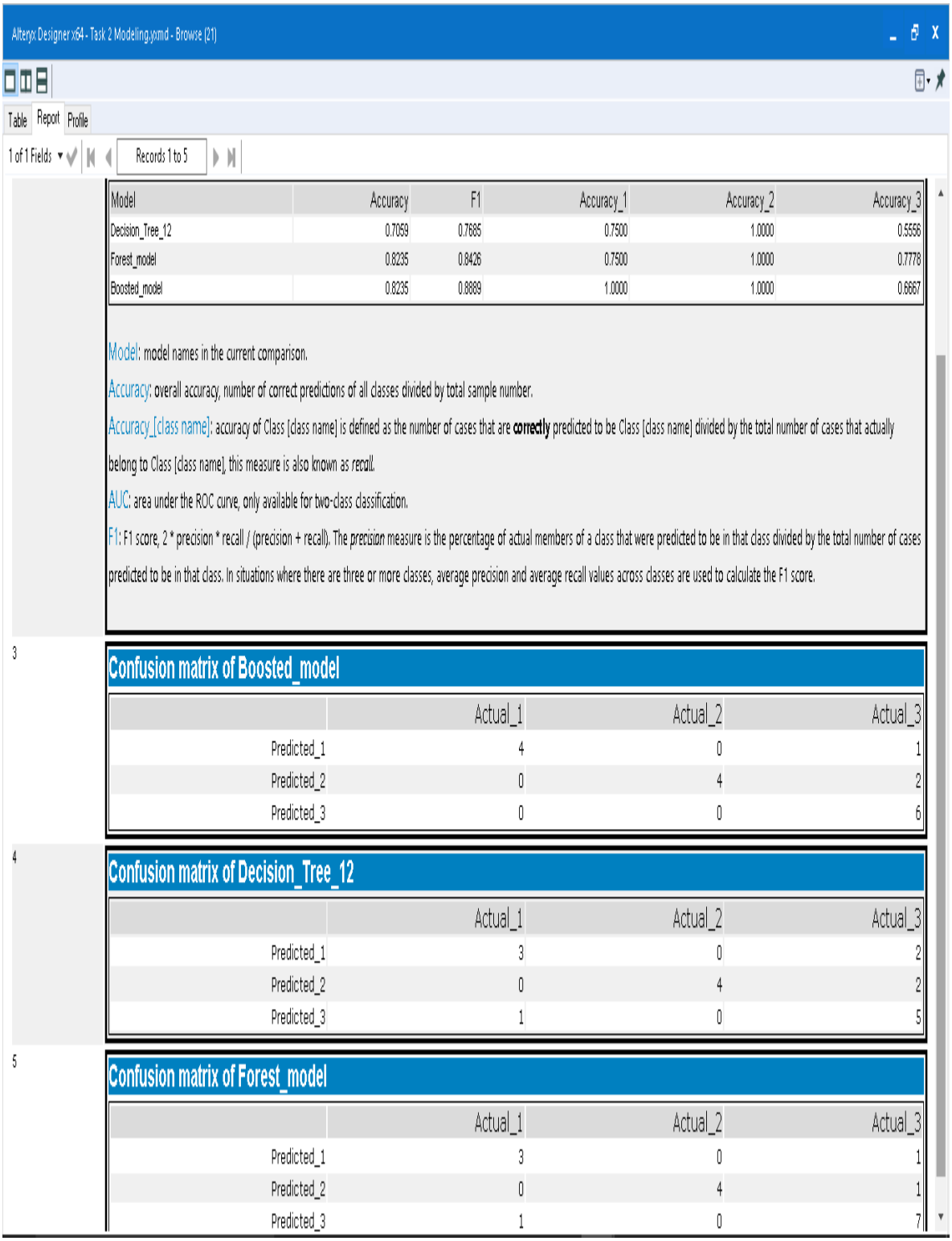
2. How many stores fall into each store format?
   There are 23 stores in store format (Cluster) 1, 29 stores in store format 2 and 33 stores in store format 3

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Ans: Stores in Cluster 1 and 3 have typically low Produce percentage of total sales while stores in Cluster 2 have higher percentage of total sales for Produce.
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.
   https://public.tableau.com/profile/obinna.iheanachor#!/vizhome/Clusters_15908119674470/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.)
   I used a Classification methodology (model) to predict since we are trying to predict the formats (which is a classification problem) for the stores and it's a Data-Rich problem as we have relevant data for the stores.
   I trained the data using Decision Tree, Forest and Boosted Models on the 85 stores with cluster values and using the accuracy, classification reports, ROC Graphs and accuracy in cluster categories I chose the Boosted Model as it performed better and used it to predict on the remaining 10 stores.



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?
Ans: ETS(M,N,M)

Our trend line does not exhibit linear behavior so we will use a None method.
The seasonality changes in magnitude each year so a multiplicative method is necessary.
The error changes in magnitude as the series goes along so a multiplicative method will be used.
This leaves us with an ETS(M, N, M) model.
For the ARIMA model, I used use an ARIMA(0, 1, 1)(0, 1, 1)12 since there are seasonal components found in the time series. After differencing the lags showed MA behaviour.
I chose the ETS model as it performed better than its ARIMA counterpart on both In-sample and validation data as it had lower values for RMSE , MAPE and ME.

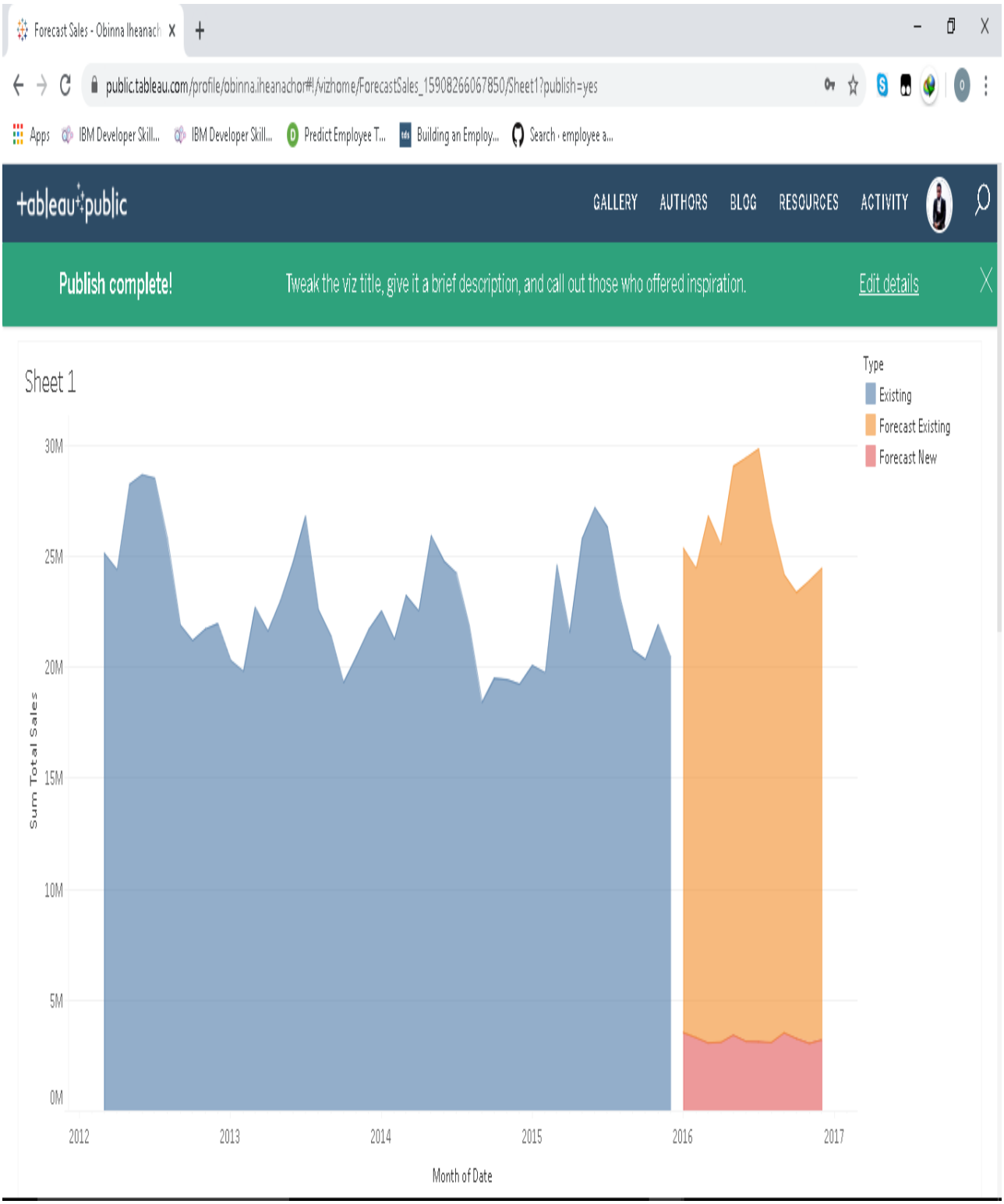
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	-492238.83	792197.3	735878.2	-2.1992	3.3098	0.433

Actual and Forecast Values

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	3544025	21829060.03
Feb 16	3317211	21146329.63
Mar 16	3081738	23735686.93
April 16	3107630	22409515.28
May 16	3430206	25621828.72
June 16	3148614	26307858.04
July 16	3132783	26705092.55
Aug 16	3099740	23440761.32
Sep 16	3532822	20640047.31
Oct 16	3274603	20086270.46
Nov 16	3057921	20858119.95
Dec 16	3217559	21255190.24



https://public.tableau.com/profile/obinna.iheanachor#!/vizhome/ForecastSales_15908266067850/Sheet1?publish=yes

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.