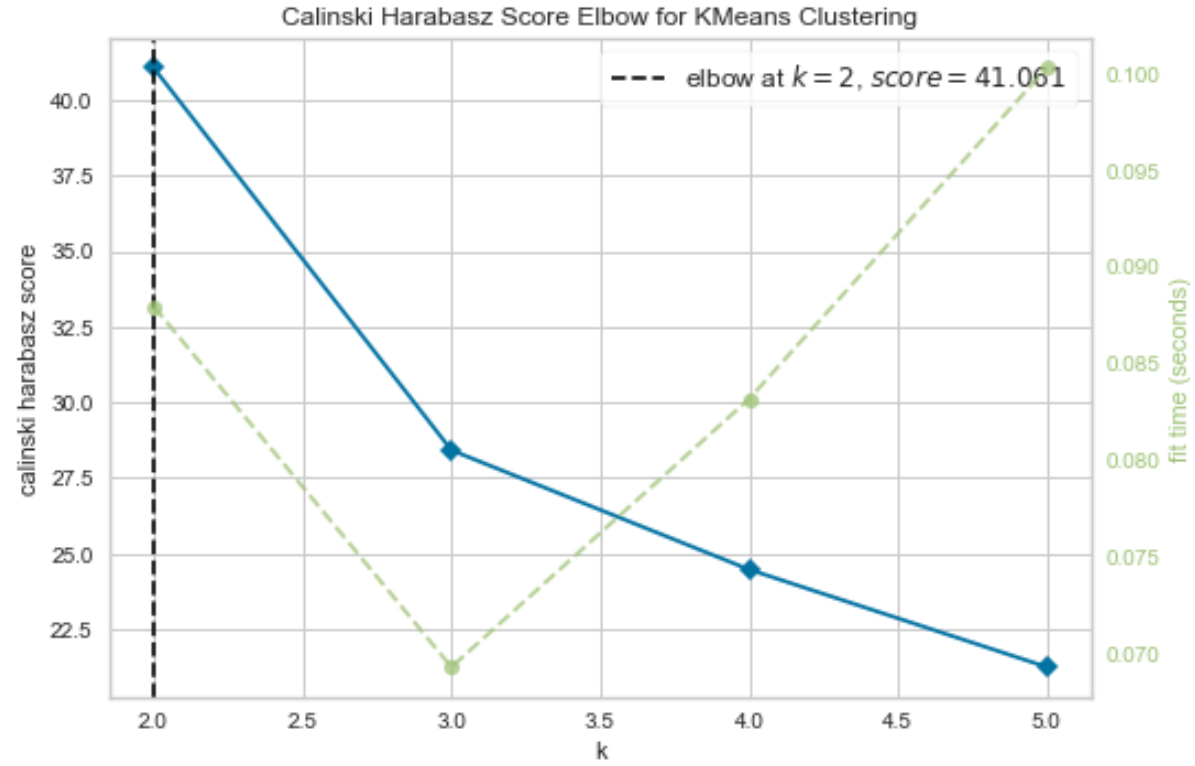
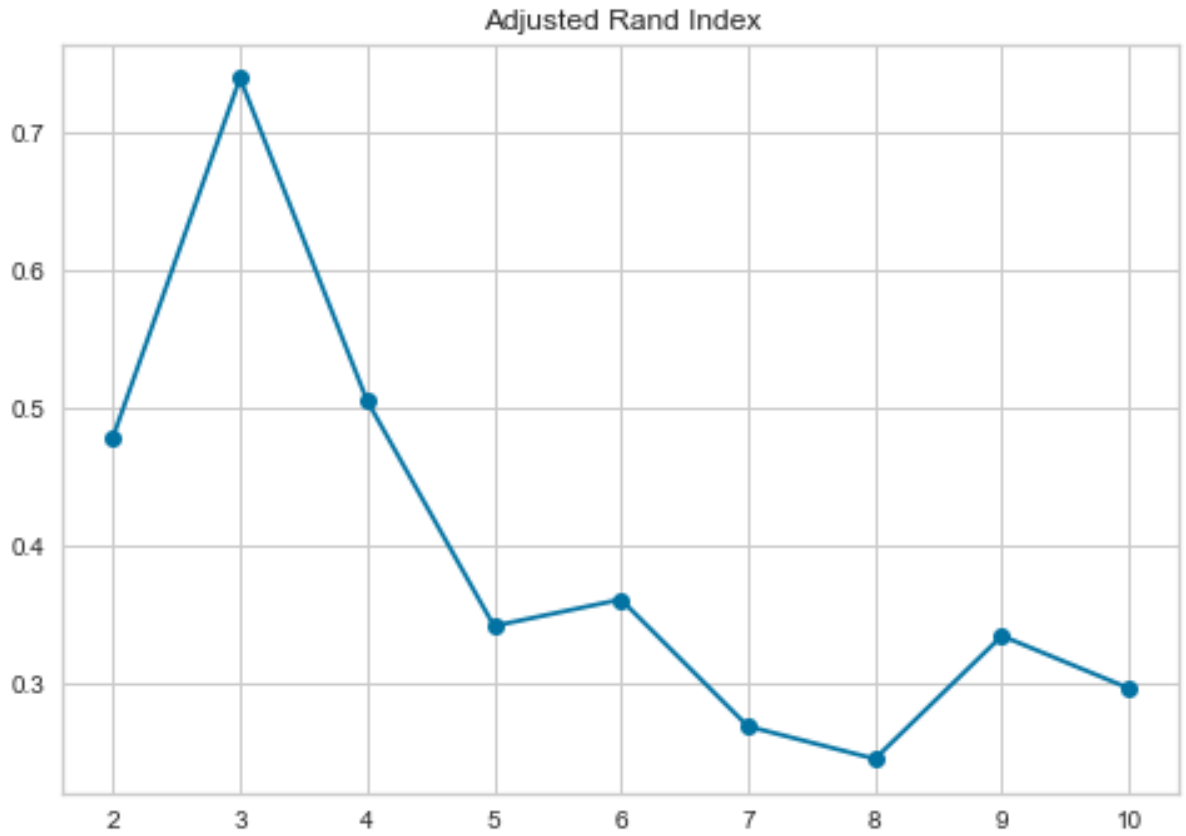
COMBINING PREDICTIVE ANALYTICS (CAPSTONE PROJECT)

Task 1: Determine Store Formats for Existing Stores

What is the optimal number of store formats? How did you arrive at that number?
The optimal number of store formats is 3, this can be shown by using the Calinski Harabasz Score and the Silhouette Visualization Method. The Visuals are shown below:





How many stores fall into each store format?
After applying K Means Clustering, the following results were obtained for the clusters of the stores:

Store Format	No. of Stores
Cluster 1	25
Cluster 2	35
Cluster 3	25

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

Cluster 2 has a higher Total Sales amount than Cluster 1 and Cluster 3.

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.

The link for the Tableau Public File can be found here:
https://public.tableau.com/app/profile/margaret.awojide/viz/UdacityProjectTask1/Sheet1

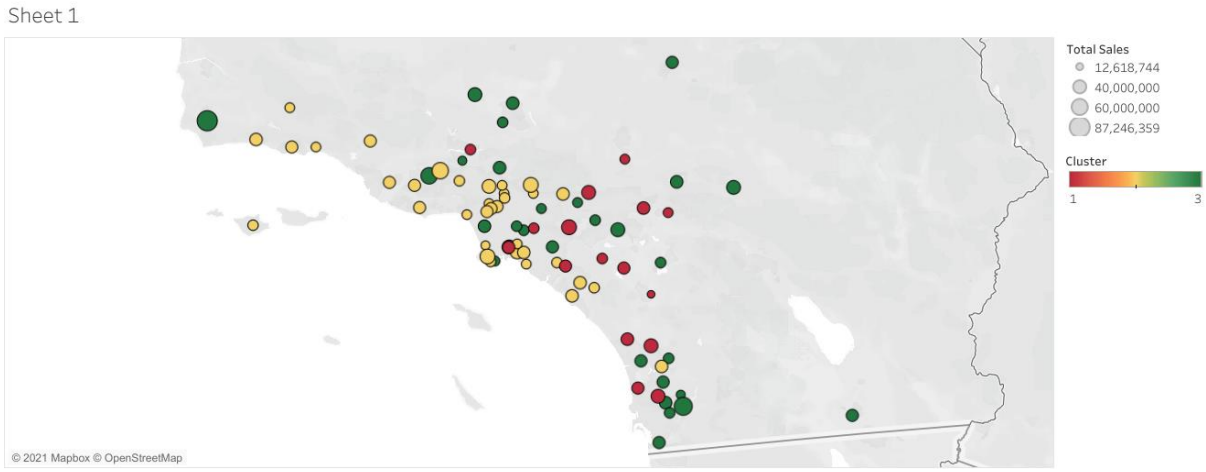
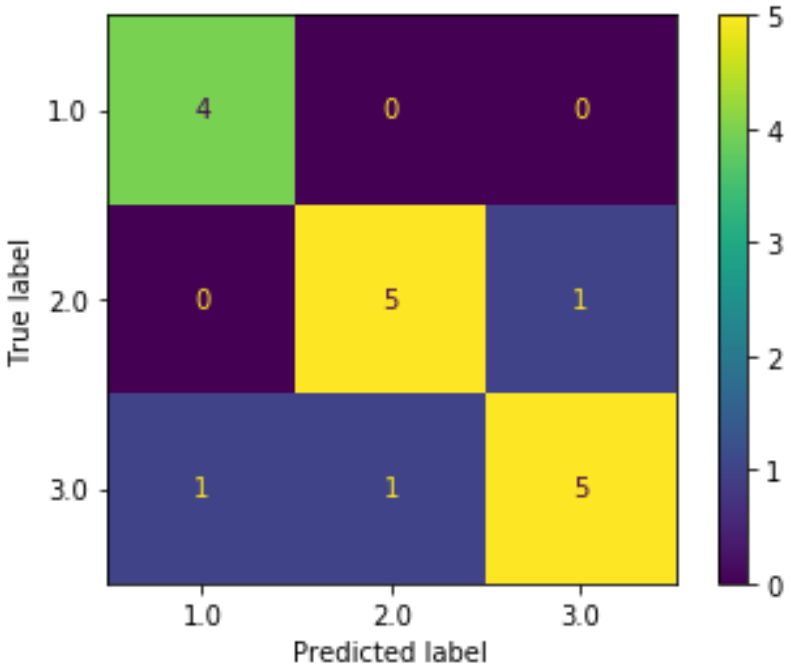


Tableau Visualization

Task 2: Store Format for New Stores

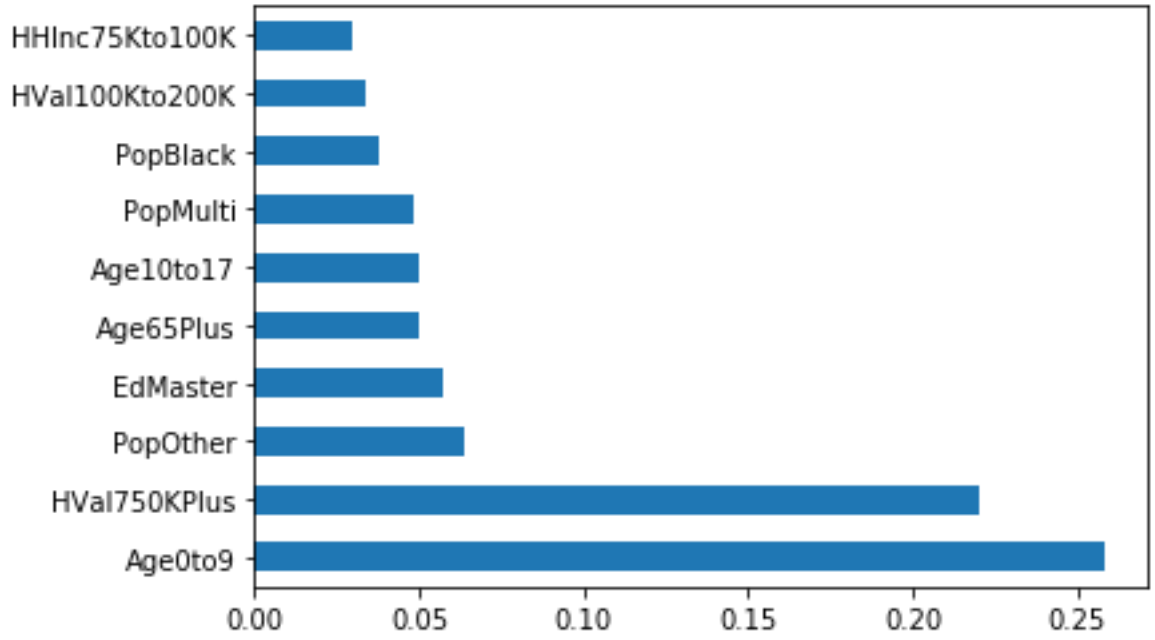
- What methodology did you use to predict the best store format for the new store and why did you choose the methodology?

A 20% validation sample was created and 3 models were tested for the best store format: Decision Trees, Random Forest and the Gradient Boost Model. The Gradient Boost Model had the highest Model Accuracy Score(0.823) and F1 Score(0.830) and was selected as the model to be used for prediction of clusters.



- What are the 3 most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.

After calculating using the Gradient Boost Model, it was discovered that the 3 most important variables are: Age0to9, HVaI750KPlus and PopOther as seen in the feature importance bar chart below. The chart below shows the top 10 important variables using the Gradient Boost Model.



- What format do each of the 10 new stores fall into? Please provide a data table

The new store format for the 10 new stores are given in the table below:

STORE	STORE FORMAT
S0086	3
S0087	2
S0088	2
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

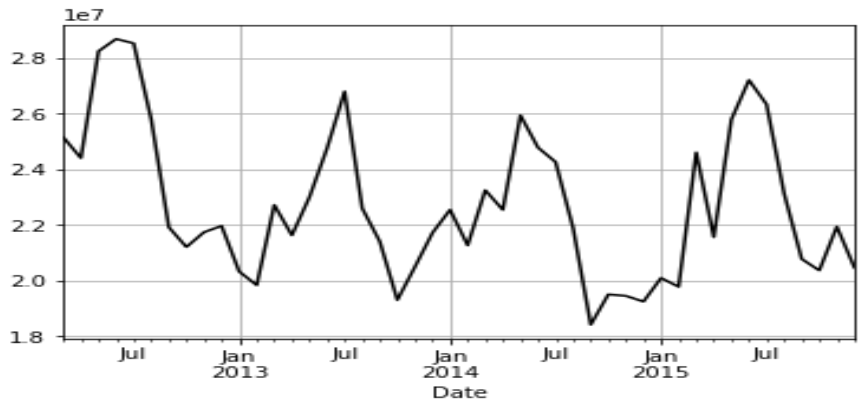
Task 3: Forecasting

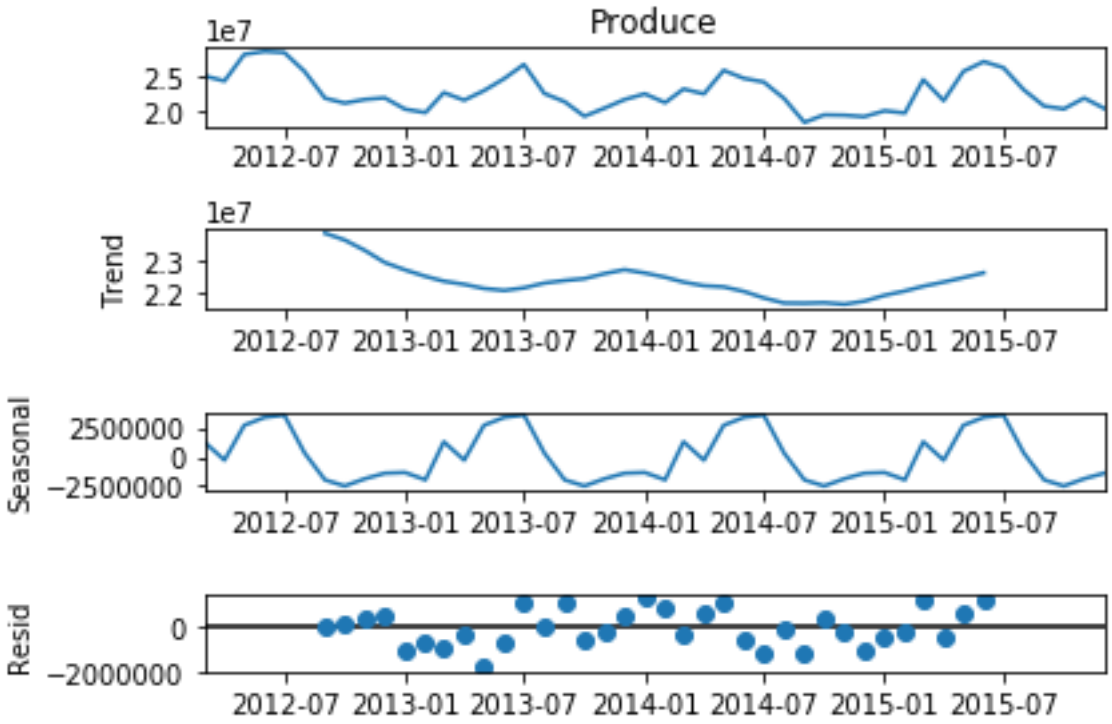
For forecasting, the ARIMA model was applied to forecast the product sales for the months of 2016 for the existing stores and new stores to be created. The time series ARIMA model used was the ARIMA (1,0,1) model.

Since we do not have data for the Produce Sales of the new stores, the values were forecasted using on the basis of the clusters they fall under.

The Augmented Dickey-Fuller test was used to check for the stationarity of the data.

Statistical Results for the Existing Stores Forecast





SARIMAX Results

Dep. Variable:	Produce	No. Observations:	46
Model:	SARIMAX(1, 0, 1)	Log Likelihood	-737.491
Date:	Fri, 24 Jun 2022	AIC	1480.983
Time:	20:44:46	BIC	1486.469
Sample:	03-01-2012	HQIC	1483.038
	- 12-01-2015		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9962	0.012	83.239	0.000	0.973	1.020
ma.L1	-0.0415	0.153	-0.272	0.786	-0.341	0.258
sigma2	4.309e+12	3.11e-15	1.38e+27	0.000	4.31e+12	4.31e+12

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	0.50
Prob(Q):	0.92	Prob(JB):	0.78
Heteroskedasticity (H):	1.28	Skew:	0.17
Prob(H) (two-sided):	0.64	Kurtosis:	2.62

Month	Existing Stores	New Stores
Jan-16	20440363.28	2672239.26
Feb-16	20362969.31	2661383.56
Mar-16	20285868.39	2650572.09
Apr-16	20209059.39	2639804.70
May-16	20132541.22	2629081.19
Jun-16	20056312.77	2618401.38
Jul-16	19980372.95	2607765.10
Aug-16	19904720.66	2597172.17

Sep-16	19829354.82	2586622.42
Oct-16	19754274.33	2576115.66
Nov-16	19679478.13	2565651.71
Dec-16	19604965.12	2555230.42

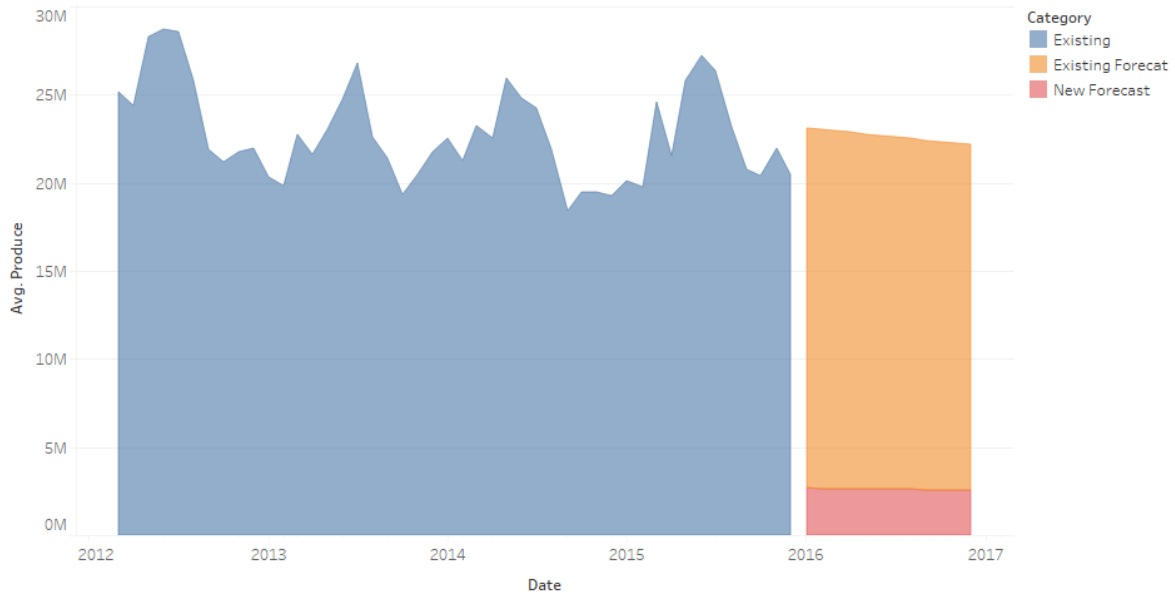


Tableau Link