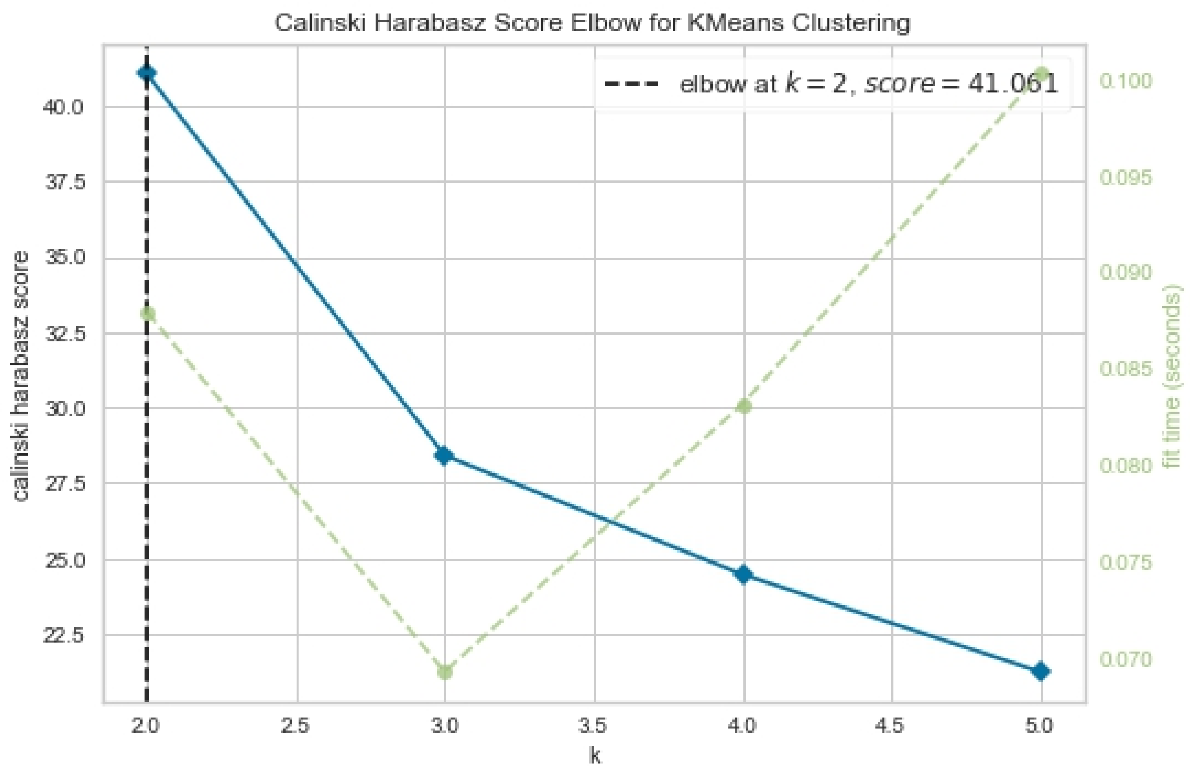
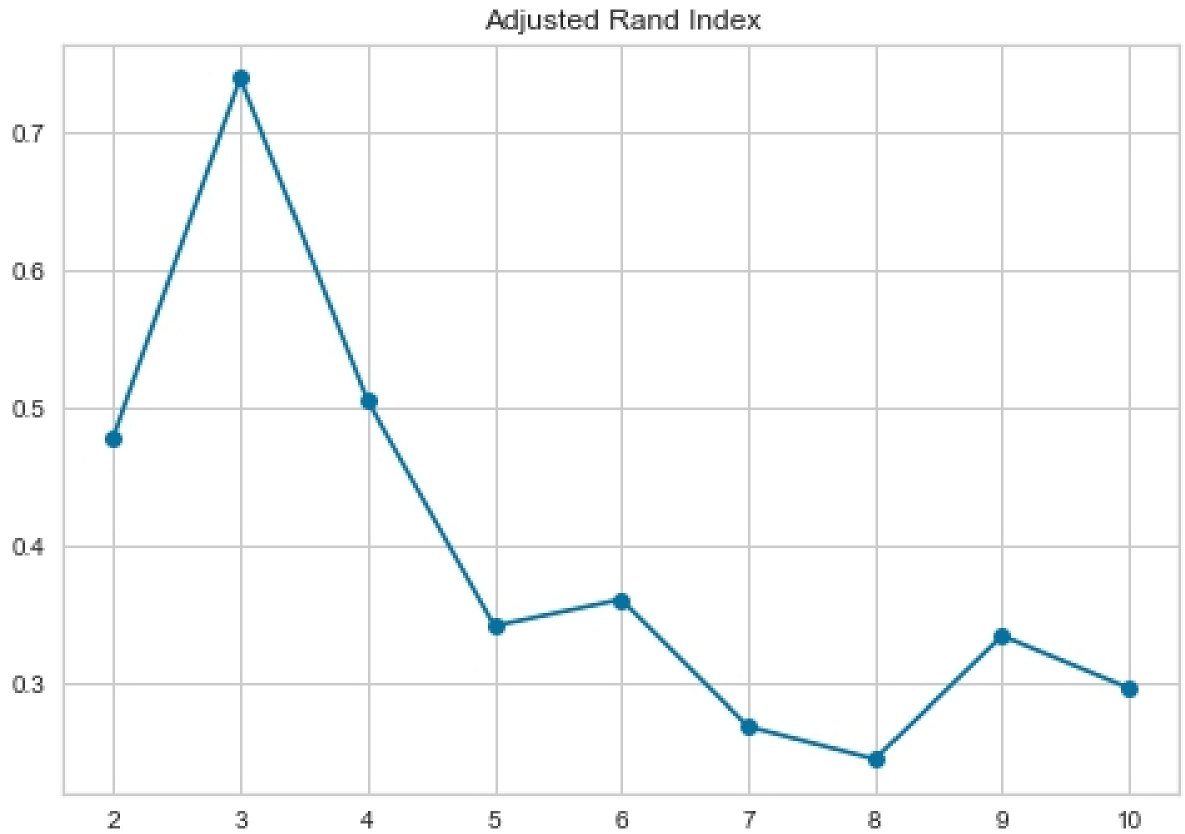
COMBINING PREDICTIVE ANALYTICS (CAPSTONE PROJECT)

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

Sheet 1

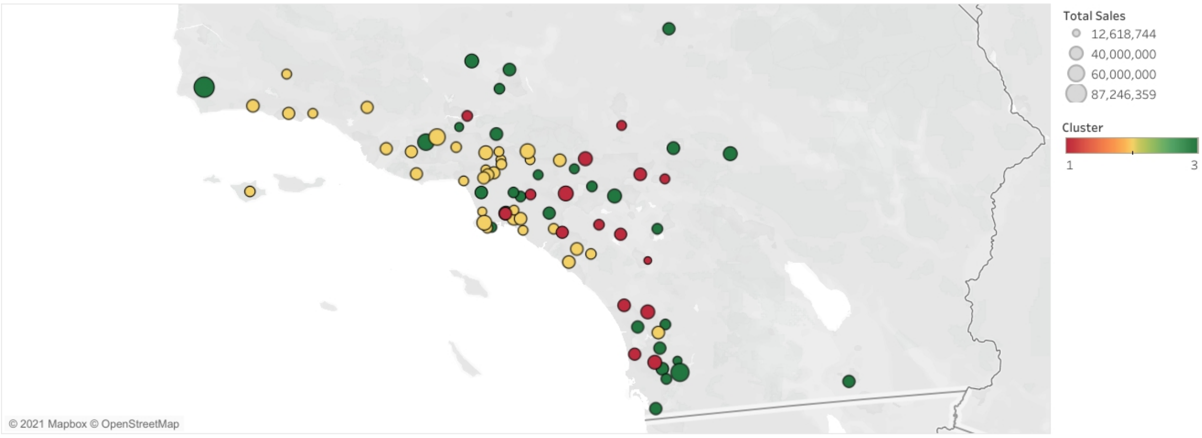
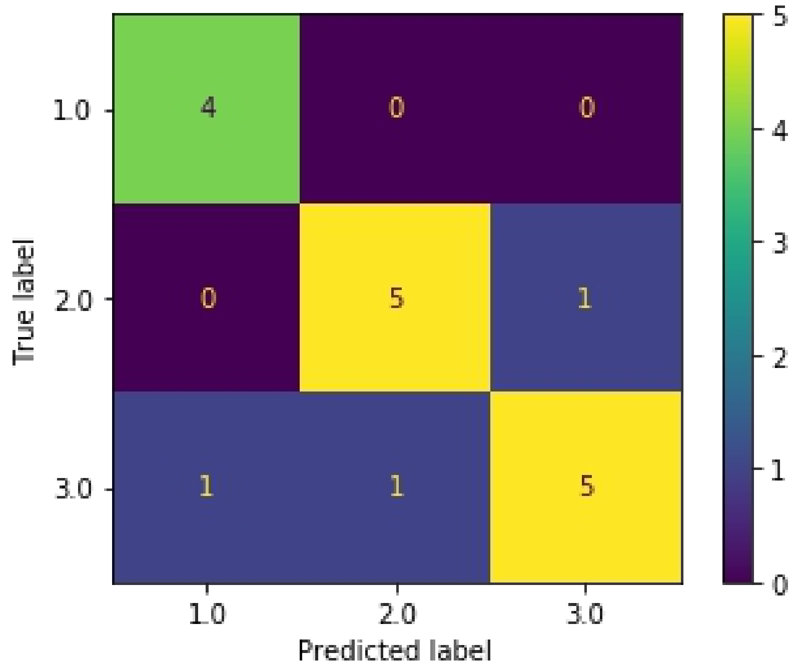


Tableau Visualization

Task 2: Store Format for New Stores

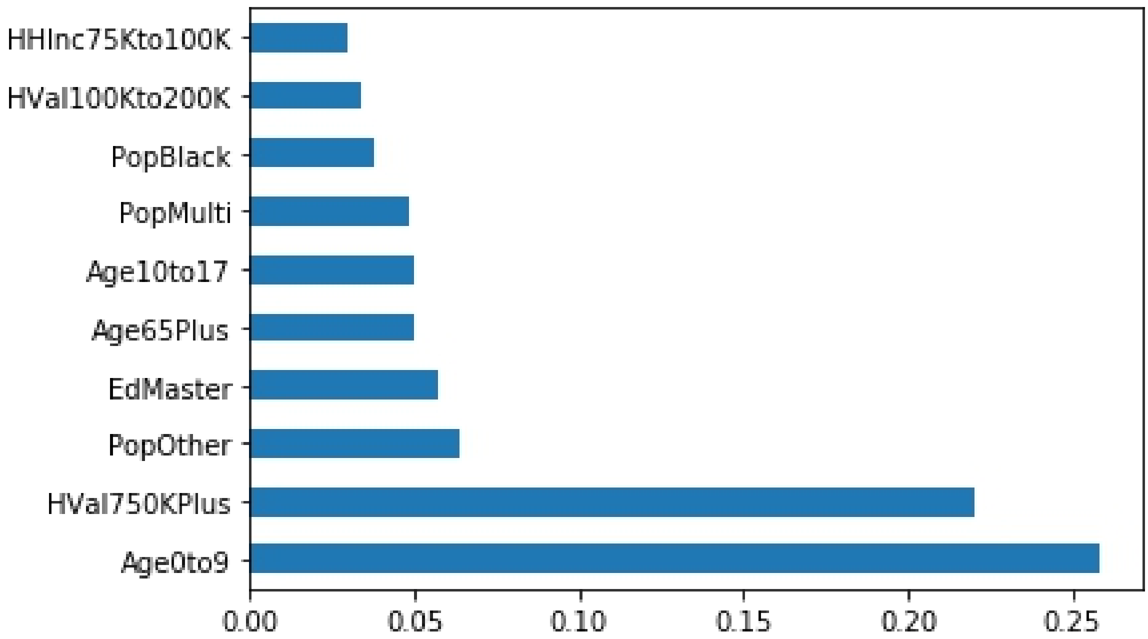
❖ What methodology did you use to predict the best store format for the new store and why did you choose the methodology?

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, HVal750KPlus 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

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?

Comparing the results of both the ETS and ARIMA models, the ETS Model performs better than the ARIMA models (based on Accuracy Measures) and has predicted values closer to the actual values.

Comparison of Time Series Models

Actual and Forecast Values:

Actual	Arima_Model
19444753.17	21031463.85798
21936906.81	21165512.05495
21962976.75	21286462.81556
21715706.67	21395595.84997
19240384.75	21494065.8318
20462899.3	21582914.61506

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
Arima_Model	-532064.6	1291405	1121404	-2.8793	5.5696	0.5969

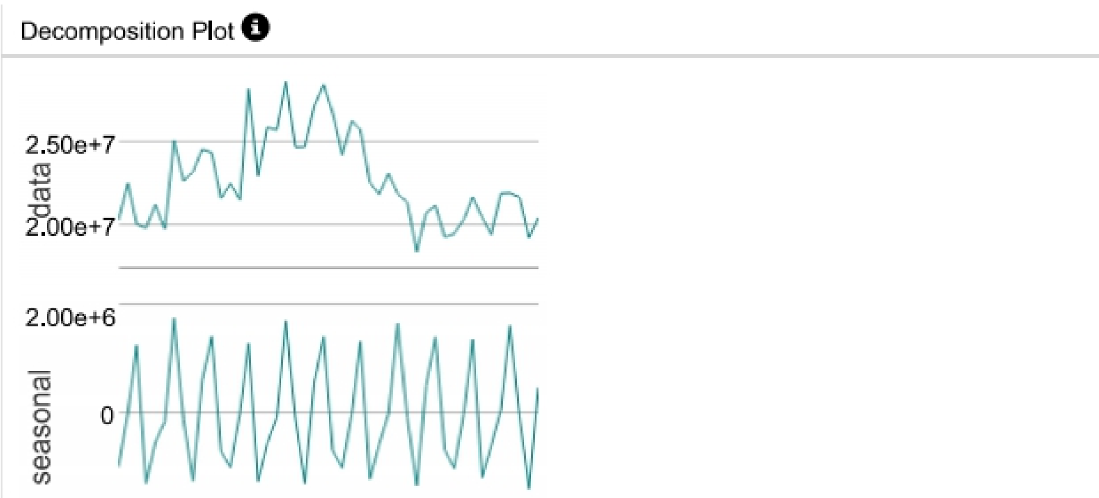
Comparison of Time Series Models

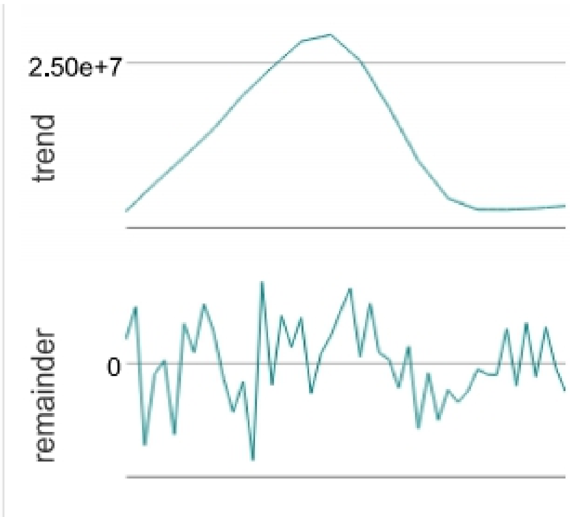
Actual and Forecast Values:

Actual	ETS_Model
19444753.17	20673939.65687
21936906.81	20673939.65687
21962976.75	20673939.65687
21715706.67	20673939.65687
19240384.75	20673939.65687
20462899.3	20673939.65687

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_Model	119998.3	1151267	1077926	0.27	5.2045	0.5738





From the Decomposition Plot, the Seasonality depicts a Multiplicative nature, The trend is not linear (None) and the Remainder/Error is Multiplicative. So, we use the ETS(M,N,M)

Model.

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	Existing Stores	New Stores
Jan-16	21,277,864.022	2603262.30
Feb-16	19,072,850.584	2508877.75
Mar-16	18,718,638.042	2989457.78
Apr-16	19,576,816.645	2849287.16
May-16	21,277,761.073	3224711.21
Jun-16	18,988,118.937	3269622.51
Jul-16	19,495,186.494	3288334.00
Aug-16	20,160,315.610	2937302.49
Sep-16	21,916,422.870	2606592.39
Oct-16	20,266,649.375	2536270.35
Nov-16	19,368,334.942	2631292.65
Dec-16	20,793,989.583	2586562.09