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

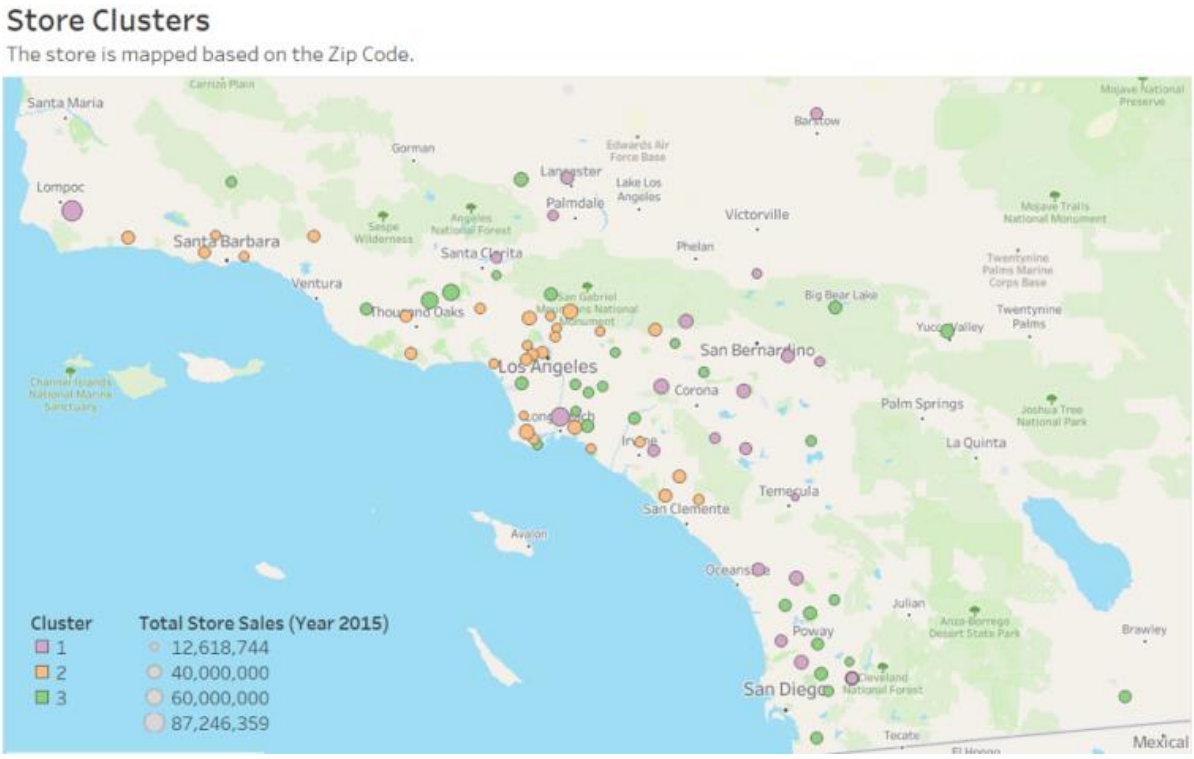
1. What is the optimal number of store formats? How did you arrive at that number?

   The optimal number of store formats is 3. I ran K-Means clustering model and used the median and spread of the Adjusted Rand Indices and Calinski-Harabasz Indices. It shows that 3 clusters are the optimal method because the box-whisker plots in the Adjusted Rand Indices show how tight the indices for each data point are within each other. Even though it looks like cluster 2 is the optimal number of clusters, it is actually 3 because the variance is too big for 2 clusters, while we see more compactness and still high median values when we have 3 clusters
2. How many stores fall into each store format?
   Cluster 1: 23 Stores

   Cluster 2: 29 Stores

   Cluster 3: 33 Stores
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Based on the summary report of the K-Means Clustering solution, one way that the clusters differ from one another could be: considering the percentage of sales by category of each store, cluster 1 sells more in general merchandise; cluster 2 sells more in produce and floral; and cluster 3 sells more in deli and meat; etc.

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.



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.)
   Boosted Model was used to predict the best store format for the new stores. Boosted Model is the best because it has a higher accuracy

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

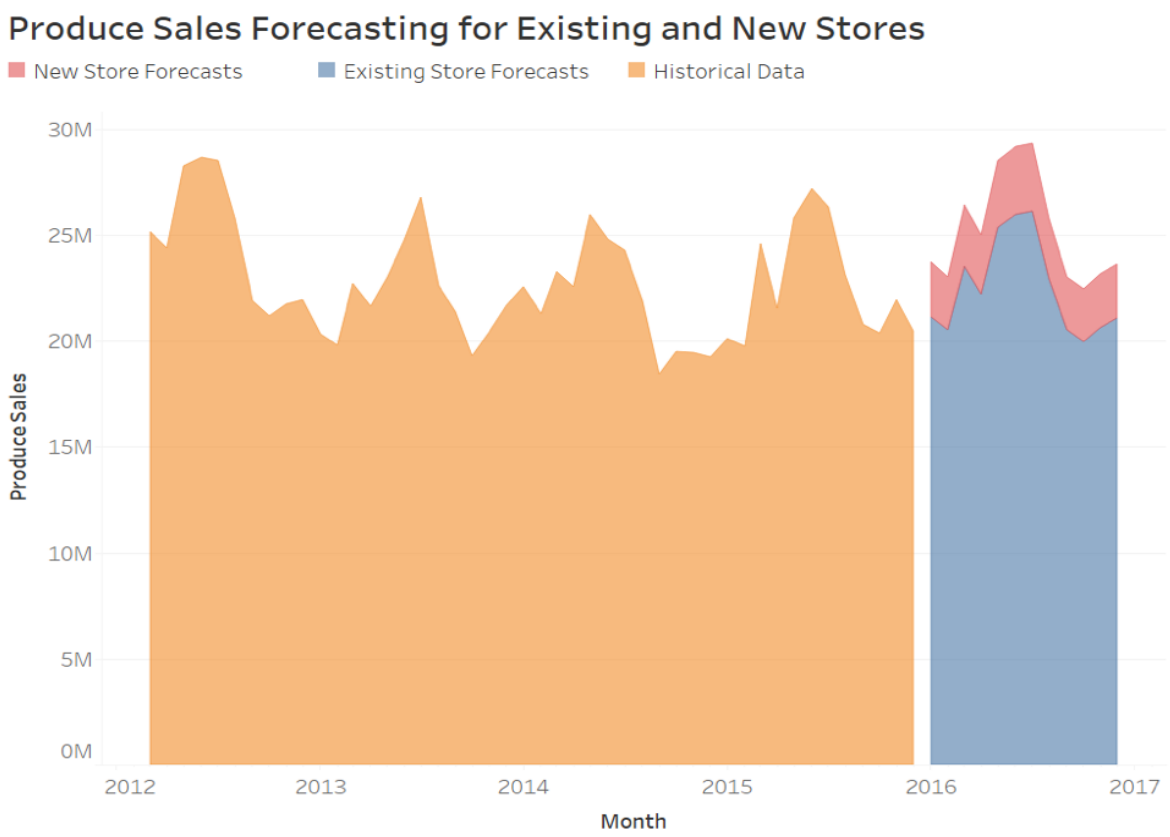
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

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) is the model that I selected to forecast the produce sales for new and existing stores.
Use storesalesdata file as the data source. Add a Summarize tool to sum the produce sales (data type: double) group by year and month.
Then from the 46 records, filter the last 6 records as a holdout sample. I use the first 40 records to train the models.
Train ETS and ARIMA models. The optimal option for ETS model is ETS(M,N,M) and for ARIMA is ARIMA(1,0,0)(1,1,0)[12].
Add TS Compare tool to obtain the forecast error measurements against the holdout sample for each model.
Compare the forecast error measurements against the holdout sample of ETS and ARIMA and select the model with the lower forecast error measurements. ETS(M,N,M) turns out to be the better one. Please see the error measures comparison below

3. 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 Store
2016-01	2,588,250	21,136,642
2016-02	2,499,159	20,507,039
2016-03	2,916,908	23,506,566
2016-04	2,791,560	22,208,406
2016-05	3,156,890	25,380,148
2016-06	3,200,940	25,966,799
2016-07	3,224,858	26,113,793
2016-08	2,861,958	22,899,286
2016-09	2,534,353	20,499,584
2016-11	2,578,336	20,602,666
2016-12	2,561,917	21,073,222



The data visualization of my forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

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.