Project: Predictive Analytics Capstone

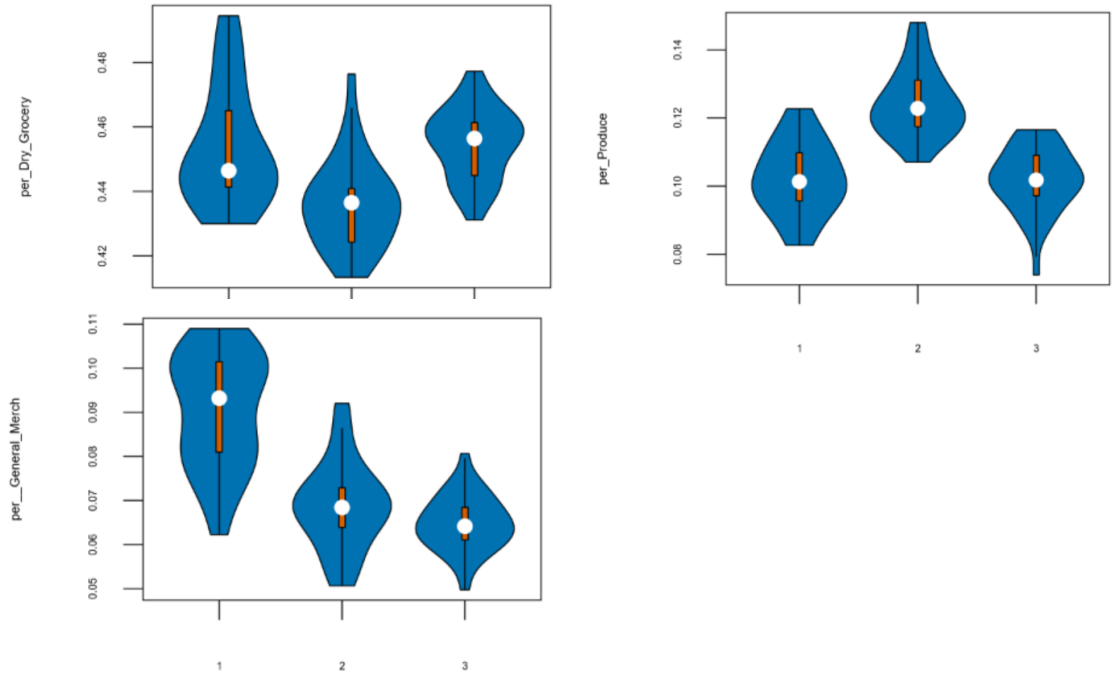
Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

Task 1: Determine Store Formats for Existing Stores

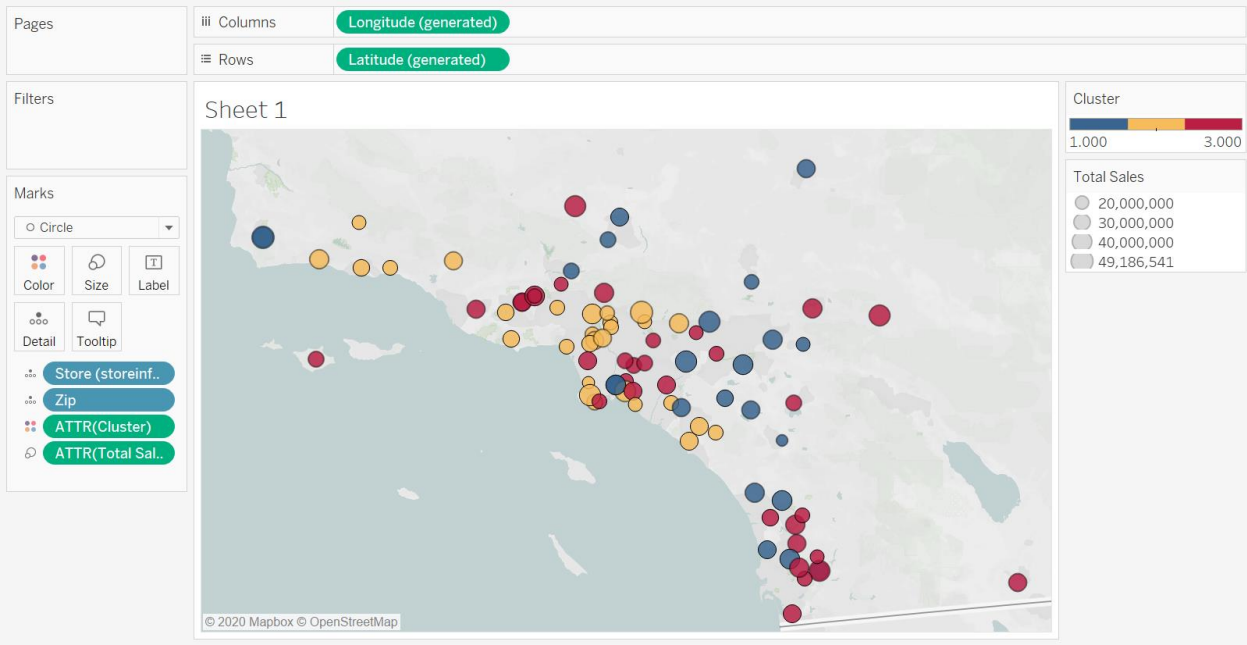
1. What is the optimal number of store formats? How did you arrive at that number?
   We applied k-means clustering, and according to the model results, 3 store formats seems to be the optimal number.

2. How many stores fall into each store format?
   23, 29, and 33 stores.

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   The clusters differ from one another in percentage sales per category per store, For example, looking at the violin plots below, we can see how the distributions of percentage sales for Dry Groceries, Produce and General Merchandise are different from one cluster to another.



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.)

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667

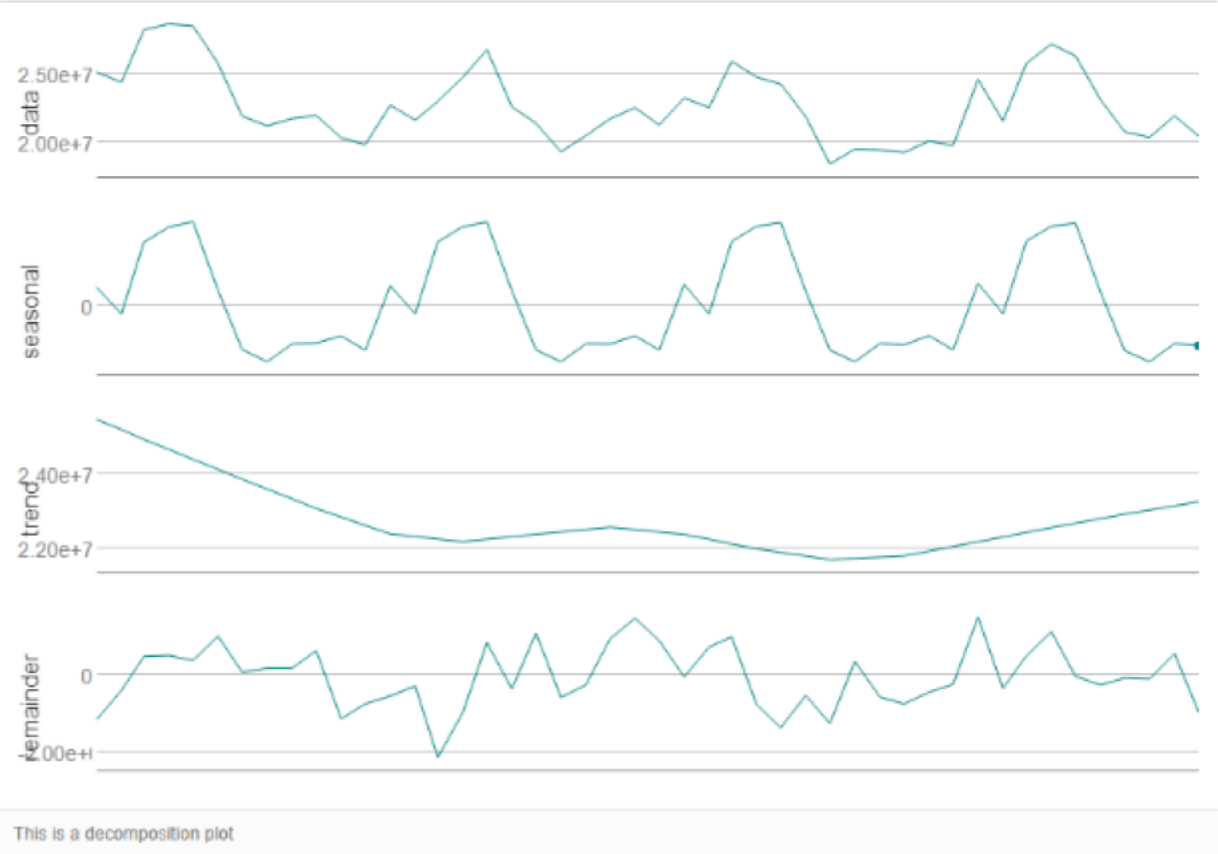
Although the Accuracy for each for the Boosted Model and the Forest Model are the same, we decided to use the Boosted model, since it has a higher F1 score.

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?



The decomposition plot suggests that we should use the Error multiplicatively, since the Remainder (error) plot does not have constant variance over time. We should not use the Trend since it changes direction multiple times within the time period. Lastly, the seasonality's peaks and valleys do not seem change that much, so we decided to try both MNM, and MNA, where the seasonality is additive in one experiment and multiplicative in the other, we compared the results and the MNM model yielded better results in the Actual vs Predicted values and Accuracy Measures table.

ETS (MNM) Actual vs Forecast Values & Accuracy Measures:

ARIMA(1,0,0)(1,1,0)[12] Actual vs Forecast Values:

Comparing Accuracy Measures for ARIMA(1,0,0)(1,1,0)[12] and ETS(MNM)

ARIMA

Actual and Forecast Values:

Actual	ARIMA
26338477.15	27997835.66704
23130626.6	23946058.0479
20774415.93	21751347.9177
20359980.58	20352513.12727
21936906.81	20971835.14524
20462899.3	21609110.45558

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	-604232.3	1050239	928412	-2.6156	4.0942	0.5463

ETS

Actual and Forecast Values:

Actual	MNM
26338477.15	26860639.57443
23130626.6	23468254.49595
20774415.93	20668464.64495
20359980.58	20054544.07631
21936906.81	20752503.51996
20462899.3	21328386.80965

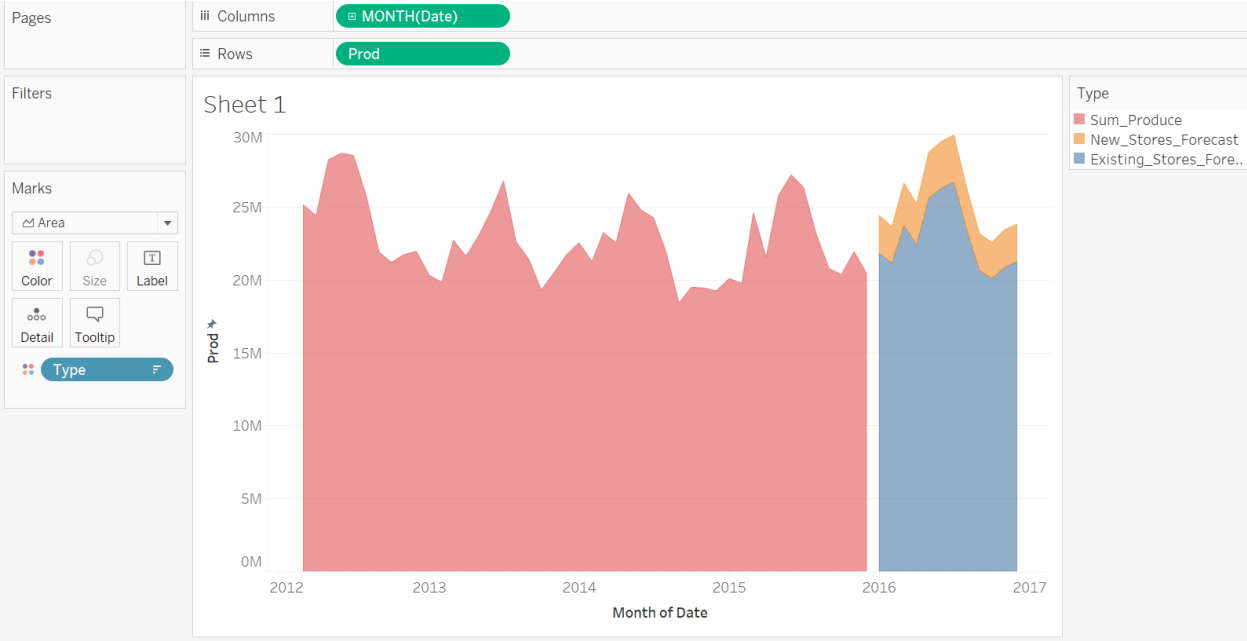
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
MNM	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257

Looking at the accuracy measures we decided that we should use the ETS(MNM) model for forecasting, since all ETS accuracy measures seem to be better than ARIMA.

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.

Year	Month	New stores (C1 * 3 + C2*6 + C3)	Existing Stores	New and Existing
2016	**1**	2588356.558	21829060.03	24417416.59
2016	**2**	2498567.174	21146329.63	23644896.81
2016	**3**	2919067.025	23735686.94	26654753.96
2016	**4**	2797280.083	22409515.28	25206795.37
2016	**5**	3163764.859	25621828.73	28785593.58
2016	**6**	3202813.289	26307858.04	29510671.33
2016	**7**	3228212.242	26705092.56	29933304.8
2016	**8**	2868914.812	23440761.33	26309676.14
2016	**9**	2538372.267	20640047.32	23178419.59
2016	**10**	2485732.285	20086270.46	22572002.75
2016	**11**	2583447.594	20858119.96	23441567.55
2016	**12**	2562181.7	21255190.24	23817371.94



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.