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?

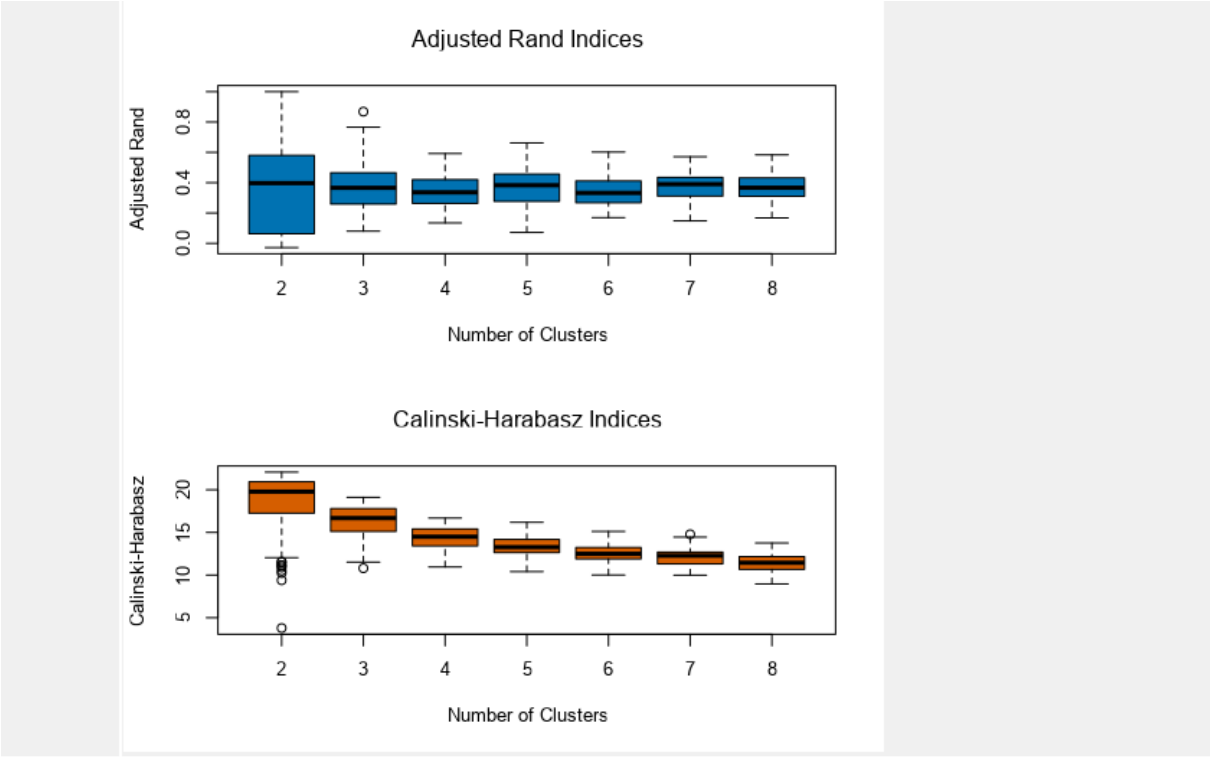
2. How many stores fall into each store format?

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

Answer:

1. The optimal number of store format is 3.
   This was because, I performed a diagnostic test using K-means to get the optimal number of clusters (segments). The results for the adjusted rand and Calinski-Harabasz as shown below:



Using the medium and spread indices to select the best cluster, we can see that

even if number 2 has the highest median, the spread is wide. The better option is to go for Number 3 which has a close median to 2 and a more compact spread. Which is better for the clustering.

2. The number of stores that fall into each clusters are as follows:

of 1 Fields | Records 1 to 9

Record Report

Summary Report of the K-Means Clustering Solution cluster

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Percent_Dry_Grocery + Percent_Dairy + Percent_Frozen_Food + Percent_Meat + Percent_Produce + Percent_Floral + Percent_Deli + Percent_Bakery + Percent_General_Merchandise, the.data)), k = 3, nre
FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

Cluster	Size	Ave Distance	Max Distance
1	23	2.320539	3.55145
2	29	2.540086	4.475132
3	33	2.115045	4.9262

Convergence after 12 iterations.

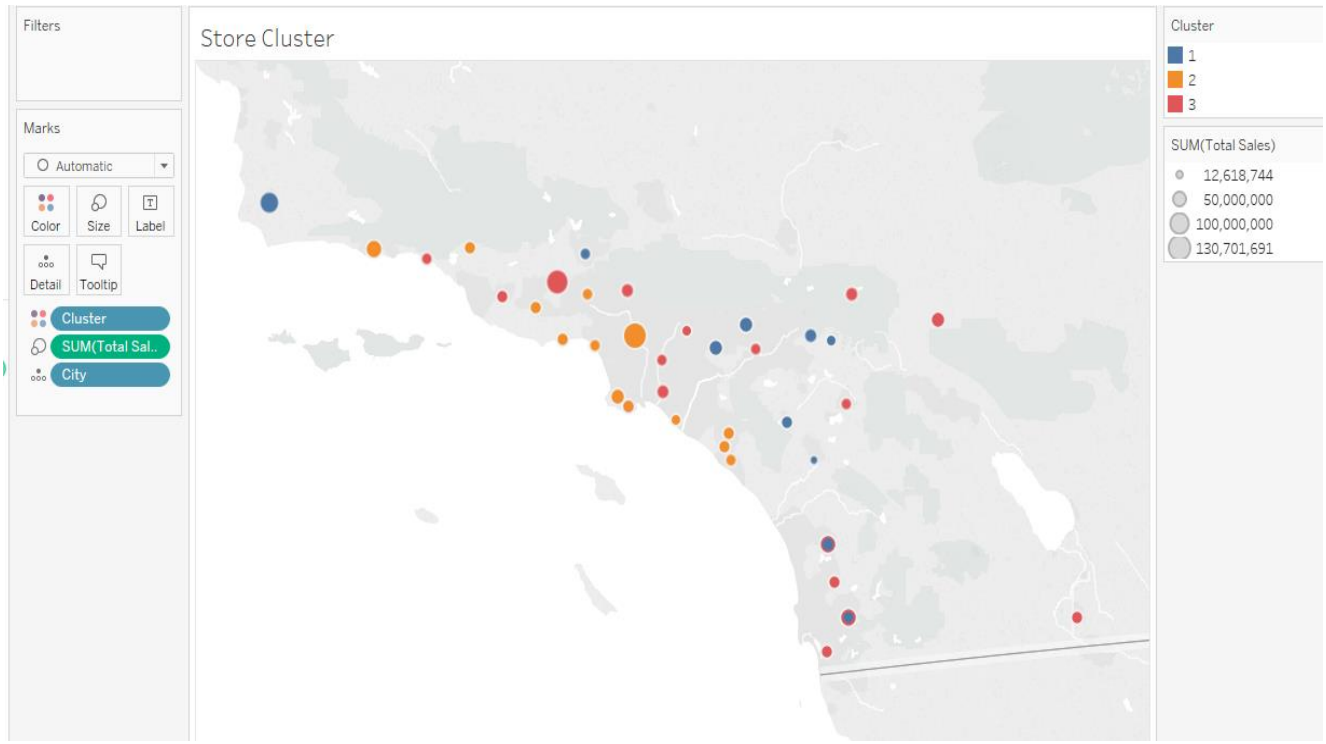
Cluster 1: 23, cluster 2: 29, and cluster 3: 33

3.

	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Perce
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-(
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	

	Percent_Bakery	Percent_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

4. From the result of the clustering model, the clusters vary if we look at the total sales in each category, For instance, looking at produce below, There is a good correlation between total sales of Floral for cluster two stores, not the same For clusters 1 & 3



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.)

ANSWER:

I used the Boosted model to predict the best store format for the three stores. After comparing the Decision tree, Forest and Boosted model, with the validation set, the Boosted model resulted in the best accuracy. As we see below, the F1 score which is a weighted average of the precision and recall is the highest for boosted trees:

of 1 Fields | Records 1 to 5

Record | Layout

1

Model Comparison Report

2

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
Boosted	0.8235	0.8889	1.0000	1.0000	0.6667
Random_Forest	0.8235	0.8426	0.7500	1.0000	0.7778

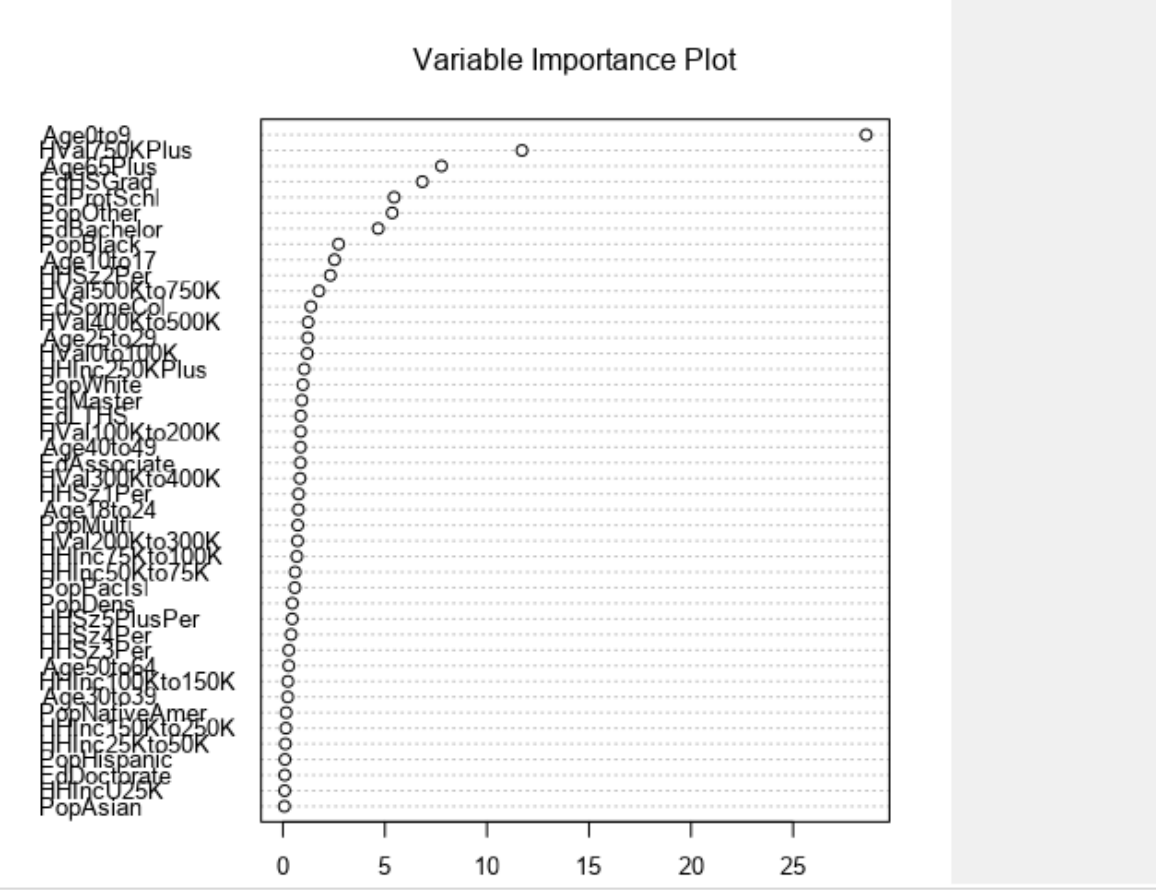
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

3

Confusion matrix of Boosted

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Variable importance plot for boosted model



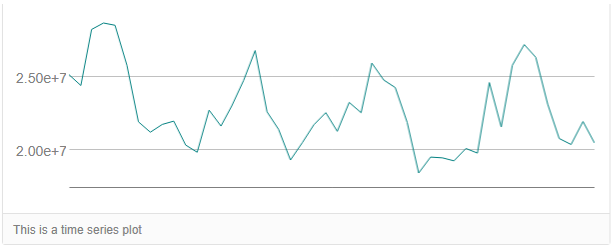
2. What format do each of the 10 new stores fall into? Please fill in the table below.

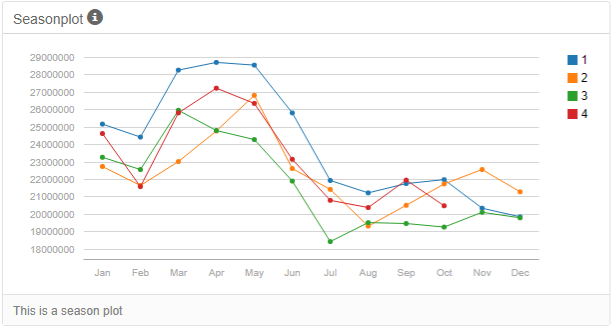
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

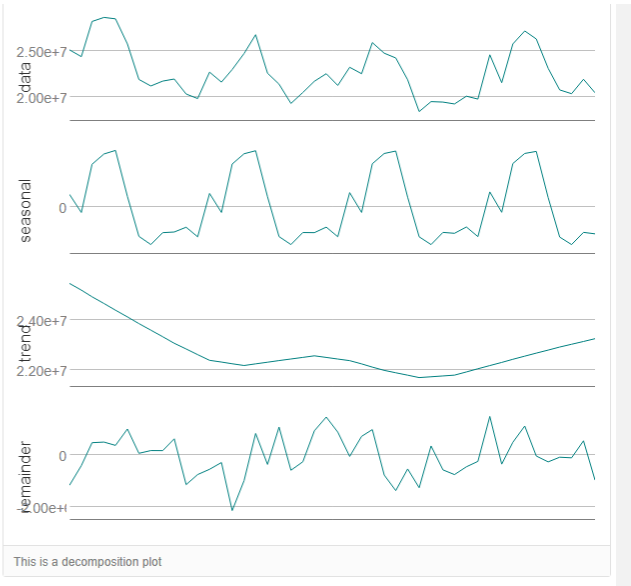
Task 3: Predicting Produce Sales

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?

ANSWER:







I used the ETS (M,N,M) . The reason I picked ETS(M,N,M) Over ARIMA (0,1,2)(0,1,1)[12] is because, even if the Arima model had a lower AIC, the ETS model produced lower errors than the ARIMA model. As shown below

Comparison of Time Series Models

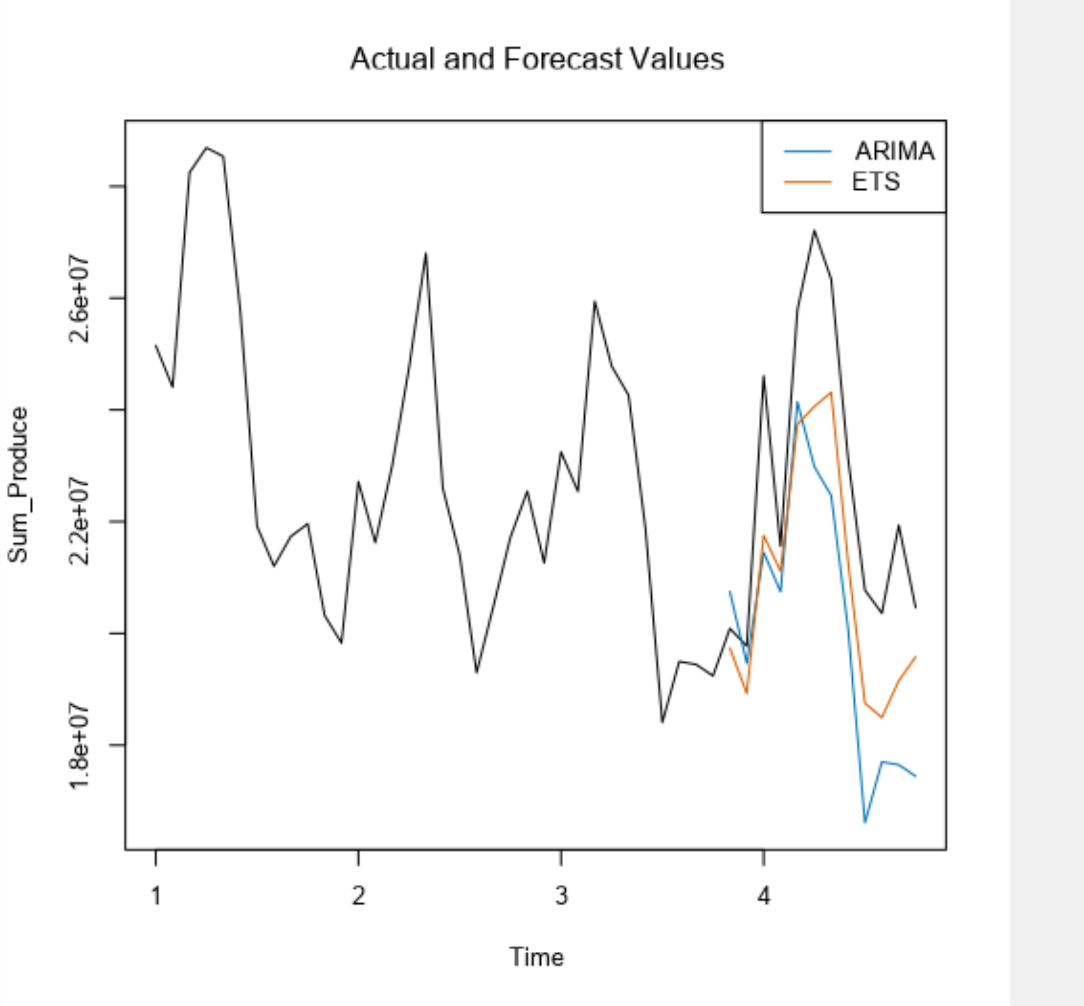
Actual and Forecast Values:

Actual	ARIMA	ETS
20088529.29	20747631.87903	19738718.27739
19772333.34	19465586.61903	18924817.83078
24608406.71	21450342.11903	21753128.98345
21559729.45	20745161.41903	21113320.07324
25792074.59	24146220.24903	23741225.58402
27212464.15	22985351.92903	24061034.3359
26338477.15	22466291.08903	24314393.56628
23130626.6	20083162.35903	21278378.06874
20774415.93	16610437.07903	18751579.10844
20359980.58	17700745.44903	18493815.39472
21936906.81	17647926.66903	19151249.71285
20462899.3	17443558.24903	19579559.13419

Accuracy Measures:

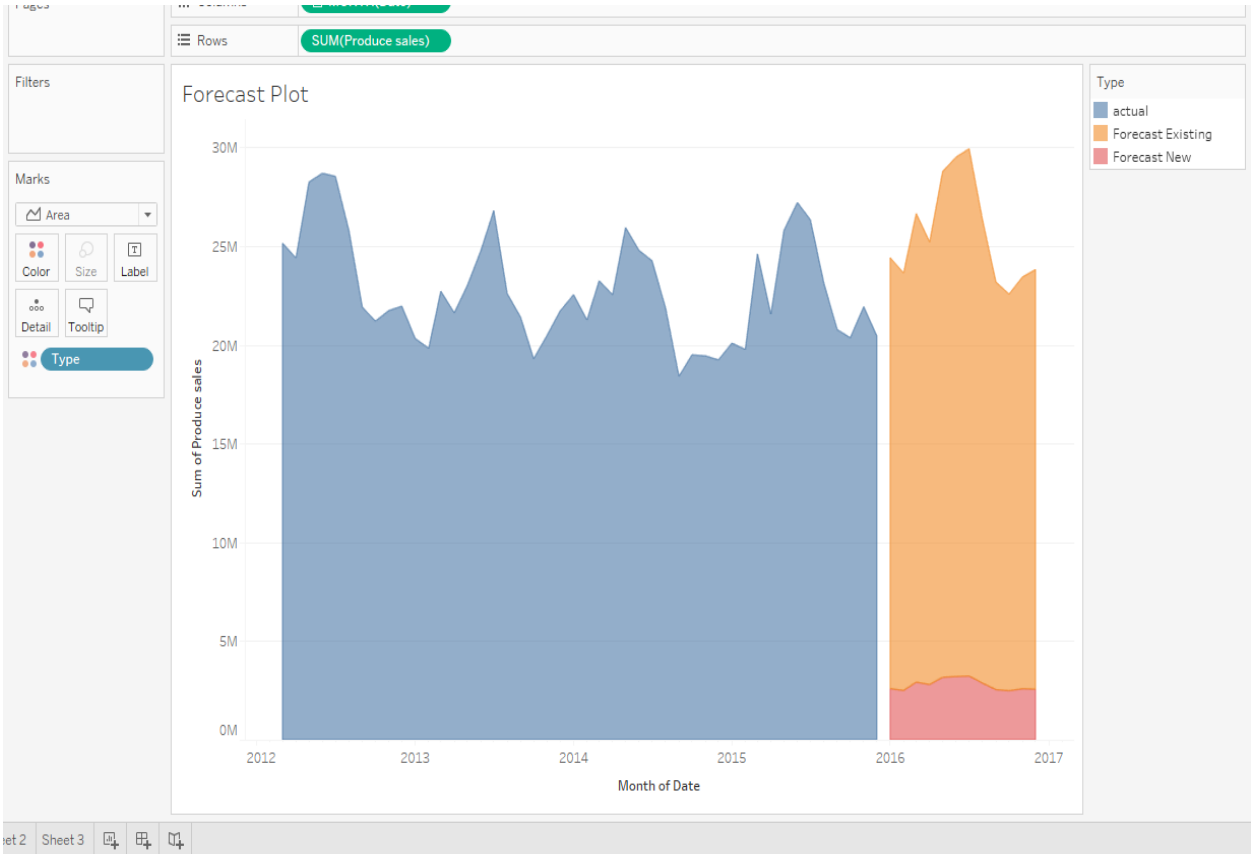
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	2545369	2999244	2655219	11.0071	11.5539	1.6988
ETS	1761302	1978476	1761302	7.5704	7.5704	1.1269

The graph below also shows that ETS follows the trend closer than the Arima model



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.

Date	Forecast_Existing_Stores	Forecast_New_stores
2016-1	21829060.03	2588356.558
2016-2	21146329.63	2498567.174
2016-3	23735686.94	2919067.025
2016-4	22409515.28	2797280.083
2016-5	25621828.73	3163764.859
2016-6	26307858.04	3202813.289
2016-7	26705092.56	3228212.242
2016-8	23440761.33	2868914.812
2016-9	20640047.32	2538372.267
2016-10	20086270.46	2485732.285
2016-11	20858119.96	2583447.594
2016-12	21255190.24	2562181.7



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.