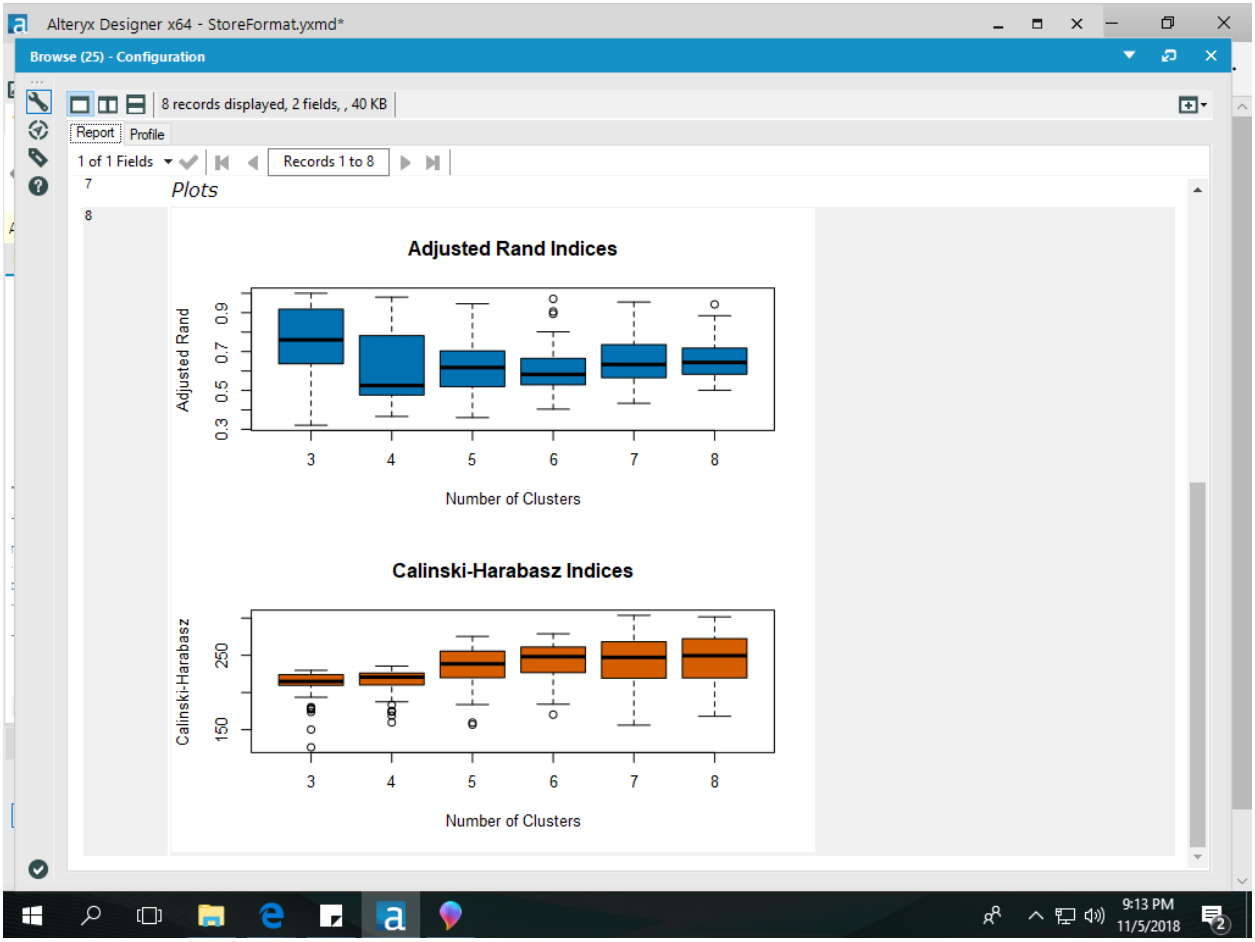
Project: Predictive Analytics Capstone

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 is 3. Adjusted Rand indices and Calinski-Harabasz indices indicates the optimal cluster is the one having higher median smaller variation. Looking athe the box whisker plots below, 3 seems to be the optimal number of cluster.

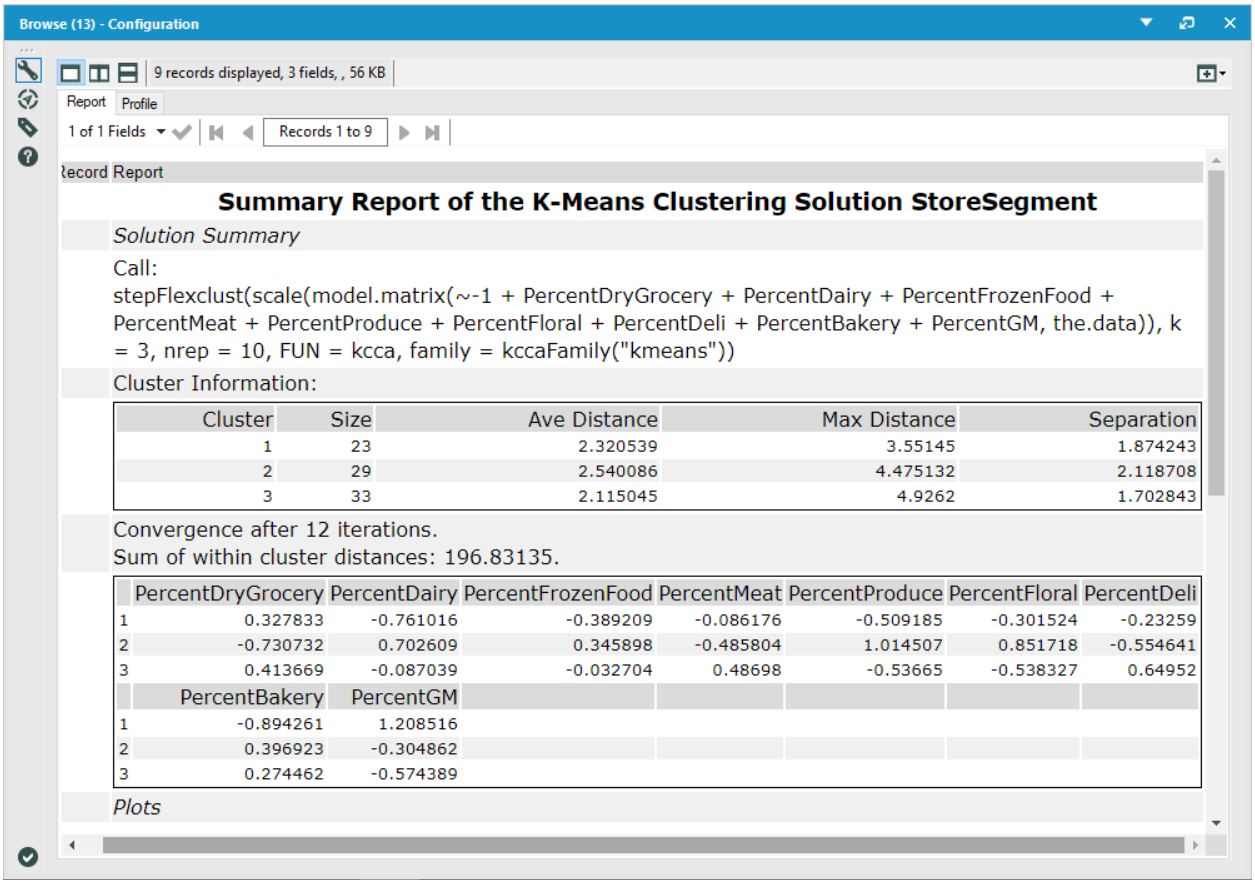


2. How many stores fall into each store format?

   23 stores fall on Cluster 1, 29 on Cluster 2 and 33 stores on Cluster 3 as depicted by the K means clustering below.

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

one another?

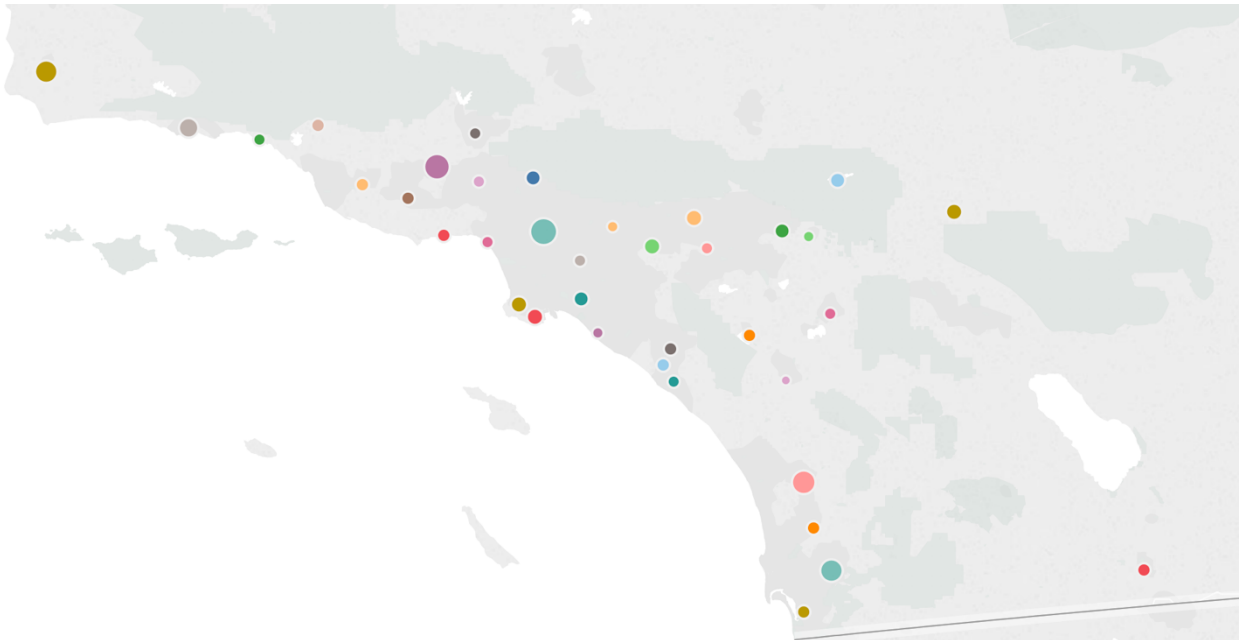


The clusters differ mainly in the size and the separation

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.

   https://public.tableau.com/profile/jaya.k1740#!/vizhome/JK-CapstoneProject-StoreInfoSales/Sheet1?publish=yes

   It will contain a graph like below indicating the city,storeid and the total sales of the store in the city.

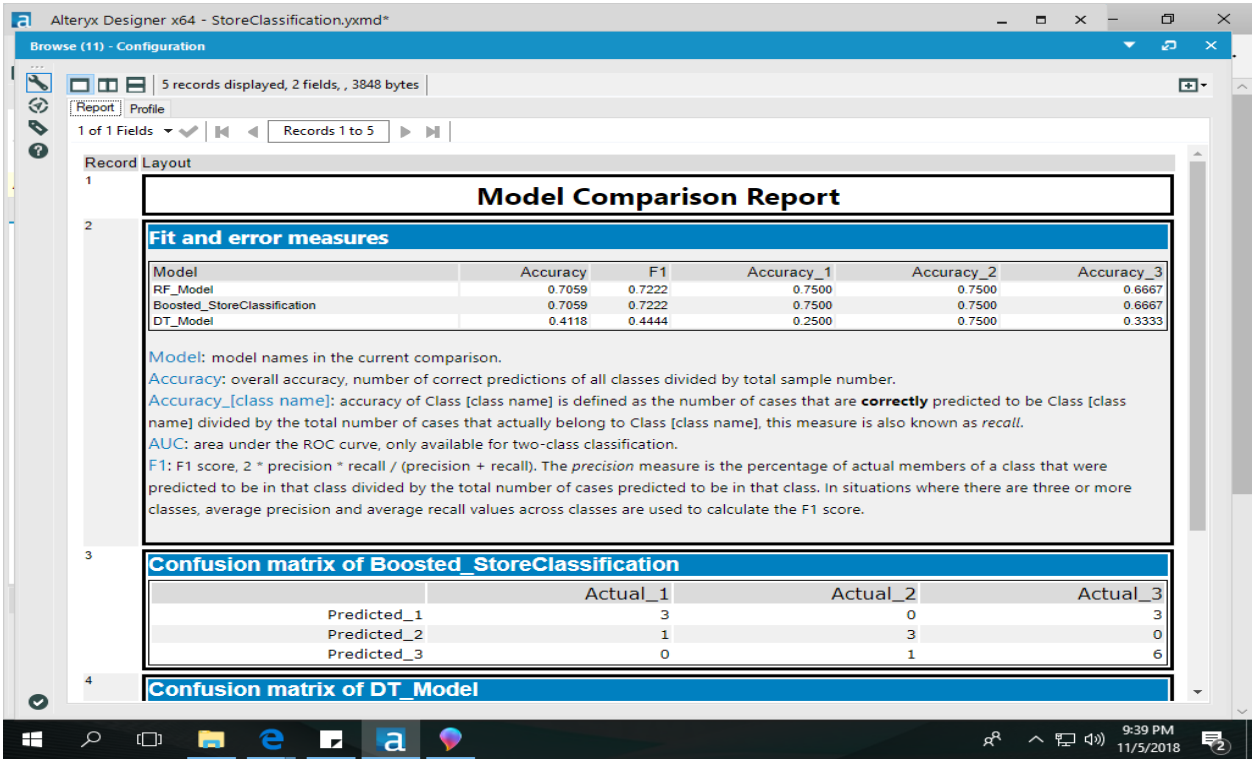


Task 2: Formats for New Stores

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.)

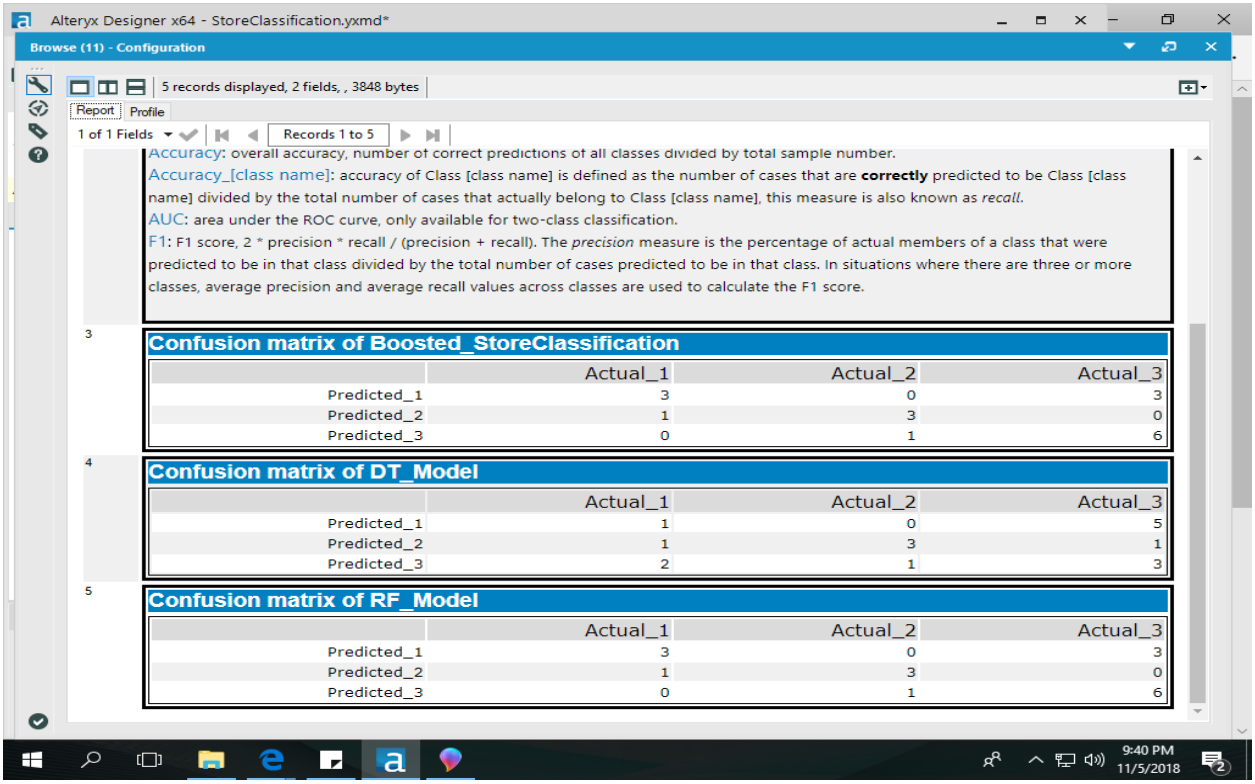
   Used the Decision Tree, Random Forrest and Boosted Model to check the validation of the sample. Below are the model comparison report. Decision Tree gave accuracy of 41.18%, Random Forrest 70.5 % and Boosted Model gave 70.5% accuracy. Based on the data, chose the Random Forrest to classify the new stores.

   Below is the confusion matrix of all the 3 models:



Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
RF_Model	0.7059	0.7222	0.7500	0.7500	0.6667
Boosted_StoreClassification	0.7059	0.7222	0.7500	0.7500	0.6667
DT_Model	0.4118	0.4444	0.2500	0.7500	0.3333

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	3
Predicted_2	1	3	0
Predicted_3	0	1	6



	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	3
Predicted_2	1	3	0
Predicted_3	0	1	6

	Actual_1	Actual_2	Actual_3
Predicted_1	1	0	5
Predicted_2	1	3	1
Predicted_3	2	1	3

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	3
Predicted_2	1	3	0
Predicted_3	0	1	6

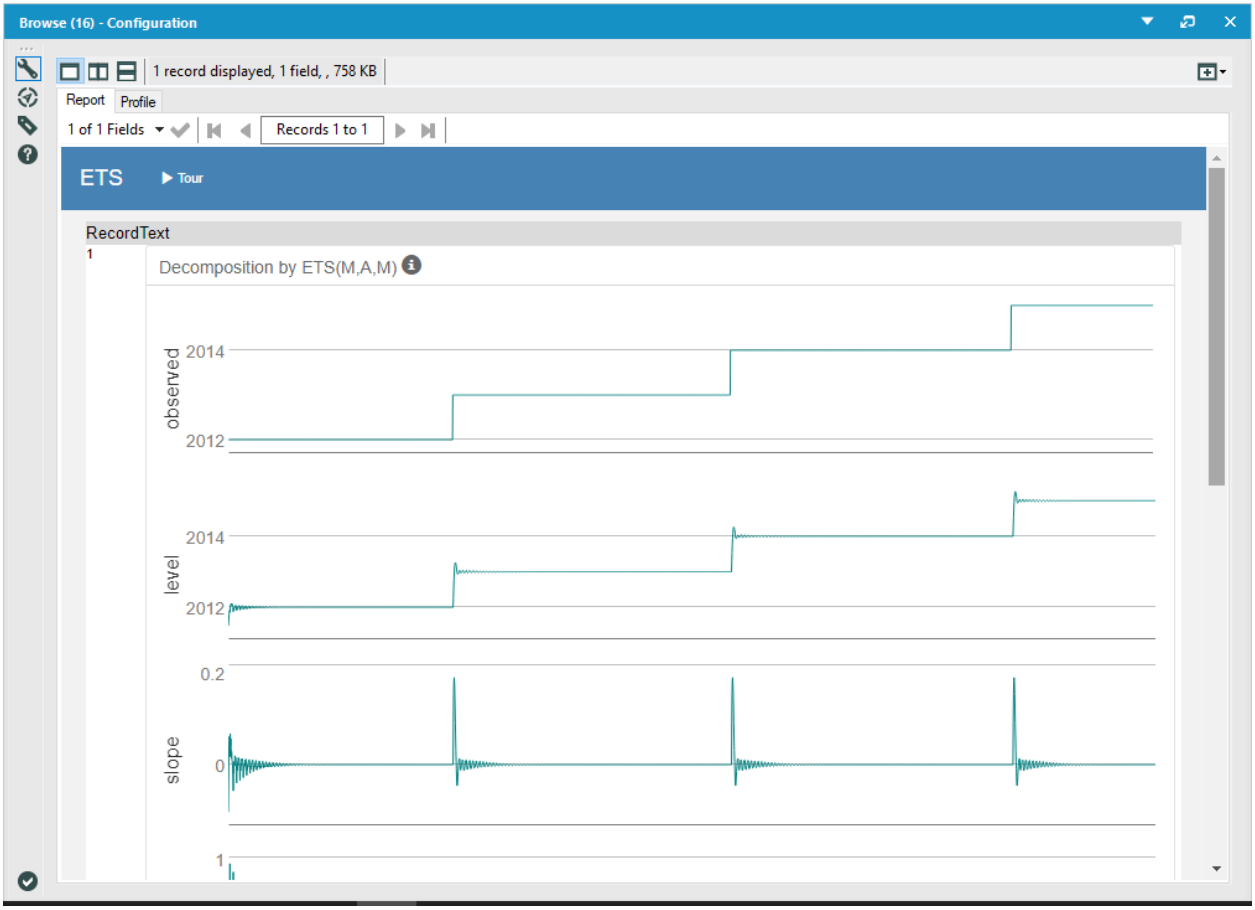
2. What format do each of the 10 new stores fall into? Please fill in the table below.
   Used the Random Forrest Model to classify the new stores and found the following result:

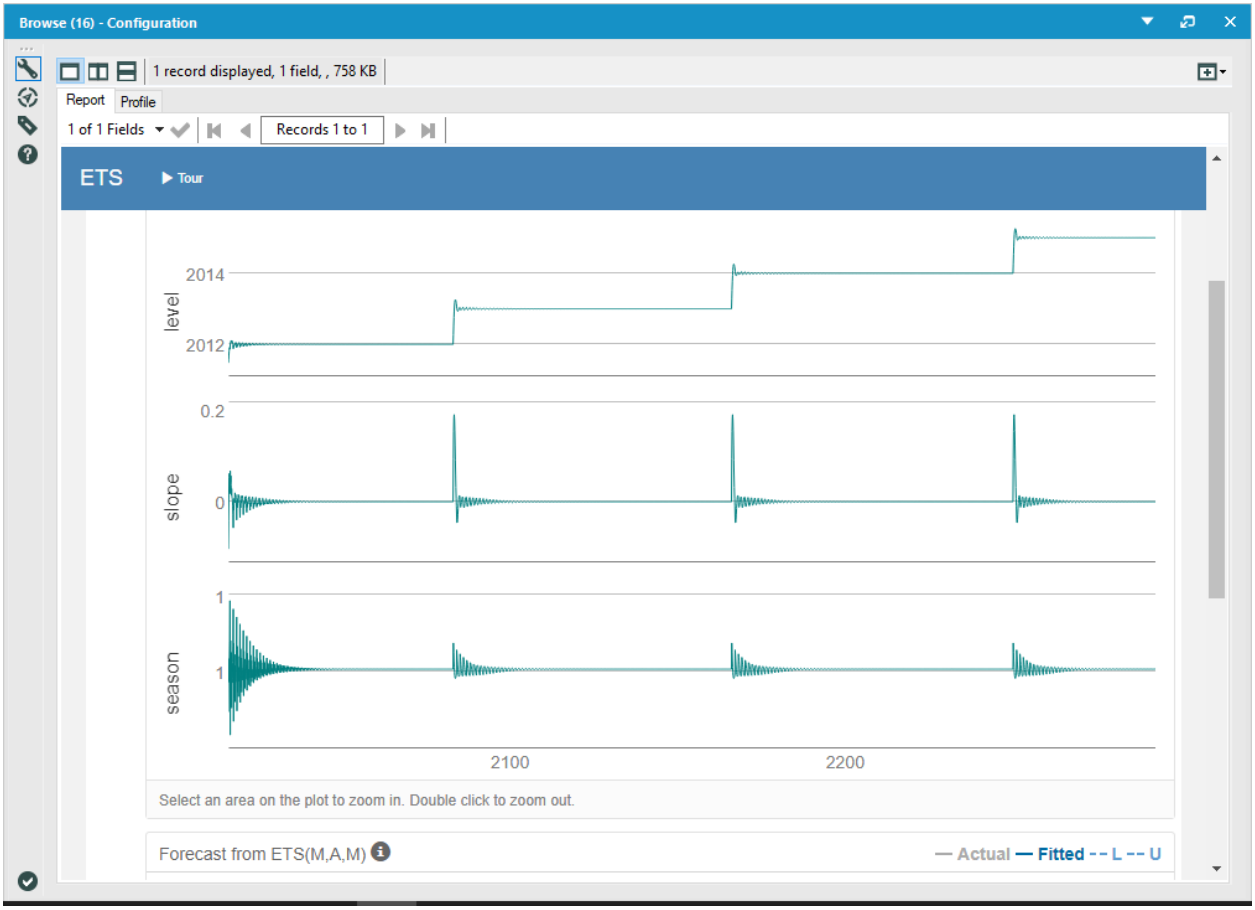
Store Number	Segment
S0086	3
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	3

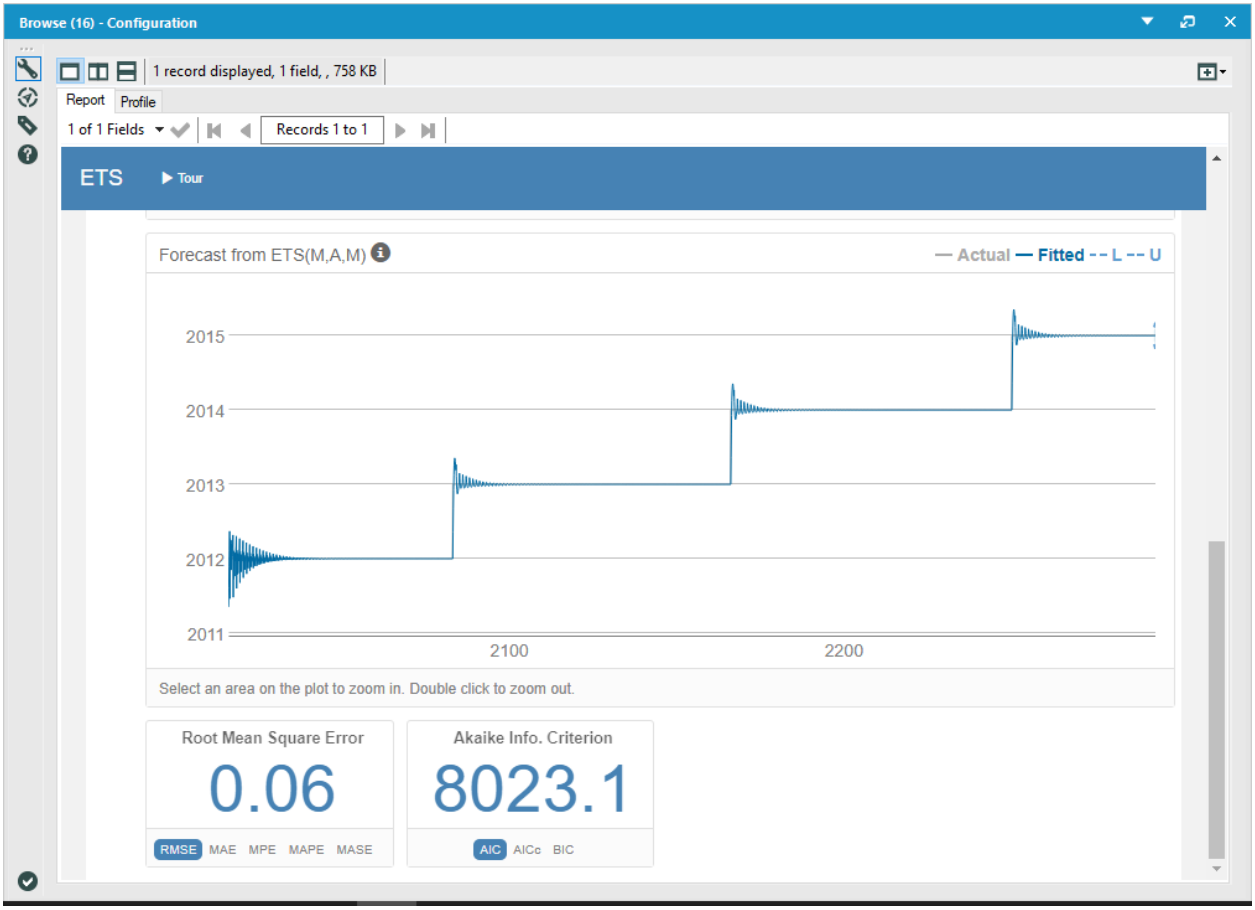
Task 3: Predicting Produce Sales

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?

Got the following for ETS :

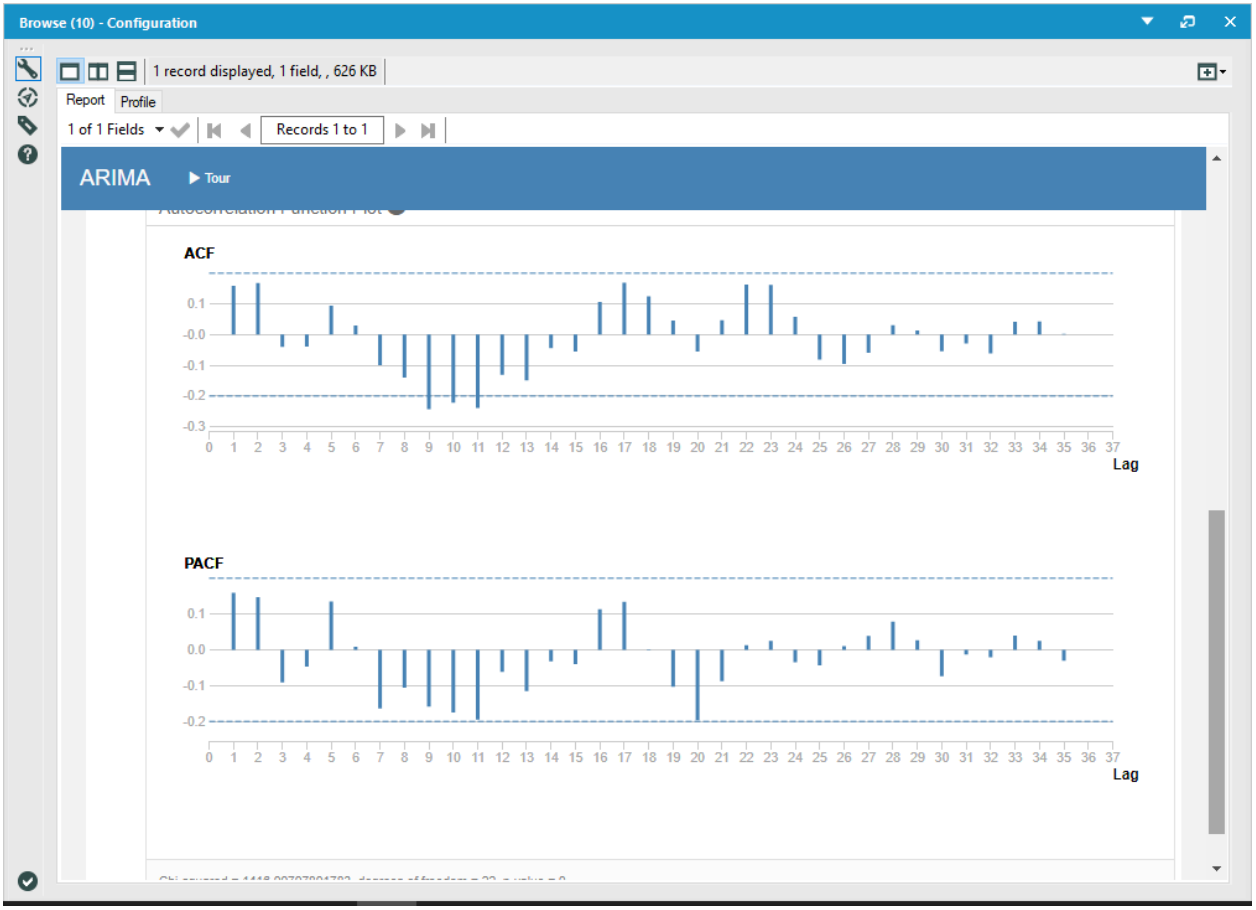

Browse (16) - Configuration
1 record displayed, 1 field, , 758 KB
Report
Profile
1 of 1 Fields
Records 1 to 1
ETS
Tour
RecordText
1
Decomposition by ETS(M,A,M)
observed
2014
2012
level
2014
2012
slope
0.2
0
1


Browse (16) - Configuration
1 record displayed, 1 field, , 758 KB
Report
Profile
1 of 1 Fields
Records 1 to 1
ETS
Tour
level
2014
2012
slope
0.2
0
season
1
1
2100
2200
Select an area on the plot to zoom in. Double click to zoom out.
Forecast from ETS(M,A,M)
Actual
Fitted
L
U



Got the following for ARIMA:


Browse (10) - Configuration
1 record displayed, 1 field, , 626 KB
Report
Profile
1 of 1 Fields
Records 1 to 1
ARIMA
Tour
Forecast from ARIMA(0,1,1)(0,1,1)[12]
— Actual — Fitted -- L -- U
3.5
3
2.5
2
1.5
1
0.5
2100
2200
Select an area on the plot to zoom in. Double click to zoom out.
Root Mean Square Error
0.72
Akaike Info. Criterion
7309.1



The AIC of ETS is slightly higher than that of Arima but the RMSE is significantly low at 6%. The AIC is comparable for both ETS and ARIMA. Since the RMSE is significantly low, I went ahead with the ETS model.

Following is the decomposition plot for ETS:

Browse (11) - Configuration

9 records displayed, 2 fields, , 45 KB

Report | Profile

1 of 1 Fields | Records 1 to 9

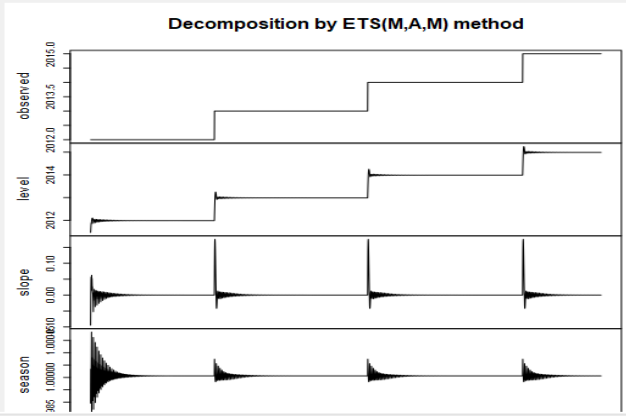
Record Report

1

Plots of Time Series Exponential Smoothing Model ETS_Produce

In statistics, a time series is a sequence of data points measured at successive points in time spaced at uniform intervals. Examples of time series are the daily closing value of a stock market index or the annual flow volume of a river. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.

2



Decomposition Plot separates time series data into several components. Decomposition method is often used to yield information about time series components i.e. trend, cycle, seasonal, etc.

- Observed: This is the actual data.
- Level: This is the overal baseline without seasonal trends.
- Slope: This is the rate of change associated with the Level.
- Season: This shows the seasonal trend of the data.

Not all of the above components will occur each time.

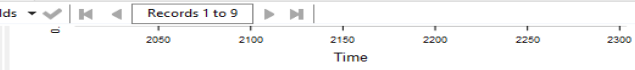
5:18 PM
11/3/2018

Browse (11) - Configuration

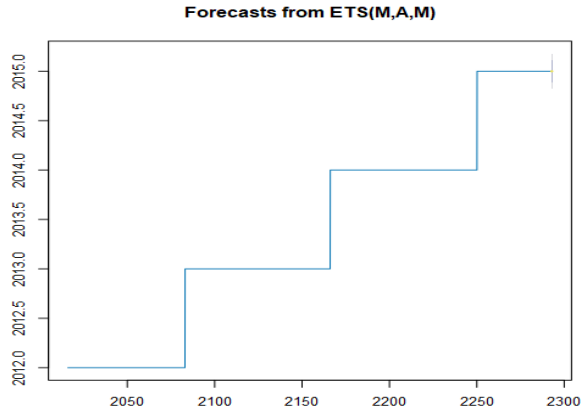
9 records displayed, 2 fields, , 45 KB

Report | Profile

1 of 1 Fields | Records 1 to 9

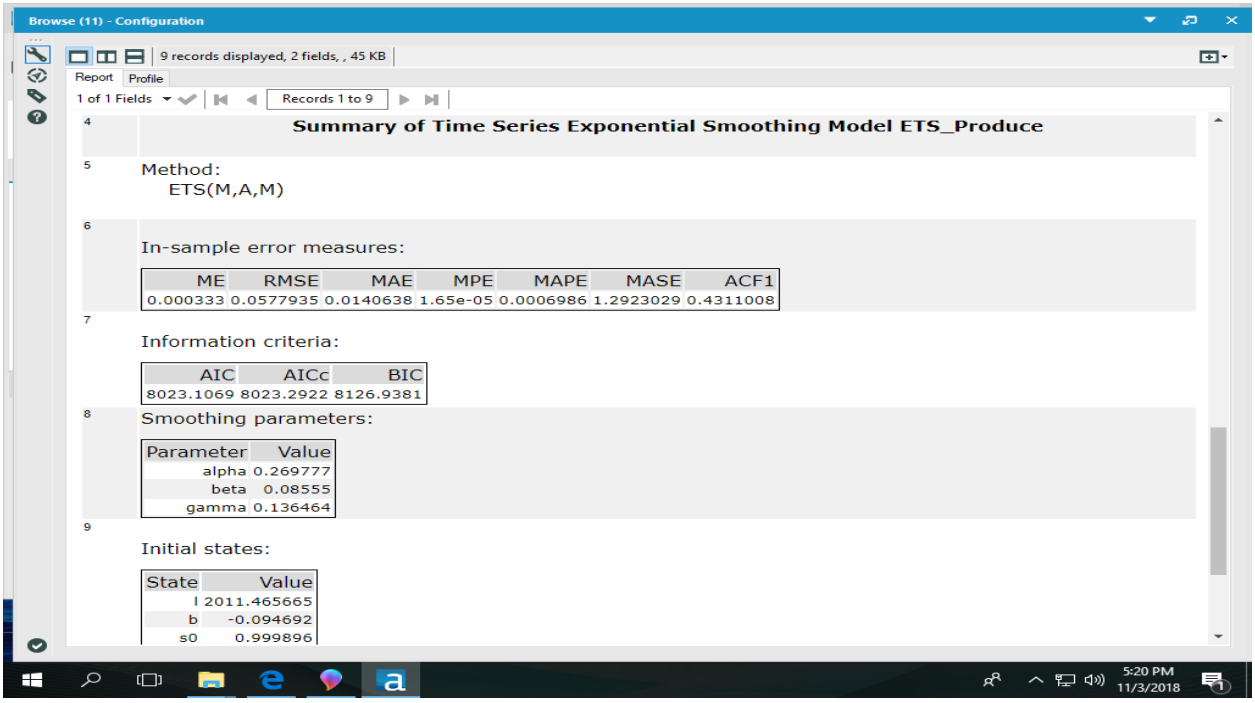


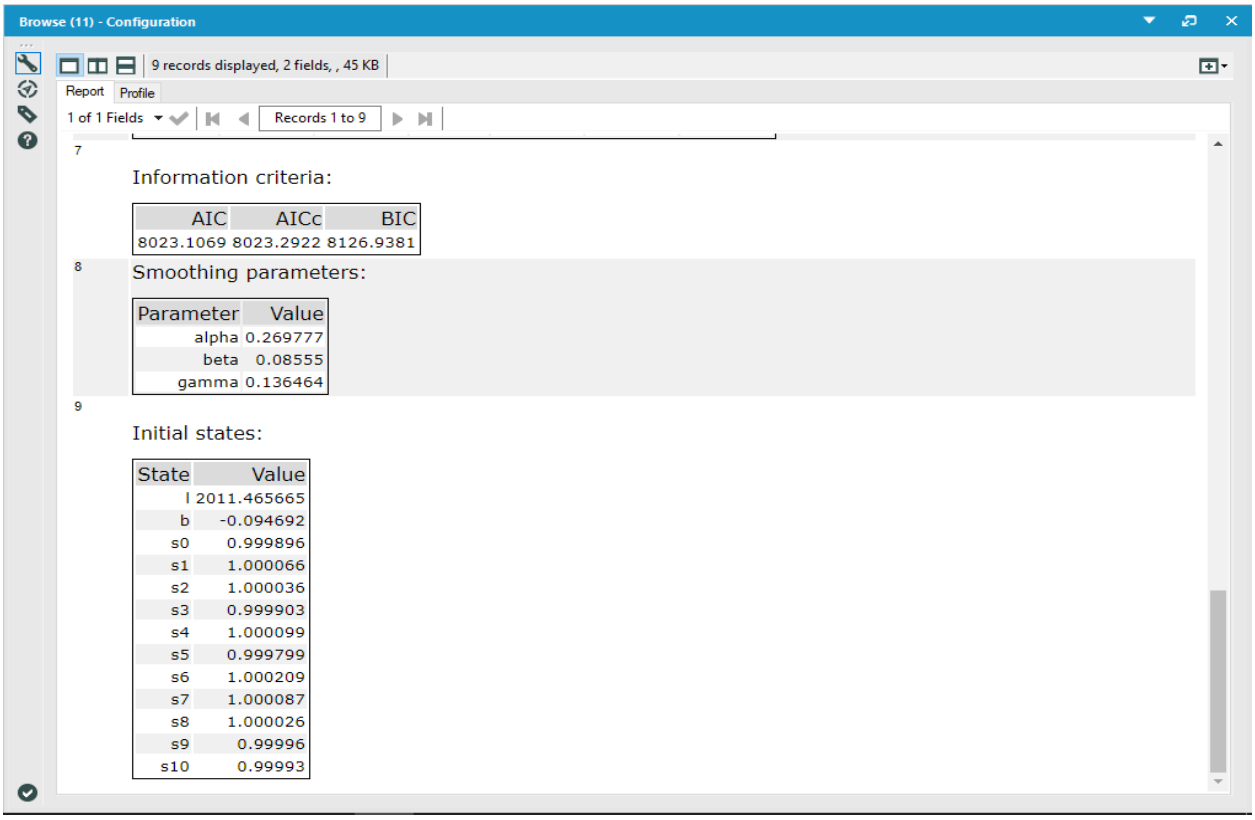
3



The Forecast Plot shows the historic data in black and the expected value in blue. The orange in the plot shows the 90% confidence interval, and the yellow shows the 95% confidence interval.

5:19 PM
11/3/2018


Browse (11) - Configuration
9 records displayed, 2 fields, , 45 KB
Report
Profile
1 of 1 Fields
Records 1 to 9
Summary of Time Series Exponential Smoothing Model ETS_Produce
Method:
ETS(M,A,M)
In-sample error measures:
ME RMSE MAE MPE MAPE MASE ACF1
0.000333 0.0577935 0.0140638 1.65e-05 0.0006986 1.2923029 0.4311008
Information criteria:
AIC AICc BIC
8023.1069 8023.2922 8126.9381
Smoothing parameters:
Parameter Value
alpha 0.269777
beta 0.08555
gamma 0.136464
Initial states:
State Value
l 2011.465665
b -0.094692
s0 0.999896
5:20 PM
11/3/2018


Browse (11) - Configuration
9 records displayed, 2 fields, , 45 KB
Report
Profile
1 of 1 Fields
Records 1 to 9
Information criteria:
AIC AICc BIC
8023.1069 8023.2922 8126.9381
Smoothing parameters:
Parameter Value
alpha 0.269777
beta 0.08555
gamma 0.136464
Initial states:
State Value
l 2011.465665
b -0.094692
s0 0.999896
s1 1.000066
s2 1.000036
s3 0.999903
s4 1.000099
s5 0.999799
s6 1.000209
s7 1.000087
s8 1.000026
s9 0.99996
s10 0.99993

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.

Store	Forecasted 2016
S0001	2284388.7
S0002	1755293.15
S0003	3792143.03
S0004	2802811.88
S0005	3648309.02
S0006	3524245.59
S0007	3192688.95
S0008	3609692.16
S0009	4210541.86
S0010	2639823.51
S0011	2205228.62
S0012	4639016.01
S0013	3629467.14
S0014	3311035.35
S0015	3921200.02
S0016	2065999.38
S0017	4215854.25
S0018	2646907.99
S0019	1865858.21
S0020	2615310
S0021	3217947.64
S0022	2625382.14
S0023	1893525.06
S0024	2709800.59
S0025	3498500.3
S0026	2411984.05
S0027	3511115.94
S0028	2017244.48
S0029	3333528.94
S0030	2965450.25
S0031	2575738.86
S0032	2331785.09
S0033	3730920.43
S0034	3561114.71
S0035	1615522.5

S0036	1748605.79
S0037	3562066.35
S0038	2171273.05
S0039	1579637.47
S0040	4543001.51
S0041	2402327.77
S0042	3843105.62
S0043	3136766.86
S0044	3735116.2
S0045	4446110.8
S0046	5002506.61
S0047	4593568.85
S0048	3440749.53
S0049	2510813.07
S0050	6382729.42
S0051	3917645.91
S0052	6637175.18
S0053	2745412.25
S0054	4386911.15
S0055	4400262.02
S0056	2572158.37
S0057	3587804.57
S0058	2897813.17
S0059	3490732.83
S0060	3516216.8
S0061	4329100.12
S0062	2882942.49
S0063	2822801.12
S0064	2806409.24
S0065	4244599.91
S0066	3021856.3
S0067	2926805.74
S0068	4775191.33
S0069	3279238.46
S0070	4013986.34
S0071	3135443.78
S0072	4147374.1
S0073	4465628.93
S0074	1908571.21
S0075	3567736.58

S0076	2605311.13
S0077	3178498.13
S0078	2343553.96
S0079	2349084.45
S0080	1042516.44
S0081	1715208.66
S0082	3544714.74
S0083	2231752.09
S0084	2173133.49
S0085	2699498.21
S0086	4223874.61
S0087	2578230.89
S0088	2906224.87
S0089	2585732.82
S0090	2663185.25
S0091	1447863.75
S0092	3198681.11
S0093	3791742.27
S0094	2227386.25
S0095	2373096.17

The tableau link for the visualization of the stores showing the existing total and forecast for all 95 stores.

https://public.tableau.com/profile/jaya.k1740#!/vizhome/JK-CapstoneProject2/Sheet1?publish=yes

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.