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. After performing k-means cluster diagnostics, I found that the compactness (interquartile range) and distinctness (median) is best for number of clusters =3. Using the median and spread of the Adjusted Rand and CH (Calinski-Harabasz) indices, it is clear that 3 clusters is the most optimal number of store formats because the box-whisker plots show how tight the indices for each data point are within each other. The below box and whisker plots demonstrate that cluster 3 has the best combination of maximum median value and least interquartile range.

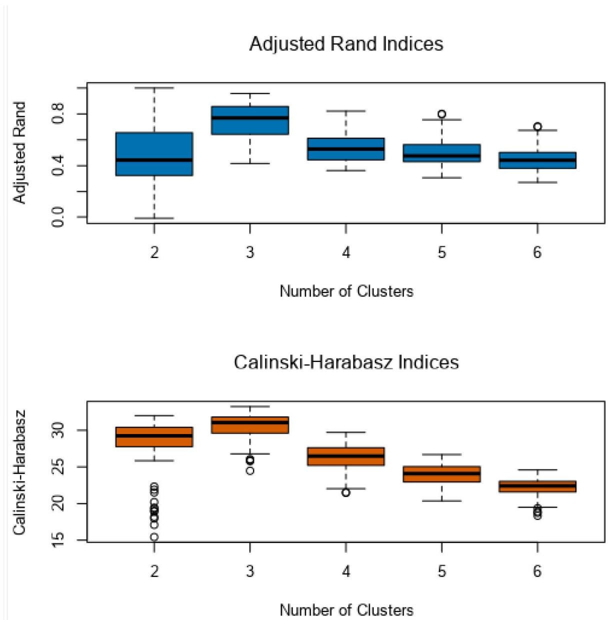


Fig 1- Box and Whisker Plot of AR and CH Indices

2. How many stores fall into each store format?

   Store Format 1(**Cluster 1**) has **23**, Store Format 2(**Cluster 2**) has **29** and Store Format 3(**Cluster 3**) has **33** stores, respectively.

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	Perc_Dry_Grocery	Perc_Diary	Perc_Sum_Frozen_Food	Perc_Sum_Meat	Perc_Sum_Produce	Perc_Sum_Floral	Perc_Sum_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Perc_Sum_Bakery	Perc_Sum_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Figure 2- K-means Cluster Diagnostics

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

   Each of the clusters differ from one another in terms of percentage of sales in each product category. The above cluster model results in Figure 2 shows that cluster 1 has the maximum percentage sales for the product category General Merchandise. Similarly, cluster 2 has leading product sales in Produce and Cluster 3 has maximum product sales for category Deli.

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.

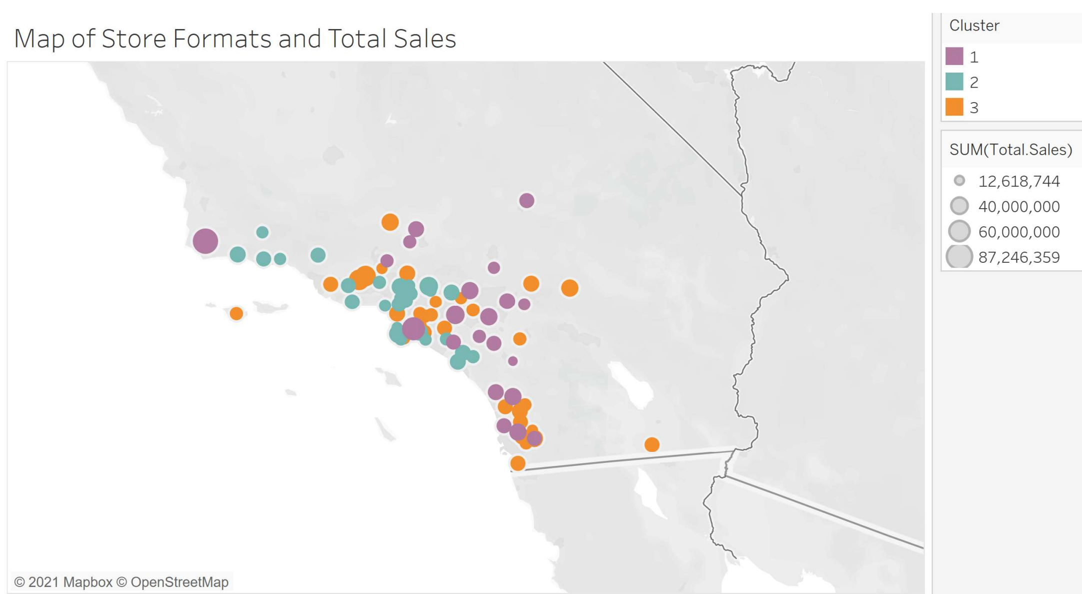
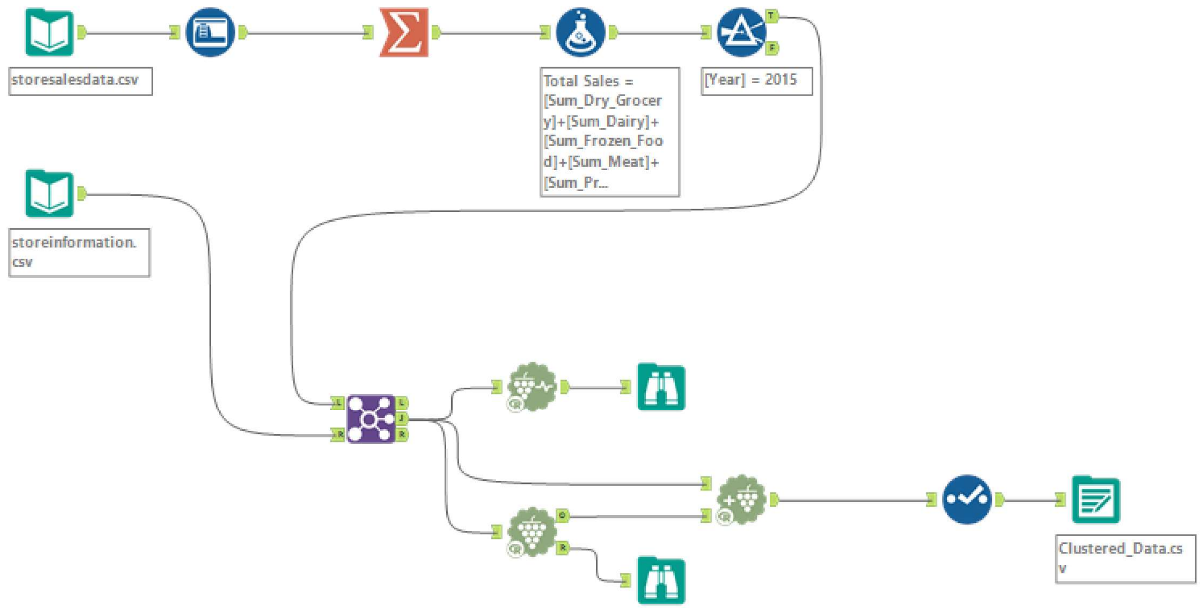


Fig 3- Tableau Visualization of Clustered Store Formats

Entire workflow for Task 1



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.)

   I selected Boosted model over Random Forest model since it has the best F1 score of 85.43% given both have same overall accuracy. Also, from the confusion matrix we can see that boosted model predicted 100% correctly for cluster 1 and cluster 2, however random forest model falls short in predicting correctly for cluster 1.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7327	0.6000	0.6667	0.8333
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000
Forest_Model	0.8235	0.8251	0.7500	0.8000	0.8750

Fig 4- Model Comparison

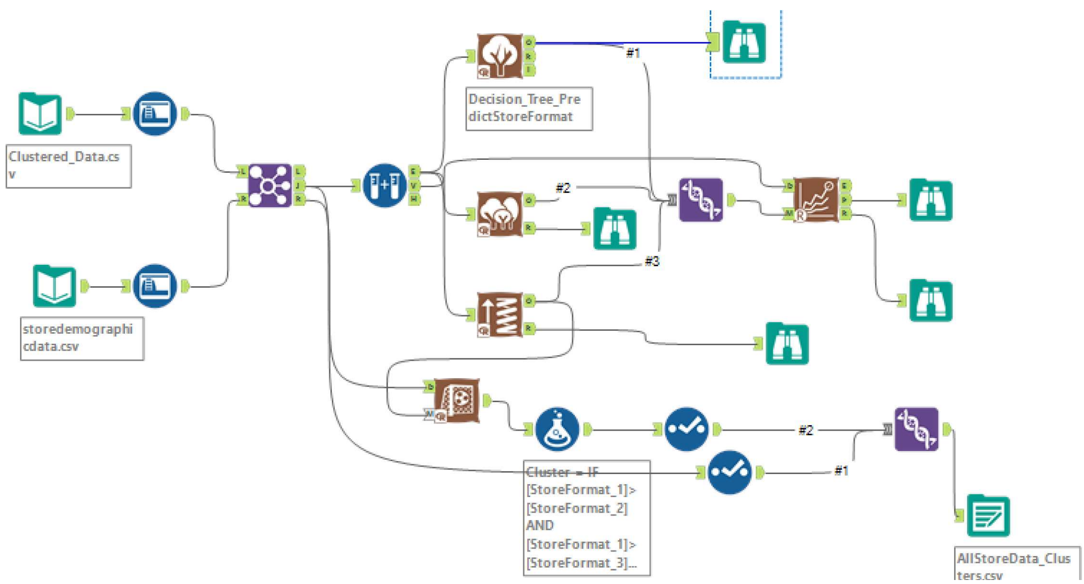
Confusion matrix of Boosted_Model			
	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Confusion matrix of Decision_Tree			
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of Forest_Model			
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

Fig 5- Confusion Matrix

Entire Workflow for Task 2



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

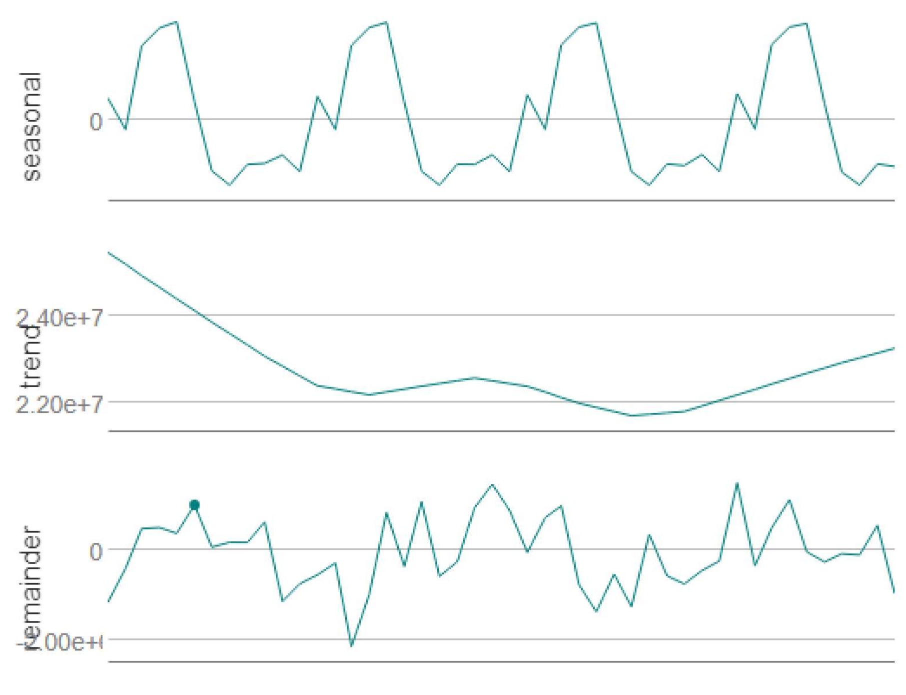
Task 3: Predicting Produce Sales

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?

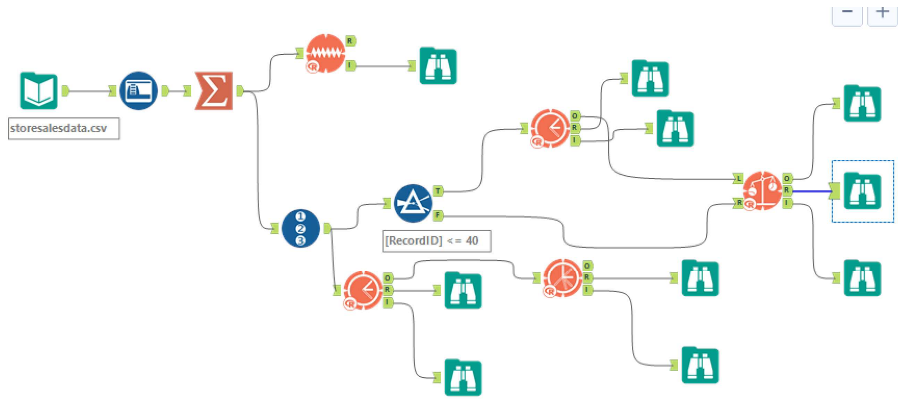
I have compared both ETS and ARIMA models using TS Compare tool to decide the best model for forecasting produce sales.

Before building ETS model, I used time series decomposition plot to observe error, trend and seasonality of a time series. The error line is fluctuating between high and low values, so a multiplicative method will be used. The trend line does not demonstrate linear plot hence I have used a None. The seasonality changes in magnitude with the time series so a multiplicative method should be used. Hence, we have an ETS(M, N, M) model.

The error, trend and seasonal plots can be seen in the below figure.



ETS Model Alteryx Workflow



The results from the ETS model are impressive. Below figures show the in-sample measures, AIC value and comparison of actual and fitted values. We will select the the ETS model with minimum AIC value =1279 to compare it with the next section of building an ARIMA model.

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3502.9443415	969051.6076376	787577.7006835	-0.1381187	3.4677635	0.4396486	0.0077488

Information criteria:

AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

Smoothing parameters:

Parameter	Value
alpha	0.674884
gamma	0.000203

Fig 6 -ETS model measures

Fig 7- Forecast by ETS model

Actual and Forecast Values:

Actual	ETS
26338477.15	26860639.57444
23130626.6	23468254.49595
20774415.93	20668464.64495
20359980.58	20054544.07631
21936906.81	20752503.51996
20462899.3	21328386.80965

Fig 8- Actual and ETS Forecasted Values

For the ARIMA model, I checked that our time series has a trend or seasonality component, so it must be made stationary before we can use ARIMA model to forecast. After differencing the lags, I found that we have seasonal autocorrelation negative suggesting MA model. Hence, I used an ARIMA (0, 1, 1) (0, 1, 1)12 since there we have seasonal components found in the time series.

The below plots show this behavior.

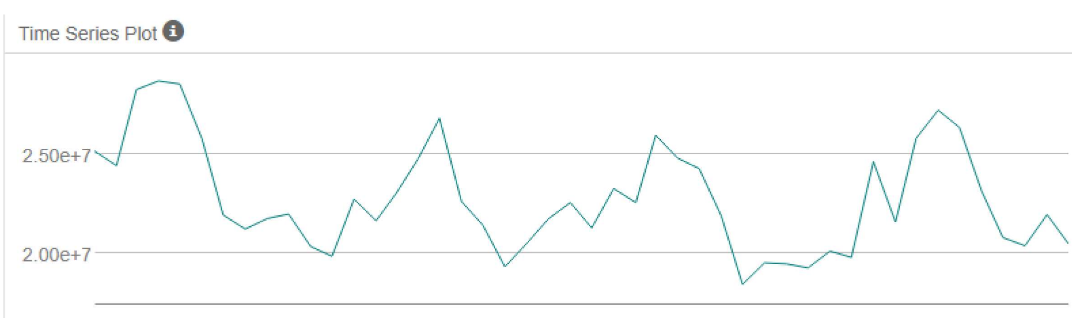


Fig 9- Time Series plot showing non-stationarity.

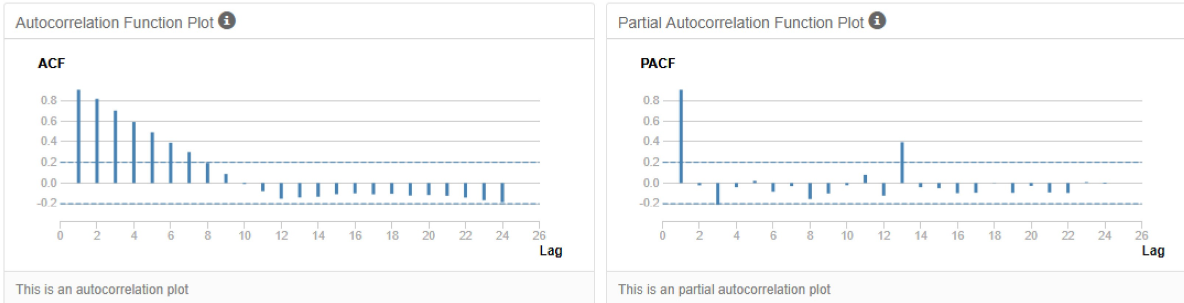
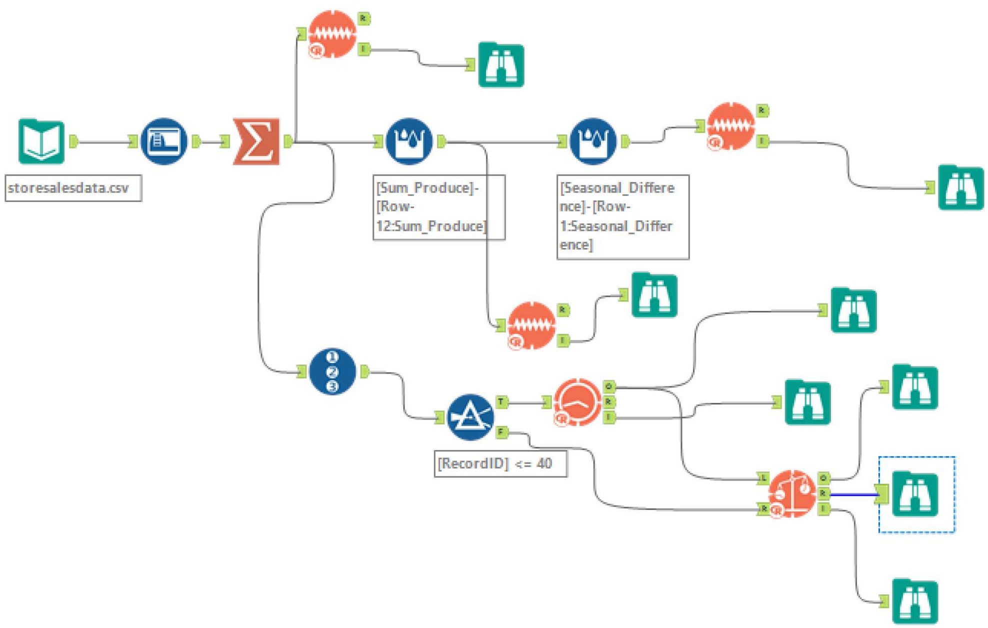


Fig 10- ACF and PACF Plots after differencing

ARIMA Model Alteryx Workflow



Comparison of Actual and Forecast Values

Actual and Forecast Values:

Actual	ARIMA
26338477.15	27182961.16627
23130626.6	24073582.27177
20774415.93	21223756.4441
20359980.58	20648299.23319
21936906.81	21205988.81004
20462899.3	21622151.40814

Fig 11- Actual and ARIMA Forecasted values

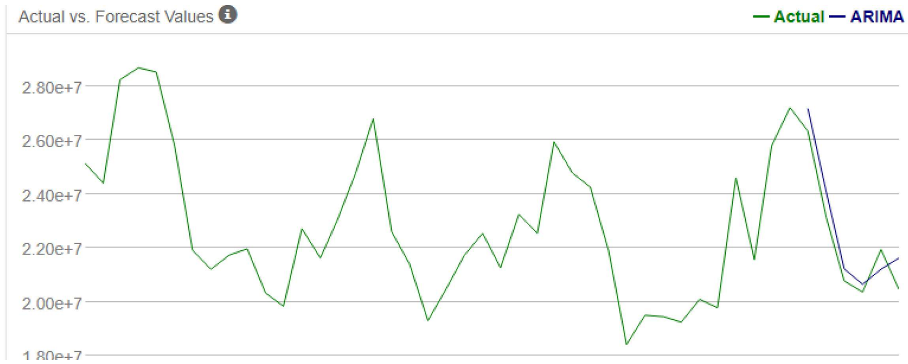


Fig 12- Forecast by ARIMA model

For comparing the two models, I used external validation to determine the model by comparing forecasted values with holdout sample. After analyzing the error measures such as RMSE, MPE, MAPE and ME and the accuracy measures, I chose the ETS model as it performed better than ARIMA on these metrics. We can also see from the tables of actual and forecasted values that ETS forecasting is closer to the actual values compared to ARIMA forecasted values.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	-492238.8	792197.3	735878.2	-2.1992	3.3098	0.433

Accuracy Measures:

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

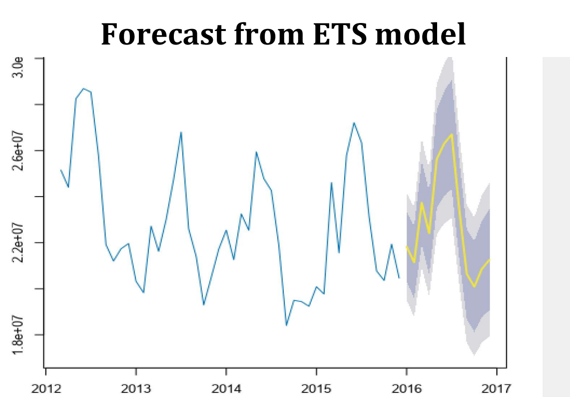
Fig 13- Accuracy Measures of ETS and ARIMA

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.

The total sales of existing stores have been calculated by grouping year, month and sum produce. The forecasted sales of new stores is calculated using each of the clusters and then multiplying the results by the number of new stores in that cluster. Then adding all these forecasts together on the same months to get a total forecast for all the new stores.

Date	New Store Sales Forecast	Existing Stores Forecast
Jan 2016	2,563,357.91	21,829,060.03
Feb 2016	2,483,924.72	21,146,329.63
March 2016	2,910,944.14	23,735,686.93
April 2016	2,764,881.86	22,409,515.28
May 2016	3,141,305.86	25,621,828.72
June 2016	3,195,054.20	26,307,858.04
July 2016	3,212,390.95	26,705,092.55
August 2016	2,852,385.76	23,440,761.32
September 2016	2,521,697.18	20,640,047.31
October 2016	2,466,750.89	20,086,270.46
November 2016	2,557,744.58	20,858,119.95
December 2016	2,530,510.80	21,255,190.24
Total Annual Sales	3,22,64,995.07	27,40,35,760.52

Table 1- Forecasted sales of existing and new stores by month for year 2016



Forecasting Alteryx Workflow

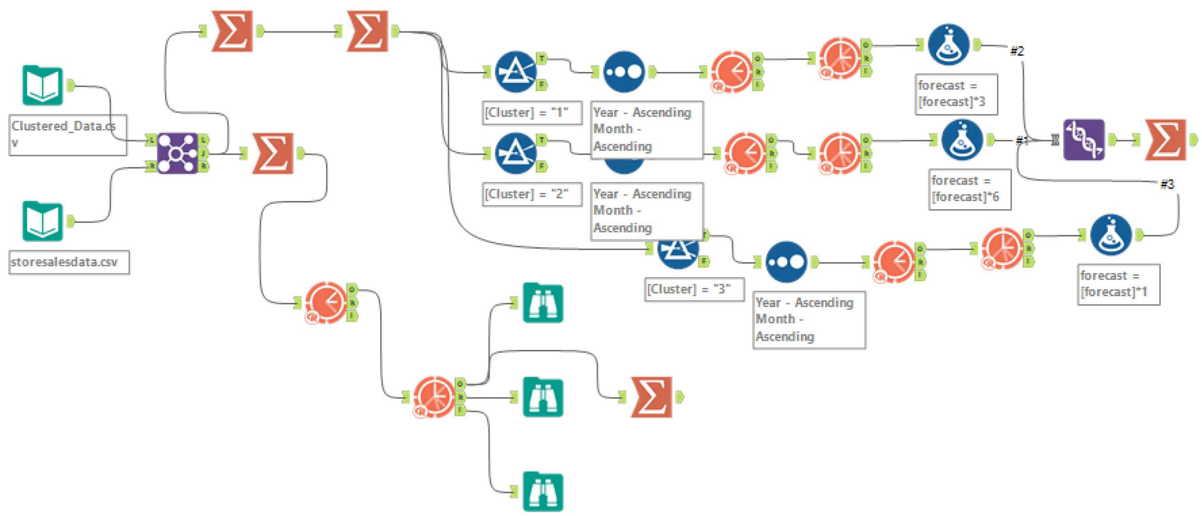
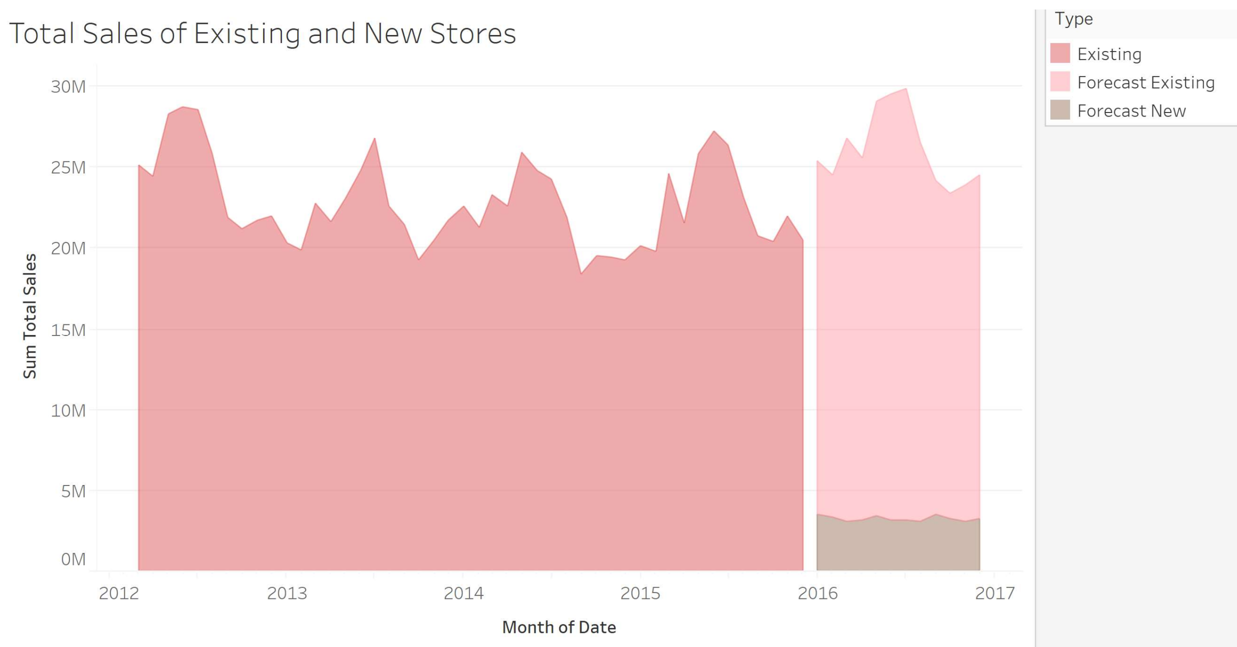


Tableau visualization



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.