Project: Predictive Analytics Capstone

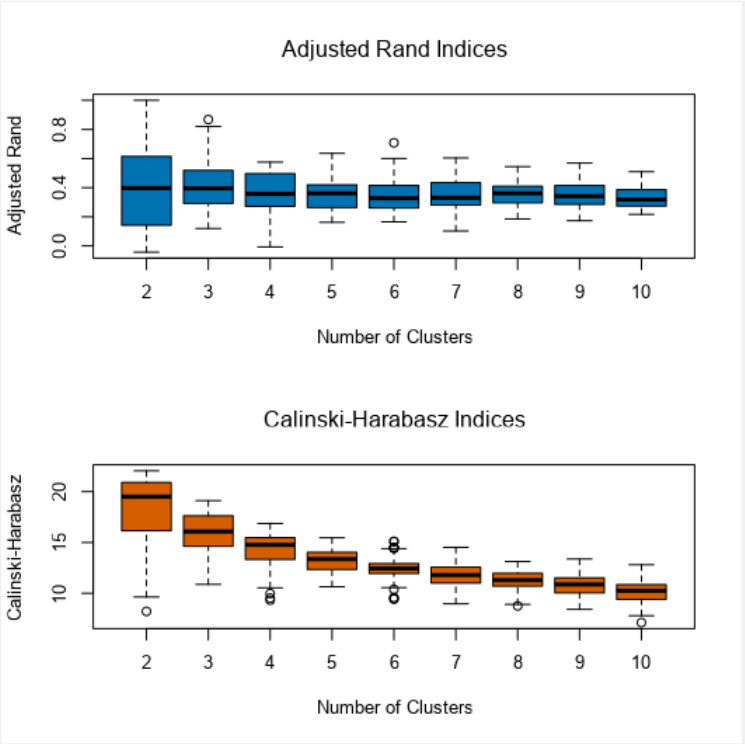
Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

Task 1: Determine Store Formats for Existing Stores

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

It has been determined to implement 3 store formats.

Using the K-Centroid Cluster Diagnosis reports for the K-Means method, has been observed that with 3 Clusters the AR and CH indices have the higher median and smaller variation at the same time per the whisker plot below.



2. How many stores fall into each store format?

Using the K-Centroid Cluster Analysis for K-Means, has been obtained the next report indicating the next number of stores per cluster id.

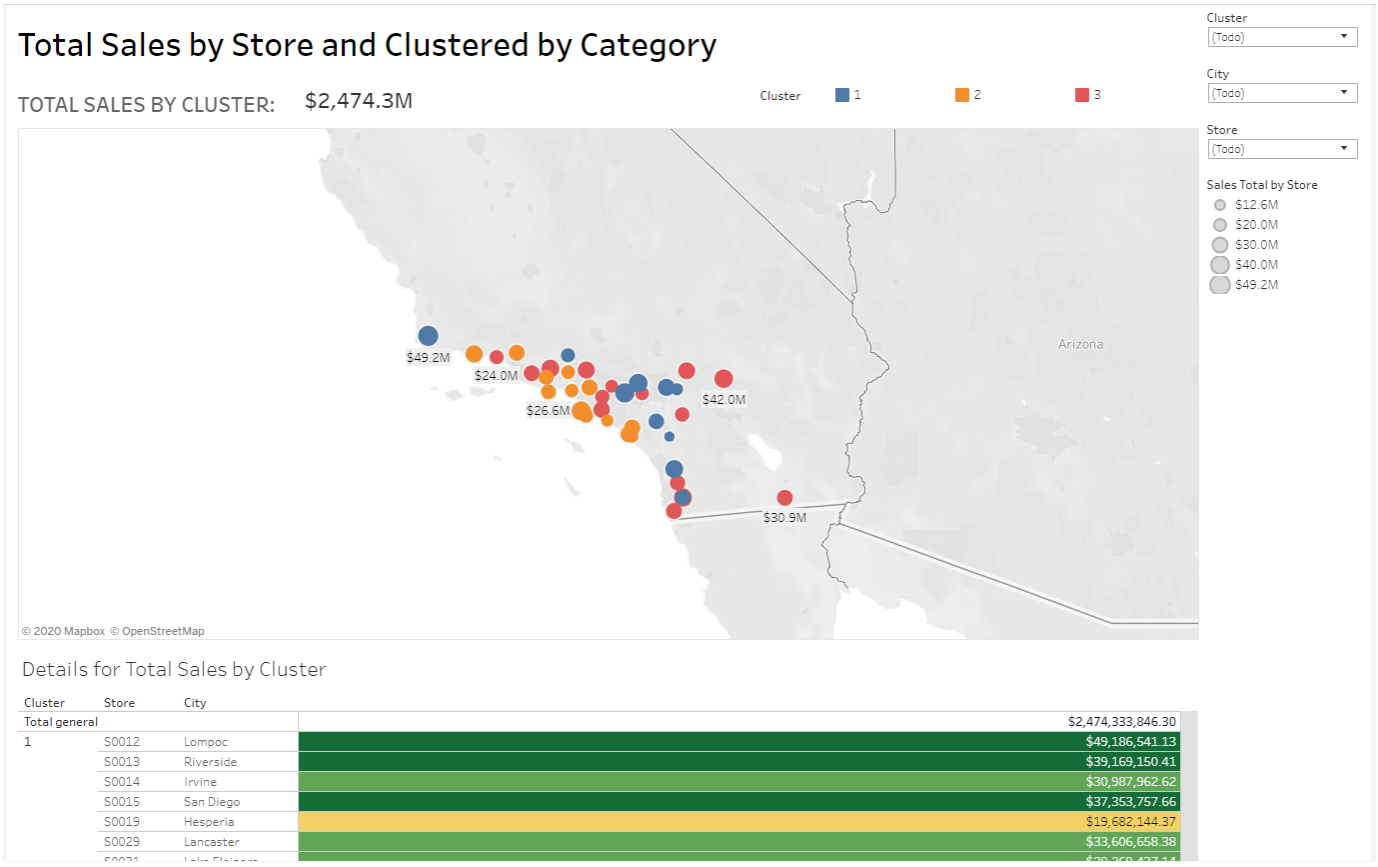
Cluster	Size
1	23
2	29
3	33

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

This difference for every cluster is given by the Distance and Separation. And, when comparing cluster values for same category, by the higher and lower values.

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

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.



You can see the tableau dashboard using the next link:

https://public.tableau.com/profile/felix.hernandez8665#!/vizhome/Task1_Clustered_Stores/DashboardStoreSalesClustered?publish=yes

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.)

It has been used the suggested configuration for sampling and random seeds, and comparing the accuracy level for the 3 models is the same, but also considering the F1 Score the best performer is Boosted Model (BM) as per the table and confusion matrix below.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT	0.8235	0.8426	0.7500	1.0000	0.7778
RF	0.8235	0.8426	0.7500	1.0000	0.7778
BM	0.8235	0.8889	1.0000	1.0000	0.6667

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.

Confusion matrix of BM

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

Confusion matrix of DT

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

Confusion matrix of RF

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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 ETS (M, N, M) Model has given a more accurate forecast per the report generated for Time Series Comparison, mainly considering the lower values for RMSE and MASE, and a nearest results compared against the Hold Out validation for 6 months.

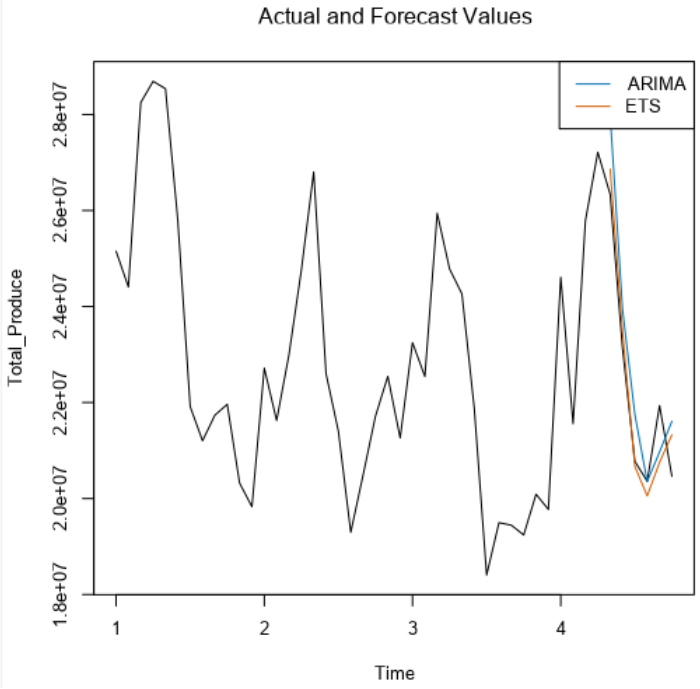
Comparison of Time Series Models

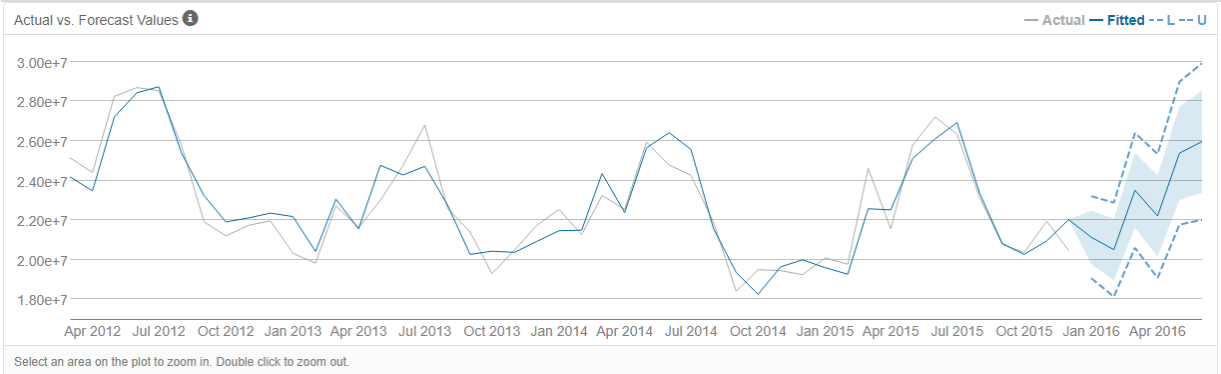
Actual and Forecast Values:

Actual	ARIMA	ETS
26338477.15	27997835.63764	26860639.57444
23130626.6	23946058.0173	23468254.49595
20774415.93	21751347.87069	20668464.64495
20359980.58	20352513.09377	20054544.07631
21936906.81	20971835.10573	20752503.51996
20462899.3	21609110.41054	21328386.80965

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257





Record	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	2016	1	21136641.781775	23208185.028684	22491151.105472	19782132.458079	19065098.534867
2	2016	2	20507039.12384	22880476.575432	22058946.457752	18955131.789929	18133601.672248
3	2016	3	23506565.982355	26405361.061884	25401986.205209	21611145.7595	20607770.902825
4	2016	4	22208405.755153	25340024.343149	24256061.087492	20160750.422815	19076787.167158
5	2016	5	25380147.771963	28988615.88472	27739598.24942	23020697.294506	21771679.659206
6	2016	6	25966799.465113	29918337.661734	28550571.413033	23383027.517192	22015261.268491

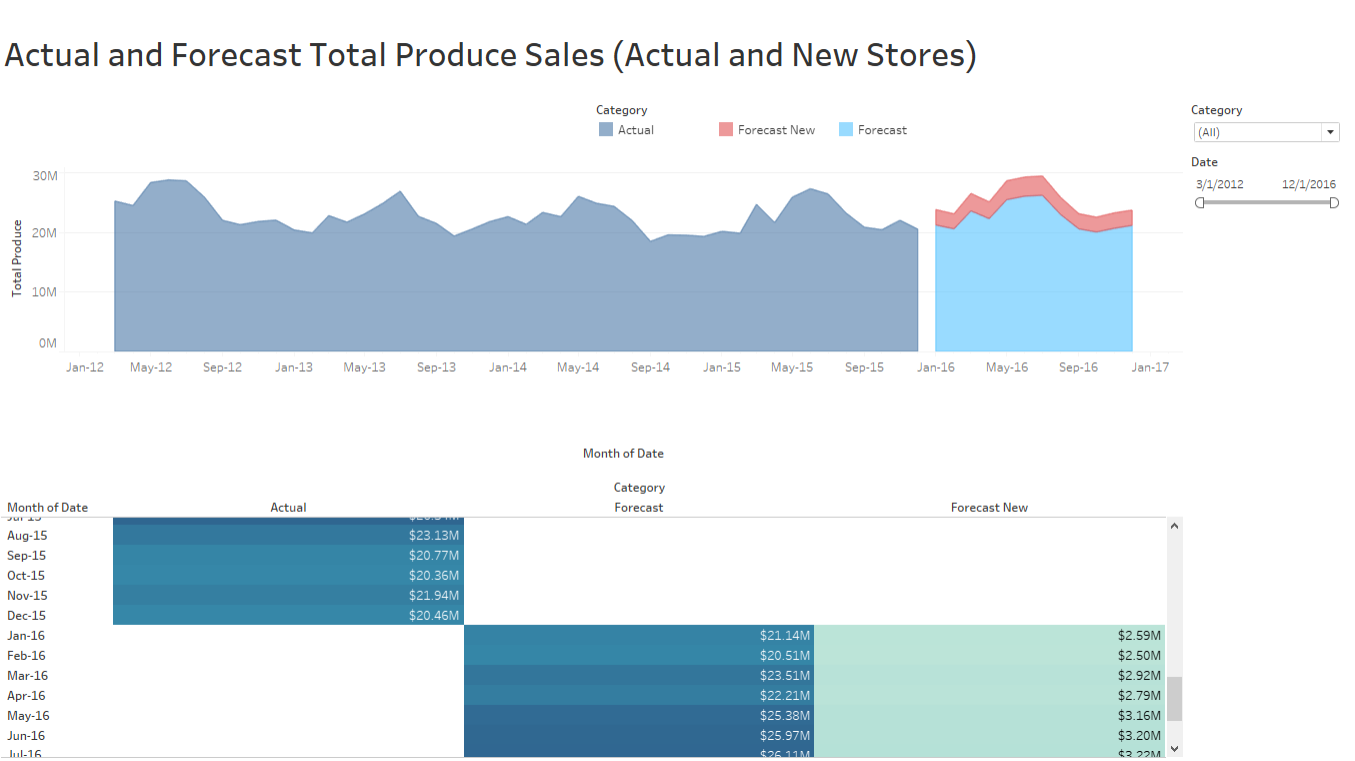
3. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

The table with forecast for new stores and existing stores can be seen below

Forecast Total Produce Sales

	Category	
Month of Date	New Stores	Existing Stores
Jan-16	$2.59M	$21.14M
Feb-16	$2.50M	$20.51M
Mar-16	$2.92M	$23.51M
Apr-16	$2.79M	$22.21M
May-16	$3.16M	$25.38M
Jun-16	$3.20M	$25.97M
Jul-16	$3.22M	$26.11M
Aug-16	$2.86M	$22.90M
Sep-16	$2.53M	$20.50M
Oct-16	$2.48M	$19.97M
Nov-16	$2.58M	$20.60M
Dec-16	$2.56M	$21.07M

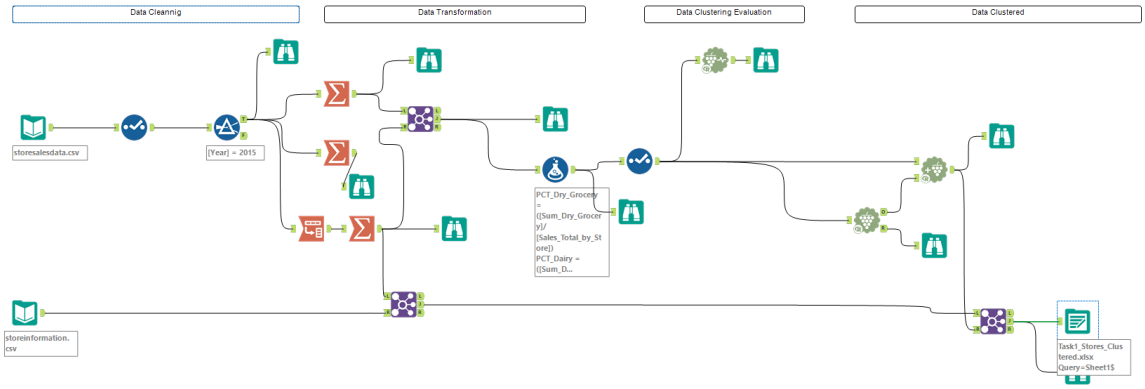
The tableau dashboard can be seen below.



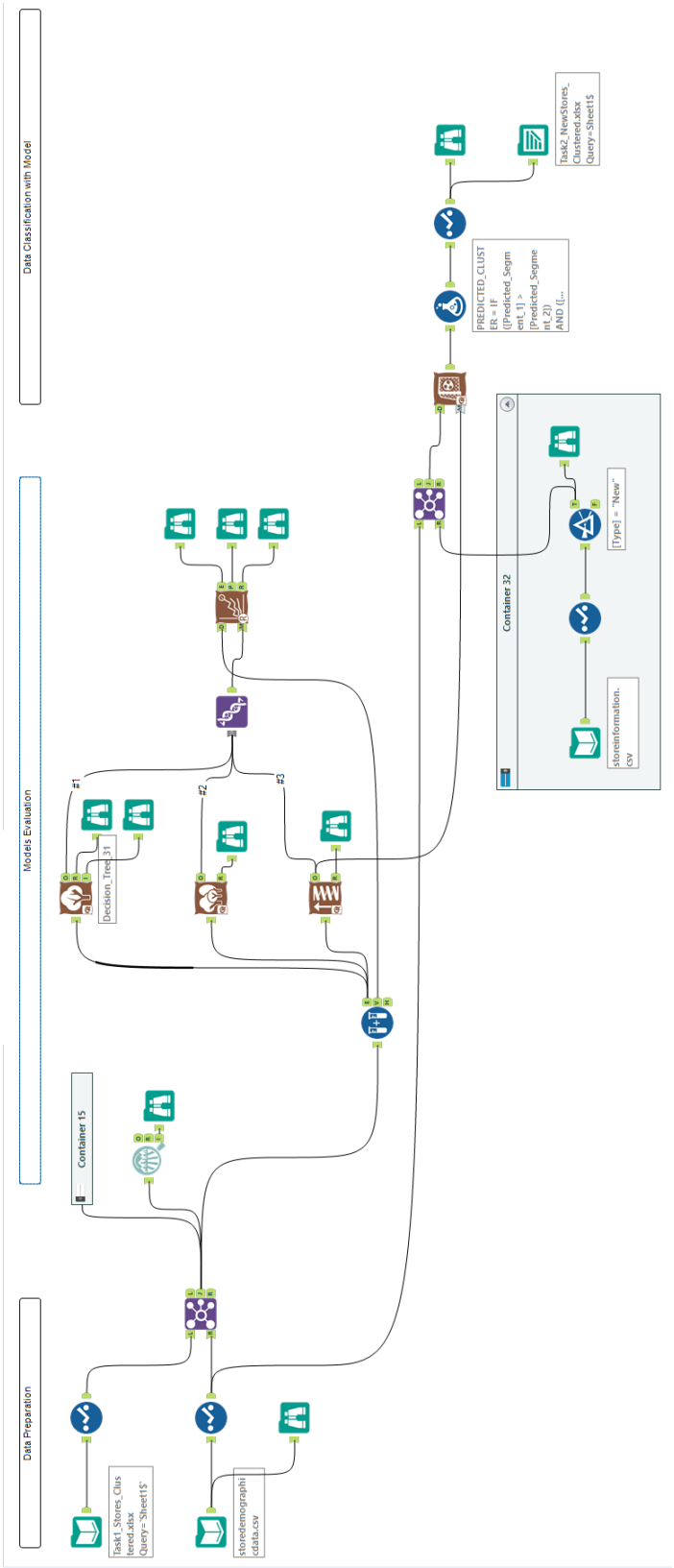
You can see the tableau dashboard using the next link:

https://public.tableau.com/profile/felix.hernandez8665#!/vizhome/Task3_TotalSales_for_Actual_and_New_Stores_with_Forecast/TotalProduceSalesActualandNewStores?publish=yes

Task 1 Alteryx Workflow:

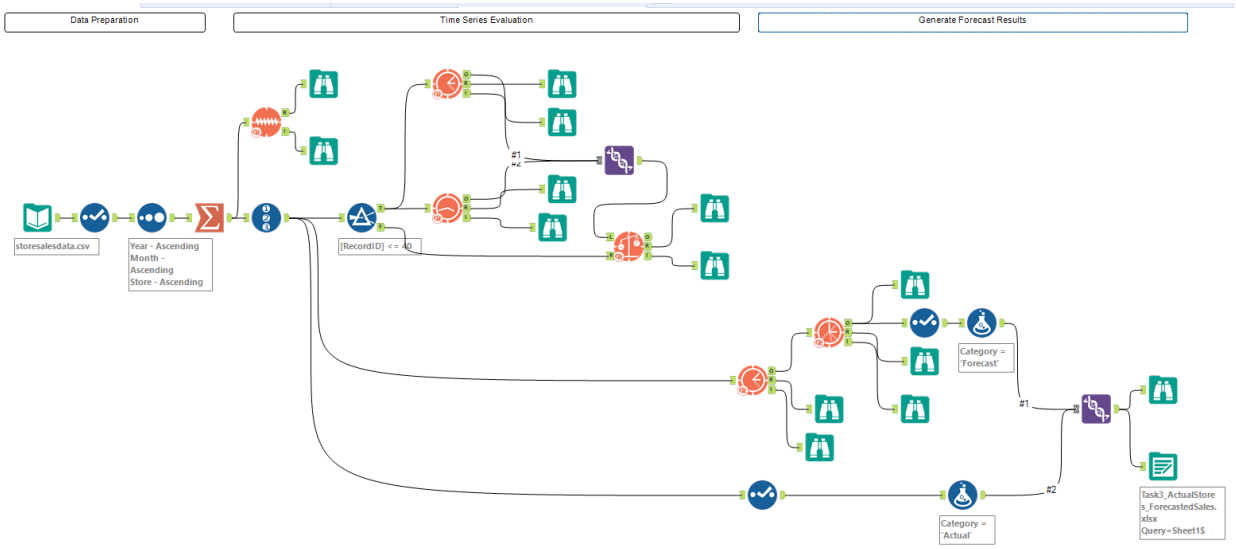


Task 2 Alteryx Workflow:



Task 3 Alteryx Workflows:

Existing Stores:



New Stores: