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 store formats is 3. According to the Adjusted Rand and Calinski-Harabasz Indices, the highest value for the types of probable existing formats is the best one, and one may observe that 3 clusters are at the top. The following whisker plot indicates the highest indices for quantity-quality clustering selection:

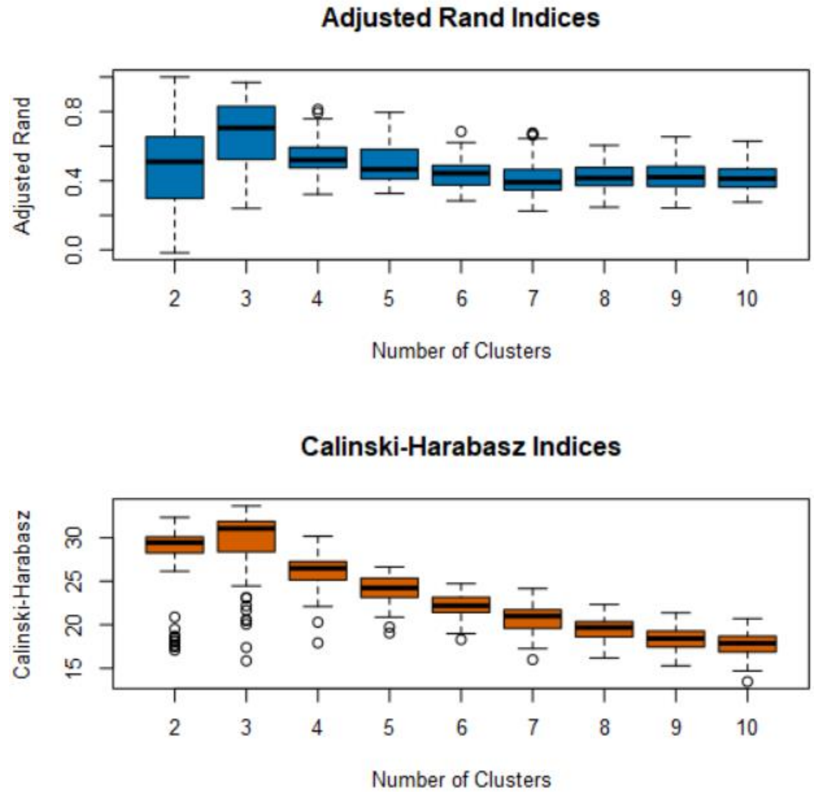


Figure 1: Box and Whisper plot for Indices Selection
Source: My Own + Alteryx, K-Centroid Cluster Analysis (2018)

2. How many stores fall into each store format?

Table 1: Cluster Size, Max & Avg. Distance and Separation

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

Source: My Own + Alteryx, K-Centroid Cluster Analysis (2018)

Format 1: 23 stores.
Format 2: 29 stores.
Format 3: 33 stores.

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

- Quantity of the variables.
- Compactness of the cluster variable.
- Distance between the clusters.

Table 2: Cluster placement for products

	Percent_Dry_Grocery	Percent_Diarry	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_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

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

Source: My own + Alteryx Tool

By visualizing the table, one must say that the clusters are not opposites, but just different in characteristics, there is no negative values. There are big values like in dry grocery and also small values like in floral.

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.

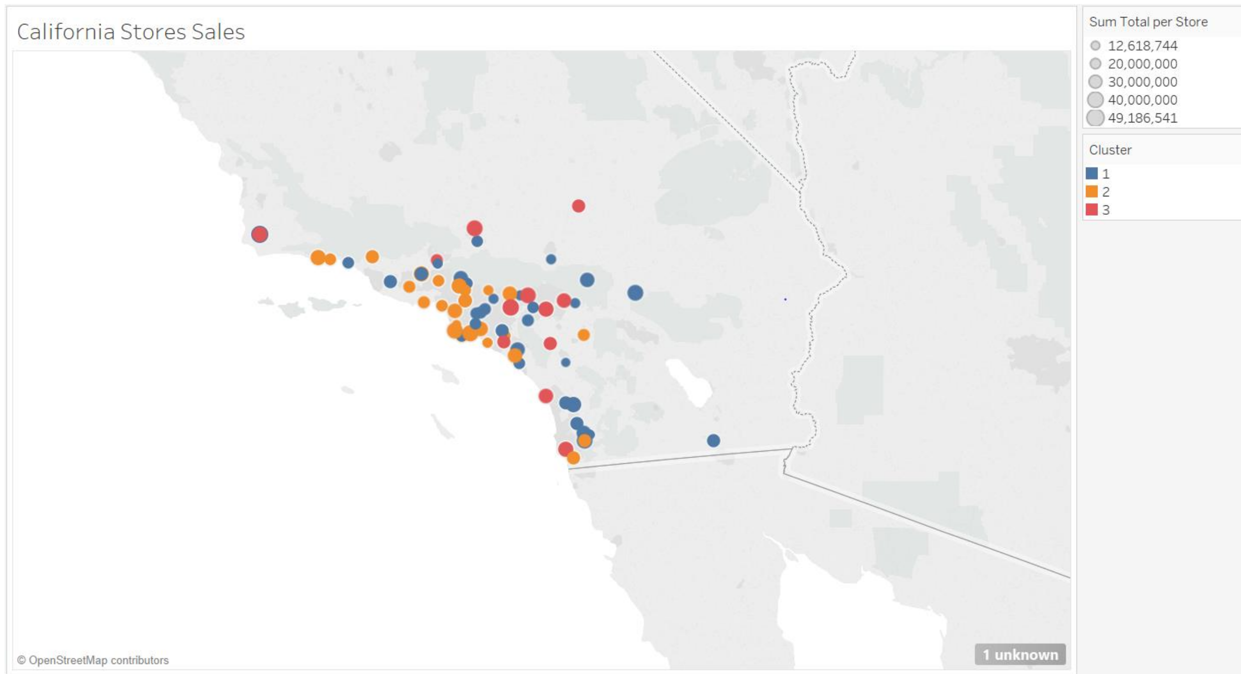


Figure 2: Map South California With Store Sales
Source: My Own + Tableau Public Tool (2018)

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.)

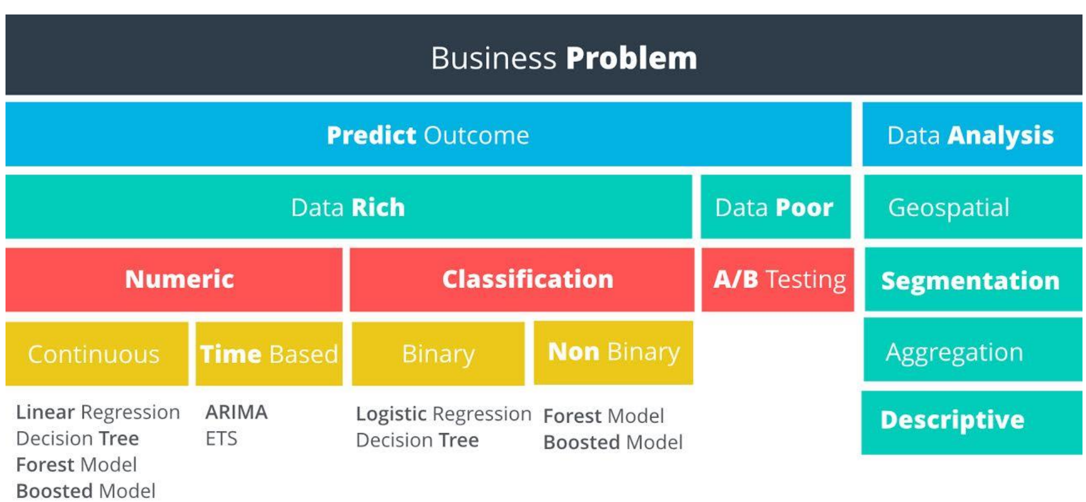


Figure 3: The Business Problem Framework
Source: Udacity.com (2018)

By observing the business problem framework and analyzing the business problem, we may conclude the following:

1. We need a model that predicts Outcome.
2. We have a data rich problem.
3. The outcome has 3 possible outcomes, so we need a classification type model.
4. It is Non-Binary, more than two outcomes:
   4.1 Forest Model.
   4.2 Boosted Model.
   4.3 Decision Tree.

Knowing this, we have to decide between these three models which one predicts best.

We run the Alteryx tool with the three Models, where we used a comparison tool to compare the three models and see which one of them produced the best outcomes.

Table 3: Comparison Tool Accuracy and Errors

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Cluster_Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Cluster_Forest	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Model_Cluster	0.8235	0.8889	1.0000	1.0000	0.6667

Source: My Own + Alteryx Tool (2018)

Forest Model and Boosted Model represent the best options because they share the same overall accuracy of 0.8235, nevertheless the Boosted Model has a higher test accuracy because it has a higher F1 score.

In the matrix shown above, the boosted model also seems to perform better with the cluster 1 and 2 accuracies, but in cluster three is fairly lower than the forest model.

The choice is 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?

When choosing an appropriate forecasting model, one has to make a time series decomposition where one may observe seasonality, trend and error plots. This will provide enough information in order to make an appropriate ETS model. The following figures displays the time series decomposition plot:

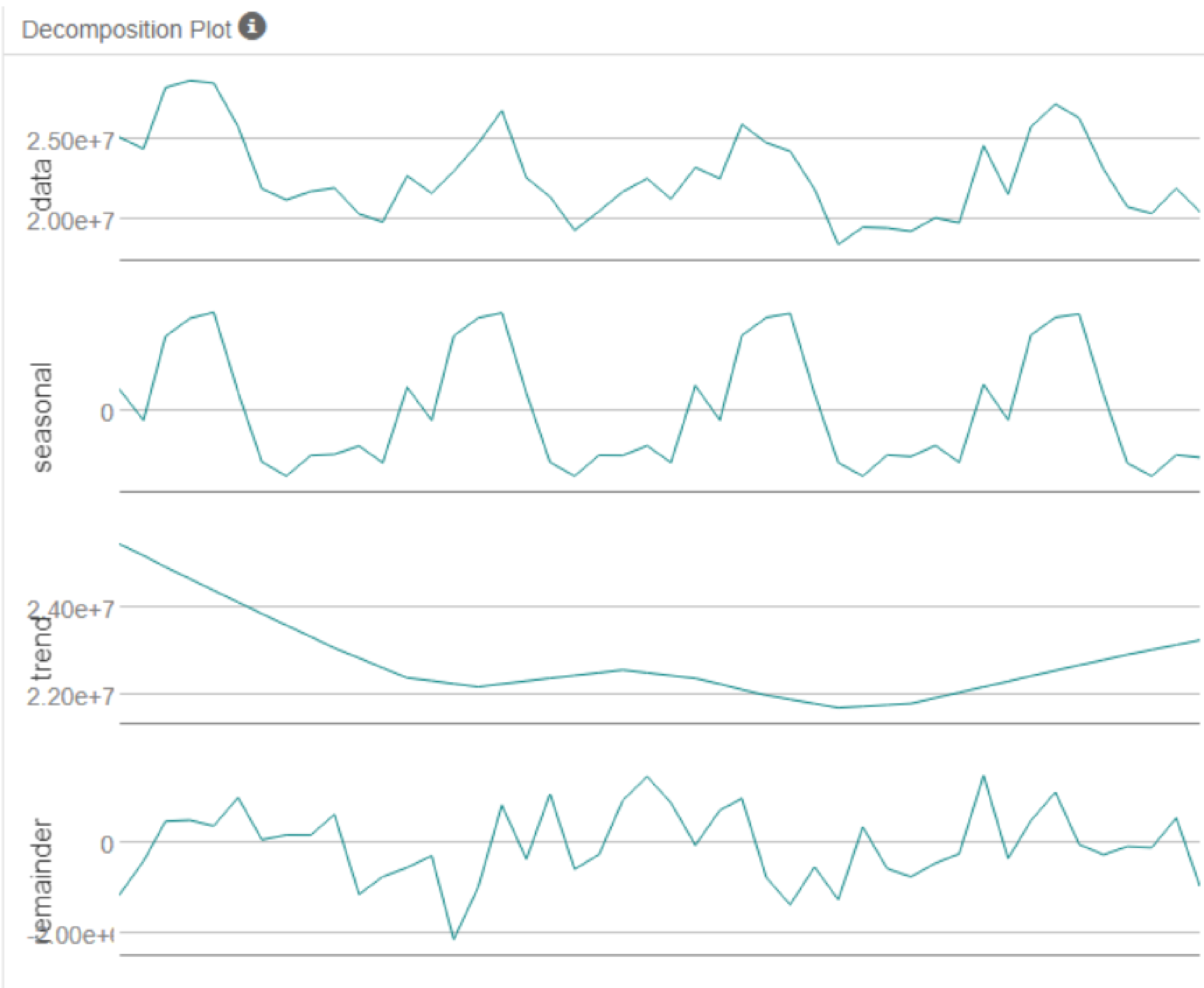


Figure 4: Time-Series Decomposition Plot
Source: My Own + Alteryx Tool (2018)

For an ETS Model:

- Error should be used multiplicatively, since it varies through time noticeably.

- There is no trend component to be used, since there is no clear evidence of linear or exponential trend, just curves.
- For the seasonal component it should be a multiplicative component, since their slight increase through seasons.

Then it is a ETS(M,N,M) type of model.

ARIMA models need a Autocorrelation Function Plot and a Partial Autocorrelation Function Plot in order to find the components.

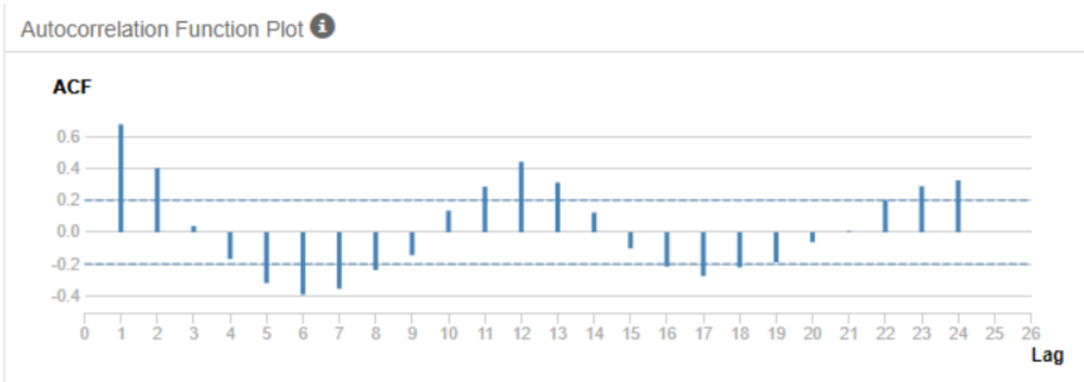


Figure 5: Autocorrelation Function Plot
Source: My Own + Alteryx Tool (2018)

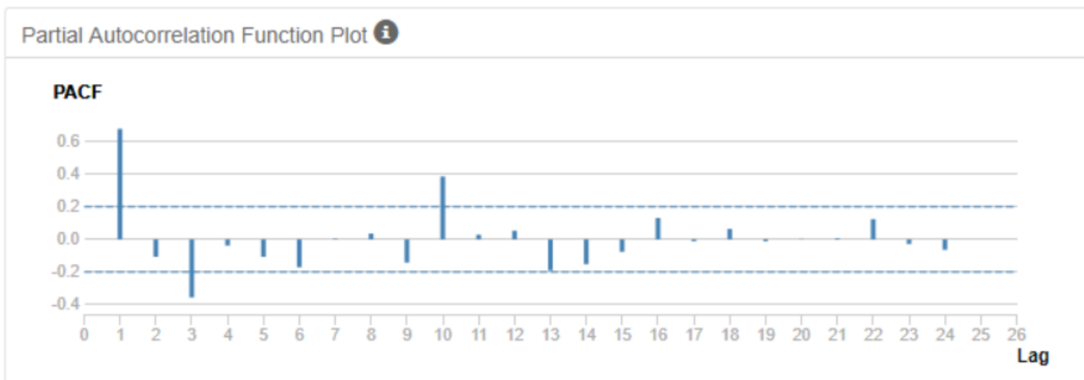


Figure 6: Partial Autocorrelation Function Plot
Source: My Own + Alteryx Tool (2018)

By analyzing the ACF and PACF plots, there has to be a differencing process, since ACF do not show stationary tendency. Since there is a seasonality component, there is also the need to do seasonality differencing. The process is made through Alteryx.

- The m for the ARIMA seasonal differencing is 12 (monthly time series)

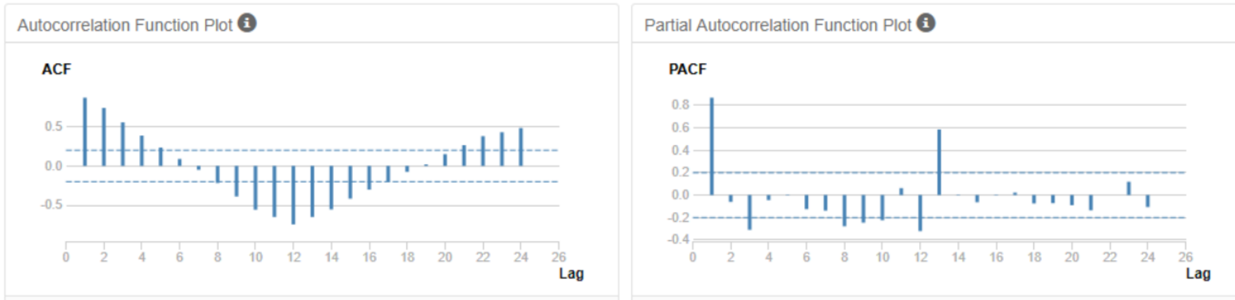


Figure 7: ACF and PACF After Seasonal Differencing
Source: My Own + Alteryx Tool

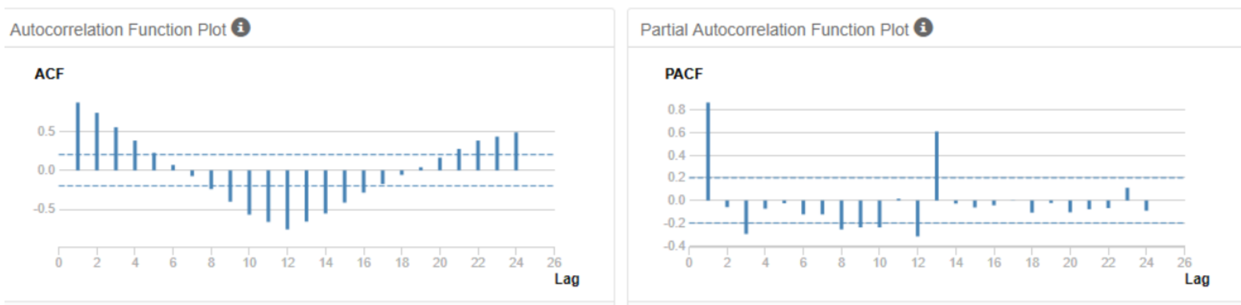


Figure 8: ACF and PACF, after Second Seasonal Differencing
Source: My Own + Alteryx Tool

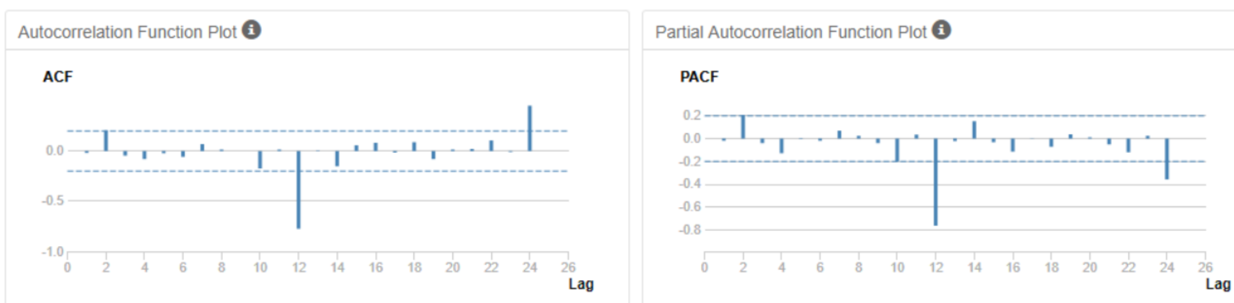


Figure 9: ACF and PACF, after First Differencing
Source: My Own + Alteryx Tool

We conclude:

- Two seasonal differencing where made and one standard differencing was made.
- P or Q in the seasonal terms are 0, no clear pattern was observed in ACF and PACF.
- There is negative autocorrelation in Lag-1, so a p=1 is selected.

Then, we conclude that ARIMA (1,1,0)(0,2,0)12 is going to be used.

After applying both models with Alteryx tool and using the TS compare tool, now we have enough information to choose our best model.

Table 4: Accuracy Measures of ETS

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_Forecast	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA

Source: My Own + Alteryx Tool (2018)

Table 5: Accuracy Measures of ARIMA

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	903785.4	1817617	1628701	3.9501	7.4258	0.9583	NA

Source: My Own + Alteryx Tool (2018)

By looking at the MASE and RMSE, we may observe that the forecast made by the ETS model are far superior 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.

Table 6: Actual and Forecast Values of ETS model, Holdout Sample

Actual and Forecast Values:

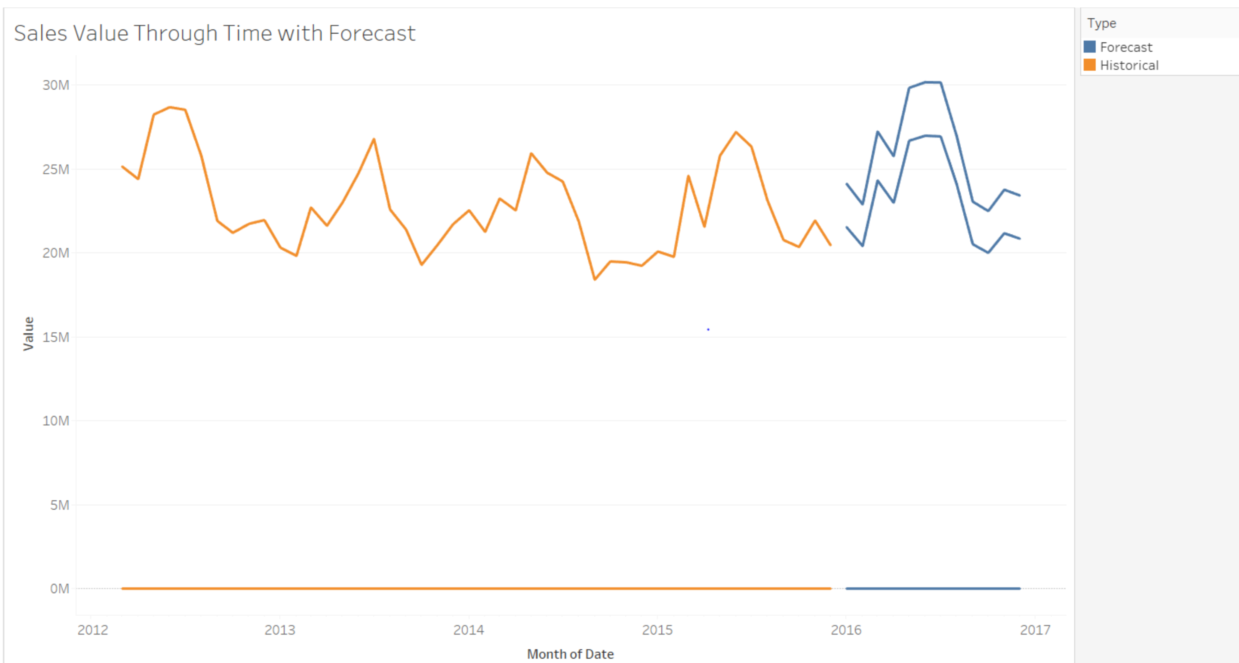
Actual	ETS_Forecast
26338477.15	26907095.61191
23130626.6	22916903.07434
20774415.93	20342618.32222
20359980.58	19883092.31778
21936906.81	20479210.4317
20462899.3	21211420.14022

Source: My Own + Alteryx Tool (2018)

Table 7: Forecasted Values for every Period

Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
11	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
12	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
1	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981697
2	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
3	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
4	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
5	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
6	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
7	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
8	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
9	21177435.485839	23994279.191514	23019270.585553	19335600.386124	18360591.780163
10	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

Source: My Own + Alteryx Tool (2018)



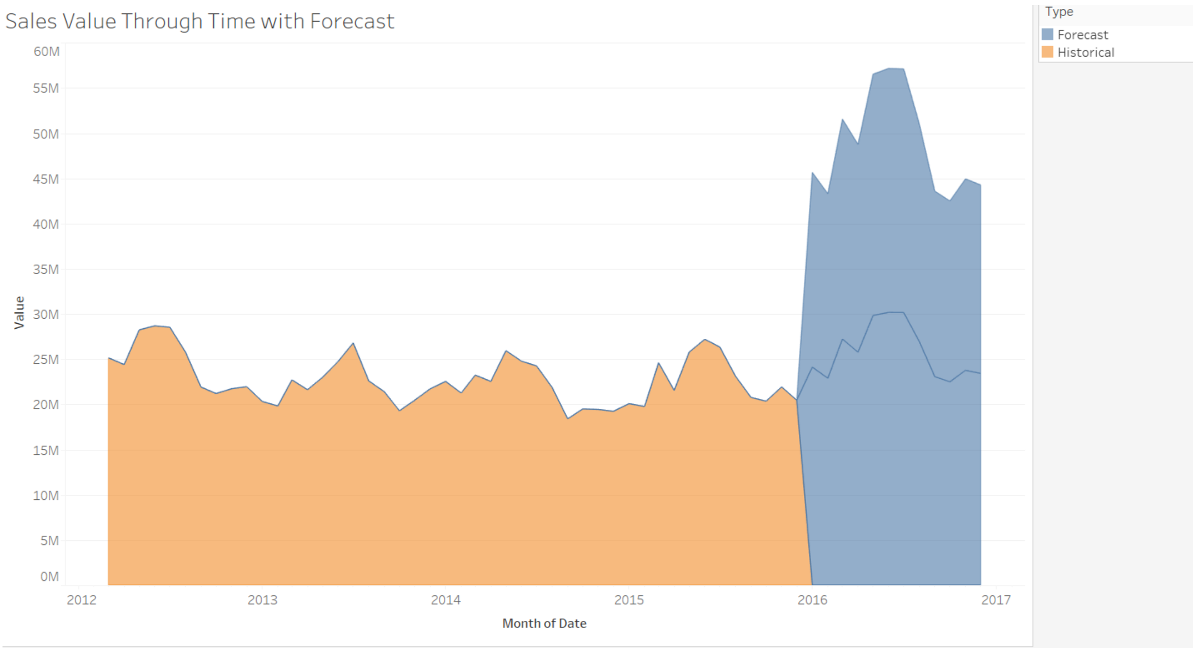


Figure 10: Forecast Plot Predicted
Source: My Own + Tableau Tool (2018)

Month	New Stores	Existing Stores
1	2,587,450.851495	21,539,936.007
2	2,477,352.892393	20,413,770.60136
3	2,913,185.23625	24325953.097628
4	2,775,745.609767	22993466.348585
5	3,150,866.835326	26691951.419156
6	3,188,922.00336	26989964.010552
7	3,214,745.646251	26948630.764764
8	2,866,348.663392	24091579.349106
9	2,538,726.84886	20523492.408643
10	2,488,148.287462	20011748.6686
11	2,595,270.386448	21177435.485839
12	2,573,396.62905	20855799.10961

The Existing stores where predicted by adding up the stores through months of the data and passing them through the ETS(M,N,M) model. The new stores, on the other hand, where predicted by classifying the data through clustering. Sales where averaged, separated depending on their store format, while finally arriving at a value. That value was multiplied according the number of new stores that fell in each cluster, and summed up for each month.

Images:

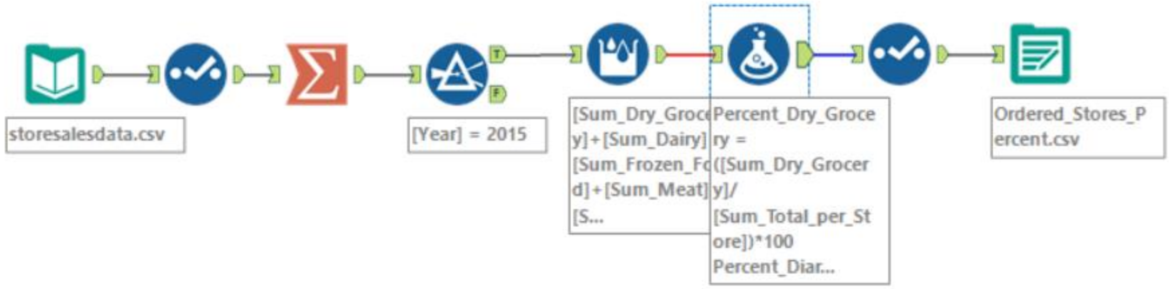


Figure 11: Workflow 1
Source: My Own + Alteryx Tool

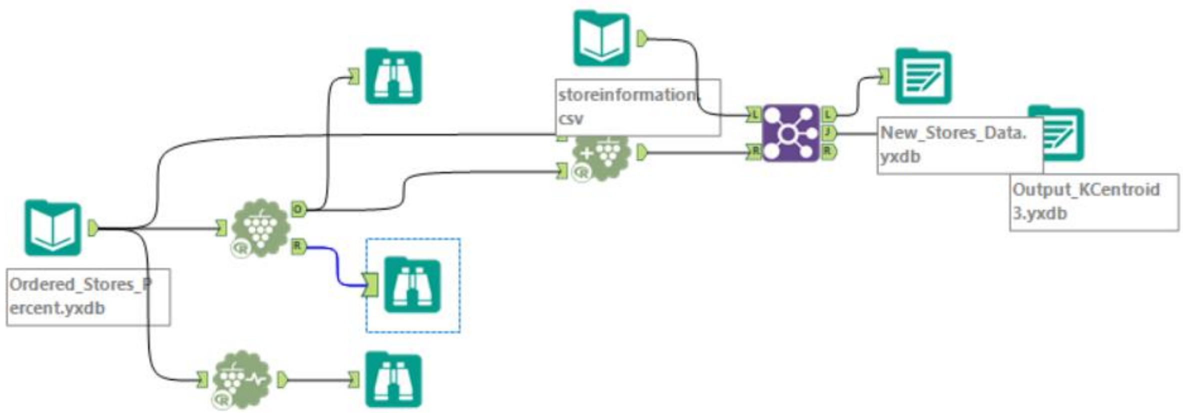


Figure 12: Workflow 2
Source: My Own + Alteryx Tool

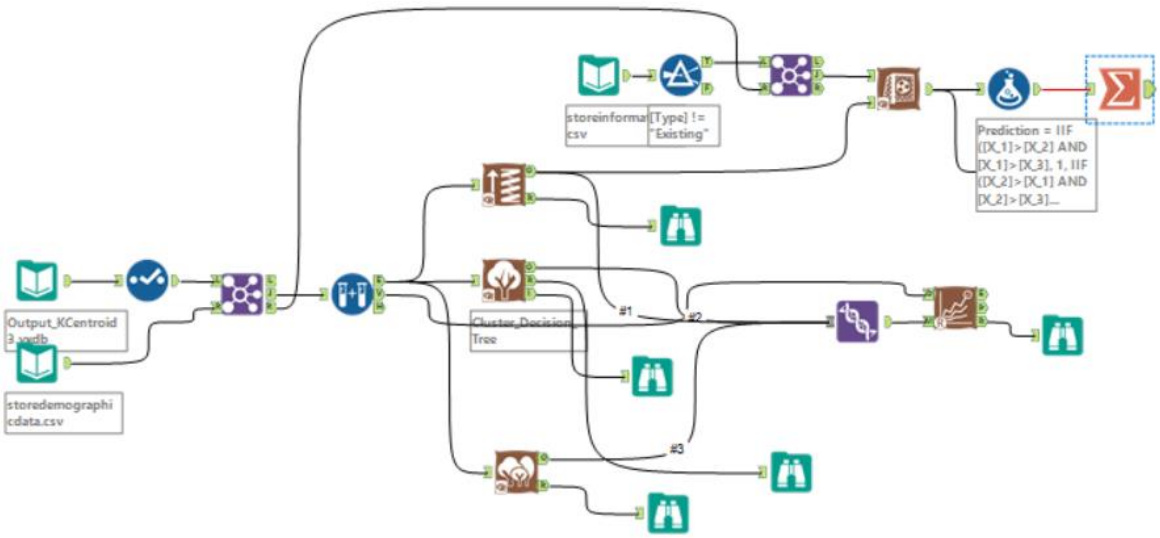


Figure 13: Workflow 3
Source: My Own + Alteryx Tool

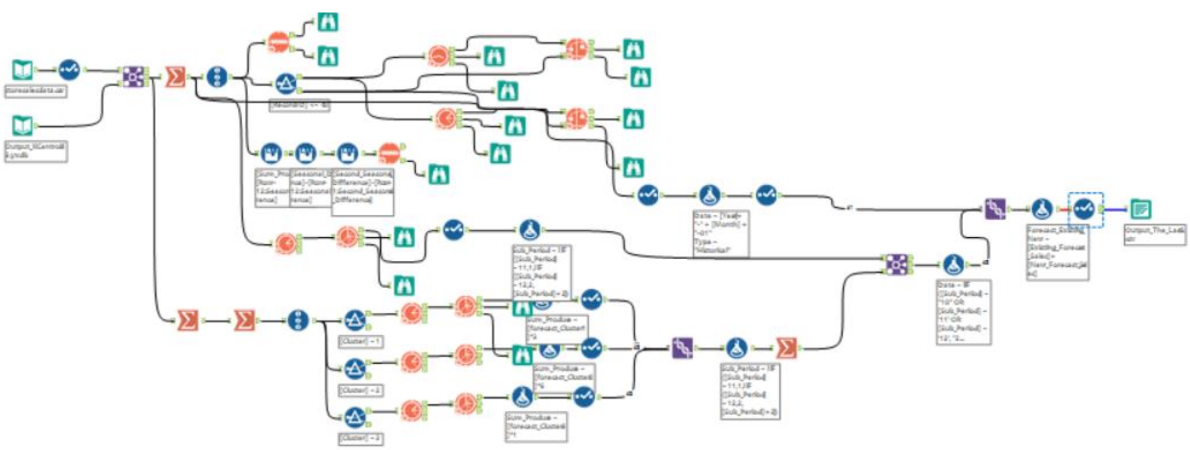


Figure 14: Workflow 4
Source: My Own + Alteryx Tool