Udacity Business Analyst Nanodegree Project: Predictive Analytics Capstone

Jason Grenig

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 formats is 3. This is because 3 clusters yields the highest median value in both indices, while also having good compactness and distinctness in the plots.

K-Means Cluster Assessment Report

Summary Statistics

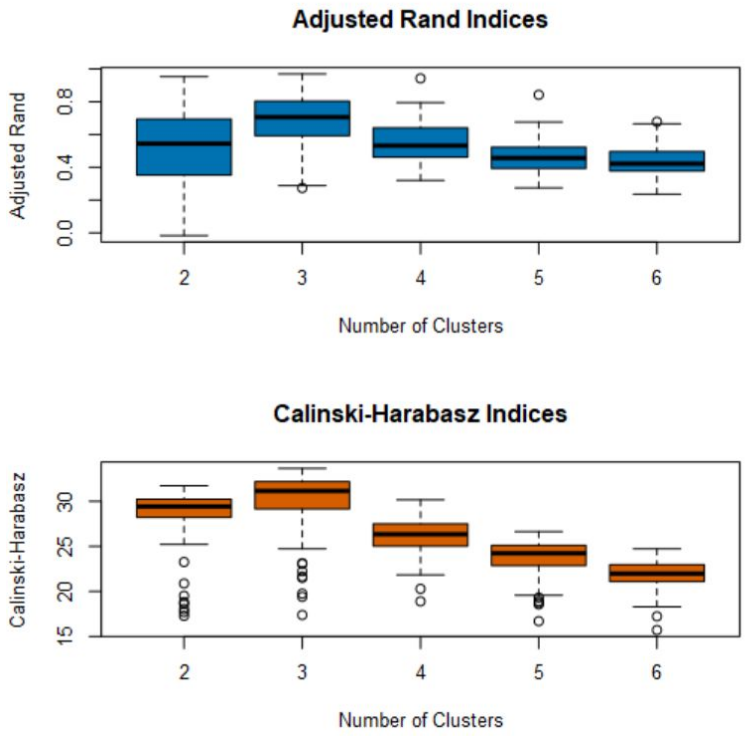
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.016485	0.27351	0.31976	0.274316	0.235718
1st Quartile	0.35943	0.594017	0.46406	0.39294	0.377774
Median	0.544023	0.705326	0.53195	0.456588	0.421798
Mean	0.524263	0.69161	0.548167	0.470346	0.435429
3rd Quartile	0.694147	0.800179	0.635682	0.520656	0.493589
Maximum	0.952939	0.969034	0.942222	0.841981	0.677532

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	17.281	17.38103	18.89398	16.69676	15.71092
1st Quartile	28.22121	29.21236	25.03471	22.86498	21.10249
Median	29.4157	31.14178	26.33467	24.22188	21.96958
Mean	28.56936	30.07118	26.18037	23.72205	21.92474
3rd Quartile	30.21867	32.17467	27.4999	25.09459	22.95561
Maximum	31.71569	33.63781	30.1583	26.63063	24.72038

Alteryx K-Means Cluster Assessment Report



Alteryx K-Means Plots of the AR and CH Indices

2. How many stores fall into each store format?
   Cluster 1 has 23 stores, Cluster 2 has 29 stores, Cluster 3 has 33 stores:

Cluster Information:

Cluster	Size
1	23
2	29
3	33

Alteryx K-Centroids Cluster Analysis Results

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Looking at the Convergence after 12 iterations section, the more positive the number, the more sales for that particular category. So we can see differences in the clusters as follows:
   - Cluster 1 has more General Merchandise sales.
   - Cluster 2 has more Produce and Dairy sales.
   - Cluster 3 has more Meat and Deli sales.

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.

	Pct_Dry_Grocery	Pct_Dairy	Pct_Frozen_Food	Pct_Meat	Pct_Produce	Pct_Floral	Pct_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

	Pct_Bakery	Pct_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Alteryx K-Centroids Cluster Analysis Results

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.

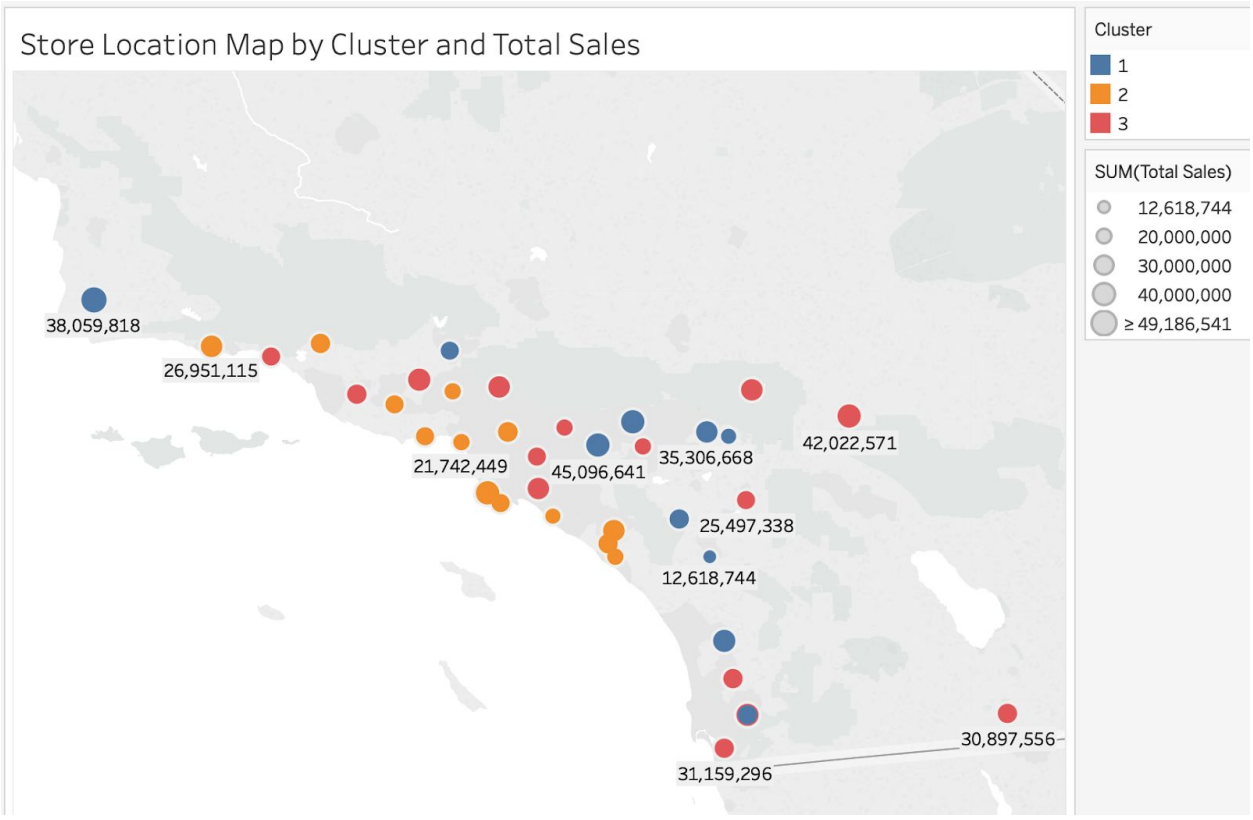


Tableau Visualization of Store Location, Cluster, and Size (Total Sales)

Tableau Public file:
https://public.tableau.com/profile/jason.grenig#!/vizhome/StoreLocationMapbyClusterandTotalSales/StoreLocationMap?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.)
   I chose the Boosted Model. Even though the accuracy is the same for all 3 models, the Boosted Model has a higher F1 score. F1 Score is the weighted average of Precision and Recall. Precision being the ratio of correctly predicted positive observations to the total predicted positive observations, and Recall being the ratio of correctly predicted positive observations to the all observations in actual class - yes. So in this measure, the Boosted model's higher F1 score indicates it's a more accurate model.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_model	0.8235	0.8426	0.7500	1.0000	0.7778
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Model	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 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 DT_model

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

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

Alteryx Model Comparison Report for Decision Tree, Forest, and Boosted Models

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?

We'll need to compare ETS vs ARIMA to see which is better to use for our forecasting. Let's start by measuring the ETS model first. We can see from the Decomposition Plot that the error is pretty randomly distributed around the mean, so it should be applied multiplicatively (m). The trend is not clear, so nothing to apply (n). Then there is seasonality present, so we'll apply this multiplicatively (m).

The ETS Model for our comparison will be: **ETS(m,n,m).**

Alteryx Time Series Plot Interactive Output

Next, we can use the TS Compare Tool on our ETS model to see how it did predicting the last 6 months of our holdout sample. Based on the Accuracy Measures (see below image), the MASE is 0.3822, which is a good sign because it's less than the 1.0 threshold for usability. We'll move on to building the ARIMA model and then do a head to head comparison.

Accuracy Measures:

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

Alteryx Time Series Comparison of ETS Model Accuracy

For the ARIMA model, first let's take a look at the Time Series Plot to see if the data are stationary. Since it's not centered around 0 on the y-axis, it's not stationary. We'll need to do seasonal differencing to stationarize it.

Alteryx Time Series Plot

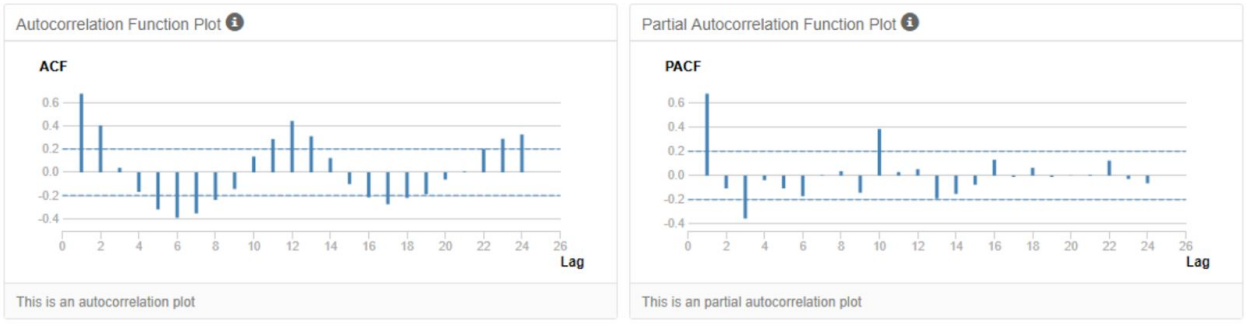
After seasonal differencing, the data are not centered around 0. Let's see if another go around will do it (first seasonal differencing).

Alteryx Time Series Plot after Seasonal Differencing

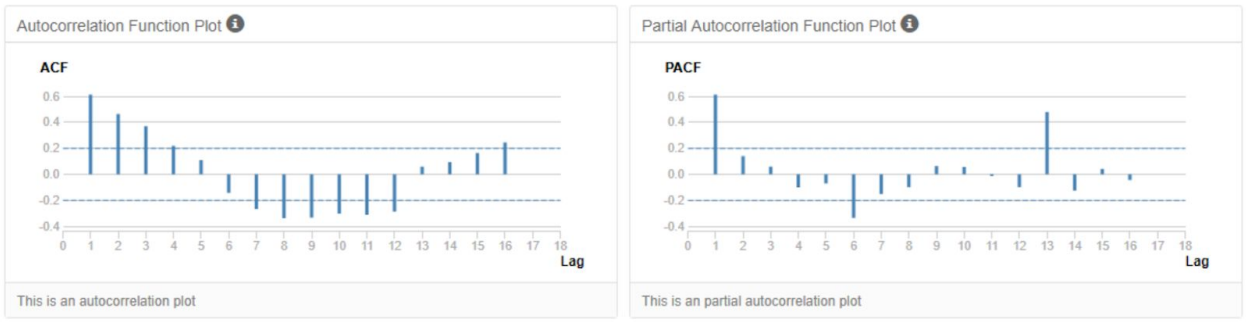
The first seasonal differencing looks much more stationarized:

Alteryx Time Series Plot after First Seasonal Differencing

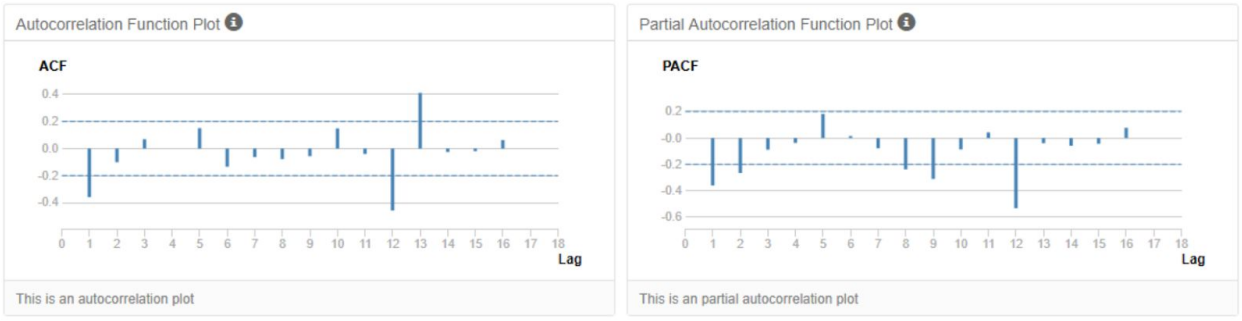
Next, we'll figure out our model terms, ARIMA(p,d,q)(P,D,Q)m, by looking at the ACF and PACF plots.



Alteryx ACF and PACF results for original TS Plot



Alteryx ACF and PACF results for Seasonal Differencing



Alteryx ACF and PACF results for the First Seasonal Differencing

For the non-seasonal terms (p,d,q):
Both the ACF and PACF have negative correlation at lag-1, which indicates an MA signature. So we'll use terms of q(2). The p term will be 0, as there is no AR signature. We used first seasonal differencing, so d(1) is used.

For the seasonal terms (P,D,Q):
Neither AR or MA signature is present, so we have P(0) and Q(0).
We used first seasonal differencing, so D(1) is used.

We have monthly data, so we use m = 12 for the number of periods.

Our terms will be ARIMA(0,1,2)(0,1,0)12

Now let's compare accuracy measures between the models:
The MASE value is lower on the ETS Model, which indicates ETS is more accurate.
The RMSE value is also lower on the ETS model, meaning this forecast has a narrower range of possible values. We should use the ETS model for our forecast.

Accuracy Measures:

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

Alteryx Time Series Comparison of ETS Model Accuracy

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_0_1_2__0_1_0_12	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909

Alteryx Time Series Comparison of ARIMA Model Accuracy

2. 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	Exisiting Stores Forecast	New Stores Forecast
Jan-16	$21,539,936	$2,626,198
Feb-16	$20,413,771	$2,529,186
Mar-16	$24,325,953	$2,940,264
Apr-16	$22,993,466	$2,774,135
May-16	$26,691,951	$3,165,320
Jun-16	$26,989,964	$3,203,286
Jul-16	$26,948,631	$3,244,464
Aug-16	$24,091,579	$2,871,488
Sep-16	$20,523,492	$2,552,418
Oct-16	$20,011,749	$2,482,837
Nov-16	$21,177,435	$2,597,780
Dec-16	$20,855,799	$2,591,815

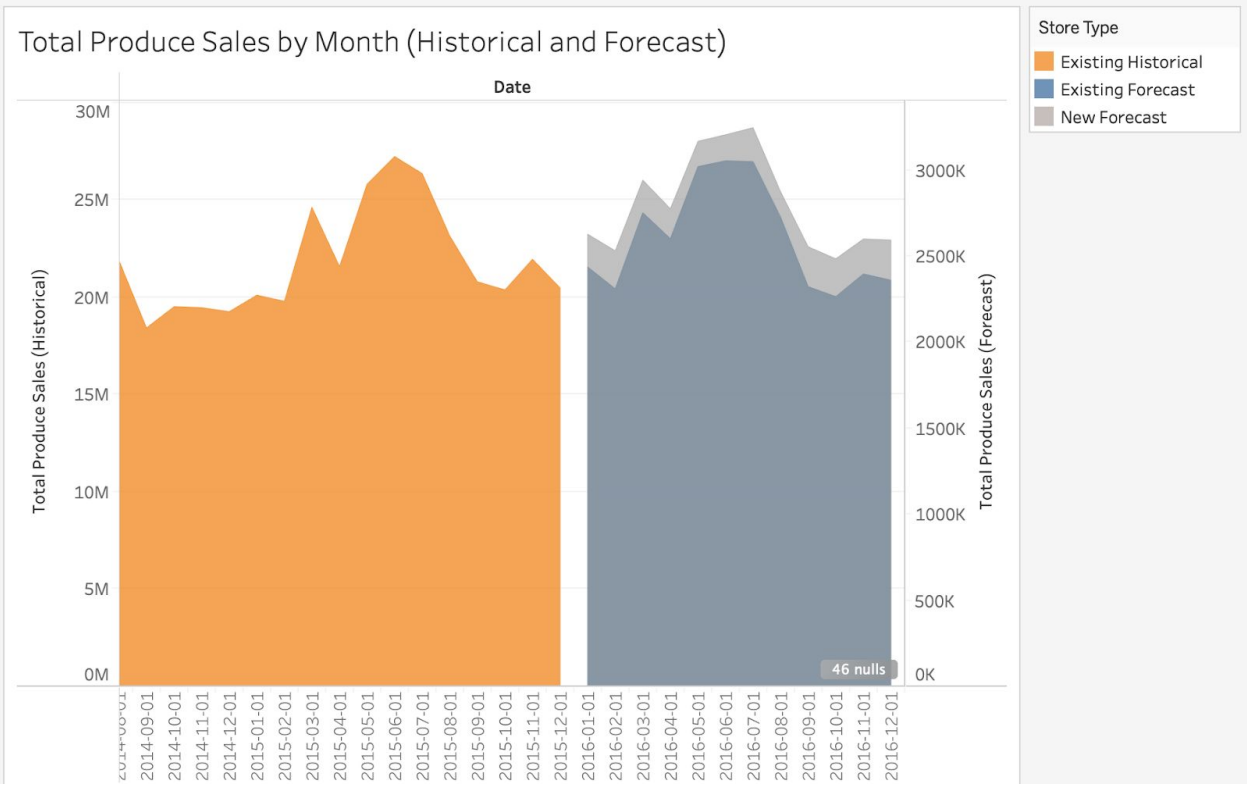
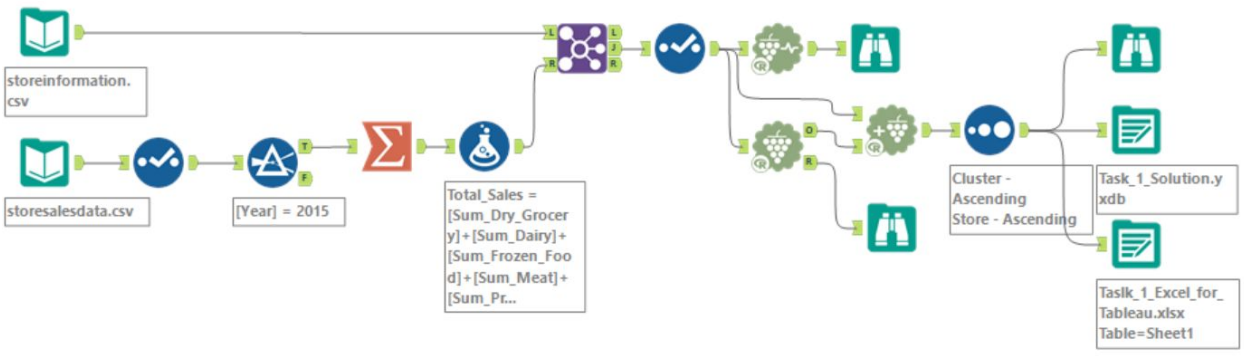


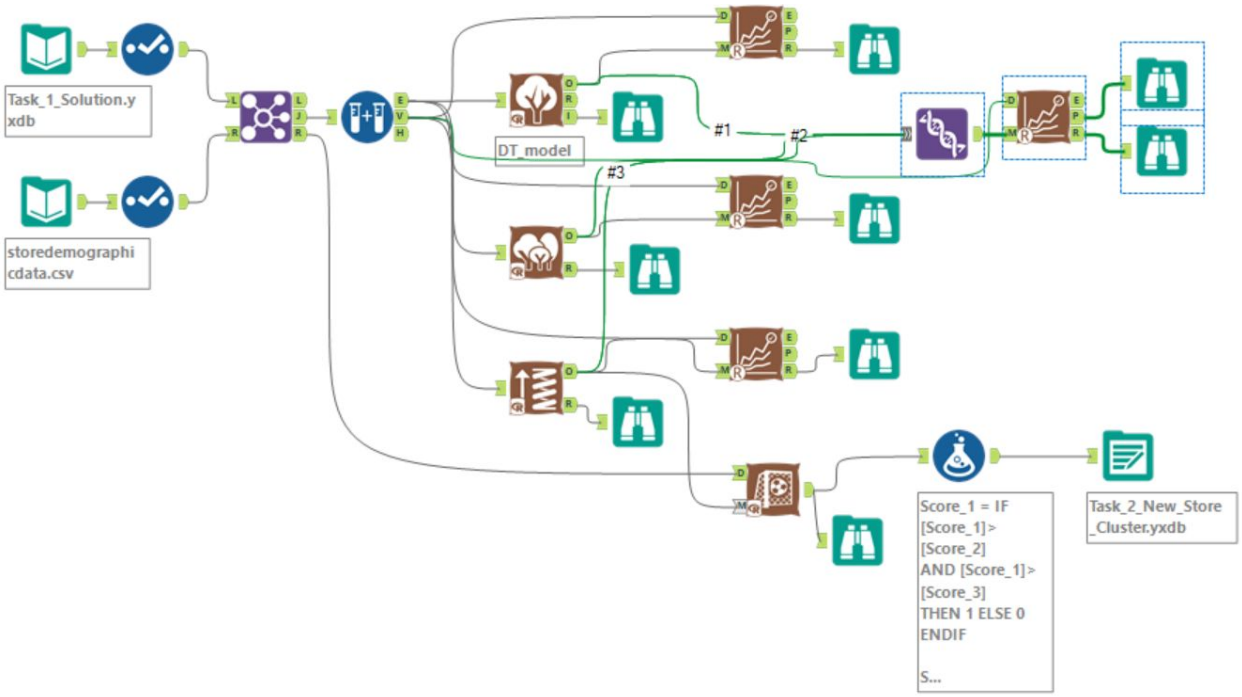
Tableau View of Historical Data, Existing Store Forecast and New Store Forecast

https://public.tableau.com/profile/jason.grenig#!/vizhome/Task3-TotalProduceSalesbyMonthHistoricalandForecast/TotalProduceSalesbyMonthHistoricalandForecast?publish=yes

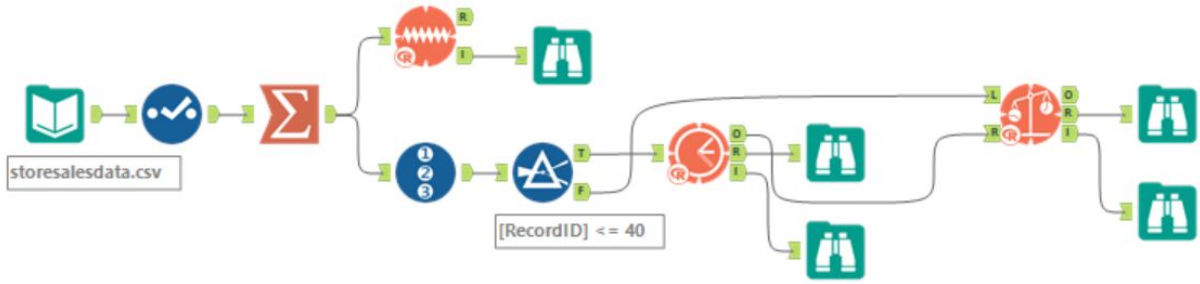
Alteryx Workflows:



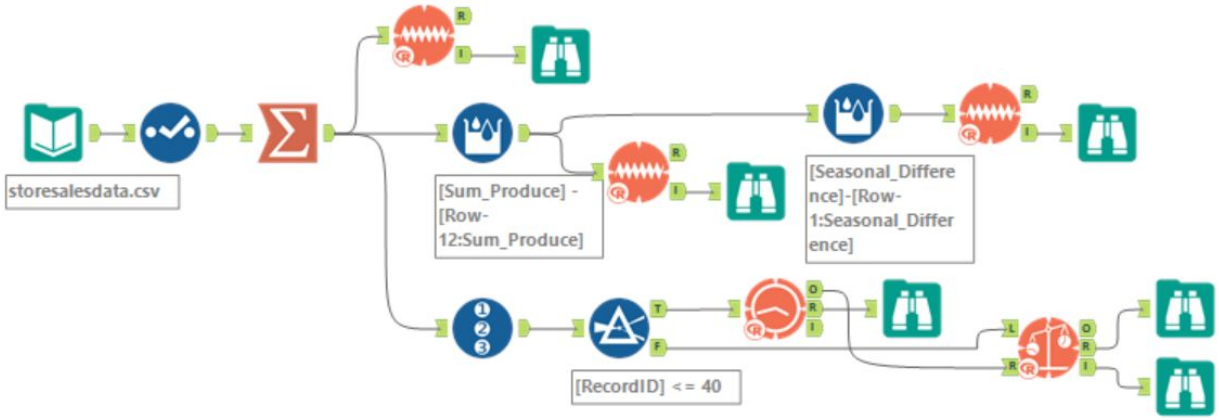
Alteryx Workflow for Task 1 - Get Clusters for Existing Stores



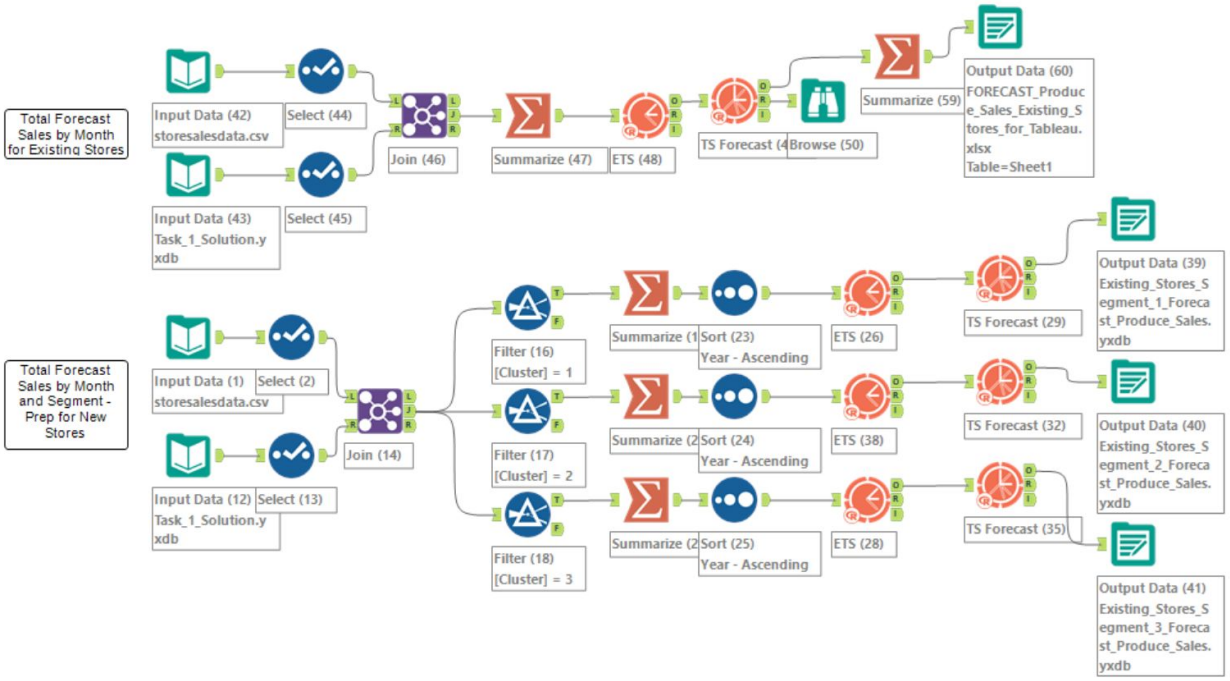
Alteryx Workflow for Task 2 - Get Clusters for New Stores



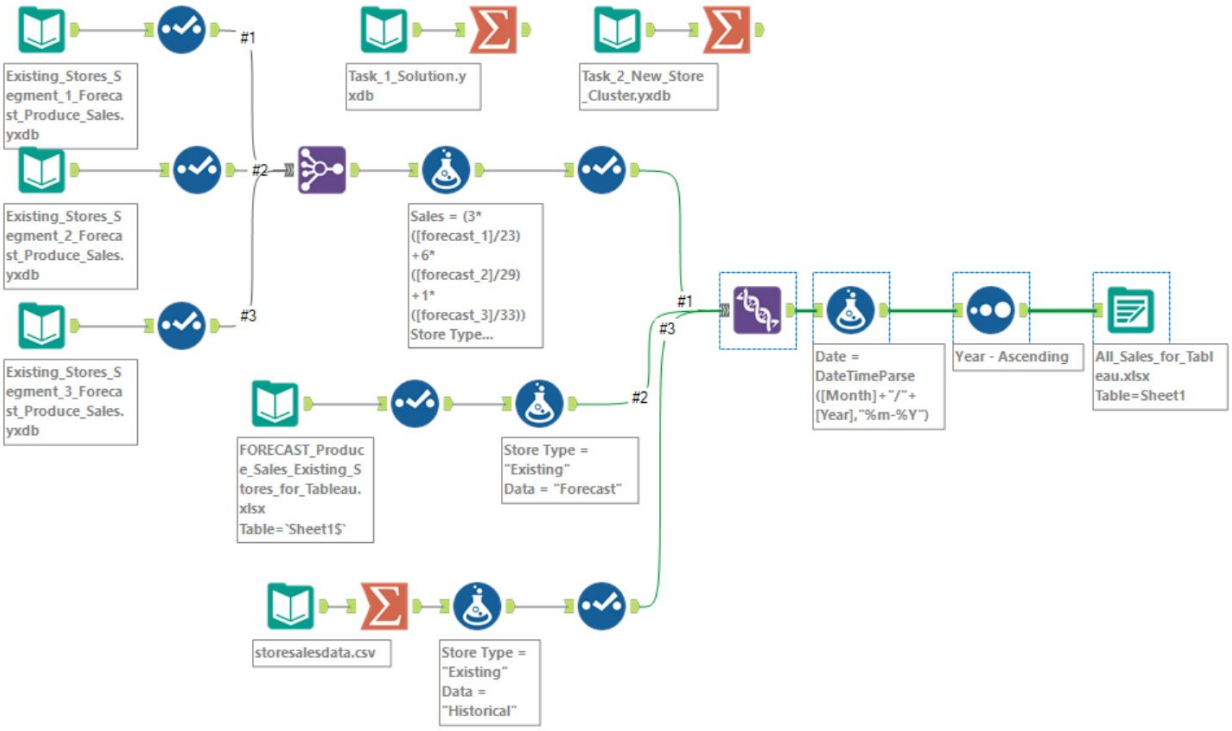
Alteryx Workflow for Task 3 - ETS Model Workflow



Alteryx Workflow for Task 3 - ARIMA Model Workflow



Alteryx Workflows: Total Forecast Sales for Existing Stores and Preparing New Store Segments



Alteryx Workflow for Total Forecast Sales for New Stores and Combining All Historical Sales