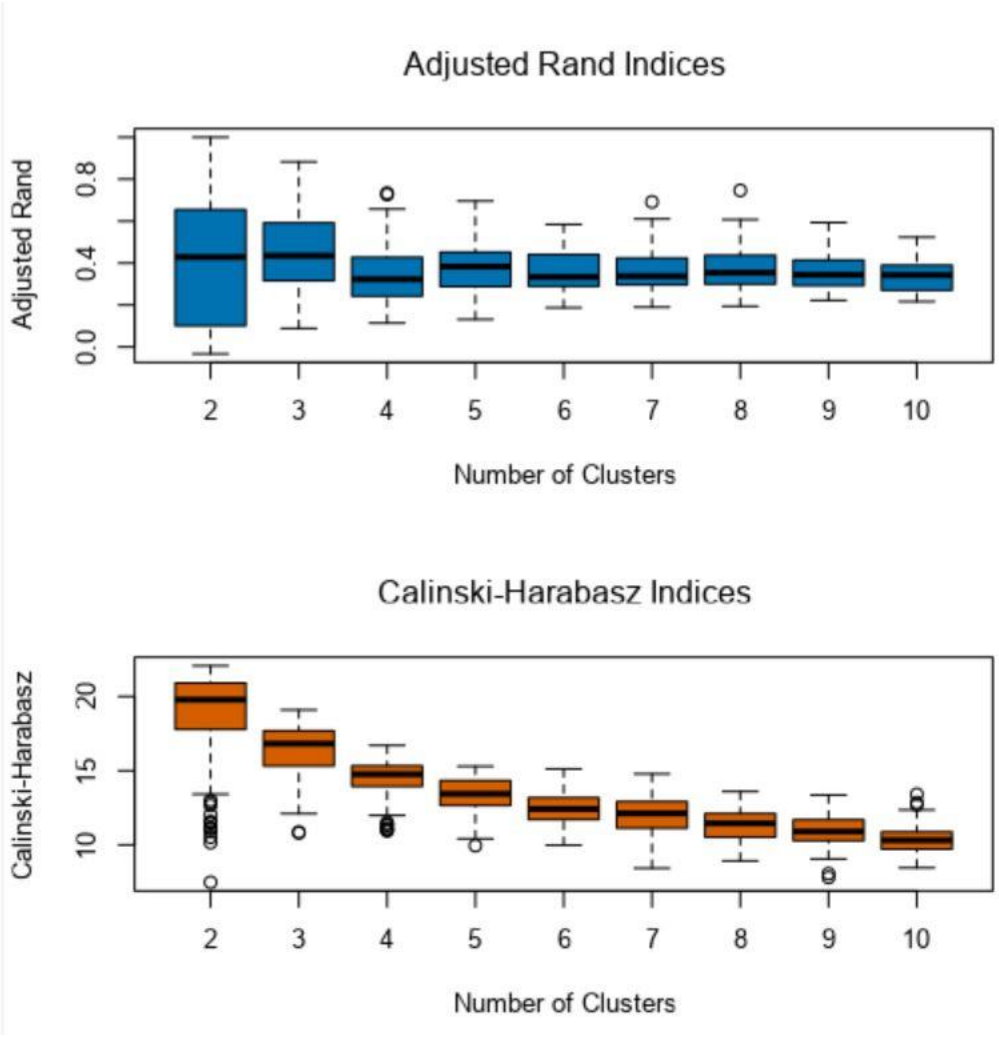
Christopher W. Hong

Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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



Based on the results of Adjusted Rand Indices and Calinski-Harabasz Indices, the optimal number of store formats is either 2 or 3. However, the IQRs of both indices of having 2 clusters are larger than those of having 3. Larger IQR means greater variance. Thus, the optimal number of store formats should be 3.

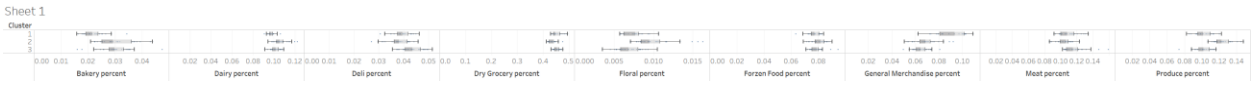
2. How many stores fall into each store format?

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

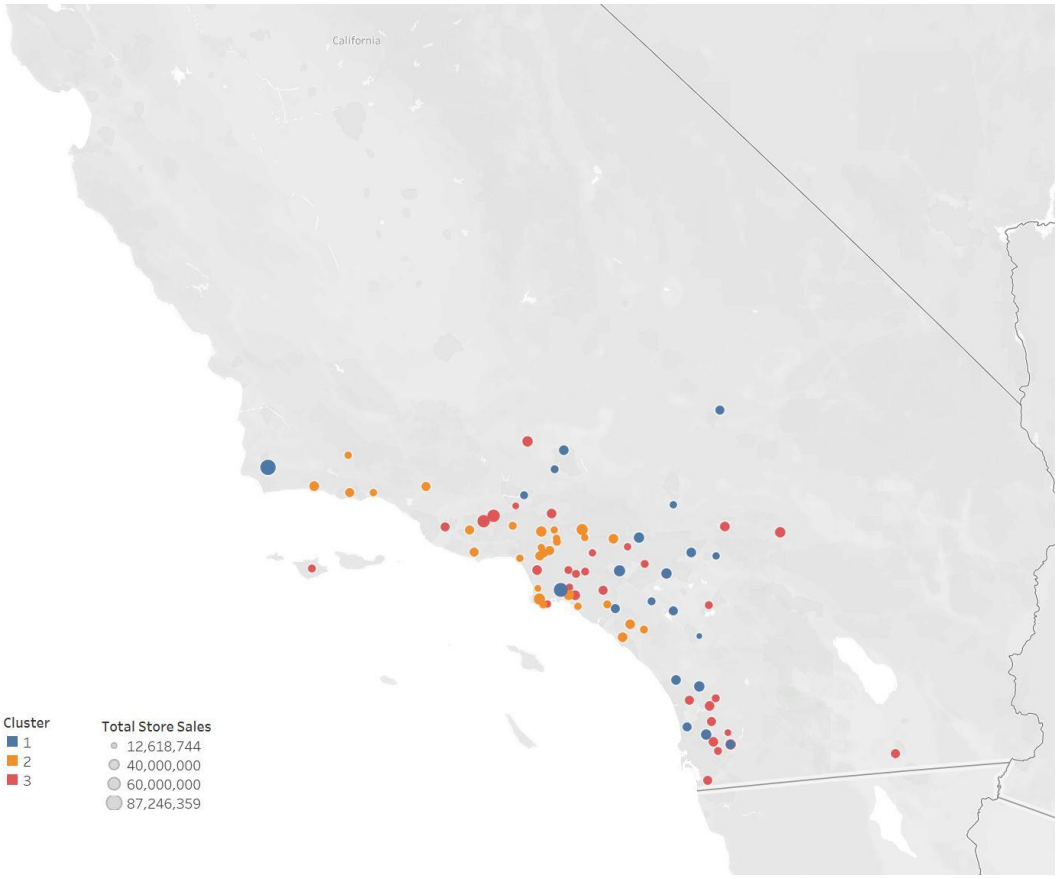
There are 23, 29 and 33 stores in store format 1, 2 and 3, respectively.

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



Base on the distribution of categories of store clusters, Cluster 1's General Merchandise percentage is much higher than Cluster 2 and 3; Cluster 2's Produce and Floral percentages are higher than others; and Cluster 3's Deli percentage is higher than others.

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/chris5448#!/vizhome/StoreClustersDistribution/Dashboard1?publish=yes

Christopher W. Hong

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.)

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
Random_Forest	0.8235	0.8251	0.7500	0.8000	0.8750
Boosted	0.8235	0.8543	0.8000	0.6667	1.0000

Based on the above model comparison report, Boosted model would be chosen to predict the best store format for the new stores since it has the highest accuracy rate and F1 score.

2. What format do each of the 10 new stores fall into? Please fill in the table below.

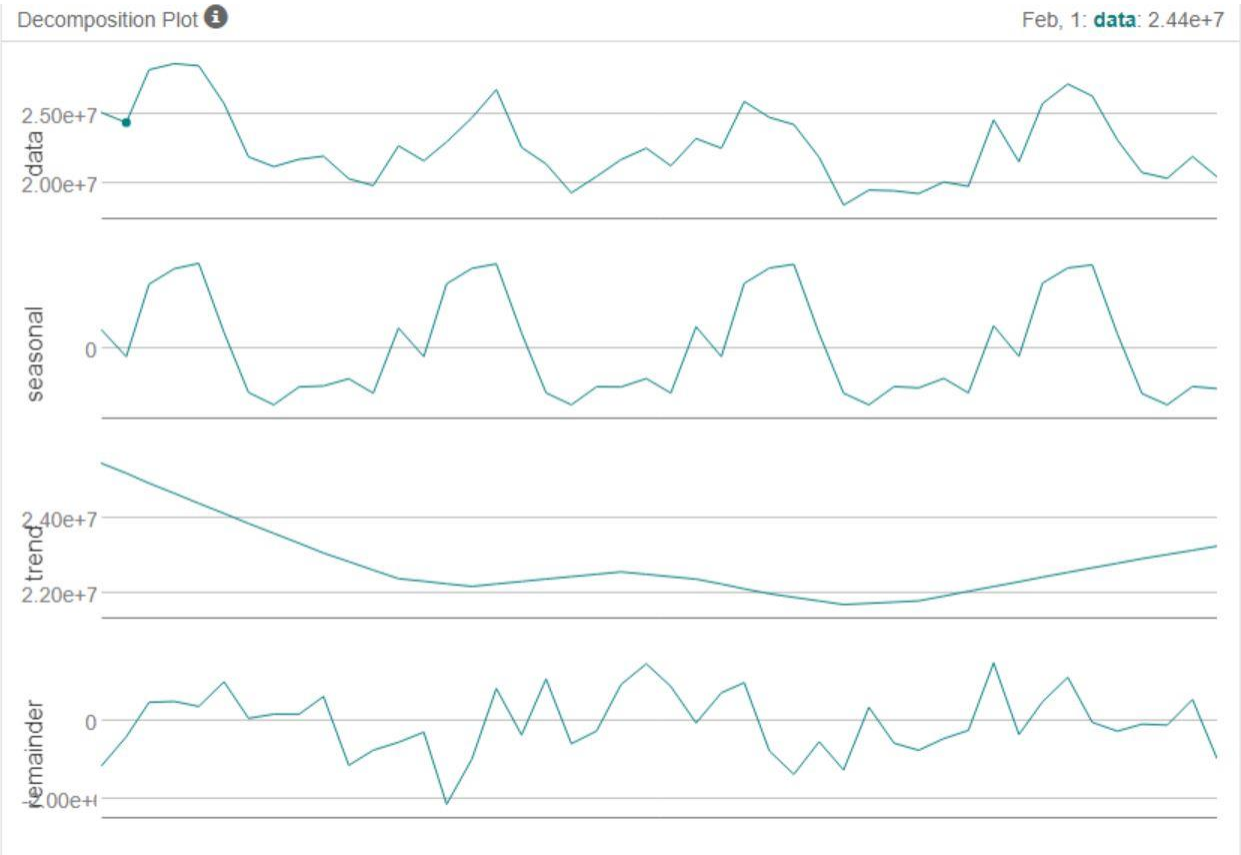
Cluster_1	Cluster_2	Cluster_3	Predicted_Cluster
0.348417	0.013522	0.638061	3
0.078987	0.804431	0.116582	2
0.486943	0.064498	0.448559	1
0.026597	0.935435	0.037968	2
0.019654	0.939601	0.040745	2
0.887418	0.003833	0.108749	1
0.028199	0.94173	0.030071	2
0.857561	0.005592	0.136847	1
0.00871	0.955864	0.035426	2
0.080423	0.641377	0.2782	2

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

Christopher W. Hong

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?



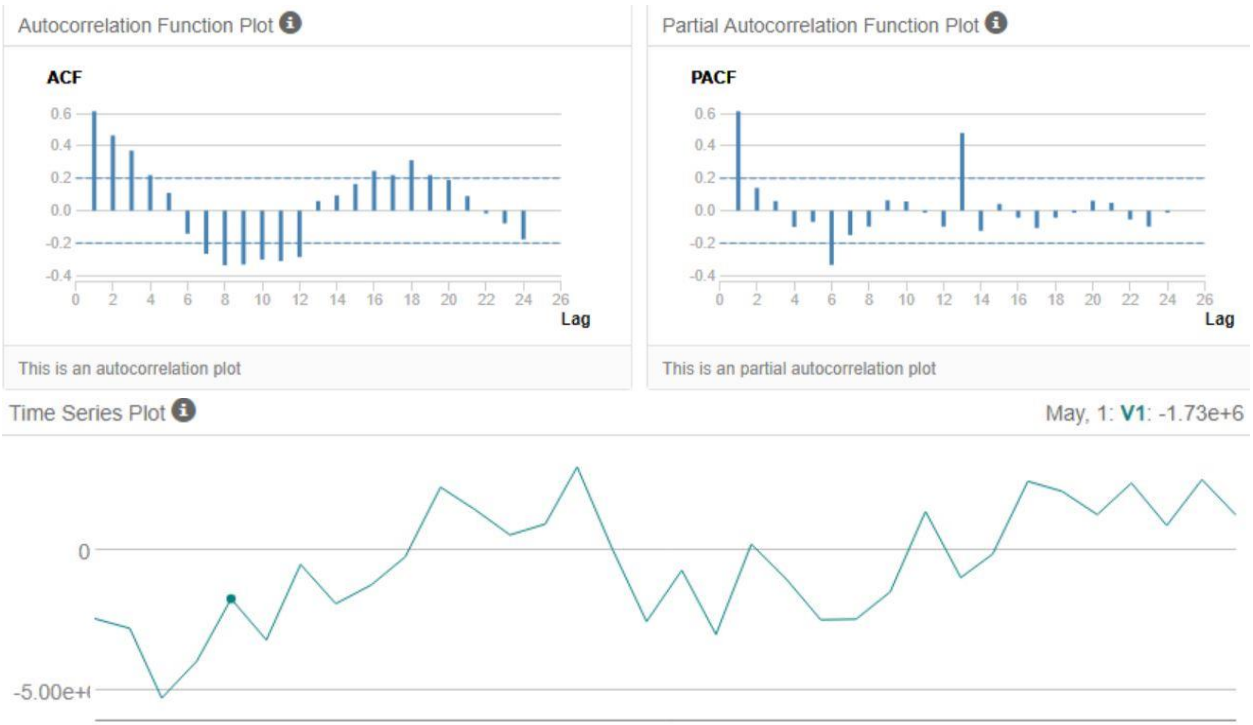
ETS:

Error: By referring to the above time series decomposition plot, looking at the error (labeled remainder) component, the valleys are shrinking and the peaks are slightly growing over time. Thus, the error should be multiplicative.

Trend: Looking at the trend component, there is no clear trend spotted, so the trend should be None.

Seasonality: Looking at the seasonal component, the line fluctuates with similar intervals over time, so there is presence of seasonality. The peeks are slightly shrinking overtime, so the seasonality should be multiplicative.

Configure ETX: base on the above ETS analysis, the final ETS model should be ETS (M, N, M).



ARIMA:

Check stationarity: The ACF plot shows a slow decay towards 0 correlation and the PACF plot shows significant correlation after lag1. Plus, the time series plot does not show any constant mean and variance. Thus, seasonal differencing is necessary.

Difference: After the first differencing, from the ACF, PACF and time series plots shown above, the time series is stationary.

Select AR and MA terms: Looking at the above ACF and PACF plots again, the significant negative autocorrelations at lag1 in both plots suggest MA 1 term; also, the significant negative lags at 12 and cut off at 24 suggests seasonal MA 1 term. Thus, the final ARIMA model should be ARIMA (0, 1, 1) (0, 1, 1) [12].

Accuracy Measures:

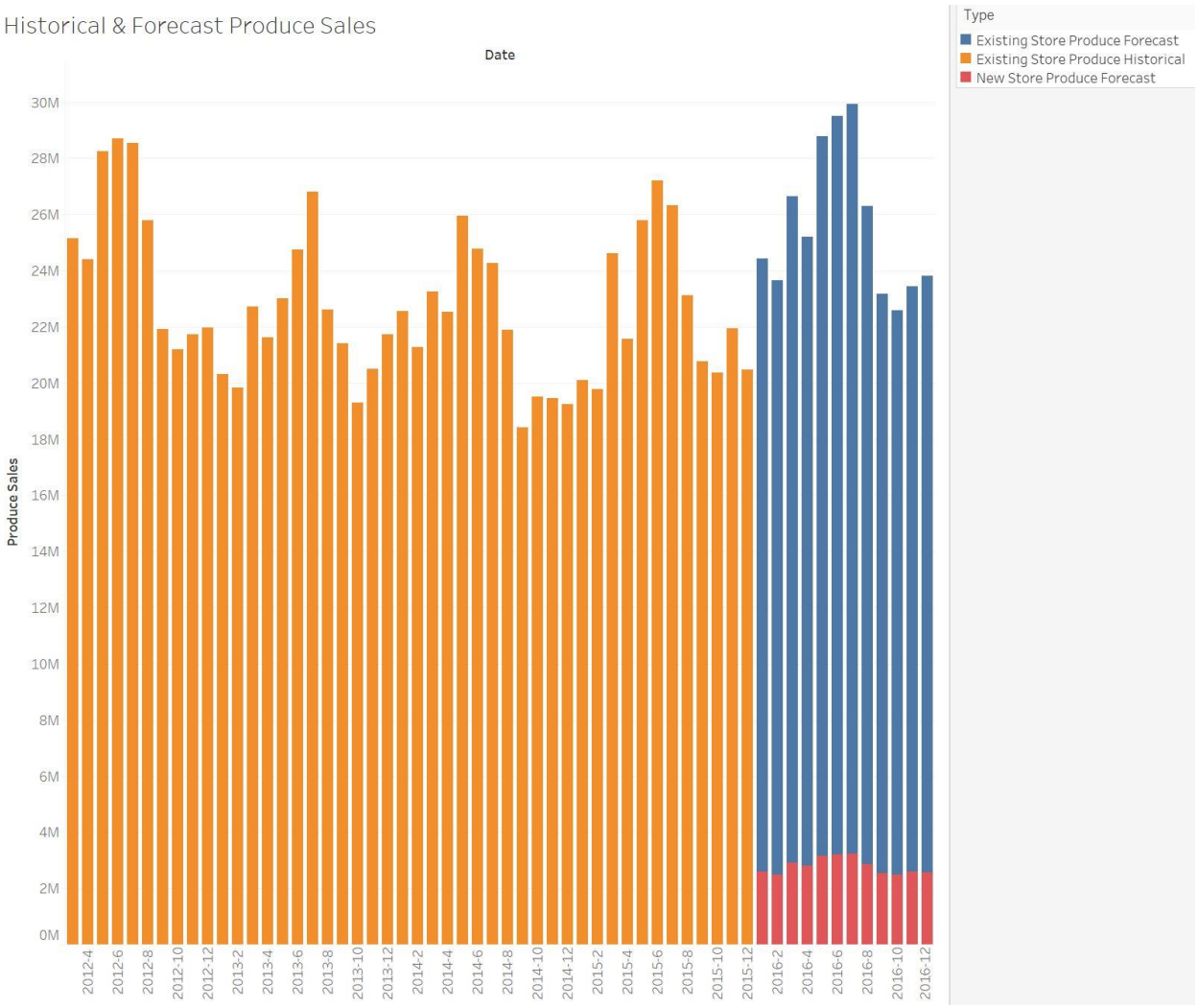
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_MNM	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	-492238.83	792197.3	735878.2	-2.1992	3.3098	0.433

The above accuracy measures on the validation set show that ETS model outperforms the ARIMA one on all metrics. Therefore, the ETS model would be used to forecast future sales.

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.

Month	New Stores	Existing Stores
Jan-16	$2,588,357	$21,829,060
Feb-16	$2,498,567	$21,146,330
Mar-16	$2,919,067	$23,735,687
Apr-16	$2,797,280	$22,409,515
May-16	$3,163,765	$25,621,829
Jun-16	$3,202,813	$26,307,858
Jul-16	$3,228,212	$26,705,093
Aug-16	$2,868,915	$23,440,761
Sep-16	$2,538,372	$20,640,047
Oct-16	$2,485,732	$20,086,270
Nov-16	$2,583,448	$20,858,120
Dec-16	$2,562,182	$21,255,190

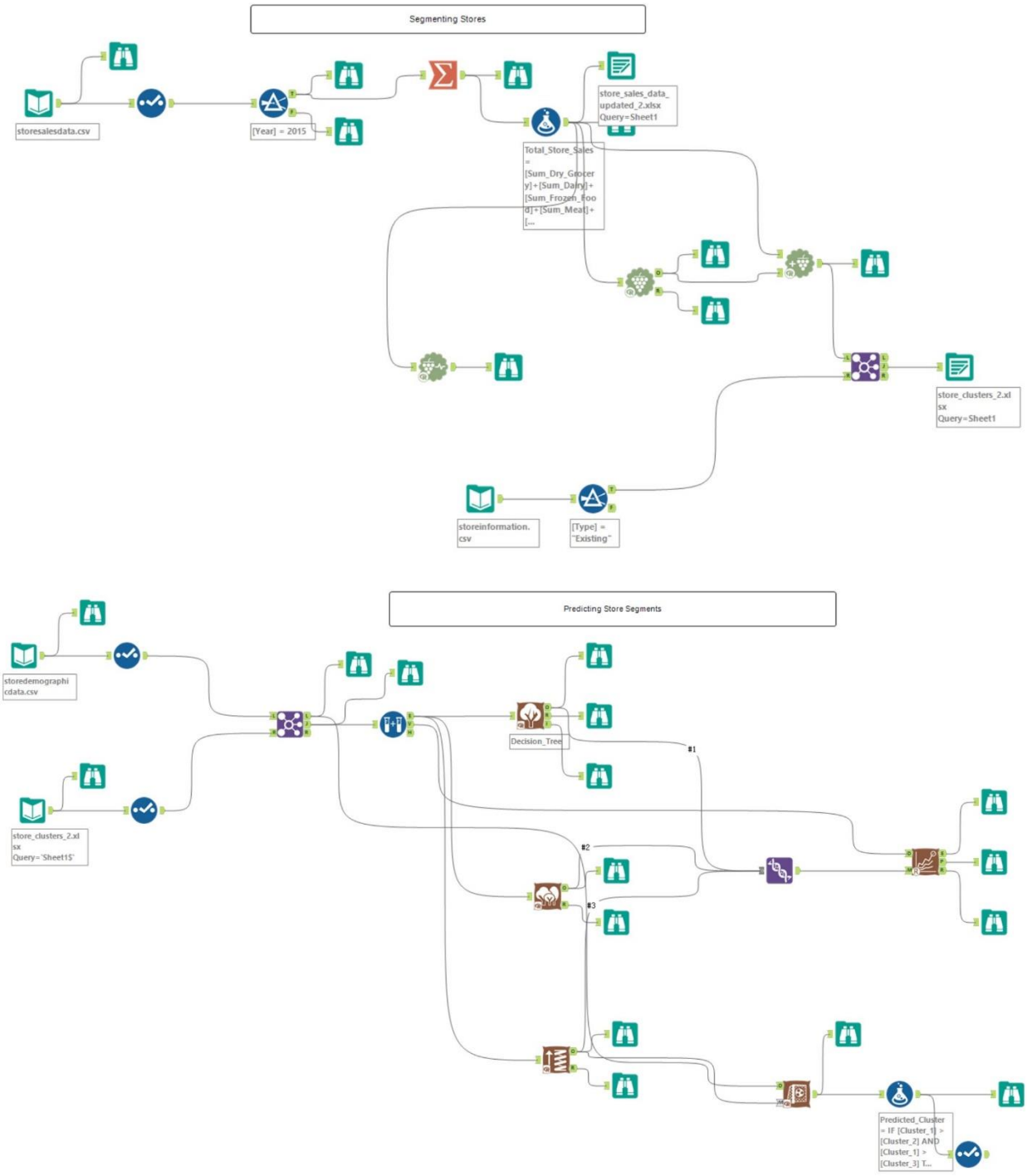
Christopher W. Hong



https://public.tableau.com/profile/chris5448#!/vizhome/HistoricalForecastProduceSales/HistoricalForecastProduceSales?publish=yes

Christopher W. Hong

Alteryx workflow:



Christopher W. Hong