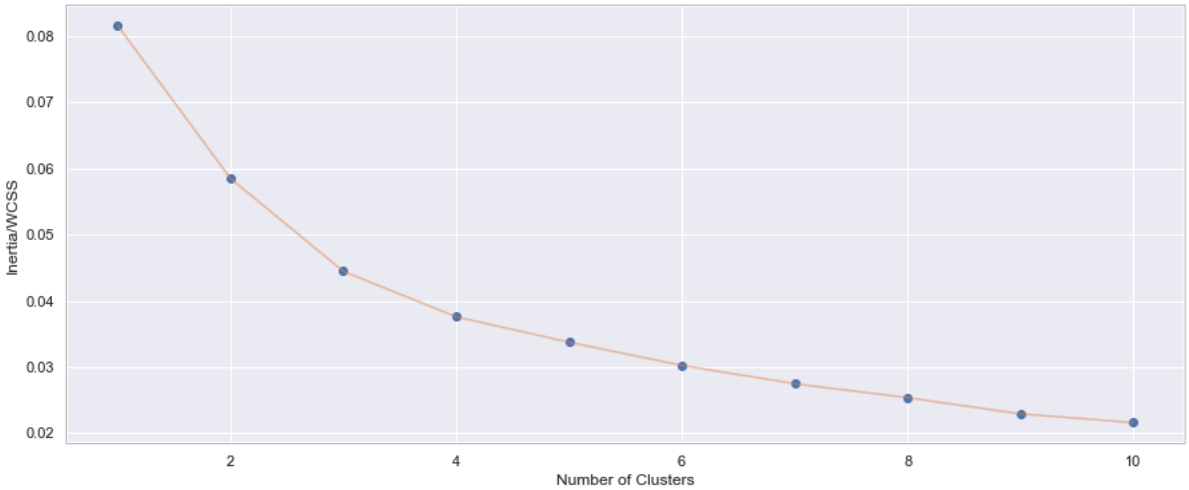
Project #7
Combining Predictive Techniques

Claudia Dai

March 2019

1 Determine Store Format for Existing Stores

In order to determine store formats for the existing store, we turn to k-means clustering. The main input for k-means clustering is the number of clusters, or the number of store formats. The optimum number of store formats is determined by minimizing the cluster sum of squares (WCSS). As the number of clusters increase, the WCSS keeps decreasing. First, WCSS decreases steeply – and at some point the rate of decrease slows down, which we call the "elbow". The elbow is a good indication of the optimum number of clusters we should use.



According to our analysis, the optimum number of store formats seems to be around 3 formats. We also take a look at the silhouette coefficient and the calinski harabaz score to confirm our first observation. We cannot use the adjusted rand score here, since we are not working with a dataset that has a target variable, so we cannot compute true from predicted. See documentation here.

The silhouette distance shows to which extent the distance between objects of the same class differ from the mean distance between objects from different clusters. The silhouette coefficient takes values ranging [-1,1], where a value close to -1 corresponds to bad clustering results, while values closer to 1 correspond to dense and well-defined clusters.

The Calinski Harabaz score is also known as the variance-ratio criterion. It looks at the ratio between the within-cluster and the between-cluster dispersions. The higher this score is, the better is the clustering solution.

```
silhouette coefficient for `2` clusters => 0.2465
silhouette coefficient for `3` clusters => 0.2889
silhouette coefficient for `4` clusters => 0.2594
silhouette coefficient for `5` clusters => 0.2300
silhouette coefficient for `6` clusters => 0.2064
silhouette coefficient for `7` clusters => 0.2139
silhouette coefficient for `8` clusters => 0.2218
silhouette coefficient for `9` clusters => 0.2103
silhouette coefficient for `10` clusters => 0.2250

calinski harabaz score for `2` clusters => 32.7890
calinski harabaz score for `3` clusters => 34.2235
calinski harabaz score for `4` clusters => 31.6282
calinski harabaz score for `5` clusters => 28.1571
calinski harabaz score for `6` clusters => 25.7408
calinski harabaz score for `7` clusters => 24.9849
calinski harabaz score for `8` clusters => 24.6480
calinski harabaz score for `9` clusters => 23.8399
calinski harabaz score for `10` clusters => 24.0989
```

Based on the silhouette coefficient, 3 clusters have the best score, and based on the Calinski Harabaz score, also 3 clusters have the best score.

Running k-means with 3 clusters yield the following result: 23 stores fall under the store format 1, 29 stores under store format 2, and 33 stores under store format 3.

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

Cluster 1 is the smallest in size compared to the other two. Cluster 2 has the biggest separation value, meaning that cluster 2 is more separated from the other clusters than the other clusters are. Cluster 3 is the most compact cluster due to the smallest average distance and the biggest size. Cluster 3 also has the farthest outlier when looking at the max distance.

2 Formats for New Stores

We try the following algorithms: decision tree, random forest and gradient boosting. We use a 20% validation sample with random seed of 3 to test and compare the models.

Decision tree yielded an accuracy score of 0.7059 with 5 misclassified samples.

	precision	recall	f1-score	support
1	0.50	1.00	0.67	4
2	0.83	0.83	0.83	6
3	1.00	0.43	0.60	7
micro avg	0.71	0.71	0.71	17
macro avg	0.78	0.75	0.70	17
weighted avg	0.82	0.71	0.70	17

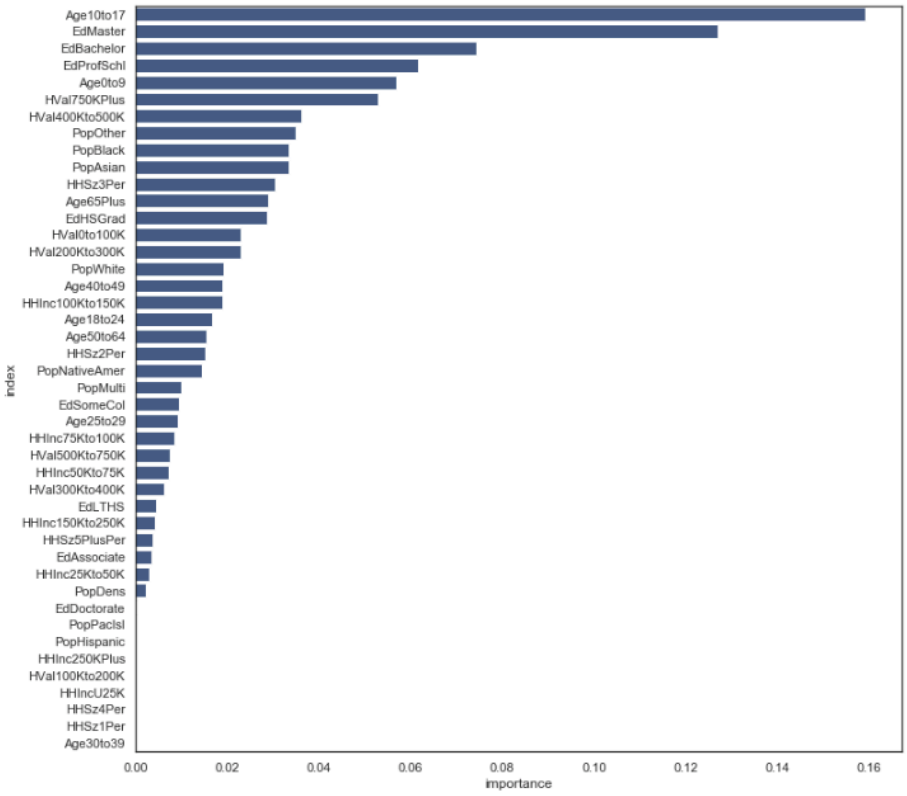
Random forest yielded an accuracy of 0.8235 with 3 misclassified samples.

	precision	recall	f1-score	support
1	0.80	1.00	0.89	4
2	0.83	0.83	0.83	6
3	0.83	0.71	0.77	7
micro avg	0.82	0.82	0.82	17
macro avg	0.82	0.85	0.83	17
weighted avg	0.83	0.82	0.82	17

Finally, gradient boosting also yielded an accuracy of 0.8235 with 3 misclassified samples.

	precision	recall	f1-score	support
1	0.80	1.00	0.89	4
2	0.83	0.83	0.83	6
3	0.83	0.71	0.77	7
micro avg	0.82	0.82	0.82	17
macro avg	0.82	0.85	0.83	17
weighted avg	0.83	0.82	0.82	17

Gradient boosting scores the same as random forest in the amount of least misclassified samples, the highest accuracy, and in precision, recall and f-1 score. We decide to use gradient boosting algorithm. The most important features are Age10to17, EdMaster, and EdBachelor.

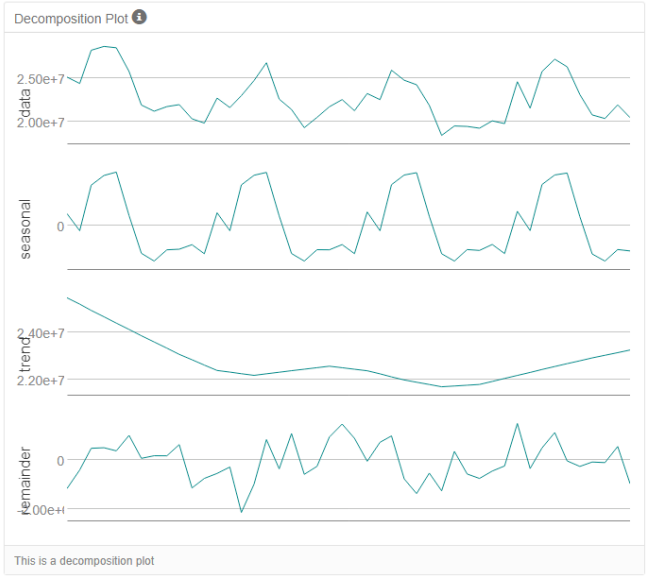


The 10 new stores fall into the following segments.

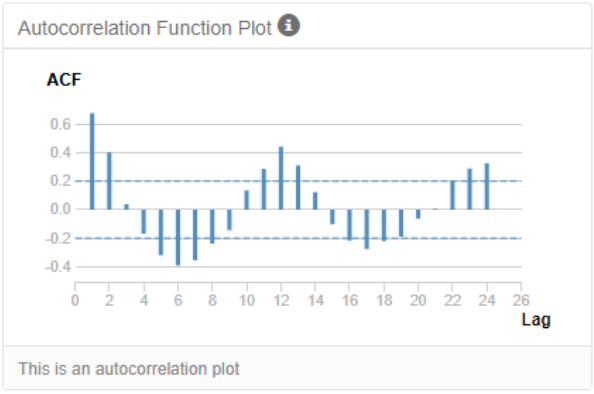
Store Number	Segment
S0086	3
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

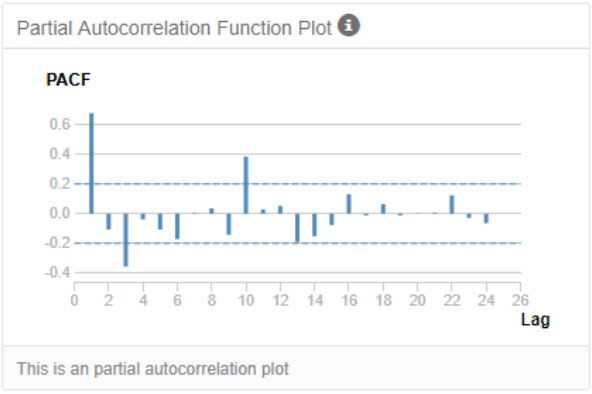
3 Predicting Produce Sales – with Alteryx

In Alteryx, the ETS model is ETS(M,N,M), because the seasonality is marginally decreasing over time (as opposed to Python's analysis, where the seasonality remains stagnant).



For the ARIMA model, we have lag-2 and ARIMA(0,1,2)(0,1,0) since we take the seasonal difference and seasonal first difference to make the TS stationary.





Our error metrics indicate the following for ETS:

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788

Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

The error metrics for ARIMA indicate:

Information Criteria:

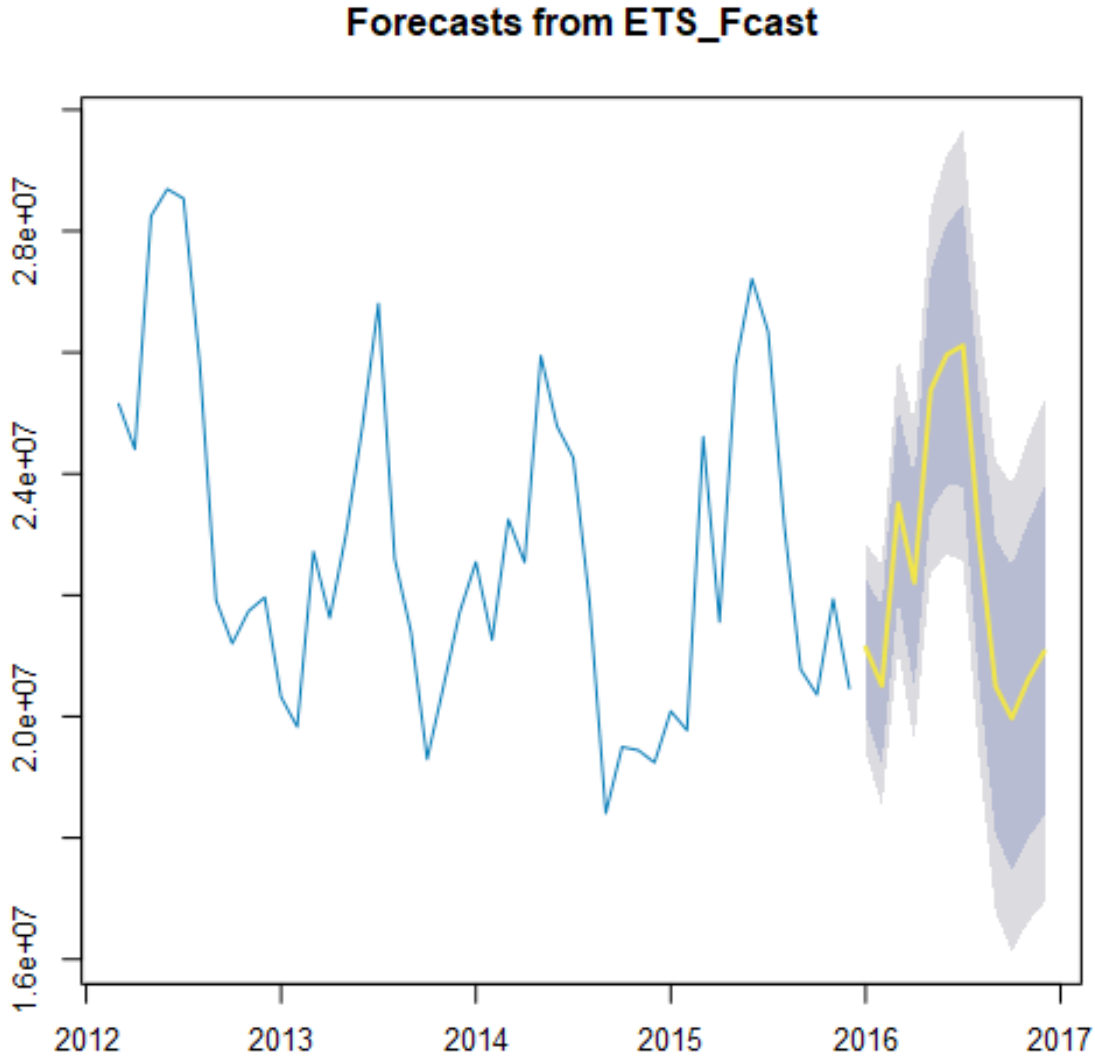
AIC	AICc	BIC
880.4445	881.4445	884.4411

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

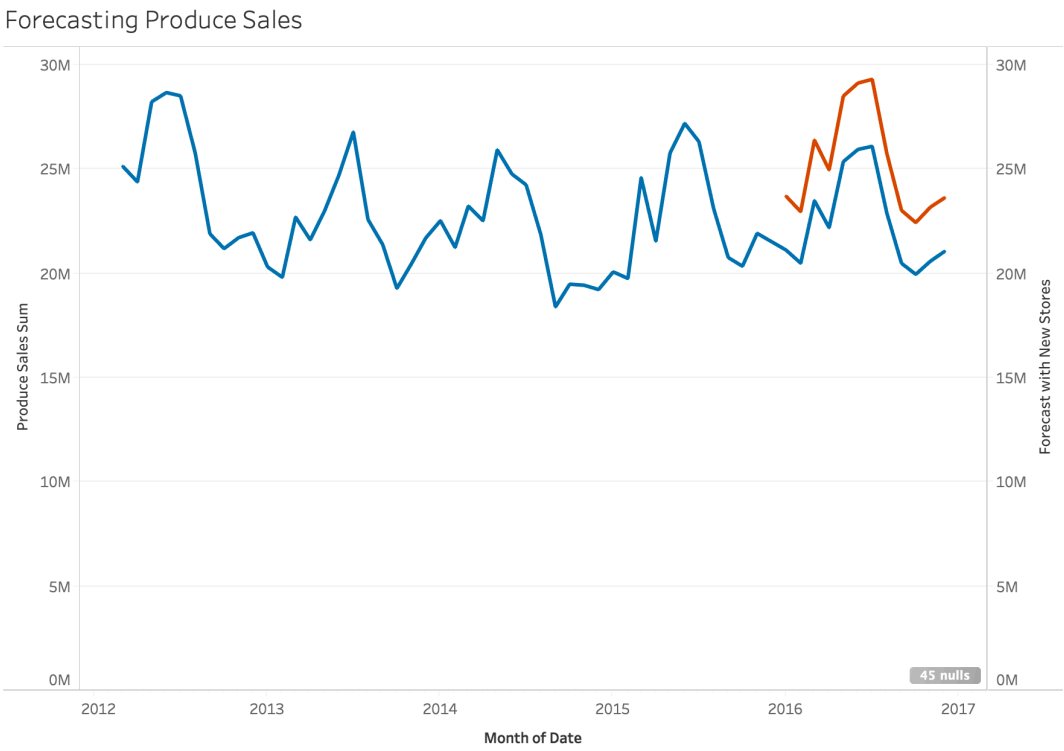
ETS has an RMSE of 1,020,596.904, while ARIMA has an RMSE of 1,042,209.853. ETS' MASE is 0.4507, while ARIMA's MASE is 0.4120.

Here are the actual and forecasted values with 80% and 90% confidence levels:



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21136208.135109	22863751.647268	22265788.122301	20006628.147918	19408664.622951
2016	2	20506604.689889	22485979.825084	21800848.524632	19212360.855146	18527229.554694
2016	3	23506131.457397	25923604.543644	25086832.145154	21925430.769639	21088658.371149
2016	4	22207971.238436	24819551.269971	23915591.635728	20500350.841144	19596391.206902
2016	5	25376698.322185	28385663.710055	27344155.037671	23409241.606699	22367732.934316
2016	6	25963559.446576	29258459.785154	28117978.976999	23809139.916154	22668659.107998
2016	7	26113357.20163	29660962.648063	28433011.720628	23793702.682632	22565751.755197
2016	8	22904671.917667	26542287.656104	25283181.003148	20526162.832187	19267056.179231
2016	9	20499151.00121	24219766.868399	22931930.953799	18066371.048621	16778535.134021
2016	10	19970808.947309	23811395.340529	22482033.410444	17459584.484174	16130222.554089
2016	11	20602232.29737	24592072.351437	23211048.483736	17993416.111005	16612392.243304
2016	12	21072786.922156	25209451.080778	23777606.230281	18367967.61403	16936122.763534

Month	New Stores	Existing Stores
Jan 2016	2,584,384	21,136,208
Feb 2016	2,470,874	20,506,605
Mar 2016	2,906,308	23,506,131
Apr 2016	2,771,532	22,207,971
May 2016	3,145,849	25,376,698
Jun 2016	3,183,909	25,963,559
Jul 2016	3,213,978	26,113,357
Aug 2016	2,858,247	22,904,672
Sep 2016	2,538,174	20,499,151
Oct 2016	2,483,550	19,970,809
Nov 2016	2,593,089	20,602,232
Dec 2016	2,570,200	21,072,787



Appendix A: Predicting Produce Sales – with Python

We compare the performance of a non-dampened ETS(M,N,A) model with a SARIMAX(2,1,3)x(0,1,0,12) model.

The ETS model has a MASE of 0.4355.

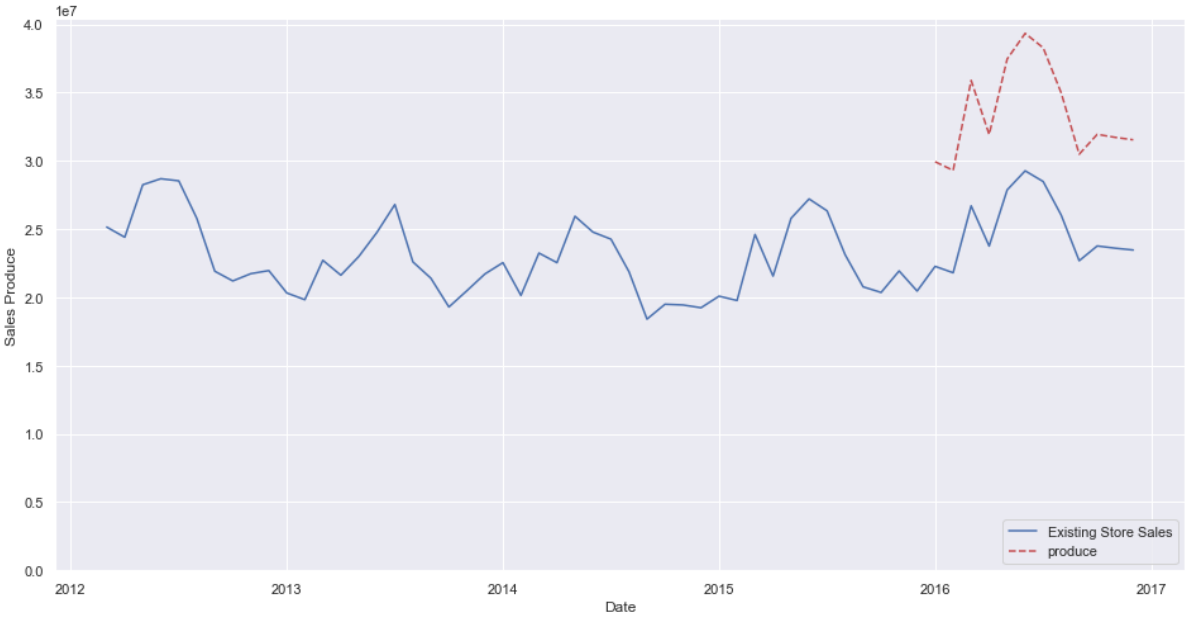
Metric	MSE	RMSE	MDAE	MAE	MSLE	MAPE	R2
Value	813680594046.156	902042.457	791492.891	765136.943	0.002	3.340	0.816

The Arima model has a MASE of 0.3449.

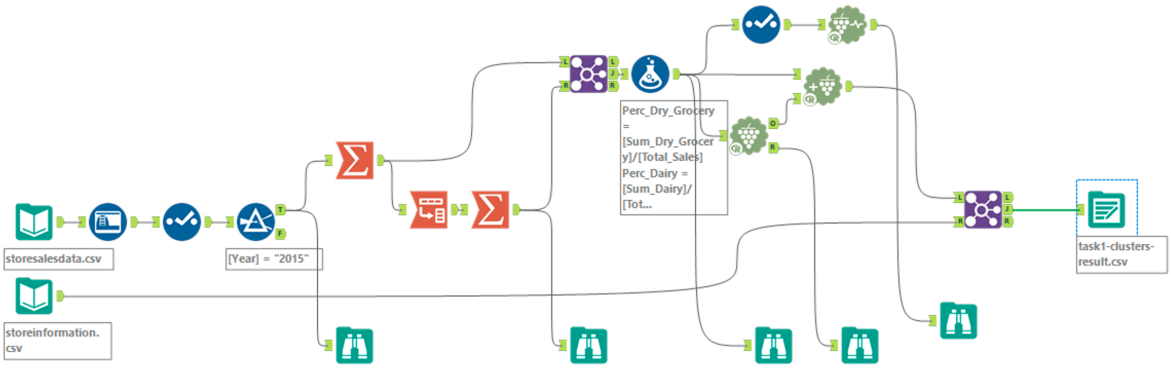
Metric	MSE	RMSE	MDAE	MAE	MSLE	MAPE	R2
Value	557404368485.785	746595.184	629454.341	605963.132	0.001	2.861	0.874

We managed to get a MASE of 0.4355 for ETS(M,N,A) while we got a MASE of 0.3449 for our ARIMA model. ETS has an RSME of 902,042.457 while ARIMA has an RSME of 746,595.184. Therefore, we choose the Arima model.

Month	New Stores	Existing Stores
Jan 2016	7,661,736.586	22273762.788
Feb 2016	7,499,705.210	21802714.431
Mar 2016	9,186,249.120	26705738.513
Apr 2016	8,171,767.123	23756494.423
May 2016	9,587,469.090	27872142.283
Jun 2016	10,068,649.383	29271002.135
Jul 2016	9,798,381.641	28485295.199
Aug 2016	8,952,828.553	26027151.578
Sep 2016	7,802,970.577	22684350.178
Oct 2016	8,175,330.519	23766853.724
Nov 2016	8,118,912.677	23602838.997
Dec 2016	8,072,964.501	23469261.084



Appendix B: Alteryx Workflow for Task 1 Clustering



Appendix C: Alteryx Workflow for Task 3 Model Comparison