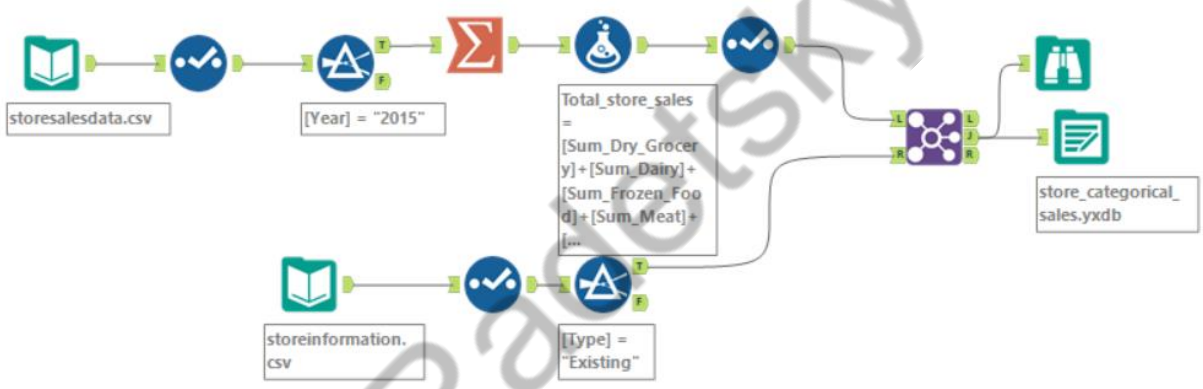
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

The optimal number of store formats is 3. I arrived at this number by using percentages of sales in different categories as variables for the cluster model and then validating the optimal number of clusters with the K-Centroid Diagnostics tool. The step-by-step are as followed:

Step 1: Preparing percentage of each product category per store data set, expressed as percentages of total sales per store.



Workflow 1: Preparing percentage of each product category per store

As the result of Workflow 1, I obtained the data set saved in the 'store_categorical_sales.yxdb' as shown in Table 1:

13 of 13 Fields | Cell Viewer

Record	Store	Total_store_sales	Percent_Dr...	Percent_Dairy	Percent_Fro...	Percent_Meat	Percent_Pro...	Percent_Floral	Percent_Deli	Percent_Bakery	Percen...	Zip	Type
1	S0001	23508945.82	46.13	10.31	7.72	10.77	9.72	0.68	4.35	3.55	6.77	92027	Existing
2	S0002	17334619.57	45.75	10.64	7.88	11.49	10.13	0.74	3.98	2.97	6.41	92040	Existing
3	S0003	30240661.99	42.13	10.24	6.9	11.47	12.54	0.96	4.18	3.61	7.97	90026	Existing
4	S0004	27913890.97	45.46	9.71	8.03	12.77	10.04	0.61	4.18	3.45	5.75	92078	Existing
5	S0005	27825886.17	44.02	10.63	8.63	10.19	13.11	0.89	3.54	2.26	6.74	90019	Existing
6	S0006	34625420.87	45.98	9.86	7.44	10.69	10.18	0.83	4.63	3.54	6.84	92071	Existing
7	S0007	30863087.31	44.21	10.21	8.92	11.51	10.34	0.78	4.63	3.02	6.37	93010	Existing
8	S0008	29181879.77	42.85	10.87	8.03	10.69	12.37	0.79	3.26	2.24	8.9	90020	Existing
9	S0009	40678620.31	44.23	10.04	7.71	11.89	10.35	0.73	4.42	2.58	8.07	93536	Existing
10	S0010	21703928.12	43.14	11.5	8.03	10.71	12.16	1.04	3.71	3.81	5.89	93108	Existing
11	S0011	22395673.62	45.38	9.4	7.19	12.14	9.85	0.79	4.87	3.24	7.14	91752	Existing
12	S0012	49186541.13	46.6	9.73	7.44	11.44	9.43	0.68	4.51	1.86	8.32	93436	Existing

Table 1: Percentage of each category per store

Grace Radetsky 2020

Step 2: With the K-Centroids Diagnostics to validate the optimal number of clusters/store formats. I will use reiterate the diagnostics test with 3 different clustering methods: K-Means, K-Medians, and Neutral Gas. And then analyze the Adjusted Rand and CH indices resulted from each method. I will be looking for the number of clusters that have the highest median Adjusted Rand and CH indices while having the most compacted spread.

K-Means

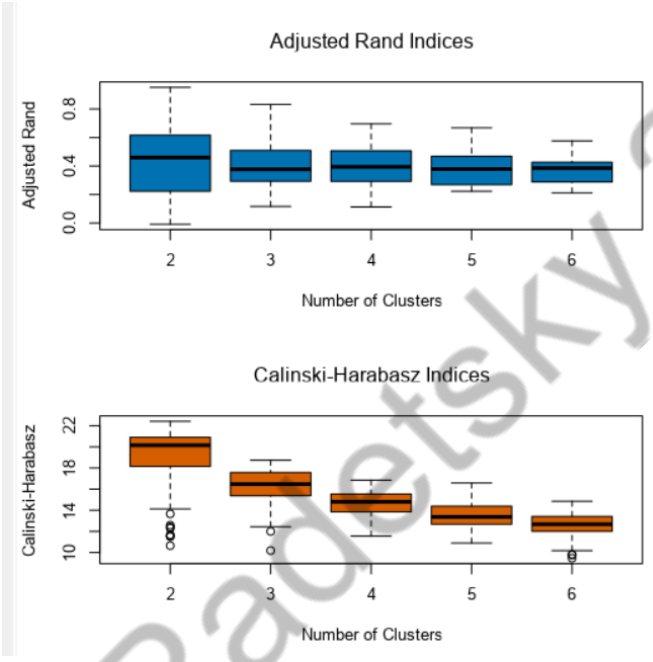


Figure 1: Performance indices of multiple clustering models using K-Mean method

With the K-Mean method, two-clusters-outcome seems to have the highest median Adjusted rand and CH indices. However, the spread of Adjusted Rand indices is too wide and there are so many outlier CH values. Therefore, two-clusters-outcome is eliminated.

The three-clusters-outcome seems to be the most optimal as the median Adjusted Rand and CH indices are ones of the highest while the spreads for those indices are very narrow.

K-Median

Grace Radetsky 2020

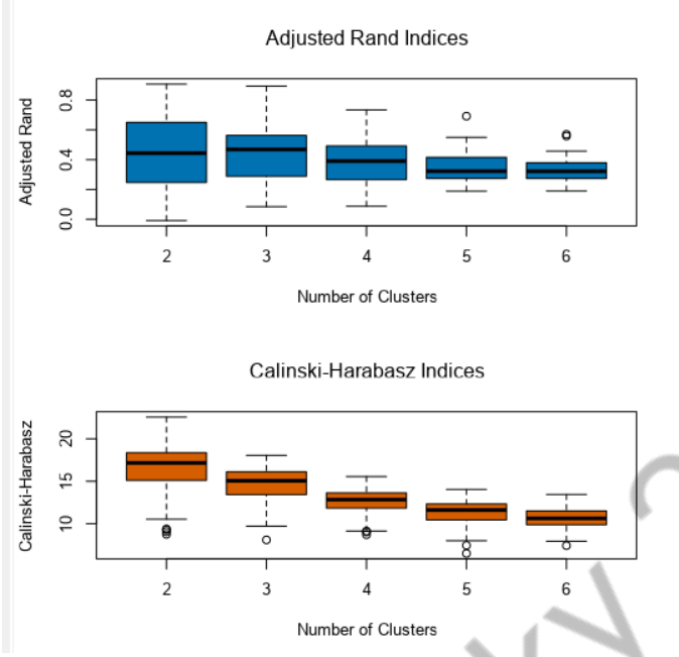


Figure 2: Performance indices of multiple clustering models using K-Median method

Similar to the result using the K-Means, K-median method also indicates that 3 is the most optimal number of clusters

Neutral Gas

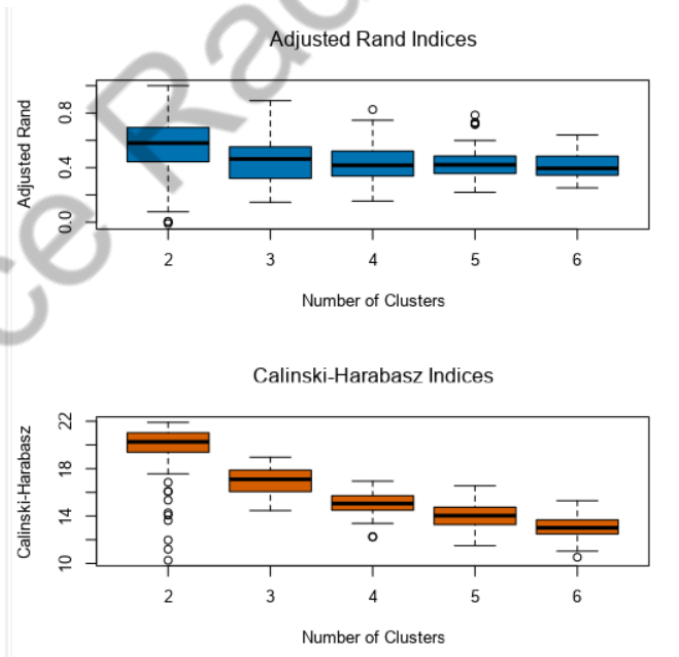
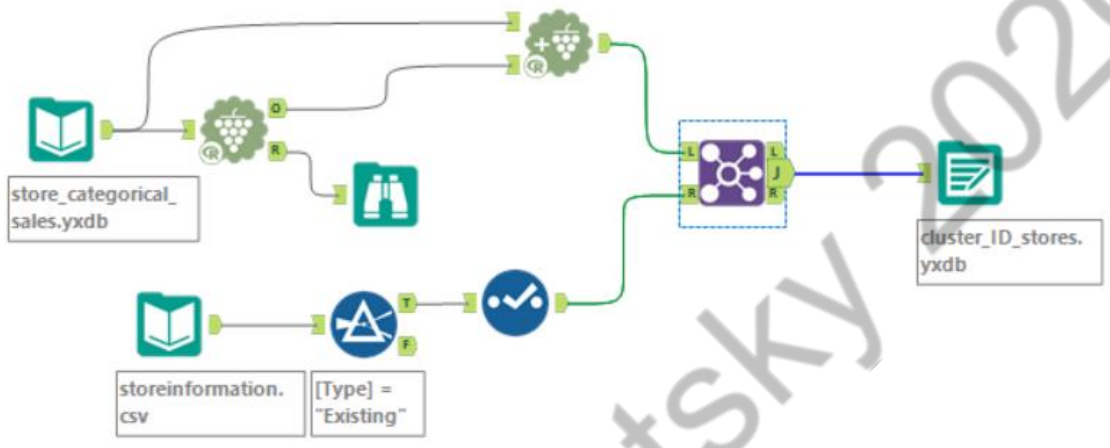


Figure 3: Performance indices of multiple clustering models using Neutral Gas method

Grace Radetsky 2020

The Neural Gas clustering method also confirms the finding of the two above methods. Therefore, I will put the 3-cluster model into production.

Step 3: I will use the 'store_categorical_sales.yxdb' data set obtained in step 1 to start the clustering process. As a result, stores will be grouped into 3 different groups representing three store formats.



Workflow 2: Clustering existing stores by categorical sales percentage

1.2. How many stores fall into each store format?

As a result of Workflow 2, stores will be grouped into 3 different groups representing three store formats as shown in Table 2

Store	Percent_ Dry_Groc er	Percent_ Dairy	Percent_ Frozen	Percent_ Meat	Percent_ Produce	Percent_ Floral	Percent_ Deli	Percent_ Bakery	Percent_ Gen_Mer	luster
S0001	46.13	10.31	7.72	10.77	9.72	0.68	4.35	3.55	6.77	1
S0002	45.75	10.64	7.88	11.49	10.13	0.74	3.98	2.97	6.41	1
S0003	42.13	10.24	6.9	11.47	12.54	0.96	4.18	3.61	7.97	2
S0004	45.46	9.71	8.03	12.77	10.04	0.61	4.18	3.45	5.75	1
S0005	44.02	10.63	8.63	10.19	13.11	0.89	3.54	2.26	6.74	2
S0006	45.98	9.86	7.44	10.69	10.18	0.83	4.63	3.54	6.84	1
S0007	44.21	10.21	8.92	11.51	10.34	0.78	4.63	3.02	6.37	1
S0008	42.85	10.87	8.03	10.69	12.37	0.79	3.26	2.24	8.9	2

Table 2: Clustering results for existing stores

25 stores belong to cluster/ format 1; 35 stores belong to cluster/format 2; and 25 stores belong to cluster/format 3

Grace Radetsky 2020

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

The Analysis Report 1 summarizes the performance of the model:

Summary Report of the K-Means Clustering Solution Clusterbycategorysales

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Percent_Dry_Grocer + Percent_Dairy + Percent_Frozen + Percent_Meat + Percent_Produce + Percent_Floral + Percent_Deli + Percent_Bakery + Percent_Gen_Mer, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.100598	4.823985	2.193986
2	35	2.475232	4.410756	1.9441
3	25	2.287649	3.582763	1.723182

Convergence after 8 iterations.
Sum of within cluster distances: 196.33929.

	Percent_Dry_Grocer	Percent_Dairy	Percent_Frozen	Percent_Meat	Percent_Produce	Percent_Floral	Percent_Deli
1	0.527735	-0.215553	-0.261252	0.61422	-0.654825	-0.664545	0.824602
2	-0.594456	0.655638	0.434322	-0.384239	0.81275	0.718829	-0.461802
3	0.304503	-0.702341	-0.346799	-0.076285	-0.483025	-0.341816	-0.178079

	Percent_Bakery	Percent_Gen_Mer
1	0.428574	-0.674798
2	0.312775	-0.328927
3	-0.866459	1.135296

Analysis Report 1: Summary Report of the K-Means clustering solution (cluster by category sales)

Looking at the report, we can see that Meat, Produce, Floral, Deli, and General Merchandize are the variables that show the most contrast between clusters, denoting by a high positive value and a high negative value. For example, looking at the variable Percent_Produce, cluster 2 has 0.81275 while cluster 1 shows -0.654825. This suggests that store format 1 and store format 2 has opposite percentage of produce and store format 1 may potentially has the largest portion of Produce products.

Similarly, I found that Cluster 1 potentially has the highest percentages of meat and Deli products. Cluster 2 potentially has the highest percentages of Produce and Floral products. Cluster 3 potentially has the highest percentages of general merchandise products.

However, the above indicators are not conclusive to show which cluster has the most of which products. Therefore, I will plot the 'store_categorical_sales.yxdb' with product categorical percentages of total sales into scatterplots to validate.

Grace Radetsky 2020

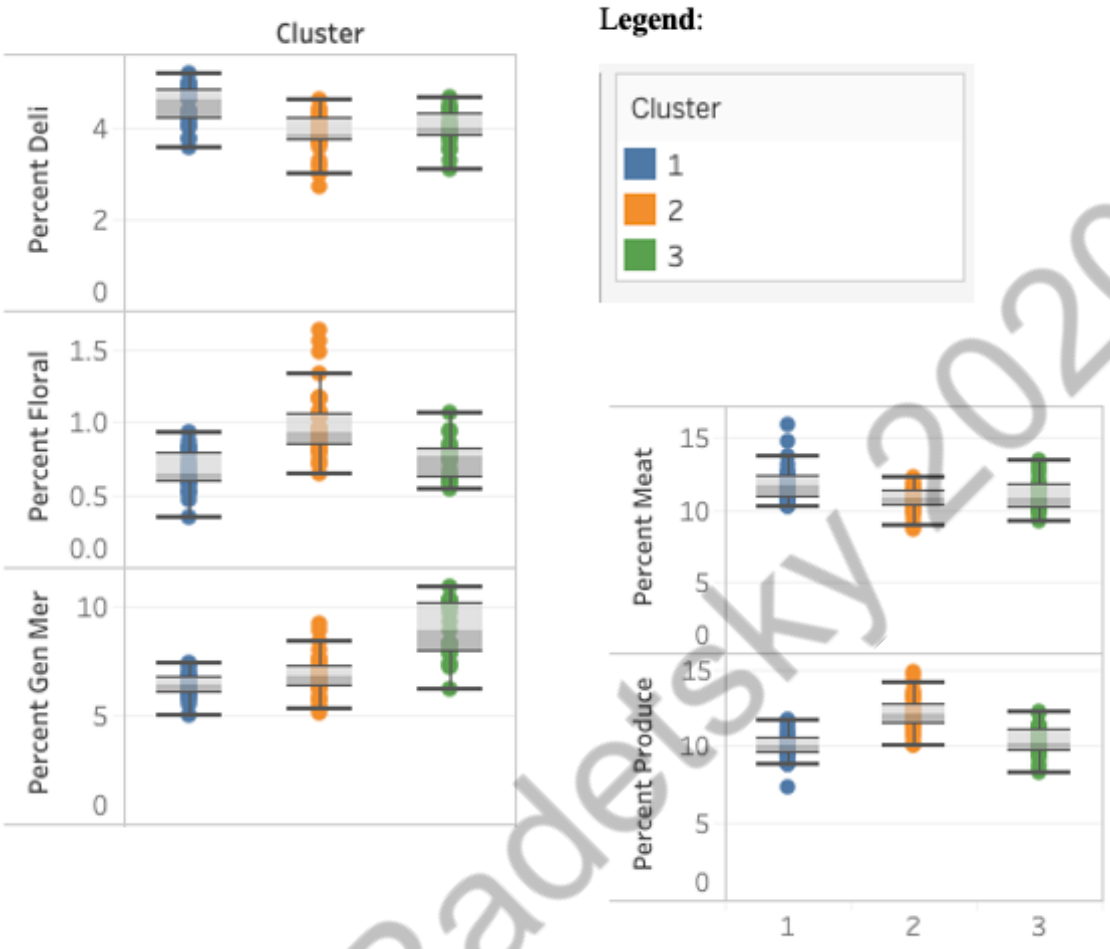


Figure 4 Scatterplots of 3 clusters by five categorical sales

The scatterplots in Figure 1 confirmed the above conclusion.

Grace Radetsky 2020

1.4. 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.

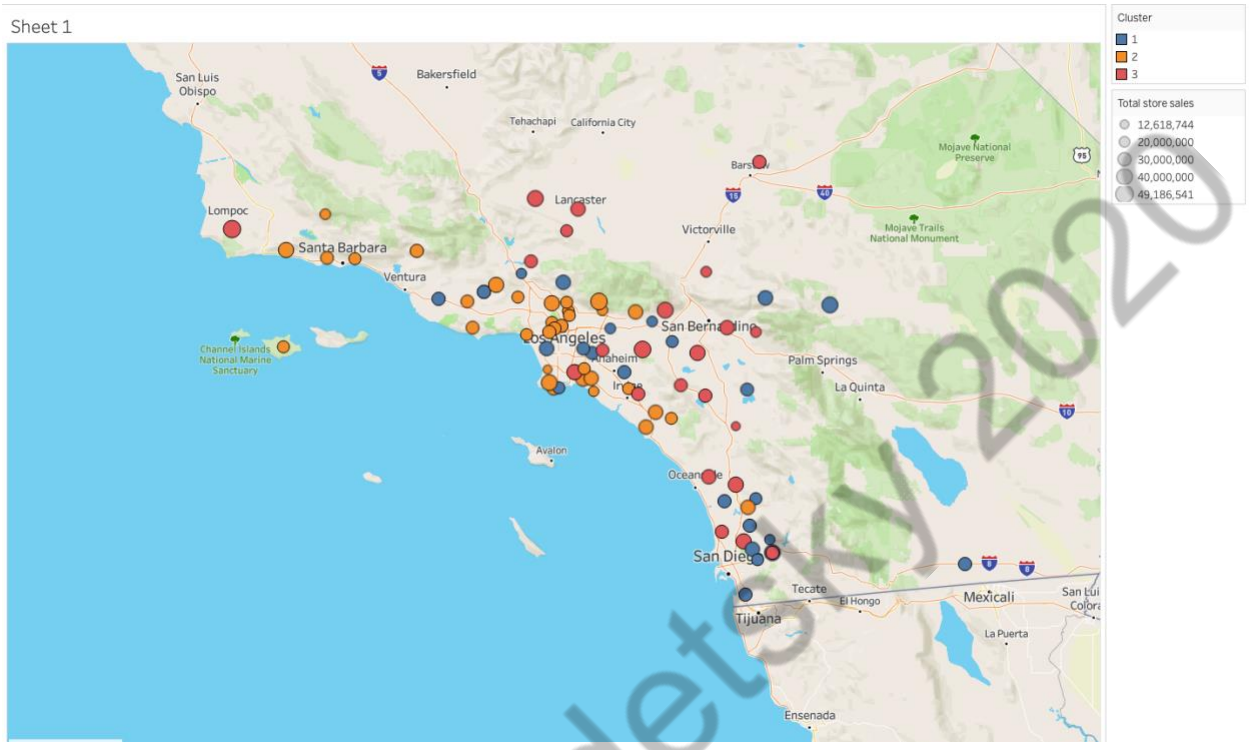


Figure 5: Clusters by location and size

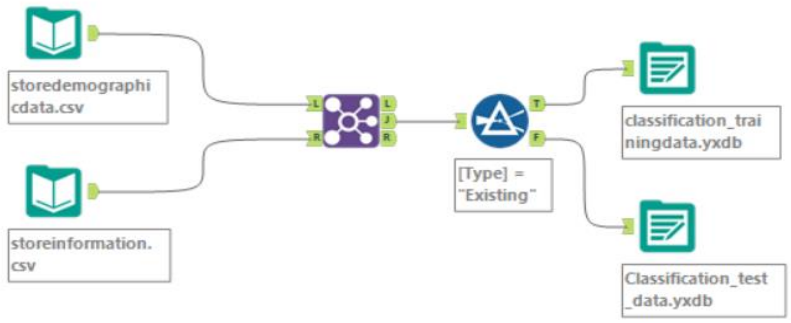
Task 2: Formats for New Stores

2.1. 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.)

Since we are trying to predict an outcome in which new stores are segmented into three formats, I will treat this as a predictive classification problem where Decision tree, Forest Mode, and Boosted Model are good candidates for. Since this is a non-binary prediction, I will not try a Logistic Regression Model. After comparing models, Boosted Model stood out as the best model.

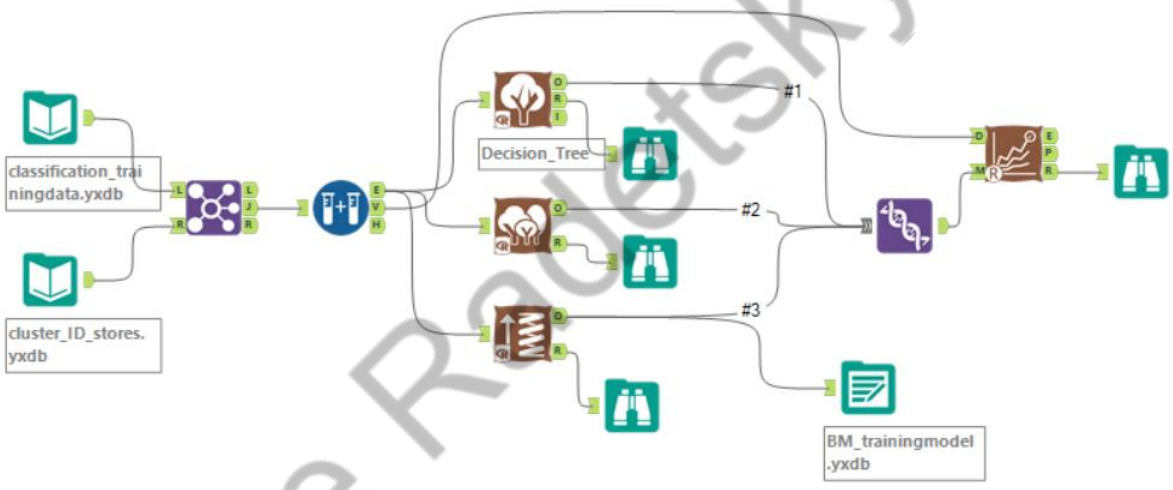
Step 1: Create two separate files containing demographic information about the existing stores and the new stores data set. The existing stores data set will be used as to train the model to make prediction about the new stores data set.

Grace Radetsky 2020



Workflow 3: Preparing training data set from existing stores and new stores data set for classification

Step 2: Configure three different predictive models with 20% data being withheld for validation, and Random Seed being 3. After that, I use a Model Comparison tool to determine the best model.



Workflow 4: Comparing classification models

The following report is the result of Workflow 4:

Grace Radetsky 2020

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.6471	0.6667	0.5000	1.0000	0.5000
Forest_Model	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted_Model	0.7647	0.8333	0.5000	1.0000	1.0000

Confusion matrix of Boosted_Model

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	0
Predicted_2	2	5	0
Predicted_3	2	0	4

Confusion matrix of Decision_Tree

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	2
Predicted_2	3	5	0
Predicted_3	1	0	2

Confusion matrix of Forest_Model

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	2	5	0
Predicted_3	2	0	3

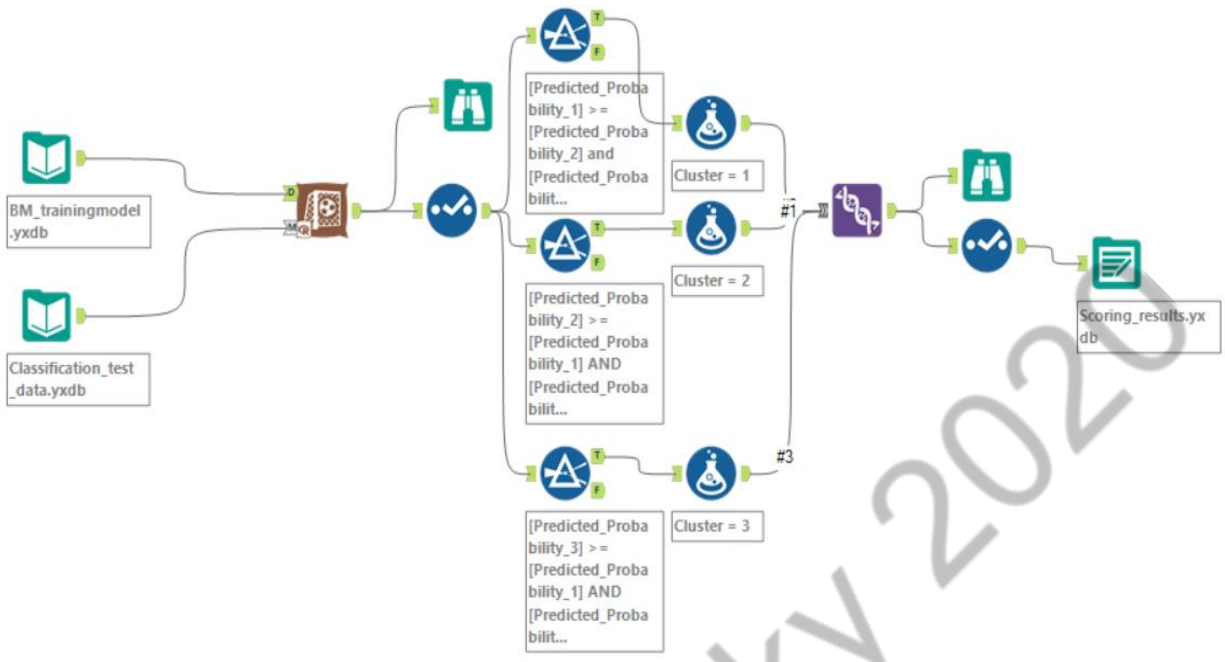
Analysis Report 2: Model comparison report (Decision Tree, Forest Model, and Boosted Model)

From the Analysis Report 2, it is clear that Boosted Model has the highest overall accuracy rate when being compared against the validation data set (at 76.47%). This model also has the highest F1 score (at 83.33%), which means this model has the best balance between precision (exactness) and recall (completeness). Therefore, I will choose the Boosted Model to predict store formats for new stores.

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

The workflow for scoring new stores is:

Grace Radetsky 2020



Workflow 5: Classifying new stores into three clusters

As the result of Workflow 5, the predicted formats for the new stores are:

Record	Store	Cluster
1	S0086	1
2	S0087	2
3	S0088	3
4	S0089	2
5	S0090	2
6	S0091	3
7	S0092	2
8	S0093	3
9	S0094	2
10	S0095	2

Table 3: Clustering results for 10 new stores

Grace Radetsky 2020

Task 3: Predicting Produce Sales

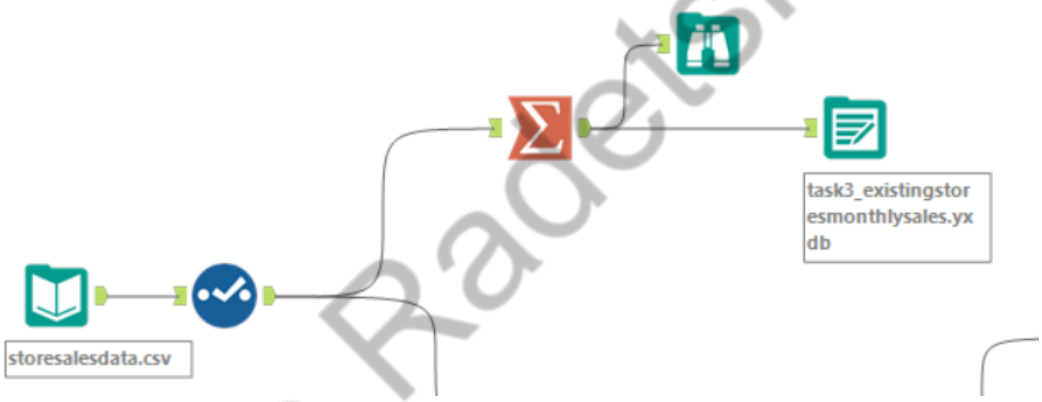
3.1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

I decided to use the ETS (m,n,m) model over the ARIMA(0,1,1)(0,1,1)[12] model for both forecasting existing and new stores as the ETS model has smaller error measurements such as MASE.

To arrive at that decision, I took the following steps:

For existing stores:

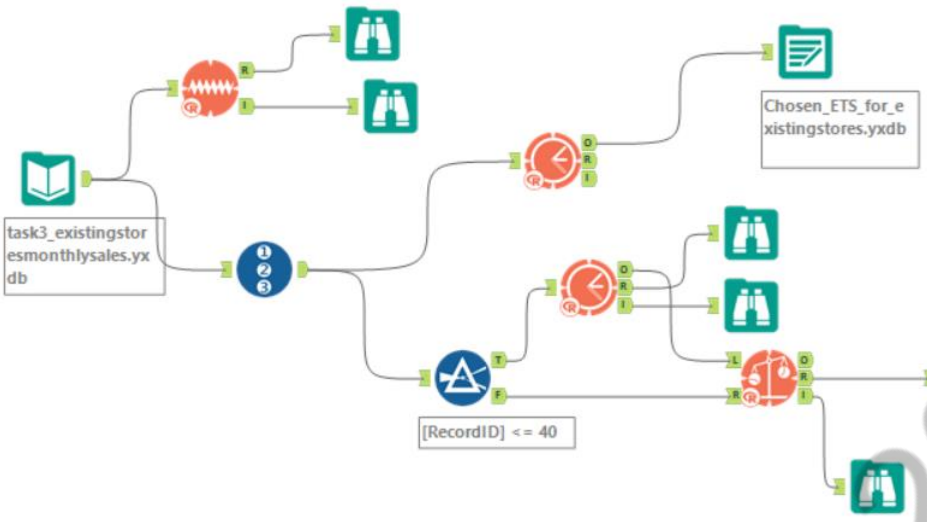
Step 1: I prepared the monthly sales of produce products for all existing stores. To do that, I use a Summarize tool to group the 'storesalesdata.csv' data by Year, then by Month, and Sum by Produce Products. The workflow for that is below:



Workflow 6: Aggregating monthly sales of produce products for all existing stores

Step 2a: Then I used the following workflow to investigate the time series, choosing the configuration for the ETS model.

Grace Radetsky 2020



Workflow 7: Exploring existing stores' time series, choosing and validating an ETS model

In Workflow 7, I used a TS plot tool to visualize the time series and its components.

Figure 6: Line graphs of existing stores' time series and components

In terms of the Error component (also known as remainder in the figure), the error varies without a specific pattern, there for the error will be applied Multiplicatively (denoted by m in the model). There is no clear trend over the span of the time series, the sales decreased significantly, hitting a low, then recover slowly. Therefore, I will not apply the trend component. The seasonal plot seems to show a clear seasonality pattern with slightly increasing magnitude every season.

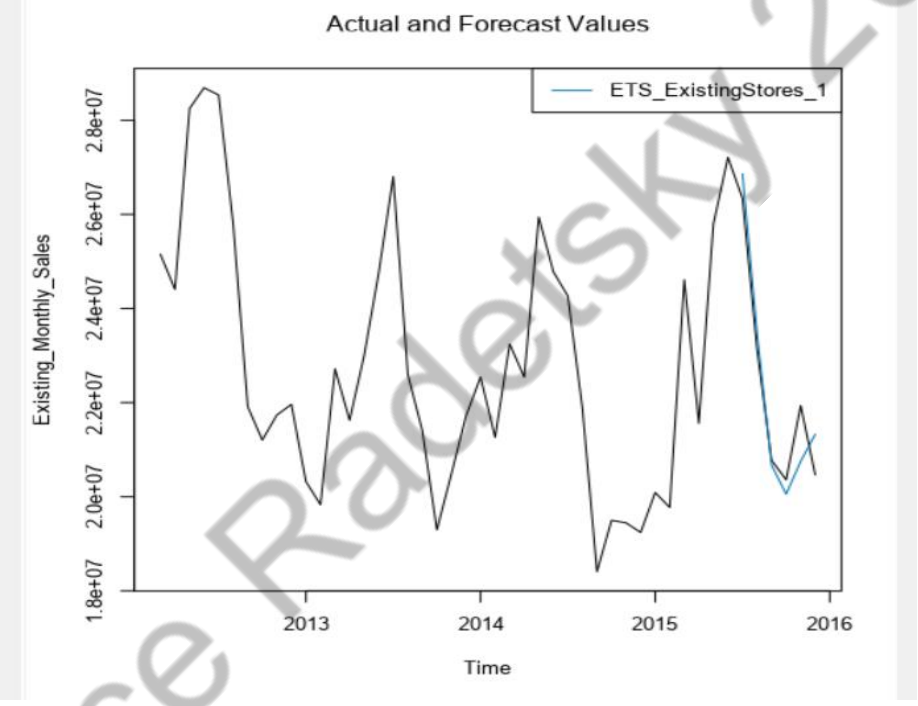
Grace Radetsky 2020

Since the dividing line between additive and multiplicative applications are very thin, I'd use Alteryx's auto setting. As the result, Alteryx chose to apply the Seasonality component multiplicatively. In a nutshell, the ETS candidate is ETS (m, n, m).

After configuring the ETS (m,n,m) model and withholding the final 6 months for validation, I got the following report, which I will use later on to compare against the ARIMA model.

Accuracy Measures:

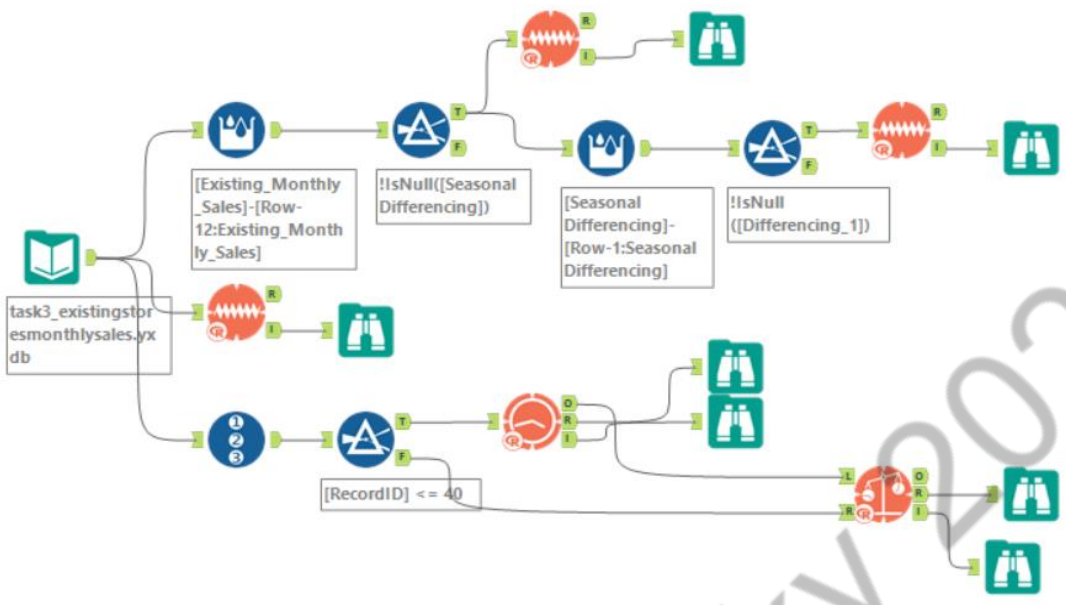
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_ExistingStores_1	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257



Analysis Report 3: Accuracy measures of ETS(m,n,m) model on existing stores data set

Step 2b: Investing the time series and choose the configuration for the ARIMA model. I use the following workflow:

Grace Radetsky 2020



Workflow 8: Exploring existing stores' time series, choosing and validating an ARIMA model

After using a TS Plot, it is clear that the time series is not yet stationary because there is still seasonality. Therefore, I will take the first seasonality differencing by subtracting values of a period with values of 12 months earlier.

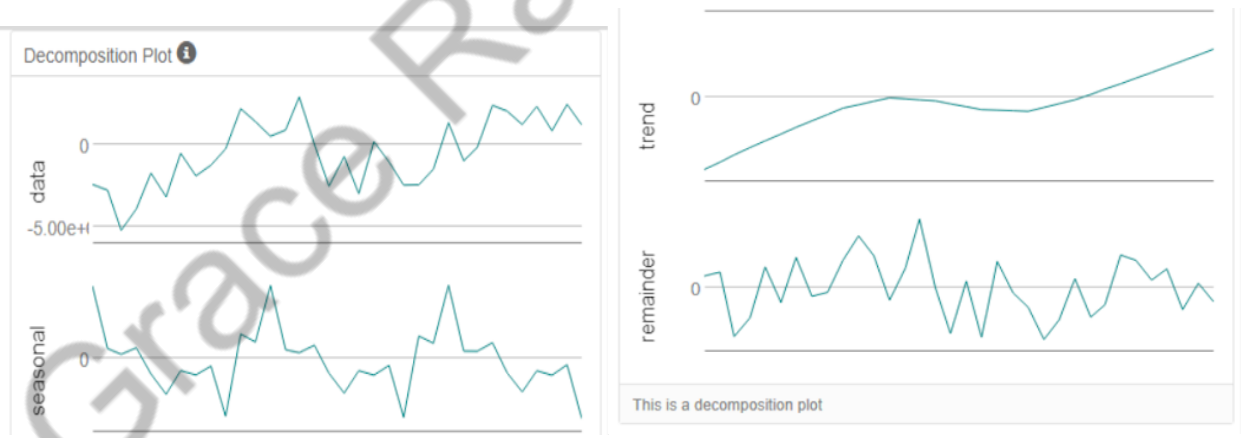


Figure 7: Visualization of existing stores' time series after a seasonal differencing

It seems that the time series of first seasional differencing values is still not stationary as the seasonality is still obvious and the mean of the time series is not constant. Therefore, I will take another non-seasonal differencing by subtracting values of a period with those of the prior period. As the result:

Grace Radetsky 2020

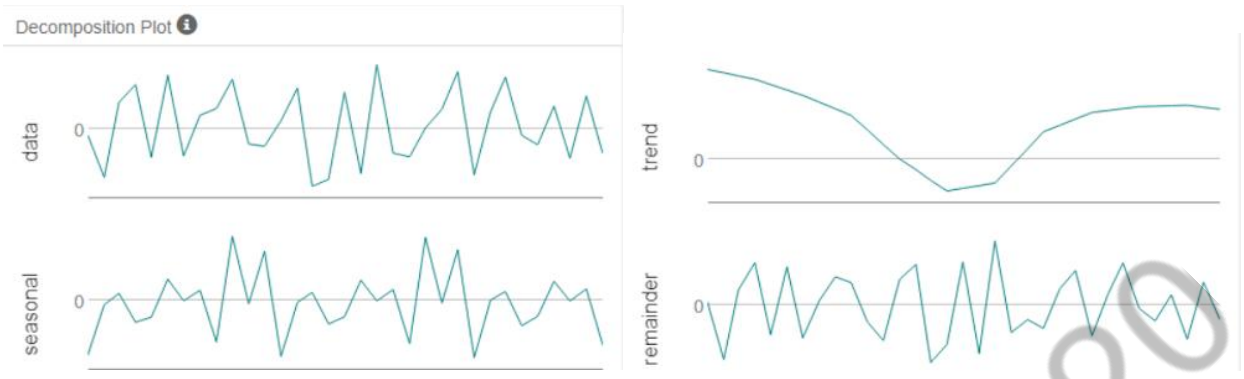


Figure 8: Visualization of existing stores' time series after a seasonal differencing and a non-seasonal differencing

From Figure 8, the data now is more stationary as the seasonality pattern is less obvious and the mean and standard deviation of the series are more constant. I could take another non-seasonal differencing to completely stationarize the time series but to simplify the process, I will stop here.

Figure 9 below shows the Autocorrection Function (ACF) plot and the Partial Autocorrection Plot (PACF) after a seasonal differencing and a non-seasonal differencing.

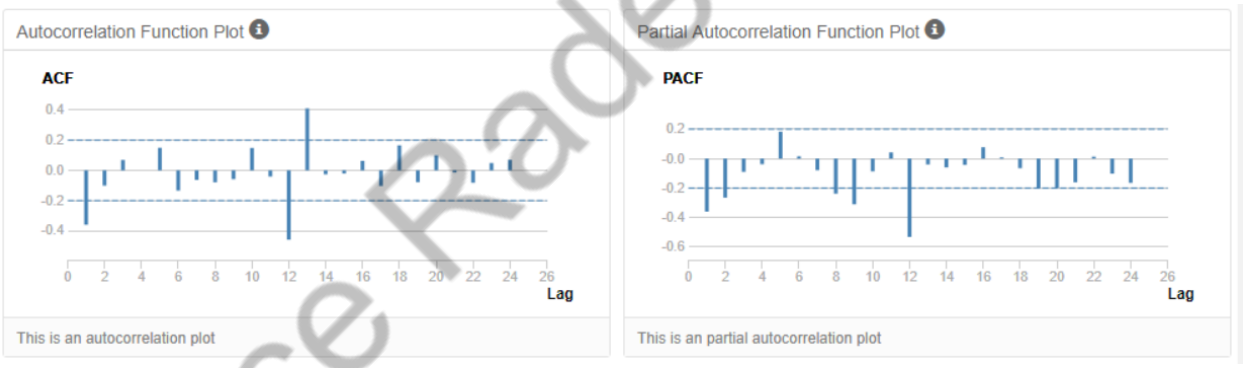


Figure 9: ACF and PACF for existing stores' time series after a seasonal differencing and a non-seasonal differencing

The ARIMA model will be denoted by the format ARIMA (p,d,q)(P,D,Q)[12]

From Figure 9, there are statistically significant spikes at lag 1 and lag 12 (the two seasonal lags), there for I will use Q = 1 to explain for the seasonal autocorrelation. Both the ACF and PACF show negative autocorrelations at lag 1 and then autocorrelations gradually cut off to zero. Therefore, I will use q = 1 to explain for the non-seasonal autocorrelation. Since I took one seasonal differencing and one non-seasonal differencing, d and D will be 1 and 1, respectively.

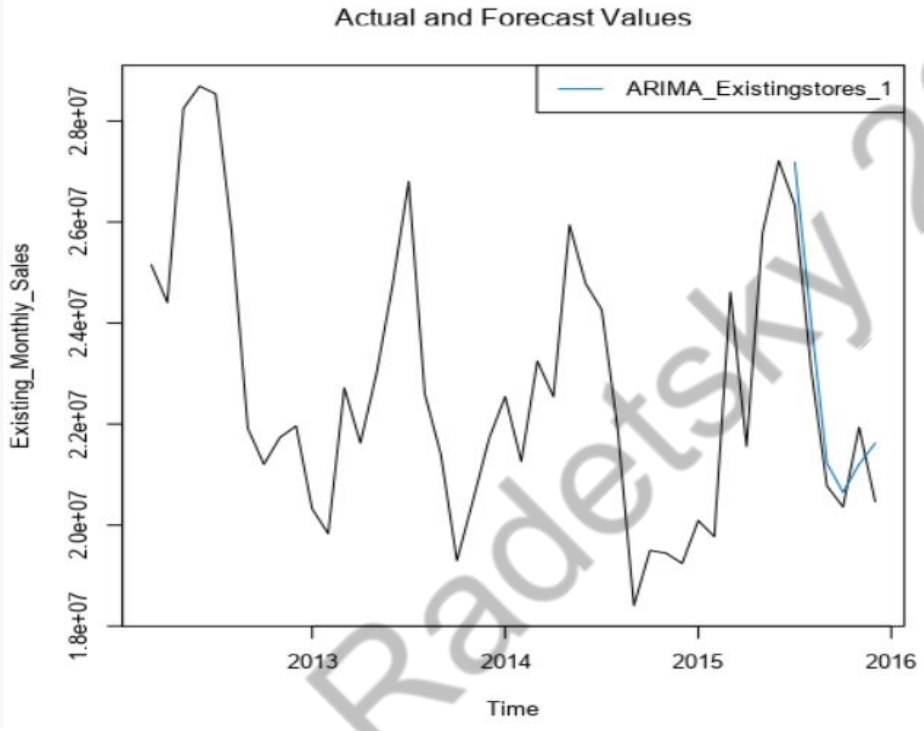
In short, the chosen ARIMA model is denoted by ARIMA(0,1,1)(0,1,1)[12]

Grace Radetsky 2020

After configuring the ARIMA(0,1,1)(0,1,1)[12] model and withholding the final 6 months for validation, I got the following report:

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_Existingstores_1	-492238.8	792197.3	735878.2	-2.1992	3.3098	0.433



Analysis Report 4: Accuracy measures of ARIMA(0,1,1)(0,1,1)[12] model on existing stores data set

Step 3: Comparing the ETS and the ARIMA model:

Based on Analysis Report 3 and Analysis Report 4 above, it appears that ETS has better accuracy of forecasts with all smaller error measurements. The ETS model has a smaller MASE value than the ARIMA (at 0.3257 and 0.433 respectively), which means the ETS model reduced more errors from a naive model compared to the ARIMA model.

Therefore, I will use the ETS (m,n,m) model to make prediction for the 2016 sales of existing stores.

For new stores:

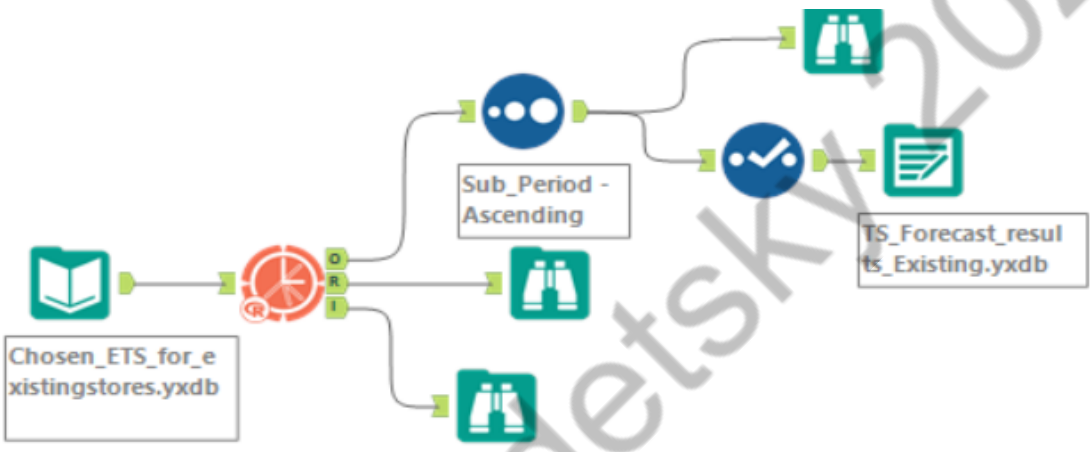
Grace Radetsky 2020

I will reiterate the time series exploration, model selection and validation process for new stores. And I will choose ETS(m,n,m) to make prediction for the 2016 sales of new stores.

3.2. 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.

For existing stores:

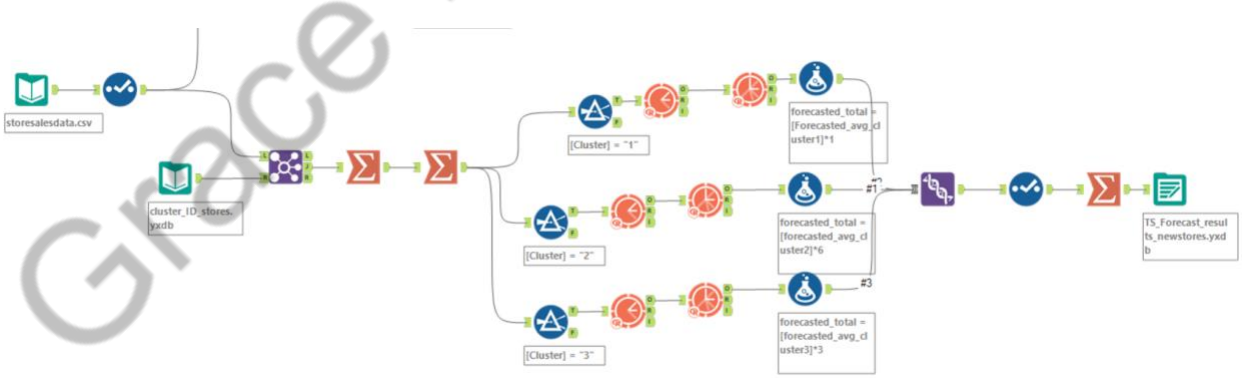
The following workflow shows the prediction for existing stores using the TS forecast tool:



Workflow 9: Predicting existing stores' 2016 sales with the ETS(m,n,m) model

For new stores:

The following workflows show the prediction for new stores using the TS forecast tool:



Workflow 10: Predicting new stores' 2016 sales with the ETS(m,n,m) model

As the results, the 2016 forecasted sales are:

Grace Radetsky 2020

Month	2016 forecast for existing stores	2016 forecast for new stores
1	21,829,060	2,563,358
2	21,146,330	2,483,925
3	23,735,687	2,910,944
4	22,409,515	2,764,882
5	25,621,829	3,141,306
6	26,307,858	3,195,054
7	26,705,093	3,212,391
8	23,440,761	2,852,386
9	20,640,047	2,521,697
10	20,086,270	2,466,751
11	20,858,120	2,557,745
12	21,255,190	2,530,511
Grand Total	274,035,761	33,200,949

Table 4: Predicted 2016 sales for existing and new stores

To visualize historical data, existing stores' 2016 forecasted sales, and new stores' 2016 forecasted sales.

Grace Radetsky 2020

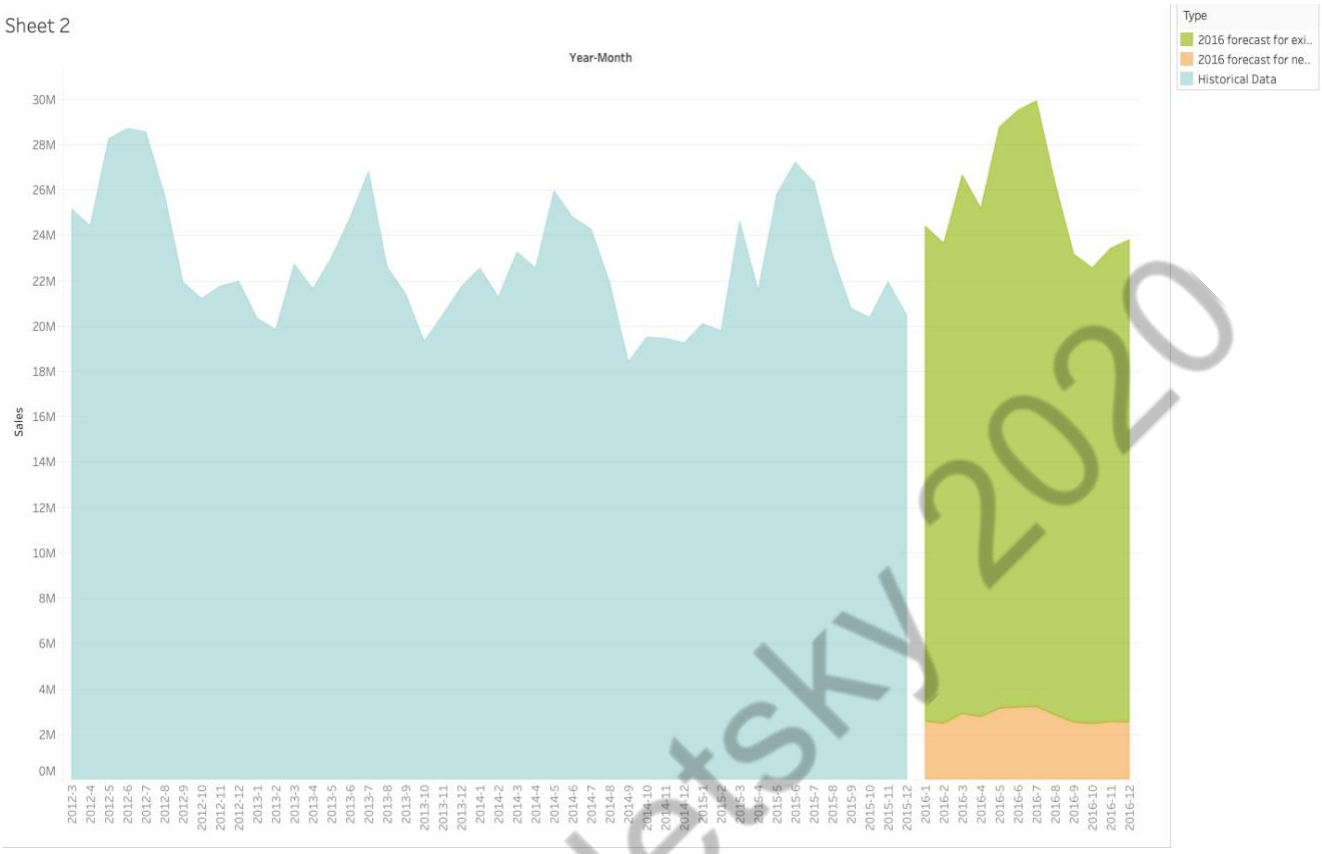


Figure 10: Historical data and 2016 forecasted sales

Grace Radetsky 2020