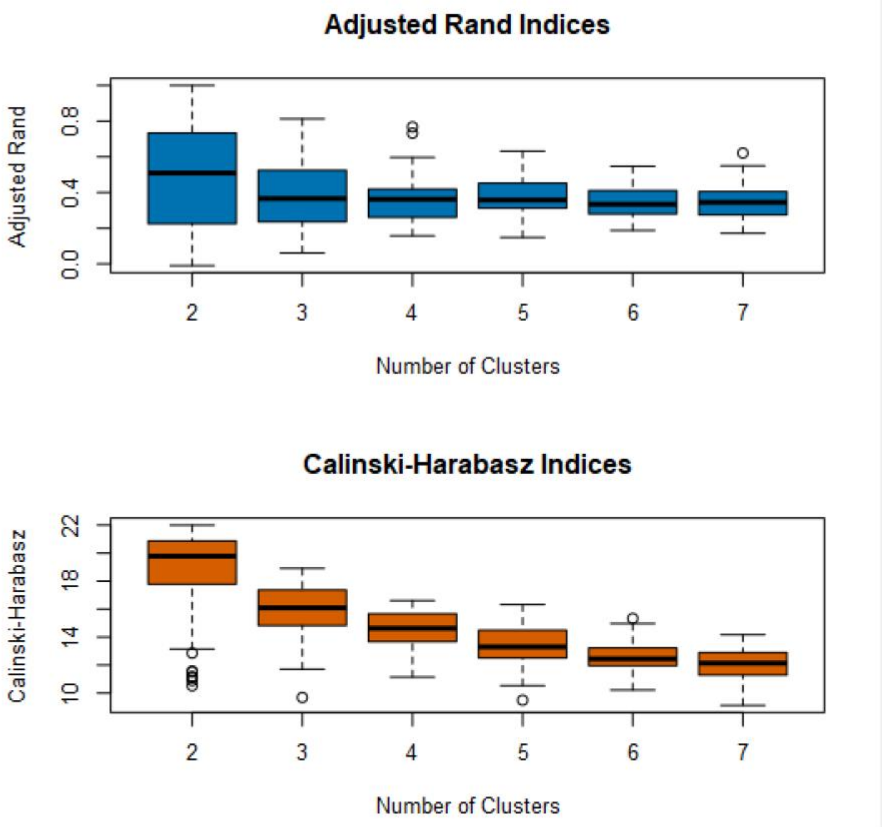
Predictive Analytics Capstone

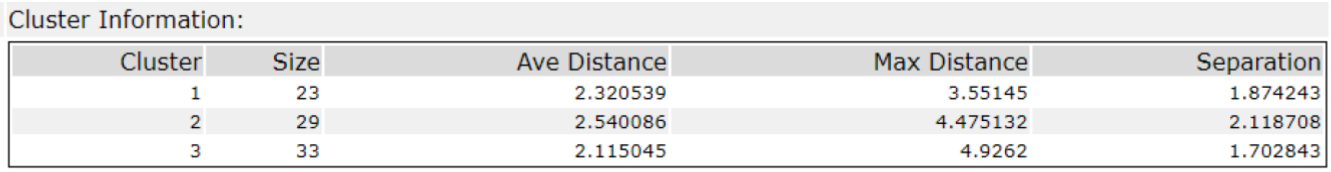
Determine Store Formats for Existing Stores

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

Using a K-centroids diagnostic tool, the Adjusted Rand and Calinski-Harabasz indices indicate that 3 is the optimal number of store formats, or clusters. While 2 clusters had the highest median values for both indices, the wider spread its the box-and-whiskers plots captured too much variance between IQR's. On the other hand, the AR and C-H indices for 3 clusters struck the optimal balance between higher median values for stability and distinctness of the clusters respectively, with relatively compact box-and-whisker plots to minimize variance.



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

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

% of Meat sales are greater than % Produce Sales in Store Formats 1 and 3 whereas % of Produce sales are greater than % of Meat sales in Store Format 2.

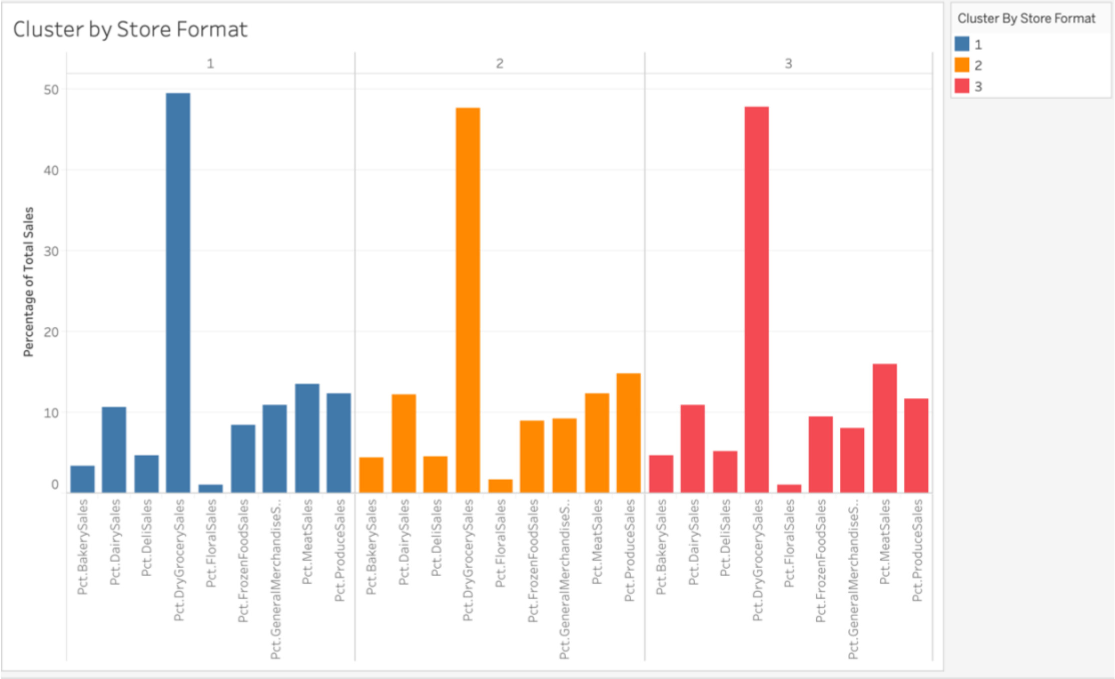
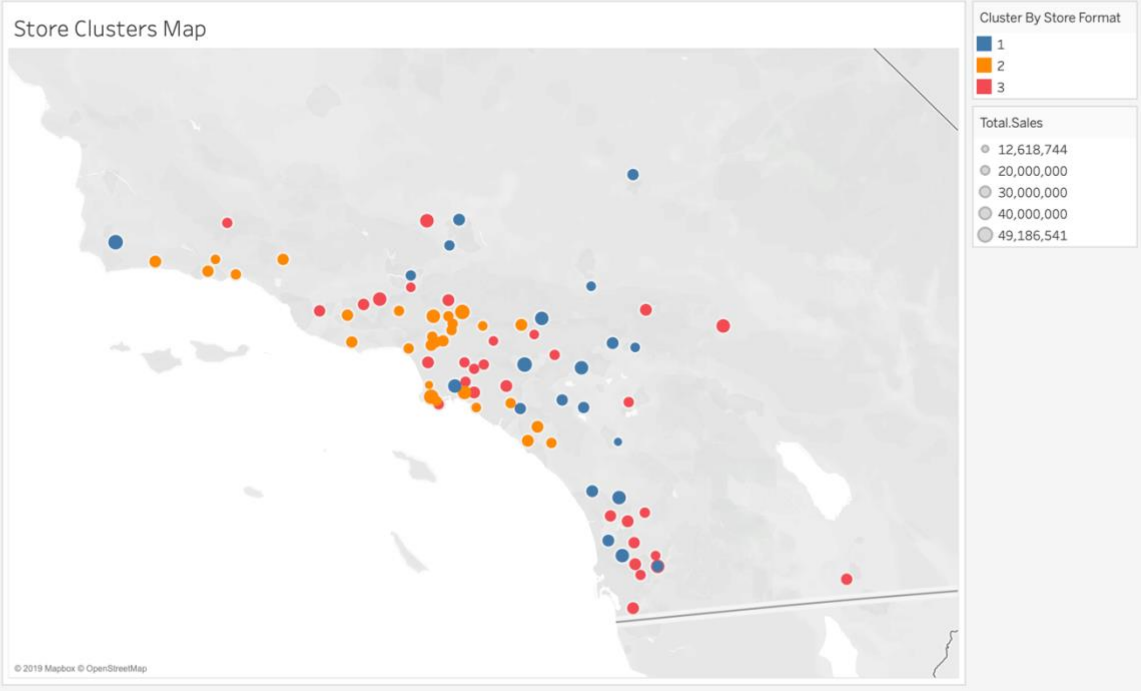


Tableau Visualization of Store Clusters by Format and Total Sales:

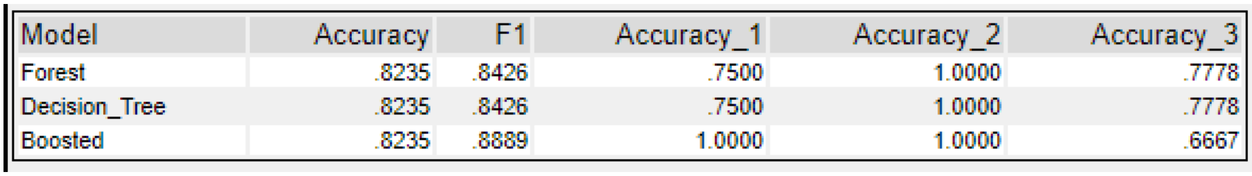


https://public.tableau.com/profile/andrew.chun7496#!/vizhome/StoreClustersMap/StoreClustersMap?publish=yes

Formats for New Stores

What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? Use a 20% validation sample with Random Seed = 3 to test differences in models.

Looking at the results of the Model Comparison tool, all of the models had the same percent accuracy of 82.35% for predicting the store format based on the demographic predictor variables. However, the Boosted Model had the highest F1 score at .8889 so I decided to use it over the other models. The Boosted Model identified stores in Cluster 1 and 2 with 100% accuracy, which provides much confidence in its ability to predict the best format for the new stores.



Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest	.8235	.8426	.7500	1.0000	.7778
Decision_Tree	.8235	.8426	.7500	1.0000	.7778
Boosted	.8235	.8889	1.0000	1.0000	.6667

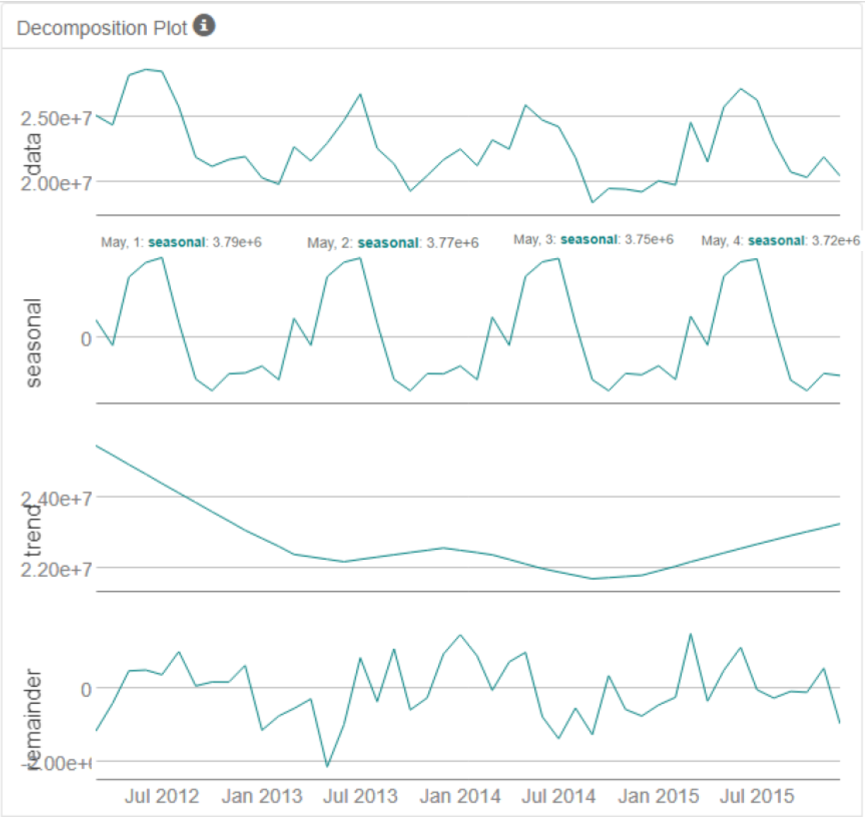
What format do each of the 10 new stores fall into?

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

Predicting Produce Sales

What type of ETS or ARIMA model did you use for each forecast? How did you come to that decision?

Using the TS Plot tool to generate a decomposition plot shown below shows the time series broken down into its three components: trend, seasonality, and error.

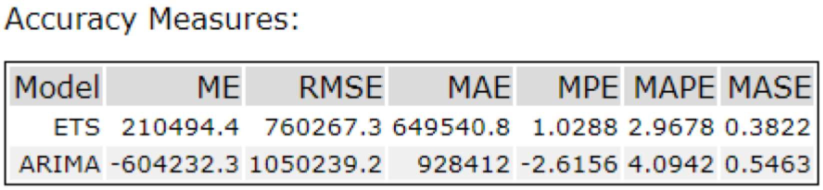


-The error plot appears to fluctuate inconsistently over time suggesting we should apply it multiplicatively.

-The trend line varies over time, initially decreasing, then increasing towards the end of the series, suggesting we should exclude trend in our ETS analysis.

-The seasonal component appears to be decreasing ever so slightly over time. Hovering a mouse over the interactive plot of the individual peaks confirmed this observation, suggesting we should apply it multiplicatively. Having seasonality suggests that any ARIMA models used for analysis will require seasonal differencing.

Shown below are the forecast error measurements:



Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
ARIMA	-604232.3	1050239.2	928412	-2.6156	4.0942	0.5463

When looking at the model's ability to predict the 6-month holdout sample, the **ETS(M,N,M)** model had better predictive qualities in every metric.

The lower values for **ME** and **RMSE** indicate that for the ETS model, the average difference as well as the percent difference between actual and forecasted values is smaller in magnitude, indicating a greater likelihood that the forecasted produce sales will predict the actual produce sales in the future.

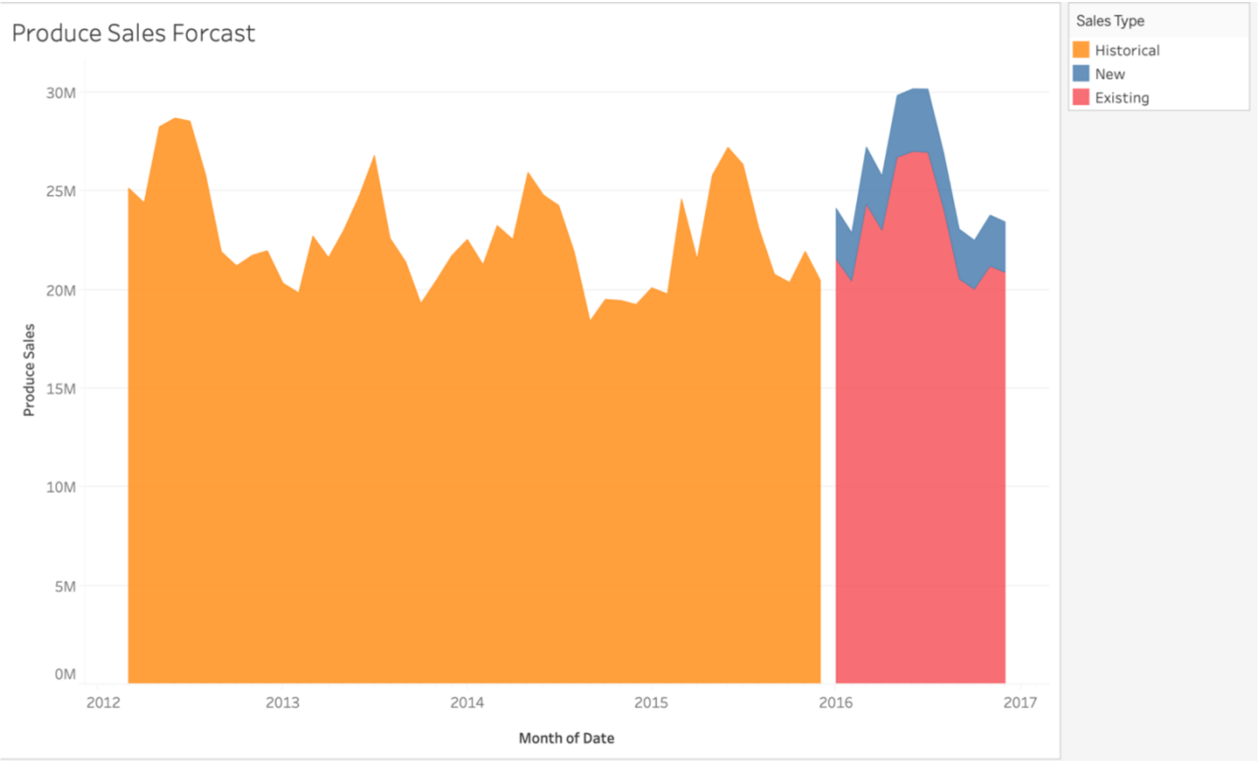
Furthermore, the negative **ME** and **MPE** values for the **ARIMA(1,0,0)(1,1,0)[12]** model suggest that the model is biased towards predicting lower future produce sales, providing a bearish outlook for the stores.

Thus, I chose to use the **ETS(M,N,M) model.**

Produce Sales Forecasts for New and Existing Stores

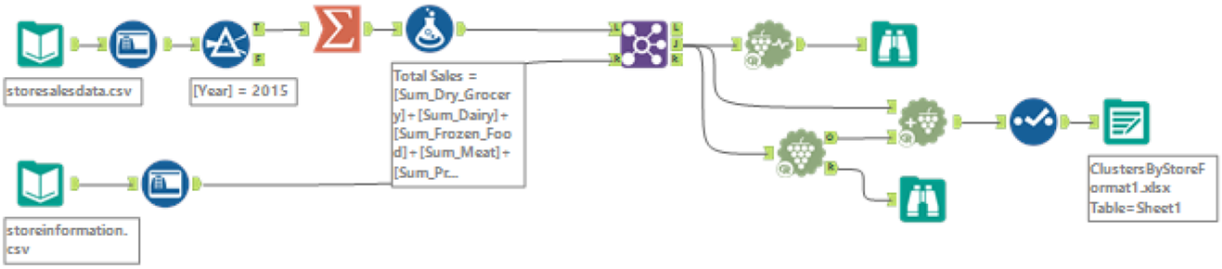
Month	New Stores	Existing Stores
Jan-16	2,584,383.53	21,539,936.01
Feb-16	2,470,873.92	20,413,770.60
Mar-16	2,906,307.87	24,325,953.10
Apr-16	2,771,532.13	22,993,466.35
May-16	3,145,848.57	26,691,951.42
Jun-16	3,183,909.28	26,989,964.01
Jul-16	3,213,977.72	26,948,630.76
Aug-16	2,858,247.21	24,091,579.35
Sep-16	2,538,173.64	20,523,492.41
Oct-16	2,483,550.17	20,011,748.67
Nov-16	2,593,089.19	21,177,435.49
Dec-16	2,570,200.44	20,855,799.11

Tableau Visualization of Produce Sales (Monthly):

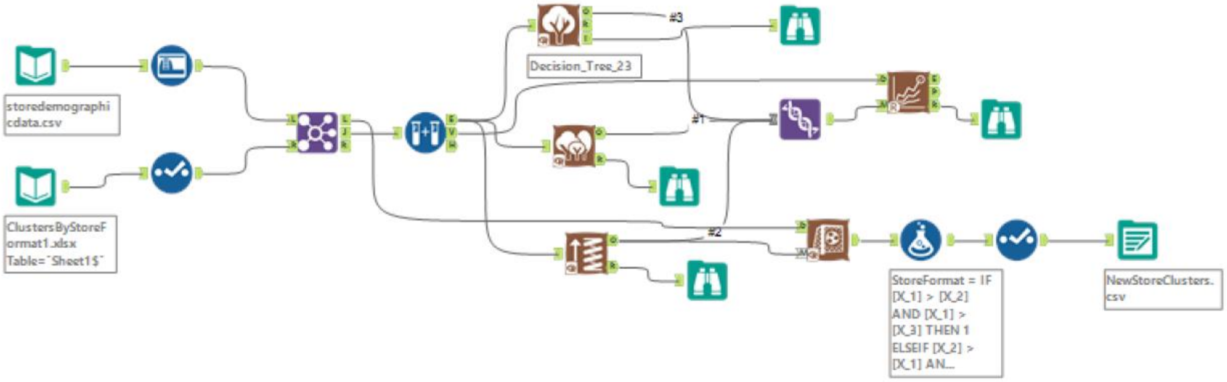


Alteryx Workflows:

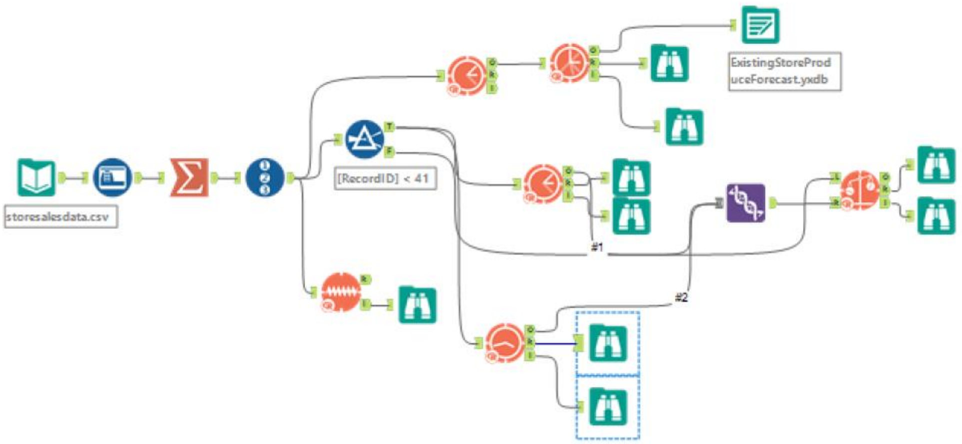
Store Clusters:

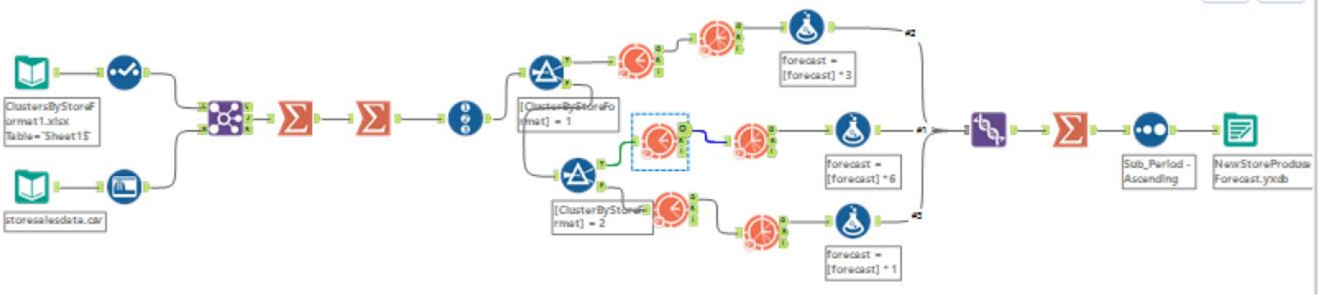

storesalesdata.csv
[Year] = 2015
Total Sales = [Sum_Dry_Grocery]+[Sum_Dairy]+[Sum_Frozen_Food]+[Sum_Meat]+[Sum_Pr...
storeinformation.csv
ClustersByStoreFormat1.xlsx Table=Sheet1

Model Classification:


storedemographicdata.csv
ClustersByStoreFormat1.xlsx Table=`Sheet1$`
Decision_Tree_23
#3
#1
#2
StoreFormat = IF [X_1] > [X_2] AND [X_1] > [X_3] THEN 1 ELSEIF [X_2] > [X_1] AN...
NewStoreClusters.csv

ETS/ARIMA models:




ClustersByStoreFormat1.xlsx Table=`Sheet1$`
storesalesdata.csv
[ClusterByStoreFormat] = 1
[ClusterByStoreFormat] = 2
forecast = [forecast] * 3
forecast = [forecast] * 6
forecast = [forecast] * 1
#2
#1
#3
Sub_Period - Ascending
NewStoreProduceForecast.yxdb