Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

Business Problem

A company currently has 85 grocery stores. Up until now, the company has treated all stores similarly, shipping the same amount of product to each store. This is beginning to cause problems as stores are suffering from product surpluses in some product categories and shortages in others. To remedy the product surplus and shortages, the company wants to introduce different store formats with no more than 40 individual stores per store format. The purpose of this report is to provide analytical support to make decisions about store formats and inventory planning. To do so, historical sales data is required for each store. All analytics will be conducted in in Alteryx Designer 2018.3 (Colorado, USA) and data visualisation will be conducted in Tableau Desktop 2018.3 (Washington, USA).

Methods

Historical sales data came pre categorised as “Dry Groceries” , “Dairy”, “Frozen Food”, “Meat”, “Produce”, “Floral” and “Deli”. After first calculating each category sales as a percentage of total sales and standardising the results to zero mean and unit variance, a K-means cluster analysis was conducted. Two to six cluster solutions were tested with 100 bootstrap replications. To account for the random starting seed, 10 starting seeds were explored and the best solution out of the set of solutions was retained and presented below.

To determine the optimal number of clusters, the adjusted rand index (ARI) and the Calinski-Harabasz Index (CH) were utilised. These metrics measure cluster stability and the compactness and distinctness of the clusters respectively.

Results

The results for this analysis suggest that the optimal number of clusters is 2 or 3 clusters (fig 1). While a 2 cluster solution has higher median values for both the CH and ARI metrics, it has a large spread of values suggesting an unstable result. Furthermore, a 2 cluster solution would result in more than 40 stores in a single cluster which is contrary to management's request. As such 3 cluster was therefore chosen as the optimal number of store formats.

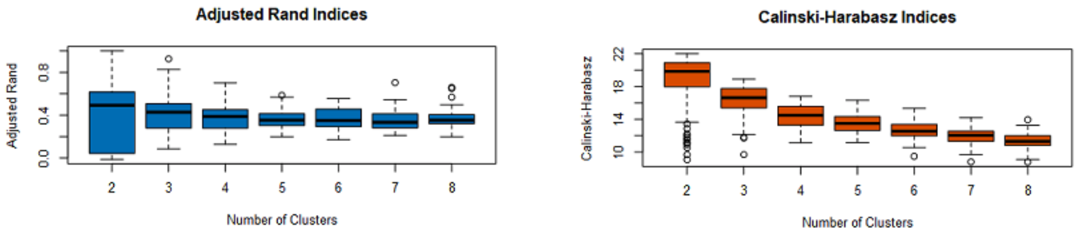
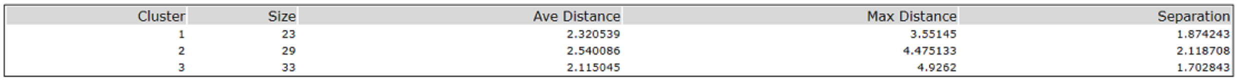


Fig (1) Adjusted Rand Index and Calinski-Harabasz Index for different cluster solutions.

Using a three cluster solution, cluster 1 had 23 stores, cluster 2 has 29 stores, while cluster 3 has 33 stores (table 1).

Table (1) Cluster Members and Cluster distance metrics



Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475133	2.118708
3	33	2.115045	4.9262	1.702843

Cluster 1 appears to be characterised by a high percentage of sales coming from general merchandise. Cluster 2 is characterised by high produce and floral sales along with relatively low dry grocery sales. Cluster 3 is characterised by relatively high deli sales (fig 2).

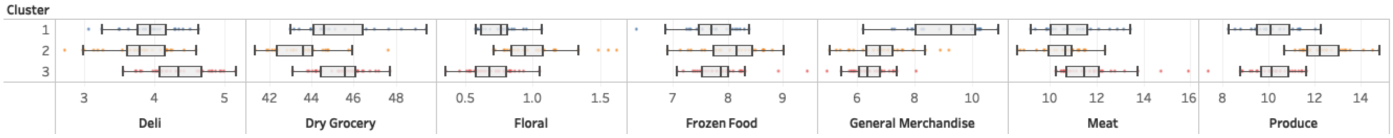
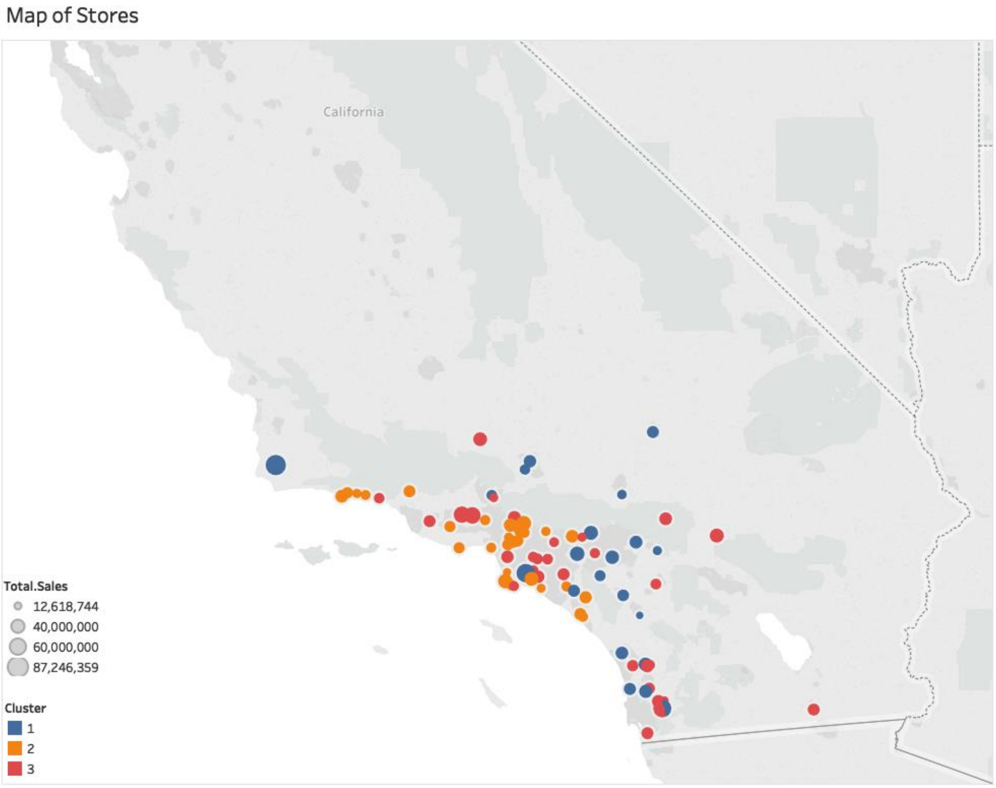


Fig (2) Cluster Characterisation

Store location along with cluster assignment (colour) and information on total sales (marker size) is presented in figure 3.



Fig(3) Tableau visualization of store location

Task 2: Formats for New Stores

Business Problem

The grocery store chain has 10 new stores opening up at the beginning of the year. The company wants to determine which store format each of the new stores should have. To do so, the demographic and socioeconomic characteristics of the population that resides in the area of both the established stores and new stores will be used to train a model to predict which of the store formats identified in task 1 should be assigned to the new stores.

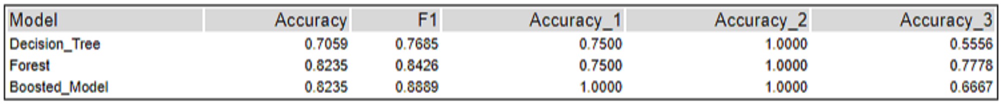
Methods

Data was randomly split into *Estimation (80%) and Validation (20%) samples and three models were trained (Decision Tree, Forest Model, Boosted Model).*

Results

Output from the model comparison tool indicate that both the forest and boosted model performed equally well in terms of overall accuracy. However when the F1 score is considered (which represents the trade-off between precision and recall), it appears that the boosted model performs better (Table 2). When exploring the variable importance plot for the boosted model the features 'Age0to9', '*HVal750KPlus'* and '*EdHSGrad'* had the largest influence on the model (fig 4).

Table (2) Summary of Model Performance



Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Forest	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667

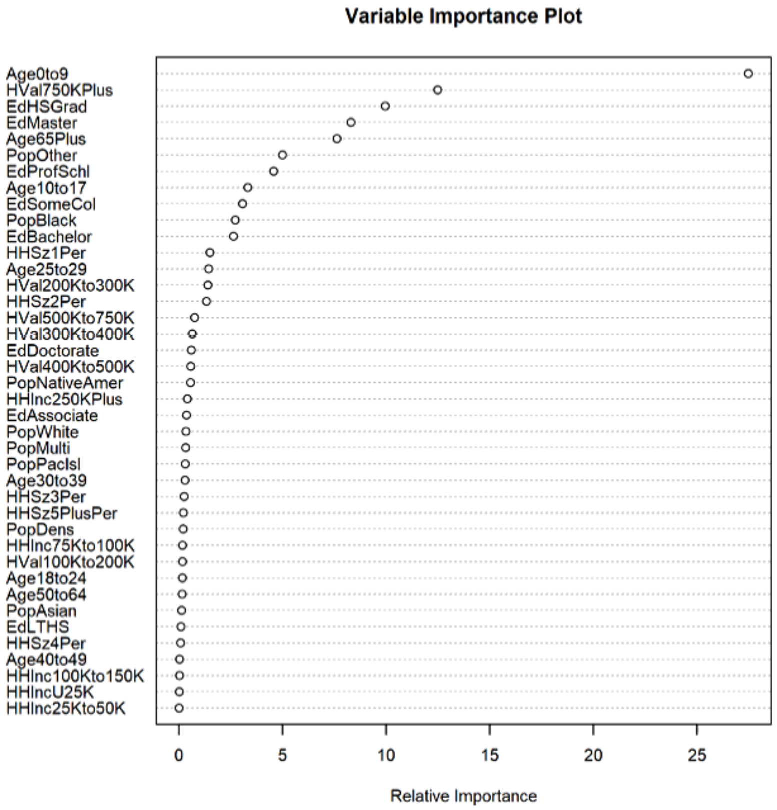


Fig (4) Variable importance of the boosted model

Utilizing the boosted model, the new stores were classified into the three store segments (table 3).

Table (3). Store Classification

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

The company has requested a monthly forecast for produce sales for the full year of 2016 for both existing and new stores. To do so, historical monthly produce sales is required for existing stores. As the supplied historic data contains monthly sales measured over a continuous time period with sequential and equal time intervals and only one data point per time unit, the historic data is sufficient to carry out a time series analysis. To assess the performance of our time series analysis, a comparison of models will be conducted on a holdout sample. This holdout sample will contain the most recent 6 months of the historical sales data. All analysis will be conducted in Alteryx 2018.3 (Colorado, USA).

Methods: Identifying the appropriate time series model.

A decomposition plot of the time series demonstrates a seasonal component of the data that decreases slightly over time. The fluctuating trend line indicates that the time series is non stationary with varying sales over time. Finally, the remainder component in the time series decomposition plot demonstrates an error that varies in magnitude (fig 5).

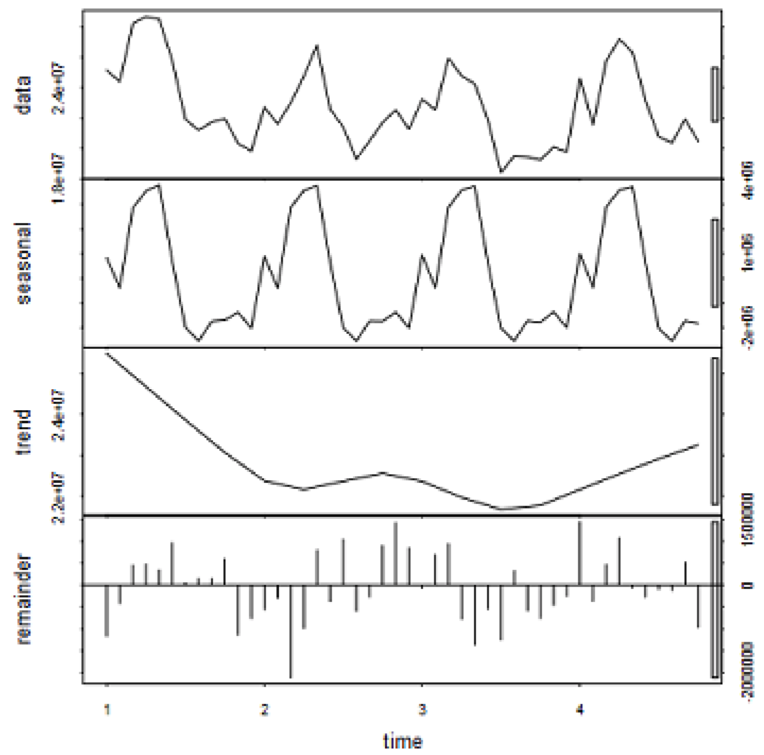


Figure (5) Time Series Decomposition plot.

Both an ETS$_{(M,N,A)}$ and ETS$_{(M,N,M)}$ model was utilized. The multiplicative error term was chosen due to the non-constant variance over time. No trend term was utilised due to the lack of a clear trend. Finally, as the seasonality component changed slightly over time, a model with additive and multiplicative terms were examined.

For the ARIMA model, the data was seasonally differenced to remove the effects of seasonality. Examination of autocorrelation function plot (ACF) depicted similar autocorrelation patterns as the original time series (fig 6).

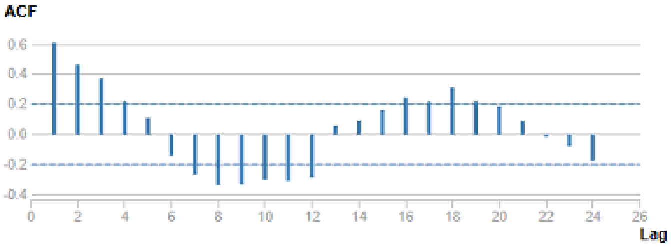


Fig (6) Autocorrelation function plot for the seasonally differenced time series

In order to make the times series stationary, the first and second seasonal difference were calculated. Examination of the first seasonal difference depicts a time series with approximately consistent mean and variance (fig 7). Furthermore, the ACF plot depicts that most of the autocorrelation has been removed from the data (fig 8). In addition the autocorrelations have no pattern and the lag-1 autocorrelation is negative suggesting that the series does not need a higher order of differencing.[1] To confirm, the second seasonal difference was examined which demonstrated an ACF value at lag-1 that was more than -0.5 negative, suggesting that the data is over differenced at the second seasonal difference.

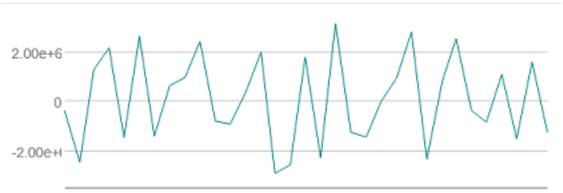


Fig (7) First seasonal difference time series

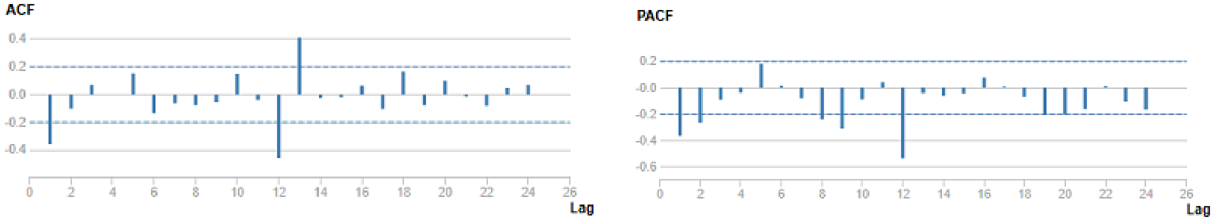


Fig (8) Autocorrelation function and Partial auto correlation function plots for the first seasonally differenced time series.

Taking the first seasonal difference data, the ACF and “Partial auto correlation function” (PACF) plots illustrate a negative correlation at 12 months so a seasonal MA term is required. The negative lag-1 autocorrelation in both the ACF and PACF plots indicate a $MA_{(1)}$ model term. The resultant ARIMA model can be described as $ARIMA_{(0,1,1)(0,1,1)[12].}$

When examining the out of sample errors for the three models, it can be seen that the $ETS_{(M,N,M)}$ outperforms the other models (table 4). As such, this model will be used to forecast future sales.

Table (4) Out of sample errors for the three models

Model	ME	RMSE	MAE	MPE	MAPE	MASE
$ETS_{(M,N,M)}$	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
$ETS_{(M,N,A)}$	-947831.4	1103108.1	970225.1	-4.3472	4.4493	0.5709
ARIMA	-492238.8	792197.3	735878.2	-2.1992	3.3098	0.433

Methods: Sales forecast

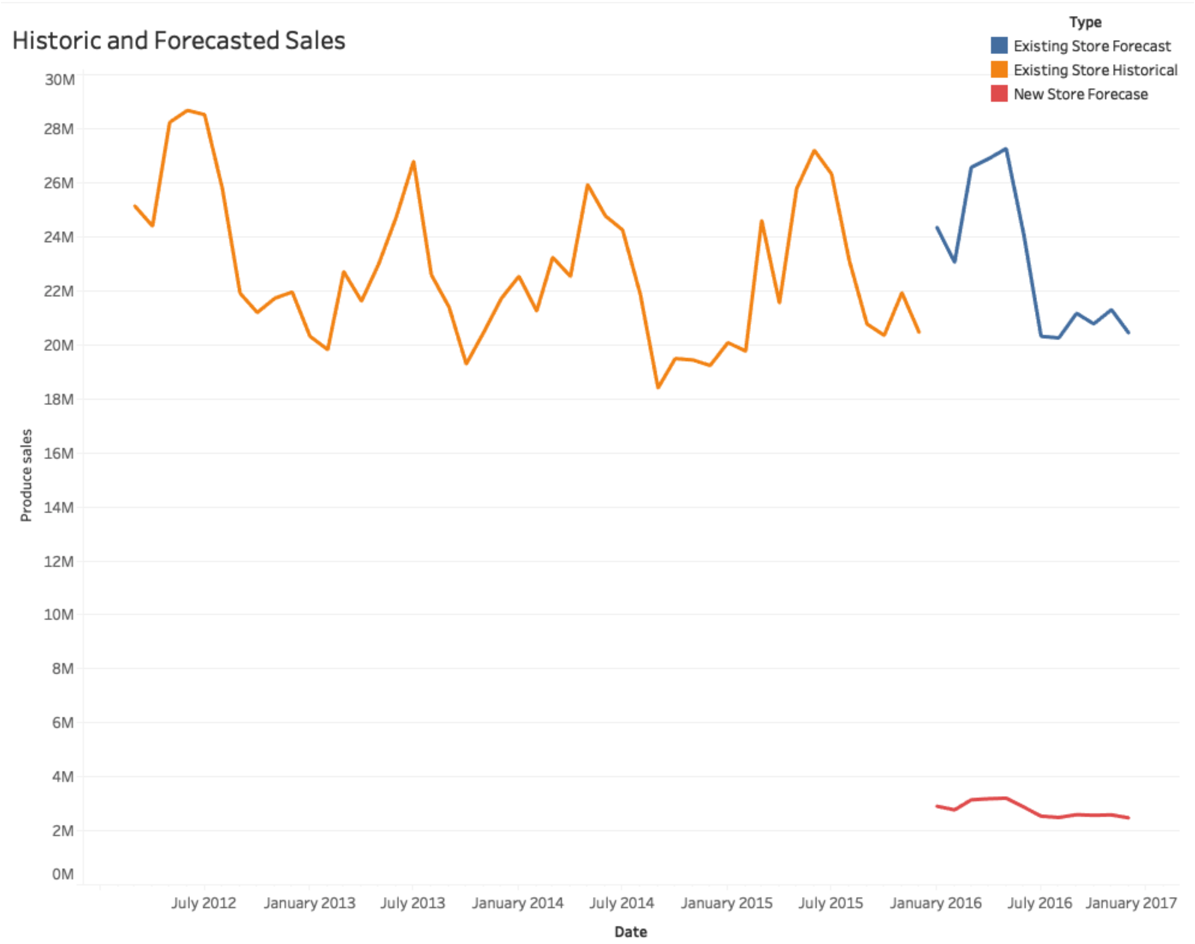
To forecast produce sales for existing stores, historic sales data was aggregated to a monthly level for each year. This data was then used in and $ETS_{(M,N,M)}$ model to forecast the following 12 month period. To forecast produce sales for new stores, the same historic data was utilised, however it was also aggregated to a cluster level so that average produce sales for each store type could be used to build the $ETS_{(M,N,M)}$ model. The forecasted average produce sales was then multiplied by the number of new stores in each cluster and summed to produce a total produce sales forecast for all new stores.

Results

Forecasted produce sales for both new and existing stores is presented in table 5 and visualised in fig 9 below.

Table (5) Forecasted Produce Sales for New and Existing Stores

Month	New Stores	Existing Stores
Jan-16	$ 2,908,915	$ 24,363,101
Feb-16	$ 2,774,783	$ 23,063,366
Mar-16	$ 3,147,545	$ 26,580,547
Apr-16	$ 3,186,404	$ 26,907,439
May-16	$ 3,208,297	$ 27,272,670
Jun-16	$ 2,876,155	$ 24,073,820
Jul-16	$ 2,541,834	$ 20,320,128
Aug-16	$ 2,493,126	$ 20,259,912
Sep-16	$ 2,596,560	$ 21,173,676
Oct-16	$ 2,573,948	$ 20,781,724
Nov-16	$ 2,589,908	$ 21,306,733
Dec-16	$ 2,477,721	$ 20,443,624



Fig(9). Historical and Forecasted Produce Sales

Conclusion

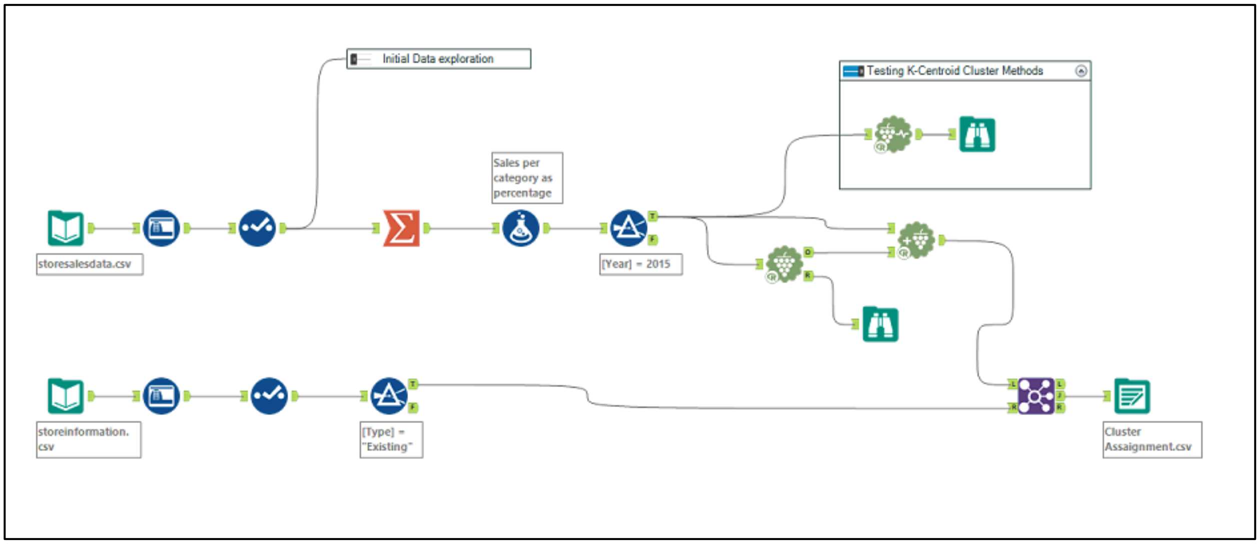
This report identified the presence of three store formats which will help remedy product surplus and shortages currently experienced across the chain by optimising the goods received by each store. Further, with an additional ten stores opening, this report was able to classify the new stores using demographic and socioeconomic data available for both the new and existing store areas. It is anticipated that this will facilitate the company planning which store format the new stores should receive. Finally, the company was particularly interested in forecasting produce sales for the next 12 months. Using the store formats, it was forecast that March, April and May will have the highest produce sales for both existing and new stores. This information can help management to make data driven strategic decisions. Full Alteryx Workflow is presented in appendix 1.

References

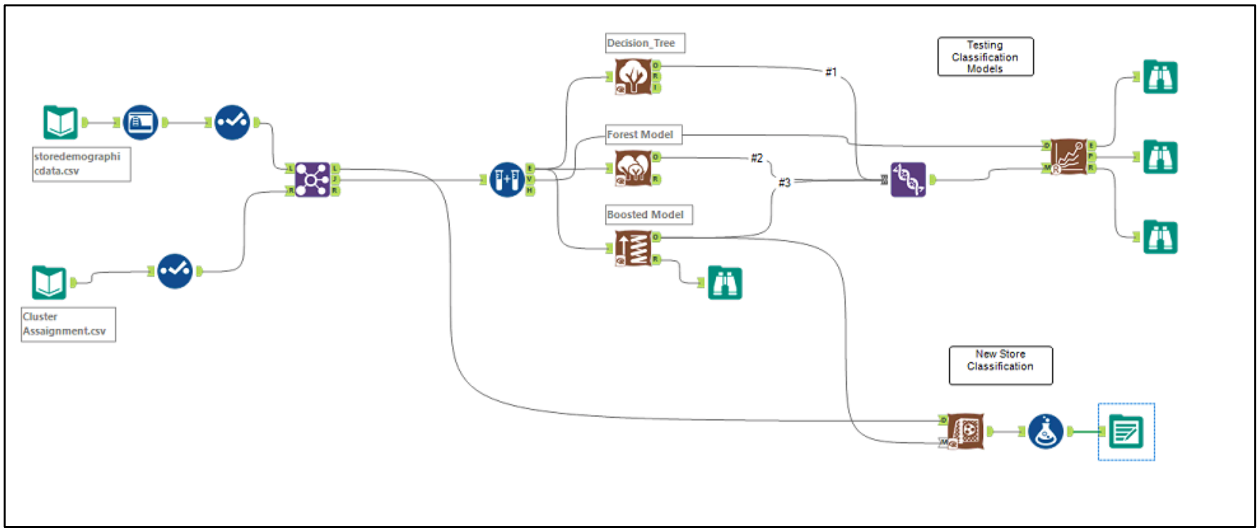
1. Hyndman R, Athanasopoulos G. Forecasting: principles and practice. 2018.

Appendix 1

Identifying existing store clusters (Task 1)

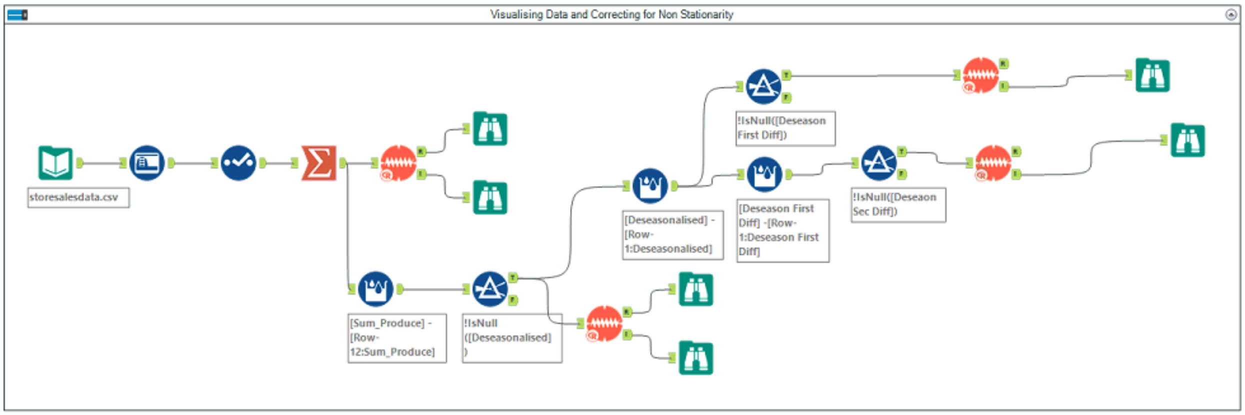


Classifying new stores (Task 2)

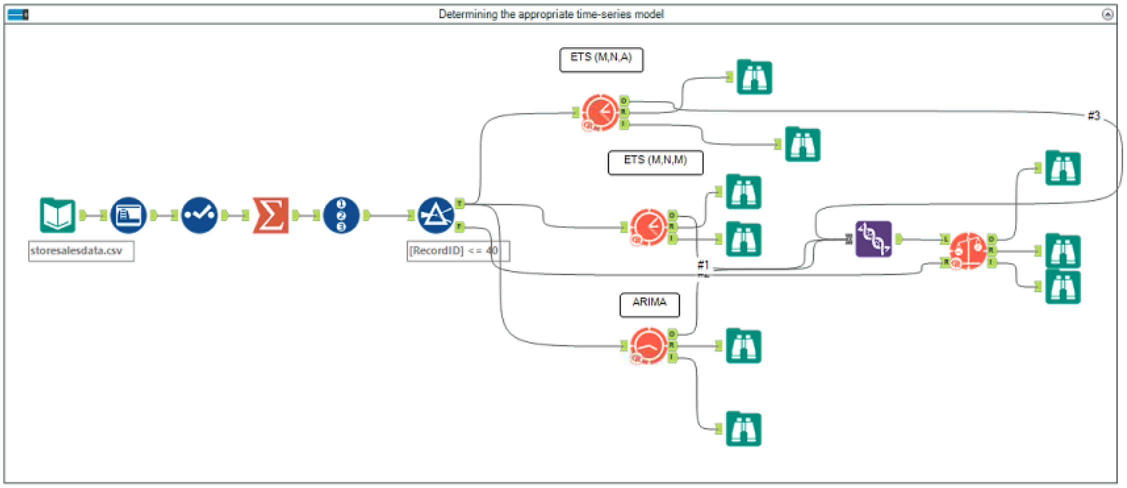


Forecasting Produce Sales (Task 3)

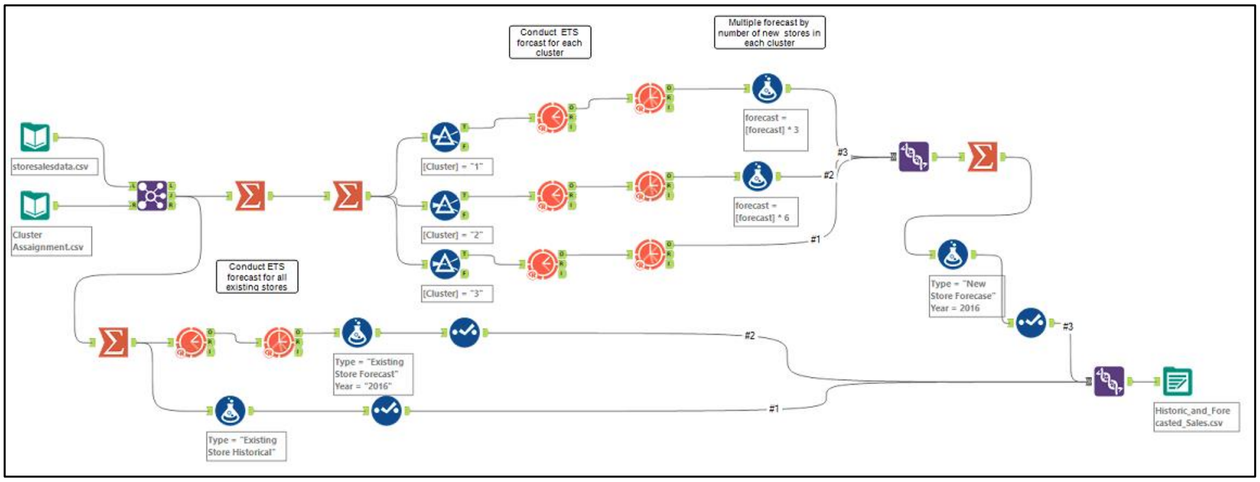
Part A: Visualising data and correcting for non-stationarity


Visualising Data and Correcting for Non Stationarity
storesalesdata.csv
!IsNull([Deseason First Diff])
[Deseasonalised] - [Row-1:Deseasonalised]
[Deseason First Diff] -[Row-1:Deseason First Diff]
!IsNull([Deseaon Sec Diff])
[Sum_Produce] - [Row-12:Sum_Produce]
!IsNull ([Deseasonalised])

Part B: Determining the appropriate model



Part C: Conduct forecast for new and existing stores


Conduct ETS forcast for each cluster
Multiple forecast by number of new stores in each cluster
storesalesdata.csv
Cluster Assignment.csv
[Cluster] = "1"
[Cluster] = "2"
[Cluster] = "3"
forecast = [forecast] * 3
forecast = [forecast] * 6
#3
#2
#1
Conduct ETS forecast for all existing stores
Type = "New Store Forecase" Year = 2016
#3
Type = "Existing Store Forecast" Year = "2016"
#2
#1
Type = "Existing Store Historical"
Historic_and_Forecasted_Sales.csv