Table of Contents

Project: Predictive Analytics Capstone

Business Issue

Task 1

Your company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products. 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. You've been asked to provide analytical support to make decisions about store formats and inventory planning. To remedy the product surplus and shortages, the company wants to introduce different store formats. Each store format will have a different product selection in order to better match local demand. The actual building sizes will not change, just the product selection and internal layouts. The terms "formats" and "segments" will be used interchangeably throughout this project.

Task 2

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. However, we don't have sales data for these new stores yet, so we'll have to determine the format using each of the new store's demographic data.

Task 3

Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast. You've been asked to prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores.

Task 1: Determine Store Formats for Existing Stores

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

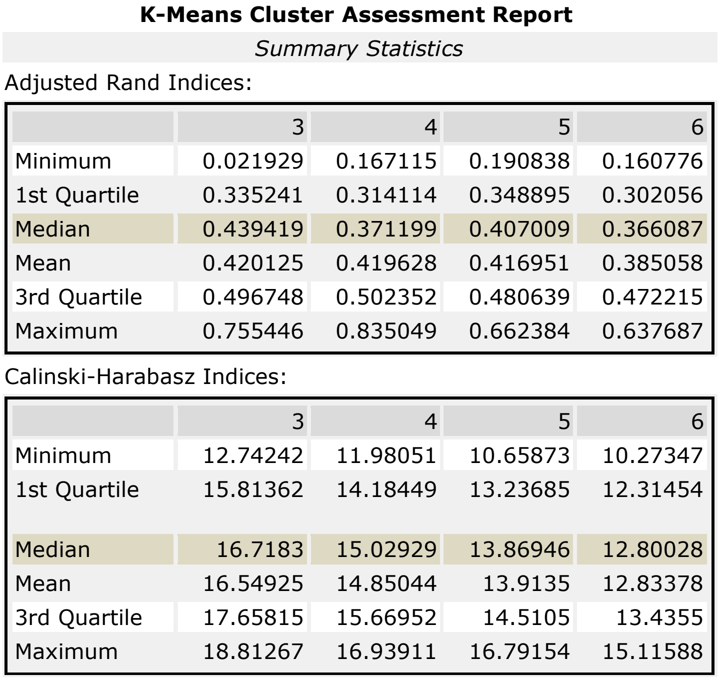
In order to determine the optimal number of store formats, we have been given two datasets: StoreSalesData.csv and StoreInformation.csv files. To start with, we can observe that the first file contains daily sales information for each product category relative to the 85 existing stores while the StoreInformation.cvs file includes additional details concerning the store location and type (existing or new) for the complete list of 95 point of sales.

Let's start our analysis by setting the right data type for each field. Additionally, we need to perform some data wrangling to structure the available information in a way that can be used for our analysis. We have been asked to use 2015 sales data, therefore, we can easily filter out these datapoints with the help of the Filter Tool. To compute the percentage of sales per category per store that we are going to use to identify the best number of clusters, we can bring in the Formula and the Summarize Tools. More specifically, we should:

- Use the Formula Tool to compute the Total_Sales by summing the sales values of all product categories.
- Use the Summarize Tool to group sales data by Store, Year and to sum the Total_Sales as well as to sum the sales values within each product category. All this sales information will be nicely displayed per Store and Year 2015. Data now start to take shape and we end up with 85 records, matching the number of existing stores.

With these aggregations, we can compute the percentage of sales per category by dividing the sum of sales per product category by the Total_Sales value that incorporates all sales categories. By multiplying the end results by 100, we get the percentage value of sales per category, grouped by store. Let's proceed by blending the datasets together through a Join. This will be particularly useful to display our final results in Tableau.

We are now ready to determine the optimal number of clusters with the K- Centroids Diagnostic Tool which outputs the following Summary Statistics report and plots:



K-Means Cluster Assessment Report

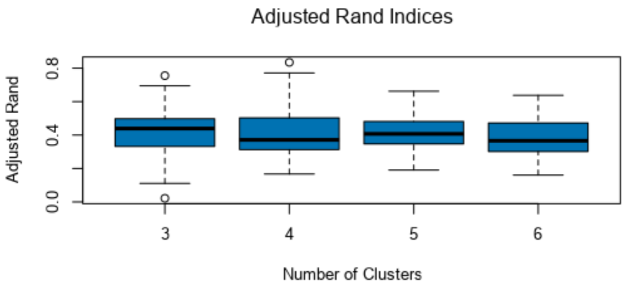
Summary Statistics

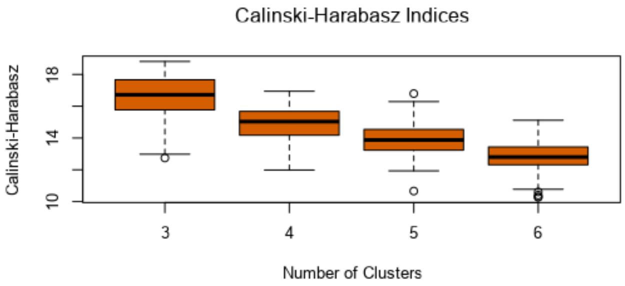
Adjusted Rand Indices:

	3	4	5	6
Minimum	0.021929	0.167115	0.190838	0.160776
1st Quartile	0.335241	0.314114	0.348895	0.302056
Median	0.439419	0.371199	0.407009	0.366087
Mean	0.420125	0.419628	0.416951	0.385058
3rd Quartile	0.496748	0.502352	0.480639	0.472215
Maximum	0.755446	0.835049	0.662384	0.637687

Calinski-Harabasz Indices:

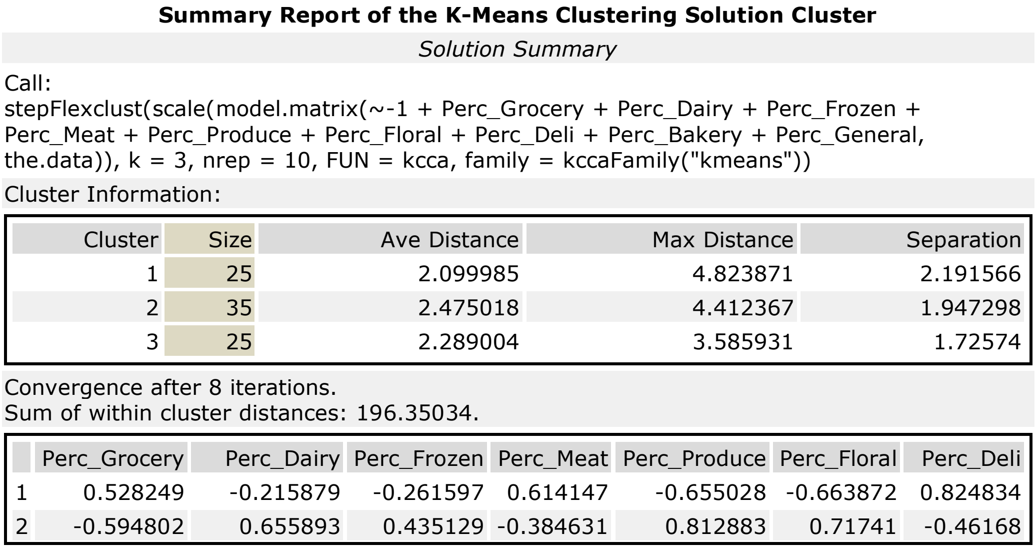
	3	4	5	6
Minimum	12.74242	11.98051	10.65873	10.27347
1st Quartile	15.81362	14.18449	13.23685	12.31454
Median	16.7183	15.02929	13.86946	12.80028
Mean	16.54925	14.85044	13.9135	12.83378
3rd Quartile	17.65815	15.66952	14.5105	13.4355
Maximum	18.81267	16.93911	16.79154	15.11588





The AR and CH indices plots seem to suggest that the highest median is for 3 clusters. This observation is confirmed by the Summary Statistics report where the Adjusted Rand Index and the Calinski-Harabasz Index display the highest median values for a number of 3 clusters. Moreover, the spread is also minimized for 3 clusters. Both the compactness and high median values for this option point out that 3 is the optimal number of clusters.

2. *How many stores fall into each store format?*
Now that we know that the optimal number of clusters is 3, we can perform the clustering analysis with the K-Centroids Cluster Analysis Tool to gauge the size of each segment and to consequently assign each existing store to a specific store format. To perform the analysis, we will be using the K-Means Clustering Method which outputs the following results:



Summary Report of the K-Means Clustering Solution Cluster

Solution Summary

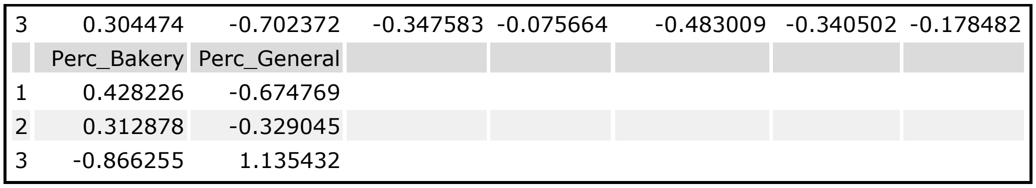
```
Call:
stepFlexclust(scale(model.matrix(~-1 + Perc_Grocery + Perc_Dairy + Perc_Frozen +
Perc_Meat + Perc_Produce + Perc_Floral + Perc_Deli + Perc_Bakery + Perc_General,
the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))
```

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

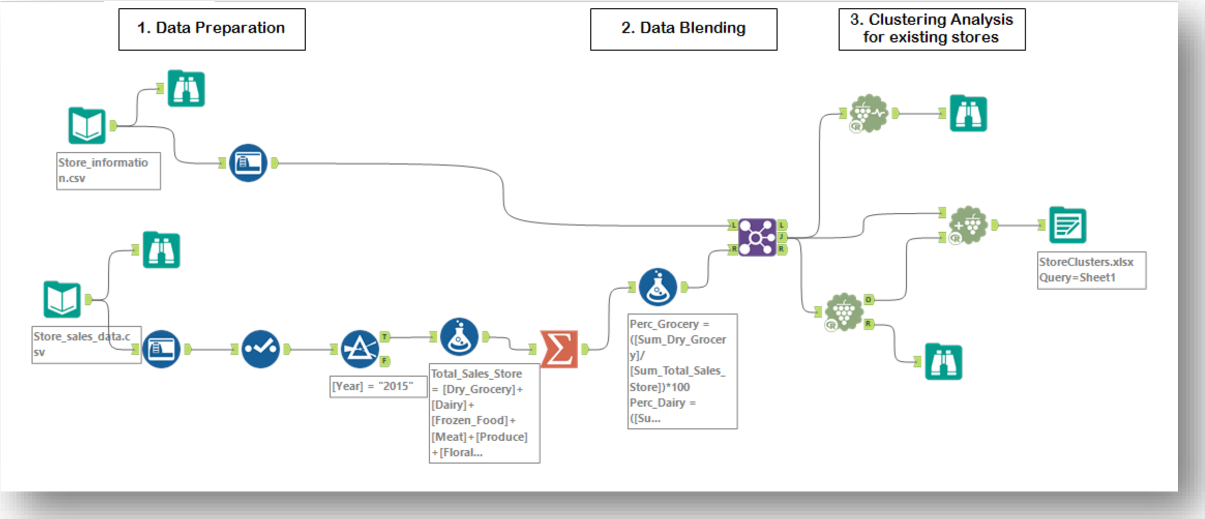
Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

	Perc_Grocery	Perc_Dairy	Perc_Frozen	Perc_Meat	Perc_Produce	Perc_Floral	Perc_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168

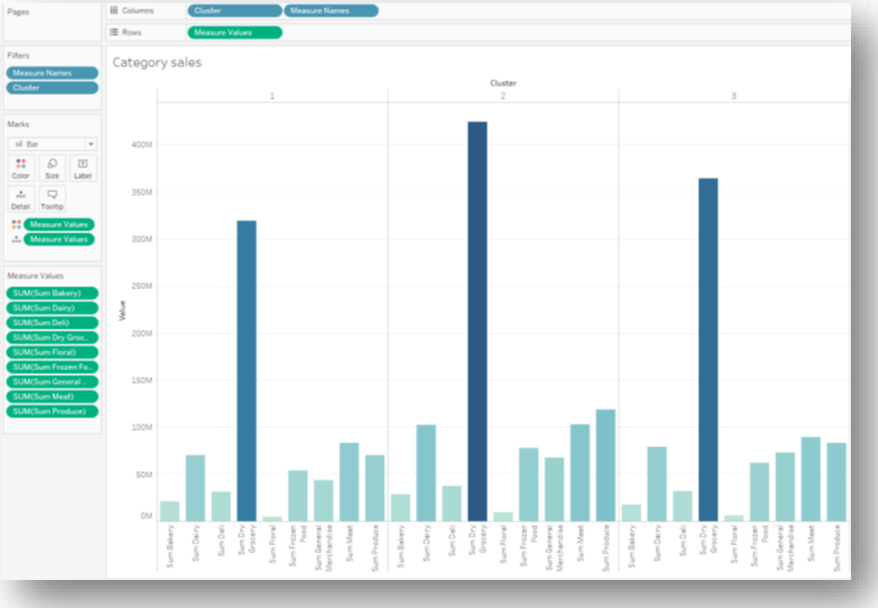


3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482
	Perc_Bakery	Perc_General					
1	0.428226	-0.674769					
2	0.312878	-0.329045					
3	-0.866255	1.135432					

From this report, we can conclude that we will have 3 clusters of the following size: 25,35,25. We can append each cluster ID to the existing stores by means of the Append Cluster Tool.
Here below you may find the complete workflow for Task 1.

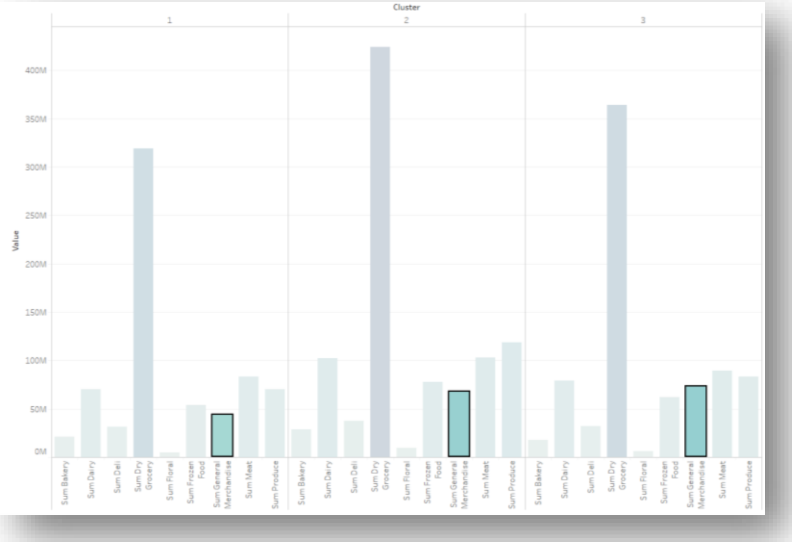

1. Data Preparation
2. Data Blending
3. Clustering Analysis for existing stores
Store_information.csv
Store_sales_data.csv
[Year] = "2015"
Total_Sales_Store = [Dry_Grocery]+ [Dairy]+ [Frozen_Food]+ [Meat]+[Produce] +[Floral...
Perc_Grocery = ([Sum_Dry_Grocery]/ [Sum_Total_Sales_Store])*100 Perc_Dairy = ([Su...
StoreClusters.xlsx Query=Sheet1

3. *Based on the results of the clustering model, what is one way that the clusters differ from one another?*
   First and foremost, we notice that Cluster 2 has the biggest size while Cluster 1 and 3 have equivalent sizes.
   Having a closer look at the identified segments, we observe that stores belonging to different clusters may have different inventory needs. For example, the category Dry_Grocery is a best seller across all clusters, however, only stores grouped into Cluster 2 exceed 400M in sales. Similarly, Cluster 1 has the lowest sales results for the category Floral, despite being the same size as Cluster 3.

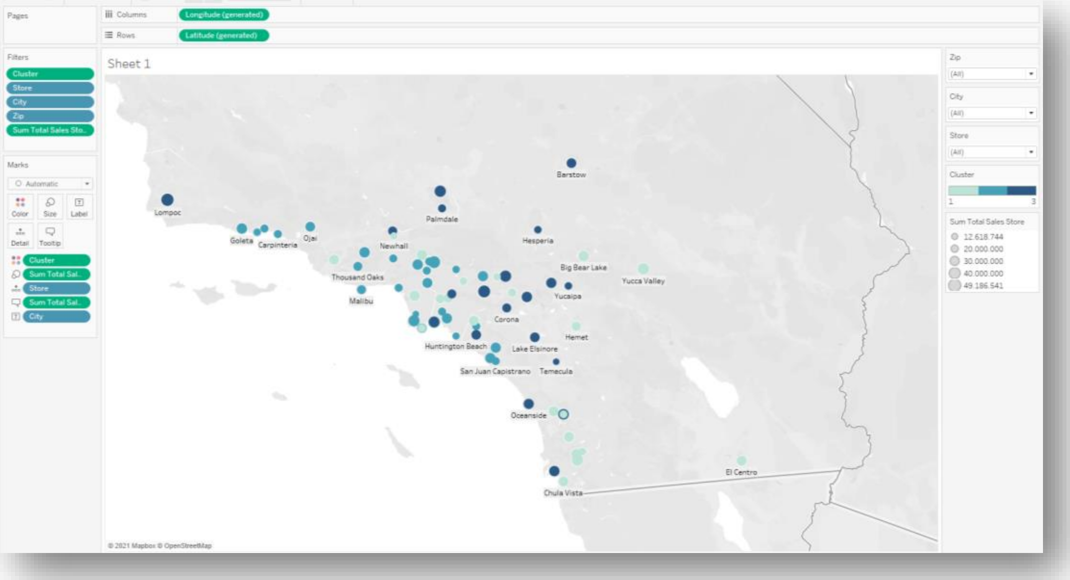


Furthermore, as shown in the table and bar chart below, stores in Cluster 3 sell more General_Merchandise items as compared to stores in Cluster 1 or 2, even though the number of stores in Cluster 2 is higher than the number of stores comprised in Cluster 3.

ClusterID	Sales_General_Merchandise
1	$ 43.941.026,31
2	$ 67.359.288,03
3	$ 72.779.160,27



4. *Please provide a Tableau visualization that shows the location of the stores, uses color to show cluster, and size to show total sales*.
The 3 store formats have been encoded with color while total sales are displayed in ascending order in the legend section on the left.
From the map, we can see that stores are located in different cities in California. In the close-up below, we may notice that stores belonging to Cluster 1 look dominant in the southern area.



Columns: Longitude (generated)
Rows: Latitude (generated)

Filters: Cluster, Store, City, Zip, Sum Total Sales Sto..

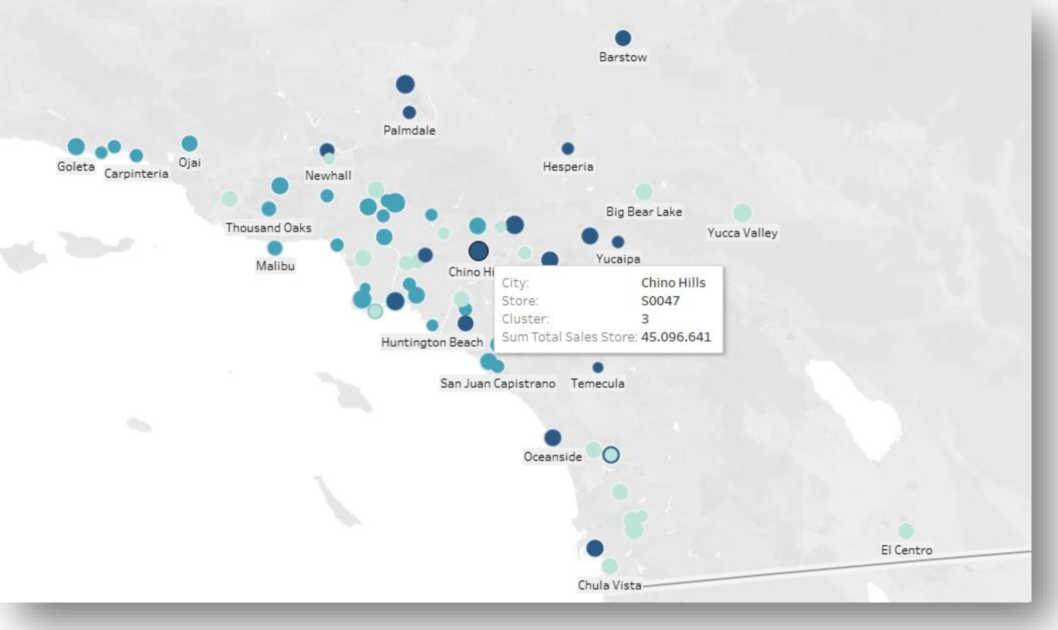
Marks: Automatic
Color: Cluster
Size: Sum Total Sal..
Detail: Store
Tooltip: Sum Total Sal..
Label: City

Zip: (All)
City: (All)
Store: (All)

Cluster: 1 – 3

Sum Total Sales Store
- 12.618.744
- 20.000.000
- 30.000.000
- 40.000.000
- 49.186.541

Larger points on the map indicate high sales volumes. When hovering over one of these points of sales, we can easily view the total sales amount, the store number as well as the city where the store is located.



City: Chino Hills
Store: S0047
Cluster: 3
Sum Total Sales Store: 45.096.641

Task 2: Formats for New Stores

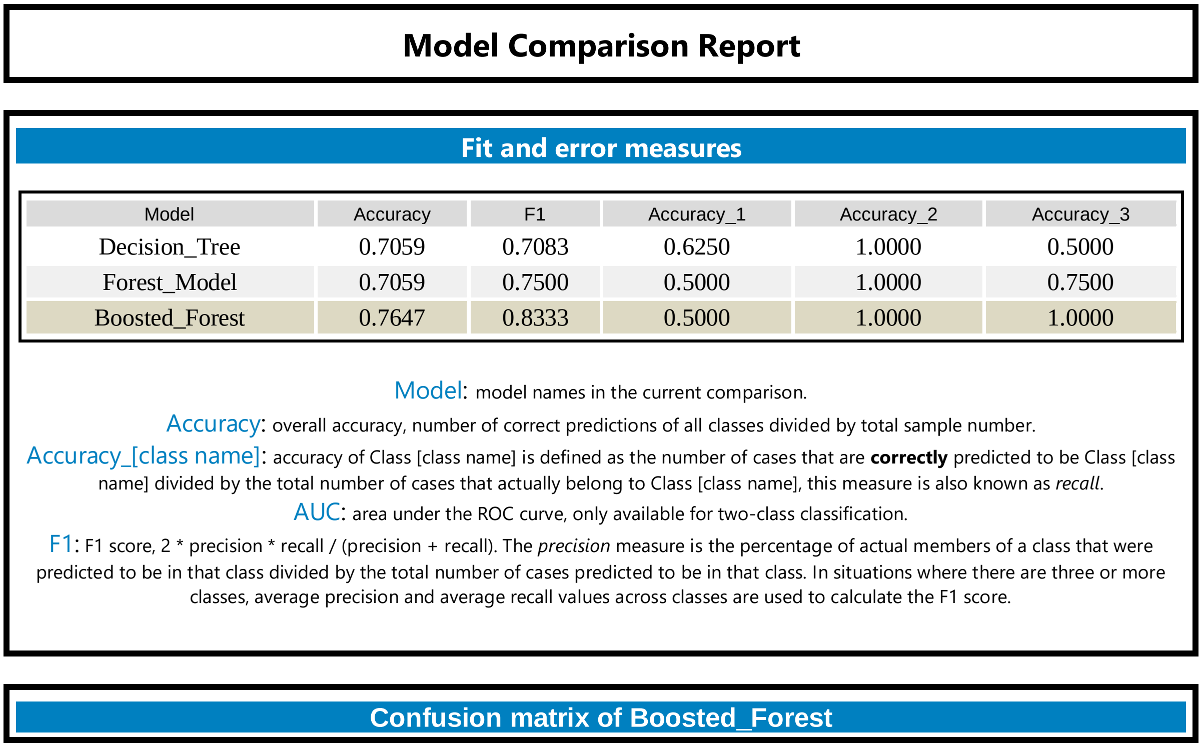
1. *What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)*

   Now that we have assigned a ClusterID to existing stores, it's time to group the 10 new stores into the identified segments. We are going to build up a model that predicts which segment a store falls into based on a set of demographic and socioeconomic predictor variables that relate to the population residing in the area around each new store. This information can be found in the StoreDemographicData.csv file.

   In this context, the target variable is the categorical variable ClusterID which can take up to 3 different values. This means that we are dealing with a non-binary classification problem. Therefore, we are going to develop three different models: a Decision Tree, a Forest Model and a Boosted Model. Afterwards, we can compare their results to understand which model displays the best fit for our business case. In a final step, we will score the winner model to assign each new store to the proper ClusterID.

   Before starting our modelling process, we need to prepare data, choosing the right data formats, and join the StoreDemographicData.csv file containing the predictor variables with our newly created list of existing stores matched to their ClusterIds.

   Equally important, we should create a holdout sample (20%) with the Create Samples Tool that we will later use as validation sample. After building up our models with 80% of records belonging to the estimation sample and after validating the three model outputs against the holdout sample, we obtain the following results:



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7083	0.6250	1.0000	0.5000
Forest_Model	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted_Forest	0.7647	0.8333	0.5000	1.0000	1.0000

Model: model names in the current comparison.

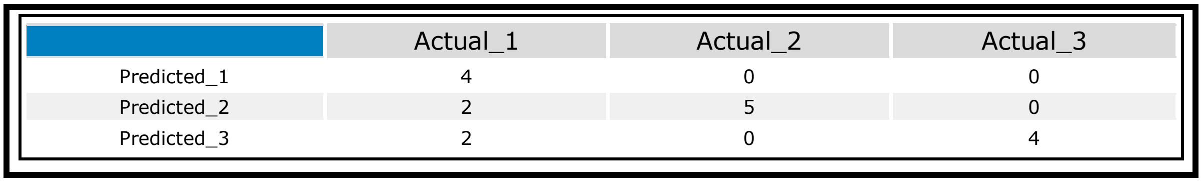
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

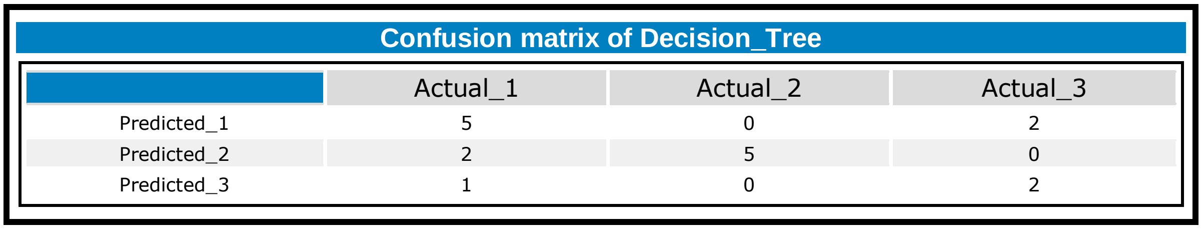
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boosted_Forest



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

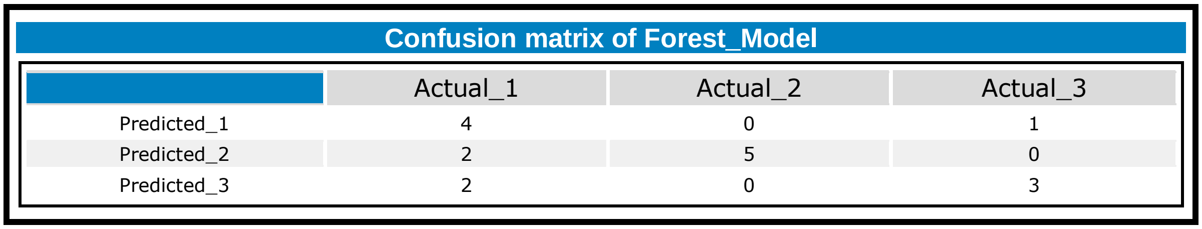
4



Confusion matrix of Decision_Tree

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

5



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

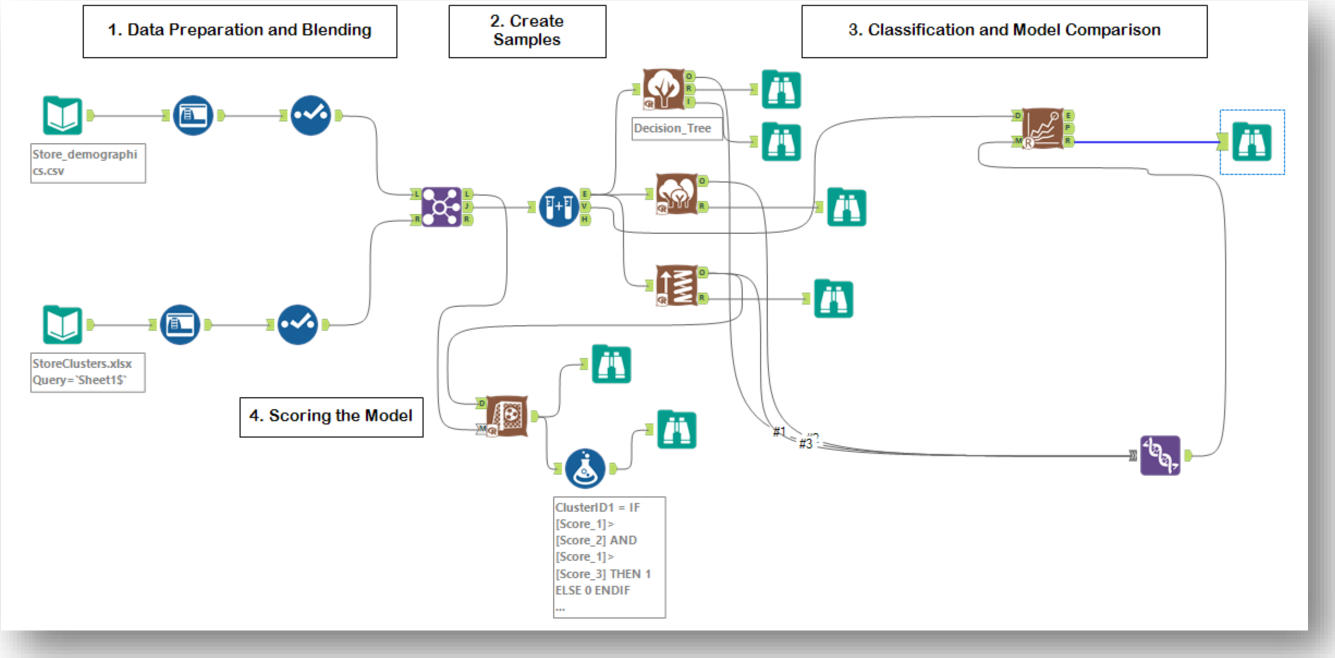
By unioning the model results, we can easily compare them side by side: the report above suggests that the Boosted Model has the highest overall accuracy as well as the highest F1 score, which is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Overall, it does a better job at predicting the cluster where each new store will fall into. For this reason, we are going to use this model to score new stores data.

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

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

ClusterID1 contains 1 new store while ClusterID2 and ClusterID3 include respectively 6 and 3 new stores.

On the following page you may find the complete workflow for Task 2.


1. Data Preparation and Blending
2. Create Samples
3. Classification and Model Comparison
Store_demographics.csv
Decision_Tree
StoreClusters.xlsx Query=`Sheet1$`
4. Scoring the Model
ClusterID1 = IF [Score_1]> [Score_2] AND [Score_1]> [Score_3] THEN 1 ELSE 0 ENDIF
...

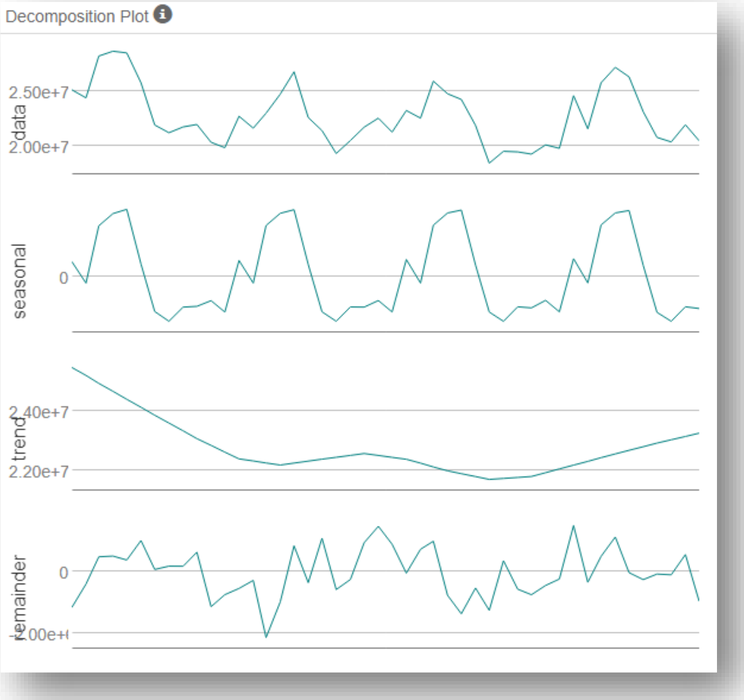
Task 3: Predicting Produce Sales

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?*

   To forecast Produce_Sales for 12 periods (months) for existing stores we should first aggregate Produce_Sales across all stores by month. We can perform this aggregation by summing Produce_Sales and grouping them by Year and Month.

   We have been asked to use a 6-month range as holdout sample, therefore, we will use the RecordId Tool and the Filter Tool to isolate RecordIds that are less or equal to RecordId40. This range will constitute our estimation sample.

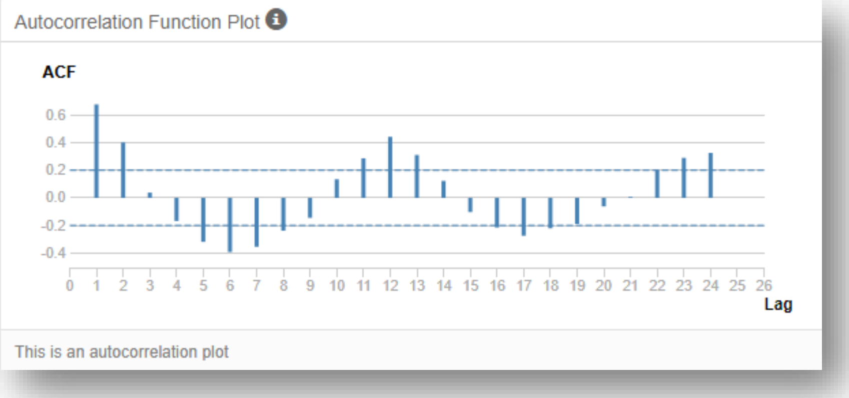
   Before starting to build our ETS and ARIMA model, however, we should have a closer look at the time series by means of a TS Plot Tool which outputs the decomposition plot shown below:

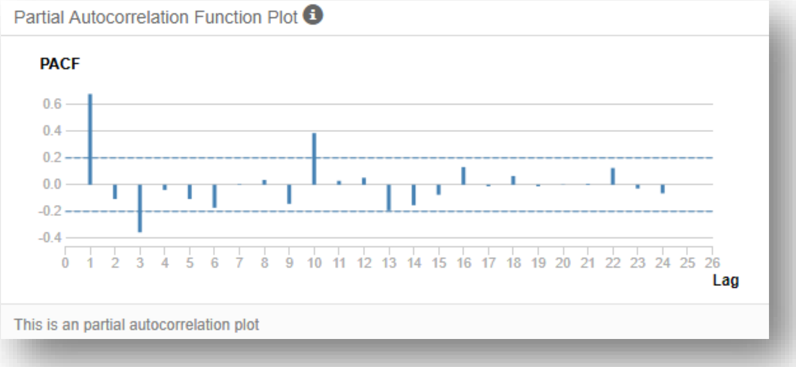


By analyzing the decomposition plot, we look at:

- Remainder: It looks like the error is growing and shrinking over time, so we will apply it multiplicatively (M).
- Trend: We can't identify a consistent, linear trend. At the beginning we detect a downward trend followed by an upward change. As there is no single detectable behavior, we should configure it as None (N).
- Seasonality: seasonal peaks seem to slowly decrease over time, suggesting a change in magnitude. We can apply it multiplicatively (M).

Let's now turn to the ACF and PACF plots:





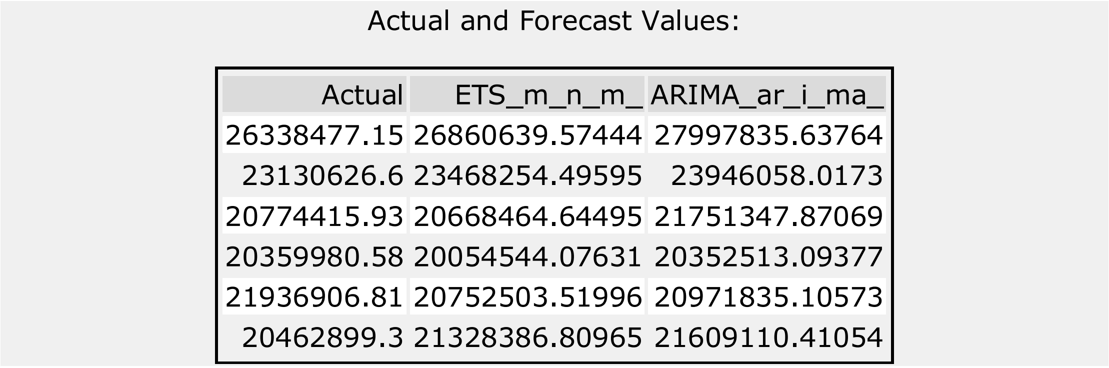
Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

By looking at these plots, we may tentatively infer the numbers of AR and/or MA terms that are needed. We can see a positive correlation for Lag-1 which indicates AR(1) or P(1). Since we can't detect a trend, there is no point in adding non-seasonal differencing, therefore d(0). The PACF plot cuts off sharply at Lag-1 while there is a more gradual decay in the ACF plot, which leads us to MA(0) or q(0). Furthermore, since we have seasonality in our series we would need to seasonally difference the series by applying D(1) term. In our case, Alteryx suggests that the optimal model is ARIMA(1,0,0)(1,1,0)[12].

Subsequently, we will validate both the ETS and ARIMA models against the validation sample and compare their outputs side by side.

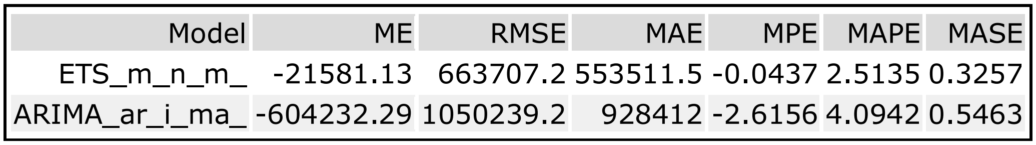
Comparison of Time Series Models



Actual and Forecast Values:

Actual	ETS_m_n_m_	ARIMA_ar_i_ma_
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

Accuracy Measures:



Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_m_n_m_	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA_ar_i_ma_	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

Overall, the ETS Model seems more accurate since its forecasted values are close enough to the actual sales values. Lastly, the ETS Model shows smaller error values across almost all error measures such as RMSE, MAPE or MASE.

Given this fact, we will apply this model to forecast Produce_Sales for both existing and new stores.

To forecast Produce_Sales for new stores, we should aggregate sales data in a way that we obtain the average monthly Total_Produce_Sales per store for each cluster. This will be the target field of our ETS Model. To achieve this objective, we need to use two Summarize Tools as follows:

With the first Summarize Tool we can:

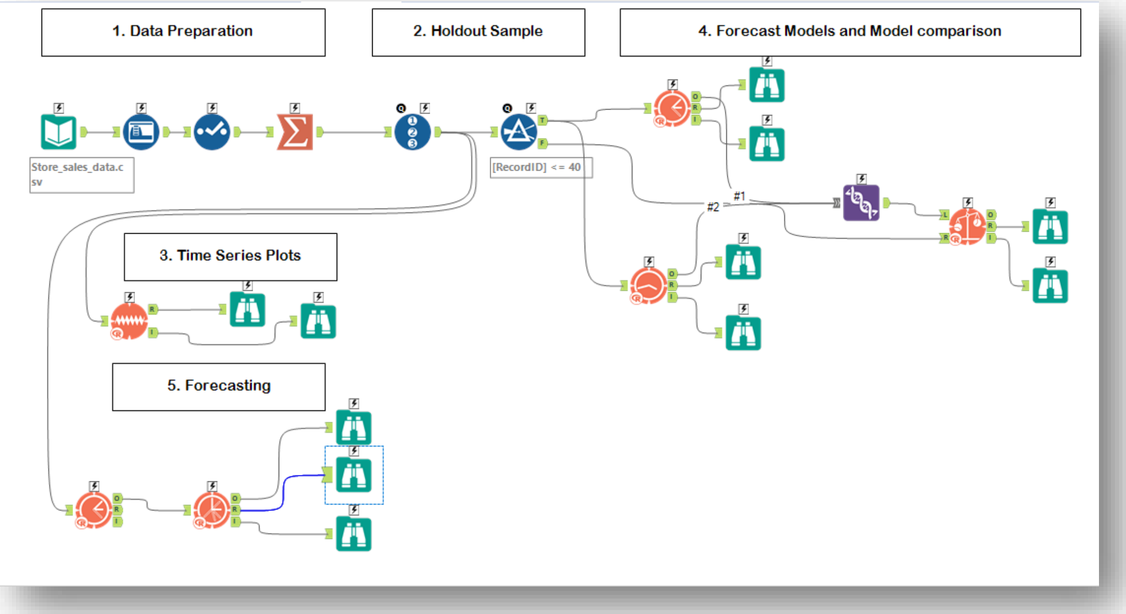
- Sum the Produce_Sales values.
- Group them by Store, Cluster, Year and Month.

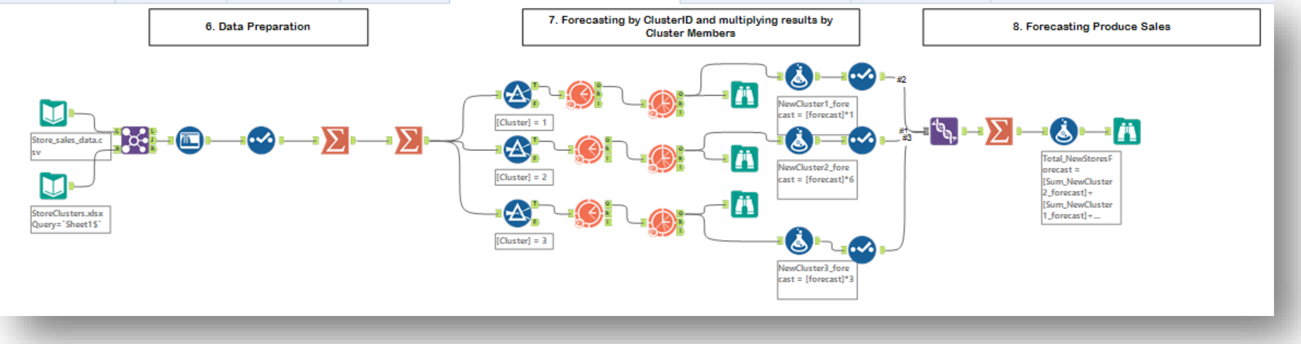
With the second Summarize Tool we can:

- Calculate the average of Sum_Produce
- Group this value by Cluster, Year and Month

Next, we will apply the model to each cluster separately and predict the future Produce_Sales values by simply multiplying the model forecasts by the number of new stores predicted to fall into that specific cluster, as revealed by the analysis performed to complete Task 2. Eventually, by summing the new stores produce sales forecasts for each of the segments we arrive at the final forecast for all new stores.

Here below you may find the complete workflows for Task 3:




6. Data Preparation
7. Forecasting by ClusterID and multiplying results by Cluster Members
8. Forecasting Produce Sales
Store_sales_data.csv
StoreClusters.xlsx Query=`Sheet1$`
[Cluster] = 1
[Cluster] = 2
[Cluster] = 3
NewCluster1_forecast = [forecast]*1
NewCluster2_forecast = [forecast]*6
NewCluster3_forecast = [forecast]*3
#2
#1
#3
Total_NewStoresForecast = [Sum_NewCluster2_forecast]+ [Sum_NewCluster1_forecast]+...

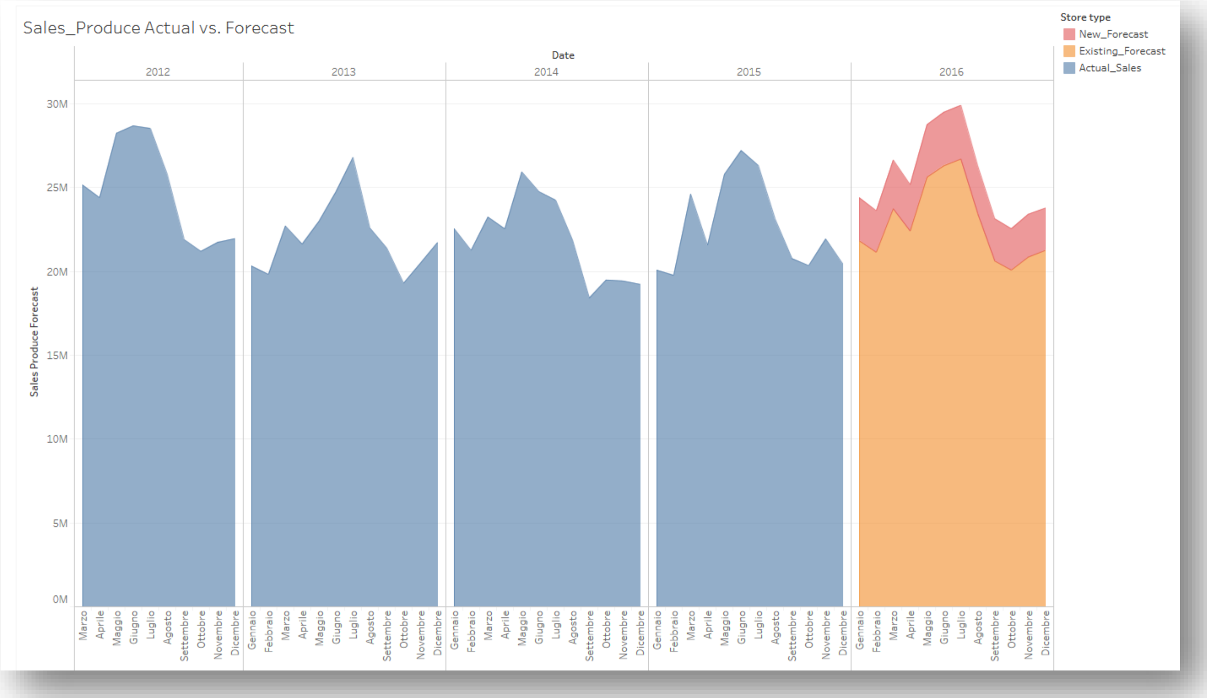
2. *Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.*

Date	Store type	Produce_Sales
01/01/2016	New_Forecast	$ 2.563.357,91
01/02/2016	New_Forecast	$ 2.483.924,73
01/03/2016	New_Forecast	$ 2.910.944,15
01/04/2016	New_Forecast	$ 2.764.881,87
01/05/2016	New_Forecast	$ 3.141.305,87
01/06/2016	New_Forecast	$ 3.195.054,20
01/07/2016	New_Forecast	$ 3.212.390,95
01/08/2016	New_Forecast	$ 2.852.385,77
01/09/2016	New_Forecast	$ 2.521.697,19
01/10/2016	New_Forecast	$ 2.466.750,89
01/11/2016	New_Forecast	$ 2.557.744,59
01/12/2016	New_Forecast	$ 2.530.510,81
01/01/2016	Existing_Forecast	$ 21.829.060,03
01/02/2016	Existing_Forecast	$ 21.146.329,63
01/03/2016	Existing_Forecast	$ 23.735.686,94
01/04/2016	Existing_Forecast	$ 22.409.515,28
01/05/2016	Existing_Forecast	$ 25.621.828,73
01/06/2016	Existing_Forecast	$ 26.307.858,04
01/07/2016	Existing_Forecast	$ 26.705.092,56
01/08/2016	Existing_Forecast	$ 23.440.761,33
01/09/2016	Existing_Forecast	$ 20.640.047,32
01/10/2016	Existing_Forecast	$ 20.086.270,46
01/11/2016	Existing_Forecast	$ 20.858.119,96
01/12/2016	Existing_Forecast	$ 21.255.190,24
01/03/2012	Actual_Sales	$ 25.151.525,84
01/04/2012	Actual_Sales	$ 24.406.048,39

01/05/2012	Actual_Sales	$	28.249.539,01
01/06/2012	Actual_Sales	$	28.691.364,32
01/07/2012	Actual_Sales	$	28.535.707,45
01/08/2012	Actual_Sales	$	25.793.520,64
01/09/2012	Actual_Sales	$	21.915.641,66
01/10/2012	Actual_Sales	$	21.203.562,52
01/11/2012	Actual_Sales	$	21.736.158,96
01/12/2012	Actual_Sales	$	21.962.976,75
01/01/2013	Actual_Sales	$	20.322.683,64
01/02/2013	Actual_Sales	$	19.829.620,75
01/03/2013	Actual_Sales	$	22.717.069,85
01/04/2013	Actual_Sales	$	21.625.385,04
01/05/2013	Actual_Sales	$	23.000.152,40
01/06/2013	Actual_Sales	$	24.755.406,20
01/07/2013	Actual_Sales	$	26.803.105,56
01/08/2013	Actual_Sales	$	22.600.217,01
01/09/2013	Actual_Sales	$	21.401.265,74
01/10/2013	Actual_Sales	$	19.296.578,09
01/11/2013	Actual_Sales	$	20.489.773,49
01/12/2013	Actual_Sales	$	21.715.706,67
01/01/2014	Actual_Sales	$	22.544.458,38
01/02/2014	Actual_Sales	$	21.262.413,12
01/03/2014	Actual_Sales	$	23.247.168,62
01/04/2014	Actual_Sales	$	22.541.987,92
01/05/2014	Actual_Sales	$	25.943.046,75
01/06/2014	Actual_Sales	$	24.782.178,43
01/07/2014	Actual_Sales	$	24.263.117,59
01/08/2014	Actual_Sales	$	21.879.988,86
01/09/2014	Actual_Sales	$	18.407.263,58
01/10/2014	Actual_Sales	$	19.497.571,95
01/11/2014	Actual_Sales	$	19.444.753,17
01/12/2014	Actual_Sales	$	19.240.384,75
01/01/2015	Actual_Sales	$	20.088.529,29
01/02/2015	Actual_Sales	$	19.772.333,34
01/03/2015	Actual_Sales	$	24.608.406,71
01/04/2015	Actual_Sales	$	21.559.729,45
01/05/2015	Actual_Sales	$	25.792.074,59
01/06/2015	Actual_Sales	$	27.212.464,15
01/07/2015	Actual_Sales	$	26.338.477,15
01/08/2015	Actual_Sales	$	23.130.626,60
01/09/2015	Actual_Sales	$	20.774.415,93

01/10/2015	Actual_Sales	$	20.359.980,58
01/11/2015	Actual_Sales	$	21.936.906,81
01/12/2015	Actual_Sales	$	20.462.899,30

Here below you may find the Tableau visualization with the Actual_Sales or past sales values (in blue) next to the forecasted sales values for both existing (in orange) and new stores (in red) within the Produce category, encoded with the respective colors. Sales data are displayed across different years on a monthly basis.


Sales_Produce Actual vs. Forecast
Date
Store type
New_Forecast
Existing_Forecast
Actual_Sales
Sales Produce Forecast
30M
25M
20M
15M
10M
5M
0M
Marzo
Aprile
Maggio
Giugno
Luglio
Agosto
Settembre
Ottobre
Novembre
Dicembre
Gennaio
Febbraio
Marzo
Aprile
Maggio
Giugno
Luglio
Agosto
Settembre
Ottobre
Novembre
Dicembre
Gennaio
Febbraio
Marzo
Aprile
Maggio
Giugno
Luglio
Agosto
Settembre
Ottobre
Novembre
Dicembre
Gennaio
Febbraio
Marzo
Aprile
Maggio
Giugno
Luglio
Agosto
Settembre
Ottobre
Novembre
Dicembre
Gennaio
Febbraio
Marzo
Aprile
Maggio
Giugno
Luglio
Agosto
Settembre
Ottobre
Novembre
Dicembre