# Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it here:
https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

## Task 1: Determine Store Formats for Existing Stores

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

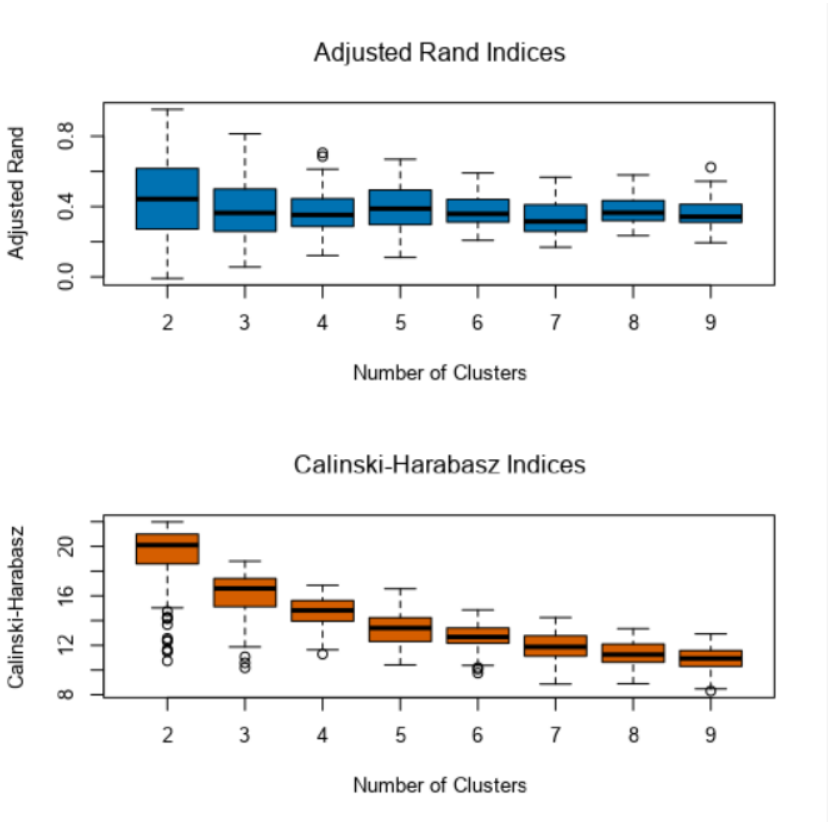



Figure 1.1: Adjusted Rand and Calinski-Harabasz (CH) Indices

The optimal number of store formats is 3.

In Figure 1.1, the Adjusted Rand Index has similar median values, and the cluster solutions with higher median values appear to have higher interquartile range values. In this case, the Adjusted Rand Index does not produce a conclusive result.

On the other hand, the Calinski-Harabasz (CH) Index shows that the 2-cluster solution produces the highest median value. However, this cluster solution has the largest range

and a considerable number of outliers. The 3-cluster solution is the best option as it has the second highest median and few outliers. Its range is much smaller than that of the 2-cluster solution, and its interquartile range is not much larger than those of the others.

2. How many stores fall into each store format?

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 |

Figure 1.2: Cluster Information of the K-Means Clustering Solution

Cluster 1 and 3 have 25 stores each, while cluster 2 has 35 stores.

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

| | Pct_Dry_Grocery | Pct_Dairy | Pct_Frozen_Food | Pct_Meat | Pct_Produce | Pct_Floral | Pct_Deli |
|---|---|---|---|---|---|---|---|
| 1 | 0.528249 | -0.215879 | -0.261597 | 0.614147 | -0.655027 | -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 | -0.178481 |

| | Pct_Bakery | Pct_General_Merchandise |
|---|---|---|
| 1 | 0.428226 | -0.674769 |
| 2 | 0.312878 | -0.329045 |
| 3 | -0.866255 | 1.135432 |

Figure 1.3: Variable Comparisons

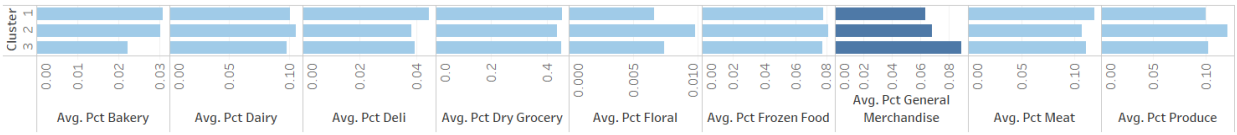



Figure 1.4: Average Percentage Sales for All Categories

Figure 1.3 shows that in Pct_General_Merchandise cluster 3 has the largest positive value, and cluster 1 has the largest negative value. Looking at Figure 1.4, we can see that on average stores in cluster 3 sell more general merchandise products.

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.

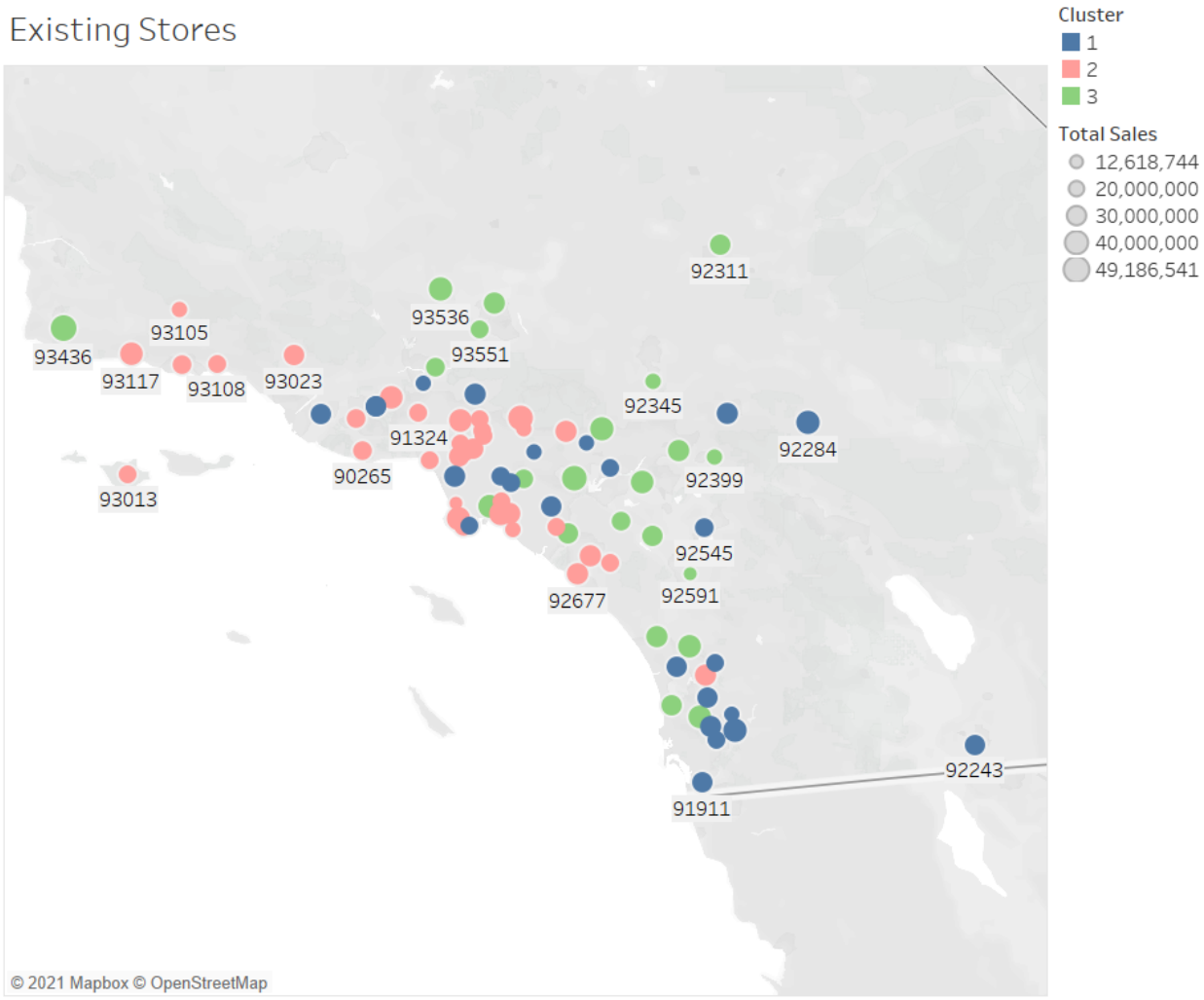



Figure 1.5: Map Showing the Locations of Existing Stores

Figure 1.5 visualises Cluster 1, 2 and 3 in blue, pink and green respectively. The sizes of the circles represent total sales, while the labels represent zip codes.

## 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.)

   As this is a non-binary classification problem, three decision models are used: a decision tree, a forest model and a boosted model. These three models are then compared to find out which one fits the data best.

| Record | Layout |
|---|---|
| 1 | **Model Comparison Report** |
| 2 | **Fit and error measures** |

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| DT_NewStore | 0.6471 | 0.6667 | 0.5000 | 1.0000 | 0.5000 |
| FM_NewStore | 0.7059 | 0.7500 | 0.5000 | 1.0000 | 0.7500 |
| BM_NewStore | 0.7059 | 0.7500 | 0.5000 | 1.0000 | 0.7500 |

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.

3 **Confusion matrix of BM_NewStore**

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

4 **Confusion matrix of DT_NewStore**

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

5 **Confusion matrix of FM_NewStore**

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

Figure 2.1: Model Comparison Report for Decision Tree Model (DT_NewStore), Forest

Model (FM_NewStore) and Boosted Model (BM_NewStore)

As we can see in Figure 2.1, accuracy of segments 1 and 2 are the same for all the models (Decision Tree, Forest Model and Boosted Model). Forest Model and Boosted Model have the highest overall accuracy, F1 score and accuracy of segment 3. Therefore, either Forest Model or Boosted Model can be used to predict the best store format for the new stores. I decided to choose the Boosted Model.

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

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

Segment 1 has 1 new store, Segment 2 has 6 new stores, and Segment 3 has 3 new stores.

## 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?

ETS(M,N,M) is used for each forecast.

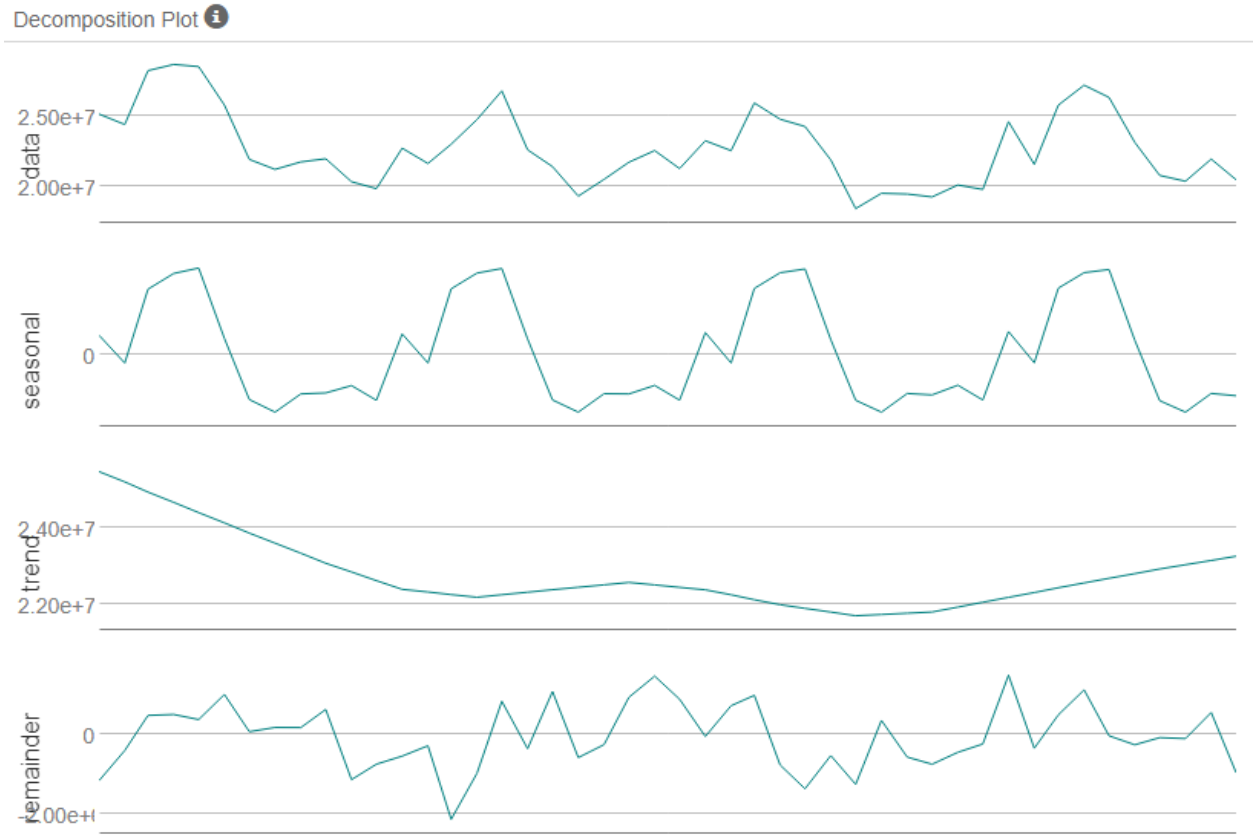



Figure 3.1: Decomposition Plot

Figure 3.1 shows a time series broken down into its three components: seasonality, trend and error. As the seasonality plot changes in magnitude (i.e. the peaks are decreasing), the seasonality is applied multiplicatively (M). On the other hand, the time series does not have a trend (N) as the line is decreasing significantly, increasing a little, decreasing a little and then increasing considerably. Additionally, as the error plot fluctuates between larger and smaller errors and therefore does not have a constant variance, the error is applied multiplicatively (M).

| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|
| 3502.9443415 | 969051.6076376 | 787577.7006835 | -0.1381187 | 3.4677635 | 0.4396486 | 0.0077488 |

Figure 3.2: In-sample error measures of ETS(M,N,M)

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

Figure 3.3: In-sample error measures of ARIMA(1,0,0)(1,1,0)[12]

(As for the ARIMA model, the non-seasonal component has an autoregressive term of 1,

a differencing term of 0 and a moving average term of 0. In addition, its seasonal component has 12 periods in each season, an autoregressive term of 1, a differencing term of 1 and a moving average term of 0.)

When comparing the in-sample error measures in Figure 3.2 and 3.3, we focus on RMSE (Root Mean Squared Error) and MASE (Mean Absolute Scaled Error). RMSE shows how many standard deviations from the mean the forecasted values fall, whereas MASE measures the relative reduction in error compared to a naive model. The ETS model has a lower RMSE and a slightly higher MASE.

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

Figure 3.4: Accuracy Measures of ETS(M,N,M) and ARIMA(1,0,0)(1,1,0)[12]

Looking at the models' ability to predict the holdout sample (in Figure 3.4), we see that the ETS model has better predictive qualities. The ME (Mean Error), RMSE, MAE (Mean Absolute Error), MPE (Mean Percentage Error), MAPE (Mean Absolute Percentage Error) and MASE of the ETS model are closer to 0.

In-sample errors are training errors. The training data is used to estimate any parameters of a forecasting method, while the test data is used to evaluate its accuracy.

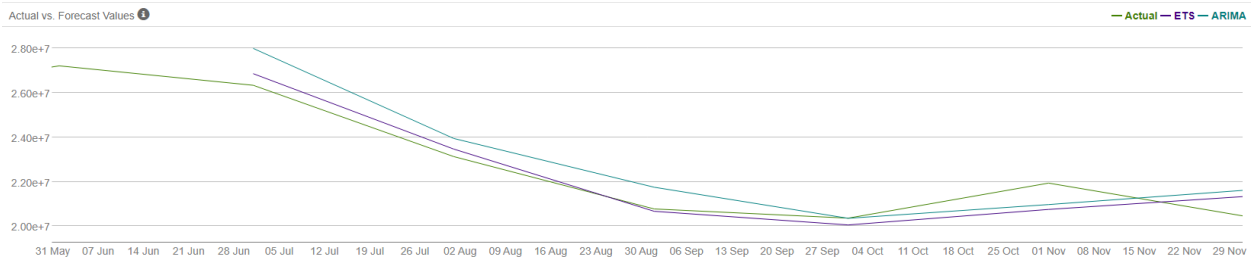



Figure 3.5: Actual vs Forecast Values

Figure 3.5 compares the forecasts provided by the ETS and ARIMA models by plotting them along with the actual values. It shows that the ETS forecasts on average are closer to the actual produce sales.

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.

| Month | New Stores | Existing Stores |
|---|---|---|

| | | |
|---|---|---|
| Jan 2016 | 2,563,357.91 | 21,136,641.78 |
| Feb 2016 | 2,483,924.73 | 20,507,039.12 |
| Mar 2016 | 2,910,944.15 | 23,506,565.98 |
| Apr 2016 | 2,764,881.87 | 22,208,405.76 |
| May 2016 | 3,141,305.87 | 25,380,147.77 |
| Jun 2016 | 3,195,054.20 | 25,966,799.47 |
| Jul 2016 | 3,212,390.95 | 26,113,792.57 |
| Aug 2016 | 2,852,385.77 | 22,899,285.77 |
| Sep 2016 | 2,521,697.19 | 20,499,583.91 |
| Oct 2016 | 2,466,750.89 | 19,971,242.82 |
| Nov 2016 | 2,557,744.59 | 20,602,665.92 |
| Dec 2016 | 2,530,510.81 | 21,073,222.08 |

The table above shows our produce forecasts for existing and new stores.

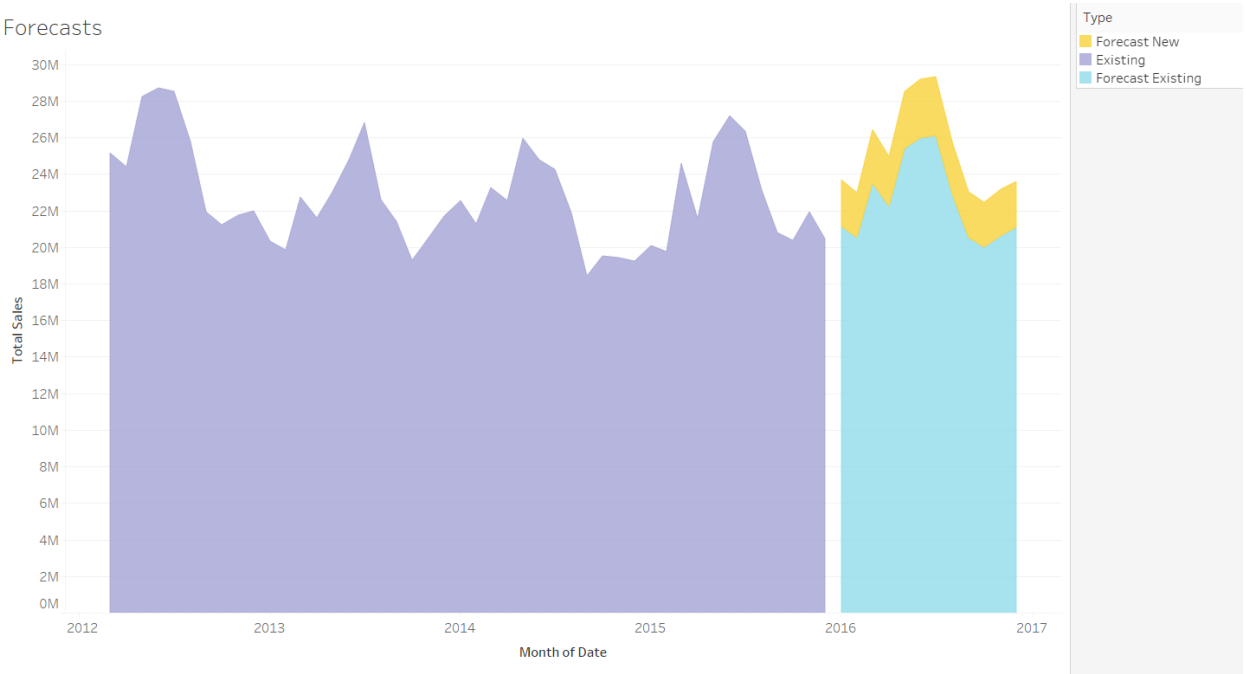



Figure 3.6: Visualisation of Forecasts

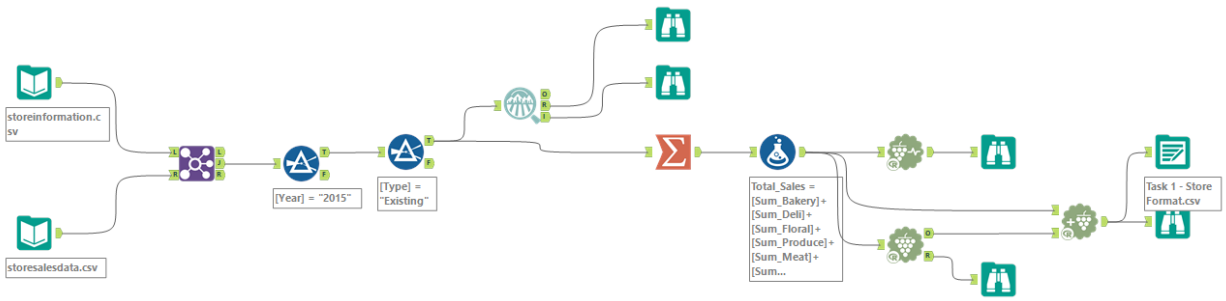
Figure 3.6 shows historical data, existing stores forecasts and new stores forecasts.

## Before you submit

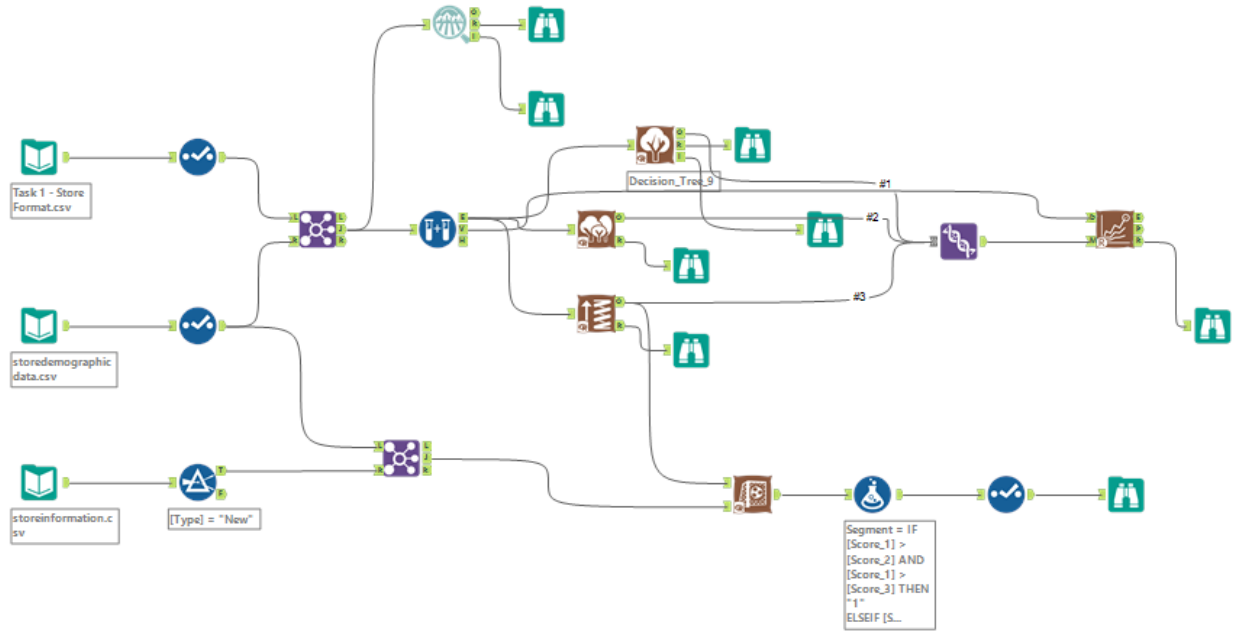
Please check your answers against the requirements of the project dictated by the rubric. Reviewers will use this rubric to grade your project.

# Appendix

## Alteryx workflow of Task 1:




## Alteryx workflow of Task 2:



Alteryx workflow of Task 3: