

# Predictive Analytics for Business

# Project #6 [Combining Predictive Techniques](https://classroom.udacity.com/nanodegrees/nd008-mena-connect/parts/e2b3ff5a-3706-4554-a8b2-8ae78dbed63a#lesson-card-scroll-target-33aed7ac-5333-4051-a15b-19a9ab499e27)

Name: Marwan Saeed Alsharabbi

Date: 17-Feb-2020

2020

**Introduction**

**Time Series Forecasting Summary**

**ETS Models**

ETS models are designed to forecast time series data by observing the trend and seasonality patterns in a time series, and projecting those trends into the future.

**STEP 1: TIME SERIES DECOMPOSITION PLOT**

A time series decomposition plot allows you to observe the seasonality, trend, and error/remainder terms of a time series.

Useful Alteryx Tool: TS Plot

**STEP 2: DETERMINE ERROR, TREND, AND SEASONALITY**

An ETS model has three main components: error, trend, and seasonality. Each can be applied either additively, multiplicatively, or not at all.

Trend - If the trend plot is linear then we apply it additively (A). If the trend line grows or shrinks exponentially, we apply it multiplicatively (M). If there is no clear trend, no trend component is included (N).

Seasonal - If the peaks and valleys for seasonality are constant over time, we apply it additively (A). If the size of the seasonal fluctuations tends to increase or decrease with the level of time series, we apply it multiplicatively (M). If there is no seasonality, it is not applied (N).

Error -  If the error plot has constant variance over time (peaks and valleys are about the same size), we apply it additively (A). If the error plot is fluctuating between large and small errors over time, we apply it multiplicatively (M).

Useful Alteryx Tool: TS Plot

**STEP 3: BUILD AND VALIDATE THE ETS MODEL**

Build the ETS model using the components determined in step 2. You can use internal and external validation to validate the quality of the model.

Internal validation: Look at in-sample error measures, particularly RMSE (Root-Mean-Square Error) and MASE (Mean Absolute Scaled Error).

External validation: Determine the accuracy measures by comparing the forecasted values with the holdout sample. This is especially important for comparing ETS models to other types of models, such as ARIMA.

Pick the ETS model with lowest AIC value. If the AIC values are comparable, use calculated errors to pick one that minimizes error the most. Many software tools will automate the selection of the model by minimizing AIC.

Useful Alteryx Tools: ETS, TS Compare

**STEP 4: FORECAST!**

Use the best ETS model to forecast for the desired time period. Make sure to add the holdout sample back into the model. Plot the results along with 80% and 95% confidence intervals.

Useful Alteryx Tool: TS Forecast

**ARIMA Models**

Summary: ARIMA which stands for Autoregressive Integrated Moving Average helps you forecast data for seasonal and nonseasonal data

**STEP 1: TIME SERIES DECOMPOSITION PLOT**

A time series decomposition plot allows you to observe the seasonality, trend, and error/remainder terms of a time series.

Useful Alteryx tool: TS Plot

**STEP 2: DETERMINE THE ARIMA TERMS**

Nonseasonal ARIMA models are displayed in the terms (p,d,q) which stand for p - periods to lag for, d - number of transformations used to make the data stationary, q - lags of the error component

Stationary - mean and variance are constant over time vs Non-Stationary - mean and variance change over time

Differencing - take the value in the current period and subtract it by the value from the previous period. You might have to do this several times to make the data stationary. This is the Integrated component which is d in the model terms.

Autocorrelation - How correlated a time series is with its past values, if positive at Lag-1 then AR if negative then MA

Partial Autocorrelation - The correlation between 2 variables controlling for the values of another set of variables. If the partial autocorrelation drops of quickly then AR terms, if it slowly decays then MA

Seasonal ARIMA models are denoted (p,d,q)(P,D,Q)m

These models may require seasonal differencing in addition to non-seasonal differencing. Seasonal differencing is when you subtract the value from a year previous of the current value.

Useful Alteryx tool: TS Plot

**STEP 3: BUILD AND VALIDATE THE ARIMA MODEL**

Build the ARIMA model using the terms determined in step 2. You can use internal and external validation to validate the quality of the model.

Internal validation: Look at in-sample error measures, particularly RMSE (Root-Mean-Square Error) and MASE (Mean Absolute Scaled Error).

External validation: Determine the accuracy measures by comparing the forecasted values with the holdout sample. This is especially important for comparing ARIMA models to other types of models, such as ETS.

Pick the ARIMA model with lowest AIC value. If the AIC values are comparable, use calculated errors to pick one that minimizes error the most. Many software tools will automate the selection of the model by minimizing AIC.

Useful Alteryx tools: ARIMA, TS Compare

**STEP 4: FORECAST!**

Use the best ARIMA model to forecast for the desired time period. Make sure to add the holdout sample back into the model. Plot the results along with 80% and 95% confidence intervals.

Useful Alteryx tool: TS Forecast

**Segmentation and Clustering**

**K-Centroid Clustering**

Summary: Cluster analysis identifies cohesive subgroups of observations within a dataset. It allows us to reduce a large number of observations into a smaller number of clusters.

**STEP 1: SELECT APPROPRIATE VARIABLES**

The first step is to understand the objectives for segmentation. Then, choose the appropriate variables that provide the information needed for clustering. A sophisticated cluster analysis cannot compensate for the poor choice of attributes.

**STEP 2: DATA PREPARATION**

Numeric data: Cluster analyses requires numeric data. Many non-numeric variables can be converted to numeric ones. Make sure to remove outliers as clustering algorithms are highly sensitive to outliers.

Variable reduction: This step often requires variable reduction techniques to combine variables that revolve around a particular theme. A common method is Principal Component Analysis (PCA), which reduces a set of related variables into few principal components (PCs) that explain most of the variances in the data. Rule of thumb is to use PCs that account for ~80% variance.

Scaling the data: Standardizing each variable using the z-score ensures that the results are not overly sensitive to variables with higher values.

Useful Alteryx Tool: Principal Components

**STEP 3: DETERMINE THE NUMBER OF CLUSTERS**

Use the AR and CH indices to determine the optimal method and number of clusters. Use a box and whisker plot. The higher the median and smaller the variation the better. Remember, clustering is an iterative process and may require comparing several models to arrive at a good solution.

Useful Alteryx Tool: K-Centroid Cluster Diagnostics

**STEP 4: CREATE THE CLUSTERING MODEL**

Select the variables, standardization process, clustering method, and number of clusters that gave the best solution. Create the cluster model and append the clusters to the dataset.

Useful Alteryx Tool: K-Centroid Cluster Analysis

**STEP 5: VISUALIZE AND VALIDATE RESULTS**

Visualization helps us determine the meaning and usefulness of the clustering solution. Use summary statistics to understand difference among clusters.

Validate the results: You can use internal validation and/or external validation. Plot the distribution of the validation variable for each cluster using box and whisker plot to visualize the differences.

Useful Alteryx Tool: Append Clusters

**Data Visualizations in Tableau**

**I. Connecting to Data**

In this section, you will get started with importing data into Tableau. Tableau public has fewer options, but paid versions of Tableau are quite extensive connecting directly to databases and cloud based data storage systems.

**II. Combining Data**

In this section, you will learn how to connect data from multiple sources for use in your visuals. If you are comfortable with SQL joins, this section should be second nature.

**III. Worksheets**

The visuals you create will be stored in worksheets. This is the template we will be working in for this course.

**IV. Aggregations**

Tableau performs aggregations of our data by default. In this section, you will learn more about how to work with different aggregations, as well as how to break your aggregations into a more granular level of the data.

**V. Hierarchies**

Hierarchies allow you to 'drill' into your data and questions at different levels. One of the easiest ways to think of hierarchies is in relation to time. You could look at your data at a year, month, day, hour, or another level. Moving across these levels is considered working with hierarchies.

You can also perform hierarchical calculations in other ways. Imagine you have a different company, with different departments, and teams within those departments. This creates a hierarchy that you might want to analyze at different levels.

**VI. Marks & Filters**

Filtering is one of the most powerful techniques in creating dashboards. This relates to the marks portion of a dashboard, which controls the colors, shapes and other attributes of our data. You can think of this like a WHERE statement in SQL used to filter your data to only the parts you are interested in for a specific question.

**VII. Show Me**

The Show Me portion of Tableau controls what your ending visual looks like. There are a lot of options here. In most cases, Tableau will guess what visual you want to create, but sometimes you might have your own ideas for implementation.

**VIII. Small Multiples & Dual Axis**

Small multiples & dual charts are a way to visualize data that needs to share an axis for comparison purposes. [This](http://www.evolytics.com/blog/tableau-201-how-to-make-small-multiples/) and [this](https://kb.tableau.com/articles/howto/dual-axis-bar-chart-multiple-measures) are great articles for explaining how these two parts of Tableau work and why you might use them.

**IX. Groups & Sets**

Groups and sets are two ways to categorize our data within a visualization. The difference between these two can be confusing, but we will see when and why you would use each.

**X. Calculated Fields**

Often you might add these fields to your dataset before adding your data to Tableau, but sometimes you want to add them to a visualization on the fly. Many of these calculated fields are things you have probably done in a spreadsheets application like finding a total or a cost per item.

**XI. Table Calculations**

Table calculations are often used to perform comparisons of our data over time or between groups. A great article on table calculations is available [here](https://www.tableau.com/about/blog/2017/2/top-10-tableau-table-calculations-65417).

**Capstone Project Overview**

The capstone project has three main tasks, each of which requires you to use skills you developed during the Nanodegree program. Once you complete all three tasks, please submit the project as a PDF.

**Tips**

**Split up your Alteryx workflows:** You will be using multiple data sources and complex tools, which can slow down the workflow runtime. Splitting up the workflows makes the process more manageable.

**Map out your work:** Before you dive into your analysis, think about the steps and plan ahead. This will reduce the amount of unnecessary work.

**Use visualizations:** Include visualizations to help explain your decisions and communicate your findings. Remember what you learned about making them look great!

**Ask for help:** Mentors and your fellow students can help you if you get stuck on something. When posting a question, make sure you include enough specificity so others can help.

**Task 1: Store Format for Existing Stores**

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.

**Task 1: Determining Store Format**

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. You’ve been asked to:

* Determine the optimal number of store formats based on sales data.
  + Sum sales data by StoreID and Year
  + Use percentage sales per category per store for clustering (category sales as a percentage of total store sales).
  + Use only 2015 sales data.
  + Use a K-means clustering model.
* Segment the 85 current stores into the different store formats.
* Use the StoreSalesData.csv and StoreInformation.csv files.

**Note:**

PCA is not used in this project.

**Task 1 Submission**

1. What is the optimal number of store formats? How did you arrive at that number?
2. How many stores fall into each store format?
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
4. Please provide a map created in Tableau that shows the location of the existing stores, uses color to show cluster, and size to show total sales. Make sure to include a legend! Feel free to simply copy and paste the map into the submission template.

**Task 2: Store Format for New Stores**

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.

**Project: Predictive Analytics Capstone**

**Business Problem #1: Store Format for Existing Stores**

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.

**Determining Store Format**

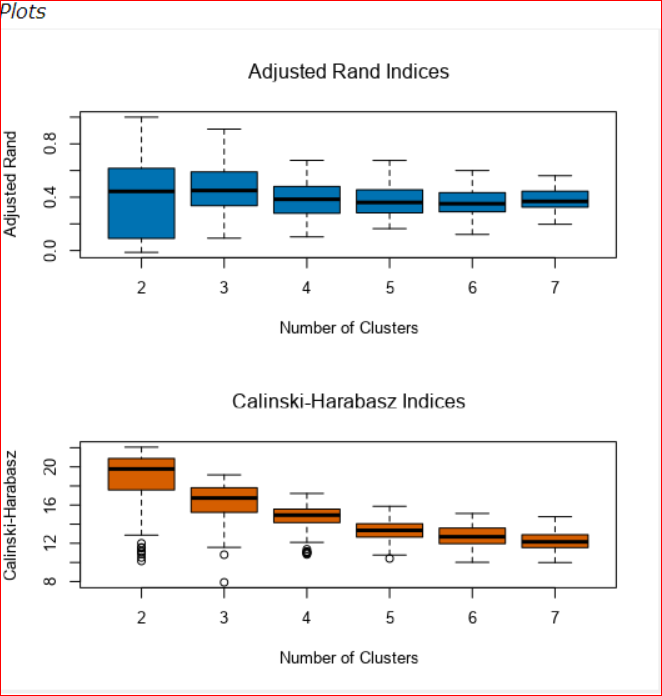
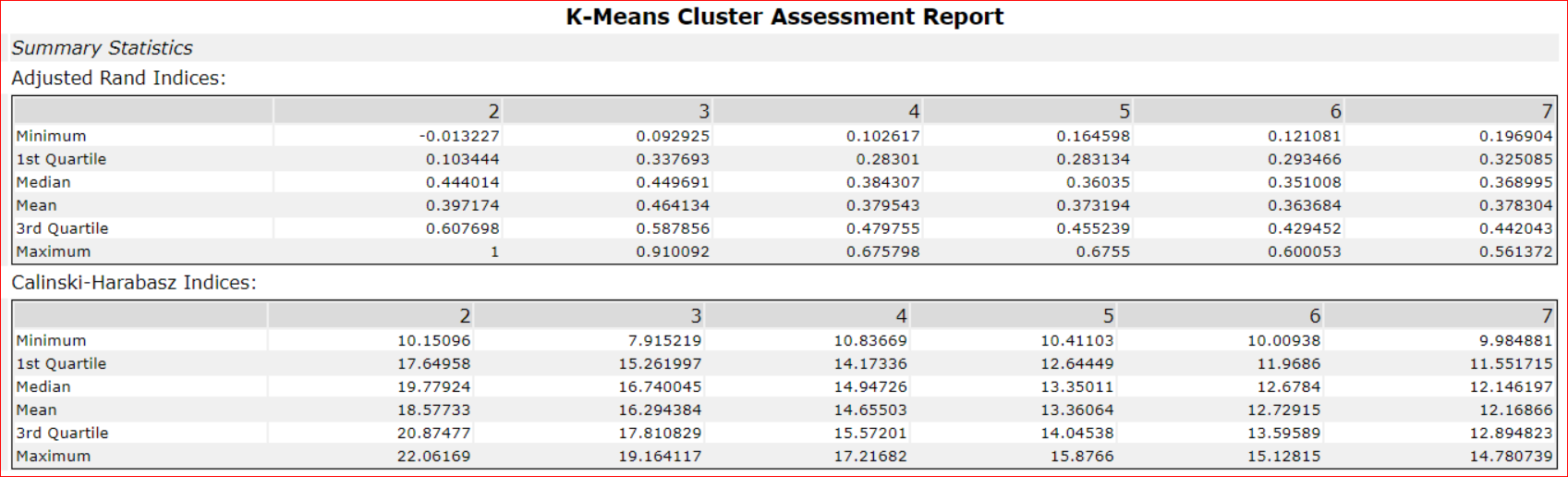
To remedy the product surplus and shortages, the company wants to introduce different store formats. Each store format will have a different product selection in order to better match local demand. The actual building sizes will not change, just the product selection and internal layouts. The terms "formats" and "segments" will be used interchangeably throughout this project. You’ve been asked to:

* Determine the optimal number of store formats based on sales data.
* Sum sales data by StoreID and Year
* Use percentage sales per category per store for clustering (category sales as a percentage of total store sales).
* Use only 2015 sales data.
* Use a K-means clustering model.
* Segment the 85 current stores into the different store formats.
* Use the StoreSalesData.csv and StoreInformation.csv files.

Task 1: Determine Store Formats for Existing Stores

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

A K-centroids analysis was conducted using K-means method to determine the number of clusters. According to our K-means assessment, Adjusted Rand Indices, and Calinski-Harabasz Indices, the optimal number of store formats is three as both the indices projected the highest median value at such and has smaller variation in its spread. see the report and plot



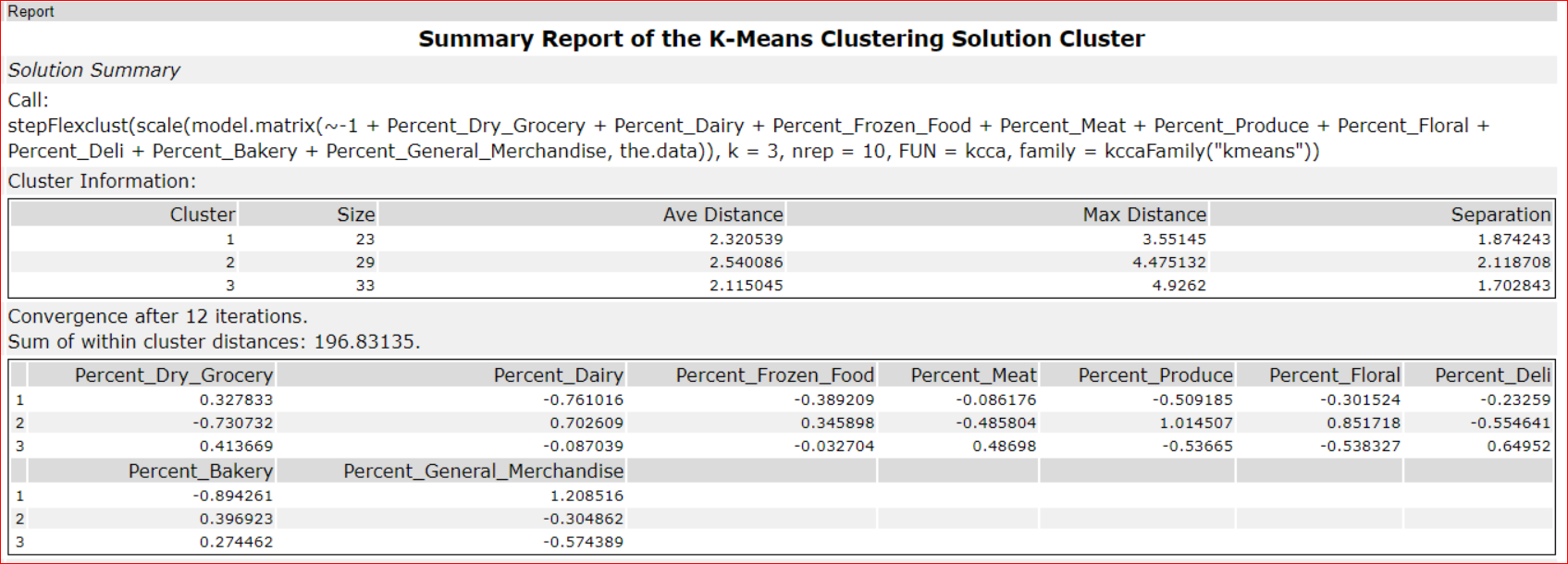
AR The higher the index, the better the stability of the block.

CH The higher the index, the better the differentiation and grouping.

According to the K-Means analysis, both Adjusted Rand Indices and Calinski-Harabasz Indices shows highest median value at 2,3, indicating that the optimal number of store formats is 3.

1. How many stores fall into each store format?

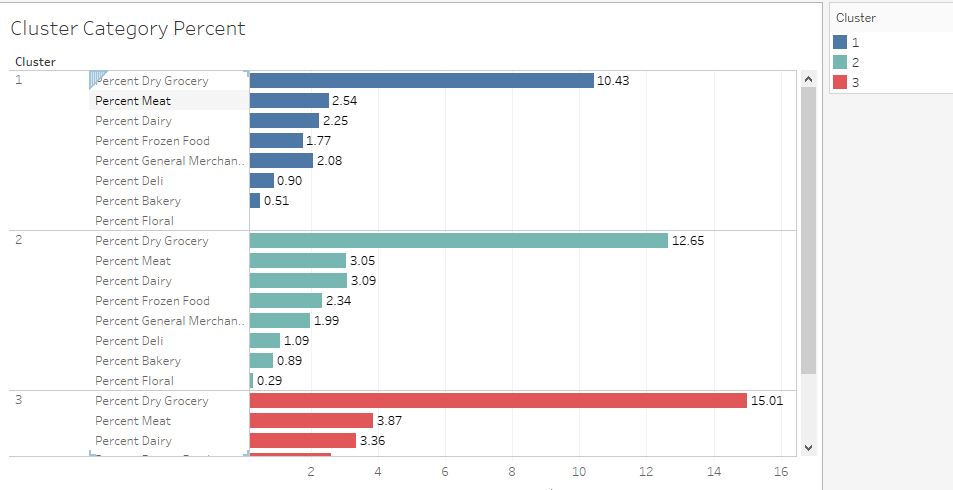
According to the Cluster Information, Cluster 1 has 23 stores, cluster 2 has 29 stores and cluster 3 has 33 stores



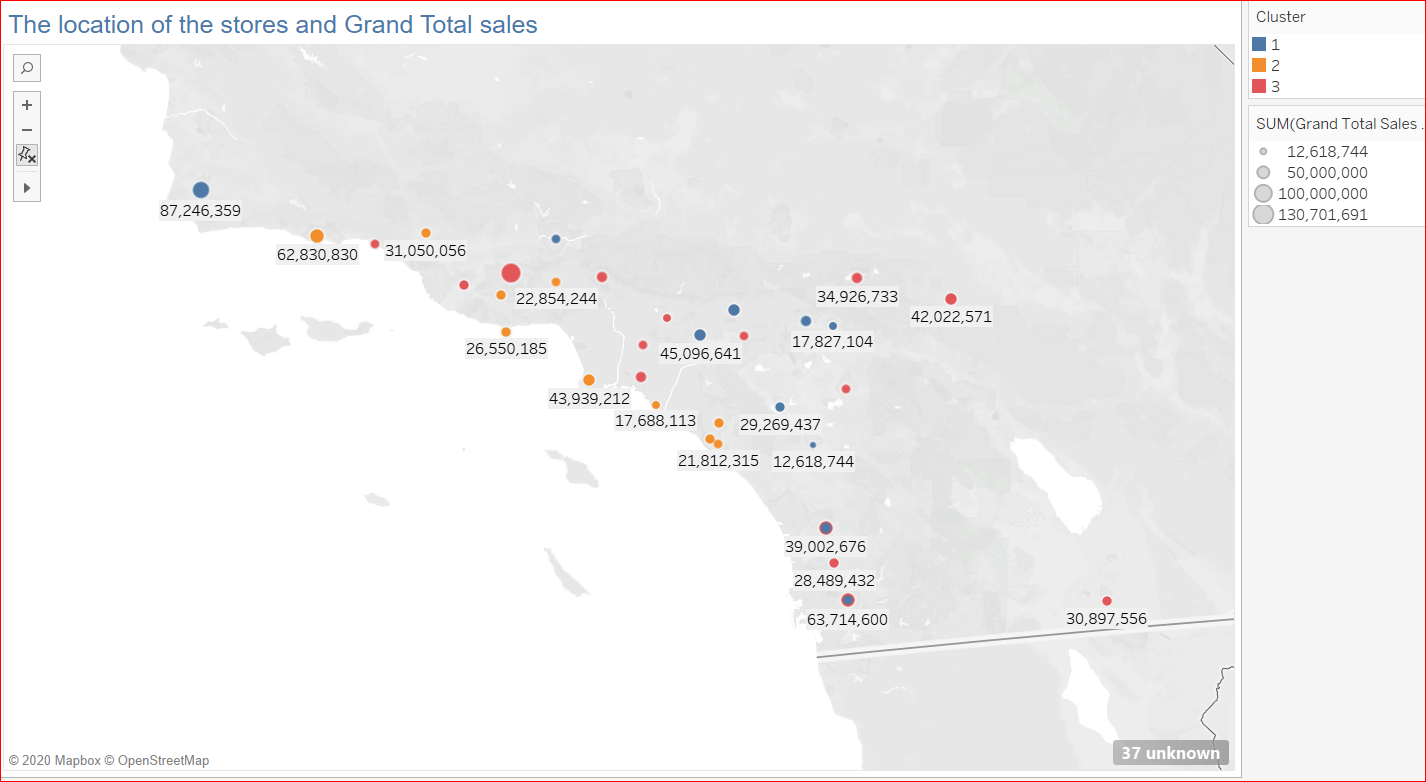
.

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

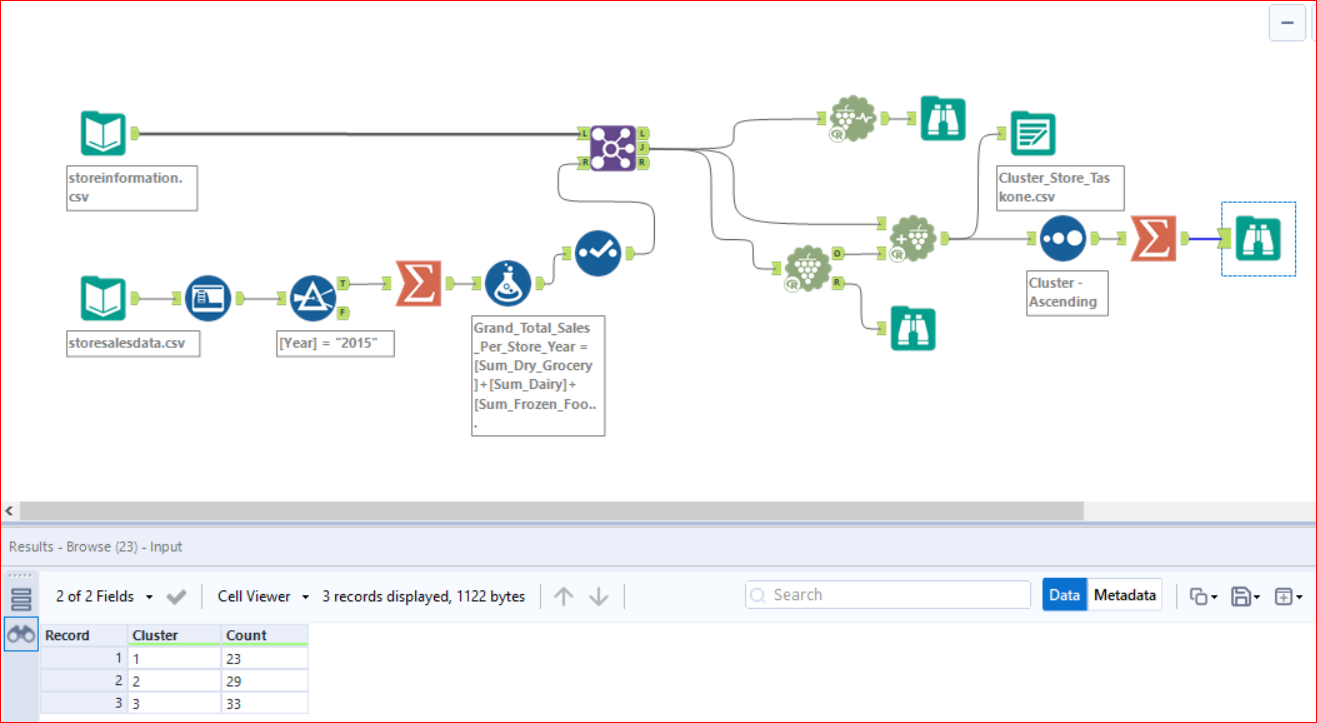
Based on the Category Share by cluster and category total sales by cluster plots, Cluster 3 has the widest range of the sales and its sales are largest for most categories Percentage of total sales highest on Percent Dry Grocery cluster 2 ,3 see the plot



1. 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.



**Workflow Task1**



**Task 2: Determine the Store Format for New Stores**

You’ve been asked to:

* Develop a model that predicts which segment a store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store.
* Use a 20% validation sample with *Random Seed* = 3 when creating samples with which to compare the accuracy of the models. Make sure to compare a decision tree, forest, and boosted model.
* Use the model to predict the best store format for each of the 10 new stores.
* Use the StoreDemographicData.csv file, which contains the information for the area around each store.
* **Note:** In a real world scenario, you could use PCA to reduce the number of predictor variables. However, there is no need to do so in this project. You can leave all predictor variables in the model.

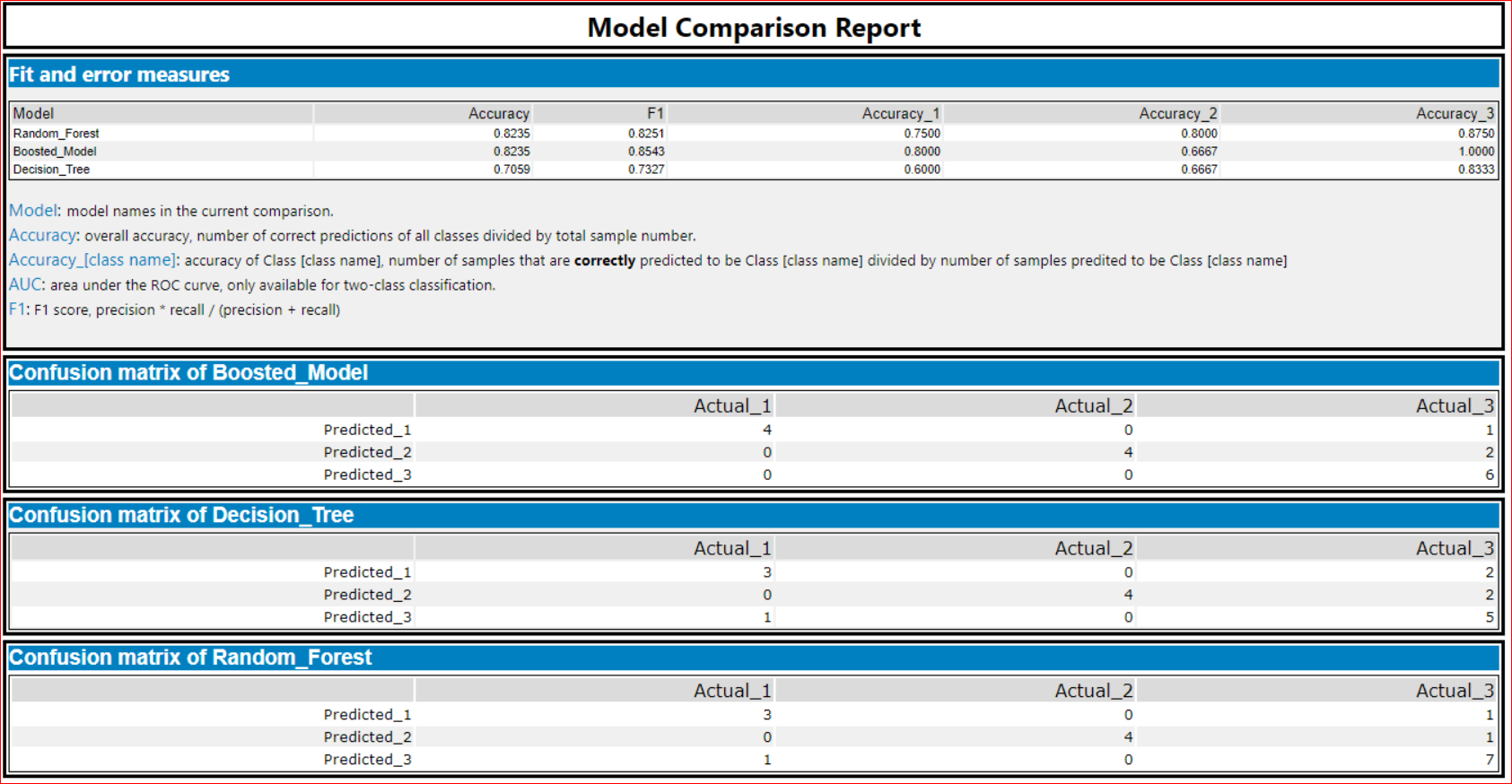
**Task 2 Submission**

* What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology?
* What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.
* What format do each of the 10 new stores fall into? Please provide a data table.

Task 2: Formats for New Stores

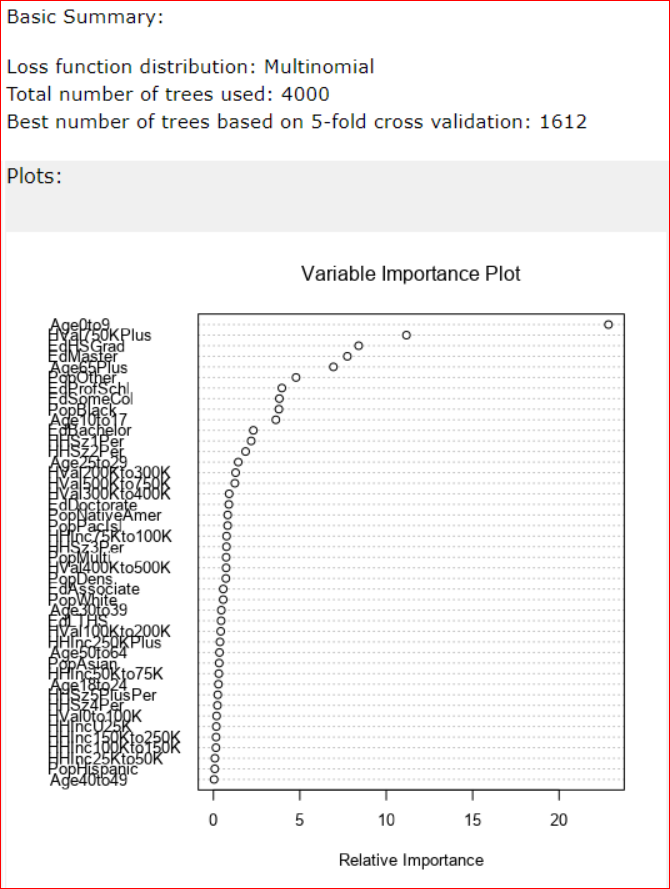
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

In any classification problem you will need to set an estimation sample 80 and a validation sample 20 of your data. This helps us compare different classification models to see which better fit the data. The comparison result made me choose boosted model have the best results in Accuracy = 0.8235, F1= 0.8543 and Accuracy\_1= 0.8000



Model Comparison Report, Boosted Model is appropriate to use for lassifying format of the new stores

**Report for Boosted Model Boosted\_Model**



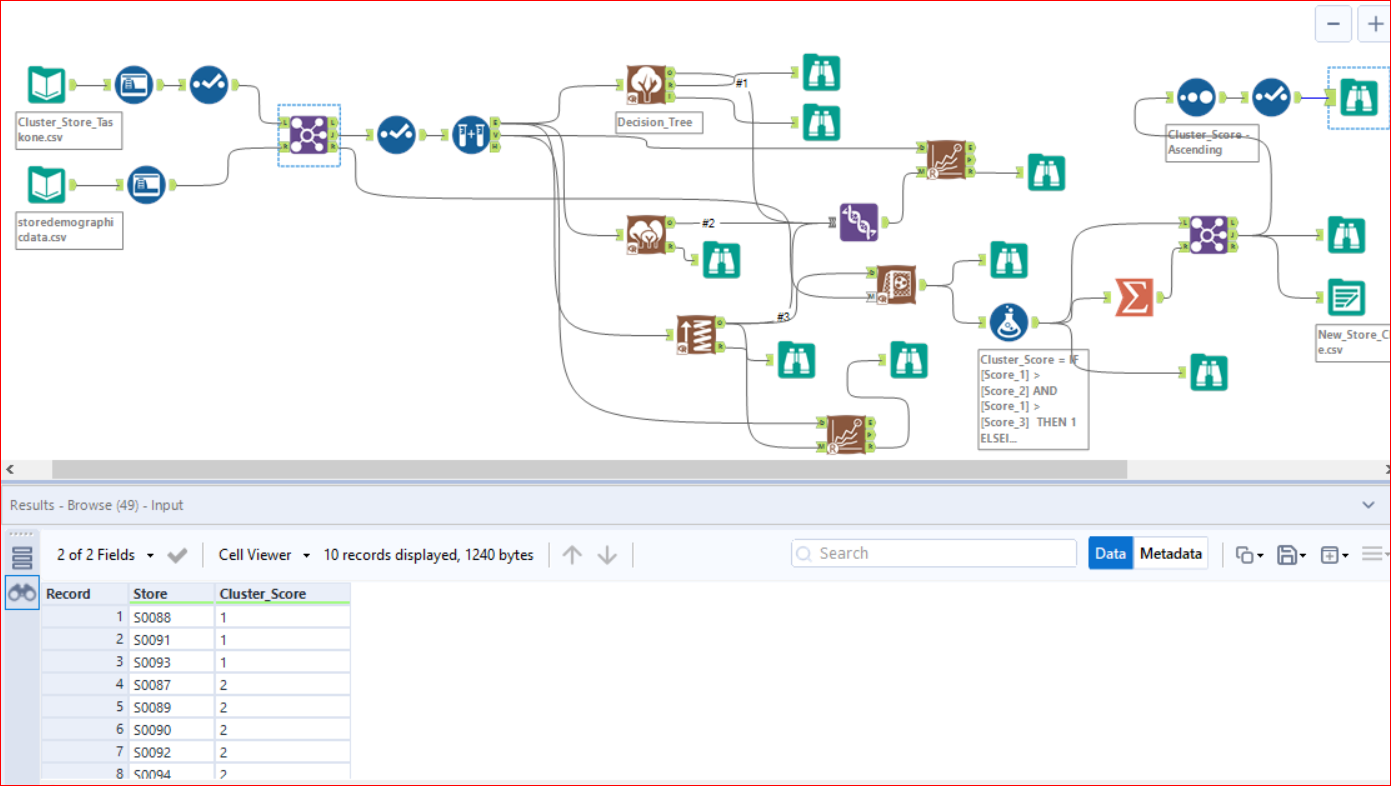
The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum to 100, and the value for each field gives the relative percentage importance of that field to the overall model.

*Ave0to9* , *HVal750KPlus* and *EdHSGrad* are the three most important variables.

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

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

**Workflow Task 2:**



**Task 3: Forecasting**

Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast.

**Task 3: Forecasting Produce Sales**

You’ve been asked to prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores. To do so, follow the steps below.

**Note:** Use a 6 month holdout sample for the TS Compare tool (this is because we do not have that much data so using a 12 month holdout would remove too much of the data)

**Step 1:** To forecast produce sales for existing stores you should aggregate produce sales across all stores by month and create a forecast.

**Step 2:** To forecast produce sales for new stores:

* Forecast **produce sales (not total sales)** for the average store (rather than the aggregate) for each segment.
* Multiply the average store produce sales forecast by the number of new stores in that segment.
* For example, if the forecasted average store produce sales for segment 1 for March is 10,000, and there are 4 new stores in segment 1, the forecast for the new stores in segment 1 would be 40,000.
* Sum the new stores produce sales forecasts for each of the segments to get the forecast for all new stores.

**Step 3:** Sum the forecasts of the existing and new stores together for the total produce sales forecast.

## 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?

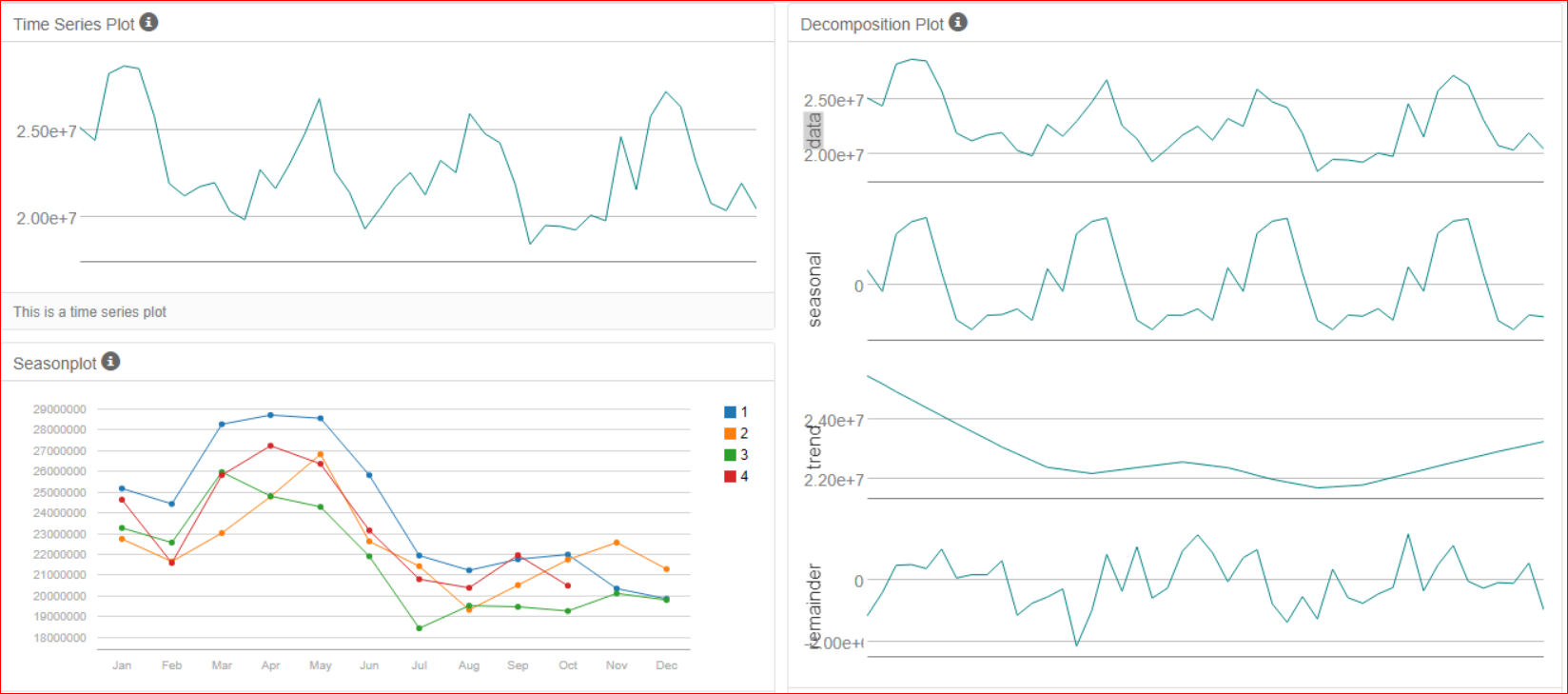
The decision to use ETS or ARIMA model can be clarified with Time Series model with a holdout sample of 12 months.

Based on the decomposition plot, our ETS(M,N,M) models shows the following:

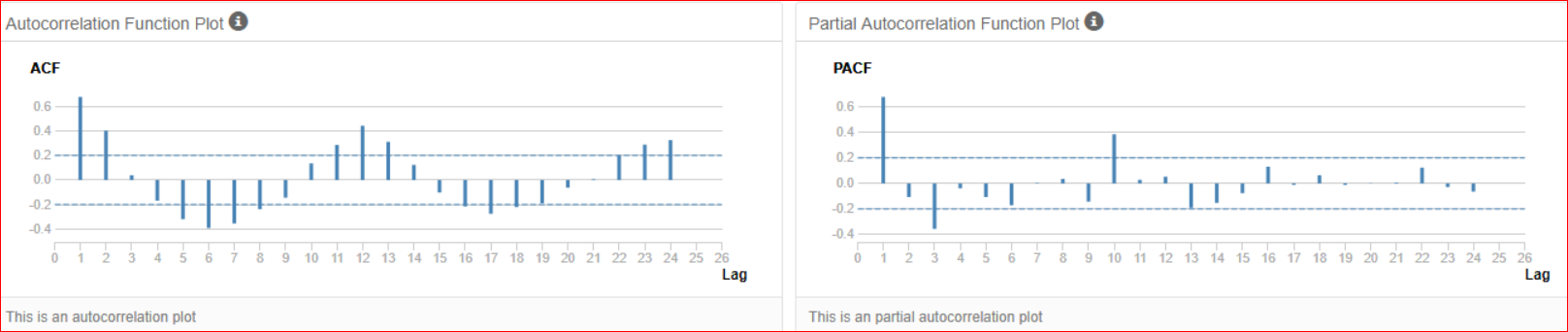
(1) The seasonality has an increasing trend and multiplicative as the peaks change over time.

(2) The trend is zero as the trend seems inconsistent.

(3) The error is irregular and multiplicative since the errors are abruptly growing and shrinking over time.

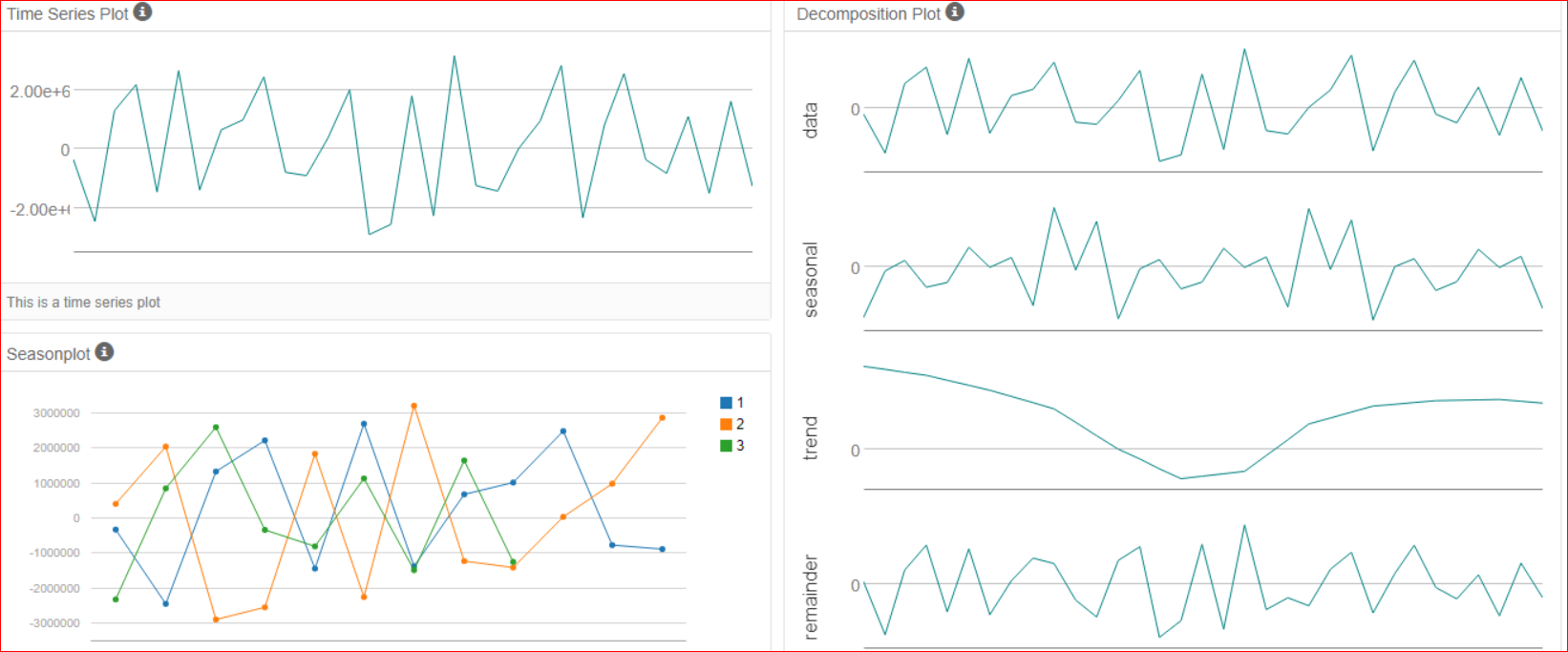
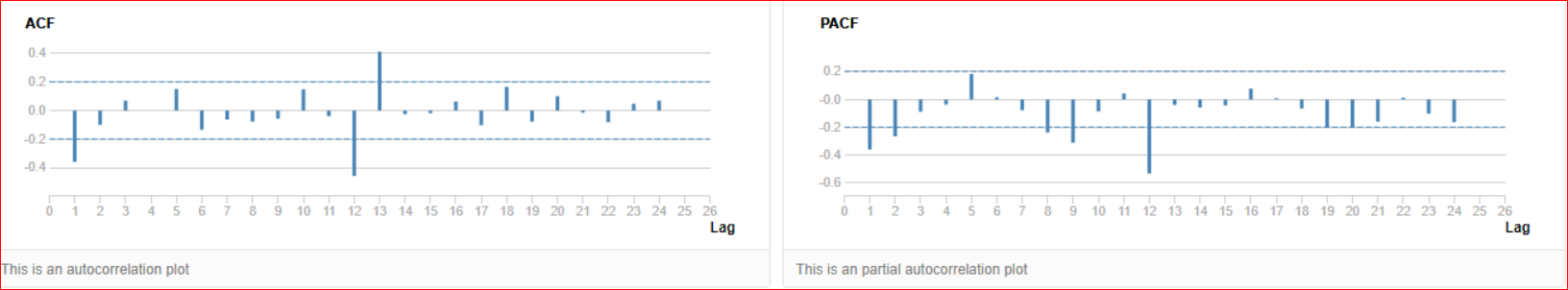


ETS(M,N,M) with no dampening is used for ETS model.



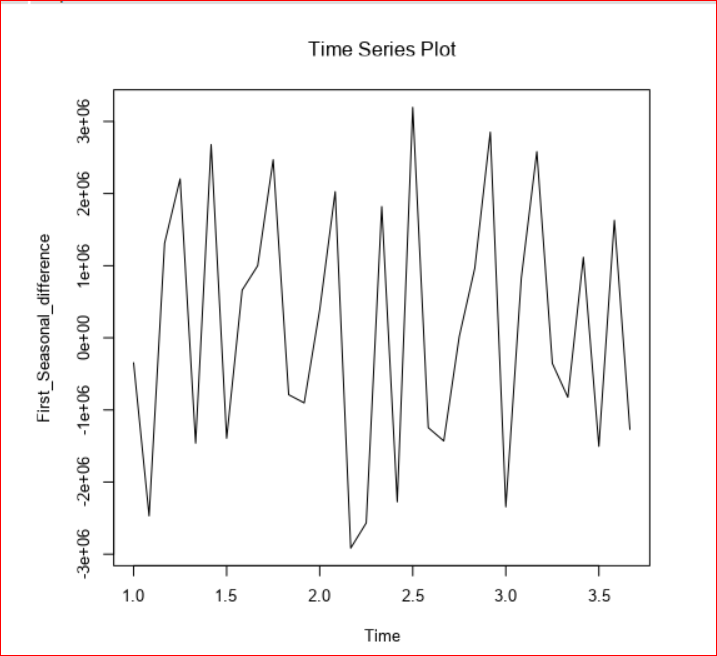
Autocorrelation plots after adding the AR term in the seasonal part of the model.

Determine Trend, Seasonal, and Error components**:**



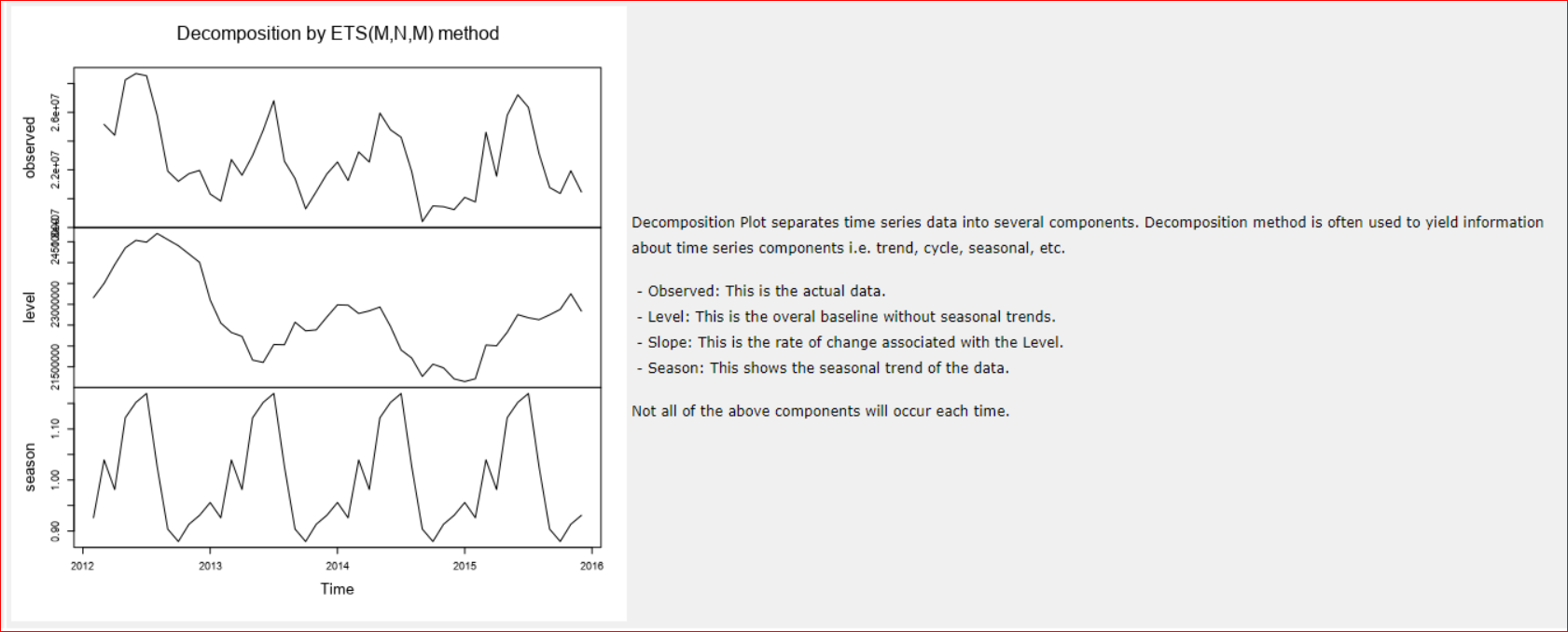
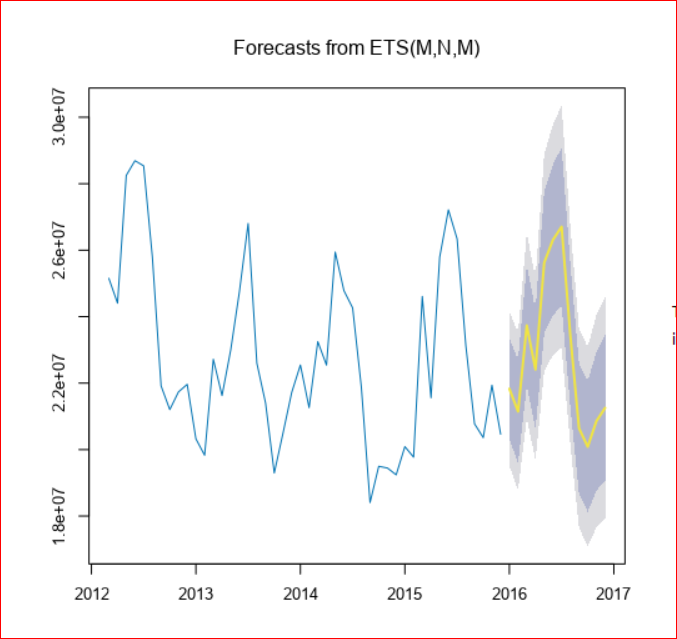
Autocorrelation plots after applying the ARIMA(1,0,0)(1,1,0)[12] model.

The decomposition plot shows our time series broken down into its three components: trend seasonal and the error. Each of these components makes up our time series and helps us confirm what we saw in our initial time series plot

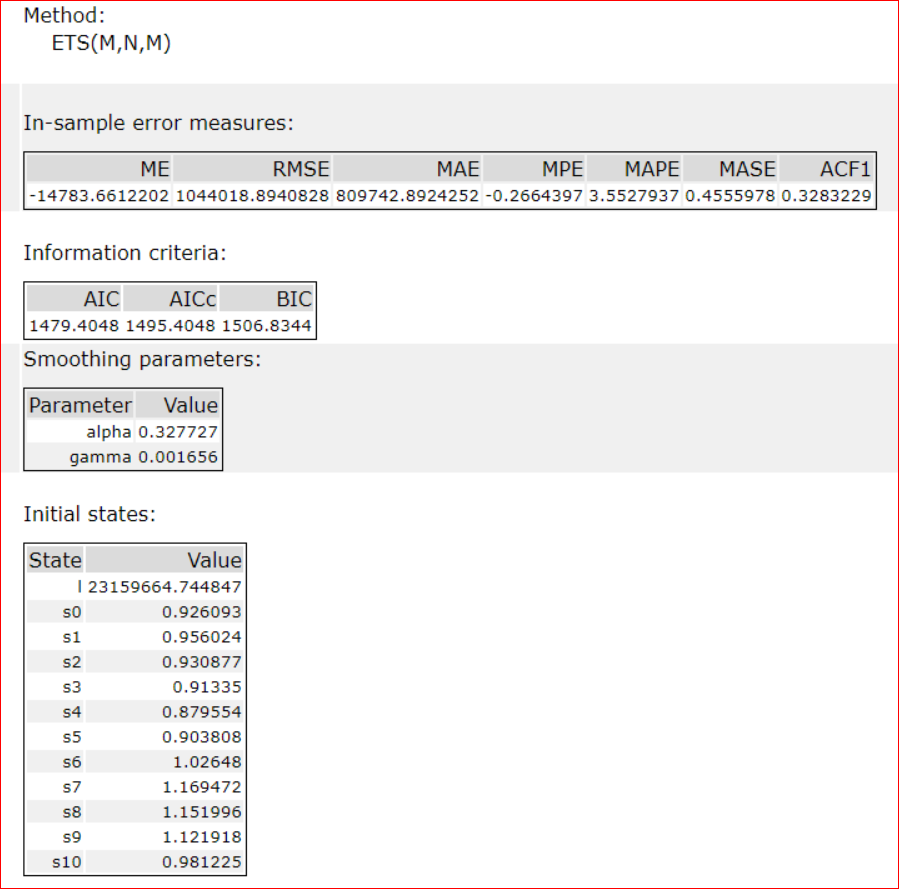


|  |
| --- |
| Plots of Time Series Exponential Smoothing Model ETS1 |
| In statistics, a time series is a sequence of data points measured at successive points in time spaced at uniform intervals. Examples of time series are the daily closing value of a stock market index or the annual flow volume of a river. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. |

The Forecast Plot shows the historic data in black and the expected value in blue. The orange in the plot shows the 90% confidence interval, and the yellow shows the 95% confidence interval.

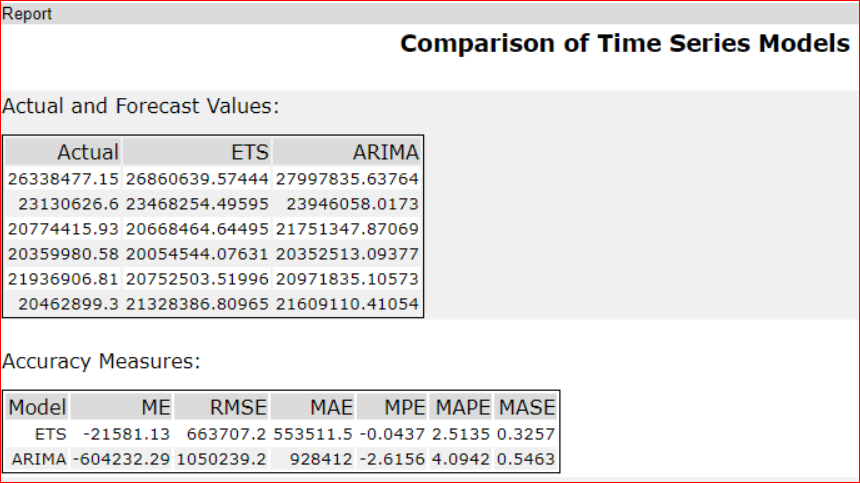
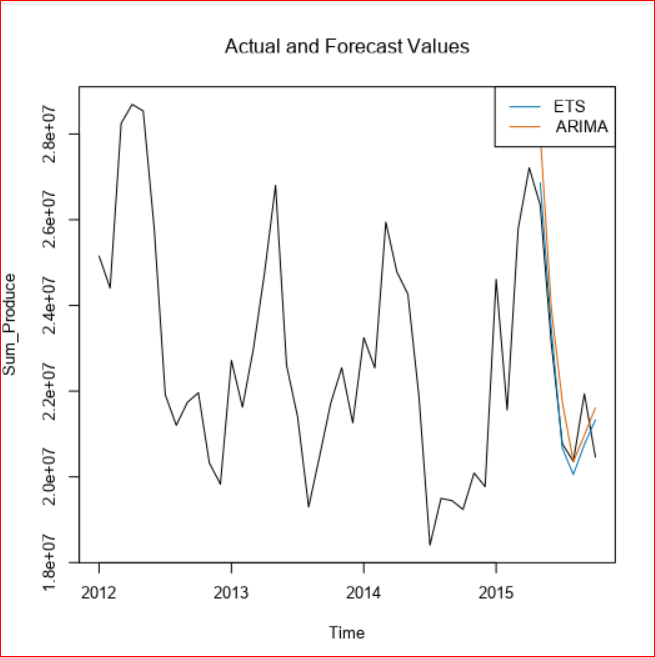


**Summary of Time Series Exponential Smoothing Model ETS1**



When fitting a forecasting model we can use a series of identifiers that help us choose the best model

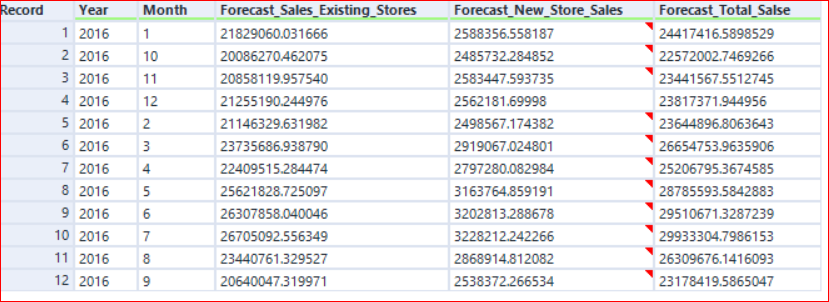
**Comparison** ARIMA/ETS



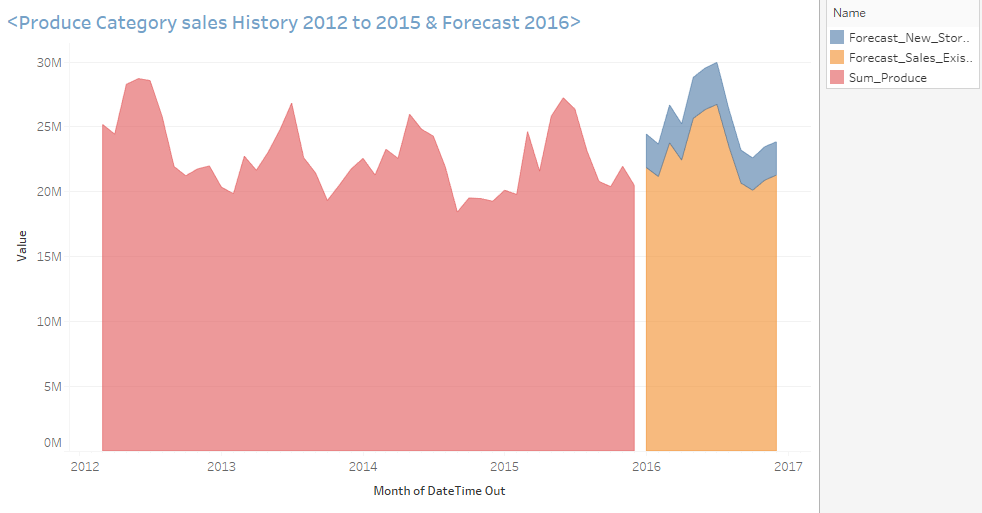
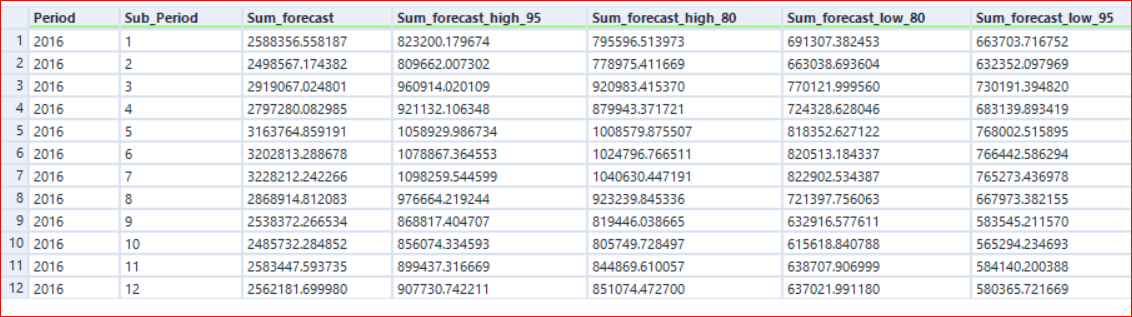
The model for each forecast was ETS(M,N,M), which compared to ARIMA(1,0,0)(1,1,0)[12] in terms of forecast error measurements against the holdout sample obtained from the TS Compare tool, the results were better (especially lower RMSE and MASE as indicated in Tables 4 for ETS model and 5 for ARIMA model). Root Mean Squared Error (RMSE) represents the sample standard deviation of the differences between predicted values and observed values. Mean Absolute Scaled Error (MASE) is defined as the mean absolute error of the model divided by the mean absolute value of the first difference of the series.

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.

The forecasted values for produce, monthly in 2016 for new and existing stores, table down shows the historical data together with these forecasts.

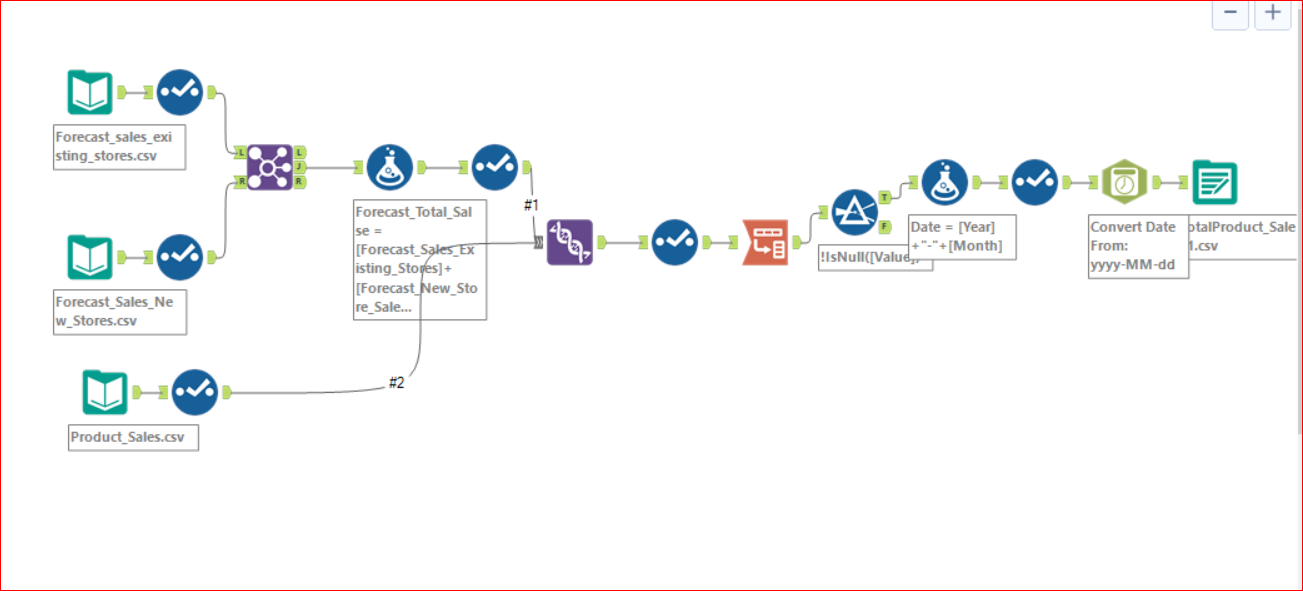
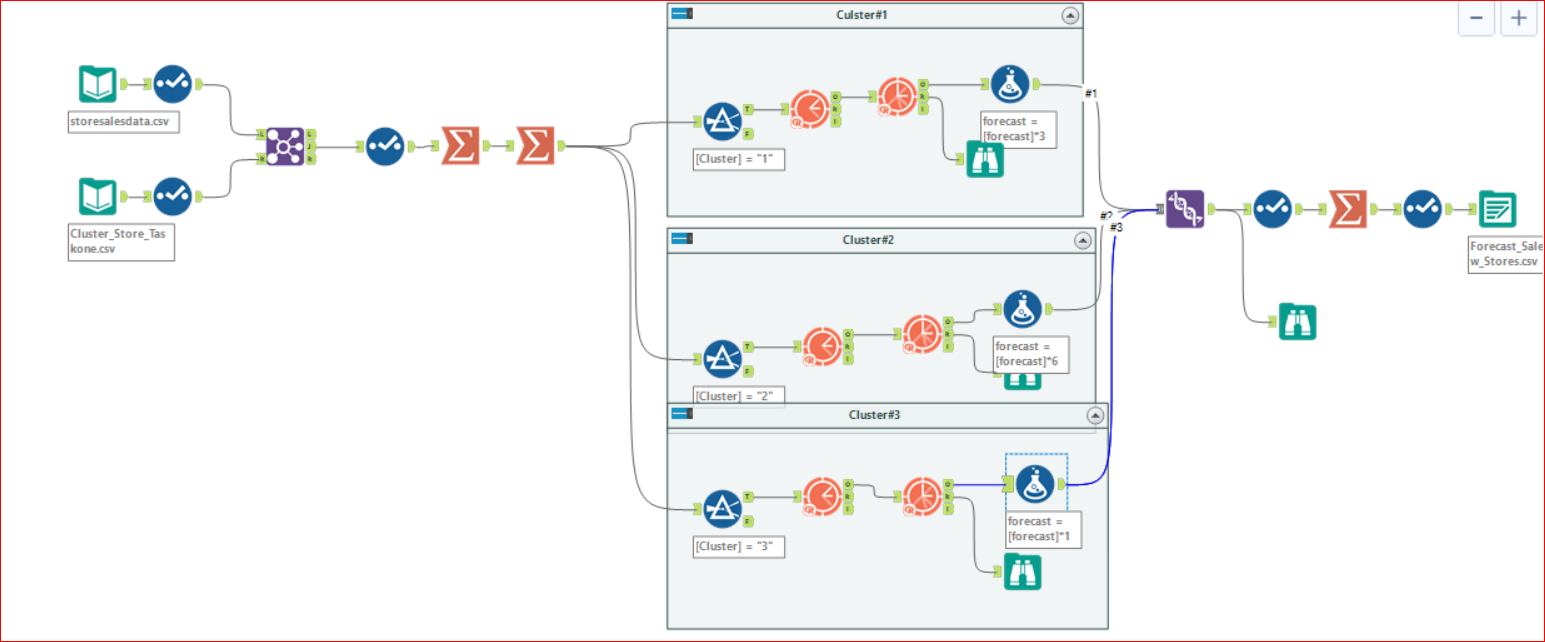
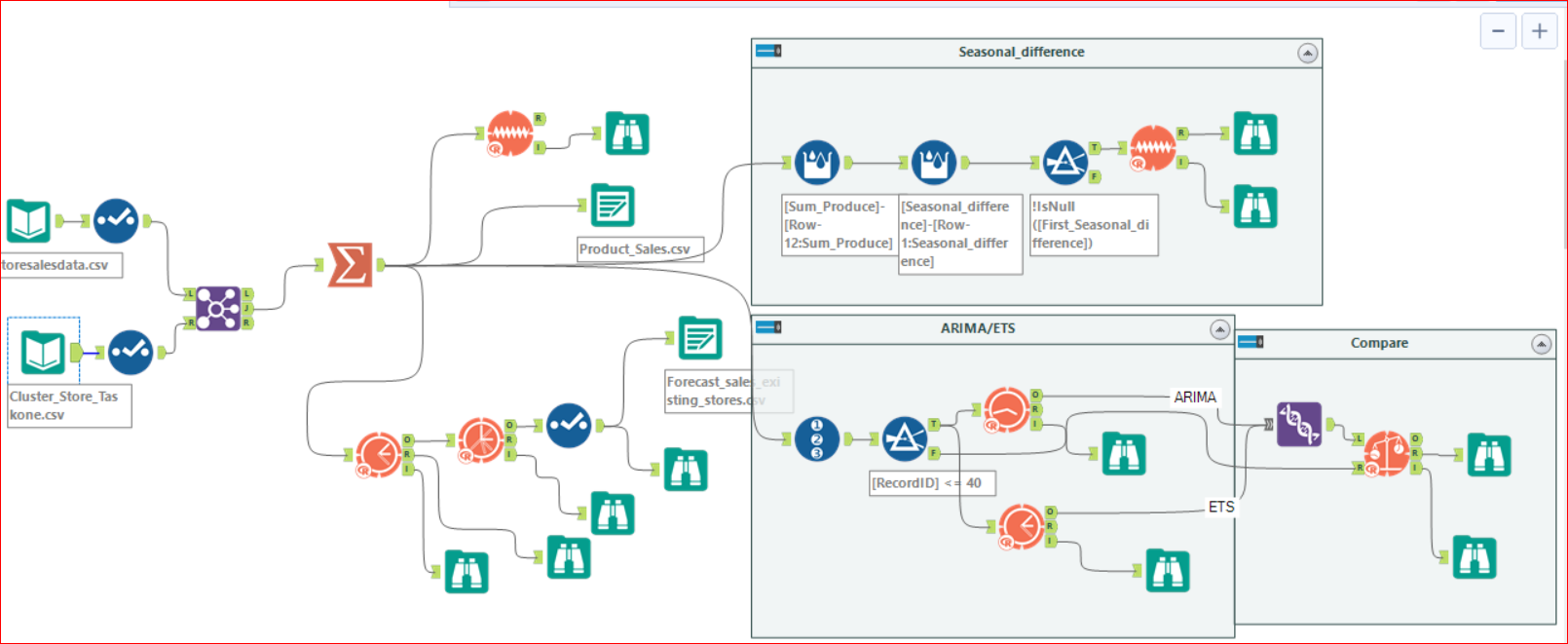


Forecast results using 95% and 80% confidence intervals:



Produce values for existing stores from 2012 to 2015 and forecasted produce values for existing and new stores monthly in 2016.

**Workflow Task3**



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.

*I hope to be home to the project requirements despite the valuable information that we learned from the lessons and also the project, but we do not know the exact correct result*

Help resources Forums: <https://knowledge.udacity.com>

**I wish success to all.**

M*arwan Saeed* *Alsharabbi*