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

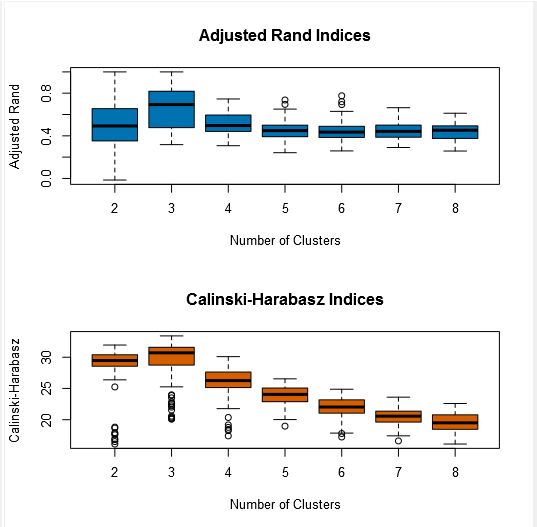
* Determine the optimal number of store formats based on sales data.
  + Use percentage sales per category per store for clustering (category sales as a percentage of total store sales).
  + Use only 2015 sales data.
  + Use a K-means clustering model.
* Segment the 85 current stores into the different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

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

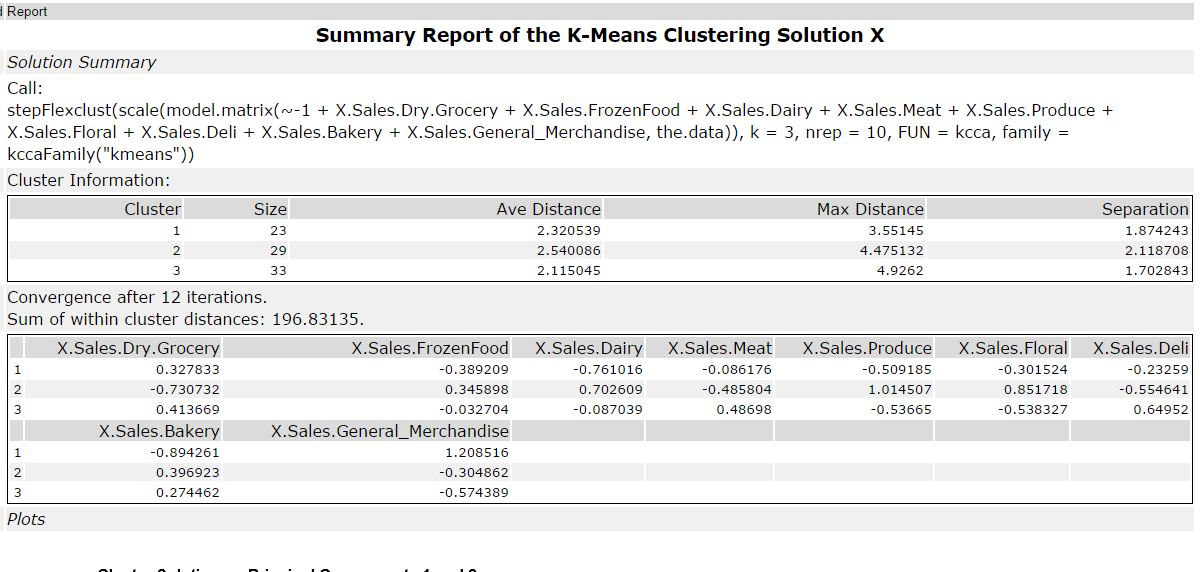
Using K-mean clustering model , using year 2015 sales data, we can see that it is suggested that the optimal number of store format would **3 different formats**.

We can see from K-mean centroid diagnostics, Adjusted Rand Indices and Calinski Harabasz indices shows highest mean and average in 3 cluster.

Therefore , we will select 3 as optimal number of store format



1. How many stores fall into each store format?



23 Stores Falls into format 1

29 Stores Falls into format 2

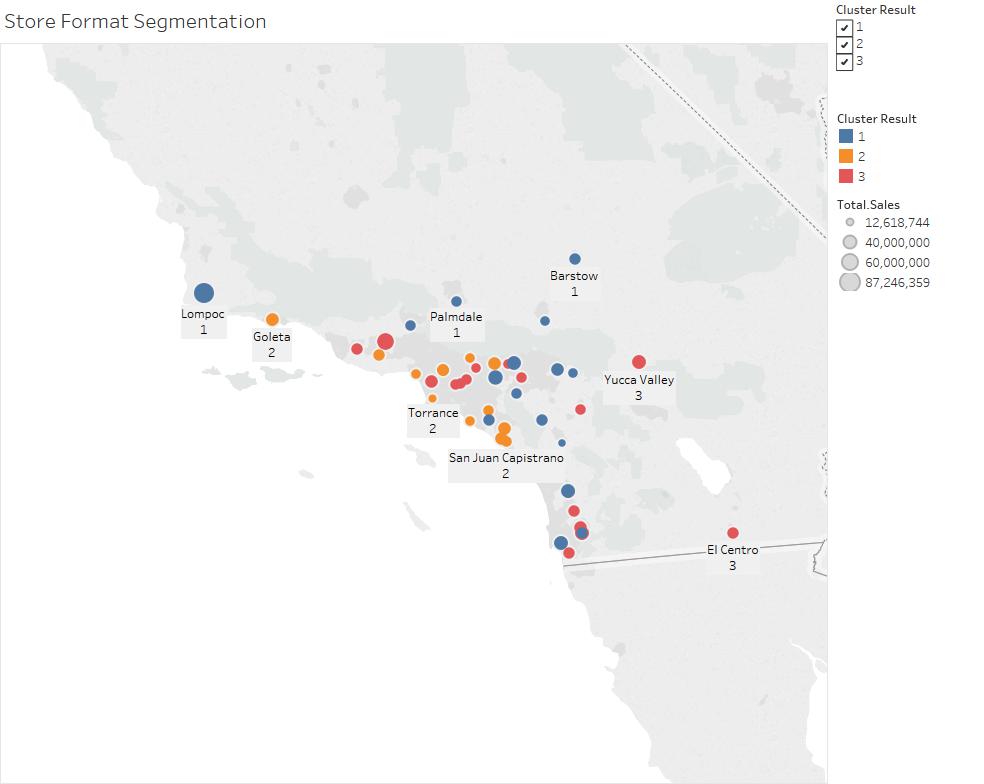
33 Stores Falls Into format 3

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

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Format | Dry Grocery | Frozen Food | Dairy | Meat | Produce | Floral | Deli | Bakery | G.Merchandise |
| 1 | High | High | Low | Med | Low | Low | Low | Low | High |
| 2 | Low | Low | High | Low | High | High | Low | High | Low |
| 3 | High | Low | Med | High | Low | Low | High | High | Low |

I can generalize 3 types format into, Dry, Frozen &General Merchandise focus, Fresh n Flowery Focus, the 3 one is Meat & Eat.

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.Task 1 Guide

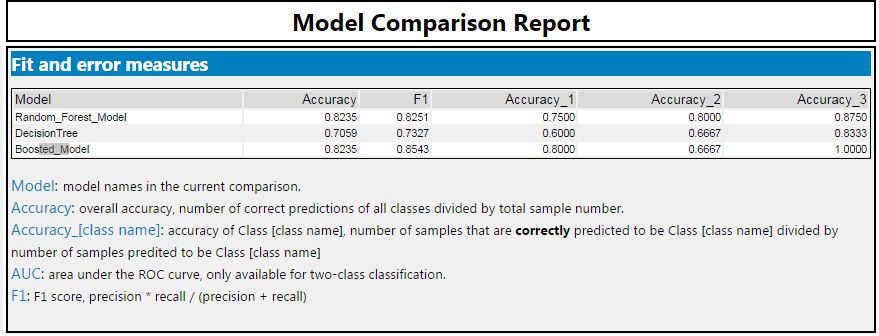


Based on the location we can see Cluster 2 Fresh and Flowery locates in more coastal area where Cluster 1 Dry, Frozen & General Merchandising more locates in the inner area, while the rest is in cluster 3 Meat & Eat

## Task 2: Formats for New Stores

* 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 the each of 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.

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.)



Based on Model Comparison report, Random Forest Model and Boosted Model Has same accuracy in predicting the clusters, but in the end I will choose Boosted Model Since it has higher F1 score compared by random forest model

.

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

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

## Task 3: Predicting Produce Sales

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

Step 2: To forecast sales for new stores:

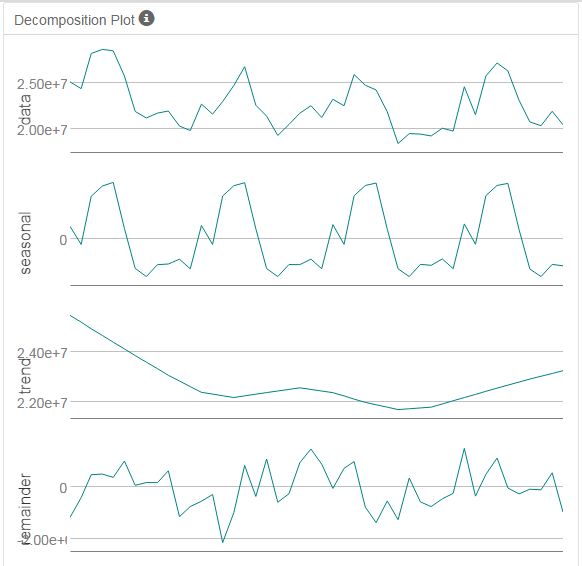
* Forecast produce sales for the average store (rather than the aggregate) for each segment.
* Multiply the average store sales forecast by the number of new stores in that segment.
* For example, if the forecasted average store sales for segment 1 for March is 10,000, and there are 4 new stores in segment 1, the forecast for the new stores in segment 1 would be 40,000.
* Sum the new stores sales forecasts for each of the segments to get the forecast for all new stores.

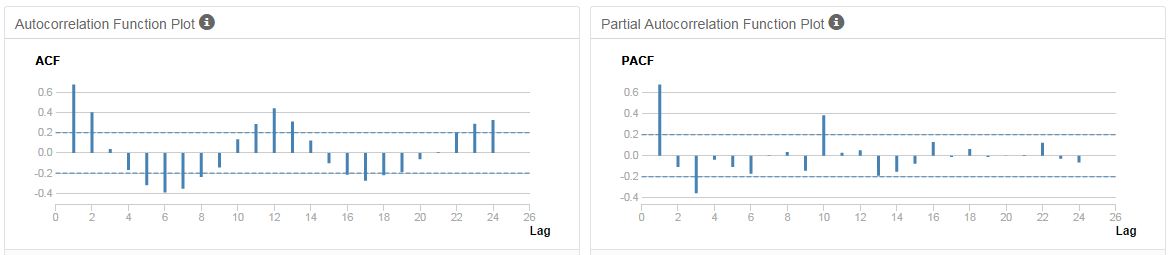
Step 3: Sum the forecasts of the existing and new stores together for the total produce sales forecast.

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?

2. Please provide a Tableau Dashboard (saved as a Tableau Public file) that includes a table and a plot of the three monthly forecasts; one for existing, one for new, and one for all stores. Please name the tab in the Tableau file "Task 3".

**ETS Decomposition Plot**

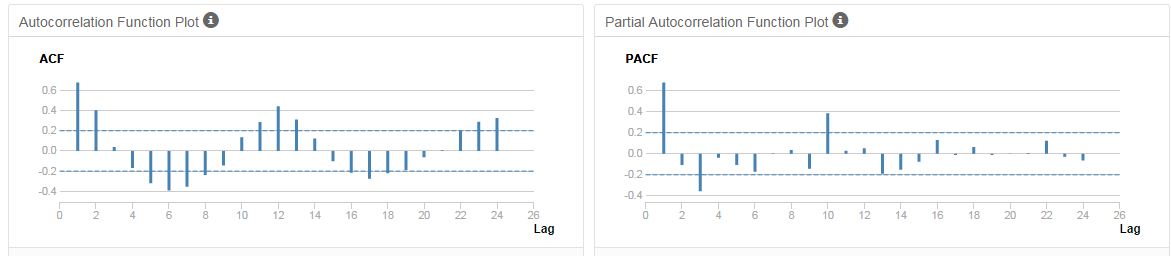




Based on the decomposition plot, the seasonal are very slightly decreasing over time, there is no trend, and increasing in error. This would suggest a ETS (M,N,M) model

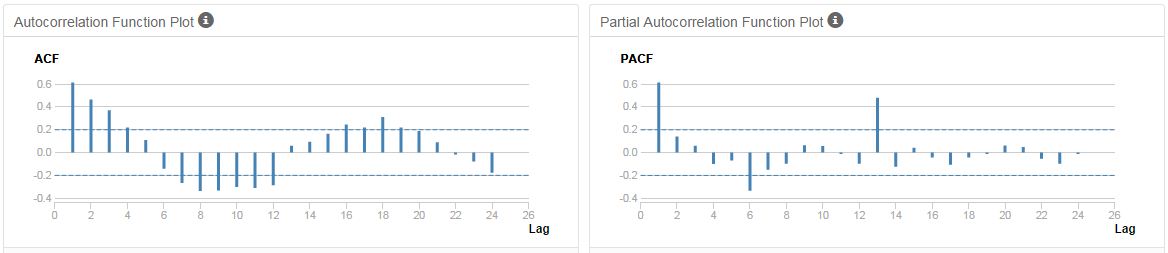
**ARIMA Method Result**

**Time Series ACF & PACF**



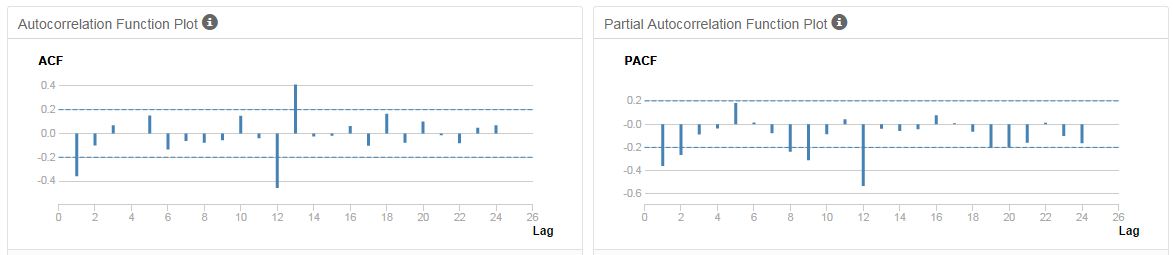
We should be able to see that the ACF presents slowly decaying serial correlations towards 0 with increases at the seasonal lags. Since serial correlation is high we will need to seasonally difference the series

**Seasonal Difference ACF & PACF**

****

We can see that the seasonal difference presents similar ACF and PACF results as the initial plots without differencing, only slightly less correlated. In order to remove correlation we will need to difference further.

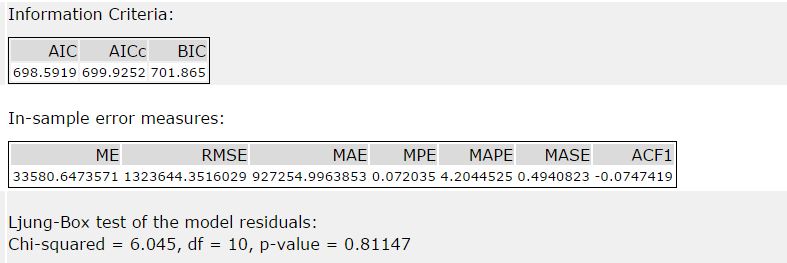
**Seasonal First Difference ACF & PACF**

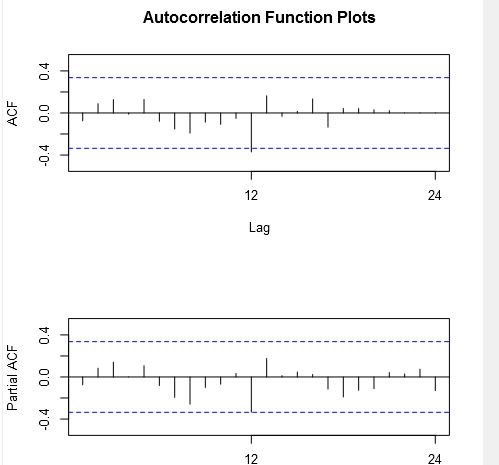


The seasonal lags (lag 12, 24, etc.) in the ACF and PACF do not have any significant correlation so there will be no need for seasonal autoregressive or moving average terms. This means the P & Q would be zero. And since the know that the forecast is monthly, we found that m would be 12.

Suggested ARIMA model : ARIMA(1,0,0)(0,1,0)[12]

**ARIMA(1,0,0)(0,1,0)[12] Result**

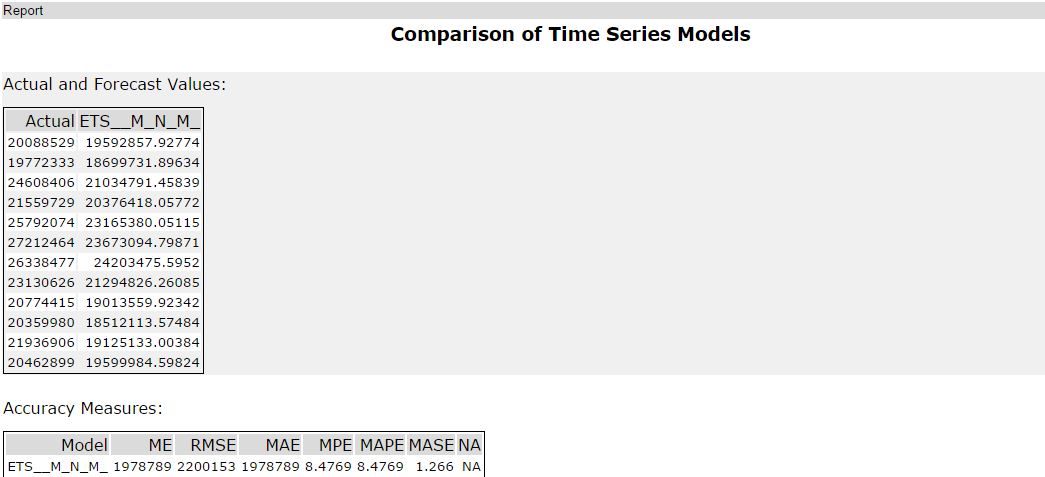
****



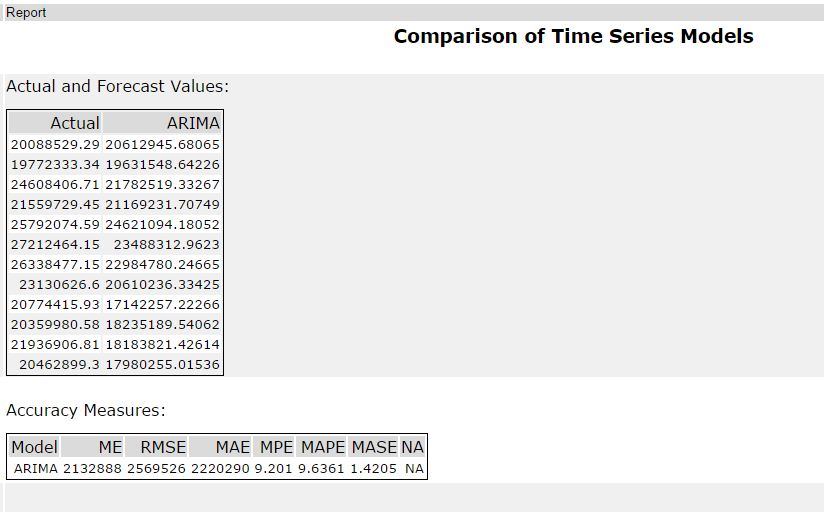
The ACF and PACF results for the ARIMA (0,1,1) (0,1,0) 12 model show no significantly correlated lags suggesting no need for adding additional AR() or MA() terms

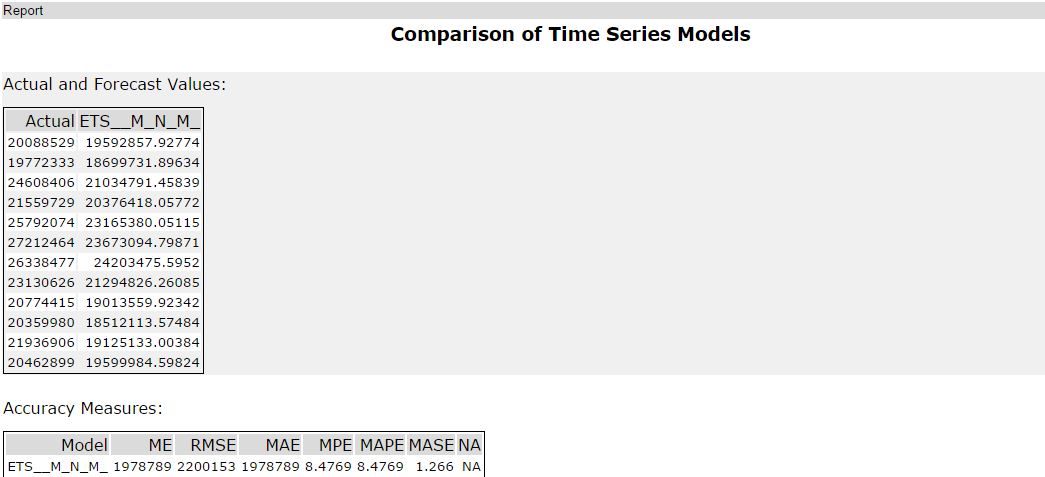
**ETS (M,N,M) & ARIMA (0,1,1) (0,1,0) 12 TS comparison result**

**ETS M,N,M TS comparison**



**ARIMA TS (0,1,1) (0,1,0) 12 comparison**



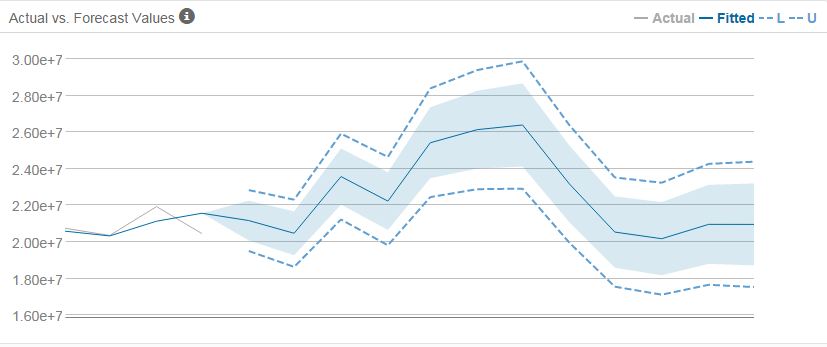


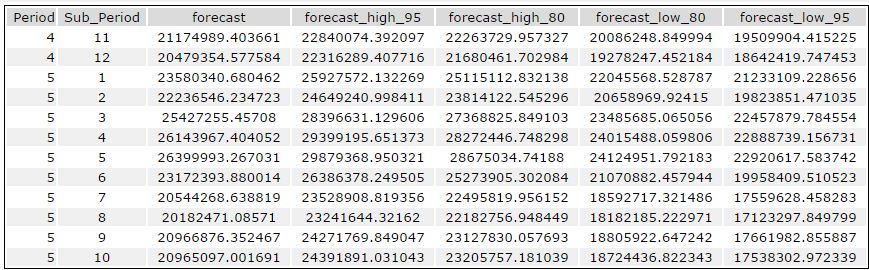
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **ME** | **RMSE** | **MAE** | **MPE** | **MAPE** | **MASE** |
| ETS\_\_M\_N\_M\_ | 1978789 | 2200153 | 1978789 | 8.4769 | 8.4769 | 1.266 |
| ARIMA | 2132888 | 2569526 | 2220290 | 9.201 | 9.6361 | 1.4205 |

Based on TS comparison on both model, we can see the result for both model are suite similar, but ETS (M,N,M) has the overall lower errors across all the variables. Mean Error Average Percentage Error (MPE) for ETS is less at 8.47% while the ARIMA model gives 9.20% errors. And Mean Absolute Scale Error (MASE) for ETS is also lower at 1.266 while the ARIMA models gives 1.42 errors. Ideally we should look for model that gives error less than MASE 1.0, that can be done by digging for more data, but since we only have 3-4 years monthly sales data, so this model is good enough.

And in this case, for the forecasting we should use the **ETS model** as our forecasting model.

**ETS Forecast result**





Final Sales Forecast – Existing Stores

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Period | forecast | forecast\_high\_95 | forecast\_high\_80 | forecast\_low\_80 | forecast\_low\_95 |
| Jan-16 | $21,174,989 | $22,840,074 | $22,263,730 | $20,086,249 | $19,509,904 |
| Feb-16 | $20,479,355 | $22,316,289 | $21,680,462 | $19,278,247 | $18,642,420 |
| Mar-16 | $23,580,341 | $25,927,572 | $25,115,113 | $22,045,569 | $21,233,109 |
| Apr-16 | $22,236,546 | $24,649,241 | $23,814,123 | $20,658,970 | $19,823,851 |
| May-16 | $25,427,255 | $28,396,631 | $27,368,826 | $23,485,685 | $22,457,880 |
| Jun-16 | $26,143,967 | $29,399,196 | $28,272,447 | $24,015,488 | $22,888,739 |
| Jul-16 | $26,399,993 | $29,879,369 | $28,675,035 | $24,124,952 | $22,920,618 |
| Aug-16 | $23,172,394 | $26,386,378 | $25,273,905 | $21,070,882 | $19,958,410 |
| Sep-16 | $20,544,269 | $23,528,909 | $22,495,820 | $18,592,717 | $17,559,628 |
| Oct-16 | $20,182,471 | $23,241,644 | $22,182,757 | $18,182,185 | $17,123,298 |
| Nov-16 | $20,966,876 | $24,271,770 | $23,127,830 | $18,805,923 | $17,661,983 |
| Dec-16 | $20,965,097 | $24,391,891 | $23,205,757 | $18,724,437 | $17,538,303 |

Sales forecast for new stores

since all new stores locates in the cluster 1, we will filter monthly sales from all cluster 1 stores and aggregates it into sum of sales.

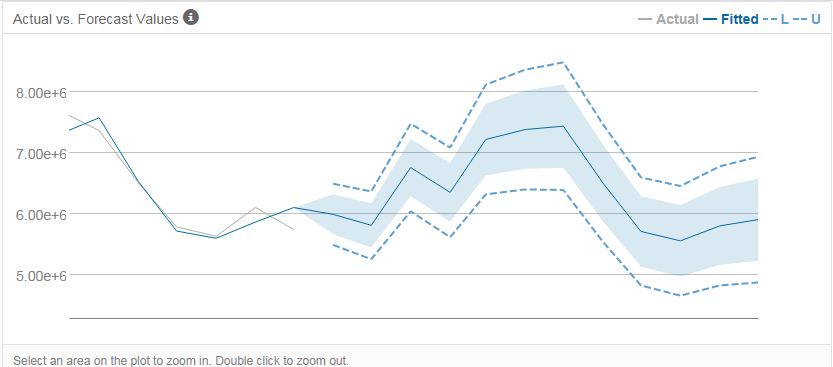
Step 1. Aggregate monthly sales of stores in cluster 1

Step 2. Conduct ETS model forecast for next 12 months

Step 3. Averaged it by number of cluster 1 existing stores

Step 4. Multiply it by number of new stores (10)

**ETS Cluster 1 Forecast Result**



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sub\_Period | forecast | forecast\_high\_95 | forecast\_high\_80 | forecast\_low\_80 | forecast\_low\_95 |
| Jan-16 | $5,993,193 | $6,496,568 | $6,322,332 | $5,664,054 | $5,489,818 |
| Feb-16 | $5,812,153 | $6,369,475 | $6,176,566 | $5,447,740 | $5,254,831 |
| Mar-16 | $6,763,382 | $7,483,488 | $7,234,234 | $6,292,530 | $6,043,277 |
| Apr-16 | $6,354,144 | $7,091,847 | $6,836,502 | $5,871,786 | $5,616,442 |
| May-16 | $7,224,490 | $8,127,480 | $7,814,923 | $6,634,057 | $6,321,501 |
| Jun-16 | $7,388,125 | $8,372,924 | $8,032,050 | $6,744,200 | $6,403,327 |
| Jul-16 | $7,443,989 | $8,494,453 | $8,130,851 | $6,757,127 | $6,393,524 |
| Aug-16 | $6,510,524 | $7,477,529 | $7,142,815 | $5,878,233 | $5,543,519 |
| Sep-16 | $5,711,427 | $6,600,092 | $6,292,494 | $5,130,359 | $4,822,761 |
| Oct-16 | $5,556,692 | $6,458,852 | $6,146,583 | $4,966,801 | $4,654,532 |
| Nov-16 | $5,802,982 | $6,782,815 | $6,443,661 | $5,162,304 | $4,823,149 |
| Dec-16 | $5,905,241 | $6,939,290 | $6,581,369 | $5,229,113 | $4,871,192 |

Since we have known that Cluster 1 consist 23 stores , we will divide forecast to 23, and then multiply the result by 10

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sub\_Period | forecast | Divide by 23 | Multiply by 10 | New Stores Forecasts |
| Jan-16 | $5,993,193 | $260,573.61 | $2,605,736.06 | $2,605,736.06 |
| Feb-16 | $5,812,153 | $252,702.31 | $2,527,023.12 | $2,527,023.12 |
| Mar-16 | $6,763,382 | $294,060.10 | $2,940,600.95 | $2,940,600.95 |
| Apr-16 | $6,354,144 | $276,267.15 | $2,762,671.47 | $2,762,671.47 |
| May-16 | $7,224,490 | $314,108.27 | $3,141,082.71 | $3,141,082.71 |
| Jun-16 | $7,388,125 | $321,222.83 | $3,212,228.33 | $3,212,228.33 |
| Jul-16 | $7,443,989 | $323,651.69 | $3,236,516.88 | $3,236,516.88 |
| Aug-16 | $6,510,524 | $283,066.26 | $2,830,662.64 | $2,830,662.64 |
| Sep-16 | $5,711,427 | $248,322.90 | $2,483,228.98 | $2,483,228.98 |
| Oct-16 | $5,556,692 | $241,595.31 | $2,415,953.08 | $2,415,953.08 |
| Nov-16 | $5,802,982 | $252,303.57 | $2,523,035.69 | $2,523,035.69 |
| Dec-16 | $5,905,241 | $256,749.60 | $2,567,496.04 | $2,567,496.04 |

Sales Forecast Summary

|  |  |  |  |
| --- | --- | --- | --- |
| **Period** | **Existing Stores** | **New Stores** | **All Stores** |
| 16-Jan | $21,174,989 | $2,605,736.06 | $23,780,725 |
| 16-Feb | $20,479,355 | $2,527,023.12 | $23,006,378 |
| 16-Mar | $23,580,341 | $2,940,600.95 | $26,520,942 |
| 16-Apr | $22,236,546 | $2,762,671.47 | $24,999,217 |
| 16-May | $25,427,255 | $3,141,082.71 | $28,568,338 |
| 16-Jun | $26,143,967 | $3,212,228.33 | $29,356,195 |
| 16-Jul | $26,399,993 | $3,236,516.88 | $29,636,510 |
| 16-Aug | $23,172,394 | $2,830,662.64 | $26,003,057 |
| 16-Sep | $20,544,269 | $2,483,228.98 | $23,027,498 |
| 16-Oct | $20,182,471 | $2,415,953.08 | $22,598,424 |
| 16-Nov | $20,966,876 | $2,523,035.69 | $23,489,912 |
| 16-Dec | $20,965,097 | $2,567,496.04 | $23,532,593 |

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.