Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?
   The optimal number of store formats is 3 which I arrived on basis of results of adjusted rand indices and C-H indices generated by workflow.

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.02152	0.2941	0.2633	0.2877	0.3025	0.2497	0.3043
1st Quartile	0.3111	0.6158	0.4196	0.3965	0.3724	0.3812	0.3853
Median	0.4759	0.735	0.5094	0.4631	0.4476	0.4196	0.4365
Mean	0.5062	0.7214	0.5274	0.4753	0.4599	0.4527	0.4474
3rd Quartile	0.8185	0.823	0.6197	0.53	0.5357	0.5247	0.5063
Maximum	1	0.969	0.9184	0.7502	0.6878	0.6773	0.6124

	9	10
Minimum	0.2604	0.2748
1st Quartile	0.4035	0.3728
Median	0.4503	0.4464
Mean	0.4572	0.4364
3rd Quartile	0.5157	0.4906
Maximum	0.6371	0.602

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	15.3	18.18	19.48	19.68	16.36	15.88	16.39
1st Quartile	28.26	30.33	25	23.01	21.38	20.09	18.84
Median	29.32	31.1	26.43	24.5	22.35	21.07	19.82
Mean	28.08	30.56	26.24	24.04	22.2	20.92	19.77
3rd Quartile	30.13	32.2	27.57	25.19	22.96	21.96	20.78
Maximum	31.58	33.57	30.14	26.85	24.69	24.23	23.16

	9	10
Minimum	16.21	14.33
1st Quartile	18.1	17.03
Median	18.97	17.97
Mean	19	17.87
3rd Quartile	19.92	18.83
Maximum	22.01	20.62

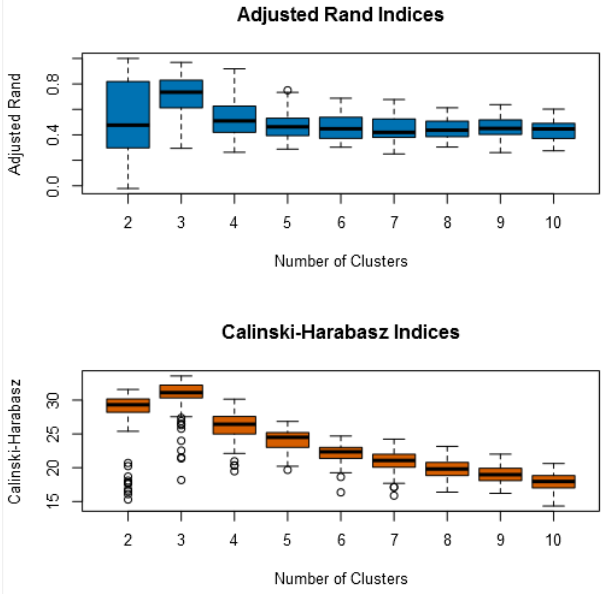
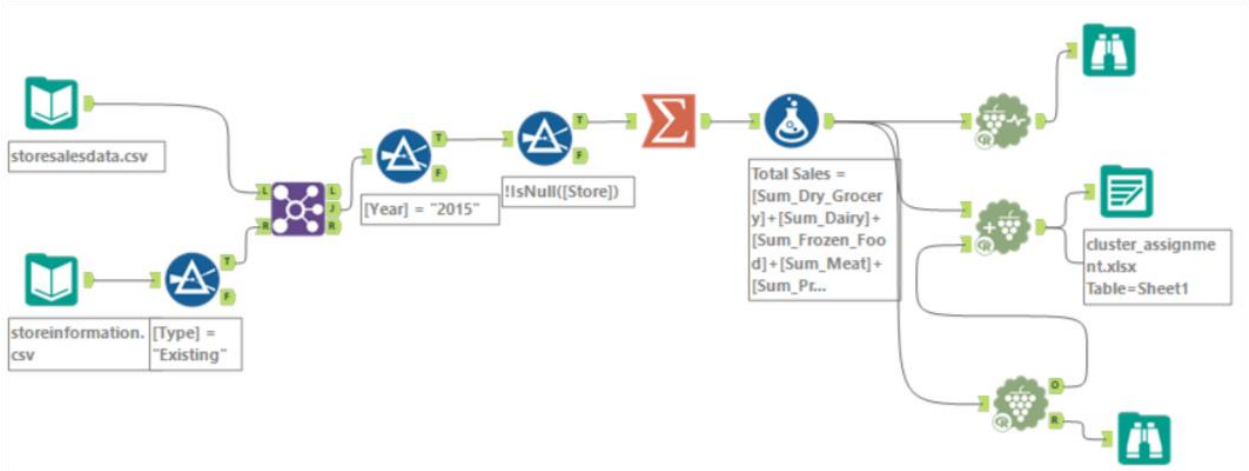


Fig. 1.1 adj rand indices & Calinski-Harabasz indices plots, which used to find optimal no of clusters.



Calculating no of clusters that should be used in K-mean clustering based on above figure the following features can be pointed out, By combining features obtained from both adj rand indices & C-H indices, cluster 3 is chosen as It has highest median value for both indices & its narrow IQR.

2. How many stores fall into each store format?
   In cluster 1, 23 existing stores. in cluster 2, 29 existing stores. In cluster 3, 33 existing stores.

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Fig 1.2 report for no if existing stores fall into each cluster

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

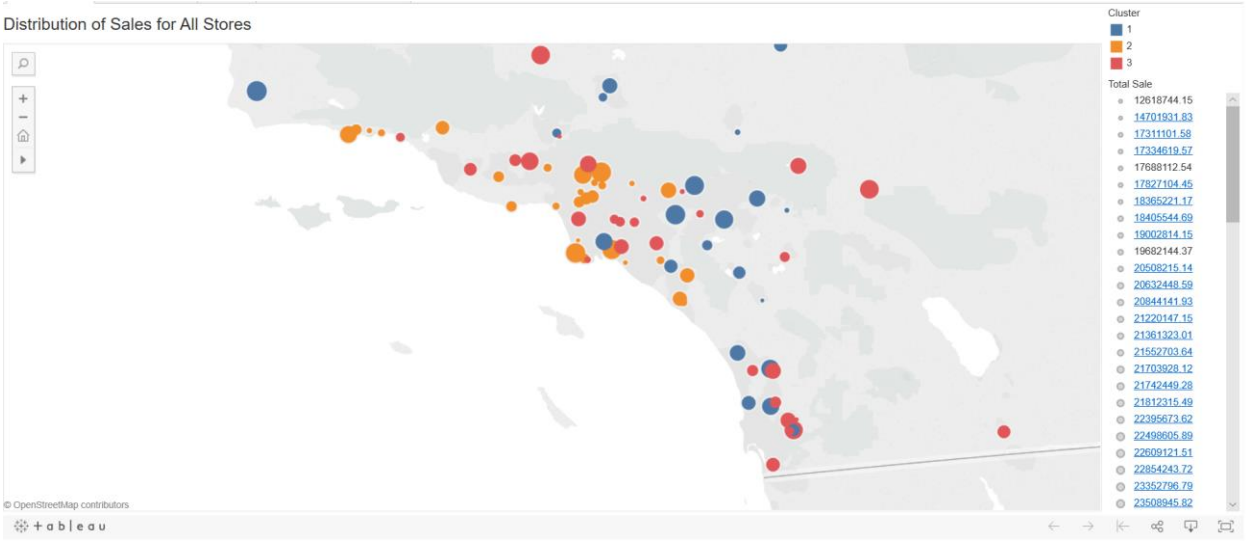


Fig 1.3 Total sakes distribution of all stores in all clusters, Cluster 1 showed in BLUE, Cluster 2 in ORANGE & Cluster 3 in GRAY.

Here circle size represents the amount of total sale for each store.

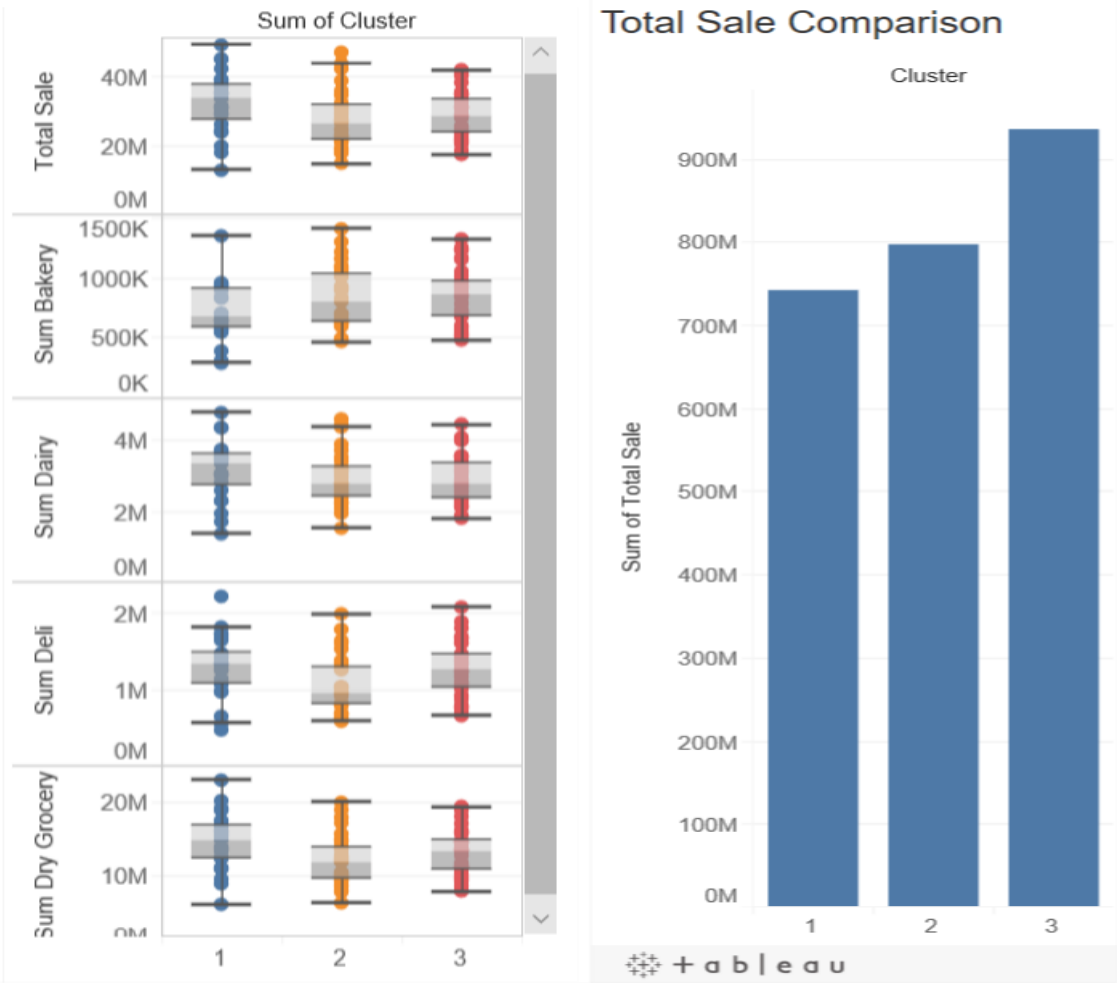


Fig. 1.4 whiskey-boxplot of total sales, sum of bakery, sum of deli, sum of dry grocery & sum of dairy for each cluster(LEFT). Bar plot shows the total no of sales of each other(RIGHT)

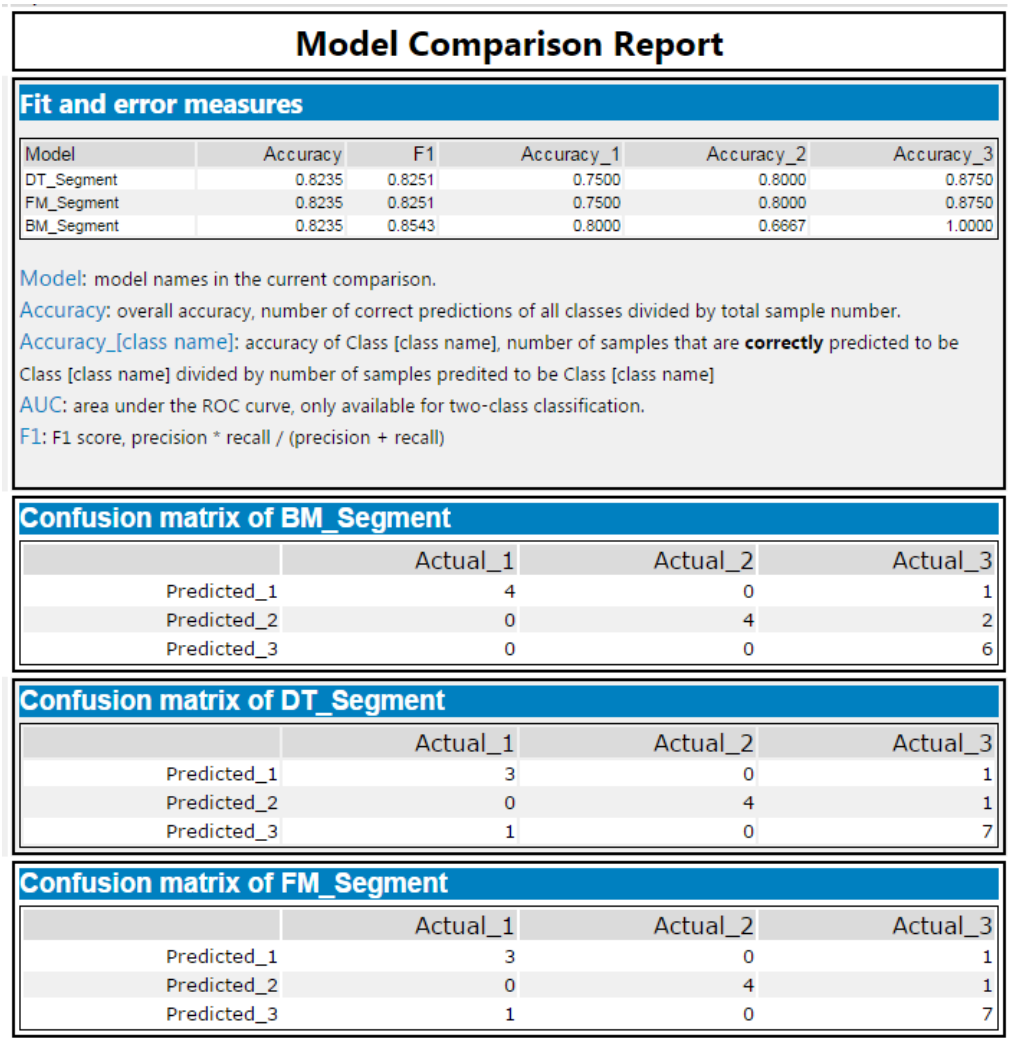
Based on fig 1.3 & 1.4, we can say that cluster 1 sold more dairy, cluster 2 sold more bakery. Cluster 3 sold more in total in comparison to other 2 clusters.

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.
   Fig 1.3
   https://public.tableau.com/profile/harsh.varudkar#!/vizhome/Project7_118/Location_sales

   Fig 1.4
   https://public.tableau.com/profile/harsh.varudkar#!/vizhome/Project7_118/Whiskey_boxplots_sumsales

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.)



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_Segment	0.8235	0.8251	0.7500	0.8000	0.8750
FM_Segment	0.8235	0.8251	0.7500	0.8000	0.8750
BM_Segment	0.8235	0.8543	0.8000	0.6667	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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of BM_Segment

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Confusion matrix of DT_Segment

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

Confusion matrix of FM_Segment

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

Fig 2.1 Model comparison between decision tree, forest model & boosted model

From the report of model comparison, Accuracy of Decision tree model is 0.8251, forest model 0.8251 and boosted model 0.8535. based on accuracy and F1 score, the boosted model can be selected as it has higher value in F1 metric. Assignment of new stores is shown in Table 2.1

From variable importance plot of forest model we can find 3 important predictor variables as follows Age0to9, HVal750K, EdHSGrad.

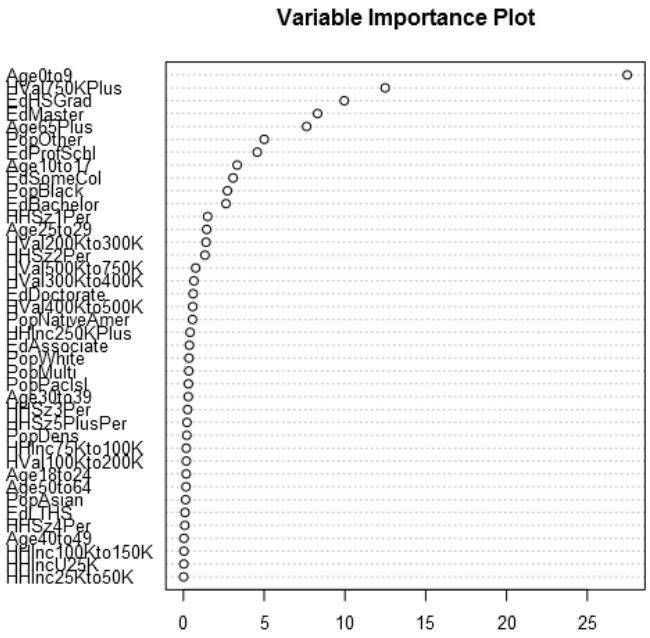


Fig. 2.2 Variable importance plot obtained by running forest model.

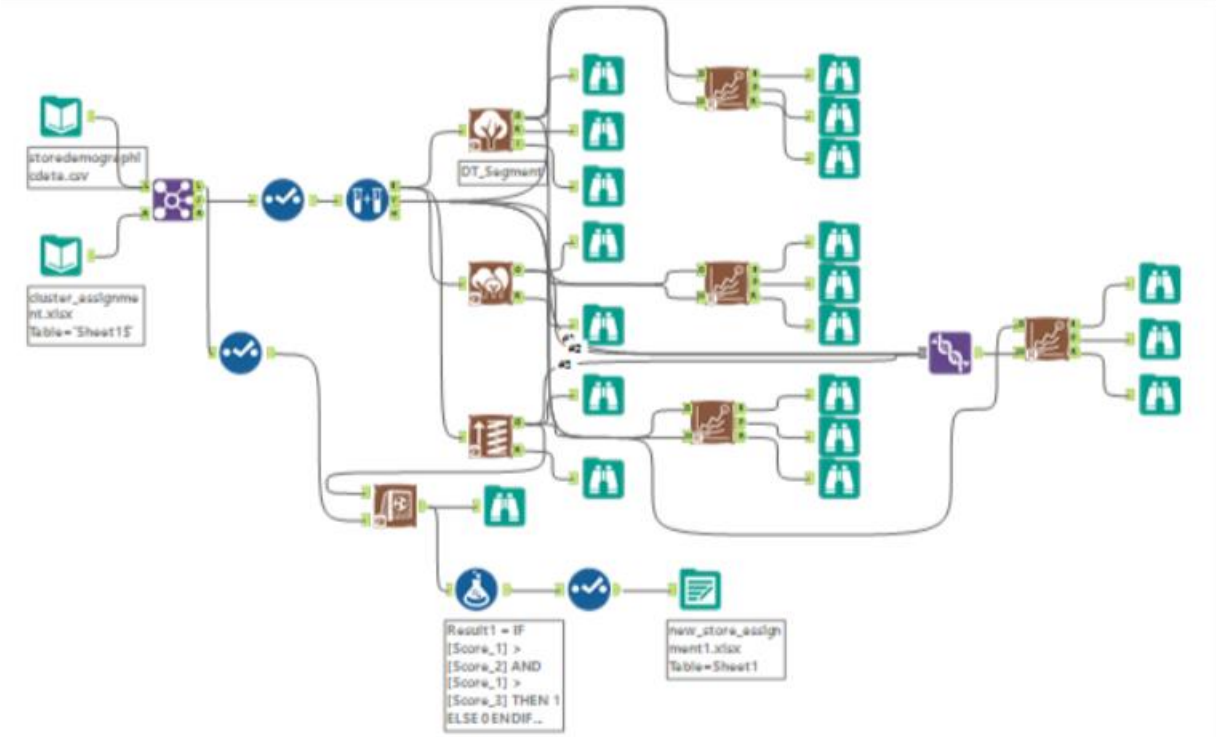


Fig 2.3 workflow used in assigning cluster for new stores.

2. 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

Table 2.1 Cluster assignment for new stores

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

Fig.2.4 Snapshot of Output data (new_store_assignment.xlsx)

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?

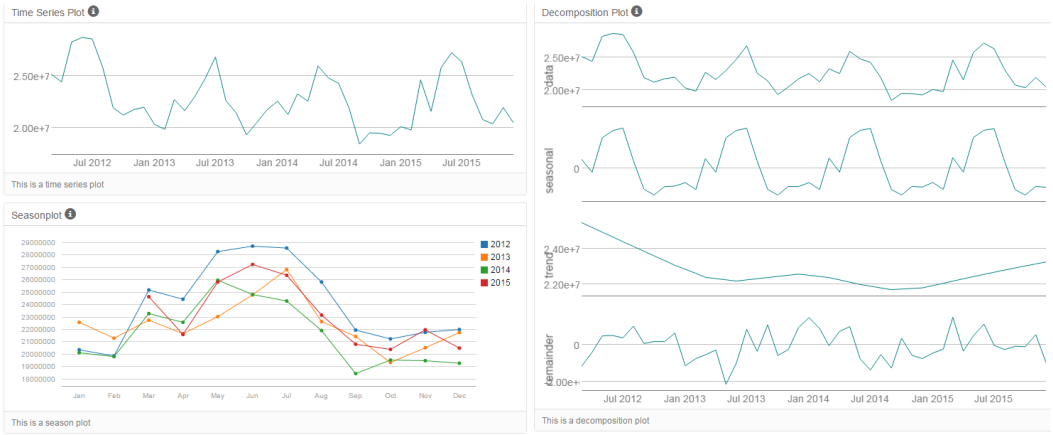


Fig 3.1 Time series plot, season plot, decomposition plot for dataset without differencing

From fig 3.1 we can say that dataset is not stationary and hence differencing will need to be performed.

ETS (M,N,M)

Decomposition plot was used to find features of ETS model. As per decomposition plot, seasonal component does show to increase & hence should be used multiplicatively. Trend component does not show to have any particular behavior and hence neither multiplicatively nor additive should be used. Error shows variation effect was selected by using auto option. The final results show that n-dampening seems to have a better accuracy in forecasting.

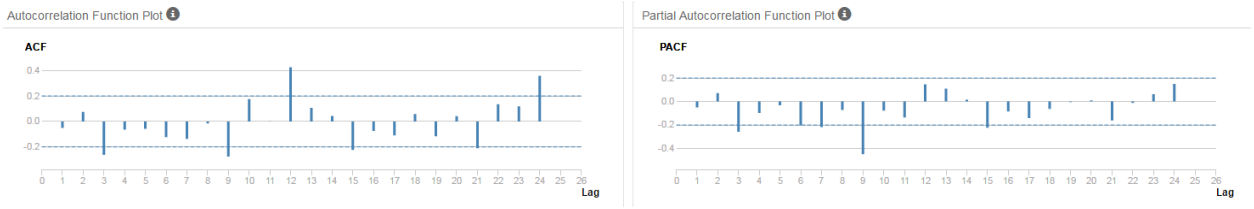


Fig 3.2 ACF & PACF plot of non-seasonal component of ARIMA with one differencing

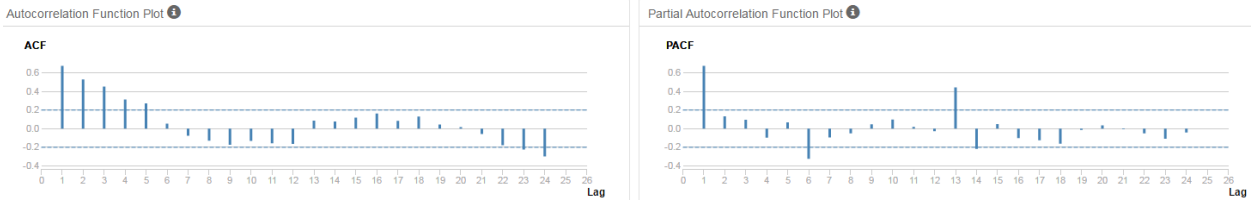


Fig 3.3 ACF & PACF plot of seasonal component of ARIMA

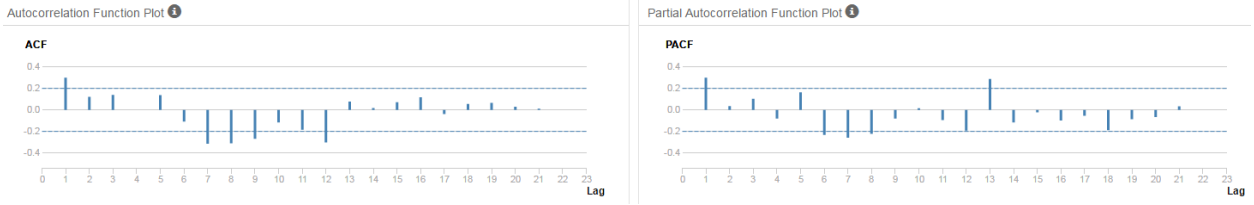


Fig 3.4 ACF & PACF plot after taking first differencing of seasonal component of ARIMA

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	1068766	1590916	1337409	4.372	5.7523	0.833	NA
ARIMA	1303043	2181554	1874870	5.3495	8.2524	1.1678	NA

Fig 3.5 Accuracy measures between two-time series models

ARIMA(0,1,2)

The set of (0,1,2)(0,1,0) has been taken for ARIMA model. The parameter determined for ARIMA are based on ACF & PACF plots given in fig 3.2, 3.3, 3.4.

For non-seasonal, it takes one time diffrecing in order to obtain a stationary series, I=1. ACF plot shows negative correlation at lag 1 which state the use of MA term. In this case MA=2 is used as there is a lag 2 as per fig 3.2 for seasonal, I=1 hence there is one seasonal differencing. No particular signature for both MR & AR terms so MR=0 & AR=0.

As per results generated from running two time series models against hold-out sample ETS(M,N,M) model has lower RMSE and a lower MASE value. Whereas RMSE is 1068766 and MASE is 0.833 for ETS model. ARIMA model has RMSE 2181554 and MASE 1.1678. hence ETS model is used to forecast total sale values of next 12 months for all existing stores and average sales of all segment as listed in table 2.2. hold-out sample has 12 months of data as forecat will be done for next 12 months.

Period	Sub_Period	forecast
2016	1	21539936
2016	2	20413771
2016	3	24325953
2016	4	22993466
2016	5	26691951
2016	6	26989964
2016	7	26948631
2016	8	24091579
2016	9	20523492
2016	10	20011749
2016	11	21177435
2016	12	20855799

Table 2.2 Forecasted sale values in next 12 months for all existing stores

Period	Sub_Period	forecast
2016	1	21539936.01
2016	2	20413770.6
2016	3	24325953.1
2016	4	22993466.35
2016	5	26691951.42
2016	6	26989964.01
2016	7	26948630.76
2016	8	24091579.35
2016	9	20523492.41
2016	10	20011748.67
2016	11	21177435.49
2016	12	20855799.11

Fig 3.6 snapshot of Output data (forecast_produce_sales.xlsx)

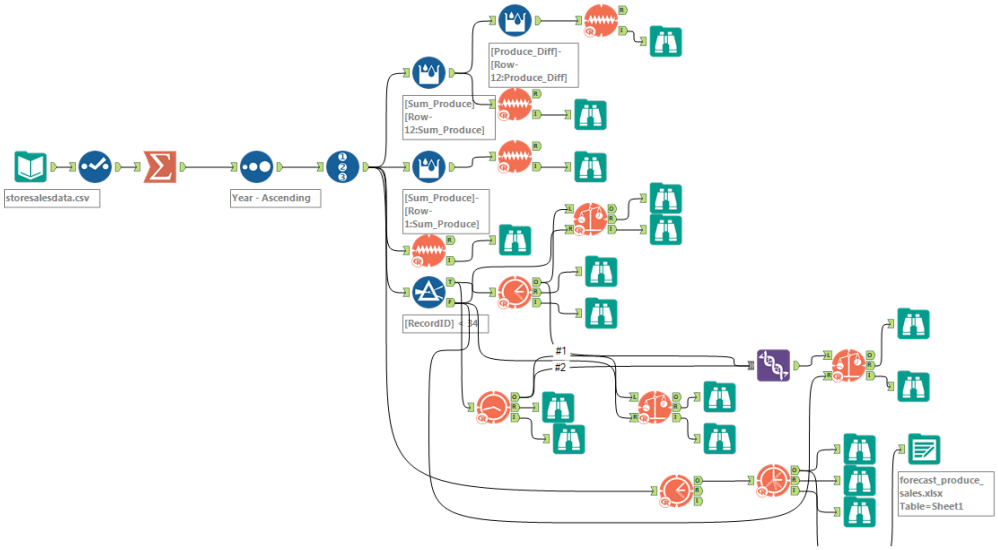


Fig 3.7 workflow for forecasting sale valur for average store in year 2016

The monthly total produce sales of all new stores are calculate by workflow mentioned in fig 3.8 & 3.9. The average sales of each segment of each month is calculated by dividing forecasted sum produce sale of each segment by the no of existing stores in segment. The monthly average sale of each segment is multiplied by no of new stores in that segment. Monthly total produce for all new stores are then calculated by summing total sales of all segment in each month.

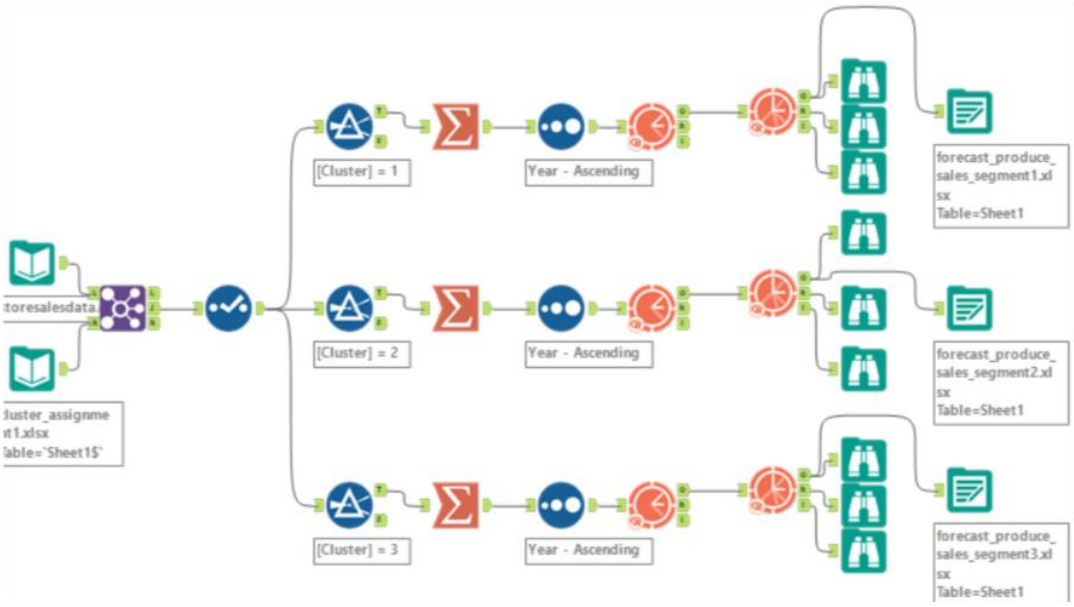


Fig 3.8 workflow for forecasting sum produce sale for each segment, later to find total produce sale of each month for new stores.

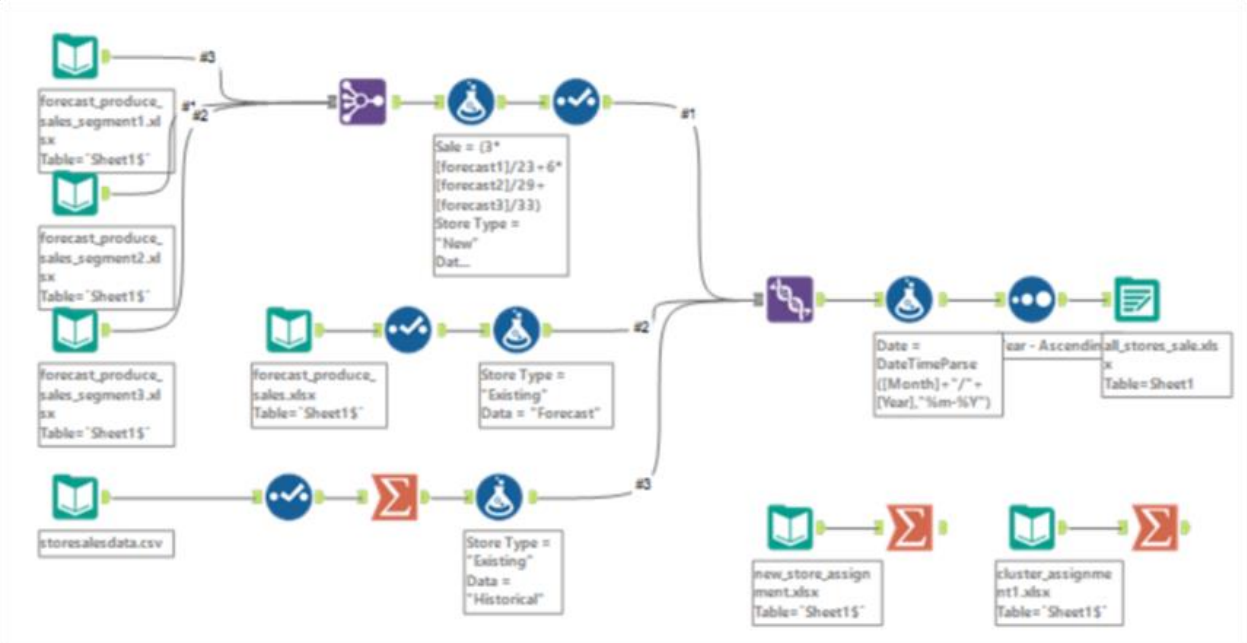
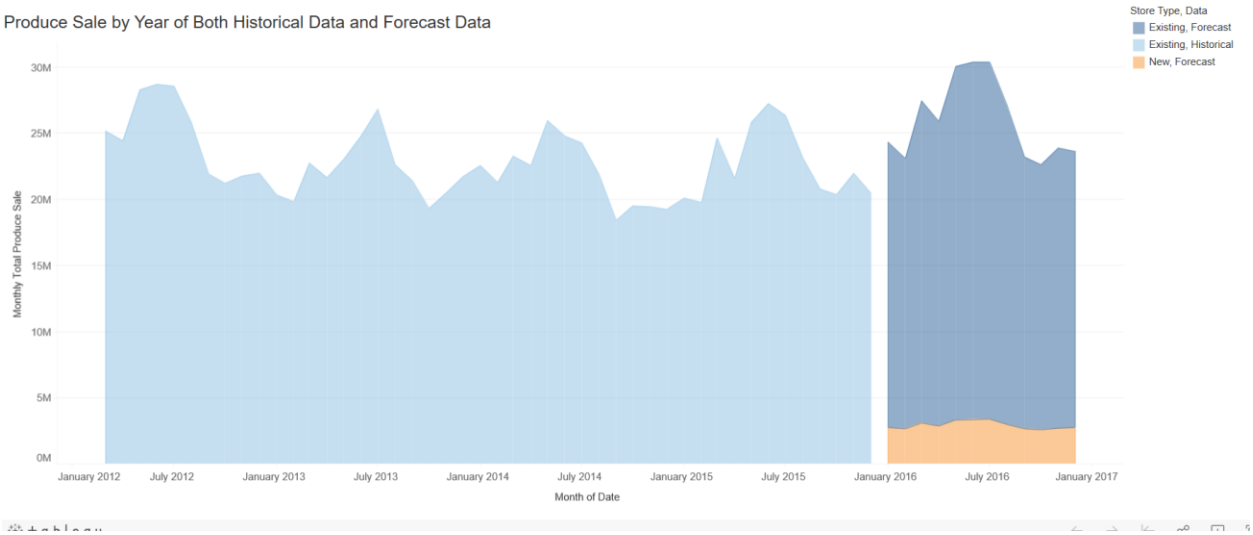


Fig 3.9 workflow used for calculating monthly total produce sales of all new stores.it is also used to generate dataset for Tableau visualization

Year	Month	Existing Stores Sale	New Stores Sale
2016	1	21,539,936	2,761,958
2016	2	20,413,771	2,656,665
2016	3	24,325,953	3,099,058
2016	4	22,993,466	2,873,607
2016	5	26,691,951	3,327,835
2016	6	26,989,964	3,356,062
2016	7	26,948,631	3,391,943
2016	8	24,091,579	2,991,383
2016	9	20,523,492	2,664,295
2016	10	20,011,749	2,588,210
2016	11	21,177,435	2,702,838
2016	12	20,855,799	2,761,943

Table 3.10 forecasted sale in next 12 months for both existing and new stores.

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.



https://public.tableau.com/profile/harsh.varudkar#!/vizhome/Project7_Task3_2/Dashboard1