UDACITY ND.Business Analytics Francihelena Uzcategui

January 15, 2019.

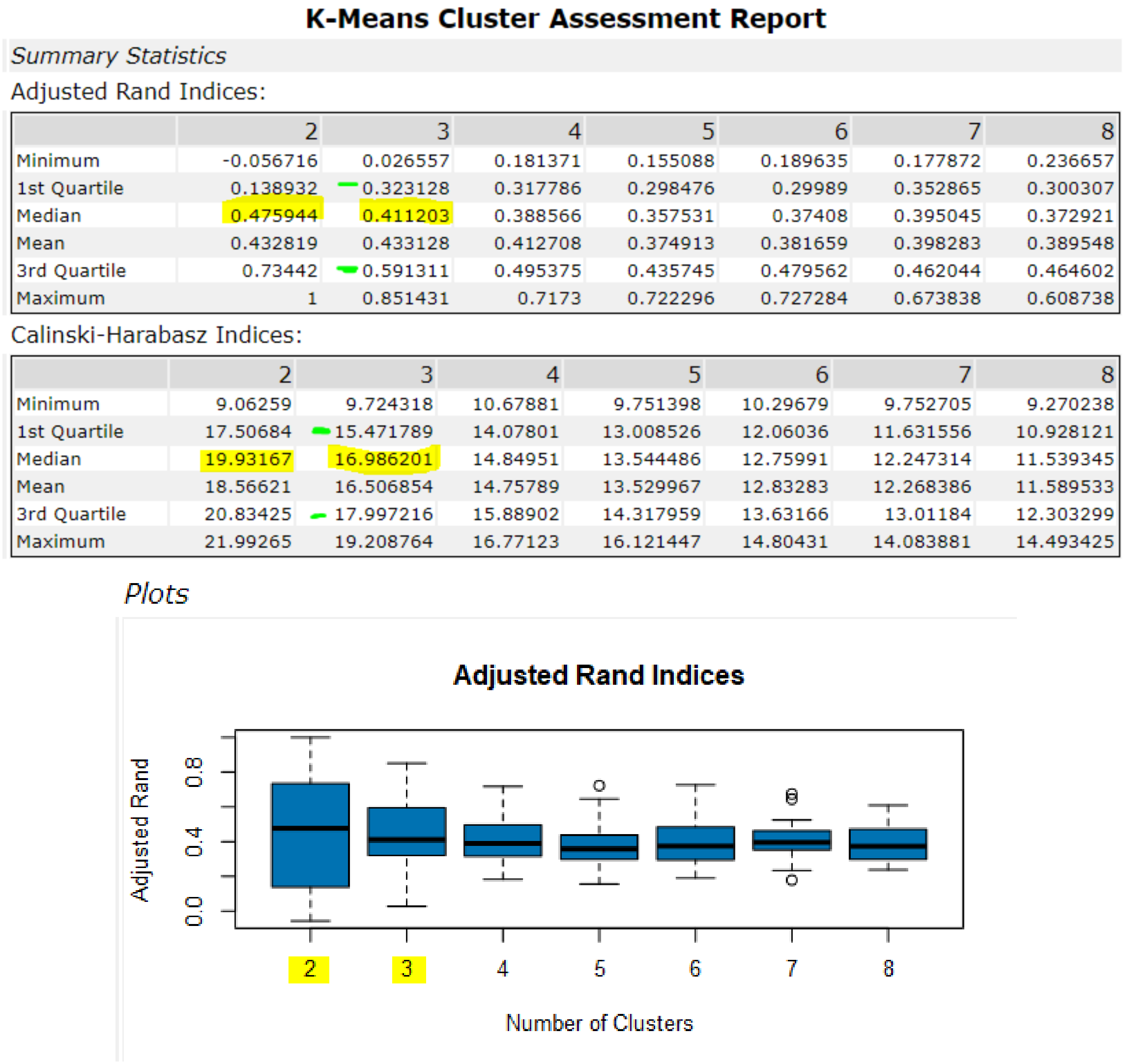
Project: Predictive Analytics Capstone

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 segments is three. We obtained this number by K-Centroids diagnostic tool and K-means method. Selecting the minimum of clusters 2 and the maximum 8, due to there are 9 fields.

K-Means Cluster assessment report includes the Rand and Calinski-Harabasz indices, these use to determine the median and spread by each cluster. The box-whisker plots shown high median values between cluster 2 and 3, although we selected the third cluster because have a tight or compact spread (green sign.)



K-Means Cluster Assessment Report

Summary Statistics

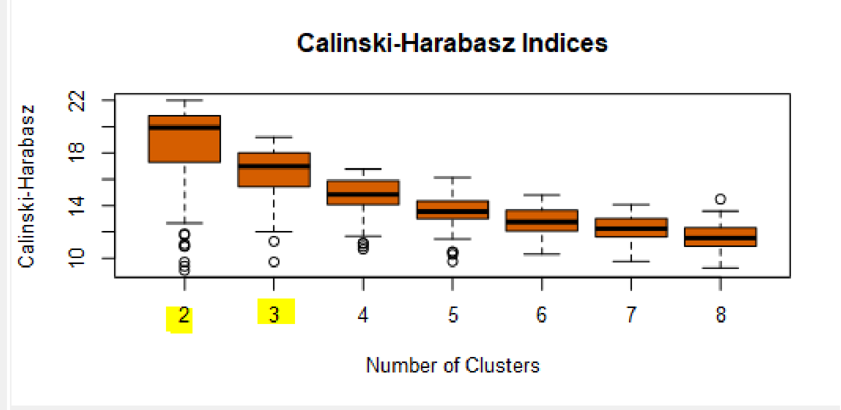
Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.056716	0.026557	0.181371	0.155088	0.189635	0.177872	0.236657
1st Quartile	0.138932	0.323128	0.317786	0.298476	0.29989	0.352865	0.300307
Median	0.475944	0.411203	0.388566	0.357531	0.37408	0.395045	0.372921
Mean	0.432819	0.433128	0.412708	0.374913	0.381659	0.398283	0.389548
3rd Quartile	0.73442	0.591311	0.495375	0.435745	0.479562	0.462044	0.464602
Maximum	1	0.851431	0.7173	0.722296	0.727284	0.673838	0.608738

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	9.06259	9.724318	10.67881	9.751398	10.29679	9.752705	9.270238
1st Quartile	17.50684	15.471789	14.07801	13.008526	12.06036	11.631556	10.928121
Median	19.93167	16.986201	14.84951	13.544486	12.75991	12.247314	11.539345
Mean	18.56621	16.506854	14.75789	13.529967	12.83283	12.268386	11.589533
3rd Quartile	20.83425	17.997216	15.88902	14.317959	13.63166	13.01184	12.303299
Maximum	21.99265	19.208764	16.77123	16.121447	14.80431	14.083881	14.493425

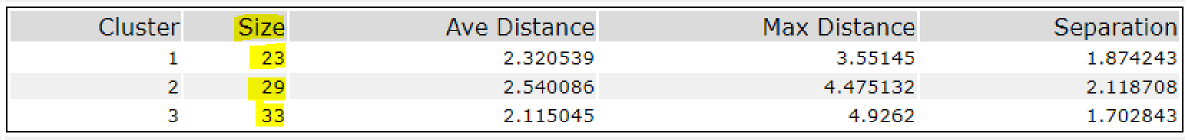
Plots



2. How many stores fall into each store format?

= According to previous three segments selected, we used the K-Centroids cluster analysis tool to test the clustering method.

The report K-Means Clustering Solution describes cluster information-topic size:



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

Which depicts a range between 20 and 35 of the number of the stores by each segment.

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

= The third cluster has the smallest Average distance 2.11, being the most compact and stable among the other.

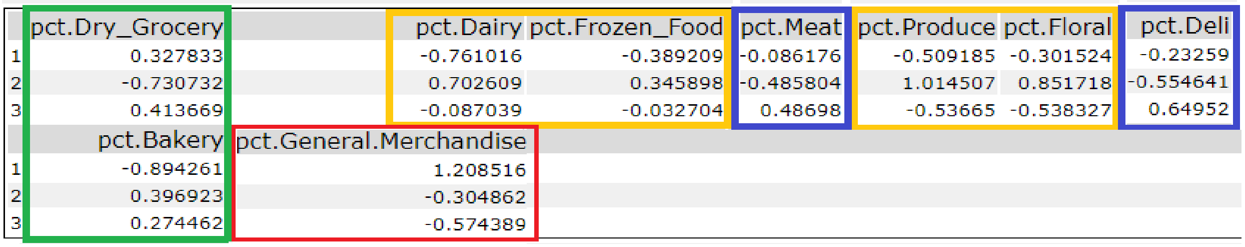
Meanwhile, the second cluster has the highest Maximum distance 4.47 from the centroid.

Finally, the third cluster has 2.11 points, being more separated from the other two clusters.

The performance of nine variables mainly is negative, just two fields Grocery and Bakery having two clusters with positive values (green box); the seven fields remaining each with two clusters have negative values.

Dairy, Frozen Food, Produce, and Floral fields (yellow boxes): first and third cluster with negative values and only the second cluster with positive values.

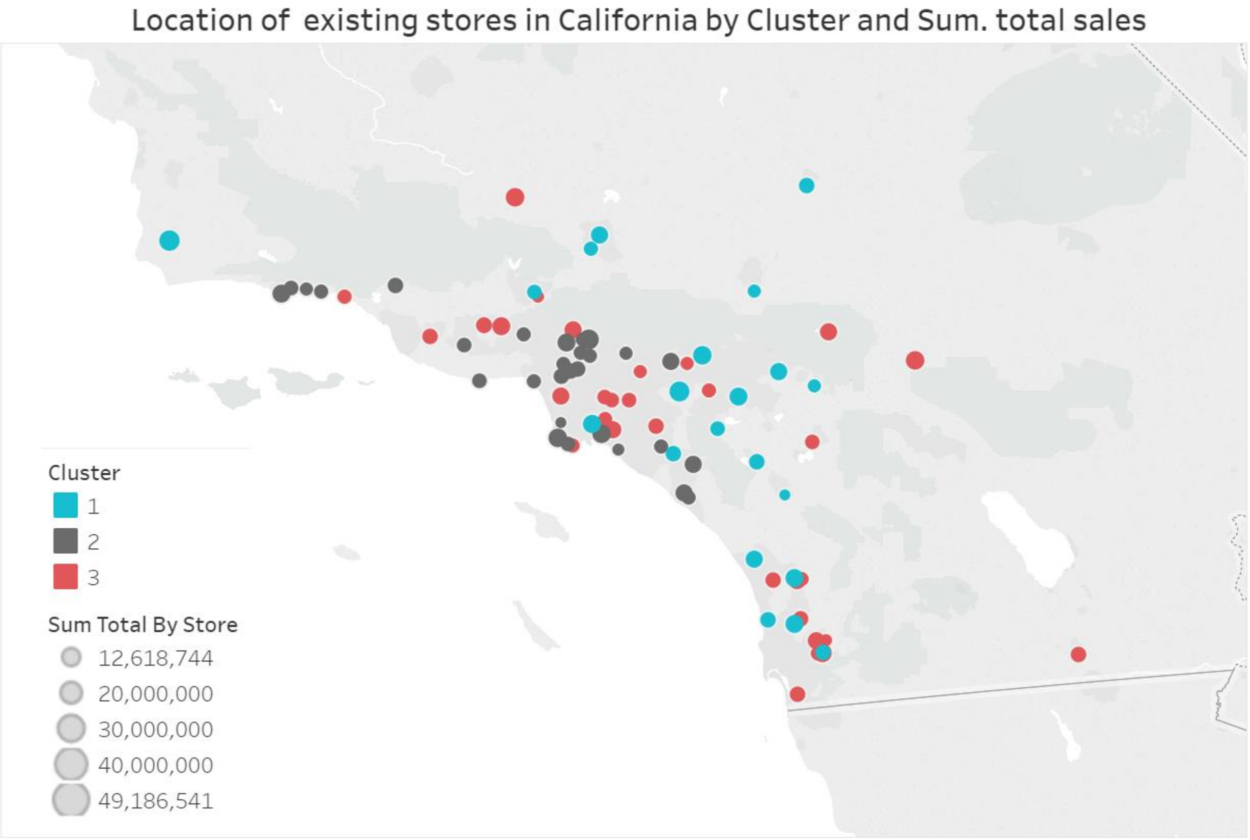
The Meat and Deli fields (blue boxes) very much alike than aforementioned: first and second cluster have negative values and just the third cluster positive.



	pct.Dry_Grocery		pct.Dairy	pct.Frozen_Food	pct.Meat	pct.Produce	pct.Floral	pct.Deli
1	0.327833		-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732		0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669		-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	pct.Bakery	pct.General.Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

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.

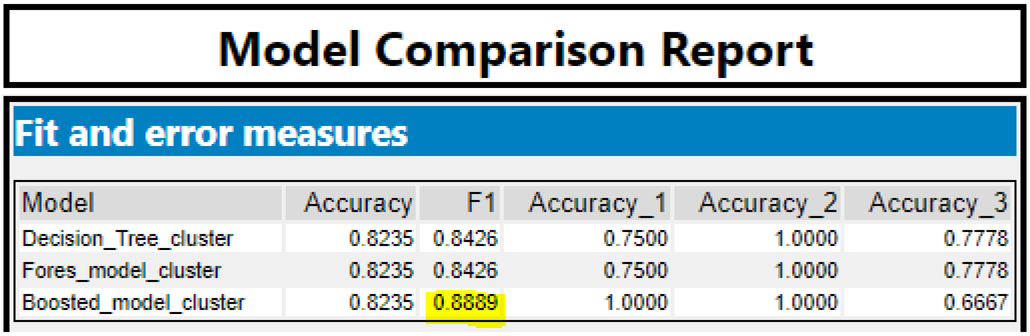


https://public.tableau.com/views/LocationofexistingstoresinCaliforniabyClusterandSum_totalsales/Sheet1?:embed=y&:display_count=yes&publish=yes

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.)

= The methodology map to solve Business problems was useful because we wanted to predict the outcome and had rich data, therefore we applied Forest, Decision tree, and Boosted models. The three models' accuracies were the same, this situation happens with small samples. Thus, the highest F1 score is for the Boosted_model_cluster 0.8889.

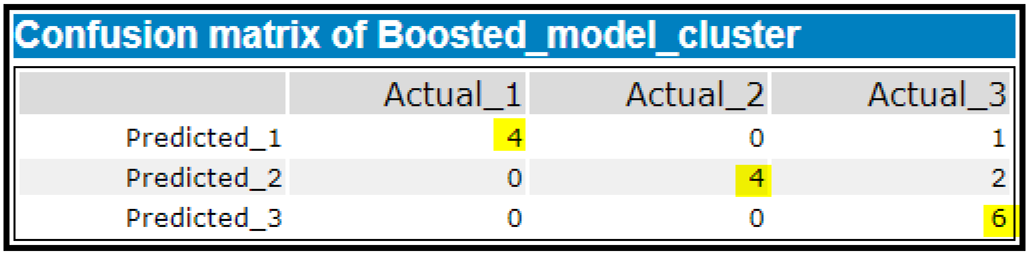


Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_cluster	0.8235	0.8426	0.7500	1.0000	0.7778
Fores_model_cluster	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_model_cluster	0.8235	0.8889	1.0000	1.0000	0.6667

The confusion matrix of Boosted_model_cluster shows the highest True positives values and high sensitivity, it means rarely fail diagnosis.



Confusion matrix of Boosted_model_cluster

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

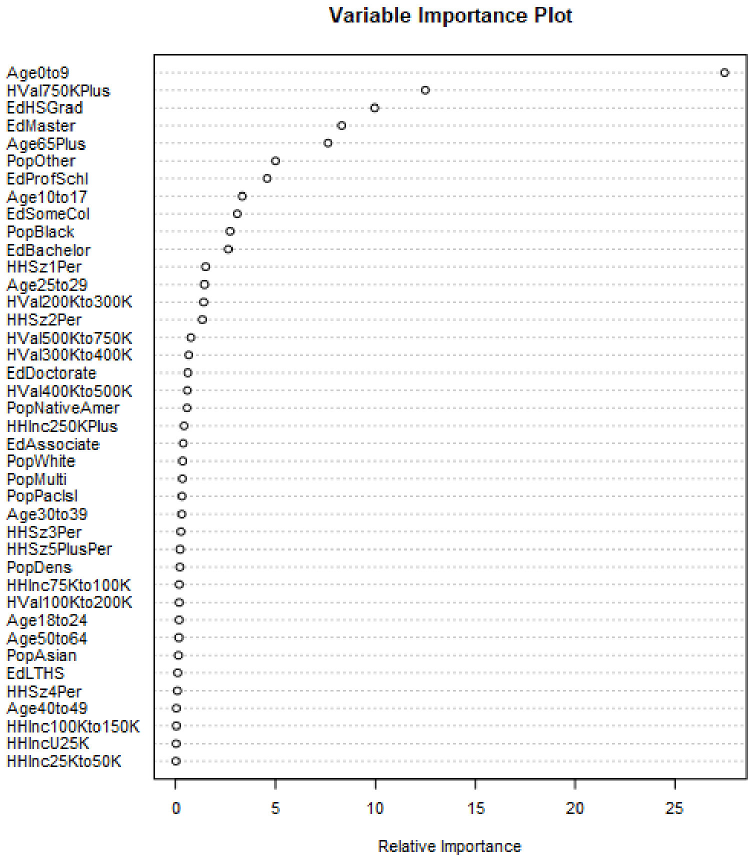
Accuracy rate 1= predicted 1/ actual 1 ⇒ 4/4= 1
Accuracy rate 2= predicted 2/ actual 2 ⇒ 4/4= 1
Accuracy rate 3= predicted 3/ actual 3 ⇒ 6/9= 0.6666

Therefore, to the above report shows the Boosted Model fits to solve the business problem because it has the highest values of accuracy by each cluster.

2. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.

= According to the Variable Importance plot, the top 3 variables are Age0to9, HVal750KPlus, and EdHSGrad, respectively. Age0to9 is the leader with more than 25 scores, thus the variable related to Children is a strong meaningful relationship among demographic indicators and store segments. A little far at 12 scores HVal750KPlus - Family Home value 750k plus, and at 10 scores the EdHSGrad-Education High School Grade.

Report for Boosted Model Boosted_model_cluster



3. What format do each of the 10 new stores fall into? Please fill in the table below.
= Following the selection of the best model to use, we filtered the 10 last stores (>85) and ran the Score tool, including the Formula tool to help the easy reading of the results by each segment.

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

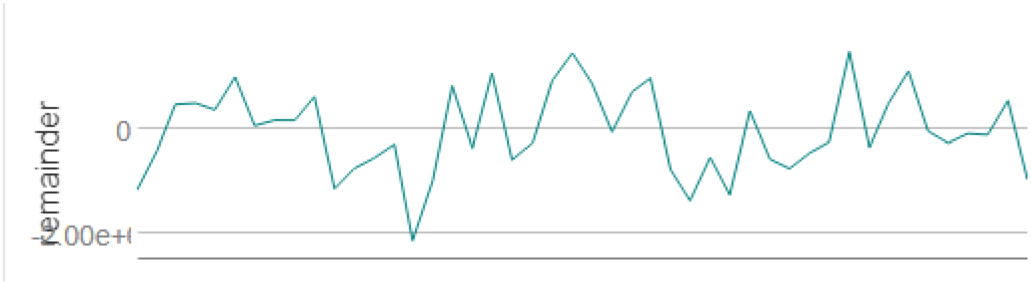
Task 3: Predicting Produce Sales

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?

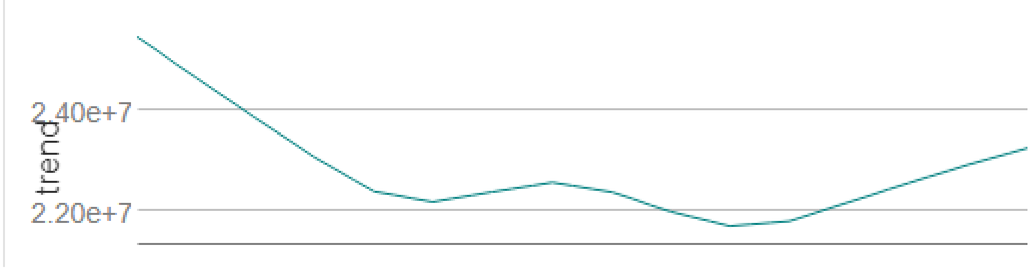
A) From the original dataset “store sales data” we compared the performance of ETS and ARIMA model to select one of them to be applied for the stores’ forecast.

- **ETS:** from the Decomposition Plot we observed the M, N, M pattern.

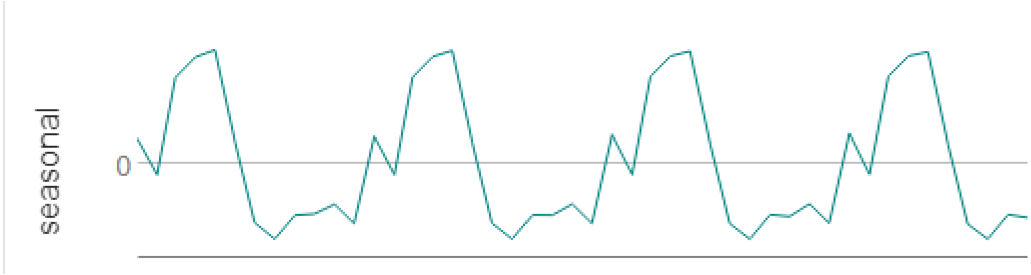
The Error plot shows variance along the years, it is fluctuating with different sizes, this means we used the error multiplicatively (M.)



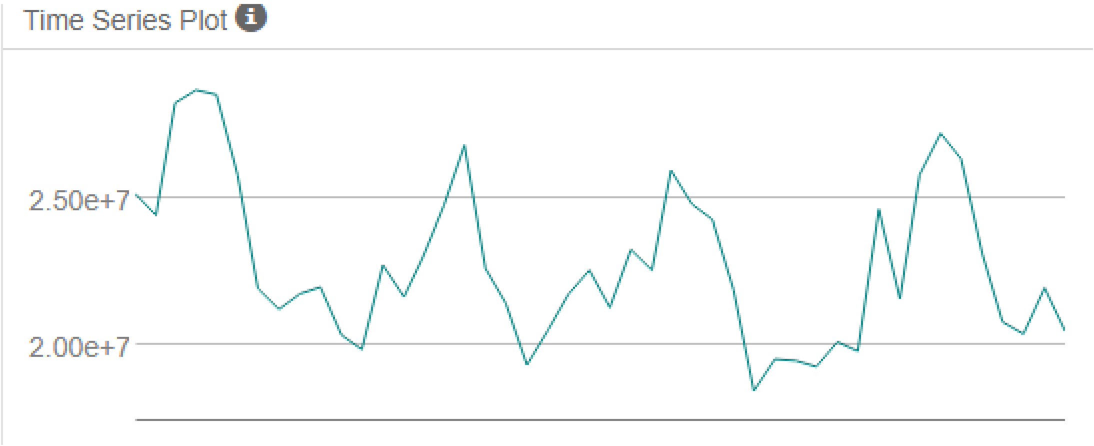
The Trend Plot, we observed the trend moves uptrend and downtrend, thus wasn’t clear the pattern, in our opinion is neutral (N.)



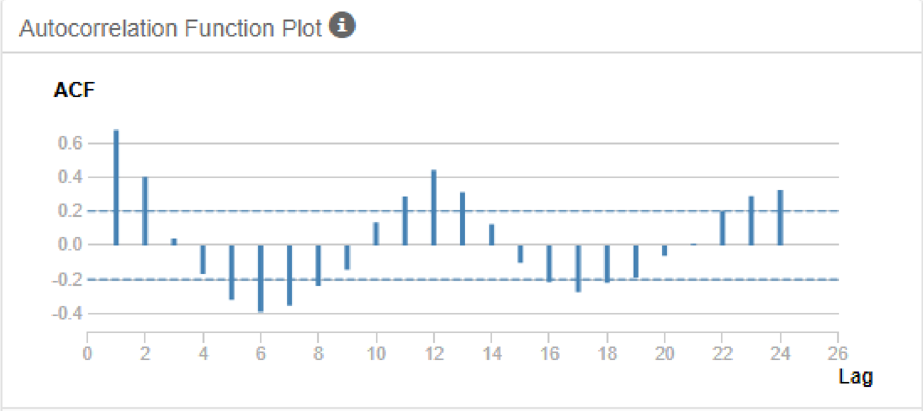
The Seasonal plot shows peaks and valleys in similar periods of time, this suggests applying seasonality in a multiplicative method (M.)

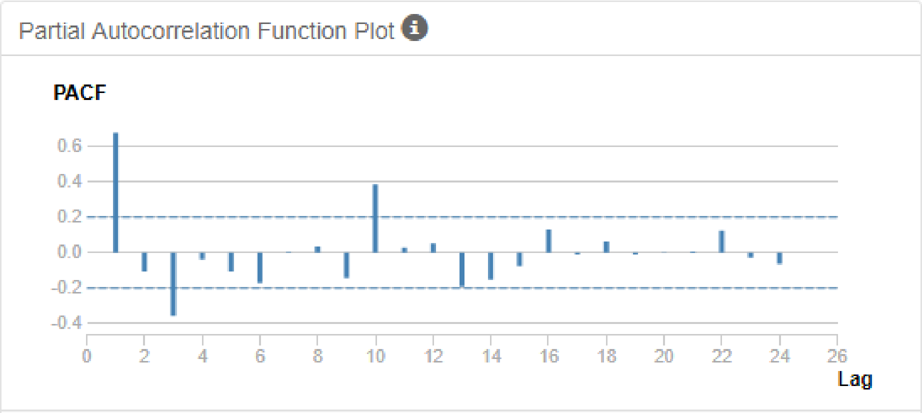


- **ARIMA:** from the first forecast from TS Plot tool the dataset showed a time series plot non-stationary because it had too many peaks and valleys, as shown in the graph below.

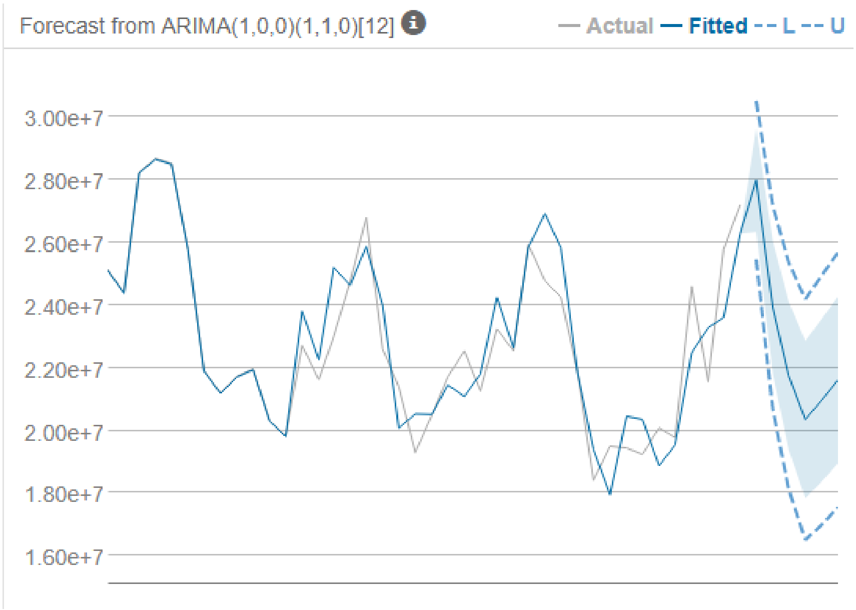


ACF and PACF graphs showed the positive correlation, being suggested the AR 1 and MA 0 terms. ARIMA (1,1,1) (0,1,0) [12] depicts that there was a positive autocorrelation at period 1. Thus, it was necessary to transform the dataset to stabilized, it was trough differencing.

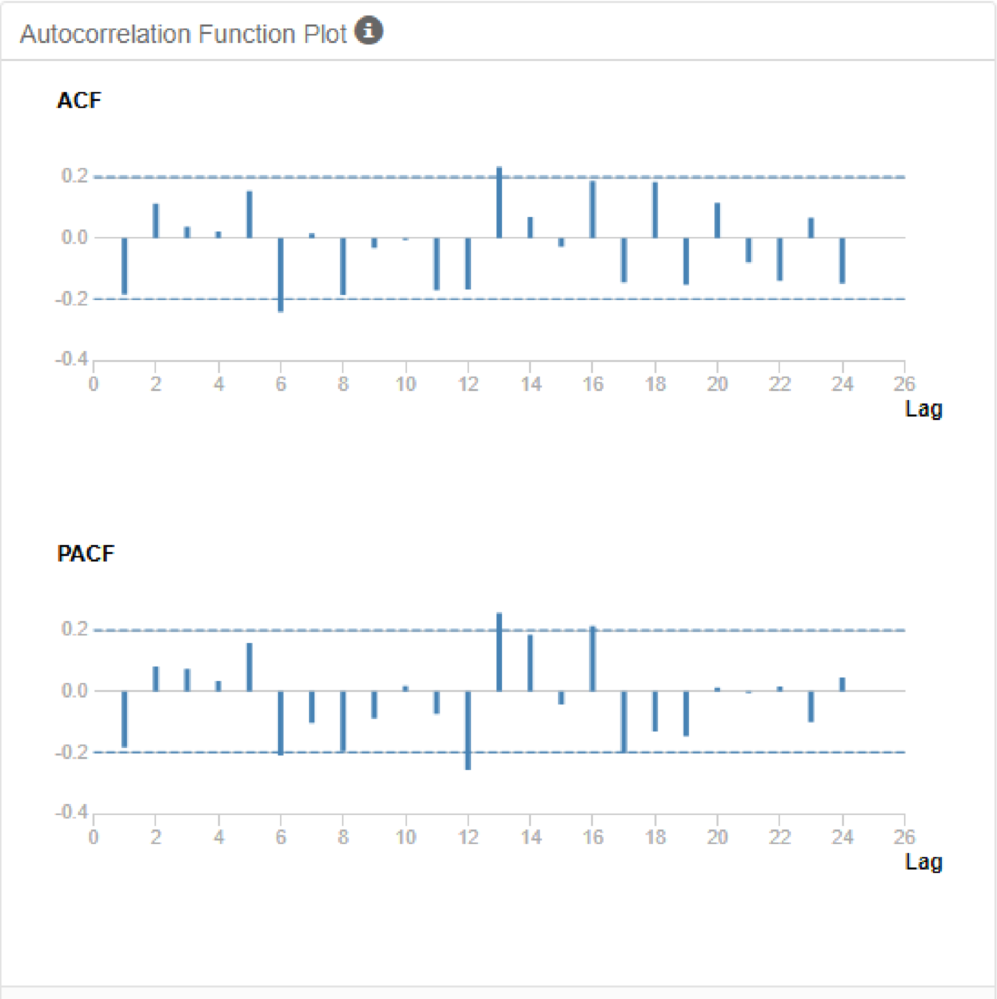




After applied the filter and ARIMA tool, we observed the time series have been decreasing towards zero, as shown the graph below. ARIMAS's terms changed respect than the first attempt. It changes to ARIMA (0,1,1) (0,1,0) [12], suggesting a negative correlation at period 1. Thus, the dataset was stabilized.

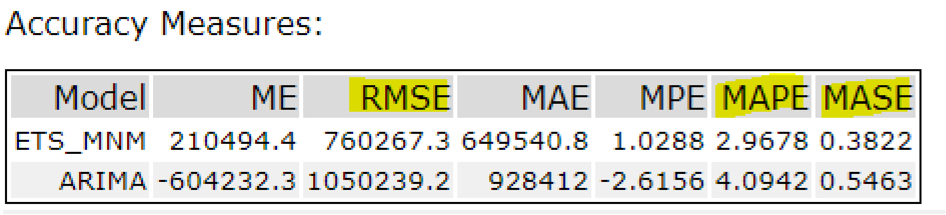


ACF and PACF graphs show a negative correlation, being suggested the AR 0 and MA 1 terms.



- **ETS vs. ARIMA from the hold out sample:**

After applied the TS compare tool to original data, we obtained the comparison of the models. ETS model in overall has the best accuracy measures (report below); especially the lowest MASE, MAPE, and RMSE. Therefore, we choose the ETS model to the forecast produce sales for the new and existing stores.

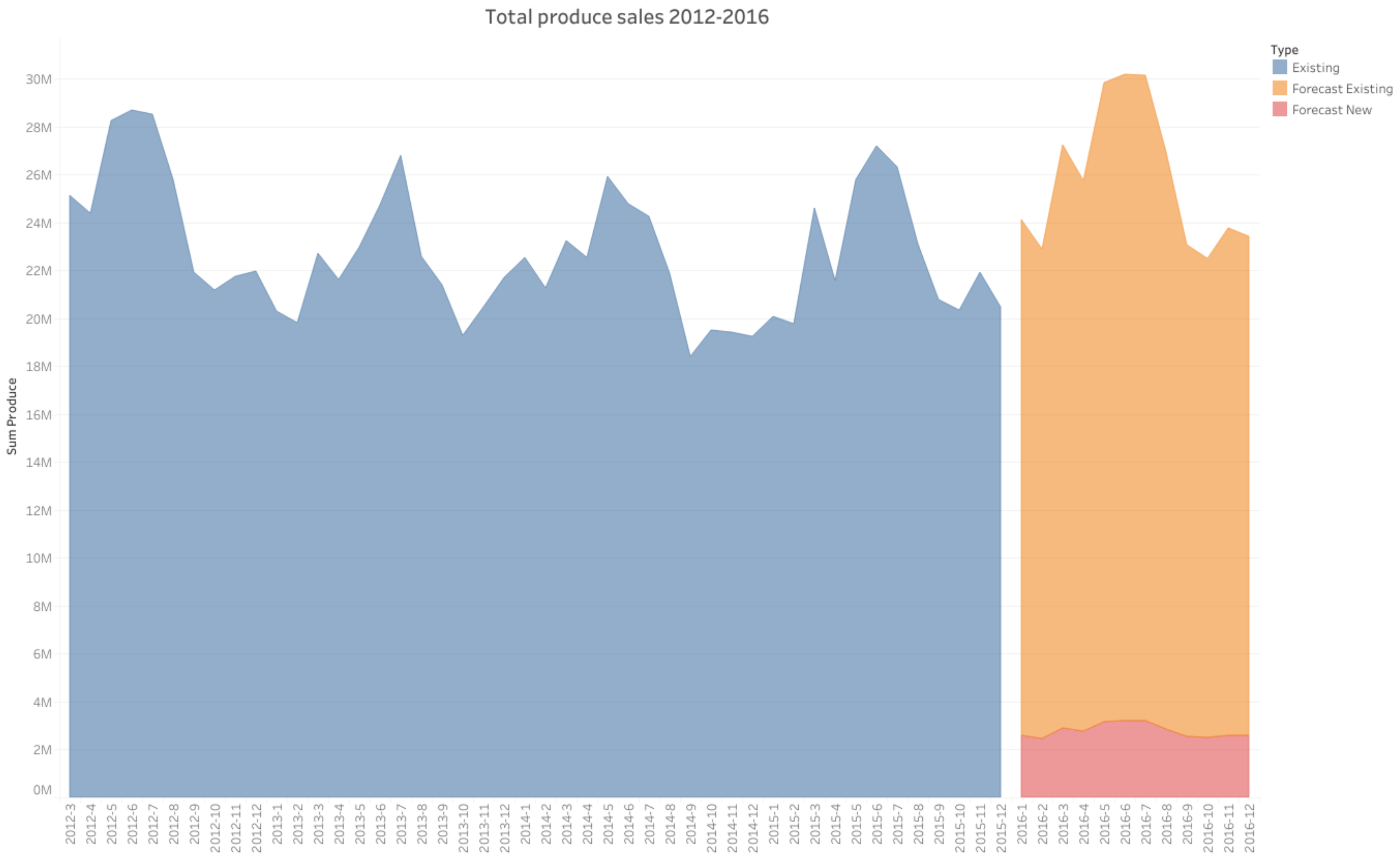


Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_MNM	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
ARIMA	-604232.3	1050239.2	928412	-2.6156	4.0942	0.5463

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
16-Jan	$ 2,587,451.00	$21,539,936.00
16-Feb	$ 2,477,353.00	$20,413,771.00
16-Mar	$ 2,913,185.00	$24,325,953.00
16-Apr	$ 2,775,746.00	$22,993,466.00
16-May	$ 3,150,867.00	$26,691,951.00
16-Jun	$ 3,188,922.00	$26,989,964.00
16-Jul	$ 3,214,746.00	$26,948,631.00
16-Aug	$ 2,866,349.00	$24,091,579.00
16-Sep	$ 2,538,727.00	$20,523,492.00
16-Oct	$ 2,488,148.00	$20,011,749.00
16-Nov	$ 2,595,270.00	$21,177,435.00
16-Dic	$ 2,573,397.00	$20,855,799.00



https://public.tableau.com/views/Totalproducesales2012-2016/Sheet1?:embed=y&:display_count=yes&publish=yes

Works Cited

Cluster Analysis Report - Meaning of Avg_Distance, Max_Distance and Separation
https://community.alteryx.com/t5/Alteryx-Designer-Discussions/Cluster-Analysis-Report-Meaning-of-Avg-Distance-Max-Distance-and/td-p/155859

Understanding the Outputs of the Decision Tree Tool
https://community.alteryx.com/t5/Alteryx-Knowledge-Base/Understanding-the-Outputs-of-the-Decision-Tree-Tool/ta-p/144773

Decision Tree Error #290: Subscript out of Bounds
https://community.alteryx.com/t5/Alteryx-Designer-Discussions/Decision-Tree-Error-290-Subscript-out-of-Bounds/td-p/124352

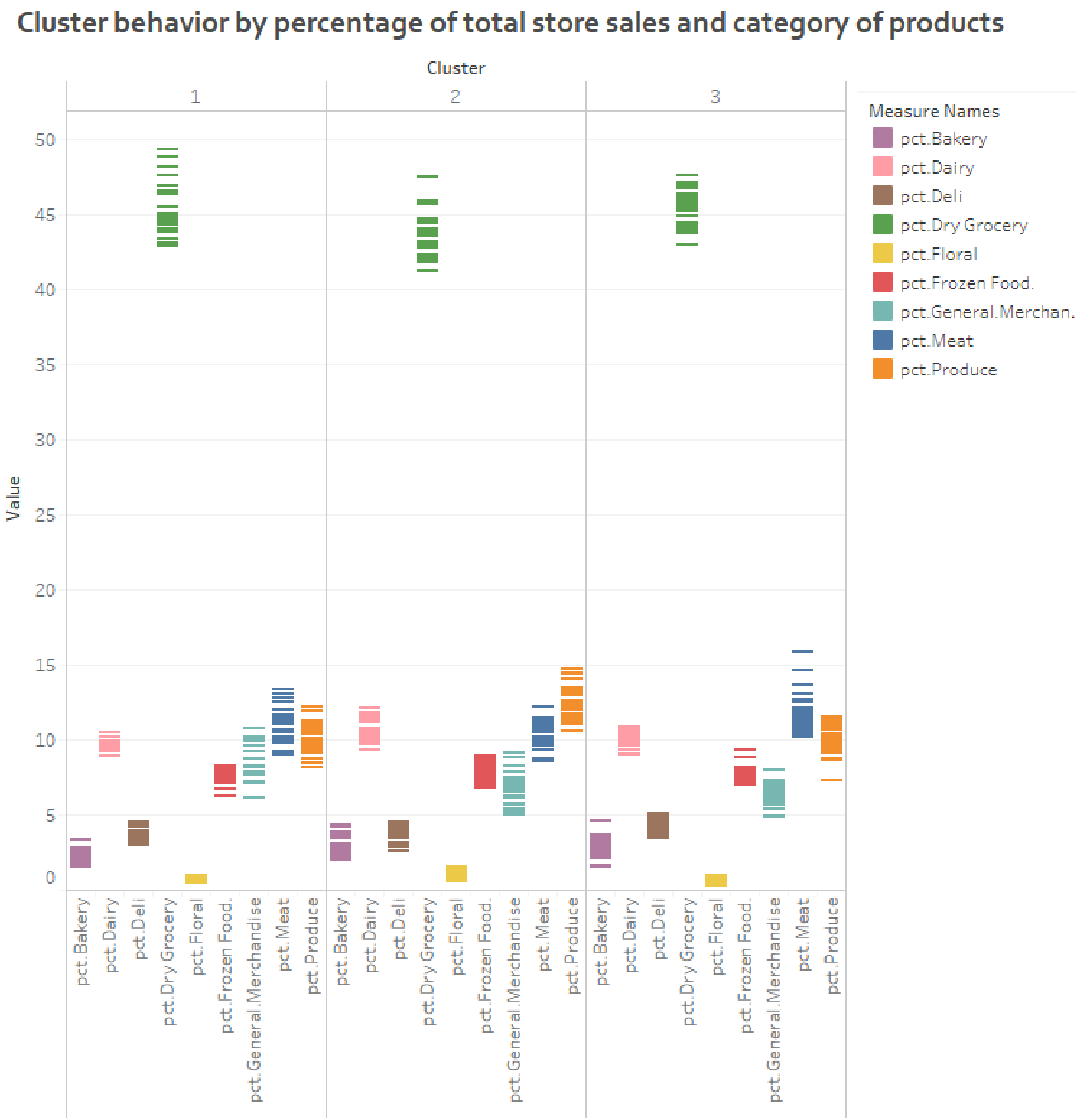
Forecast for new and existing stores
https://docs.google.com/document/d/e/2PACX-1vSOHc9V98yrO0glbn2DV0Zyzapecq4BwgR8C8oGGRe8sSIWvR8T_fMGd7a4wYIQU5eUpUOXUmPXkTC1/pub

Facts and fallacies of the AIC
https://robjhyndman.com/hyndsight/aic/

Combining Predictive Techniques
https://study-hall.udacity.com

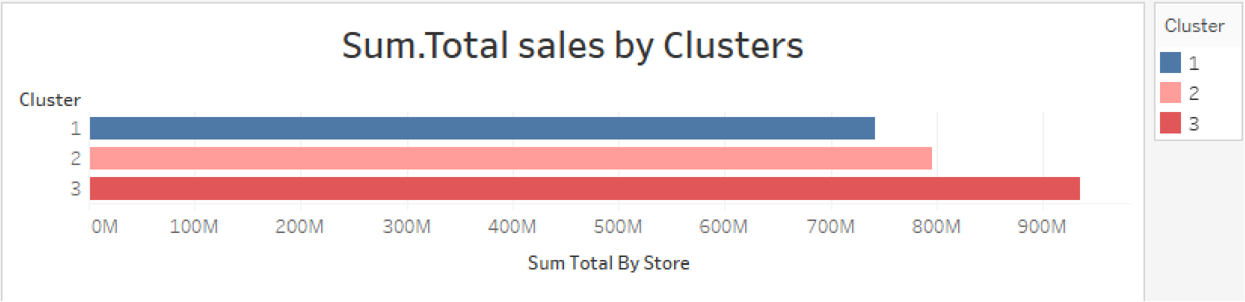
Annexes

Task 1:



https://public.tableau.com/views/Clusterbehaviorbypercentageoftotalstoresalesandcategoryofproducts/Sheet1?:embed=y&:display_count=yes&publish=yes

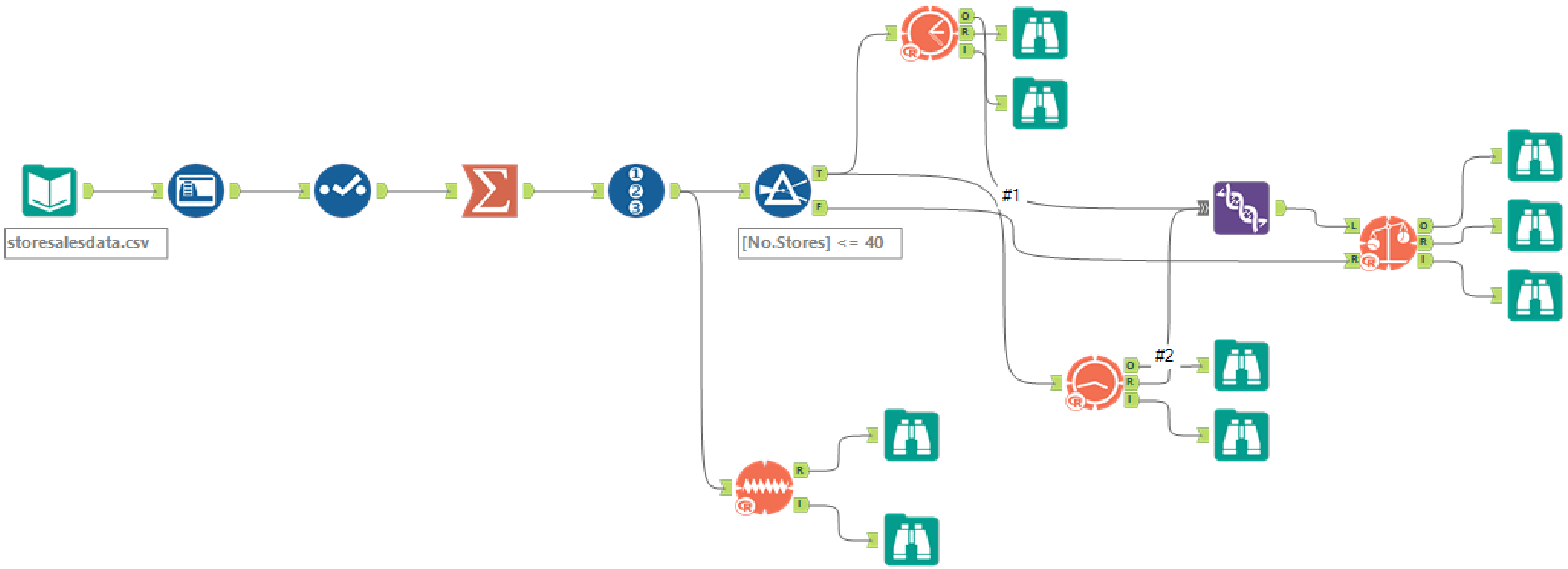
Task 1:



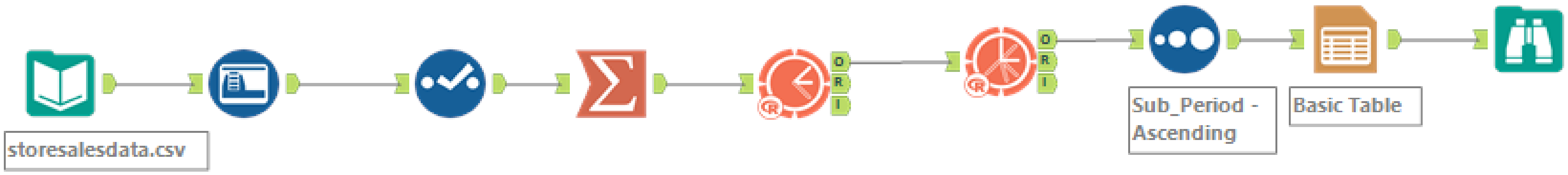
https://public.tableau.com/views/Sum_TotalsalesbyClusters/Sheet2?:embed=y&:display_count=yes&publish=yes

Task 3:

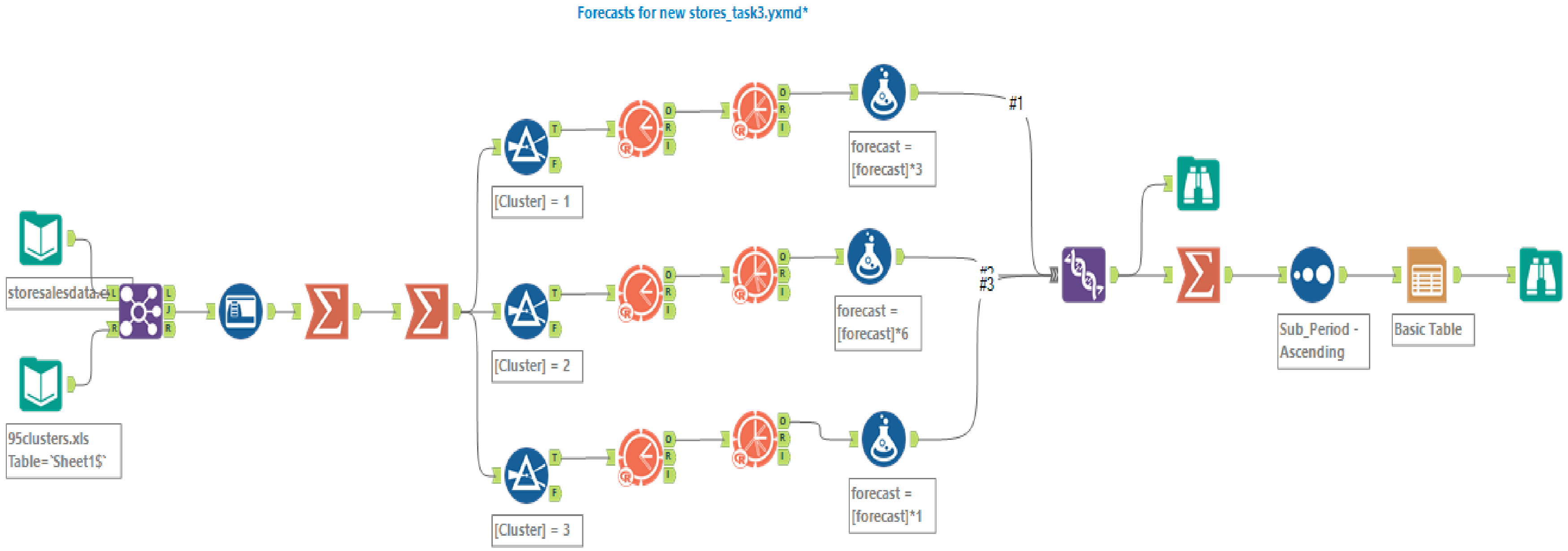
Comparison_ETS_ARIMA_task3.yxmd*

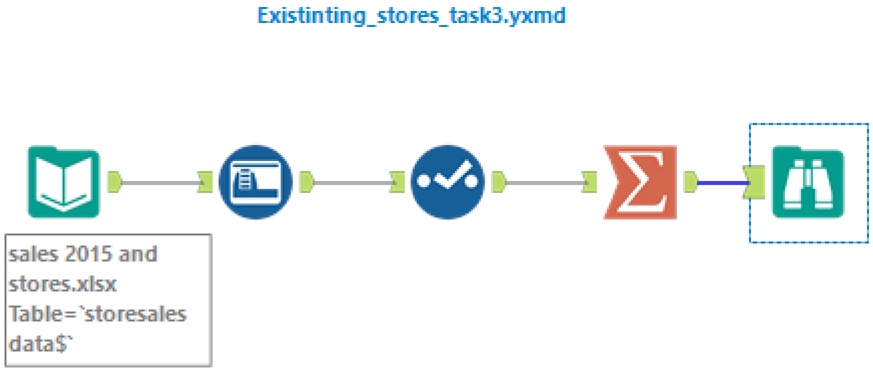


Forecast_for_existing_stores_task3.yxmd



Task 3:


Forecasts for new stores_task3.yxmd*
storesalesdata.el
95clusters.xls
Table=`Sheet1$`
L
J
R
[Cluster] = 1
[Cluster] = 2
[Cluster] = 3
T
F
O
R
I
forecast =
[forecast]*3
forecast =
[forecast]*6
forecast =
[forecast]*1
#1
#2
#3
Sub_Period -
Ascending
Basic Table


Existinting_stores_task3.yxmd
sales 2015 and
stores.xlsx
Table=`storesales
data$`