Project: Predictive Analytics Capstone | Ioanna Vasilopoulou

Task 1: Determine Store Formats for Existing Stores

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

   In order to determine the optimal number of store formats we summed the sales data by store ID and year. We used only the sales data for 2015 and we used the percentage sales per category per store for clustering (category sales as a percentage of total store sales). We chose two measurements to evaluate and identify the number of cluster that we should use: Adjusted Rand index and Calinski-Harabasz index, with K-means as the selected clusterization algorithm. Based on the indices below, the 3-clusters option has the strong median across both indices. This means that 3 clusters are more stable, compact and distinct.

K-Means Cluster Assessment Report

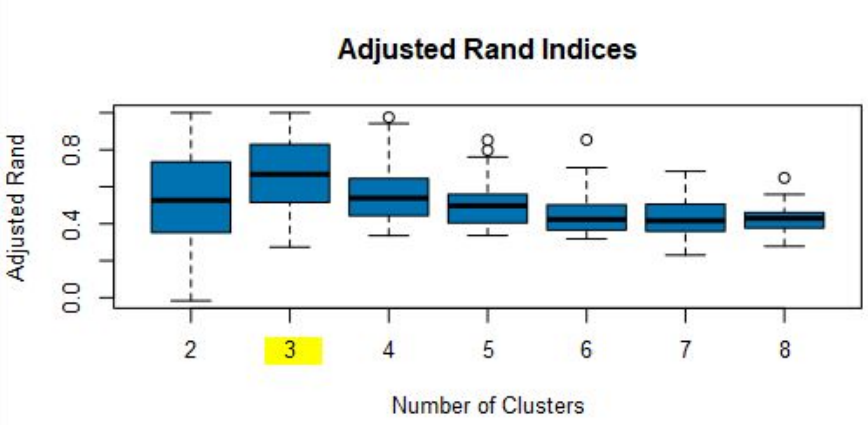
Summary Statistics

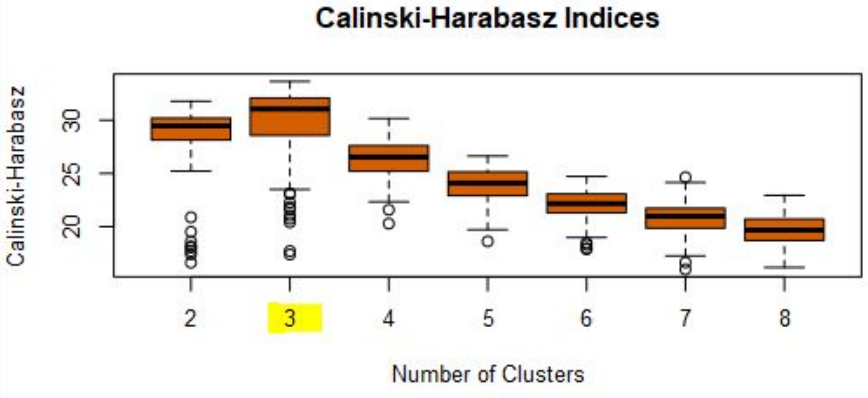
Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.016293	0.27351	0.335359	0.336327	0.318262	0.230196	0.27786
1st Quartile	0.352041	0.515917	0.445826	0.409773	0.366788	0.358895	0.377341
Median	0.526785	0.66768	0.538528	0.497192	0.423541	0.416509	0.428806
Mean	0.53781	0.664773	0.565975	0.50103	0.45115	0.432196	0.421514
3rd Quartile	0.734477	0.826692	0.644691	0.555087	0.499921	0.502931	0.458601
Maximum	1	1	0.975264	0.852076	0.8539	0.683894	0.647983

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.61829	17.38103	20.28456	18.61989	17.8746	15.98702	16.16824
1st Quartile	28.17383	28.57484	25.20913	22.93454	21.30575	19.85155	18.71365
Median	29.46587	31.05384	26.53788	24.086	22.16245	20.97743	19.6662
Mean	28.45131	29.70664	26.41806	23.87003	22.02174	20.77195	19.65973
3rd Quartile	30.17907	32.08726	27.59305	25.10099	23.06602	21.72942	20.7099
Maximum	31.78345	33.63781	30.1583	26.63063	24.72038	24.63982	22.95166





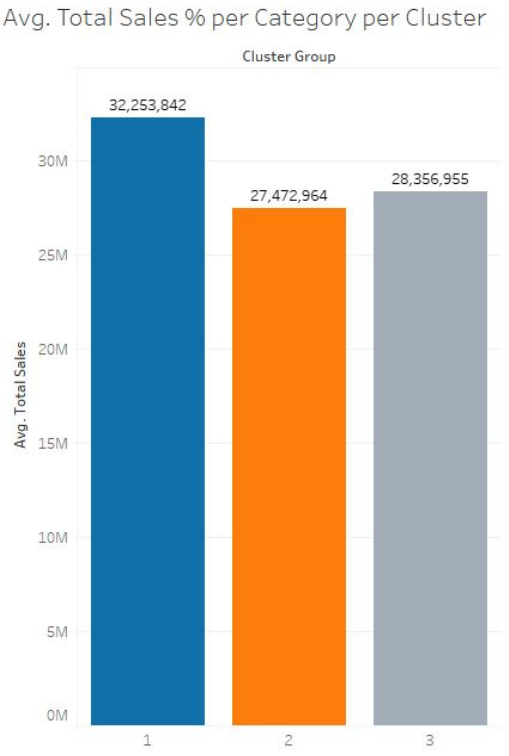
1. **How many stores fall into each store format?**
   We have 3 cluster groups with 23, 29 and 33 stores in each one, respectively.

Cluster Information:

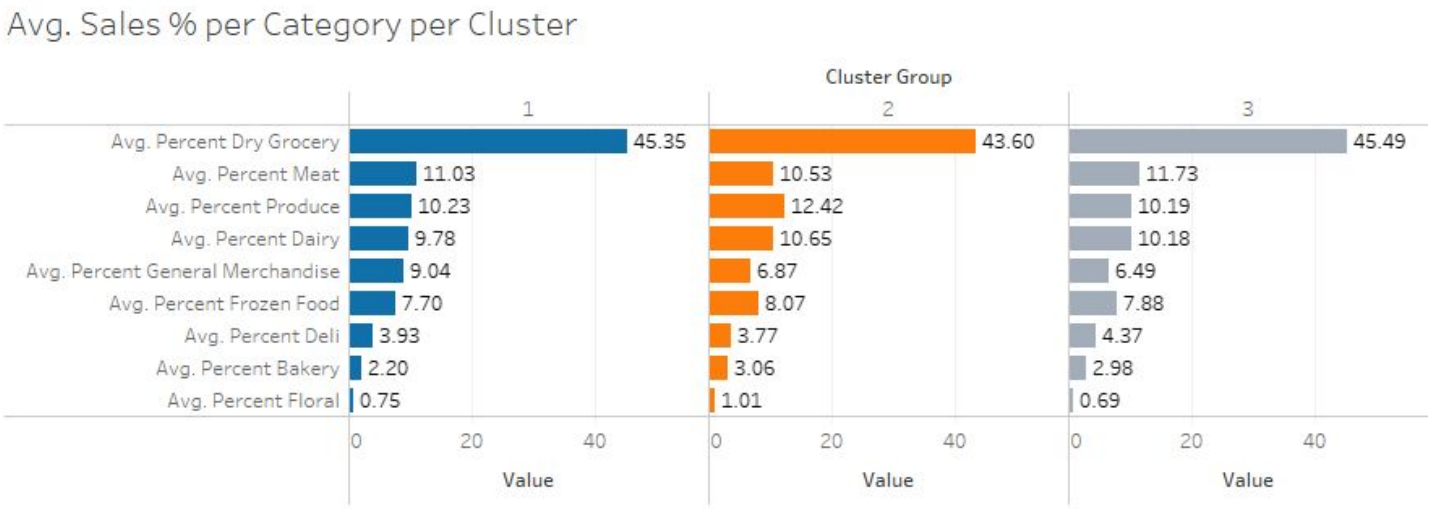
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Record #	Cluster	Count
1	1	23
2	2	29
3	3	33

2. **Based on the results of the clustering model, what is one way that the clusters differ from one another?**
   The cluster group with the highest avg. total sales (%) per category is the fist one, followed by 3 with 12.08% less avg. total sales (%). Finally, the cluster group with the lowest number of avg. sales (%) is no. 2.

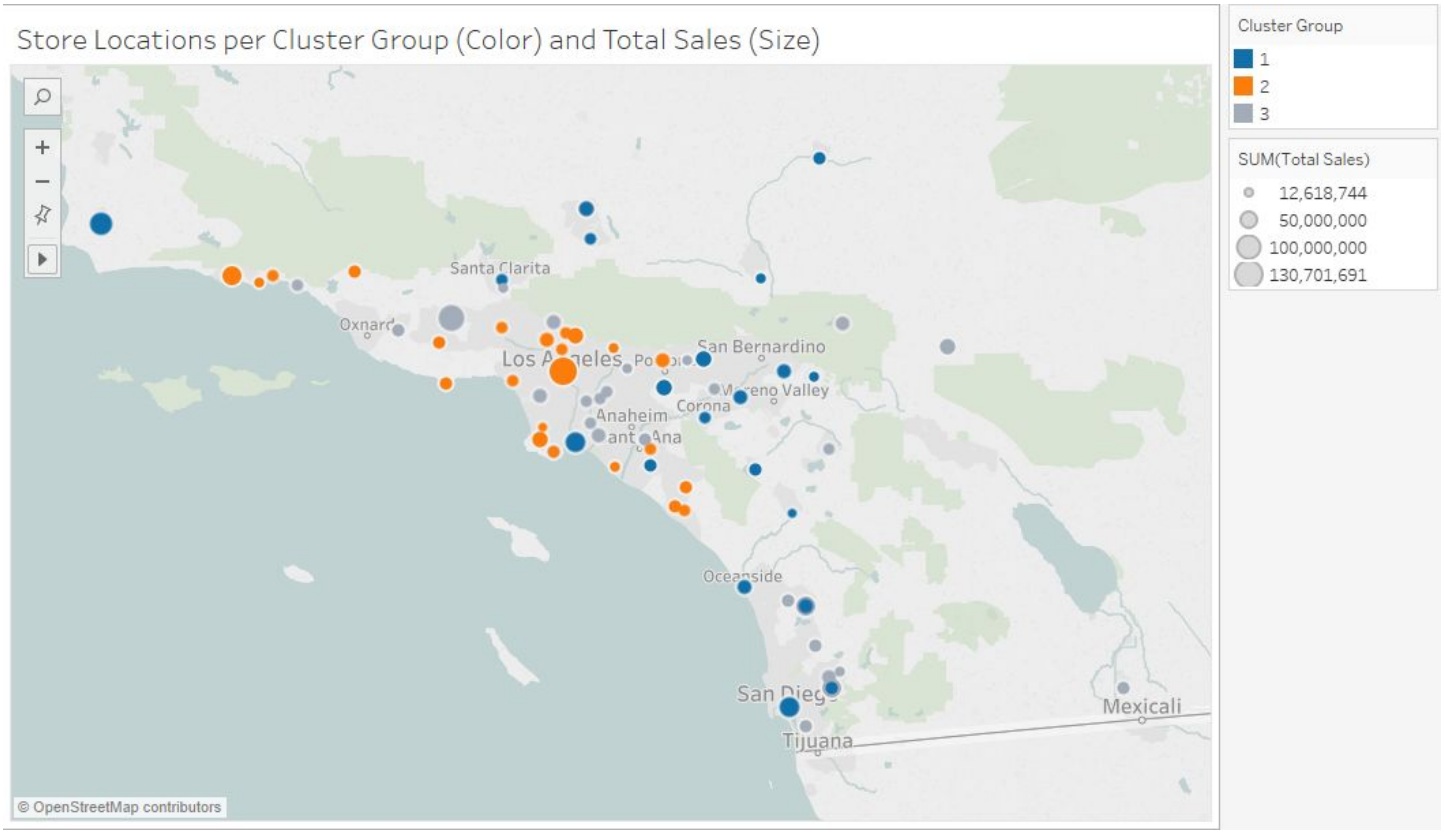


We can see some differences in the below sales categories for each cluster group. Although group 2 has the lowest avg. sales (%) overall, it has the highest avg. sales (%) in frozen food, dairy, bakery, floral and produce compared to the other 2 groups. However, the other 2 groups have higher % in meat, which might be the main reason their sales overall is higher, as meat costs more than the above categories of foods mentioned for cluster group 2.



3. **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.**

   Tableau link:
   https://public.tableau.com/profile/ioanna.vasilopoulou#!/vizhome/CombiningPredictiveTechniques_15565748021770/Task1?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.)**

   We used the Model Comparison tool to test the results from the 3 different models (Decision Tree, Boosted and Forest). Based on the Fit and Error Measures table below we can see that the accuracy for the both Boosted and Forest models is the same, at 0.8235. Decision Tree model is lower than both models, at 0.7059. F1 score, though, which is the precision measure is higher with the Boosted model at 0.8889 compared to the Forest model which was at 0.8426; hence we choose the Boosted Model to predict the best store format for the new stores.

Model Comparison Report

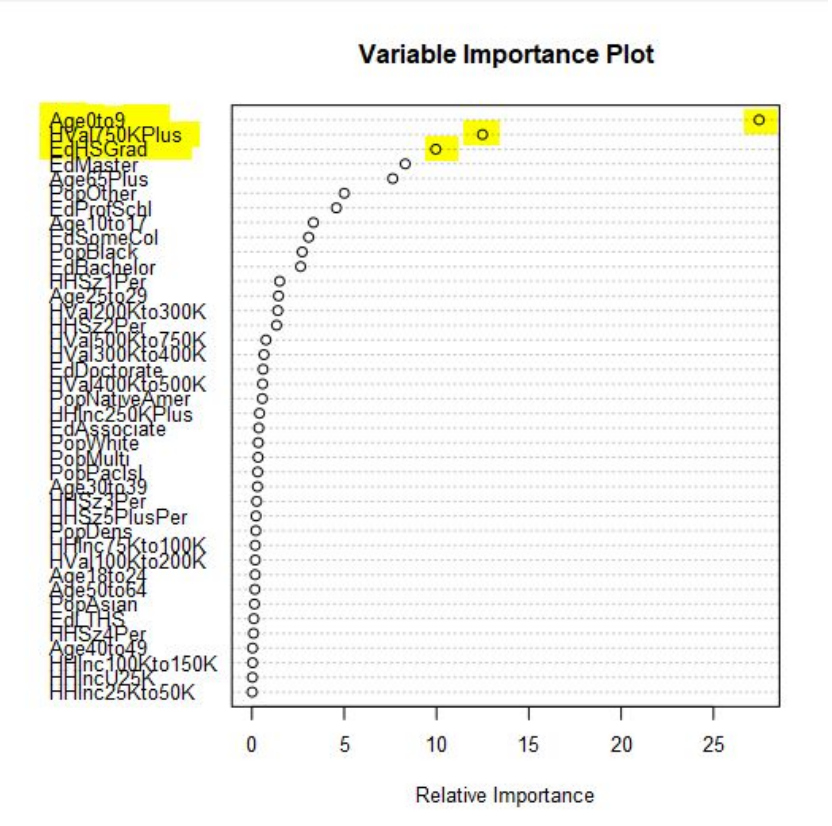
Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DecisionTreeModel	0.7059	0.7685	0.7500	1.0000	0.5556
BoostedModel	0.8235	0.8889	1.0000	1.0000	0.6667
ForestModel	0.8235	0.8426	0.7500	1.0000	0.7778

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.

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

   The 3 most important variables that help explain the relationship between demographic indicators and store formants are: Age 0 to 9, HVal750KPlus and EdHSGrad.



3. **What format do each of the 10 new stores fall into? Please fill in the table below.**
   Using the Boosted Model, the format that he 10 new stores fall into is the below:

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

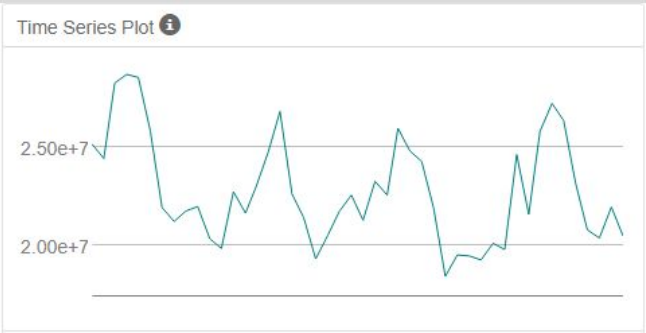
Task 3: Predicting Produce Sales

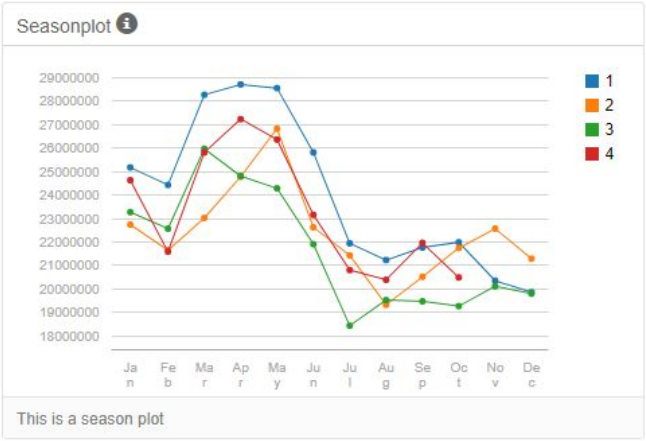
1. **What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?**

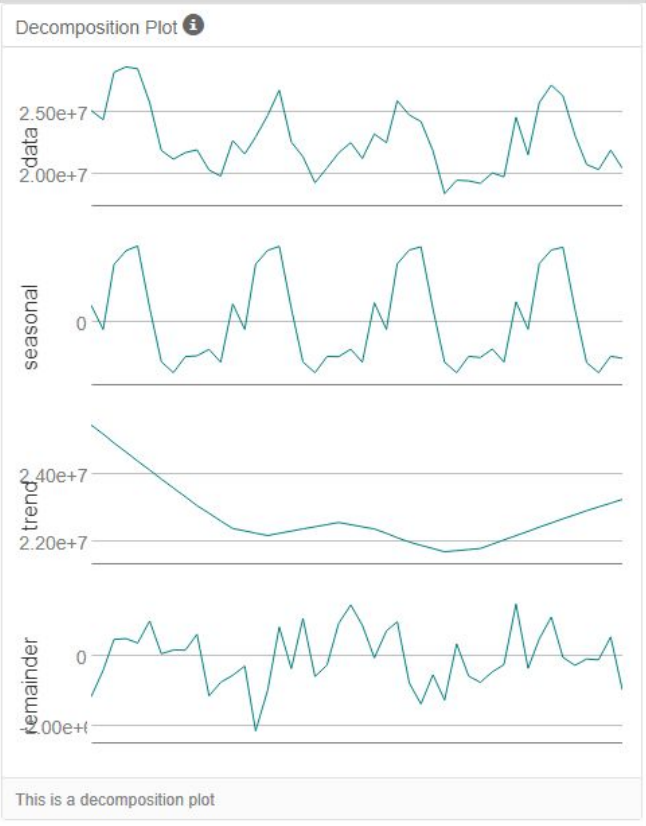
   We need to firstly investigate which model, ETS or ARIMA, would give us better results, in order to forecast the produce sales for existing and new stores. Before doing this, we need to configure these 2 models in terms of the ETS(error, trend, seasonality) and ARIMA(p,d,q)(P,D,Q)m parameters.

   Existing ETS:

   In this case, the error will be applied multiplicatively (M). As there is no clear trend, t will be none (N). Seasonality is also constant so it will be applied multicatively (M). Therefore the model we have is an ETS(M,N,M) model.

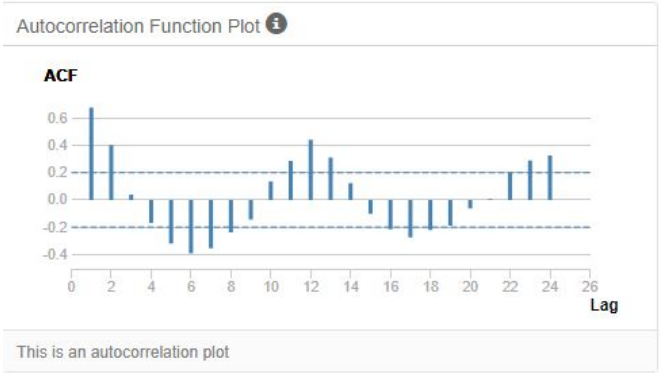


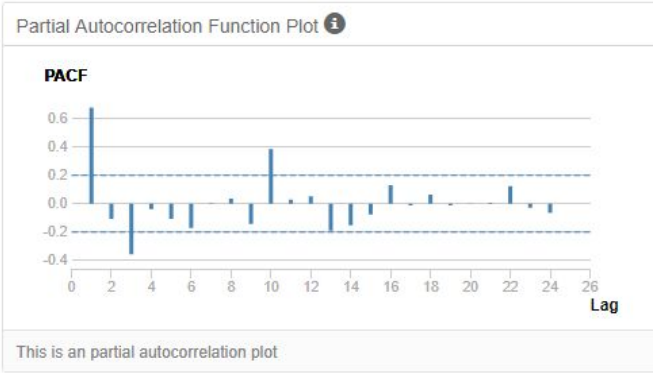




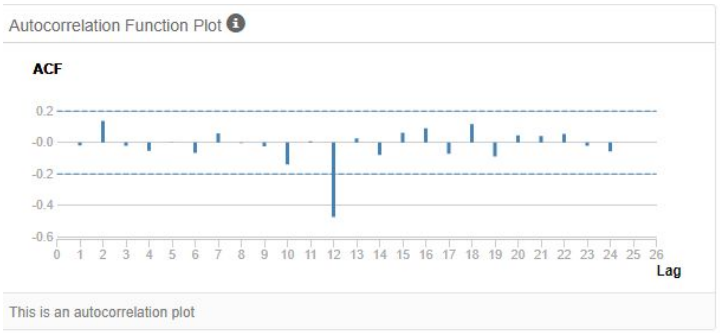
   Existing ARIMA:

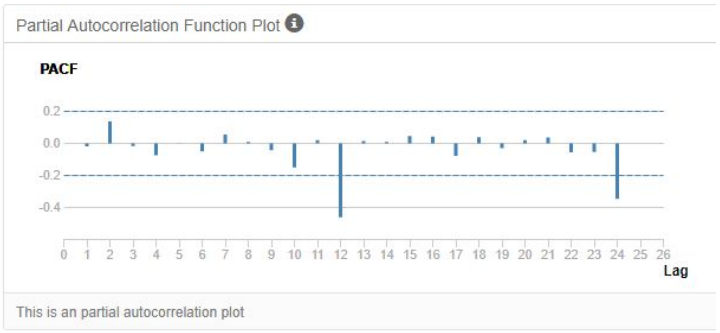
   From the ACF and PACF plots below we can see that our time series needs differencing.





After implementing one non-seasonal and one seasonal differencing the ACF and PACF plots are:





We can see that there is a negative peak at lags 12 and 24 (12-month intervals). For the seasonal terms, we can see that there is a peak negative at 12-month intervals which shows seasonal terms of MA. Therefore the model we have is an ARIMA (0, 1, 1) (0, 1, 1) 12.

By checking the values on the actual and forecast values, we can clearly conclude that the ETS model is better than the ARIMA one, as RMSE and MASE measures are inferior on the ETS model. For the ETS model, RMSE is 1,983,593 and MASE 1.2691 and for the ARIMA model, RMSE is 2,999,244 and MASE is 1.6988.

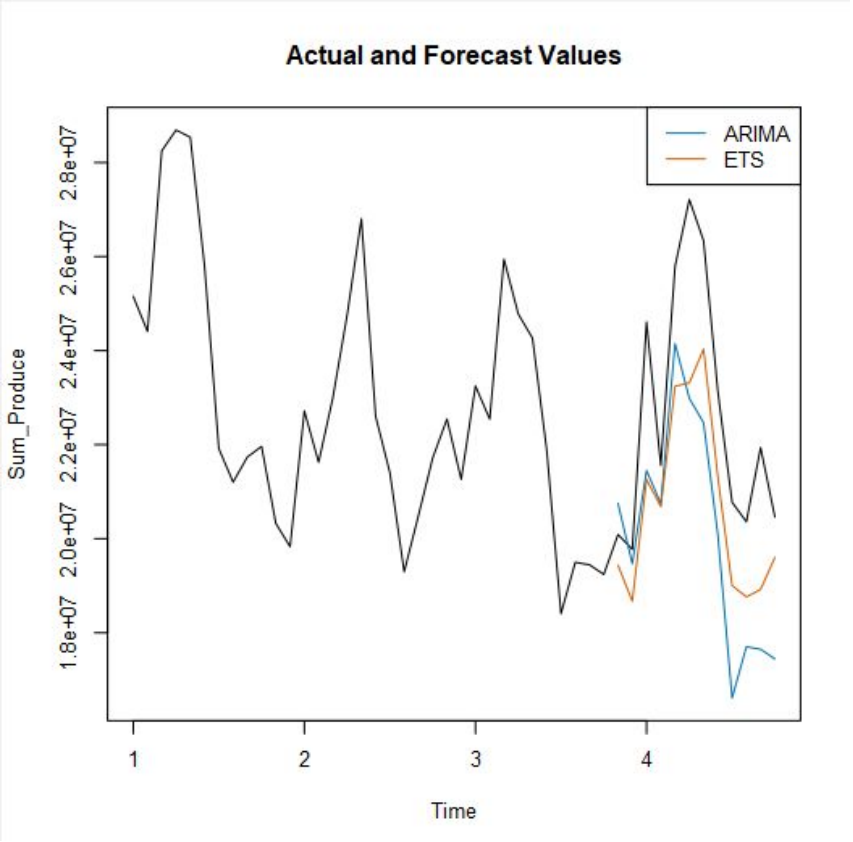
Actual and Forecast Values:

Actual	ARIMA	ETS
20088529.29	20747631.87903	19436519.35454
19772333.34	19465586.61903	18669751.66304
24608406.71	21450342.11903	21257997.96854
21559729.45	20745161.41903	20680556.00271
25792074.59	24146220.24903	23242214.11317
27212464.15	22985351.92903	23312369.36898
26338477.15	22466291.08903	24034277.68653
23130626.6	20083162.35903	21312089.66864
20774415.93	16610437.07903	19002972.03718
20359980.58	17700745.44903	18761864.90984
21936906.81	17647926.66903	18922916.61466
20462899.3	17443558.24903	19600202.20077

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	2545369	2999244	2655219	11.0071	11.5539	1.6988
ETS	1983593	2226513	1983593	8.4729	8.4729	1.2691

On the graph below we can see the time series values alongside the prediction values for both models and observe how the ETS model “behaves” more accurately than the ARIMA model for this dataset.



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.**
After creating 3 clusters to identify the average sales per store and using the ETS (M,N,M) model, we can see the forecast sales for the existing and new stores.

Month	New Stores	Existing Stores
January 2016	2,587,451	21,539,936
February 2016	2,477,353	20,413,771
March 2016	2,913,185	24,325,953
April 2016	2,775,746	22,993,466
May 2016	3,150,867	26,691,951
June 2016	3,188,922	26,989,964
July 2016	3,214,746	26,948,631
August 2016	2,866,349	24,091,579
September 2016	2,538,727	20,523,492
October 2016	2,488,148	20,011,749
November 2016	2,595,270	21,177,435
December 2016	2,573,397	20,855,799

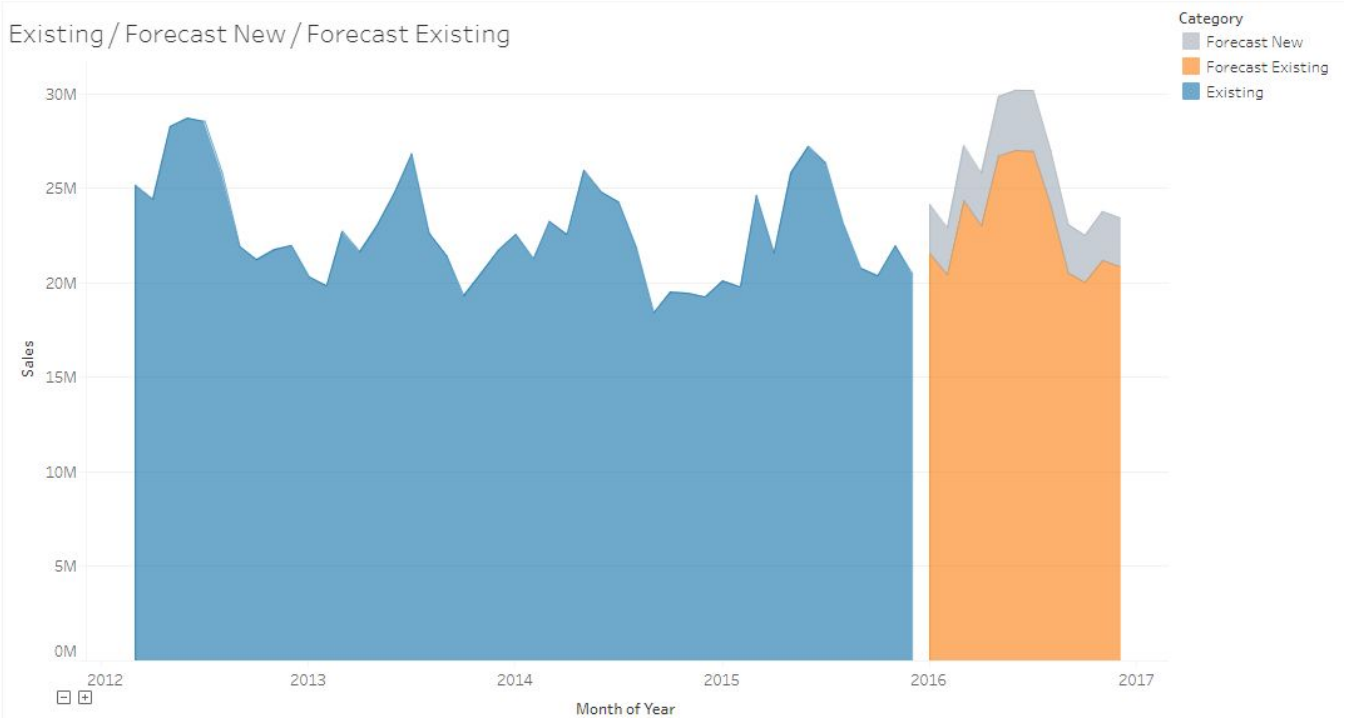


Tableau Link:
https://public.tableau.com/profile/ioanna.vasilopoulou#!/vizhome/CombiningPredictiveTechniques_15565748021770/ExistingForecastNewForecastExisting?publish=yes