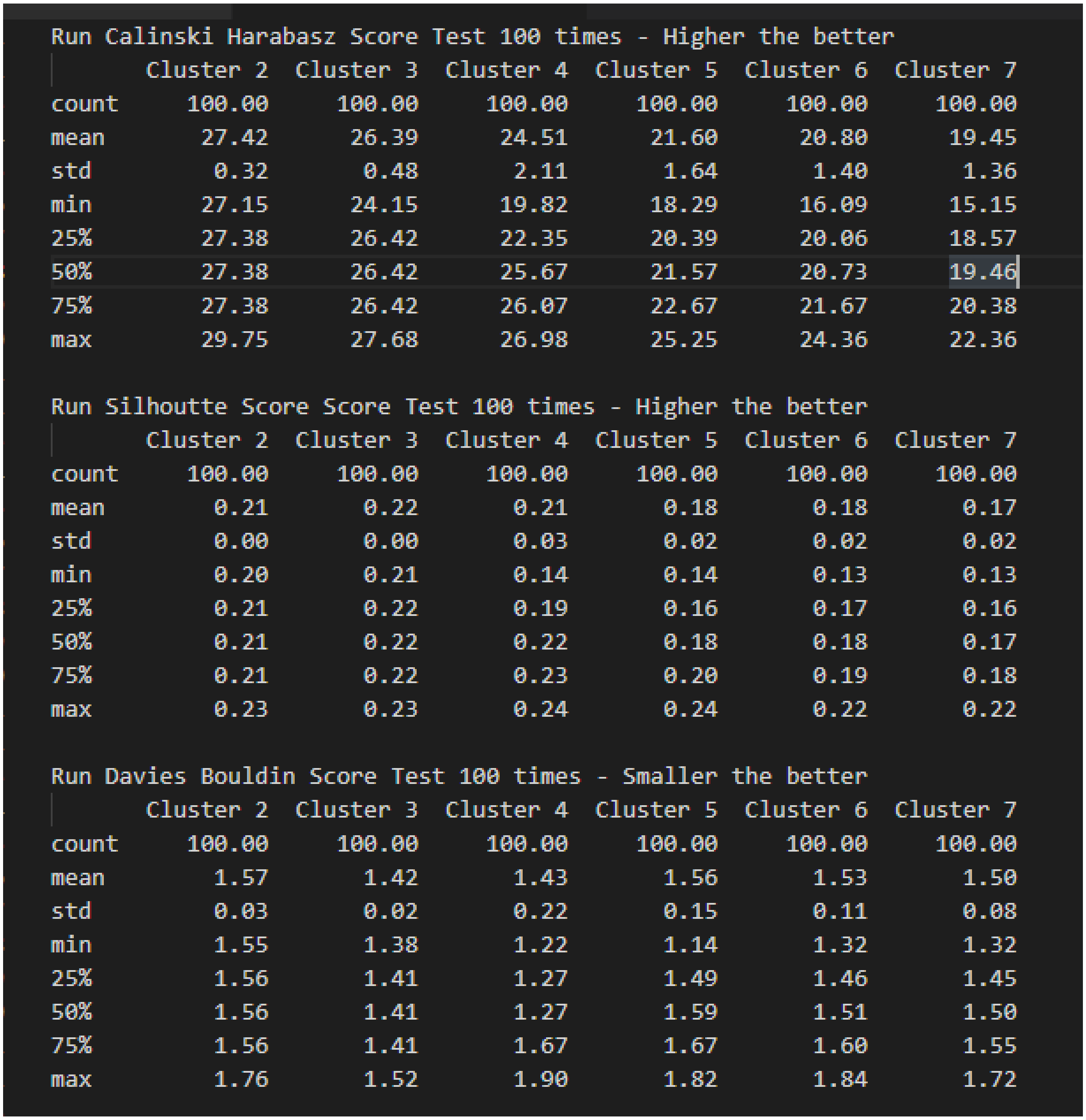
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

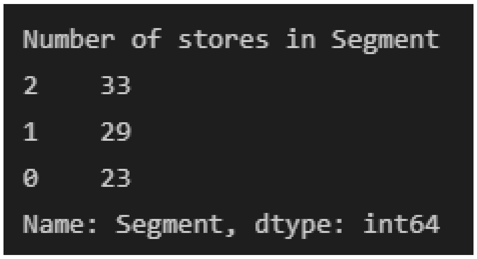
1. What is the optimal number of store formats? How did you arrive at that number?
   The optimal number of store formats is three. The the test used to select optimal number of store format.

   - ✓ Davies Bouldin Score Test show 3 is optimal store format number
   - ✓ Silhoutte Score Test show 3 is optimal store format number
   - ✓ Calinski Harabasz Score Test show 2 is optimal store format number.
   - ✓ Overall I chose 3 as the optimal number of store formats



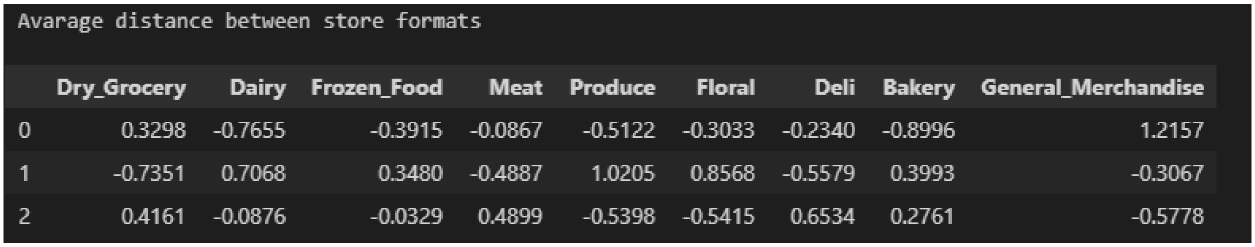
```
Run Calinski Harabasz Score Test 100 times - Higher the better
       Cluster 2  Cluster 3  Cluster 4  Cluster 5  Cluster 6  Cluster 7
count     100.00     100.00     100.00     100.00     100.00     100.00
mean       27.42      26.39      24.51      21.60      20.80      19.45
std         0.32       0.48       2.11       1.64       1.40       1.36
min        27.15      24.15      19.82      18.29      16.09      15.15
25%        27.38      26.42      22.35      20.39      20.06      18.57
50%        27.38      26.42      25.67      21.57      20.73      19.46
75%        27.38      26.42      26.07      22.67      21.67      20.38
max        29.75      27.68      26.98      25.25      24.36      22.36

Run Silhoutte Score Score Test 100 times - Higher the better
       Cluster 2  Cluster 3  Cluster 4  Cluster 5  Cluster 6  Cluster 7
count     100.00     100.00     100.00     100.00     100.00     100.00
mean        0.21       0.22       0.21       0.18       0.18       0.17
std         0.00       0.00       0.03       0.02       0.02       0.02
min         0.20       0.21       0.14       0.14       0.13       0.13
25%         0.21       0.22       0.19       0.16       0.17       0.16
50%         0.21       0.22       0.22       0.18       0.18       0.17
75%         0.21       0.22       0.23       0.20       0.19       0.18
max         0.23       0.23       0.24       0.24       0.22       0.22

Run Davies Bouldin Score Test 100 times - Smaller the better
       Cluster 2  Cluster 3  Cluster 4  Cluster 5  Cluster 6  Cluster 7
count     100.00     100.00     100.00     100.00     100.00     100.00
mean        1.57       1.42       1.43       1.56       1.53       1.50
std         0.03       0.02       0.22       0.15       0.11       0.08
min         1.55       1.38       1.22       1.14       1.32       1.32
25%         1.56       1.41       1.27       1.49       1.46       1.45
50%         1.56       1.41       1.27       1.59       1.51       1.50
75%         1.56       1.41       1.67       1.67       1.60       1.55
max         1.76       1.52       1.90       1.82       1.84       1.72
```

2. How many stores fall into each store format?



```
Number of stores in Segment
2    33
1    29
0    23
Name: Segment, dtype: int64
```

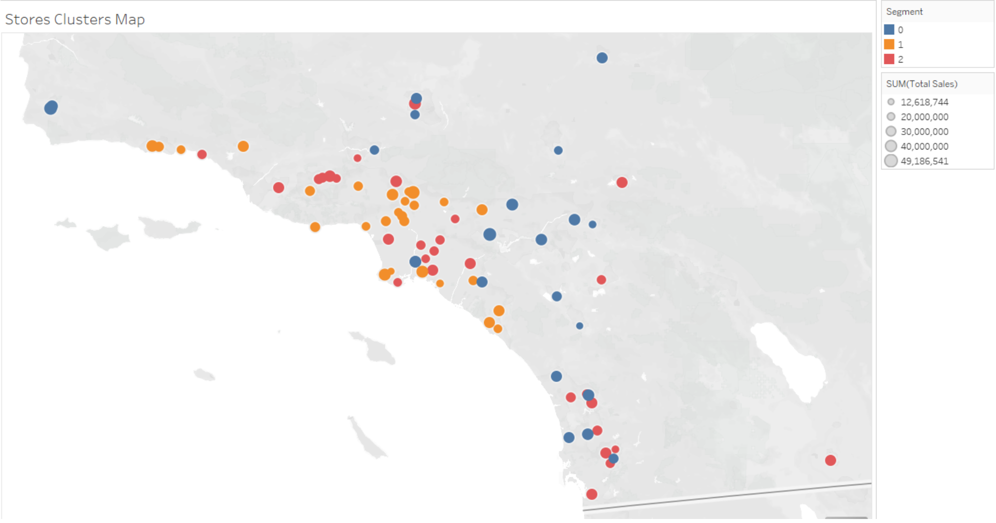
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   - ✓ Store format 1 has negative average distance in dry grocery compared to other store formats
   - ✓ Store format 0 sales more General merchandise than the rest
   - ✓ Store format 1 sales more Produce than the rest
   - ✓ Store format 2 sales more Deli than the rest



Avarage distance between store formats

	Dry_Grocery	Dairy	Frozen_Food	Meat	Produce	Floral	Deli	Bakery	General_Merchandise
0	0.3298	-0.7655	-0.3915	-0.0867	-0.5122	-0.3033	-0.2340	-0.8996	1.2157
1	-0.7351	0.7068	0.3480	-0.4887	1.0205	0.8568	-0.5579	0.3993	-0.3067
2	0.4161	-0.0876	-0.0329	0.4899	-0.5398	-0.5415	0.6534	0.2761	-0.5778

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.

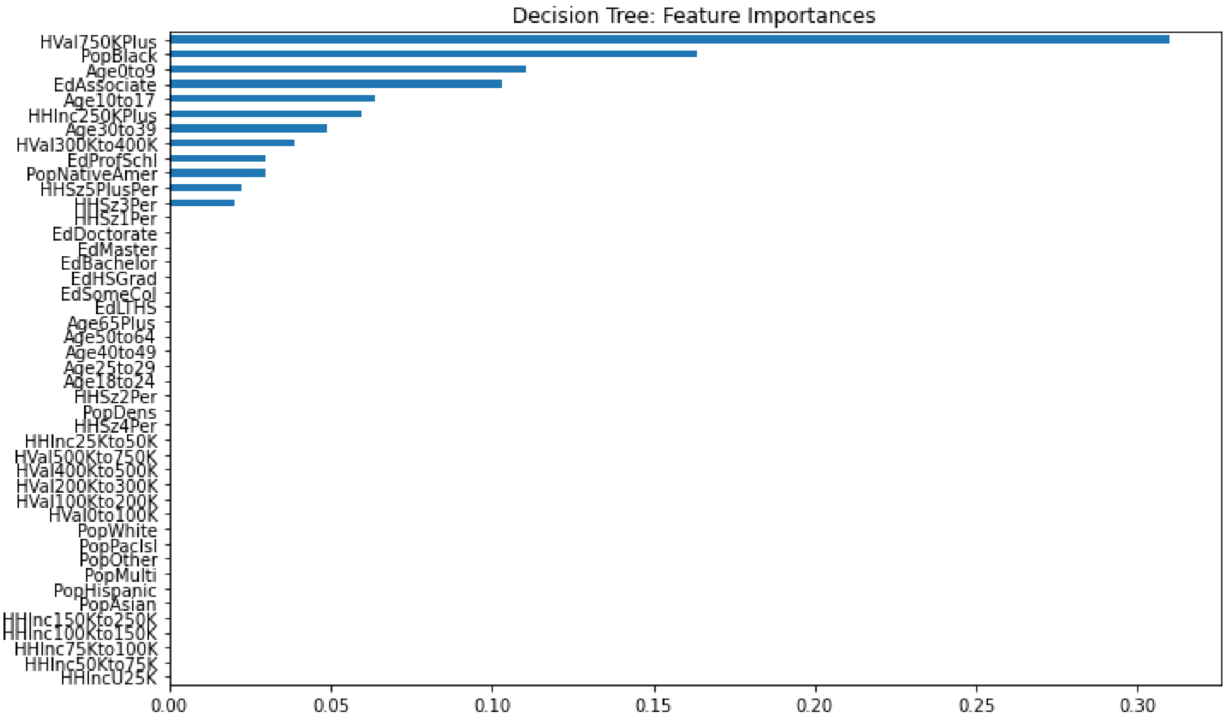


Task 2: Formats for New Stores

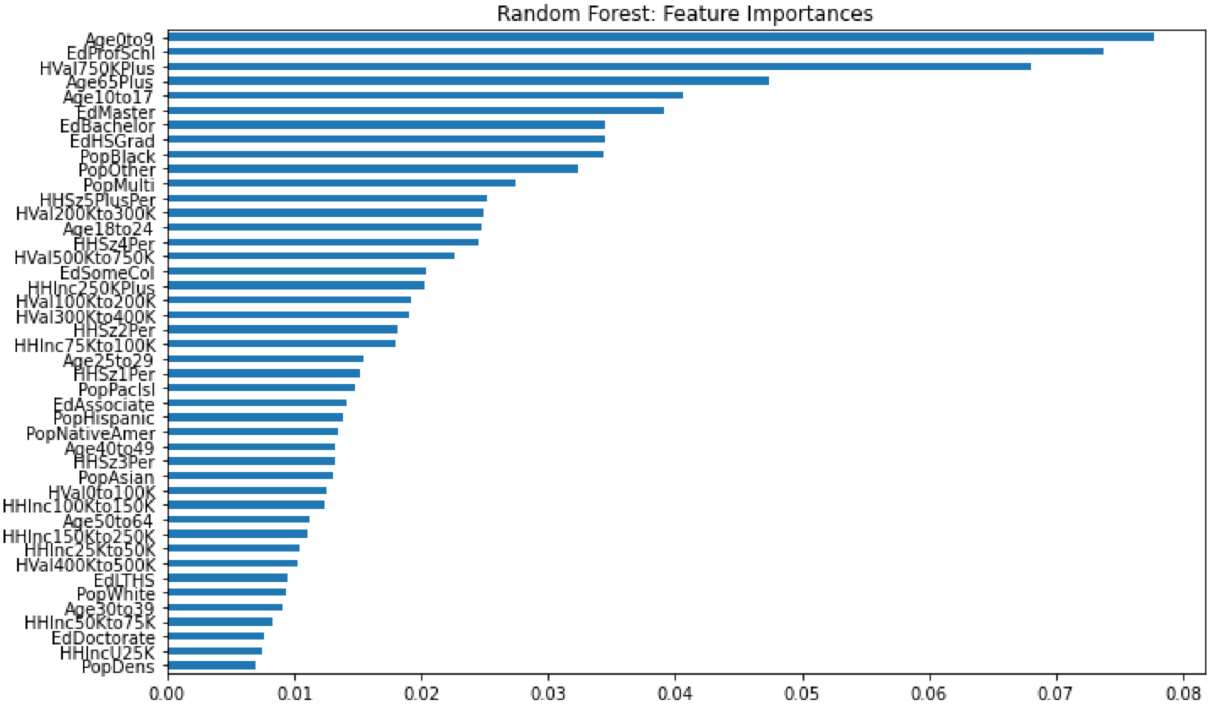
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

    ✓ I tested Decision Tree, Forest and Boosted models to find best model to segment stores

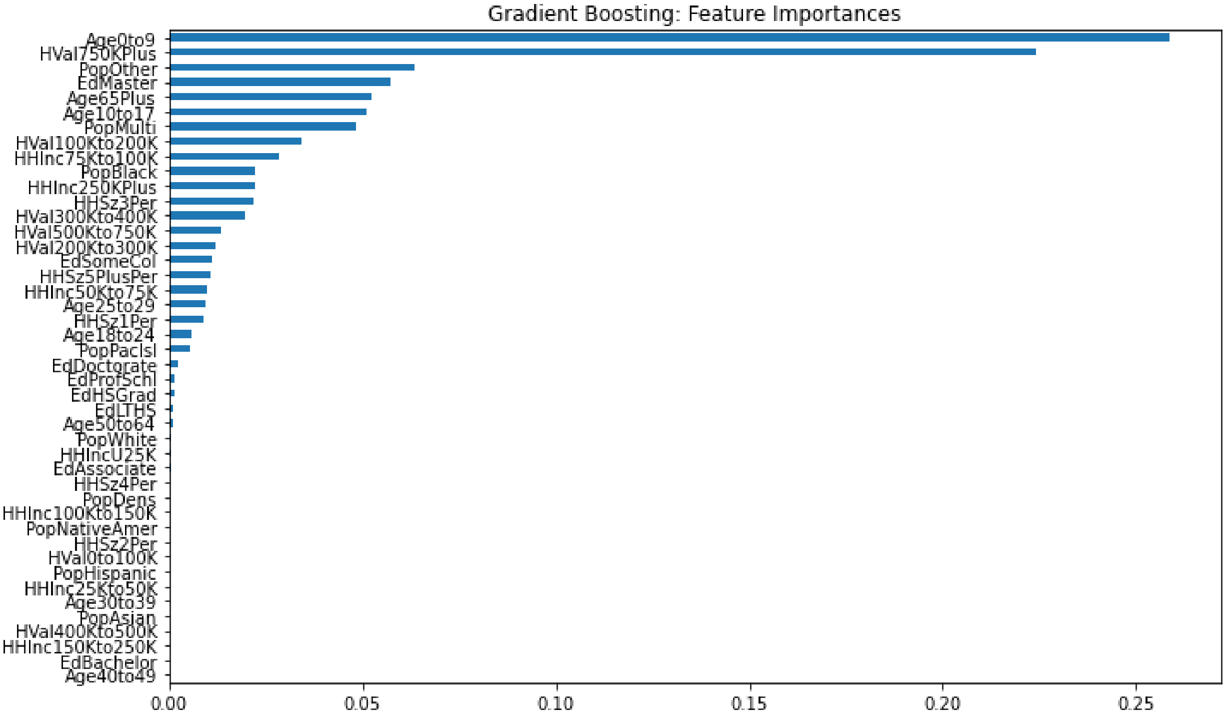
    ✓ Decision Tree Model top three important variables are HVal750KPlus, PopBlack and Age0to9.



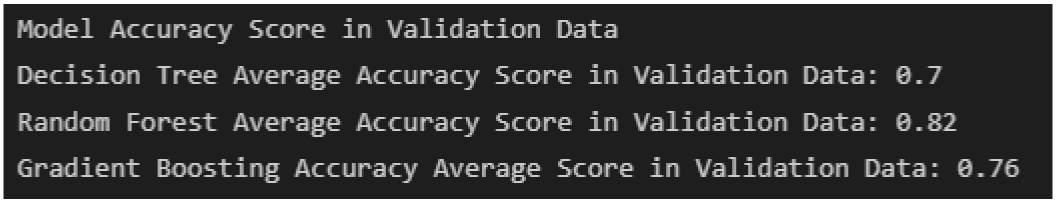
- ✓ Forest Model top three important variables are Age0to9, EdProfSchl, and HVal750KPlus.



- ✓ Boosted Model top three important variables are Age0to9, HVal750KPlus, and PopOther.



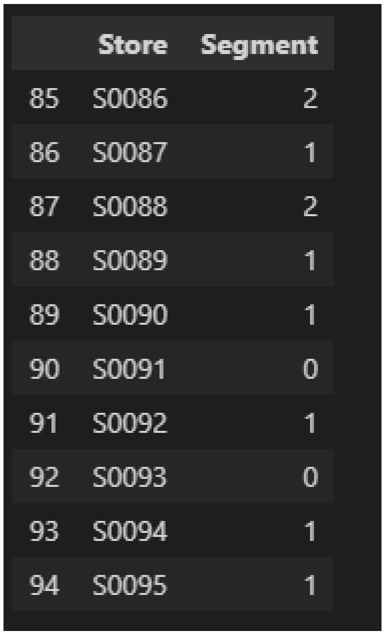
✓ I use average accuracy score on validation data to choose the best model.



```
Model Accuracy Score in Validation Data
Decision Tree Average Accuracy Score in Validation Data: 0.7
Random Forest Average Accuracy Score in Validation Data: 0.82
Gradient Boosting Accuracy Average Score in Validation Data: 0.76
```

✓ I will use Random Forest model due to higher score of 82% accuracy in validation data.

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

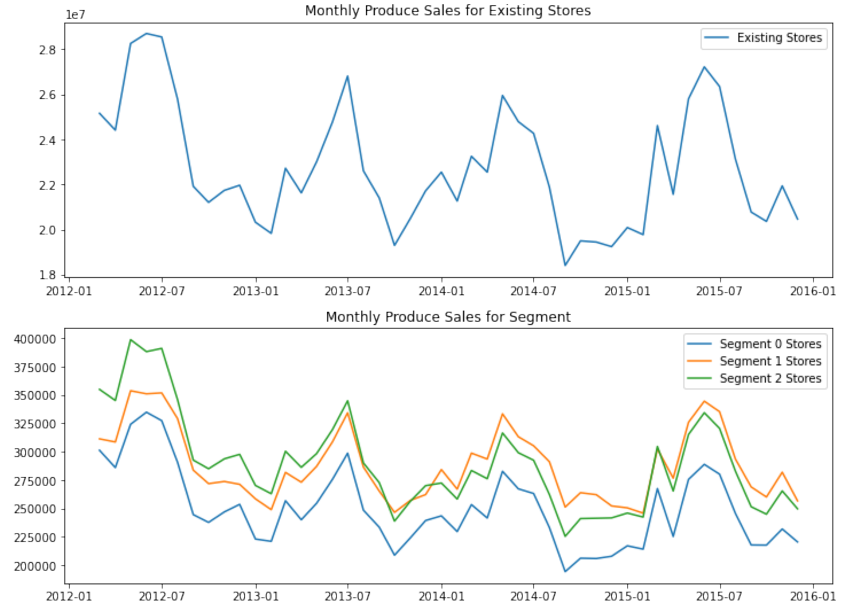


	Store	Segment
85	S0086	2
86	S0087	1
87	S0088	2
88	S0089	1
89	S0090	1
90	S0091	0
91	S0092	1
92	S0093	0
93	S0094	1
94	S0095	1

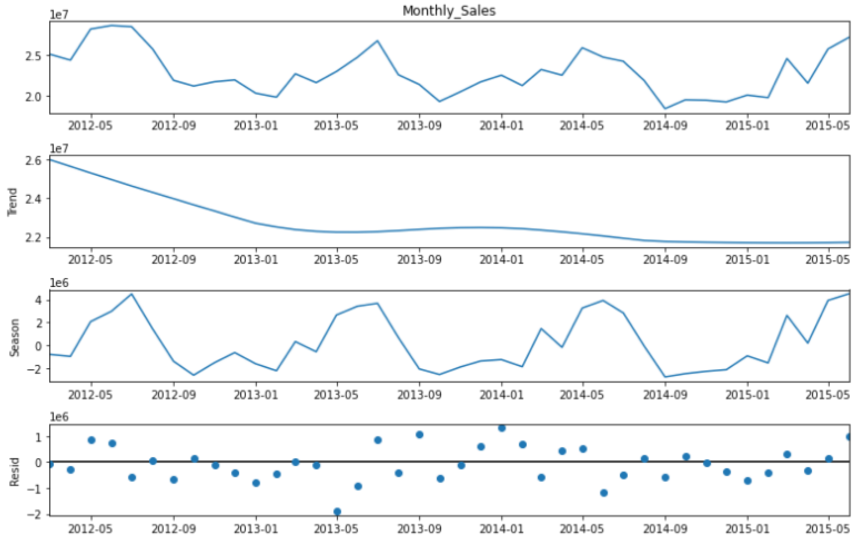
Task 3: Predicting Produce Sales

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?

   ✓ Sales Plot



   ✓ Time Decomposition Plot

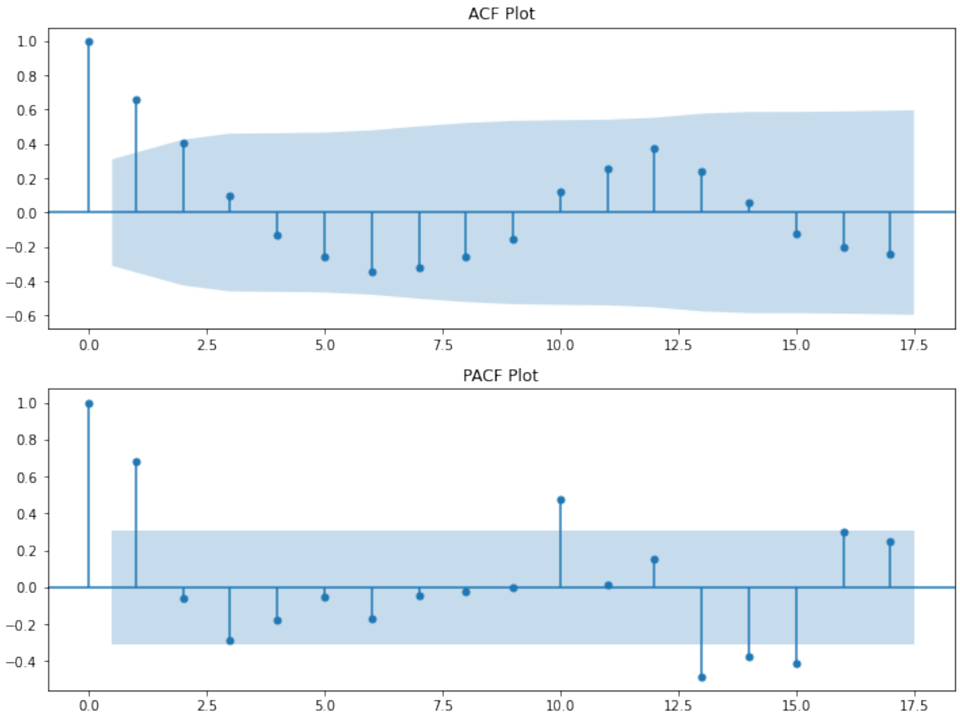


   ✓ I use ETS(M,A,A) with damped trend to compare with ARIMA model

✓ ACF and PACF Plot

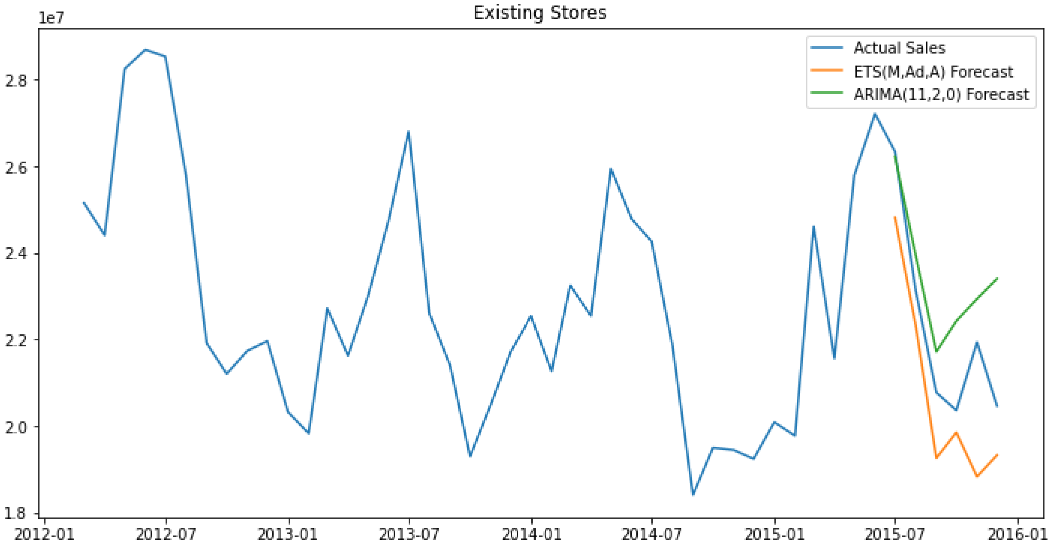
ACF plot show strong significance in lag 2. PACF show strong significance in lag 2 and 11.

ACF and PACF favor AR as both change from positive to zero in there first two lags.

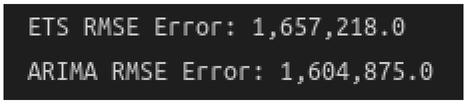


✓ I use ARIMA(11,2,0) to compare with ETS model

- ✓ Holdout sample forecast

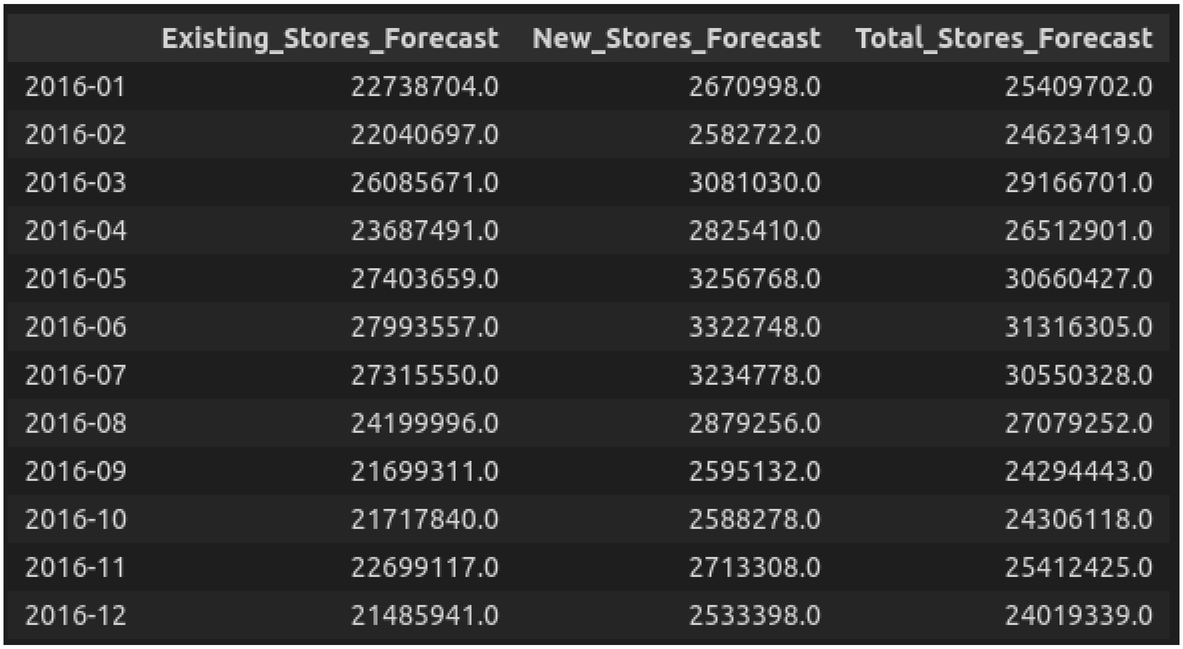


- ✓ I will chose ARIMA Model to forecast 2016 sales due to lower error and close follow up the sales trends in holdout sample



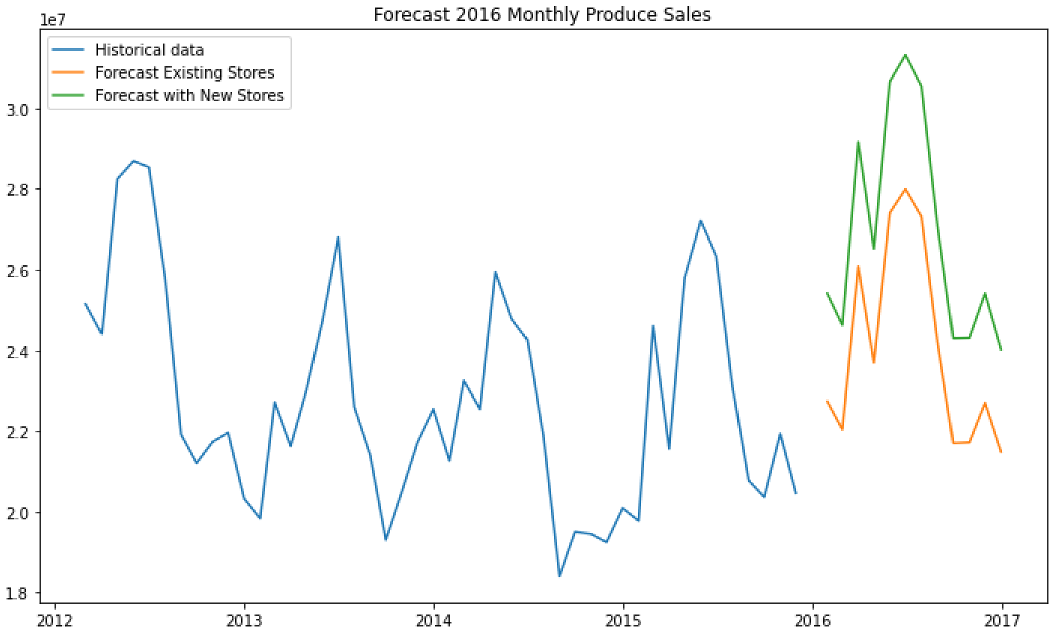
```
ETS RMSE Error: 1,657,218.0
ARIMA RMSE Error: 1,604,875.0
```

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.
   - ✓ Forecasted sales for 2016



	Existing_Stores_Forecast	New_Stores_Forecast	Total_Stores_Forecast
2016-01	22738704.0	2670998.0	25409702.0
2016-02	22040697.0	2582722.0	24623419.0
2016-03	26085671.0	3081030.0	29166701.0
2016-04	23687491.0	2825410.0	26512901.0
2016-05	27403659.0	3256768.0	30660427.0
2016-06	27993557.0	3322748.0	31316305.0
2016-07	27315550.0	3234778.0	30550328.0
2016-08	24199996.0	2879256.0	27079252.0
2016-09	21699311.0	2595132.0	24294443.0
2016-10	21717840.0	2588278.0	24306118.0
2016-11	22699117.0	2713308.0	25412425.0
2016-12	21485941.0	2533398.0	24019339.0

   - ✓ Forecast for 2016 plot


Forecast 2016 Monthly Produce Sales
1e7
Historical data
Forecast Existing Stores
Forecast with New Stores