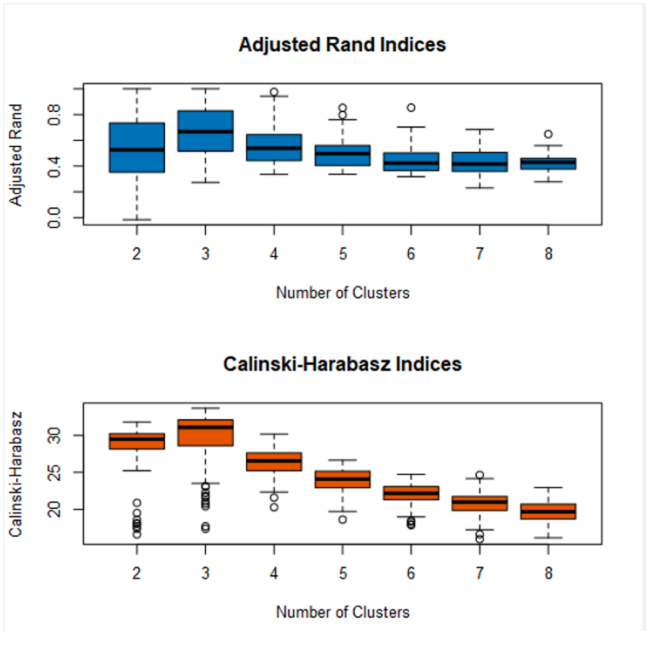
Project: Combining Predictive Techniques

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 is 3. I used a K-Means clustering method



: Awesome: Great job! Yes, the RAND and CH indices indicate that 3 clusters is optimal, so we chose 3 for the number of formats.

2. **How many stores fall into each store format?**

The analysis using K-means was performed and the distribution is presented below

Cluster	Size
1	23
2	29
3	33

: Awesome: The number of stores in each format is correct - great job!

Cluster 1, 2 and 3 have 23, 29 and 33 stores respectively

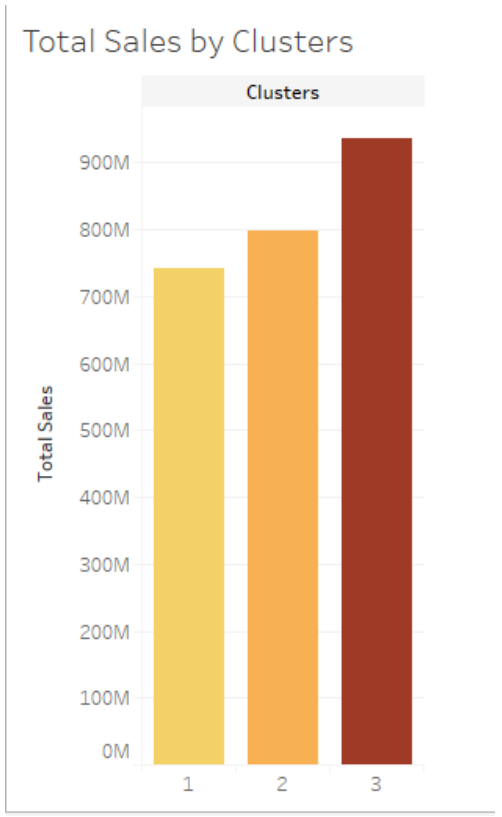
3. **Based on the results of the clustering model, what is one way that the clusters differ from one another?**

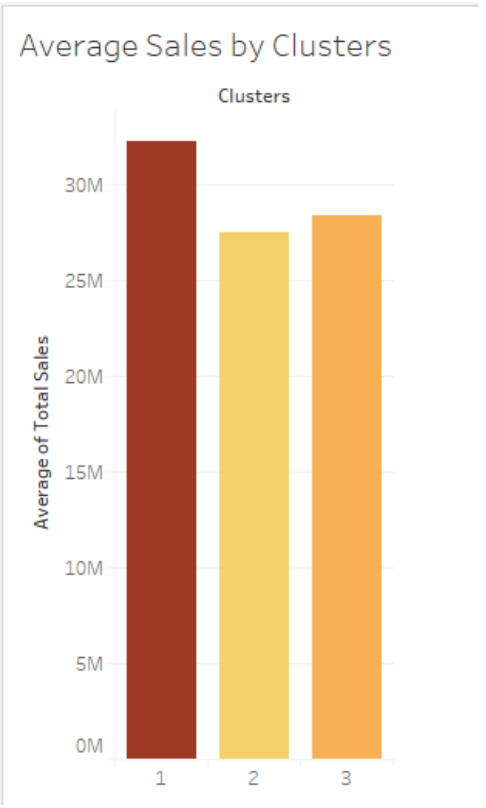
From the report of the K-Means Clustering, we can see the Cluster 1 has the lowest size which is 23 and cluster 3 has the highest size. I also observed that cluster 3 has the highest total sales while cluster 1 sells more on average sales.

: Awesome: Excellent work providing observations about the difference among the clusters in terms of total sales and size.

Cluster Information:

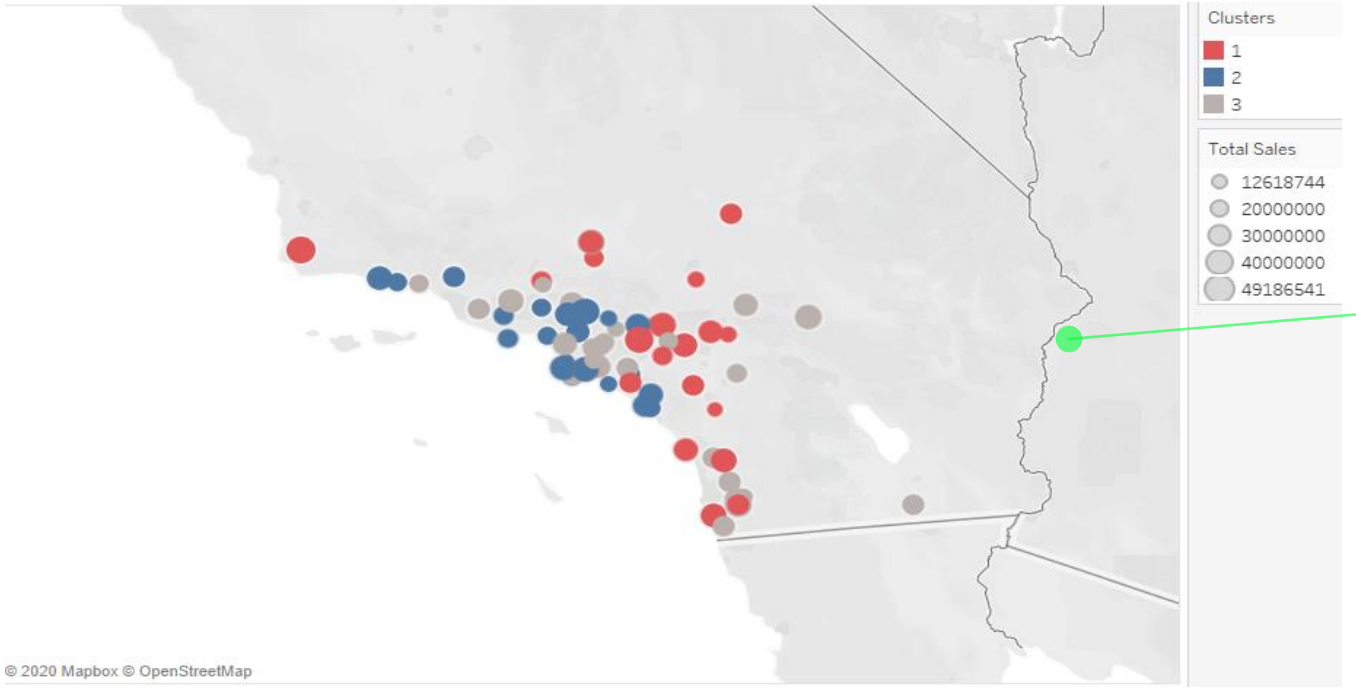
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843





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.**

The visualization is here



: Awesome: The map looks great. Color is used to show the clusters and size is used 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 three models (Decision Tree Model, Boosted Model, and Forest Model). And I used the Boosted Model to predict the best store format for the new stores after. This is based primarily on the accuracy and F1 values. Though the accuracy of Boosted Model and Forest Model are the same (82.35% each), the boosted model is better in F1 value

: Awesome: Great job! Yes, the Boosted model should be used since it has high accuracy and higher F1 score. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_Model	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778

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

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

: Awesome: The stores are correctly segmented - great job!

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?

I used ETS model for forecast. I came to this decision after making comparison between ETS and ARIMA.

The decomposition plot using TS Plot tool is presented above. In doing this, sales is aggregated in all stores per month to make a forecast. The time series is decomposed into three time series – seasonal component, the trend component and the remainder. The ETS Model is built by examining all these components.

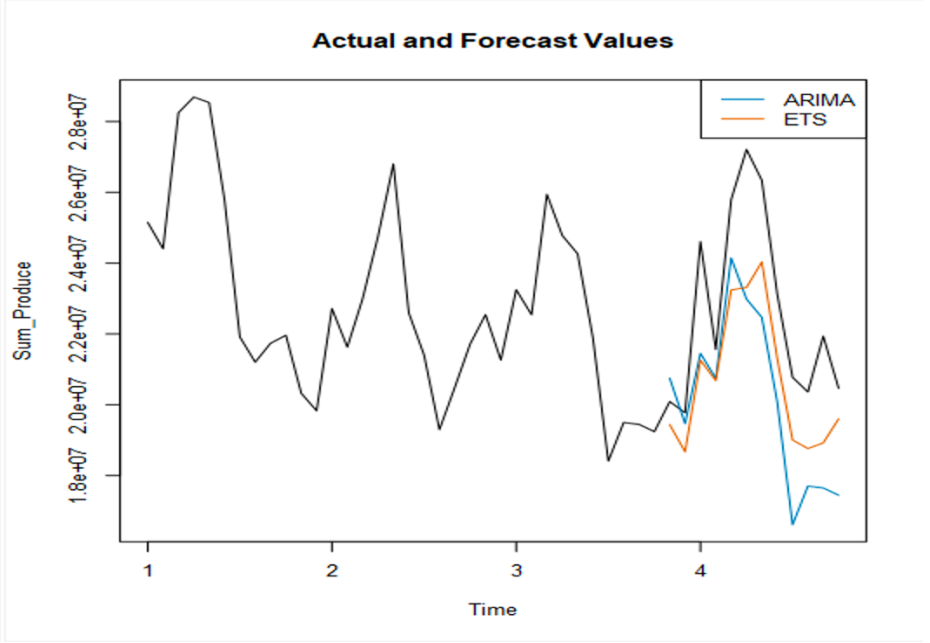
The remainder/error is increasing in variance hence it would be multiplicative. The plot also shows no clear trend and we can say at best that sales is fluctuating at somewhat similar intervals

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	2545369	2999244	2655219	11.0071	11.5539	1.6988
ETS	1761302	1978476	1761302	7.5704	7.5704	1.1269

: Awesome: Great work presenting the forecast accuracy measures against the holdout sample as a justification for selecting ETS. Yes, from the values in the table we can see that ETS has
lower errors so it has better predictive qualities.

Also on the basis of the accuracy measures and in-sample error measurements, the ETS (M.N,M) performed better than the ARIMA model. The RMSE, MASE are lower and therefore better in the ETS Model than in the ARIMA Model. Though the ARIMA appears better on the basis of Akaike Info. Criterion (AIC), the ETS Model outperformed it on many of the other indices

Forecasts of the two model was also compared in the graph below:



The black line represents the actual sales. The graph shows all time series values and forecast values for the compared models.In the test we can see how the ETS model behaves more accurately than the ARIMA model for this data set, that it, its forecasts are closer to the actual values than the ARIMA Model.

: Awesome: Great job! Yes, we should use ETS(MNM) as the type of ETS model. By looking at the decomposition plot we can see that there is quite a bit of seasonality. From plot, we can also see that the trend turns up at the end, so trend should not be applied, and it appears that the remainder changes in magnitude, so we should apply it multiplicatively.

: Suggestion: It would be good to show the tested ARIMA model. If just the last 6 records are used as a holdout sample he auto option in the ARIMA tool should normally suggest ARIMA(1,0,0)(1,1,0)[12] where seasonal differencing is applied just. And then AR terms are selected. No regular differencing is applied because in some cases AR terms can be used in the role of a non-seasonal differencing. In that case, the series go into category ”underdifferenced”. Please check the project review section for a link to rule 6 where that is stated in more details: “Rule 6: If the PACF of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is positive--i.e., if the series appears slightly “underdifferenced”--then consider adding an AR term to the model.The lag at which the PACF cuts off is the indicated number of AR terms.” Also, we do not have a trend in the series so that is why we should not apply non-seasonal differencing since it used to remove the trend. Since we don’t have trend there is no point of adding a non-seasonal differencing.

: Suggestion: Please note that normally you cannot compare the AIC from an ETS model with the AIC from an ARIMA model. The two models treat initial values differently. For example, after differencing, an ARIMA model is computed on fewer observations, whereas an ETS model is always computed on the full set of data. Please check the project review for a link where you can read more about that.

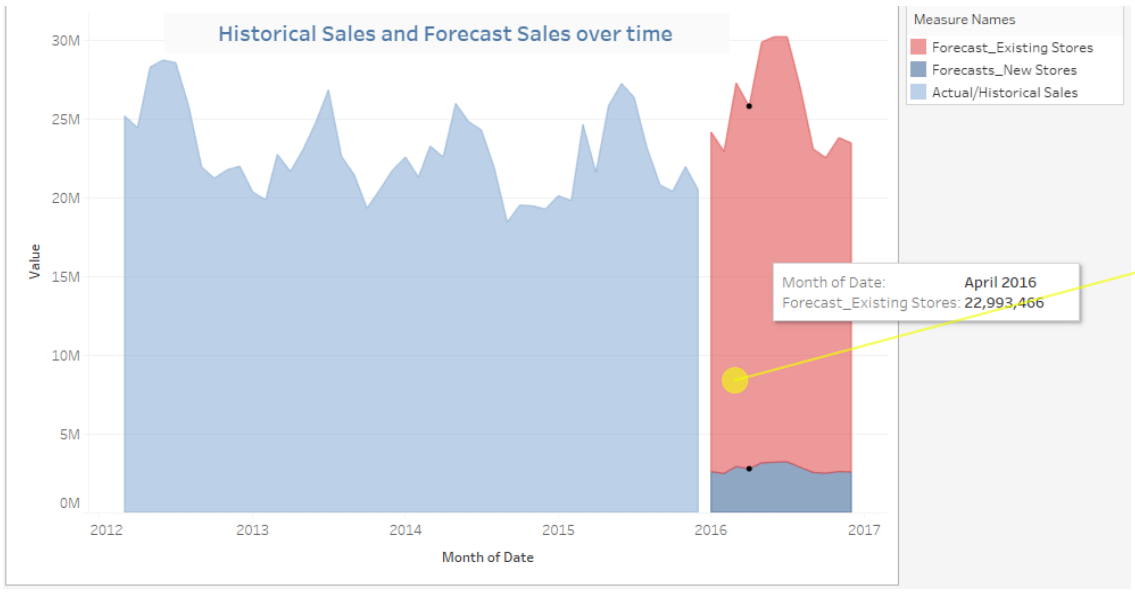
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Forecast of sales for existing and new stores is presented below:

Year of Date	Month of Date	Forecast of Sales for Existing Stores	Forecast of Sales for New Stores
2016	January	21,539,936.01	2,587,450.85
2016	February	20,413,770.60	2,477,352.89
2016	March	24,325,953.10	2,913,185.24
2016	April	22,993,466.35	2,775,745.61
2016	May	26,691,951.42	3,150,866.84
2016	June	26,989,964.01	3,188,922.00
2016	July	26,948,630.76	3,214,745.65
2016	August	24,091,579.35	2,866,348.66
2016	September	20,523,492.41	2,538,726.85
2016	October	20,011,748.67	2,488,148.29
2016	November	21,177,435.49	2,595,270.39
2016	December	20,855,799.11	2,573,396.63

: Awesome: The forecasts for the existing and new stores are within the expected range - great job! Great job plotting the results!

The visualization is here



: Suggestion: Great job with the plot! I would suggest moving the forecast for the new stores on top of the sales for existing stores because the goal of the visualization is to show the total forecasted sales so by stacking it on top it makes it more clear the impact of the new stores to the total. By the way, if you are interested in how to close the gap in the plot between the actual sales and the forecasted ones you can check the example in the project review section.