

Combining Predictive Techniques

REVIEW	HISTORY
--------	---------

Meets Specifications

BT - Boosted model, DT - Decision Tree, RF - Random Model

Thank you for your note. The BT, DT, RF models are correctly trained. Great job resolving the issues mentioned by the previous reviewer! **You have done fantastic work on this project!** Thorough and well-presented analysis! One of the things that I liked is how you used the plots of the forecasted vs actual values and the forecasting errors against the holdout sample as a justification for selecting ETS over the ARIMA model. Great job! That is one of the toughest parts for students to get right. If you want to know more about those different error values are you can read about them here - [Forecast accuracy measures](#) and [Three types of forecasts: estimation, validation, and the future.](#)

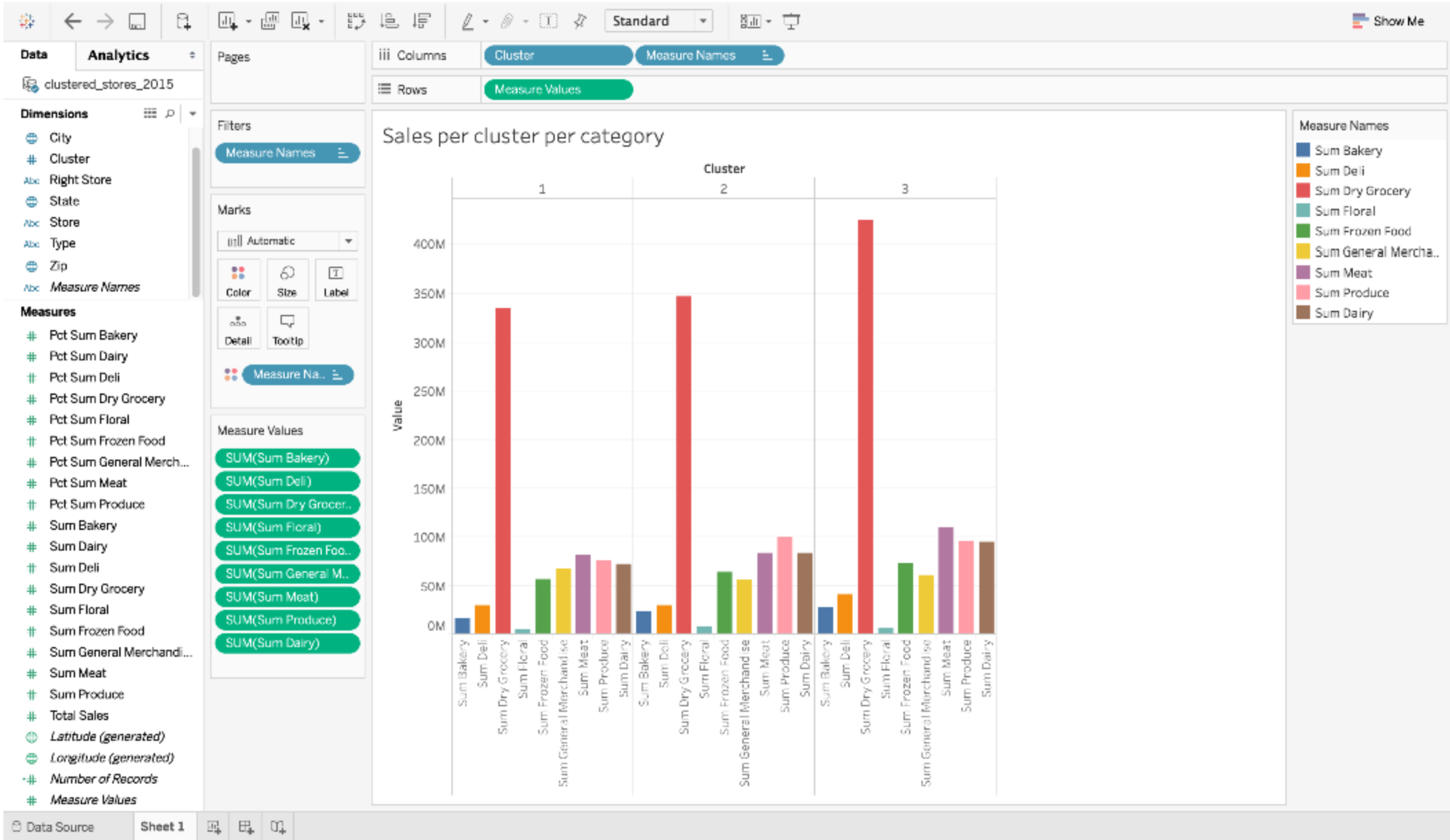
Please check the comments below for some additional suggestions and notes. I hope you will find them useful.

Congratulations on completing this program! I wish you all the best and good luck in your future endeavors! 🎉

Overall

✓	The write up is written clearly, in complete sentences, and without major typos.
✓	Several visualizations are included. All visualizations are clearly labeled and help answer the related questions.
	Great presentation of all of the plots! All visualizations are clearly labeled and help answer the related questions.

Task 1

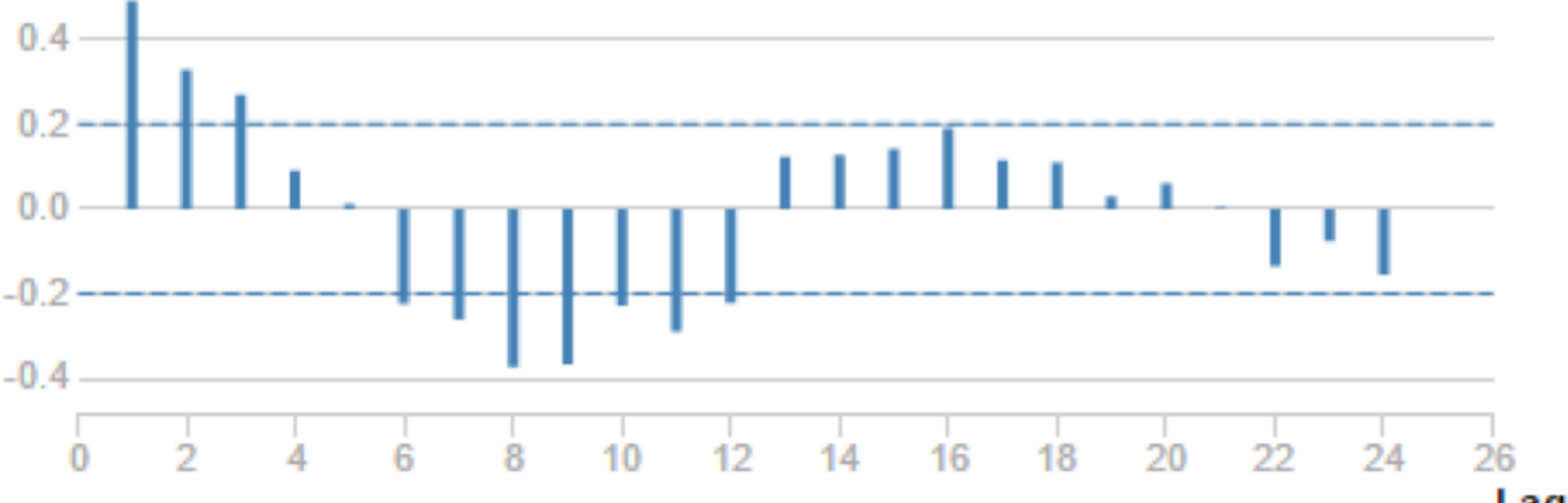
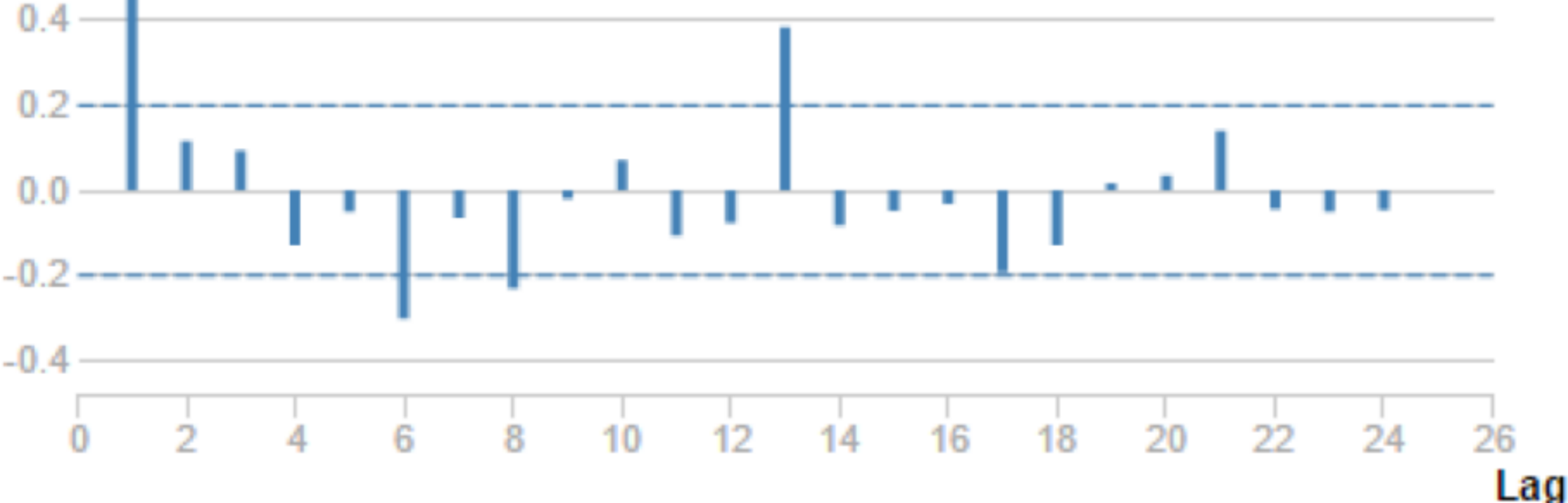
✓	Accurately identifies the correct number of formats and provides justification using the Adjusted Rand and CH indices.
	Indeed! The RAND and CH indices indicate that 3 clusters are optimal due to high median/mean value and compactness and fewer outliers, so we chose 3 for the number of formats. Great job!
✓	Identifies the correct number of stores that fall into each store format.
✓	Provides one observation about the differences among clusters, and uses the results of the clusters to provide justification.
	<p>Indeed, clusters differ in terms of size, distance, and separation. Great job!</p> <p>Suggestion:</p> <p>What we could also do here is provide a visual that shows category sales values per cluster. I am suggesting that because the z scores not always match the direction of the values and PC plot is not very easy to interpret. Meaning, the value could be negative or low but in reality, if you look at the average sales or the summed sales you can see a different picture. That is the practical thing that a manager will want to know. Here is an example:</p> 

✓	Includes a map that shows the location of the stores, uses color to show cluster, and size to show total sales. A legend is used for both color and size.
	The map looks great. Color is used to show the clusters and size is used to show total sales. Legends for color and size are presented.

Task 2

✓	States the type of classification model used and adequately justifies the choice using at least one model comparison method.
	We can use either the Boosted or Forest model since they have similar accuracy and F1 score. Great job selecting the Boosted model.
✓	Includes a table that correctly identifies the format for each of the 10 new stores.

Task 3

✓	<p>Compares and identifies the best ETS or ARIMA model to use for forecasting. Justifies the decision by showing the plot and shows forecast error measurements against the holdout sample.</p> <p>ETS(M,N,M) is the correct model ETS type of model! Well done justifying your choice by presenting the decomposition plot. Excellent work selecting ETS over ARIMA and justifying that choice by providing the forecast error measurements against the holdout sample.</p> <p>Some additional notes on the selection of the optimal ARIMA:</p> <p>For the type ARIMA model, great job selecting - 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”. Here you can read more about it - “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 non-seasonal differencing.</p> <p>Furthermore, note that as you have learned in the previous project when you have seasonality in your series you first need to seasonally difference the series by applying D(1) term. Here is how the ACF PACF after seasonal difference was applied should look like:</p> <div><div><p>ACF</p><p>Lag</p></div><div><p>This is an autocorrelation plot</p></div><div><p>Partial Autocorrelation Function Plot ⓘ</p><div><p>PACF</p><p>Lag</p></div></div></div> <p>Which can be obtained from here:</p>
---	--

Update Existing Field

Create New Field

Season Diff

Type

Double

Size

8

Num Rows

Values for Rows that don't Exist

12

NULL

Group By (Optional)

RecordID

Month

Sum_Produce

Year

^

Clear

Variables

Functions

Saved Expressions

Row -12

Row -11

Row -10

Row -9

Row -8

Row -7

Row -6

Row -5

Row -4

Row -3

Row -2

Expression:

[Sum_Produce] - [Row-12:Sum_Produce]

If you are getting errors in the previous tool check this [Knowledge post](#). And as I have explained above in our case this time you will just need a seasonal differencing and seasonal and non-seasonal AR to obtain a series where ACF and PACF results shows no significantly correlated lags suggesting no need for adding additional AR() or MA() terms.

The figure consists of two vertically stacked plots. The top plot is the Autocorrelation Function (ACF) plot, with the y-axis labeled 'ACF' ranging from -0.4 to 0.4 and the x-axis labeled 'Lag' ranging from 0 to 24. It shows a significant positive peak at lag 12. The bottom plot is the Partial Autocorrelation Function (Partial ACF) plot, with the y-axis labeled 'Partial ACF' ranging from -0.4 to 0.4 and the x-axis labeled 'Lag' ranging from 0 to 24. It shows a significant positive peak at lag 1. Both plots include dashed blue horizontal lines at approximately +/- 0.35 to indicate the confidence interval.

And if you would like to learn even more about ARIMA models check these links:

- [Rules for selecting Seasonal AR and MA](#)
- [ARMA models](#)

A table with the correct 12 month forecasts for existing and new stores is provided. A visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts is provided.

The forecasts look within the expected range! Great job! Great job with the plot! Great job placing the forecast for the new stores on top of the sales for existing stores. That is suggested 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.

Suggestion:

If you are interested in how to close the gap in the plot between the actual sales and the forecasted ones you can check this example:



RETURN TO PATH

