Combining Predictive Techniques

REVIEW

HISTORY

Meets Specifications

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 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.

Thorough and well-presented analysis! Excellent work!

✓

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.

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

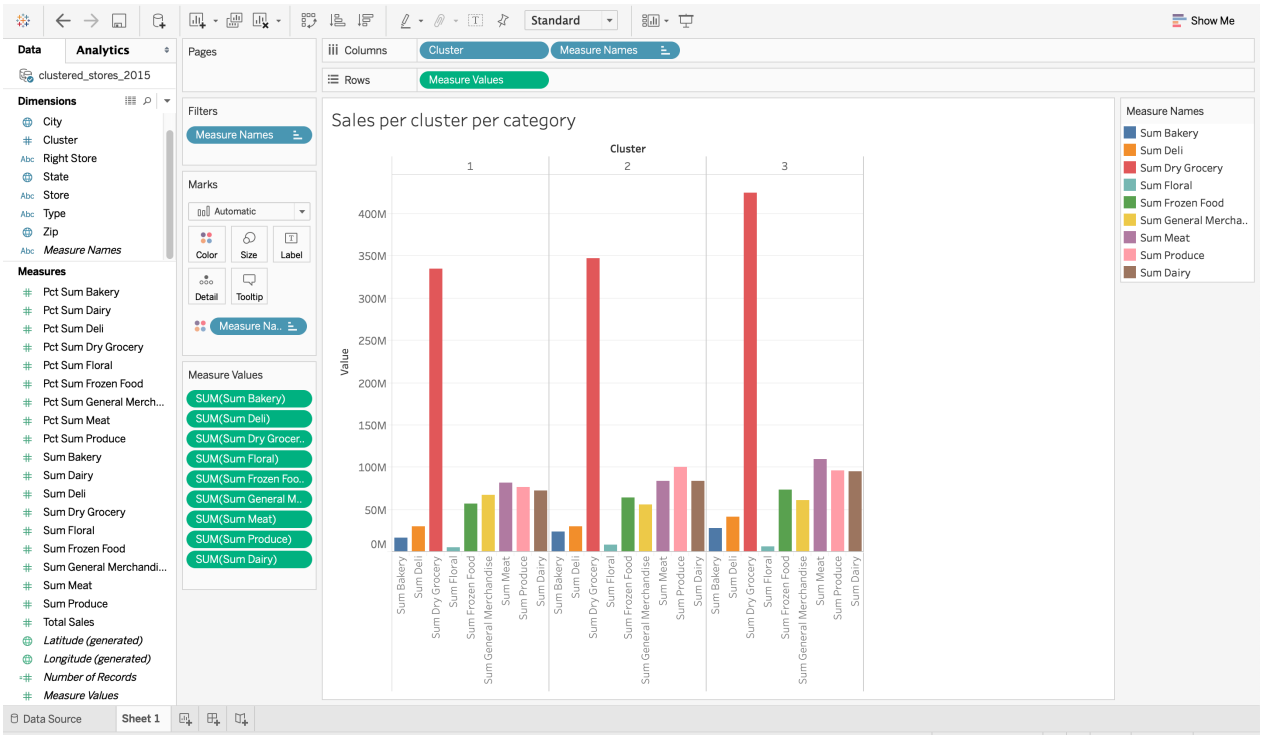
✓

Provides one observation about the differences among clusters, and uses the results of the clusters to provide justification.

Suggestion: Indeed, clusters differ in category sales. Great job!

Suggestion:

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:



✓

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 and are presented.

Task 2

✓

States the type of classification model used and adequately justifies the choice using at least one model comparison method.

Great job! Yes, both the Forest and Boosted models perform well, and either can be used. The Boosted model could be used since it has high accuracy and it also has a high F1 score. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

Here is a link about F1 score if you want to look more into it. And if you are interested in learning more about the different tree models you can take a look at this article about tree based models. Ignore the code based part, just check the theoretical part about the models.

✓

Includes a table that correctly identifies the format for each of the 10 new stores.

Task 3

✓

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! 🔠

Suggestion:

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. Here is an example how to do that:

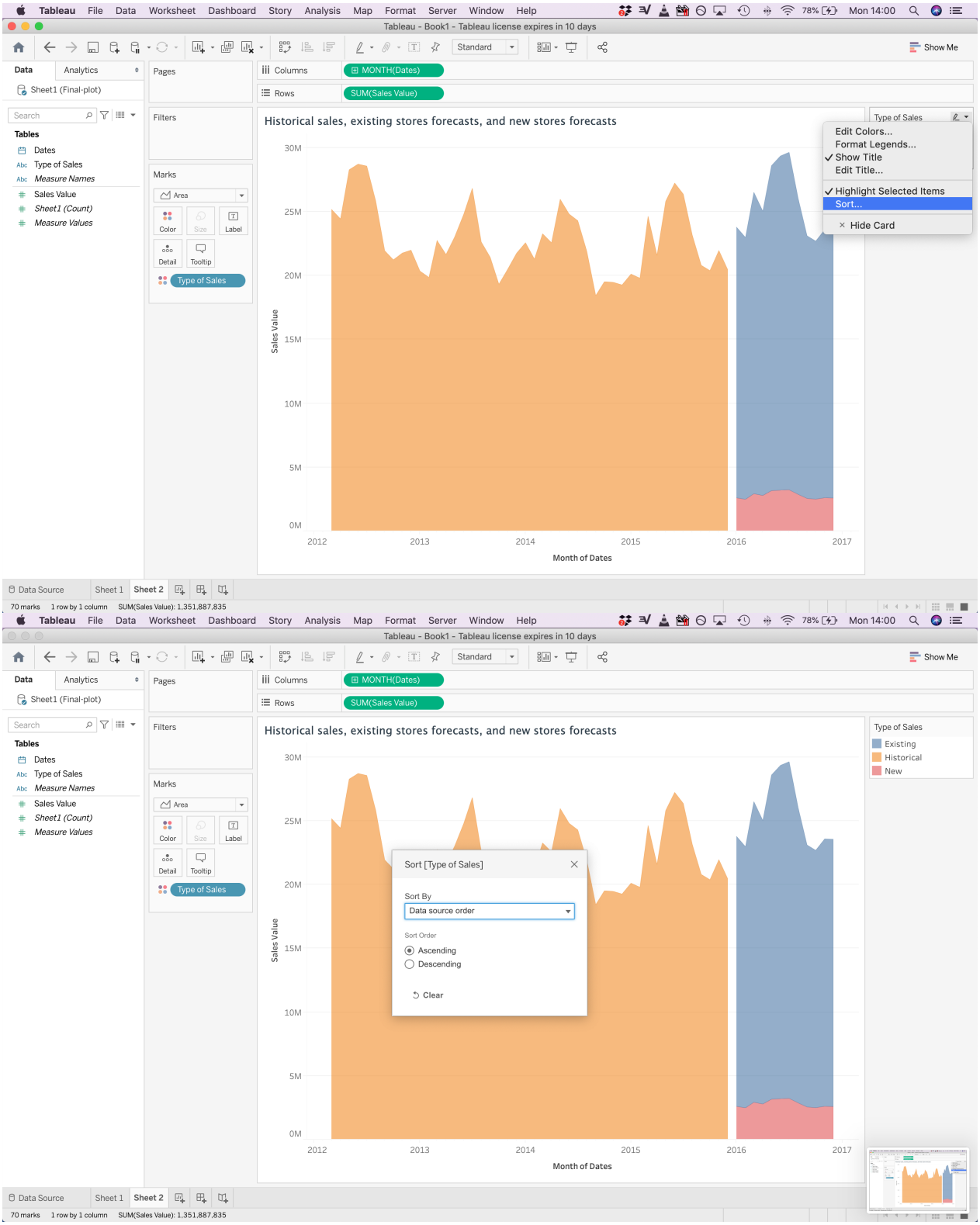
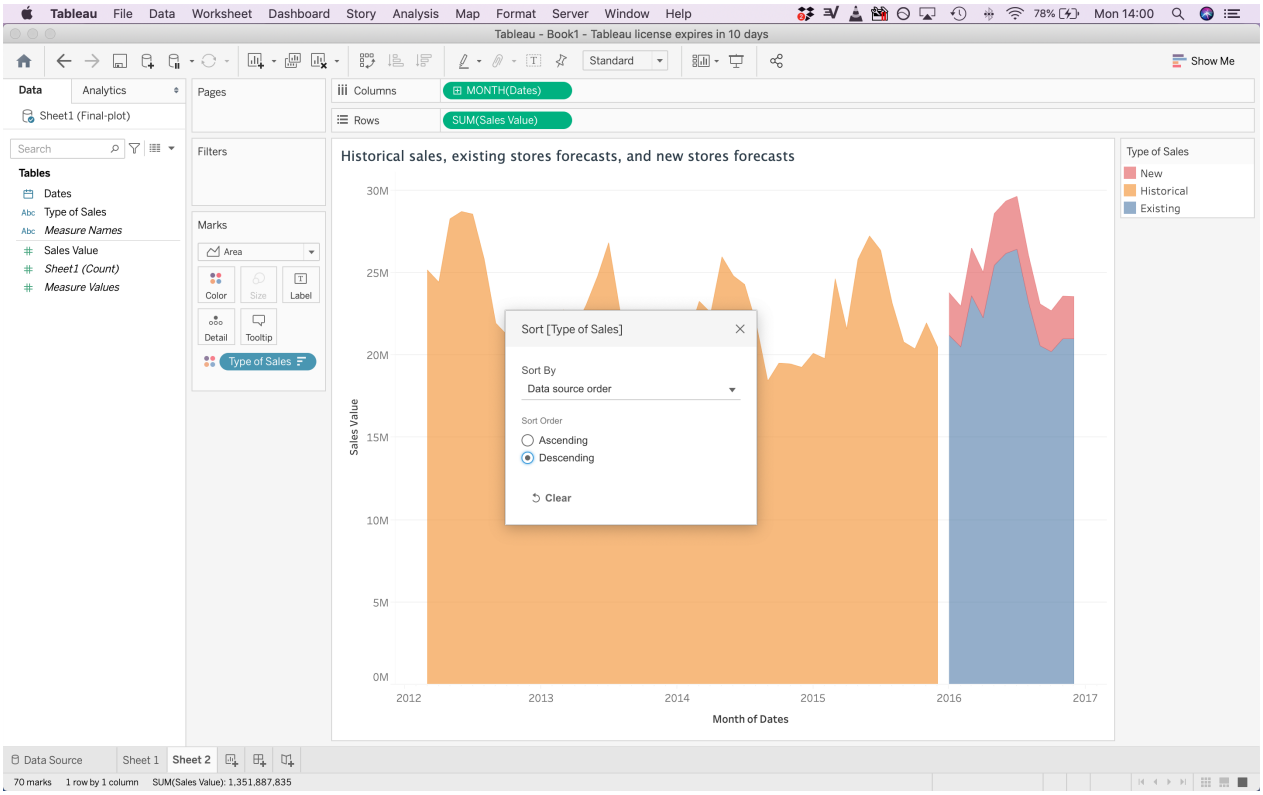


Tableau File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Tableau - Book1 - Tableau license expires in 10 days
Standard
Show Me
Data
Analytics
Sheet1 (Final-plot)
Search
Tables
Dates
Type of Sales
Measure Names
Sales Value
Sheet1 (Count)
Measure Values
Pages
Filters
Marks
Area
Color
Size
Label
Detail
Tooltip
Type of Sales
Columns
MONTH(Dates)
Rows
SUM(Sales Value)
Historical sales, existing stores forecasts, and new stores forecasts
Type of Sales
Edit Colors...
Format Legends...
Show Title
Edit Title...
Highlight Selected Items
Sort...
Hide Card
30M
25M
20M
15M
10M
5M
0M
Sales Value
2012
2013
2014
2015
2016
2017
Month of Dates
Data Source
Sheet 1
Sheet 2
70 marks
1 row by 1 column
SUM(Sales Value): 1,351,887,835
Existing
Historical
New
Sort [Type of Sales]
Sort By
Data source order
Sort Order
Ascending
Descending
Clear



✓

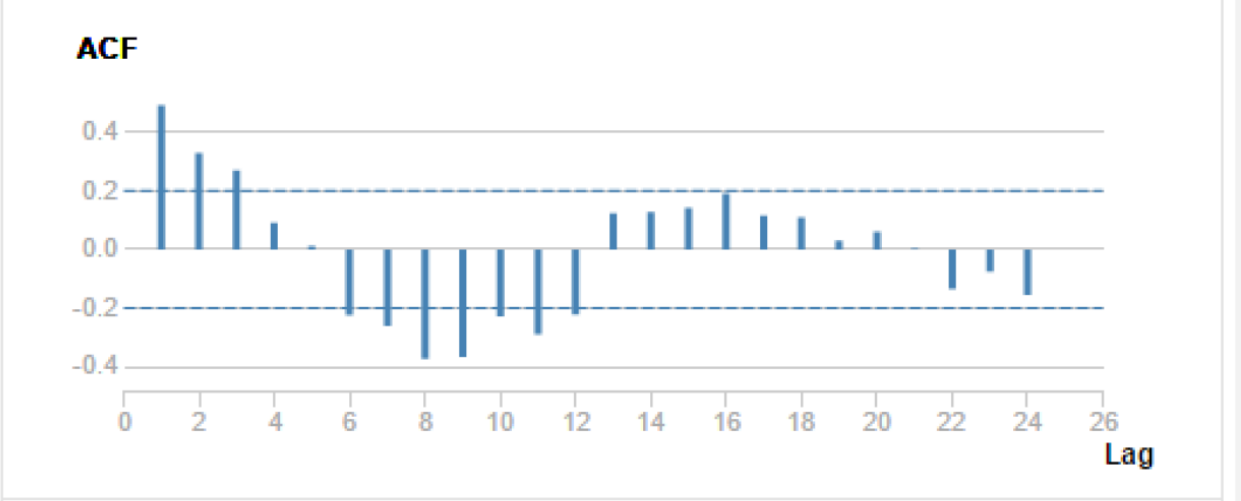
Compares and identifies the best ETS or ARIMA model to use for forecasting. Justifies the decision by showing the plot and shows forecast error measurements against the holdout sample.

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.

Some additional notes on the selection of the optimal ARIMA:

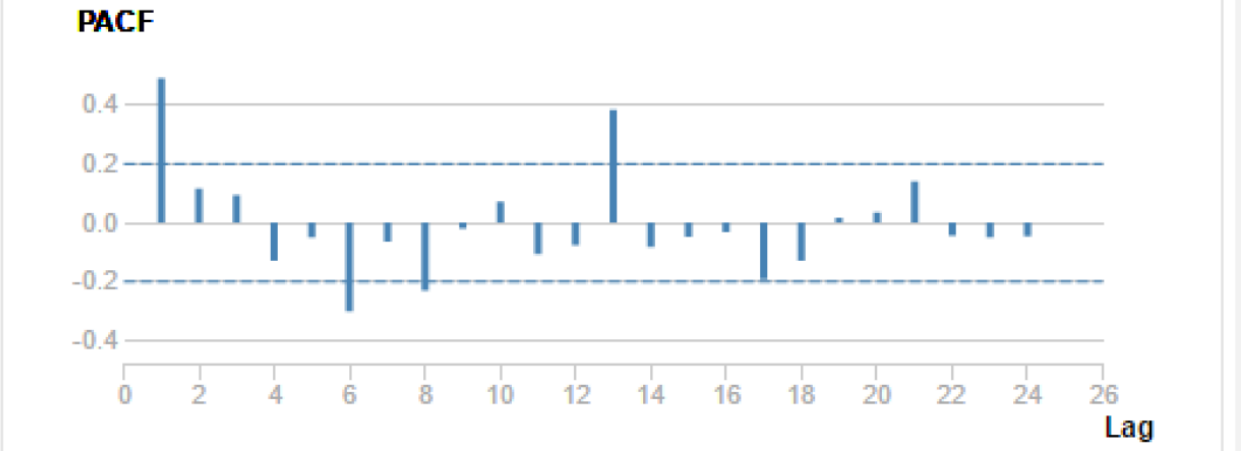
It would be good to show the tested ARIMA model as well. As the optimal ARIMA I suggest checking the auto option that the ARIMA tool gives. That model is ARIMA(1,0,0)(1,1,0)[12]. Let me give you some additional notes about why that model can be used. We consider as the optimal model the auto option - 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.

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:

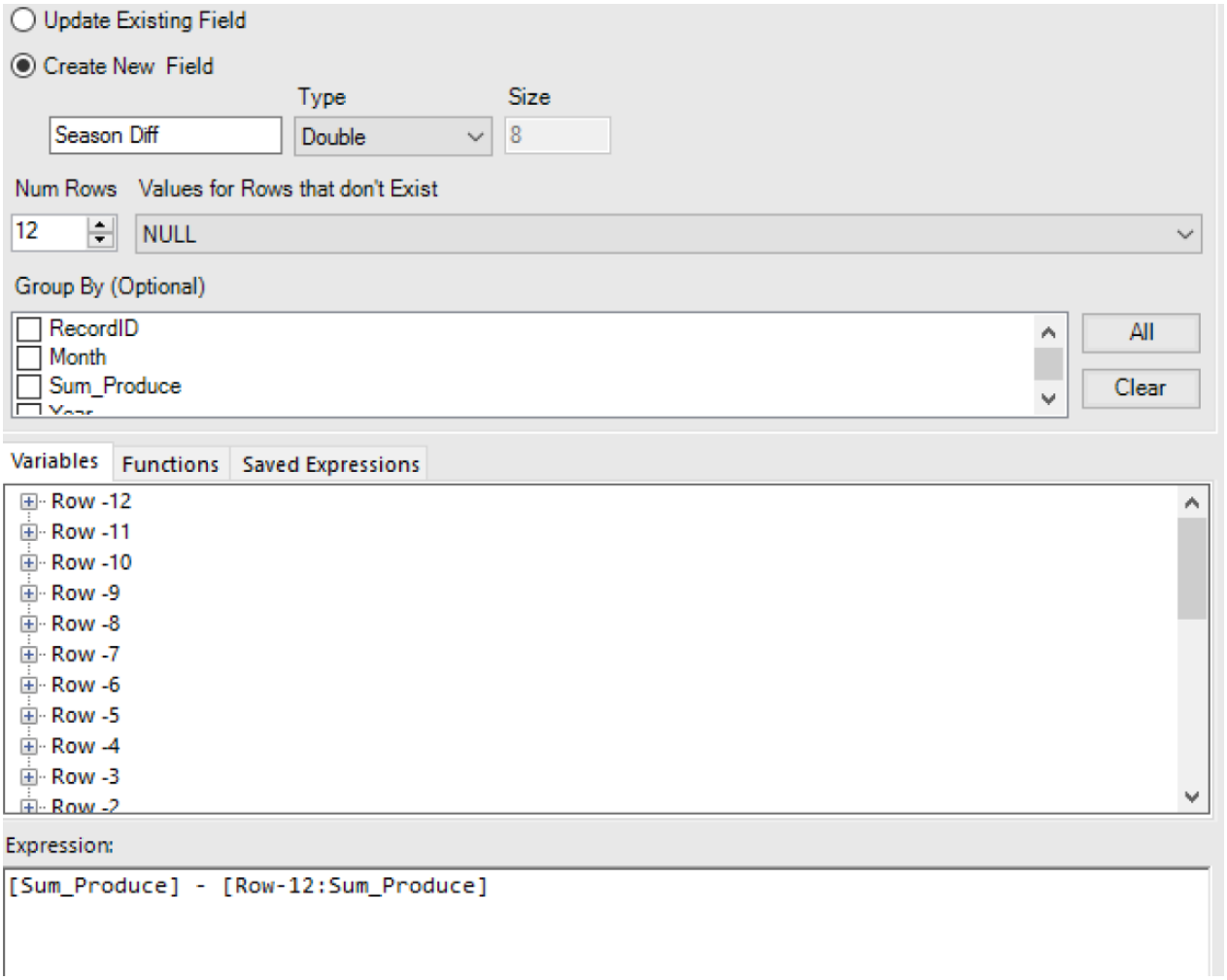


This is an autocorrelation plot

Partial Autocorrelation Function Plot

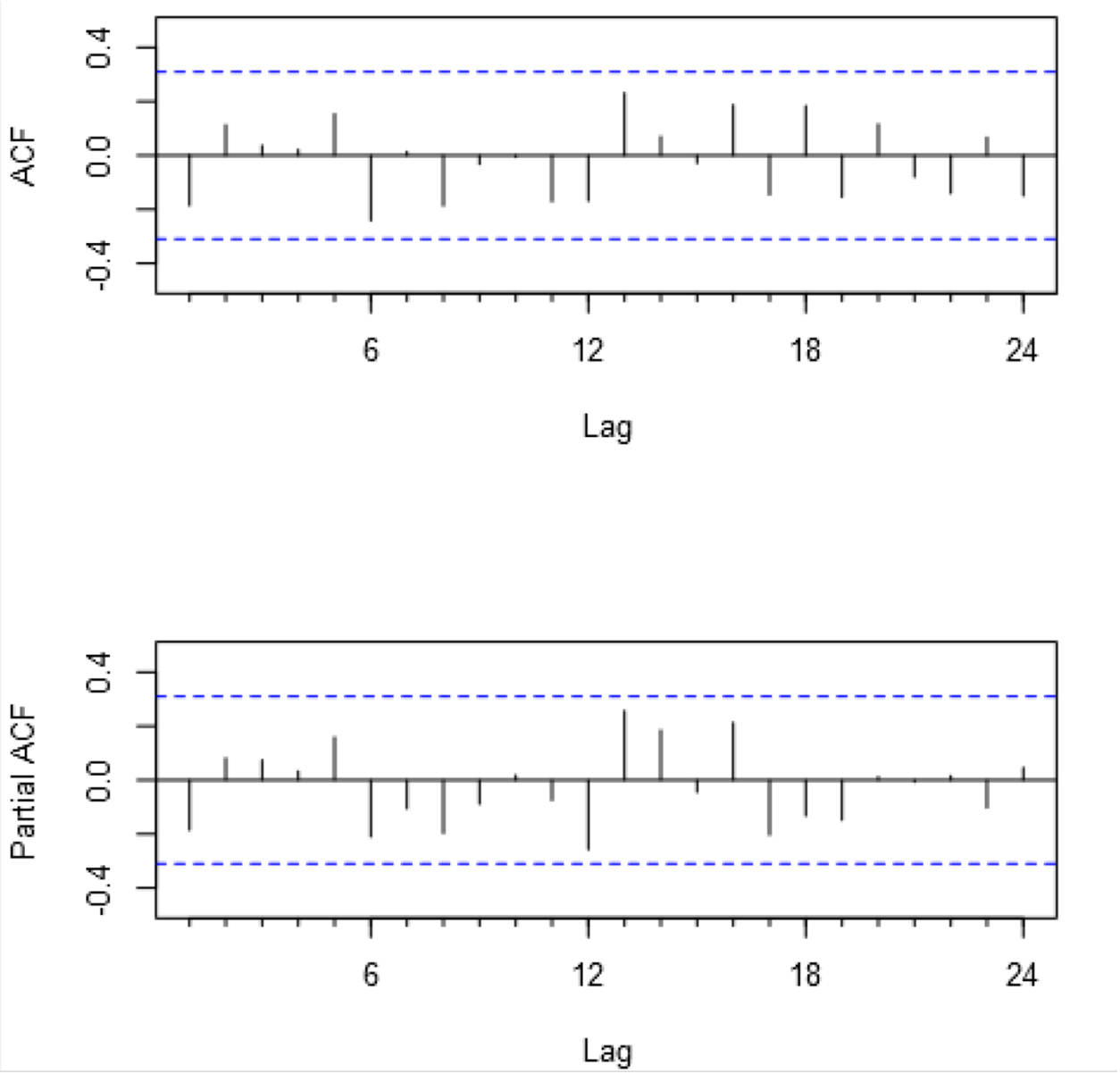


Which can be obtained from here:



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.



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

- Rules for selecting Seasonal AR and MA
- ARMA models

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

START