Draft version January 18, 2022
Typeset using LaTeX default style in AASTeX631

Getting Your Toolbox Ready for Modeling

Kevin B. Moposita[1]

[1] *Villanova University*
800 Lancaster Avenue
Villanova, PA 19085, USA

1. INTRODUCTION

The purpose of this assignment is to have students recall and draw back on previous lessons involving coding. In doing so, students will be ready to tackle future assignments. For this assignment, white noise was generated and modified so as to retrieve specific bits of data. In this paper, a brief explanation on white noise, autocorrelation, cumulative correlation are touched on along with corresponding figures.

2. ANALYSIS AND DISCUSSION

2.1. *White Noise*

For this assignment, random white noise with the parameters of a length of 1 million, a mean of 1, and a standard deviation of 0.1 was generated to start off. A time white noise time series is labelled as such due to zero correlation being present within the series. *Figure 1* depicts this initial time series. From this initial time series, another plot was created that shows the differences within the white noise distributions. For each histogram, the binning was increased. *Figure 2* demonstrates the distribution differences.

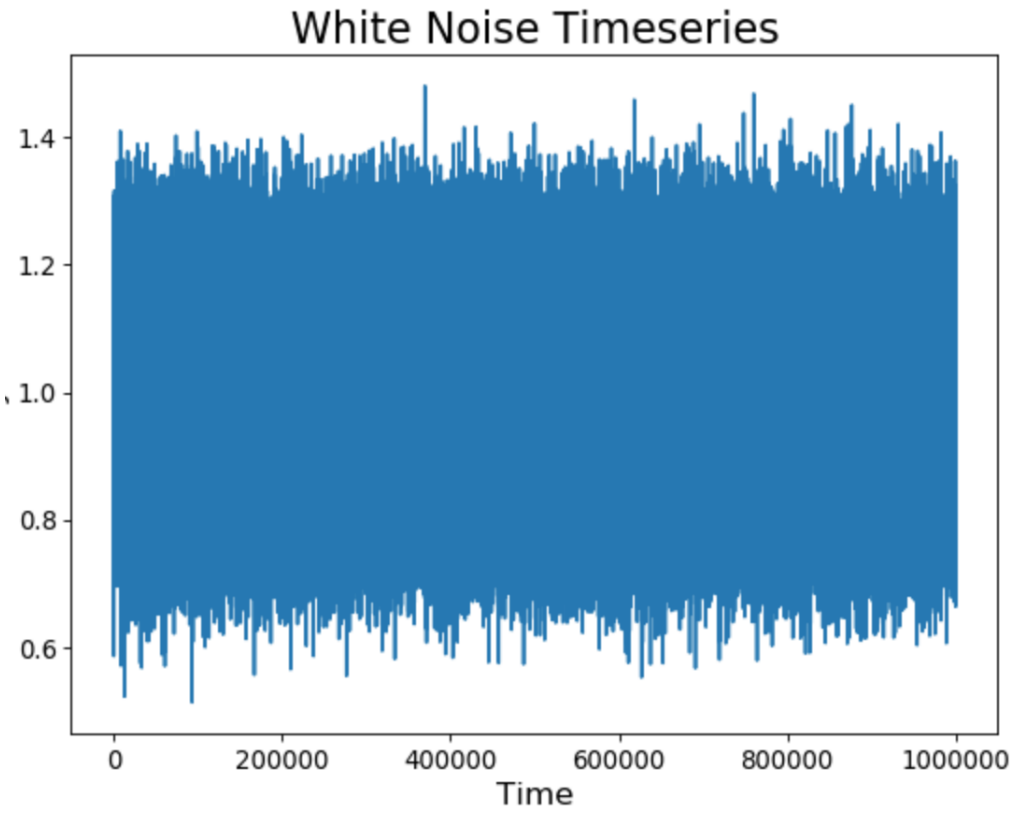

Figure 1. The initial white noise time series with the mentioned parameters.

[H]

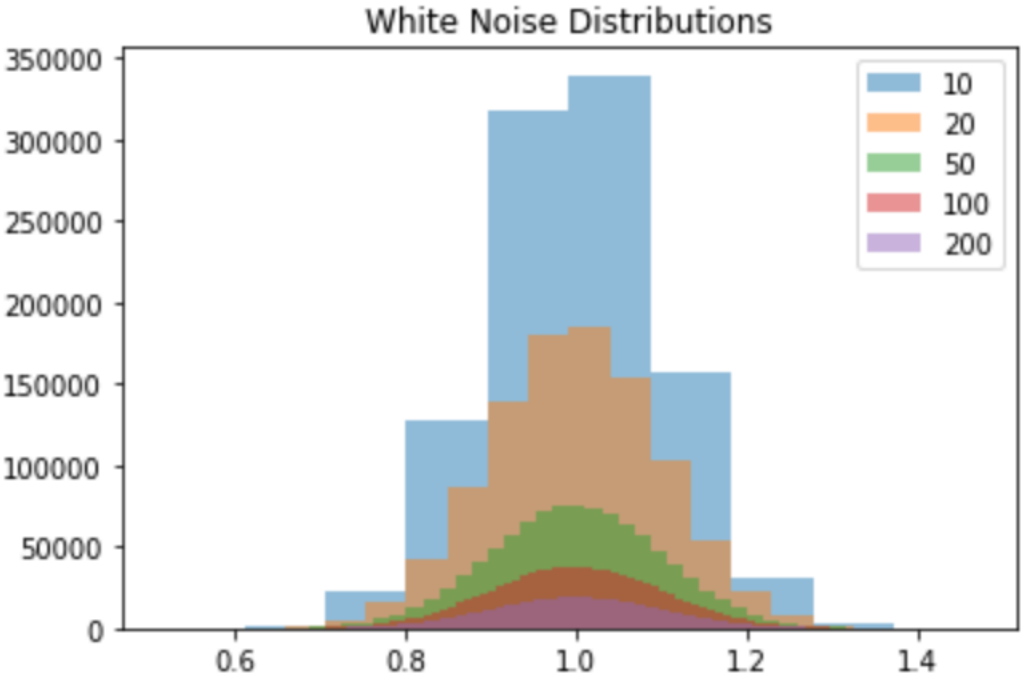

Figure 2. Depicts the white noise distribution as the number of bins increases.

2.2. *Autocorrelation*

With the creation of the white noise time series, autocorrelation can be determined. Autocorrelation is described as the correlation between two observation points within different areas of the data set. To find correlation, there is usually a lag introduced to find any matches. When autocorrelation was plotted initially, there were no peaks identified as demonstrated by *Figure 3*. There appears to be little to nothing going on. If there were a peak present, it would be indicative that high autocorrelation is present. However, because white noise is random, there is none present. That changes, however, if a signal were to be introduced every so often within the time series. Should this occur, autocorrelation would be present as shown in *Figure 4*.

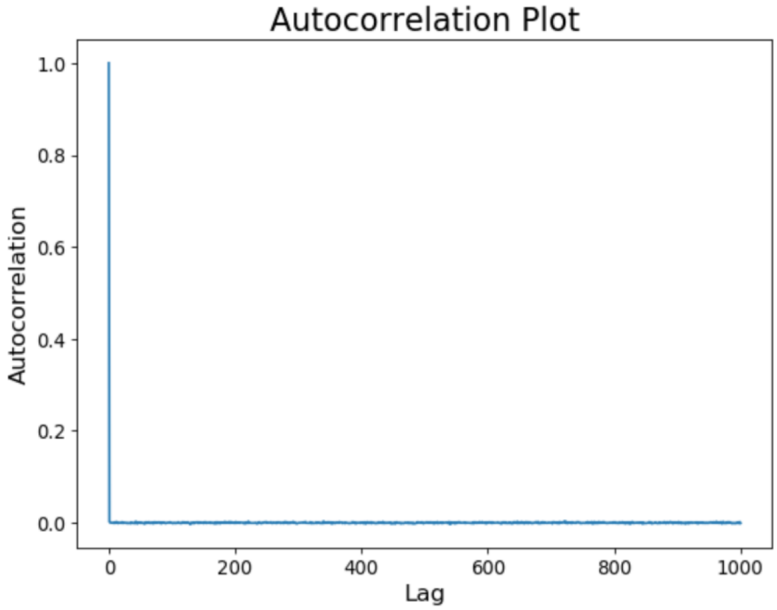

Figure 3. Caption

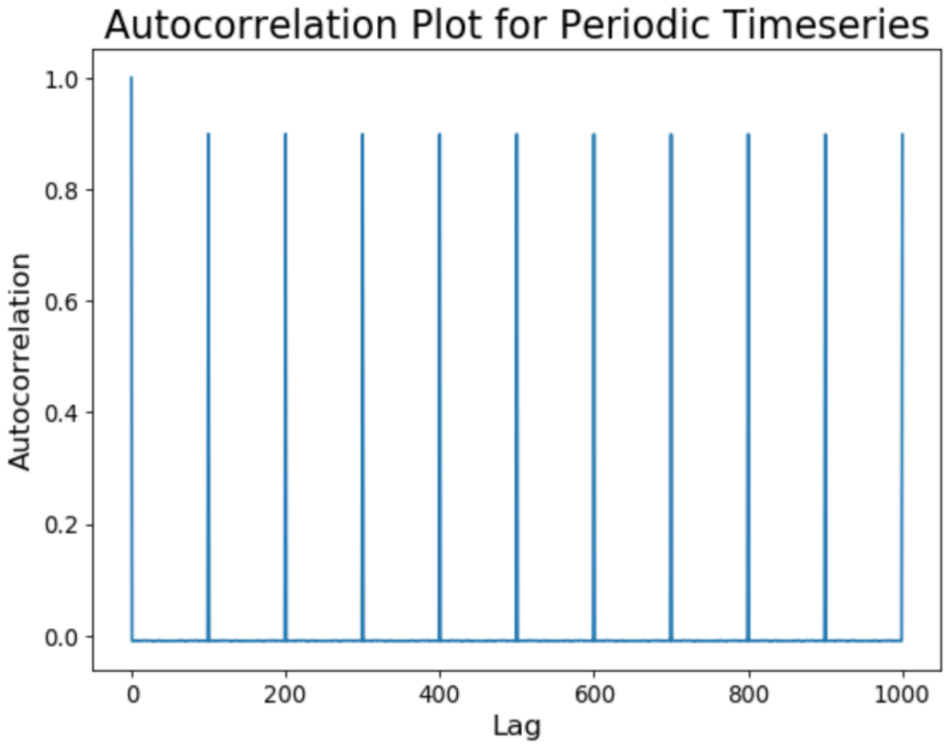

Figure 4. This image depicts the autocorrelation that would be present in a time series should a signal be present.

2.3. *Outliers and Similarity*

Just as we analyzed the autocorrelation within the original time series, outliers can also be analyzed. The outliers are meant to track where in the time series are the outliers present. *Figure 5* depicts the outliers right onto the white noise time series plot. Within the first set of outliers, most of them appear within the first sigma, 317,311 to be exact. From there, the number of outliers dramatically decreases until only 1 is present in the fifth sigma. This was repeated a few times and the numbers are present in the table below. From these trials, it can be said the greatest number of outliers will always be within the first sigma region.

Sigma	Theoretical Values	Computed Values
1	317311	316830
2	45499	44963
3	2700	2622
4	63	52
5	1	0

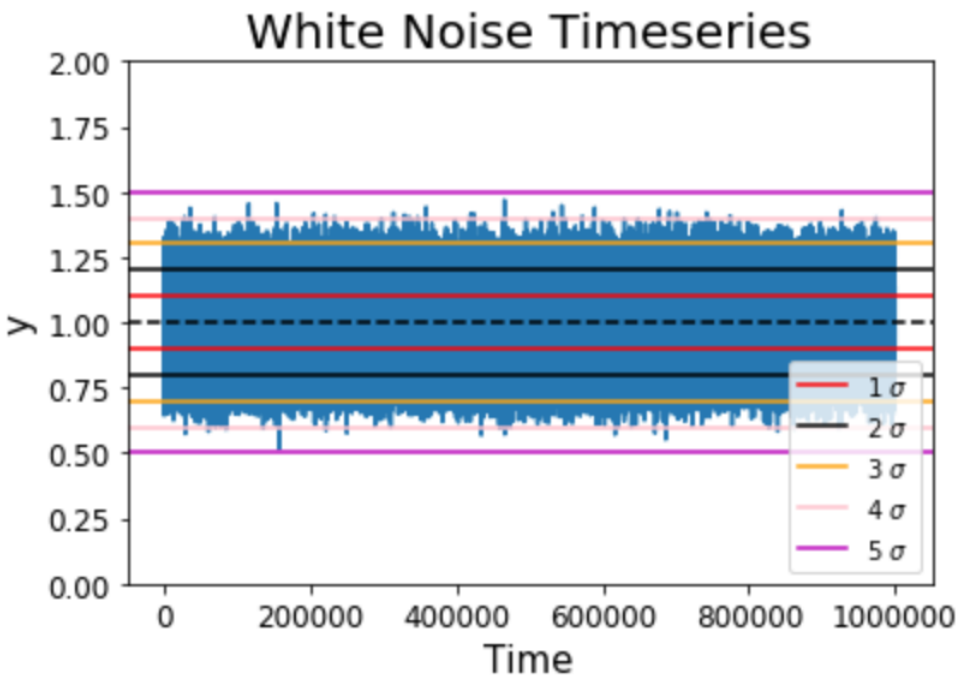

Figure 5. A visual of the outliers within the white noise time series.

To determine similarity between two consecutively generated time series, the Pearson correlation test can be useful. This statistical test is meant to measure the statistical relationship between two variables in two data sets. This would be very much useful when analyzing data. However, because Pearson's correlation is being used on white noise, we expect for the 'r' value (correlation value) to be near zero, meaning little to no similarity. *Figure 6* and *Figure 7* are the two consecutively generated time series that were analyzed by Pearson's correlation. The 'r' value from this correlation is 0.0027.

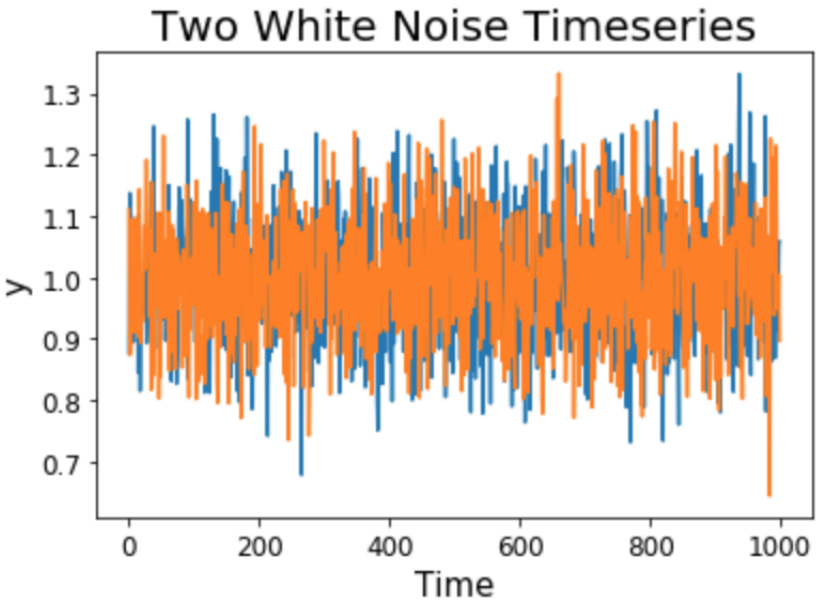

Figure 6. First generated time series

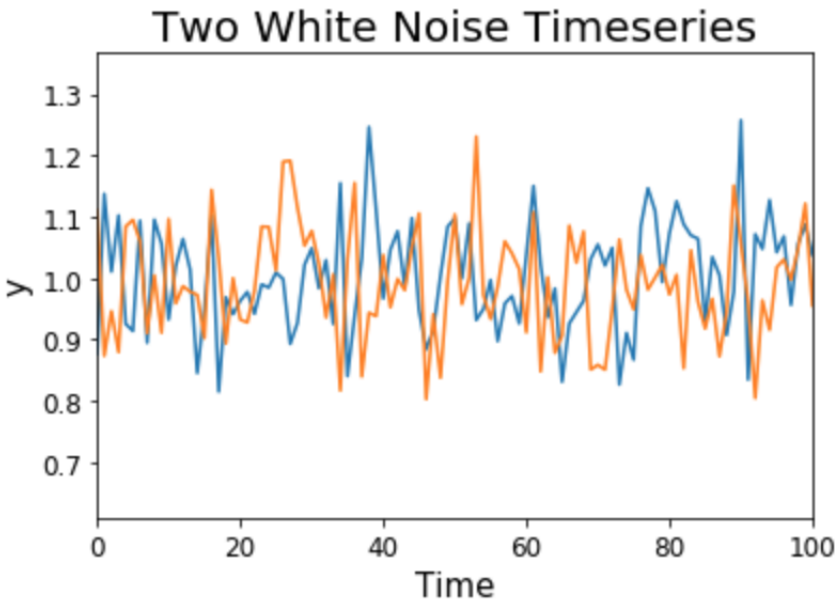

Figure 7. Second generated time series

3. CONCLUSION

The steps from this assignment were not only useful to help sharpen coding skill, but also to get an understanding of the tool set that data scientists use to better understand the data being analyzed. As this assignment shows, even when it appears as though the data reveals nothing important, a closer look often reveals this is not the case. White noise appears to cloud very relevant information and getting an understanding of it is critical if we are to keep working with it.