Since many of you were unable to download the data, I have made a smaller (much smaller) sized file for you to work on. The dataset I have attached contains time series data. THE NAME OF THE FILE IS **Class_data.nc**

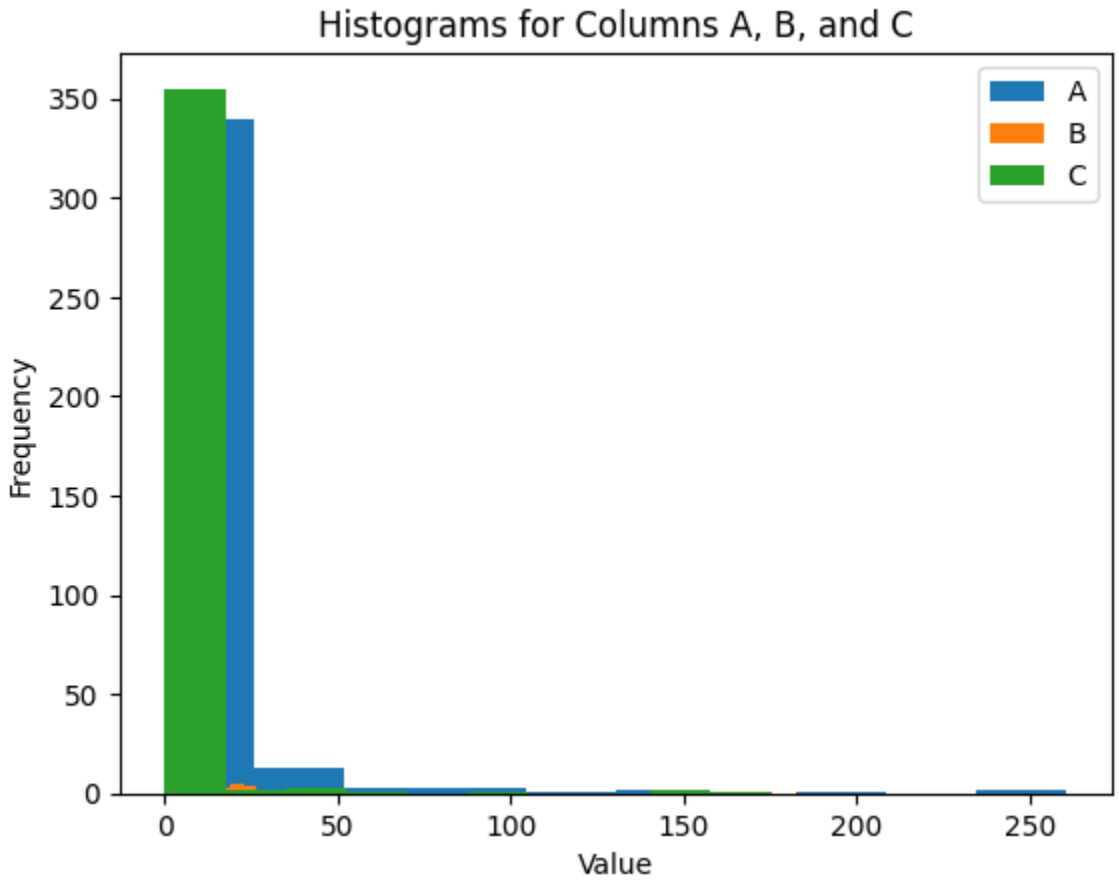
Please use the data to answer these questions (25 pts, each 5 pts):

1. Please calculate the interquartile range and the lower bound values for column "A"

```
First Quartile (Q1): 0.0
Third Quartile (Q3): 2.7853899002075195
Interquartile Range (IQR): 2.7853899002075195
Lower Bound: -4.178084850311279
```

The code is attached in For_IQR.ipynb

2. Please plot the histogram for A,B, C.



The code is attached in For_IQR.ipynb

3. Please find if the time series a is correlated with B-F.

```
Correlations between A and B-F:
B   -0.027192
C   -0.031546
D   -0.019519
E   -0.032380
F    0.009236
```

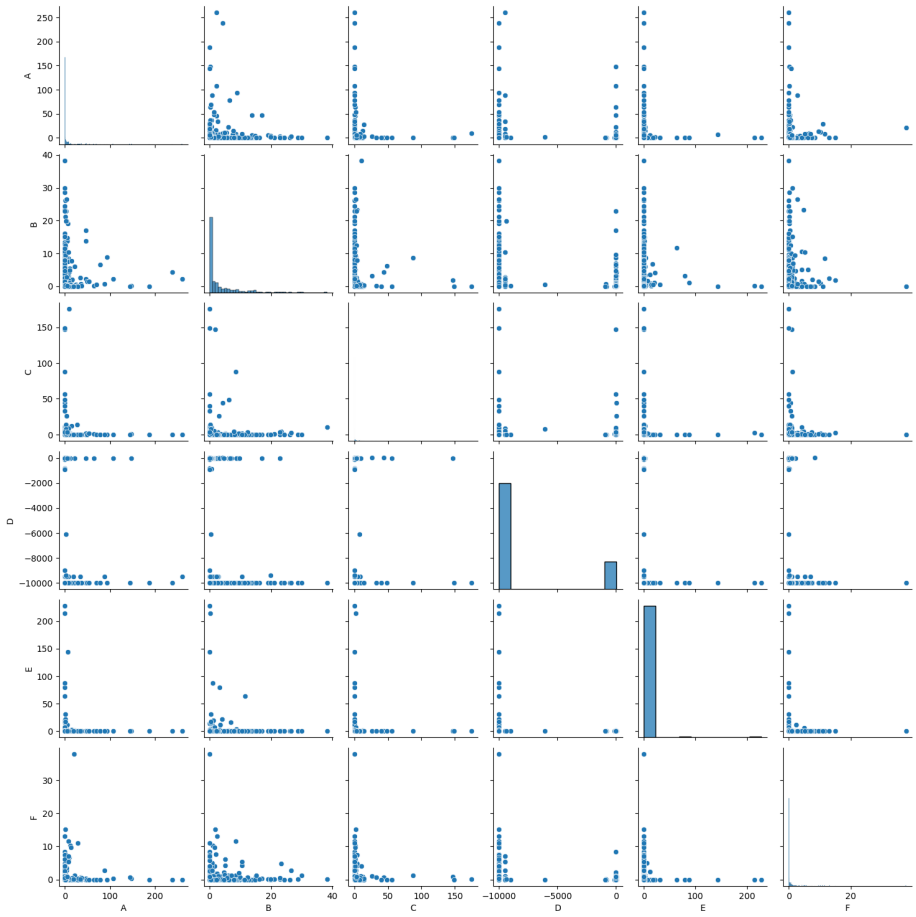
`Nothing is correlated with A.` The code is attached in For_IQR.ipynb

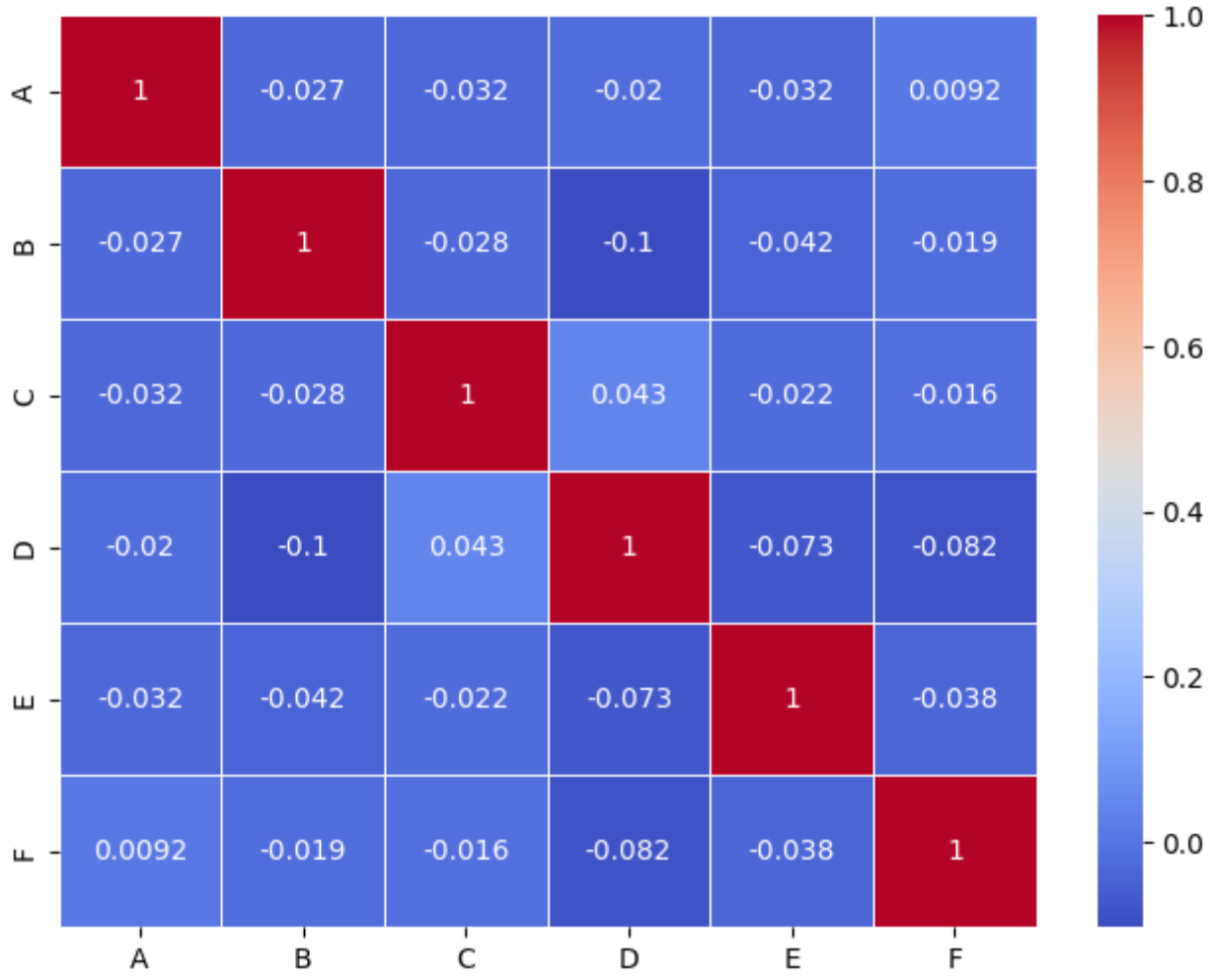
4. How do you decide if they are correlated?

   To decide if two time series are correlated, we will calculate the respective correlation coefficient and examine its value:

   - If the correlation coefficient is close to 1 or -1, it suggests a strong positive or negative relationship, respectively.
   - If the correlation coefficient is close to 0, it suggests a weak or no relationship.

5. Use these (A-F) as a matrix and plot the scatterplot matrix and heatmap.





The code is attached in For_IQR.ipynb

6. MARKOV CHAIN:

a. uses the data "A". If values are >0 mark them as 1 (presence). If values are 0 then mark them as 0 (absent). Let's name the array "XXY" 5 pts

The code is attached in markov_chain.ipynb

b. Use the column D. If values are >0 mark them as 1 (presence). If values are 0 then mark them as 0 (absent). 5 pts

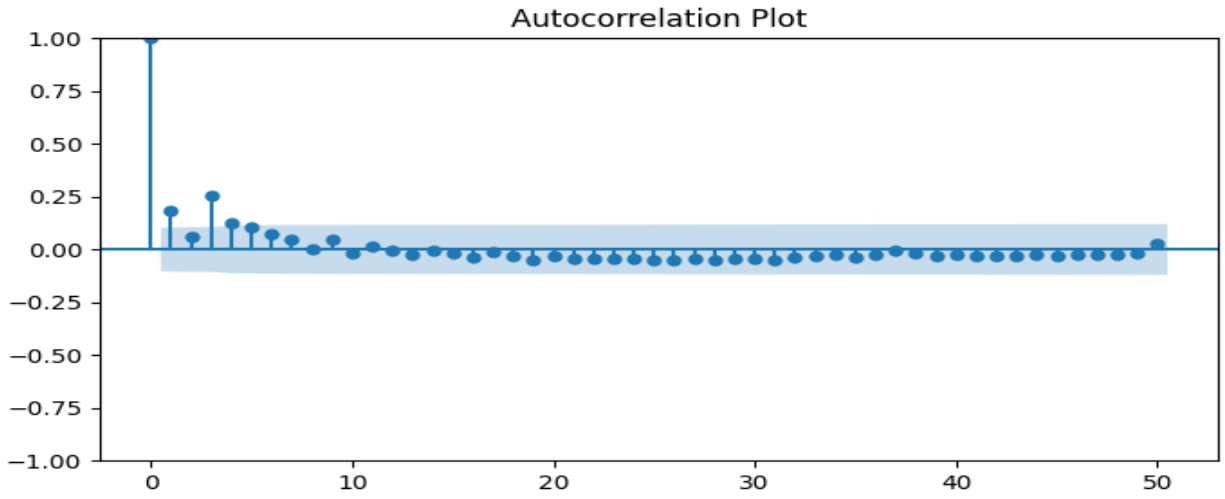
The code is attached in markov_chain.ipynb

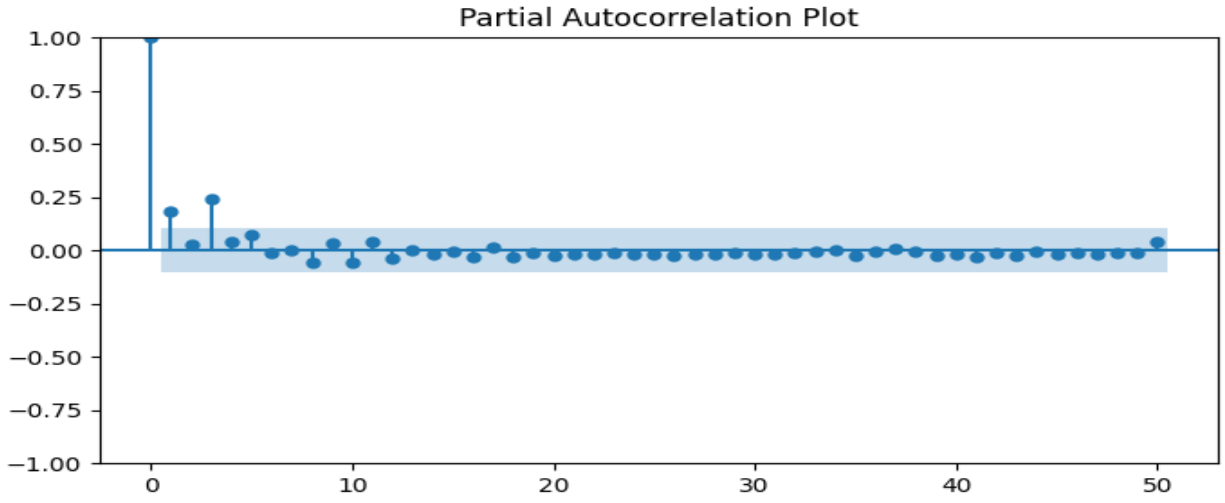
c. Are these time series the same? Use the chi-square test to answer. 10 pts

```
Chi-squared statistic: 3120.307376395728
P-value: 0.0
Observed and expected values are significantly different.
```

The code is attached in markov_chain.ipynb

7. Can you plot the autocorrelation and the partial auto-correlation plots? 5 pts





The code is attached in auto_correlation_and_partial_corelation.ipynb.

8. Run the augmented Dickey–Fuller test and show the test statistics using the residuals and the trend. 5 pts

```
ADF Test Statistics (with trend):
ADF Statistic: -7.831335862375576
P-value: 1.75355612653178e-10
Lags Used: 2
Number of Observations: 362
Critical Values:
1%: -3.9839983263172876
5%: -3.4226880872267436
10%: -3.1342229649827877

ADF Test Statistics (with constant):
ADF Statistic: -7.78155524128868
P-value: 8.390955199385318e-12
Lags Used: 2
Number of Observations: 362
Critical Values:
1%: -3.448544133483233
5%: -2.8695574079525565
10%: -2.5710411593052713
```

The code is attached in auto_correlation_and_partial_corelation.ipynb.

9. Divide the data from column “A” into training and testing. This time split by 85:15. That means we are training using 85% of the entire data and testing using 15%. 5 pts

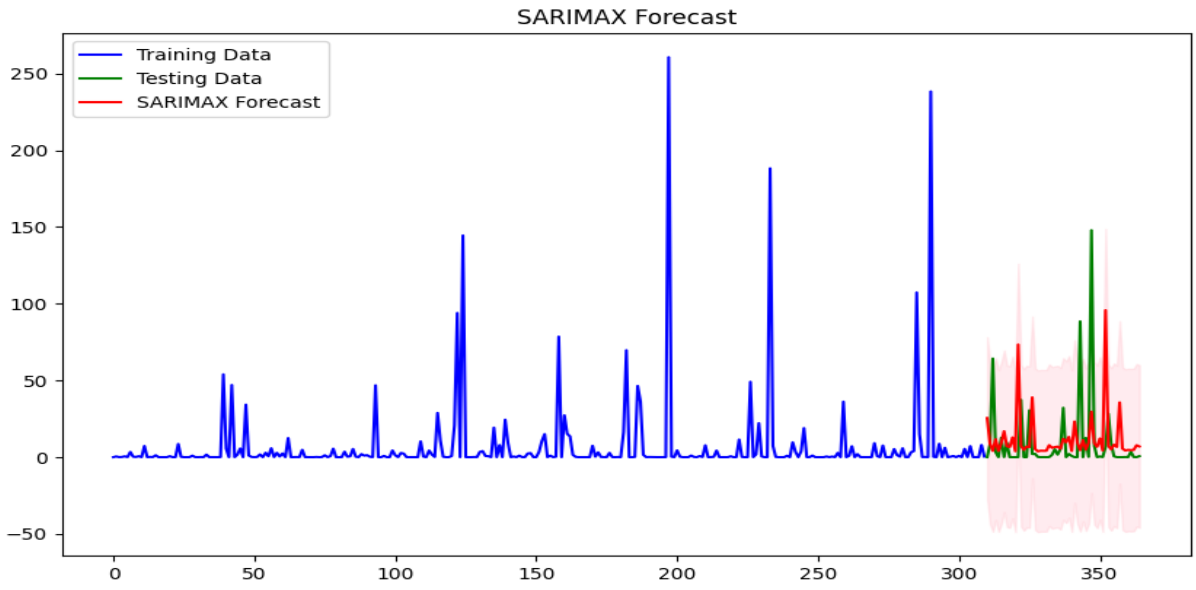
```
data_A = variable_data_A


# Split the data into training and testing sets (85:15 split)
train_data, test_data = train_test_split(data_A, test_size=0.15,
random_state=42)
```

The code is attached in auto_correlation_and_partial_corelation.ipynb.

10. Use the SARIMAX to develop your model. 10 pts

The code is attached in auto_correlation_and_partial_corelation.ipynb.

11. Please use the SARIMAX model to predict and test with your testing data. 5 pts



The code is attached in auto_correlation_and_partial_corelation.ipynb.

12. Calculate if any of the time series are related to each other. 5 pts

I have run: {'pearson', 'kendall', 'spearman'}. I have not seen any relationship. The code is attached in auto_correlation_and_partial_corelation.ipynb.

13. Let's construct another Fourier series:

Data_A=[26,27,29,30,31,32,33,33,34,35,36,37,38,40,41,42,43,45,47,49,50,53,54,57,59,62,59,57,56,55,54,53,51,49,47,49,45,44,43,41,39,37,36,35,34,32,32,31,30,27,26]

The code is attached in fourier_series.ipynb.

14. can you compare the array XXY (question 6a) with the data_A (previous line)? 10 pts

Kendall Correlation: 0.6406874037664375
Kendall P-Value: 4.049718801904716e-48
Spearman Correlation: 0.7639389069241743
Spearman P-Value: 4.769552938539591e-71
Pearson Correlation: 0.1613861544825252
Pearson P-Value: 0.0019814292551196566

Kendall and Spearman Correlations:

- Both the Kendall and Spearman correlations have high positive values, close to 1. This suggests a strong positive rank-based relationship between XXY and data_A.
- The extremely low p-values (close to zero) for both Kendall and Spearman correlations indicate that these relationships are statistically significant.

Pearson Correlation:

- The Pearson correlation also suggests a positive relationship, but the value is lower (0.1614) compared to the rank-based correlations. This indicates a weaker linear relationship.
- The p-value for the Pearson correlation is small (0.00198), indicating that the linear relationship is statistically significant, but less significant compared to the rank-based correlations.

In summary, all three correlation methods suggest a positive relationship between XXY and data_A, but the strength and nature of the relationship vary. Kendall and Spearman correlations indicate a strong positive rank-based relationship. Pearson correlation indicates a weaker linear relationship.

The code is attached in markov_chain.ipynb (last section).

15. What is the difference between a discrete data set and a continuous data set? (will ask you in the last class, so don't CHATGPT) 5 pts

The fundamental distinction between discrete and continuous data is that discrete data consists of distinct, separate values that are often countable, while continuous data encompasses an infinite range of values without gaps and can include fractional or decimal measurements.

16. Did you see any discrete data so far in this assignment? 5 pts

Yes XXY array

17. Why do we need to use the Markov chain to analyze time-series data sets that have values like 1 or 0? Can we just count or plot histograms of "1"s and "0" to answer the Markov chain? 5 pts.

Markov chains are essential for analyzing time-series data with binary values like 1 or 0 because they capture sequential patterns and dependencies. Counting and histograms don't consider the order of observations or provide predictive insights. Markov chains reveal dynamic behavior, enabling distinction between sequences with the same counts. They can represent various states, making them versatile for complex data. In essence, Markov chains offer a more comprehensive understanding of time-series data, beyond what simple counting and plotting can achieve. Counting the sequences "00001111" and "01010101" might yield the same counts of "0"s and "1"s, but Markov chains recognize their distinct temporal behaviors. In this case, the sequences may appear similar in terms of simple counting, but Markov chains reveal their unique patterns and dependencies, showing that they are not the same. The behavior of Markov chains is sensitive to the order and sequence of data, which can be critical in applications where the dynamics of the data matter.