# Introduction

Human brain is possibly the most critical part of the human body. From helping humans perform simple mundane tasks to caryying out extraordinary feats throughout history, the brain is what controls everything in humans. The brain works by sending and receiving signals throughout the body with the aid of the nerves. By analyzing such signals or electrical impulses, changes in brain activity and state can be measured. Electroencephalography, or EEG, is such method used to measure the electrical activity of the brain in which small metal electrodes are attached to the scalp surface. Analysing and interpreting such signals provides a wide range of information about the state of brain.

# Task 1

## Time Series Plot

A time series is a sequence of data which is recorded continuously over time. Time series helps in studying how variables changes or evolves through time. Generally, when plotting a time series graph, time (or date) is plotted on the X axis or the horizontal axis and the variable is plotted on the Y axis.

Plotting the signals against time provides a way of interpreting the data. Here, EEG input/output signals with their respective time in seconds has been provided. We can simply plot those signals against time to interpret the basic overview of the signals.

Below is a plot of the 4 input signals with respect to their times (in millisecond).

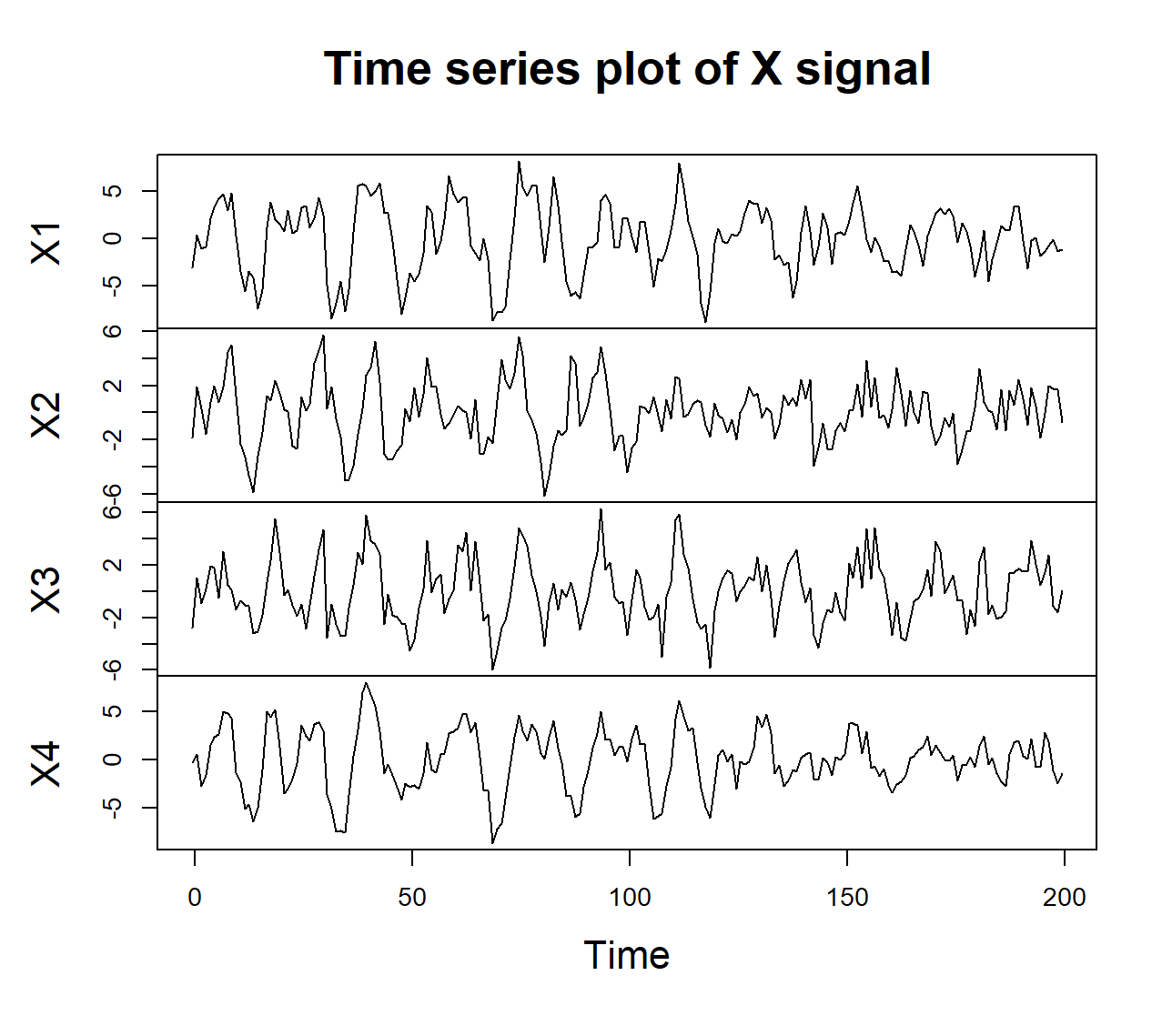


Figure 1 Time Series Plot of X (input) Signal

The above plot shows how the signals change with time. Looking at the plot, it can be seen that there are some values that deviate from the general trend of the signal. Slight spikes can be seen in all the input signals right around the same time frames which gradually transitions into somewhat relaxed state after around 120ms. All the signals have been subjected to noise.

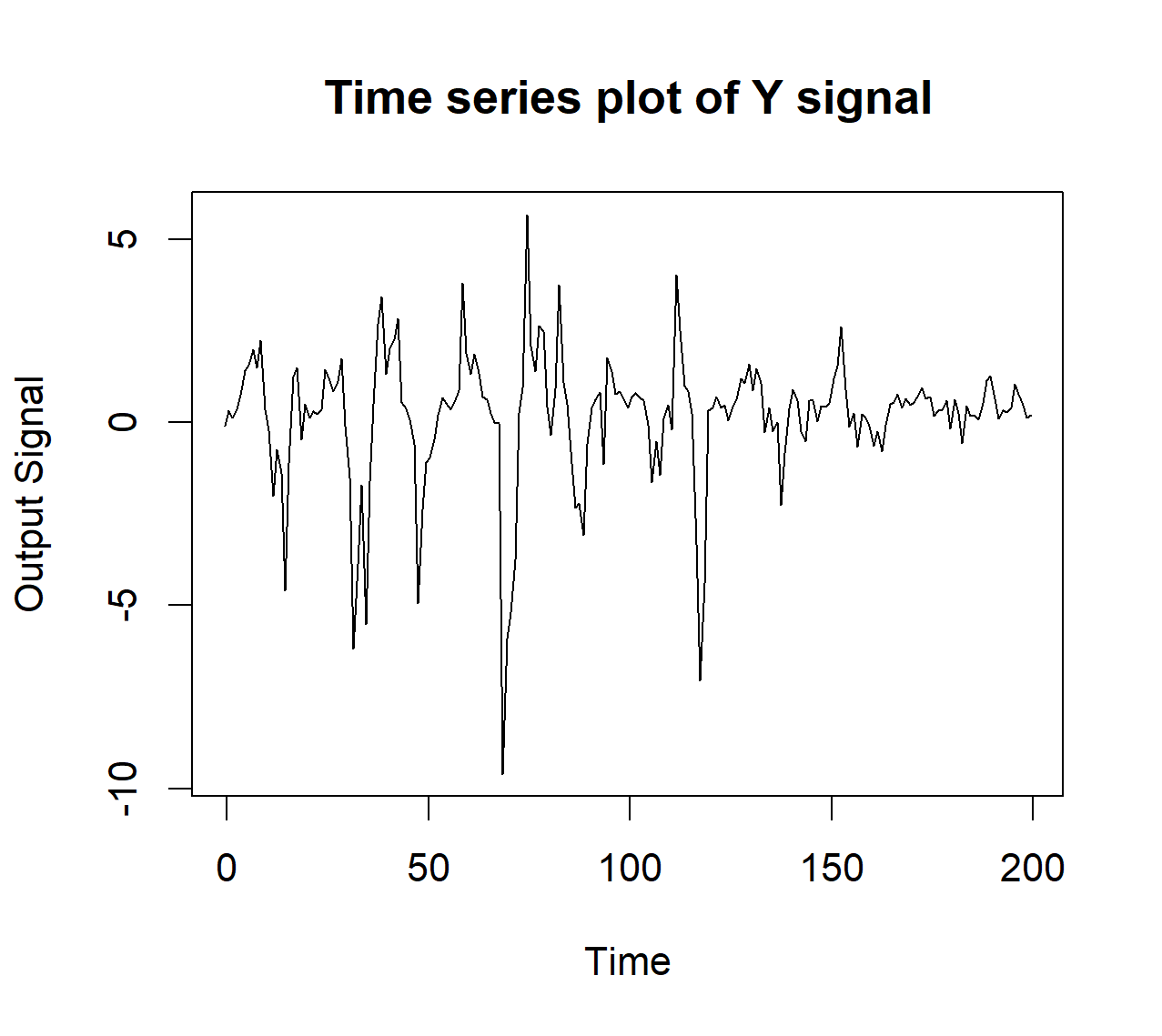


Figure 2 Time Series Plot of Y (Output) Signal

The output signal has also been subjected to some additive noise. We can see certain abrupt spikes in the signal. The abrupt fall of the signal at ~60ms can be explained looking back at *Figure 1* where the input signals X1, X3 and X4 are seen the have gone down as well at the same time. It is more or less the same at 120ms where the ouput signal has shown abrupt change at the same time where input signals X1, X3 and X4 show similar downward spike.

The general movement of the signal seems quite similar to those of input signals where the spikyness calms down quite significantly after 120ms having seen that all the input signals become significantly steady after 120ms.

## Distribution Plots of EEG Signals

On the basis of our data, we can visualize it by plotting histograms and density plots. A histogram is a bar style chart that approximately represents *numerical* datas by classifying them into bins or intervals. Density plots simply shows the density of various points in the data. The distribution shape of a dataset can be determined with the help of distribution plots. The figures below show the histogram and density plot of the input signal.

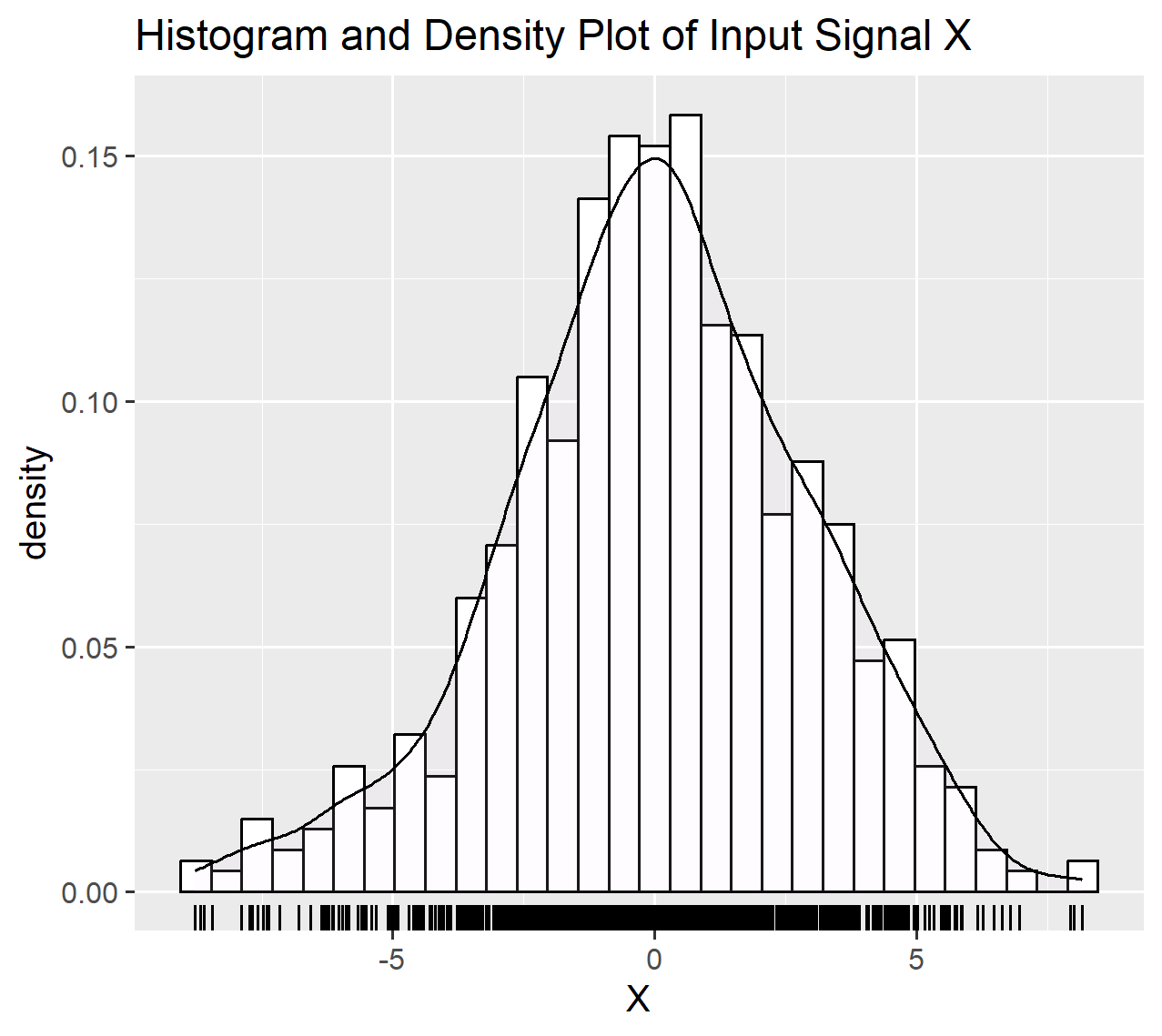
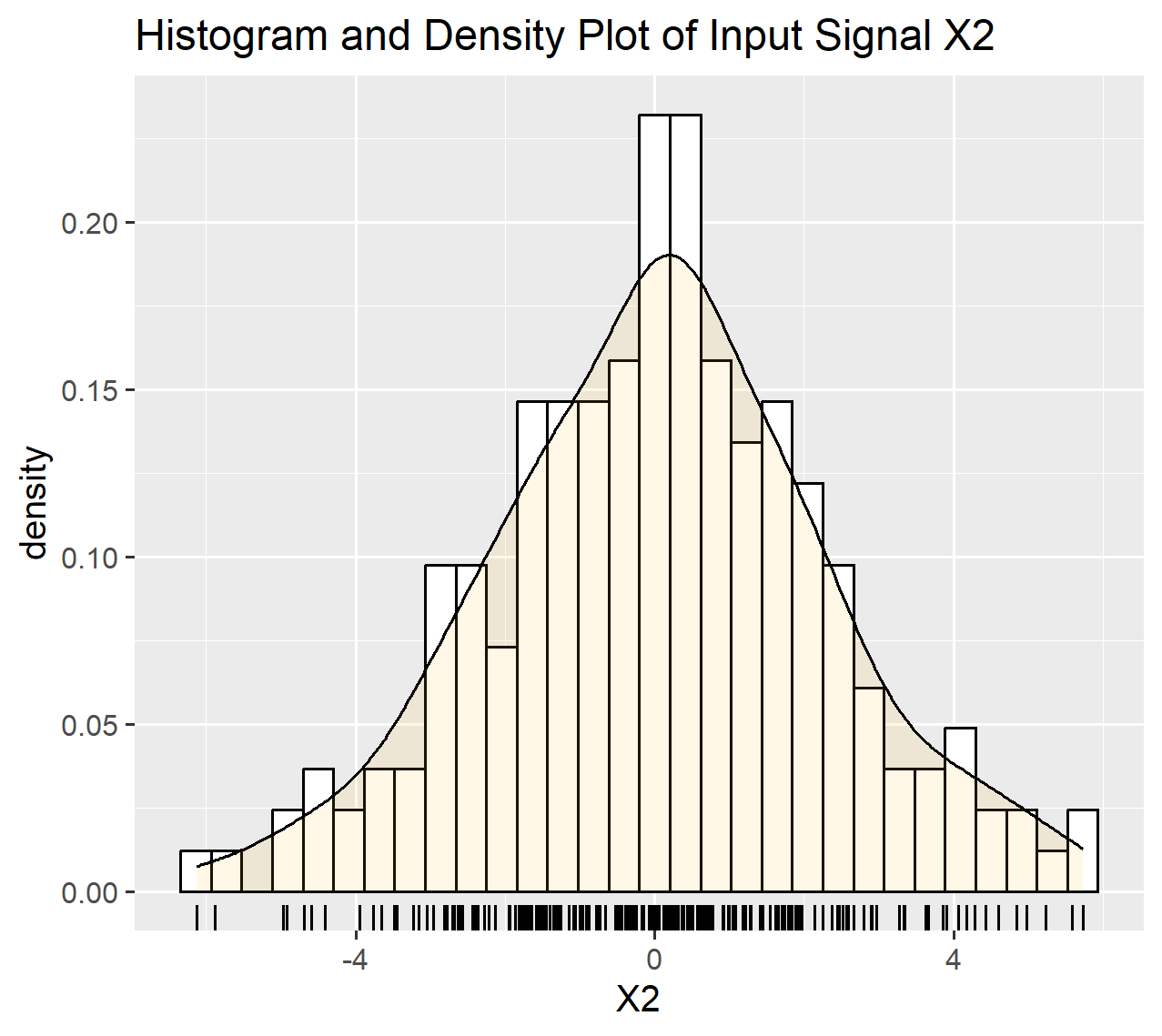
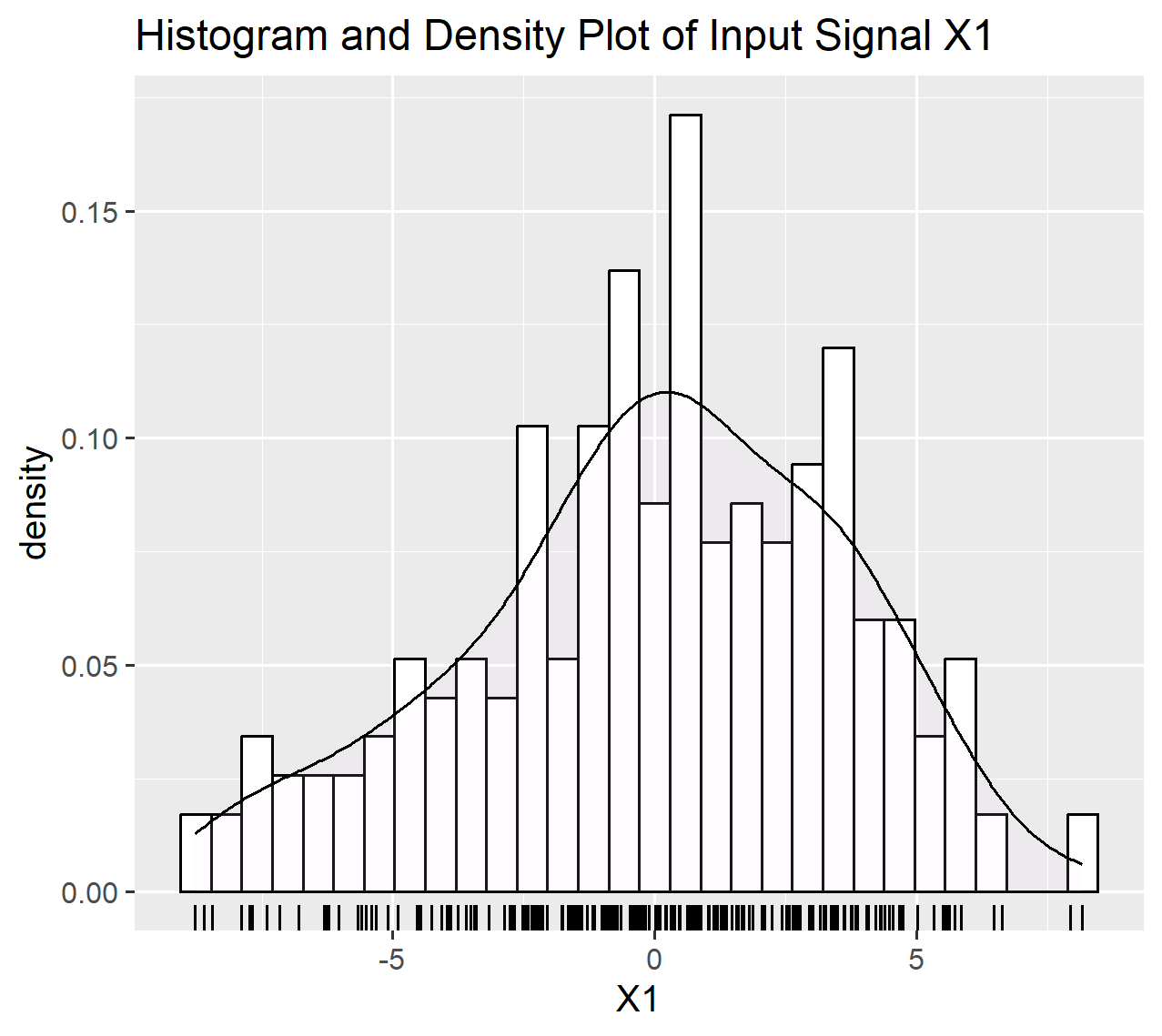
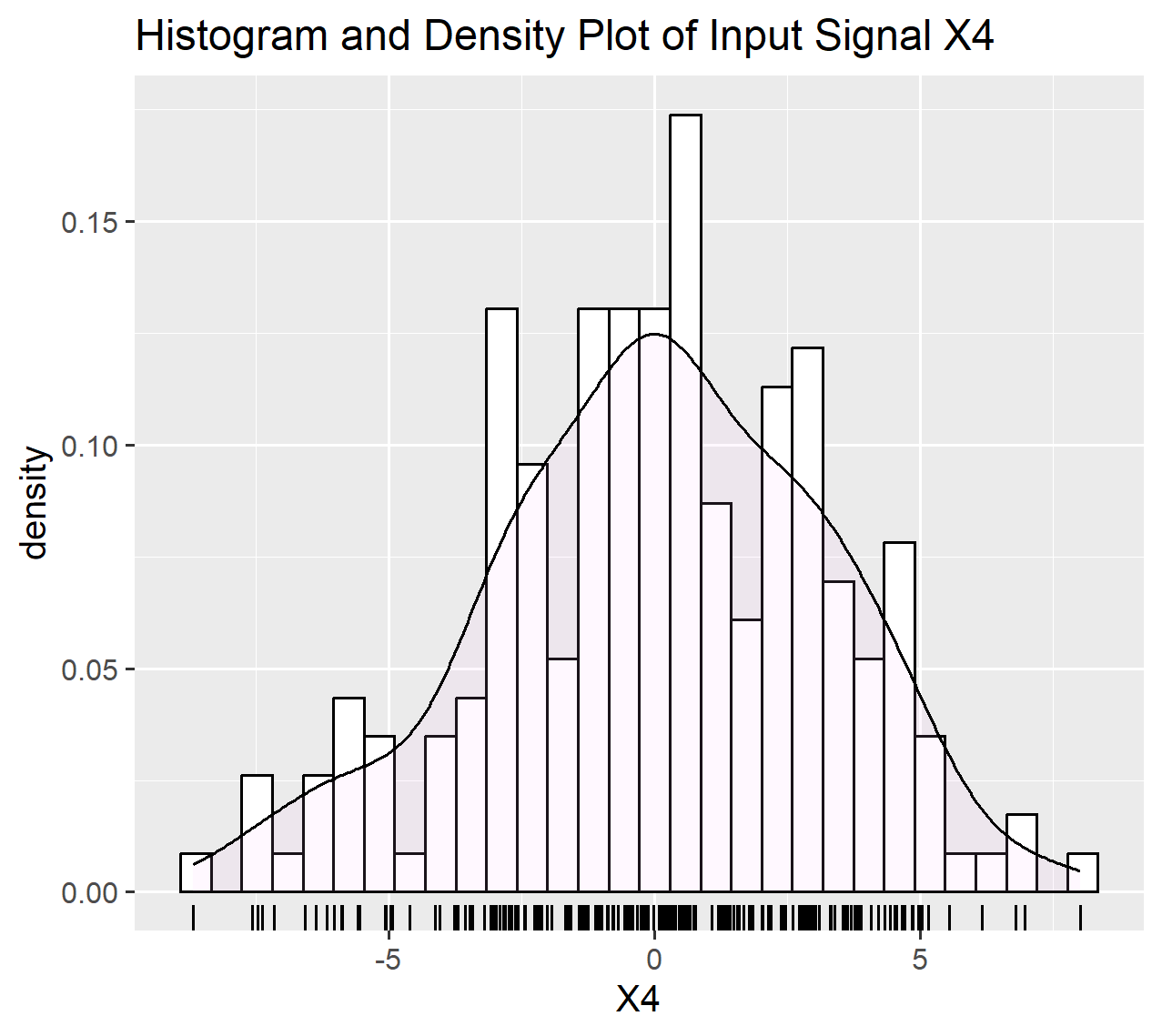
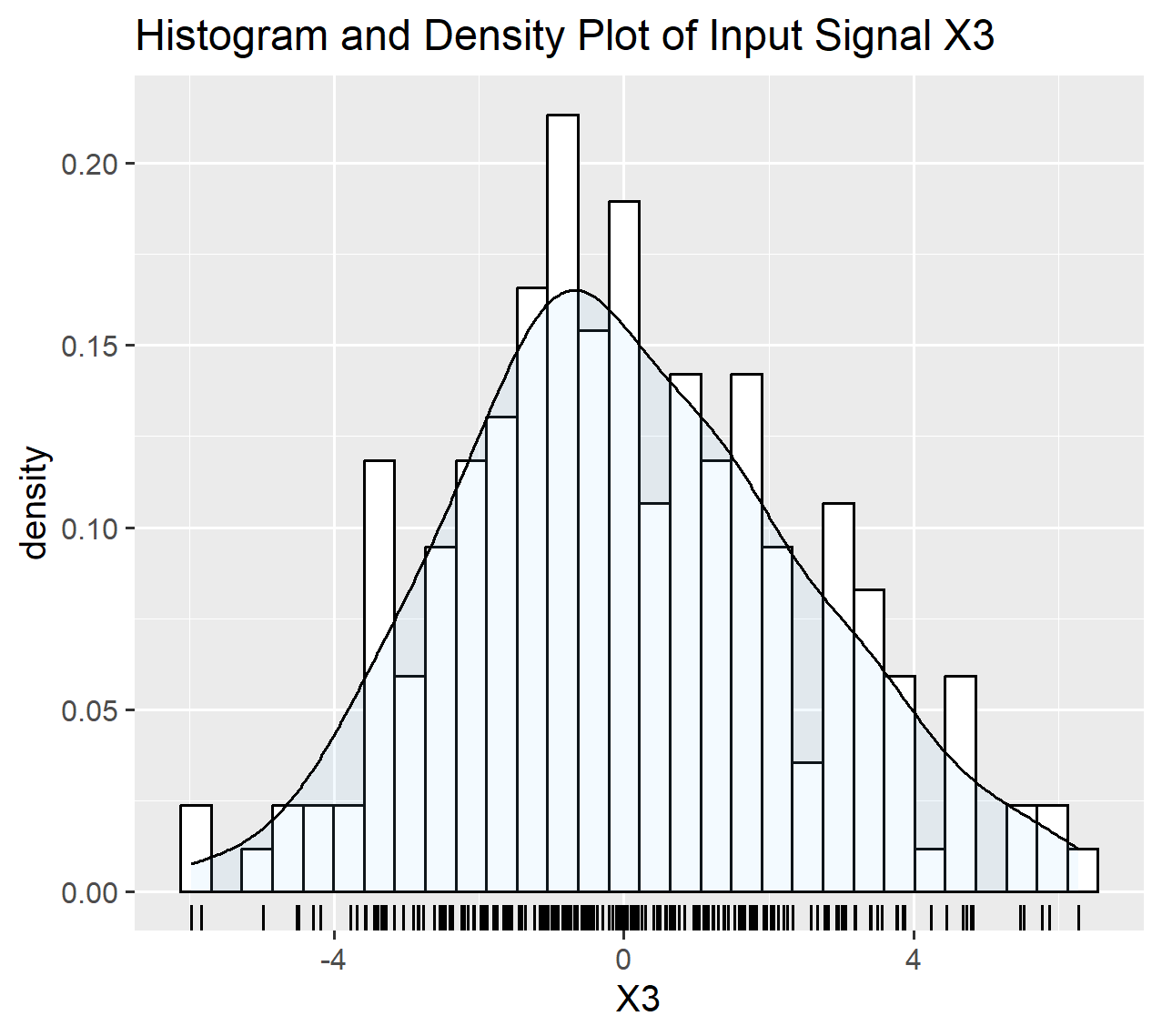


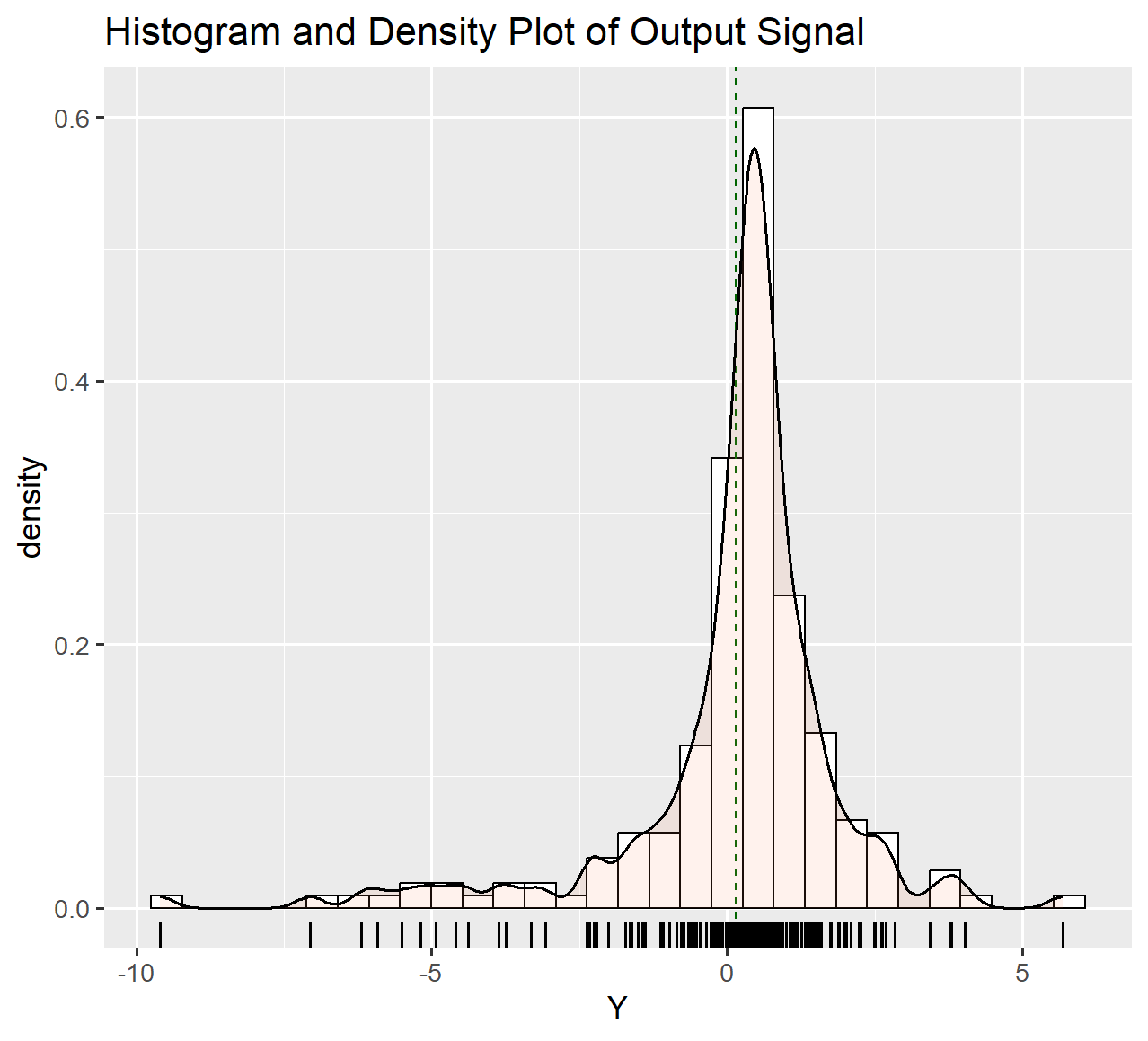
Figure 3 Plot of the Input Signal

The shape of the density plot represents a bell-shaped curve. From the figure we can see that most values are accumulated around the center and there are lesser values at the far end signifying that the signal does not have a lot of extreme values or outliers as the tails of the density plot tapers down.





The distribution shapes of all the input signals are fairly close to a bell shaped curve. They data does not have extreme outliers so there is no need to refactor the dataset.



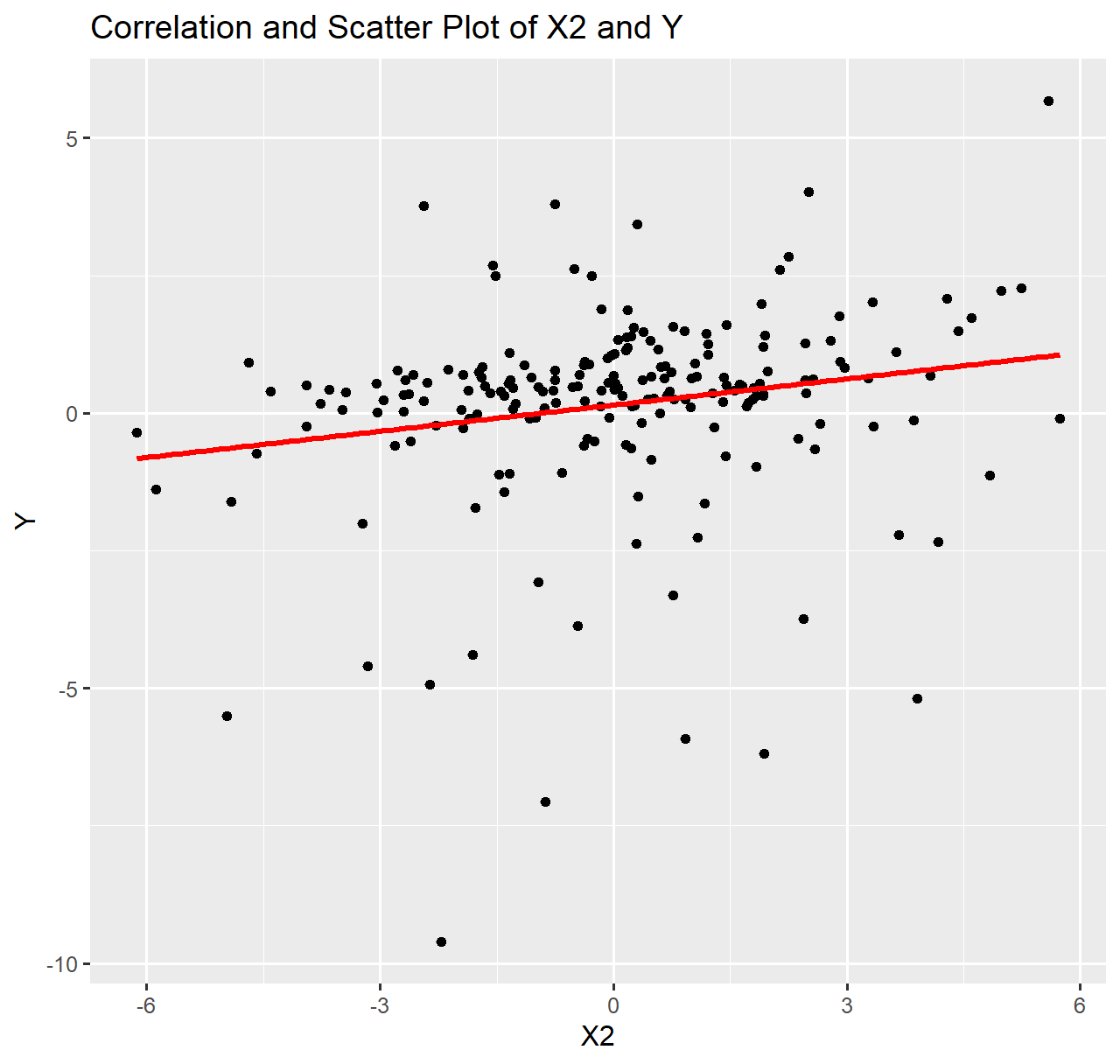
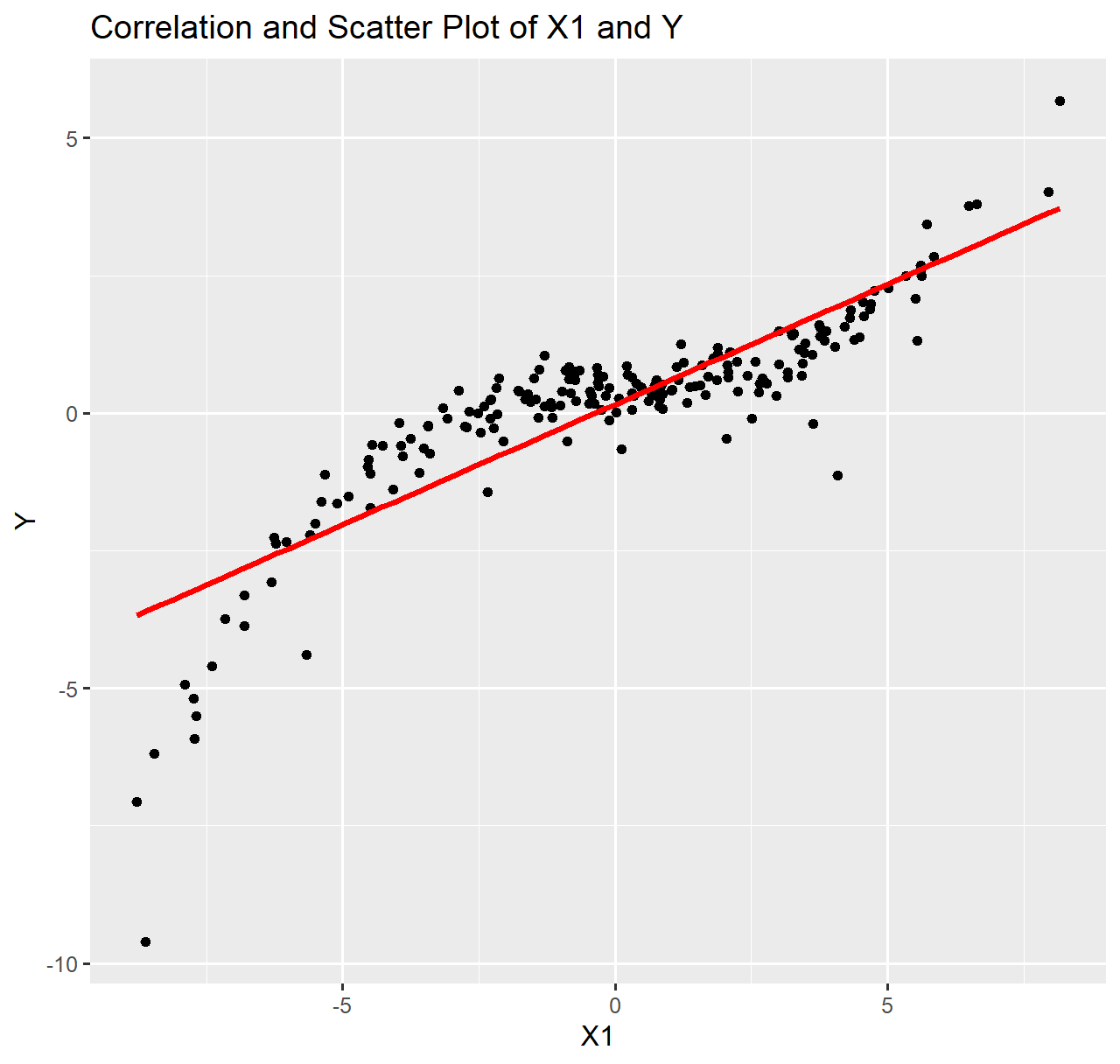
From the plot above, we can see that most values

## Correlation and Scatter Plots

Scatter Plot is a kind of plot where data points of two variables are plotted in a two-dimensional plane. Generally, the dependent variable is plotted in the Y-axis whereas the independent variable is plotted in the X-axis. This kind of plot is also called scatter chart, scattergram or scatter graph.

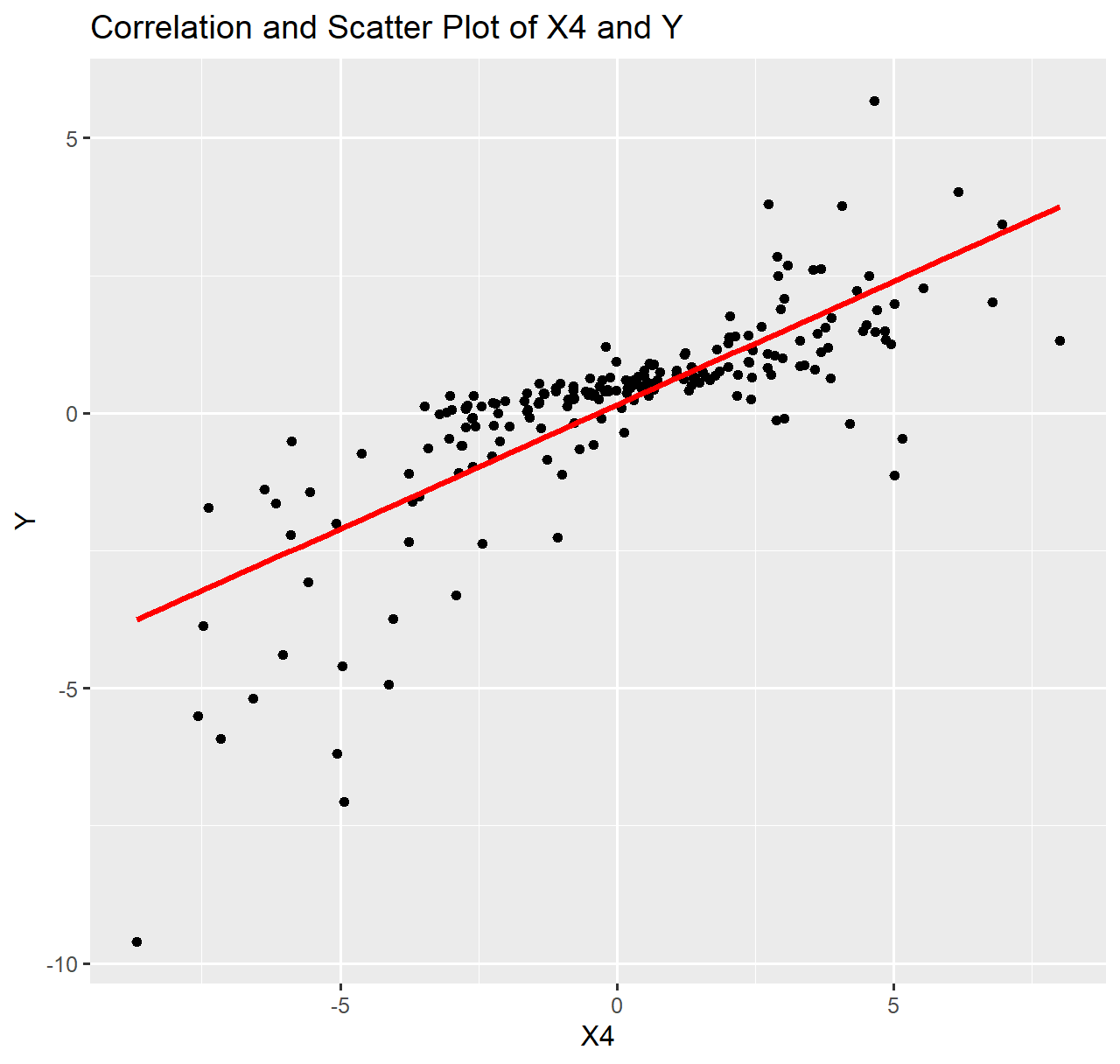
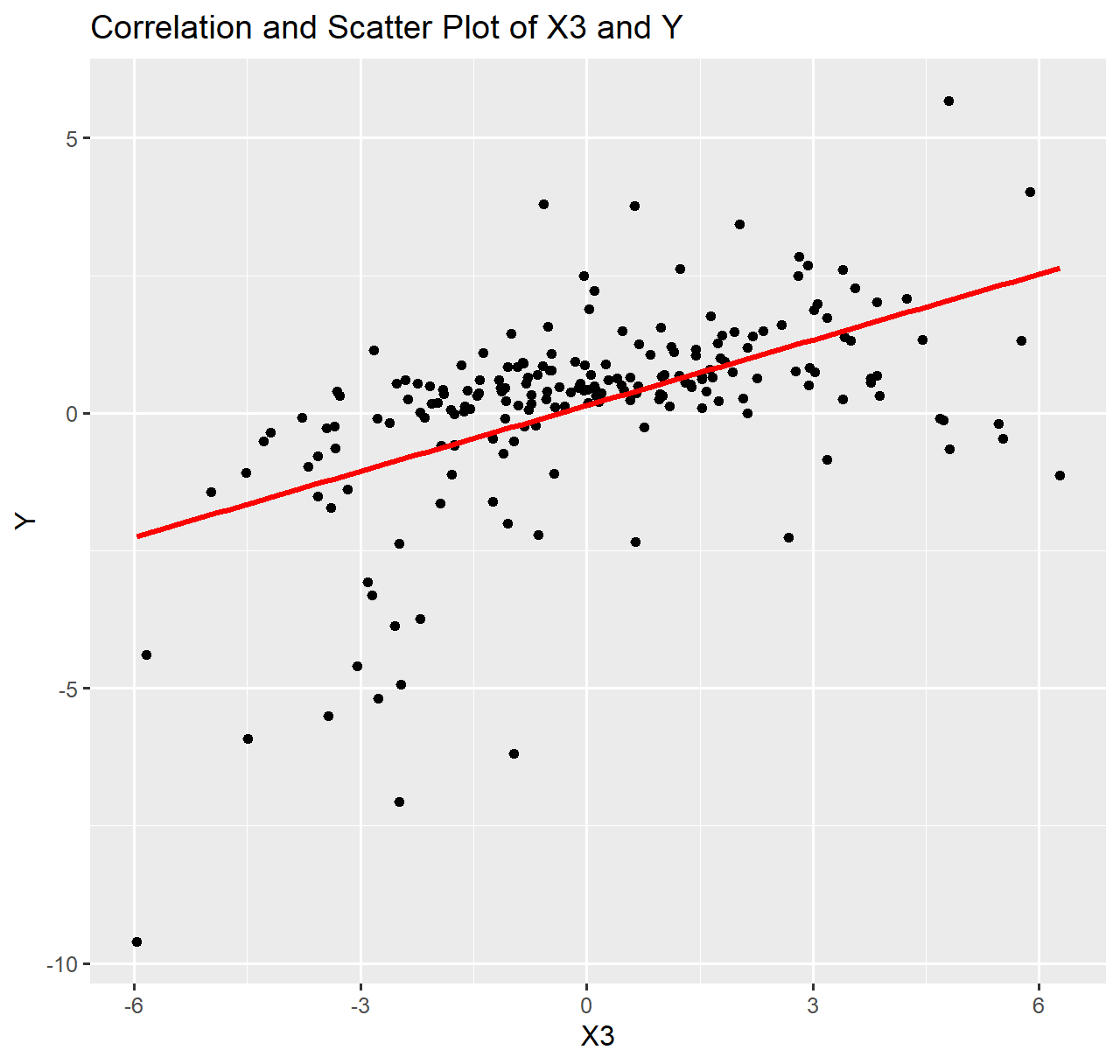
When an individual set of data point is plotted on the chart, it can be represented as a (x, y) coordinate as well. Plotting all the sets of data points on the chart, we can generally see a kind of pattern or a trend in which it either looks like rising, falling or just neutral. This kind of relation which can be seen when plotting scatter plots illustrates the degree of *Correlation* between the variables. Correlation is a common tool for describing simple relationships without making a statement about cause and effect and it also cannot accurately describe curvilinear relationships. (JMP Statitistical Discovery n.d.)

In our given dataset, the X input signals are independent variables and the Y ouput signal is a dependent variable. The scatter plot below shows the data points and the line (line of best fit) shows the correlation between the variables.



The scatter plot on the left shows the relation between the X1 or the input variable and Y or the output variable respectively. The line of best fit drawn has been drawn to to study the relationship between the variables more clearly. (Wikipedia 2023)

Both the scatter plots for X1 and X2 with respect to Y, show a positive correlation. Comparing the two, X1 and Y seem to have *higher positive correlation* than X2 and Y as the points in plot for X2 and Y are more *scattered* than in the first plot.



The points in the plot for X3 and Y are more scattered than that of X4 and Y. Both the plots show a positive correlation, whereas X4 and Y seem to have higher positive correlation.

Analysing all four scatter plots, the signals X1 and X4 seem to be have a high correlation with Y whereas for X3 and X2, the data points are a lot more scattered and even though it has a positive correlation, it is lower than that of X1 and X4.

# Task 2

## Regression – modelling the relationship between EEG signals

As we have already seen above in the scatter plots, drawing a simple linear regression line of best fit does not help much to define a model that can properly describe the relationship between the datas.

For such nonlinear data, polynomial regression model may help in defining a relationship better. Here five different nonlinear polynomial regression models have been provided. Out of the five models, the goal is the find the model which can best define the relationship.

## Task 2.1

The method of Least Squares can be used to estimate parameters by minimizing the sum of squares of the difference between the observed data and the value provided by a model.

All the provided models have some parameters which needs to be estimated. These parameters can be estimated by using the Least squares method. Least Squares method is denoted by and it can be calculated as,

where, X = input signal and Y = output signal data of the EEG signals.

For calculation of the in R, the above equation can be formulated as,

The first non-linear polynomial regression model (model 1) is,

The model above, in R would look like,

Xmodel1 <- cbind(ones,(X[,'X4']),(X[,'X1'])^2,(X[,'X1'])^3,(X[,'X2'])^4,(X[,'X1'])^4)

After creating a model, the value of for the model is calculated using the formula shown above which, in R would look like:

model1\_thetahat = solve(t(Xmodel1)%\*%Xmodel1)%\*%t(Xmodel1)%\*%Y

The table given below is the value of ‘model1\_thetahat’ calculated in R.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0.401580779 | 0.12771011 | -0.000290217 | 0.009668813 | -0.000409892 | -0.000154337 |

values for model 2,

|  |  |  |  |
| --- | --- | --- | --- |
| 0.483065688 | 0.143578928 | 0.010038614 | -0.001912836 |

values for model 3,

|  |  |  |
| --- | --- | --- |
| 0.340561975 | 0.021330543 | -0.002857744 |

values for model 4,

|  |  |  |  |
| --- | --- | --- | --- |
| 0.509013488 | 0.053322048 | 0.012067145 | -0.001855997 |

values for model 5,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0.479828463 | 0.14344607 | 0.000325464 | 0.010056276 | -0.001919875 |

After calculating the values for all the models, those values are used to calculated the of all the models.

## Task 2.2

## Computing the Model Residual Sum of Squared Errors (RSS)

The values calculated in the previous step are all estimated using the models. The actual values of Y are already present in the provided dataset. All five models have their own estimated .   
The difference between the actual value and the value predicted by the model is called error or residual. The error can be represented as,

The sum of squares of all those errors is called RSS. It measures the variance in the value of the observed data when compared to its predicted value as per the regression model as also seen in the figure below. (wallstreetmojo.com)

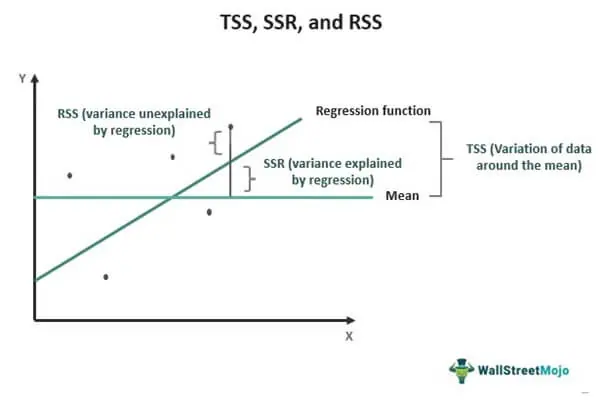


Figure 4 Residual sum of squares (wallstreetmojo.com)

It can be denoted as:

where,.

In R, the of model 1 is calculated as,

Y\_hat\_m1 = Xmodel1 %\*% model1\_thetahat

Similarly, the of all other remaining models is also calculated and then used to further calculate the RSS of all the models.

The RSS is calculated in R as,

RSS\_Model\_1=sum((Y-Y\_hat\_m1)^2)

The above equation is for calculating RSS of model 1. The RSS of remaining models is calculated in similar way.

|  |  |
| --- | --- |
| Model | RSS Value |
| Model 1 | 35.39663 |
| Model 2 | 2.139762 |
| Model 3 | 463.3124 |
| Model 4 | 20.259 |
| Model 5 | 2.135503 |

The table shows the RSS values obtained in R for all the models. Generally, when calculating RSS of a model with reference to provided potential models, the model which has the minimum RSS is said to be the model which fits the actual dataset better than other models.

Taking a look at our table, the models 2 and 5 have the most minimum values of RSS comparing to all other models. So, the regression models 2 and 5 must fit the actual dataset better than any other model present.

## Task 2.3

## Computing the Log-Likelihood Function

From the above tasks, the values of the various parameters of the models have been estimated. The parameters are a part of the model and has a great role in the output or the actual data. Likelihood method is a measure that explains how well those parameters describes the data.  
It is a measure of the likeliness of getting or estimating data that resembles the actual data, given the parameters and the model.

Log-likelihood is simply a likelihood function but it uses the logs of likelihood. The log is taken generally because it is usually computationally simpler and easier to optimize. (StatisticsHowTo n.d.)

The equation for log-likelihood function is,

Where,

is the log-likelihood.

is the total number of Y signals.

is the variance, which can be calcuated using RSS from previous task as,

The equation of log-likelihood can be represented and used in R as,

likelihood\_1=(N/2)\*(log(2\*pi))(N/2)\*(log(variance\_model1))(1/(2\*variance\_model1))\*RSS\_Model\_1

Similary calculating the log-likelihood of all the models gives the following values.

|  |  |
| --- | --- |
| Model | Likelihood Value |
| 1 | -110.6707 |
| 2 | 171.3245 |
| 3 | -369.1351 |
| 4 | -54.58995 |
| 5 | 171.5247 |

The log-likelihood values above show that for models 2 and 5, the values are close and highest. The high values of log-likelihood in those models show that they fit the dataset better than the other models.

## Taks 2.4

## Computing the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

Selecting a good model that is not too complex but can still describe the dataset well is crucial. As adding or having more parameters in a model makes it complex whereas it might not be as explanatory as other models with lesser parameters or lesser complexity.

Akaike Information Criterion (AIC) is one of the methods which helps in choosing a better model with the help of number of prameters used in a model and the log-likelihood function. It entertains parsimoniousness in the use of number of parameters in a model as it penalizes the model more if it has more parameters and also prevents overfitting of model. (Kutz, 2020)

AIC can be represented as:

where,   
k is the number of parameters used in a model.

is the log-likelihood function.

Bayesian Information Criterion (BIC) is a method similar to AIC but it has a different penalty factor. BIC can be represented as:

As both AIC and BIC provide scores based on how complex a model is and how better the model explains the data, the model which has a smaller score than other models is considered to be performing well.

The 5 regression models have different number of parameters as already seen in previous tasks. Using it and the log-likelhood calculated in previous taks, the table below shows the values of AIC for different models.

|  |  |
| --- | --- |
| Model | AIC Value |
| 1 | 233.3414 |
| 2 | -334.6489 |
| 3 | 744.2702 |
| 4 | 117.1799 |
| 5 | -333.0493 |

The table below lists the calculated values of BIC for different models.

|  |  |
| --- | --- |
| Model | BIC Value |
| 1 | 253.1613 |
| 2 | -321.4357 |
| 3 | 754.1801 |
| 4 | 130.3931 |
| 5 | -316.5328 |

Looking at the scores from the tables above, it can be said that out of all models, model 2 and model 5 have the lowest AIC and BIC scores.

Wikipedia (2023) *Scatter Plot* [online] available from < https://en.wikipedia.org/wiki/Scatter\_plot> [4 Feb 2023]

JMP Statitistical Discovery (n.d) *Correlation* [online] available from <https://www.jmp.com/  
en\_ca/statistics-knowledge-portal/what-is-correlation.html#:~:text=Correlation%20is%20a%20  
statistical%20measure,statement%20about%20cause%20and%20effect.> [4 Feb 2023]

CFA Institute (n.d.) *Residual Sum of Squares* [online] available from < https://www.wallstreetmojo.  
com/residual-sum-of-squares/> [5 Feb 2023]

G. Stephanie (n.d.) *Log Likelihood Function* [online] available from < https://www.  
statisticshowto.com/log-likelihood-function/> [5 Feb 2023]

K. Nathan (2020) *Model selection: Information criteria.* Youtube [online video] available from <https://www.youtube.com/watch?v=Gc9EzmfcSas&ab\_channel=NathanKutz> [5 Feb 2023]