Name: Shekhinamary Jebaraj
Student ID: 201740454
Module: Time Series and Spectral Analysis (MATH5802M)

TIME SERIES ANALYSIS AND FORECASTING OF SLAVE MIGRATION ACROSS THE ATLANTIC

INTRODUCTION

The dataset estimates the annual number of enslaved people transported by British ships across the Atlantic. The analysis seeks to illuminate this obscure time by identifying patterns, trends, and perhaps predicting models associated to these missions. The dataset's key components are years, which indicate the voyages' chronological periods, and num, which estimates the number of enslaved people engaged in each year. Beyond descriptive analysis, statistical and time series analysis are used to anticipate future trends.

SUMMARY OF THE FINDINGS

Upon doing an analysis of the number of slaves who embarked on a voyage across the Atlantic Ocean — on a ship under a British flag, a number of significant observations come to light. The preliminary investigation unveiled a discernible upward trajectory in the quantity of enslaved individuals over the course of time, accompanied by intermittent variations in specific years. In order to gain a deeper comprehension of the fundamental patterns, a linear regression model was utilized to exclude the linear trend, yielding residuals labeled as Y. Following that, an examination was conducted on an autoregressive (AR) process utilizing the Yule-Walker equations, and AR models of varying orders were employed for fitting purposes. Significantly, the AR(3) model demonstrated the ability to capture noteworthy patterns within the data. The frequency content of the number of slaves (X), detrended data (Y), and AR(3) residuals (Z) was analyzed using periodograms. The periodograms yielded valuable insights regarding the prevailing frequencies and fluctuations observed within the datasets. The findings of the investigation indicated that the application of detrending and autoregressive (AR) modeling had an impact on the frequency characteristics, resulting in the emergence of discernible patterns within the residuals. The examination of distinct frequencies exhibiting variations in power peaks across datasets revealed the influence of modeling decisions on the observed patterns. The ARIMA (3,0,0) model was eventually chosen as an appropriate model, and a forecast for a period of 5 years was prepared, integrating confidence ranges to accommodate for uncertainties. The projected values then indicated a downward trajectory in the upcoming five-year period, implying an anticipated reduction in the population of enslaved individuals.

DATA ANALYSIS

Plotting Original Data and Linear Trend :

The original time series plot shows a significant annual variation in the number of slaves sailing under the British flag across the Atlantic over the years. The estimated yearly trend for the number of slaves sailing under the British flag shows a mostly upward trend. Over the designated historical period, the number of embarked slaves increased steadily, as shown by the linear regression line. A clear upward trajectory of the linear regression line indicates a relatively strong trend. It also points to a steady and significant increase in the number of slaves traveling across the Atlantic.

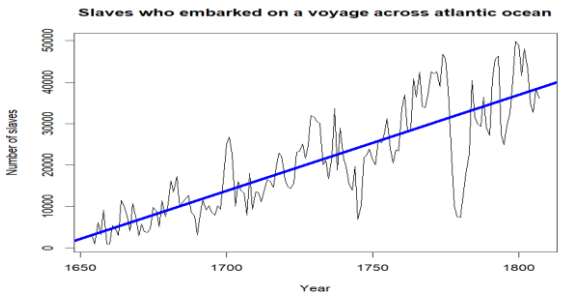

Fig 1: Original Plot

Residual Analysis :

The deviations from the linear trend are displayed in the residual plot, which also may indicate seasonality or any nonlinear patterns. The residual plot illustrates the patterns left after the removal of the linear trend.

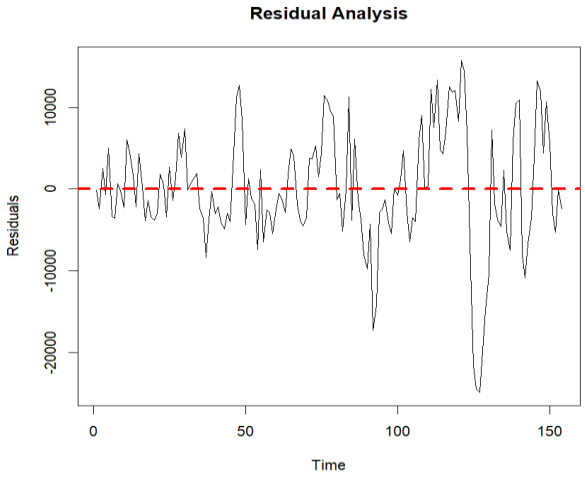

Fig 2: Residual Plot

The horizontal red line at Y=0 serves as a reference, indicating points where the data deviates either above or below the linear trend. Points above the red line represent years with higher-than-predicted numbers, while points below the line indicate lower-than-predicted numbers. A sizable amount of the variability in the annual estimated number of slaves embarked is likely captured by the linear model, as evidenced by the residual plot's relatively narrow spread around the zero line. The residuals' tight clustering suggests that the model largely accounts for the observed data, leaving little variability that cannot be explained. Also, The plot indicates a lack of distinct recurring patterns or cycles throughout the years as it is not exhibiting a clear periodic pattern and hence, seasonality is not present in the data.

ACF and PACF of residuals :

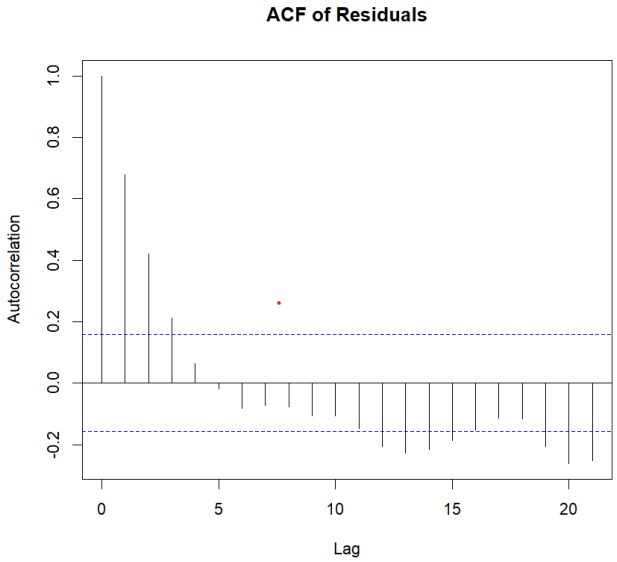

Fig 3: ACF Plot

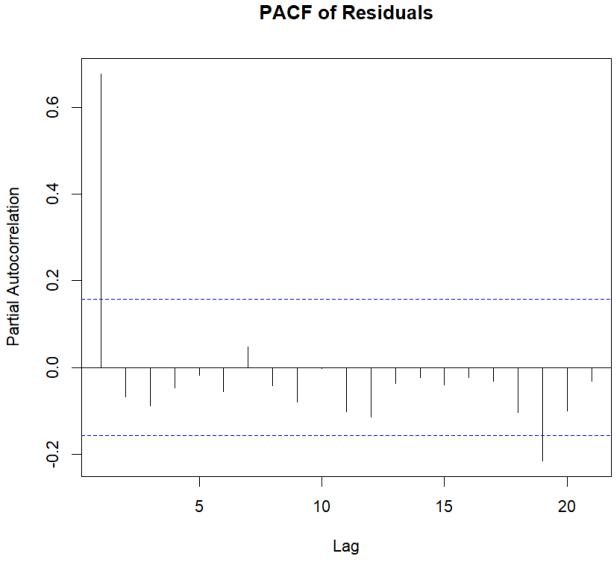

Fig 4: PACF Plot

There are two prevalent methodologies for modeling time series data, namely the Moving Average (MA) and AutoRegressive (AR) and processes.

The Moving Average (MA) process is a statistical technique utilized for the analysis of time series data, with the purpose of mitigating short-term fluctuations or noise through smoothing. The main objective of this method is to uncover the fundamental patterns present in the data by computing the mean of successive observations within a designated timeframe or interval.

The AutoRegressive (AR) process is a mathematical framework employed to characterize a time series in which each individual data point is forecasted by considering its own historical values. Stated differently, the future values of the time series are represented as a linear combination of its previous observations.

Upon examination of both the autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots, it is evident that the ACF values exhibit a gradual decrease and lack a discernible pattern of decay beyond the initial lag. This observation implies a gradual decline in autocorrelation, which is characteristic of an autoregressive (AR) process.

The partial autocorrelation function (PACF) exhibits a pronounced peak at the initial lag (lag 1) and subsequently diminishes gradually. The significant decrease observed following the initial lag in the partial autocorrelation function (PACF) also suggests the presence of an autoregressive (AR) process.

The presence of a notable autocorrelation at the initial lag in the partial autocorrelation function (PACF), coupled with the gradual decline in the autocorrelation function (ACF), indicates that an **AutoRegressive (AR) model** may be better suited to elucidate the underlying patterns observed in the data.

Fit AR Models :

For an AR(p) process, the **Yule-Walker equations** are a system of linear equations that can be solved to obtain the autoregressive coefficients as it offers a technique for estimating the parameters of an autoregressive model using the sample autocorrelation function (ACF) derived from the time series data. The utilization of this technique is advantageous for the purpose of fitting an autoregressive (AR) model to the observed data. The strength and direction of these relationships are determined by the coefficients. To fit an AR(p) model to a time series using the Yule-Walker equations in R, you can use the 'ar' function.

Arfit1 <- ar(resid1, order=1, aic=F)

Arfit2 <- ar(resid1, order=2, aic=F)

Arfit3 <- ar(resid1, order=3, aic=F)

The equations obtained are

AR(1) : $X_t = 0.67X_{t-1} + Z_t$

AR(2) : $X_t = 0.72X_{t-1} - 0.06X_{t-2} + Z_t$

AR(3) : $X_t = 0.71X_{t-1} - 0X_{t-2} - 0.08X_{t-3} + Z_t$

Each of the AR(p) models (AR(1), AR(2), and AR(3)) uses the estimated values for σ to represent the variance of the error term. These numbers show how well the models fit together and are acquired during the fitting process. In our case

For AR(1): `σ estimated as 28766227`

For AR(2): `σ estimated as 28820780`

For AR(3): `σ estimated as 28786148`

The accuracy of the model's forecasts can be evaluated using these numbers. When the value of σ is smaller, it means that the model is fitting the data better and can explain more of the variability.

Looking at the residual and squared residual correlograms is one way to evaluate the AR(p) models' data fitting abilities. The correlogram is a visual depiction of the autocorrelation function (ACF), which quantifies the degree of correlation between a time series and its lagged values at various time lags.

The correlogram's horizontal axis (X-axis) shows temporal delays. Every data point on the x-axis has a lag value, which represents the amount of time units since the correlation calculation.

Autocorrelation magnitude is shown on the vertical axis (Y-axis). A complete negative correlation is represented by -1 on the y-axis, a complete positive correlation by 1, and no correlation by 0.

To show confidence limits, plots often have dotted lines or shaded areas. Points outside these borders may be statistically significant.

Interpreting Correlograms :

At varying lags, autocorrelation intensity and sign are shown by correlogram bar height and position. A bar outside confidence bounds may indicate a strong association at that latency.

Negative lags indicate prospective linkage, while positive lags show historical association.

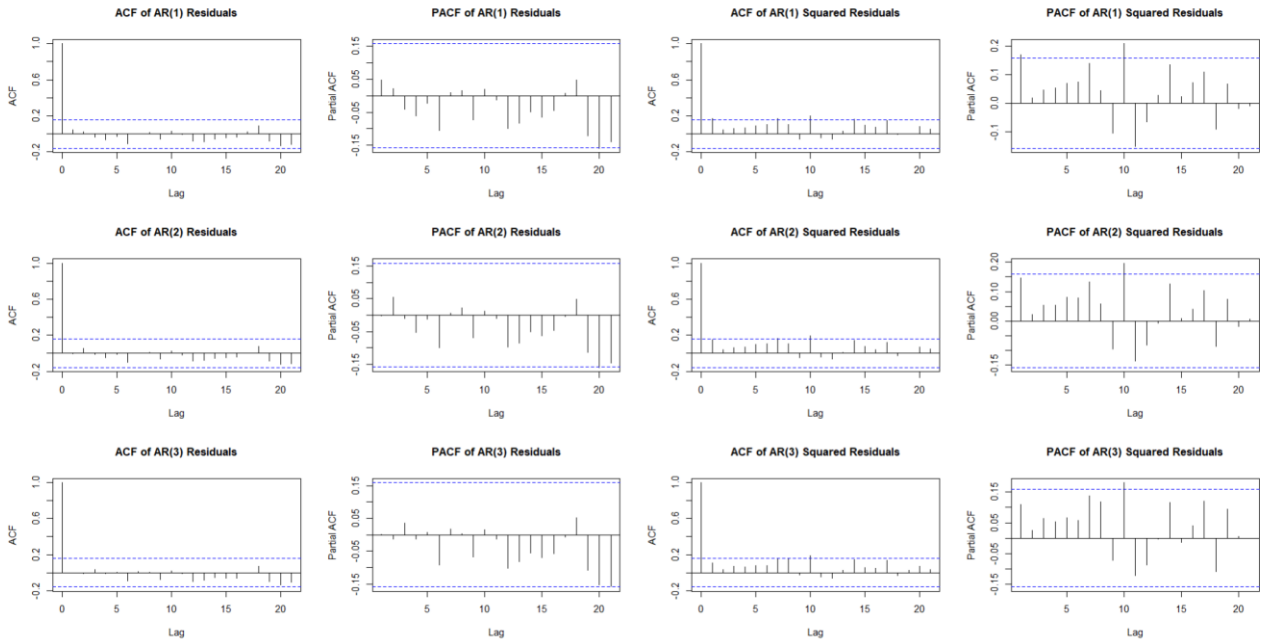

Fig 5: Correlograms of ACF/PACF of AR(1), AR(2), AR(3) and its squared residuals

AR(1) :

There is a gradual decline in autocorrelations (ACF), indicating a slow decline. Positive correlation (0.047) with the prior observation is shown by the autocorrelation coefficient at lag 1.

Partially autocorrelated functions (PACF) tend to drop after the initial lag, implying that the primary lag captures most of the dependency.

$AR(1)^2$:

Autocorrelations characterize the squared first-order autoregressive process (AR(1)2). The autocorrelations gradually decrease, with a positive autocorrelation at lag 1 (0.170). The pattern is weaker than in the AR(1) model.

Similar to partial autocorrelations (PACF), lag 1 has a statistically significant value of 0.170.

AR(2) :

Order 2 autoregressive model (AR(2)): The autocorrelation function (ACF) shows diminishing autocorrelations. At lag 1, the autocorrelation is negative (-0.003), reversing the direction of the AR(1) model. The observed pattern is more complex.

Partial autocorrelations (PACF) with alternating positive and negative values are also complex.

$AR(2)^2$:

In the squared AR(2) model, autocorrelations gradually decrease, with a positive autocorrelation at lag 1 (0.147). The observed pattern is weaker than AR(2).

A statistically significant partial autocorrelation (PACF) value of 0.147 at lag 1 matches the pattern.

AR(3) :

In the AR(3) model, autocorrelations (ACF) have a more complex pattern with oscillations. At lag 1, the autocorrelation coefficient is insignificant, but at lag 2, it is 0.001.

Partial autocorrelations (PACF) have a similar complicated pattern of positive and negative values.

$AR(3)^2$:

In the squared AR(3) model, autocorrelations drop steadily. Note the 0.110 positive autocorrelation at lag 1. The observed pattern is weaker than AR(3).

Similar to partial autocorrelations (PACF), lag 1 has a statistically significant value of 0.110.

Overall assessment, based on a comprehensive evaluation:

The data is captured by the autoregressive model of order 1 (AR(1)), which shows positive autocorrelation at lag 1.

Alternating positive and negative autocorrelations give the autoregressive model of order 2 (AR(2)) complexity.

The autoregressive model of order 3 (AR(3)) adds complexity with oscillating patterns and alternating autocorrelations.

According to the autocorrelation function (ACF) and partial autocorrelation function (PACF), the **AR(3) model is the best choice** because it captures significant autocorrelations at a lag of 3 and significant partial autocorrelations at longer lags.

Periodograms :

The spectrum of a time series signal can be visually represented by a periodogram. Finding the frequencies in a data set is a common task in signal processing and analyzing time series. A periodogram is a graphical representation of a signal's power distribution across all of its frequencies.

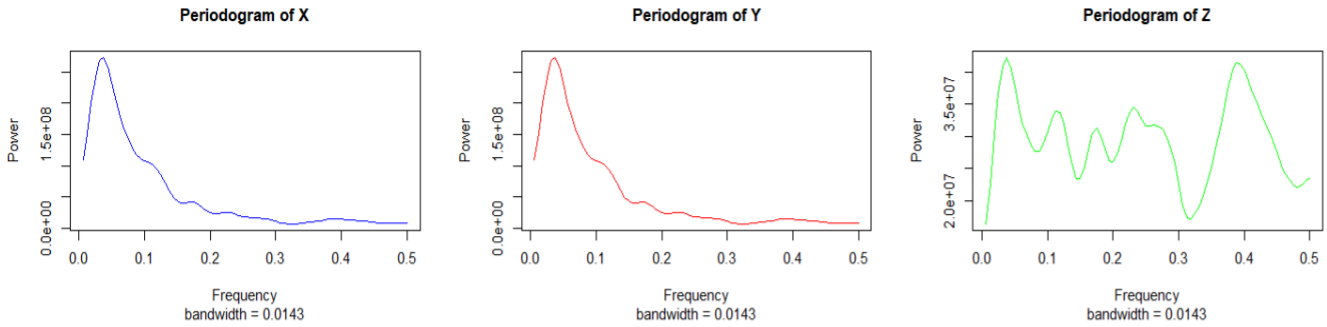

Fig 6: Periodograms of X, Y and Z

The variable X represents the initial dataset, denoted as num. The variable Y represents the detrended dataset. Lastly, the variable Z represents the residuals obtained using an autoregressive model of order 3 (AR(3)).

The process of interpreting the results obtained from the periodograms include analyzing the frequencies and their related power values for each dataset (X, Y, Z).

The following are comprehensive remarks for each of them –

Periodogram of X :

The frequencies in the periodogram are uniformly distributed, thereby including the entire range of frequencies.

The observed values span a range of 0.00625 to 0.5, denoting the frequency of cycles per observation.

The power numbers for X indicate the magnitude or intensity of each frequency component.

The presence of peaks in power indicates the presence of dominating frequencies within the original data.

Periodogram of Y :

In a manner akin to X, frequencies that are uniformly distributed and represent the entire spectrum of cycles per observation.

The presence of peaks in power within the detrended data, subsequent to the removal of linear trends, signifies the prevalence of dominant frequencies.

Periodogram of Z :

The frequencies are uniformly distributed, spanning the entire range of cycles per observation.

The presence of peaks in power within the residuals derived from the AR(3) model signifies the existence of dominant frequencies.

In-depth Comparative Analysis :

The disparities in power between X and Y underscore the significance of detrending.

The presence of peaks in the power values of Y may indicate the existence of inherent cyclical patterns that cannot be accounted for by a linear trend.

The comparison between X and Z highlights the significance of AR(3) modeling in determining variations in power.

The presence of peaks in the power values of Z may suggest the existence of frequencies that were not accounted for by the AR(3) model.

The power disparities between Y and Z are indicative of the collective impact resulting from the application of detrending and AR(3) modeling techniques.

The presence of peaks in the power values of Z could potentially indicate the existence of residual patterns following the application of both detrending and AR(3) modeling techniques.

Also, The observed frequency for X, Y, and Z consistently reaches a maximum value of 0.0375 cycles each observation. The findings of this study suggest that the primary frequency observed in the dataset, as determined by the peak with the highest power, remains consistent following the application of detrending (Y) and AR(3) modeling (Z).

Now, for the chosen model we are refitting the AR parameters using the command *arima.* We do this for several reasons such as

The utilization of a distinct implementation permits the validation of outcomes derived from alternate methodologies, such as the Yule-Walker equations. This practice aids in maintaining the uniformity and precision of the calculated parameters.

Various functions or methods may employ diverse algorithms and optimization strategies to estimate autoregressive (AR) parameters. By employing an alternative function, such as ARIMA, to re-estimate the model, one can evaluate the resilience of the obtained outcomes.

ARIMA Model Fitting :

To fit an ARIMA model, the following code was used –

arima_model<-arima(data_1, order=c(3,0,0), include.mean = FALSE)

arima_model

summary(arima_model) – gives the summary of the model

On running this, the output provides an overview of the coefficients and other relevant factors pertaining to the fitted ARIMA(3,0,0) model

```
Coefficients:
      ar1     ar2     ar3
   0.8812  0.0242  0.0709
s.e.  0.0800  0.1076  0.0806
```

The calculated coefficients for the autoregressive terms are denoted as ar1, ar2, and ar3.

The variable "s.e." denotes the standard errors of the coefficients that correspond to the given data.

σ^2 estimated as 33177667: log likelihood = -1553.42, aic = 3114.84

σ^2 represents the estimated variance of the white noise process.

The log likelihood and AIC (Akaike Information Criterion) serve as metrics for evaluating the goodness of fit of a model to the data, where a lower AIC value signifies a superior fit.

Summary Information :

The summary information presents comprehensive facts regarding the different elements of the fitted model, encompassing coefficients, residuals, and model specifications.

The elements encompassed under this framework consist of many attributes, namely the length, class, and mode. These attributes pertain to coefficients (coef), residuals (residuals), and other relevant components.

```
Length    Class       Mode
coef         3    -none- numeric
sigma2       1    -none- numeric
var.coef     9    -none- numeric
mask         3    -none- logical
loglik       1    -none- numeric
aic          1    -none- numeric
arma         7    -none- numeric
residuals  154      ts   numeric
call         4    -none- call
series       1    -none- character
code         1    -none- numeric
n.cond       1    -none- numeric
nobs         1    -none- numeric
model       10    -none- list
```

Forecast :

Finally, using the command predict on the result of the ar fit, forecast for the next 5 years has been obtained (i.e. 1808–1812)

The code used to predict the data was –

```
forecast <- predict(arima_model, n.ahead = 5)

forecast_summary <- cbind(data_1,
                          forecast$pred,
                          forecast$pred - 1.96 * forecast$se,
                          forecast$pred + 1.96 * forecast$se)

colnames(forecast_summary) <- c("Original", "Forecast", "Lower CI", "Upper CI")
```

According to the ARIMA(3,0,0) model, the projected trend in the number of slaves embarked on journeys across the Atlantic Ocean over the upcoming five-year period (1808-1812) is as follows -

	Original	Forecast	Lower CI	Upper CI
1808	NA	35080.96	23791.35	46370.57
1809	NA	34497.69	19450.54	49544.85
1810	NA	33809.09	16255.60	51362.59
1811	NA	33114.04	13384.85	52843.23
1812	NA	32443.55	10816.10	54071.01

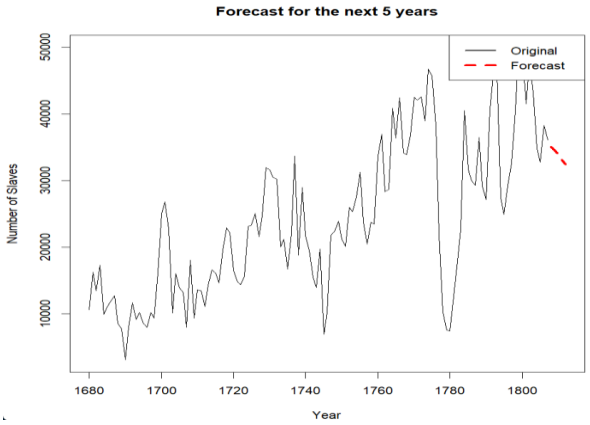

Fig 7: Forecast Plot for the next 5 years (1808-1812)

The projected figures for the years 1808-1812 indicate a general upward trajectory in the quantity of enslaved individuals engaged on maritime journeys. Nevertheless, it is important to highlight that the predicted data suggests a downward trend over the next five years, indicating a **potential decrease** in the number of individuals subjected to enslavement. The projected values, accompanied with their corresponding upper and lower confidence ranges, are presented for each year. In the year 1808, the projected quantity of enslaved individuals was estimated to be 35,080.96, accompanied by a lower confidence interval of 23,791.35 and an upper confidence range of 46,370.57. The ensuing years exhibit same patterns.

It is imperative to acknowledge that the projected values are accompanied by a degree of uncertainty, as evidenced by the inclusion of confidence intervals. As the interval widens, the level of uncertainty pertaining to the forecast increases. The projected values offer valuable insights into the anticipated trajectory of the quantity of slaves embarked, facilitating comprehension and readiness for likely forthcoming patterns.

UNIVERSITY OF LEEDS

You must sign this (digitally signing with your name is acceptable) and include it with each piece of work you submit.

I am aware that the University defines plagiarism as presenting someone else's work, in whole or in part, as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the UK) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the School's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Student Signature	Shekhinamary Jebaraj	Date	24/11/2023
Student Name	Shekhinamary Jebaraj	Student Number	201740454