PROJECT 3

MATH 5900 – ADVANCED DATA ANALYSIS

April 14, 2024

TAIWO JEGEDE

E00755963

PART I

I. Three descriptive statistics

i. The missingness rates: This shows us the percentage of missing values for each variable, so we can see how much data is missing.

ii. Patterns of missingness: This involves examining how missing values are distributed across variables and observations, and this can help reveal if missingness is related to certain variables or observations.

iii. Comparison of complete vs incomplete cases: Comparing the distributions of variables between complete and incomplete observations. Differences may indicate the data is not MCAR.

II. Visuals that can be used to explore missingness

i. One way is to use a Dot chart, by using a dot chart, we can get a a clear indication of the proportion of missing values for each variable, which then allows us to quickly identify variables with high levels of missingness.

ii. Data Matrix Plot: A data matrix plot provides a comprehensive view of missingness across the entire dataset, representing missing values as empty cells within a matrix-like structure. This plot shows a matrix or heatmap representation of the missing data patterns, revealing complex relationships in how the missing values are distributed.

III. Differences between MAR and MNAR

Missing at Random (MAR):

The probability of data being missing depends on other observed variables in the dataset, but not on the specific missing values themselves.

For example, if men are more likely to report their age than women, the missingness is related to the observed gender variable, so the data is MAR.

Methods like multiple imputation can work well under the MAR assumption, as the missingness can be explained by the observed data.

Missing Not at Random (MNAR):

The probability of data being missing depends on the specific missing values themselves, even after accounting for other observed variables. For example, if the sickest patients are more likely to drop out of a study then the missingness is related to the unobserved values, so the data is MNAR. MNAR patterns are more challenging to identify and handle, as the missingness cannot be fully explained by the observed data.

So in essence, under MAR, the missingness can be accounted for by the observed data, while under MNAR, the missingness depends on the unobserved, missing values themselves.

IV. Why we would have bias in our data

If we ignore and delete missing data from the study, then it is expected that we have bias in our effects estimates and hypothesis tests becuase:

Loss of Statistical Power: Deleting observations with missing data reduces the overall sample size, which can lead to a loss of statistical power. This means that our ability to detect significant effects, if they exist, would be reduced.

Systematic Differences: If the observations with missing data differ systematically from the observations with complete data, deleting the missing data would result in a biased sample. This could lead to biased estimates of the relationships between the variables.

Violation of Assumptions: Many statistical methods, such as regression analysis, rely on the assumption that the data is missing at random (MAR) or missing completely at random (MCAR). Ignoring missing data and deleting observations would violate these assumptions, leading to biased parameter estimates and invalid statistical inferences.

PART II

PURPOSE OF THE ANALYSIS

The primary purpose of this analysis is to address the challenges posed by missing values in the dataset, including missing outcome values (student enrollment), values missing not at random, and missing non-continuous variables. By effectively handling these missing data issues, the analysis will then aim to investigate the trends and characteristics of student enrollment in Nigerian universities. Like how has student enrollment in Nigerian universities changed over the years, and what trends can be observed?

To address the research question, the analysis will employ appropriate missing data imputation techniques, such as multiple imputation, to handle the various types of missing values present in the dataset. Once the missing data issues have been resolved, the analysis will proceed to explore the trends in student enrollment and the relationships between enrollment and university characteristics, such as PhD-granting status and public/private status.

DESCRIPTION OF THE DATA

The dataset provided contains information on various universities in Nigeria, including their country, region, income group, founding year, closure year, public/private status, coordinates, latitude, longitude, PhD-granting status, master's-granting status, bachelor's-granting status, number of divisions, total number of fields, number of unique fields, specialized status, merger status, and IAU ID.

The key variables of interest for this analysis are:

students5_estimated: The estimated number of students enrolled in the university.

year: The year for which the student enrollment data is provided.

phd_granting: A binary variable indicating whether the university grants PhD degrees 1 or not 0.

private01: A binary variable indicating whether the university is private 1 or public 0.

iau_id: A unique identifier for each university.

b_granting: A binary variable indicating whether the university grants bachelor's degrees 1 or not 0.

m_granting: A binary variable indicating whether the university grants master's degrees 1 or not 0.

PROPOSED ANALYSIS

To address the research questions, the following analyses will be conducted:

Descriptive Analysis of Missing Data:

Examining the proportion of missing data in the dataset, including the percentage of missing values for each variable. I will also use the flux plot to visualize the patterns of missing data, and this will help to identify the variables and observations with the highest rates of missing values. Because we are only interested in certain columns and not all, I will subset the data so that I can only focus on the columns that answer all the research questions which has already been identified above already.

Missing Data Imputation:

After the exploration of the missing values has given us enough context, then the appropriate missing data imputation techniques will be applied such as multiple imputation, to

handle the various types of missing values present in the dataset, including missing outcome values (student enrollment), values missing not at random, and missing non-continuous variables. I will then evaluate the performance of the imputation methods using appropriate diagnostics, such as comparing the distributions of the original and imputed data using the mean, variance.

ANALYSIS RESULT

The analysis of the missing data in my dataset revealed several key findings, The variables `iau_id` and `private01` had 100% completeness, with no missing values.

For `m_granting` and `b_granting`, they had a relatively low percentage of observed values, with only 16.12% completeness. The variables `phd_granting` and `year` also had 100% completeness, with no missing values, for `students5_estimated` had a majority of observed values, with 94.10% completeness.

In this case, the p-value is 0, indicating that under the null hypothesis (that the data is MCAR), we would never expect to observe a test statistic as extreme as the one calculated from the data. Therefore, we reject the null hypothesis and conclude that the data is not missing completely at random.

After performing multiple imputation on the dataset, the analysis compared the means of the imputed variables (`m_granting`, `b_granting`, and `students5_estimated`) with the means of the observed variables, for `m_granting` and `b_granting`, the imputed means across multiple datasets were notably lower than the means of the observed variables. This suggests that the imputation process may have introduced a downward bias in estimating these variables.

For `students5_estimated`, the imputed means across multiple datasets were higher than the mean of the observed variable. This indicates a potential upward bias in estimating the number of students.

The analysis also compared the variances of the imputed variables (`m_granting`, `b_granting`, and `students5_estimated`) with the variances of the observed variables. For `m_granting`, the imputed variances across multiple datasets were slightly higher than the variance of the observed variable, indicating some variability in the imputation results. For `b_granting`, the imputed variances across multiple datasets were notably higher than the variance of the observed variable. This suggests that the imputation process may have introduced additional variability in estimating this variable. Similarly, for `students5_estimated`, the imputed variances across multiple datasets were higher than the variance of the observed variable. This indicates increased variability in the imputed estimates of the number of students.

I also provided the trace plot and the density plot, and if we take a look at the density plot, especially for students5_estimated, we will see that the blue and red lines are somewhat together, which shows us how well the imputation worked.

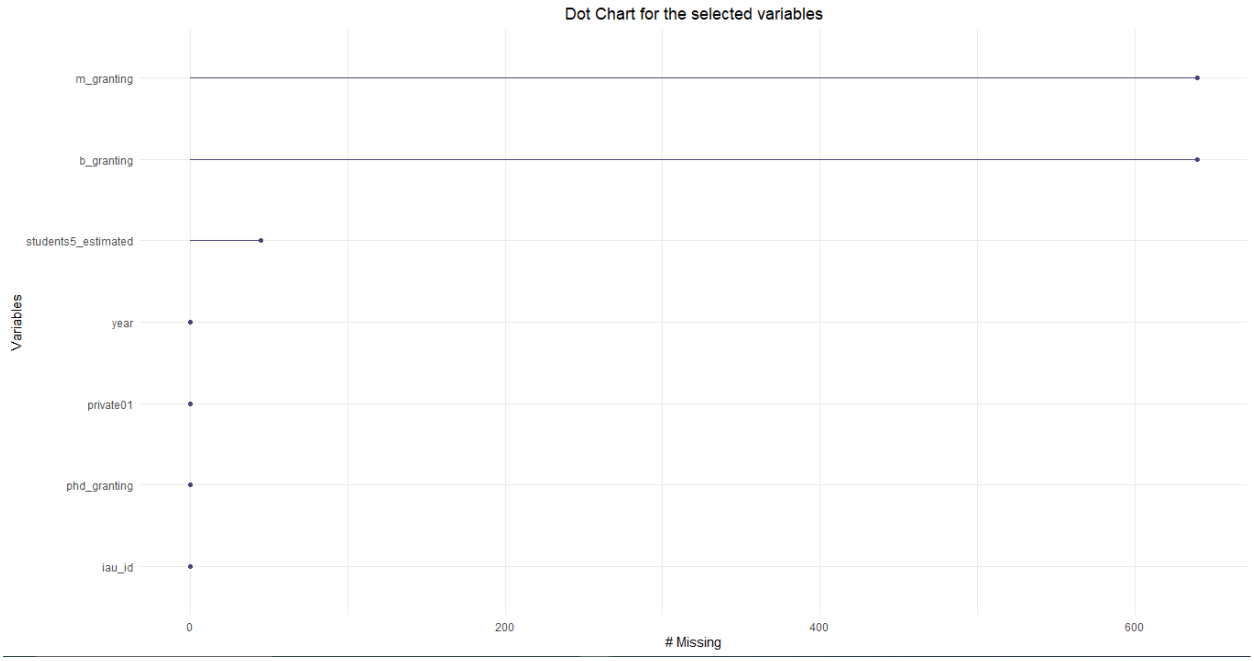
CONCLUSION

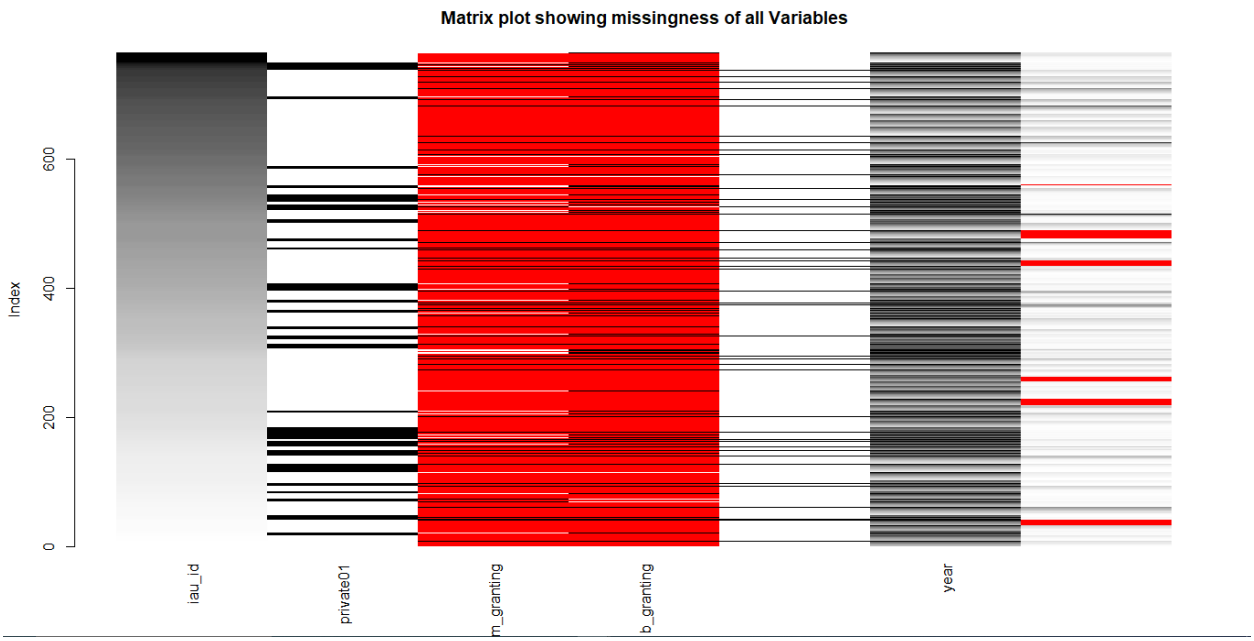
The analysis of the dataset provided me valuable insights into the trends and characteristics of student enrollment in the country's higher education system. The analysis focused on addressing the various missing data issues in the dataset, including missing outcome values (student enrollment) and values missing not at random. By employing multiple imputation techniques, the analysis was able to handle these missing data challenges and prepare the dataset for further investigation. These findings highlight the importance of carefully evaluating the

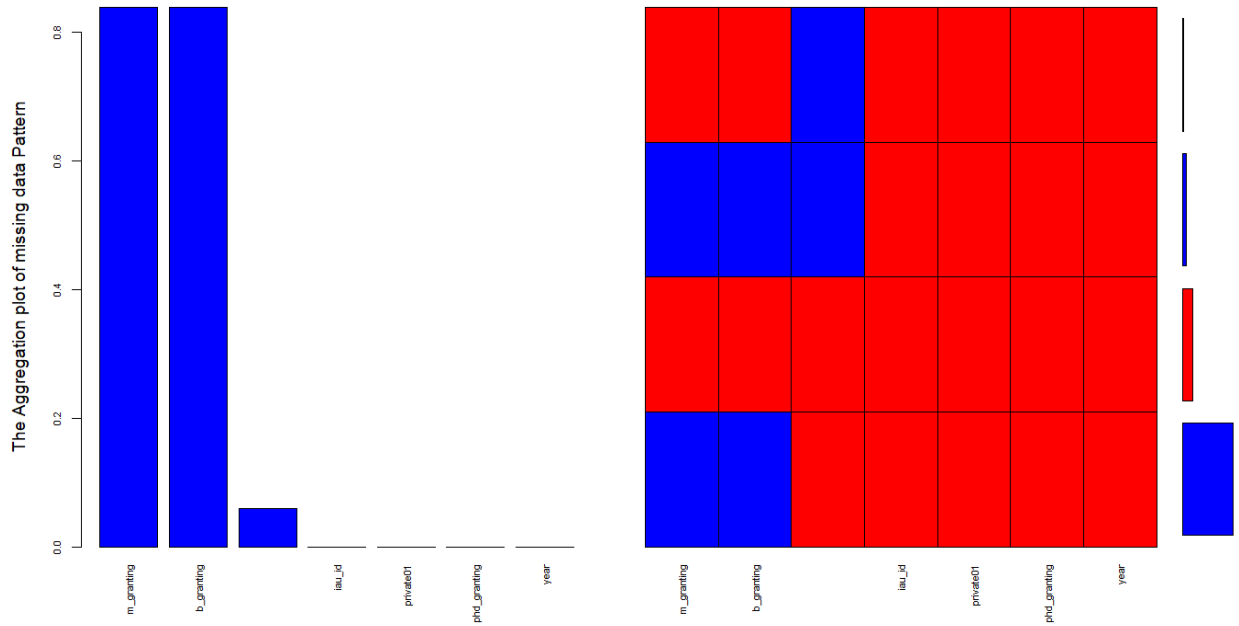
imputation process and considering potential biases and variability introduced during the handling of missing data in the dataset.

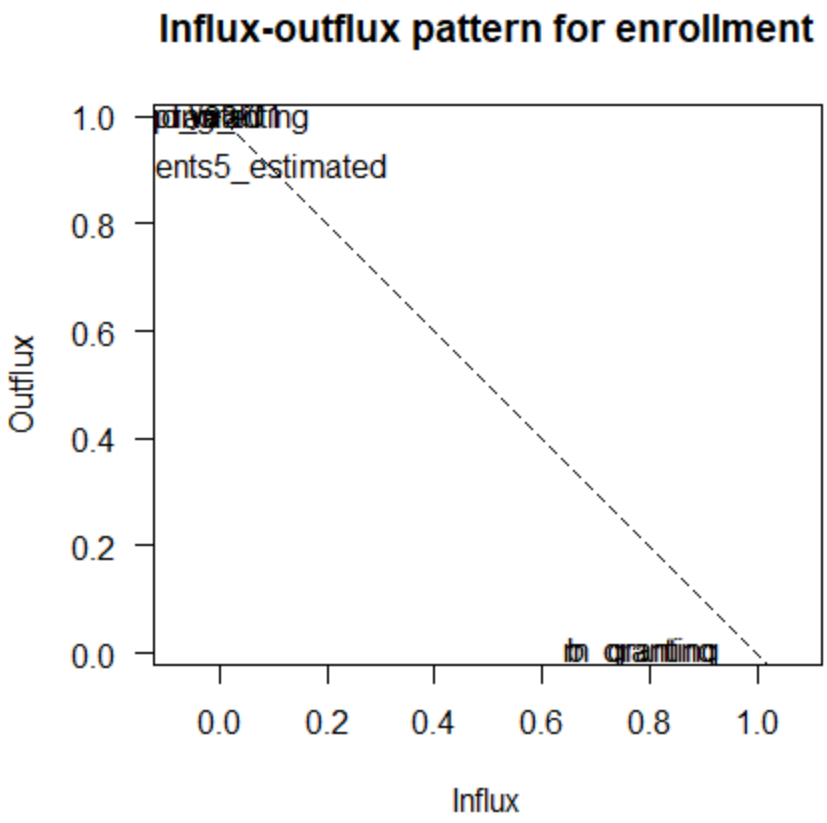
APPENDIX

```
> flux(enrollments)
                       pobs     influx     outflux      ainb        aout       fico
iau_id               1.0000000 0.00000000 1.000000000 0.0000000 0.289427698 0.84403670
private01            1.0000000 0.00000000 1.000000000 0.0000000 0.289427698 0.84403670
m_granting           0.1612058 0.78660359 0.003018868 0.8226562 0.005420054 0.03252033
b_granting           0.1612058 0.78660359 0.003018868 0.8226562 0.005420054 0.03252033
phd_granting         1.0000000 0.00000000 1.000000000 0.0000000 0.289427698 0.84403670
year                 1.0000000 0.00000000 1.000000000 0.0000000 0.289427698 0.84403670
students5_estimated  0.9410223 0.04681275 0.904150943 0.6962963 0.278087279 0.83426184
>
```




Matrix plot showing missingness of all Variables
Index
0
200
400
600
iau_id
private01
m_granting
b_granting
year

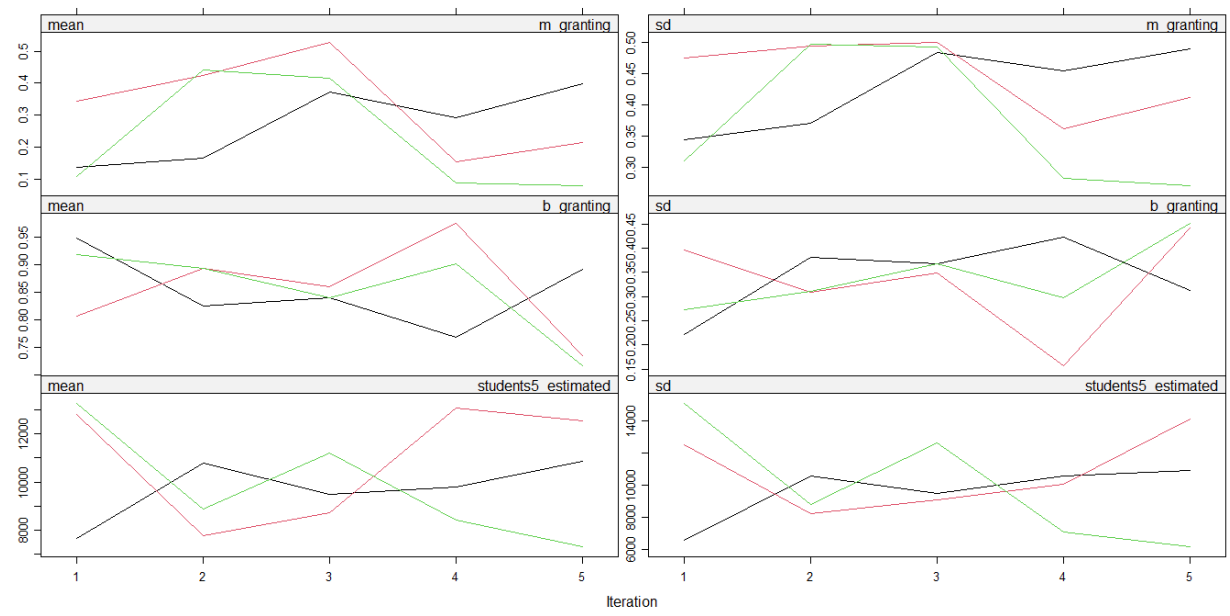

The Aggregation plot of missing data Pattern
0.0
0.2
0.4
0.6
0.8
m_granting
b_granting
iau_id
private01
phd_granting
year

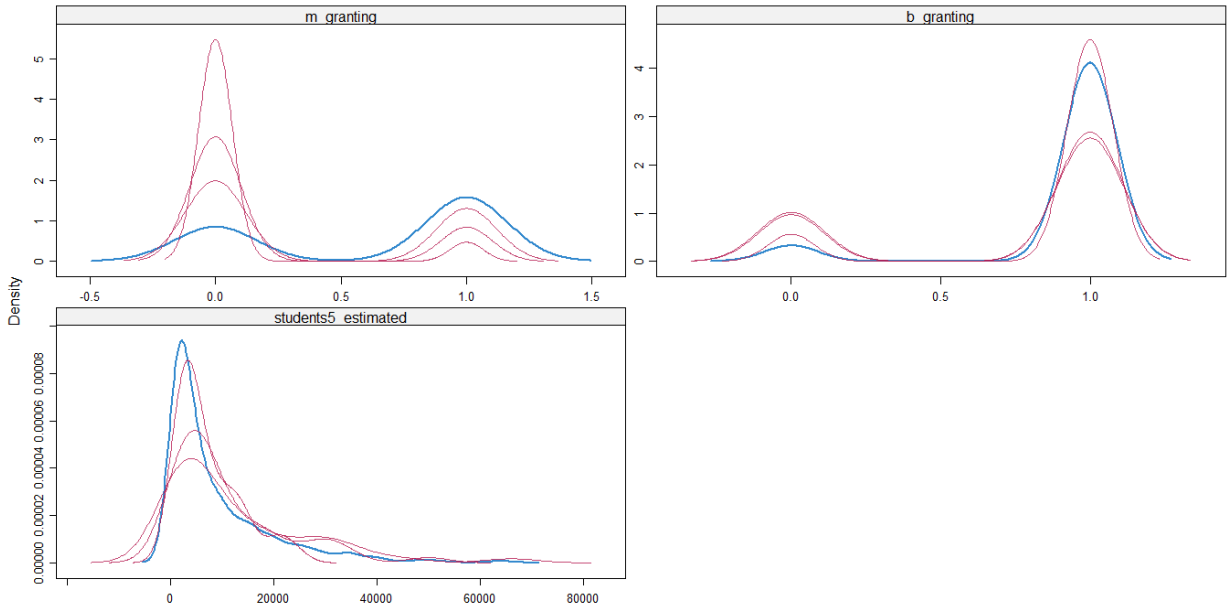



```
> mcar_test(enrollments)
# A tibble: 1 × 4
  statistic    df p.value missing.patterns
      <dbl> <dbl>   <dbl>            <int>
1      448.    15       0                4
> 
```

```
> # MEANS #
> sapply(m_granting_imputed,mean)
        1         2         3
0.3968750 0.2156250 0.0796875
> mean(enrollments$m_granting,na.rm=TRUE)
[1] 0.6504065
> sapply(b_granting_imputed,mean)
       1        2        3
0.890625 0.734375 0.715625
> mean(enrollments$b_granting,na.rm=TRUE)
[1] 0.9268293
> sapply(students5_estimated ,mean)
Error: object 'students5_estimated' not found
> sapply(students5_estimated_imputed ,mean)
       1        2        3
10849.53 12532.27  7310.20
> mean(enrollments$students5_estimated ,na.rm=TRUE)
[1] 8810.37
```

```
> # VARIANCES: IMPUTED SMALLER #
> sapply(m_granting_imputed,var)
         1          2          3
0.23973983 0.16939554 0.07345217
> var(enrollments$m_granting,na.rm=TRUE)
[1] 0.2292416
> sapply(b_granting_imputed,var)
         1          2          3
0.09756455 0.19537363 0.20382433
> var(enrollments$b_granting,na.rm=TRUE)
[1] 0.06837265
> sapply(students5_estimated_imputed,var)
        1         2         3
119660140 198742615  37814217
> var(enrollments$students5_estimated,na.rm=TRUE)
[1] 106364566
>
```





CODE

enrollment = read.csv('Nigeria dataset.csv')

View(enrollment)

(proportionMissing = sum(is.na(enrollment))/prod(dim(enrollment)))

```
# Missingness by Variable
missing_by_variable <- colMeans(is.na(enrollment))
print(missing_by_variable)

enrollments <- enrollment[, c( "iau_id", "private01", "m_granting", "b_granting",
                               "phd_granting", "year", "students5_estimated")]

(proportionMissing = sum(is.na(enrollments))/prod(dim(enrollments)))

# Missingness by Variable
missing_by_variable <- colMeans(is.na(enrollments))
print(missing_by_variable)

data_summary <- sapply(enrollments, function(x) {
  data_length <- length(x)
  num_missing <- sum(is.na(x))
  num_observed <- data_length - num_missing
  percent_complete <- (num_observed / data_length) * 100
  percent_observed <- percent_complete  # In this context, observed = complete
  percent_unobserved <- (num_missing / data_length) * 100

  return(c(Percent_Complete = percent_complete,
           Percent_Observed = percent_observed,
           Percent_Unobserved = percent_unobserved))
})


# Print the summary
print(data_summary)
```

```
aggr_plot <- aggr(enrollments, col=c('red','blue'), numbers=TRUE, sortVars=TRUE,
                  labels=names(enrollments), cex.axis=.7, gap=3, ylab=c("The Aggregation plot of missing data Pattern", ""))

library('ggplot2')

gg_miss_var(enrollments) +
  ggtitle('Dot Chart for the selected variables') + theme(plot.title = element_text(hjust = 0.5))

matrixplot(enrollments) + title('Matrix plot showing missingness of all Variables')

library('VIM')
library('mice')
flux(enrollment)

fluxplot(enrollment)

summary(enrollment)
install.packages("naniar")
library("naniar")
mcar_test(enrollments)
num_imputations <- 3

# Perform multiple imputation
imputed_data <- mice(enrollment, m = num_imputations)
```

```
summary(imputed_data)

## IMPUTED VALUES ##
(imputed_values = imputed_data$imp)

m_granting_imputed = imputed_values$m_granting
b_granting_imputed = imputed_values$b_granting
students5_estimated_imputed = imputed_values$students5_estimated
# MEANS #
sapply(m_granting_imputed,mean)
mean(enrollments$m_granting,na.rm=TRUE)

sapply(b_granting_imputed,mean)
mean(enrollments$b_granting,na.rm=TRUE)

sapply(students5_estimated_imputed ,mean)
mean(enrollments$students5_estimated ,na.rm=TRUE)
# VARIANCES: IMPUTED SMALLER #
sapply(m_granting_imputed,var)
var(enrollments$m_granting,na.rm=TRUE)
sapply(b_granting_imputed,var)
var(enrollments$b_granting,na.rm=TRUE)
sapply(students5_estimated_imputed,var)
var(enrollments$students5_estimated,na.rm=TRUE)
plot(imputed_data)
densityplot(imputed_data)
```