MATH 5900 – ADVANCED DATA ANALYTICS

ASSIGNMENT 5

TAIWO JEGEDE – E00755963

APPLIED DATA SCIENCE

SOFTWARE USED: R

QUESTION 1

I

An appropriate model to be used for the research interest is given by

$$\text{CumulativeGPA}=\beta0+\beta1\times \text{deptclim}+\beta2\times \text{equity}+\beta3\times \text{deptsupp}+\beta4\times \text{discrim}+\beta5\times \text{dptsocl}+\beta6$$

$$\times \text{FrstGener}+\beta7\times \text{race}+\beta8\times \text{children}+\epsilon$$

CummulativeGPA is the dependent variable, representing the cumulative GPA of the graduate student.

deptclim, equity, deptsupp, discrim, and dptsocl are the independent variables representing the general departmental climate, equity, support, discrimination, and social relations, respectively.

FrstGener is a binary independent variable representing first-generation status (1 if the student is first-generation, 0 otherwise).

race is a categorical independent variable representing race/ethnicity.

children is the number of children living with the student, representing the family responsibility of the student.

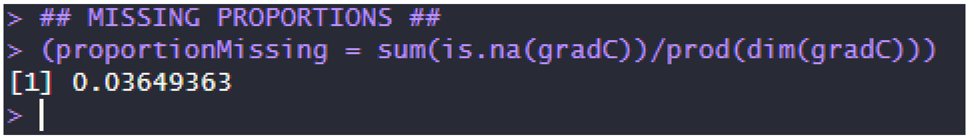
$\beta0,\beta1,\beta2,\ldots,\beta8$ are the coefficients to be estimated.

ϵ represents the error term, capturing the unexplained variance in the dependent variable.

II

Non-visual descriptive statistics

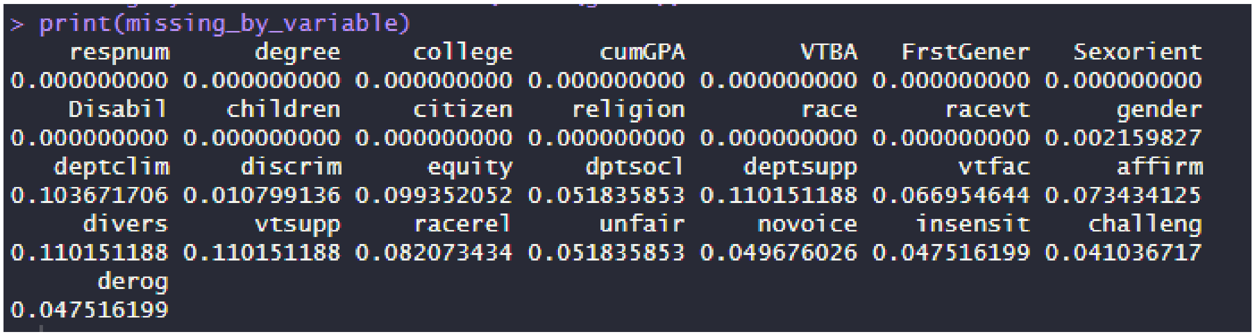
i. **Getting the proportion of the missing data**



```
> ## MISSING PROPORTIONS ##
> (proportionMissing = sum(is.na(gradC))/prod(dim(gradC)))
[1] 0.03649363
> 
```

I used the code above to get the proportion of missing values in the dataset. And from my output, approximately 3.65% of the data is missing. This gives us an overall understanding of how much data is missing relative to the total dataset.

ii. **Missingness by the Variables**



```
> print(missing_by_variable)
    respnum      degree     college      cumGPA        VTBA   FrstGener   Sexorient 
0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 
    Disabil    children     citizen    religion        race      racevt      gender 
0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.002159827 
   deptclim     discrim      equity     dptsocl    deptsupp       vtfac      affirm 
0.103671706 0.010799136 0.099352052 0.051835853 0.110151188 0.066954644 0.073434125 
     divers      vtsupp     racerel      unfair     novoice    insensit    challeng 
0.110151188 0.110151188 0.082073434 0.051835853 0.049676026 0.047516199 0.041036717 
      derog 
0.047516199 
```

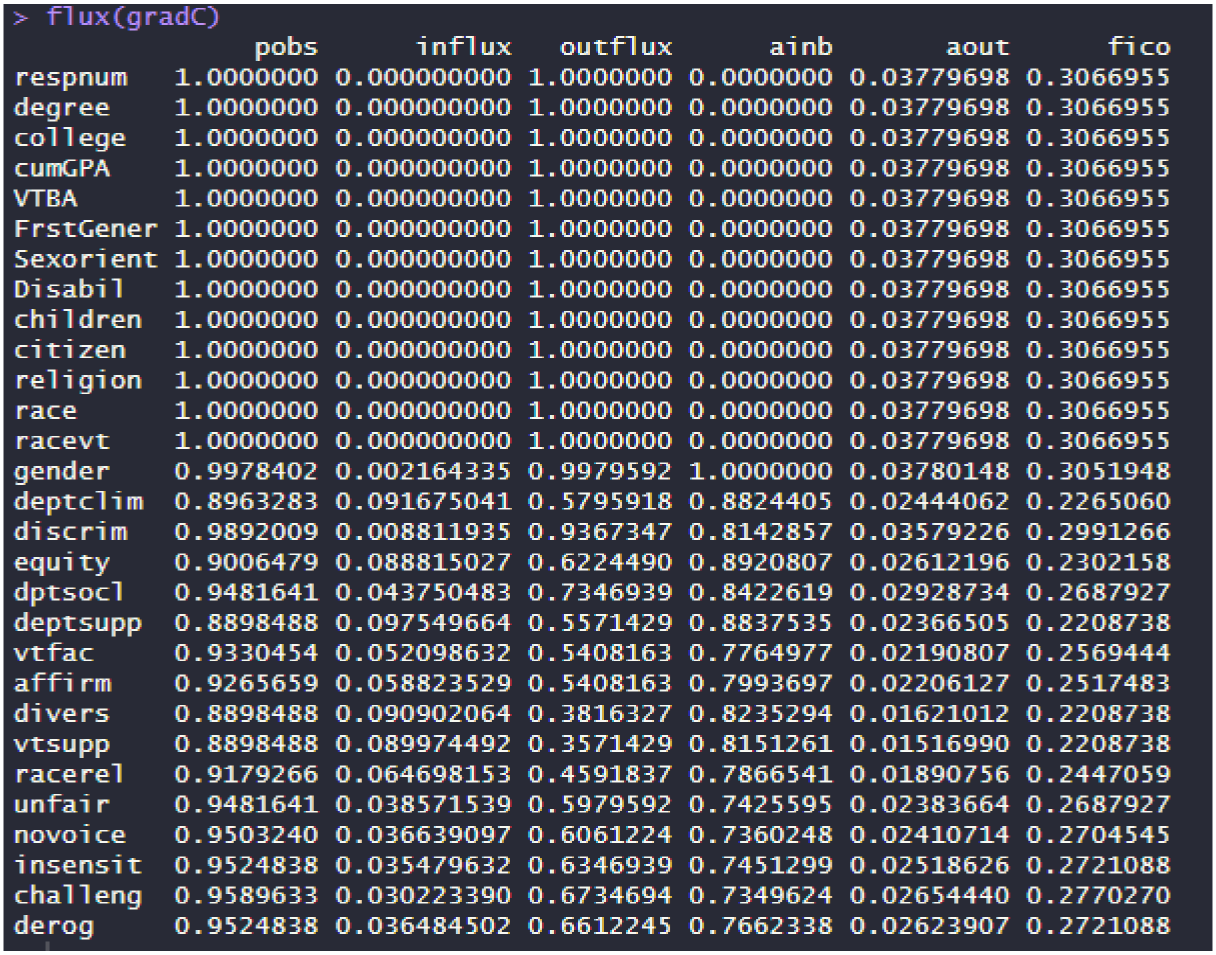
respnum: There are no missing values for this variable. This suggests that all records have a response number associated with them.

degree, college, cumGPA, VTBA, FrstGener, Sexorient, Disabil, children, citizen, religion, race, racevt: Similarly, there are no missing values for these variables. This indicates that data on these demographic and background characteristics are complete for all records.

gender: There are very few missing values (0.22%). This suggests that most records have information on the gender of the respondents, but there are a few cases where this information is missing.

deptclim, discrim, equity, dptsocl, deptsupp, vtfac, affirm, divers, vtsupp, racerel, unfair, novoice, insensit, challeng, derog: These variables have varying proportions of missing values ranging from approximately 4% to 11%. This indicates that a significant portion of the data is missing for these variables. It may be necessary to handle missing data appropriately before conducting further analysis involving these variables.

iii. Influx Coefficient



```
> flux(gradC)
              pobs      influx   outflux      ainb       aout      fico
respnum   1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
degree    1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
college   1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
cumGPA    1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
VTBA      1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
FrstGener 1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
Sexorient 1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
Disabil   1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
children  1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
citizen   1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
religion  1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
race      1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
racevt    1.0000000 0.000000000 1.0000000 0.0000000 0.03779698 0.3066955
gender    0.9978402 0.002164335 0.9979592 1.0000000 0.03780148 0.3051948
deptclim  0.8963283 0.091675041 0.5795918 0.8824405 0.02444062 0.2265060
discrim   0.9892009 0.008811935 0.9367347 0.8142857 0.03579226 0.2991266
equity    0.9006479 0.088815027 0.6224490 0.8920807 0.02612196 0.2302158
dptsocl   0.9481641 0.043750483 0.7346939 0.8422619 0.02928734 0.2687927
deptsupp  0.8898488 0.097549664 0.5571429 0.8837535 0.02366505 0.2208738
vtfac     0.9330454 0.052098632 0.5408163 0.7764977 0.02190807 0.2569444
affirm    0.9265659 0.058823529 0.5408163 0.7993697 0.02206127 0.2517483
divers    0.8898488 0.090902064 0.3816327 0.8235294 0.01621012 0.2208738
vtsupp    0.8898488 0.089974492 0.3571429 0.8151261 0.01516990 0.2208738
racerel   0.9179266 0.064698153 0.4591837 0.7866541 0.01890756 0.2447059
unfair    0.9481641 0.038571539 0.5979592 0.7425595 0.02383664 0.2687927
novoice   0.9503240 0.036639097 0.6061224 0.7360248 0.02410714 0.2704545
insensit  0.9524838 0.035479632 0.6346939 0.7451299 0.02518626 0.2721088
challeng  0.9589633 0.030223390 0.6734694 0.7349624 0.02654440 0.2770270
derog     0.9524838 0.036484502 0.6612245 0.7662338 0.02623907 0.2721088
```

For variables where the influx is close to 0, such as 'respnum', 'degree', 'college', 'cumGPA', 'VTBA', 'FrstGener', and 'Sexorient', there is little to no change in missingness between consecutive variables. This suggests that missingness for these variables is relatively stable throughout the dataset.

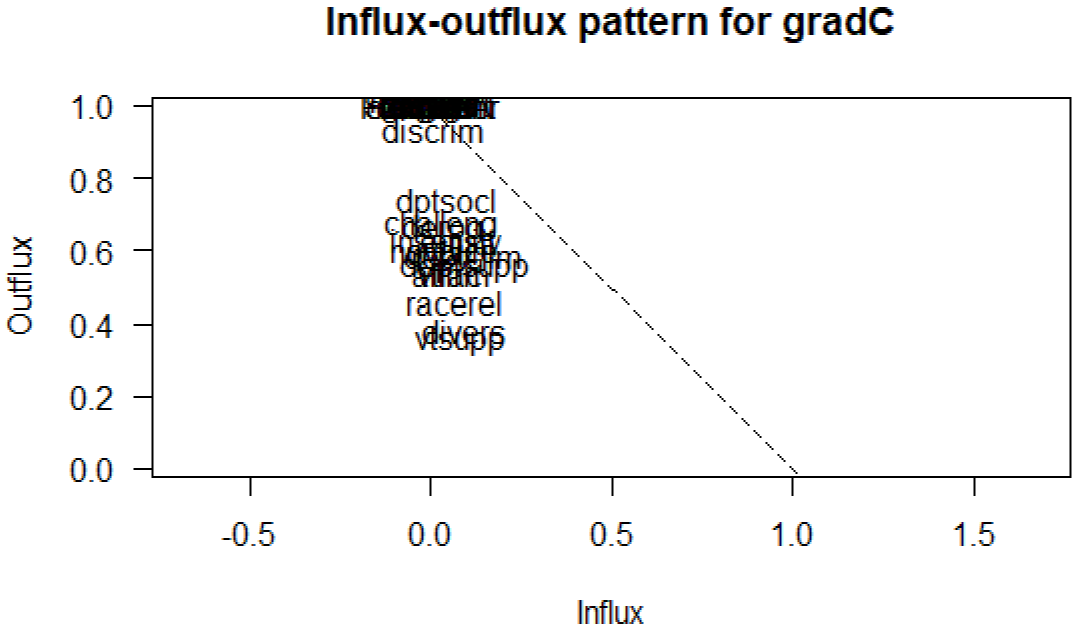
For 'gender', there is a small positive influx (0.22%), indicating a slight increase in missingness compared to the previous variable. However, this increase is minimal.

For variables like 'deptclim', 'discrim', 'equity', 'dptsocl', 'deptsupp', 'vtfac', 'affirm', 'divers', 'vtsupp', 'racerel', 'unfair', 'novoice', 'insensit', 'challeng', and 'derog', there are larger influx values, indicating more substantial changes in missingness between consecutive variables. This means that we can obtain one value for each variable, describing the amount of information about the variable available from other variables.

III

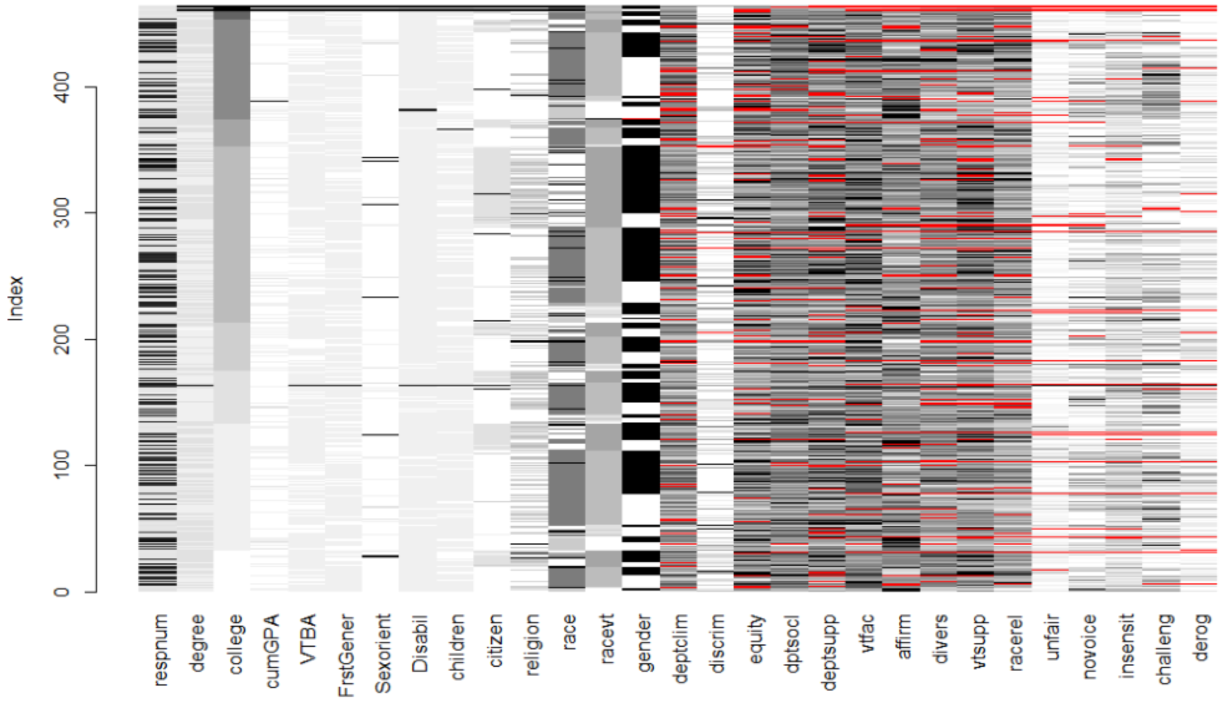
Visual descriptive statistics

i. **Influx/Outflux**



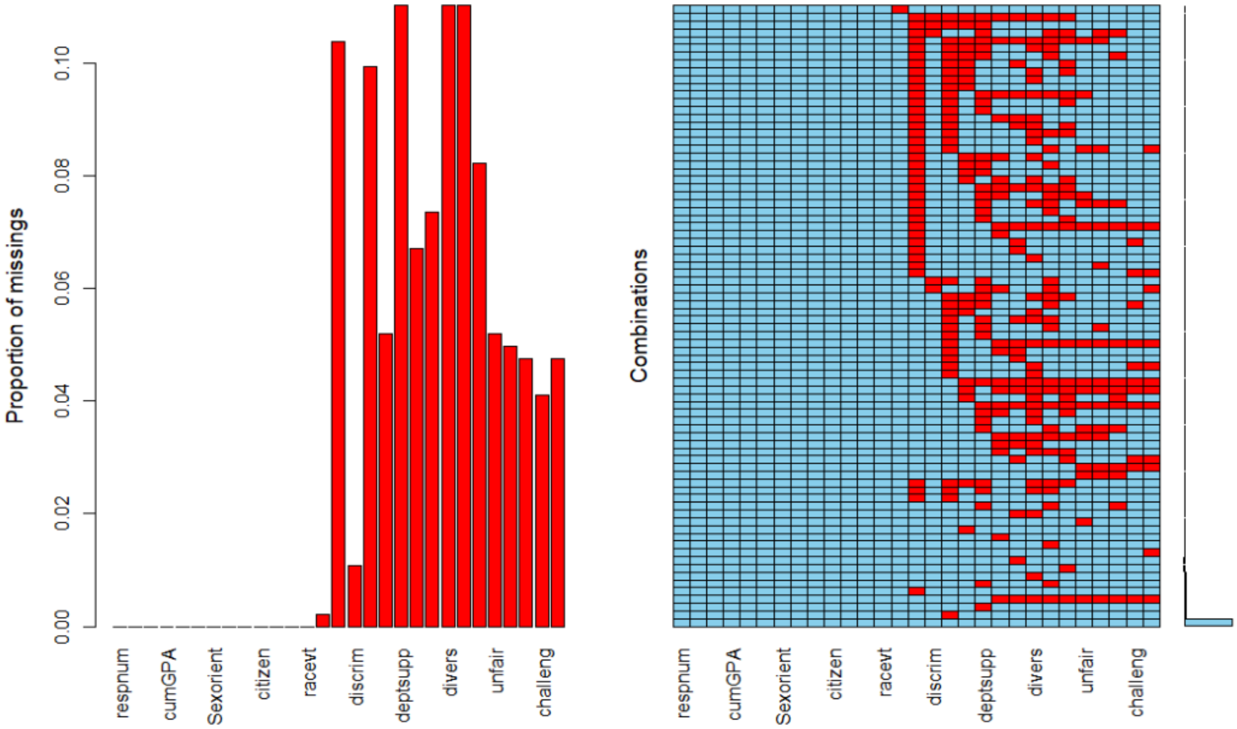
From the flux plot above, we can see that there is no variable with an influx of 1, however there are those that are close to one, which means that they will need information from other variables. The opposite applies to Outflux, we have variables that are 1 or close to one, these variables will give out information to other variables.

ii. Matrix Plot



The above visualization is a data matrix plot, and from our visualization, we can see that some of the variables like deptclim, discrim, equity etc are somewhat color coded and this means that they have missing values.

iii. Aggregation Plot



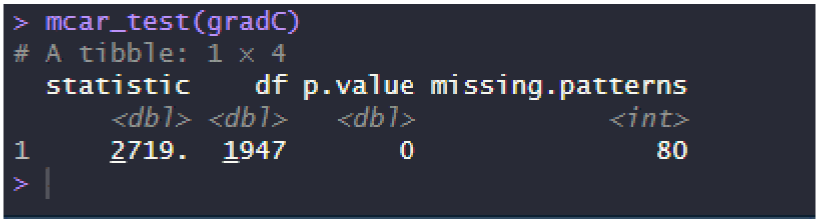
The aggregation plot shows us two colors, and from these colors we can figure out which one of our variables have missing values. So when they are blue it means they are not missing, and if they are red it means that they are missing.

IV

One test we could use to access the nature of the missingness:

```
install.packages("naniar")
library("naniar")
mcar_test(gradC)
```

Output:



```
> mcar_test(gradC)
# A tibble: 1 × 4
  statistic    df p.value missing.patterns
      <dbl> <dbl>   <dbl>            <int>
1     2719.  1947       0               80
> 
```

Interpretation:

The extremely small p-value (close to zero) indicates strong evidence against the null hypothesis of the data being missing completely at random (MCAR). Therefore, we reject the null hypothesis in favor of the alternative hypothesis, suggesting that the missingness in the dataset is not completely random.

V

The data has a small amount of missingness (3.65%) but it's not random. When we look at the missingness by variables, it is not actually consistent across all variables, some have none, and some have up to 10% missing data. And from what we got from our MCAR Test, the test confirms that the missingness is not random. This means that there is a high chance the data is missing at random or missing not at random.

QUESTION 2

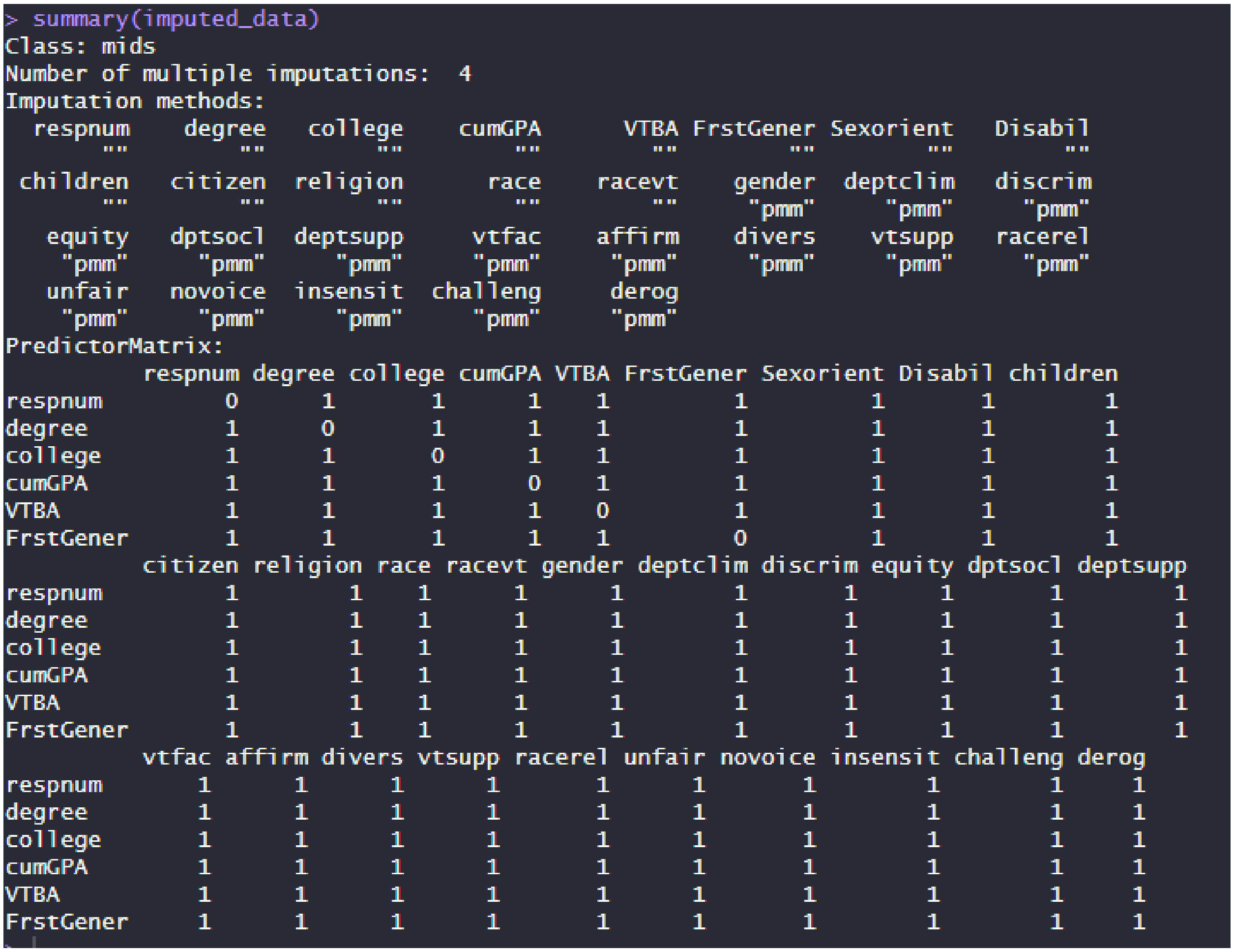
I

Given the nature of the missingness in the dataset, particularly the evidence against missing data being completely at random (MCAR), a suitable method for analyzing the data while properly accounting for the missing values would be multiple imputation.

II

Before I did the multiple imputation, I first had to make sure we only did the multiple imputation for variables that had missing values. And also ensured that we did not have columns that were missing completely.

Output:



```
> summary(imputed_data)
Class: mids
Number of multiple imputations:  4
Imputation methods:
  respnum    degree   college    cumGPA      VTBA FrstGener Sexorient   Disabil
       ""        ""        ""        ""        ""        ""        ""        ""
 children   citizen  religion      race    racevt    gender  deptclim   discrim
       ""        ""        ""        ""        ""     "pmm"     "pmm"     "pmm"
   equity   dptsocl  deptsupp     vtfac    affirm    divers    vtsupp   racerel
    "pmm"     "pmm"     "pmm"     "pmm"     "pmm"     "pmm"     "pmm"     "pmm"
   unfair   novoice  insensit  challeng     derog
    "pmm"     "pmm"     "pmm"     "pmm"     "pmm"
PredictorMatrix:
          respnum degree college cumGPA VTBA FrstGener Sexorient Disabil children
respnum         0      1       1      1    1         1         1       1        1
degree          1      0       1      1    1         1         1       1        1
college         1      1       0      1    1         1         1       1        1
cumGPA          1      1       1      0    1         1         1       1        1
VTBA            1      1       1      1    0         1         1       1        1
FrstGener       1      1       1      1    1         0         1       1        1
          citizen religion race racevt gender deptclim discrim equity dptsocl deptsupp
respnum         1        1    1      1      1        1       1      1       1        1
degree          1        1    1      1      1        1       1      1       1        1
college         1        1    1      1      1        1       1      1       1        1
cumGPA          1        1    1      1      1        1       1      1       1        1
VTBA            1        1    1      1      1        1       1      1       1        1
FrstGener       1        1    1      1      1        1       1      1       1        1
          vtfac affirm divers vtsupp racerel unfair novoice insensit challeng derog
respnum       1      1      1      1       1      1       1        1        1     1
degree        1      1      1      1       1      1       1        1        1     1
college       1      1      1      1       1      1       1        1        1     1
cumGPA        1      1      1      1       1      1       1        1        1     1
VTBA          1      1      1      1       1      1       1        1        1     1
FrstGener     1      1      1      1       1      1       1        1        1     1
```

From our output, we can see “pmm”, which means that the type of imputation was actually based on a linear regression model. Also, we notice there are some columns with “ “, this means that no imputation was done on these columns because they didn’t have any missing value. The next thing we will be doing is running diagnosis to see if the approach we used “multiple imputation” worked.

III

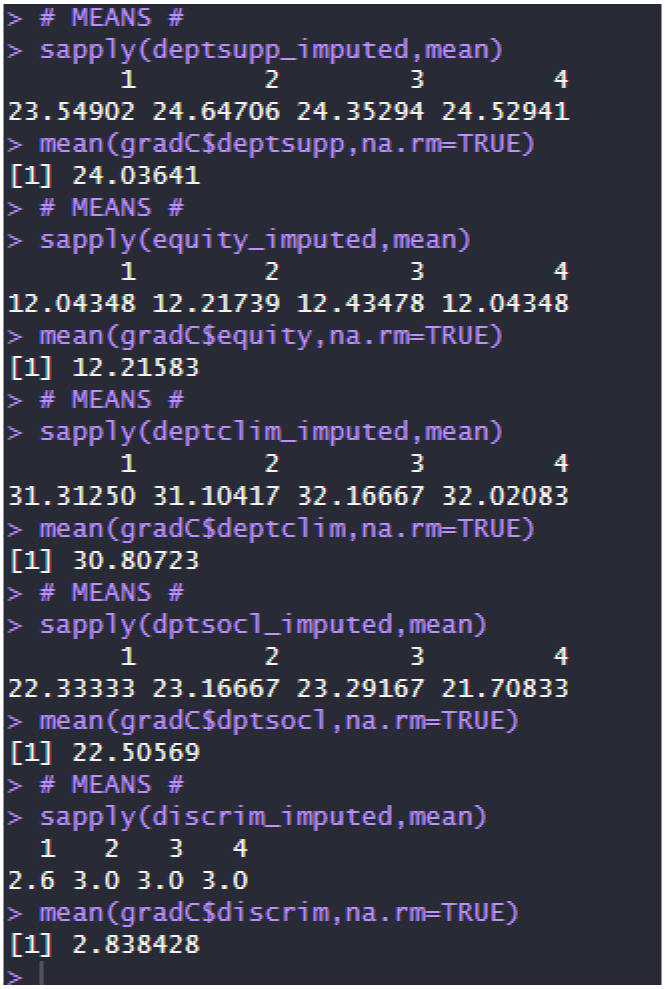
Multiple Imputation Diagnostics:

We compare the imputed data and the observed data, using

```
## IMPUTED VALUES ##

(imputed_values = imputed_data$imp)

deptsupp_imputed = imputed_values$deptsupp

equity_imputed = imputed_values$equity

deptclim_imputed = imputed_values$deptclim

dptsocl_imputed = imputed_values$dptsocl

discrim_imputed = imputed_values$discrim
```

After which we then check for the equality of the mean and of the variance. Also we use plots like trace plots and density plots amongst others.

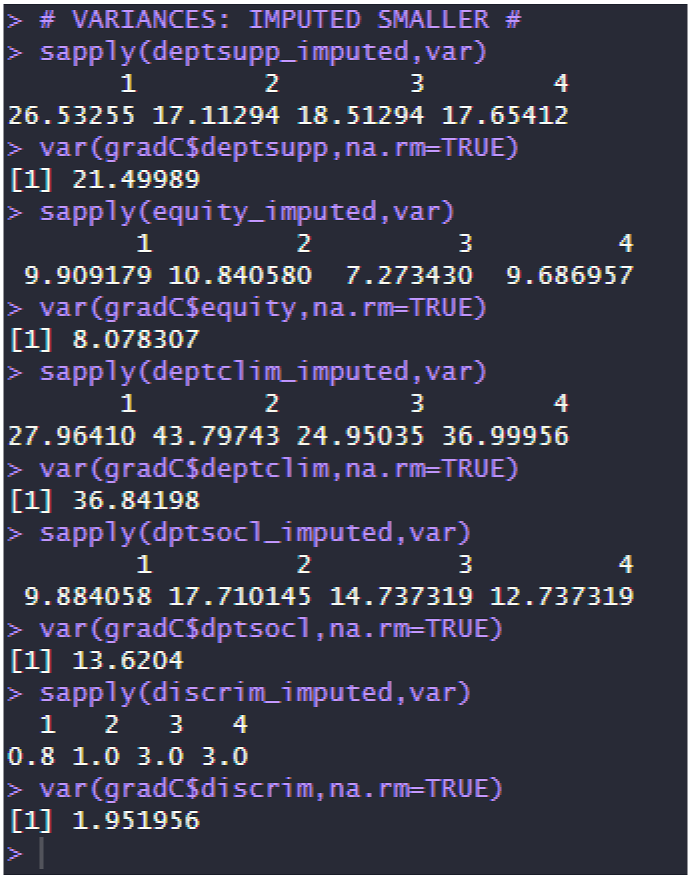
MEAN



```
> # MEANS #
> sapply(deptsupp_imputed,mean)
       1        2        3        4 
23.54902 24.64706 24.35294 24.52941 
> mean(gradC$deptsupp,na.rm=TRUE)
[1] 24.03641
> # MEANS #
> sapply(equity_imputed,mean)
       1        2        3        4 
12.04348 12.21739 12.43478 12.04348 
> mean(gradC$equity,na.rm=TRUE)
[1] 12.21583
> # MEANS #
> sapply(deptclim_imputed,mean)
       1        2        3        4 
31.31250 31.10417 32.16667 32.02083 
> mean(gradC$deptclim,na.rm=TRUE)
[1] 30.80723
> # MEANS #
> sapply(dptsocl_imputed,mean)
       1        2        3        4 
22.33333 23.16667 23.29167 21.70833 
> mean(gradC$dptsocl,na.rm=TRUE)
[1] 22.50569
> # MEANS #
> sapply(discrim_imputed,mean)
  1   2   3   4 
2.6 3.0 3.0 3.0 
> mean(gradC$discrim,na.rm=TRUE)
[1] 2.838428
> 
```

From the above output of the diagnostics using the mean, when we observe the result, we can see some similarities between the means, there is no outrageous values present, they all happen to be within a plausible range. Now we check for the Variance.

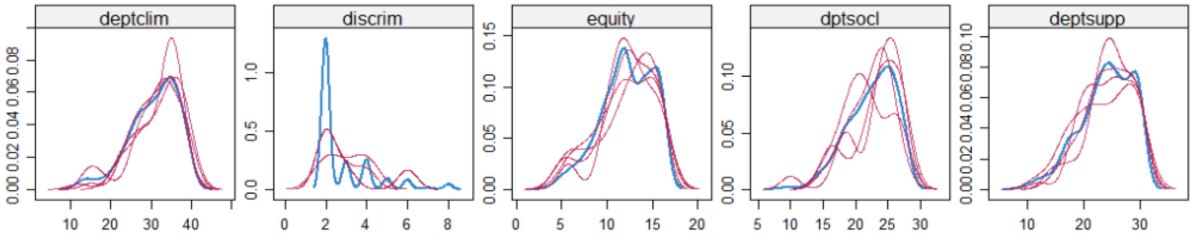
VARIANCE



```
> # VARIANCES: IMPUTED SMALLER #
> sapply(deptsupp_imputed,var)
       1        2        3        4
26.53255 17.11294 18.51294 17.65412
> var(gradC$deptsupp,na.rm=TRUE)
[1] 21.49989
> sapply(equity_imputed,var)
        1         2         3         4
 9.909179 10.840580  7.273430  9.686957
> var(gradC$equity,na.rm=TRUE)
[1] 8.078307
> sapply(deptclim_imputed,var)
       1        2        3        4
27.96410 43.79743 24.95035 36.99956
> var(gradC$deptclim,na.rm=TRUE)
[1] 36.84198
> sapply(dptsocl_imputed,var)
        1         2         3         4
 9.884058 17.710145 14.737319 12.737319
> var(gradC$dptsocl,na.rm=TRUE)
[1] 13.6204
> sapply(discrim_imputed,var)
  1   2   3   4
0.8 1.0 3.0 3.0
> var(gradC$discrim,na.rm=TRUE)
[1] 1.951956
>
```

From the above variance diagnostics that i did, the diagnostic indicates that for most variables, the imputed variances tend to be smaller than the variances of the observed data. However, for the 'discrim' variable, the results are mixed. This comparison suggests that the imputation method may affect the variability of the data, potentially influencing subsequent analyses and interpretations.

USING PLOTS



From the above density plot, we have two colors, blue and red. The red lines are the different imputations that we did, we did four imputations. The blue line represents the density plot (Value of Imputation), when we take a look at deptclim, equity and deptsupp for example, we can see that there is a closeness between the blue and red line, and this just tells us that the imputation is good, unlike discrim where there is a bit of a divergence between the blue and red lines.

QUESTION 3

I

Missing data pattern is basically just the art of visualizing missing values that exists in a dataset. So this pattern could have a systematic arrangement or distribution of missing values within a dataset. It describes how missing values are distributed across variables or observations, indicating the structure of missingness in the data.

Why it is important to investigate missing data patterns

It is very important that we investigate the missing data patterns because it helps us to understand the missing data mechanism, like if it is Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). Also, when we understand the pattern, we can now choose which kind of imputation would be the best. It also helps us in identifying potential data quality issues, detecting hidden relationships between variables, and improving the reliability of statistical analyses.

II

Missing Completely At Random (MCAR):

In MCAR, the probability of missingness is unrelated to both observed and unobserved data. This means that missing values occur randomly and independently of any variables in the dataset. In other words, missingness is completely unpredictable and occurs by chance.

Example: A graduate students who did not take a quiz. Their score will be missing, but it won't be because of what their final quiz score is or what they should have gotten.

Missing At Random (MAR):

Here, the missingness is independent of unobserved values, but it can actually depend on observed values. So basically, the missingness is related to the observed data but not to the unobserved data.

Example: People could fill out a form, maybe a survey and decide to leave certain questions unanswered, maybe income or gender information. While income is missing, the likelihood of missingness depends on the observed variable (education level), making it MAR.

Missing Not At Random (MNAR):

Missingness here can depend on the unobserved values, so the probability of missingness depends on the missing values themselves, even after considering observed variables. This means that the missingness is related to unobserved data or variables not included in the analysis.

Example: In a timed quiz, there were some questions that were incomplete, in a case like this there will be no score for the incomplete questions, so it is not missing at random.

III

Influx and outflux are terms used when we want to describe the flow of missing data within a dataset. They provide information about how missing values enter and exit the dataset.

Influx:

Influx refers to the proportion of cases where missing values enters the dataset.

Higher values (close to 1) means that there is missing information, which is contained in other variables. So this mean that it uses information from other variables.

Outflux:

Outflux refers to the proportion of cases where missing values exit the dataset.

Higher values, close to 1 indicates that it is going to provide information for other variables in the dataset.

When we incorporate information about influx and outflux into subsequent analyses, it allows us to better understand and account for missingness in the data, which leads to more accurate and reliable analysis, when we understand the influx and outflux of missing data, we can pick the appropraite imputation method to use. For example, if there is a very high influx in our dataset (values of 1), our multiple imputation will not work, we will have to remove those columns first.

IV

Traditionally, whenever we want to replace missing values, we often use the single estimate (like mean/median) and this creates a false sense of certainty and ignores the inherent variability in the missing data. Which is why we use MICE, bec works in a loop. It builds a separate regression model for each variable with missing values. Each model predicts missing values for its assigned variable based on the existing data points in all other variables (including previously imputed values). These imputed values are then used as predictors in the next round of modeling for another variable. This creates a "chain" of dependencies, where each variable benefits from the most recent estimates.

Multiple Imputations: The process iterates for a set number of times, creating multiple completed datasets. Each dataset reflects a plausible scenario for the missing values.

ADVANTAGE

The key advantage of the Chained Equations method is that it avoids the issue of "replacing" a missing random variable with a single value by incorporating uncertainty into the imputation process. Instead of imputing a single value for each missing value, Chained Equations imputes multiple plausible values based on the observed values of other variables. By iteratively updating imputed values based on regression models, the method captures the uncertainty associated with missing data and produces imputed datasets that reflect the variability in the missing values. This allows for more accurate parameter estimates and standard errors in subsequent analyses, as the uncertainty introduced by the missing data is appropriately accounted for.