MATH 5900: ADVANCED DATA ANALYSIS

Project 2: Missing Data Analysis

Noah Borquaye

May 2023

1 Purpose of Analysis

The target or the response variable for this data set (Calc2), which is the main criteria to determine if a student will enrol in Calculus 2 or not in the following semester. Our client(a university) wants us to develop a model for future predictions. Analysis of the data revealed that the data contains missing values and some of the entries contain inappropriate characters Thus, prior to building our predictive model, we aim to account for the missingness in the data by performing imputation.

2 Description of the data

The CalculusGrades data set has 900 rows and 9 columns (variables). Five of the variables (GPA, Exam1, Exam2, Exam3, Exam) are numeric while four (Prev, Major, TOD, Calc2) are categorical. Here are the attributes for the Data:

1. **Prev**: True(1) if already taken a Calculus 1 Course.
2. **Major**: One of the following:
   - **STEM** = Science, Technology Engineering, or Mathematics.
   - **PMED** = Pre med.
   - **BUSN** = Business, Accounting, Finance, or Marking.
   - **ARHU** = Arts or the Humanities.
   - **OTHR** = None of the majors listed above.
3. **TOD**: Either Morning, Evening or Afternoon.
4. **GPA**: GPA on a 4 point scale 4.0.
5. **Exam1**: Score on First Exam in current Calculus Course.
6. **Exam2**: Score on Second Exam in current Calculus Course.
7. **Exam3**: Score on Third Exam in current Calculus Course.
8. **Exam4**: Score on Fourth Exam in current Calculus Course.

9. **Calc2** : This is the response or outcome. Calc2 is True (1) if student enrolled in Calculus 2 the next semester or False (0) if they did not.

The data was obtained from an open online dataset (Kaggle).

3 Proposed Analysis

We start our analysis by first performing exploratory analysis of our data. This process enables us to 'know' our data well. Here we check if there are missing values and which variables contain the missing values; we also check for the distribution of the data by plotting barplot. in plotting barplot, we first convert our data into a binary matrix.

Again,. We also perform non-visual and visual exploration of the data set to enable us have a fair idea about the nature of missingness. To determine the nature of missingness, we employ the Little's MCAR test.

Furthermore, we perform imputation to account for the missingness in the data using MICE (Multivariate Imputation by Chained Equations). We will perform relevant diagnistics such a plotting of imputed values to evaluate our method of analysis.

Finally, we assess the performance of our imputation by comparing simple descritives such as mean and variance between original and imputed data.

4 Analysis of Results

4.1 Exploratory Analysis

When we carefully observed our data some of the entries in the variable, Major contained inappropriate entries. We identified these and treated them as missing (i.e we replaced them with NA). We also observed that only three of the variables has missing data; Major has 87 missing values, Exam3 has 166 missing values and Exam4 has 418 missing values. In performing non-visual exploration of the data, we look at the proportion of the data that is

complete, observed and missing. From Table 1 (see appendix), we have 43.70% of complete data,

92.54% observed data and only 7.46% missing data. Again, from Table 2, flux plot shows that the variables, TOD, GPA, Exam1, Exam2 and Calc2 each has zero influx value and a perfect outflux value (outflux = 1.0). This indicates that these variables can provide information about other variables and no information about these variables can be provided by other variables. We also observe from table 2 that Exam4 recorded the highest influx value (0.3894) and the lowest out- flux value (0.2265). This means that Exam4 can provide least information about other variables but more information about it can be provided by other variables. This is evident that Exam4 contains the highest missing values.

Moreover, by exploring our data visually, we considered aggregation plot, which visually shows each pattern of missingness, along with frequency. and data matrix plot, which also shows which variables have missing data. From Figure 2, aggregate plot shows that the variables, Major, Exam3 and Exam4 has missing data, since the columns of these variables have red solid color. We observe that the level of missingness is relatively low (between the range of 0 to 5%). From Figure 3, the data matrix plot yielded the same results as the aggregate plot.

4.2 Little's Test

Now, we assess the nature of the missingness in the data using the Little's MCAR test. The test uses a chi-square statistic to assess the fit of the observed data to the expected data under the MCAR assumption.

H_0: data is missing completely at random (MCAR).

H_A: data is not missing completely at random (MCAR).

Since the $p - value = 0.2125774 > \alpha = 0.05$, we fail to reject H_0 and conclude that the data is missing completely at random (MCAR).

4.3 Imputation of Data

Missing data was imputed by performing five iterations, with the variable Major, set as a factor. We observed that imputation took effect in the variables Major numeric, Exam3 and Exam4, as shown in Table 3.

From Figure 4, imputation after 5 iterations, the plot shows no clear trend, thet is we observe a mixed up pattern which indicates that imputation was successful. Figure 5, a histogram of inputed values of the variable Major shows that imputation at each iteration worked pretty well as the bars at each level of iteration are almost the same. On the contrary, figure 7 shows that imputation in Exam4 worked fairly.

4.4 Assessing the Performance of our Imputation Method

By assessing the the performance of our imputation method, we compare simple descriptives between the original and Imputed data. From table 4, we observe that the mean for the Exam3 in original data is 77.33 and that of imputation ranges between 77.80 to 79.66. This implies the means of the imputed values are almost the same for all the five iterations and are very close to the means of the original data. Also, the mean for the Exam4 in original data is 74.43 and that of imputation ranges between 73.38 to 73.97, which indicates that imputation good for Exam4.

Similarly, from table 5, we observe that the variance for the Exam3 in original data is 169.49 and that of imputation ranges between 165.35 to 146.60. So variance at each level of iteration

is smaller than that of the original data, which is indicates that our imputation worked very well for Exam3. The variance for Exam4 in original data is 255.85 and its variance after first imputation was found to be 256.59 and reduced at gradually to 262.93 at the third iteration and then decreased again to 250.31 at the fifth iteration. These comparisons tell us that generally, our imputation worked.

5 Conclusion

Analysis of the result shows that the data is missing completely at random as we failed to reject the null hypothesis by the Little's test. Thus the missingness in the data is independent of both observed and unobserved responses . Plot of histogram of imputed values revealed that our imputation worked very well, especially for the variable, Major. When we compare the means and variances of the original and the imputed data, we notice that the means are close and the variance of the imputed data are generally lower as compared to that of the original data and we conclude that our method of imputation worked.

6 Appendix

List of Tables

Table 1: Non-Visual Descriptive Statistics.

Descriptive Statistics for Missing Data

No. of cases	1000	
No. of complete cases	437	(43.70%)
No. of incomplete cases	563	(56.30%)
No. of values	9000	
No. of observed values	8329	(92.54%)
No. of missing values	671	(7.46%)
No. of variables	9	
No. of missing values across all variables		
Mean	74.56	(7.46%)
SD	141.34	(14.13%)
Minimum	0.00	(0.00%)
P25	0.00	(0.00%)
P75	87.00	(8.70%)
Maximum	418.00	(41.80%)

Table 2: Flux plot.

	pobs <dbl>	influx <dbl>	outflux <dbl>	ainb <dbl>	aout <dbl>	fico <dbl>
Prev	1.000	0.00000000	1.0000000	0.0000000	0.08387500	0.5630000
Major	0.913	0.07864089	0.8092399	0.9410920	0.07434283	0.5213582
TOD	1.000	0.00000000	1.0000000	0.0000000	0.08387500	0.5630000
GPA	1.000	0.00000000	1.0000000	0.0000000	0.08387500	0.5630000
Exam1	1.000	0.00000000	1.0000000	0.0000000	0.08387500	0.5630000
Exam2	1.000	0.00000000	1.0000000	0.0000000	0.08387500	0.5630000
Exam3	0.834	0.14935767	0.6274218	0.9367470	0.06309952	0.4760192
Exam4	0.582	0.38936247	0.2265276	0.9697967	0.03264605	0.2491409
Calc2	1.000	0.00000000	1.0000000	0.0000000	0.08387500	0.5630000

Table 3: Imputation of missing values.

iter	imp	variable		
1	1	Major_numeric	Exam3	Exam4
1	2	Major_numeric	Exam3	Exam4
1	3	Major_numeric	Exam3	Exam4
1	4	Major_numeric	Exam3	Exam4
1	5	Major_numeric	Exam3	Exam4
2	1	Major_numeric	Exam3	Exam4
2	2	Major_numeric	Exam3	Exam4
2	3	Major_numeric	Exam3	Exam4
2	4	Major_numeric	Exam3	Exam4
2	5	Major_numeric	Exam3	Exam4
3	1	Major_numeric	Exam3	Exam4
3	2	Major_numeric	Exam3	Exam4
3	3	Major_numeric	Exam3	Exam4
3	4	Major_numeric	Exam3	Exam4
3	5	Major_numeric	Exam3	Exam4
4	1	Major_numeric	Exam3	Exam4
4	2	Major_numeric	Exam3	Exam4
4	3	Major_numeric	Exam3	Exam4
4	4	Major_numeric	Exam3	Exam4
4	5	Major_numeric	Exam3	Exam4
5	1	Major_numeric	Exam3	Exam4
5	2	Major_numeric	Exam3	Exam4
5	3	Major_numeric	Exam3	Exam4
5	4	Major_numeric	Exam3	Exam4
5	5	Major_numeric	Exam3	Exam4

Table 4: Means of original vs imputed data.

```
       1        2        3        4        5
77.80120 76.53614 75.97590 77.36747 79.61446
[1] 77.33573
       1        2        3        4        5
73.38278 72.70335 74.11005 74.59330 73.97368
[1] 74.43643
```

Table 5: Variance of original vs imputed dat

```
       1        2        3        4        5
165.3481 193.4744 169.0540 136.6581 146.6020
[1] 169.4934
       1        2        3        4        5
256.5917 251.6768 244.8128 262.9325 250.3135
[1] 255.854
```

List of Figures

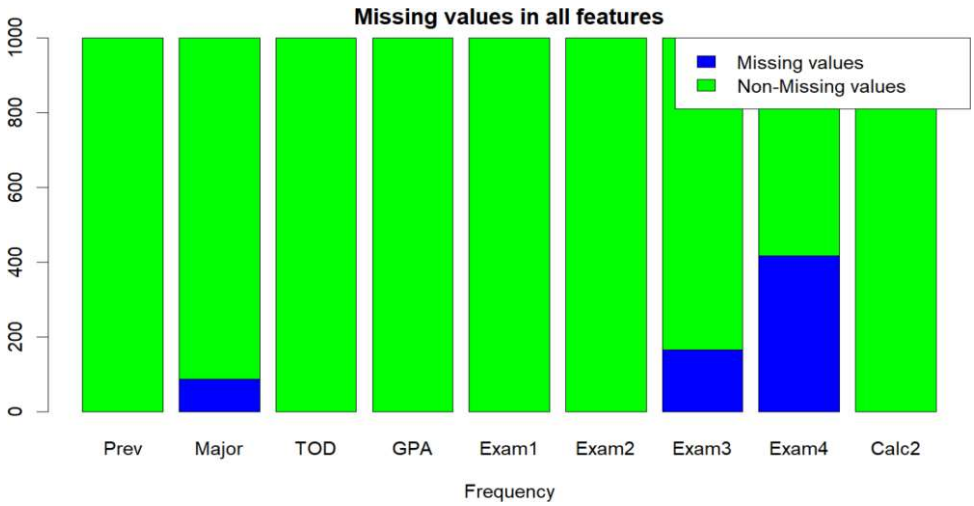

Figure 1: Barplot.

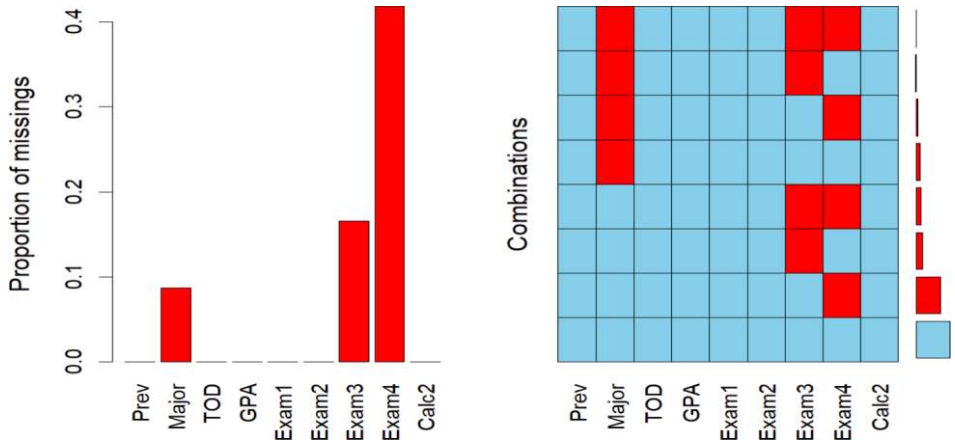

Figure 2: Aggregate p

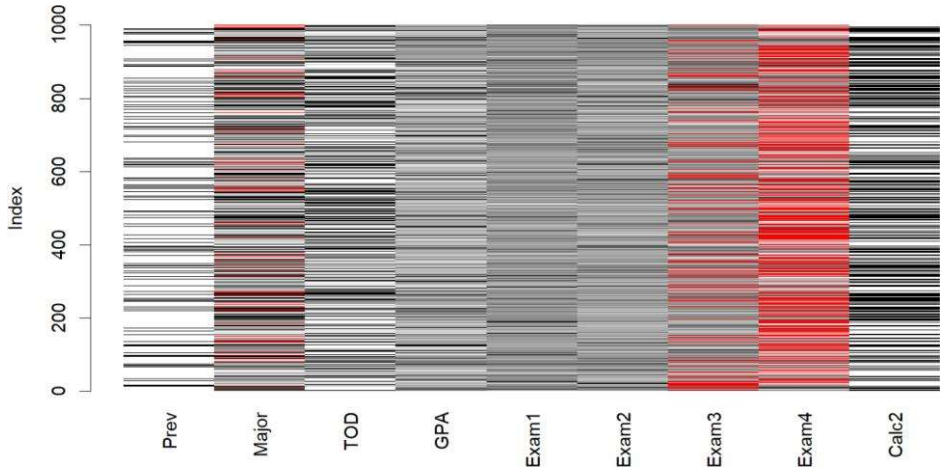

Figure 3: Data Matrix plot

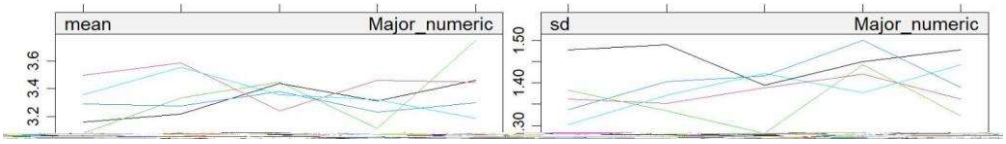

Figure 4: Barplot.

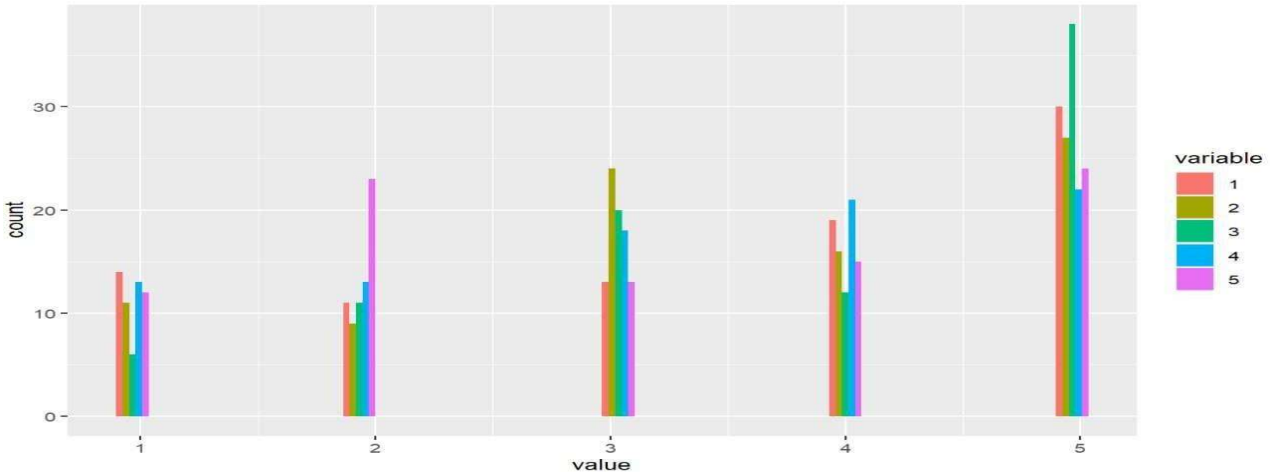

Figure 5: Histogram of Major numeric Imputed.

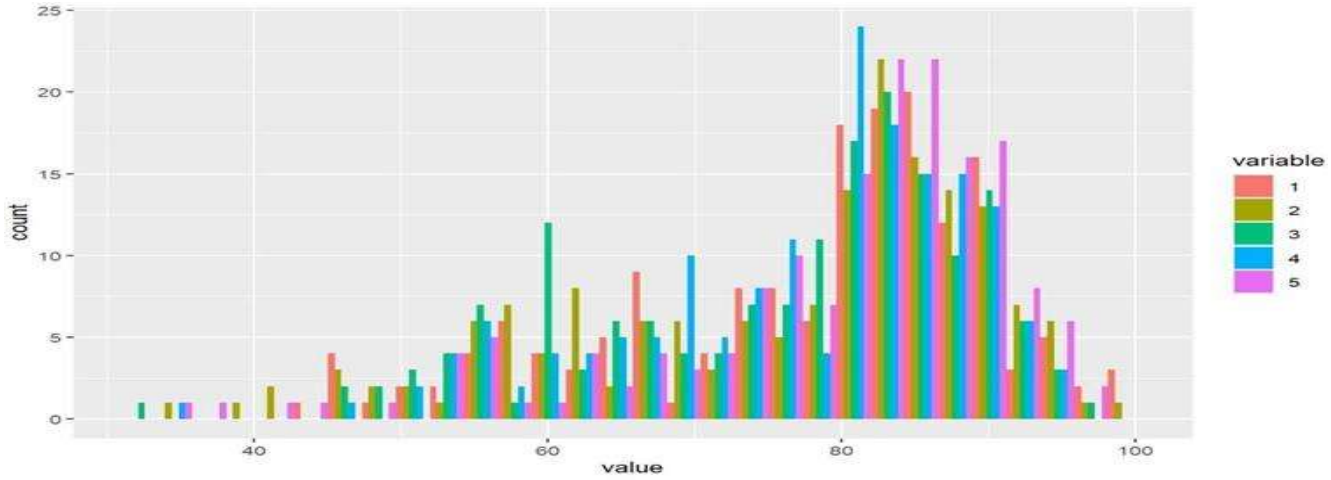

Figure 6: Histogram of Exam3 Imputed.

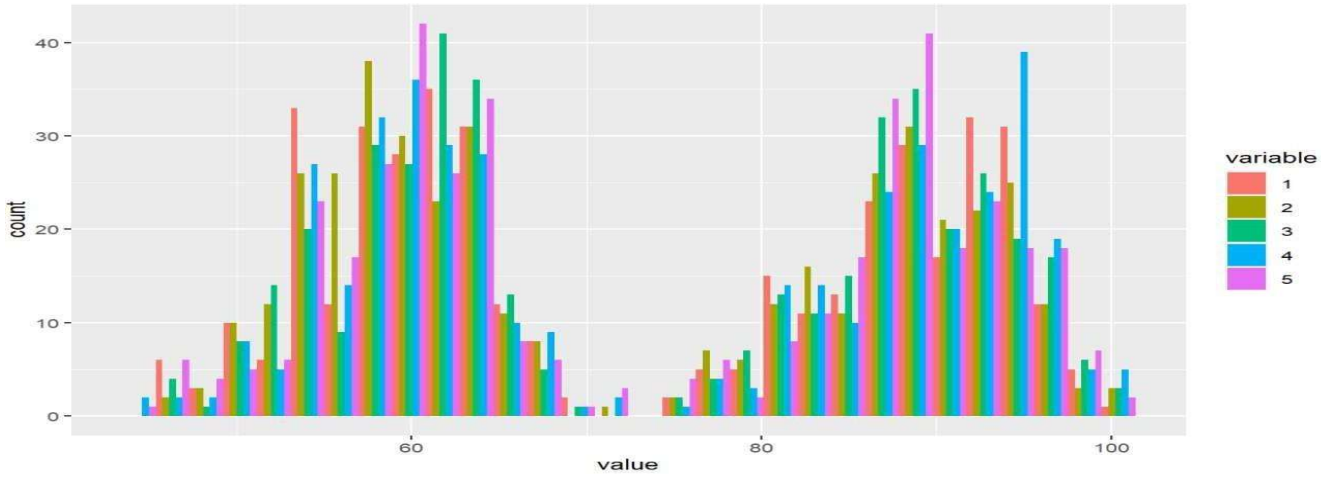

Figure 7: Histogram of Exam4 Imputed.

Relevant codes

```
% Loading Data
CalcGradesData = read.csv('CalculusGrades.csv',header=TRUE)
% Replacing characters with 'NA'
CalcGradesData ¡- CalcGradesData CalcGradesData
%function convert dataframe to binary TRUE/FALSE matrix
toBinaryMatrix ¡- function(CalcGradesData ) m¡-c()
for(i in colnames(CalcGradesData )) x¡-sum(is.na(CalcGradesData [,i])) missing value countm¡-append(m function
call
binMat = toBinaryMatrix(CalcGradesData )
barplot(binMat, main = "Missing values in all features",xlab = "Frequency", col = c("blue","green"))
legend("topright", c("Missing values","Non-Missing values"),
fill = c("blue","green"))
% Non-Visual exploration
na.descript(CalcGradesData)
% Fluxplot
flux(CalcGradesData)
%Visual exploration
aggr(CalcGradesData)
matrixplot(CalcGradesData)
% Convert categorical variable with levels into numeric
CalcGradesDataMajor_numeric = as.numeric(factor(CalcGradesDataMajor )) CalcGradesData$TOD_nu
as.numeric(factor(CalcGradesData$TOD))
print(CalcGradesData)
%Littles's Test
CalcmodelData = as.data.frame(cbind(CalcGradesData$Prev, CalcGradesData$Major numeric, Calc
colnames(CalcmodelData) = c("Prev", "Major_numeric", "TOD_numeric", "GPA", "Exam1", "Exam2,
"Exam3
```

```
littleTest =
mcartest(CalcmodelData)littleTest$p.value
%Imputation
set.seed(12345)
imputation = mice(CalcmodelData, m = 5)
summary(imputation)
% Imputed Values
imputed values = imputation$imp
Major_numeric_imputed = imputed_values$Major numeric Exam3
imputed_= imputed values$Exam3
Exam4_imputed = imputed values$Exam4
% Histogram of Imputed values
ggplot(melt(as.data.frame(Major numeric imputed)), aes(value, fill = variable))+geom histogram(po
"dodge", title = "histogramforMajor numeric imputed")
ggplot(melt(as.data.frame(Exam3 imputed)),
aes(value, fill = variable)) +
geom_histogram(position = "dodge")
ggplot(melt(as.data.frame(Exam4 imputed)), aes(value,
fill = variable)) +
geom_histogram(position = "dodge")
% Comparison of Descriptives Between Original and Imputed Data
%MEANSsapply(Exam3 imputed, mean)mean(CalcmodelData$Exam3, na.rm = TRUE)
sapply(Exam4 imputed, mean)
mean(CalcmodelData$Exam4, na.rm = TRUE)
sapply(Exam3 imputed, var)
var(CalcmodelData$Exam3, na.rm = TRUE)
sapply(Exam4 imputed, var)
var(CalcmodelData$Exam4, na.rm = TRU
```