University of San Carlos
DCISM

# Problem Set 4

MATH 3109

**PREPARED BY:**
Woodrow A. Fajardo

**SUBMITTED TO:**
Prof. Elmer S. Poliquit

**MATH-3109**
**NOVEMBER 2025**

# 1 Raw Data

RAW DATA income.csv

```
setwd("~/School/MATH-3109/PROBSET-4")

raw_data = read.csv("income.csv")
print(raw_data)
View(raw_data)
```

RAW DATA

```
   Age   Income Education Gender Satisfaction PurchaseAmount
1   56 71935.67       PhD   Male         8.60         155.18
2   NA 73080.72    Master   Male         7.73             NA
3   NA 13413.41       PhD Female         5.86             NA
4   25 59051.62      <NA> Female           NA         143.81
5   38 46234.34       PhD   <NA>         9.69         247.50
6   56 47541.99    Master Female         6.46             NA
7   36       NA    Master Female         3.48         222.89
8   40 72304.71      <NA> Female         3.67         115.79
9   28       NA      <NA> Female         2.49         216.34
10  28 55333.27       PhD Female         1.14         195.94
11  41       NA  Bachelor   Male           NA         223.39
12  53       NA       PhD Female         4.55         236.81
13  57 45599.01       PhD   <NA>         3.64         161.01
14  41 49552.42  Bachelor Female         1.13         157.81
15  NA 51426.89  Bachelor Female           NA         192.47
```

# 2 Data Cleaning

a. Identify the missing values in the dataset.

Finding Missing Values

```
na_indices <- which(is.na(raw_data), arr.ind = TRUE)

missing_locations <- data.frame(
    Row_Number = na_indices[, 1],
    Column_Name = colnames(raw_data)[na_indices[, 2]]
)

missing_locations <- missing_locations[order(missing_locations$Row_Number), ]

print("Detailed Missing Value Locations:")
formatted_list <- paste("Row", missing_locations$Row_Number, "-", missing_locations$Column_Name)
print(formatted_list)

print(missing_locations)
View(missing_locations)
```

MISSING VALUES LIST

```
Row_Number    Column_Name
2               Age
2               PurchaseAmount
3               Age
3               PurchaseAmount
4               Satisfaction
6               PurchaseAmount
7               Income
9               Income
11              Income
11              Satisfaction
12              Income
15              Age
15              Satisfaction
```

b. What variables have missing values?

Finding Variables containg missing values

```r
vars_with_missing <- colnames(raw_data)[colSums(is.na(raw_data)) > 0]

cat("Variables containing missing values:\n")
cat(vars_with_missing, sep = "\n")
```

VARIABLES with MISSING VALUES

```
Variables containing missing values:
Age
Income
Satisfaction
PurchaseAmount
```

c. What type of missingness (MCAR, MAR, MNAR) might be present?

From what I can see, the missing data seems to be Missing Completely at Random (MCAR). There isn't any obvious pattern linking the missing spots to specific variables. Since the dataset is so small, the missing values just look like they are scattered randomly throughout the list.

## 3 Imputation

a. Choose one imputation method (mean, regression, kNN, or multiple imputation) and justify your choice.

I decided to use three different methods to fill in the missing data, depending on the type of variable. For the number-based columns like Age, Income, Satisfaction, and Purchase Amount, I used Predictive Mean Matching (PMM). For Gender, I used logistic regression because it only has two options, and for Education, I used polytomous regression since it has multiple categories.

Imputation Methods

```
raw_data$Gender <- factor(raw_data$Gender, levels=c("Male", "Female"))
raw_data$Education <- factor(raw_data$Education, levels=c("PhD", "Master",
    "Bachelor", "Highschool"))

init <- mice(raw_data, maxit=0)

methods <- init$method
methods["Gender"] <- "logreg"
methods["Education"] <- "polyreg"
methods["Age"] <- "pmm"
methods["Income"] <- "pmm"
methods["Satisfaction"] <- "pmm"
methods["PurchaseAmount"] <- "pmm"
```

b. Impute the values using your own method.

Imputation of Data

```
imputed_data <- mice(raw_data, method = methods, m = 5, seed = 123,
    printFlag=FALSE)

complete_data <- complete(imputed_data, 1)
print(complete_data)
```

COMPLETE IMPUTED DATA

```
   Age   Income Education Gender Satisfaction PurchaseAmount
1   56 71935.67       PhD   Male         8.60         155.18
2   36 73080.72    Master   Male         7.73         157.81
3   57 13413.41       PhD Female         5.86         222.89
4   25 59051.62    Master Female         1.14         143.81
5   38 46234.34       PhD   Male         9.69         247.50
6   56 47541.99    Master Female         6.46         192.47
7   36 51426.89    Master Female         3.48         222.89
8   40 72304.71    Master Female         3.67         115.79
9   28 45599.01    Master Female         2.49         216.34
10  28 55333.27       PhD Female         1.14         195.94
11  41 51426.89  Bachelor   Male         9.69         223.39
12  53 47541.99       PhD Female         4.55         236.81
13  57 45599.01       PhD Female         3.64         161.01
14  41 49552.42  Bachelor Female         1.13         157.81
15  41 51426.89  Bachelor Female         2.49         192.47
```

## 4 Exploration

a. Generate summary statistics before and after imputation.

Summary Stat of RAW DATA

```
print(summary(raw_data))
```

STATISTICS OUTPUT

```
      Age            Income            Education     Gender
 Min.   :25.00   Min.   :13413   PhD       :6   Male  : 3
 1st Qu.:34.00   1st Qu.:46888   Master    :3   Female:10
 Median :40.50   Median :51427   Bachelor  :3   NA's  : 2
 Mean   :41.58   Mean   :53225   Highschool:0
 3rd Qu.:53.75   3rd Qu.:65494   NA's      :3
 Max.   :57.00   Max.   :73081
 NA's   :3       NA's   :4

  Satisfaction   PurchaseAmount
 Min.   :1.130   Min.   :115.8
 1st Qu.:3.232   1st Qu.:157.2
 Median :4.110   Median :194.2
 Mean   :4.870   Mean   :189.1
 3rd Qu.:6.777   3rd Qu.:223.0
 Max.   :9.690   Max.   :247.5
 NA's   :3       NA's   :3
```

Summary Stat of COMPLETE DATA

```
print(summary(complete_data))
```

STATISTICS OUTPUT

```
      Age           Income            Education     Gender
 Min.   :25.0   Min.   :13413   PhD       :6   Male  : 4
 1st Qu.:36.0   1st Qu.:46888   Master    :6   Female:11
 Median :41.0   Median :51427   Bachelor  :3
 Mean   :42.2   Mean   :52098   Highschool:0
 3rd Qu.:54.5   3rd Qu.:57192
 Max.   :57.0   Max.   :73081

  Satisfaction   PurchaseAmount
 Min.   :1.130   Min.   :115.8
 1st Qu.:2.490   1st Qu.:157.8
 Median :3.670   Median :192.5
 Mean   :4.784   Mean   :189.5
 3rd Qu.:7.095   3rd Qu.:222.9
 Max.   :9.690   Max.   :247.5
```

The imputation worked well. It filled in the missing numbers without changing the overall "story" or patterns found in the original raw data.

b. Plot histograms or boxplots to compare distributions pre- and post- imputation.

## 4.1 Age Comparison Histogram Plot

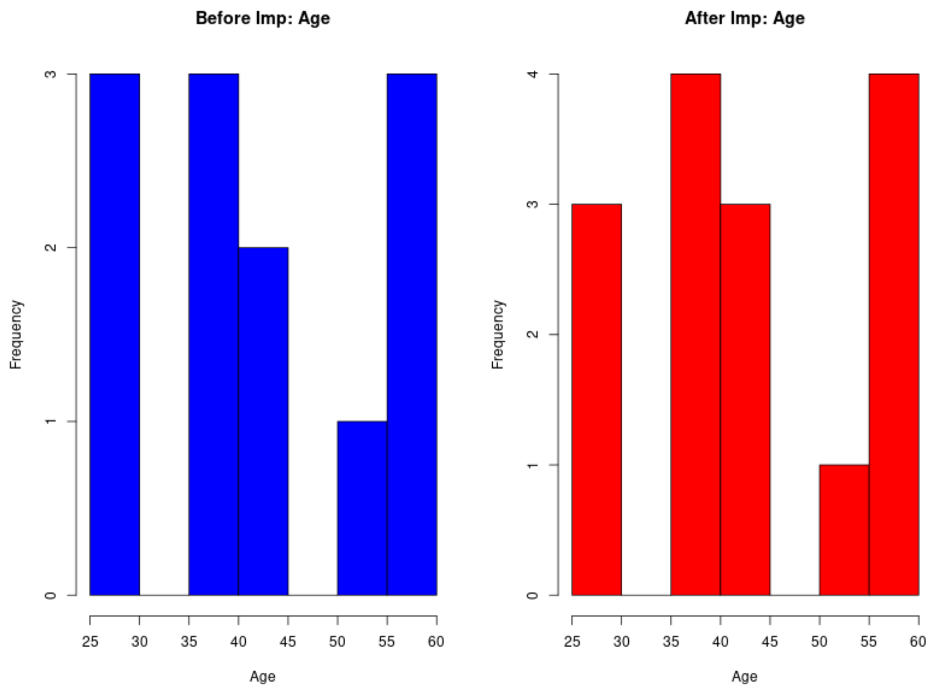


**FIGURE 1**
Age Before and After Imputation Histogram Plot.

Age

```
par(mfrow = c(1, 2))
hist(raw_data$Age,
     main = "Before Imp: Age",
     xlab = "Age",
     col = "blue",
     border = "black")
hist(complete_data$Age,
     main = "After Imp: Age",
     xlab = "Age",
     col = "red",
     border = "black")
```

## 4.2 Income Comparison Histogram Plot

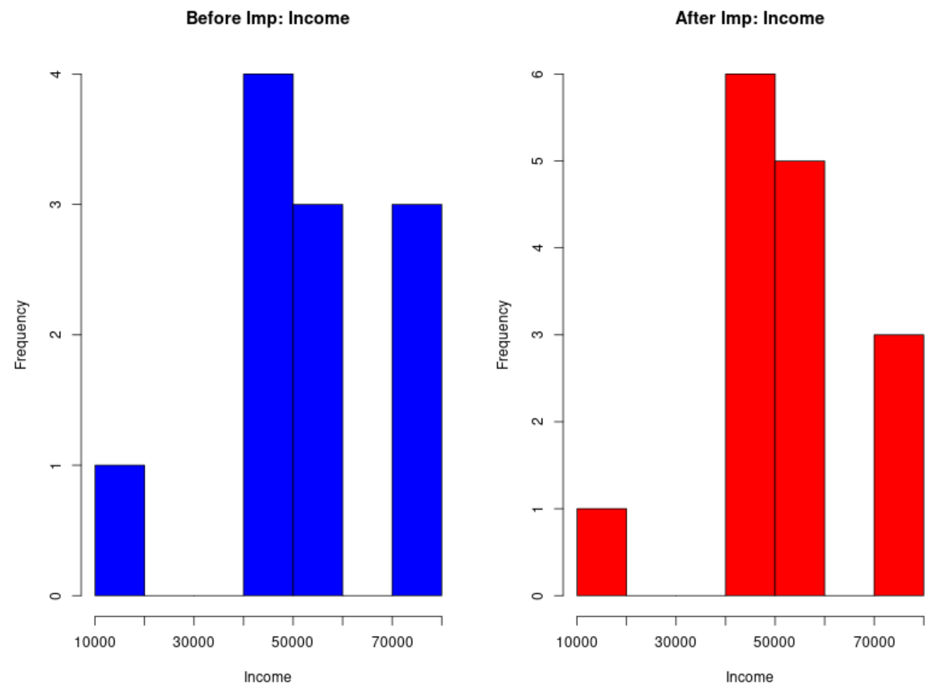


**FIGURE 2**
Income Before and After Imputation Histogram Plot.

Income

```
par(mfrow = c(1, 2))
hist(raw_data$Income,
    main = "Before Imp: Income",
    xlab = "Income",
    col = "blue",
    border = "black")
hist(complete_data$Income,
    main = "After Imp: Income",
    xlab = "Income",
    col = "red",
    border = "black")
```

## 4.3 PurchaseAmount Comparison Histogram Plot

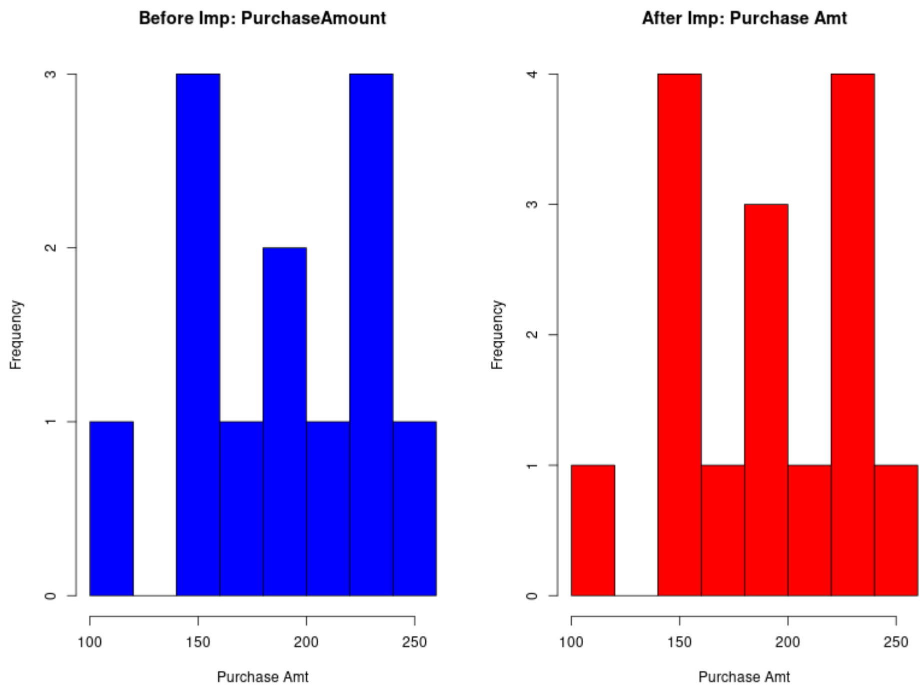


**FIGURE 3**
PurchaseAmount Before and After Imputation Histogram Plot.

Purchase Amount

```
par(mfrow = c(1, 2))
hist(raw_data$PurchaseAmount,
     main = "Before Imp: PurchaseAmount",
     xlab = "Purchase Amt",
     col = "blue",
     border = "black")
hist(complete_data$PurchaseAmount,
     main = "After Imp: Purchase Amt",
     xlab = "Purchase Amt",
     col = "red",
     border = "black")
```

## 4.4 Satisfaction Comparison Histogram Plot

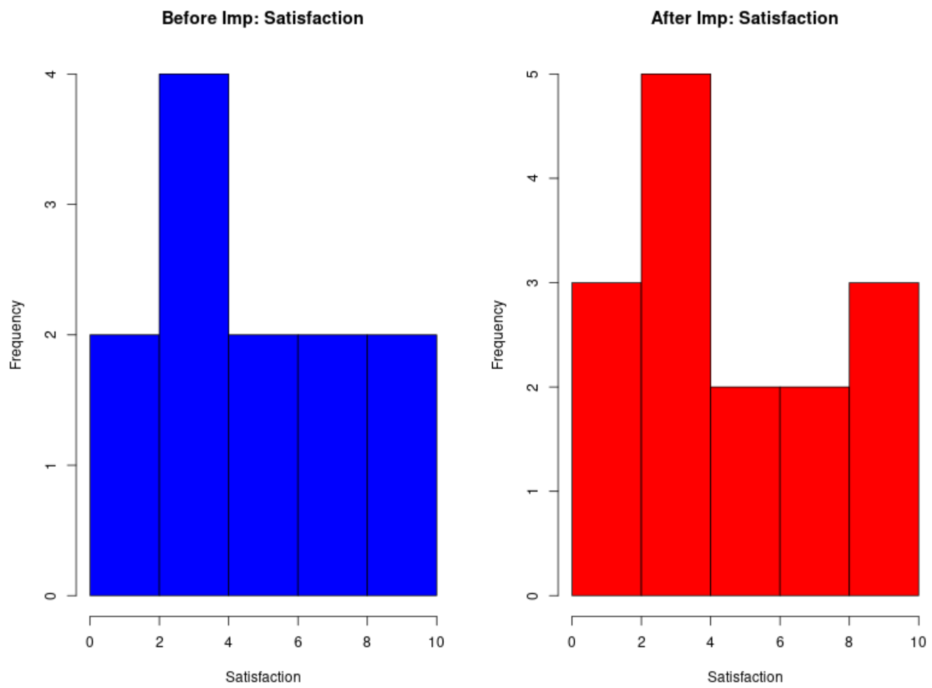


**FIGURE 4**
Satisfaction Before and After Imputation Histogram Plot.

Satisfaction

```
par(mfrow = c(1, 2))
hist(raw_data$Satisfaction,
     main = "Before Imp: Satisfaction",
     xlab = "Satisfaction",
     col = "blue",
     border = "black"
)
hist(complete_data$Satisfaction,
     main = "After Imp: Satisfaction",
     xlab = "Satisfaction",
     col = "red",
     border = "black"
```

## 4.5 Gender Comparison Bar Plot

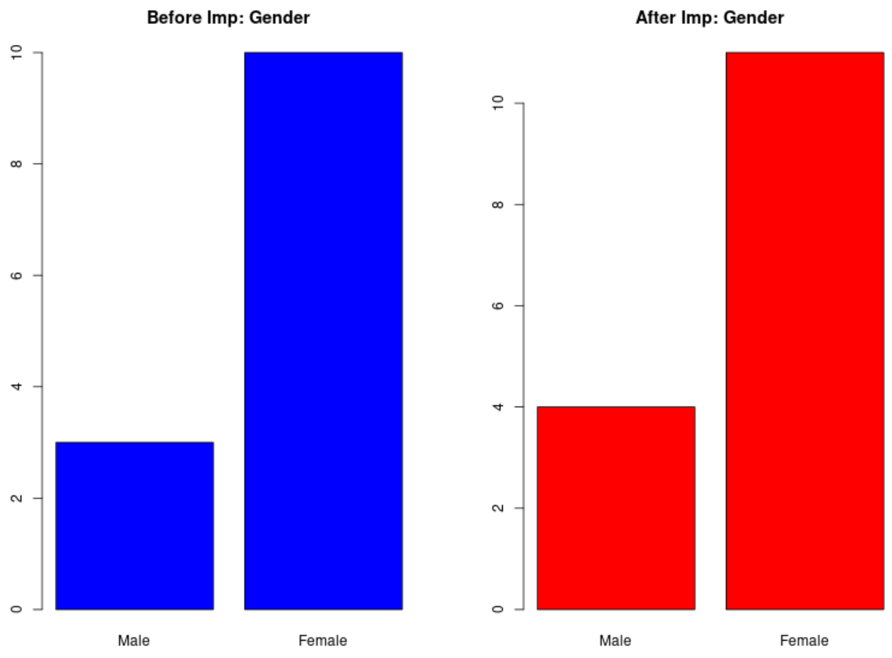


**FIGURE 5**
Gender Before and After Imputation Bar Plot.

Gender

```r
par(mfrow = c(1, 2))
barplot(table(raw_data$Gender),
    main = "Before Imp: Gender",
    col = "blue")
barplot(table(complete_data$Gender),
    main = "After Imp: Gender",
    col = "red")
```

### 4.6 Education Comparison Bar Plot

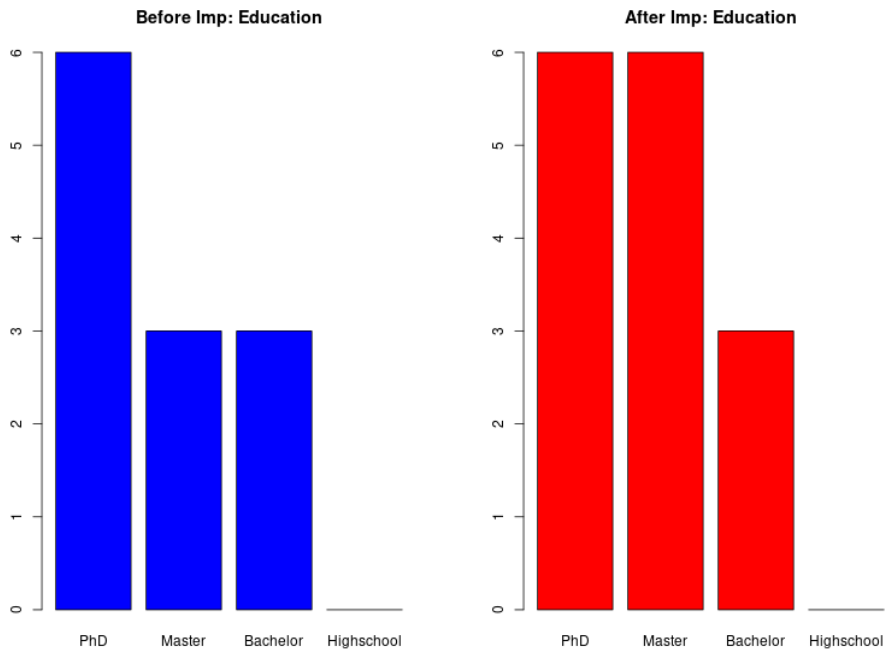


**FIGURE 6**
Education Before and After Imputation Bar Plot.

Education

```
par(mfrow = c(1, 2))
barplot(table(raw_data$Education),
    main = "Before Imp: Education",
    col = "blue")
barplot(table(complete_data$Education),
    main = "After Imp: Education",
    col = "red")
```

### 4.7 Interpretations

Created histograms and bar plots to compare the data before and after filling in the missing values. Looking at the charts, the new data keeps the same shape as the original data. This shows that the missing values were filled in successfully without messing up the original distribution.

## 5 Simulate Synthetic Data

a. Based on the imputed dataset, stimulate a synthetic dataset with $n = 100$.

Synthetic Data Generate

```r
make_syndata <- function(complete_data, method_vec, n_syn = 100){
    na_data <- raw_data[rep(NA, n_syn), ]
    combined <- rbind(complete_data, na_data)
    imp_syn_data <- mice(combined, method = method_vec, m = 1, maxit = 1, seed =
     ↪  123, printFlag = FALSE)
    syn_data <- complete(imp_syn_data, 1)
    syn_data[(nrow(complete_data) + 1):nrow(syn_data), ]
}

synthetic_data <- make_syndata(complete_data, methods, n_syn = 100)
print(tail(synthetic_data, 10))
```

Generated Synthetic Data

```
      Age   Income Education Gender Satisfaction PurchaseAmount
NA.90  41 51426.89       PhD Female         2.49         216.34
NA.91  40 51426.89    Master   Male         6.46         236.81
NA.92  53 71935.67    Master   Male         9.69         157.81
NA.93  36 46234.34  Bachelor Female         3.67         247.50
NA.94  57 46234.34    Master   Male         7.73         192.47
NA.95  28 49552.42  Bachelor   Male         6.46         236.81
NA.96  56 55333.27    Master   Male         9.69         195.94
NA.97  57 47541.99    Master   Male         8.60         223.39
NA.98  28 59051.62  Bachelor   Male         9.69         247.50
NA.99  25 59051.62       PhD Female         3.67         222.89
```

b. Ensure similar distribution and structure to the original imputed data.

Summary Stat of RAW DATA

```r
print(summary(raw_data))
```

STTISTICS OUTPUT

```
     Age            Income            Education      Gender  
Min.   :25.00   Min.   :13413   PhD        :6   Male  : 3  
1st Qu.:34.00   1st Qu.:46888   Master     :3   Female:10  
Median :40.50   Median :51427   Bachelor   :3   NA's  : 2  
Mean   :41.58   Mean   :53225   Highschool:0               
3rd Qu.:53.75   3rd Qu.:65494   NA's       :3               
Max.   :57.00   Max.   :73081                               
NA's   :3       NA's   :4                                   

Satisfaction   PurchaseAmount
Min.   :1.130   Min.   :115.8  
1st Qu.:3.232   1st Qu.:157.2  
Median :4.110   Median :194.2  
Mean   :4.870   Mean   :189.1  
3rd Qu.:6.777   3rd Qu.:223.0  
Max.   :9.690   Max.   :247.5  
NA's   :3       NA's   :3      
```

Summary Stat of SYNTHETIC DATA

```r
print(summary(synthetic_data))
```

STATISTICS OUTPUT

```
      Age            Income             Education      Gender   
 Min.   :25.00   Min.   :13413   PhD        :29   Male  :46  
 1st Qu.:36.00   1st Qu.:46234   Master     :42   Female:54  
 Median :41.00   Median :51427   Bachelor   :29              
 Mean   :42.18   Mean   :52578   Highschool: 0              
 3rd Qu.:56.00   3rd Qu.:55333                              
 Max.   :57.00   Max.   :73081                              

 Satisfaction   PurchaseAmount 
 Min.   :1.130   Min.   :115.8  
 1st Qu.:2.490   1st Qu.:192.5  
 Median :4.550   Median :216.3  
 Mean   :5.199   Mean   :203.3  
 3rd Qu.:7.730   3rd Qu.:223.4  
 Max.   :9.690   Max.   :247.5  
```

The process did a good job of copying the stats for the numerical variables, like Age and Income, the ranges and averages stayed about the same. However, something shifted with the categories. The original data had uneven numbers for Gender and Education, but the new synthetic dataset balanced them out almost perfectly.

## 6 Regression Analysis

a. Use `Purchase Amount` as the outcome.
b. Predictors: `Age`, `Income`, `Satisfaction`, `Gender`, `Education`
c. Interpret Summary Results.

COMPLETE DATA REGRESSION

```
complete_regression <- lm(PurchaseAmount ~ Age + Income + Satisfaction + Gender
 ↪ + Education,
                                          data = complete_data)
print(summary(complete_regression))
```

REGRESSION SUMMARY OUTPUT

```
Call:
lm(formula = PurchaseAmount ~ Age + Income + Satisfaction + Gender +
    Education, data = complete_data)

Residuals:
    Min      1Q  Median      3Q     Max
-40.471  -8.694  -1.920   7.119  37.842

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.267e+02  9.466e+01   2.395   0.0435 *
Age               -2.955e+00  1.304e+00  -2.267   0.0531 .
Income            -1.072e-03  8.618e-04  -1.244   0.2486
Satisfaction       1.924e+01  9.865e+00   1.950   0.0870 .
GenderFemale       9.231e+01  6.767e+01   1.364   0.2097
EducationMaster   -3.764e+01  2.518e+01  -1.495   0.1732
EducationBachelor -6.761e+00  2.057e+01  -0.329   0.7508
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.89 on 8 degrees of freedom
Multiple R-squared:  0.7076,    Adjusted R-squared:  0.4883
F-statistic: 3.226 on 6 and 8 DF,  p-value: 0.06462
```

SYNTHETIC DATA REGRESSION

```
synthetic_regression <- lm(PurchaseAmount ~ Age + Income + Satisfaction + Gender
 ↪ + Education,
                                        data = synthetic_data)
print(summary(synthetic_regression))
```

REGRESSION SUMMARY OUTPUT

```
Call:
lm(formula = PurchaseAmount ~ Age + Income + Satisfaction + Gender +
    Education, data = synthetic_data)

Residuals:
    Min      1Q  Median      3Q     Max
-51.980 -16.896   1.313  17.227  44.868

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.226e+02  1.993e+01  16.186  < 2e-16 ***
Age               -6.077e-01  2.716e-01  -2.238 0.027622 *
Income            -9.543e-04  2.239e-04  -4.262 4.86e-05 ***
Satisfaction      -2.114e+00  1.748e+00  -1.210 0.229385
GenderFemale      -3.860e+01  9.565e+00  -4.036 0.000112 ***
EducationMaster   -2.751e+01  7.492e+00  -3.672 0.000401 ***
EducationBachelor -5.407e-01  7.205e+00  -0.075 0.940345
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.98 on 93 degrees of freedom
Multiple R-squared:  0.5131,    Adjusted R-squared:  0.4817
F-statistic: 16.34 on 6 and 93 DF,  p-value: 9.215e-13
```

1000 SYNTHETIC DATA REGRESSION

```
synthetic1000_data <- make_syndata(complete_data, methods, n_syn = 1000)
synthetic1000_regression <- lm(PurchaseAmount ~ Age + Income + Satisfaction + Gender + Education,
                                                     data = synthetic1000_data)
print(summary(synthetic1000_regression))
```

REGRESSION SUMMARY OUTPUT

```
Call:
lm(formula = PurchaseAmount ~ Age + Income + Satisfaction + Gender +
    Education, data = synthetic1000_data)

Residuals:
   Min     1Q Median     3Q    Max
-63.06 -14.83   0.91  16.08  60.76

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.940e+02  6.274e+00  30.919  < 2e-16 ***
Age               -2.415e+00  8.554e-02 -28.235  < 2e-16 ***
Income            -3.455e-04  6.368e-05  -5.425 7.29e-08 ***
Satisfaction       1.503e+01  5.390e-01  27.894  < 2e-16 ***
GenderFemale       7.641e+01  2.773e+00  27.551  < 2e-16 ***
EducationMaster   -3.862e+01  1.962e+00 -19.691  < 2e-16 ***
EducationBachelor -6.761e+00  2.324e+00  -2.910   0.0037 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.13 on 993 degrees of freedom
Multiple R-squared:  0.5944,    Adjusted R-squared:  0.5919
F-statistic: 242.5 on 6 and 993 DF,  p-value: < 2.2e-16
```

POOLED MODEL REGRESSION

```
pooled_model <- with(imputed_data, lm(PurchaseAmount ~ Age + Income + Satisfaction
                                                     + Gender + Education))
pooled_results <- pool(pooled_model)
print(summary(pooled_results))
```

REGRESSION POOLED SUMMARY OUTPUT

```
               term      estimate    std.error  statistic       df    p.value
1       (Intercept) 240.780120493 1.158604e+02  2.0781912 6.128225 0.08196649
2               Age  -1.389056914 1.879782e+00 -0.7389457 2.194601 0.53089093
3            Income  -0.000924375 1.236205e-03 -0.7477523 4.092732 0.49528832
4      Satisfaction   9.459600865 1.231774e+01  0.7679656 3.269060 0.49417910
5      GenderFemale  28.879693995 7.892727e+01  0.3659026 3.114476 0.73788660
6   EducationMaster -27.814550761 5.081571e+01 -0.5473613 2.229740 0.63400953
7 EducationBachelor  -3.169597038 3.072495e+01 -0.1031604 6.192437 0.92109579
```

## 6.1 Interpretation

This analysis really shows why choosing the right model and having enough data matters. In my first attempt with a single imputed dataset, being female looked like it increased spending by about 92.31. But when I tried it with a small synthetic dataset ($n = 100$), it got it wrong and predicted a decrease instead. When I used a larger synthetic dataset ($n = 1000$) or the Pooled Model (which averages results from five datasets), the prediction went back to being positive. The Pooled Model gave a more conservative estimate of an increase of 28.87. It seems like the Pooled Model is the most reliable here because it smooths out extreme values, though the small size of the original data still makes it hard to be completely sure about the results.