University of San Carlos
DCISM

Final Culminating Activity

MATH 3109

PREPARED BY:
Woodrow A. Fajardo

SUBMITTED TO:
Prof. Elmer S. Poliquit

MATH-3109
NOVEMBER 2025

Contents

Raw Data based from Kaggle:

1 Introduction

1.1 Project Objective

1.2 Literature Review

1.3 Raw Data

STATISTICAL MODELING

```r
library(mice)
library(corrplot)
library(ggplot2)
library(splines)
library(performance)
library(see)
library(ggpubr)

setwd("~/School/MATH-3109/FINAL-1")
raw_data <- read.csv("raw_dataset.csv")
View(raw_data)
head(raw_data)
```

TERMINAL OUTPUT

```
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
  ↪  DiabetesPedigreeFunction Age Outcome
1           6     148            72            35       0 33.6
↪  0.627  50       1
2           1      85            66            29       0 26.6
↪  0.351  31       0
3           8     183            64             0       0 23.3
↪  0.672  32       1
4           1      89            66            23      94 28.1
↪  0.167  21       0
5           0     137            40            35     168 43.1
↪  2.288  33       1
6           5     116            74             0       0 25.6
↪  0.201  30       0
```

2 Data Cleaning

2.1 Converting Zeroes Values to NA

STATISTICAL MODELING

```r
colSums(raw_data == 0)
```

TERMINAL OUTPUT

```
Pregnancies                 Glucose            BloodPressure
 ↪  SkinThickness                 Insulin                       BMI
 ↪  DiabetesPedigreeFunction                    Age                   Outcome
          NA                        NA                        NA
           ↪  NA                        NA                        NA
           ↪  0                         0                       500
```

STATISTICAL MODELING

```
col_w_zeroes <- c("Pregnancies", "Glucose", "BloodPressure", "SkinThickness",
 ↪  "Insulin", "BMI")
raw_data[col_w_zeroes] <- lapply(raw_data[col_w_zeroes], function(x) ifelse(x ==
 ↪  0, NA, x))

summary(raw_data)
```

TERMINAL OUTPUT

```
  Pregnancies        Glucose      BloodPressure    SkinThickness      Insulin
   ↪  BMI        DiabetesPedigreeFunction      Age            Outcome
 Min.   : 1.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.00   Min.   : 14.00
 ↪  Min.   :18.20   Min.   :0.0780           Min.   :21.00   Min.   :0.000
 1st Qu.: 2.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00   1st Qu.: 76.25
 ↪  1st Qu.:27.50   1st Qu.:0.2437           1st Qu.:24.00   1st Qu.:0.000
 Median : 4.000   Median :117.0   Median : 72.00   Median :29.00   Median :125.00
 ↪  Median :32.30   Median :0.3725           Median :29.00   Median :0.000
 Mean   : 4.495   Mean   :121.7   Mean   : 72.41   Mean   :29.15   Mean   :155.55
 ↪  Mean   :32.46   Mean   :0.4719           Mean   :33.24   Mean   :0.349
 3rd Qu.: 7.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00   3rd Qu.:190.00
 ↪  3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00   3rd Qu.:1.000
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.00
 ↪  Max.   :67.10   Max.   :2.4200           Max.   :81.00   Max.   :1.000
 NA's   :111      NA's   :5       NA's   :35       NA's   :227     NA's   :374
 ↪  NA's   :11
```

2.2 MICE Imputation

STATISTICAL MODELING

```
init <- mice(raw_data, maxit = 0)
methods <- init$method #$
methods[col_w_zeroes] <- "pmm"
imputed_data <- mice(raw_data, method = methods, m = 5, maxit = 5, seed = 123)
final_data <- complete(imputed_data, 1)
head(final_data)
```

View full R script output here.

STATISTICAL MODELING

```
anyNA(final_data)
```

TERMINAL OUTPUT

```
anyNA(final_data)
```

2.3 Imputation Validation

Refer to Appendix Section for the full R script.

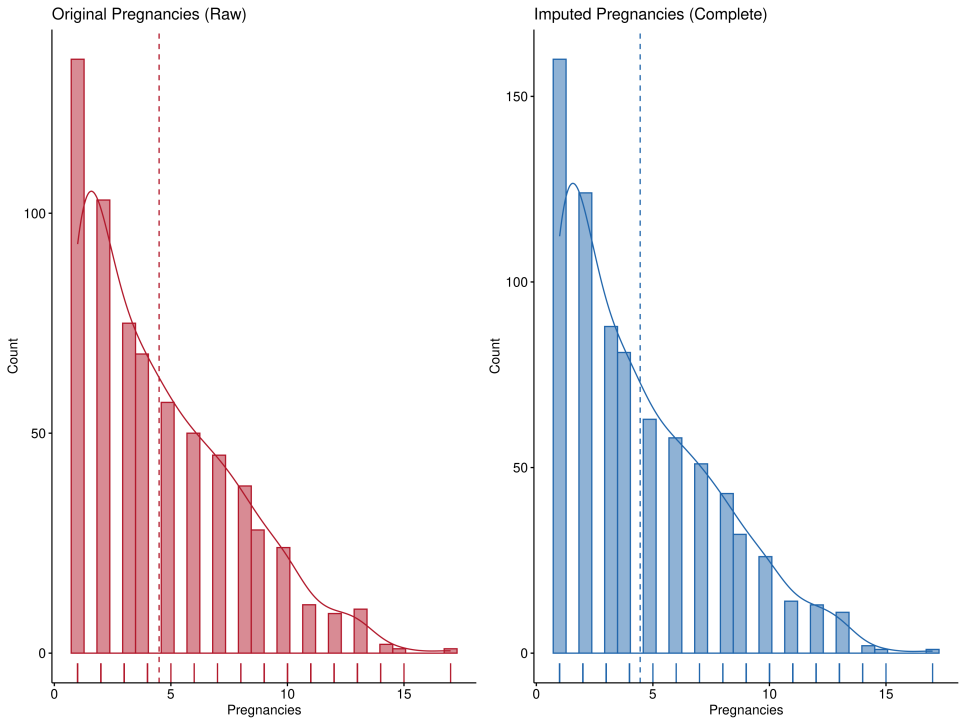

FIGURE 1
A standard centered figure.

Source: Generated by LaTeX

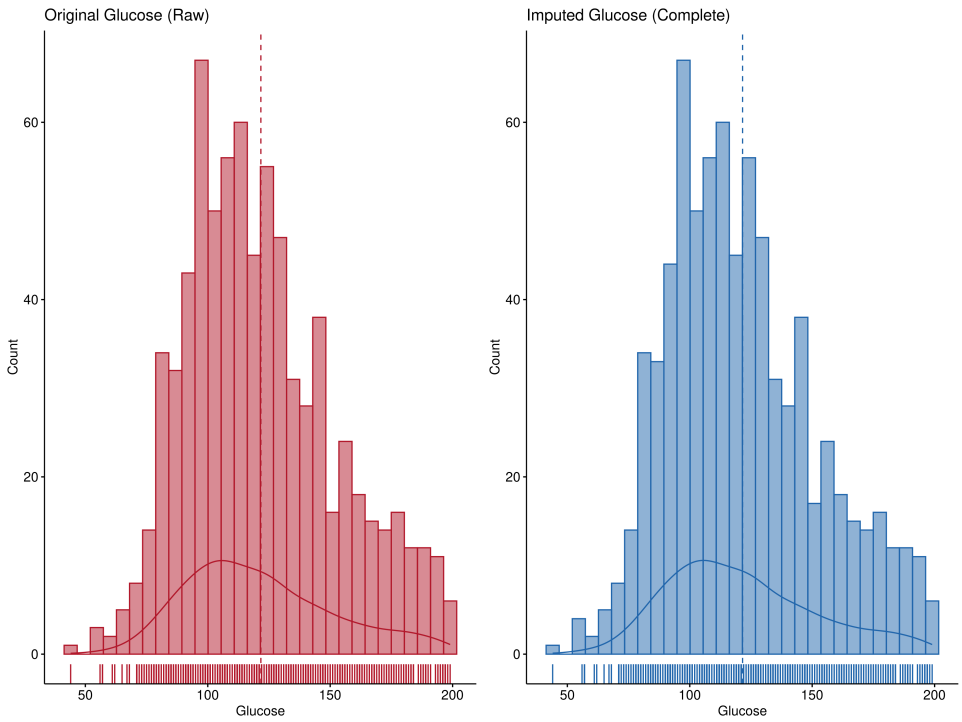

FIGURE 2

A standard centered figure.

Source: Generated by LaTeX

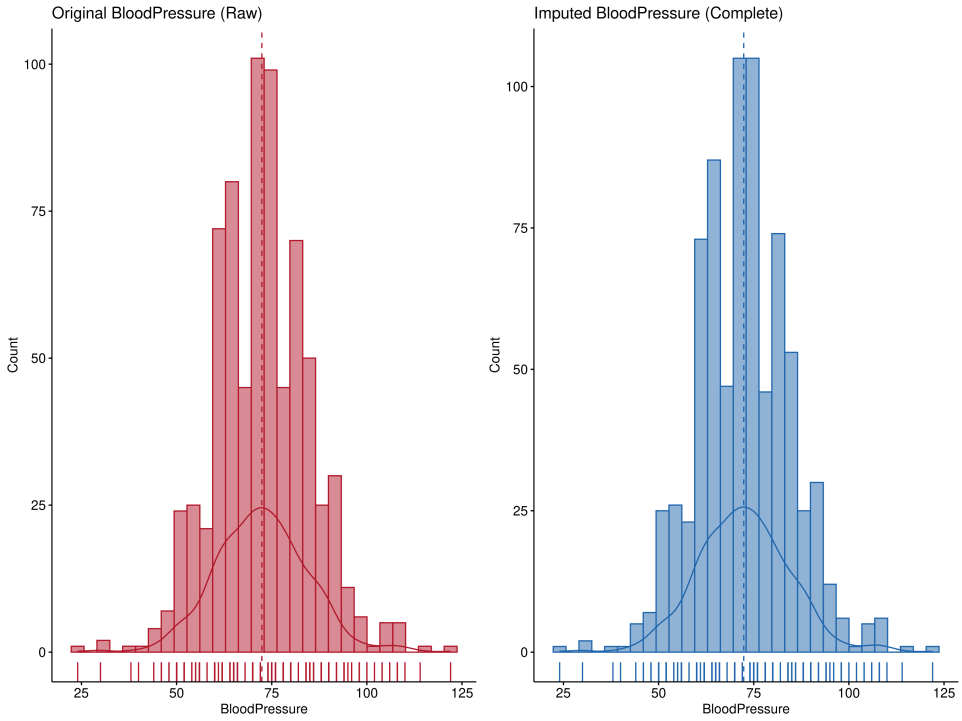

FIGURE 3

A standard centered figure.

Source: Generated by LaTeX

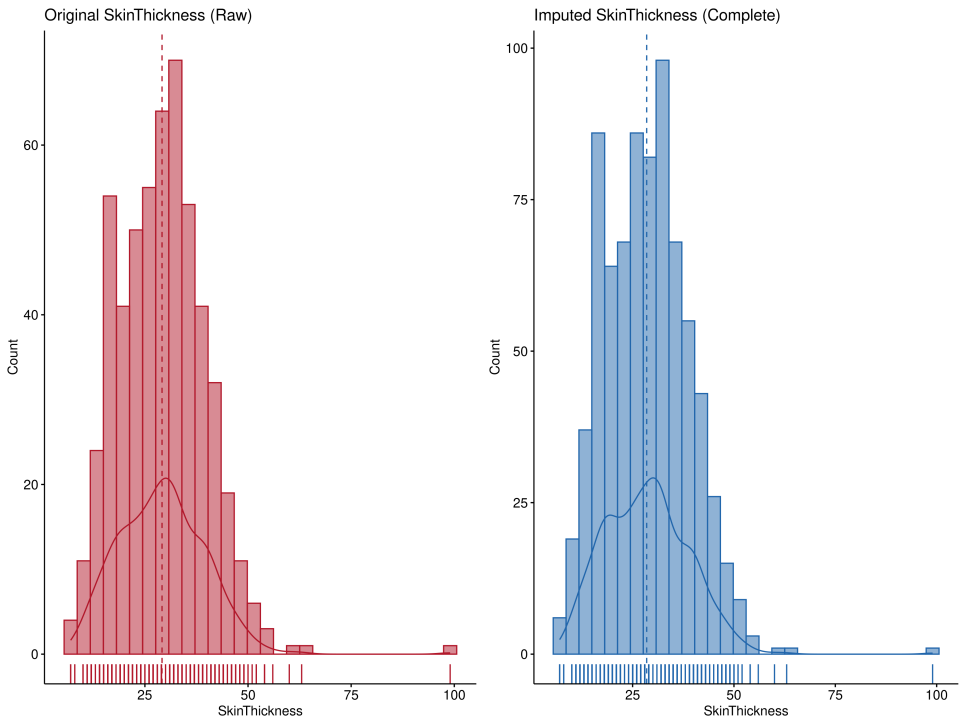

FIGURE 4

A standard centered figure.

Source: Generated by LaTeX

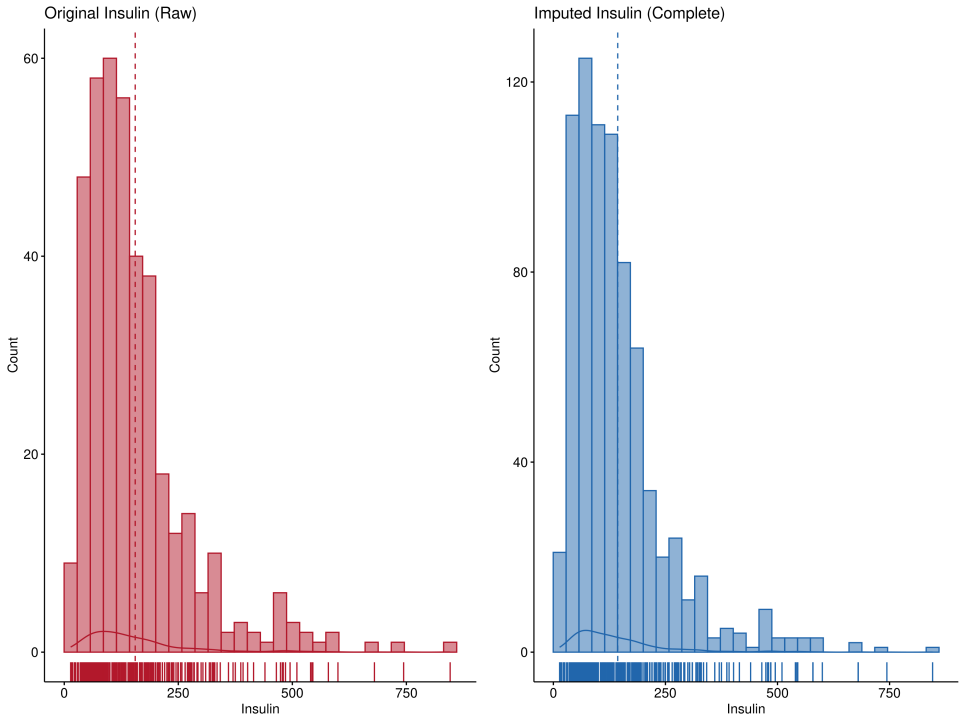

FIGURE 5

A standard centered figure.

Source: Generated by LaTeX

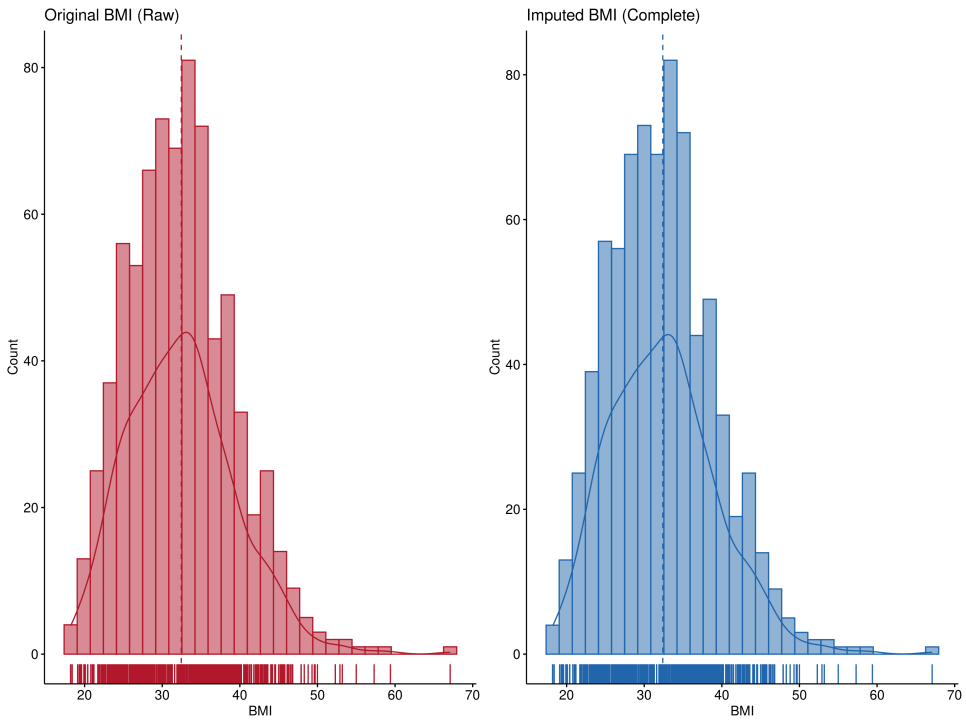

FIGURE 6

A standard centered figure.

Source: Generated by LaTeX

3 Exploratory Analysis

3.1 Correlation Matrix

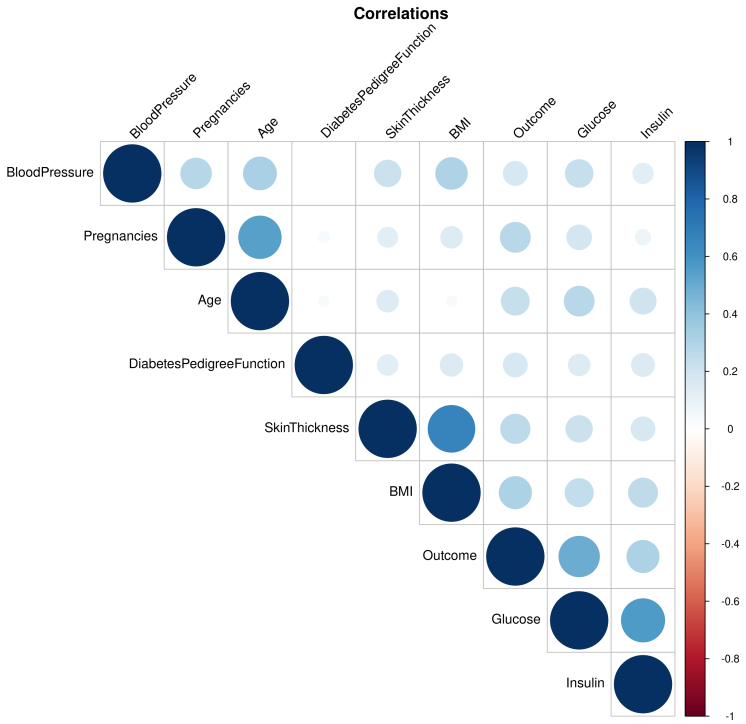

FIGURE 7
A standard centered figure.

Source: Generated by LaTeX

STATISTICAL MODELING

```
cor_matrix <- cor(final_data, use = "complete.obs")
corrplot(cor_matrix, method = "circle", type = "upper", order = "hclust",
        tl.col = "black", tl.srt = 45, title = "Correlations", mar=c(0,0,1,0))
```

3.2 Analyzing Linearity

4 Modelling Logistic Regression

4.1 Model Specification

STATISTICAL MODELING

```
set.seed(123)
sample_index <- sample(1:nrow(final_data), 0.8 * nrow(final_data))
train_data <- final_data[sample_index, ]
test_data <- final_data[-sample_index, ]

log_model <- glm(Outcome ~ Glucose + BMI + Age + Pregnancies +
↪  DiabetesPedigreeFunction,
                data = train_data, family= "binomial")

summary(log_model)
```

TERMINAL OUTPUT

```
glm(formula = Outcome ~ Glucose + BMI + Age + Pregnancies +
↪  DiabetesPedigreeFunction,
    family = "binomial", data = train_data)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -9.200623   0.809675 -11.363  < 2e-16 ***
Glucose                   0.037413   0.004099   9.128  < 2e-16 ***
BMI                       0.079814   0.016719   4.774 1.81e-06 ***
Age                       0.009594   0.010003   0.959 0.337506
Pregnancies               0.130485   0.037653   3.465 0.000529 ***
DiabetesPedigreeFunction  0.659625   0.321580   2.051 0.040247 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 796.42  on 613  degrees of freedom
Residual deviance: 568.82  on 608  degrees of freedom
AIC: 580.82

Number of Fisher Scoring iterations: 5
```

4.2 Odds Ratio Interpretation

STATISTICAL MODELING

```
1 exp(coef(log_model))
```

TERMINAL OUTPUT

```
  (Intercept)                 Glucose                     BMI
  ↪  Age             Pregnancies DiabetesPedigreeFunction
0.0001009765            1.0381215630            1.0830855110
↪  1.0096399484            1.1393803946            1.9340669910
```

4.3 Model Performance

5 Linear Regression (Predict Insulin)

STATISTICAL MODELING

```
1 linear_simple <- lm(Insulin ~ Glucose + BMI + Age, data = final_data)
2 summary(linear_simple)
```

TERMINAL OUTPUT

```
Call:
lm(formula = Insulin ~ Glucose + BMI + Age, data = final_data)

Residuals:
    Min      1Q  Median      3Q     Max
-289.02  -45.91  -14.70   25.01  567.44

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -167.6456    19.1120  -8.772  < 2e-16 ***
Glucose        1.8543     0.1125  16.481  < 2e-16 ***
BMI            2.0651     0.4791   4.310 1.84e-05 ***
Age            0.5693     0.2841   2.004   0.0455 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89.01 on 764 degrees of freedom
Multiple R-squared:  0.3401,    Adjusted R-squared:  0.3375
F-statistic: 131.2 on 3 and 764 DF,  p-value: < 2.2e-16
```

STATISTICAL MODELING

```r
linear_spline <- lm (Insulin ~ Glucose + BMI + bs(Age, degree = 3), data =
↪  final_data)
summary(linear_spline)
```

TERMINAL OUTPUT

```
Call:
lm(formula = Insulin ~ Glucose + BMI + bs(Age, degree = 3), data = final_data)

Residuals:
    Min      1Q  Median      3Q     Max
-299.47  -47.78  -14.42   26.00  554.90

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -151.3071    18.2230  -8.303 4.61e-16 ***
Glucose                 1.8612     0.1123  16.571  < 2e-16 ***
BMI                     2.2633     0.4871   4.646 3.98e-06 ***
bs(Age, degree = 3)1  -57.2504    32.8240  -1.744   0.0815 .
bs(Age, degree = 3)2   80.9994    50.5390   1.603   0.1094
bs(Age, degree = 3)3    5.4944    62.7432   0.088   0.9302
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 88.82 on 762 degrees of freedom
Multiple R-squared:  0.3447,    Adjusted R-squared:  0.3404
F-statistic: 80.16 on 5 and 762 DF,  p-value: < 2.2e-16
```

5.1 B-Spline Justification

Refer to Appendix Section for the full R script.

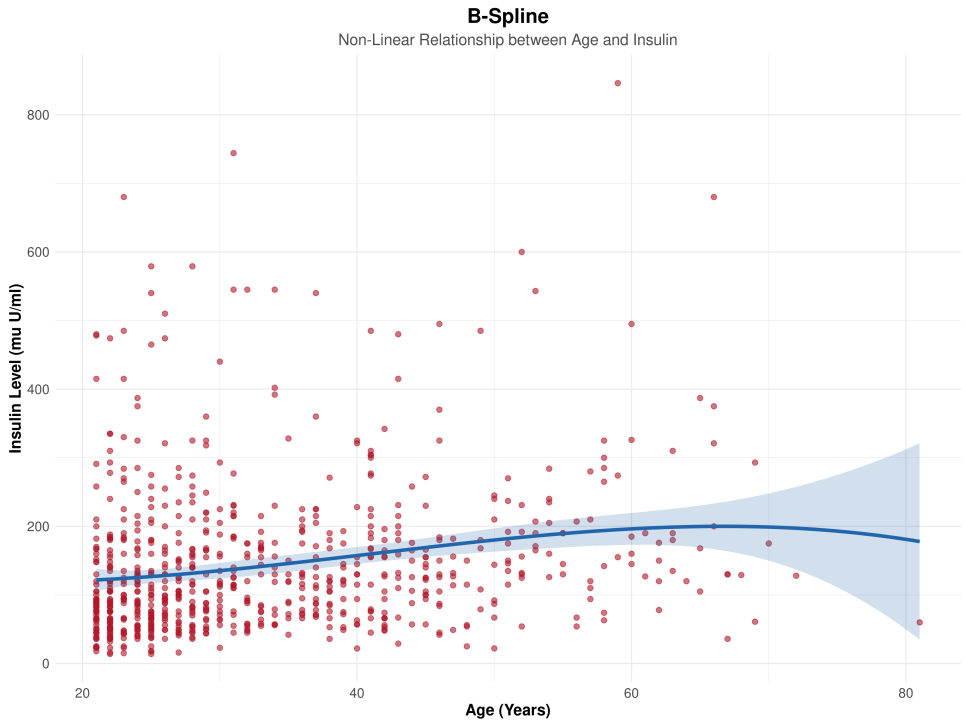

FIGURE 8

A standard centered figure.

Source: Generated by LaTeX

5.2 Comparison Linear vs. B-Spline

STATISTICAL MODELING

```
linear_simple <- lm(Insulin ~ Glucose + BMI + Age, data = final_data)
linear_spline <- lm (Insulin ~ Glucose + BMI + bs(Age, degree = 3), data =
  ↪  final_data)
anova(linear_simple, linear_spline)
```

TERMINAL OUTPUT

```
Analysis of Variance Table

Model 1: Insulin ~ Glucose + BMI + Age
Model 2: Insulin ~ Glucose + BMI + bs(Age, degree = 3)
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1    764 6052997
2    762 6010880  2     42116 2.6696 0.06993 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.3 Residual Diagnostics

FIGURE 9

A standard centered figure.

Source: Generated by LaTeX

STATISTICAL MODELING

```
image_rend <- check_model(linear_spline)
png("splien_model.png", width = 12, height = 9, units = "in", res = 300)
print(image_rend)
dev.off()
```

6 Discussions

6.1 Validation of Literature Claims

6.2 Key Findings

6.3 Limitations (if any)

7 Appendix

HISTOGRAM COMPARISON CODE EXAMPLE

```
p1 <- gghistogram(raw_data, x = "Pregnancies",
    title = "Original Pregnancies (Raw)",
    xlab = "Pregnancies", ylab = "Count",
    fill = "#b2182b",
    color = "#b2182b",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

p2 <- gghistogram(final_data, x = "Pregnancies",
    title = "Imputed Pregnancies (Complete)",
    xlab = "Pregnancies", ylab = "Count",
    fill = "#2166ac",
    color = "#2166ac",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

ggarrange(p1, p2, ncol = 2, nrow = 1)
```

STATISTICAL MODELING

```
p1 <- gghistogram(raw_data, x = "Glucose",
    title = "Original Glucose (Raw)",
    xlab = "Glucose", ylab = "Count",
    fill = "#b2182b",
    color = "#b2182b",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

p2 <- gghistogram(final_data, x = "Glucose",
    title = "Imputed Glucose (Complete)",
    xlab = "Glucose", ylab = "Count",
    fill = "#2166ac",
    color = "#2166ac",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

ggarrange(p1, p2, ncol = 2, nrow = 1)
```

STATISTICAL MODELING

```r
p1 <- gghistogram(raw_data, x = "BloodPressure",
    title = "Original BloodPressure (Raw)",
    xlab = "BloodPressure", ylab = "Count",
    fill = "#b2182b",
    color = "#b2182b",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

p2 <- gghistogram(final_data, x = "BloodPressure",
    title = "Imputed BloodPressure (Complete)",
    xlab = "BloodPressure", ylab = "Count",
    fill = "#2166ac",
    color = "#2166ac",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

ggarrange(p1, p2, ncol = 2, nrow = 1)
```

STATISTICAL MODELING

```r
p1 <- gghistogram(raw_data, x = "SkinThickness",
    title = "Original SkinThickness (Raw)",
    xlab = "SkinThickness", ylab = "Count",
    fill = "#b2182b",
    color = "#b2182b",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

p2 <- gghistogram(final_data, x = "SkinThickness",
    title = "Imputed SkinThickness (Complete)",
    xlab = "SkinThickness", ylab = "Count",
    fill = "#2166ac",
    color = "#2166ac",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

ggarrange(p1, p2, ncol = 2, nrow = 1)
```

STATISTICAL MODELING

```r
p1 <- gghistogram(raw_data, x = "Insulin",
    title = "Original Insulin (Raw)",
    xlab = "Insulin", ylab = "Count",
    fill = "#b2182b",
    color = "#b2182b",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

p2 <- gghistogram(final_data, x = "Insulin",
    title = "Imputed Insulin (Complete)",
    xlab = "Insulin", ylab = "Count",
    fill = "#2166ac",
    color = "#2166ac",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

ggarrange(p1, p2, ncol = 2, nrow = 1)
```

STATISTICAL MODELING

```r
p1 <- gghistogram(raw_data, x = "BMI",
    title = "Original BMI (Raw)",
    xlab = "BMI", ylab = "Count",
    fill = "#b2182b",
    color = "#b2182b",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

p2 <- gghistogram(final_data, x = "BMI",
    title = "Imputed BMI (Complete)",
    xlab = "BMI", ylab = "Count",
    fill = "#2166ac",
    color = "#2166ac",
    add = "mean", rug = TRUE, add_density = TRUE,
    ggtheme = theme_pubr()
)

ggarrange(p1, p2, ncol = 2, nrow = 1)
```

STATISTICAL MODELING

```r
ggplot(final_data, aes(x = Age, y = Insulin)) +
  geom_point(color = "#b2182b", alpha = 0.6, size = 2) +
  geom_smooth(method = "lm",
              formula = y ~ bs(x, degree = 3),
              color = "#2166ac",
              fill = "#2166ac",
              alpha = 0.2,
              size = 1.5) +

  labs(title = "B-Spline",
       subtitle = "Non-Linear Relationship between Age and Insulin",
       x = "Age (Years)",
       y = "Insulin Level (mu U/ml)") +

  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5, size = 18),
    plot.subtitle = element_text(hjust = 0.5, color = "gray30", size = 14),
    axis.title = element_text(face = "bold", size = 14),
    axis.text = element_text(size = 12))
```