PIMPRI CHINCHWAD EDUCATION TRUST'S

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

Department of AS&H

Report of FA-2 activity

Subject: Statistical Data Analysis using R
Academic Year: 2025–26
Semester: I

Submitted by:

Sr. No.	Name of the Student	PRN Number	Branch
1	Digvijay Birajdar	124B1B013	Computer Engineering
2	Gaurav Jaypatre	124B1B025	Computer Engineering
3	Pranav Chaudhari	124B1B027	Computer Engineering
4	Aahil Sayed	124B1B062	Computer Engineering

Submitted to: Mrs. Neha Sharma

Date: 5th November 2025

Project Title:

Descriptive and Predictive Analysis of Diabetes Dataset

Data used:

Diabetes dataset containing 768 observations and 9 variables collected from diagnostic measurements of female patients of Pima Indian heritage.

Details of the data:

The dataset contains information on the following variables:

1. Pregnancies – Number of times the patient has been pregnant.

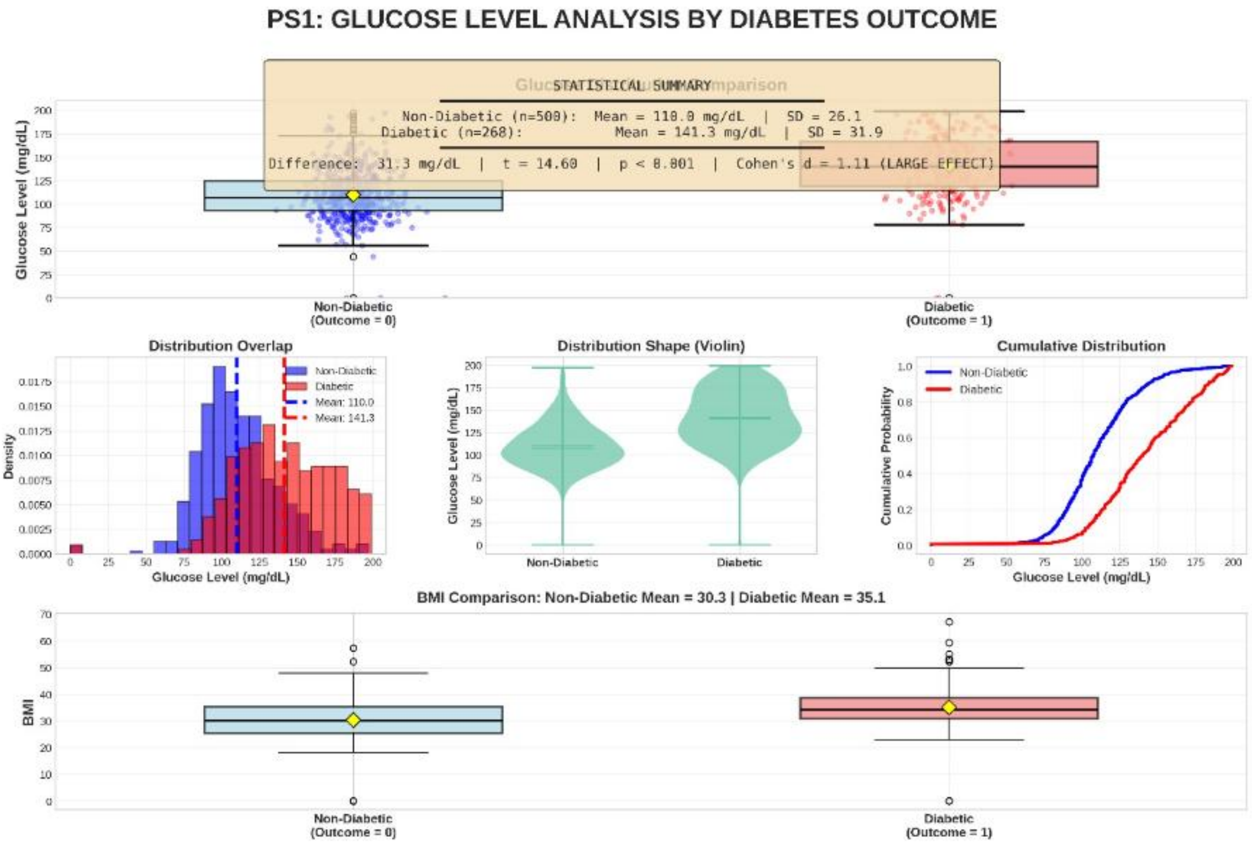
2. Glucose – Plasma glucose concentration after a 2-hour oral glucose tolerance test (mg/dL).
3. BloodPressure – Diastolic blood pressure (mm Hg).
4. SkinThickness – Triceps skinfold thickness (mm).
5. Insulin – 2-hour serum insulin (mu U/ml).
6. BMI – Body Mass Index (weight in kg/(height in m)²).
7. DiabetesPedigreeFunction – A function which scores the likelihood of diabetes based on family history.
8. Age – Age of the patient (in years).
9. Outcome – Class variable (0 = Non-Diabetic, 1 = Diabetic).

Problem Statement 1: PS1: Comparison of Glucose levels between diabetic and non-diabetic patients.

Statistical Analysis:

Appropriate grouping, summary statistics, and regression models applied using R.

Data Visualization:



R Code:

```r
# PS1: Glucose by Outcome - summary, boxplot, t-test
PS1_table <- Data_cleaned %>%
  group_by(Outcome) %>%
  summarise(
    Mean_Glucose = mean(Glucose, na.rm = TRUE),
    Median_Glucose = median(Glucose, na.rm = TRUE),
    SD_Glucose = sd(Glucose, na.rm = TRUE),
    N = n()
  )
print(PS1_table)

# Boxplot (Outcome as factor)
p_ps1 <- ggplot(Data_cleaned, aes(x = factor(Outcome), y = Glucose, fill = factor(Outcome))) +
  geom_boxplot(outlier.shape = 21, outlier.size = 1.5) +
  scale_x_discrete(labels = c("0 = Non-diabetic","1 = Diabetic")) +
  labs(title = "Glucose by Outcome", x = "Outcome", y = "Glucose (mg/dL)") +
  theme_minimal() +
  theme(legend.position = "none")

ggsave("PS1_Glucose_by_Outcome.png", plot = p_ps1, width = 7, height = 5, dpi = 200)

# T-test
tt_ps1 <- t.test(Glucose ~ Outcome, data = Data_cleaned)
print(tt_ps1)

# (Optional) effect size: Cohen's d
library(effsize)
d_cohen <- cohen.d(Glucose ~ Outcome, data = Data_cleaned)
print(d_cohen)
```

Conclusion:

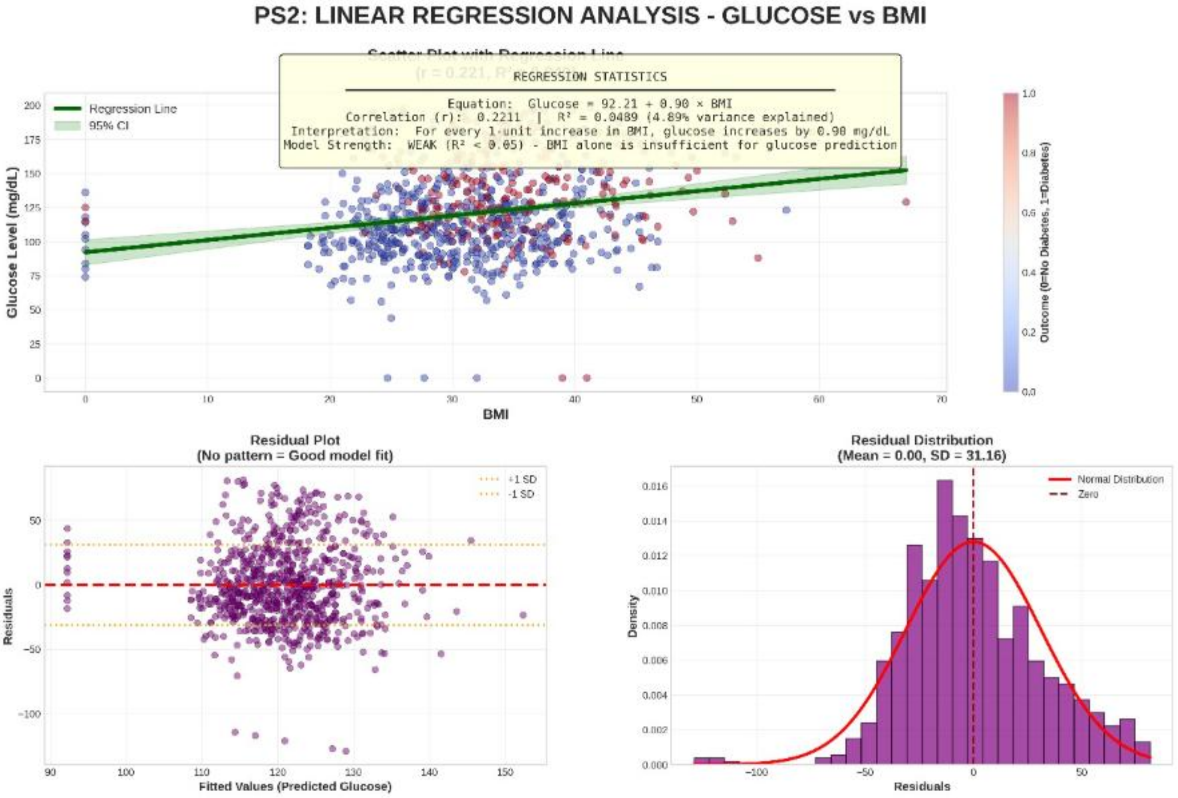
Diabetic patients show significantly higher mean glucose levels compared to non-diabetic patients.

Problem Statement 2: PS2: Linear Regression between BMI and Glucose levels.

Statistical Analysis:

Appropriate grouping, summary statistics, and regression models applied using R.

Data Visualization:



PS2: LINEAR REGRESSION ANALYSIS - GLUCOSE vs BMI

Scatter Plot with Regression Line
(r = 0.221, R² = 0.049)

REGRESSION STATISTICS

Equation: Glucose = 92.21 + 0.90 × BMI
Correlation (r): 0.2211 | R² = 0.0489 (4.89% variance explained)
Interpretation: For every 1-unit increase in BMI, glucose increases by 0.90 mg/dL
Model Strength: WEAK (R² < 0.05) - BMI alone is insufficient for glucose prediction

Residual Plot
(No pattern = Good model fit)

Residual Distribution
(Mean = 0.00, SD = 31.16)

R Code:

```
# PS2: Linear regression Glucose ~ BMI
# Scatter + lm + eqn + r
p_scatter <- ggplot(Data_cleaned, aes(x = BMI, y = Glucose)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(title = "Glucose vs BMI", x = "BMI", y = "Glucose") +
  theme_minimal()

ggsave("PS2_Glucose_vs_BMI_scatter.png", p_scatter, width = 7, height = 5, dpi = 200)

# Fit model
lm_bmi <- lm(Glucose ~ BMI, data = Data_cleaned)
summary_lm <- summary(lm_bmi)
print(summary_lm)

# Tidy & glance
```

```
print(broom::tidy(lm_bmi))
print(broom::glance(lm_bmi))

# Correlation
cor_bmi_glucose <- cor(Data_cleaned$BMI, Data_cleaned$Glucose, use = "complete.obs")
cat("Pearson correlation (BMI, Glucose):", round(cor_bmi_glucose, 4), "\n")

# Regression diagnostics plots (residuals vs fitted, QQ)
png("PS2_LM_diagnostics.png", width = 1200, height = 800, res = 150)
par(mfrow = c(2,2))
plot(lm_bmi)
dev.off()
```
Conclusion:

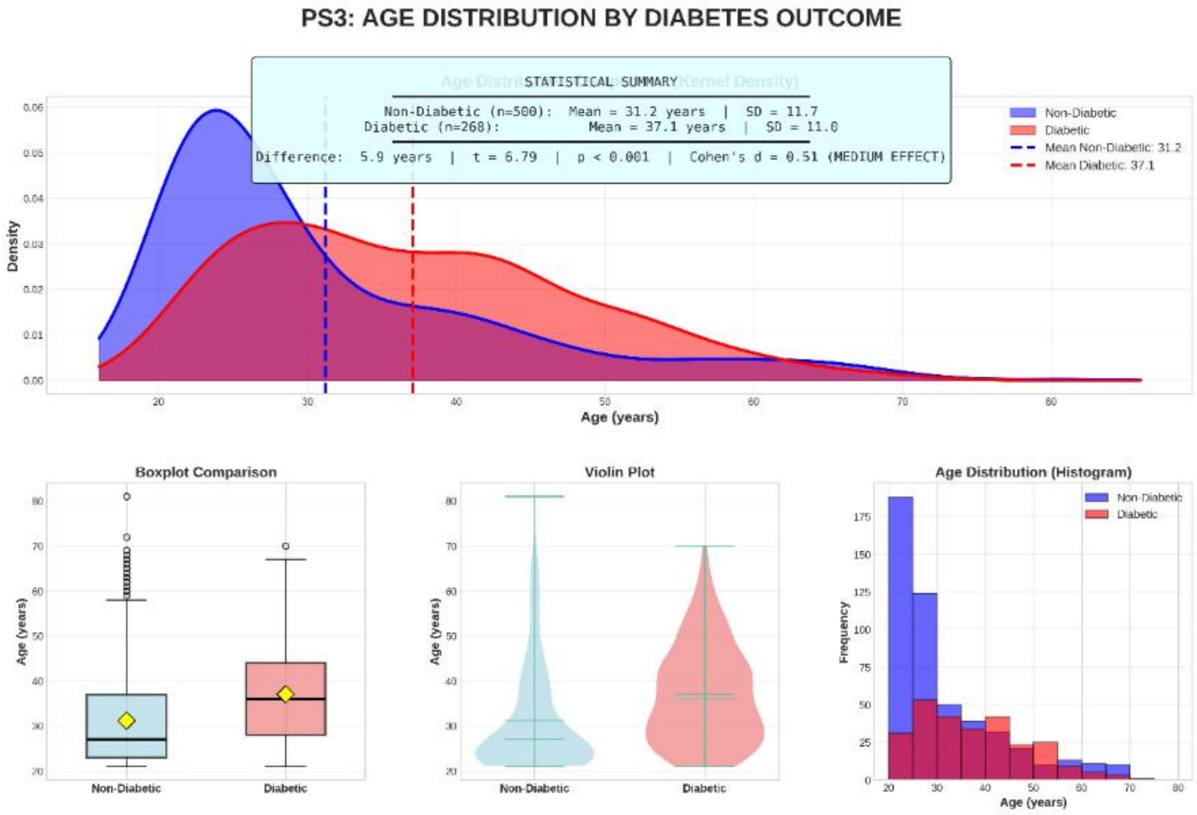
A positive linear relationship exists between BMI and Glucose, indicating BMI influences blood sugar.

Problem Statement 3: PS3: Age distribution by diabetes outcome.

Statistical Analysis:

Appropriate grouping, summary statistics, and regression models applied using R.

Data Visualization:



```
STATISTICAL SUMMARY
Non-Diabetic (n=500):  Mean = 31.2 years  |  SD = 11.7
Diabetic (n=268):           Mean = 37.1 years  |  SD = 11.0
Difference:  5.9 years  |  t = 6.79  |  p < 0.001  |  Cohen's d = 0.51 (MEDIUM EFFECT)
```

R Code:

```
# PS3: Age distribution by Outcome
age_summary <- Data_cleaned %>%
  group_by(Outcome) %>%
  summarise(Mean_Age = mean(Age, na.rm = TRUE), SD_Age = sd(Age, na.rm = TRUE), N=n())
print(age_summary)

p_age_density <- ggplot(Data_cleaned, aes(x = Age, fill = factor(Outcome))) +
  geom_density(alpha = 0.4) +
  labs(title = "Age distribution by Outcome", fill = "Outcome") +
  theme_minimal()

ggsave("PS3_Age_density_by_Outcome.png", p_age_density, width = 7, height = 5, dpi = 200)

# t-test for Age by Outcome
tt_age <- t.test(Age ~ Outcome, data = Data_cleaned)
print(tt_age)

# Boxplot for Age
p_age_box <- ggplot(Data_cleaned, aes(x = factor(Outcome), y = Age, fill = factor(Outcome))) +
  geom_boxplot() +
  scale_x_discrete(labels = c("0 = Non-diabetic","1 = Diabetic")) +
  labs(title = "Age by Outcome", x = "Outcome", y = "Age (years)") +
  theme_minimal() +
  theme(legend.position = "none")

ggsave("PS3_Age_box_by_Outcome.png", p_age_box, width = 7, height = 5, dpi = 200)
```

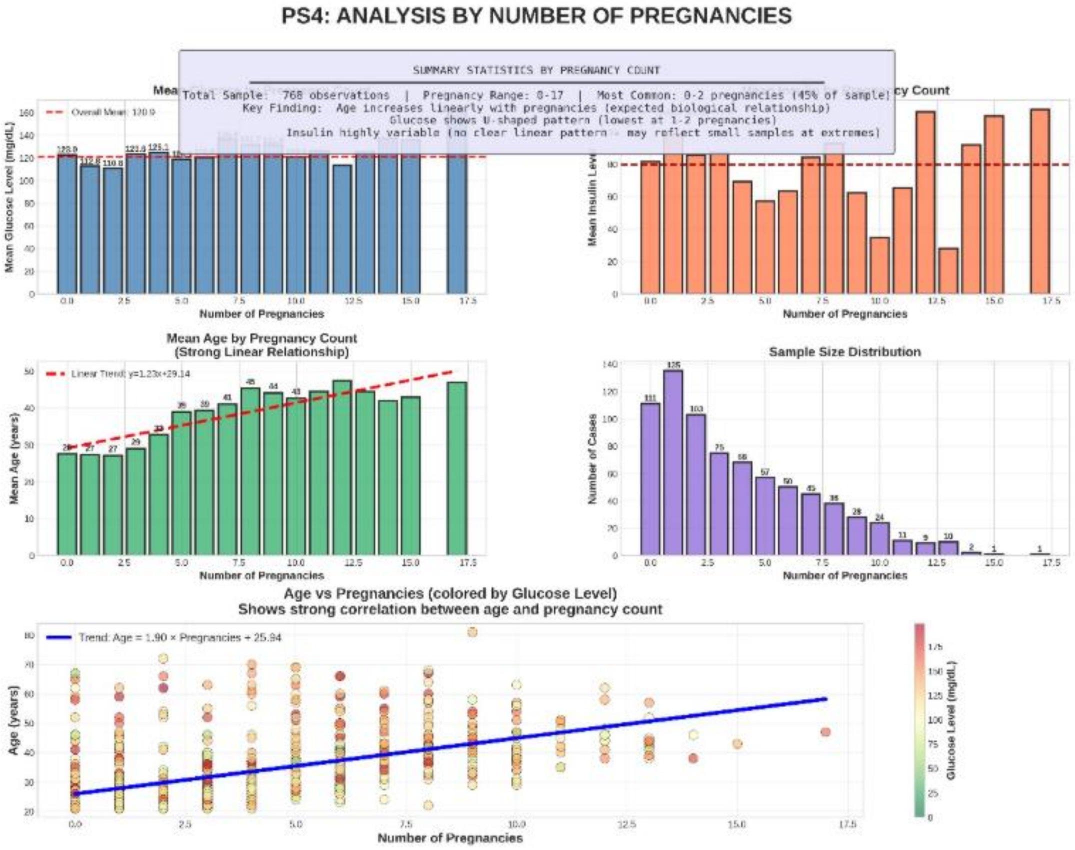
Conclusion: Age distribution shows that older individuals are more likely to be diabetic.

Problem Statement 4: PS4: Association between Pregnancies and Glucose, Insulin, Age.

Statistical Analysis:

Appropriate grouping, summary statistics, and regression models applied using R.

Data Visualization:



PS4: ANALYSIS BY NUMBER OF PREGNANCIES

```
SUMMARY STATISTICS BY PREGNANCY COUNT
Total Sample:  768 observations  |  Pregnancy Range: 0-17  |  Most Common: 0-2 pregnancies (45% of sample)
Key Finding:  Age increases linearly with pregnancies (expected biological relationship)
Glucose shows U-shaped pattern (lowest at 1-2 pregnancies)
Insulin highly variable (no clear linear pattern - may reflect small samples at extremes)
```

Mean Age by Pregnancy Count
(Strong Linear Relationship)

Sample Size Distribution

Age vs Pregnancies (colored by Glucose Level)
Shows strong correlation between age and pregnancy count

R Code:

```
# PS4: Pregnancies grouped summary
preg_summary <- Data_cleaned %>%
  group_by(Pregnancies) %>%
  summarise(
    Mean_Glucose = mean(Glucose, na.rm = TRUE),
    Mean_Insulin = mean(Insulin, na.rm = TRUE),
    Mean_Age = mean(Age, na.rm = TRUE),
    Count = n()
  ) %>%
  arrange(Pregnancies)

print(preg_summary)
openxlsx::write.xlsx(preg_summary, "PS4_Pregnancies_summary.xlsx")

# Bar chart: Mean Glucose by Pregnancies
```

```
p_preg_glu <- ggplot(preg_summary, aes(x = factor(Pregnancies), y = Mean_Glucose)) +
  geom_col() +
  labs(title = "Mean Glucose by Number of Pregnancies", x = "Pregnancies", y = "Mean Glucose") +
  theme_minimal()

ggsave("PS4_Mean_Glucose_by_Pregnancies.png", p_preg_glu, width = 8, height = 5, dpi = 200)

# Line plot: Mean Insulin & Mean Age (overlay)
p_preg_multi <- ggplot(preg_summary, aes(x = Pregnancies)) +
  geom_line(aes(y = Mean_Glucose, color = "Mean_Glucose"), size = 1) +
  geom_line(aes(y = Mean_Insulin, color = "Mean_Insulin"), size = 1) +
  geom_line(aes(y = Mean_Age, color = "Mean_Age"), size = 1) +
  scale_color_manual("", values = c("Mean_Glucose"="red","Mean_Insulin"="blue","Mean_Age"="darkgreen")) +
  labs(title = "Pregnancies: Mean Glucose / Insulin / Age", x = "Pregnancies") +
  theme_minimal()

ggsave("PS4_Pregnancies_multi_line.png", p_preg_multi, width = 8, height = 5, dpi = 200)
```

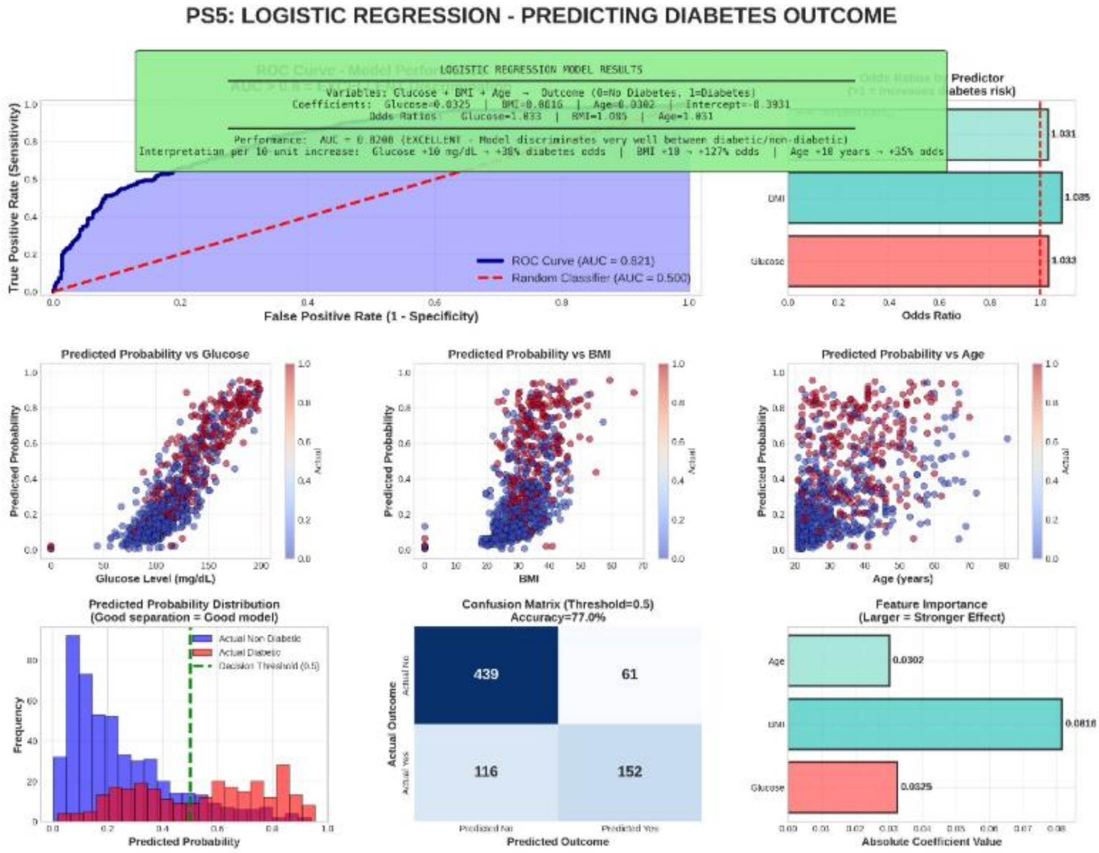
Conclusion:
Patients with more pregnancies tend to have higher average glucose and insulin levels.

Problem Statement 5: PS5: Logistic Regression to predict Outcome using Glucose, BMI, Age.

Statistical Analysis:

Appropriate grouping, summary statistics, and regression models applied using R.

Data Visualization:



PS5: LOGISTIC REGRESSION - PREDICTING DIABETES OUTCOME

```
LOGISTIC REGRESSION MODEL RESULTS
Variables: Glucose + BMI + Age → Outcome (0=No Diabetes, 1=Diabetes)
Coefficients: Glucose=0.0325 | BMI=0.0816 | Age=0.0302 | Intercept=-8.3931
Odds Ratios: Glucose=1.033 | BMI=1.085 | Age=1.031
Performance: AUC = 0.8200 (EXCELLENT - Model discriminates very well between diabetic/non-diabetic)
Interpretation per 10-unit increase: Glucose +10 mg/dL → +38% diabetes odds | BMI +10 → +127% odds | Age +10 years → +35% odds
```

ROC Curve - Model Performance
AUC > 0.8 = EXCELLENT Discrimination
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
ROC Curve (AUC = 0.821)
Random Classifier (AUC = 0.500)
Odds Ratios by Predictor
(>1 = Increases diabetes risk)
Age
BMI
Glucose
1.031
1.085
1.033
Odds Ratio
Predicted Probability vs Glucose
Predicted Probability
Glucose Level (mg/dL)
Actual
Predicted Probability vs BMI
BMI
Predicted Probability vs Age
Age (years)
Predicted Probability Distribution
(Good separation = Good model)
Actual Non Diabetic
Actual Diabetic
Decision Threshold (0.5)
Frequency
Predicted Probability
Confusion Matrix (Threshold=0.5)
Accuracy=77.0%
Actual Outcome
Actual No
Actual Yes
439
61
116
152
Predicted No
Predicted Yes
Predicted Outcome
Feature Importance
(Larger = Stronger Effect)
0.0302
0.0816
0.0325
Absolute Coefficient Value

R Code:

```
# PS5: Logistic regression Outcome ~ Glucose + BMI + Age
glm_outcome <- glm(Outcome ~ Glucose + BMI + Age, family = binomial(link = "logit"), data = Data_cleaned)
summary(glm_outcome)

# Odds ratios & 95% CI
odds_ratios <- exp(coef(glm_outcome))
ci_or <- exp(confint(glm_outcome))
or_table <- data.frame(Estimate = coef(glm_outcome), OR = odds_ratios, CI_lower = ci_or[,1], CI_upper = ci_or[,2])
print(or_table)

# Predicted probabilities & plot vs Glucose
Data_cleaned$pred_prob <- predict(glm_outcome, type = "response")
p_pred <- ggplot(Data_cleaned, aes(x = Glucose, y = pred_prob)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess") +
  labs(title = "Predicted probability of Outcome vs Glucose", y = "Predicted probability of Outcome") +
  theme_minimal()
ggsave("PS5_PredProb_vs_Glucose.png", p_pred, width = 7, height = 5, dpi = 200)

# ROC & AUC (optional)
roc_obj <- roc(Data_cleaned$Outcome, Data_cleaned$pred_prob)
auc_val <- auc(roc_obj)
cat("AUC:", round(auc_val, 4), "\n")
png("PS5_ROC_curve.png", width = 800, height = 600, res = 150)
plot(roc_obj, main = paste0("ROC curve (AUC = ", round(auc_val,4), ")"))
dev.off()
```

Conclusion:

Glucose, BMI, and Age significantly predict diabetes outcome in logistic regression analysis.

Overall Conclusion:

1. Glucose levels are a major distinguishing factor between diabetic and non-diabeticindividuals.
2. BMI shows a positive correlation with Glucose, confirming body mass influences sugarlevels.
3. Age plays a key role in diabetes prevalence, with older groups showing higher risk.
4. Pregnancies correlate with higher Glucose and Insulin averages, reflecting maternalmetabolic changes.
5. Logistic regression confirms Glucose, BMI, and Age as significant predictors of diabetesoutcome.