EXAM 1

RAJ SHAH

2025-03-05

```r
# Load Required Libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(moderndive)
```

```
## Warning: package 'moderndive' was built under R version 4.4.3
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.4.2
```

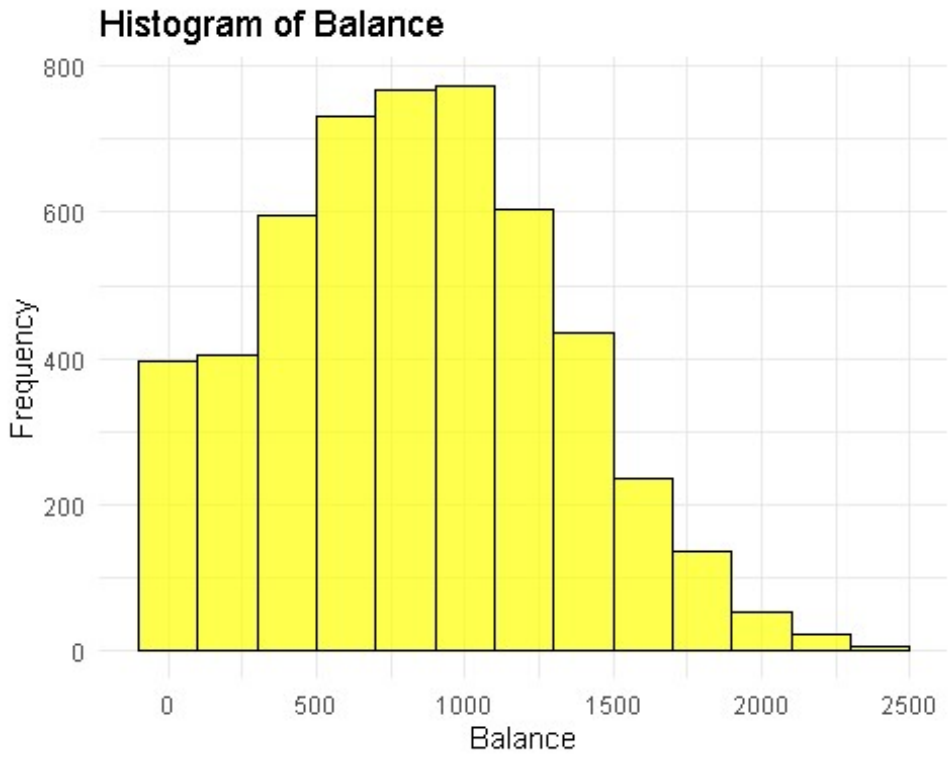
```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

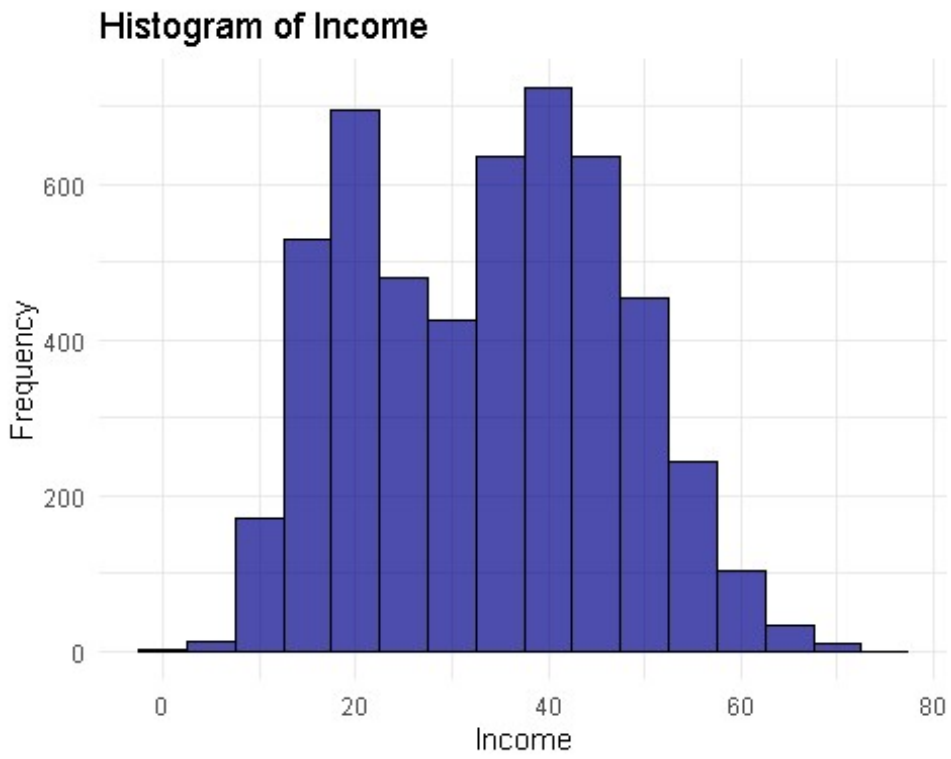
```r
# Load the Data
data <- read.csv("C:/Users/rajsh/OneDrive/Desktop/Inference Data Science
291/EXAM/Default.csv")

# Display first few rows
head(data)
```

```
##   student income balance default
## 1      No 60.204       0       0
## 2      No 39.318     265       0
## 3     Yes 12.147    1947       0
## 4      No 55.574    1914       1
## 5      No 48.837    1281       0
## 6      No 45.552    1447       0
```

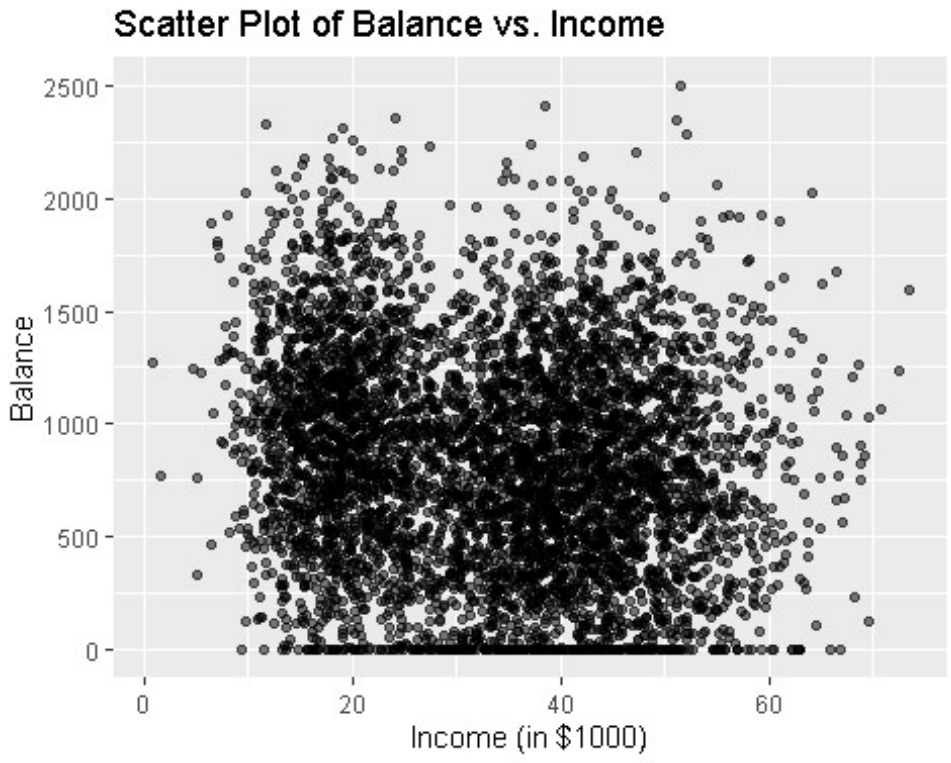
```r
# 1. Histogram of Balance and Income
# Histogram for balance
ggplot(data, aes(x = balance)) +
  geom_histogram(binwidth = 200, fill = "yellow", color = "black", alpha =
0.7) +
  labs(title = "Histogram of Balance", x = "Balance", y = "Frequency") +
  theme_minimal()
```




```r
# Histogram for income
ggplot(data, aes(x = income)) +
  geom_histogram(binwidth = 5, fill = "darkblue", color = "black", alpha =
0.7) +
  labs(title = "Histogram of Income", x = "Income", y = "Frequency") +
  theme_minimal()
```




```
# Explanation:
# The histogram of balance helps visualize its distribution. It may be right-
skewed,
# meaning most balances are low, with some higher balances stretching the
distribution.
# The histogram of income helps assess if income is normally distributed or
skewed.

# 2. Scatter Plot of Balance vs. Income
ggplot(data, aes(x = income, y = balance)) +
  geom_point(alpha = 0.5) +
  labs(title = "Scatter Plot of Balance vs. Income", x = "Income (in $1000)",
y = "Balance")
```

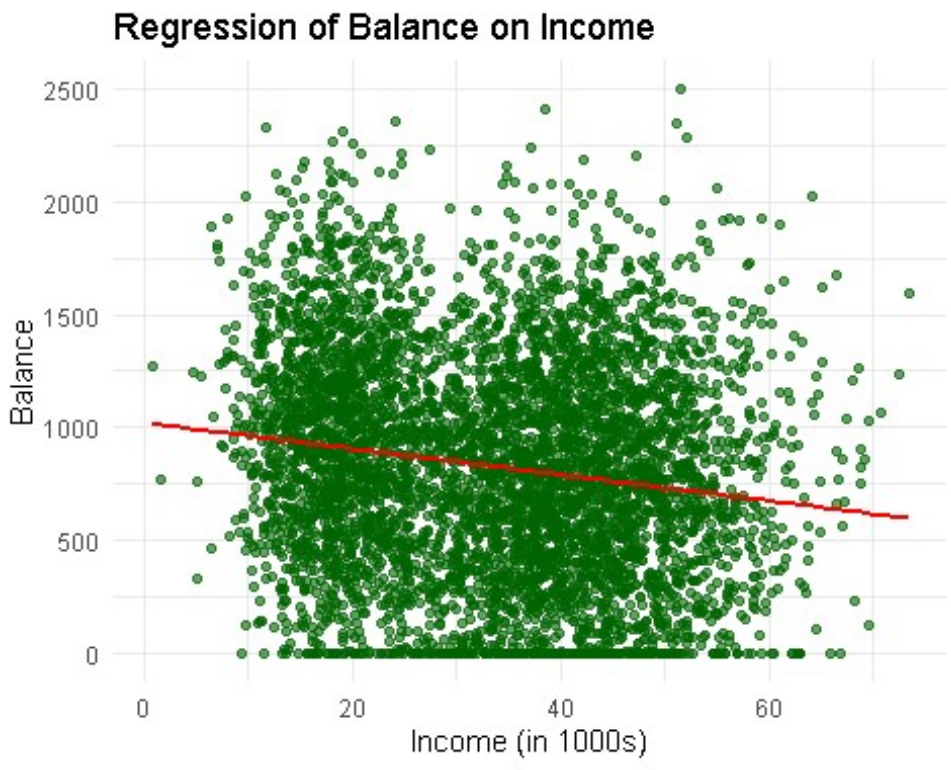



```
# Explanation:
# The scatter plot visually assesses the relationship between balance and
income.
# The regression line indicates a weak negative trend, meaning as income
increases,
# balance slightly decreases.

# 3. Correlation Between Balance and Income
correlation <- cor(data$income, data$balance, use = "complete.obs")
print(paste("Correlation between balance and income:", correlation))

## [1] "Correlation between balance and income: -0.159232694479796"

# Explanation:
# The correlation coefficient is -0.159, which indicates a weak negative
relationship.
# This means that income and balance move in opposite directions, but only
slightly.

# 4. Simple Regression of Balance on Income
model <- lm(balance ~ income, data = data)
summary(model)

##
## Call:
## lm(formula = balance ~ income, data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -967.15 -362.86   -7.45  326.04 1774.51
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1020.7665    17.9691   56.81   <2e-16 ***
## income        -5.7525     0.4967  -11.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 476.1 on 5155 degrees of freedom
## Multiple R-squared:  0.02536,    Adjusted R-squared:  0.02517
## F-statistic: 134.1 on 1 and 5155 DF,  p-value: < 2.2e-16
```

```r
# Scatter plot with regression line
ggplot(data, aes(x = income, y = balance)) +
  geom_point(color = "darkgreen", alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Regression of Balance on Income", x = "Income (in 1000s)", y
= "Balance") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
# Explanation:
# The regression equation: Balance = 1020.77 - 5.75 * Income
```

```r
# The slope (-5.75) means that for every $1000 increase in income, balance
decreases by $5.75.
# The p-value is <2e-16, indicating that the relationship is statistically
significant.
# However, R-squared = 2.54%, meaning income explains only 2.54% of balance
variability.

# 5. Expected Change in Balance for $1000 Increase in Income
slope <- coef(model)[2]
print(paste("Expected change in balance for $1000 increase in income:",
slope))
```

```
## [1] "Expected change in balance for $1000 increase in income: -
5.75250272647959"
```

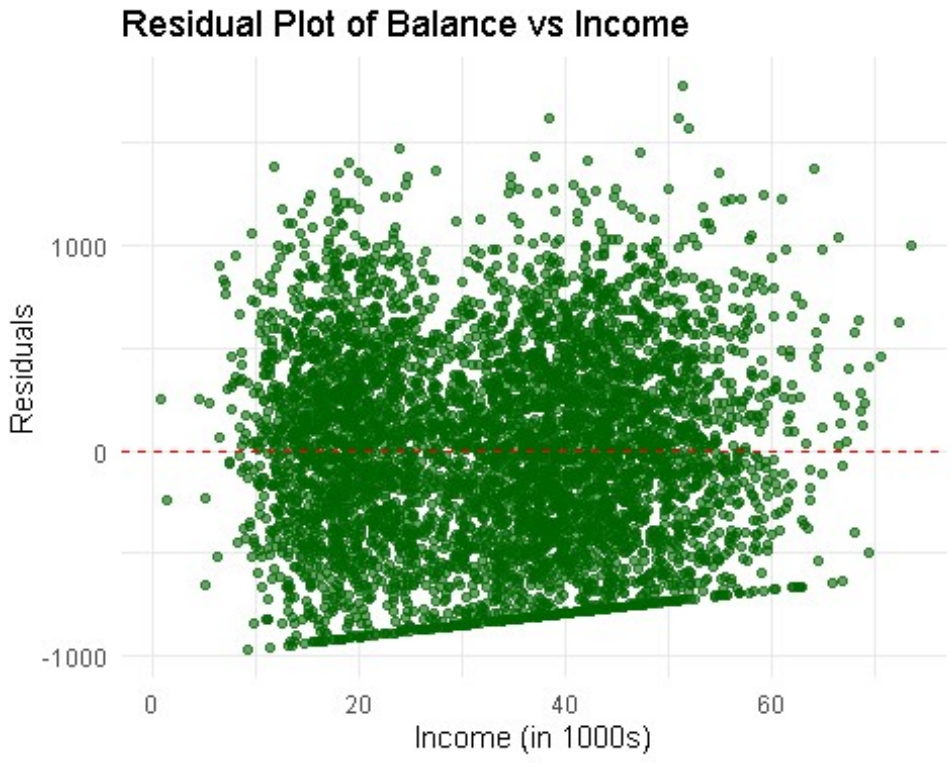
```r
# Explanation:
# The slope = -5.75, meaning that for every $1000 increase in income,
# the expected balance decreases by $5.75.

# 6. Predict Balance for Income = 40K and 80K
pred_40K <- predict(model, data.frame(income = 40))
pred_80K <- predict(model, data.frame(income = 80))
print(paste("Predicted balance for income = 40K:", pred_40K))
```

```
## [1] "Predicted balance for income = 40K: 790.666406304172"
```

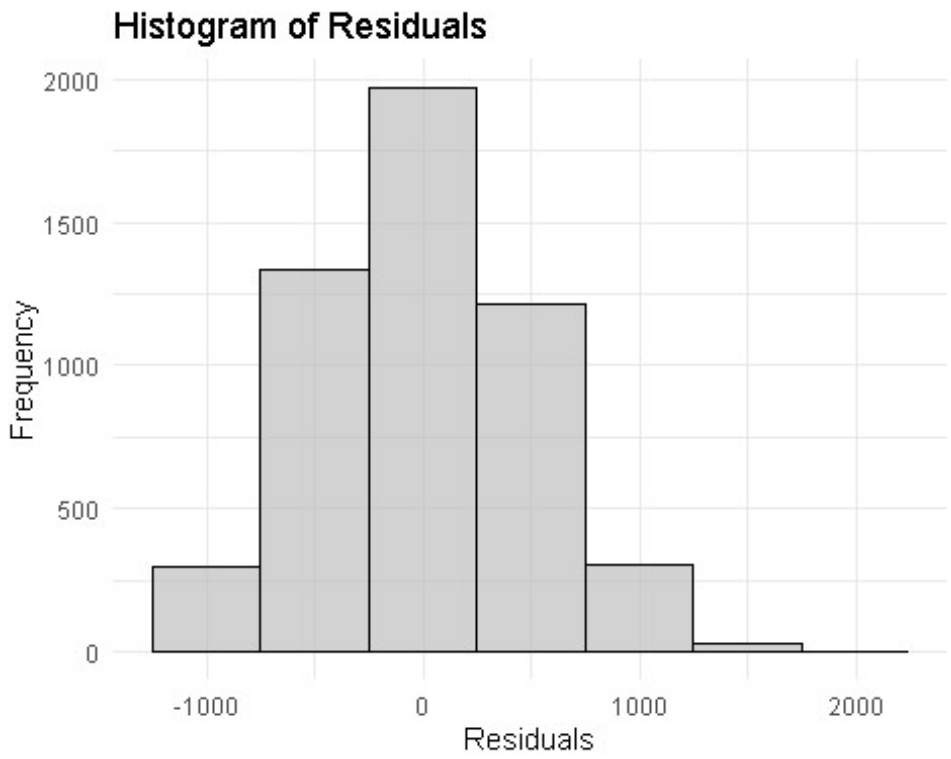
```r
print(paste("Predicted balance for income = 80K:", pred_80K))
```

```
## [1] "Predicted balance for income = 80K: 560.566297244989"
```

```r
# Explanation:
# For income = 40K, predicted balance = 790.67.
# For income = 80K, predicted balance = 560.57.
# Since R-squared is low (2.54%), income is not a strong predictor of
balance.

# 7. Residual Analysis
# Compute residuals
data$residuals <- model$residuals

# Residual plot
ggplot(data, aes(x = income, y = residuals)) +
  geom_point(color = "darkgreen", alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residual Plot of Balance vs Income", x = "Income (in 1000s)",
y = "Residuals") +
  theme_minimal()
```




```r
# Histogram of residuals
ggplot(data, aes(x = residuals)) +
  geom_histogram(binwidth = 500, fill = "grey", color = "black", alpha = 0.7)
+
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency") +
  theme_minimal()
```

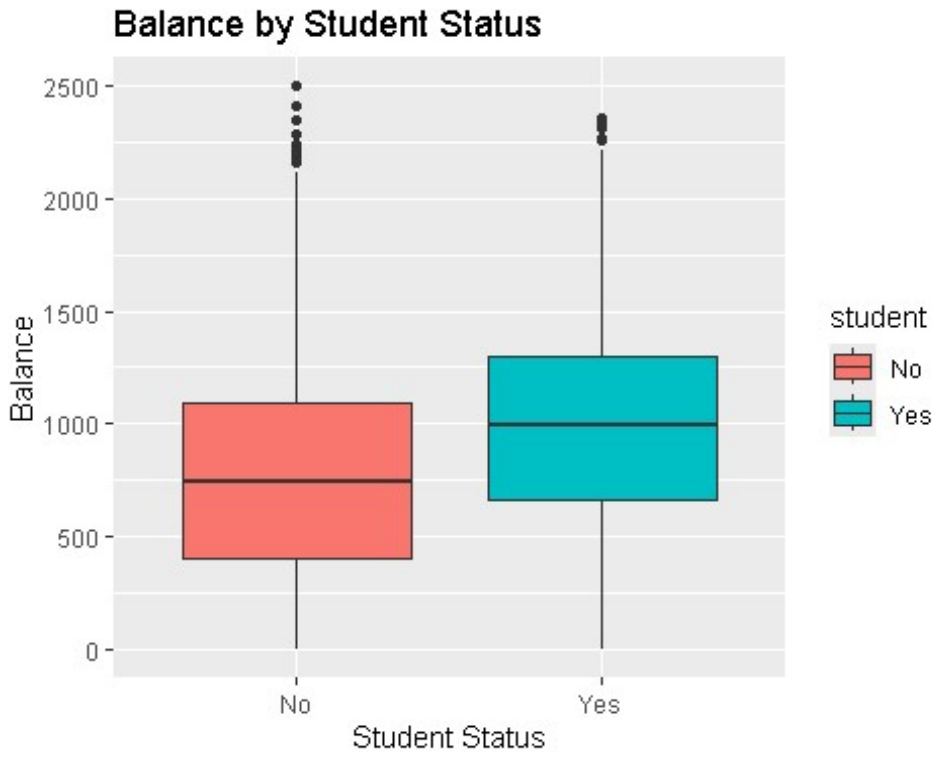



```r
# QQ plot for normality check
ggplot(data, aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "QQ Plot of Residuals") +
  theme_minimal()
```


```r
# Explanation:
# The residual plot helps check for non-linearity or heteroscedasticity.
# A random pattern means assumptions are met; a funnel shape suggests
heteroscedasticity.
# The histogram and QQ plot check if residuals follow a normal distribution.

# 8. Percentage of Variability Explained by the Model
r_squared <- summary(model)$r.squared
print(paste("Percentage of total variability explained by the model:",
r_squared * 100, "%"))
```

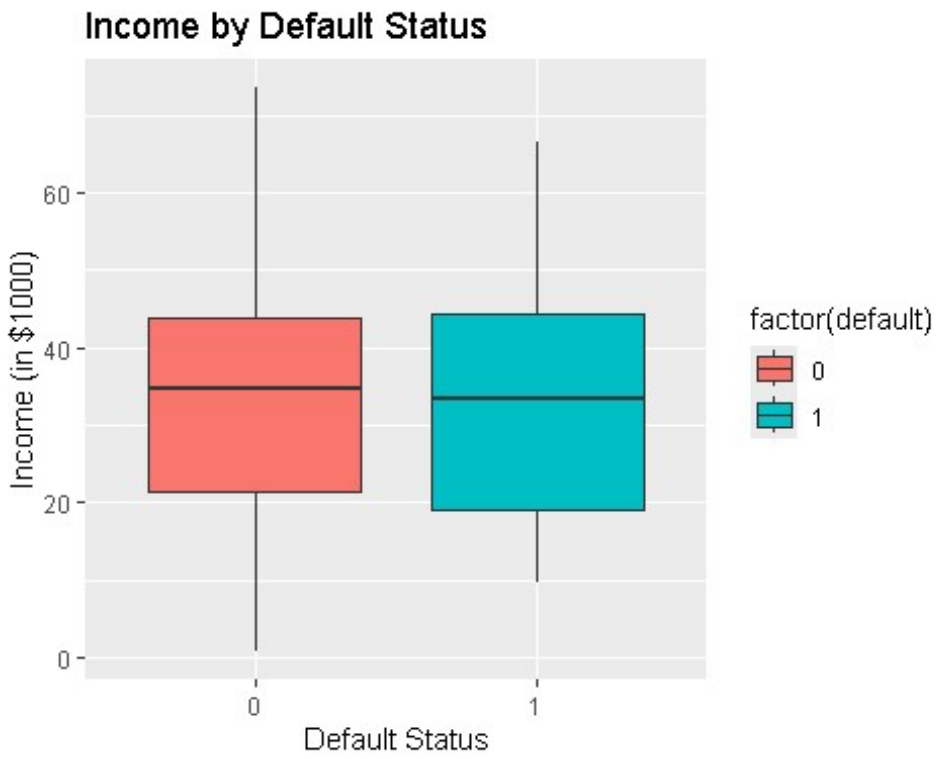
```
## [1] "Percentage of total variability explained by the model:
2.53550509912962 %"
```

```r
# Explanation:
# R-squared = 2.54%, meaning income explains only 2.54% of the variation in
balance.
# This suggests that balance is influenced by other factors.

# 9. Side-by-Side Boxplot of Balance by Student Status
ggplot(data, aes(x = student, y = balance, fill = student)) +
  geom_boxplot() +
  labs(title = "Balance by Student Status", x = "Student Status", y =
"Balance")
```




```
# Explanation:
# The boxplot compares balance distributions for students and non-students.
# If student balances are higher, the median line will be higher for
students.

# 10. Regression of Balance on Student Status
student_model <- lm(balance ~ student, data = data)
summary(student_model)

##
## Call:
## lm(formula = balance ~ student, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -989.54 -351.54  -12.02  320.46 1737.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  761.023      7.784   97.77   <2e-16 ***
## studentYes   228.513     14.448   15.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 470.9 on 5155 degrees of freedom
## Multiple R-squared:  0.04628,    Adjusted R-squared:  0.0461
## F-statistic: 250.2 on 1 and 5155 DF,  p-value: < 2.2e-16
```

```r
# Explanation:
# Regression equation: Balance = 761.02 + 228.51 * StudentYes
# Intercept (761.02) = Average balance for non-students.
# Student coefficient (228.51) = Students have, on average, $228.51 higher
balance.
# P-value (<2e-16) indicates a statistically significant difference.

# 11. Reference Level Change in Regression
data$student <- relevel(factor(data$student), ref = "Yes")
student_model_ref <- lm(balance ~ student, data = data)
summary(student_model_ref)

##
## Call:
## lm(formula = balance ~ student, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -989.54 -351.54  -12.02  320.46 1737.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   989.54      12.17   81.30   <2e-16 ***
## studentNo    -228.51      14.45  -15.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 470.9 on 5155 degrees of freedom
## Multiple R-squared:  0.04628,    Adjusted R-squared:  0.0461
## F-statistic: 250.2 on 1 and 5155 DF,  p-value: < 2.2e-16

# Explanation:
# Swapping the reference level: Balance = 989.54 - 228.51 * StudentNo
# Now, the intercept represents the average balance for students ($989.54).
# The coefficient (-228.51) shows that non-students have lower balances by
$228.51.

# 12. Boxplot of Income by Default Status & Summary Statistics
ggplot(data, aes(x = factor(default), y = income, fill = factor(default))) +
  geom_boxplot() +
  labs(title = "Income by Default Status", x = "Default Status", y = "Income
(in $1000)")
```




```r
summary_stats <- data %>%
  group_by(default) %>%
  summarise(
    count = n(),
    mean_income = mean(income, na.rm = TRUE),
    median_income = median(income, na.rm = TRUE)
  )

print(summary_stats)
```

```
## # A tibble: 2 × 4
##   default count mean_income median_income
##     <int> <int>       <dbl>         <dbl>
## 1       0  4992        33.7          34.7
## 2       1   165        32.7          33.5
```

```r
# Explanation:
# The boxplot compares income distributions for defaulters (default = 1) and
non-defaulters (default = 0).
# If median income is lower for defaulters, this suggests income might be
linked to default risk.
# Summary Statistics:
# Non-defaulters: Mean income = $33.7K, Median income = $34.7K.
# Defaulters: Mean income = $32.7K, Median income = $33.5K.
# The difference is small, suggesting income alone might not be a strong
predictor of default.
```