Exploratory Data Analysis (EDA)

Logistic Regression I

Learning Objectives

1. Identify significant predictor variables.
2. Check for correlations between predictor variables.

ACQUIRE: Credit Default Dataset

- Dataset has 13330 observations with 3 predictor variables and 1 response variable

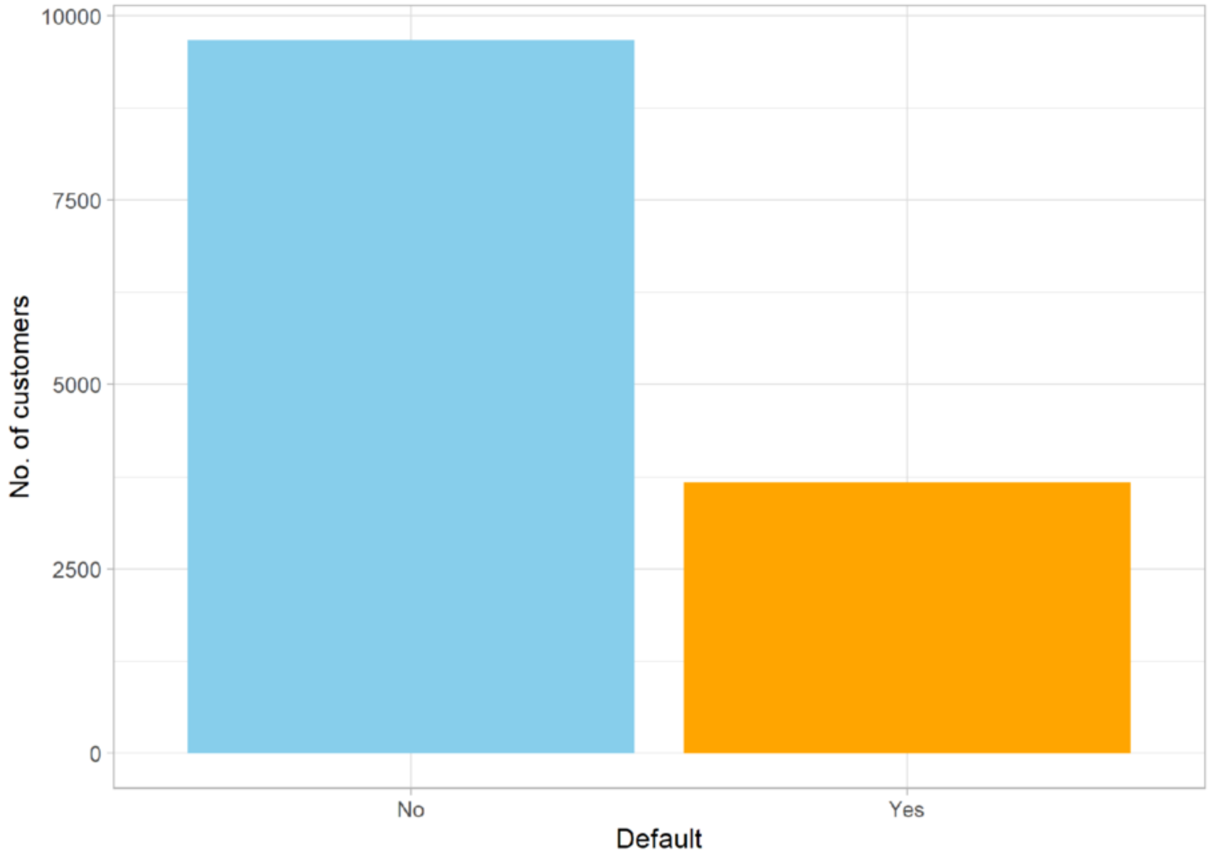
```
str(default_data)
```

```
'data.frame': 13330 obs. of  5 variables:
$ balance  : num  1288 1238 1530 1628 1465 ...
$ income   : num  44253 14863 30004 17547 58700 ...
$ student  : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 2 2 2 1 ...
$ default  : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
$ p_default: num  1 1 1 1 1 1 1 1 1 1 ...
```

ANALYSE: Credit Default Dataset

How many customers default?

- About 30% of the customers default, or the odds of default is $0.3/(1-0.3) = 3/7$.



ANALYSE: Credit Default Dataset

Code

```
ggplot(default_data) +
        geom_bar(aes(x=default, fill=default)) +
        labs(y="No. of customers", x="Default") +
        scale_fill_manual(values = c("skyblue", "orange")) +
        theme_light() +
        theme(legend.position = "none")
```

ANALYSE: Credit Default Dataset

What proportion of customers default?

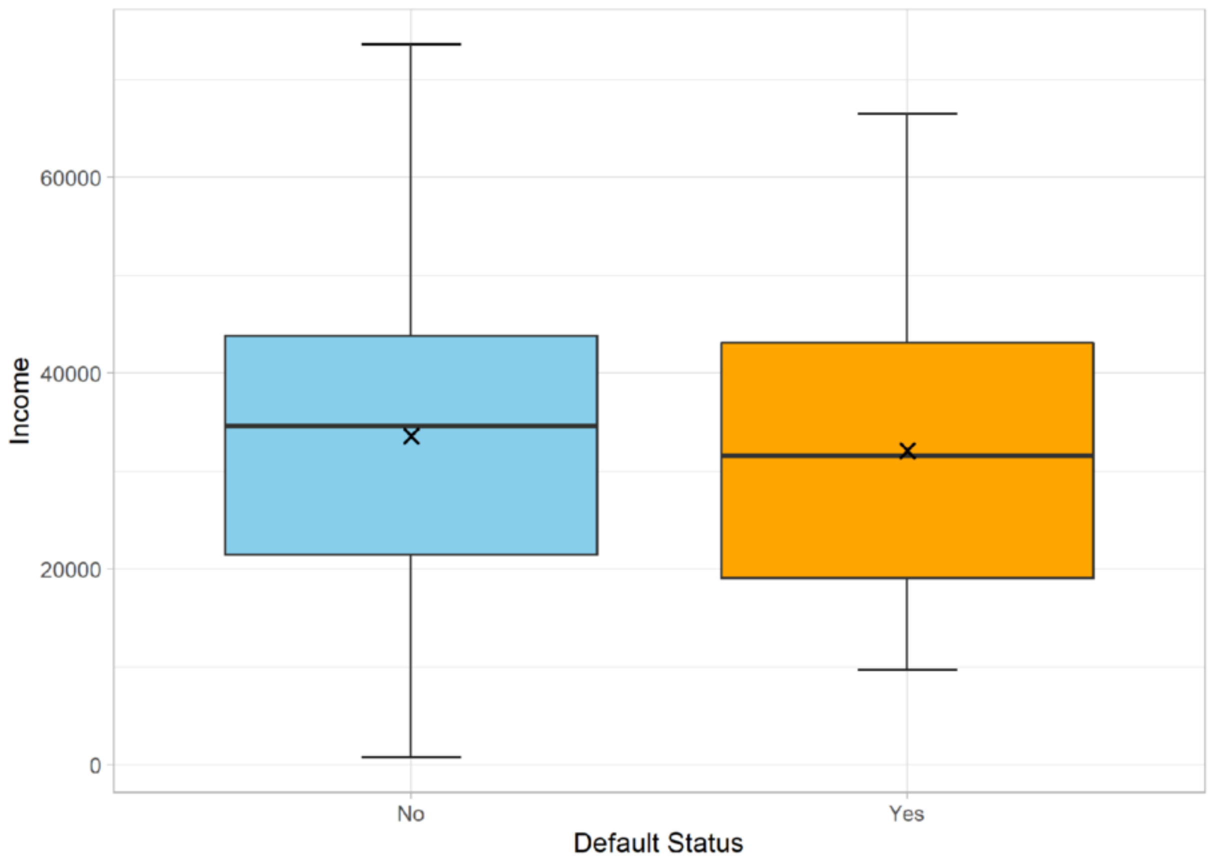
- 27.5% of the customers in the dataset default on their credit card payments.

```
default_data %>%
        group_by(default) %>%
        summarise(prop = n()/nrow(default_data))
```

```
# A tibble: 2 x 2
default  prop
<fct>   <dbl>
1 No      0.725
2 Yes     0.275
```

Predictor Variable: Income

- Distribution of income is similar between defaulters and non-defaulters. So, `income` not likely to be a good predictor of `default`.



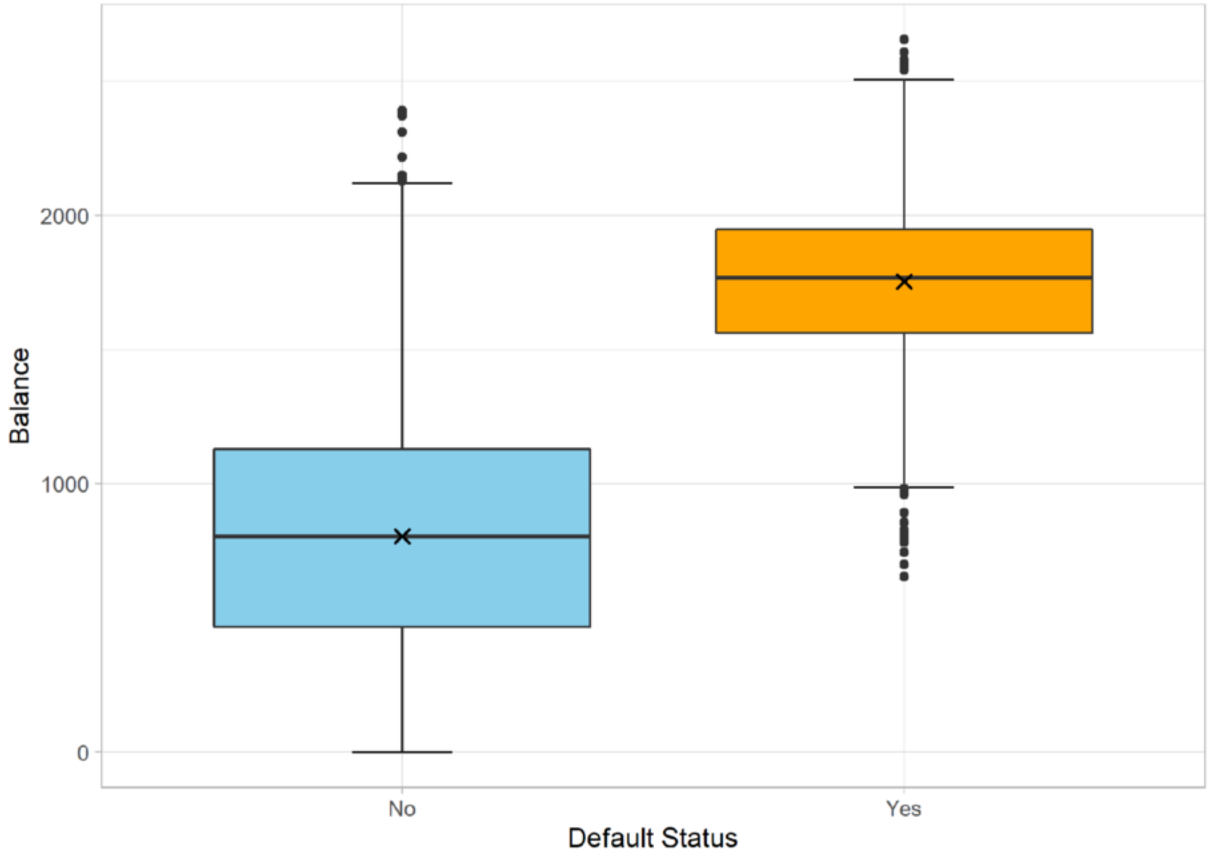
Predictor Variable: Income

Code

```
ggplot(default_data, aes(x=default, y=income)) +
        stat_boxplot(geom = 'errorbar', width = 0.2) +
        geom_boxplot(aes(fill=default)) +
        stat_summary(fun="mean", shape=4) +
        scale_fill_manual(values = c("skyblue", "orange")) +
        theme_light() +
        theme(legend.position = "none")+
        labs(x="Default Status",y="Income")
```

Predictor Variable: Balance

- Defaulters have a higher balance compared to non-defaulters. So, `balance` could be a good predictor of `default`.



Predictor Variable: Balance

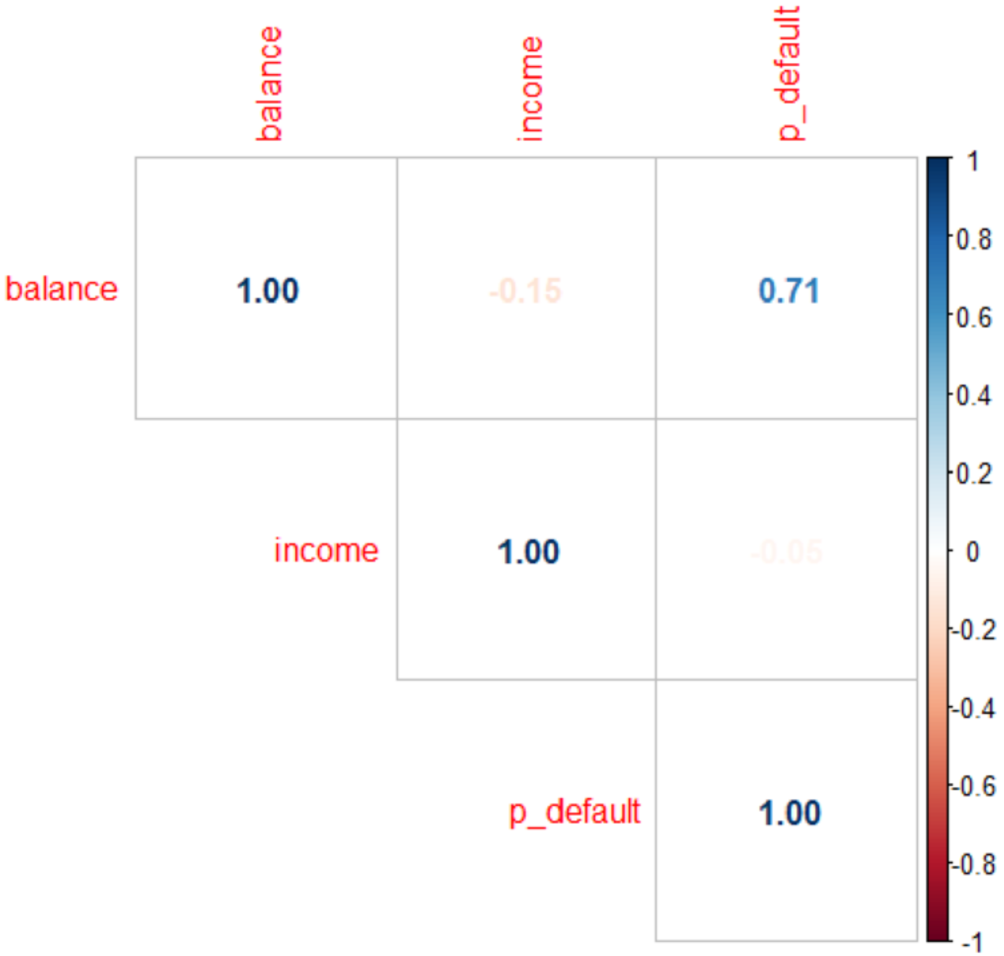
Code

```
ggplot(default_data, aes(x=default,y=balance)) +
        stat_boxplot(geom = 'errorbar', width = 0.2) +
        geom_boxplot(aes(fill=default)) +
        stat_summary(fun="mean", shape=4) +
        scale_fill_manual(values = c("skyblue", "orange")) +
        theme_light() +
        theme(legend.position = "none")+
        labs(x="Default Status", y="Balance")
```

Predictor Variables: Income, Balance

Correlation

- As `default` is categorical response variable, use point biserial correlation in the correlation matrix.
- `balance` has high positive correlation (0.71) with `default`.
- `income` has weak negative correlation (-0.05) with `default`.



Predictor Variables: Income, Balance

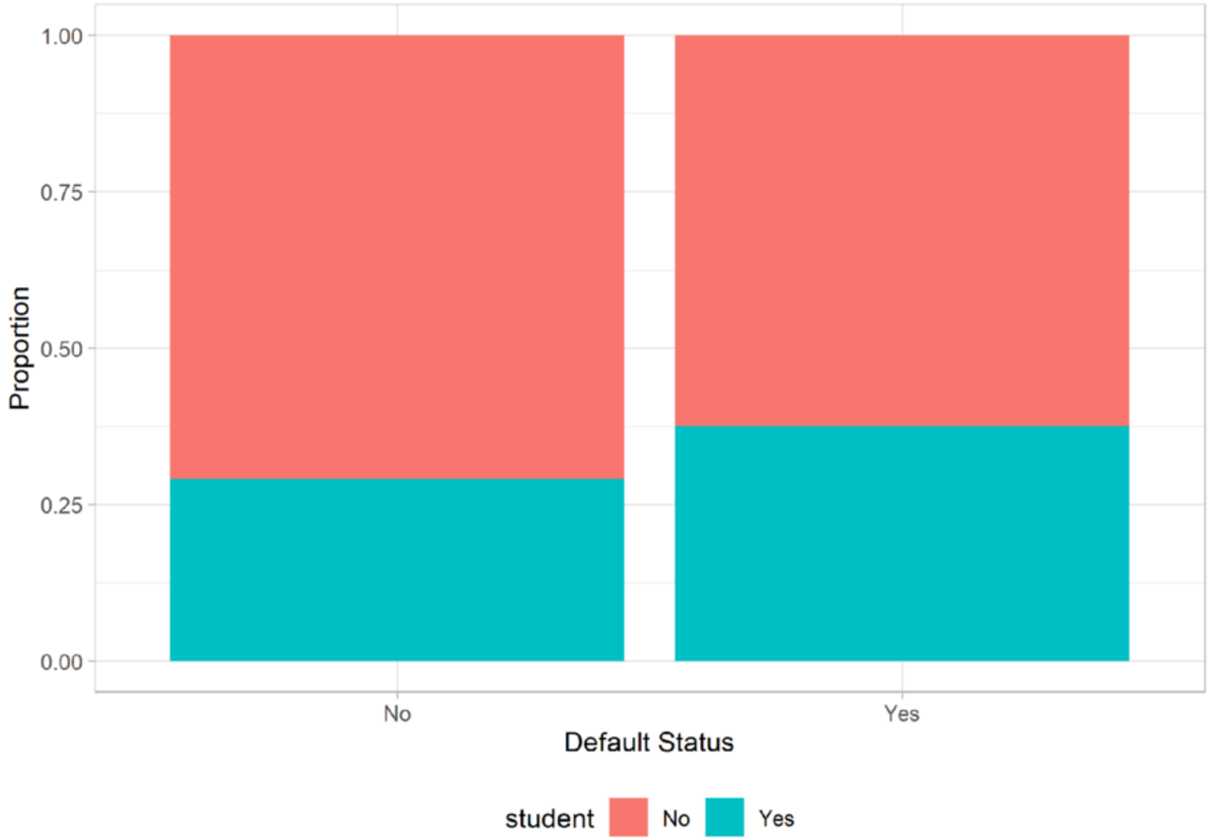
Correlation

```
correlation <- cor(default_data[,c(1:2,5)])
corrplot(corr = correlation, method = 'number', type = 'upper')
correlation
```

```
             balance      income   p_default
balance    1.0000000 -0.14538678  0.71313963
income    -0.1453868  1.00000000 -0.04947877
p_default  0.7131396 -0.04947877  1.00000000
```

Predictor Variable: Student

- Proportion of students among defaulters is higher than among non-defaulters. So, `student` can be a good predictor of `default`.



Predictor Variable: Student

Code

```
#Stacked bar plot: Student and Default
ggplot(default_data) +
        geom_bar(aes(x=default,fill=student), position = "fill") +
        labs(y="Proportion", x="Default Status") +
        theme_light() +
        theme(legend.position = "bottom")
```

Predictor Variable: Student

- Since both `student` and `default` are categorical variables, use a Chi-squared test.
- The p-value is below 0.05, and so there is significant association between `student` and `default`.

```
chisq.test(default_data$student, default_data$default)
```

```
Pearson's Chi-squared test with Yates' continuity correction

data:  default_data$student and default_data$default
X-squared = 86.497, df = 1, p-value < 2.2e-16
```

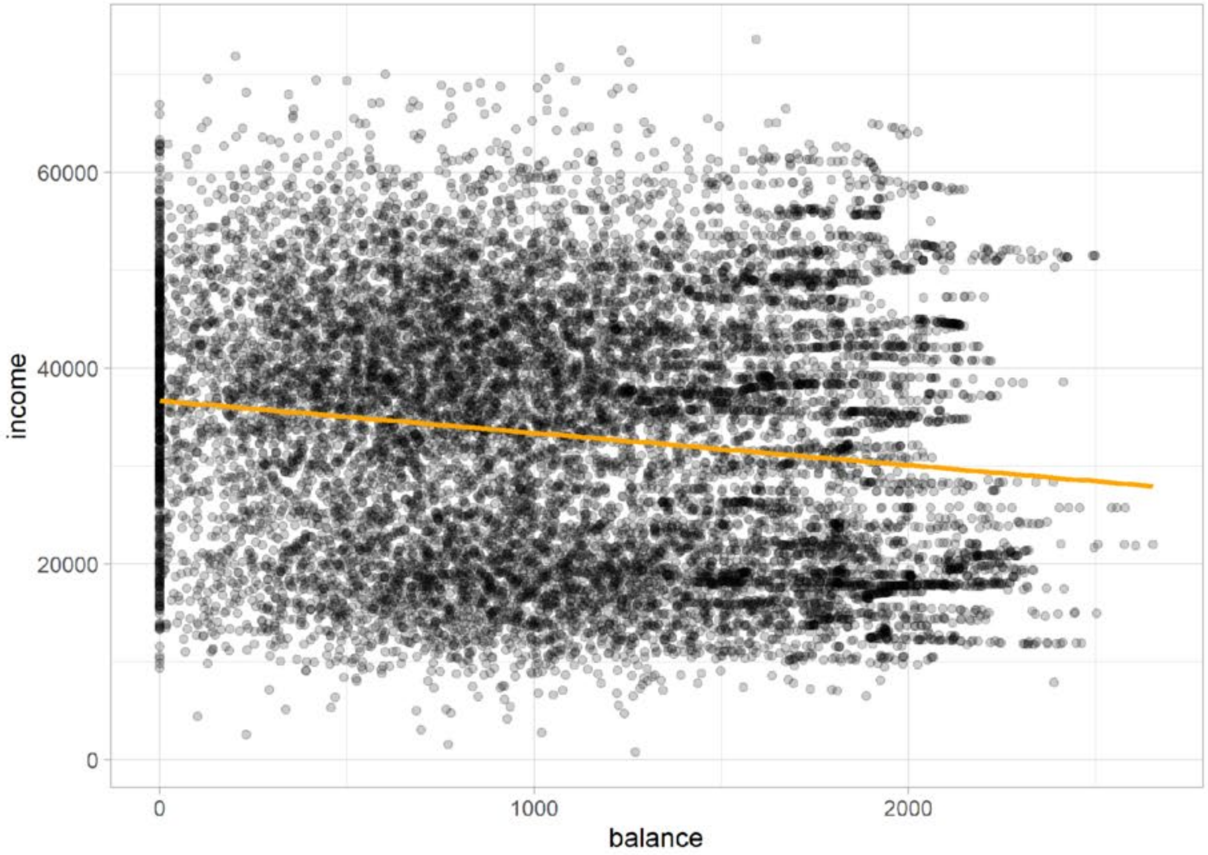
Correlations Between Predictor Variables

Income and Balance

- Pearson's correlation: Weak negative correlation between `income` and `balance`.

```
cor(default_data$balance, default_data$income)
```

```
[1] -0.1453868
```



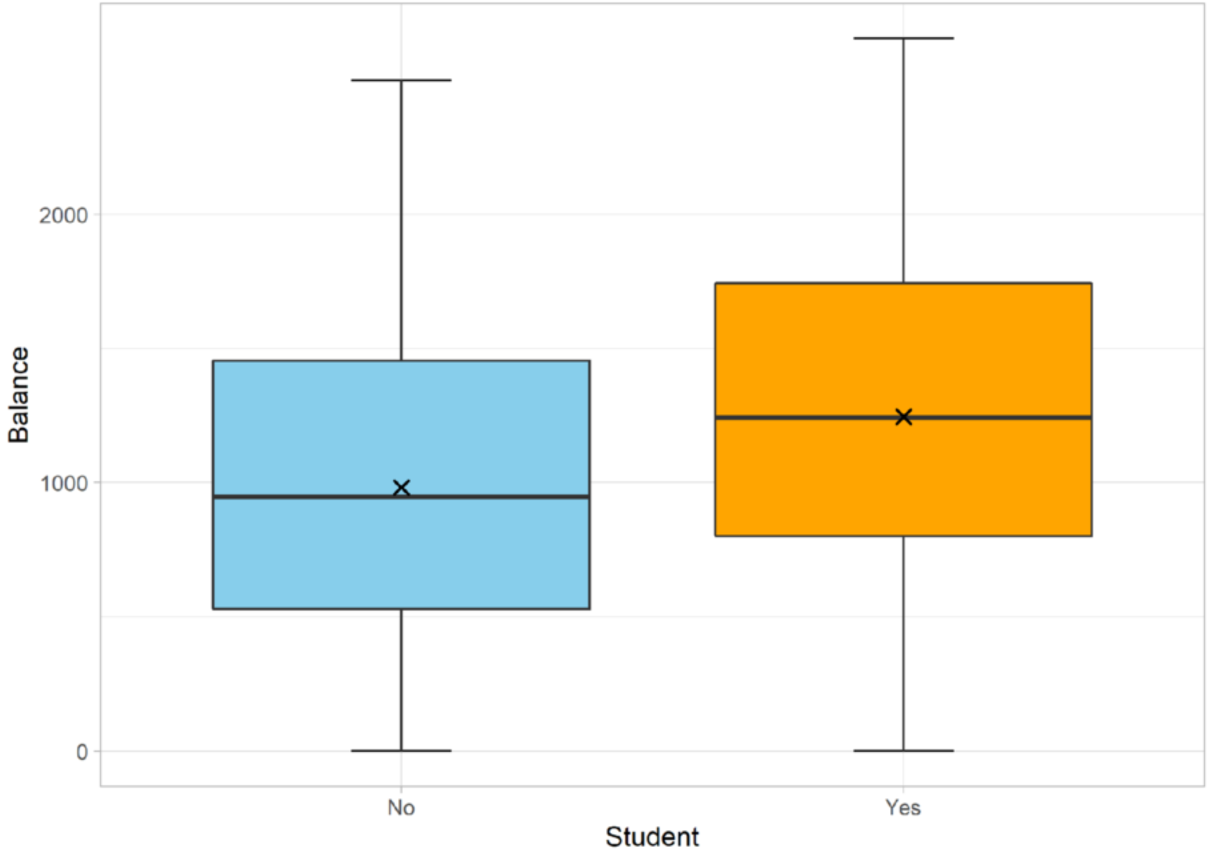
Correlations Between Predictor Variables

Student vs Balance

- Point biserial correlation: Weak positive correlation between `student` and `balance`.

```
cor(default_data$balance, as.numeric(default_data$student))
```

```
[1] 0.2062442
```



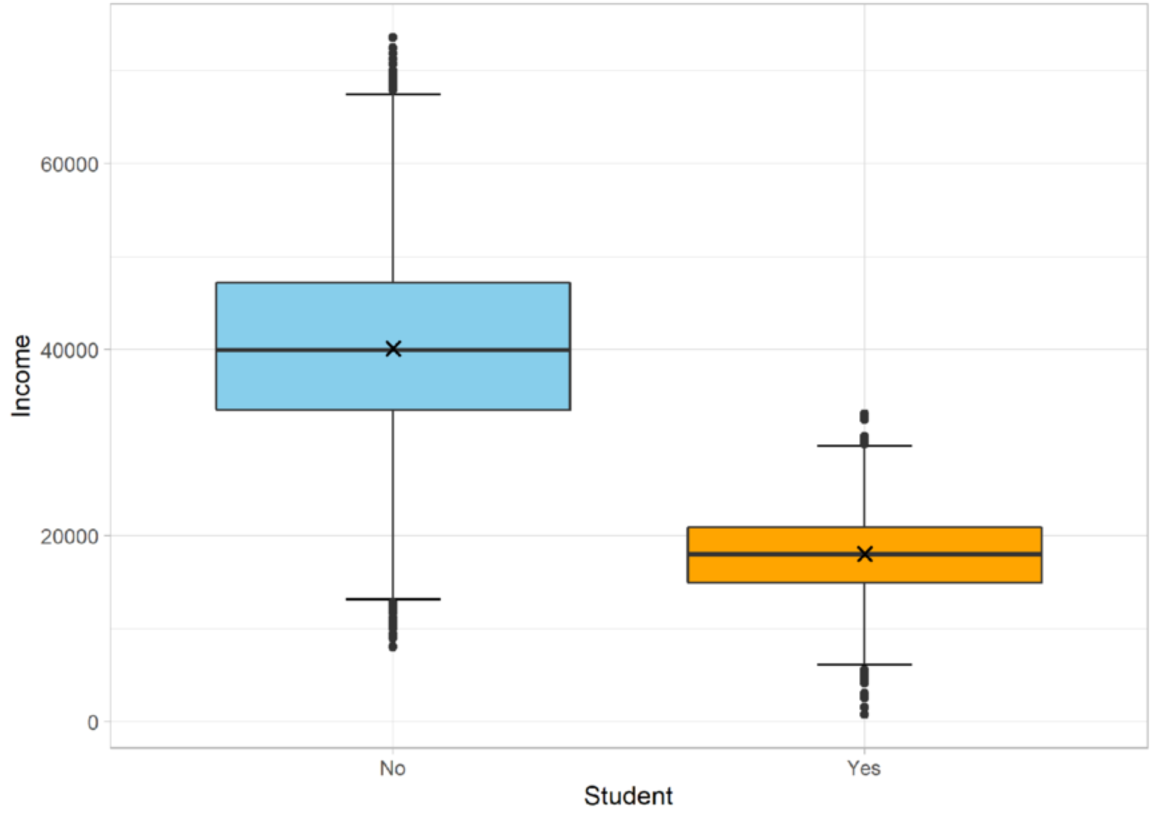
Correlations Between Predictor Variables

Student vs Income

- Point biserial correlation: Strong negative correlation between student and income.

```
cor(default_data$income, as.numeric(default_data$student))
```

```
[1] -0.7624316
```



Summary

- Significant predictors:
  - `student` and `balance` are significant predictors
  - `income` is not a strong predictor
- Correlations between predictors:
  - `balance` has weak correlation with `income` and `student`
  - `income` has strong correlation with `student`