Group_27

Stage 1: Exploratory Data Analysis (EDA)

1. Import data and packages

```
library(dplyr)
library(ggplot2)
library(janitor)
library(car)

# Import dataset
clean_data <- read.csv("C:\\Users\\2980157G\\Downloads\\cleaned_dataset27_1.1.csv")
```

2. Check data structure and summary statistics

```
# Check structure
str(clean_data)
```

```
'data.frame':	1376 obs. of  8 variables:
 $ Age           : int  37 49 20 64 32 58 42 20 42 25 ...
 $ Education     : chr  "Bachelors" "College Education" "College Education" "Bachelors" ...
 $ Marital_Status: chr  "Married" "Divorced/Spouse Absent/Separated/Widowed" "Never-married"
 $ Occupation    : chr  "White-Collar" "Blue-Collar" "Service" "White-Collar" ...
 $ Sex           : chr  "Male" "Male" "Male" "Male" ...
 $ Hours_PW      : int  80 45 45 55 42 40 50 14 40 40 ...
 $ Nationality   : chr  "United-States" "United-States" "United-States" "United-States" ...
 $ Income        : chr  ">50K" "<=50K" "<=50K" ">50K" ...
```

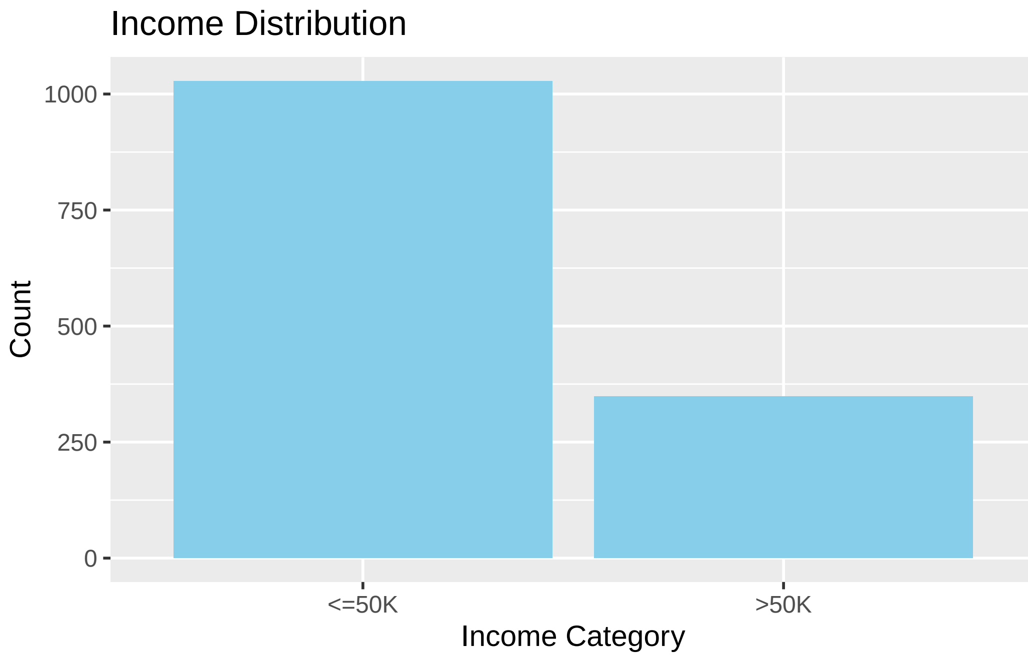
```r
# Summary statistics
summary(clean_data)
```

```
      Age         Education         Marital_Status      Occupation
 Min.   :17.00   Length:1376        Length:1376        Length:1376
 1st Qu.:28.00   Class :character   Class :character   Class :character
 Median :38.00   Mode  :character   Mode  :character   Mode  :character
 Mean   :38.89
 3rd Qu.:47.00
 Max.   :90.00
     Sex               Hours_PW     Nationality           Income
 Length:1376        Min.   : 3.00   Length:1376        Length:1376
 Class :character   1st Qu.:40.00   Class :character   Class :character
 Mode  :character   Median :40.00   Mode  :character   Mode  :character
                    Mean   :41.18
                    3rd Qu.:45.00
                    Max.   :99.00
```

The dataset contains 1376 observations with 8 variables. The data structure confirms that "Age" and "Hours_PW" are numerical, while other variables are categorical. The summary statistics show that the median age is 38, and the average work hours per week is around 41. Some extreme values are present, such as a maximum age of 90 and a maximum work hour of 99, which may require further investigation.

3. Visualize the income distribution

```r
# Bar plot for income distribution
ggplot(clean_data, aes(x = Income)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Income Distribution", x = "Income Category", y = "Count")
```



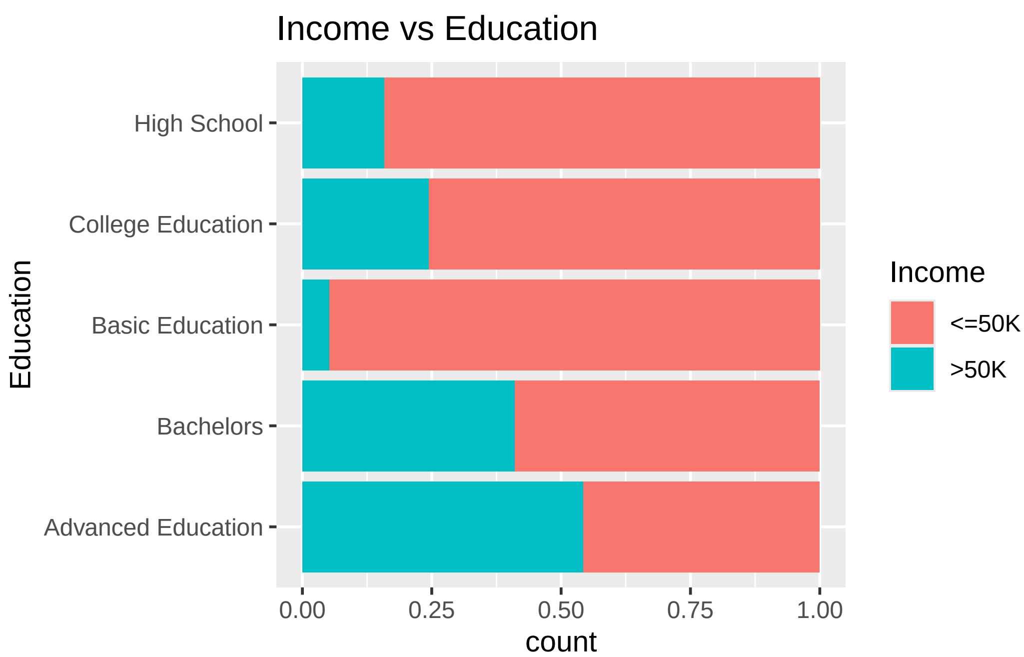
The chart shows that the majority of individuals have an income of $50K or less. This indicates a class imbalance in the target variable, which could influence the performance of classification models.

4. Explore categorical variables' relationship with income

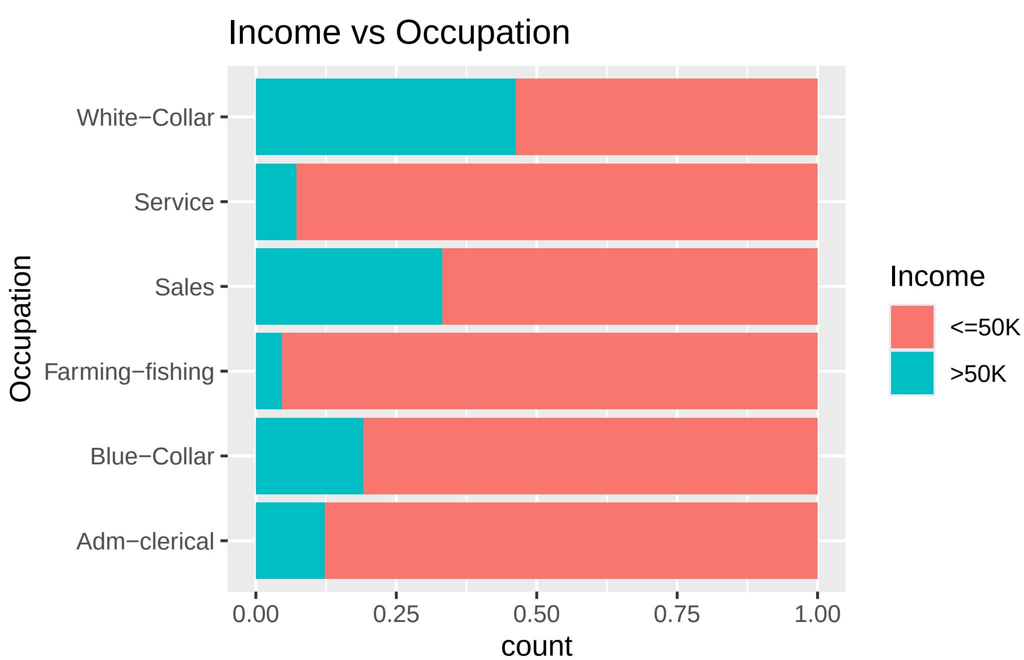
```r
# Stacked bar plots for categorical variables
categorical_vars <- c("Education", "Occupation", "Sex", "Marital_Status")

lapply(categorical_vars, function(var) {
  ggplot(clean_data, aes_string(x = var, fill = "Income")) +
    geom_bar(position = "fill") +
    coord_flip() +
    labs(title = paste("Income vs", var))
})
```

```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.
```

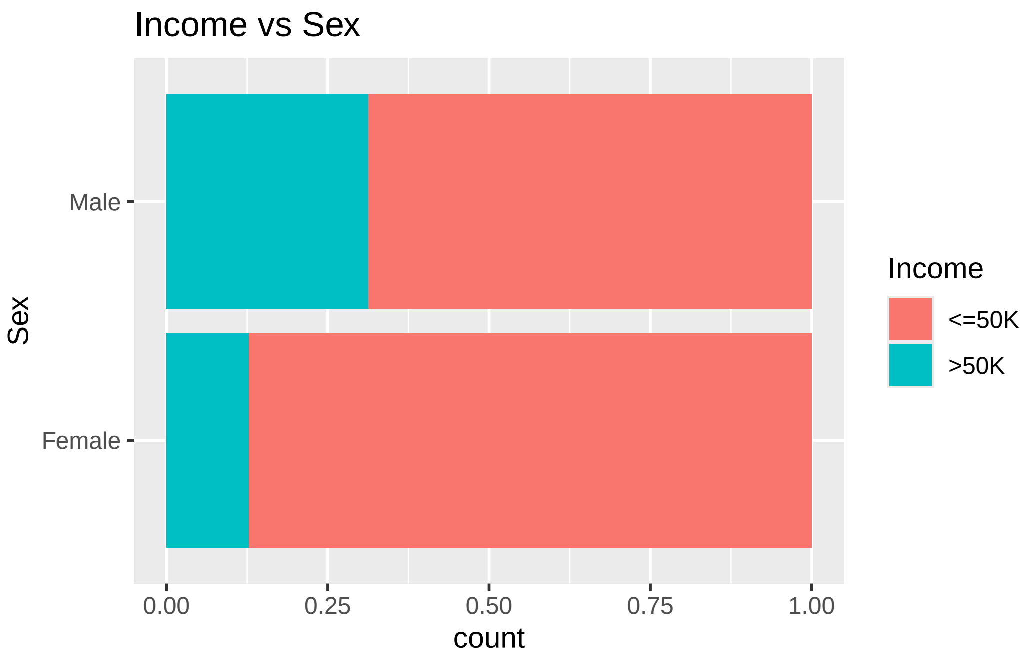
```
[[1]]
```


Income vs Education
High School
College Education
Basic Education
Bachelors
Advanced Education
Education
0.00
0.25
0.50
0.75
1.00
count
Income
<=50K
>50K

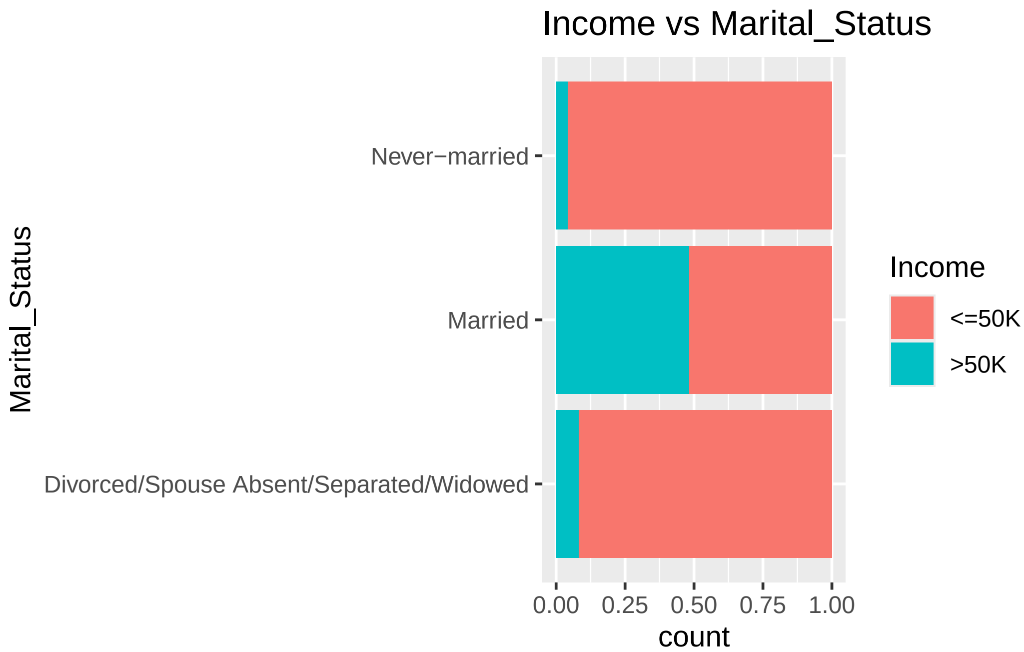
[[2]]


Income vs Occupation
White−Collar
Service
Sales
Farming−fishing
Blue−Collar
Adm−clerical
Occupation
0.00
0.25
0.50
0.75
1.00
count
Income
<=50K
>50K

[[3]]


Income vs Sex
Male
Female
Sex
0.00
0.25
0.50
0.75
1.00
count
Income
<=50K
>50K

[[4]]



- **Income vs. Education**

Higher education levels are associated with a higher proportion of individuals earning more than 50K. Advanced education and bachelor's degree holders have a significantly larger share of high-income earners compared to those with only basic or high school education.

- **Income vs. Occupation**

White-collar and sales jobs have a higher percentage of individuals earning >50K compared to blue-collar, service, and farming occupations. This suggests that occupation type plays a crucial role in income level.

- **Income vs. Sex**

A higher percentage of males earn more than 50K compared to females. The income gap between genders suggests possible structural or occupational differences affecting earnings.

- **Income vs. Marital Status**

Married individuals have a significantly higher proportion of high-income earners compared to those who are never married or divorced/separated/widowed. This could indicate that marital stability is associated with higher earnings, possibly due to dual-income households or other economic advantages.

5. Check category balance

```r
# Check proportion of income categories
prop.table(table(clean_data$Income))
```

```
   <=50K     >50K 
0.747093 0.252907 
```

The dataset is imbalanced, with about 75% of individuals earning 50K and 25% earning >50K. This imbalance may affect model performance and should be considered when building predictive models.

6. Establish Logistic Regression Model

```r
# Convert Income to binary (0 = <=50K, 1 = >50K)
clean_data <- clean_data %>%
  mutate(Income = ifelse(Income == ">50K", 1, 0))


# Logistic regression model
model <- glm(Income ~ Age + Education + Marital_Status + Occupation + Sex + Hours_PW + Natio
             data = clean_data,
             family = binomial(link = "logit"))
summary(model)
```

```

Call:
glm(formula = Income ~ Age + Education + Marital_Status + Occupation + 
    Sex + Hours_PW + Nationality, family = binomial(link = "logit"), 
    data = clean_data)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 -5.090717   0.714610  -7.124 1.05e-12 ***
Age                          0.025900   0.007460   3.472 0.000517 ***
EducationBachelors          -0.437211   0.302472  -1.445 0.148329    
EducationBasic Education    -2.765322   0.630528  -4.386 1.16e-05 ***
```

```
EducationCollege Education  -0.692475   0.305703  -2.265 0.023501 *
EducationHigh School        -1.077803   0.316860  -3.402 0.000670 ***
Marital_StatusMarried        2.750066   0.274474  10.019  < 2e-16 ***
Marital_StatusNever-married -0.268038   0.355729  -0.753 0.451155
OccupationBlue-Collar       -0.040757   0.331201  -0.123 0.902062
OccupationFarming-fishing   -2.770325   0.864437  -3.205 0.001352 **
OccupationSales              0.594309   0.339475   1.751 0.080003 .
OccupationService           -0.390250   0.393379  -0.992 0.321175
OccupationWhite-Collar       1.204002   0.310606   3.876 0.000106 ***
SexMale                     -0.267497   0.233600  -1.145 0.252165
Hours_PW                     0.044293   0.007863   5.633 1.77e-08 ***
NationalityUnited-States    -0.105831   0.301564  -0.351 0.725633
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1556.27  on 1375  degrees of freedom
Residual deviance:  966.68  on 1360  degrees of freedom
AIC: 998.68

Number of Fisher Scoring iterations: 6
```

Individuals working longer hours per week tend to fall into the >$50K income group more frequently. This suggests a positive correlation between hours worked and income level.

7. Check for multicollinearity

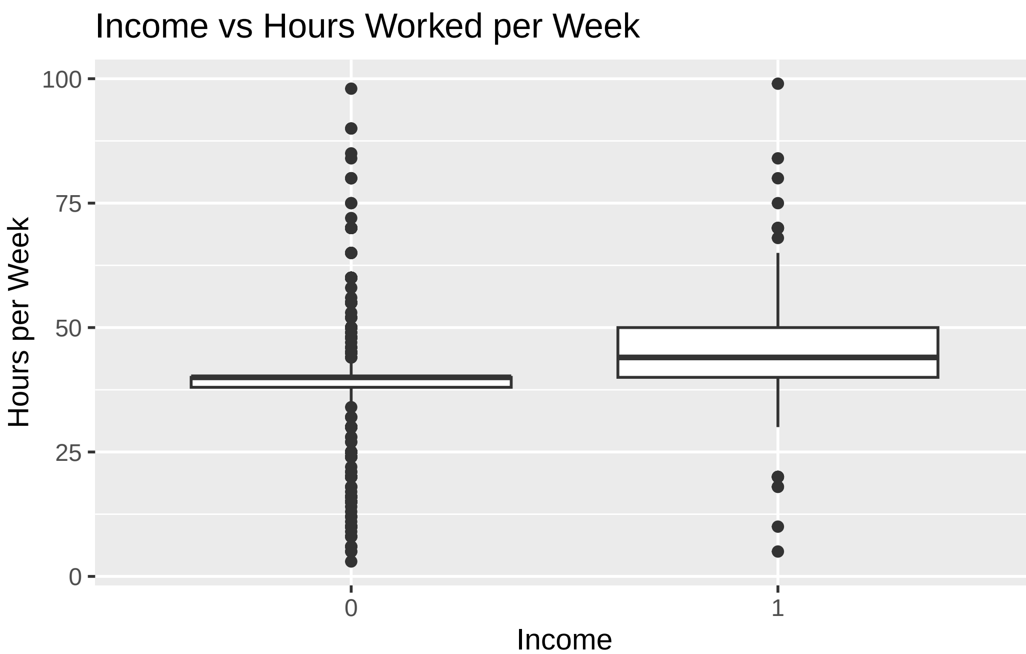
```r
# Variance Inflation Factor (VIF) test
vif(model)
```

```
                   GVIF Df GVIF^(1/(2*Df))
Age            1.142520  1        1.068887
Education      1.620551  4        1.062204
Marital_Status 1.522199  2        1.110754
Occupation     1.834595  5        1.062561
Sex            1.464875  1        1.210320
Hours_PW       1.143694  1        1.069436
Nationality    1.064644  1        1.031816
```

The VIF values are all below 2, indicating that there is no significant multicollinearity among the predictors. This means the variables are not highly correlated, and no immediate adjustments are needed.

8. Visualize numeric variables against income

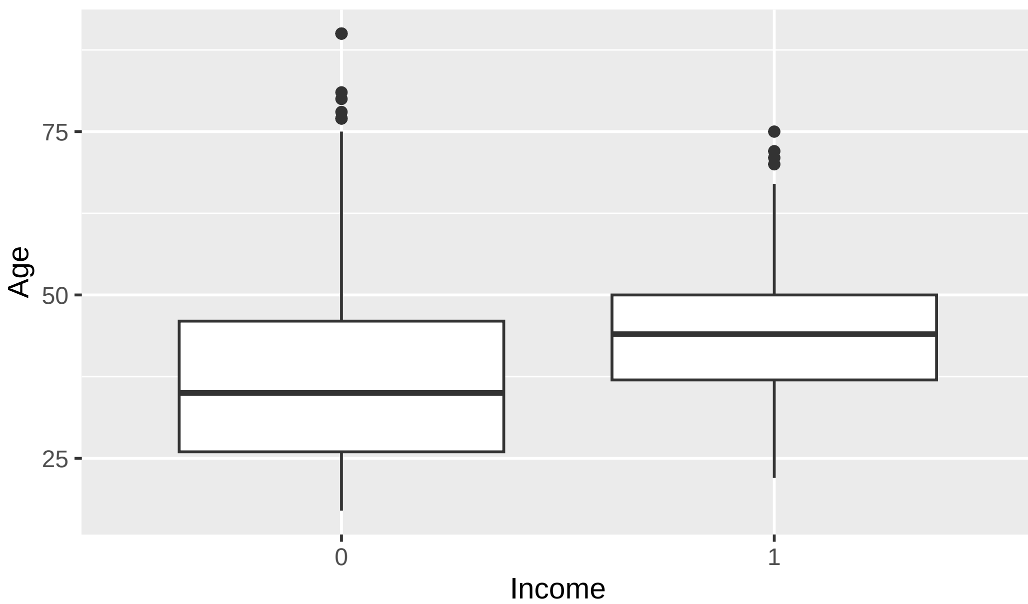
```
# Box plot for Hours per Week
ggplot(clean_data, aes(x = as.factor(Income), y = Hours_PW)) +
  geom_boxplot() +
  labs(title = "Income vs Hours Worked per Week", x = "Income", y = "Hours per Week")
```



Individuals with higher income (>50K) tend to work more hours per week on average. The median work hours are higher for this group, and there are fewer individuals working very few hours compared to the lower-income group. However, there are some outliers working extreme hours in both income groups.

```
# Box plot for Age
ggplot(clean_data, aes(x = as.factor(Income), y = Age)) +
  geom_boxplot() +
  labs(title = "Income vs Age", x = "Income", y = "Age")
```

Income vs Age



Higher-income individuals tend to be older on average. The median age of the >50K income group is higher than that of the 50K group. The distribution also shows a wider age range among high earners, though both groups have some outliers at older ages.