PROJECT REPORT

American International University-Bangladesh

Course name – Introduction to Data science

Faculty – kamrun Naher koli

SECTION -E

GROUP-I

GROUP MEMBER:

Name	Id
Md. Easin Khandokar	22-46515-1
Provakar Sagor	22-46639-1
Sadia Afroz Shuprova	22-46462-1

Dataset

The dataset explores factors affecting an individual's annual income, influenced by attributes like education, age, gender, and occupation. It contains 16 columns, with 14 features describing demographics and personal information, and a target variable income categorized as <=50K or >50K. This widely used KNN dataset serves as an excellent example for data preprocessing and machine learning practice. Analysis can reveal patterns, such as older and more educated individuals tending to earn higher incomes. It provides a foundation for building predictive models to estimate income levels based on personal attributes.

Dataset link: https://www.kaggle.com/datasets/wenruliu/adult-income-dataset

Import library

Code

```
library(dplyr)

library(ggplot2)

library(caret)

library(reshape2)

library(Amelia)
```

This code loads important R libraries to work with data and model it, dplyr, ggplot2, and reshape2 help to manipulate data and visualize it, and caret is employed to work with machine-learning processes. Amelia allows management and input of missing data.

Import Dataset

code

df <- read.csv ("C:/Users/asus/OneDrive/Desktop/adult_income/adult.csv", stringsAsFactors = FALSE)

print(head(df))

```
> print(head(df))
  age workclass fnlwgt    education educational.num
1  25   Private 226802         11th               7
2  38   Private  89814      HS-grad               9
3  28 Local-gov 336951   Assoc-acdm              12
4  44   Private 160323 Some-college              10
5  18         ? 103497 Some-college              10
6  34   Private 198693         10th               6
      marital.status        occupation  relationship  race
1      Never-married Machine-op-inspct     Own-child Black
2 Married-civ-spouse   Farming-fishing       Husband White
3 Married-civ-spouse   Protective-serv       Husband White
4 Married-civ-spouse Machine-op-inspct       Husband Black
5      Never-married                 ?     Own-child White
6      Never-married     Other-service Not-in-family White
  gender capital.gain capital.loss hours.per.week
1   Male            0            0             40
2   Male            0            0             50
3   Male            0            0             40
4   Male         7688            0             40
5 Female            0            0             30
6   Male            0            0             30
  native.country income
1  United-States  <=50K
2  United-States  <=50K
3  United-States   >50K
4  United-States   >50K
5  United-States  <=50K
6  United-States  <=50K
```

print(str(df))

```
1  United-States  <=50K
2  United-States  <=50K
3  United-States   >50K
4  United-States   >50K
5  United-States  <=50K
6  United-States  <=50K
> print(str(df))
'data.frame':	48842 obs. of  15 variables:
 $ age            : int  25 38 28 44 18 34 29 63 24 55 ...
 $ workclass      : chr  "Private" "Private" "Local-gov" "Private" ...
 $ fnlwgt         : int  226802 89814 336951 160323 103497 198693 227026 104626 369667 104996 ...
 $ education      : chr  "11th" "HS-grad" "Assoc-acdm" "Some-college" ...
 $ educational.num: int  7 9 12 10 10 6 9 15 10 4 ...
 $ marital.status : chr  "Never-married" "Married-civ-spouse" "Married-civ-spouse" "Married-civ-spouse" ...
 $ occupation     : chr  "Machine-op-inspct" "Farming-fishing" "Protective-serv" "Machine-op-inspct" ...
 $ relationship   : chr  "Own-child" "Husband" "Husband" "Husband" ...
 $ race           : chr  "Black" "White" "White" "Black" ...
 $ gender         : chr  "Male" "Male" "Male" "Male" ...
 $ capital.gain   : int  0 0 0 7688 0 0 0 3103 0 0 ...
 $ capital.loss   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ hours.per.week : int  40 50 40 40 30 30 40 32 40 10 ...
 $ native.country : chr  "United-States" "United-States" "United-States" "United-States" ...
 $ income         : chr  "<=50K" "<=50K" ">50K" ">50K" ...
NULL
>
```

The dataset contains 48,842 observations and 15 variables, including demographic, work, and income-related attributes. Most variables are numeric or categorical, reflecting typical features used for income classification tasks.

print(summary(df))

```
      age         workclass             fnlwgt          education          educational.num marital.status    
 Min.   :17.00   Length:48842       Min.   :  12285   Length:48842       Min.   : 1.00   Length:48842      
 1st Qu.:28.00   Class :character   1st Qu.: 117551   Class :character   1st Qu.: 9.00   Class :character  
 Median :37.00   Mode  :character   Median : 178145   Mode  :character   Median :10.00   Mode  :character  
 Mean   :38.64                      Mean   : 189664                      Mean   :10.08                     
 3rd Qu.:48.00                      3rd Qu.: 237642                      3rd Qu.:12.00                     
 Max.   :90.00                      Max.   :1490400                      Max.   :16.00                     
  occupation        relationship           race              gender           capital.gain    capital.loss   
 Length:48842       Length:48842       Length:48842       Length:48842       Min.   :    0   Min.   :   0.0  
 Class :character   Class :character   Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :    0   Median :   0.0  
                                                                             Mean   : 1079   Mean   :  87.5  
                                                                             3rd Qu.:    0   3rd Qu.:   0.0  
                                                                             Max.   :99999   Max.   :4356.0  
 hours.per.week  native.country        income         
 Min.   : 1.00   Length:48842       Length:48842      
 1st Qu.:40.00   Class :character   Class :character  
 Median :40.00   Mode  :character   Mode  :character  
 Mean   :40.42                                        
 3rd Qu.:45.00                                        
 Max.   :99.00                                        
>
```

The summary shows reasonable distributions for age, education level, and work hours, while variables like capital.gain and capital.loss are highly skewed with many zeros. Most categorical fields contain diverse but well-structured values suitable for modeling.

Unique Value

code

print(unique(df$workclass))

print(unique(df$education))

print(unique(df$occupation))

print(unique(df$relationship))

print(unique(df$race))

print(unique(df$gender))

print(unique(df$income))

```
> print(unique(df$workclass))
[1] "Private"          "Local-gov"        "?"                "Self-emp-not-inc" "Federal-gov"
[6] "State-gov"        "Self-emp-inc"     "Without-pay"      "Never-worked"
> print(unique(df$education))
 [1] "11th"         "HS-grad"      "Assoc-acdm"   "Some-college" "10th"         "Prof-school"  "7th-8th"
 [8] "Bachelors"    "Masters"      "Doctorate"    "5th-6th"      "Assoc-voc"    "9th"          "12th"
[15] "1st-4th"      "Preschool"
> print(unique(df$occupation))
 [1] "Machine-op-inspct" "Farming-fishing"   "Protective-serv"   "?"                 "Other-service"
 [6] "Prof-specialty"    "Craft-repair"      "Adm-clerical"      "Exec-managerial"   "Tech-support"
[11] "Sales"             "Priv-house-serv"   "Transport-moving"  "Handlers-cleaners" "Armed-Forces"
> print(unique(df$relationship))
[1] "Own-child"      "Husband"        "Not-in-family"  "Unmarried"      "Wife"           "Other-relative"
> print(unique(df$race))
[1] "Black"              "White"              "Asian-Pac-Islander" "Other"              "Amer-Indian-Eskimo"
> print(unique(df$gender))
[1] "Male"   "Female"
> print(unique(df$income))
[1] "<=50K" ">50K"
>
```

The categorical variables show a wide range of distinct values, including several unknown entries marked as “?”. Key fields like gender and income have simple categories, while attributes such as workclass, education, and occupation are more diverse, reflecting varied population backgrounds.

Drop Column

Code

df <- df %>% select(-fnlwgt)

print(head(df))

```
  age workclass    education educational.num     marital.status        occupation  relationship  race gender
1  25   Private         11th               7      Never-married Machine-op-inspct     Own-child Black   Male
2  38   Private      HS-grad               9 Married-civ-spouse   Farming-fishing       Husband White   Male
3  28 Local-gov   Assoc-acdm              12 Married-civ-spouse   Protective-serv       Husband White   Male
4  44   Private Some-college              10 Married-civ-spouse Machine-op-inspct       Husband Black   Male
5  18         ? Some-college              10      Never-married                 ?     Own-child White Female
6  34   Private         10th               6      Never-married     Other-service Not-in-family White   Male
  capital.gain capital.loss hours.per.week native.country income
1            0            0             40  United-States  <=50K
2            0            0             50  United-States  <=50K
3            0            0             40  United-States   >50K
4         7688            0             40  United-States   >50K
5            0            0             30  United-States  <=50K
6            0            0             30  United-States  <=50K
>
```

The unnecessary variable fnlwgt was removed to simplify the dataset and reduce noise. The updated dataset now contains only the relevant features for further analysis.

Missing value

code

df[df == "?"] <- NA

colSums(is.na(df))

```
            age       workclass       education educational.num  marital.status      occupation    relationship 
              0            2799               0               0               0            2809               0 
           race          gender    capital.gain    capital.loss  hours.per.week  native.country          income 
              0               0               0               0               0             857               0 
```

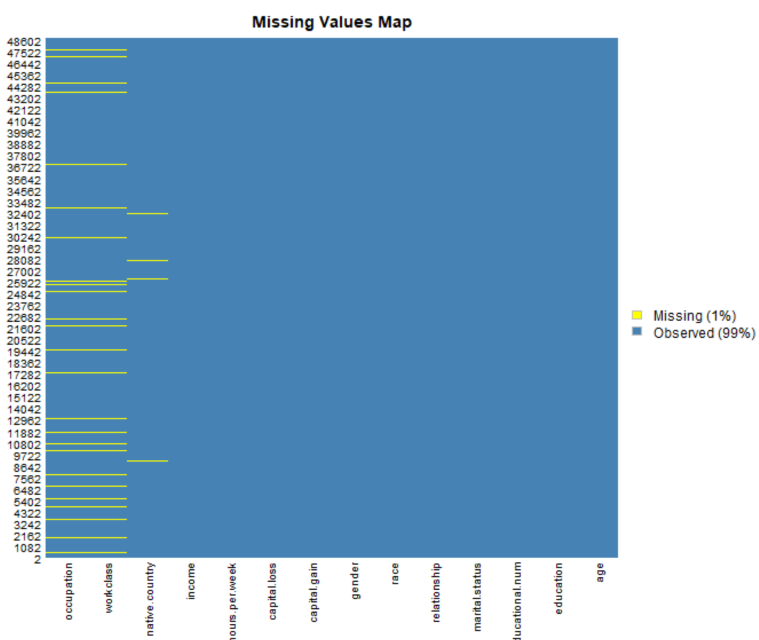
head(df)

```
  age workclass    education educational.num     marital.status        occupation  relationship  race gender
1  25   Private         11th               7      Never-married Machine-op-inspct     Own-child Black   Male
2  38   Private      HS-grad               9 Married-civ-spouse   Farming-fishing       Husband White   Male
3  28 Local-gov   Assoc-acdm              12 Married-civ-spouse   Protective-serv       Husband White   Male
4  44   Private Some-college              10 Married-civ-spouse Machine-op-inspct       Husband Black   Male
5  18      <NA> Some-college              10      Never-married              <NA>     Own-child White Female
6  34   Private         10th               6      Never-married     Other-service Not-in-family White   Male
  capital.gain capital.loss hours.per.week native.country income
1            0            0             40  United-States  <=50K
2            0            0             50  United-States  <=50K
3            0            0             40  United-States   >50K
4         7688            0             40  United-States   >50K
5            0            0             30  United-States  <=50K
6            0            0             30  United-States  <=50K
>
```

All placeholder “?” values were converted to proper NA entries to accurately represent missing data. A column-wise NA count was generated to identify variables requiring further cleaning or imputation.

df[!complete.cases(df),]

missmap(df, main = "Missing Values Map", col = c("yellow", "steelblue"), legend = TRUE)

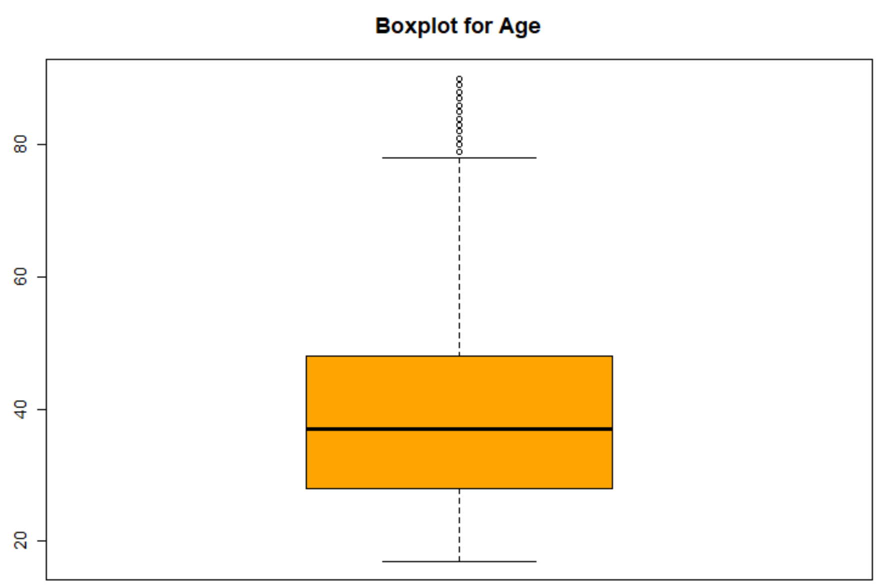


Rows with missing values were extracted to inspect incomplete records. A visual **missing values map** was plotted, clearly highlighting columns with gaps for targeted imputation.

Detect outliers in 'age' column

code

```
par(mar = c(5,4,4,2))


boxplot(df$age,
    main = "Boxplot for Age",
    col = "orange",
    cex.main = 1.5,
    cex.axis = 1.2)
```



A boxplot of the age variable was created to visualize its distribution and detect outliers. The plot highlights the median, quartiles, and extreme age values in the dataset.

Convert numeric to categorical

Code

```
df$Age_group <- cut(df$age,
        breaks = c(0, 25, 45, 65, 100),
        labels = c("Young", "Adult", "Middle-aged", "Senior"))



df$Workload <- ifelse(df$hours.per.week > 40, "Full-time", "Part-time")
df
```

```
  age workclass    education educational.num     marital.status        occupation  relationship  race gender
1  25   Private         11th               7      Never-married Machine-op-inspct     Own-child Black   Male
2  38   Private      HS-grad               9 Married-civ-spouse   Farming-fishing       Husband White   Male
3  28 Local-gov   Assoc-acdm              12 Married-civ-spouse   Protective-serv       Husband White   Male
4  44   Private Some-college              10 Married-civ-spouse Machine-op-inspct       Husband Black   Male
5  18      <NA> Some-college              10      Never-married              <NA>     Own-child White Female
6  34   Private         10th               6      Never-married     Other-service Not-in-family White   Male
  capital.gain capital.loss hours.per.week native.country income Age_group  Workload hours_norm
1            0            0             40  United-States  <=50K     Young Part-time  0.3979592
2            0            0             50  United-States  <=50K     Adult Full-time  0.5000000
3            0            0             40  United-States   >50K     Adult Part-time  0.3979592
4         7688            0             40  United-States   >50K     Adult Part-time  0.3979592
5            0            0             30  United-States  <=50K     Young Part-time  0.2959184
6            0            0             30  United-States  <=50K     Adult Part-time  0.2959184
>
```

New categorical features were created: Age_group segments individuals by age ranges, and Workload classifies them based on weekly working hours. These transformations facilitate easier analysis and visualization of demographic and work patterns.

Normalize a continuous variable

Code

df$hours_norm <- (df$hours.per.week - min(df$hours.per.week, na.rm=TRUE)) /

 (max(df$hours.per.week, na.rm=TRUE) - min(df$hours.per.week, na.rm=TRUE))

df

```
  age workclass    education educational.num     marital.status        occupation  relationship  race gender
1  25   Private         11th               7      Never-married Machine-op-inspct     Own-child Black   Male
2  38   Private      HS-grad               9 Married-civ-spouse   Farming-fishing       Husband White   Male
3  28 Local-gov   Assoc-acdm              12 Married-civ-spouse   Protective-serv       Husband White   Male
4  44   Private Some-college              10 Married-civ-spouse Machine-op-inspct       Husband Black   Male
5  18      <NA> Some-college              10      Never-married              <NA>     Own-child White Female
6  34   Private         10th               6      Never-married     Other-service Not-in-family White   Male
  capital.gain capital.loss hours.per.week native.country income Age_group  Workload hours_norm
1            0            0             40  United-States  <=50K     Young Part-time  0.3979592
2            0            0             50  United-States  <=50K     Adult Full-time  0.5000000
3            0            0             40  United-States   >50K     Adult Part-time  0.3979592
4         7688            0             40  United-States   >50K     Adult Part-time  0.3979592
5            0            0             30  United-States  <=50K     Young Part-time  0.2959184
6            0            0             30  United-States  <=50K     Adult Part-time  0.2959184
>
```

The **hours.per.week** variable was normalized to a 0–1 scale in the new **hours_norm** column. This standardization allows fair comparison and modeling alongside other features.

Remove duplicate rows

Code

df <- df %>% distinct()

sum(duplicated(df))

```
> sum(duplicated(df))
[1] 0
>
```

Duplicate records were removed from the dataset, ensuring each observation is unique. A check confirmed that no duplicates remain, maintaining data integrity for analysis.

Filtering the data

Code

df_filtered <- df %>% filter(age > 40)

df_filtered

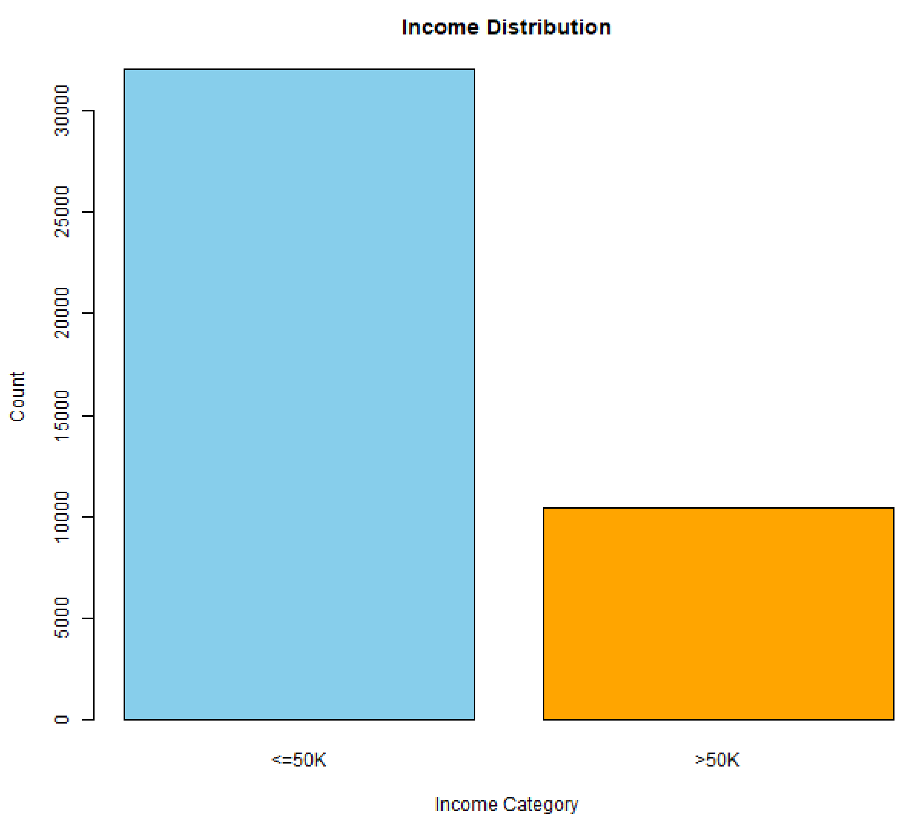
```
  age        workclass    education educational.num     marital.status        occupation relationship  race
1  44          Private Some-college              10 Married-civ-spouse Machine-op-inspct      Husband Black
2  63 Self-emp-not-inc  Prof-school              15 Married-civ-spouse    Prof-specialty      Husband White
3  55          Private      7th-8th               4 Married-civ-spouse      Craft-repair      Husband White
4  65          Private      HS-grad               9 Married-civ-spouse Machine-op-inspct      Husband White
5  58             <NA>      HS-grad               9 Married-civ-spouse              <NA>      Husband White
6  48          Private      HS-grad               9 Married-civ-spouse Machine-op-inspct      Husband White
  gender capital.gain capital.loss hours.per.week native.country income   Age_group  Workload hours_norm
1   Male         7688            0             40  United-States   >50K       Adult Part-time 0.39795918
2   Male         3103            0             32  United-States   >50K Middle-aged Part-time 0.31632653
3   Male            0            0             10  United-States  <=50K Middle-aged Part-time 0.09183673
4   Male         6418            0             40  United-States   >50K Middle-aged Part-time 0.39795918
5   Male            0            0             35  United-States  <=50K Middle-aged Part-time 0.34693878
6   Male         3103            0             48  United-States   >50K Middle-aged Full-time 0.47959184
>
```

A subset of the dataset was created containing only individuals older than 40. This filtered data enables focused analysis on the middle-aged and senior population segments.

Handle Imbalanced Dataset

Code

```
table(df$income)


barplot(table(df$income),
    col = c("skyblue", "orange"),
    main = "Income Distribution",
    xlab = "Income Category",
    ylab = "Count")
```

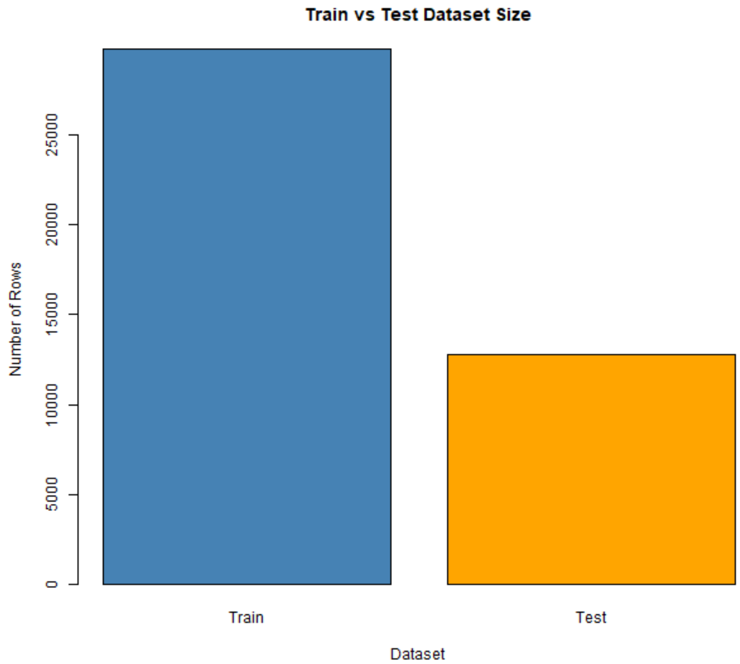


The income distribution was summarized and visualized using a bar plot. It clearly shows the count of individuals earning <=50K versus >50K, highlighting class imbalance in the dataset.

Train-Test Split

Code

```
set.seed(123)
trainIndex <- createDataPartition(df$income, p=0.7, list=FALSE)

train <- df[trainIndex, ]
test  <- df[-trainIndex, ]

dim(train)
dim(test)

sizes <- c(nrow(train), nrow(test))
names(sizes) <- c("Train", "Test")

barplot(sizes,
    col = c("steelblue", "orange"),
    main = "Train vs Test Dataset Size",
    ylab = "Number of Rows",
    xlab = "Dataset")
```

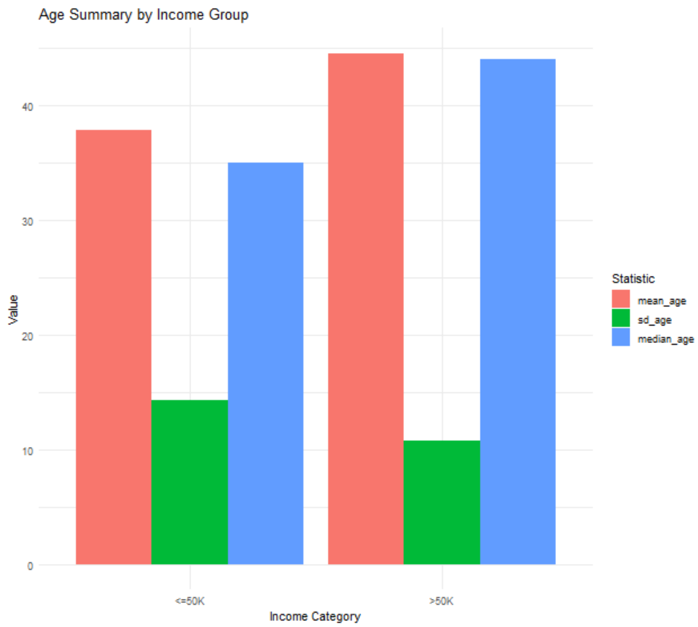


The dataset was split into 70% training and 30% testing sets to prepare for model building and evaluation. A bar plot visualizes the number of rows in each subset, confirming the partition sizes.

Descriptive Statistics

Code

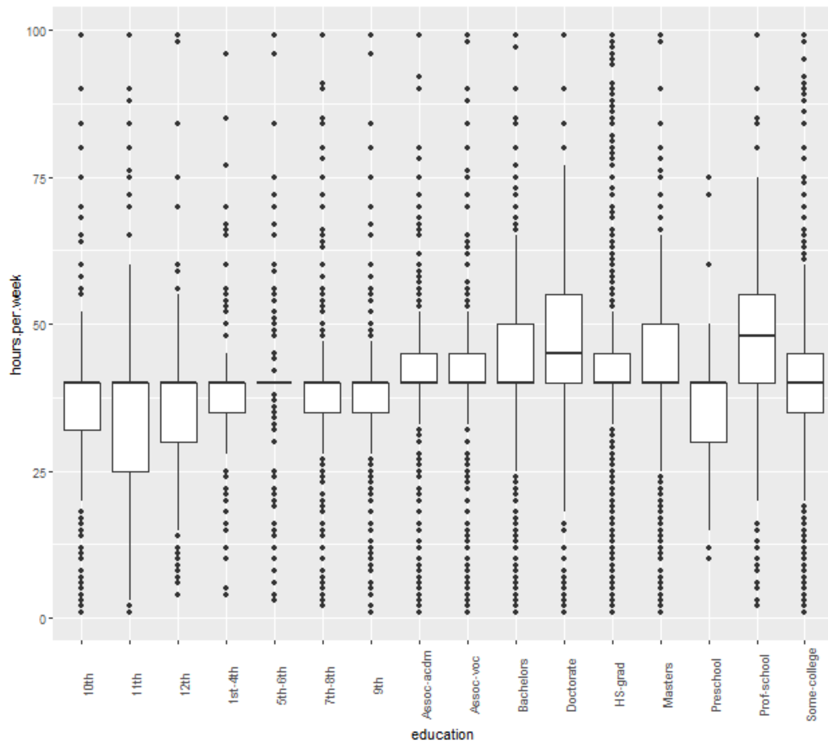
```
age_summary <- df %>%
 group_by(income) %>%
 summarise(
  mean_age = mean(age, na.rm = TRUE),
  sd_age = sd(age, na.rm = TRUE),
  median_age = median(age, na.rm = TRUE)
 )
age_summary

age_long <- melt(age_summary,
          id.vars = "income",
          variable.name = "Statistic",
          value.name = "Value")

ggplot(age_long, aes(x = income, y = Value, fill = Statistic)) +
 geom_bar(stat = "identity", position = "dodge") +
 labs(title = "Age Summary by Income Group",
    x = "Income Category",
    y = "Value") +
 theme_minimal()
```



The age distribution was summarized by income category, calculating mean, standard deviation, and median ages. Individuals earning >50K are generally older (mean 44.3) than those earning <=50K (mean 36.9), with slightly less age variability. A bar plot was created to visually compare these statistics across income groups. This highlights that higher income is associated with older age in the dataset.

visualize education to hours.per.week

```
ggplot(df, aes(x = education, y = hours.per.week)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle=90))
```



A boxplot was created to examine weekly working hours across education levels. It shows variations and potential outliers in hours, highlighting how education may influence workload patterns.