Bike Buyers Analysis

Arrhat Maharjan

2025-03-28

Task 1. Data Cleaning and Preparation

In this section, we'll clean the bike buyers dataset and prepare it for analysis. This includes handling missing values, checking for outliers, and ensuring proper data types.

```r
library(ggplot2)
library(corrplot)

# task 1 - data cleaning and preparation
# load the dataset
bike_buyers.dataset <- read.csv("./bike_buyers.csv", stringsAsFactors = TRUE)

# replace the empty value with NA
bike_buyers.dataset[bike_buyers.dataset == ""] <- NA

# make a copy of the dataset
dataset <- bike_buyers.dataset

# structure of the dataser
str(dataset)
```

```
## 'data.frame':    1000 obs. of  13 variables:
##  $ ID              : int  12496 24107 14177 24381 25597 13507 27974 19364 22155 19280 ...
##  $ Marital.Status  : Factor w/ 3 levels "","Married","Single": 2 2 2 3 3 2 3 2 NA 2 ...
##  $ Gender          : Factor w/ 3 levels "","Female","Male": 2 3 3 NA 3 2 3 3 3 3 ...
##  $ Income          : int  40000 30000 80000 70000 30000 10000 160000 40000 20000 NA ...
##  $ Children        : int  1 3 5 0 0 2 2 1 2 2 ...
##  $ Education       : Factor w/ 5 levels "Bachelors","Graduate Degree",..: 1 4 4 1 1 4 3 1 5 4 ...
##  $ Occupation      : Factor w/ 5 levels "Clerical","Management",..: 5 1 4 4 1 3 2 5 1 3 ...
##  $ Home.Owner      : Factor w/ 3 levels "","No","Yes": 3 3 2 3 2 3 NA 3 3 3 ...
##  $ Cars            : int  0 1 2 1 0 0 4 0 2 1 ...
##  $ Commute.Distance: Factor w/ 5 levels "0-1 Miles","1-2 Miles",..: 1 1 4 5 1 2 1 1 5 1 ...
##  $ Region          : Factor w/ 3 levels "Europe","North America",..: 1 1 1 3 1 1 3 1 3 1 ...
##  $ Age             : int  42 43 60 41 36 50 33 43 58 NA ...
##  $ Purchased.Bike  : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 1 2 2 1 2 ...
```

```r
#summary of the dataset
summary(dataset)
```

```
##        ID          Marital.Status    Gender          Income            Children
```

```
##  Min.   :11000          :  0          :  0   Min.   : 10000   Min.   :0.00
##  1st Qu.:15291   Married:535    Female:489   1st Qu.: 30000   1st Qu.:0.00
##  Median :19744   Single :458    Male  :500   Median : 60000   Median :2.00
##  Mean   :19966   NA's   :  7    NA's  : 11   Mean   : 56268   Mean   :1.91
##  3rd Qu.:24471                               3rd Qu.: 70000   3rd Qu.:3.00
##  Max.   :29447                               Max.   :170000   Max.   :5.00
##                                              NA's   :6        NA's   :8
##               Education            Occupation  Home.Owner      Cars
##  Bachelors          :306   Clerical      :177      :  0   Min.   :0.000
##  Graduate Degree    :174   Management    :173   No  :314   1st Qu.:1.000
##  High School        :179   Manual        :119   Yes :682   Median :1.000
##  Partial College    :265   Professional  :276   NA's:  4   Mean   :1.455
##  Partial High School: 76   Skilled Manual:255              3rd Qu.:2.000
##                                                            Max.   :4.000
##                                                            NA's   :9
##    Commute.Distance           Region         Age        Purchased.Bike
##  0-1 Miles :366     Europe       :300   Min.   :25.00   No :519
##  1-2 Miles :169     North America:508   1st Qu.:35.00   Yes:481
##  10+ Miles :111     Pacific      :192   Median :43.00
##  2-5 Miles :162                         Mean   :44.18
##  5-10 Miles:192                         3rd Qu.:52.00
##                                         Max.   :89.00
##                                         NA's   :8
```

```r
#check number of NAs
colSums(is.na(dataset))
```

```
##               ID   Marital.Status           Gender           Income
##                0                7               11                6
##         Children        Education       Occupation       Home.Owner
##                8                0                0                4
##             Cars Commute.Distance           Region              Age
##                9                0                0                8
##   Purchased.Bike
##                0
```

```r
# omit the row with any NA values
dataset <- na.omit(dataset)

# drop unused factor levels
dataset <- droplevels(dataset)

#check number of NAs
colSums(is.na(dataset))
```

```
##               ID   Marital.Status           Gender           Income
##                0                0                0                0
##         Children        Education       Occupation       Home.Owner
##                0                0                0                0
##             Cars Commute.Distance           Region              Age
##                0                0                0                0
##   Purchased.Bike
##                0
```

```r
# save cleaned dataset
write.csv(dataset, "new_data.csv", row.names = FALSE)

# structure of cleaned dataset
str(dataset)
```

```
## 'data.frame':    952 obs. of  13 variables:
##  $ ID             : int  12496 24107 14177 25597 13507 19364 22173 12697 25323 23542 ...
##  $ Marital.Status  : Factor w/ 2 levels "Married","Single": 1 1 1 2 1 1 1 2 1 2 ...
##  $ Gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 1 1 2 2 ...
##  $ Income          : int  40000 30000 80000 30000 10000 40000 30000 90000 40000 60000 ...
##  $ Children        : int  1 3 5 0 2 1 3 0 2 1 ...
##  $ Education       : Factor w/ 5 levels "Bachelors","Graduate Degree",..: 1 4 4 1 4 1 3 1 4 4 ...
##  $ Occupation      : Factor w/ 5 levels "Clerical","Management",..: 5 1 4 1 3 5 5 4 1 5 ...
##  $ Home.Owner      : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 2 1 1 2 1 ...
##  $ Cars            : int  0 1 2 0 0 0 2 4 1 1 ...
##  $ Commute.Distance: Factor w/ 5 levels "0-1 Miles","1-2 Miles",..: 1 1 4 1 2 1 2 3 2 1 ...
##  $ Region          : Factor w/ 3 levels "Europe","North America",..: 1 1 1 1 1 1 3 3 1 3 ...
##  $ Age             : int  42 43 60 36 50 43 54 36 35 45 ...
##  $ Purchased.Bike  : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 2 1 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:48] 4 7 9 10 13 28 50 99 111 118 ...
##   ..- attr(*, "names")= chr [1:48] "4" "7" "9" "10" ...
```

The structure stays the same after cleaning the data.

The data cleaning process involved:

1. Converting empty strings to NA values
2. Removing rows with any NA values
3. Dropping unused factor levels
4. Saving the cleaned dataset for future use

Task 2. Summary of Variables

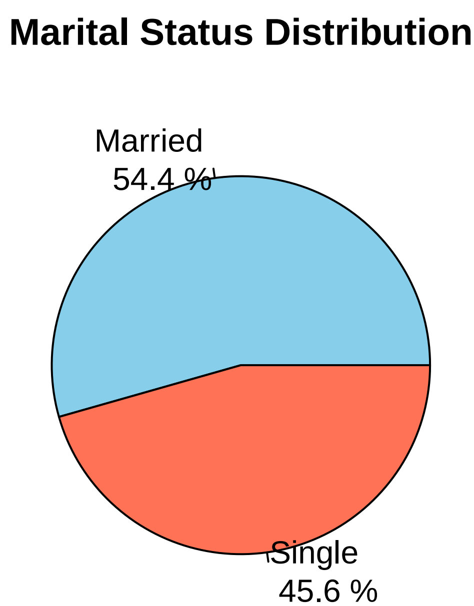
In this section, we'll look at the variables in our dataset using descriptive statistics and visualizations to understand the data.

Summary of Variables:

ID (Numerical): Unique identifier (not useful for analysis). 952 entries after cleaning the data.

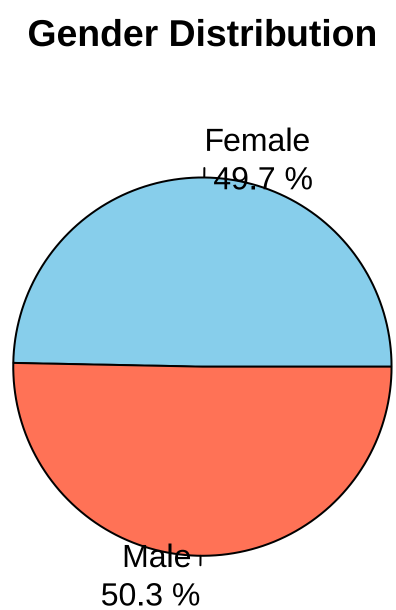
Marital.Status (Categorical): Represents the marital status of an individual. 518 Married individuals, 434 Single.

```
# define color pallette for our visual representation
colors <- c("skyblue", "coral1","darkseagreen","mediumpurple","darkorange1")

# pie chart marital status distribution
marital_count <- table(dataset$Marital.Status)
marital_percent <- round(100 * marital_count / sum(marital_count), 1)
martial_label <- paste(names(marital_count), "\n", marital_percent, "%")
pie(marital_count,
    labels = martial_label,
    col = colors,
    main = "Marital Status Distribution")
```



Gender (Categorical): Represents the gender of the individual. 473 Females, 479 Males.

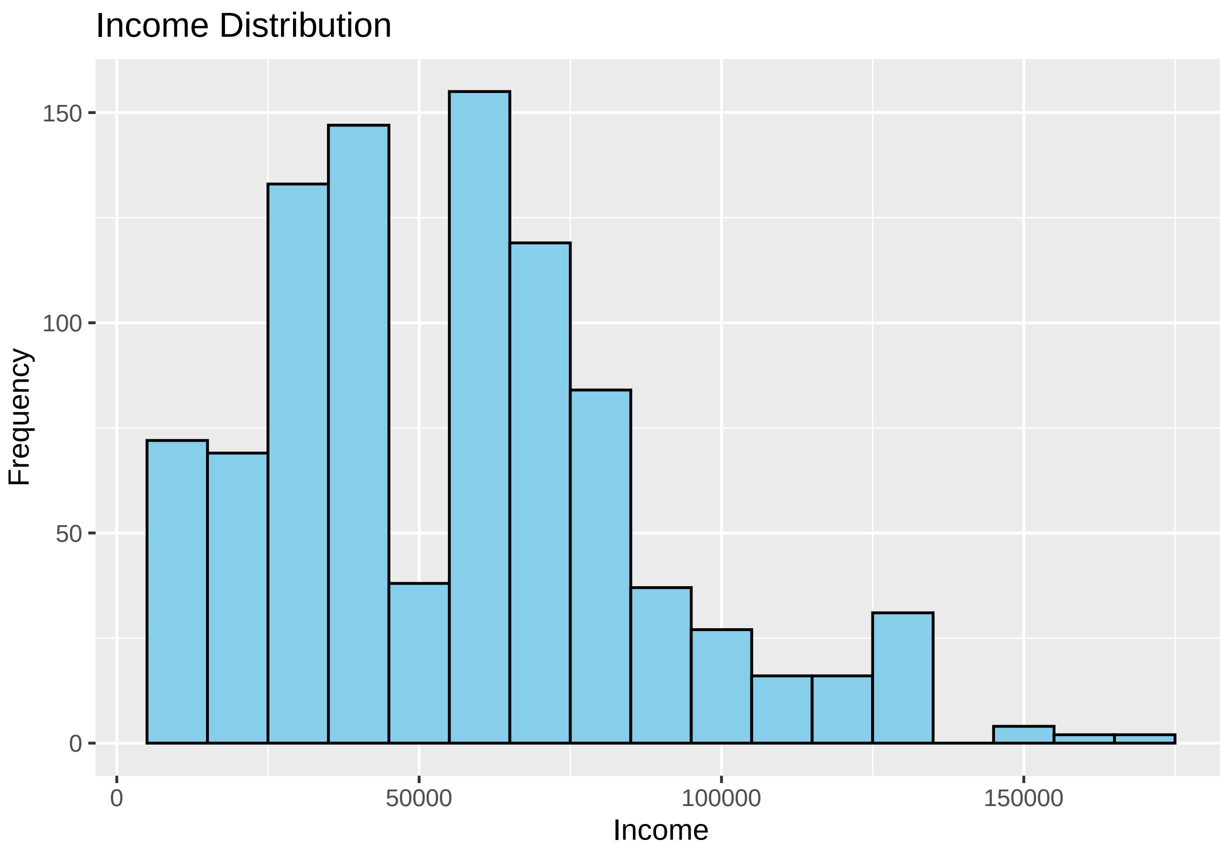
```r
# pie chart gender distribution
gender_count <- table(dataset$Gender)
gender_percent <- round(100 * gender_count / sum(gender_count), 1)
gender_label <- paste(names(gender_count), "\n", gender_percent, "%")
pie(gender_count,
    labels = gender_label,
    col = colors,
    main = "Gender Distribution")
```



Marital.Status and Gender have a fairly balanced distribution with marital (49.7% Female, 50.3% Male) and marital status (54.4% Married, 45.6% Single).

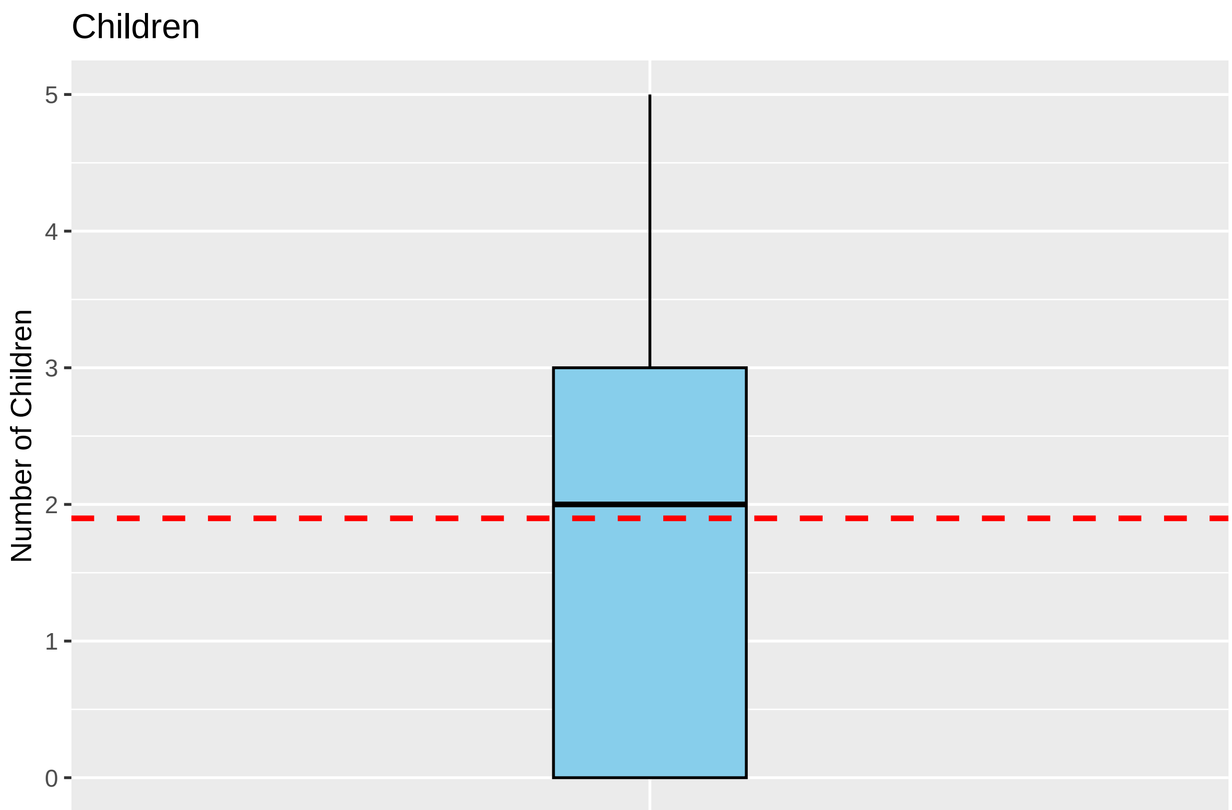
Income (Numerical): Annual income of the individual (continuous). Ranges from $10,000 to $170,000, with a median of $60,000 and mean $55,903.

```
# histogram income distribution
ggplot(dataset, aes(x = Income)) +
  geom_histogram(
    binwidth = 10000,
    fill = "skyblue",
    color = "black",
    alpha = 1
  ) +
  labs(title = "Income Distribution", x = "Income", y = "Frequency")
```



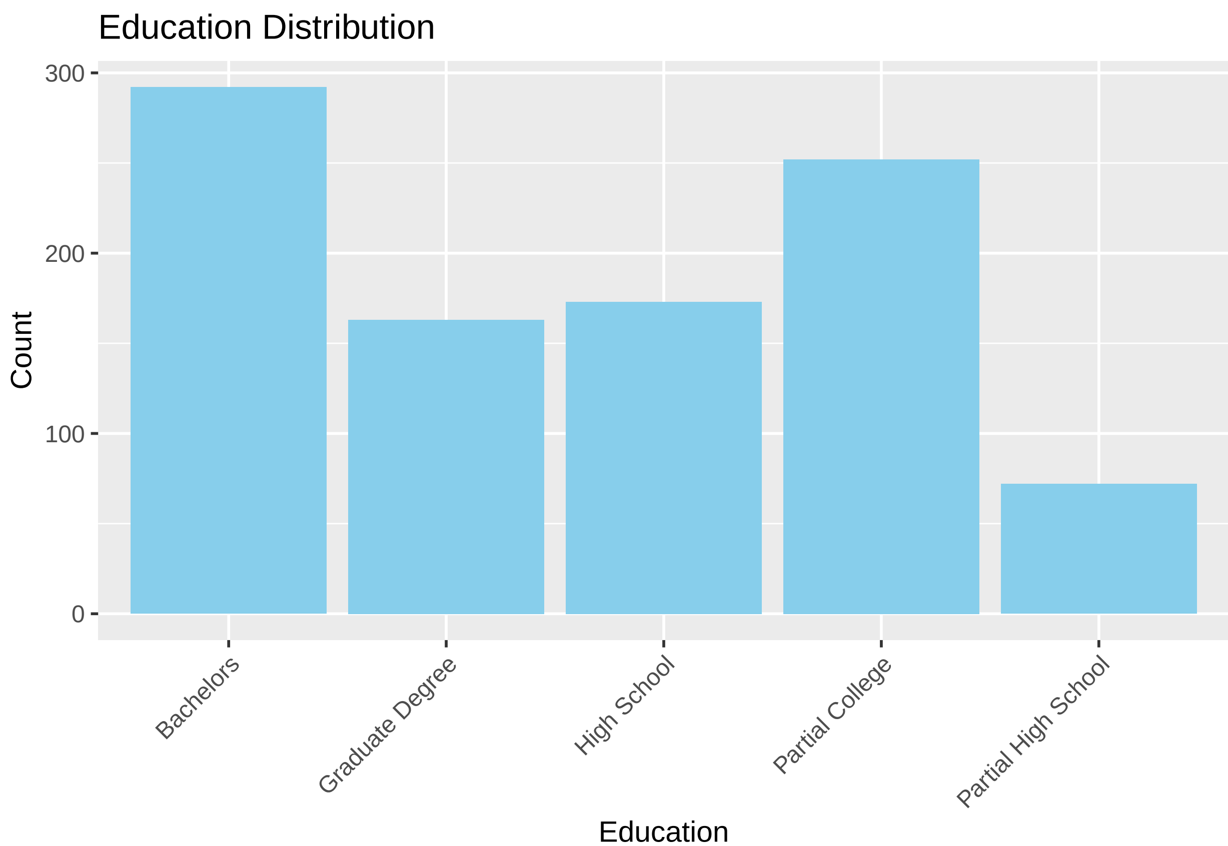
Children (Numerical): Number of children (discrete). Minimum 0, maximum 5, with an average of 1.89 children per household.

```r
# box plot children per household
ggplot(dataset, aes(x = "", y = Children)) +
  geom_boxplot(fill = "skyblue",
               color = "black",
               width = 0.2) +
  geom_hline(
    aes(yintercept = mean(Children)),
    color = "red",
    linetype = "dashed",
    linewidth = 1
  ) +
  labs(title = "Children", y = "Number of Children") +
  theme(axis.title.x = element_blank())
```



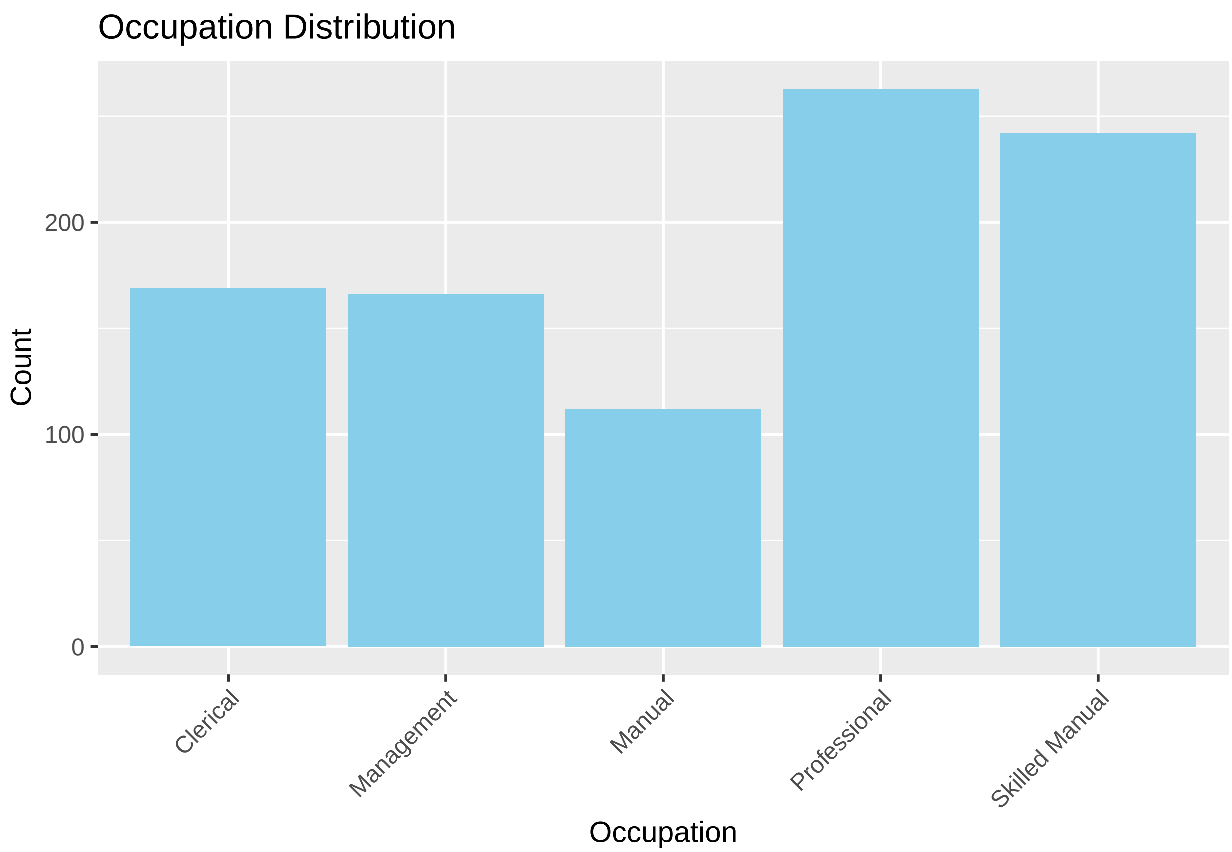
Education (Categorical): Indicates the highest level of education completed. Most individuals have a Bachelor's degree (30.7%/292) or Partial College education (26.5%/252).

```
# bar chart education distribution
ggplot(dataset, aes(x = Education)) +
  geom_bar(fill = "skyblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Education Distribution", x = "Education", y = "Count")
```



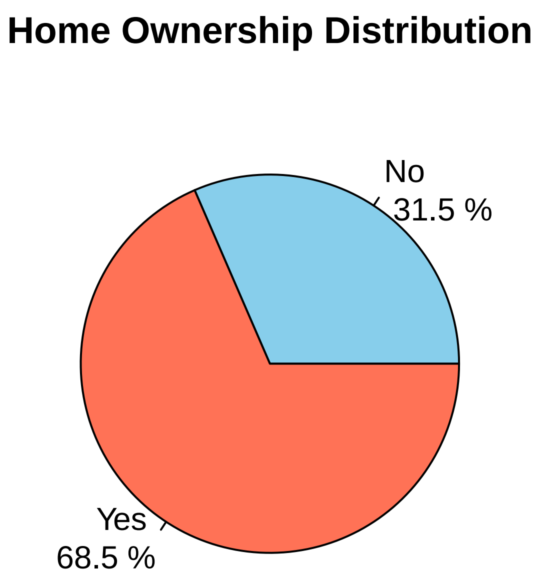
Occupation (Categorical): Job category of the individual.Professionals (27.6%/263) and Skilled Manual workers (25.4%/242) are the most common occupations.

```r
# bar chart occupation distribution
ggplot(dataset, aes(x = Occupation)) +
  geom_bar(fill = "skyblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Occupation Distribution", x = "Occupation", y = "Count")
```



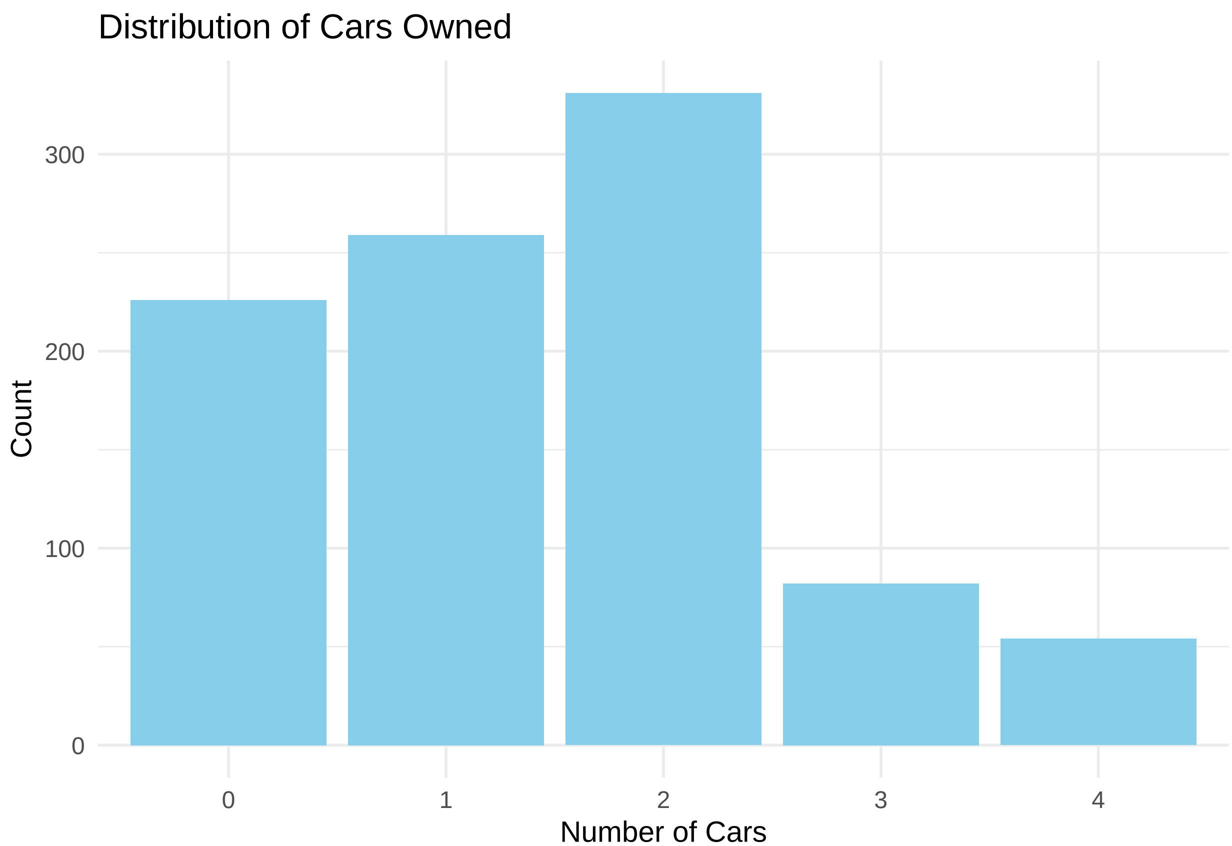
Home.Owner (Categorical): Indicates whether the individual owns a home. Most individuals (68.5%/652) own their homes while the rest(300) do not.

```r
# pie chart for home ownership
home_owner_count <- table(dataset$Home.Owner)
home_owner_percent <- round(100 * home_owner_count / sum(home_owner_count), 1)
home_owner_label <- paste(names(home_owner_count), "\n", home_owner_percent, "%")
pie(home_owner_count,
    labels = home_owner_label,
    col = colors,
    main = "Home Ownership Distribution")
```



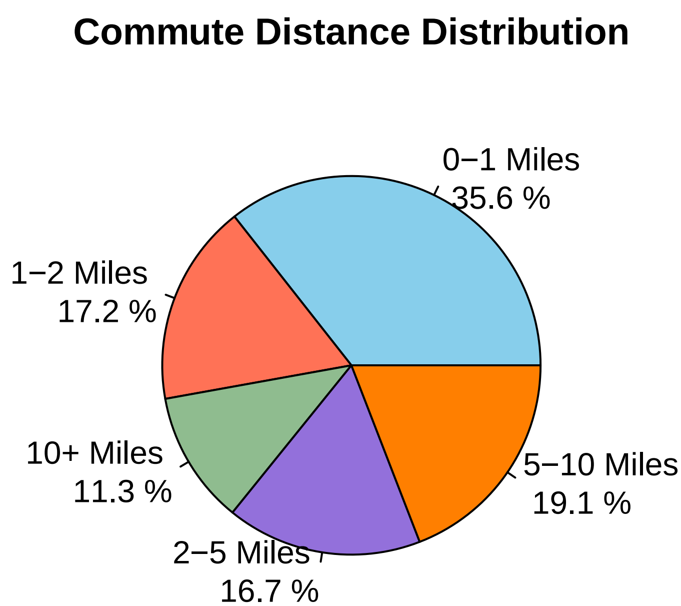
Cars (Numerical): Number of cars owned (discrete). Most individuals own 1 or 2 cars, with a maximum being 4.

```
# bar chart for number of cars owned
ggplot(dataset, aes(x = factor(Cars))) +
  geom_bar(fill = "skyblue") +
  labs(title = "Distribution of Cars Owned",
       x = "Number of Cars",
       y = "Count") +
  theme_minimal()
```



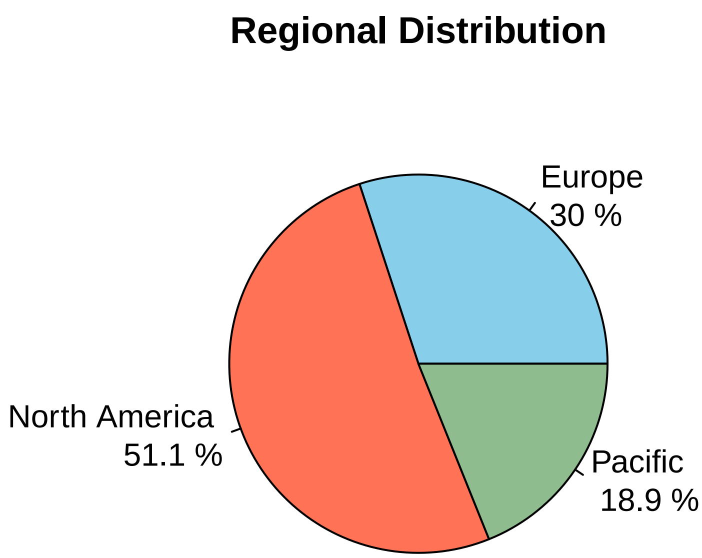
Commute.Distance (Categorical): Represents how far the person commutes daily. Most individuals (35.6%/339) have a short commute of 0-1 miles, followed by 5-10 miles (19.1%/182).

```
# pie chart for commute distance
commute_count <- table(dataset$Commute.Distance)
commute_percent <- round(100 * commute_count / sum(commute_count), 1)
commute_label <- paste(names(commute_count), "\n", commute_percent, "%")
pie(commute_count,
    labels = commute_label,
    col = colors,
    main = "Commute Distance Distribution")
```



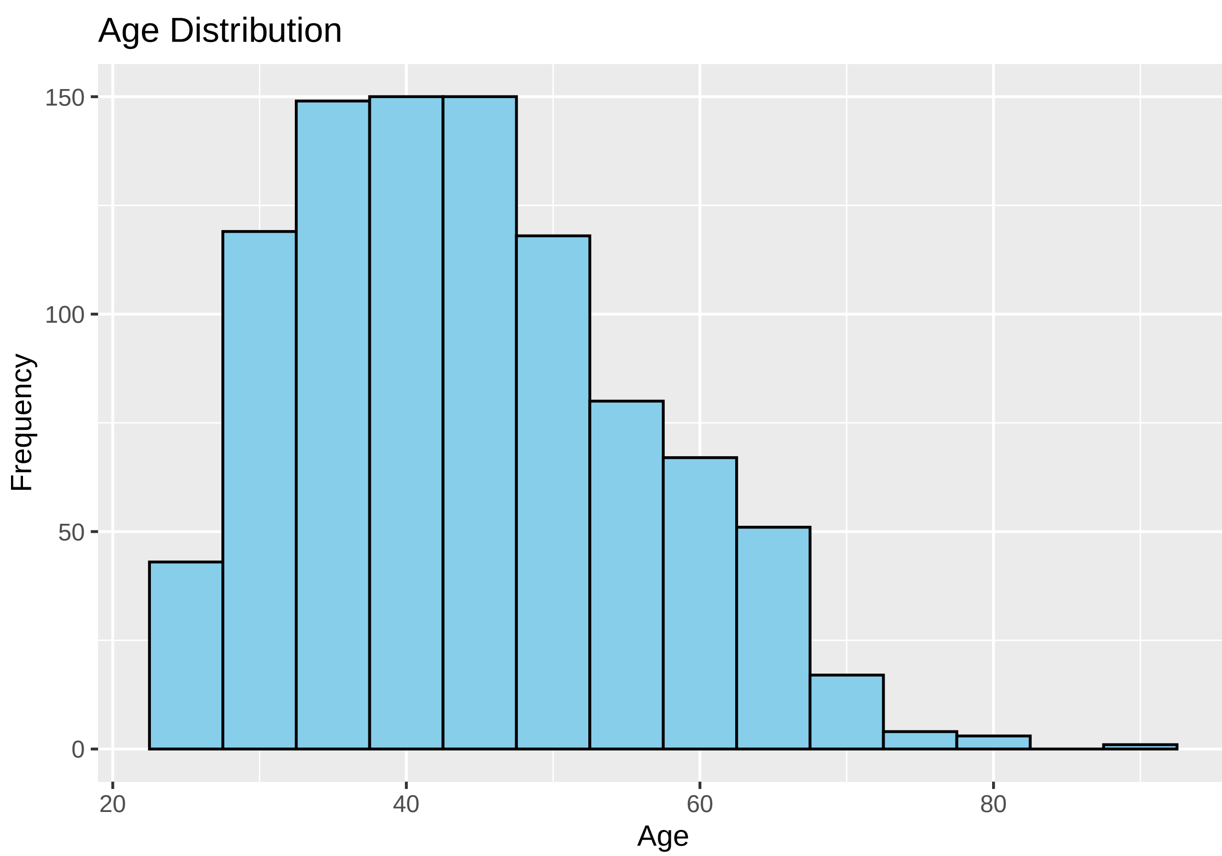
Region (Categorical): The geographical region where the individual lives. The dataset is skewed toward North America (51.1%/486), with Europe (30.0%/286) and Pacific (18.9%/180) having fewer representatives.

```r
# pie chart for region
region_count <- table(dataset$Region)
region_percent <- round(100 * region_count / sum(region_count), 1)
region_label <- paste(names(region_count), "\n", region_percent, "%")
pie(region_count,
    labels = region_label,
    col = colors,
    main = "Regional Distribution")
```



Age (Numerical): Age of the individual (continuous). Ranges from 25 to 89, with a median age of 43 and mean of 44.26. Majority of the individuals in middle-age.

```r
# histogram age distribution
ggplot(dataset, aes(x = Age)) +
  geom_histogram(
    binwidth = 5,
    fill = "skyblue",
    color = "black",
    alpha = 1
  ) +
  labs(title = "Age Distribution", x = "Age", y = "Frequency")
```



Purchased.Bike (Categorical): Indicates whether the individual purchased a bike.The data shows a fairly balanced distribution with 47.9%/456 of individuals having purchased a bike and 52.1%/496 not having done so.

```r
# pie chart for bike purchase
bike_purchase_count <- table(dataset$Purchased.Bike)
bike_purchase_percent <- round(100 * bike_purchase_count / sum(bike_purchase_count), 1)
bike_purchase_label <- paste(names(bike_purchase_count), "\n", bike_purchase_percent, "%")
pie(bike_purchase_count,
    labels = bike_purchase_label,
    col = colors,
    main = "Bike Purchase Distribution")
```

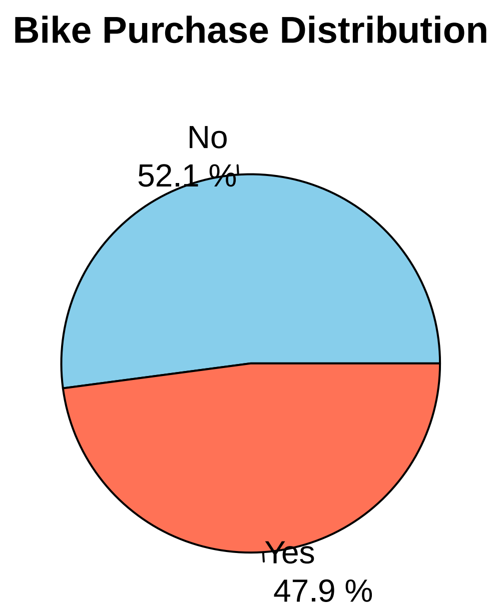


Table 1: Summary of Variables

Variable	Description
ID	952 entries after cleaning the data
Marital Status	518 Married (54.4%), 434 Single (45.6%)
Gender	473 Females (49.7%), 479 Males (50.3%)
Income	Range: $10,000-$170,000, Median: $60,000, Mean: $55,903
Children	Range: 0-5, Mean: 1.89 children per household
Education	Most common: Bachelors (30.7%), Partial College (26.5%)
Occupation	Most common: Professional (27.6%), Skilled Manual (25.4%)
Home Owner	68.5% own a home, 31.5% do not
Cars	Most own 1 or 2 cars, maximum: 4
Commute Distance	Most common: 0-1 Miles (35.6%), 5-10 Miles (19.1%)
Region	North America: 51.1%, Europe: 30.0%, Pacific: 18.9%
Age	Range: 25-89 years, Median: 43, Mean: 44.26
Purchased Bike	47.9% purchased a bike, 52.1% did not

Task 3. Income Analysis

a. Income Distribution and Statistics

```r
# summary statistics income
summary_stats <- data.frame(
  Mean = mean(dataset$Income),
  Median = median(dataset$Income),
  Variance = var(dataset$Income),
  SD = sd(dataset$Income)
)
summary_stats
```

```
##       Mean Median  Variance       SD
## 1 55903.36  60000 951443858 30845.48
```

The distribution is right-skewed (positively skewed). Right-skewed nature can be confirmed from the mean ($55,903.36) being lower than the median ($60,000). Peak frequency occurs around $60,000 and followed by $40,000. Low frequency observations at higher income levels above $100,000 and very low after $140,000.

b. Bike Ownership by Income Level

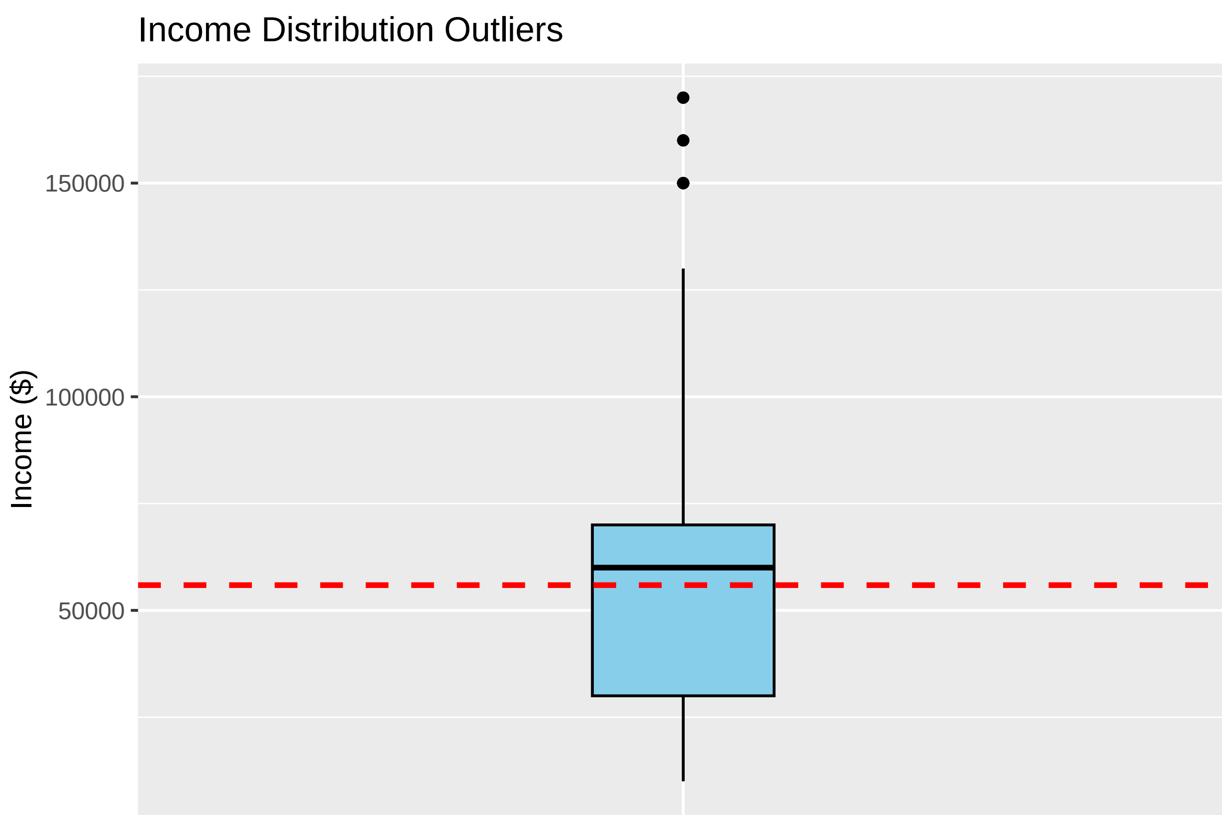
```r
# income ranges
income_groups <- cut(
  dataset$Income,
  breaks = c(0, 40000, 80000, 120000, 170000),
  labels = c(
    "Low 0-40k",
    "Medium 40k-80k",
    "High 80-120k",
    "Very High 120-170k"
  ),
  include.lowest = TRUE
)

# bike by income summary
bikebyincome_summary <- do.call(rbind, by(dataset, income_groups, function(x) {
  data.frame(
    Total = nrow(x),
    Purchased = sum(x$Purchased.Bike == "Yes"),
    Not_Purchased = sum(x$Purchased.Bike == "No")
  )
}))
bikebyincome_summary
```

```
##                    Total Purchased Not_Purchased
## Low 0-40k            421       194           227
## Medium 40k-80k       396       192           204
## High 80-120k          96        50            46
## Very High 120-170k    39        20            19
```

Highest number of bike owners are low and medium income individuals. Although not much difference, high income individuals have a slightly higher bike ownership rate.

c. Income Outliers

```r
# box plot income outliers
ggplot(dataset, aes(x = "", y = Income)) +
  geom_boxplot(fill = "skyblue",
               color = "black",
               width = 0.2) +
  geom_hline(
    aes(yintercept = mean(Income)),
    color = "red",
    linetype = "dashed",
    linewidth = 1
  ) +
  labs(title = "Income Distribution Outliers", y = "Income ($)") +
  theme(axis.title.x = element_blank())
```

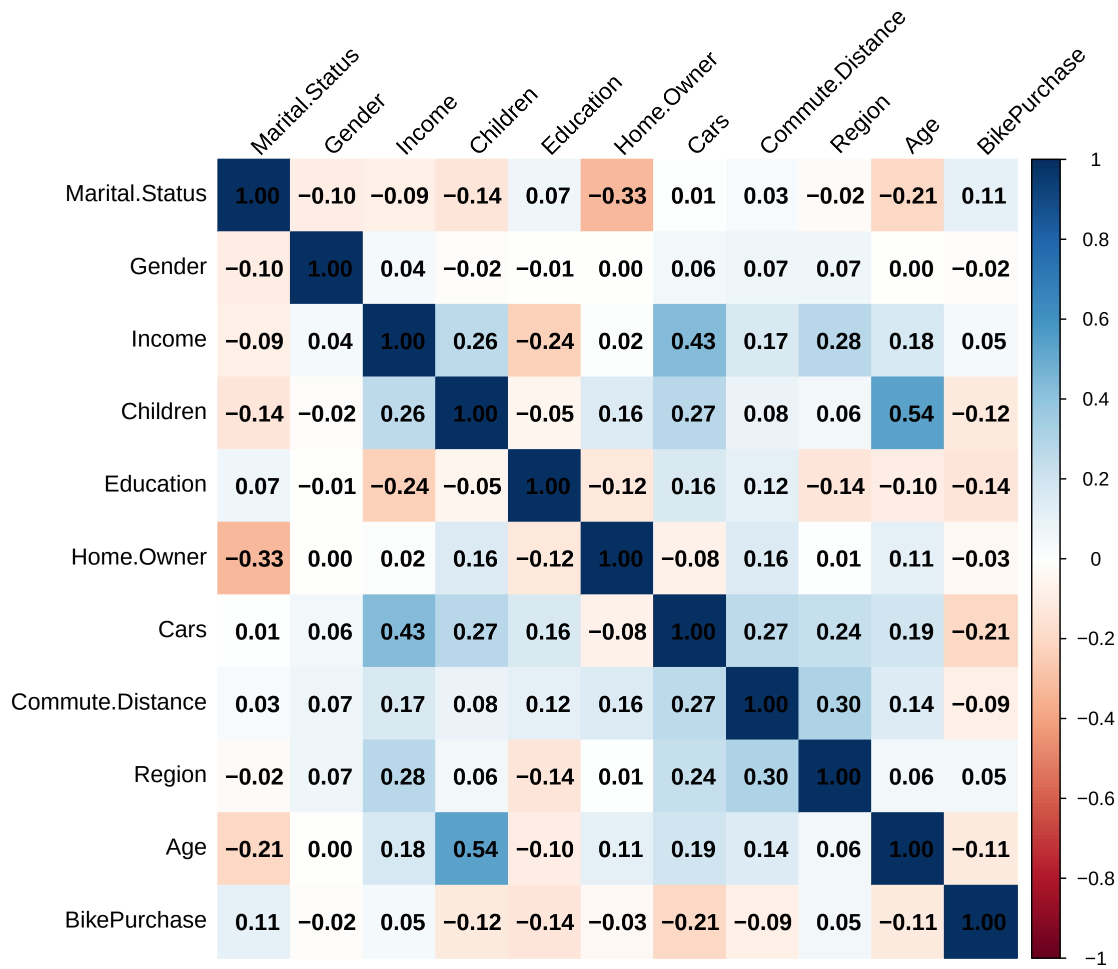


Most outliers are in the high income range.

d. Correlation with Bike Purchase

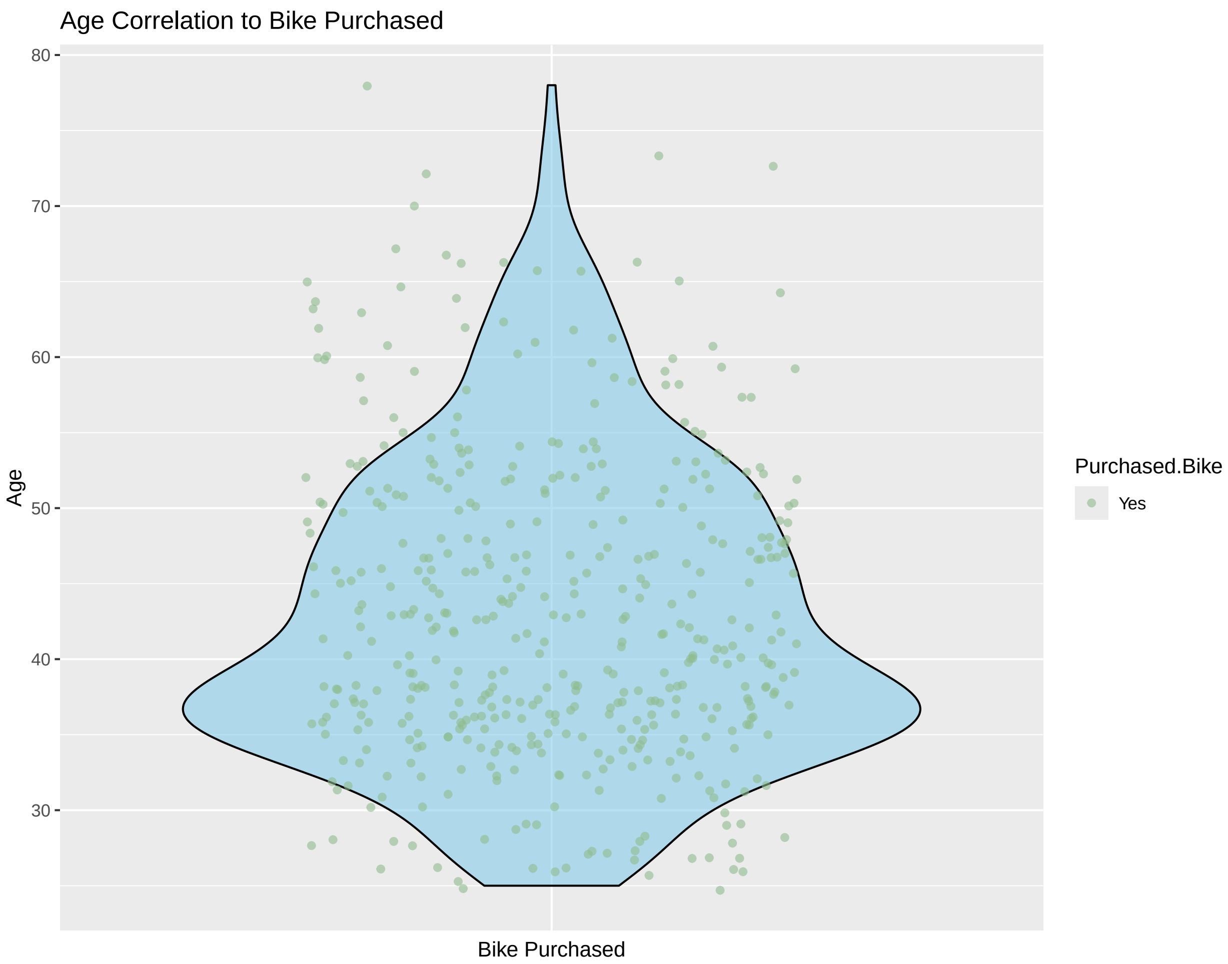
```r
# data frame bike purchase correlation
cor_data <- data.frame(
  Marital.Status = as.numeric(factor(dataset$Marital.Status)),
  Gender = as.numeric(factor(dataset$Gender)),
  Income = as.numeric(as.character(dataset$Income)),
  Children = as.numeric(as.character(dataset$Children)),
  Education = as.numeric(factor(dataset$Education)),
  Home.Owner = as.numeric(factor(dataset$Home.Owner)),
  Cars = as.numeric(as.character(dataset$Cars)),
  Commute.Distance = as.numeric(factor(dataset$Commute.Distance)),
  Region = as.numeric(factor(dataset$Region)),
  Age = as.numeric(as.character(dataset$Age)),
  BikePurchase = ifelse(dataset$Purchased.Bike == "Yes", 1, 0)
)

# correlation matrix
correlation_matrix <- cor(cor_data)
```

```r
# correlation graph
corrplot(
  correlation_matrix,
  method = "color",
  type = "full",
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 45
)
```



We can see that the highest correlation the attribute Purchased.Bike has other than itself is the field Age(0.54) followed by Cars(0.43); and the lowest being Home.Owner(-0.33).

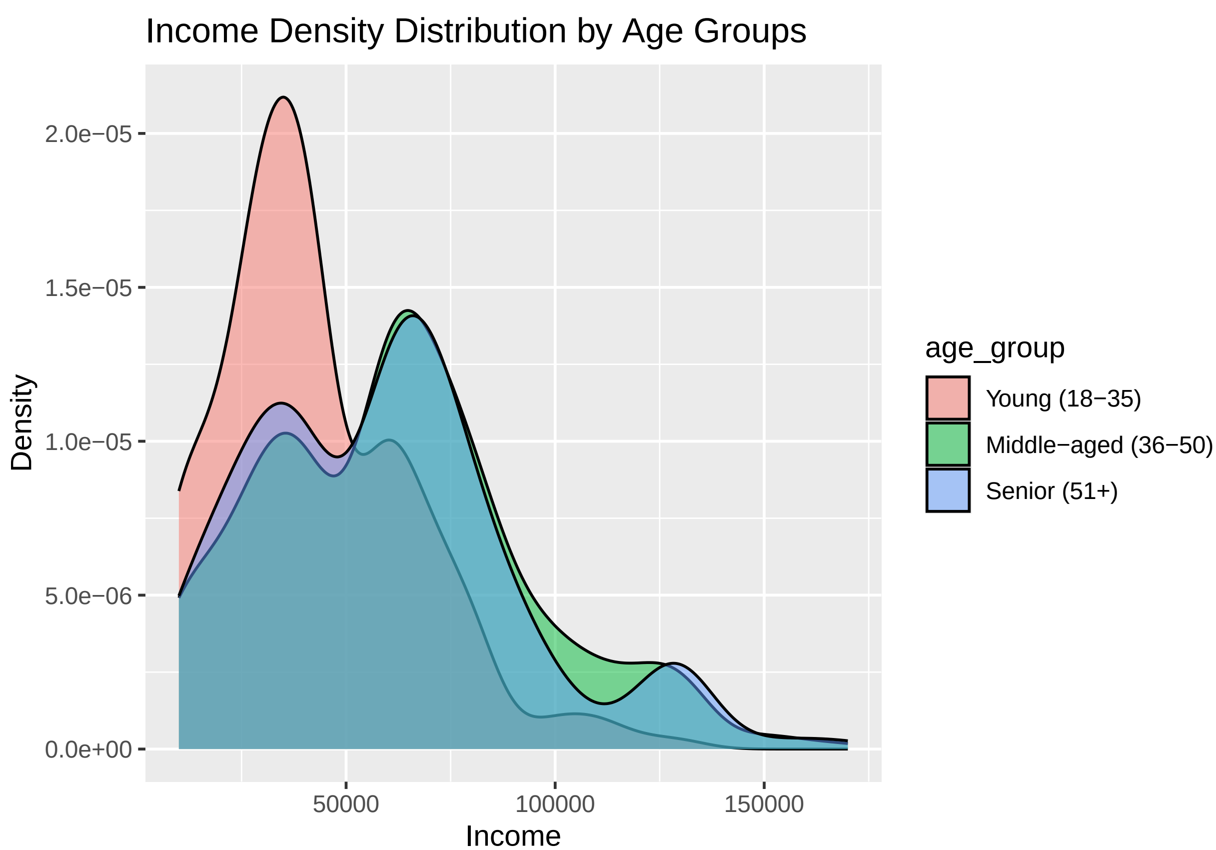
```r
#scatter plot age by bike
ggplot(subset(dataset, Purchased.Bike == "Yes"), aes(x = Purchased.Bike, y = Age, color = Purchased.Bike
  geom_violin(fill = colors[1], color = "black", alpha = 0.6) +
  geom_jitter(width = 0.3, alpha = 0.6) +
  labs(title = "Age Correlation to Bike Purchased", x = "Bike Purchased", y = "Age") +
  scale_color_manual(values = c(colors[3])) +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```



We can see in the visualization that the younger individuals have a higher number of bike owners.

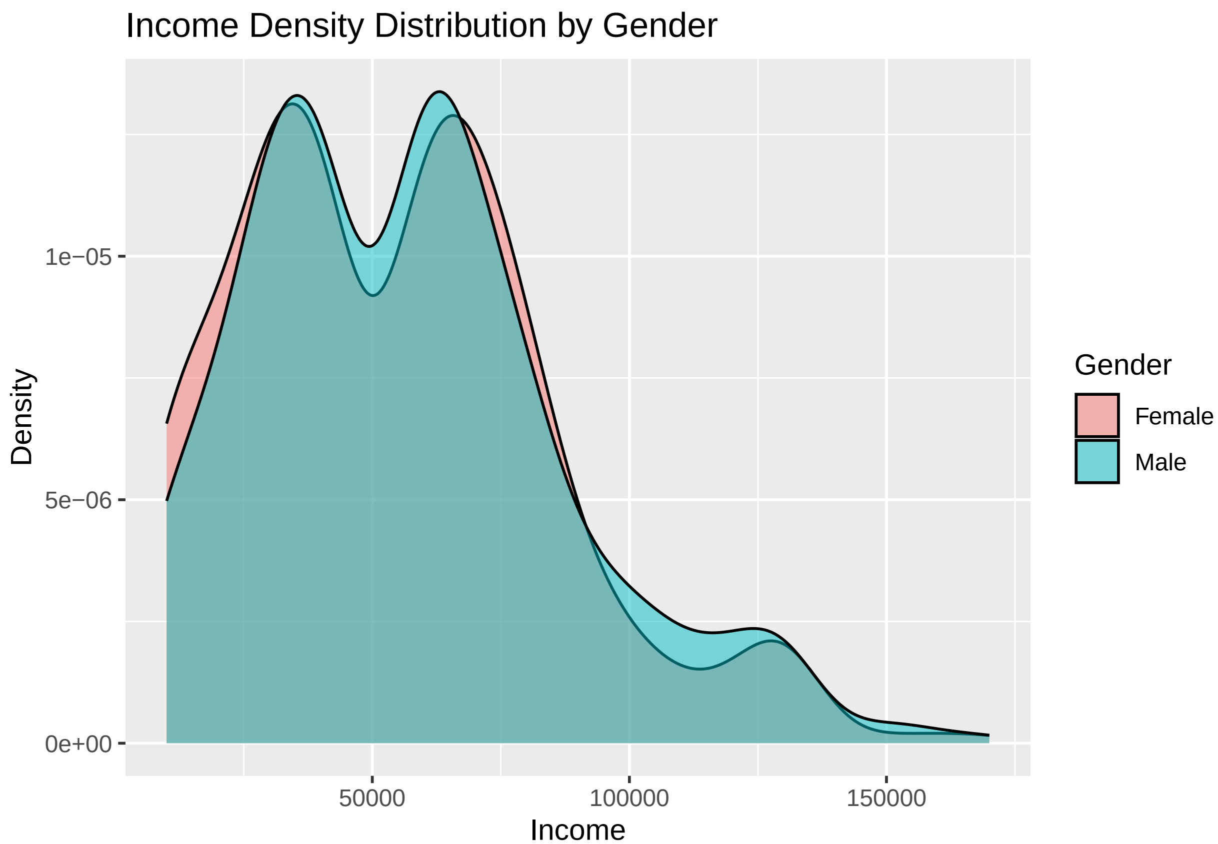
Task 4. Income Distribution Compared to Age and Gender

```r
# age group dataframe
age_group <- cut(
  dataset$Age,
  breaks = c(0, 35, 50, 100),
  labels = c("Young (18-35)", "Middle-aged (36-50)", "Senior (51+)")
)

# density plot income by age groups
ggplot(dataset, aes(x = Income, fill = age_group)) +
  geom_density(alpha = 0.5) +
  labs(title = "Income Density Distribution by Age Groups", x = "Income", y = "Density")
```



We can see the younger individual mostly occupy the lower income category with the high density.

```r
# density plot income by gender
ggplot(dataset, aes(x = Income, fill = Gender)) +
  geom_density(alpha = 0.5) +
  labs(title = "Income Density Distribution by Gender", x = "Income", y = "Density")
```



The income between Male and Female is fairly balanced with a slight difference around certain income level like 0-25000 being higher for Female and 100000-125000 for Male.