group10

group10

Table of contents

1 Introduction

The primary objective of this analysis is to investigate the relationship between net income from interest, dividends, and profit from capital investments (`py090n`)(unemployment benefits?) and the predictors `gender`, `citizenship`, `hsize` (household size), and `age` in the region of West Austria. Descriptive statistics will be used to understand the distribution and relationships among the variables in this subset of the EUSILC-P dataset.

Methods of analysis include univariate visualizations, bivariate comparisons, and the exploration of potential interactions among predictors to guide subsequent regression modeling.

2 Data Collection

The dataset originates from the EUSILC-P survey, which collects comprehensive social and economic data.

- **Survey Type**: Longitudinal survey
- **Data Characteristics**:
  - Variables: `py090n`, `gender`, `citizenship`, `hsize`, `age`
  - Scale Levels: Continuous (`py090n`, `age`, `hsize`), Categorical (`gender`, `citizenship`)
  - Missing Values: Handled using imputation where necessary.

Below is the R code used for loading and preparing the data:

5 univariate, 5 bivariate , 4 joint

```
library(dplyr)
```

```
Warning: package 'dplyr' was built under R version 4.3.3
```

```
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
```

```
library(simFrame)
```

```
Warning: package 'simFrame' was built under R version 4.3.3
```

```
Loading required package: Rcpp
```

```
Warning: package 'Rcpp' was built under R version 4.3.2
```

```
Loading required package: lattice
```

```
Loading required package: parallel
```

```r
data(eusilcP)
```

```r
head(eusilcP)
```

	hidregionhsizeeqsizeeqIncomepidid agegenderecoStatcitizenshippy010npy050npy090npy100npy110npy120npy130npy140nhy040nhy050nhy070nhy080nhy090nhy110nhy130nhy145nmain														
39993	Upper Austria 2	1.5	11128.45002501male 1	Other	16692.67 0.00 0	0	0	0	0	0	0.00 0	0	0.00 0.00 0.00 0.00	TRUE
39994	Upper Austria 2	1.5	11128.45002402female 4	AT	0.00 0.00 0.00 0	0	0	0	0	0	0.00 0	0	0.00 0.00 0.00 0.00	FALSE
31024	Styria 2	1.5	19694.85005701female 1	AT	0.00 12564.59 0.00	0	0	0	0	0	0.00 0	0	0.00 0.00 93.63 -187.26	FALSE
31025	Styria 2	1.5	19694.85005302male 1	AT	16884.06 0.00 0	0	0	0	0	0	0.00 0	0	0.00 0.00 93.63 -187.26	TRUE
29071	Styria 1	1.0	5066.24000301female 6	AT	0.00 0.00 5066.24 0	0	0	0	0	0	0.00 0	0	0.00 0.00 0.00 0.00	TRUE
41342	Upper Austria 3	1.8	31480.01003201male 1	AT	25047.09 0.00 0	0	0	0	0	0	7167.39 0		31.15 1340.90 0.00	TRUE

```r
# Filter dataset to include only entries from the West Austria region
west_austria <- eusilcP %>%
  filter(region %in% c("Vorarlberg", "Tyrol", "Salzburg", "Upper Austria"))

west_austria <- west_austria %>% select(gender, citizenship, hsize, age, py090n)

# Transform hsize to integer
west_austria$hsize <- as.integer(as.character(west_austria$hsize))

# Summarize to identify potential data quality issues
summary(west_austria)
```

```
   gender       citizenship       hsize            age             py090n
 male  :10555   AT    :15763   Min.   :1.000   Min.   :-1.00   Min.   :    0.0
 female:11121   EU    :  430   1st Qu.:2.000   1st Qu.:20.00   1st Qu.:    0.0
                Other: 1335    Median :3.000   Median :39.00   Median :    0.0
                NA's : 4148    Mean   :3.324   Mean   :38.89   Mean   :  375.1
                               3rd Qu.:4.000   3rd Qu.:56.00   3rd Qu.:    0.0
                               Max.   :9.000   Max.   :94.00   Max.   :26589.4
                                                               NA's   :4148
```

Filtered data contains a lot of NAs. Dropping it all together may damage the possible underlying patterns. Let's check what are those values exaclty, maybe there is some relations between NAs

```
rows_with_NA <- west_austria %>% filter(if_any(everything(), is.na))
summary(rows_with_NA)
```

```
   gender      citizenship      hsize            age             py090n
male  :2200   AT    :   0   Min.   :2.000   Min.   :-1.000   Min.   : NA
female:1948   EU    :   0   1st Qu.:4.000   1st Qu.: 4.000   1st Qu.: NA
              Other:    0   Median :4.000   Median : 8.000   Median : NA
              NA's :4148    Mean   :4.355   Mean   : 7.829   Mean   :NaN
                            3rd Qu.:5.000   3rd Qu.:12.000   3rd Qu.: NA
                            Max.   :9.000   Max.   :15.000   Max.   : NA
                                                             NA's   :4148
```

Interestingly, all the data, which contains at lease one NA value in a row consists of children (age summary ranges between -1 and 15), which explains NAs in py090n as it represents unemployment benefits(children are not eligible for unemployment benefits). Also, all the NAs in citizenship are children respectively. Maybe this subset of data directly represents all of the children from the data. Let's check everyone with age less than 16.

```
rows_with_age_below_16 <- west_austria %>% filter(age < 16)
summary(rows_with_age_below_16)
```

```
   gender      citizenship      hsize            age             py090n
male  :2200   AT    :   0   Min.   :2.000   Min.   :-1.000   Min.   : NA
female:1948   EU    :   0   1st Qu.:4.000   1st Qu.: 4.000   1st Qu.: NA
              Other:    0   Median :4.000   Median : 8.000   Median : NA
              NA's :4148    Mean   :4.355   Mean   : 7.829   Mean   :NaN
                            3rd Qu.:5.000   3rd Qu.:12.000   3rd Qu.: NA
                            Max.   :9.000   Max.   :15.000   Max.   : NA
                                                             NA's   :4148
```

Indeed, our hypothesis have been confirmed as we can directly see that all persons below 16 are the same persons from previous analysis as it includes the same NA's. So it will make sense to totally remove this subset of the data as it represents children, who are not eligible for unemployment benefits

```r
west_austria <- west_austria[complete.cases(west_austria), ]

summary(west_austria)
```

```
   gender      citizenship       hsize           age            py090n       
male  :8355   AT    :15763   Min.   :1.00   Min.   :16.00   Min.   :    0.0  
female:9173   EU    :  430   1st Qu.:2.00   1st Qu.:32.00   1st Qu.:    0.0  
              Other: 1335   Median :3.00   Median :45.00   Median :    0.0  
                            Mean   :3.08   Mean   :46.24   Mean   :  375.1  
                            3rd Qu.:4.00   3rd Qu.:60.00   3rd Qu.:    0.0  
                            Max.   :9.00   Max.   :94.00   Max.   :26589.4  
```

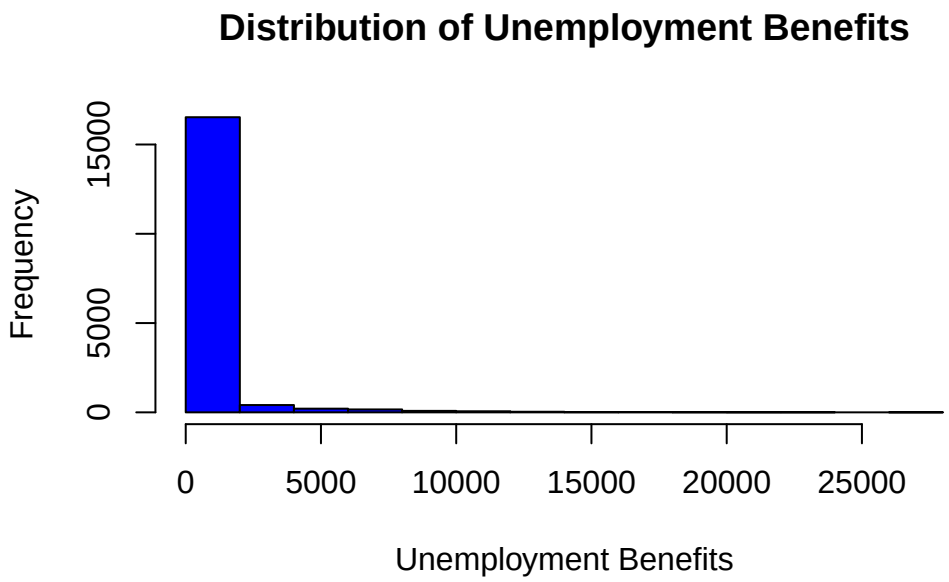
3 Descriptive Analysis of the Sample

3.1 Univariate Analysis

3.1.1 Net Income (py090n)

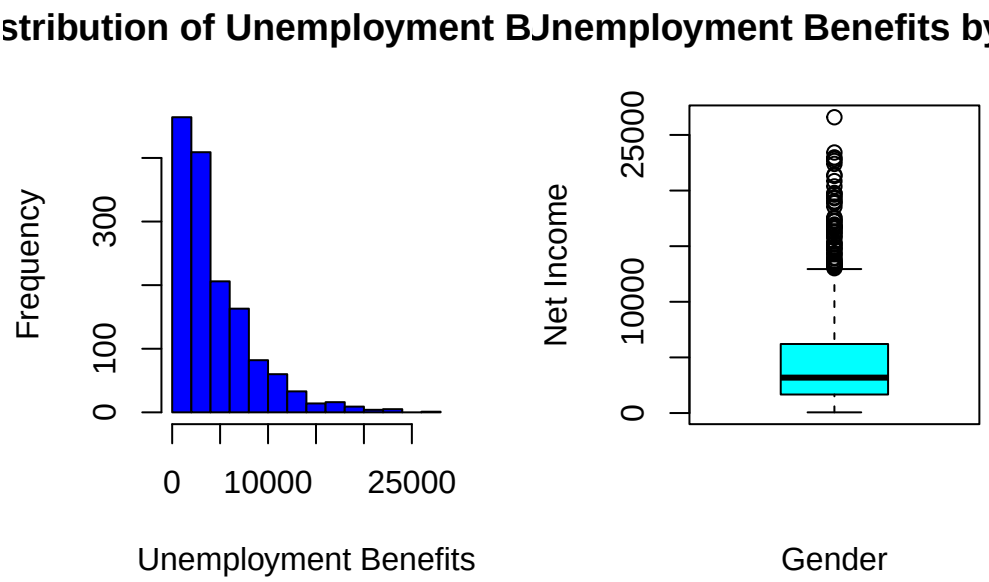
Net income represents the earnings from interest, dividends, and capital investments. Its distribution helps understand the financial background of individuals in West Austria.

```r
# Create a histogram to visualize the distribution of net income
hist(west_austria$py090n,
     col = "blue",
     main = "Distribution of Unemployment Benefits",
     xlab = "Unemployment Benefits",
     ylab = "Frequency")
```



As there are too many persons with 0 unemployment benefits, plotting with zero value included does not provide a lot of information, we can plo it excluding zero value

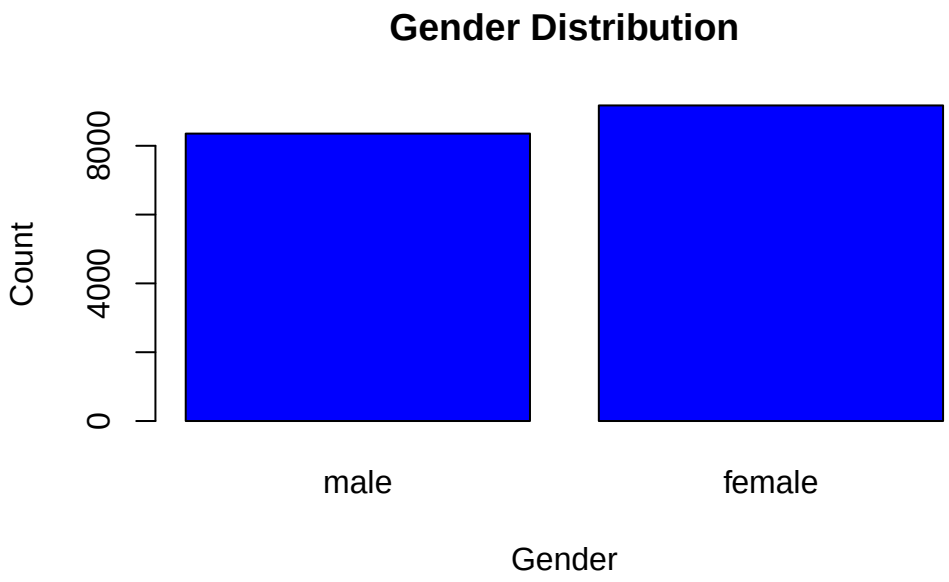
```
par(mfrow = c(1, 2))

# Histogram for non-zero unemployment benefits
hist(west_austria$py090n[west_austria$py090n != 0],
     col = "blue",
     main = "Distribution of Unemployment Benefits",
     xlab = "Unemployment Benefits",
     ylab = "Frequency")

# Boxplot for non-zero unemployment benefits by gender
boxplot(west_austria$py090n[west_austria$py090n != 0],
        col = "cyan",
        main = "Unemployment Benefits by Gender",
        xlab = "Gender",
        ylab = "Net Income")
```



3.1.2 Gender Distribution

The `gender` variable indicates whether individuals are male or female. The distribution provides insight into the gender representation in the dataset.

```
# Create a bar plot to show the distribution of genders in the dataset
table_gender <- table(west_austria$gender)
barplot(table_gender,
        main = "Gender Distribution",
        xlab = "Gender",
        ylab = "Count",
        col = "blue")
```

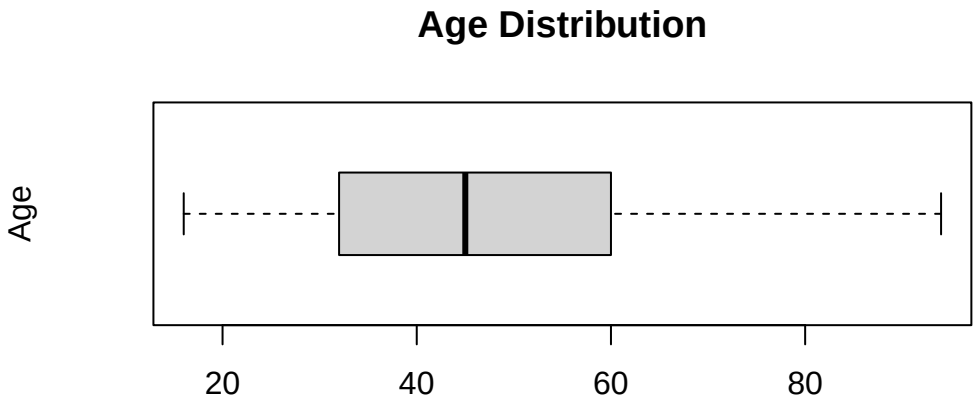


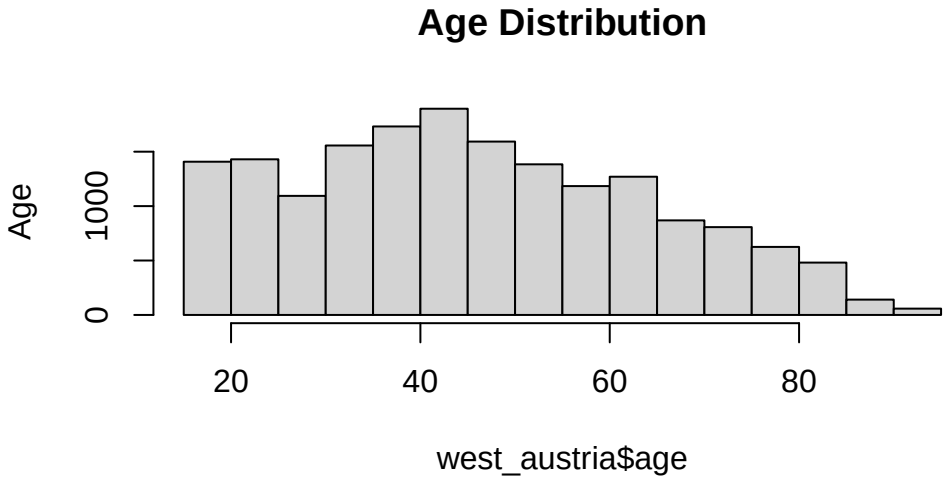
Both categories have approximately equal distribution.

3.1.3 Age Distribution

Age represents the individual's age at the time of the survey. Its distribution gives insights into the demographic structure of the dataset.

```
par(mfrow = c(2, 1))
boxplot(west_austria$age,
        horizontal = TRUE,
        main = "Age Distribution",
        ylab = "Age")
hist(west_austria$age,
        main = "Age Distribution",
        ylab = "Age")
```

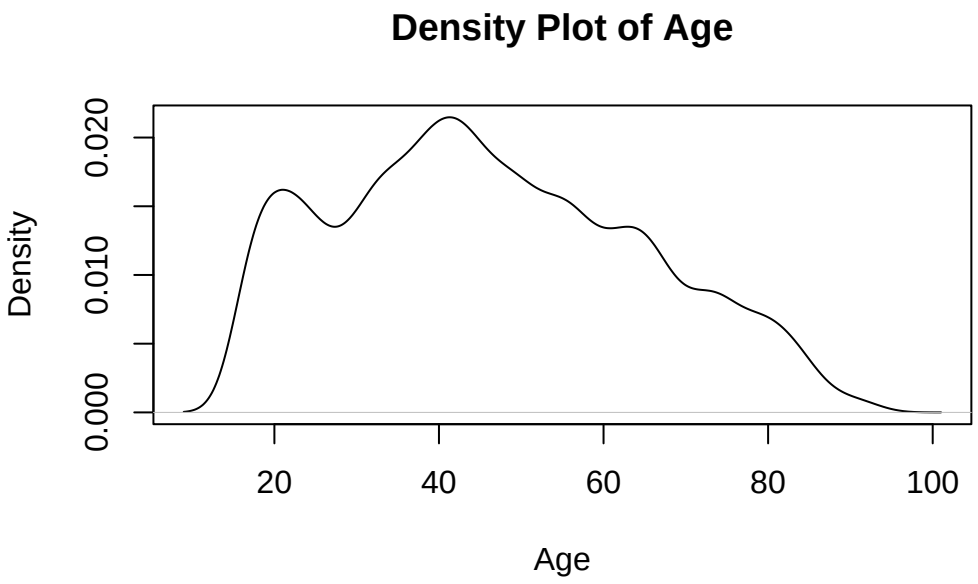





```
# Summarize the age variable
summary(west_austria$age)
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
16.00   32.00   45.00   46.24   60.00   94.00
```

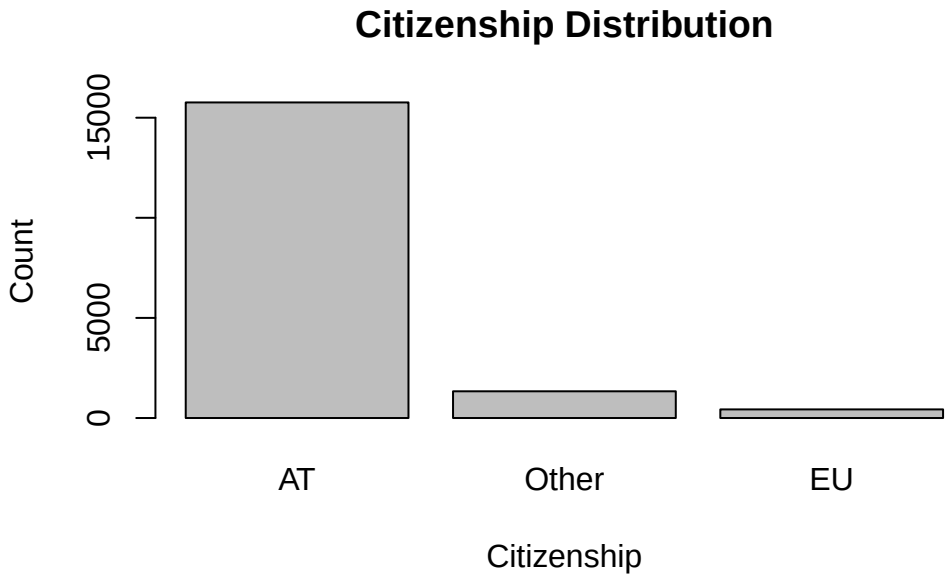
```r
# Create a density plot to visualize the distribution of age fill graph
plot(density(west_austria$age),
     main = "Density Plot of Age",
     xlab = "Age",
     ylab = "Density"
     )
```



3.1.4 Citizenship

The `citizenship` variable differentiates between Austrian citizens and foreigners. This distribution helps understand the dataset's demographic diversity.

```r
# Create a bar plot to show the distribution of citizenship statuses
# including the citizenship labels in the axis
# label NA as unknown
# sort by count, citizenship is a factor
table_citizenship <- table(west_austria$citizenship)
barplot(sort(table_citizenship, decreasing = TRUE),
        main = "Citizenship Distribution",
        xlab = "Citizenship",
        ylab = "Count")
```



3.1.5 Household Size (`hsize`)

Household size represents the number of people in a household. Its distribution is essential to analyze living conditions.

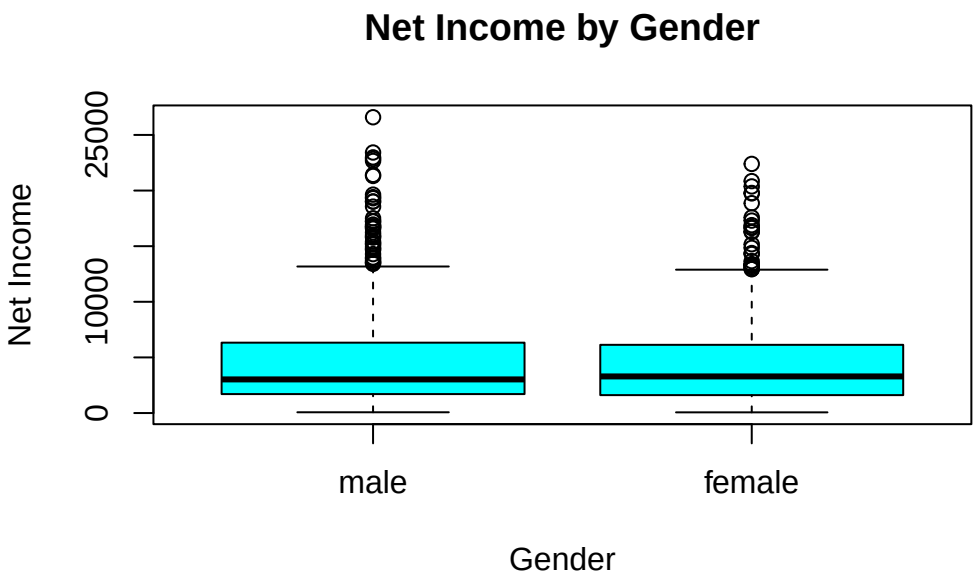
```
# Create a histogram to visualize the distribution of household sizes
table_hsize <- table(west_austria$hsize)
barplot(table_hsize,
        main = "Household Size Distribution",
        xlab = "Household Size",
        ylab = "Count")
```

3.2 Bivariate Analysis

3.2.1 Gender and Net Income

This comparison helps understand the income distribution across genders.

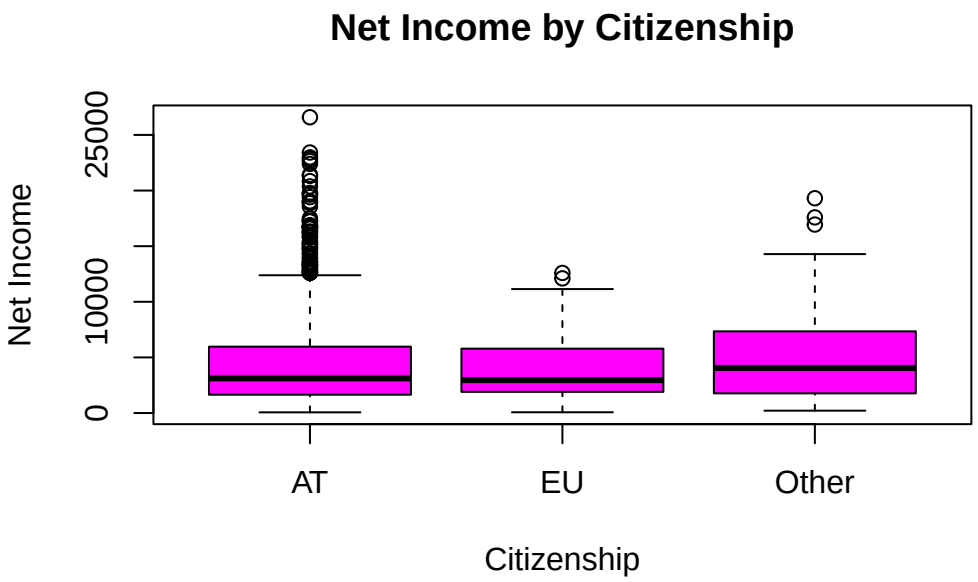
```
# Create boxplots to compare net income across genders
boxplot(py090n ~ gender,
        data = west_austria,
        subset = py090n != 0,
        col = "cyan",
        main = "Net Income by Gender",
        xlab = "Gender",
        ylab = "Net Income")
```



3.2.2 Citizenship and Net Income

This analysis highlights the income differences between Austrian citizens and foreigners.

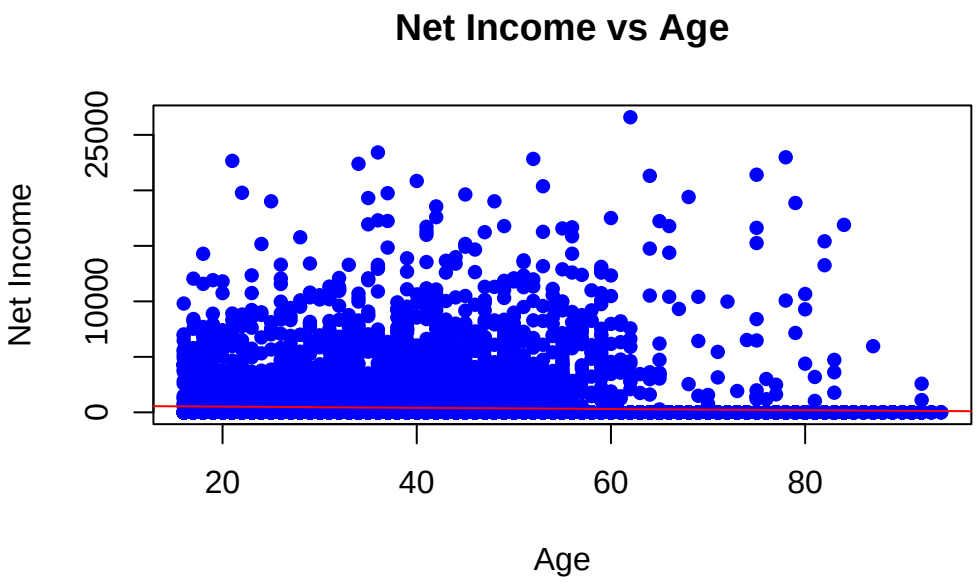
```
# Create boxplots to compare net income across citizenship statuses
boxplot(py090n ~ citizenship,
        data = west_austria,
        subset = py090n != 0,
        col = "magenta",
        main = "Net Income by Citizenship",
        xlab = "Citizenship",
        ylab = "Net Income")
```



3.2.3 Age and Net Income

Exploring this relationship helps identify trends or patterns in income with respect to age.

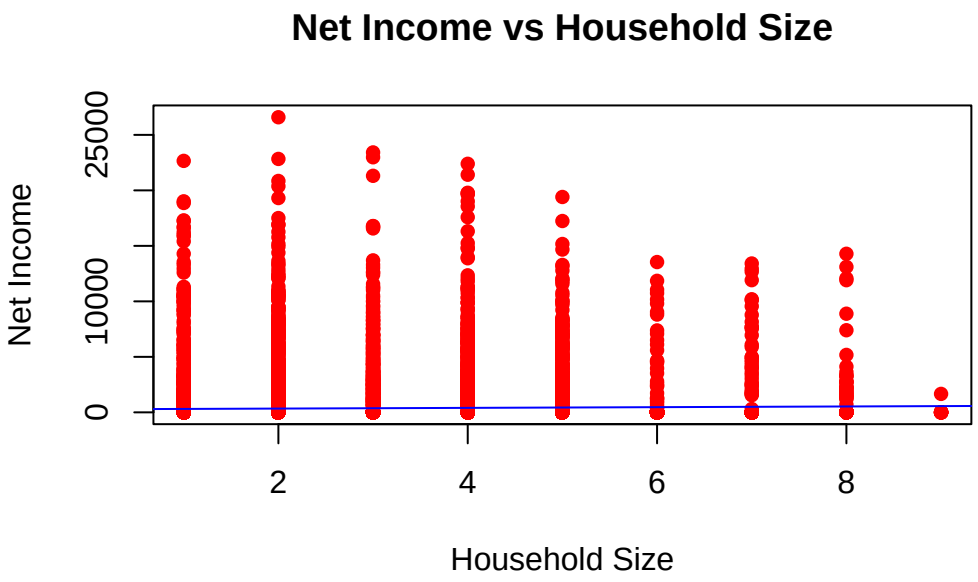
```
# Create a scatter plot to explore the relationship between age and net income
plot(west_austria$age, west_austria$py090n,
     col = "blue",
     pch = 16,
     main = "Net Income vs Age",
     xlab = "Age",
     ylab = "Net Income")
abline(lm(py090n ~ age, data = west_austria), col = "red")
```



3.2.4 Household Size and Net Income

Analyzing this relationship provides insights into how income varies with household size.

```
# Create a scatter plot to explore the relationship between household size and net income
plot(west_austria$hsize, west_austria$py090n,
     col = "red",
     pch = 16,
     main = "Net Income vs Household Size",
     xlab = "Household Size",
     ylab = "Net Income")
abline(lm(py090n ~ hsize, data = west_austria), col = "blue")
```



4 Summary

The descriptive analysis reveals key patterns in the dataset. The distribution of net income is skewed, and there are noticeable differences in income across genders and citizenships. Age and household size appear to have linear relationships with income. These findings set the stage for deeper inferential analysis in the final report.