Interim report: EUSILC-P Dataset Analysis

Anton Shapovalov, Richard Maria

Table of contents


```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
```

```
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ---------------------------------------- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(simFrame)
```

```
Warning: Paket 'simFrame' wurde unter R Version 4.4.2 erstellt
```

```
Lade nötiges Paket: Rcpp
Lade nötiges Paket: lattice
Lade nötiges Paket: parallel
```

```
library(forcats)
library(gridExtra)
```

```
Warning: Paket 'gridExtra' wurde unter R Version 4.4.2 erstellt
```

```
Attache Paket: 'gridExtra'

Das folgende Objekt ist maskiert 'package:dplyr':

    combine
```

```
library(patchwork)
```

```
Warning: Paket 'patchwork' wurde unter R Version 4.4.2 erstellt
```

```
data(eusilcP)

# modify default theme
theme_set(theme_light() + theme(plot.title = element_text(hjust = 0.5)))
```

1 Introduction

The primary objective of this analysis is to investigate the relationship between net income from unemployment benefits in national currency (€) per year and the predictors `gender`, `citizenship`, `hsize` (household size), and `age` (in years) in the region of West Austria. We focus primarily on the group of people receiving benefits, 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: `benefits` (*py090n*), `gender`, `citizenship`, `hsize`, `age`
  - Scale Levels: Numerical(`benefits`, `age`, `hsize`), Categorical (`gender`, `citizenship`)
  - Missing Values: Handled using imputation where necessary.

5 univariate, 5 bivariate , 4 joint

```
# 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))

# rename py090n to benefits
west_austria <- west_austria %>% rename(benefits = py090n)

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

```
   gender      citizenship       hsize            age           benefits
 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            benefits
 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 `benefits` 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            benefits
 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

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

summary(west_austria)
```

```
    gender      citizenship       hsize            age            benefits
 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
```

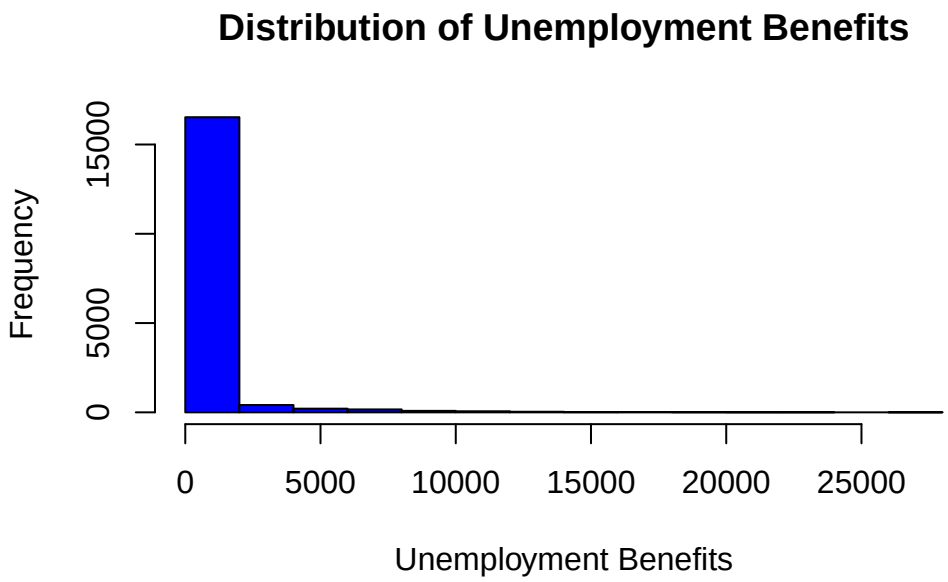
3 Descriptive Analysis

3.1 Univariate Analysis

3.1.1 Net Income from Unemployed Benefits (`benefits`)

Net income from unemployment benefits (**benefits**) is the primary variable of interest. A histogram is used to visualize the distribution of benefits.

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



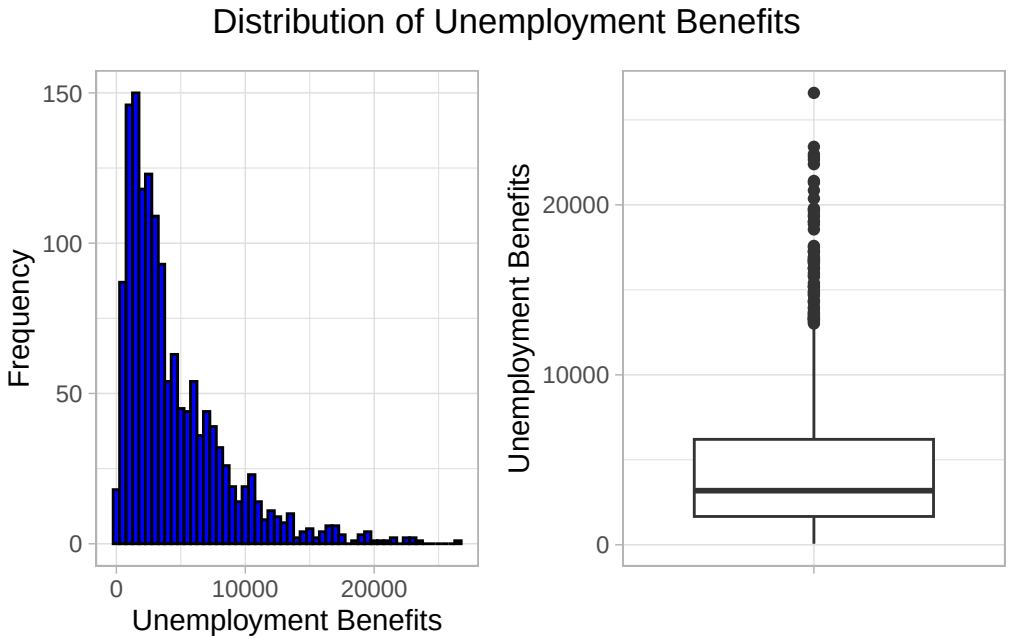
As there are too many persons with 0 unemployment benefits, plotting with zero value included does not provide a lot of information, we can plot it excluding zero values.

```
west_austria_filtered <- subset(west_austria, benefits != 0)

# Histogram for non-zero unemployment benefits
p1 <- ggplot(west_austria_filtered, aes(x = benefits)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black") +
  labs(x = "Unemployment Benefits", y = "Frequency")

# Boxplot for non-zero unemployment benefits by gender
p2 <- ggplot(west_austria_filtered, aes(x = "", y = benefits)) +
  geom_boxplot() +
  labs(x="", y="Unemployment Benefits")

# Arrange plots side by side
(p1 | p2) + plot_annotation(title = "Distribution of Unemployment Benefits")
```




```
summary(west_austria_filtered$benefits)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   61.9  1666.6  3184.3  4485.4  6204.5 26589.4
```

The distribution of non zero net income from unemployment benefits is left-skewed, with a mean of €4485.4 and a median of €3184.3. The majority of individuals receive benefits below €6204.5.

3.1.2 Age Distribution

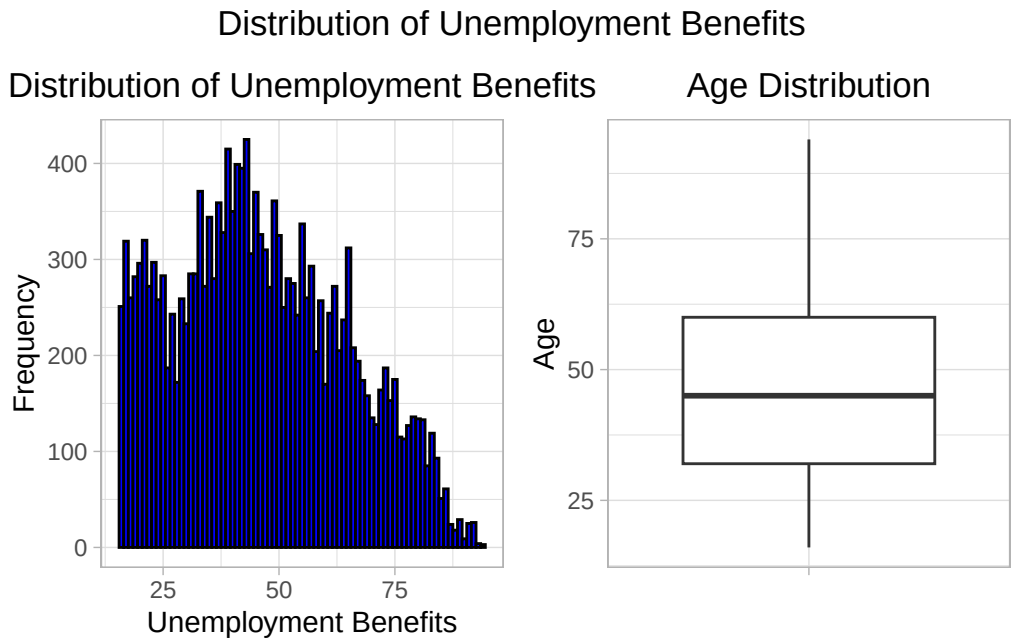
Age represents the individual's age at the time of the survey. Its distribution gives insights into the demographic structure of the dataset.

```
# Histogram for age distribution
p1 <- ggplot(west_austria, aes(x = age)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  labs(title = "Distribution of Unemployment Benefits",
       x = "Unemployment Benefits", y = "Frequency")

# Boxplot for age distribution
```

```r
p2 <- ggplot(west_austria, aes(x = "", y = age)) +
  geom_boxplot() +
  labs(title = "Age Distribution",
       x = "", y = "Age")

# Arrange plots side by side
(p1 | p2) + plot_annotation(title = "Distribution of Unemployment Benefits")
```




```r
summary(west_austria$age)
```

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

The age distribution has a median of 45 years, with a mean of 46.24 years. The majority of individuals are between 32 and 60 years old, with a minimum age of 16 and a maximum age of 94. The distribution is slightly right-skewed.

```r
summary(west_austria_filtered$age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   28.00   39.00   39.36   49.00   92.00
```

The age distribution has a mean age of 39.36 years. The majority of individuals are between 28 and 49 years old. The lower bound of 16 comes from the fact that children are not eligible for unemployment benefits. The oldest person in the dataset is 92 years old. This could indicate an error in the data, as people with that age usually do not receive unemployment benefits.

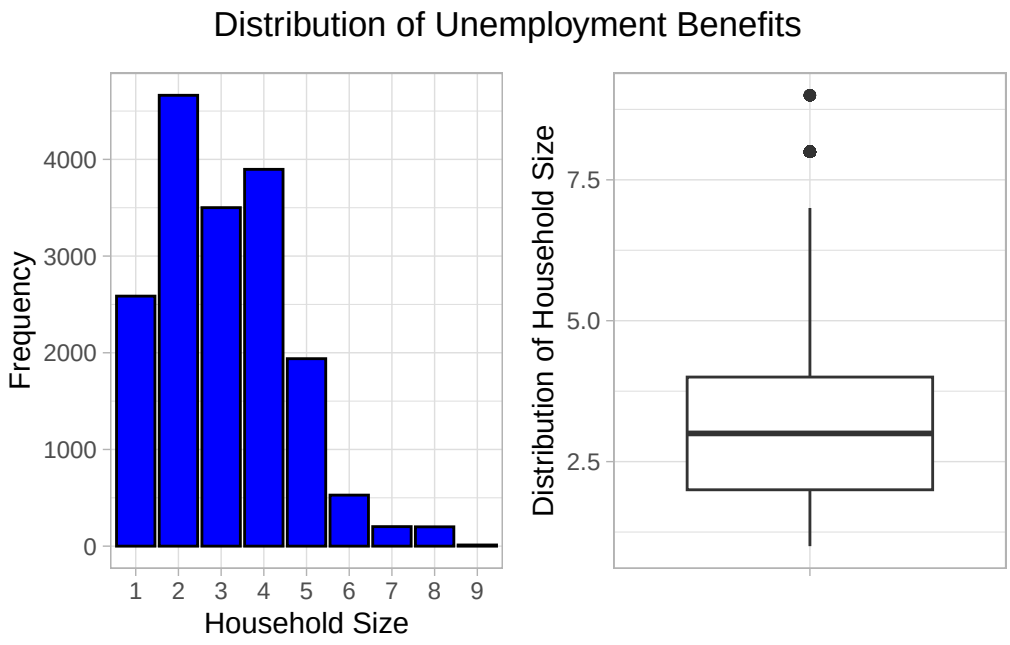
3.1.3 Household Size (`hsize`)

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

```
# Histogram for houshold size
p1 <- ggplot(west_austria, aes(x = factor(hsize))) +
  geom_bar(fill = "blue", color = "black") +
  labs(x = "Household Size", y = "Frequency")

# Boxplot for household size
p2 <- ggplot(west_austria, aes(x = "", y = hsize)) +
  geom_boxplot() +
  labs(x="", y="Distribution of Household Size")

# Arrange plots side by side
(p1 | p2) + plot_annotation(title = "Distribution of Unemployment Benefits")
```



```r
summary(west_austria$hsize)
```

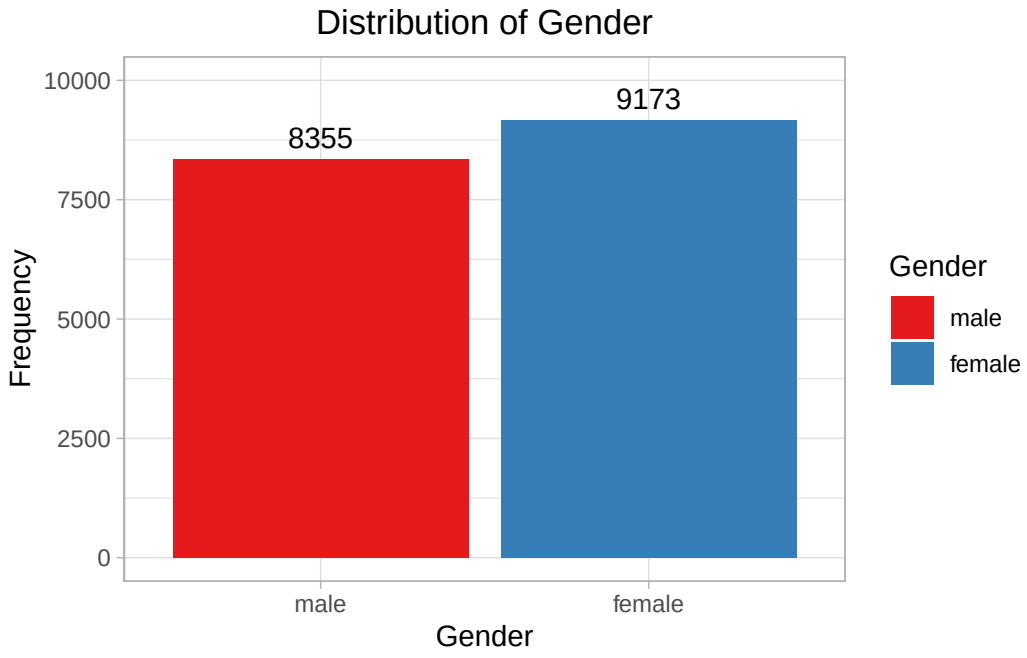
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.00    3.00    3.08    4.00    9.00
```

The distribution of household size in western austria has a median of 3 people per household. The majority of households have between 2 and 4 members. The minimum value of 1 indicates that some individuals live alone. A household size of more than 4 is less common.

3.1.4 Gender Distribution

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

```r
# Create a bar plot to show the distribution of genders in the dataset
table_gender <- table(west_austria$gender)
ggplot(west_austria, aes(x = gender, fill = gender)) +
  geom_bar() +
  scale_fill_brewer(palette = "Set1") +
  labs(title = "Distribution of Gender", x = "Gender", y = "Frequency", fill="Gender") +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.5)+
  ylim(0, 10000)
```



```r
summary(west_austria$gender)
```

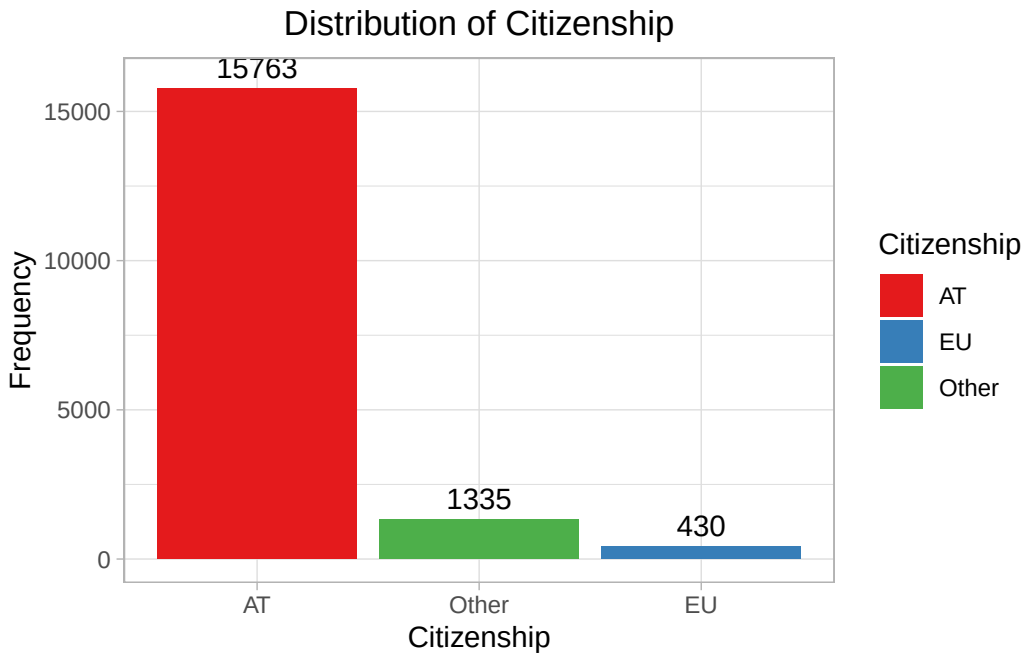
```
  male female
  8355   9173
```

Both categories have approximately equal distribution. This balance is essential for ensuring representativeness in the dataset.

3.1.5 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)
ggplot(west_austria, aes(x = fct_infreq(citizenship), fill = citizenship)) +
  geom_bar() +
  scale_fill_brewer(palette = "Set1") +
  labs(title = "Distribution of Citizenship", x = "Citizenship", y = "Frequency", fill="Citi
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.5) +
  ylim(0, 16000)
```




```
summary(west_austria$citizenship)
```

```
   AT    EU Other
15763   430  1335
```

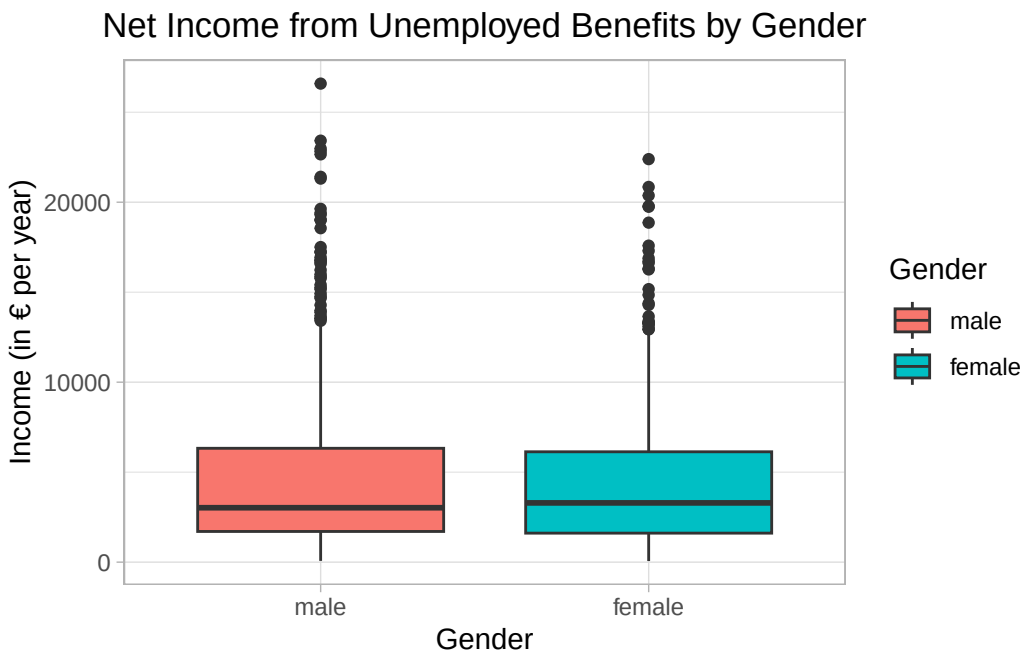
The majority of individuals in the dataset are Austrian citizens, followed by citizens from other countries outside the European Union. A smaller proportion of individuals are from other countries inside the EU.

3.2 Bivariate Analysis

3.2.1 Gender and Net Income from Unemployed Benefits

This comparison helps understand the income distribution across genders.

```
# Create boxplots to compare net income across genders
ggplot(west_austria_filtered, aes(x = gender, y = benefits, fill = gender)) +
  geom_boxplot() +
  labs(title = "Net Income from Unemployed Benefits by Gender",
       x = "Gender",
       y = "Income (in € per year)",
       fill="Gender")
```



On a first glance it seems that there is no significant difference between the unemployed benefits between male and female. Let's check the summary statistics to confirm this.

```
bartlett.test(benefits ~ gender, data = west_austria)
```

```
    Bartlett test of homogeneity of variances

data:  benefits by gender
Bartlett's K-squared = 35.371, df = 1, p-value = 2.725e-09
```

The p-value is less than 0.05, which indicates that the variances of the two groups are significantly different. Therefore, we should use the Welch's t-test to compare the means of the two groups.

```
t.test(benefits ~ gender, data = west_austria)
```

```
    Welch Two Sample t-test

data:  benefits by gender
t = -0.36706, df = 17106, p-value = 0.7136
```

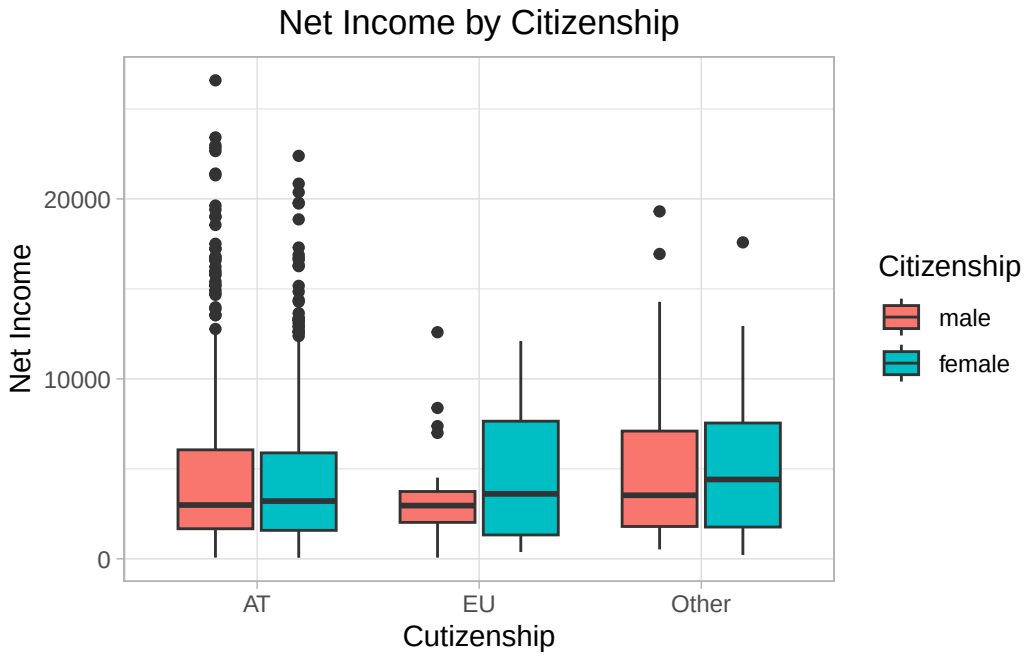
```
alternative hypothesis: true difference in means between group male and group female is not e
95 percent confidence interval:
 -59.70875  40.87338
sample estimates:
  mean in group male mean in group female
            370.2212             379.6389
```

The t-test results suggest no significant difference in the average benefits between male and female groups. While the sample means differ slightly, the variation within the data is too large relative to the difference for it to be meaningful in a statistical sense.

3.2.2 Citizenship and Net Income from Unemployed Benefits

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

```r
# Create boxplots to compare net income across citizenship statuses
ggplot(west_austria_filtered, aes(x = citizenship, y = benefits, fill = gender)) +
  geom_boxplot() +
  labs(title = "Net Income by Citizenship",
       x = "Cutizenship",
       y = "Net Income",
       fill = "Citizenship")
```



In this plot we see great differences between the groups citizenship, gender and the unemployed benefits. The group of Austrian citizens has a higher median and a smaller range of values compared to the other groups. The group of citizens from other countries inside the EU has the lowest median and the highest range of values. The group of citizens from other countries outside the EU has a median between the other two groups and a range of values similar to the group of citizens from other countries inside the EU.

```
bartlett.test(benefits ~ citizenship, data = west_austria)
```

```
    Bartlett test of homogeneity of variances

data:  benefits by citizenship
Bartlett's K-squared = 301.61, df = 2, p-value < 2.2e-16
```

The p-value is less than 0.05, which indicates that the variances of the three groups are significantly different. Therefore, we should use the Welch's ANOVA to compare the means of the three groups.

```
oneway.test(benefits ~ citizenship, data = west_austria)
```

```
    One-way analysis of means (not assuming equal variances)

data:  benefits and citizenship
F = 16.705, num df = 2.00, denom df = 919.66, p-value = 7.475e-08
```

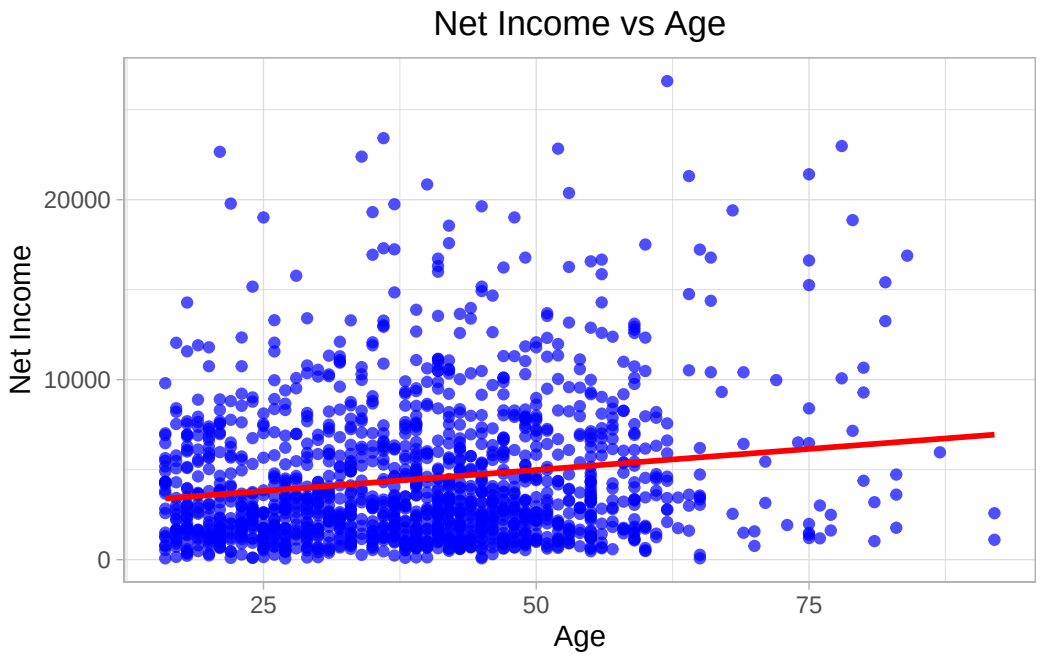
The p-value is less than 0.05, which indicates that there is a significant difference in the average benefits between the three groups.

3.2.3 Age and Net Income

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

```
ggplot(west_austria_filtered, aes(x = age, y = benefits)) +
  geom_point(color = "blue", alpha = 0.7) +  # Scatter points
  geom_smooth(method = "lm", color = "red", se = FALSE) +  # Regression line
  labs(title = "Net Income vs Age",
       x = "Age",
       y = "Net Income")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



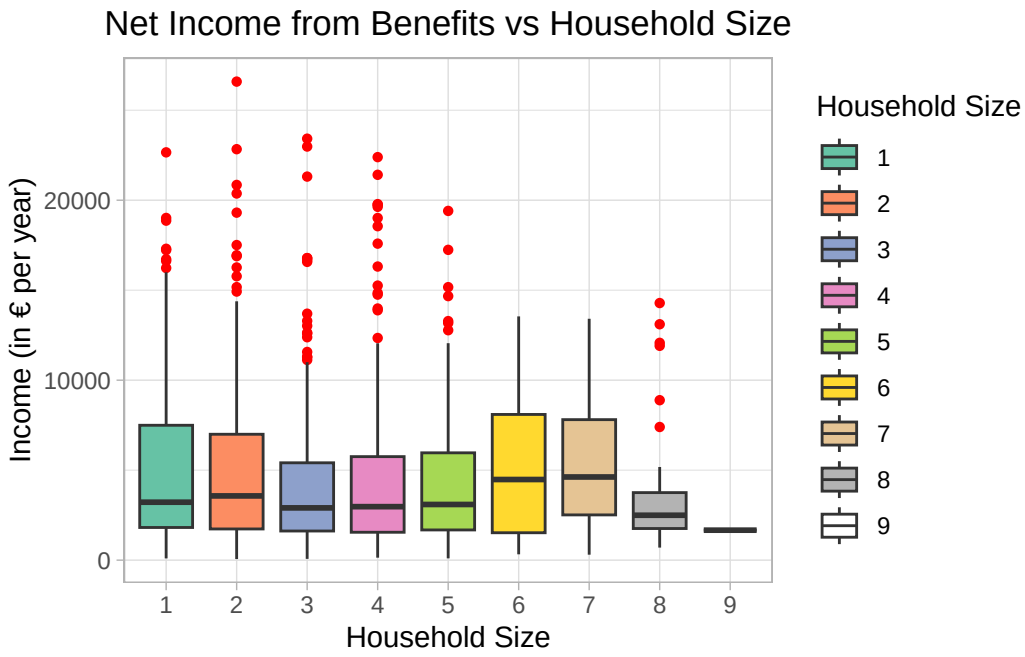
The scatter plot shows a slight positive relationship between age and net income from unemployment benefits. The regression line indicates that older individuals tend to have slightly higher benefits.

3.2.4 Household Size and Net Income from Benefits

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
ggplot(west_austria_filtered, aes(x = factor(hsize), y = benefits, fill = factor(hsize))) +
  geom_boxplot(outlier.color = "red", outlier.shape = 16) +
  scale_fill_brewer(palette = "Set2") +  # Adjust fill colors
  labs(title = "Net Income from Benefits vs Household Size",
       x = "Household Size",
       y = "Income (in € per year)",
       fill = "Household Size")
```

```
Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum for palette Set2 is
Returning the palette you asked for with that many colors
```



The boxplot shows that the median net income from benefits is highest for households with 6-7 members. The range of net income is also wider for households with fewer members. This could be due to the presence of outliers in households with fewer people.

```
bartlett.test(benefits ~ hsize, data = west_austria)
```

```
    Bartlett test of homogeneity of variances

data:  benefits by hsize
Bartlett's K-squared = 204.16, df = 8, p-value < 2.2e-16
```

Bartlett's test indicates that the variances of the nine groups are significantly different. Therefore, we should use the Welch's ANOVA to compare the means of the nine groups.

```
oneway.test(benefits ~ hsize, data = west_austria)
```

```
    One-way analysis of means (not assuming equal variances)

data:  benefits and hsize
F = 3.5272, num df = 8.0, denom df = 260.9, p-value = 0.0006799
```

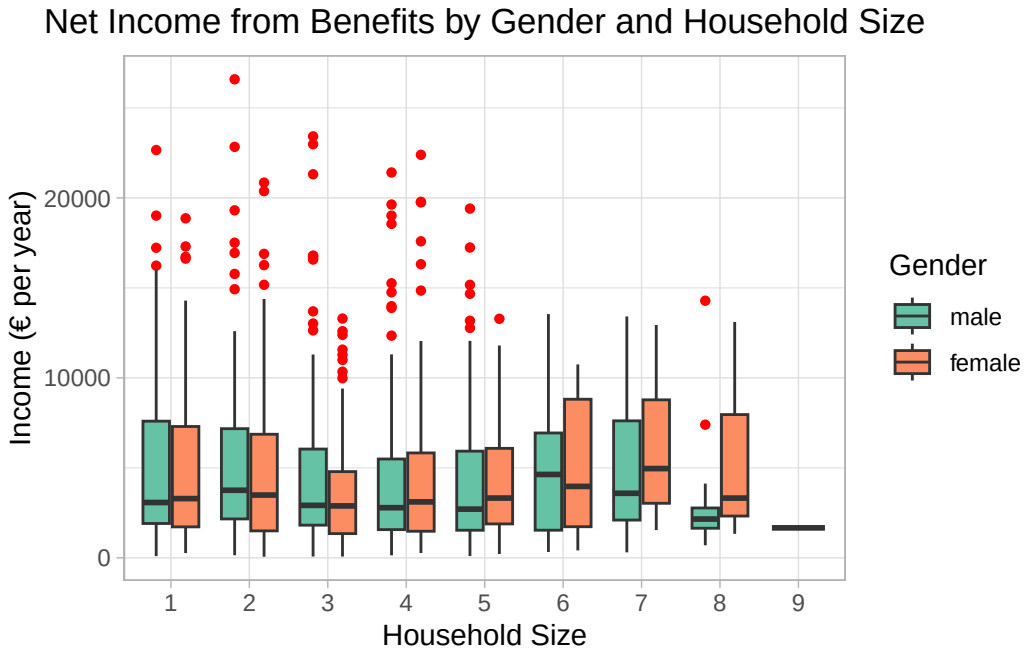
With a p-value of 0.0006799, we can reject the null hypothesis that the means of the nine groups are equal. This suggests that there is a significant difference in the average benefits across different household sizes.

3.2.5 Joint variables

3.2.6 Household Size, Gender and Net Income from Benefits

Analyzing the relationship of Gender, Household Size and Net Income from Benefits.

```
p1 <- ggplot(west_austria_filtered, aes(x = factor(hsize), y = benefits, fill = gender)) +
  geom_boxplot(outlier.color = "red", outlier.shape = 16) +
  labs(title = "Net Income from Benefits by Gender and Household Size",
       x = "Household Size",
       y = "Income (€ per year)",
       fill = "Gender") +
  scale_fill_brewer(palette = "Set2")
p1
```



This plot shows the relationship between Gender, Household Size and Benefits. In this representation it is hard to tell which household size has the highest benefits for unemployed people. But what we can see is, that female have a higher median of benefits than male in households

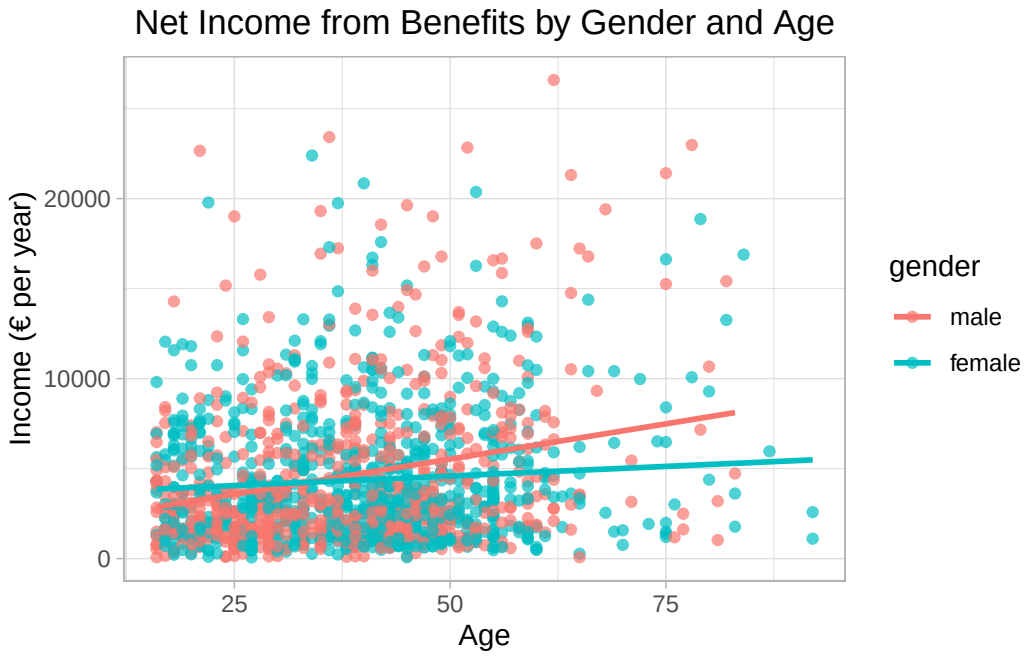
with 1, 2, 4, 7 and 8 members. In households with 3, 5 and 6 members male have a higher median benefit. We also can see that no matter how big the household size is, male have a higher range of benefits than female persons.

3.2.7 Gender, Age and Net Income from Benefits

Analyzing the relationship between Gender, Age and Net Income from Benefits including regression lines for male and female property of the gender variable.

```
p2 <- ggplot(west_austria_filtered, aes(x = age, y = benefits, color = gender)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Net Income from Benefits by Gender and Age",
       x = "Age",
       y = "Income (€ per year)",
       fill = "Gender")
p2
```

```
`geom_smooth()` using formula = 'y ~ x'
```



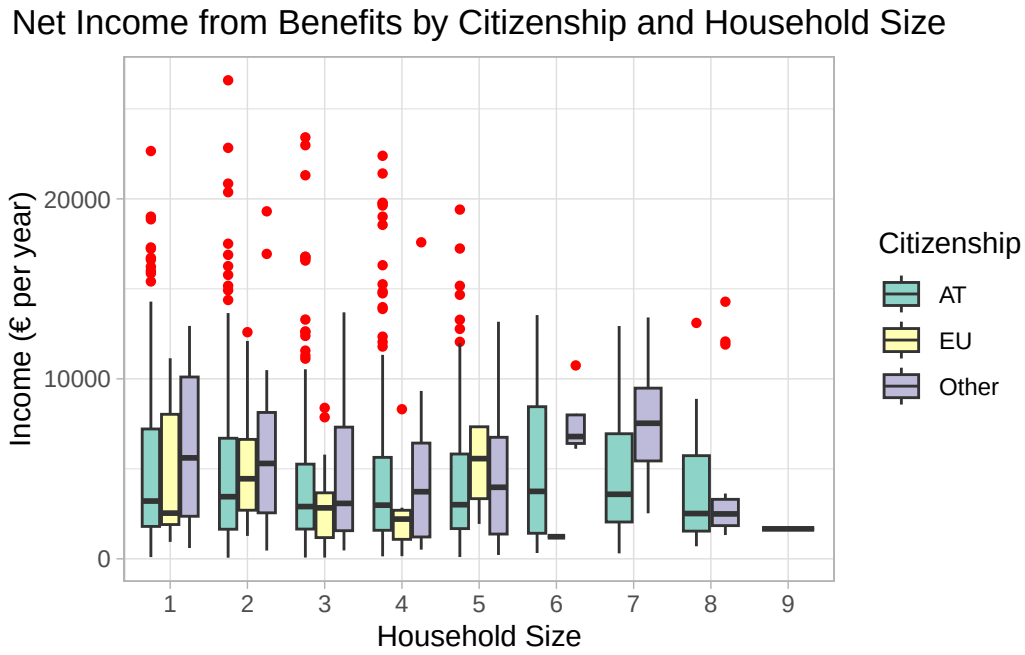
The regression line for females begins at a higher intercept compared to the regression line for males. However, the male regression line has a steeper slope, surpassing the female regression

line at approximately 35 years of age. As unemployed benefits in austria is based on income from the past, this could indicate that female persons have a lower income than male.

3.2.8 Citizenship, Household Size and Net Income from Benefits

Analyzing the relationship between Citizenship, Household Size and Net Income from Benefits.

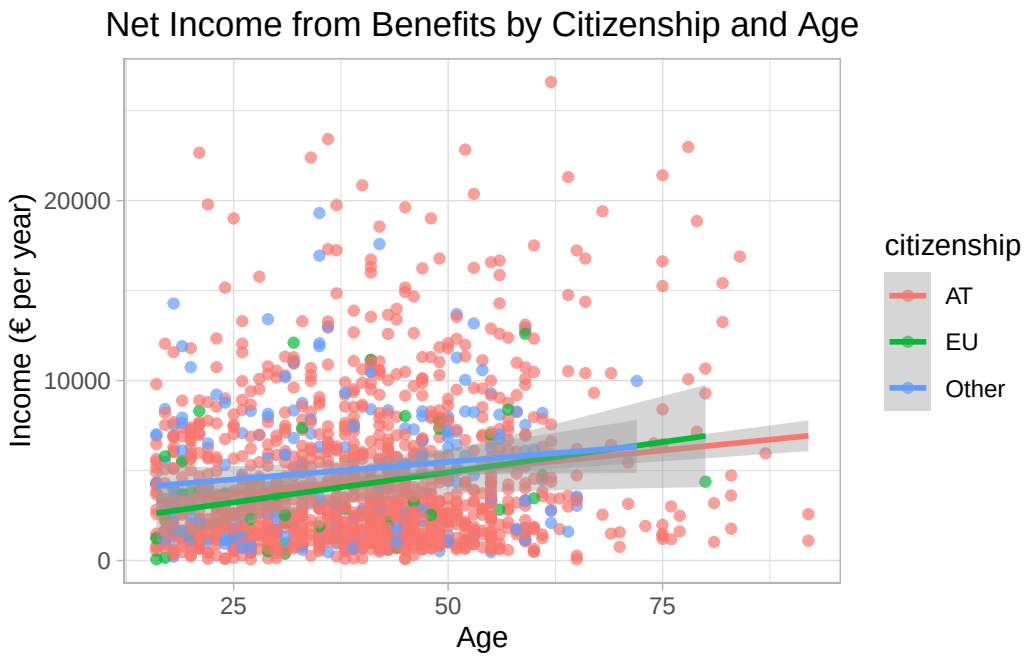
```
p3 <- ggplot(west_austria_filtered, aes(x = factor(hsize), y = benefits, fill = citizenship))
  geom_boxplot(outlier.color = "red", outlier.shape = 16) +
  labs(title = "Net Income from Benefits by Citizenship and Household Size",
       x = "Household Size",
       y = "Income (€ per year)",
       fill = "Citizenship") +
  scale_fill_brewer(palette = "Set3")
p3
```



3.2.9 Citizenship, Age and Net Income from Benefits

Analyzing the relationship between Citizenship, Age and Net Income from Benefits.

```r
p4 <- ggplot(west_austria_filtered, aes(x = age, y = benefits, color = citizenship)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = TRUE, formula = y~x) +
  labs(title = "Net Income from Benefits by Citizenship and Age",
       x = "Age",
       y = "Income (€ per year)",
       fill = "Citizenship")
p4
```



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.