Interim report: EUSILC-P Dataset Analysis

Anton Shapovalov, Richard Maria

Table of contents

1 Introduction

The primary objective of this analysis is to investigate the relationship between unemployment benefits in € per year and the following variables: `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. It is worth noting that the dataset used in this analysis is synthetically generated based on real Austrian EU-SILC data. While the survey itself is longitudinal, the synthetic subset utilized in this study represents data from 2006.

The subset used for analysis includes the following variables:

- **Numerical Variables**:
  - `py090n` (renamed to *benefits* for convenience): Unemployment benefits €.
  - `hsize`: Household size.
  - `age`: Age in years.
- **Categorical Variables**:
  - `gender`: Gender of the individual.
  - `citizenship`: Citizenship status.

A detailed analysis of the data and the cleaning steps undertaken are outlined in the following sections.

```
   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
```

The filtered data contains a significant number of NAs. Dropping them altogether may obscure potential underlying patterns. Let's examine these values more closely to determine if there are any relationships between the NAs.

```
   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 rows containing at least one NA value correspond to children (age ranges from -1 to 15). This explains the NAs in the `benefits` variable, as children are not eligible for unemployment benefits. Similarly, all NAs in the `citizenship` variable are also associated with children. It is possible that this subset of data exclusively represents children. To confirm this, let's examine all individuals with an 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 has been confirmed, as all individuals under the age of 16 are the same as those identified in the previous analysis, containing the same NAs. Therefore, it makes sense to completely remove this subset of the data, as it represents children who are not eligible for unemployment benefits.

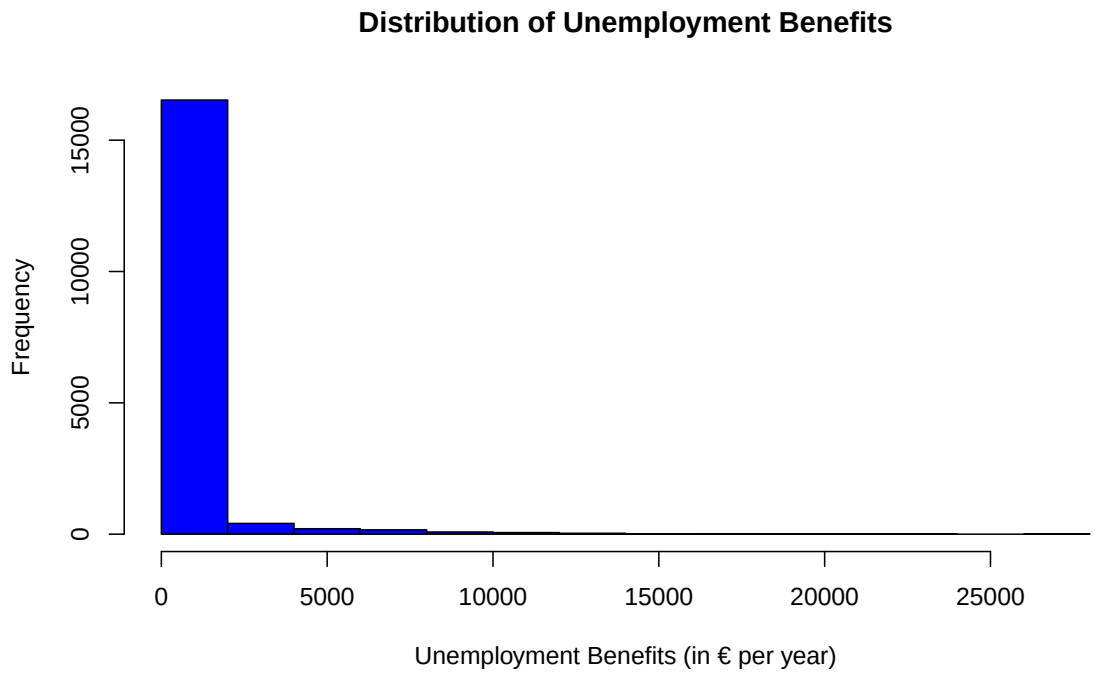
```
   gender      citizenship       hsize            age            benefits
 male  :8355   AT    :15763   Min.   : 1.00   Min.   :16.00   Min.   :    0.0
 female:9173   EU    :  430   1st Qu.: 2.00   1st Qu.:32.00   1st Qu.:    0.0
               Other: 1335    Median : 3.00   Median :45.00   Median :    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

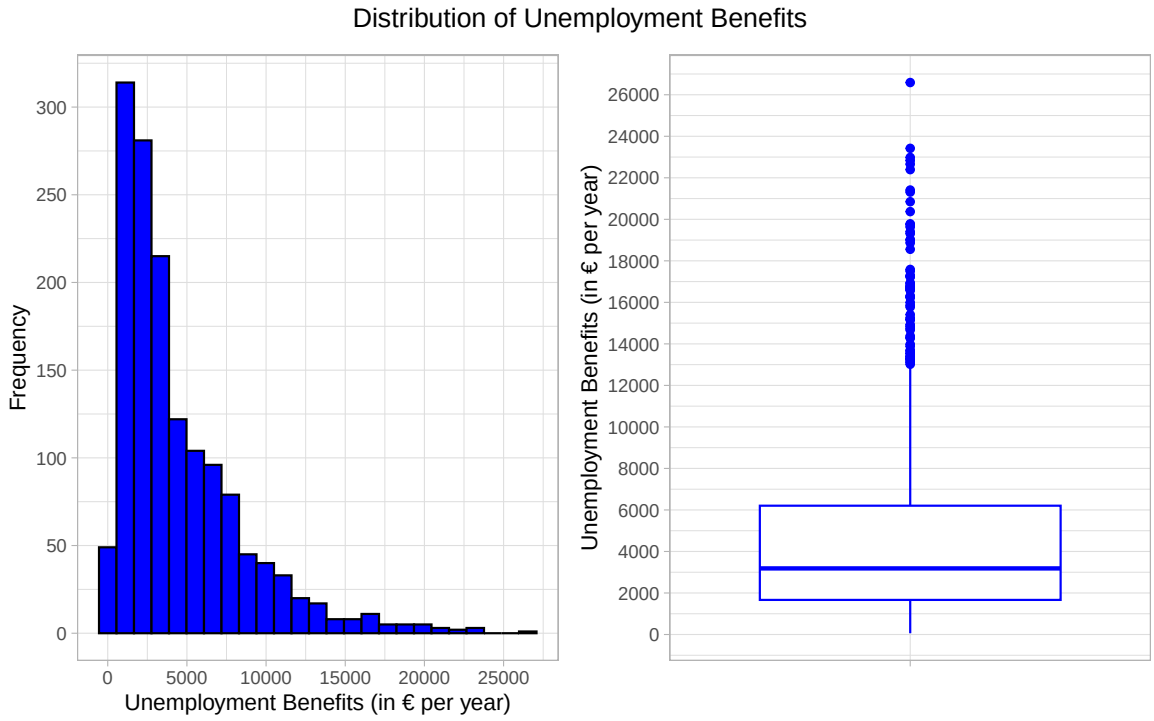
3.1 Univariate Plots

3.1.1 Unemployed Benefits (`benefits`)

Unemployment benefits (`benefits`) is the primary variable of interest. A histogram is used to visualize the distribution of benefits.



Since there are too many individuals with 0 unemployment benefits, including these values in the plot does not provide much meaningful information. Therefore, we can create the plot excluding zero values.

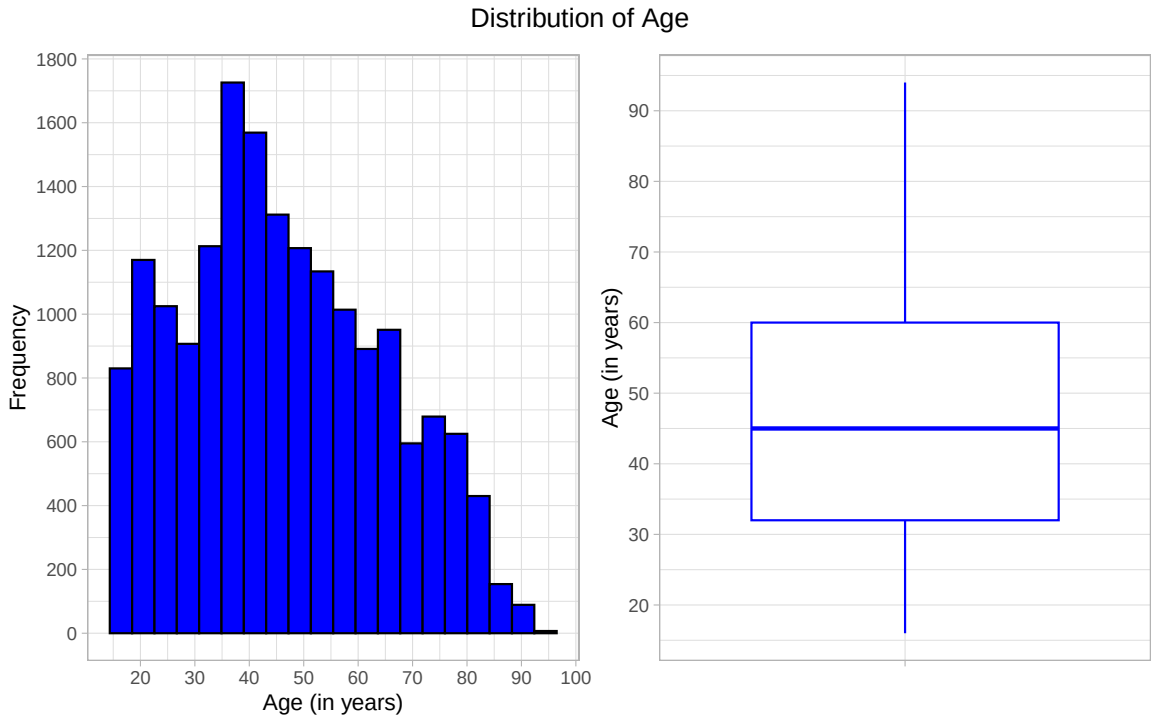



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
61.9  1666.6  3184.3  4485.4  6204.5 26589.4
```

The distribution of non-zero net income from unemployment benefits is right-skewed, with a mean of €4,485.4 and a median of €3,184.3. The histogram indicates that the `benefits` variable follows a log-normal distribution, which may suggest a log transformation for future model building. Additionally, we observe a significant number of outliers at the higher end of the variable's distribution.

3.1.2 Age (`age`)

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

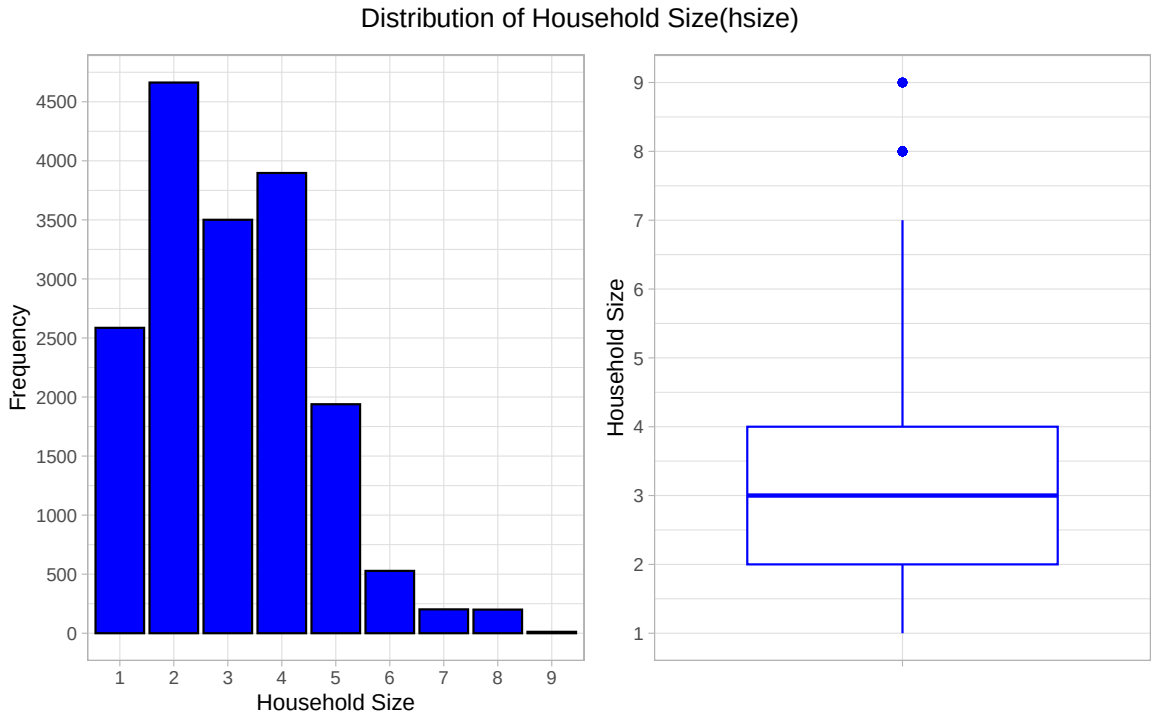



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
16.00   32.00   45.00   46.24   60.00   94.00
```

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. The lower bound of 16 comes from the fact that children are not eligible for unemployment benefits. The oldest person in the dataset is 94 years old. This could indicate an error in the data, as people with that age usually do not receive unemployment benefits. Age box plot does not show any outliers.

3.1.3 Household Size (`hsize`)

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




```
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. Households of size 8 and 9 can be considered outliers.

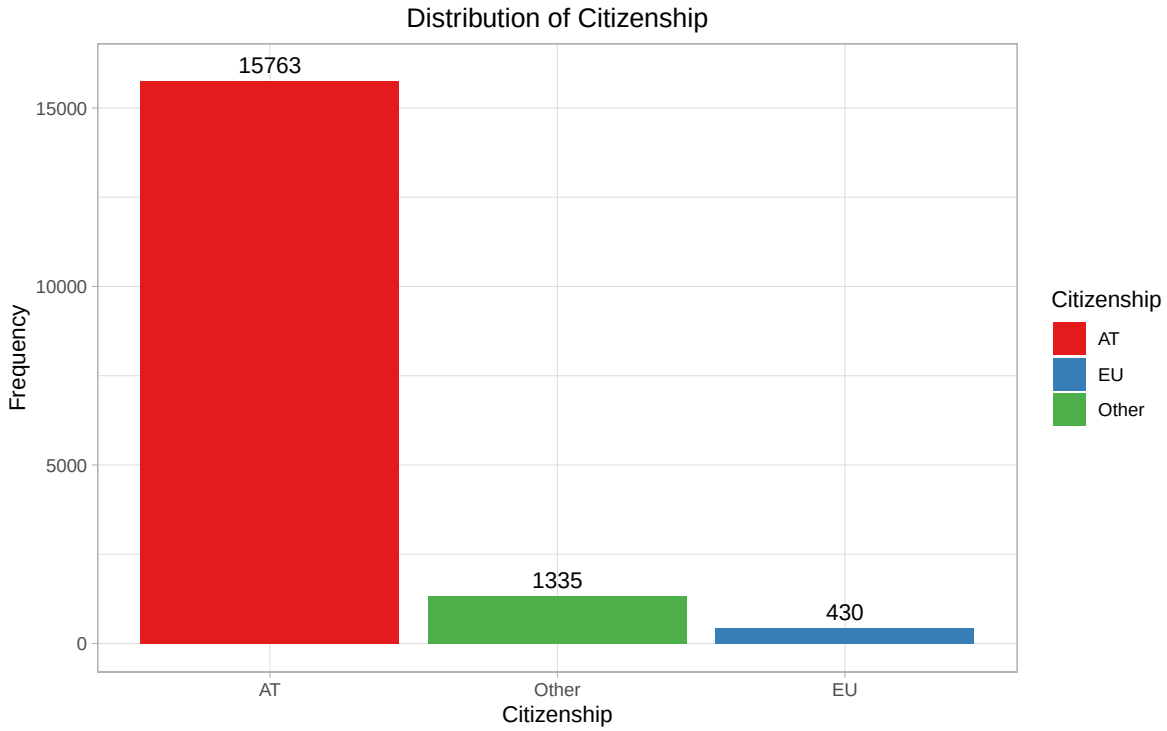
3.1.4 Gender (`gender`)

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

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

3.1.5 Citizenship (`citizenship`)

The `citizenship` variable differentiates between Austrian citizens, EU foreigners and third-country nationals. This distribution helps understand the dataset's demographic diversity.

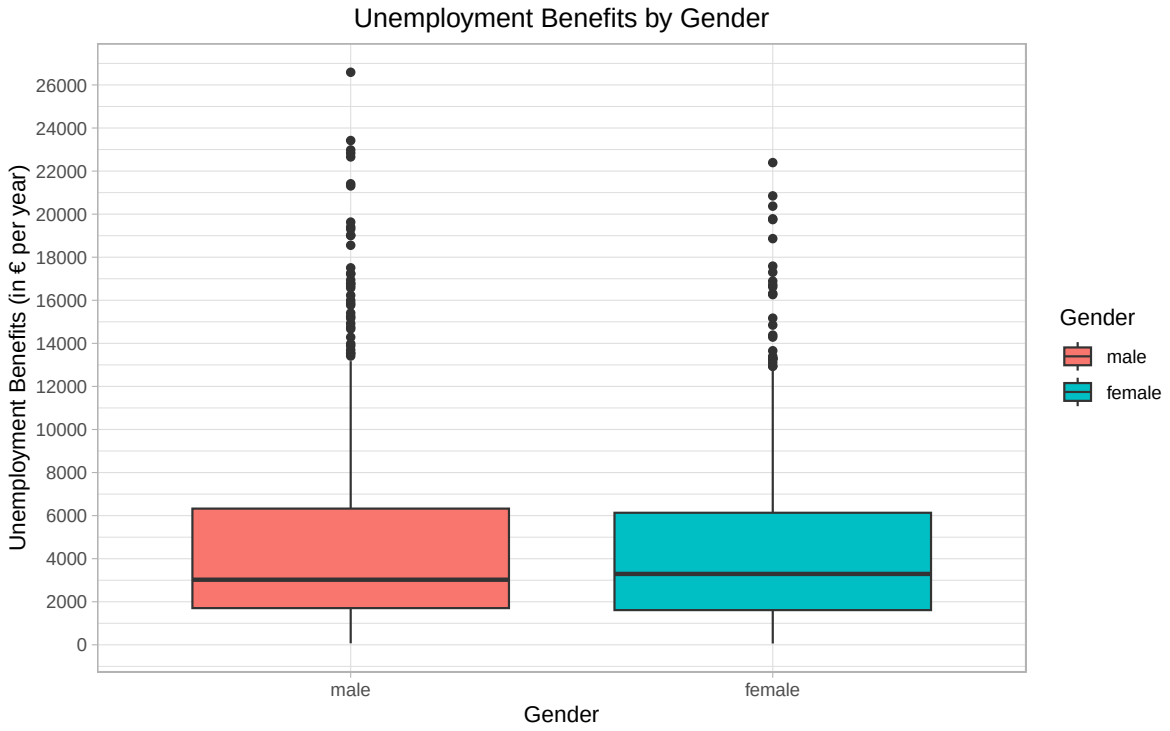


The majority of individuals in the dataset are Austrian citizens, followed by citizens from other countries outside the European Union. A smaller proportion of individuals are from other countries inside the EU.

3.2 Bivariate Plots

3.2.1 Gender x Unemployment Benefits

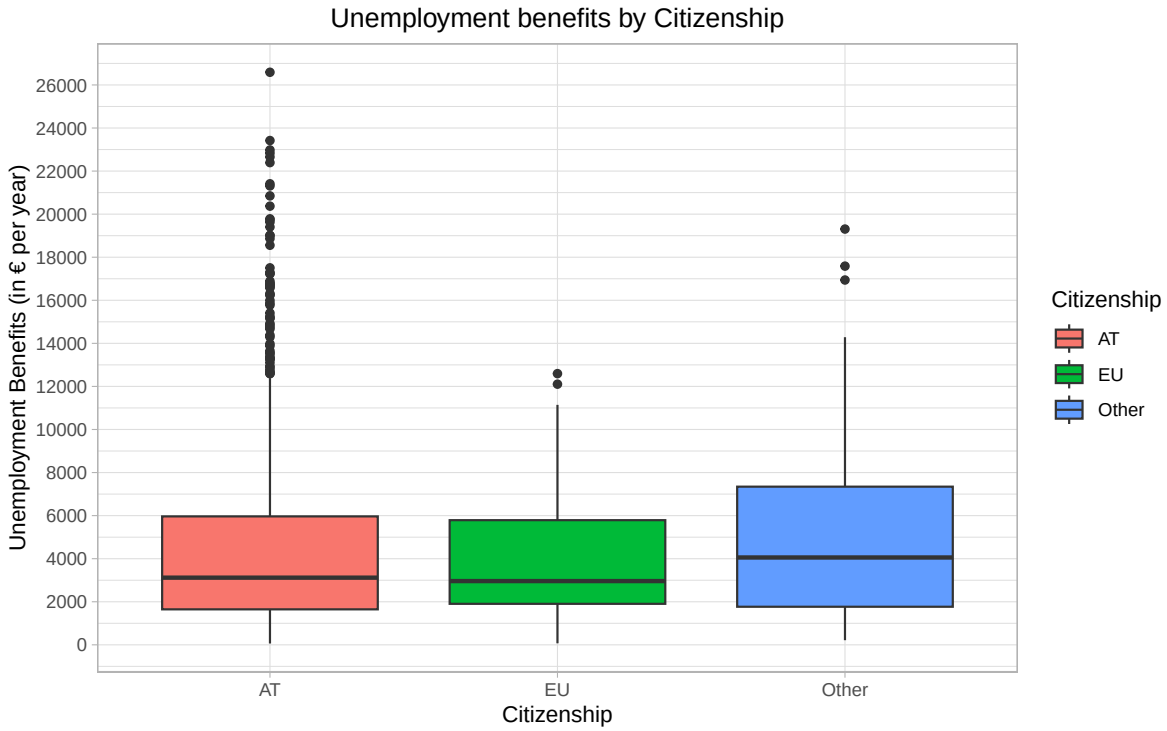
This comparison helps understand the income distribution across genders.



The plot does not indicate significant differences between the two groups. Both groups exhibit numerous outliers, with those in the male group appearing slightly more pronounced.

3.2.2 Citizenship x Unemployment Benefits

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



The "Other" median is noticeably higher than other two categories and "AT" has a large amount of outliers, which may be attributed to highly unbalanced distribution of the given categories.

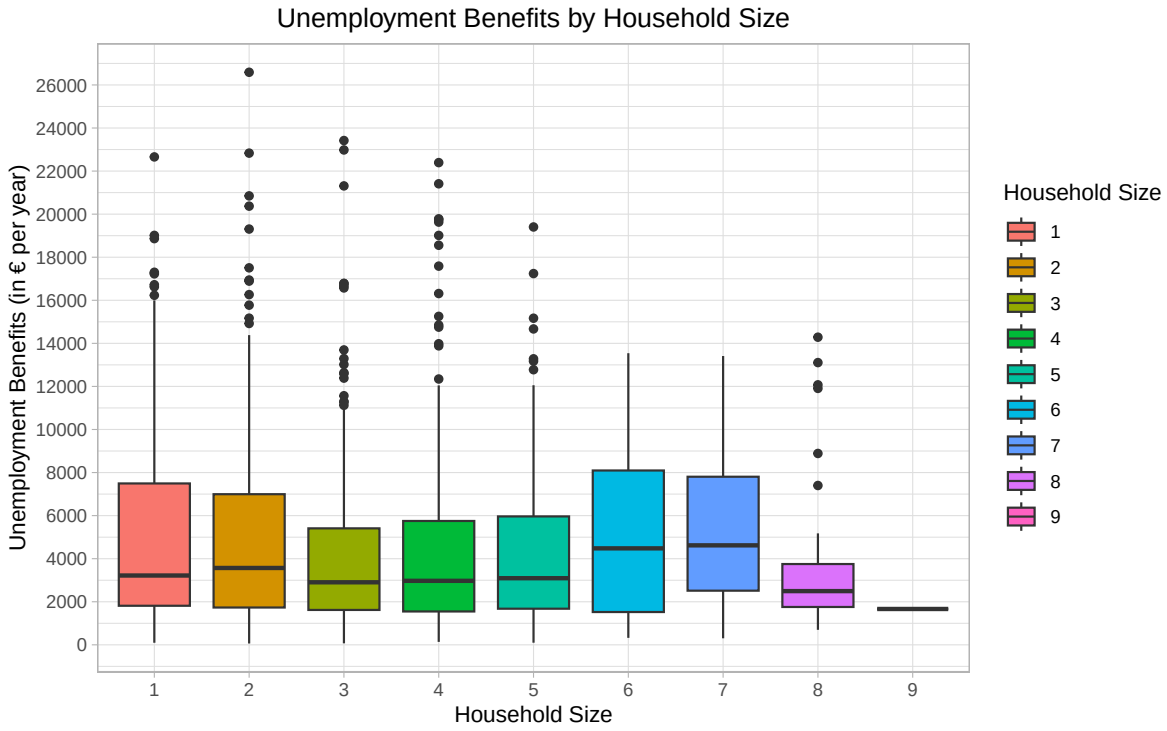
3.2.3 Age x Unemployment Benefits

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

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 x Unemployment Benefits

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

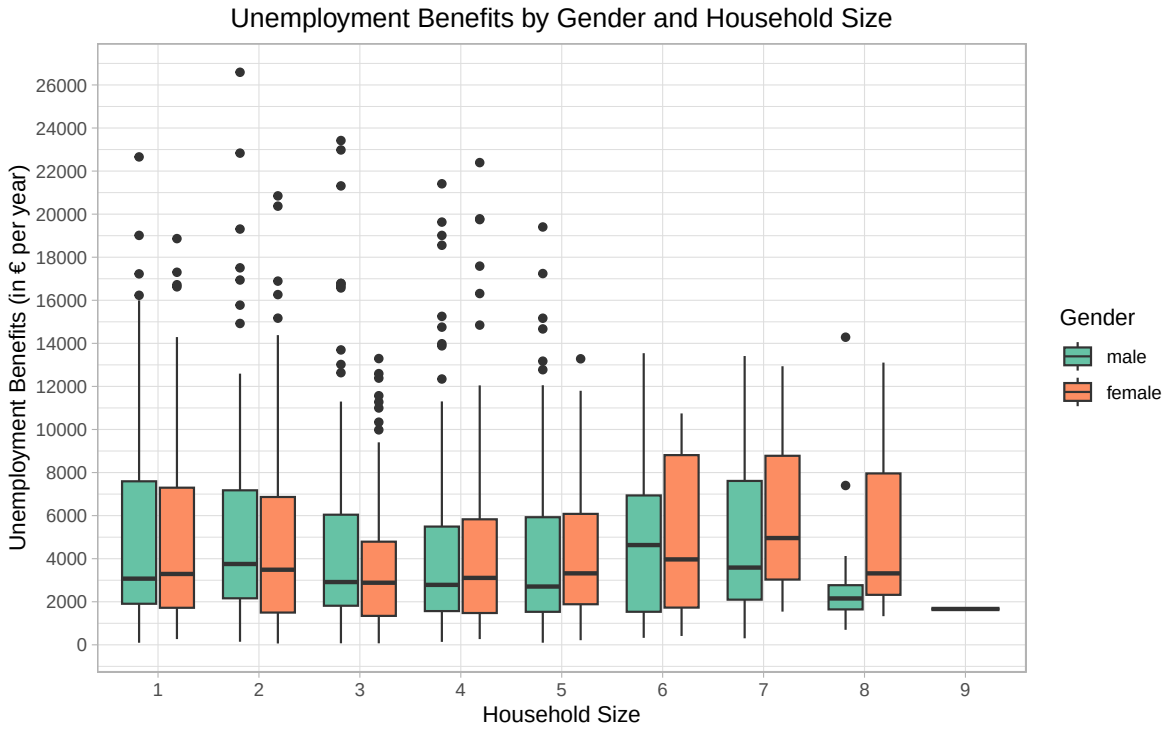


The boxplot shows that the median net income from benefits is highest for households with 6-7 members. The range of net income is also wider for households with fewer members. This could be due to the presence of outliers in households with fewer people.

3.3 Joint Plots

3.3.1 Gender x Household Size x Unemployment Benefits

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



This plot shows the relationship between Gender, Household Size, and Unemployment. An immediate observation is an outlier with a household size of 9, representing a single female sample. Apart from this, greater variability can be observed in smaller households.

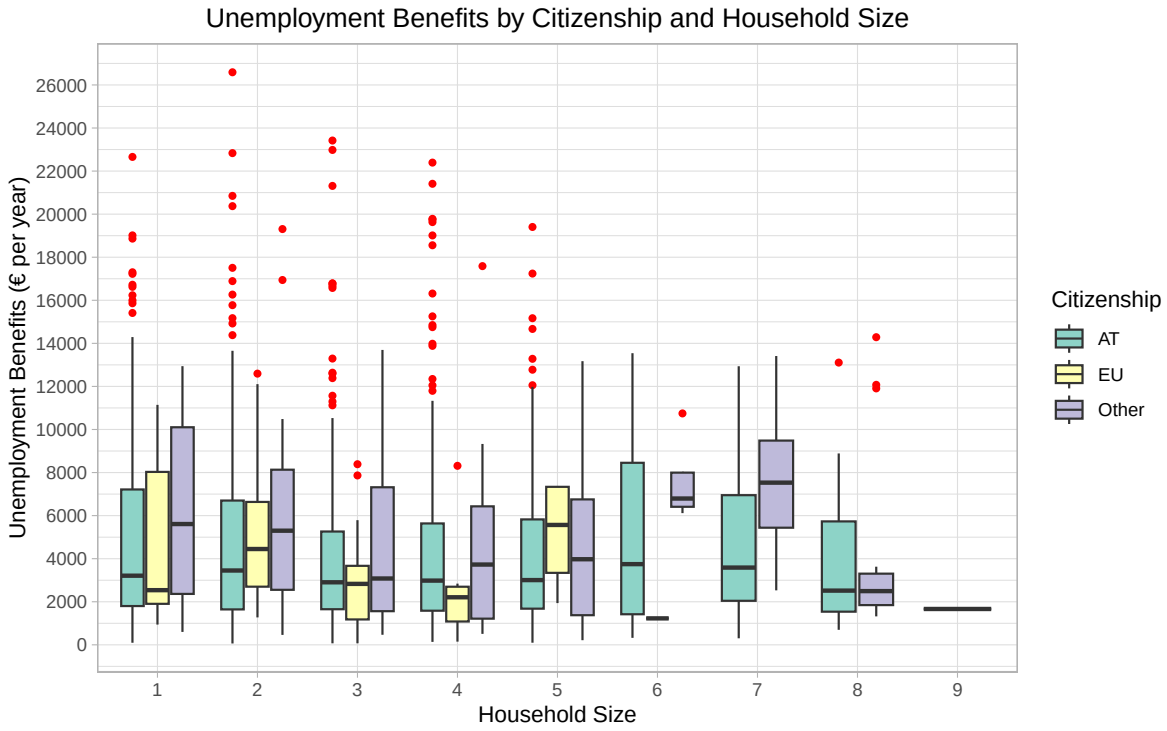
3.3.2 Gender x Age x Unemployment Benefits

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

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. Both groups have similar Standard error bands.

3.3.3 Citizenship x Household Size x Unemployment Benefits

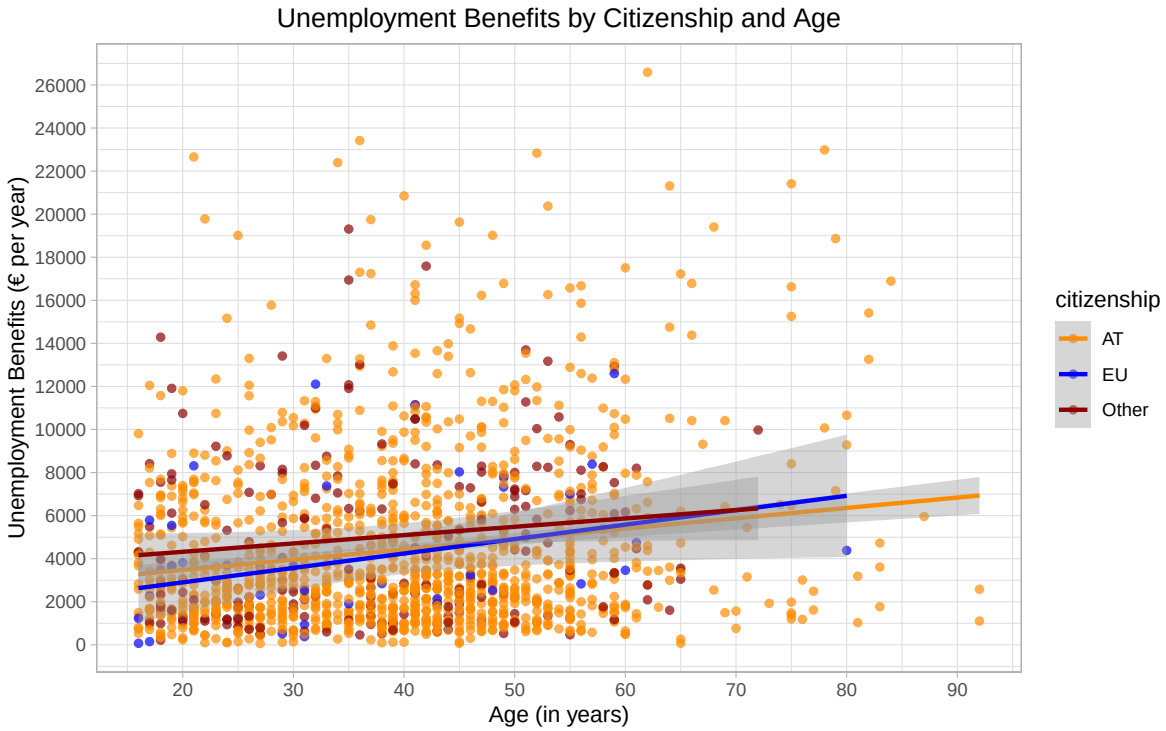
Analyzing the relationship between Citizenship, Household Size and Unemployment Benefits.



We once again observe the previously mentioned outlier with a household size of 9, and generally, there is a lot of variability in the smaller household sizes. It is also worth noting that the 'EU' category is absent for household sizes greater than 6.

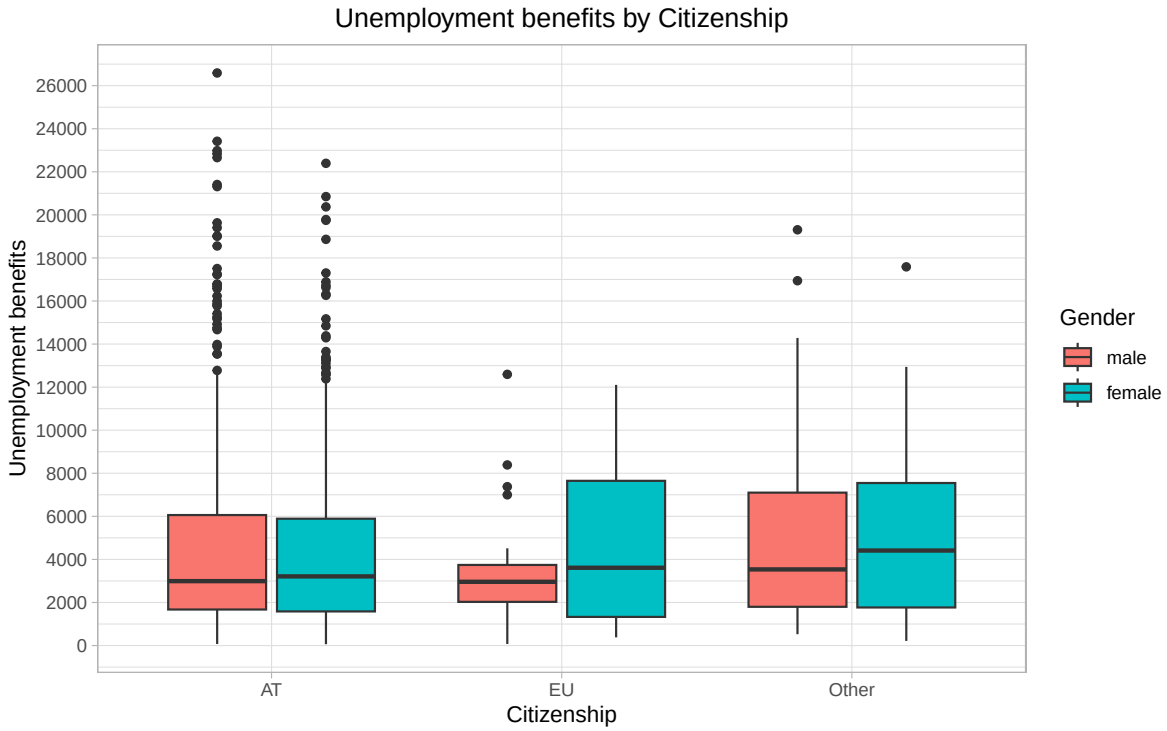
3.3.4 Citizenship x Age x Unemployment Benefits

Analyzing the relationship between Citizenship, Age and Unemployment Benefits.



The intercept is highest for the 'Other' category, followed by 'AT' and 'EU'. Interestingly, the slope relationships are reversed, with 'EU' visually having the steepest slope, followed by 'AT' and 'Other'. However, all three groups indicate that unemployment benefits increase with age.

3.3.5 Citizenship x Gender x Unemployment Benefits



This plot demonstrates approximately similar medians of unemployment benefits across citizenship groups. However, the range of values and the number of outliers are largest for the Austrian group, which may be influenced by the unbalanced distribution. Additionally, it is worth noting that the median unemployment benefits for females are higher in each citizenship group.

3.4 Descriptive Analysis Summary

The key insights, which can be helpful in the model building are following:

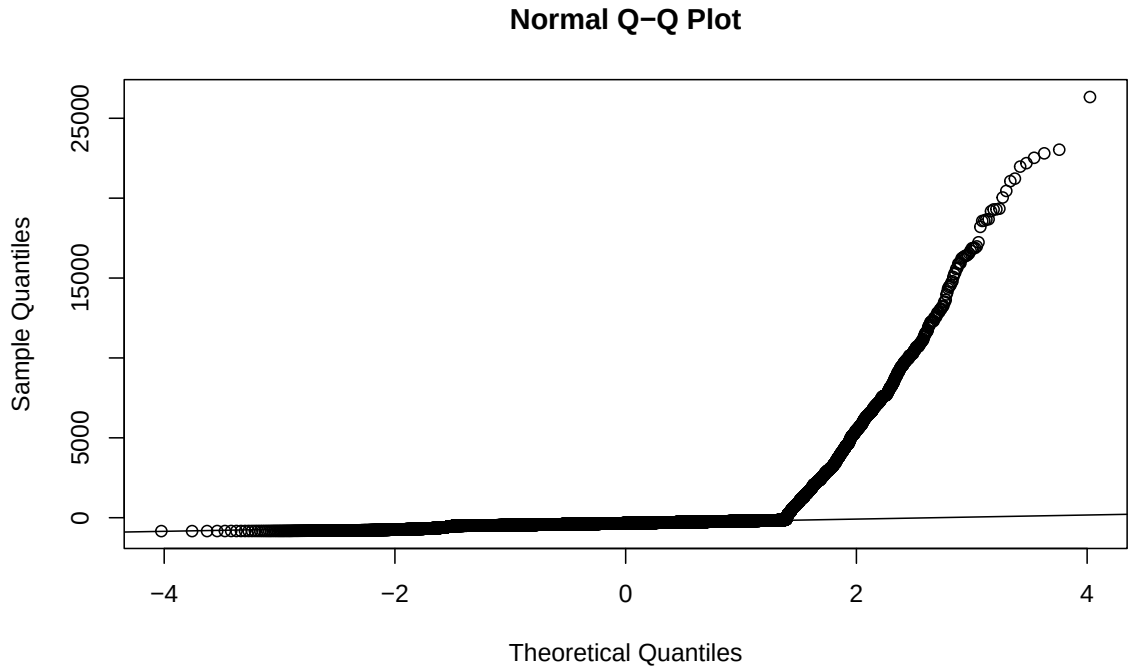
- Unemployment benefits variable is very unbalanced and has a lot of people with zero benefits and seems to follow log normal distribution, therefore log1p may be a good candidate for transformation.
- Household size and Age are also slightly skewed, which may lead to optional transformation, however only household size has outliers.
- Citizenship is also highly unbalanced, which may significantly influence the model
- Unemployment benefits seem to positively correlate with age.

- Age seems to have interaction effect with gender and citizenship
- Probably an interaction effect of citizenship with gender

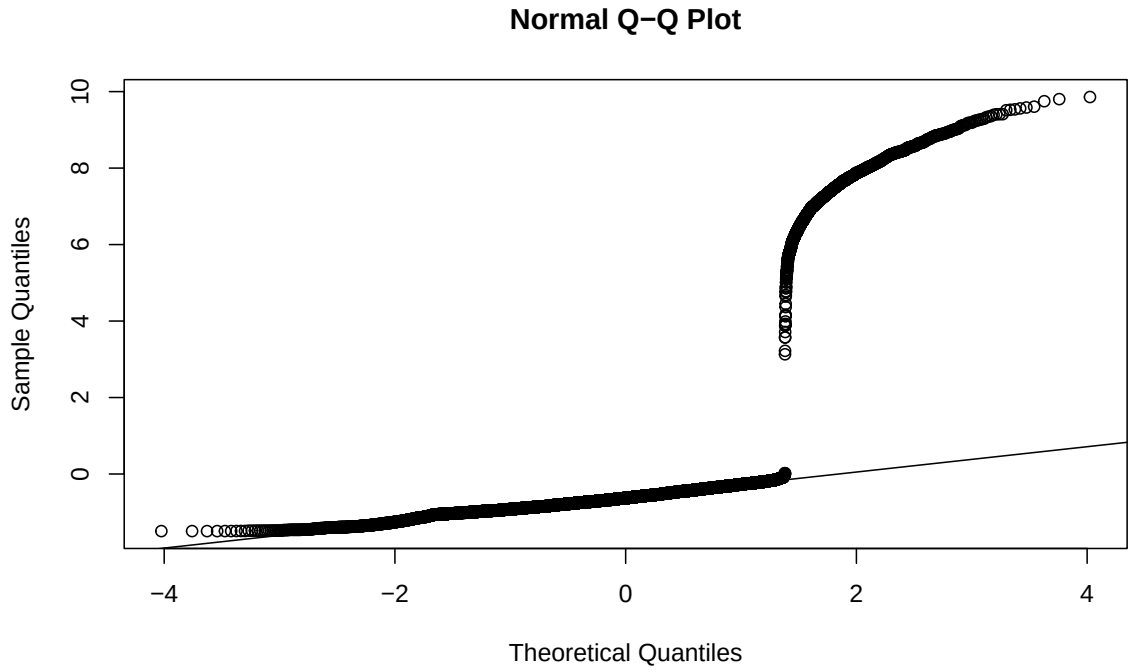
4 Regression modelling

```
Call:
lm(formula = benefits ~ age + hsize + gender + citizenship, data = west_austria)

Residuals:
    Min      1Q  Median      3Q     Max
 -839.2  -427.9  -347.2  -255.8 26329.9

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      575.3601    57.6537   9.980  < 2e-16 ***
age               -5.0010     0.7712  -6.484 9.15e-11 ***
hsize             -2.8727     9.4442  -0.304    0.761
genderfemale      26.2066    25.6433   1.022    0.307
citizenshipEU     62.1372    82.5676   0.753    0.452
citizenshipOther 323.4106    48.8086   6.626 3.55e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1687 on 17522 degrees of freedom
Multiple R-squared:  0.005995,  Adjusted R-squared:  0.005711
F-statistic: 21.14 on 5 and 17522 DF,  p-value: < 2.2e-16
```

We can immediately observe that R-squared is extremely low and equals to approximately 0.006, which means that the model explain only 0.6 percent of data. The given situation urges us to check for the distribution of residuals, as the model is highly unfit.



As it was implied the residuals are non-normally distributed especially at a higher end of distribution, which may signal for log transforming the target variable



Still not normally distributed residuals

5 Conclusion and criticism

5.1 Summary

5.2 Possible problems

5.3 Generalizability of the findings

5.4 Possible further questions