Regression Project – Group 3 (Intermediate Report)

Reema Aboluhom and Muzammil Muhammad

2025-09-12

Table of contents

0.1

1 1. Introduction

- The aim of this intermediate report is to explore determinants of income in South Austria.
- We focus on data management and descriptive statistics before conducting regression modelling.
- The response variable is net employment income (`py010n`).
- Explanatory variables include gender, citizenship, household size, and age.
- For the intermediate report we restrict to data management and descriptive statistics.

1.1

2 2. Data collection and description

- Source: EU-SILC (European Union Statistics on Income and Living Conditions), Austria.
- Type of data: survey data, representative sample of private households.
- Data format: cross-sectional microdata with social, demographic, and income information.
- Variables used:
  - `py010n` — employment income (numeric)
  - `age` — age in years (numeric)
  - `hsize` — household size (categorical converted to numeric)

  - `gender` — male / female (categorical)
  - `citizenship` — grouped nationality categories (categorical)
  - `region` — Austrian federal region

- Missing value handling:
  - Keep only positive values of income (`py010n > 0`)
  - Convert `hsize` to numeric
  - Remove observations with missing values

- Subsetting:
  - Only individuals living in **Styria** and **Carinthia** (South Austria, NUTS-1 region)

3 #Running code

4 3. Load packages

#options(repos = c(CRAN = https://cloud.r-project.org))

```
suppressPackageStartupMessages({
  library(tidyverse)
  library(readxl)
  library(simFrame)
  library(dplyr)        # data manipulation
  library(ggplot2)      # visualization
  library(tidyr)        # data tidying
  library(forcats)      # factor management
  library(effects)  # effect plots
  library(gt)
})
```

5 4. Load data and select variables

```r
data(eusilcP)

dat = eusilcP

str(eusilcP)
```

```
'data.frame':   58654 obs. of  28 variables:
 $ hid         : int  1 1 2 2 3 4 4 4 5 6 ...
 $ region      : Ord.factor w/ 9 levels "Burgenland"<"Lower Austria"<..: 6 6 5 5 5 6 6 6 3 2
 $ hsize       : Factor w/ 9 levels "1","2","3","4",..: 2 2 2 2 1 3 3 3 1 5 ...
 $ eqsize      : num  1.5 1.5 1.5 1.5 1 1.8 1.8 1.8 1 2.6 ...
 $ eqIncome    : num [1:58654(1d)] 11128 11128 19695 19695 5066 ...
  ..- attr(*, "dimnames")=List of 1
  .. ..$ : chr [1:58654] "2592313" "2592313" "2045000" "2045000" ...
 $ pid         : int  1 2 1 2 1 1 2 3 1 1 ...
 $ id          : chr  "0000101" "0000102" "0000201" "0000202" ...
 $ age         : num  25 24 57 53 30 32 33 8 77 34 ...
 $ gender      : Factor w/ 2 levels "male","female": 1 2 2 1 2 1 2 1 2 2 ...
 $ ecoStat     : Factor w/ 7 levels "1","2","3","4",..: 1 4 1 1 6 1 1 NA 5 2 ...
 $ citizenship: Factor w/ 3 levels "AT","EU","Other": 3 1 1 1 1 1 1 NA 1 1 ...
 $ py010n      : num  16693 0 0 16884 0 ...
 $ py050n      : num  0 0 12565 0 0 ...
 $ py090n      : num  0 0 0 0 5066 ...
 $ py100n      : num  0 0 0 0 0 ...
 $ py110n      : num  0 0 0 0 0 0 0 NA 0 0 ...
 $ py120n      : num  0 0 0 0 0 0 0 NA 0 0 ...
 $ py130n      : num  0 0 0 0 0 0 0 NA 0 0 ...
 $ py140n      : num  0 0 0 0 0 0 0 NA 0 0 ...
 $ hy040n      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ hy050n      : num  0 0 0 0 0 ...
 $ hy070n      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ hy080n      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ hy090n      : num  0 0 0 0 0 ...
 $ hy110n      : num  0 0 0 0 0 ...
 $ hy130n      : num  0 0 93.6 93.6 0 ...
 $ hy145n      : num  0 0 -187 -187 0 ...
 $ main        : logi  TRUE FALSE FALSE TRUE TRUE TRUE ...
```

```r
head(eusilcP)
```

```
      hid        region hsize eqsize eqIncome pid      id age gender ecoStat
```

```
39993   1 Upper Austria     2    1.5 11128.45   1 0000101  25   male       1
39994   1 Upper Austria     2    1.5 11128.45   2 0000102  24 female       4
31004   2        Styria     2    1.5 19694.85   1 0000201  57 female       1
31005   2        Styria     2    1.5 19694.85   2 0000202  53   male       1
29071   3        Styria     1    1.0  5066.24   1 0000301  30 female       6
41322   4 Upper Austria     3    1.8 31480.01   1 0000401  32   male       1
      citizenship   py010n   py050n  py090n py100n py110n py120n py130n py140n
39993       Other 16692.67     0.00    0.00      0      0      0      0      0
39994          AT     0.00     0.00    0.00      0      0      0      0      0
31004          AT     0.00 12564.59    0.00      0      0      0      0      0
31005          AT 16884.06     0.00    0.00      0      0      0      0      0
29071          AT     0.00     0.00 5066.24      0      0      0      0      0
41322          AT 25047.39     0.00    0.00      0      0      0      0      0
      hy040n  hy050n hy070n hy080n hy090n  hy110n hy130n  hy145n  main
39993      0    0.00      0      0   0.00    0.00   0.00    0.00  TRUE
39994      0    0.00      0      0   0.00    0.00   0.00    0.00 FALSE
31004      0    0.00      0      0   0.00    0.00  93.63 -187.26 FALSE
31005      0    0.00      0      0   0.00    0.00  93.63 -187.26  TRUE
29071      0    0.00      0      0   0.00    0.00   0.00    0.00  TRUE
41322      0 7167.39      0      0  31.15 1349.91   0.00    0.00  TRUE
```

```
summary(eusilcP)
```

```
      hid                 region          hsize          eqsize
 Min.   :    1   Vienna       :11657   2      :14128   Min.   :1.000
 1st Qu.: 6262   Lower Austria:11127   4      :13180   1st Qu.:1.500
 Median :12465   Upper Austria:10310   3      :12429   Median :2.000
 Mean   :12488   Styria       : 8142   1      : 8602   Mean   :1.943
 3rd Qu.:18719   Tyrol        : 4796   5      : 6745   3rd Qu.:2.400
 Max.   :25000   Carinthia    : 4111   6      : 2094   Max.   :4.500
                 (Other)      : 8511   (Other): 1476
   eqIncome           pid            id                  age
 Min.   :     0   Min.   :1.00   Length:58654       Min.   :-1.00
 1st Qu.: 13539   1st Qu.:1.00   Class :character   1st Qu.:22.00
 Median : 18322   Median :2.00   Mode  :character   Median :40.00
 Mean   : 20163   Mean   :2.07                      Mean   :39.75
 3rd Qu.: 24277   3rd Qu.:3.00                      3rd Qu.:57.00
 Max.   :179946   Max.   :9.00                      Max.   :97.00

   gender         ecoStat      citizenship       py010n           py050n
 male  :28539   1      :20900   AT   :44066   Min.   :     0   Min.   : -6895
 female:30115   5      :12836   EU   : 1257   1st Qu.:     0   1st Qu.:     0
```

```
              7      : 4607   Other: 3162   Median :  2382   Median :     0
              2      : 4362   NA's :10169   Mean   :  9062   Mean   :  1288
              4      : 2921                 3rd Qu.: 16820   3rd Qu.:     0
              (Other): 2859                 Max.   :199075   Max.   :129874
              NA's   :10169                 NA's   :10169    NA's   :10169
    py090n            py100n           py110n            py120n
Min.   :    0.0   Min.   :     0   Min.   :    0.0   Min.   :    0.00
1st Qu.:    0.0   1st Qu.:     0   1st Qu.:    0.0   1st Qu.:    0.00
Median :    0.0   Median :     0   Median :    0.0   Median :    0.00
Mean   :  444.6   Mean   :  3713   Mean   :   72.9   Mean   :   51.22
3rd Qu.:    0.0   3rd Qu.:     0   3rd Qu.:    0.0   3rd Qu.:    0.00
Max.   :29887.1   Max.   :101777   Max.   :22546.8   Max.   :46398.44
NA's   :10169     NA's   :10169    NA's   :10169     NA's   :10169
    py130n            py140n             hy040n             hy050n
Min.   :    0.0   Min.   :    0.00   Min.   : -2962.5   Min.   :-11857
1st Qu.:    0.0   1st Qu.:    0.00   1st Qu.:     0.0   1st Qu.:     0
Median :    0.0   Median :    0.00   Median :     0.0   Median :     0
Mean   :  393.7   Mean   :   41.73   Mean   :   879.9   Mean   :  2826
3rd Qu.:    0.0   3rd Qu.:    0.00   3rd Qu.:     0.0   3rd Qu.:  4558
Max.   :53183.6   Max.   :18643.46   Max.   :129586.6   Max.   :118309
NA's   :10169     NA's   :10169
    hy070n             hy080n             hy090n             hy110n
Min.   :    0.00   Min.   :     0.0   Min.   :  -457.46   Min.   :    0.00
1st Qu.:    0.00   1st Qu.:     0.0   1st Qu.:     0.75   1st Qu.:    0.00
Median :    0.00   Median :     0.0   Median :    58.45   Median :    0.00
Mean   :   93.12   Mean   :   744.6   Mean   :   462.45   Mean   :   32.97
3rd Qu.:    0.00   3rd Qu.:     0.0   3rd Qu.:   234.78   3rd Qu.:    0.00
Max.   :17954.97   Max.   :124206.2   Max.   :112011.03   Max.   :14506.49

    hy130n            hy145n              main
Min.   :-5489.6   Min.   :-29519.3   Mode :logical
1st Qu.:    0.0   1st Qu.:  -256.8   FALSE:33654
Median :    0.0   Median :     0.0   TRUE :25000
Mean   :  339.1   Mean   :  -108.8
3rd Qu.:    0.0   3rd Qu.:     0.0
Max.   :40762.9   Max.   : 49768.0
```

---

5. Data preparation

- Filter for South Austria regions (Styria and Carinthia)

- Keep only positive income observations

- Convert household size to numeric

- Remove missing values

- Group citizenship categories if needed

```
dat <- eusilcP %>%

  select(py010n, gender, citizenship, hsize, age, region) %>%

  filter(region %in% c("Carinthia", "Styria")) %>%

  filter(py010n > 0) %>%

  na.omit()



dat$gender <- as.factor(dat$gender)

dat$citizenship <- as.factor(dat$citizenship)

dat$hsize <- as.numeric(as.character(dat$hsize))
```

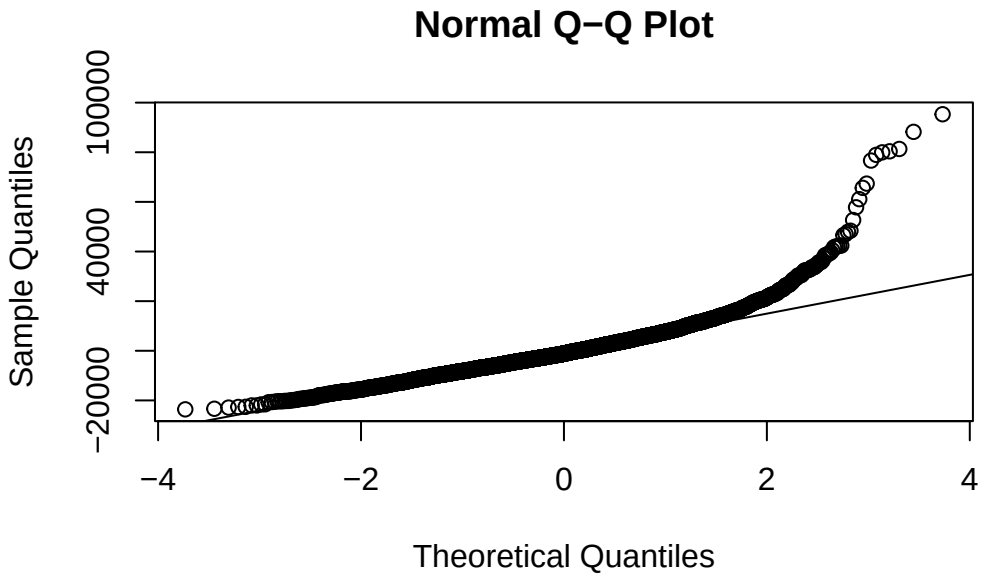
```
 # Initial model with interaction Preview for Regression Stage

model_int <- lm(py010n ~ gender * citizenship + hsize + age, data = dat)

anova(model_int)
```

```
Analysis of Variance Table

Response: py010n
                     Df     Sum Sq    Mean Sq  F value    Pr(>F)
gender                1 7.5470e+10 7.5470e+10 832.5069 < 2.2e-16 ***
citizenship           2 5.0537e+09 2.5268e+09  27.8734 9.087e-13 ***
hsize                 1 5.3097e+09 5.3097e+09  58.5714 2.318e-14 ***
age                   1 1.5135e+10 1.5135e+10 166.9498 < 2.2e-16 ***
gender:citizenship    2 2.9543e+07 1.4772e+07   0.1629    0.8496
Residuals          5259 4.7675e+11 9.0654e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

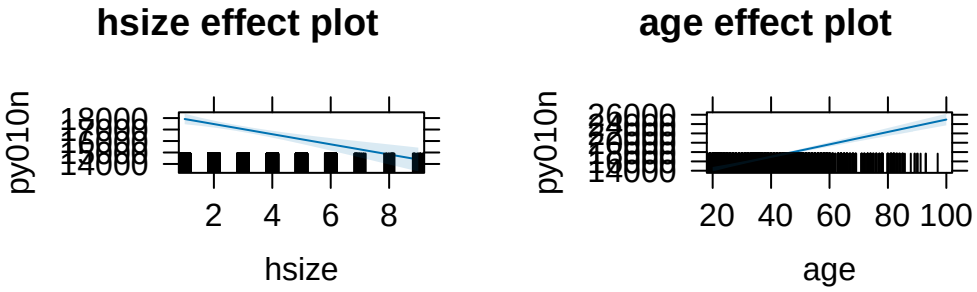
```r
summary(model_int)
```

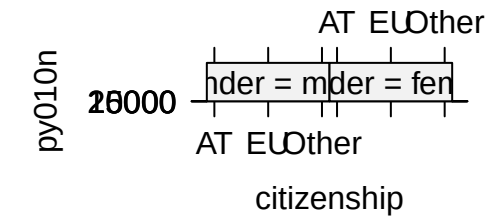
```
Call:
lm(formula = py010n ~ gender * citizenship + hsize + age, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-23598  -5912  -1132   4624  95317

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  16522.34     568.03  29.087  < 2e-16 ***
genderfemale                 -7718.53     271.50 -28.429  < 2e-16 ***
citizenshipEU                 1474.71    1516.38   0.973    0.331
citizenshipOther             -5174.37     983.76  -5.260 1.50e-07 ***
hsize                         -441.55      88.19  -5.007 5.71e-07 ***
age                            133.02      10.30  12.920  < 2e-16 ***
genderfemale:citizenshipEU     169.22    2160.86   0.078    0.938
genderfemale:citizenshipOther  860.66    1517.69   0.567    0.571
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9521 on 5259 degrees of freedom
Multiple R-squared:  0.1748,    Adjusted R-squared:  0.1737
F-statistic: 159.2 on 7 and 5259 DF,  p-value: < 2.2e-16
```

```r
qqnorm(residuals(model_int))

qqline(residuals(model_int))
```




```
plot(allEffects(model_int))
```





6 6. Descriptive statistics

6.1 6.1 Numeric summaries

```
# Summaries
summ_py010n <- summary(dat$py010n)
summ_age    <- summary(dat$age)
summ_hsize  <- summary(dat$hsize)


summ_py010n
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
     1.93  10066.01  16225.84  16952.35  21939.78 118362.27
```

```
summ_age
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   29.00   40.00   39.73   49.00   97.00
```

```
summ_hsize
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.00    3.00    3.19    4.00    9.00
```

6.2 6.2 Frequency tables

```
table(dat$gender)
```

```
  male female
  3004   2263
```

```
table(dat$citizenship)
```

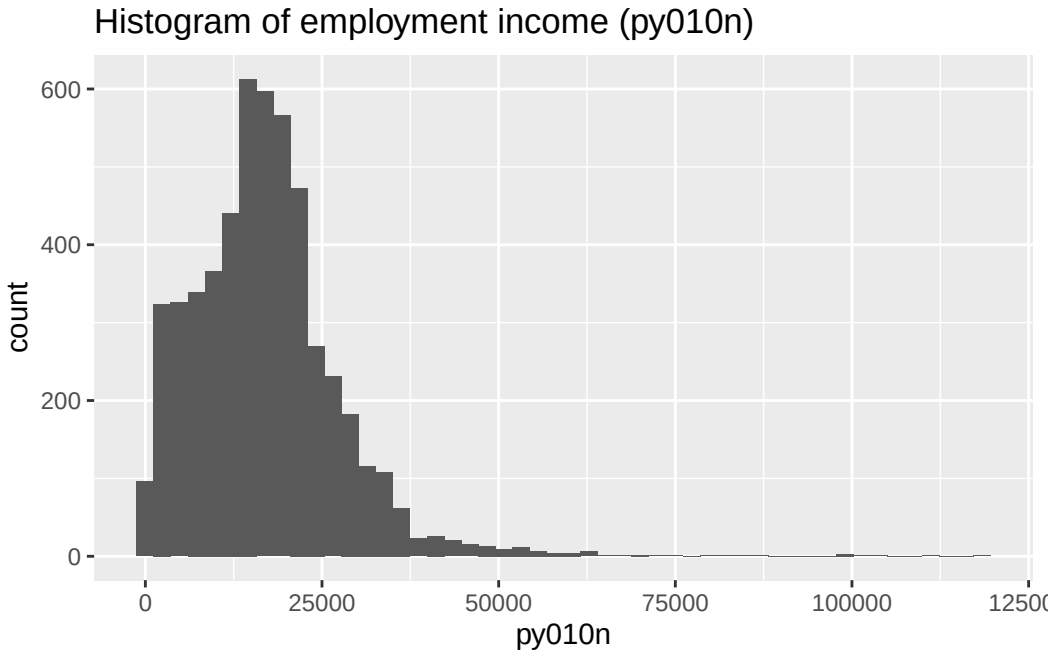
```
   AT    EU Other
 5021    79   167
```

---

7 7. Univariate visualizations

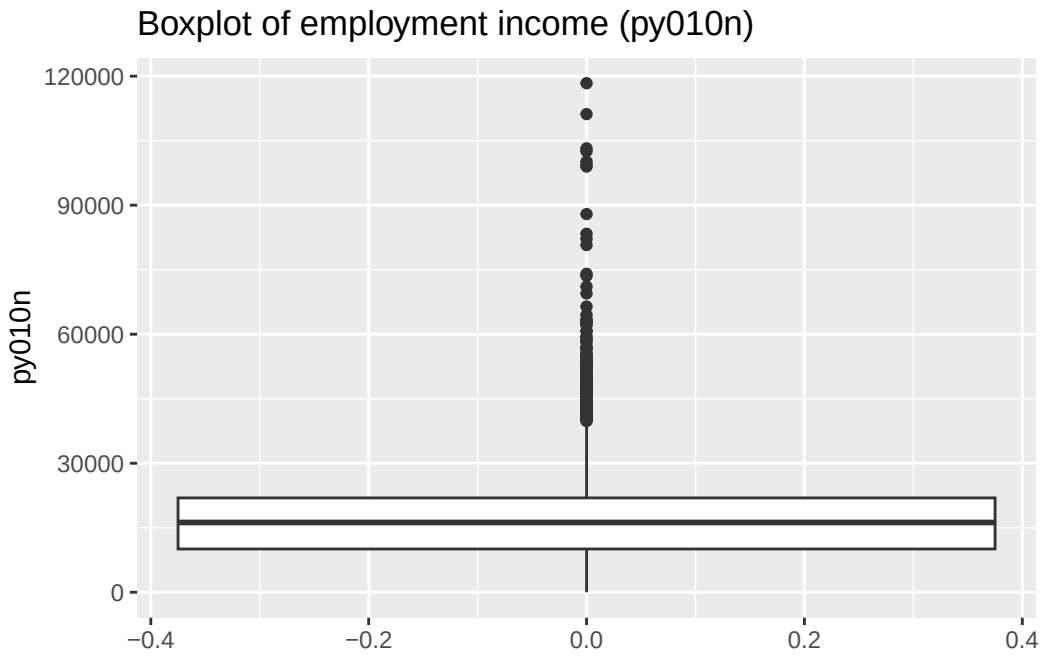
7.1 7.1 Employment income (py010n)

```
ggplot(dat, aes(x = py010n)) +
  geom_histogram(bins = 50) +
  labs(title = "Histogram of employment income (py010n)")
```




```
#The histogram of income shows a strong right skew, with a few very high earners
```

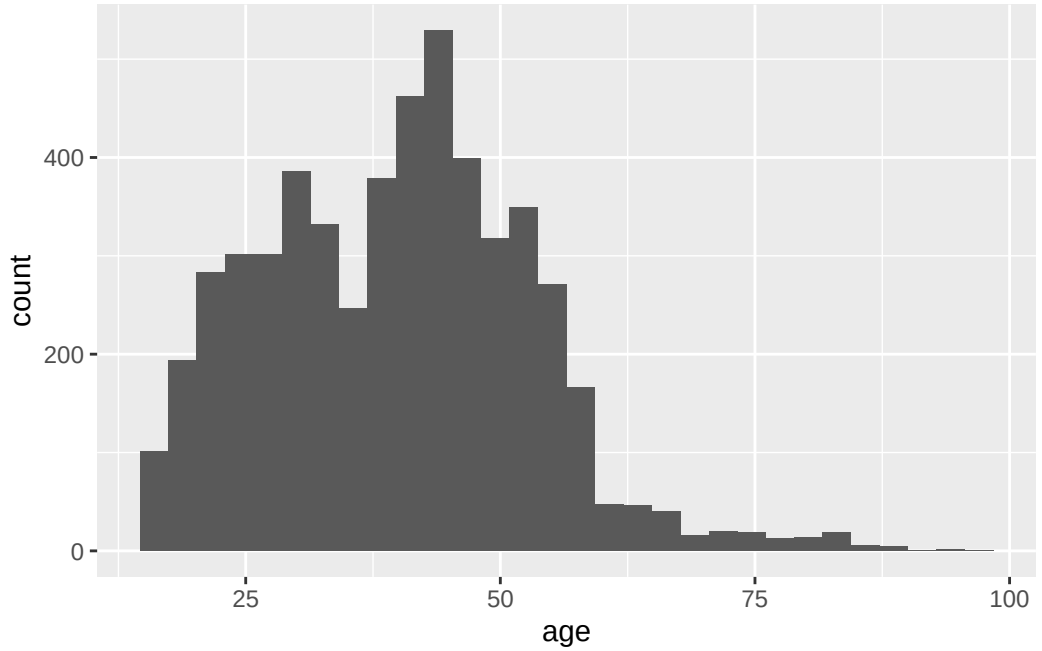
```
ggplot(dat, aes(y = py010n)) +
  geom_boxplot() +
  labs(title = "Boxplot of employment income (py010n)")
```




```
#
```

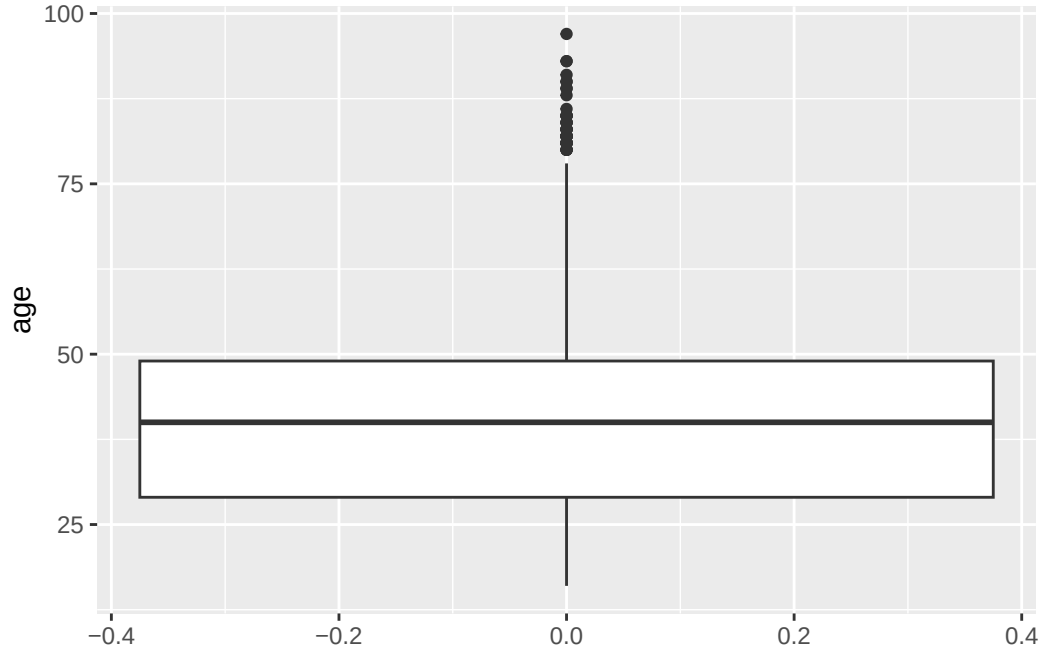
7.2 7.2 Age

```
ggplot(dat, aes(x = age)) + geom_histogram(bins = 30)
```



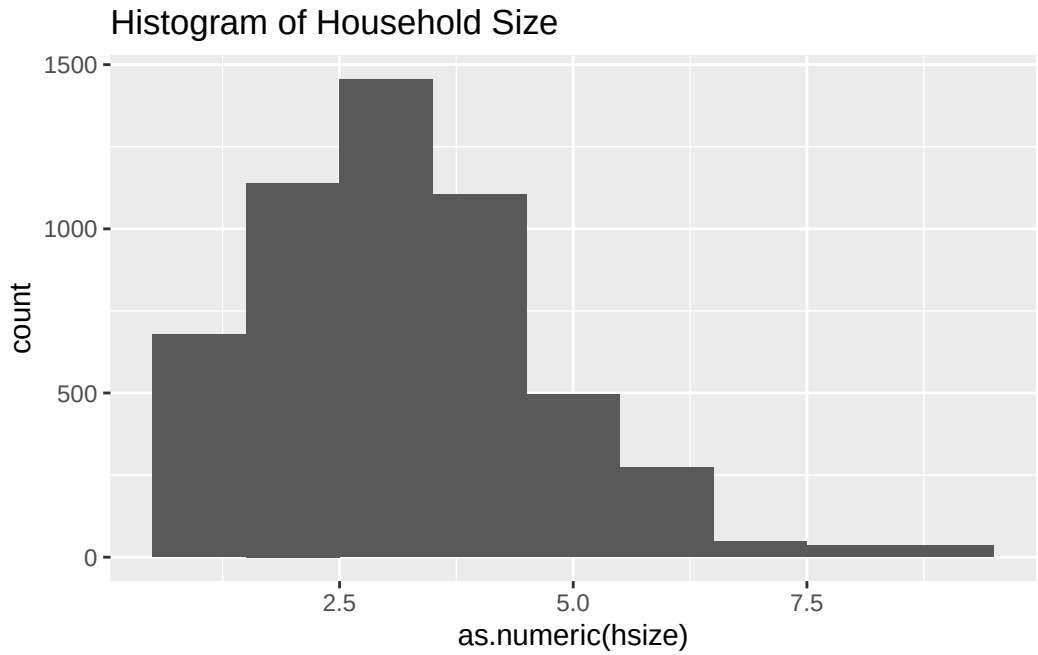

```
#Age distribution is roughly uniform with fewer very old respondents
```

```
ggplot(dat, aes(y = age)) + geom_boxplot()
```

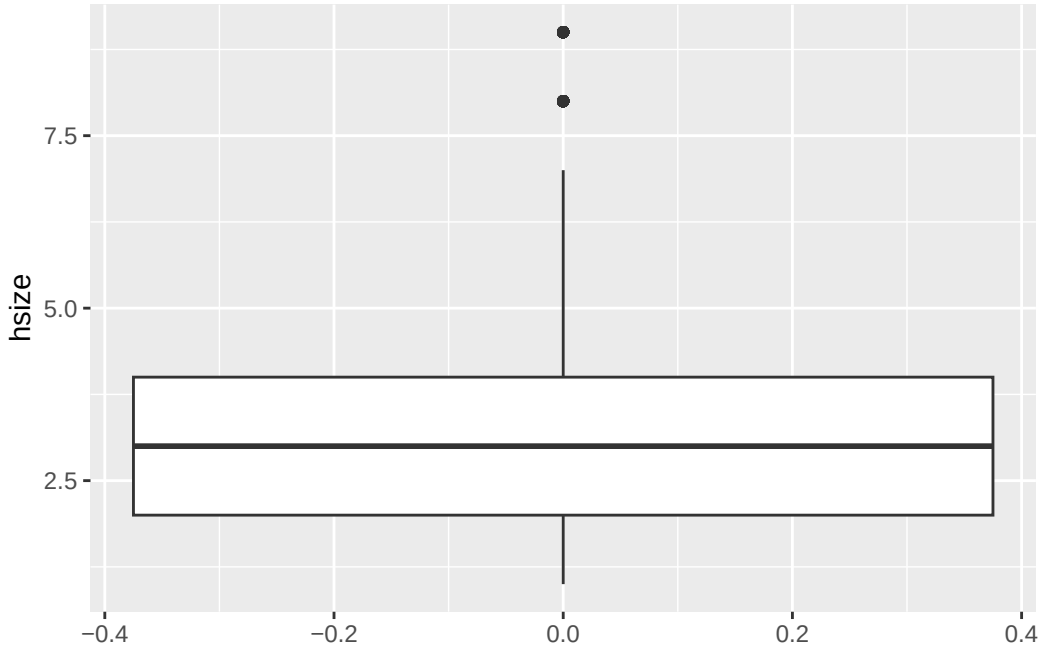


7.3 7.3 Household size (hsize)

```
ggplot(dat, aes(x = as.numeric(hsize))) +
  geom_histogram(binwidth = 1) +
  labs(title = "Histogram of Household Size")
```

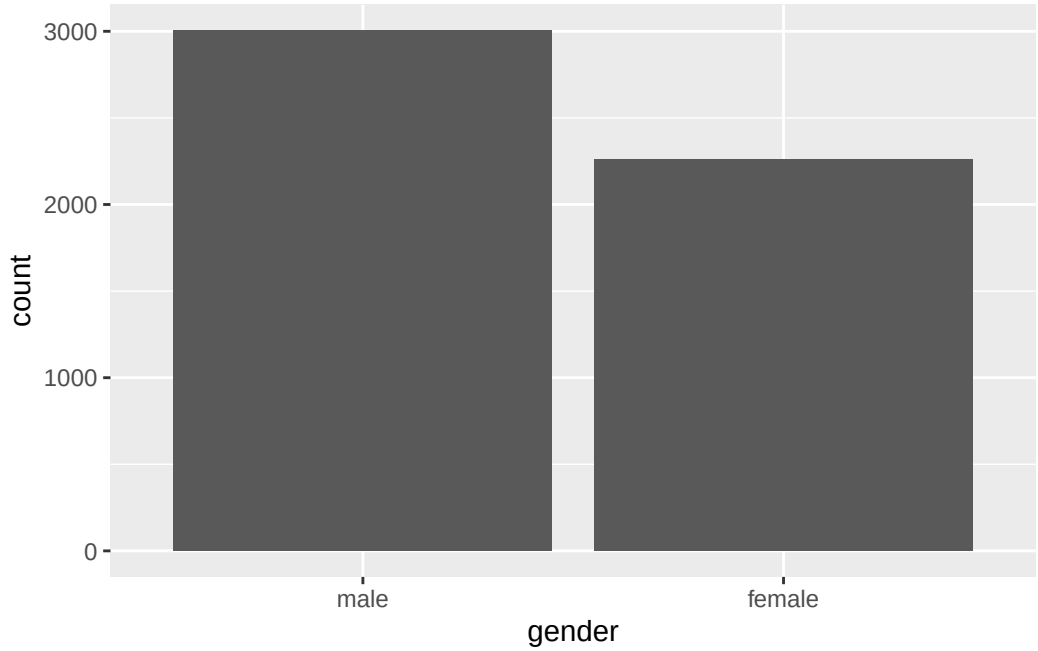



```
ggplot(dat, aes(y = hsize)) + geom_boxplot()
```



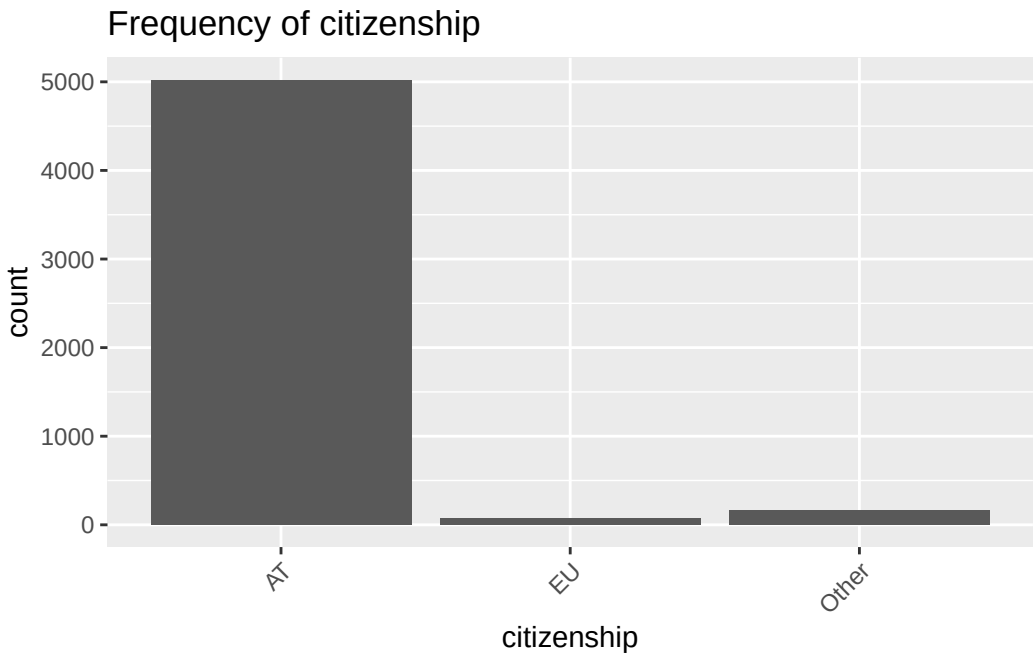
7.4 7.4 Gender

```
ggplot(dat, aes(x = gender)) + geom_bar()
```



7.5 7.5 Citizenship

```
ggplot(dat, aes(x = citizenship)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Frequency of citizenship")
```



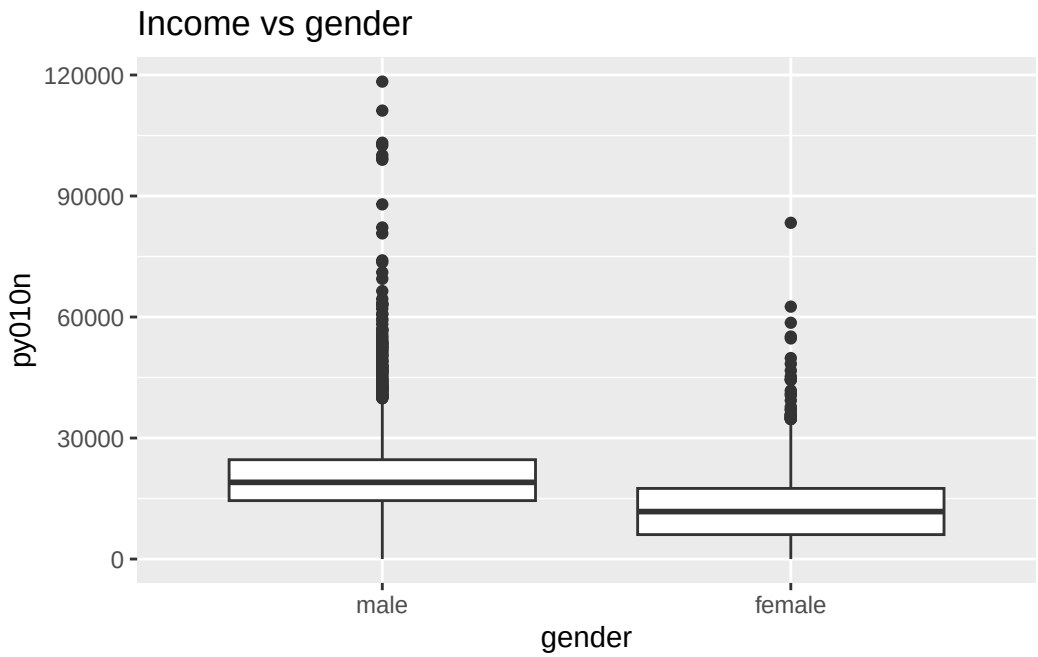

```
#Income differences by gender appear stronger among Austrian citizens than among non-citizens
```

---

8 8. Bivariate plots (predictors vs response)

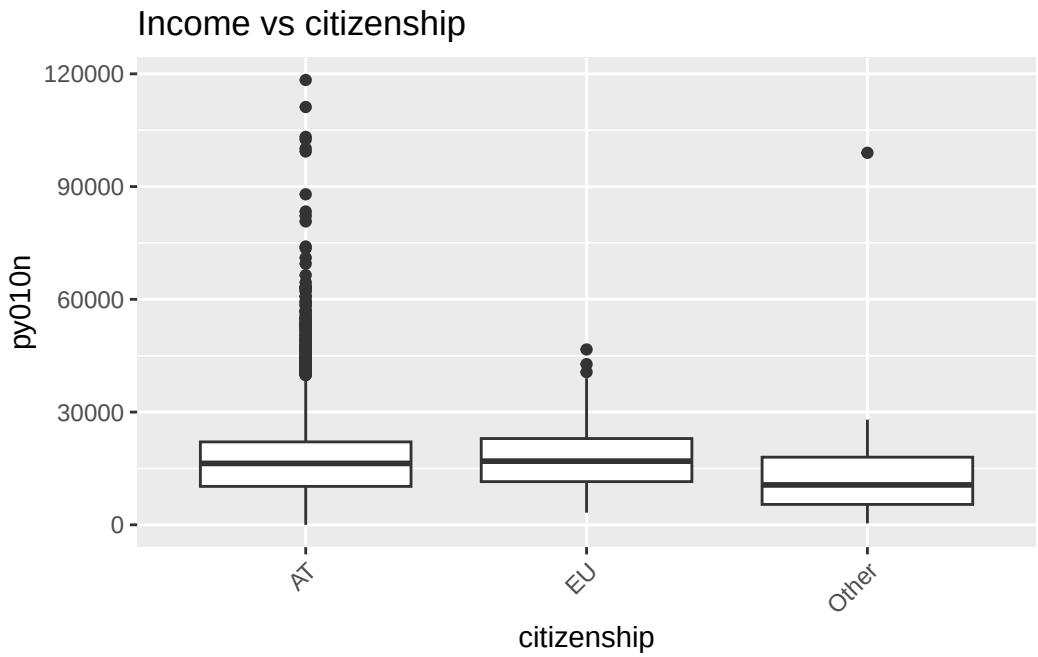
8.1 8.1 Gender vs income

```
ggplot(dat, aes(x = gender, y = py010n)) +
  geom_boxplot() +
  labs(title = "Income vs gender")
```



8.2 8.2 Citizenship vs income

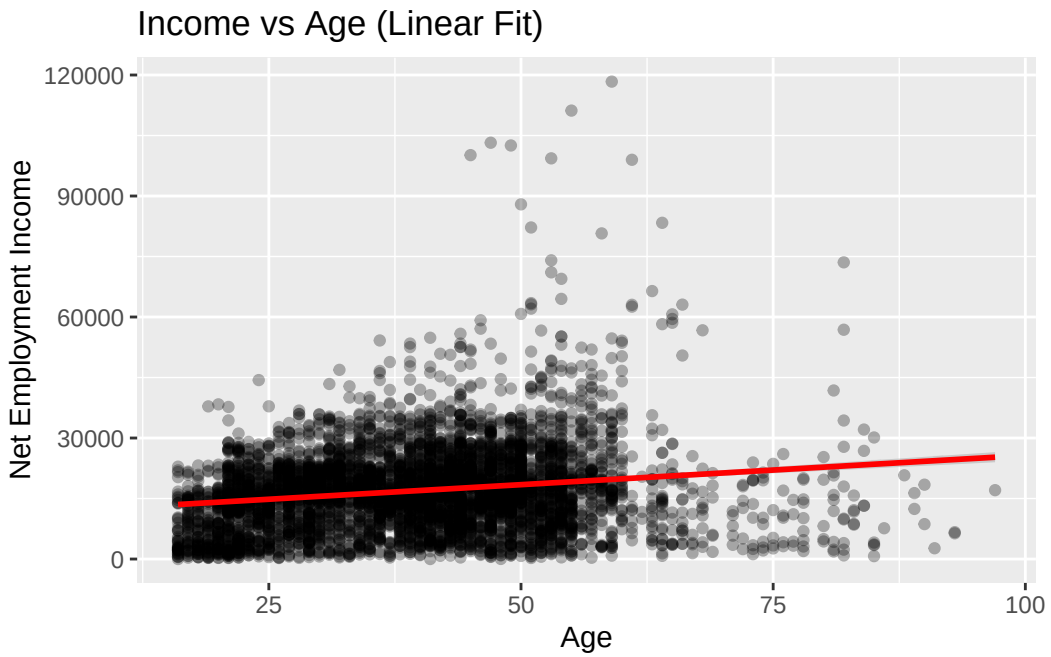
```
ggplot(dat, aes(x = citizenship, y = py010n)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Income vs citizenship")
```



8.3 8.3 Age vs income

```
ggplot(dat, aes(x = age, y = py010n)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Income vs Age (Linear Fit)",
       x = "Age",
       y = "Net Employment Income")
```

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

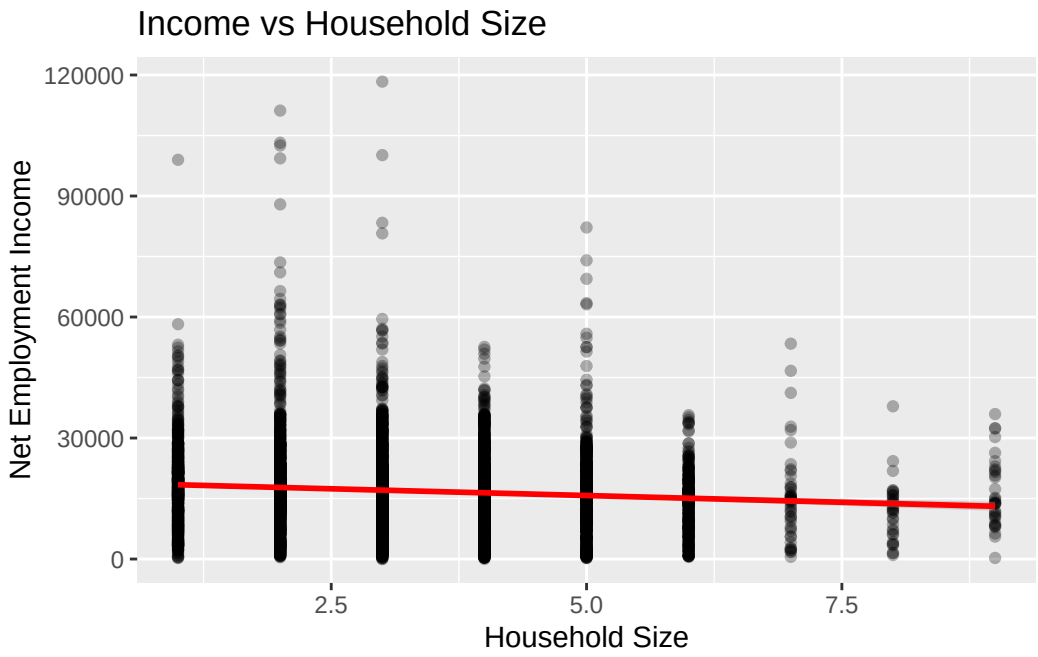



```
#The fitted linear regression line shows a positive association between age and income. On av
```

8.4 8.4 Household size vs income

```
ggplot(dat, aes(x = hsize, y = py010n)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Income vs Household Size",
       x = "Household Size",
       y = "Net Employment Income")
```

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

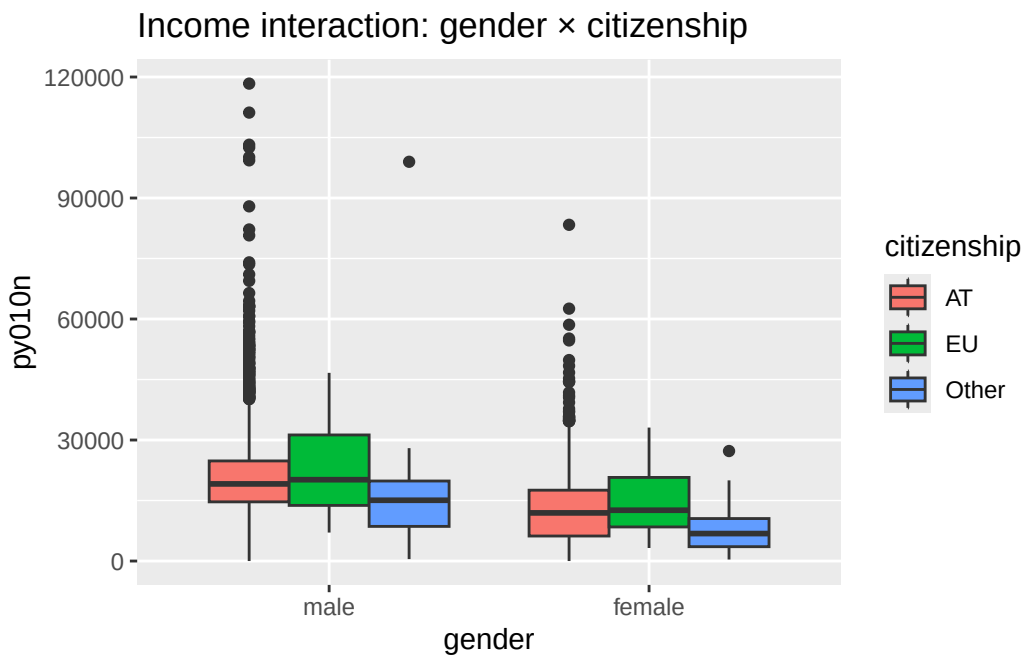


---

9 9. Interaction plots

9.1 9.1 Gender × Citizenship

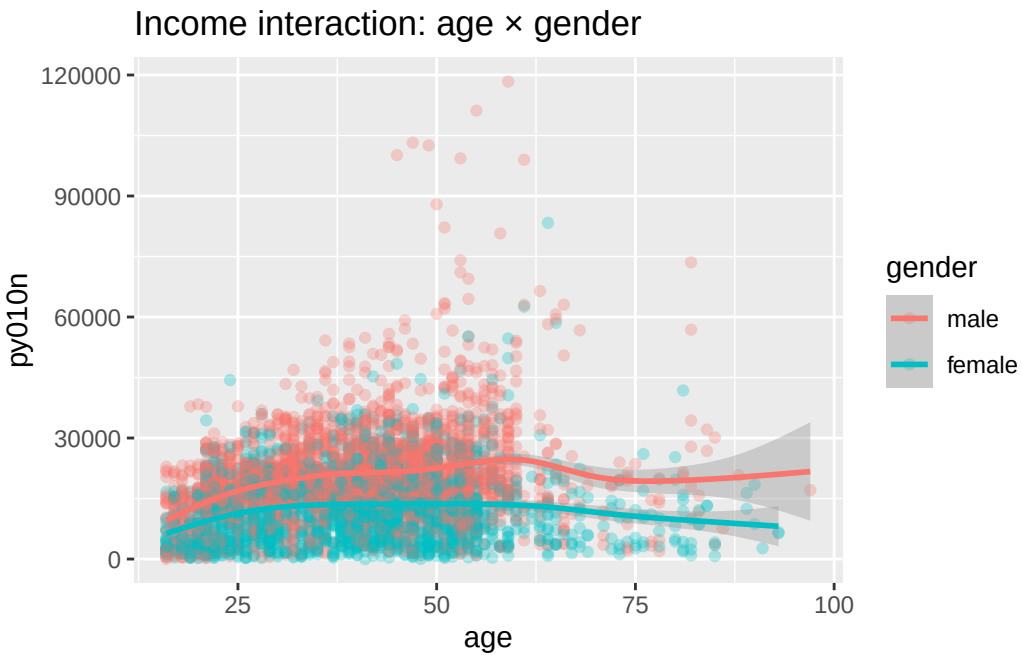
```
ggplot(dat, aes(x = gender, y = py010n, fill = citizenship)) +
  geom_boxplot(position = "dodge") +
  labs(title = "Income interaction: gender × citizenship")
```



9.2 9.2 Age × Gender

```
ggplot(dat, aes(x = age, y = py010n, color = gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth() +
  labs(title = "Income interaction: age × gender")
```

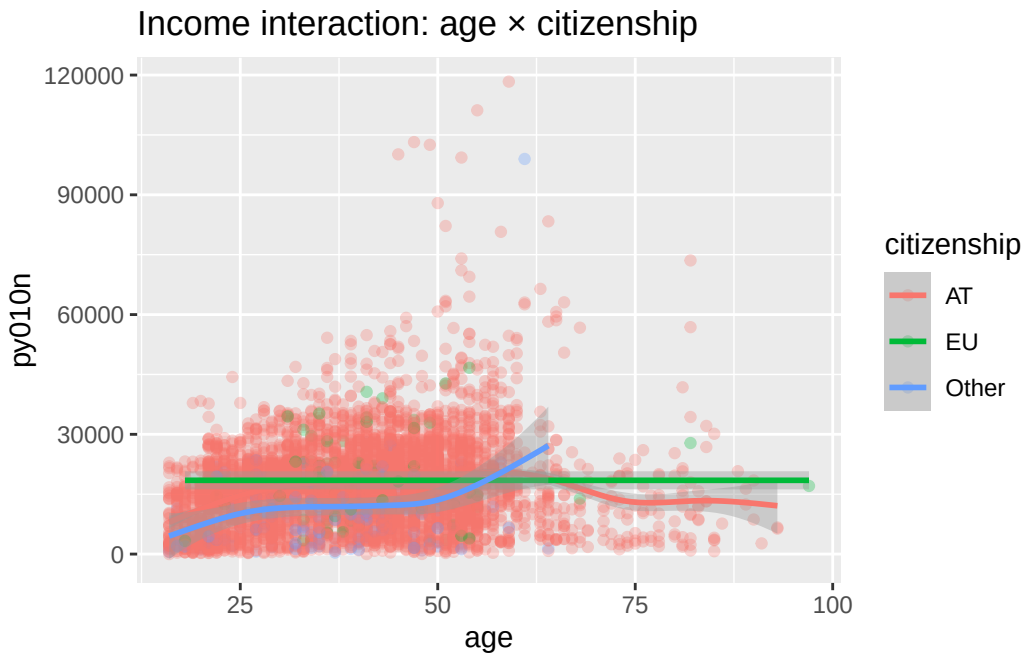
```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



9.3 9.3 Age × Citizenship

```
ggplot(dat, aes(x = age, y = py010n, color = citizenship)) +
  geom_point(alpha = 0.3) +
  geom_smooth() +
  labs(title = "Income interaction: age × citizenship")
```

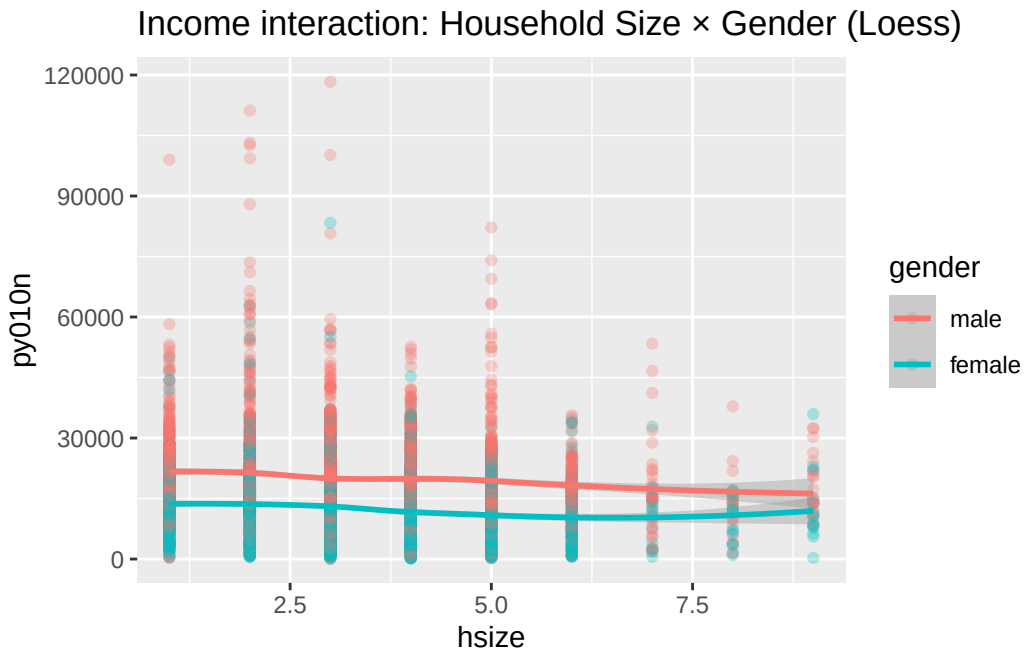
```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



9.4 9.4 Household Size × Gender

```
ggplot(dat, aes(x = hsize, y = py010n, color = gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "loess") +
  labs(title = "Income interaction: Household Size × Gender (Loess)")
```

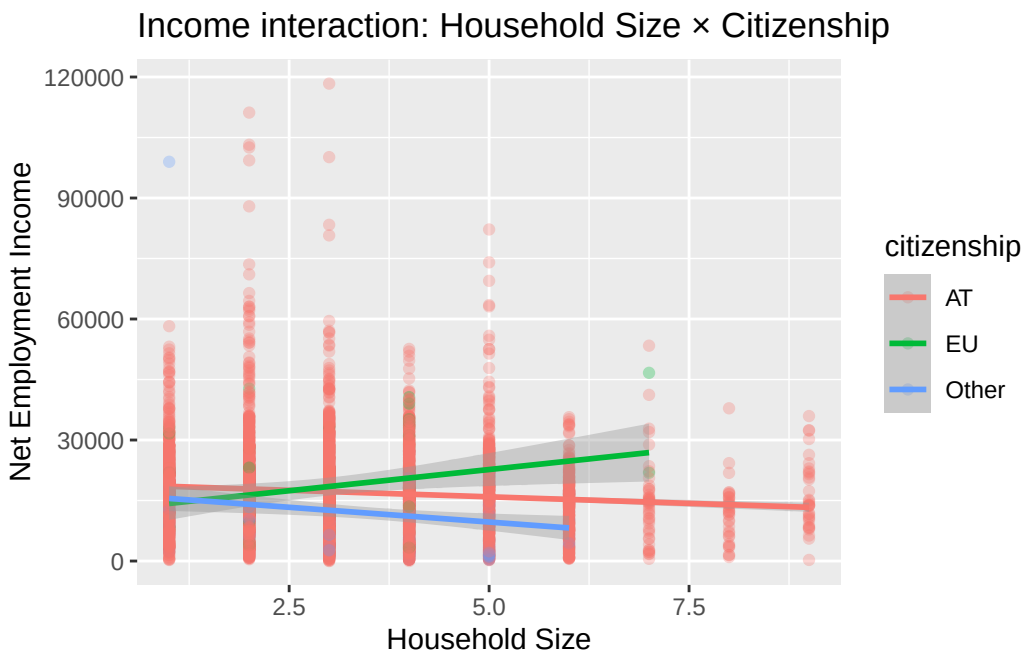
```
`geom_smooth()` using formula = 'y ~ x'
```



9.5 9.5 Household Size × Citizenship

```
ggplot(dat, aes(x = hsize, y = py010n, color = citizenship)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  labs(title = "Income interaction: Household Size × Citizenship",
       x = "Household Size",
       y = "Net Employment Income")
```

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



10

11 10. Contingency tables (categorical × categorical)

```
table(dat$gender, dat$citizenship)
```

```
          AT   EU Other
  male   2867   40    97
  female 2154   39    70
```

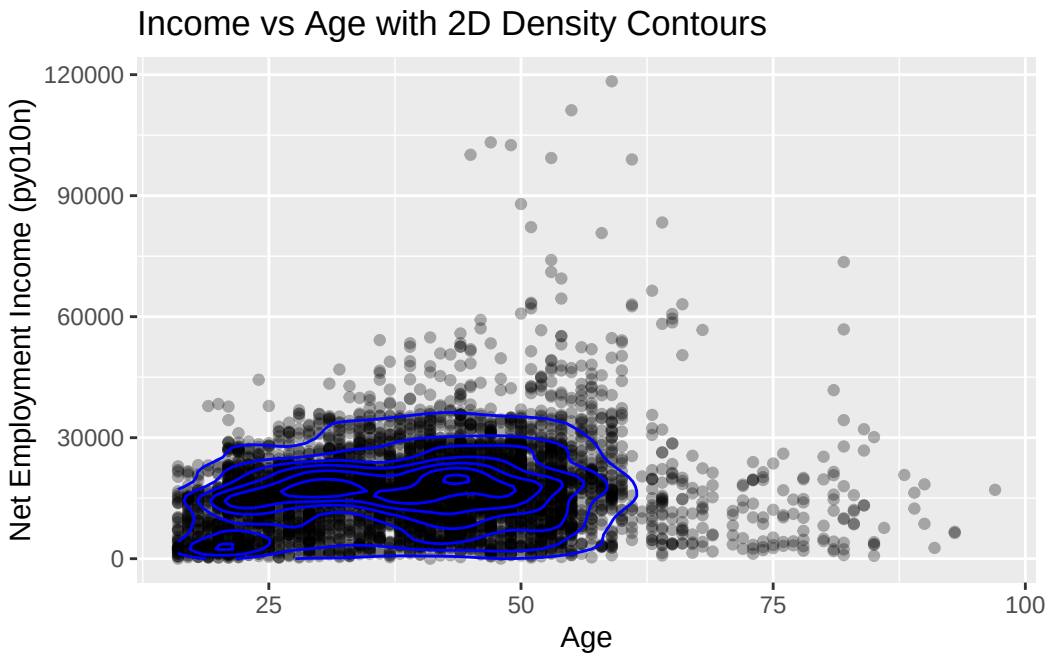
- If categories are too detailed, merge small ones:

```
dat$citizenship <- fct_lump(dat$citizenship, n = 3)

table(dat$gender, dat$citizenship)
```

```
          AT   EU Other
```

```
  male   2867   40    97
  female 2154   39    70
```

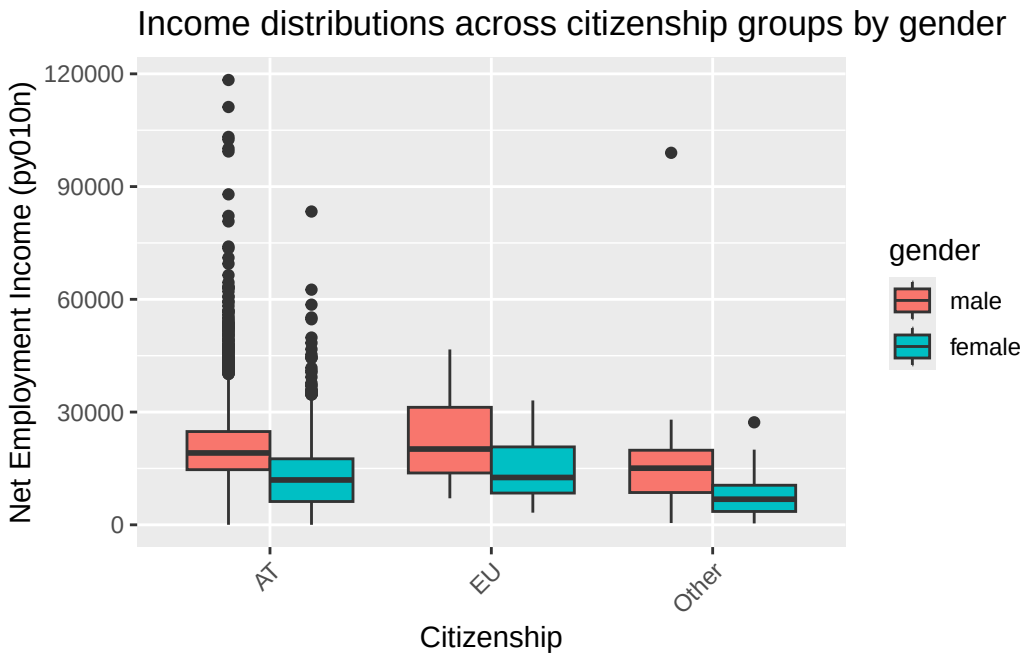
```
# Extra visualization: Age vs Income with 2D density contours
ggplot(dat, aes(x = age, y = py010n)) +
  geom_point(alpha = 0.3) +
  geom_density_2d(color = "blue") +
  labs(title = "Income vs Age with 2D Density Contours",
       x = "Age",
       y = "Net Employment Income (py010n)")
```




```
#The contingency table of gender × citizenship shows that most respondents are Austrian citi
```

Compares income distributions across citizenship groups, separately for men and women.

```
# Boxplot of income by citizenship, split by gender
ggplot(dat, aes(x = citizenship, y = py010n, fill = gender)) +
  geom_boxplot(position = "dodge") +
  labs(title = "Income distributions across citizenship groups by gender",
       x = "Citizenship",
       y = "Net Employment Income (py010n)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
#The comparison of employment income (py010n) across citizenship groups, separately for men a
```

12 11. Summary

- The income variable is highly right-skewed with outliers.
- Men typically have higher median employment income than women.
- Citizenship differences may indicate structural inequality in wages.
- Age and income show a nonlinear increasing pattern.
- Larger households do not clearly correlate with higher or lower income.
- Some interaction effects appear visible (gender × citizenship, etc.), though regression results show no significant interaction.
- The regression model explains about 17% of income variance, suggesting other unobserved factors are important.