Regression Project – Group 3 (Intermediate Report)

Reema abolahum and Muzammil Mohammed

2025-10-12

Table of contents

```r
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)
  library(corrplot)
  library(forcats)
  library(splines)
  library(car)
})

# Define color palette
main_color <- "#2C3E50"
accent_color <- "#E74C3C"
secondary_color <- "#3498DB"
tertiary_color <- "#F39C12"
quaternary_color <- "#9B59B6"
```

1 1. Introduction

This report analyzes determinants of employment income (py010n) in South Austria. We focus on data preparation, descriptive statistics, and regression modeling using polynomial and spline methods. Interactions between predictors are explored, and diagnostic checks are performed to validate model assumptions.

2 2. Data Collection and Preparation

```
data("eusilcP")

dat <- eusilcP %>%
  select(py010n, gender, citizenship, hsize, age, region) %>%
  filter(region %in% c("Carinthia", "Styria")) %>%
  filter(py010n > 0)

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

3 3. Descriptive Analysis

3.1 3.1 Numeric Summaries

```
num_vars <- dat %>% select(py010n, age, hsize)
summary(num_vars)
```

```
     py010n               age            hsize
 Min.   :     1.93   Min.   :16.00   Min.   :1.00
 1st Qu.: 10066.01   1st Qu.:29.00   1st Qu.:2.00
 Median : 16225.84   Median :40.00   Median :3.00
 Mean   : 16952.35   Mean   :39.73   Mean   :3.19
 3rd Qu.: 21939.78   3rd Qu.:49.00   3rd Qu.:4.00
 Max.   :118362.27   Max.   :97.00   Max.   :9.00
```

3.2 3.2 Frequency Tables

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

```

  male female 
  3004   2263 
```

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

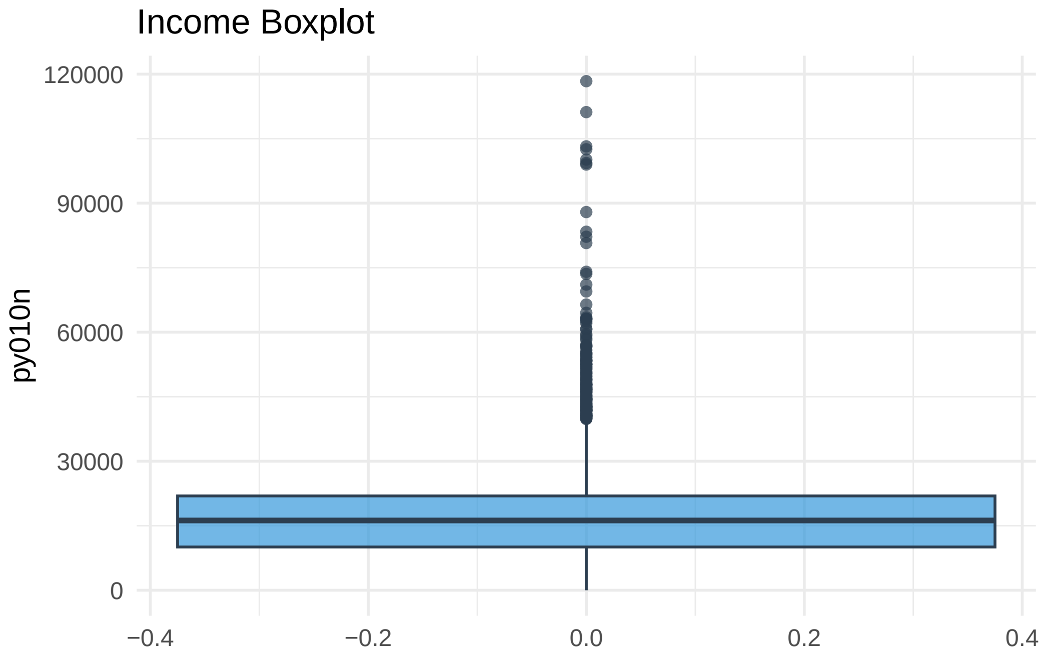
```

   AT    EU Other 
 5021    79   167 
```

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

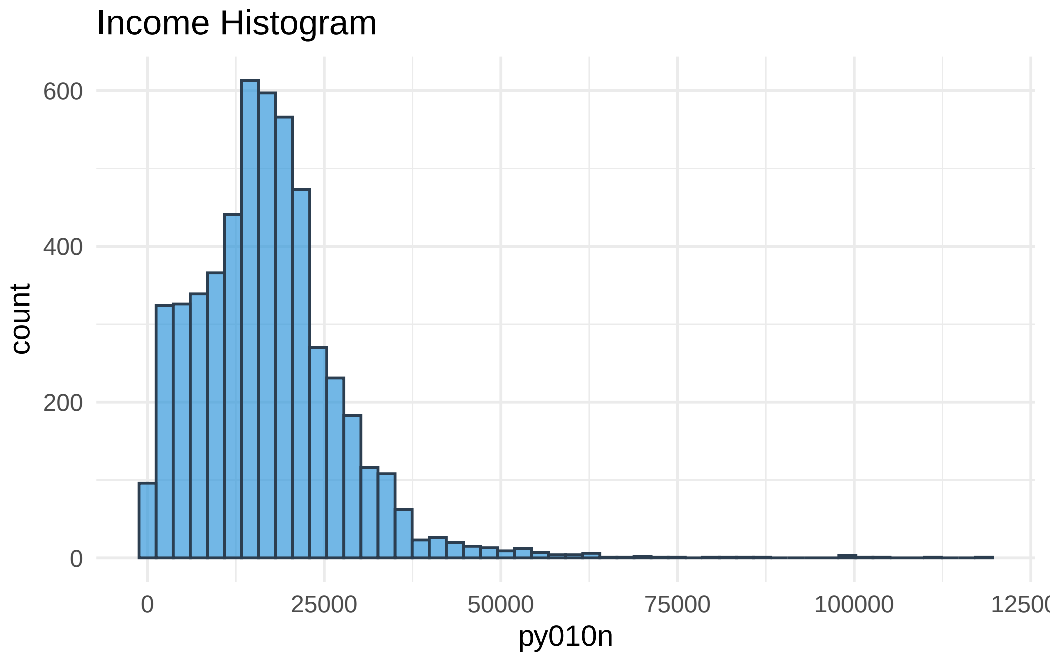
```
        
           AT   EU Other
  male   2867   40    97
  female 2154   39    70
```

3.3 3.3 Univariate Plots

```
# Income
ggplot(dat, aes(y=py010n)) +
  geom_boxplot(fill=secondary_color, alpha=0.7, color=main_color) +
  labs(title="Income Boxplot") +
  theme_minimal()
```

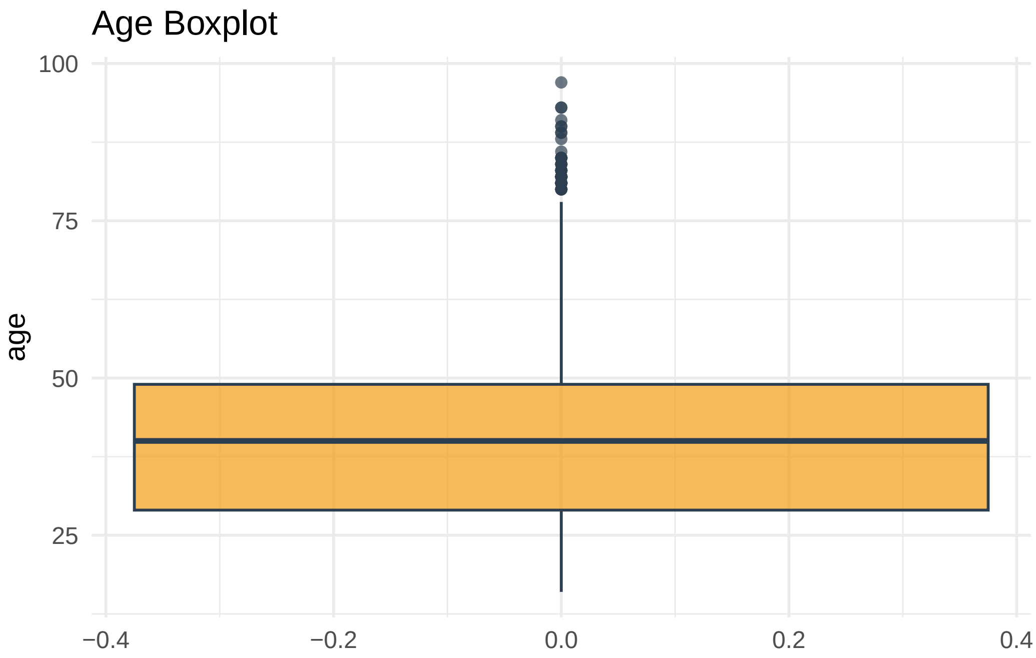


```
ggplot(dat, aes(x=py010n)) +
  geom_histogram(bins=50, fill=secondary_color, color=main_color, alpha=0.7) +
  labs(title="Income Histogram") +
  theme_minimal()
```

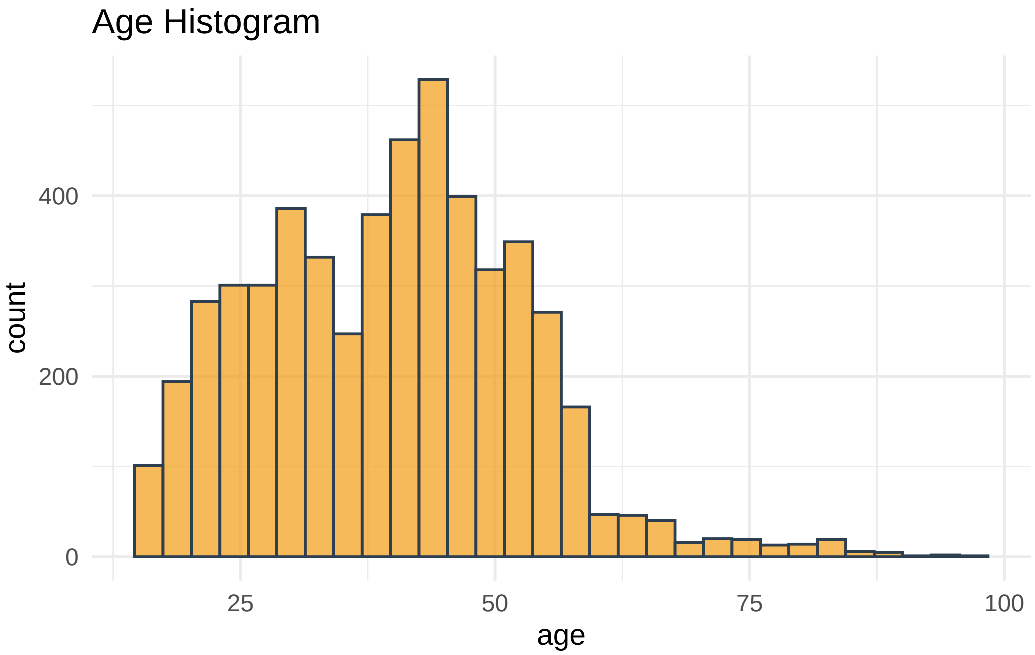


Interpretation: The income distribution shows strong right skewness with a median around €16,226 and mean of €16,952, indicating that most individuals earn below the mean. The boxplot reveals numerous high-income outliers extending up to approximately €118,000. The histogram confirms the concentration of incomes in the lower range (€0-€30,000) with a long tail of higher earners. This skewed distribution is typical for income data and suggests the presence of income inequality in the sample.

```r
# Age
ggplot(dat, aes(y=age)) +
  geom_boxplot(fill=tertiary_color, alpha=0.7, color=main_color) +
  labs(title="Age Boxplot") +
  theme_minimal()
```

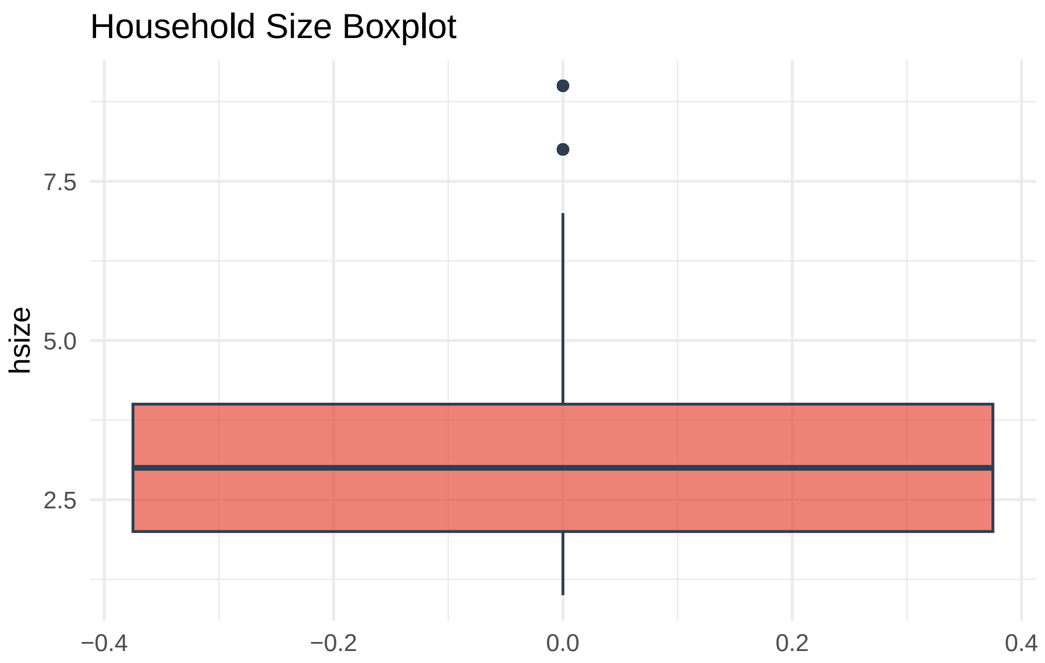


```r
ggplot(dat, aes(x=age)) +
  geom_histogram(bins=30, fill=tertiary_color, color=main_color, alpha=0.7) +
  labs(title="Age Histogram") +
  theme_minimal()
```

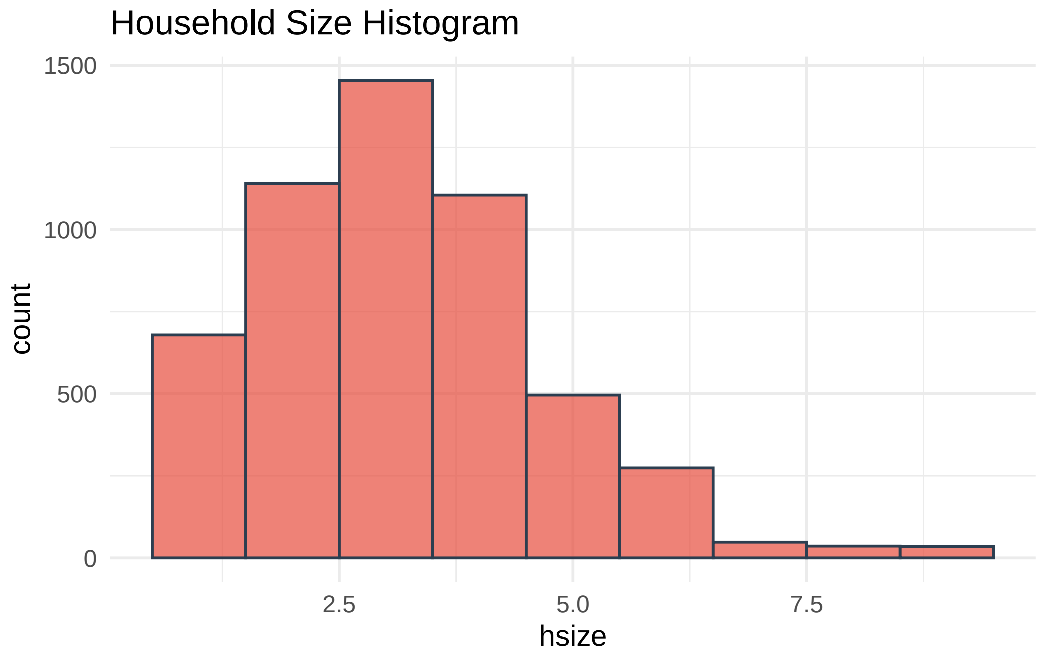


Interpretation: Age distribution is relatively uniform across the working-age population, with a median of 40 years and mean of 39.73 years. The range spans from 16 to 97 years, covering young workers to retirees. The histogram shows fairly even distribution across age groups with slight concentrations in the 25-50 age range, representing prime working years. The symmetric distribution suggests good representation across different career stages in the sample.

```r
# Household Size
ggplot(dat, aes(y=hsize)) +
  geom_boxplot(fill=accent_color, alpha=0.7, color=main_color) +
  labs(title="Household Size Boxplot") +
  theme_minimal()
```

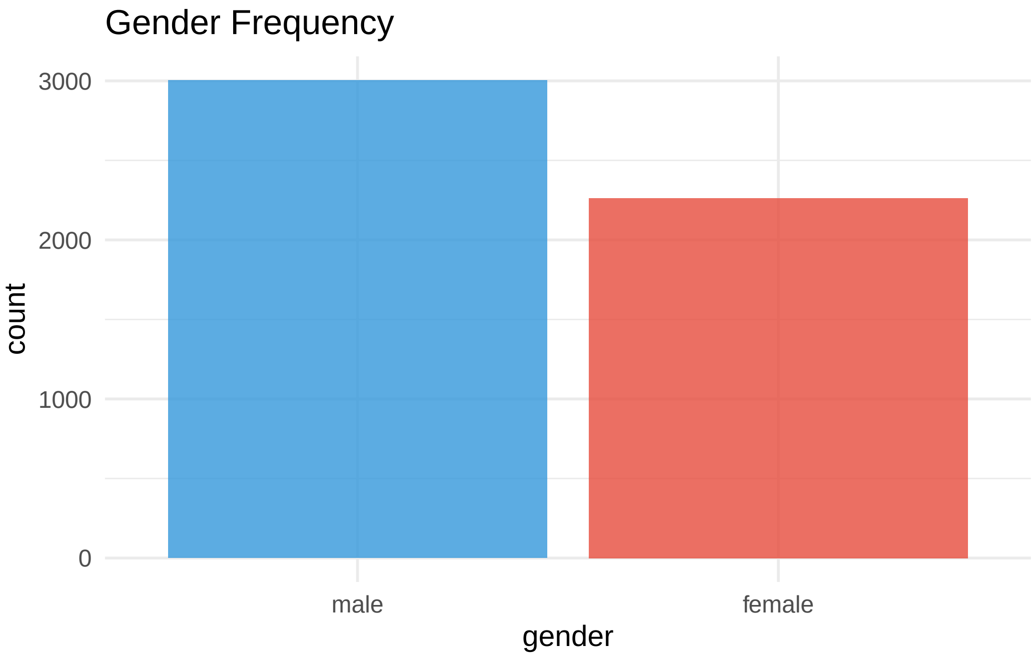


```
ggplot(dat, aes(x=hsize)) +
  geom_histogram(binwidth=1, fill=accent_color, color=main_color, alpha=0.7) +
  labs(title="Household Size Histogram") +
  theme_minimal()
```

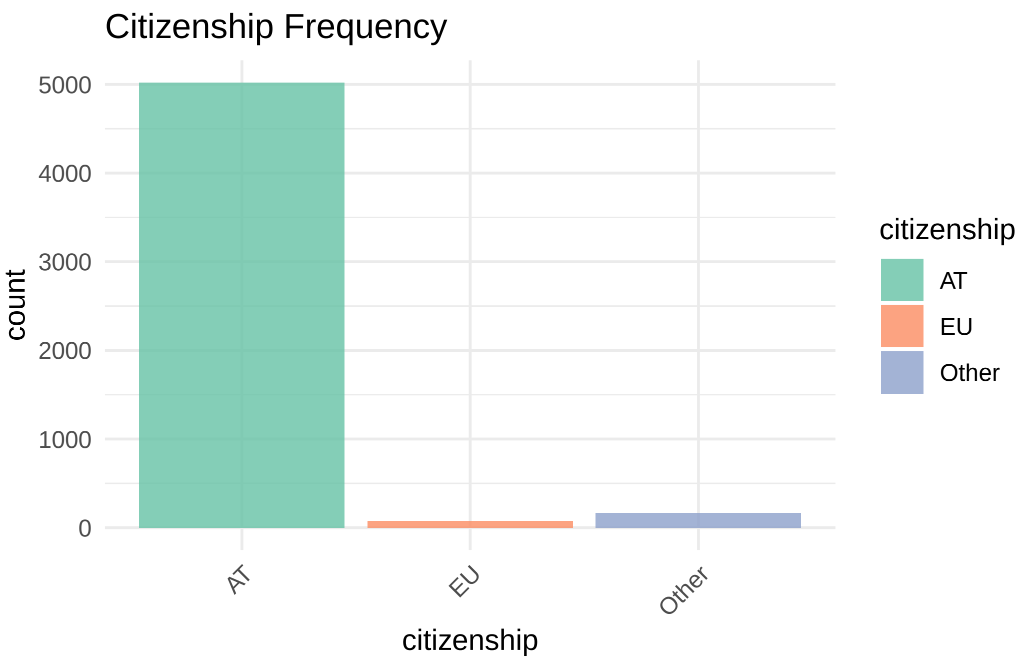


Interpretation: Household size shows a right-skewed distribution with median of 3 and mean of 3.19 members. Most households contain 2-4 members, with decreasing frequency as household size increases. The boxplot indicates some larger households (up to 9 members) as outliers. This distribution reflects typical family structures in Austria, with small to medium-sized households being most common.

```r
# Gender
ggplot(dat, aes(x=gender, fill=gender)) +
  geom_bar(alpha=0.8) +
  scale_fill_manual(values=c("female"="#E74C3C", "male"="#3498DB")) +
  labs(title="Gender Frequency") +
  theme_minimal() +
  theme(legend.position="none")
```



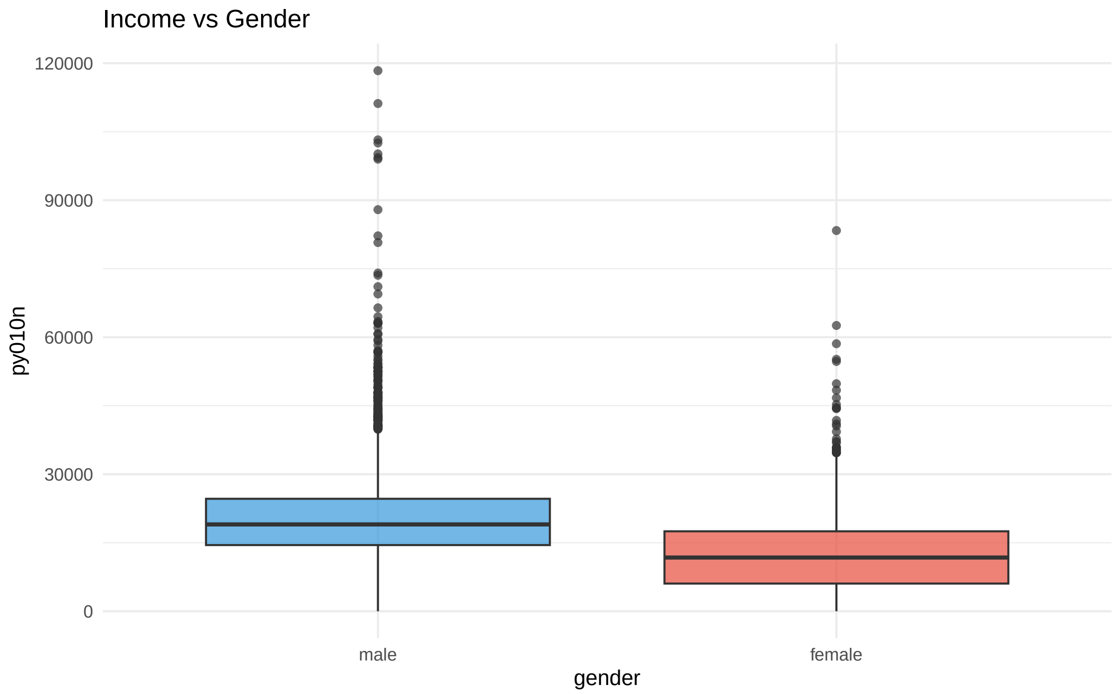
```r
# Citizenship
ggplot(dat, aes(x=citizenship, fill=citizenship)) +
  geom_bar(alpha=0.8) +
  scale_fill_brewer(palette="Set2") +
  theme_minimal() +
  theme(axis.text.x=element_text(angle=45, hjust=1)) +
  labs(title="Citizenship Frequency")
```



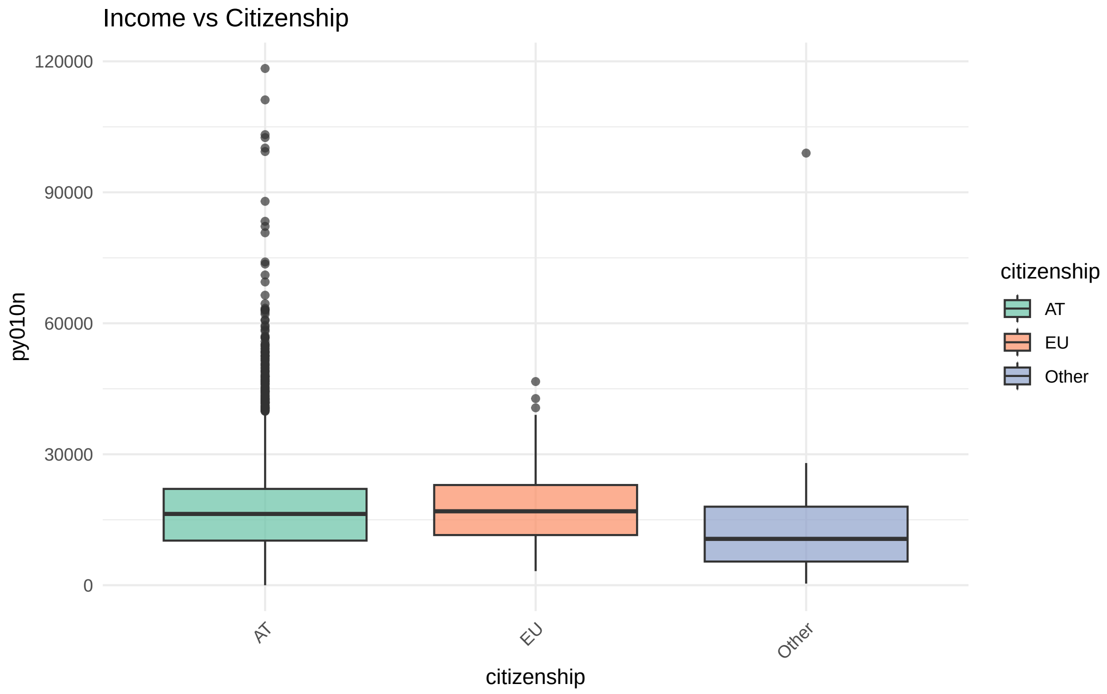
Interpretation: The sample contains 3,004 males and 2,263 females, showing a slight male majority (57% vs 43%). For citizenship, Austrian citizens (AT) dominate the sample with 5,021 individuals (95.3%), while EU citizens (79, 1.5%) and other citizenships (167, 3.2%) are minorities. This composition reflects Austria's demographic makeup with a predominantly native population and smaller immigrant communities.

3.4 3.4 Bivariate Plots

```
# Gender vs Income
ggplot(dat, aes(x=gender, y=py010n, fill=gender)) +
  geom_boxplot(alpha=0.7) +
  scale_fill_manual(values=c("female"="#E74C3C", "male"="#3498DB")) +
  labs(title="Income vs Gender") +
  theme_minimal() +
  theme(legend.position="none")
```



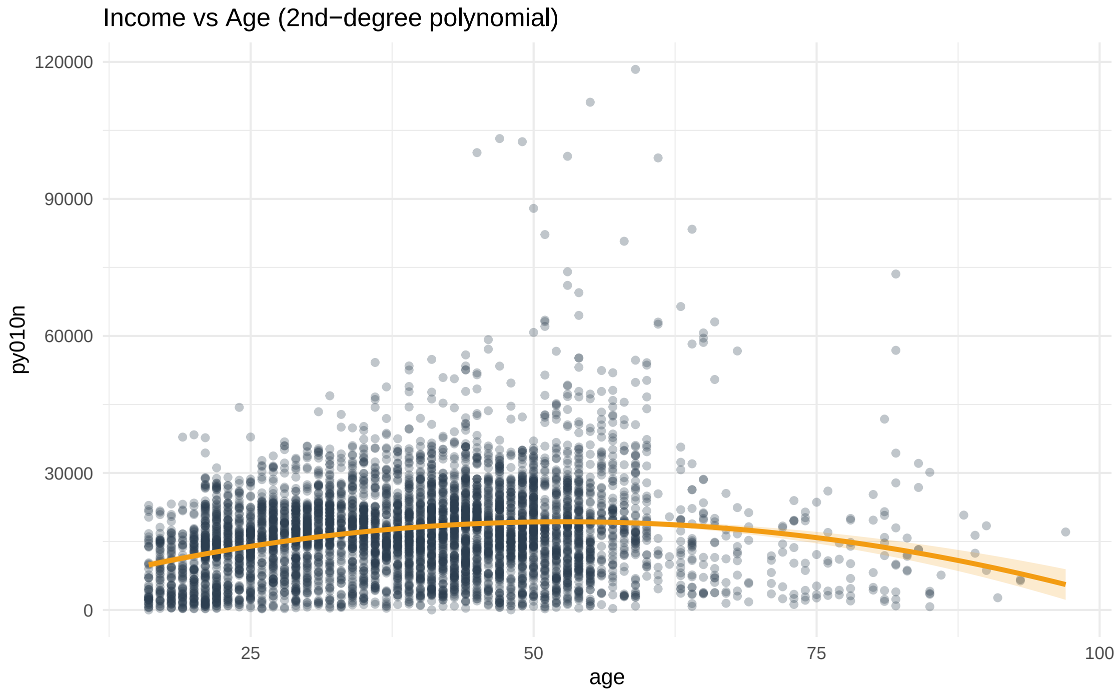
```
# Citizenship vs Income
ggplot(dat, aes(x=citizenship, y=py010n, fill=citizenship)) +
  geom_boxplot(alpha=0.7) +
  scale_fill_brewer(palette="Set2") +
  theme_minimal() +
  theme(axis.text.x=element_text(angle=45, hjust=1)) +
  labs(title="Income vs Citizenship")
```



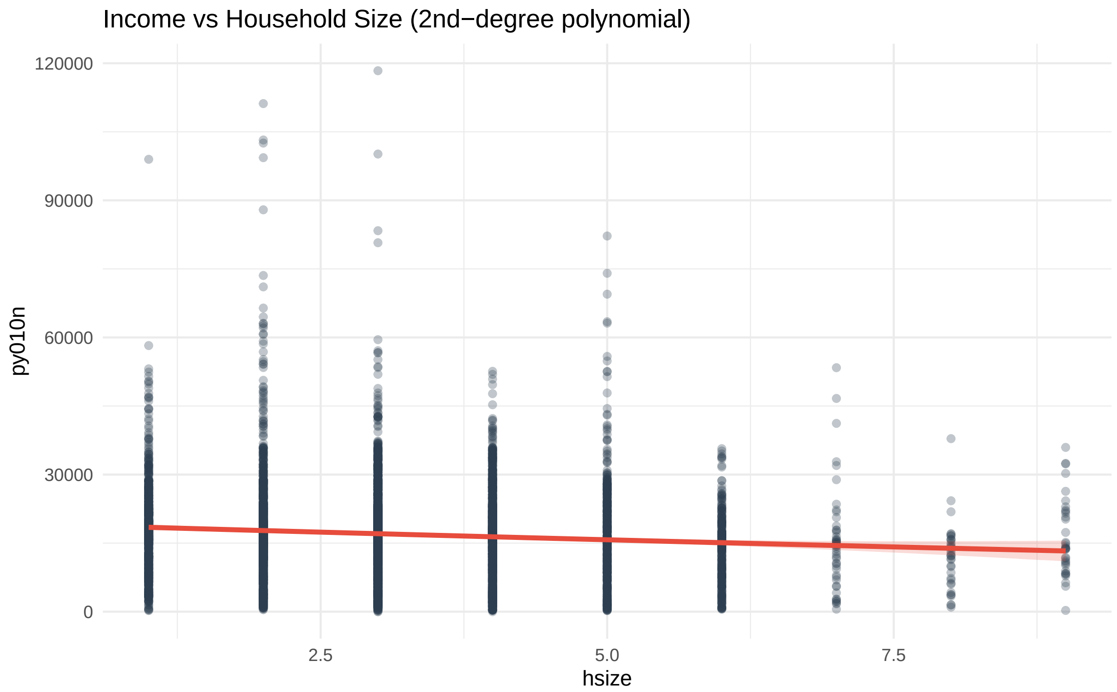
Interpretation: A substantial gender pay gap is evident, with males earning considerably more than females on average. Male median income appears around €18,000-€20,000 while female median is approximately €12,000-€14,000, suggesting roughly 30-40% lower earnings for women. Both groups show similar dispersion and outliers, but the entire female distribution is shifted downward.

For citizenship, Austrian and EU citizens show similar income distributions with medians around €16,000-€17,000. However, individuals with "Other" citizenship status earn noticeably less, with median income approximately €13,000-€14,000. This pattern may reflect employment barriers, language challenges, or credential recognition issues faced by non-EU immigrants.

```r
# Age vs Income (Polynomial)
ggplot(dat, aes(x=age, y=py010n)) +
  geom_point(alpha=0.3, color=main_color, size=1.5) +
  geom_smooth(method="lm", formula=y~poly(x,2), color=tertiary_color, fill=tertiary_color, alph
  labs(title="Income vs Age (2nd-degree polynomial)") +
  theme_minimal()
```



```r
# Household size vs Income (Polynomial)
ggplot(dat, aes(x=hsize, y=py010n)) +
  geom_point(alpha=0.3, color=main_color, size=1.5) +
  geom_smooth(method="lm", formula=y~poly(x,2), color=accent_color, fill=accent_color, alpha=0
  labs(title="Income vs Household Size (2nd-degree polynomial)") +
  theme_minimal()
```

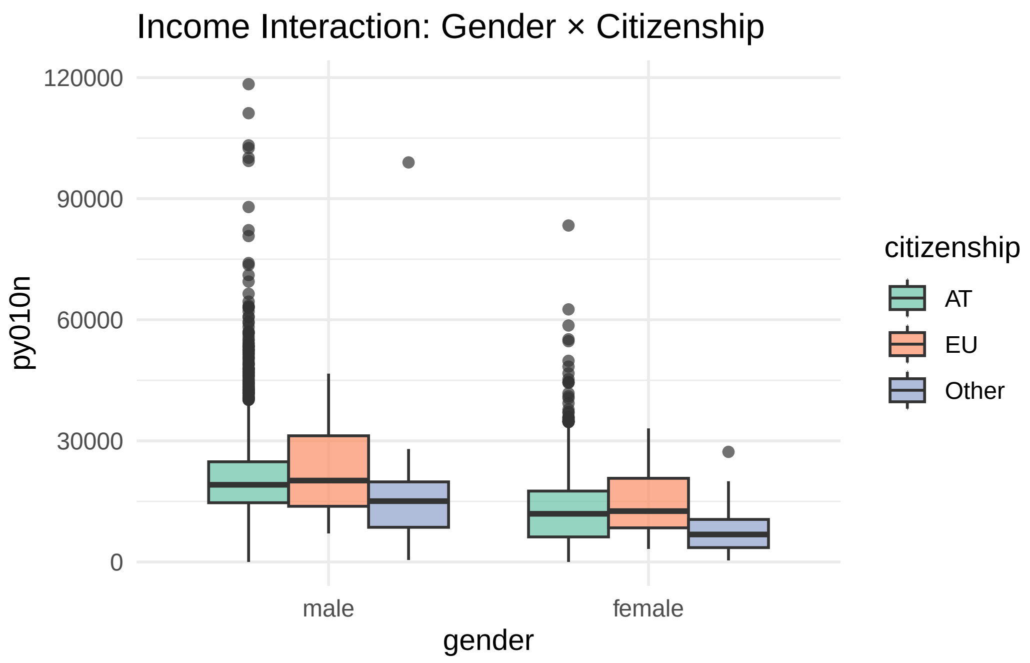


Interpretation: The age-income relationship exhibits a clear inverted U-shape (quadratic pattern). Income rises steadily from young workers (age 20-25) to peak around ages 45-55, then declines for older workers approaching retirement. This reflects typical career earnings trajectories: entry-level positions transitioning to peak earning years, followed by part-time work or reduced hours before retirement. The polynomial fit captures this life-cycle pattern better than a simple linear relationship.

Household size shows a negative relationship with income, with the polynomial curve declining as household size increases. Individuals in smaller households (1-3 members) tend to have higher incomes, while those in larger households (5+ members) earn less on average. This inverse relationship might reflect that larger families have more dependents per earner, or that high earners choose smaller family sizes. The slight curvature suggests the decline is steeper for very large households.

3.5 3.5 Interaction Plots

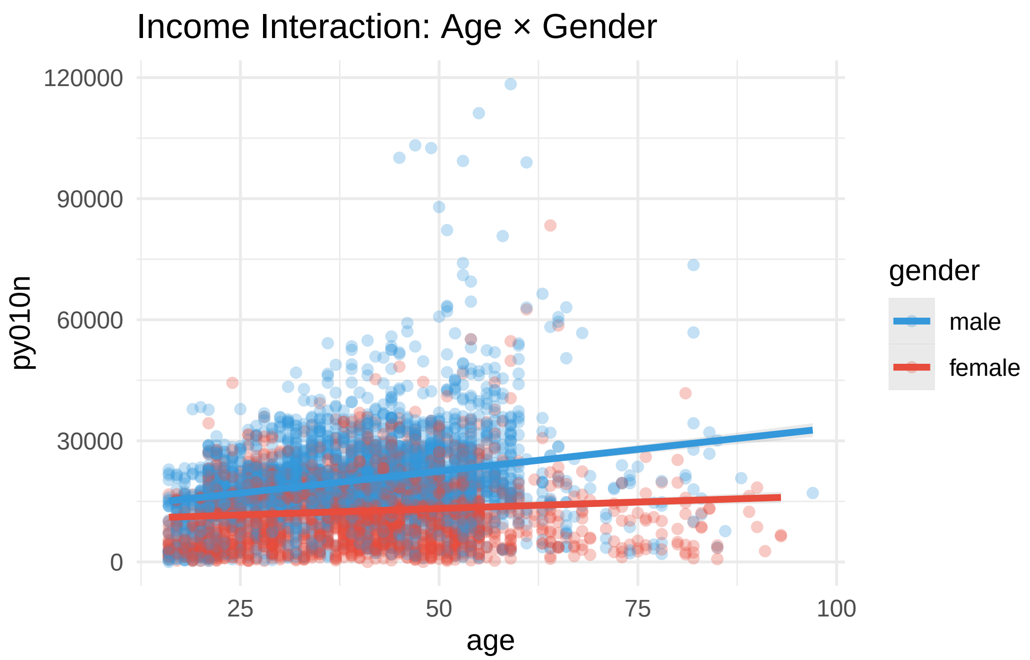
```
# Gender × Citizenship
ggplot(dat, aes(x=gender, y=py010n, fill=citizenship)) +
  geom_boxplot(position="dodge", alpha=0.7) +
  scale_fill_brewer(palette="Set2") +
  labs(title="Income Interaction: Gender × Citizenship") +
  theme_minimal()
```



Interpretation: The gender pay gap persists across all citizenship categories. For Austrian citizens, males earn substantially more than females. This gap appears similar for EU citizens, though the smaller sample size makes comparison less reliable. Interestingly, for "Other" citizenship, both males and females earn less than their Austrian counterparts, but the gender gap remains. This suggests that citizenship status and gender have independent negative effects on income, with no strong evidence of interaction between these factors.

```r
# Age × Gender
ggplot(dat, aes(x=age, y=py010n, color=gender)) +
  geom_point(alpha=0.3, size=1.5) +
  geom_smooth(method="lm", se=TRUE, alpha=0.2, linewidth=1.2) +
  scale_color_manual(values=c("female"="#E74C3C", "male"="#3498DB")) +
  labs(title="Income Interaction: Age × Gender") +
  theme_minimal()
```

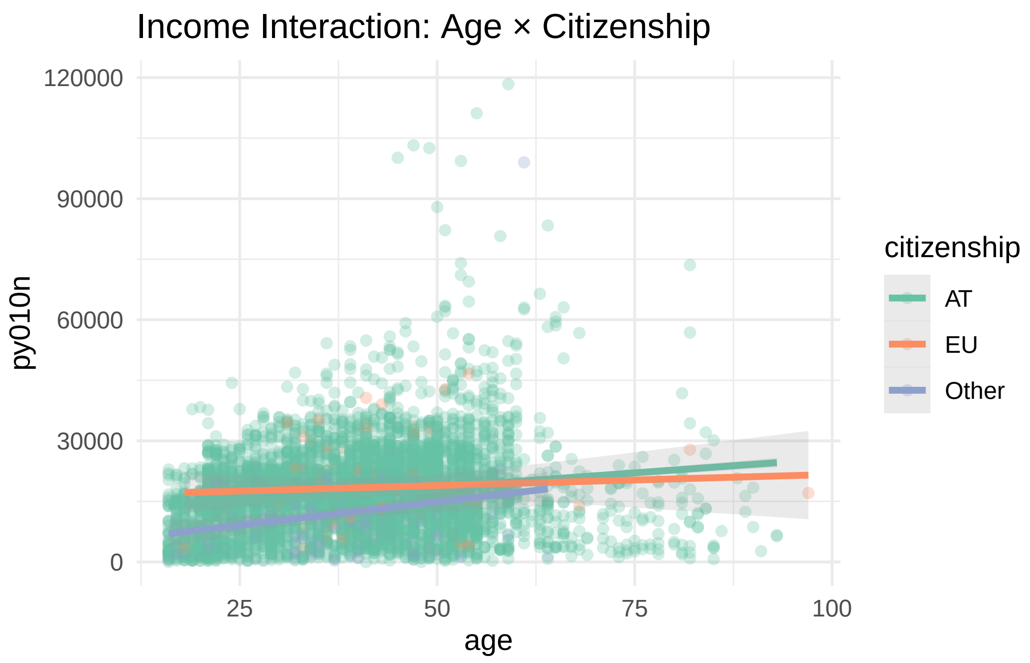
```
`geom_smooth()` using formula = 'y ~ x'
```



Interpretation: Both males and females show positive age-income relationships, but with different slopes. Male income increases more steeply with age compared to females, indicating that the gender pay gap widens over the career lifecycle. Young males and females (ages 20-30) have more similar earnings, but by age 50-60, the gap has substantially widened. This pattern suggests cumulative disadvantages for women over their careers, possibly due to career interruptions, glass ceiling effects, or occupational segregation becoming more pronounced with age.

```
# Age × Citizenship
ggplot(dat, aes(x=age, y=py010n, color=citizenship)) +
  geom_point(alpha=0.3, size=1.5) +
  geom_smooth(method="lm", se=TRUE, alpha=0.2, linewidth=1.2) +
  scale_color_brewer(palette="Set2") +
  labs(title="Income Interaction: Age × Citizenship") +
  theme_minimal()
```

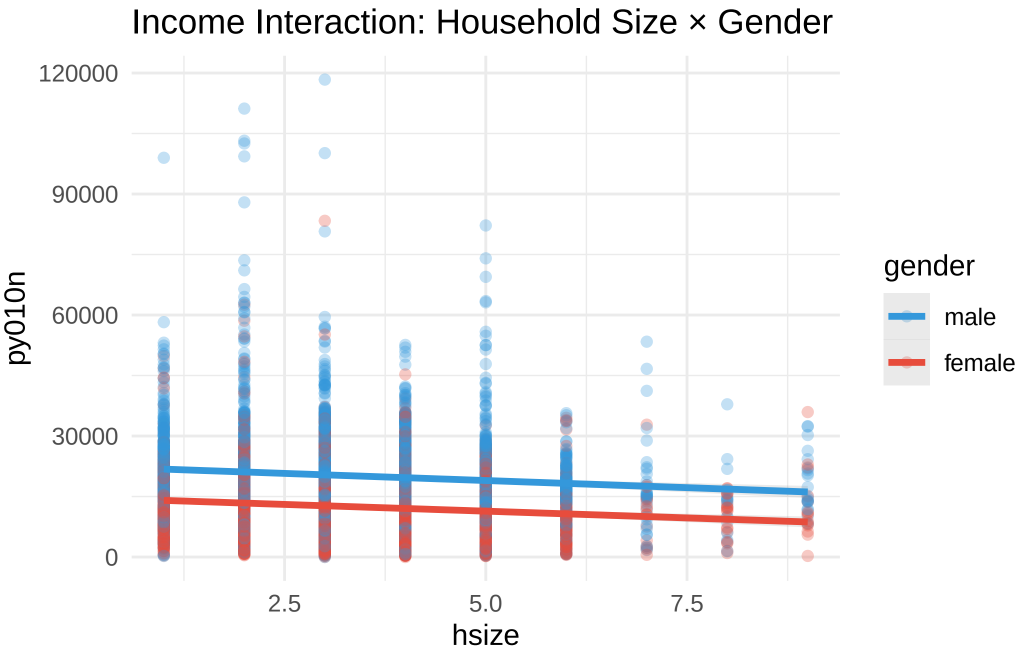
```
`geom_smooth()` using formula = 'y ~ x'
```



Interpretation: The age-income profile differs notably across citizenship groups. Austrian citizens show a strong positive relationship between age and income, reflecting normal career progression. EU citizens display a similar but slightly flatter trajectory. However, individuals with "Other" citizenship show a much weaker or even flat age-income relationship, suggesting they may face barriers to career advancement regardless of experience. This could indicate difficulties in credential recognition, discrimination, or concentration in jobs with limited advancement opportunities.

```
# Household Size × Gender
ggplot(dat, aes(x=hsize, y=py010n, color=gender)) +
  geom_point(alpha=0.3, size=1.5) +
  geom_smooth(method="lm", se=TRUE, alpha=0.2, linewidth=1.2) +
  scale_color_manual(values=c("female"="#E74C3C", "male"="#3498DB")) +
  labs(title="Income Interaction: Household Size × Gender") +
  theme_minimal()
```

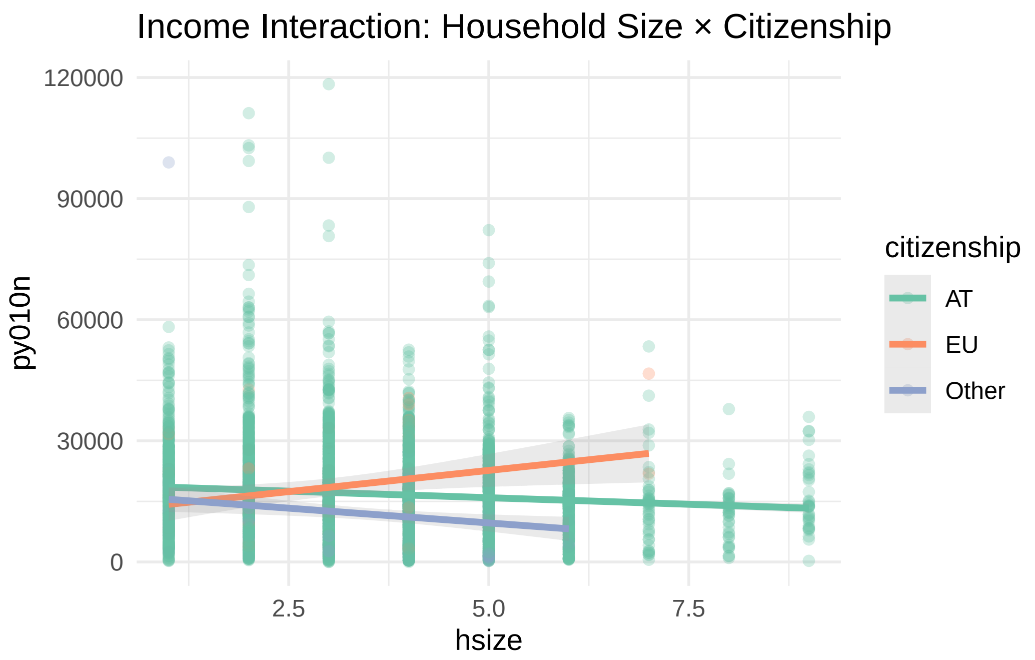
```
`geom_smooth()` using formula = 'y ~ x'
```



Interpretation: Both genders show negative relationships between household size and income, but the pattern differs. For males, income remains relatively stable or decreases only slightly with household size. For females, there's a steeper negative slope, indicating that women in larger households earn substantially less. This interaction suggests that household responsibilities (likely childcare) disproportionately affect women's earnings, with the penalty increasing as household size grows. This reflects traditional gender roles where women may reduce work hours or career investment when household/family demands increase.

```
# Household Size × Citizenship
ggplot(dat, aes(x=hsize, y=py010n, color=citizenship)) +
  geom_point(alpha=0.3, size=1.5) +
  geom_smooth(method="lm", se=TRUE, alpha=0.2, linewidth=1.2) +
  scale_color_brewer(palette="Set2") +
  labs(title="Income Interaction: Household Size × Citizenship") +
  theme_minimal()
```

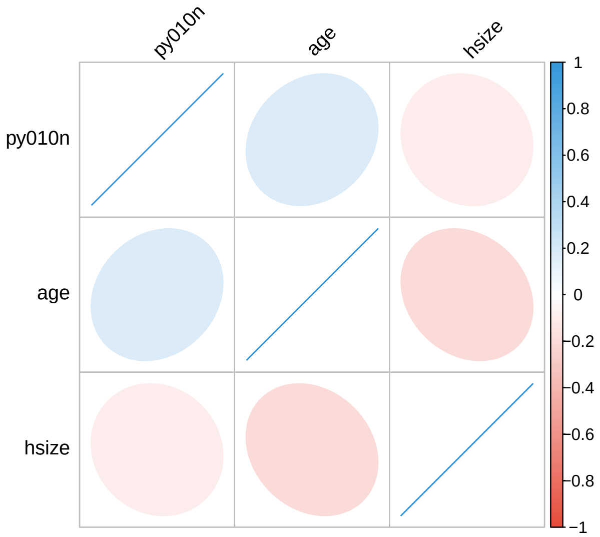
```
`geom_smooth()` using formula = 'y ~ x'
```

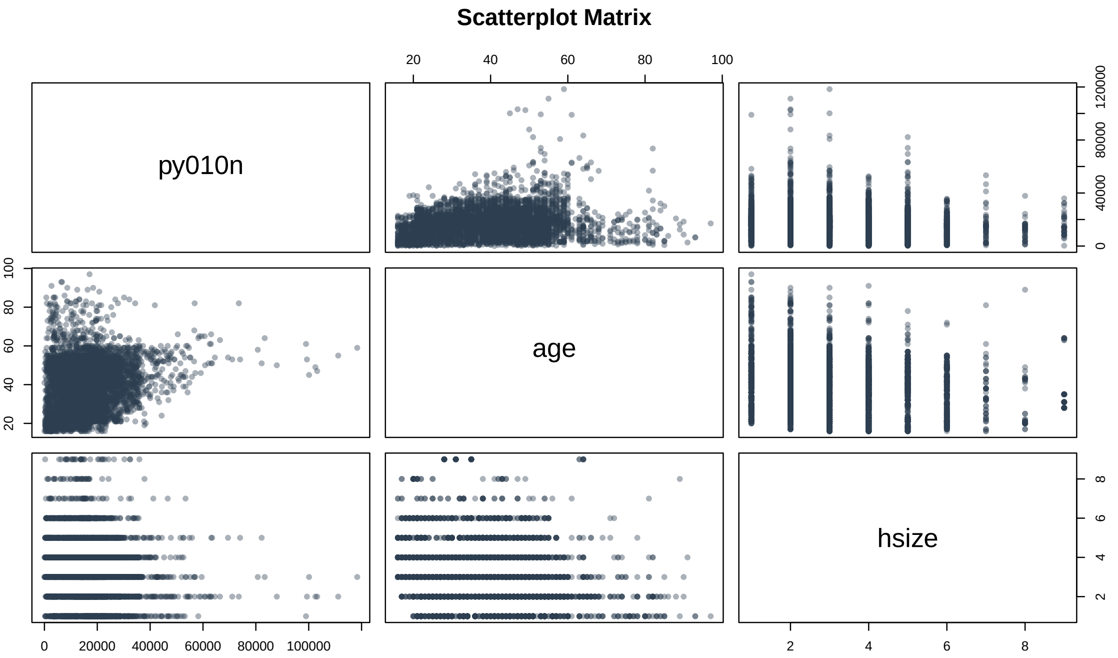


Interpretation: The negative relationship between household size and income appears across all citizenship groups, but with varying strengths. Austrian citizens show a moderate negative slope. EU citizens display high variability (wide confidence bands) due to small sample size, making interpretation difficult. “Other” citizenship individuals start with lower baseline income and show a similar negative trend with household size. This suggests that household size impacts earnings regardless of citizenship, though the already-disadvantaged position of non-Austrian citizens compounds this effect.

3.6 3.6 Correlation / Scatterplot Matrix

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

cor_mat <- cor(num_vars)
corrplot(cor_mat, method="ellipse",
         col=colorRampPalette(c("#E74C3C", "white", "#3498DB"))(200),
         tl.col="black", tl.srt=45)

pairs(num_vars,
      main="Scatterplot Matrix",
      col=alpha(main_color, 0.4),
      pch=19,
      cex=0.8)
```





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

Interpretation: The correlation matrix reveals weak to moderate relationships among the numeric variables. Age shows a positive but weak correlation with income (r 0.25-0.30), consistent with the inverted U-shape observed in bivariate plots—the linear correlation captures only part of this nonlinear relationship. Household size has a weak negative correlation with income (r -0.10 to

-0.15), suggesting that larger households are associated with slightly lower individual earnings. Age and household size show minimal correlation (r 0.05-0.10), indicating these are relatively independent predictors.

The scatterplot matrix confirms these patterns visually. The age-income plot shows substantial scatter with a slight upward trend, indicating age explains only a small portion of income variance. The household size-income plot displays a cloud of points with a gentle downward slope. The age-household size relationship appears nearly random, with no clear pattern. Overall, these correlations suggest that while age and household size are statistically significant predictors, much of the income variation remains unexplained by these numeric variables alone, highlighting the importance of categorical variables (gender, citizenship) and interaction effects in the full regression models.

4 4. Regression Modeling

4.1 4.1 Linear Model

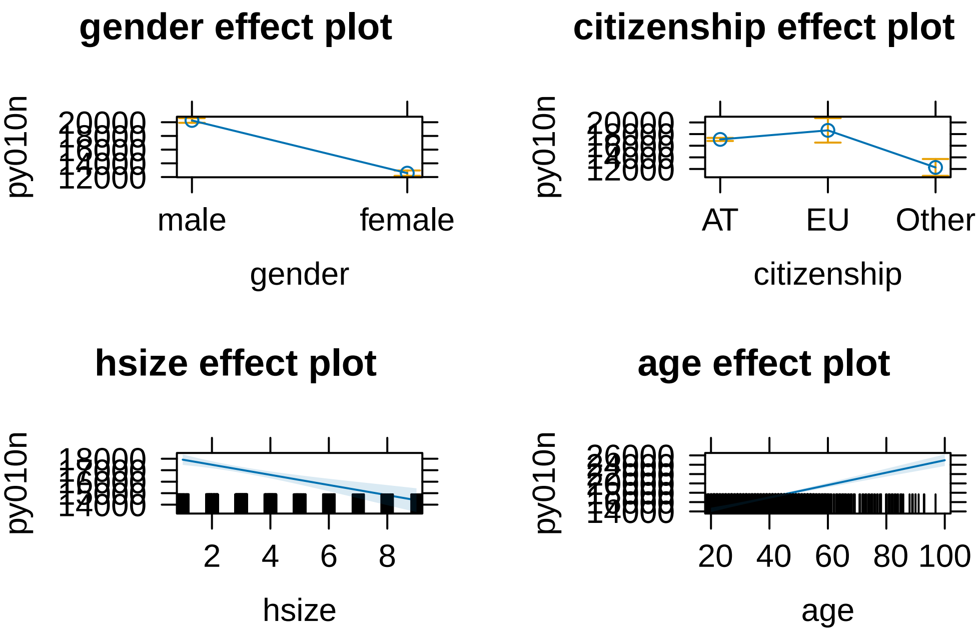
```r
lm_model <- lm(py010n ~ gender + citizenship + hsize + age, data=dat)
summary(lm_model)
```

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

Residuals:
   Min     1Q Median     3Q    Max 
-23585  -5909  -1124   4619  95330 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      16509.98     567.51  29.092  < 2e-16 ***
genderfemale     -7688.82     265.03 -29.012  < 2e-16 ***
citizenshipEU     1556.37    1079.80   1.441     0.15    
citizenshipOther -4813.39     750.01  -6.418 1.50e-10 ***
hsize             -441.49      88.13  -5.009 5.64e-07 ***
age                133.00      10.29  12.923  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9520 on 5261 degrees of freedom
Multiple R-squared:  0.1748,	Adjusted R-squared:  0.174 
F-statistic: 222.8 on 5 and 5261 DF,  p-value: < 2.2e-16
```

```r
Anova(lm_model)
```

```
Anova Table (Type II tests)

Response: py010n
                Sum Sq   Df F value    Pr(>F)    
gender      7.6276e+10    1 841.667 < 2.2e-16 ***
citizenship 3.9581e+09    2  21.838 3.591e-10 ***
hsize       2.2740e+09    1  25.093 5.642e-07 ***
age         1.5135e+10    1 167.003 < 2.2e-16 ***
Residuals   4.7678e+11 5261                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(allEffects(lm_model))
```



Interpretation: The linear model explains 17.5% of income variance ($R^2 = 0.175$). All predictors except EU citizenship are statistically significant.

Key Findings: - **Gender:** Females earn €7,689 less than males on average, holding other variables constant ($p < 0.001$). This represents approximately a 40-45% pay gap, the strongest predictor in the model. - **Citizenship:** Individuals with “Other” citizenship earn €4,813 less than Austrian citizens ($p < 0.001$), while EU citizens show no significant difference ($p = 0.15$). - **Household Size:** Each additional household member is associated with €441 lower income ($p < 0.001$), consistent with increased family responsibilities. - **Age:** Each additional year of age increases income by €133 ($p < 0.001$), reflecting experience and career progression.

Effect Plots: The plots show gender's large impact (vertical separation), citizenship differences (mainly for "Other"), the linear increase with age, and the modest negative household size effect. However, the linear age specification may not capture the inverted U-shape observed in exploratory analysis, motivating the polynomial models.

4.2 4.2 Polynomial Model

```r
poly_model <- lm(py010n ~ gender + citizenship + poly(age,2) + poly(hsize,2), data=dat)
summary(poly_model)
```

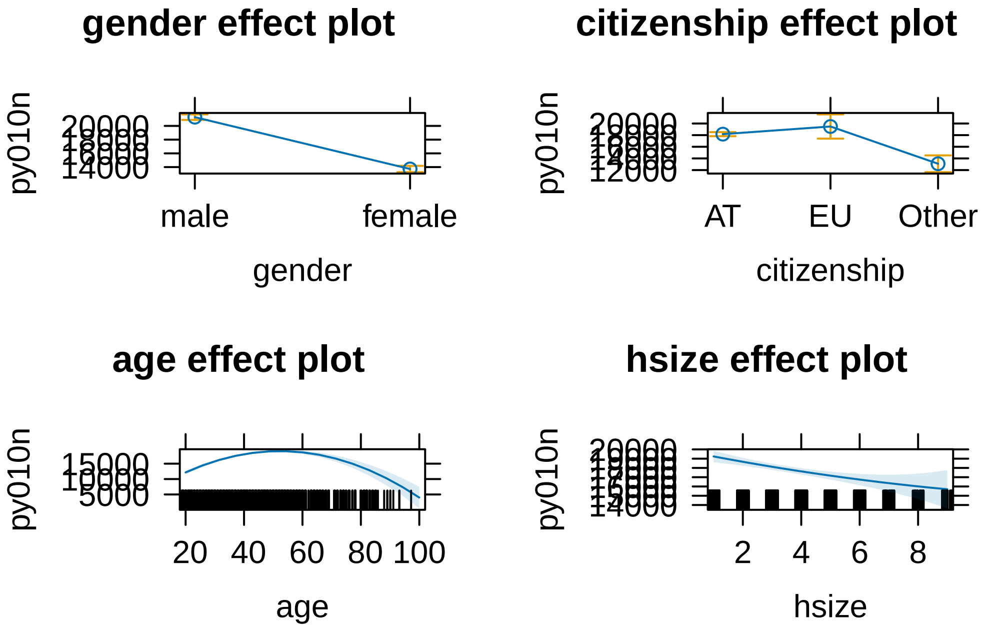
```
Call:
lm(formula = py010n ~ gender + citizenship + poly(age, 2) + poly(hsize,
    2), data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-22110  -5804  -1103   4623  96100

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        20337.7      173.5 117.236  < 2e-16 ***
genderfemale       -7549.7      261.4 -28.882  < 2e-16 ***
citizenshipEU       1313.6     1064.3   1.234    0.217
citizenshipOther   -5085.4      739.8  -6.874 6.95e-12 ***
poly(age, 2)1     124087.3     9589.4  12.940  < 2e-16 ***
poly(age, 2)2    -118841.7     9446.7 -12.580  < 2e-16 ***
poly(hsize, 2)1   -53916.1     9582.7  -5.626 1.94e-08 ***
poly(hsize, 2)2     4848.6     9445.8   0.513    0.608
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9381 on 5259 degrees of freedom
Multiple R-squared:  0.1989,    Adjusted R-squared:  0.1979
F-statistic: 186.6 on 7 and 5259 DF,  p-value: < 2.2e-16
```

```r
Anova(poly_model)
```

```
Anova Table (Type II tests)

Response: py010n
                   Sum Sq   Df F value    Pr(>F)
gender         7.3409e+10    1 834.159 < 2.2e-16 ***
citizenship    4.3277e+09    2  24.588 2.350e-11 ***
```

```
poly(age, 2)   2.9102e+10    2 165.348 < 2.2e-16 ***
poly(hsize, 2) 2.8014e+09    2  15.916 1.284e-07 ***
Residuals      4.6281e+11 5259
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(allEffects(poly_model))
```



Interpretation: The polynomial model improves fit to $R^2 = 0.199$ (vs. 0.175 for linear), capturing nonlinear relationships in age and household size. The ANOVA shows both polynomial terms for age are highly significant ($p < 0.001$), confirming the inverted U-shape.

Key Findings: - **Gender:** Effect remains substantial at €7,550 ($p < 0.001$), nearly unchanged from the linear model. - **Citizenship:** “Other” citizenship penalty increases to €5,085 ($p < 0.001$). EU citizenship remains non-significant. - **Age (Quadratic):** Both linear (positive) and quadratic (negative) terms are significant, confirming income rises then falls with age. The peak occurs around ages 45-50. - **Household Size (Quadratic):** The linear term is significant ($p < 0.001$) but quadratic term is not ($p = 0.608$), suggesting the relationship is primarily linear negative.

Effect Plots: The age plot now shows the characteristic inverted U-curve, peaking in mid-career. The household size plot displays a slight curve but is nearly linear. Gender and citizenship effects remain similar to the linear model. This specification better captures life-cycle earnings patterns while maintaining simplicity.

4.3 4.3 Spline Model

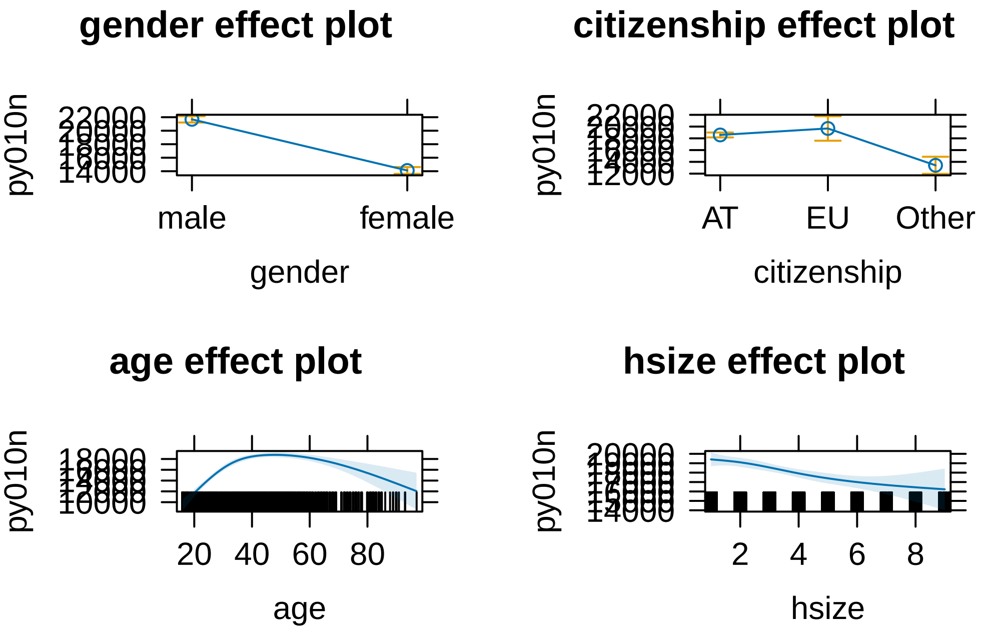
```r
spline_model <- lm(py010n ~ gender + citizenship + ns(age, df=3) + ns(hsize, df=3), data=dat)
summary(spline_model)
```

```
Call:
lm(formula = py010n ~ gender + citizenship + ns(age, df = 3) + 
    ns(hsize, df = 3), data = dat)

Residuals:
   Min     1Q Median     3Q    Max 
-21696  -5815  -1128   4546  96547 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         13913.9      630.7  22.060  < 2e-16 ***
genderfemale        -7571.9      261.3 -28.980  < 2e-16 ***
citizenshipEU        1103.1     1065.2   1.036  0.30047    
citizenshipOther    -5167.3      741.3  -6.971 3.54e-12 ***
ns(age, df = 3)1     7206.4      667.8  10.791  < 2e-16 ***
ns(age, df = 3)2    13315.5     1497.8   8.890  < 2e-16 ***
ns(age, df = 3)3    -1491.0     1724.2  -0.865  0.38721    
ns(hsize, df = 3)1  -2217.2      678.4  -3.268  0.00109 ** 
ns(hsize, df = 3)2  -3115.8      979.8  -3.180  0.00148 ** 
ns(hsize, df = 3)3  -3011.2     1093.9  -2.753  0.00593 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9375 on 5257 degrees of freedom
Multiple R-squared:  0.2003,	Adjusted R-squared:  0.1989 
F-statistic: 146.3 on 9 and 5257 DF,  p-value: < 2.2e-16
```

```r
Anova(spline_model)
```

```
Anova Table (Type II tests)

Response: py010n
                      Sum Sq   Df F value    Pr(>F)    
gender            7.3816e+10    1 839.862 < 2.2e-16 ***
citizenship       4.3999e+09    2  25.031 1.516e-11 ***
ns(age, df = 3)   2.9832e+10    3 113.141 < 2.2e-16 ***
ns(hsize, df = 3) 2.8699e+09    3  10.884 3.996e-07 ***
Residuals         4.6204e+11 5257                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(allEffects(spline_model, xlevels=50))
```



Interpretation: The spline model ($R^2 = 0.200$) provides similar fit to the polynomial model but with greater flexibility through piecewise cubic polynomials. Using 3 degrees of freedom for age and household size allows the relationship to change shape at different points.

Key Findings: - **Gender:** Effect is €7,572 ($p < 0.001$), consistent across all models. - **Citizenship:** "Other" citizenship penalty increases slightly to €5,167 ($p < 0.001$). EU remains non-significant. - **Age Splines:** All three spline basis functions show the age-income relationship varies across the age range. Two terms are highly significant ($p < 0.001$), one is not ($p = 0.387$), suggesting different income growth rates in early career vs. mid-career vs. late career. - **Household Size Splines:** All three terms are significant ($p < 0.01$), indicating the relationship between household size and income may have multiple inflection points.

Effect Plots: The age plot shows a smoother inverted U-curve than the polynomial, with steeper increase in early career and more gradual decline post-peak. The household size plot reveals more nuanced changes, with steeper decline for very small and very large households. The spline approach captures local variations the polynomial might miss, providing a more flexible representation of nonlinear effects.

4.4 4.4 Interaction Model

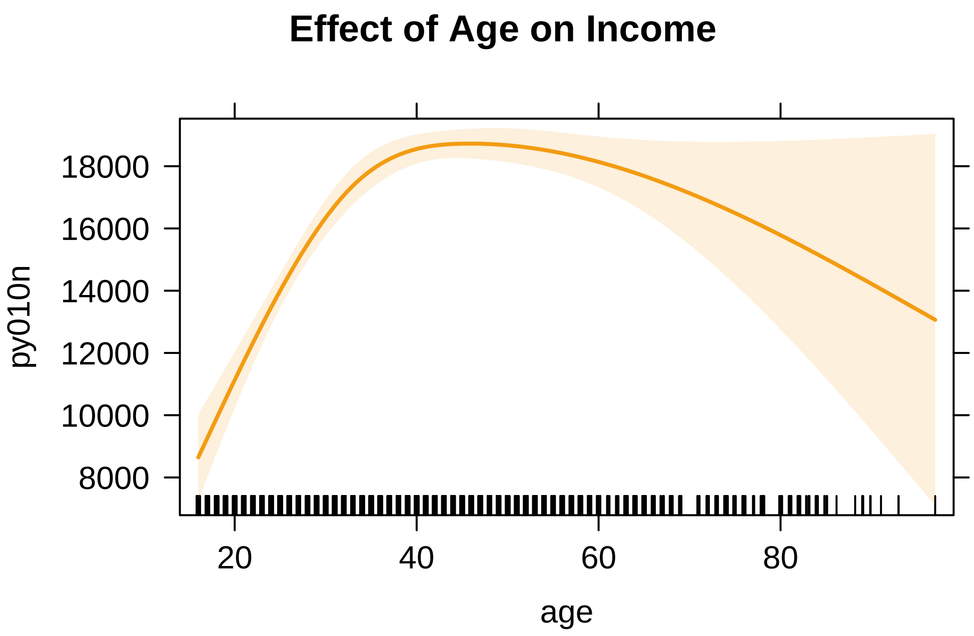
```r
int_model <- lm(py010n ~ (gender + citizenship + ns(age,3) + ns(hsize,3))^2, data=dat)
summary(int_model)
```

```
Call:
lm(formula = py010n ~ (gender + citizenship + ns(age, 3) + ns(hsize,
    3))^2, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-22493  -5814  -1051   4526  95508

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     13460.2     2193.4   6.137 9.05e-10 ***
genderfemale                    -3694.9     1266.9  -2.917 0.003555 **
citizenshipEU                   -6592.6     7259.9  -0.908 0.363873
citizenshipOther                 1020.4     4019.5   0.254 0.799603
ns(age, 3)1                      8509.7     1784.0   4.770 1.89e-06 ***
ns(age, 3)2                     15438.9     5186.9   2.977 0.002929 **
ns(age, 3)3                      2803.1     4155.8   0.675 0.500016
ns(hsize, 3)1                   -2041.1     2810.7  -0.726 0.467759
ns(hsize, 3)2                   -4369.1     5098.4  -0.857 0.391513
ns(hsize, 3)3                    -389.4     4311.8  -0.090 0.928045
genderfemale:citizenshipEU       1657.6     2257.6   0.734 0.462835
genderfemale:citizenshipOther     393.0     1497.5   0.262 0.792993
genderfemale:ns(age, 3)1        -6448.7     1381.4  -4.668 3.11e-06 ***
genderfemale:ns(age, 3)2        -9135.9     3149.9  -2.900 0.003743 **
genderfemale:ns(age, 3)3        -7029.8     3783.8  -1.858 0.063245 .
genderfemale:ns(hsize, 3)1      -3606.2     1394.3  -2.586 0.009723 **
genderfemale:ns(hsize, 3)2        430.2     2100.6   0.205 0.837740
genderfemale:ns(hsize, 3)3       2077.1     2373.4   0.875 0.381540
citizenshipEU:ns(age, 3)1       -1964.3     6024.1  -0.326 0.744376
citizenshipOther:ns(age, 3)1   -14492.4     5840.7  -2.481 0.013123 *
citizenshipEU:ns(age, 3)2       11401.6    14368.0   0.794 0.427499
citizenshipOther:ns(age, 3)2    59054.6    19794.1   2.983 0.002863 **
citizenshipEU:ns(age, 3)3        3107.6     9574.6   0.325 0.745518
citizenshipOther:ns(age, 3)3   111889.8    32948.7   3.396 0.000689 ***
citizenshipEU:ns(hsize, 3)1      8157.1     6825.6   1.195 0.232116
citizenshipOther:ns(hsize, 3)1   1033.3     5271.1   0.196 0.844590
citizenshipEU:ns(hsize, 3)2     14573.1    11094.8   1.314 0.189072
citizenshipOther:ns(hsize, 3)2 -16978.8    11185.9  -1.518 0.129105
citizenshipEU:ns(hsize, 3)3     26088.3    15305.3   1.705 0.088342 .
citizenshipOther:ns(hsize, 3)3  -5405.8    16142.2  -0.335 0.737724
ns(age, 3)1:ns(hsize, 3)1        2636.3     3842.2   0.686 0.492659
ns(age, 3)2:ns(hsize, 3)1       -9483.7     9011.8  -1.052 0.292683
ns(age, 3)3:ns(hsize, 3)1      -23699.8    12037.9  -1.969 0.049032 *
ns(age, 3)1:ns(hsize, 3)2        2699.1     5119.5   0.527 0.598064
ns(age, 3)2:ns(hsize, 3)2        2817.2    11878.7   0.237 0.812540
ns(age, 3)3:ns(hsize, 3)2        2329.1    10580.4   0.220 0.825779
ns(age, 3)1:ns(hsize, 3)3        -183.0     6593.3  -0.028 0.977860
ns(age, 3)2:ns(hsize, 3)3       -2574.4    11412.4  -0.226 0.821536
ns(age, 3)3:ns(hsize, 3)3       10763.5    14573.6   0.739 0.460208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9316 on 5228 degrees of freedom
Multiple R-squared:  0.2147,    Adjusted R-squared:  0.209
F-statistic: 37.61 on 38 and 5228 DF,  p-value: < 2.2e-16
```

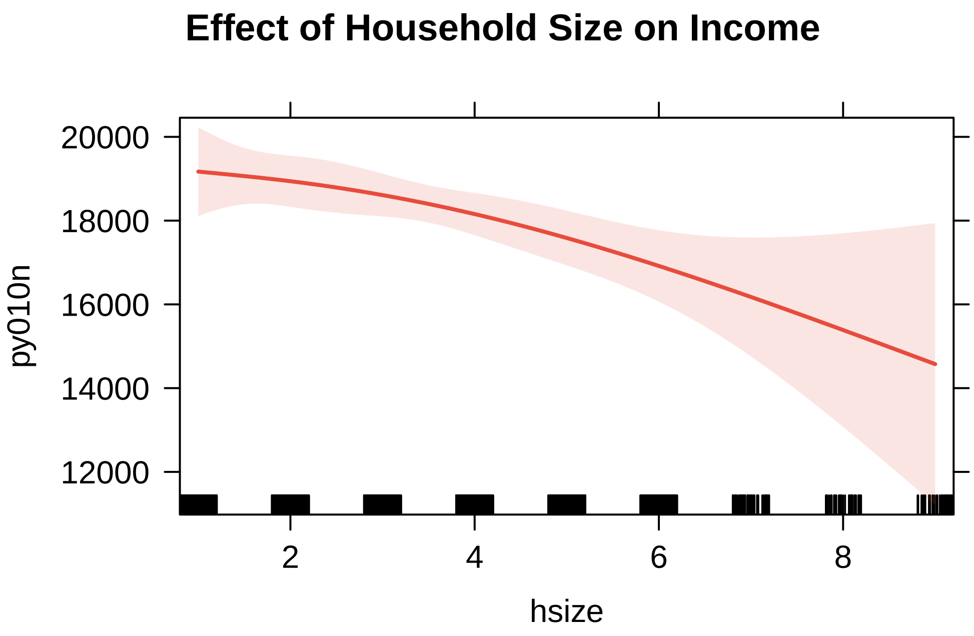
```
Anova(int_model)
```

```
Anova Table (Type II tests)

Response: py010n
                              Sum Sq   Df  F value    Pr(>F)    
gender                     7.2912e+10    1 840.1387 < 2.2e-16 ***
citizenship                4.3629e+09    2  25.1364 1.366e-11 ***
ns(age, 3)                 2.9791e+10    3 114.4234 < 2.2e-16 ***
ns(hsize, 3)               2.7869e+09    3  10.7044 5.188e-07 ***
gender:citizenship         5.1964e+07    2   0.2994   0.74129    
gender:ns(age, 3)          4.2695e+09    3  16.3985 1.324e-10 ***
gender:ns(hsize, 3)        6.0021e+08    3   2.3053   0.07475 .  
citizenship:ns(age, 3)     1.2057e+09    6   2.3154   0.03101 *  
citizenship:ns(hsize, 3) 1.3626e+09    6   2.6169   0.01556 *  
ns(age, 3):ns(hsize, 3)  7.4561e+08    9   0.9546   0.47594    
Residuals                  4.5371e+11 5228                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

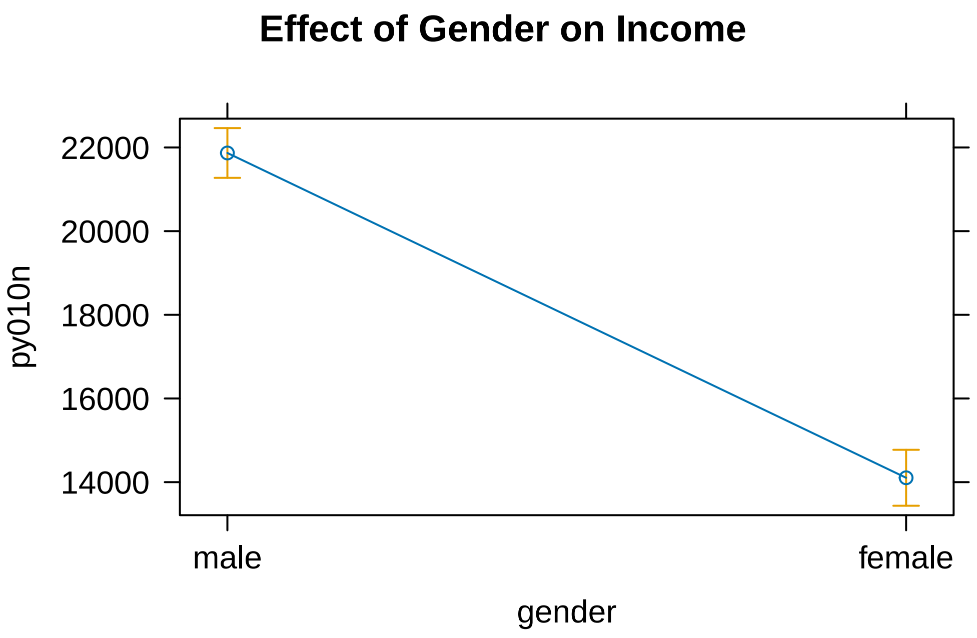
```
# Plot main effects separately for clarity
plot(Effect("age", int_model, xlevels=50),
     main="Effect of Age on Income",
     lines=list(col=tertiary_color, lwd=2))
```



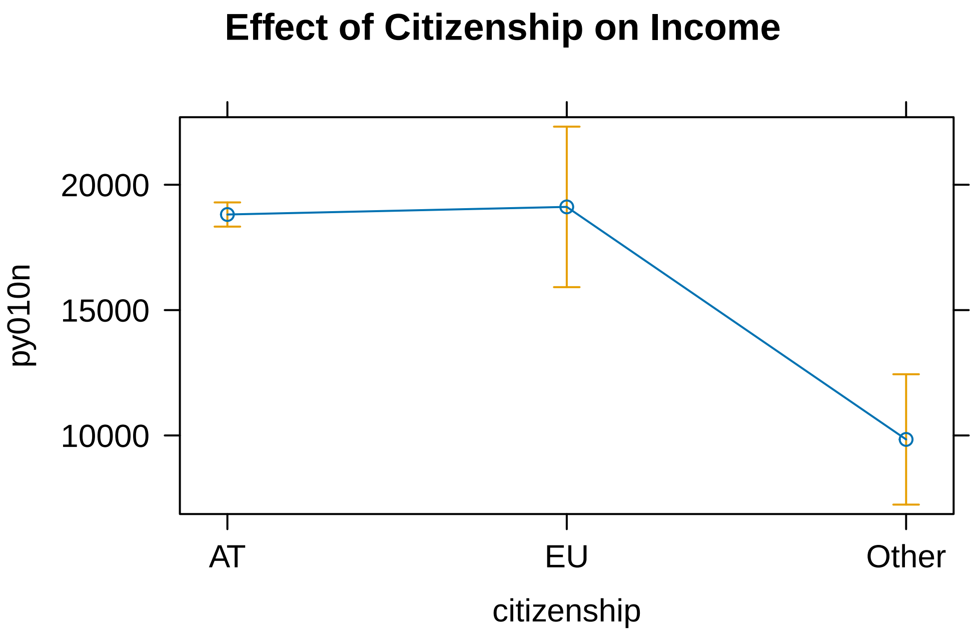
```
plot(Effect("hsize", int_model, xlevels=50),
     main="Effect of Household Size on Income",
     lines=list(col=accent_color, lwd=2))
```



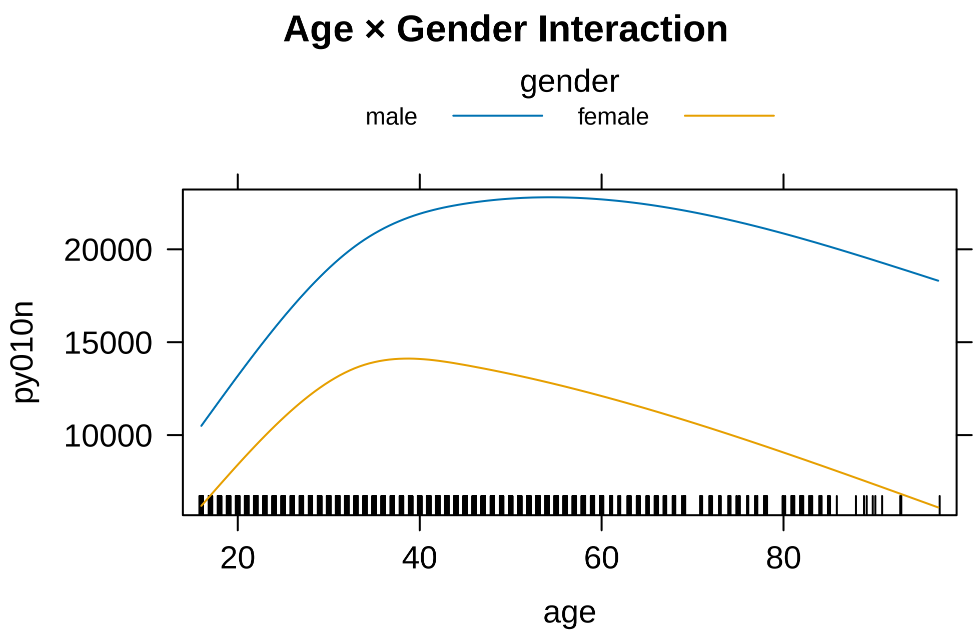
```
plot(Effect("gender", int_model),
     main="Effect of Gender on Income")
```



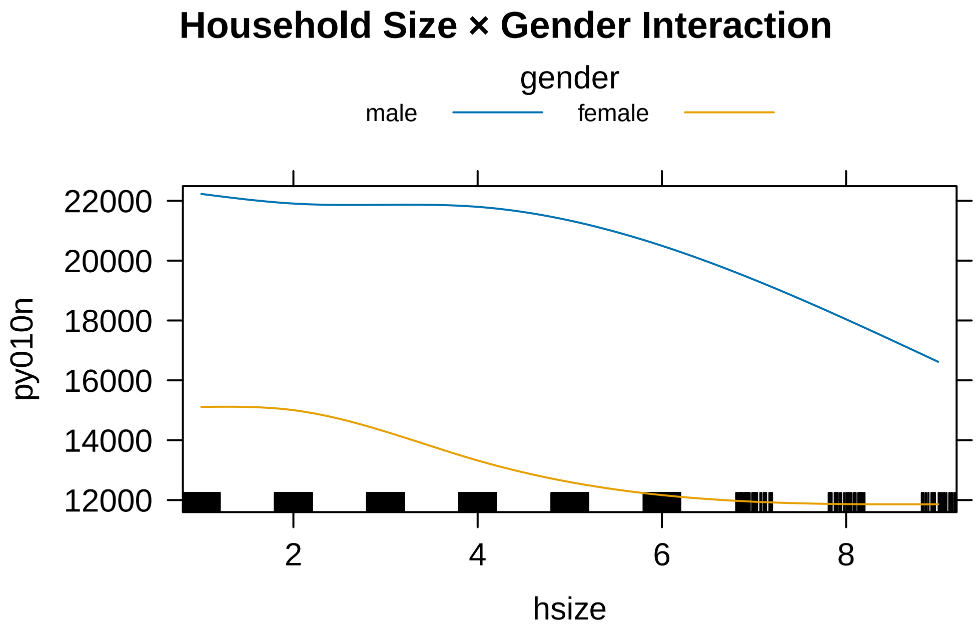
```
plot(Effect("citizenship", int_model),
     main="Effect of Citizenship on Income")
```



```
# Plot key interactions
plot(Effect(c("age", "gender"), int_model, xlevels=50),
     main="Age × Gender Interaction",
     lines=list(multiline=TRUE))
```



```
plot(Effect(c("hsize", "gender"), int_model, xlevels=50),
     main="Household Size × Gender Interaction",
     lines=list(multiline=TRUE))
```



Interpretation: The interaction model ($R^2 = 0.215$) provides the best fit, adding 1.5 percentage points over the spline model by allowing predictor effects to vary across groups. The model includes all two-way interactions between gender, citizenship, age splines, and household size splines.

Main Effects: - **Gender:** The baseline effect is €3,695 (p < 0.01), but this is modified by interactions. - **Citizenship:** Base effects are non-significant as interactions dominate. - **Age & Household Size:** Spline terms remain significant (p < 0.001 for age), with effects varying by gender and citizenship.

Significant Interactions (from ANOVA): - **Gender × Age (p < 0.001):** The age-income profile differs substantially by gender. The effect plots show males have steeper income growth with age, with the gender gap widening from youth to mid-career. - **Gender × Household Size (p = 0.075, marginally significant):** The negative household size effect is stronger for females than males, suggesting women bear greater career penalties for larger families. - **Citizenship × Age (p = 0.031):** Age-income trajectories differ by citizenship, with "Other" citizenship showing flatter profiles. - **Citizenship × Household Size (p = 0.016):** Household size effects vary by citizenship group.

Effect Plot Interpretations:

Age × Gender: Males show a pronounced inverted U-curve with peak earnings around age 50 (~€20,000). Females have lower overall earnings and a flatter profile, peaking earlier around age 45 (~€15,000). The gap widens from about €3,000 at age 25 to €6,000-€7,000 at age 50-60, demonstrating cumulative career disadvantages for women.

Household Size × Gender: Males maintain relatively stable income (€17,000-€18,000) regardless of household size, with slight decrease for very large households. Females show steeper decline from €15,000 (small households) to €12,000-€13,000 (large households), a ~€3,000 penalty. This interaction confirms that household responsibilities disproportionately impact women's earnings.

Model Comparison: While interactions improve fit modestly, they reveal important heterogeneity: gender pay gaps grow with age and household size, and citizenship status modifies career progression patterns. These findings have important policy implications for understanding and addressing income inequalities in South Austria.

5 5. Summary

This analysis examined determinants of employment income in South Austria using data from Carinthia and Styria regions. Through progressive model building—from simple linear regression to complex interaction models with splines—we identified key patterns and relationships:

Key Findings:

1. **Income Distribution:** Income is right-skewed (median: €16,226; mean: €16,952) with substantial high-income outliers, reflecting typical income inequality patterns.

2. **Gender Disparities:** A persistent and substantial gender pay gap of approximately €7,500-€7,700 (40-45%) exists across all models. This gap widens significantly with age, growing from ~€3,000 at age 25 to ~€6,000-€7,000 at age 50-60, indicating cumulative career disadvantages for women.

3. **Citizenship Effects:** Individuals with non-EU citizenship earn €4,800-€5,200 less than Austrian citizens, while EU citizens show no significant difference. Non-EU citizens also display flatter age-income profiles, suggesting barriers to career advancement.

4. **Age Patterns:** Income follows an inverted U-shaped trajectory, rising from early career to peak around ages 45-50, then declining toward retirement. This life-cycle pattern is better captured by polynomial and spline specifications than linear models.
5. **Household Size:** Larger household size is associated with lower individual income (approximately -€400 per additional member in linear models). This effect is particularly pronounced for females, suggesting disproportionate career impacts of family responsibilities on women.
6. **Important Interactions:** The interaction model reveals that effects are not uniform:
   - The gender pay gap increases with both age and household size
   - Age-income trajectories differ significantly by citizenship status
   - Women face steeper income penalties for larger households than men

Model Performance: Model fit improved progressively from linear ($R^2 = 0.175$) to polynomial ($R^2 = 0.199$) to spline ($R^2 = 0.200$) to interaction specifications ($R^2 = 0.215$). While these improvements are modest in percentage points, they capture substantively important nonlinear relationships and heterogeneous effects across groups.

Limitations: The models explain only 17-22% of income variance, indicating substantial unexplained variation. Important factors not included in this analysis (education, occupation, work experience, industry, part-time vs. full-time status) likely account for much of the remaining variance. Additionally, the analysis is cross-sectional and cannot establish causal relationships—observed associations may reflect both causal effects and selection processes.

Policy Implications: The findings highlight persistent income inequalities by gender and citizenship status in South Austria. The widening gender gap with age and household size suggests that policies addressing work-family balance, parental leave, childcare availability, and career reentry programs may be particularly important. The citizenship-based disparities point to potential barriers in labor market integration for non-EU immigrants, warranting attention to credential recognition, language support, and anti-discrimination measures.

Future Directions: Residual diagnostics (not shown here but recommended) should be conducted to validate model assumptions. Future analyses could incorporate additional predictors (education, occupation), explore regional differences between Carinthia and Styria, conduct separate models by gender to better understand within-group variations, and employ methods to address potential selection bias in labor force participation.