

# Regression Project – Group 3 (Intermediate Report)

Reema abolahum and Muzammil Mohammed

2025-10-12

## Table of contents

<b>1</b>	<b>1. Introduction</b>	<b>2</b>
<b>2</b>	<b>2. Data Collection and Preparation</b>	<b>2</b>
<b>3</b>	<b>3. Descriptive Analysis</b>	<b>3</b>
3.1	3.1 Numeric Summaries . . . . .	3
3.2	3.2 Frequency Tables . . . . .	3
3.3	3.3 Univariate Plots . . . . .	4
3.4	3.4 Bivariate Plots . . . . .	11
3.5	3.5 Interaction Plots . . . . .	15
3.6	3.6 Correlation / Scatterplot Matrix . . . . .	20
<b>4</b>	<b>4. Regression Modeling</b>	<b>22</b>
4.1	4.1 Linear Model . . . . .	22
4.2	4.2 Polynomial Model . . . . .	24
4.3	4.3 Spline Model . . . . .	26
4.4	4.4 Interaction Model . . . . .	28
<b>5</b>	<b>5. Summary</b>	<b>35</b>

```
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)
```

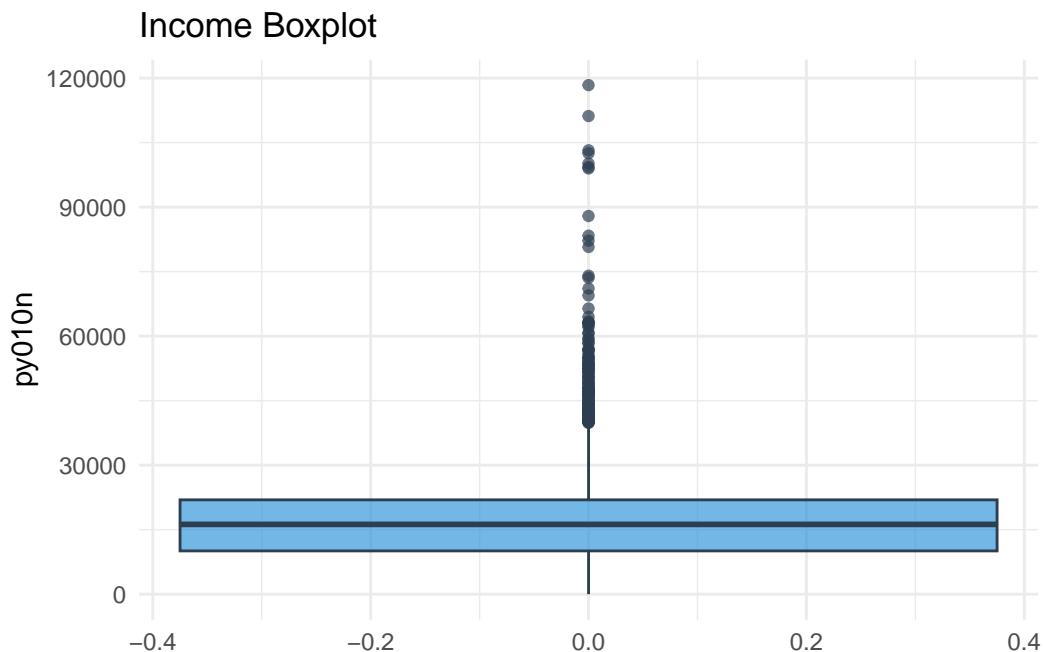
	AT	EU	Other
	5021	79	167

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

	AT	EU	Other
male	2867	40	97
female	2154	39	70

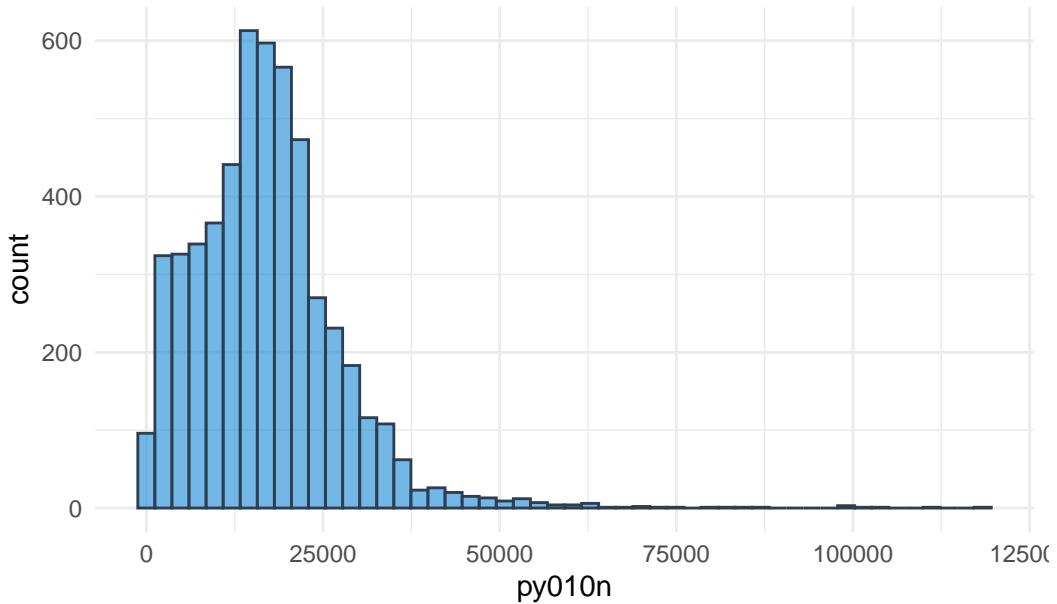
### 3.3 3.3 Univariate Plots

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

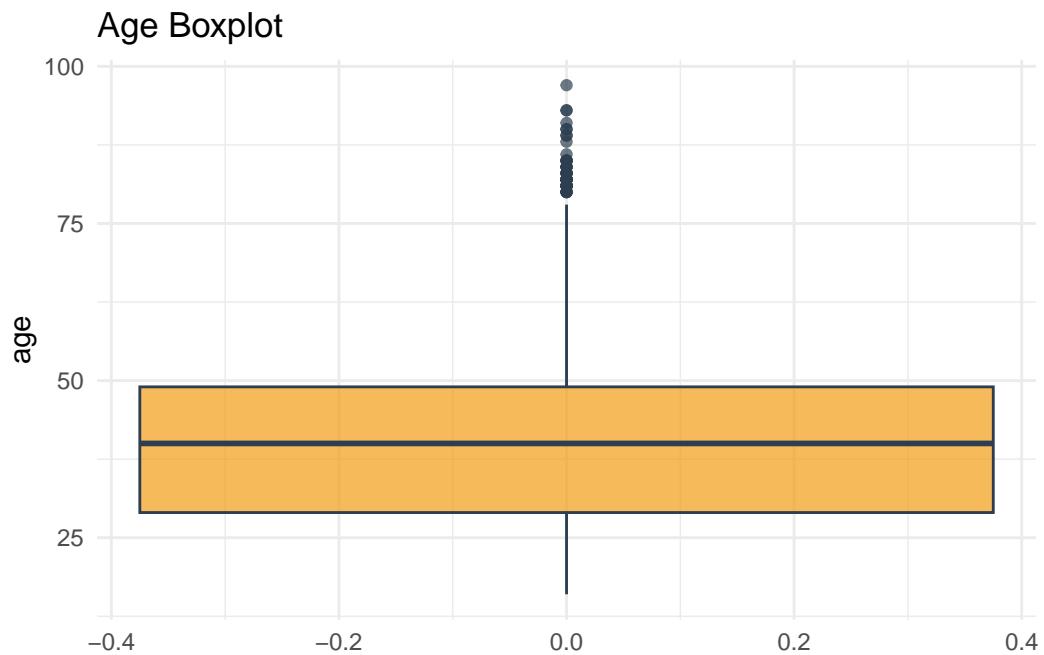


```
ggplot(dat, aes(x=py010n)) +
  geom_histogram(bins=50, fill=secondary_color, color=main_color, alpha=0.7) +
  labs(title="Income Histogram") +
  theme_minimal()
```

### Income Histogram

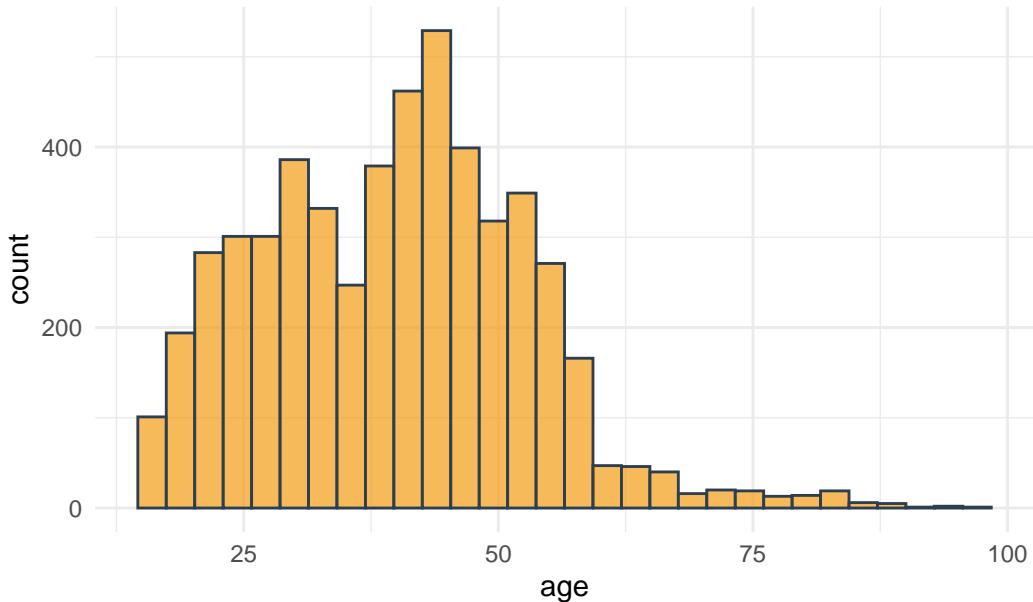


```
# Age
ggplot(dat, aes(y=age)) +
  geom_boxplot(fill=tertiary_color, alpha=0.7, color=main_color) +
  labs(title="Age Boxplot") +
  theme_minimal()
```



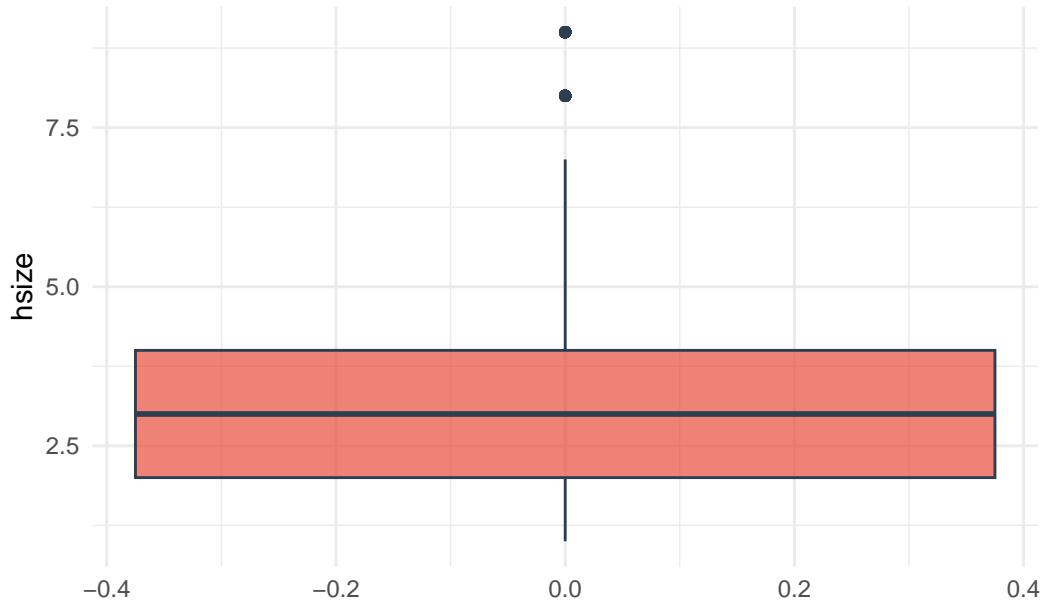
```
ggplot(dat, aes(x=age)) +  
  geom_histogram(bins=30, fill=tertiary_color, color=main_color, alpha=0.7) +  
  labs(title="Age Histogram") +  
  theme_minimal()
```

## Age Histogram

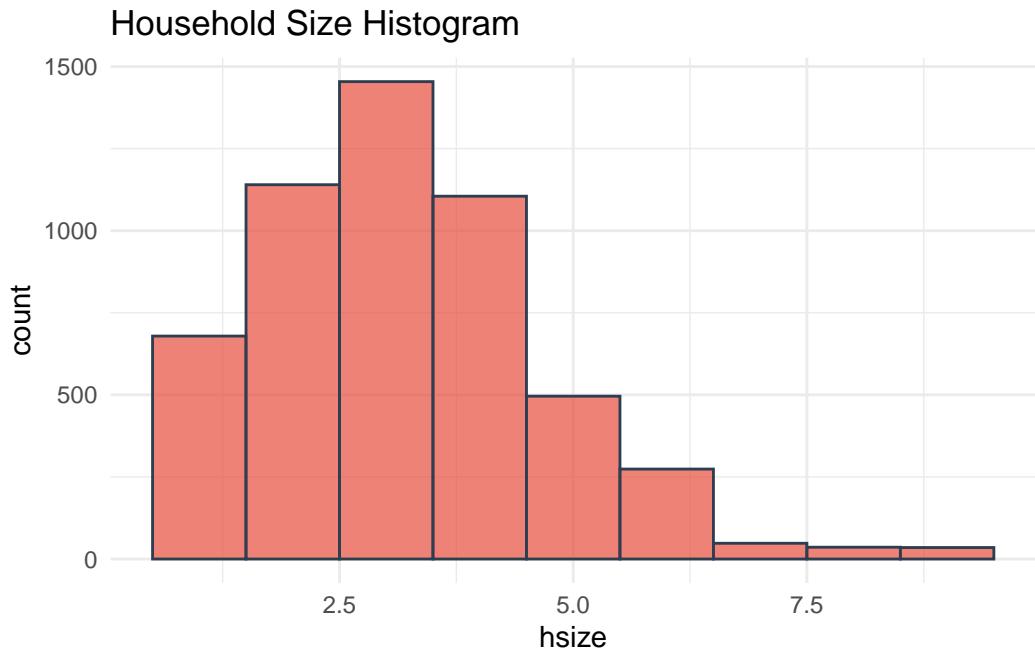


```
# Household Size
ggplot(dat, aes(y=hsize)) +
  geom_boxplot(fill=accent_color, alpha=0.7, color=main_color) +
  labs(title="Household Size Boxplot") +
  theme_minimal()
```

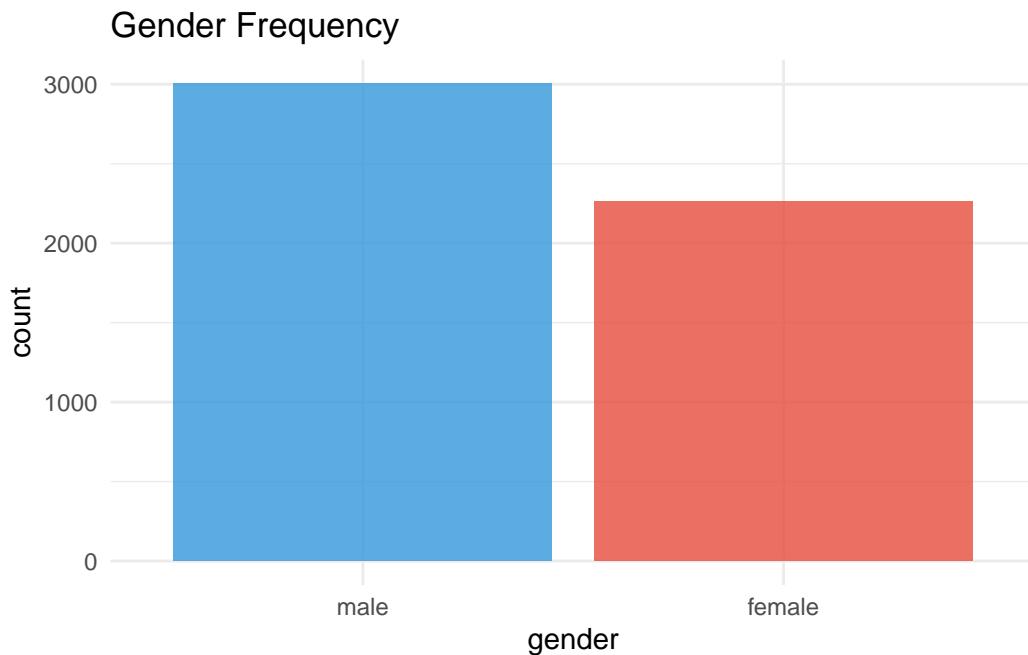
## Household Size Boxplot



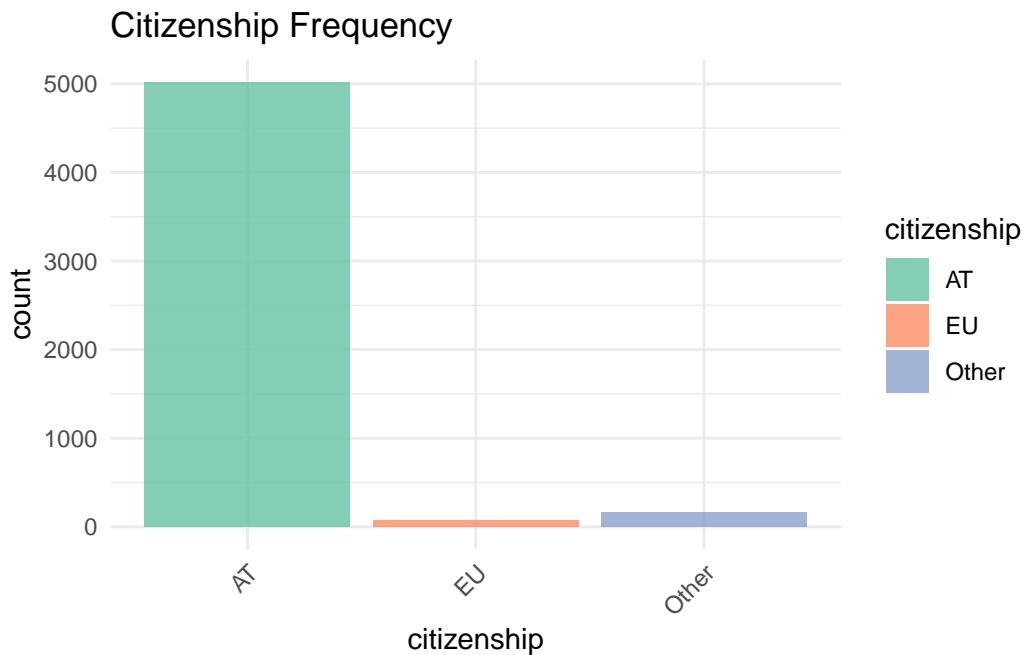
```
ggplot(dat, aes(x=hszie)) +  
  geom_histogram(binwidth=1, fill=accent_color, color=main_color, alpha=0.7) +  
  labs(title="Household Size Histogram") +  
  theme_minimal()
```



```
# 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")
```

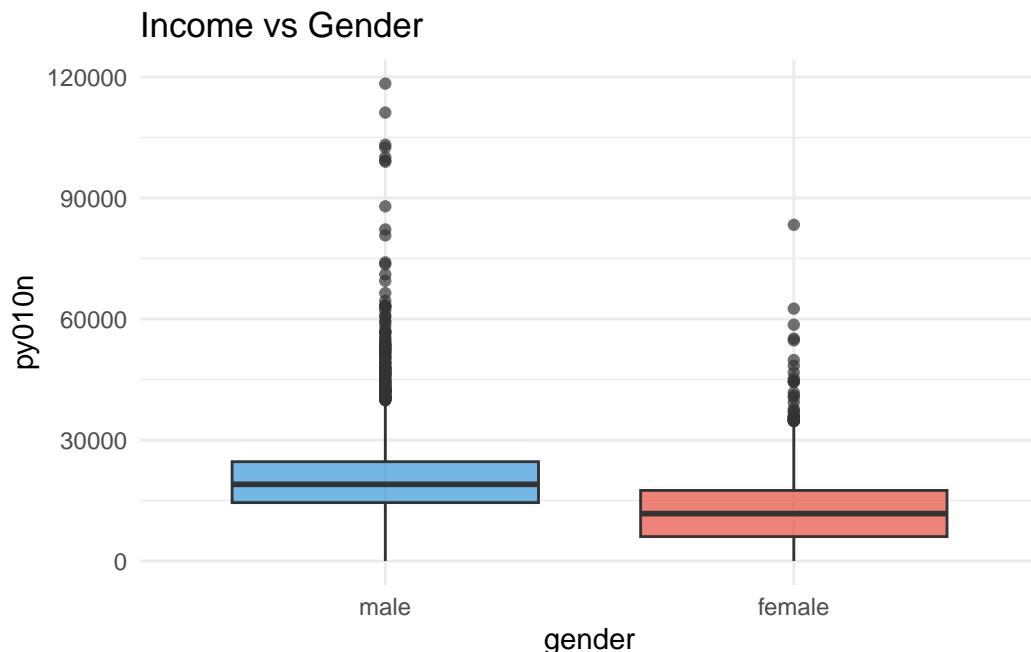


```
# Citizenship
ggplot(dat, aes(x=citizenship, fill=citizenship)) +
  geom_bar(alpha=0.8) +
  scale_fill_brewer(palette="Set2") +
  theme_minimal() +
  theme(axis.text.x=element_text(angle=45, hjust=1)) +
  labs(title="Citizenship Frequency")
```

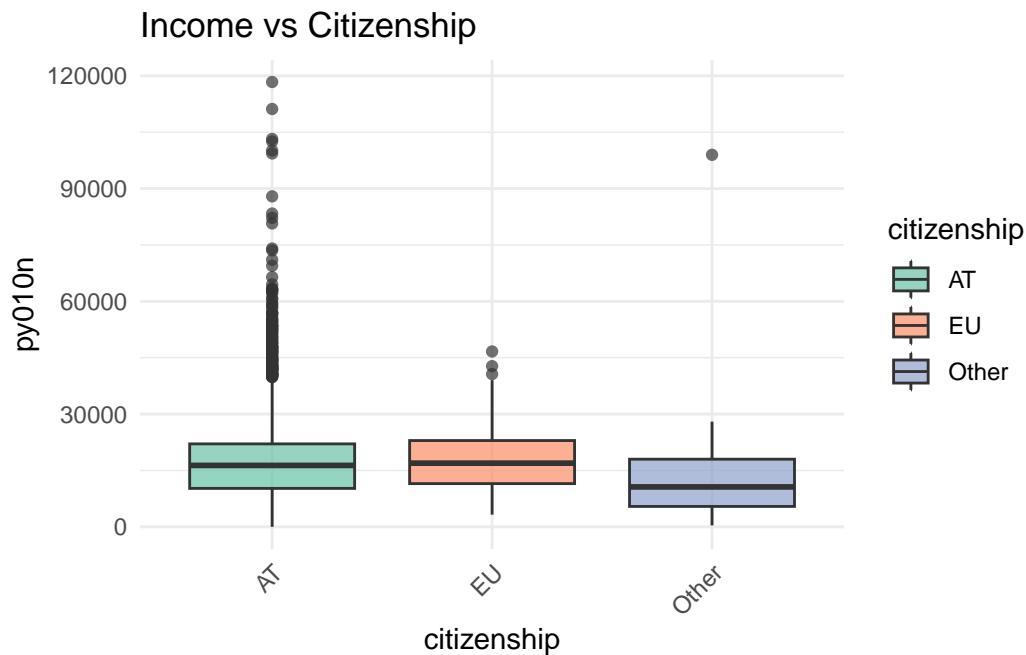


### 3.4 3.4 Bivariate Plots

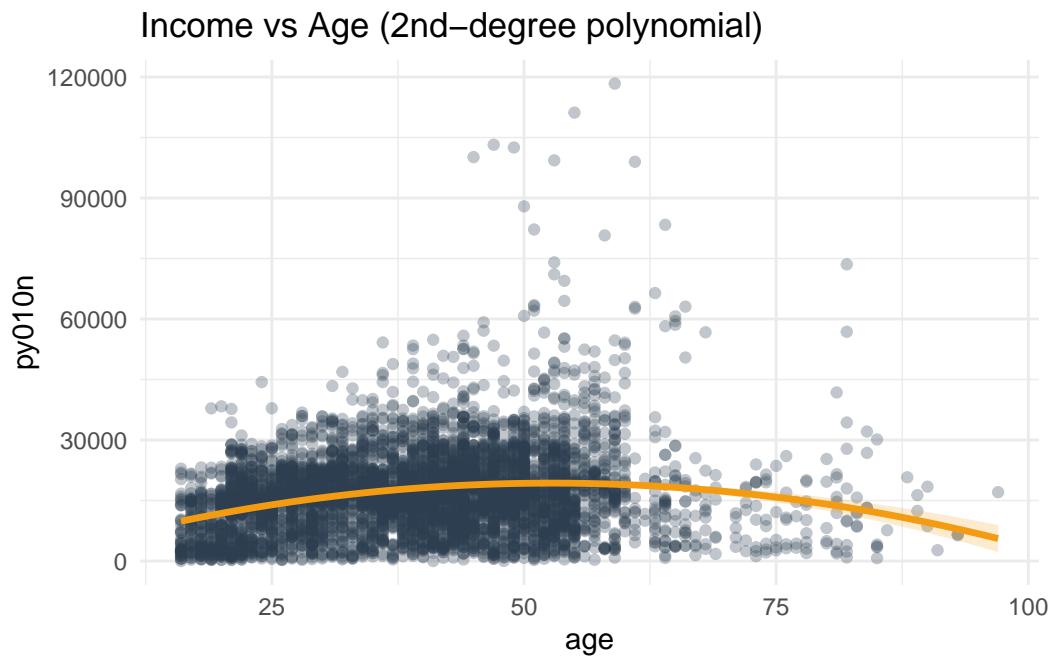
```
# Gender vs Income
ggplot(dat, aes(x=gender, y=py010n, fill=gender)) +
  geom_boxplot(alpha=0.7) +
  scale_fill_manual(values=c("female"="#E74C3C", "male"#3498DB")) +
  labs(title="Income vs Gender") +
  theme_minimal() +
  theme(legend.position="none")
```



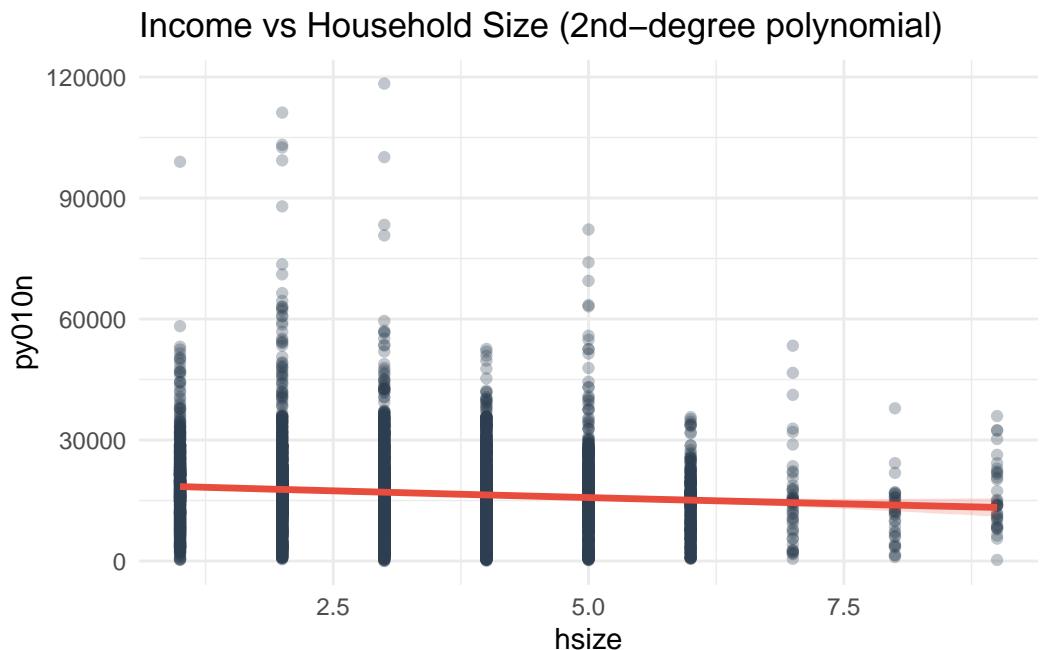
```
# Citizenship vs Income
ggplot(dat, aes(x=citizenship, y=py010n, fill=citizenship)) +
  geom_boxplot(alpha=0.7) +
  scale_fill_brewer(palette="Set2") +
  theme_minimal() +
  theme(axis.text.x=element_text(angle=45, hjust=1)) +
  labs(title="Income vs Citizenship")
```



```
# 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, a
```



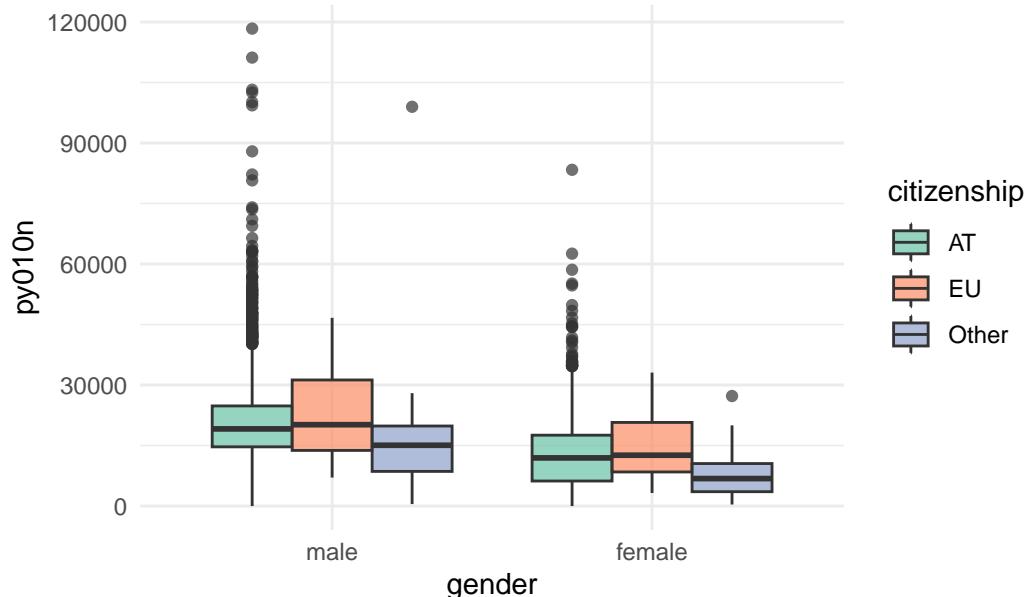
```
# Household size vs Income (Polynomial)
ggplot(dat, aes(x=hsizen, y=py010n)) +
  geom_point(alpha=0.3, color=main_color, size=1.5) +
  geom_smooth(method="lm", formula=y~poly(x,2), color=accent_color, fill=accent_color, alpha=0.5) +
  labs(title="Income vs Household Size (2nd-degree polynomial)") +
  theme_minimal()
```



### 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()
```

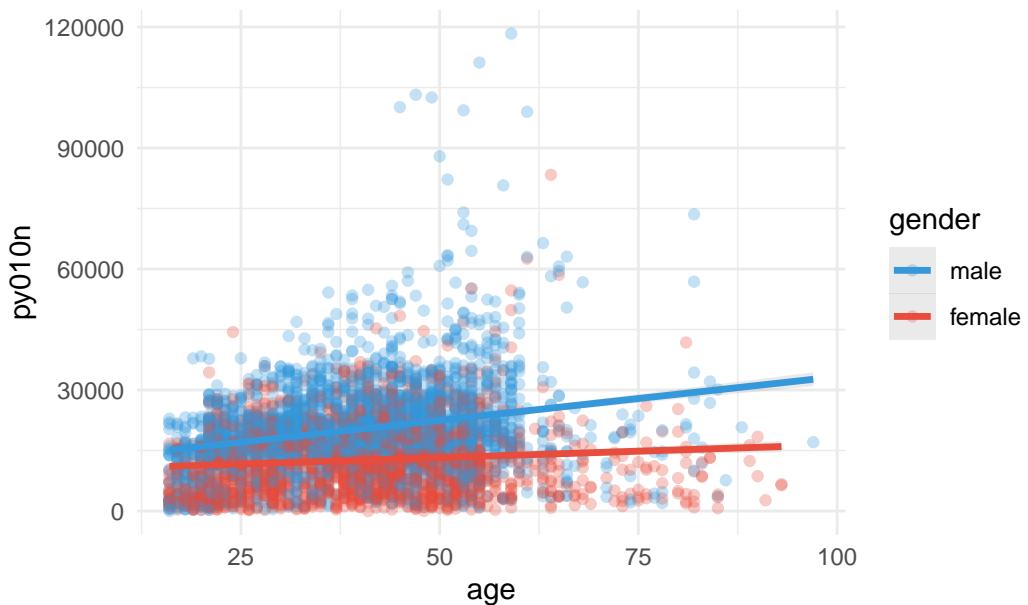
## Income Interaction: Gender × Citizenship



```
# 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'
```

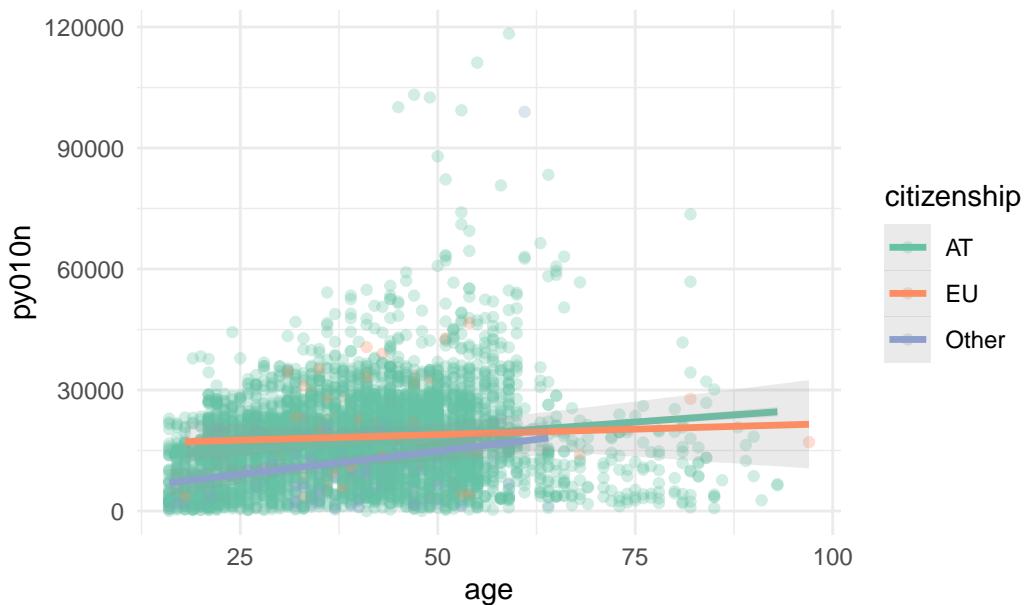
### Income Interaction: Age × Gender



```
# 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'
```

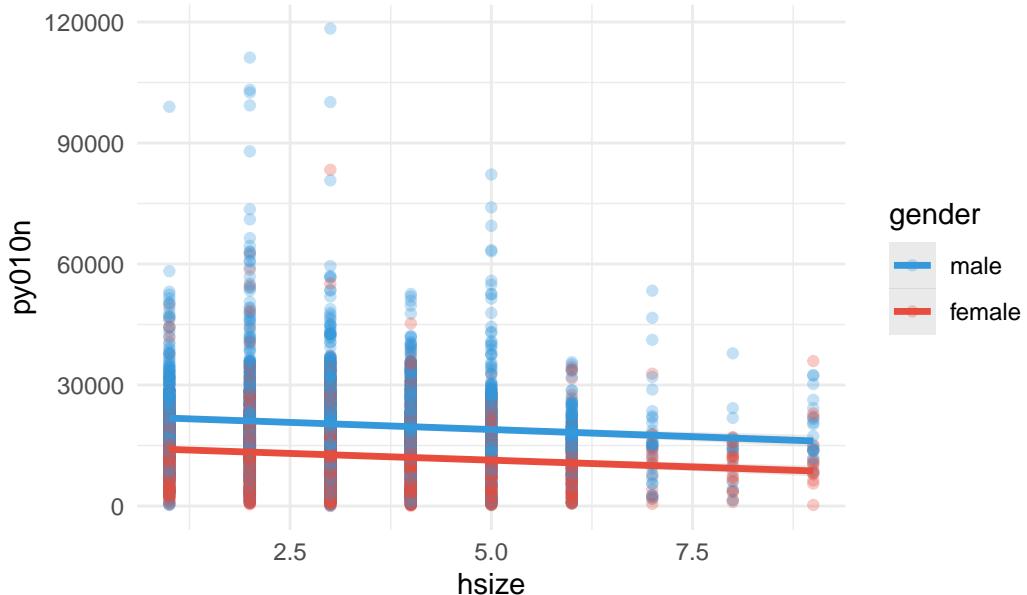
## Income Interaction: Age × Citizenship



```
# 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'

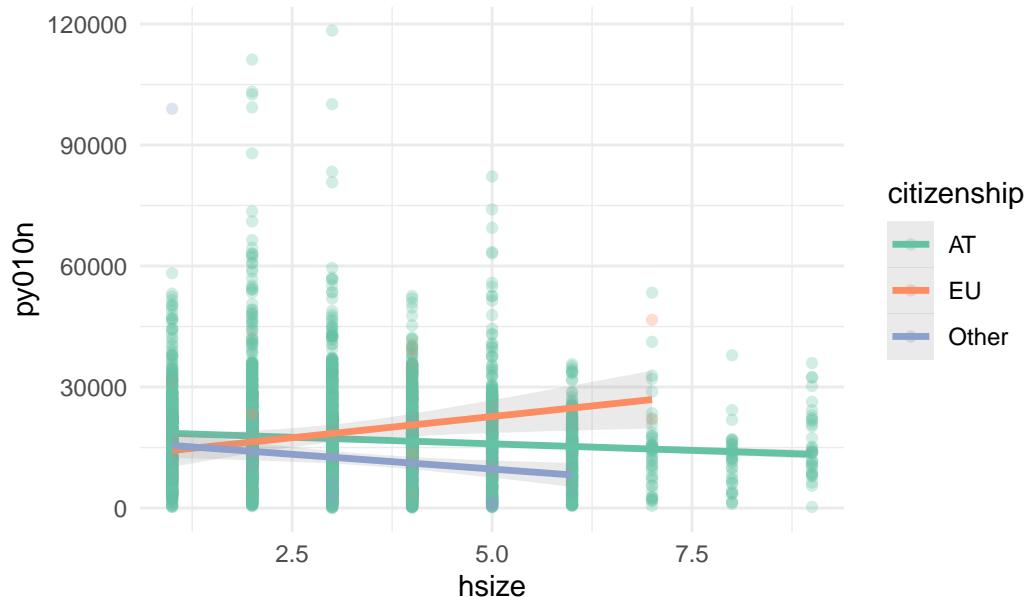
### Income Interaction: Household Size x Gender



```
# Household Size x Citizenship
ggplot(dat, aes(x=hsiz, 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 x Citizenship") +
  theme_minimal()
```

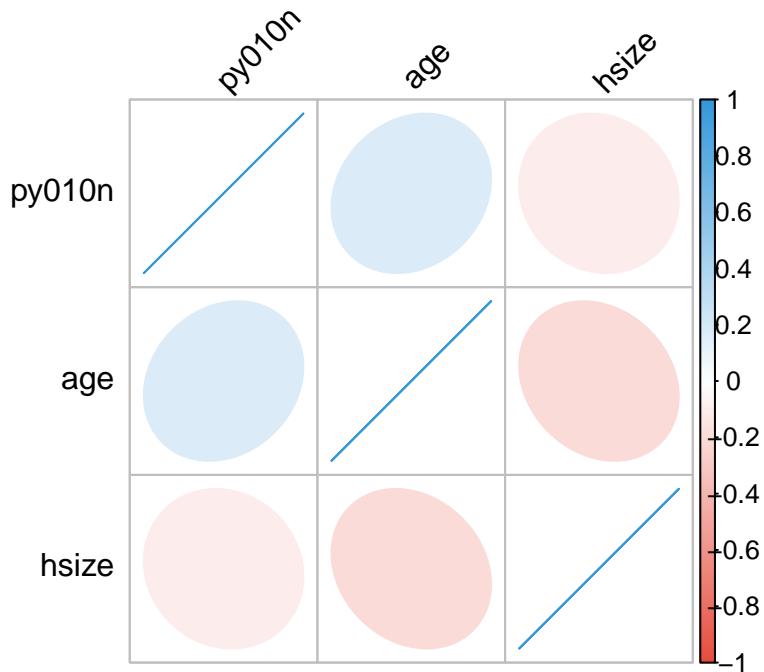
```
`geom_smooth()` using formula = 'y ~ x'
```

## Income Interaction: Household Size x Citizenship

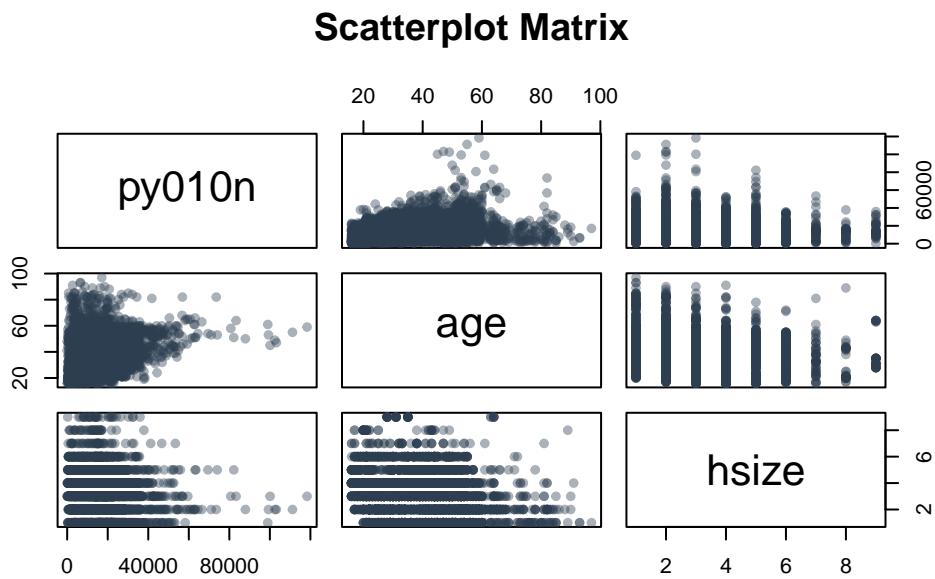


### 3.6 Correlation / Scatterplot Matrix

```
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)
```



## 4 4. Regression Modeling

### 4.1 4.1 Linear Model

```
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
```

```
Anova(lm_model)
```

Anova Table (Type II tests)

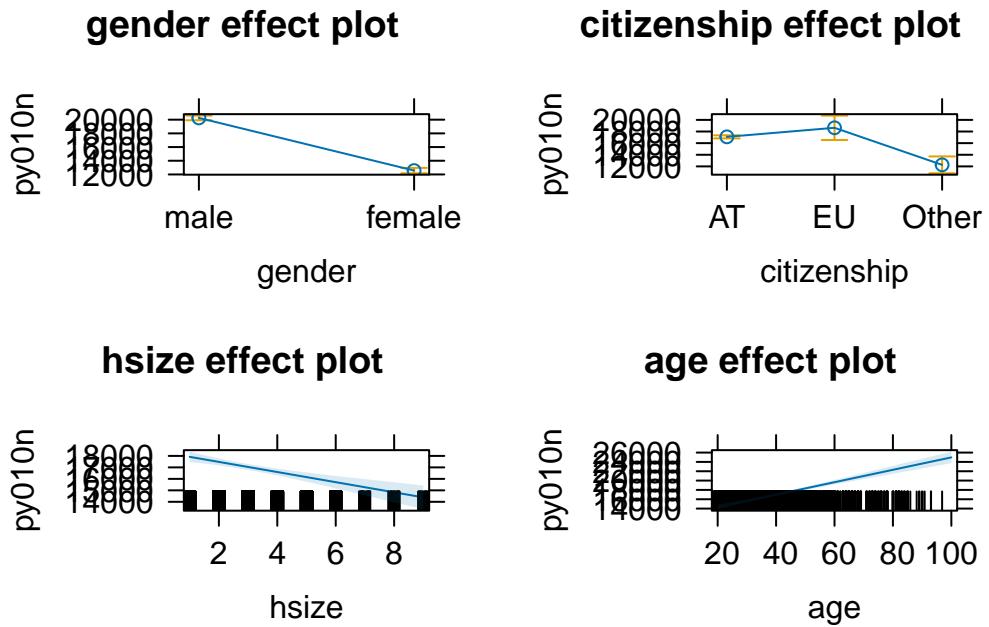
Response: py010n

	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

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



## 4.2 4.2 Polynomial Model

```
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
```

```
Anova(poly_model)
```

Anova Table (Type II tests)

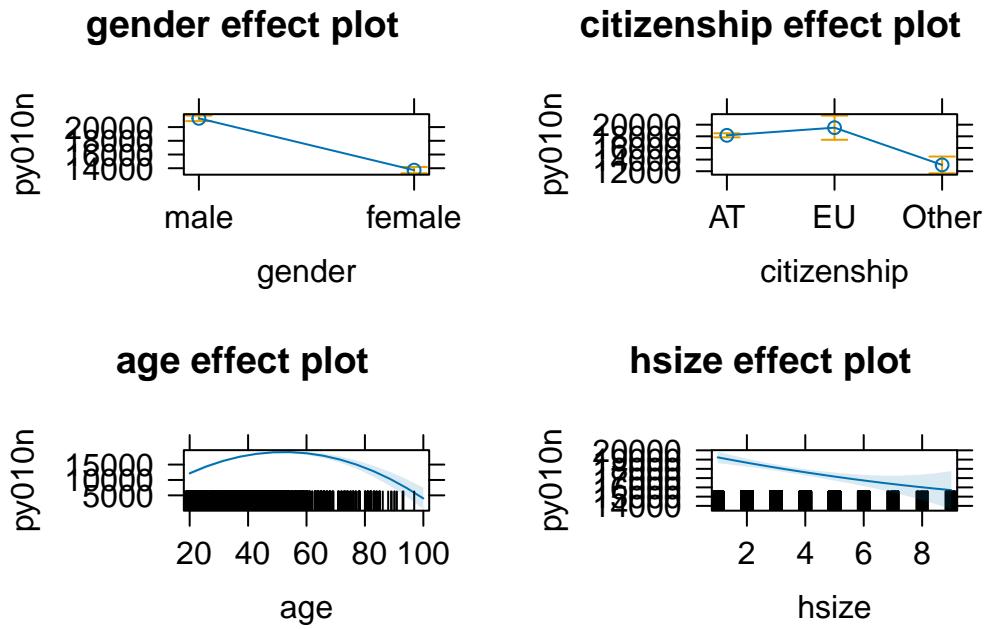
Response: py010n

	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))
```



### 4.3 Spline Model

```
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

```

### Anova(spline\_model)

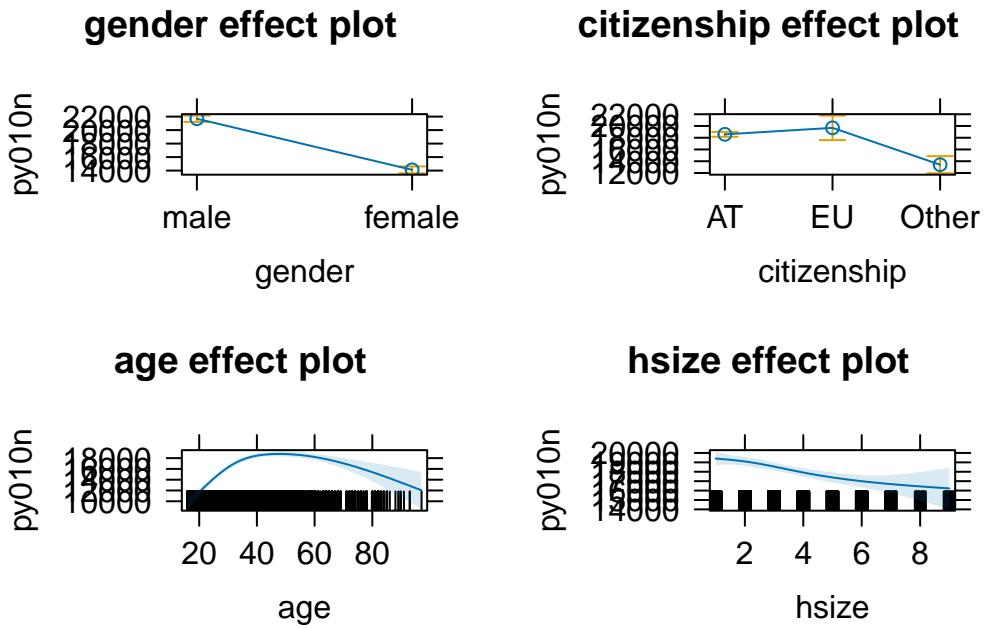
Anova Table (Type II tests)

```

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

```

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



#### 4.4 Interaction Model

```
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 ' '

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

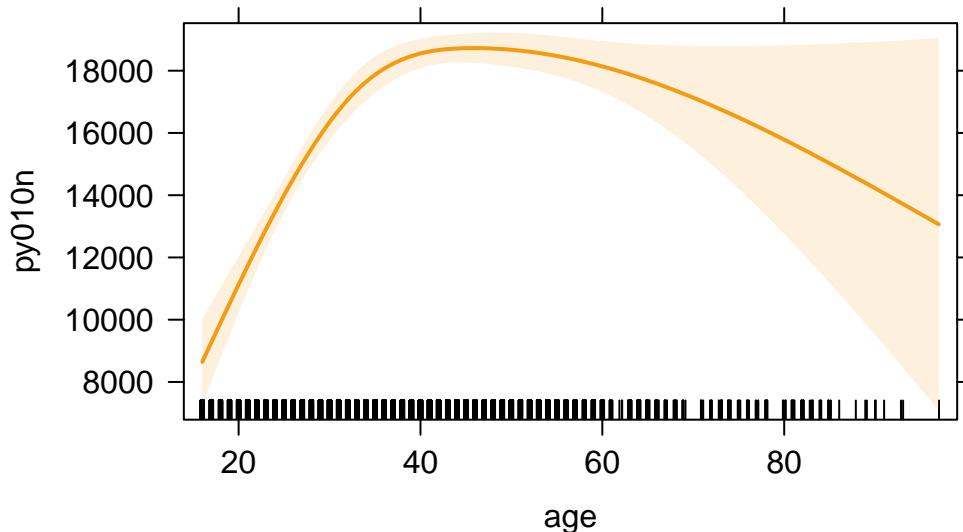
	Sum Sq	Df	F value	Pr(>F)
gender	7.2912e+10	1	840.1387	< 2.2e-16 ***
citizenship	4.3629e+09	2	25.1364	1.366e-11 ***
ns(age, 3)	2.9791e+10	3	114.4234	< 2.2e-16 ***
ns(hsize, 3)	2.7869e+09	3	10.7044	5.188e-07 ***
gender:citizenship	5.1964e+07	2	0.2994	0.74129
gender:ns(age, 3)	4.2695e+09	3	16.3985	1.324e-10 ***
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

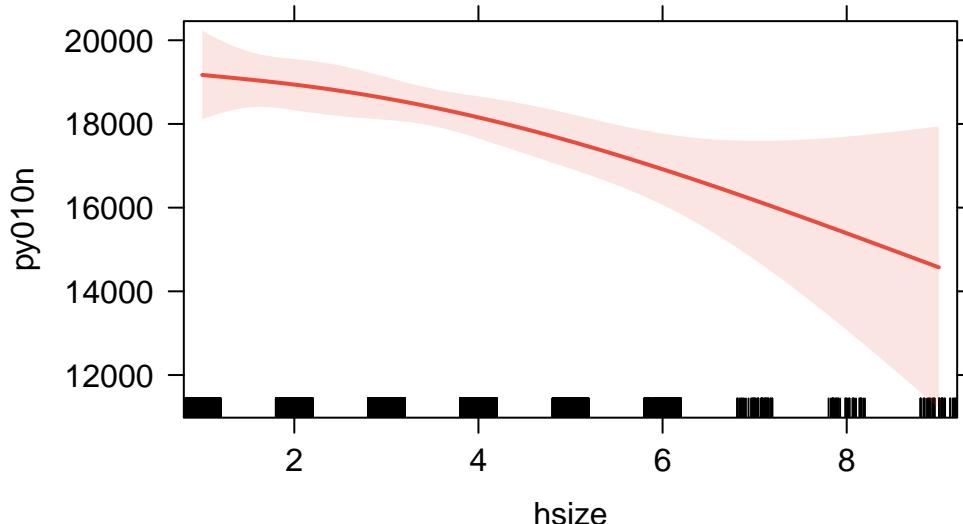
```
# Plot main effects separately for clarity
plot(Effect("age", int_model, xlevels=50),
      main="Effect of Age on Income",
      lines=list(col=tertiary_color, lwd=2))
```

## Effect of Age on Income

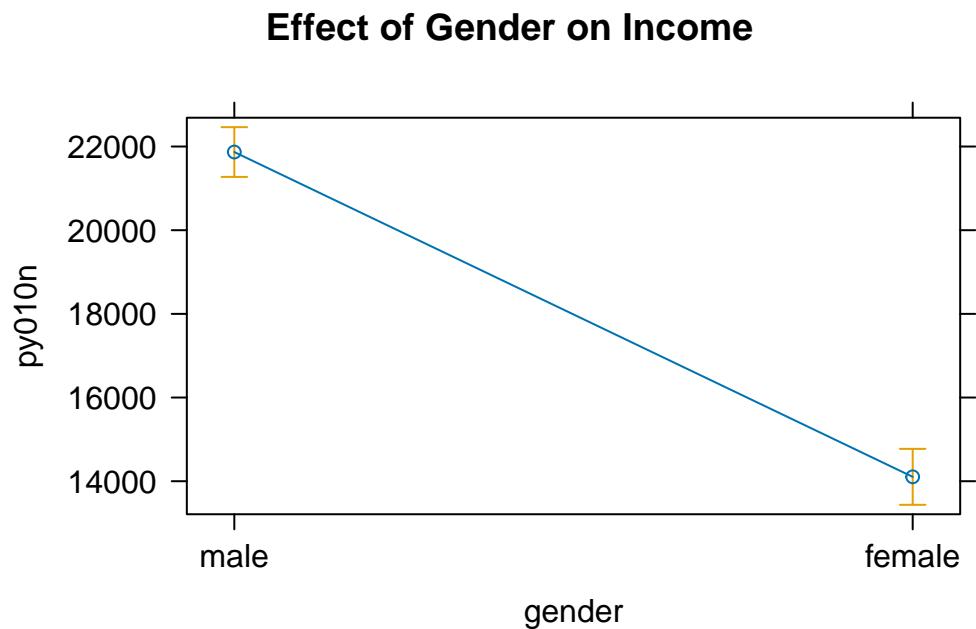


```
plot(Effect("hsize", int_model, xlevels=50),
     main="Effect of Household Size on Income",
     lines=list(col=accent_color, lwd=2))
```

## Effect of Household Size on Income

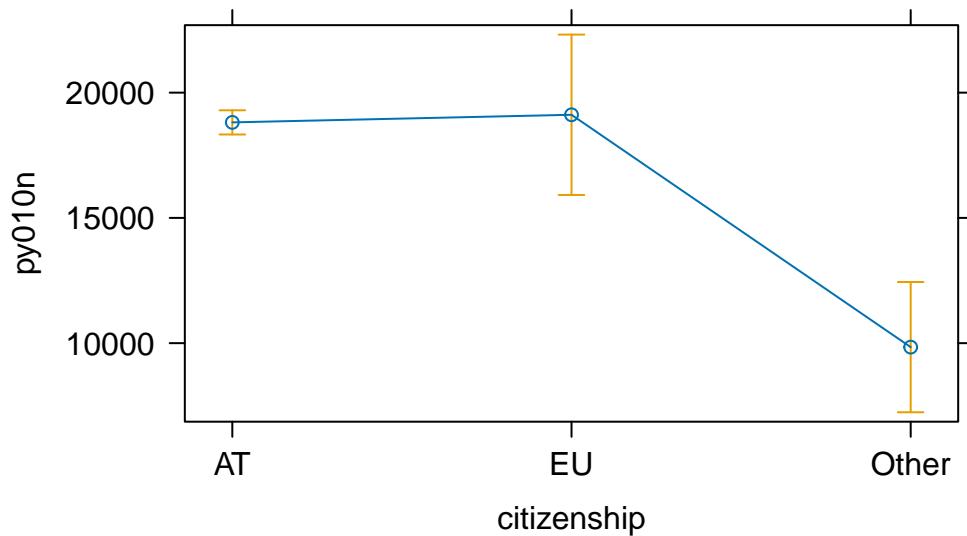


```
plot(Effect("gender", int_model),  
     main="Effect of Gender on Income")
```



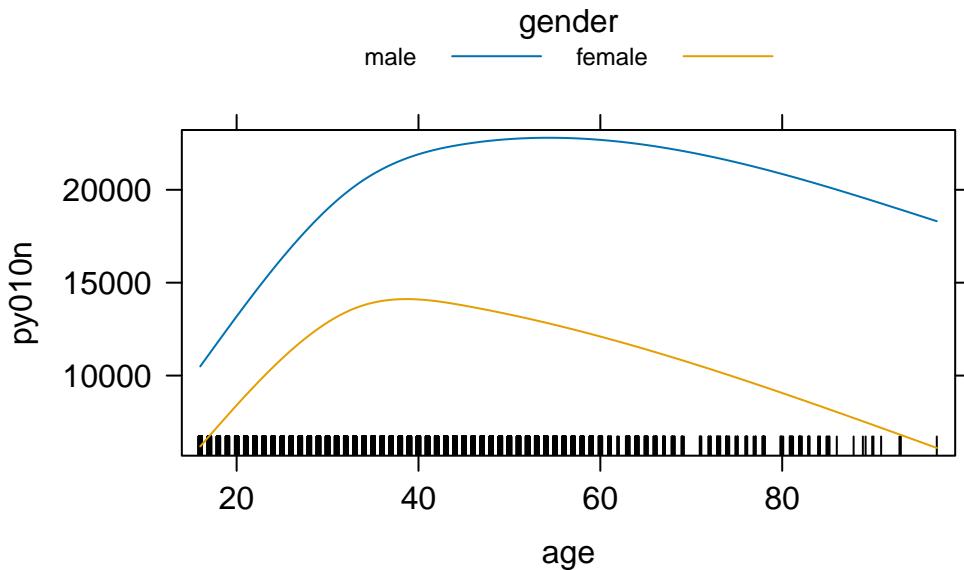
```
plot(Effect("citizenship", int_model),  
     main="Effect of Citizenship on Income")
```

## Effect of Citizenship on Income



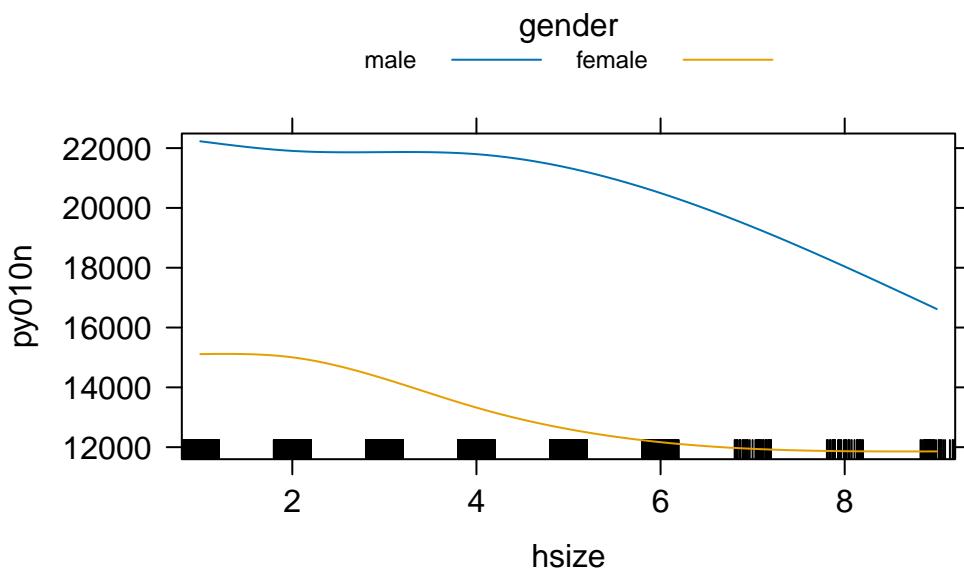
```
# Plot key interactions
plot(Effect(c("age", "gender"), int_model, xlevels=50),
      main="Age × Gender Interaction",
      lines=list(multiline=TRUE))
```

## Age × Gender Interaction



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

## Household Size × Gender Interaction



## **5 5. Summary**

This analysis shows skewed income with outliers, gender and citizenship differences, nonlinear relationships with age and household size, and relevant interactions. Polynomial and spline regressions provide flexible modeling of numeric predictors, and residual diagnostics suggest the models are adequate.