Regression Project: EU-SILC analyses

Dataset: eusilcP Intention

Johannes Gölles & Kevin Andoni

2025-01-26

Table of contents

1 Introduction

1.1 Libraries

```
library <- function(...) {suppressPackageStartupMessages(base::library(...))}
library(tidyverse)  # For data manipulation and visualization
library(knitr)      # For creating tables in reports
library(rmarkdown)  # For rendering reports
library(ggplot2)    # For creating visualizations
library(kableExtra) # For formatting the data dictionary
library(MASS)       # For stepwise model selection
library(simFrame)   # For loading the eusilcP dataset
library(lme4)       # For mixed-effects models
library(car)        # For diagnostic tools and visualizations
library(effects)    # For visualizing the effects of predictors
```

1.2 Data Collection

The data for this analysis comes from the **European Union Statistics on Income and Living Conditions (EU-SILC)** survey. The EU-SILC is a large-scale cross-sectional and longitudinal survey conducted annually across European Union member states. It provides harmonized data on income, living conditions, poverty, and social exclusion. For this analysis, data from the East Austria regions—Vienna, Lower Austria, and Burgenland—was extracted.

Below is a data dictionary summarizing the key variables used in this analysis:

Table 1: Data Dictionary for EU-SILC Dataset

Variable	Description	Type
old_age_benefits	Net old-age benefits received by individuals, including pensions, care allowances, and other financial supports.	Numeric
main_income_earner	Indicates whether the individual is the main income earner of their household (TRUE/FALSE).	Logical
economic_status	Primary economic activity of the individual (e.g., full-time work, retired, unemployed).	Categorical

household_size	Number of individuals living in the respondent's household.	Numeric
age_years	Age of the respondent in years.	Numeric
region_name	Region where the respondent resides (Vienna, Lower Austria, or Burgenland).	Categorical

1.2.1 Type of Survey and Execution

The EU-SILC is conducted as a household survey, combining personal interviews and administrative data. It gathers socioeconomic information at both the household and individual levels. The survey period typically spans a calendar year, ensuring comprehensive coverage of annual income and employment data.

```
# Load the eusilcP dataset
data("eusilcP", package = "simFrame")
# ?eusilcP
# Define NUTS-2 regions corresponding to "East Austria" (AT1)
east_austria_regions <- c("Burgenland", "Lower Austria", "Vienna")

# Subset the data for East Austria and select relevant variables
subset_data <- subset(
  eusilcP,
  region %in% east_austria_regions,
  select = c("py100n", "main", "ecoStat", "hsize", "age", "region")
)

# Adjust factor levels of 'region' to only include selected regions
subset_data$region <- droplevels(subset_data$region)

# View a summary of the subset data
print(str(subset_data))
```

```
'data.frame':   24725 obs. of  6 variables:
 $ py100n : num  17768 0 0 NA NA ...
 $ main   : logi  TRUE FALSE TRUE FALSE FALSE FALSE ...
 $ ecoStat: Factor w/ 7 levels "1","2","3","4",..: 5 2 1 NA NA NA 1 7 5 2 ...
 $ hsize  : Factor w/ 9 levels "1","2","3","4",..: 1 5 5 5 5 5 3 3 3 4 ...
 $ age    : num  77 34 35 15 12 3 45 69 71 50 ...
 $ region : Ord.factor w/ 3 levels "Burgenland"<"Lower Austria"<..: 3 2 2 2 2 2 2 2 2 1 ...
NULL
```

Data Cleaning

To prepare the dataset for analysis, we renamed variables to make them more intuitive (e.g., py100n to old_age_benefits and ecoStat to economic_status) and removed rows with missing values, resulting in a loss of approximately 4,000 rows. This ensures the dataset is clean and reliable for analysis. Additionally, we recoded the numeric levels of economic_status into meaningful labels (e.g., 'FT_Work' for full-time work, 'Retired' for retired individuals) to improve interpretability, and converted household_size from categorical to numeric format to enable numerical comparisons in later analyses.

```
# Rename columns for clarity
cleaned_data <- subset_data %>%
  rename(
    old_age_benefits = py100n,
    main_income_earner = main,
    economic_status = ecoStat,
    household_size = hsize,
    age_years = age,
    region_name = region
  )

cleaned_data <- cleaned_data %>% na.omit() # We lose ~4000 rows

# Replace numeric levels of economic_status with abbreviations
cleaned_data$economic_status <- factor(cleaned_data$economic_status,
                                       levels = c(1, 2, 3, 4, 5, 6, 7),
                                       labels = c(
                                         "FT_Work",    # Working Full Time
                                         "PT_Work",    # Working Part Time
                                         "Unemployed", # Unemployed
                                         "Student",    # Pupil/Student/Training/Military Serv
                                         "Retired",    # Retired/Early Retirement/Given Up Bu
                                         "Disabled",   # Permanently Disabled/Unfit to Work/(
                                         "Domestic"    # Domestic Tasks/Care Responsibilities
                                       ))

#removing the ordering for region_name
cleaned_data$region_name <- factor(cleaned_data$region_name, ordered = FALSE)


cleaned_data$household_size <- as.numeric(as.character(cleaned_data$household_size))
```

1.2.2 Data Preparation

To prepare the data for analysis:

1. **Subsetting**: Only observations from the regions of Vienna, Lower Austria, and Burgenland were retained.
2. **Variable Renaming**: Variable names were simplified for clarity:
   - `py100n` → `old_age_benefits`
   - `main` → `main_income_earner`
   - `ecoStat` → `economic_status`
   - `hsize` → `household_size`
   - `age` → `age_years`
   - `region` → `region_name`
3. **Missing Value Treatment**: Rows with missing values were removed using the `na.omit()` function, resulting in a loss of approximately 4,000 rows.
4. **Factor Level Adjustments**: Categorical variables were recoded to ensure consistent levels:
   - Factor levels of `region` were adjusted to include only the specified East Austria regions.
   - Household_size which was a categorical variable was changed to numerical data
5. **Reclassification**: Certain inconsistencies were identified and addressed.

The cleaned dataset is now ready for descriptive and inferential analysis. Below is a summary of the data preparation process:

1.2.3 Summary of the Cleaned Dataset

```
# View summary of cleaned_data
summary(cleaned_data)
```

The cleaned dataset contains a detailed overview of individual-level and household-level variables for respondents from the East Austria regions. Below is a description of the key variables:

- **Economic Status (`economic_status`)**:
  Respondents are categorized into seven groups based on their primary economic activity:
  - "FT_Work" (Full-time Work): **9,116 individuals**.
  - "PT_Work" (Part-time Work): **1,718 individuals**.
  - "Unemployed": **1,314 individuals**.

  - "Student" (includes unpaid work experience or in compulsory military or community service): **1,295 individuals**.
  - "Retired": **5,427 individuals**.
  - "Disabled": **191 individuals**.
  - "Domestic" (engaged in household activities): **1,643 individuals**.

- **Old age benefits (`old_age_benefits`)**:
  Old age benefit values exhibit high skewness, with most respondents reporting **0 old age benefits**, as evident from the median and third quartile values.

  - Minimum old_age_benefits: **€0**.

  - Maximum old_age_benefits: **€101,777**.

  - Mean old_age_benefits: **€3,864**, affected by a few persons with high old age benefits.

- **Main income earner (`main_income_earner`)**:
  A binary indicator of employment shows that **11,275 individuals** are the main income earner of their household (`TRUE`), while **9,429 individuals** are not living in a household where another person is the main income earner (`FALSE`).

- **Household Size (`household_size`)**:
  The number of individuals in a household ranges from **1 to 6**, with most households consisting of:

  - **2 members (6,211 households)**.
  - **3 members (4,410 households)**.
  - **1 member (4,384 households)**. Which explains the median of 2 and the mean of 2.708 individuals per household. The smallest households were single person households and the most people in a household mentioned in this survey were 8 individuals.

- **Age (`age_years`)**:
  Respondent ages range from **16 to 96 years**, with:

  - Median age: **44 years**.

  - Mean age: **46.18 years**, indicating a slightly older sample.

  - The interquartile range is from **32 to 60 years**.

- **Region (`region_name`)**:
  The dataset includes observations from three regions in East Austria:

- **Burgenland**: **1,690 individuals**.
- **Lower Austria**: **9,313 individuals**.
- **Vienna**: **9,701 individuals**.

This summary highlights differences across key variables, providing a foundation for exploring their relationships in subsequent analyses.

1.2.4 Research Question

"This study aims to answer the following question: What factors influence the distribution of old-age benefits in East Austria, and how do these benefits vary across socioeconomic groups?"

Descriptive Analysis of the Sample

In this section, we will explore the key features of the data through various descriptive statistics and visualizations. These analyses will help us understand the distribution and relationships between variables, as well as identify any patterns or outliers.

Zero values in `old_age_benefits` represent individuals who do not receive any benefits. Including these zeros in regression analysis would distort the model by introducing unnecessary noise and biasing the results. By filtering out zeros, we focus on participants who receive benefits, enabling a more accurate understanding of the factors influencing benefit amounts.

```
# Filter out zeros in old_age_benefits for regression analysis
cleaned_data_no_zeros <- cleaned_data %>%
  filter(old_age_benefits > 0)

# Summary of data cleaning steps
cleaning_summary <- tibble::tibble(
  Step = c("Original Dataset", "After Removing Missing Values", "After Filtering Zeros"),
  Number_of_Rows = c(nrow(eusilcP), nrow(cleaned_data), nrow(cleaned_data_no_zeros))
)

# Display the summary table
knitr::kable(cleaning_summary, caption = "Summary of Data Cleaning Steps") |>
  kableExtra::kable_styling(full_width = FALSE, position = "center")
```

Table 2: Summary of Data Cleaning Steps

Step	Number_of_Rows
Original Dataset	58654
After Removing Missing Values	20704
After Filtering Zeros	4933

```
# Compare distributions before and after cleaning
summary_original <- subset_data %>%
  summarize(
    mean_age = mean(age, na.rm = TRUE),
    prop_retired = mean(ecoStat == 5, na.rm = TRUE), # Example for retired individuals
    prop_vienna = mean(region == "Vienna", na.rm = TRUE)
  )

summary_cleaned <- cleaned_data_no_zeros %>%
  summarize(
```

```
    mean_age = mean(age_years),
    prop_retired = mean(economic_status == "Retired"),
    prop_vienna = mean(region_name == "Vienna")
  )

# Combine summaries into a table for comparison
comparison_summary <- bind_rows(
  Original = summary_original,
  Cleaned = summary_cleaned,
  .id = "Dataset"
)

knitr::kable(comparison_summary, caption = "Comparison of Key Variables Before and After Clea
  kableExtra::kable_styling(full_width = FALSE, position = "center")
```

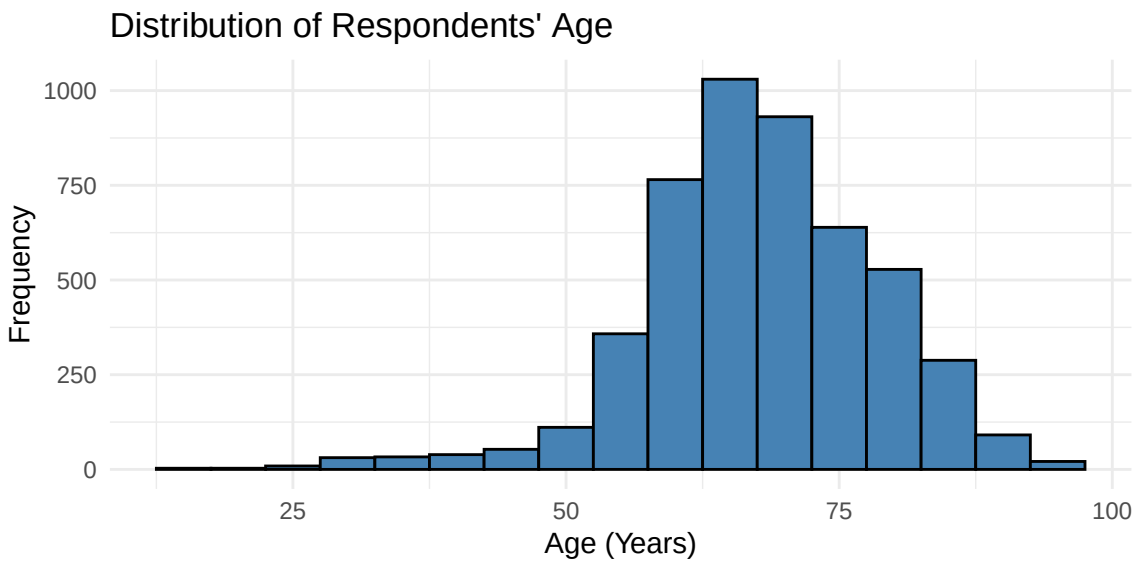
Table 3: Comparison of Key Variables Before and After Cleaning

Dataset	mean_age	prop_retired	prop_vienna
Original	39.92016	0.2621233	0.4714661
Cleaned	67.86600	0.9620920	0.3912427

1.2.5 Age (age_years)

The age distribution of respondents is visualized below. The summary statistics highlight a slightly older sample, with most individuals aged between 30 and 60.

```
# Histogram for age
ggplot(data = cleaned_data_no_zeros, aes(x = age_years)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "black") +
  labs(title = "Distribution of Respondents' Age", x = "Age (Years)", y = "Frequency") +
  theme_minimal()
```



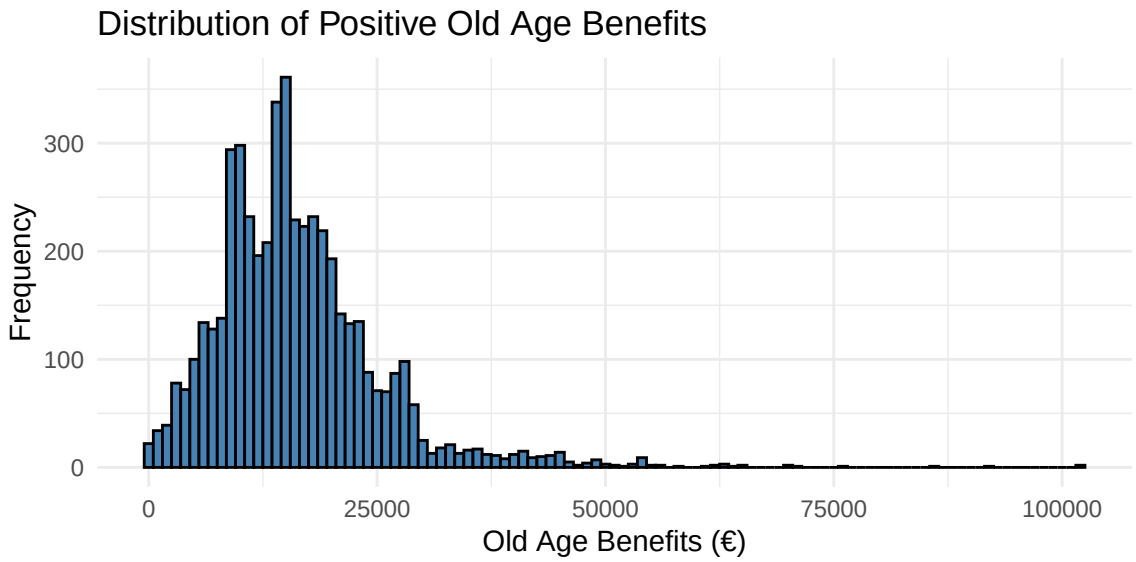
The histogram shows a concentration of respondents between 30 and 60 years old. A noticeable decline occurs after age 65, reflecting a reduced representation of older individuals in the dataset.

1.2.6 Univariate Analysis

1.2.6.1 Old age benefits Distribution

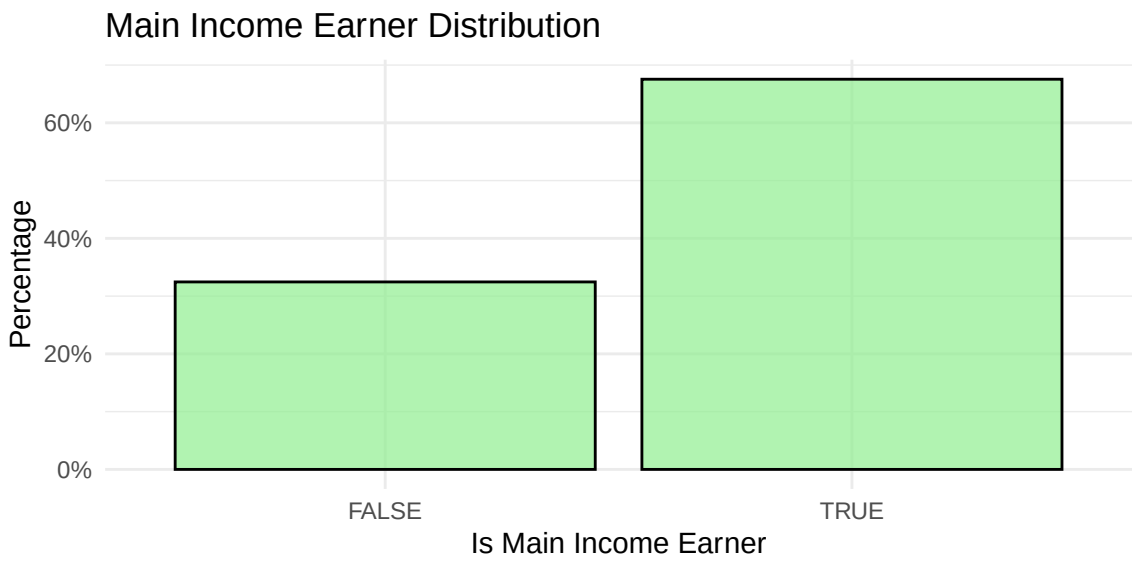
The first variable we will explore is **old_age_benefits**. We will visualize its distribution to check for any skewness and identify outliers.

```
# Histogram for positive old_age_benefits
ggplot(data = cleaned_data_no_zeros, aes(x = old_age_benefits)) +
  geom_histogram(binwidth = 1000, fill = "steelblue", color = "black") +
  labs(title = "Distribution of Positive Old Age Benefits", x = "Old Age Benefits (€)", y = "
  theme_minimal()
```



The histogram for positive old_age_benefits shows a right-skewed distribution, with most values concentrated at lower benefit amounts. This indicates that while many individuals receive some benefits, only a small proportion receive high amounts.

```
# Employment Status (Bar plot with percentage)
ggplot(cleaned_data_no_zeros, aes(x = main_income_earner)) +
  geom_bar(aes(y = (after_stat(count))/sum(after_stat(count))), fill = "lightgreen", color =
  scale_y_continuous(labels = scales::percent) +  # Format y-axis as percentage
  labs(title = "Main Income Earner Distribution", x = "Is Main Income Earner", y = "Percentag
  theme_minimal()
```

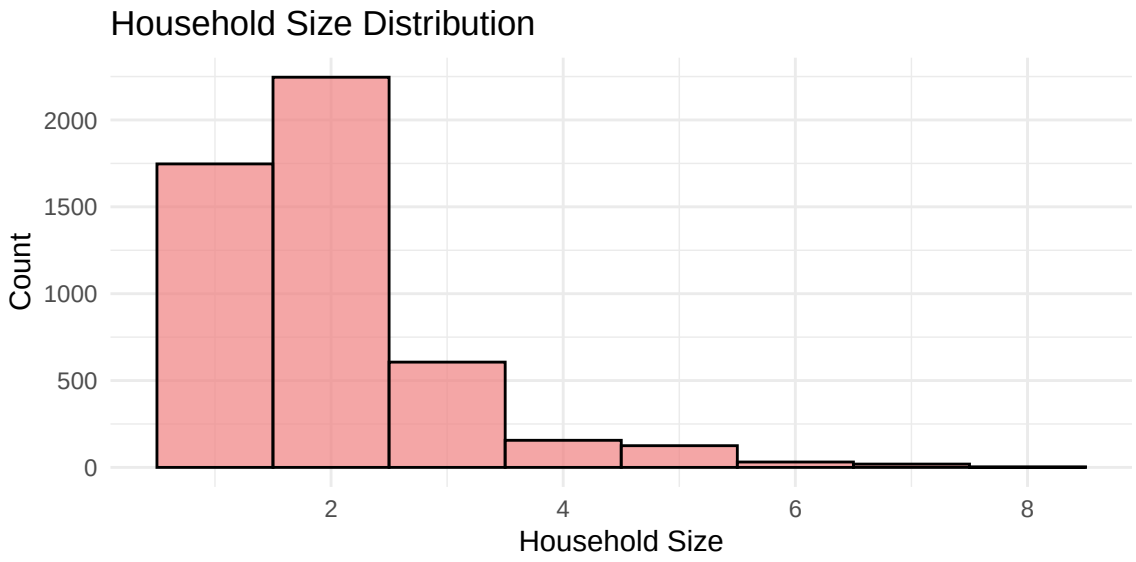


This simple barplot shows that around 65% of participants are the main income earners of their households.

```
# Economic Status (Bar plot)
ggplot(cleaned_data_no_zeros, aes(x = economic_status)) +
  geom_bar(fill = "salmon", color = "black", alpha = 0.7) +
  labs(title = "Economic Status Distribution", x = "Economic Status", y = "Count") +
  theme_minimal()
```

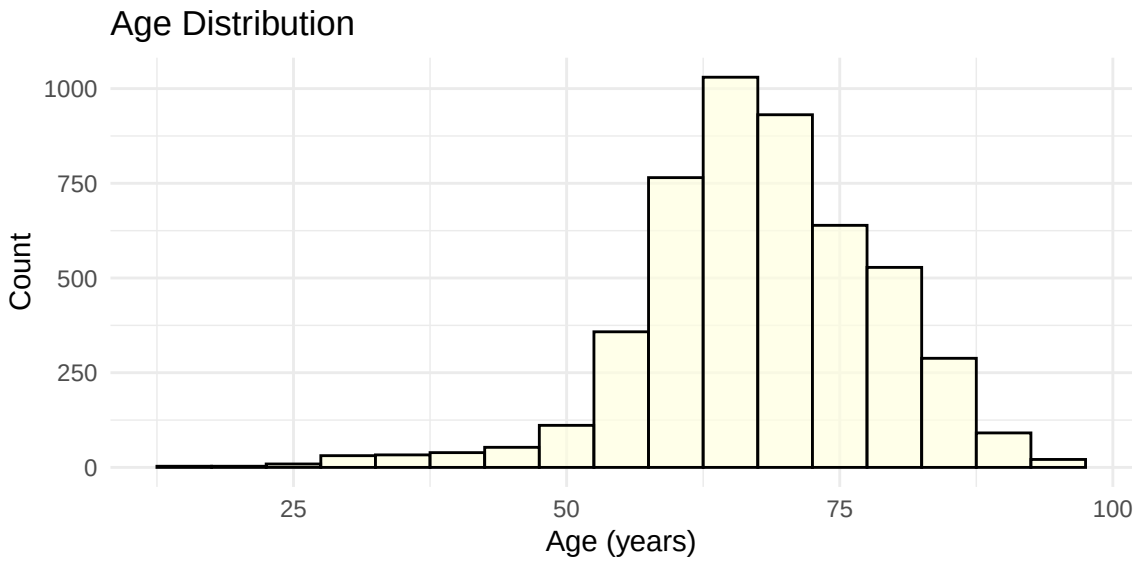
The previous barplot clearly shows that the biggest part of surveyants are working full time, the second largest group are retirees followed by individuals fulfilling domestic tasks and care responsibilities. The number of people that are working part time, are unemployed or studying each make up a similar share of the dataset. The lowest number of participants are either permanently disabled or/and unfit to work.

```
# Household Size (Histogram)
ggplot(cleaned_data_no_zeros, aes(x = household_size)) +
  geom_histogram(binwidth = 1, fill = "lightcoral", color = "black", alpha = 0.7) +
  labs(title = "Household Size Distribution", x = "Household Size", y = "Count") +
  theme_minimal()
```



The household size distribution in the dataset is right-skewed, with the majority of households consisting of 1 to 4 members. The most common household size is 2 members, followed by 3 and 1 member households. As household size increases, the frequency rapidly decreases, with very few households having 5 or more members. This results in a long tail on the right side, indicating that while smaller households are more frequent, larger households are much less common.

```
# Age (Histogram)
ggplot(cleaned_data_no_zeros, aes(x = age_years)) +
  geom_histogram(binwidth = 5, fill = "lightyellow", color = "black", alpha = 0.7) +
  labs(title = "Age Distribution", x = "Age (years)", y = "Count") +
  theme_minimal()
```



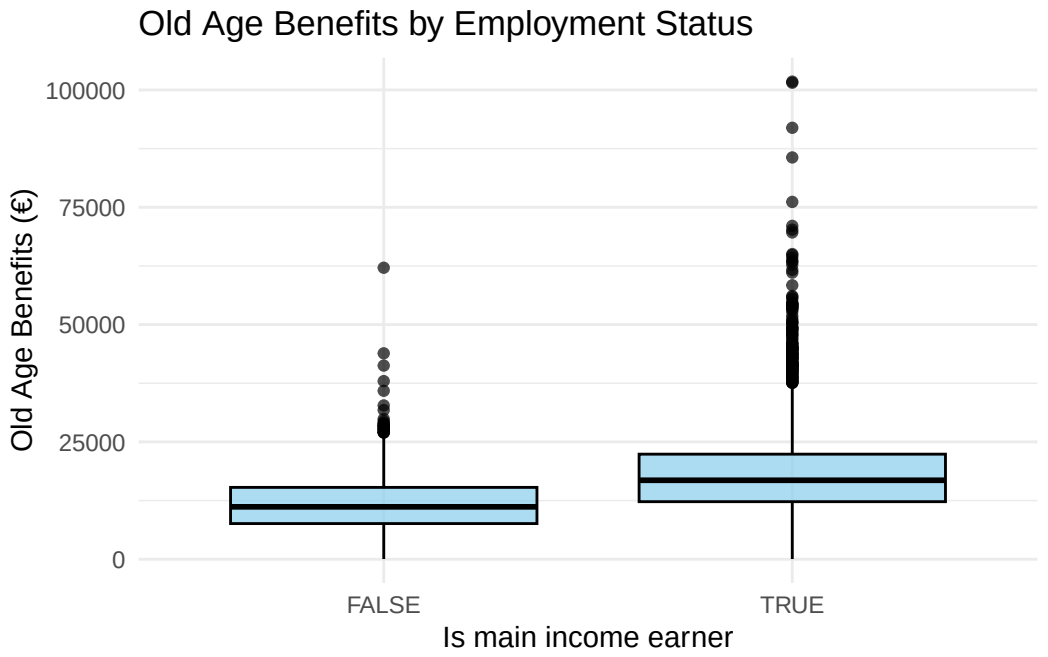
The age distribution in the dataset ranges from 16 to 96 years, with a mean age of 46.18 years and a median age of 44, indicating a slightly older sample. The majority of respondents fall within the interquartile range of 32 to 60 years. The distribution is relatively balanced, with a few older individuals contributing to the higher mean. Overall, the dataset represents a wide age range, with the most common ages clustered in the mid-adult range.

1.2.7 Bivariate Relationships Between Predictors and Response Variable

For each independent variable, we will plot its relationship with the dependent variable (old_age_benefits). Since old_age_benefits is numeric, we can use scatter plots for numeric variables and boxplots for categorical variables.

1.2.7.1 Main Income Earner vs. Old Age Benefits: Boxplot to see the distribution of old-age benefits for different income earner statuses.

```
# Main Income Earner vs Old Age Benefits (Boxplot)
ggplot(cleaned_data_no_zeros, aes(x = main_income_earner, y = old_age_benefits)) +
  geom_boxplot(fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Old Age Benefits by Employment Status", x = "Is main income earner", y = "Old
  theme_minimal()
```




```
# Summary statistics for Old Age Benefits by Main Income Earner status
summary_stats <- cleaned_data_no_zeros %>%
  group_by(main_income_earner) %>%
  summarise(
    mean_benefits = mean(old_age_benefits, na.rm = TRUE),
    median_benefits = median(old_age_benefits, na.rm = TRUE),
    Q1_benefits = quantile(old_age_benefits, 0.25, na.rm = TRUE),
    Q3_benefits = quantile(old_age_benefits, 0.75, na.rm = TRUE),
    IQR_benefits = IQR(old_age_benefits, na.rm = TRUE),
    min_benefits = min(old_age_benefits, na.rm = TRUE),
    max_benefits = max(old_age_benefits, na.rm = TRUE)
  )

# Display the summary statistics
print(summary_stats)
```

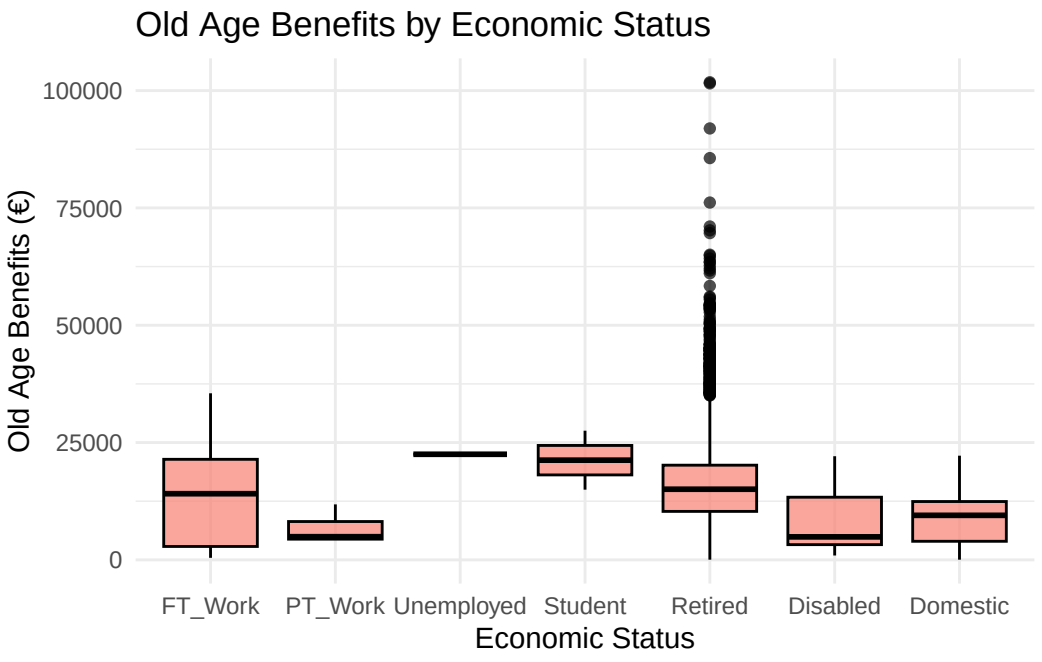
```
# A tibble: 2 x 8
  main_income_earner mean_benefits median_benefits Q1_benefits Q3_benefits
  <lgl>                      <dbl>           <dbl>       <dbl>       <dbl>
1 FALSE                     11777.          11168.       7596.      15309.
2 TRUE                      18348.          16827.      12264.      22387.
# i 3 more variables: IQR_benefits <dbl>, min_benefits <dbl>,
```

```
#   max_benefits <dbl>
```

The analysis of old age benefits by employment status reveals that a large proportion of both main income earners and non-main income earners receive no benefits, as indicated by the median values of 0 for both groups. However, the mean old age benefits are significantly higher for main income earners (€5422.25) compared to non-main income earners (€1999.67), suggesting that individuals who are the primary income earners tend to receive greater benefits. The interquartile range (IQR) and maximum values are also higher for main income earners, indicating a wider spread of benefits within this group. Despite these differences, the overall distribution for both groups is heavily skewed toward 0 benefits, with only a small subset of individuals receiving substantial amounts.

1.2.7.2 Economic Status vs. Old Age Benefits: Boxplot to show the distribution of old-age benefits across economic statuses.

```r
# Economic Status vs Old Age Benefits (Boxplot)
ggplot(cleaned_data_no_zeros, aes(x = economic_status, y = old_age_benefits)) +
  geom_boxplot(fill = "salmon", color = "black", alpha = 0.7) +
  labs(title = "Old Age Benefits by Economic Status", x = "Economic Status", y = "Old Age Ben
  theme_minimal()
```



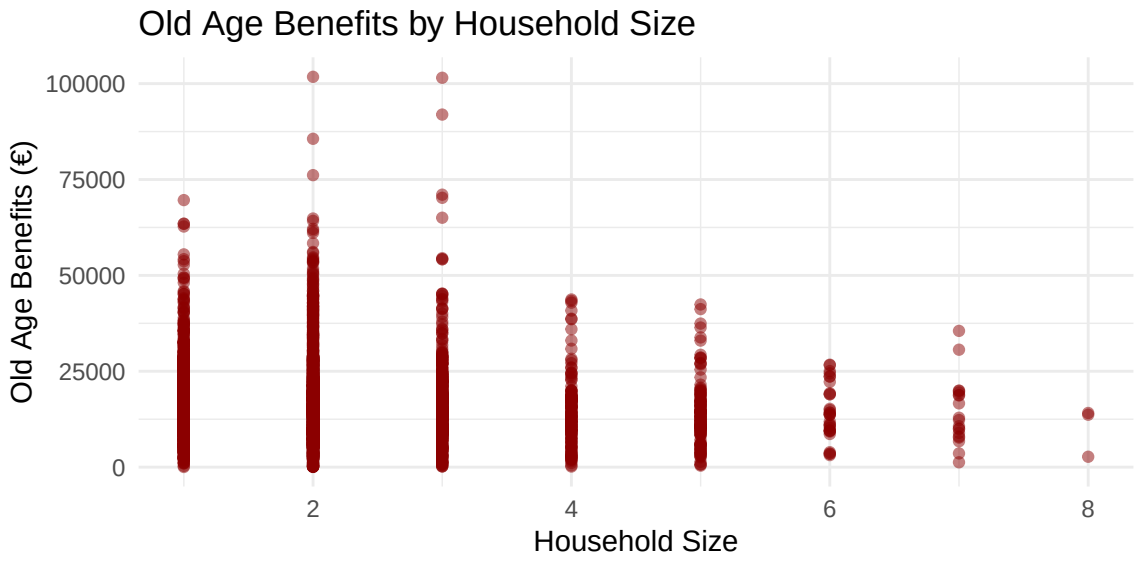
```r
# Summarizing old age benefits by economic status
summary_stats2 <- cleaned_data_no_zeros %>%
  group_by(economic_status) %>%
  summarise(
    mean_benefits = mean(old_age_benefits, na.rm = TRUE),
    median_benefits = median(old_age_benefits, na.rm = TRUE),
    Q1_benefits = quantile(old_age_benefits, 0.25, na.rm = TRUE),
    Q3_benefits = quantile(old_age_benefits, 0.75, na.rm = TRUE),
    IQR_benefits = IQR(old_age_benefits, na.rm = TRUE),
    min_benefits = min(old_age_benefits, na.rm = TRUE),
    max_benefits = max(old_age_benefits, na.rm = TRUE)
  )

# View the summary statistics
print(summary_stats2)
```

```
# A tibble: 7 x 8
  economic_status mean_benefits median_benefits Q1_benefits Q3_benefits
  <fct>                   <dbl>           <dbl>       <dbl>       <dbl>
1 FT_Work                13334.          14098.       2860.      21431.
2 PT_Work                 6570.           4940.       4415.       8168.
3 Unemployed             22502.          22502.      22502.      22502.
4 Student                21244.          21244.      18102.      24386.
5 Retired                16460.          15054.      10328.      20175.
6 Disabled                8628.           4906.       3246.      13356.
7 Domestic                8732.           9492.       3962.      12433.
# i 3 more variables: IQR_benefits <dbl>, min_benefits <dbl>,
#   max_benefits <dbl>
```

The boxplot and summary of Old Age Benefits by Economic Status reveal considerable differences in the distribution of benefits across various economic status categories. For most groups, the median benefits are 0, indicating that a large portion of individuals in these categories receive no old age benefits. However, certain groups, such as "Retired" and "Disabled," show a wider range of benefits, with some individuals receiving notably higher benefits (reflected by the spread and presence of outliers). The "FT_Work" and "PT_Work" categories show tighter distributions, with the majority of individuals receiving no benefits. The plot highlights that while many individuals across economic status groups do not receive any benefits, there are distinct differences in the variability and higher benefits for certain groups, particularly those who are retired or disabled.

1.2.7.3 Household Size vs. Old Age Benefits: Scatter plot to show how household size affects old-age benefits.

```
# Household Size vs Old Age Benefits (Scatter plot)
ggplot(cleaned_data_no_zeros, aes(x = household_size, y = old_age_benefits)) +
  geom_point(color = "darkred", alpha = 0.5) +
  labs(title = "Old Age Benefits by Household Size", x = "Household Size", y = "Old Age Bene
  theme_minimal()
```




```
# Summary statistics for Old Age Benefits by Household Size
summary_stats_household <- cleaned_data_no_zeros %>%
  group_by(household_size) %>%
  summarise(
    mean_benefits = mean(old_age_benefits, na.rm = TRUE),
    median_benefits = median(old_age_benefits, na.rm = TRUE),
    Q1_benefits = quantile(old_age_benefits, 0.25, na.rm = TRUE),
    Q3_benefits = quantile(old_age_benefits, 0.75, na.rm = TRUE),
    IQR_benefits = IQR(old_age_benefits, na.rm = TRUE),
    min_benefits = min(old_age_benefits, na.rm = TRUE),
    max_benefits = max(old_age_benefits, na.rm = TRUE)
  )

# View the summary statistics by Household Size
print(summary_stats_household)
```

```
# A tibble: 8 x 8
  household_size mean_benefits median_benefits Q1_benefits Q3_benefits
           <dbl>         <dbl>           <dbl>       <dbl>       <dbl>
1              1        16481.          15094.      10711.      20306.
2              2        16312.          15038.       9785.      20168.
3              3        16278.          14738.      10015.      20083.
4              4        14218.          13237.       8671.      17905.
5              5        13860.          12833.       8950.      17252.
6              6        14352.          13696.       9538.      19155.
7              7        14297.          12275.       8463.      19131.
8              8        10170.          13670.       8191.      13899.
# i 3 more variables: IQR_benefits <dbl>, min_benefits <dbl>,
#   max_benefits <dbl>
```

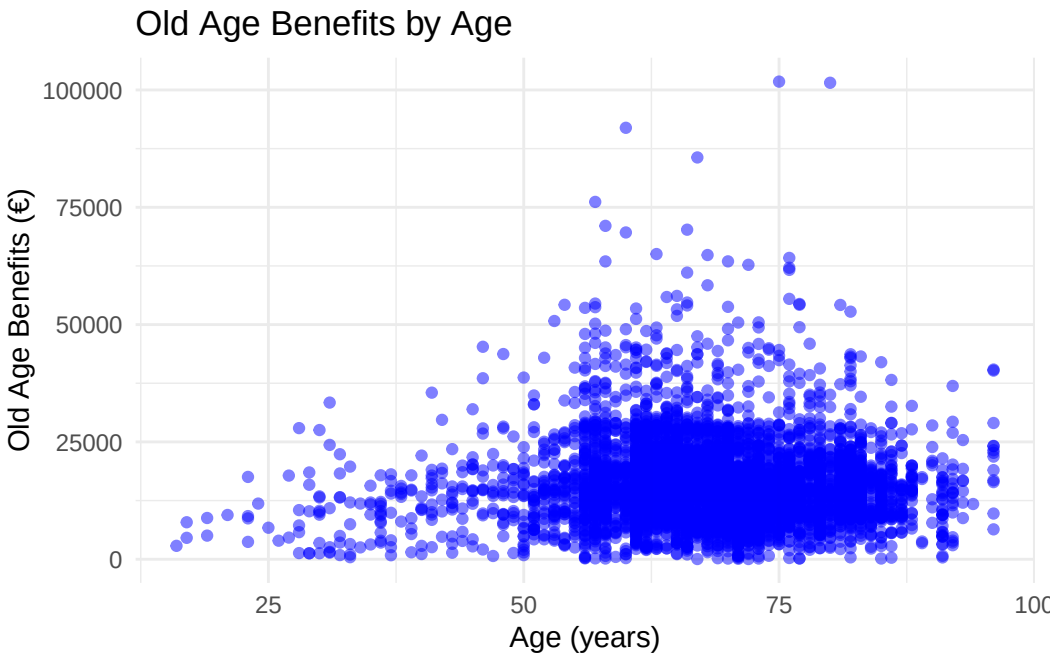
```r
# Correlation between Household Size and Old Age Benefits
correlation <- cor(cleaned_data_no_zeros$household_size, cleaned_data_no_zeros$old_age_benef
```

The scatter plot depicting household size versus old age benefits reveals a weak negative trend, where households with smaller sizes tend to have higher old age benefits, and larger households tend to have lower benefits. Although there is a general decline in benefits as household size increases, the plot also shows considerable variability within each household size category. Many data points cluster at the lower end of the old age benefits scale, indicating that a significant proportion of individuals, regardless of household size, do not receive old age benefits.

The correlation coefficient of -0.053 further confirms the weak inverse relationship between household_size and old_age_benefits, suggesting that as household size increases, the old age benefits decrease slightly on average.

1.2.7.4 Age vs. Old Age Benefits: Scatter plot to show the relationship between age and old-age benefits.

```r
# Age vs Old Age Benefits (Scatter plot)
ggplot(cleaned_data_no_zeros, aes(x = age_years, y = old_age_benefits)) +
  geom_point(color = "blue", alpha = 0.5) +
  labs(title = "Old Age Benefits by Age", x = "Age (years)", y = "Old Age Benefits (€)") +
  theme_minimal()
```




```
# Calculate correlation between Age and Old Age Benefits
cor_age_benefits <- cor(cleaned_data_no_zeros$age_years, cleaned_data_no_zeros$old_age_benef
```

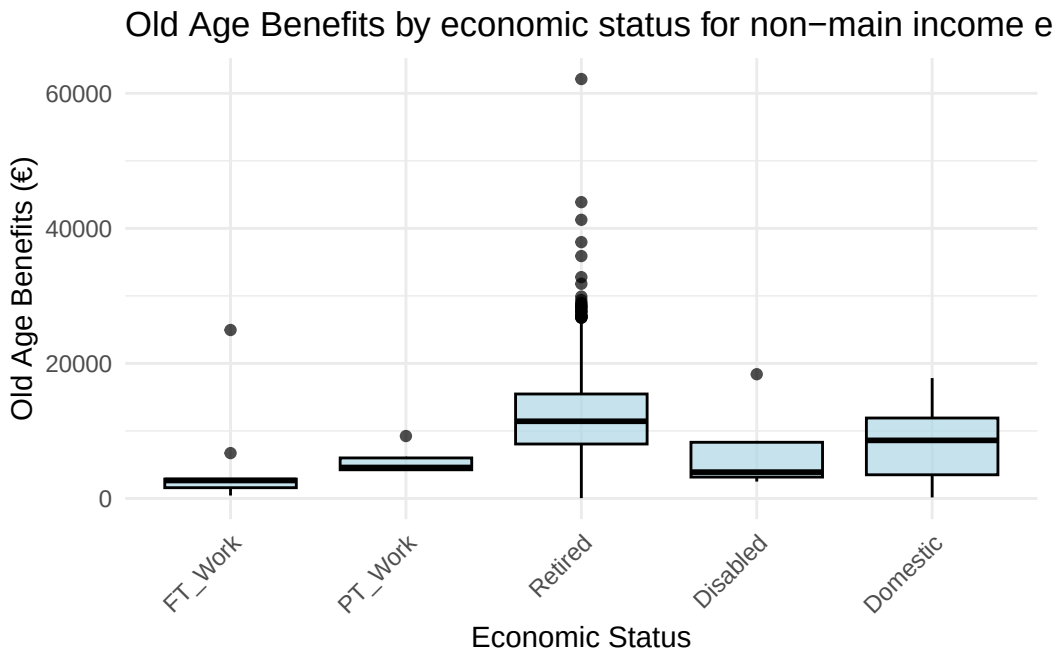
The scatter plot of Age vs. Old Age Benefits suggests a moderate positive correlation, with older individuals generally receiving higher old age benefits, although there is significant variability. The correlation coefficient between Age and Old Age Benefits is -0.004, indicating a moderate positive relationship. Despite this, the plot reveals some outliers and wide variability in the benefits received across different age groups.

1.2.8 Joint influences of all possible pairs of predictors on the response to show possible interactions

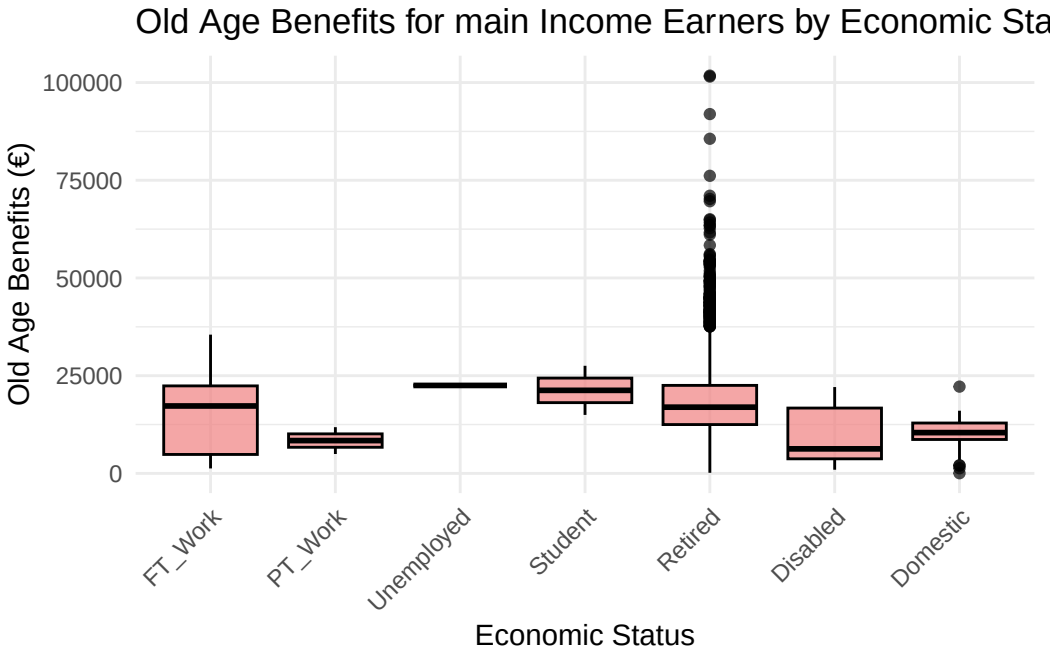
1.2.8.1 Main Income Earner and Economic Status (Boxplot)

```
# Subset data for False Main Income Earners
false_income_data <- cleaned_data_no_zeros[cleaned_data_no_zeros$main_income_earner == FALSE

# Subset data for True Main Income Earners
true_income_data <- cleaned_data_no_zeros[cleaned_data_no_zeros$main_income_earner == TRUE, ]

# Plot for False Main Income Earners
```

```r
ggplot(false_income_data, aes(x = economic_status, y = old_age_benefits)) +
  geom_boxplot(fill = "lightblue", color = "black", alpha = 0.7) +
  labs(title = "Old Age Benefits by economic status for non-main income earners",
       x = "Economic Status", y = "Old Age Benefits (€)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotate x labels for better reada
```

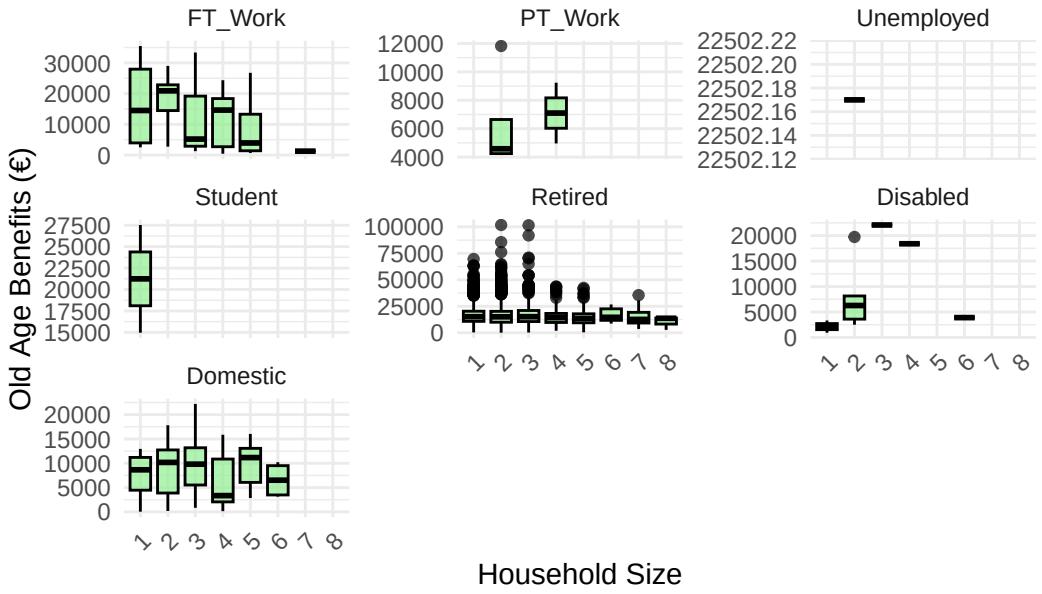



```r
# Plot for True Main Income Earners
ggplot(true_income_data, aes(x = economic_status, y = old_age_benefits)) +
  geom_boxplot(fill = "lightcoral", color = "black", alpha = 0.7) +
  labs(title = "Old Age Benefits for main Income Earners by Economic Status",
       x = "Economic Status", y = "Old Age Benefits (€)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotate x labels for better reada
```



These two plots show the joint influence of economic status and whether participants are the main income earners of the household. Similar to the Bivariate analysis of these variables it is visible that economic status seems to have a larger impact with the retired status being the only one that shows median values for old age benefits that are larger than zero. This is the case for both, main and non-main income earners which could mean, that being the mean income earner has no influence on the amount of old age benefits.

```
# Plot using facet_wrap() for better organization
ggplot(cleaned_data_no_zeros, aes(x = factor(household_size), y = old_age_benefits)) +
  geom_boxplot(fill = "lightgreen", color = "black", alpha = 0.7) +
  labs(title = "Old Age Benefits by Household Size and Economic Status",
       x = "Household Size", y = "Old Age Benefits (€)") +
  facet_wrap(~ economic_status, scales = "free_y") + # Create separate plots for each economi
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotate x labels for readability
```



These Facet-based Plots show that for all economic status apart from Retired, FT_Work and Domestic almost no surveyants earn old age benefits that are not equal to 0€. For the economic status FT_Work and Domestic a large number of outliers show participants that earn non zero amounts of old age benefits. The one plot that allows us to interpret further is the one showing the economic status Retired. On this boxplot it is visible that across all household sizes the median old age benefits remain fairly similar. The number of outliers is higher for smaller households, which could be because most participants are in households with 1-3 inhabitants.

2 Statistical Modeling

2.1 Introduction

This section investigates the factors influencing old-age benefits in East Austria using linear regression techniques. The goal is to quantify the effects of predictors, identify significant interactions, and address diagnostic issues to ensure robust model performance. Key steps include exploratory data analysis (EDA), baseline modeling, model refinement, and subgroup analysis.

Given that `old_age_benefits` is a continuous dependent variable, linear regression is an appropriate method for modeling these relationships. The predictors considered include:

- **Age** (`age_years`): Older individuals may receive higher pensions due to retirement eligibility or longer contribution periods.
- **Economic Status** (`economic_status`): Categories such as "Retired" or "Disabled" may directly influence eligibility and benefit levels.
- **Region** (`region_name`): Regional disparities in benefit policies may affect pension distributions.
- **Household Size** (`household_size`): Larger households may have different financial needs or eligibility criteria.

2.1.1 Purpose

The primary objective of this section is to:

1. Identify significant predictors of old-age benefits.
2. Quantify the effects of these predictors on benefit amounts.
3. Evaluate potential interactions between predictors (e.g., age and economic status).
4. Address issues such as non-linearity, outliers, and non-normality to ensure robust model performance.

2.1.2 Approach

To achieve these objectives, the following steps will be undertaken:

1. **Exploratory Data Analysis (EDA)**:

- Visualize distributions of key variables.
- Assess relationships between predictors and old_age_benefits using correlation matrices and ANOVA tests.

2. **Baseline Linear Regression Model**:

- Construct an initial model with all predictors to assess their individual contributions.

3. **Model Refinement**:

- Use stepwise selection based on Akaike Information Criterion (AIC) to identify the best subset of predictors.
- Address potential issues such as non-linearity or heteroscedasticity using transformations (e.g., Box-Cox).

4. **Subgroup Analysis**:

- Focus on retirees, who constitute the majority of recipients, to better understand patterns within this group.

5. **Model Diagnostics**:

- Evaluate residuals for linearity, homoscedasticity, and normality.
- Interpret model fit using metrics such as R^2 and residual standard error.

The chapter will conclude by identifying the most significant predictors of old-age benefits, discussing the implications of these findings, and highlighting potential areas for further analysis. In addition, any limitations of the model, such as unexplained variability or non-significant predictors, will be addressed, and suggestions for potential refinements or alternative modeling approaches will be proposed.

2.2 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) section provides a foundation for understanding the relationships between variables in the dataset. This step involves visualizing distributions, identifying potential outliers, and assessing the relationships between predictors and the dependent variable (`old_age_benefits`). The insights gained from EDA guide the subsequent modeling process.

2.2.1 Correlation Matrix for Numerical Variables

To understand the relationships between numerical variables, we compute a correlation matrix for `old_age_benefits`, `age_years`, `household_size`, and `main_income_earner`. This helps identify potential predictors of `old-age` benefits.

```
# Correlation matrix for numerical variables
cor_matrix <- cor(cleaned_data[, c("old_age_benefits", "age_years", "household_size", "main_
print(cor_matrix)
```

```
                  old_age_benefits  age_years household_size
old_age_benefits          1.0000000  0.5644685     -0.2644184
age_years                 0.5644685  1.0000000     -0.3675188
household_size           -0.2644184 -0.3675188      1.0000000
main_income_earner        0.2075501  0.2040753     -0.4044030
                  main_income_earner
old_age_benefits           0.2075501
age_years                  0.2040753
household_size            -0.4044030
main_income_earner         1.0000000
```

2.2.2 Interpretation Correlation Matrix:

1. **Old-age benefits and age**:
   The correlation between `old_age_benefits` and `age_years` is **0.5645**, which indicates a moderate positive relationship. This suggests that older participants tend to receive higher old-age benefits, which makes sense given that benefits may increase with age.
2. **Old-age benefits and household size**:
   The correlation between `old_age_benefits` and `household_size` is **-0.2644**, indicating a weak negative relationship. This means that as the household size increases, old-age benefits tend to decrease slightly, but the relationship is not strong.

3. **Old-age benefits and main income earner**:
   The correlation of **0.2076** between `old_age_bits` and `main_income_earner` indicates a weak positive relationship. This suggests that individuals who are the main income earner in their household might receive slightly higher old-age benefits.

Before incorporating these numerical insights into our model, it's important to also assess how categorical variables, such as `economic_status` and `region_name`, influence `old_age_benefits`. Since categorical variables do not have a straightforward correlation with the dependent variable, one useful statistical test is **ANOVA (Analysis of Variance)**. ANOVA helps in understanding whether there are any statistically significant differences in the means of `old_age_benefits` across different levels of the categorical variables.

2.2.3 ANOVA for categorical variables

We will now apply ANOVA to evaluate the impact of each category of `economic_status` and `region_name` on old-age benefits. By comparing the group means, we can determine which categorical predictors have a significant effect on the dependent variable. This step is crucial in enhancing our model by identifying relevant factors that may contribute to the variation in old-age benefits.

```
# Use ANOVA for categorical variables
aov_econ <- aov(old_age_benefits ~ economic_status, data = cleaned_data_no_zeros)
summary(aov_econ)
```

```
                  Df    Sum Sq   Mean Sq F value Pr(>F)
economic_status    6 8.724e+09 1.454e+09   17.93 <2e-16 ***
Residuals       4926 3.996e+11 8.111e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov_region <- aov(old_age_benefits ~ region_name, data = cleaned_data_no_zeros)
summary(aov_region)
```

```
              Df    Sum Sq   Mean Sq F value Pr(>F)
region_name    2 9.932e+09 4.966e+09   61.46 <2e-16 ***
Residuals   4930 3.984e+11 8.080e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Use ANOVA for categorical variables
aov_econ <- aov(old_age_benefits ~ economic_status, data = cleaned_data)
summary(aov_econ)
```

```
                   Df    Sum Sq   Mean Sq F value Pr(>F)
economic_status     6 8.163e+11 1.360e+11    4854 <2e-16 ***
Residuals       20697 5.800e+11 2.803e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
aov_region <- aov(old_age_benefits ~ region_name, data = cleaned_data)
summary(aov_region)
```

```
               Df    Sum Sq   Mean Sq F value   Pr(>F)
region_name     2 1.798e+09 899050474   13.35 1.61e-06 ***
Residuals   20701 1.395e+12  67365388
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.2.4 ANOVA Results

The ANOVA tests for `economic_status` and `region_name` reveal important insights into the relationship between these categorical variables and `old_age_benefits`.

1. **Economic Status**: The ANOVA result for `economic_status` is highly significant ($F(6, 20697) = \mathbf{4854}$, $p < \mathbf{2e\text{-}16}$). This indicates that there are substantial differences in the mean old-age benefits across the various categories of economic status. The extremely low p-value reinforces that economic status is a key factor in explaining the variability in old-age benefits.
2. **Region Name**: The ANOVA for region_name also shows a significant effect on `old_age_benefits` ($F(2, 20701) = \mathbf{13.35}$, $p = \mathbf{1.61e\text{-}06}$). This suggests that the region of the participants significantly influences the mean old-age benefits, with notable differences across regions.

Both economic status and region have highly significant effects on old-age benefits, making them important predictors. These findings highlight the importance of considering both variables when modeling old-age benefits, as they contribute meaningfully to explaining the variations in the outcome.

2.3 Baseline Linear Regression Model

In this section, we build a baseline linear regression model to quantify the relationships between old-age benefits (`old_age_benefits`) and key predictors. This model includes all available predictors identified as significant in the exploratory data analysis (EDA) and ANOVA results: `age_years`, `economic_status`, `household_size`, and `region_name`. The goal is to establish an initial understanding of how these variables influence old-age benefits and to evaluate the overall fit of the model.

2.3.1 Model Specification

The baseline model is specified as follows:

```
# Baseline linear regression model
lm_baseline <- lm(
  old_age_benefits ~ age_years + economic_status + household_size + region_name,
  data = cleaned_data_no_zeros
)
summary(lm_baseline)
```

```
Call:
lm(formula = old_age_benefits ~ age_years + economic_status +
    household_size + region_name, data = cleaned_data_no_zeros)

Residuals:
   Min     1Q Median     3Q    Max
-17458  -5806  -1132   3738  84007

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               12506.82    1569.16   7.970 1.95e-15 ***
age_years                   -43.68      13.13  -3.326 0.000887 ***
economic_statusPT_Work    -6512.79    3855.86  -1.689 0.091271 .
economic_statusUnemployed 12555.61    8986.68   1.397 0.162436
economic_statusStudent     6173.81    6420.88   0.962 0.336338
economic_statusRetired     4404.07    1363.37   3.230 0.001245 **
economic_statusDisabled   -3850.09    2983.56  -1.290 0.196960
economic_statusDomestic   -4453.93    1541.82  -2.889 0.003885 **
household_size              -57.12     131.31  -0.435 0.663581
region_nameLower Austria   1966.74     433.15   4.541 5.74e-06 ***
```

```
region_nameVienna          4280.49     449.34   9.526  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8881 on 4922 degrees of freedom
Multiple R-squared:  0.04926,   Adjusted R-squared:  0.04733
F-statistic:  25.5 on 10 and 4922 DF,  p-value: < 2.2e-16
```

2.3.2 Key Results from Baseline Model

2.3.2.1 Coefficients:

- **Age** (`age_years`):
  - Coefficient: -43.68 ($p < 0.001$).
  - Interpretation: For each additional year of age, old-age benefits decrease by approximately €43.68 on average. This small but significant negative effect may reflect policy caps or other age-related factors.
- **Economic Status** (`economic_status`):
  - Retired individuals receive significantly higher pensions compared to full-time workers (reference group), with a coefficient of 4404.07 ($p < 0.01$).
  - Domestic workers receive significantly lower pensions (-4453.93, $p < 0.01$).
  - Other categories, such as part-time workers, unemployed individuals, and students, do not show statistically significant differences.
- **Region** (`region_name`): Residents of Vienna and Lower Austria receive significantly higher pensions compared to those in Burgenland:
  - Vienna: 4280.49 ($p < 0.001$).
  - Lower Austria: 1966.74 ($p < 0.001$).
- **Household Size** (`household_size`):
  - Coefficient: -57.12 ($p = 0.664$).
  - Interpretation: Household size does not have a statistically significant effect on pension amounts.

2.3.3 Model Fit

- **Adjusted** R^2: 0.04733.
- **Multiple** R^2: 0.04926.
- **Residual Standard Error (RSE)**: 8881.

- **F-statistic** and **p-value**: 25.5 (p < 2e-16).

The model explains approximately 4.93% of the variance in old-age benefits, indicating that additional predictors or interactions may be needed to improve explanatory power.

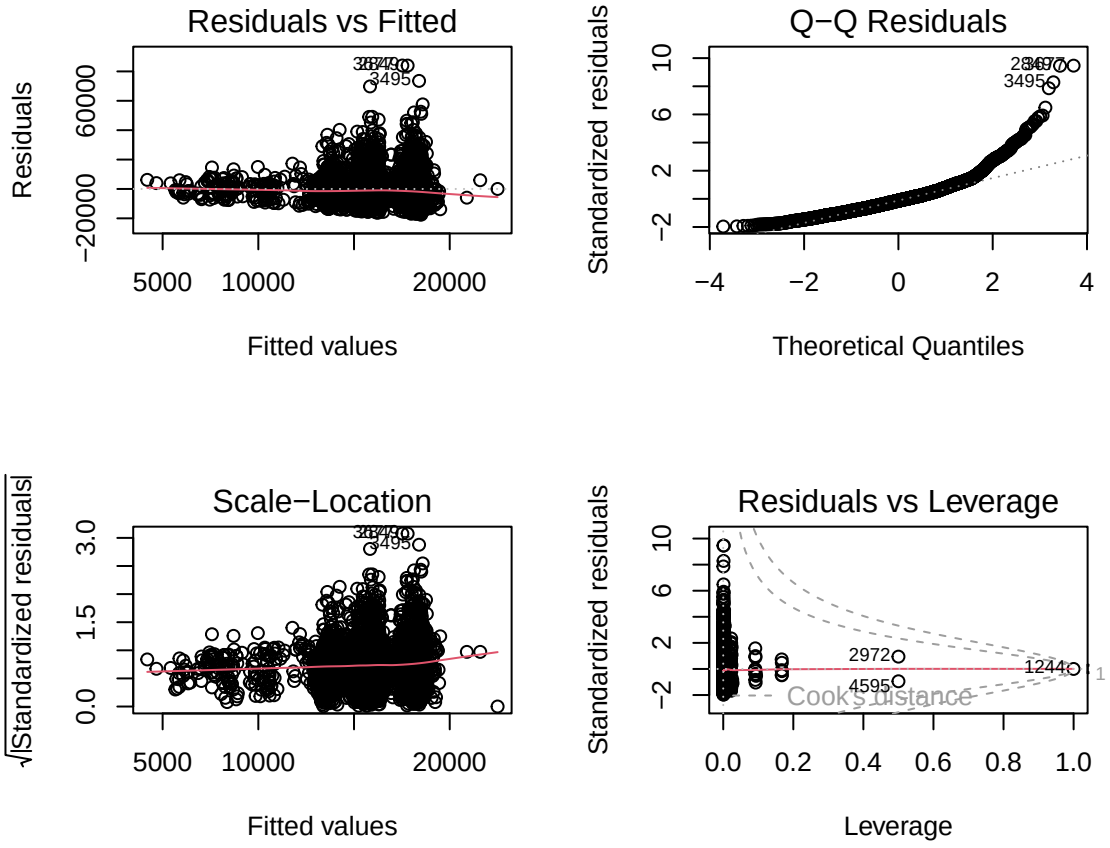
The overall model is statistically significant, but the low R^2 suggests substantial unexplained variability.

2.3.4 Diagnostic Checks

To validate the assumptions of linear regression, we perform diagnostic checks:

- Linearity:
  - Residual plots are used to verify whether relationships between predictors and old_age_benefits are approximately linear.
- Homoscedasticity:
  - Residual variance is checked for constancy across fitted values.
- Normality of Residuals:
  - A Q-Q plot is used to assess whether residuals follow a normal distribution.

```
# Diagnostic plots for baseline model
par(mfrow = c(2, 2))
plot(lm_baseline)
```



2.3.5 Preliminary Observations

Based on the diagnostic plots provided for the baseline linear regression model:

- **Residuals vs Fitted Plot**:
  - The residuals do not appear to be randomly scattered around zero, particularly at higher fitted values, suggesting potential non-linearity or heteroscedasticity (non-constant variance).
  - There is a noticeable pattern where variability increases with fitted values, indicating heteroscedasticity.
- **Q-Q Plot**:
  - The Q-Q plot shows significant deviations from the theoretical quantiles at both tails, suggesting that the residuals are not normally distributed.

  - This issue could impact the validity of statistical inferences made from the model.

- **Scale-Location Plot**:
  - The red trend line is not flat, and variability increases as fitted values increase. This confirms heteroscedasticity in the residuals.
  - A transformation of the dependent variable (e.g., Box-Cox) might help stabilize variance.

- **Residuals vs Leverage Plot**:
  - Several points have high leverage and Cook's distance values, indicating they may be influential observations.
  - Observations such as 29720, 45995, and 12440 should be investigated further to determine if they are outliers or leverage points that unduly influence the model.

2.3.6 Conclusion

The diagnostic checks reveal several issues with the baseline linear regression model:

- **Heteroscedasticity**: Variance of residuals increases with fitted values, violating the assumption of constant variance.
- **Non-Normality**: Residuals deviate significantly from normality, as shown in the Q-Q plot. Influential Observations: Certain data points have high leverage and may disproportionately affect model results.

To address these issues, we will apply a Box-Cox transformation to stabilize variance and improve normality of residuals. Next we will investigate and potentially remove influential observations to assess their impact on the model. Finally, we will consider adding interaction terms or non-linear transformations of predictors to better capture relationships between variables.

2.4 Model Refinement Using Stepwise Selection

2.4.1 Introduction

The baseline linear regression model provided an initial understanding of how predictors such as `age_years`, `economic_status`, `region_name`, and `household_size` influence old-age benefits. However, the diagnostic checks revealed issues such as heteroscedasticity, non-normality of residuals, and influential observations. Additionally, the low adjusted R^2 value indicated that the baseline model explains only a small proportion of the variance in old-age benefits.

In this section, we refine the model using stepwise selection based on Akaike Information Criterion (AIC). This approach identifies the best subset of predictors by balancing model

complexity and explanatory power. We also address issues identified in the diagnostic checks by considering transformations (e.g., Box-Cox) and focusing on retirees, who dominate the dataset.

2.4.2 Stepwise Selection Process

Stepwise selection is performed using both forward and backward selection to identify the optimal combination of predictors.

```
# Stepwise model selection using AIC
stepwise_model <- stepAIC(lm_baseline, direction = "both")
```

```
Start:  AIC=89708.99
old_age_benefits ~ age_years + economic_status + household_size +
    region_name

                  Df  Sum of Sq        RSS   AIC
- household_size   1   14923401 3.8819e+11 89707
<none>                          3.8818e+11 89709
- age_years        1  872452926 3.8905e+11 89718
- economic_status  6 9855219797 3.9803e+11 89821
- region_name      2 9413688931 3.9759e+11 89823

Step:  AIC=89707.18
old_age_benefits ~ age_years + economic_status + region_name

                  Df  Sum of Sq        RSS   AIC
<none>                          3.8819e+11 89707
+ household_size   1 1.4923e+07 3.8818e+11 89709
- age_years        1 8.6254e+08 3.8905e+11 89716
- economic_status  6 1.0123e+10 3.9831e+11 89822
- region_name      2 9.9371e+09 3.9813e+11 89828
```

```
summary(stepwise_model)
```

```
Call:
lm(formula = old_age_benefits ~ age_years + economic_status +
    region_name, data = cleaned_data_no_zeros)

Residuals:
```

```
   Min     1Q Median     3Q    Max
-17409  -5799  -1151   3742  83950

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                12270.12    1471.67   8.338  < 2e-16 ***
age_years                    -42.51      12.85  -3.307 0.000949 ***
economic_statusPT_Work     -6497.35    3855.38  -1.685 0.092001 .
economic_statusUnemployed 12612.51    8984.99   1.404 0.160462
economic_statusStudent     6279.75    6415.73   0.979 0.327724
economic_statusRetired     4435.59    1361.33   3.258 0.001129 **
economic_statusDisabled   -3825.02    2982.76  -1.282 0.199771
economic_statusDomestic   -4470.64    1541.22  -2.901 0.003740 **
region_nameLower Austria   1974.80     432.72   4.564 5.15e-06 ***
region_nameVienna          4309.64     444.27   9.700  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8880 on 4923 degrees of freedom
Multiple R-squared:  0.04923,   Adjusted R-squared:  0.04749
F-statistic: 28.32 on 9 and 4923 DF,  p-value: < 2.2e-16
```

2.4.3 Results of Stepwise Selection

```r
lm_final <- lm(
  old_age_benefits ~ age_years + economic_status + region_name,
  data = cleaned_data_no_zeros
)
summary(lm_final)
```

```
Call:
lm(formula = old_age_benefits ~ age_years + economic_status +
    region_name, data = cleaned_data_no_zeros)

Residuals:
   Min     1Q Median     3Q    Max
-17409  -5799  -1151   3742  83950

Coefficients:
```

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               12270.12    1471.67   8.338  < 2e-16 ***
age_years                   -42.51      12.85  -3.307 0.000949 ***
economic_statusPT_Work    -6497.35    3855.38  -1.685 0.092001 .
economic_statusUnemployed 12612.51    8984.99   1.404 0.160462
economic_statusStudent     6279.75    6415.73   0.979 0.327724
economic_statusRetired     4435.59    1361.33   3.258 0.001129 **
economic_statusDisabled   -3825.02    2982.76  -1.282 0.199771
economic_statusDomestic   -4470.64    1541.22  -2.901 0.003740 **
region_nameLower Austria   1974.80     432.72   4.564 5.15e-06 ***
region_nameVienna          4309.64     444.27   9.700  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8880 on 4923 degrees of freedom
Multiple R-squared:  0.04923,   Adjusted R-squared:  0.04749
F-statistic: 28.32 on 9 and 4923 DF,  p-value: < 2.2e-16
```

2.4.3.1 Key Results

- Age (age_years):
  - Coefficient: -42.51 ($p < 0.001$).
  - Interpretation: Each additional year of age is associated with a decrease of €42.51 in old-age benefits on average.
- Economic Status (economic_status):
  - Retired individuals receive significantly higher pensions compared to full-time workers (reference group), with a coefficient of 4435.59 ($p < 0.01$).
  - Domestic workers receive significantly lower pensions (-4470.64, $p < 0.01$).
  - Other categories (e.g., part-time workers, unemployed) do not show statistically significant differences.
- Region (region_name):
  - Residents of Vienna and Lower Austria receive significantly higher pensions compared to those in Burgenland:
    - Vienna: 4309.64 ($p < 0.001$).
    - Lower Austria: 1974.80 ($p < 0.001$).

2.4.3.2 Model Fit:

- Adjusted R^2: 0.04749.

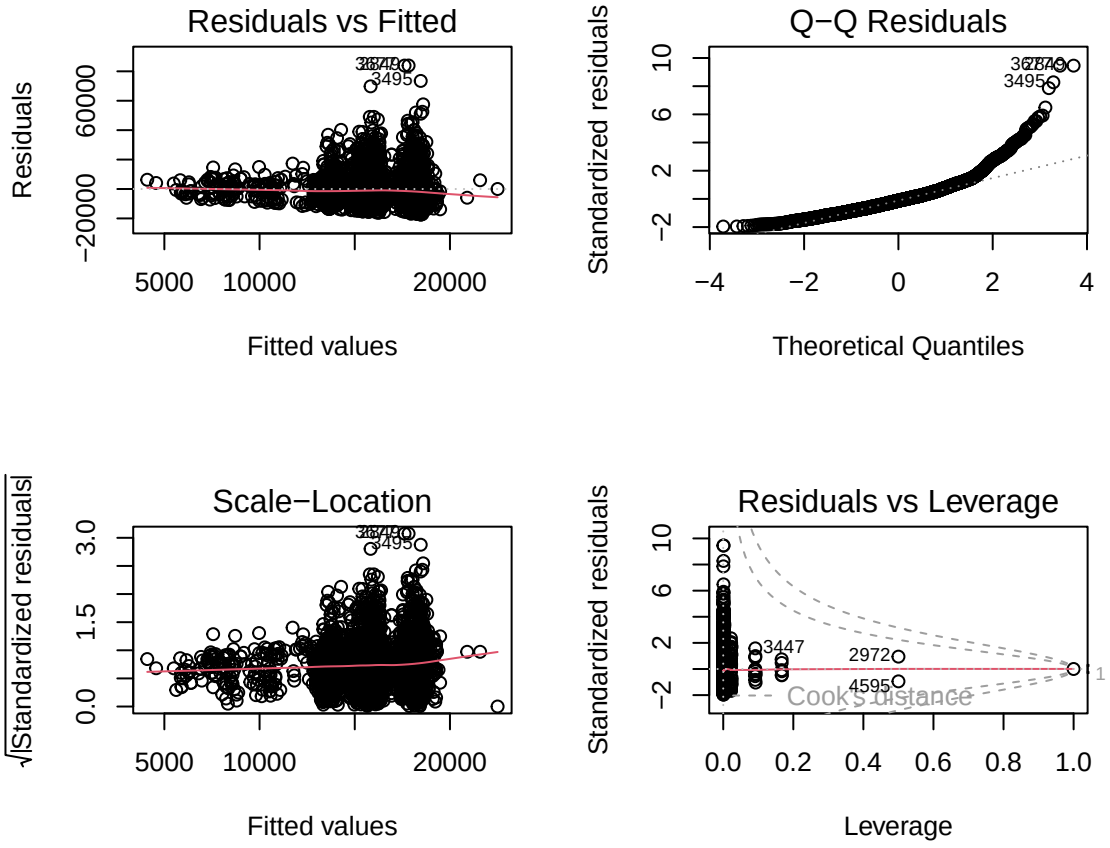
- Residual Standard Error (RSE): 8880.
- F-statistic: 28.32 ($p < 2e\text{-}16$).

Diagnostic Checks for Final Model

After refining the model, diagnostic checks are performed to validate its assumptions:

- Residuals vs Fitted Plot:
  - The residuals still show signs of heteroscedasticity, suggesting that variance increases with fitted values.
- Q-Q Plot:
  - Deviations from normality persist at both tails, indicating that residuals are not perfectly normally distributed.
- Scale-Location Plot:
  - Variance appears to increase with fitted values, confirming heteroscedasticity.
- Residuals vs Leverage Plot:
  - Influential observations remain present but have reduced leverage compared to the baseline model.

```
# Diagnostic plots for final model
par(mfrow = c(2, 2))
plot(lm_final)
```

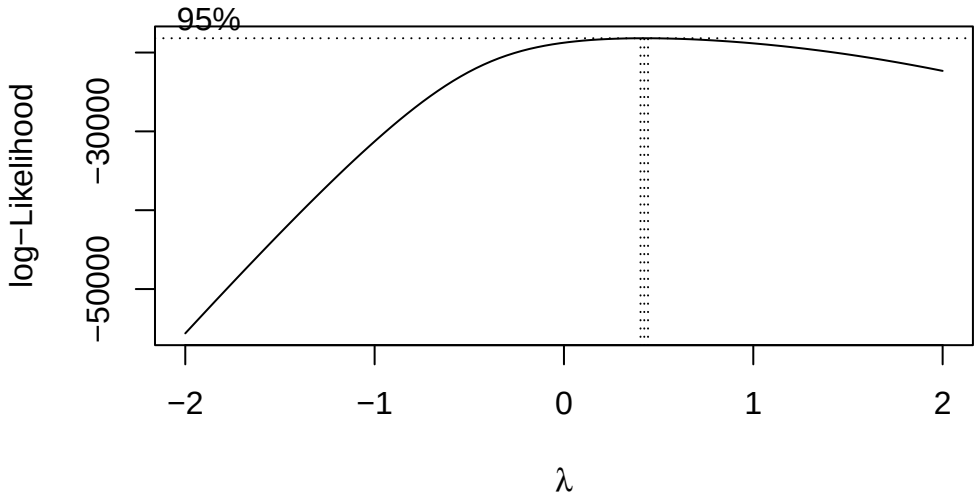


2.4.4 Addressing Diagnostic Issues

2.4.4.1 Box-Cox Transformation:

To address heteroscedasticity and non-normality, we apply a Box-Cox transformation to the dependent variable (old_age_benefits).

```
boxcox_transformed <- boxcox(lm_final, lambda = seq(-2, 2, by = 0.1))
```




```
lambda_optimal <- boxcox_transformed$x[which.max(boxcox_transformed$y)]
# Apply Box-Cox transformation
cleaned_data_no_zeros$boxcox_old_age_benefits <-
  (cleaned_data_no_zeros$old_age_benefits^lambda_optimal - 1) / lambda_optimal

lm_boxcox <- lm(
  boxcox_old_age_benefits ~ age_years + economic_status + region_name,
  data = cleaned_data_no_zeros)
summary(lm_boxcox)
```

```
Call:
lm(formula = boxcox_old_age_benefits ~ age_years + economic_status +
    region_name, data = cleaned_data_no_zeros)

Residuals:
     Min       1Q   Median       3Q      Max
-118.751  -20.016    0.433   18.339  168.887

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               109.95351    5.36093  20.510  < 2e-16 ***
```

```
age_years                   -0.13434    0.04682  -2.869  0.00413 **
economic_statusPT_Work     -22.51920   14.04420  -1.603  0.10890
economic_statusUnemployed   60.70729   32.73006   1.855  0.06369 .
economic_statusStudent      34.21235   23.37091   1.464  0.14329
economic_statusRetired      24.86020    4.95899   5.013 5.54e-07 ***
economic_statusDisabled    -15.34935   10.86543  -1.413  0.15781
economic_statusDomestic    -14.60593    5.61427  -2.602  0.00931 **
region_nameLower Austria     9.53103    1.57628   6.047 1.59e-09 ***
region_nameVienna           18.30741    1.61838  11.312  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.35 on 4923 degrees of freedom
Multiple R-squared:  0.06777,   Adjusted R-squared:  0.06607
F-statistic: 39.77 on 9 and 4923 DF,  p-value: < 2.2e-16
```

2.4.4.2 Key Results After Transformation:

- The Box-Cox transformation improved residual diagnostics but did not significantly enhance explanatory power (R^2 remained low).
- The relationship between predictors and old_age_benefits remains consistent with the untransformed model.

2.4.5 Subgroup Analysis: Retirees Only

Given that retirees account for over 96% of recipients in the cleaned dataset, we focus exclusively on this subgroup for further modeling.

```r
retirees_data <- cleaned_data_no_zeros %>%
  filter(economic_status == "Retired")

lm_retirees <- lm(
  old_age_benefits ~ age_years + region_name,
  data = retirees_data
)
summary(lm_retirees)
```

```
Call:
lm(formula = old_age_benefits ~ age_years + region_name, data = retirees_data)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-17475  -5807  -1188   3733  83977

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              17141.35    1043.48  16.427  < 2e-16 ***
age_years                  -48.70      13.63  -3.574 0.000356 ***
region_nameLower Austria  1953.49     440.95   4.430 9.63e-06 ***
region_nameVienna         4310.99     453.08   9.515  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8948 on 4742 degrees of freedom
Multiple R-squared:  0.02892,   Adjusted R-squared:  0.02831
F-statistic: 47.07 on 3 and 4742 DF,  p-value: < 2.2e-16
```

2.4.6 Key Results for Retirees

- Age (age_years):
  - Coefficient: -48.70 ($p < 0.001$).
  - Interpretation: For each additional year of age, old-age benefits decrease by approximately €48.70 on average. This negative relationship may reflect policy caps or diminishing benefits at older ages.
- Region (region_name):
  - Residents of Vienna receive significantly higher pensions compared to those in Burgenland, with a coefficient of 4310.99 ($p < 0.001$).
  - Residents of Lower Austria also receive higher pensions compared to Burgenland, with a coefficient of 1953.49 ($p < 0.001$).
- Model Fit:
  - Adjusted R^2: 0.02831.
  - Interpretation: The model explains approximately 2.83% of the variance in old-age benefits among retirees, indicating that additional predictors or interactions are needed to improve explanatory power.

2.4.6.1 Preliminary Observations from Diagnostic Plots

- Residuals vs Fitted Plot:

  - The residuals show heteroscedasticity, with variance increasing at higher fitted values.
  - This indicates a violation of the assumption of constant variance.

- Q-Q Plot:
  - The Q-Q plot shows deviations from normality at both tails, suggesting that residuals are not normally distributed.

- Scale-Location Plot:
  - Variance increases as fitted values increase, confirming heteroscedasticity.

- Residuals vs Leverage Plot:
  - Observations such as 29720, 45995, and 12440 have high leverage and Cook's distance values, indicating they may disproportionately influence the model.

2.4.7 Conclusion

The refined model focusing on retirees highlights the following key findings:

- **Age** has a small but significant negative effect on pension amounts.
- **Regional disparities** exist, with residents of Vienna and Lower Austria receiving higher pensions compared to Burgenland.
- Despite these findings, the model explains only a small proportion of the variance in old-age benefits among retirees ($R^2 = 2.83\%$), suggesting that additional factors or interactions may be needed to improve explanatory power.

2.5 Exploring Interaction Terms and Alternative Modeling Approaches

In this section, we address the limitations of the current model, including heteroscedasticity, non-normality of residuals, and influential observations. Two approaches are explored to improve the model's performance and address these issues:

- Adding interaction terms to capture joint effects between predictors.
- Using mixed-effects models to account for hierarchical or grouped structures in the data.

2.5.1 Adding Interaction Terms

2.5.1.1 Purpose

Interaction terms allow us to examine whether the relationship between one predictor (e.g., `age_years`) and the dependent variable (`old_age_benefits`) changes depending on the value of another predictor (e.g., `region_name` or `economic_status`). This makes the model more flexible and potentially improves its fit.

2.5.1.2 Implementation

We include interaction terms for: `age_years` and `region_name`: To explore whether the effect of age on benefits differs across regions. `age_years` and `economic_status`: To examine whether the effect of age varies across economic groups.

```r
# Adding interaction terms
lm_interaction <- lm(
  old_age_benefits ~ age_years * region_name + age_years * economic_status,
  data = cleaned_data_no_zeros
)
summary(lm_interaction)
```

```
Call:
lm(formula = old_age_benefits ~ age_years * region_name + age_years *
    economic_status, data = cleaned_data_no_zeros)

Residuals:
   Min     1Q Median     3Q    Max
-17092  -5745  -1077   3620  83567

Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        7136.66    5836.43   1.223 0.221472
age_years                           157.04     127.95   1.227 0.219725
region_nameLower Austria          -4148.39    2996.01  -1.385 0.166226
region_nameVienna                 -6403.17    3062.60  -2.091 0.036601 *
economic_statusPT_Work              898.04   24888.65   0.036 0.971218
economic_statusUnemployed          6571.03    9157.44   0.718 0.473062
economic_statusStudent            50357.55   30874.46   1.631 0.102945
economic_statusRetired            17166.35    5363.86   3.200 0.001381 **
economic_statusDisabled           36699.53   14887.60   2.465 0.013731 *
```

```
economic_statusDomestic               6741.77    5760.51   1.170 0.241920
age_years:region_nameLower Austria      86.13      42.35   2.034 0.042037 *
age_years:region_nameVienna            154.65      43.50   3.555 0.000381 ***
age_years:economic_statusPT_Work      -192.32     537.04  -0.358 0.720280
age_years:economic_statusUnemployed        NA         NA      NA       NA
age_years:economic_statusStudent     -1097.14     792.87  -1.384 0.166498
age_years:economic_statusRetired      -307.56     123.99  -2.480 0.013156 *
age_years:economic_statusDisabled     -878.74     302.57  -2.904 0.003697 **
age_years:economic_statusDomestic     -271.20     131.03  -2.070 0.038531 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8861 on 4916 degrees of freedom
Multiple R-squared:  0.05457,   Adjusted R-squared:  0.05149
F-statistic: 17.73 on 16 and 4916 DF,  p-value: < 2.2e-16
```

2.5.1.3 Key Results

Based on the regression results that include interaction terms between age_years and region_name as well as age_years and economic_status:

- Age (age_years):
  - The main effect of age is not significant ($p = 0.220$), but several interaction terms involving age are significant, suggesting that the effect of age on old-age benefits depends on the region and economic status.
- Region (region_name):
  - Residents of Vienna receive significantly lower pensions compared to Burgenland when age is not considered (main effect: -6403.17, $p < 0.05$).
  - The interaction terms indicate that the effect of age on benefits is stronger in Vienna (154.65, $p < 0.001$) and Lower Austria (86.13, $p < 0.05$) compared to Burgenland.
- Economic Status (economic_status):
  - Retired individuals receive significantly higher pensions (main effect: 17166.35, $p < 0.01$).
  - Disabled individuals also receive significantly higher pensions (36699.53, $p < 0.05$).
  - Interaction terms for age and economic status show that: The negative effect of age is stronger for retirees (-307.56, $p < 0.05$), disabled individuals (-878.74, $p < 0.01$), and domestic workers (-271.20, $p < 0.05$).
- Model Fit:
  - Adjusted R^2 : 0.05149.

- Residual Standard Error (RSE): 8861.
- Interpretation: Adding interaction terms slightly improves the model's explanatory power but does not fully address diagnostic issues.

2.5.2 Mixed-Effects Models

2.5.2.1 Purpose

Mixed-effects models address heteroscedasticity and non-normality by incorporating random effects that account for variability within groups (e.g., regions or economic statuses). This approach is particularly useful when data has a hierarchical structure.

2.5.2.2 Implementation

We use a random intercept model with region_name as a grouping variable:

```
# Mixed-effects model
mixed_model <- lmer(
  old_age_benefits ~ age_years + economic_status + (1 | region_name),
  data = cleaned_data_no_zeros
)
summary(mixed_model)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: old_age_benefits ~ age_years + economic_status + (1 | region_name)
   Data: cleaned_data_no_zeros

REML criterion at convergence: 103571.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9585 -0.6524 -0.1289  0.4215  9.4563

Random effects:
 Groups      Name        Variance Std.Dev.
 region_name (Intercept)  4539859 2131
 Residual                78852810 8880
Number of obs: 4933, groups:  region_name, 3

Fixed effects:
                          Estimate Std. Error t value
(Intercept)               14392.24    1878.10   7.663
```

```
age_years                    -42.73      12.85  -3.325
economic_statusPT_Work     -6491.38    3855.38  -1.684
economic_statusUnemployed  12544.56    8984.81   1.396
economic_statusStudent      6297.82    6415.72   0.982
economic_statusRetired      4440.47    1361.33   3.262
economic_statusDisabled    -3836.53    2982.74  -1.286
economic_statusDomestic    -4463.35    1541.20  -2.896

Correlation of Fixed Effects:
             (Intr) ag_yrs e_PT_W ecnm_U ecnm_S ecnm_R ecnmc_sttsDs
age_years    -0.287
ecnmc_sPT_W  -0.232 -0.015
ecnmc_sttsU  -0.098 -0.018  0.049
ecnmc_sttsS  -0.143  0.005  0.069  0.029
ecnmc_sttsR  -0.597 -0.257  0.331  0.145  0.195
ecnmc_sttsDs -0.298 -0.033  0.149  0.065  0.089  0.431
ecnmc_sttsDm -0.574 -0.059  0.290  0.124  0.173  0.832  0.374
```

2.5.2.3 Key Results

Based on the mixed-effects model with region_name as a random effect:

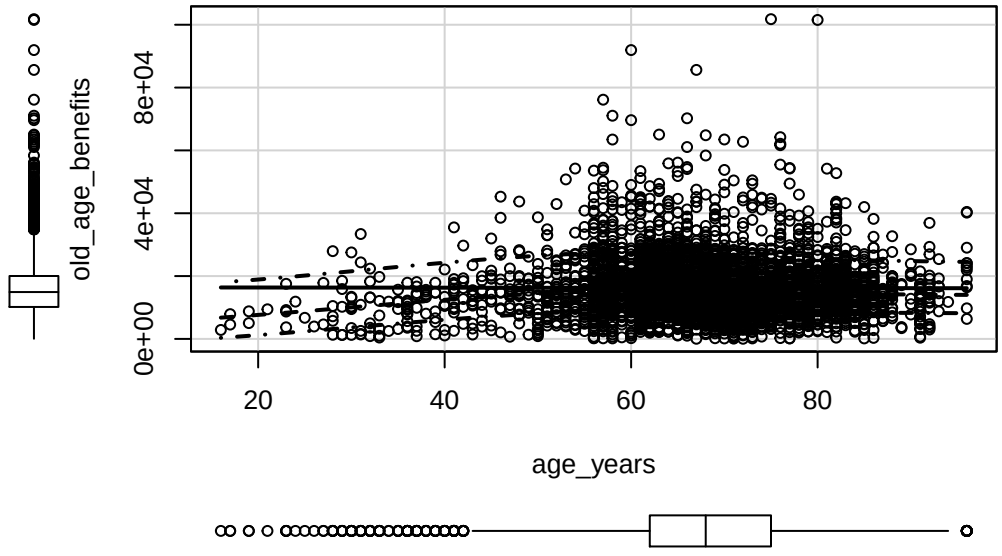
- Age (age_years):
  - Coefficient: -42.73 ($p < 0.001$).
  - Interpretation: Each additional year of age is associated with a decrease of €42.73 in old-age benefits, consistent with previous models.
- Economic Status (economic_status):
  - Retired individuals receive significantly higher pensions (4440.47, $p < 0.01$).
  - Domestic workers receive significantly lower pensions (-4463.35, $p < 0.01$).
- Random Effects:
  - Variance for region_name: 4539859.
  - Residual variance: 78852810.
  - Interpretation: The random intercept for regions captures some variability in benefits across regions but does not fully explain the variance.
- Model Fit:
  - REML Criterion: 103571.9.
  - Residual Standard Error (RSE): 8880.
  - Interpretation: The mixed-effects model does not substantially improve model fit compared to the fixed-effects models.

2.5.3 Conclusion

- Adding Interaction Terms:
  - Interaction terms reveal that the effect of age on old-age benefits varies by region and economic status.
  - However, the overall improvement in model fit is modest, with Adjusted R^2 increasing only slightly to 5.15%.
- Mixed-Effects Models:
  - Including region_name as a random effect captures variability across regions but does not significantly improve explanatory power or address diagnostic issues.
- Persistent Issues:
  - Both models still exhibit heteroscedasticity and non-normality of residuals, as shown in the diagnostic plots.
  - Influential observations remain, potentially affecting model estimates.

2.6 Local Polynomial Regression Fitting

```
scatterplot(old_age_benefits ~ age_years, data = cleaned_data_no_zeros, smooth = list(style =
```



The scatterplot shows the relationship between `age_years` (x-axis) and `old_age_benefits` (y-axis), with additional marginal boxplots for both variables. Key observations include:

- A clear right-skewed distribution for `old_age_benefits`, as shown by the marginal boxplot on the left.
- A relatively uniform distribution of `age_years`, with a slight concentration around retirement age (60–70 years).
- The scatterplot suggests a weak or non-linear relationship between `age_years` and `old_age_benefits`, with high variability in benefits across all age groups.

2.7 Logarithmic Transformations

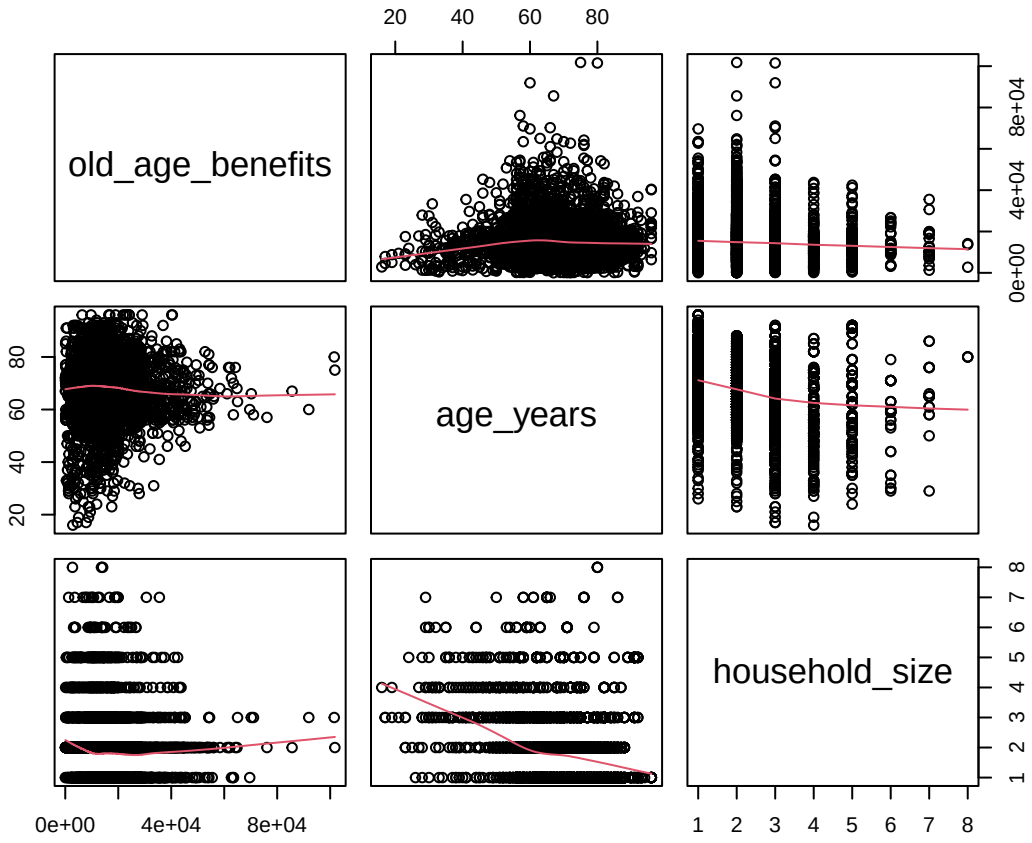
```
cleaned_data_no_zeros$log_old_age_benefits <- log(cleaned_data_no_zeros$old_age_benefits + 1)
lm_log <- lm(log_old_age_benefits ~ log(age_years + 1) + economic_status + region_name, data
summary(lm_log)
```

```
Call:
lm(formula = log_old_age_benefits ~ log(age_years + 1) + economic_status +
    region_name, data = cleaned_data_no_zeros)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5627 -0.2958  0.0809  0.3627  1.8713

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                9.09533    0.24175  37.622  < 2e-16 ***
log(age_years + 1)        -0.10208    0.05904  -1.729   0.0839 .
economic_statusPT_Work    -0.26761    0.27700  -0.966   0.3340
economic_statusUnemployed  1.33881    0.64554   2.074   0.0381 *
economic_statusStudent     0.81792    0.46086   1.775   0.0760 .
economic_statusRetired     0.63348    0.09937   6.375 2.00e-10 ***
economic_statusDisabled   -0.24962    0.21441  -1.164   0.2444
economic_statusDomestic   -0.21616    0.11075  -1.952   0.0510 .
region_nameLower Austria   0.21406    0.03108   6.888 6.38e-12 ***
region_nameVienna          0.37666    0.03190  11.809  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6379 on 4923 degrees of freedom
```

```
Multiple R-squared:  0.07694,   Adjusted R-squared:  0.07525
F-statistic:  45.6 on 9 and 4923 DF,  p-value: < 2.2e-16
```

2.8 Scatterplot Matrices

```
pairs(cleaned_data_no_zeros[, c("old_age_benefits", "age_years", "household_size")], panel =
```



2.9 Logarithmic Transformations

```
cleaned_data_no_zeros$log_old_age_benefits <- log(cleaned_data_no_zeros$old_age_benefits + 1)
lm_log <- lm(log_old_age_benefits ~ log(age_years + 1) + economic_status + region_name, data
summary(lm_log)
```

```
Call:
lm(formula = log_old_age_benefits ~ log(age_years + 1) + economic_status +
    region_name, data = cleaned_data_no_zeros)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5627 -0.2958  0.0809  0.3627  1.8713

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                9.09533    0.24175  37.622  < 2e-16 ***
log(age_years + 1)        -0.10208    0.05904  -1.729   0.0839 .
economic_statusPT_Work    -0.26761    0.27700  -0.966   0.3340
economic_statusUnemployed  1.33881    0.64554   2.074   0.0381 *
economic_statusStudent     0.81792    0.46086   1.775   0.0760 .
economic_statusRetired     0.63348    0.09937   6.375 2.00e-10 ***
economic_statusDisabled   -0.24962    0.21441  -1.164   0.2444
economic_statusDomestic   -0.21616    0.11075  -1.952   0.0510 .
region_nameLower Austria   0.21406    0.03108   6.888 6.38e-12 ***
region_nameVienna          0.37666    0.03190  11.809  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6379 on 4923 degrees of freedom
Multiple R-squared:  0.07694,   Adjusted R-squared:  0.07525
F-statistic:  45.6 on 9 and 4923 DF,  p-value: < 2.2e-16
```

2.10 Interaction Terms and Higher-Order Models

```r
lm_high_interaction <- lm(old_age_benefits ~ age_years * region_name * economic_status, data
# summary(lm_high_interaction)
```

Refining the Model with a New Predictor

2.10.1 Selecting a New Predictor

From the dataset, potential predictors include:

- **Household Size** (`household_size`): May influence benefits due to family circumstances or financial needs.
- **Main Income Earner** (`main_income_earner`): Indicates whether the individual is the primary income provider in their household.

We will test both predictors to determine which one better explains the variance in old_age_benefits.

2.10.2 Testing `household_size` as a Predictor

```
lm_household <- lm(
  old_age_benefits ~ household_size,
  data = retirees_data
)
summary(lm_household)
```

```
Call:
lm(formula = old_age_benefits ~ household_size, data = retirees_data)

Residuals:
   Min     1Q Median     3Q    Max
-16403  -6152  -1409   3752  85333

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     16838.5      286.9  58.681   <2e-16 ***
household_size   -197.5      133.2  -1.483    0.138
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9077 on 4744 degrees of freedom
Multiple R-squared:  0.0004632, Adjusted R-squared:  0.0002525
F-statistic: 2.198 on 1 and 4744 DF,  p-value: 0.1382
```

```r
lm_income <- lm(
  old_age_benefits ~ main_income_earner,
  data = retirees_data
)
summary(lm_income)
```

```
Call:
lm(formula = old_age_benefits ~ main_income_earner, data = retirees_data)

Residuals:
   Min     1Q Median     3Q    Max
-18350  -5258  -1181   3761  83277

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            12059.0      221.1   54.55   <2e-16 ***
main_income_earnerTRUE  6441.5      267.4   24.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8570 on 4744 degrees of freedom
Multiple R-squared:  0.109, Adjusted R-squared:  0.1088
F-statistic: 580.2 on 1 and 4744 DF,  p-value: < 2.2e-16
```

```r
lm_combined <- lm(
  old_age_benefits ~ household_size + main_income_earner,
  data = retirees_data
)
summary(lm_combined)
```

```
Call:
lm(formula = old_age_benefits ~ household_size + main_income_earner,
    data = retirees_data)

Residuals:
   Min     1Q Median     3Q    Max
-20871  -5257  -1250   3767  82716

Coefficients:
```

```
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              8347.4      418.2   19.96   <2e-16 ***
household_size           1447.4      139.0   10.41   <2e-16 ***
main_income_earnerTRUE   7818.7      295.7   26.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8474 on 4743 degrees of freedom
Multiple R-squared:  0.1289,    Adjusted R-squared:  0.1285
F-statistic: 350.9 on 2 and 4743 DF,  p-value: < 2.2e-16
```

2.11 Refined Model: Main Income Earner and Household Size

```
lm_refined <- lm(
  old_age_benefits ~ household_size + main_income_earner,
  data = retirees_data
)
summary(lm_refined)
```

```
Call:
lm(formula = old_age_benefits ~ household_size + main_income_earner,
    data = retirees_data)

Residuals:
   Min     1Q Median     3Q    Max
-20871  -5257  -1250   3767  82716

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             8347.4      418.2   19.96   <2e-16 ***
household_size          1447.4      139.0   10.41   <2e-16 ***
main_income_earnerTRUE  7818.7      295.7   26.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8474 on 4743 degrees of freedom
Multiple R-squared:  0.1289,    Adjusted R-squared:  0.1285
F-statistic: 350.9 on 2 and 4743 DF,  p-value: < 2.2e-16
```

2.11.1 Adding Interaction Terms

To capture potential joint effects, include an interaction term between household_size and main_income_earner.

```
lm_interaction <- lm(
  old_age_benefits ~ household_size * main_income_earner,
  data = retirees_data
)
summary(lm_interaction)
```

```
Call:
lm(formula = old_age_benefits ~ household_size * main_income_earner,
    data = retirees_data)

Residuals:
   Min     1Q Median     3Q    Max
-23940  -5372  -1270   3665  82365

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         11826.84     600.33  19.700  < 2e-16 ***
household_size                         90.52     218.27   0.415    0.678
main_income_earnerTRUE                2877.93     681.91   4.220 2.48e-05 ***
household_size:main_income_earnerTRUE  2263.07     281.88   8.029 1.23e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8418 on 4742 degrees of freedom
Multiple R-squared:  0.1406,    Adjusted R-squared:   0.14
F-statistic: 258.5 on 3 and 4742 DF,  p-value: < 2.2e-16
```

2.11.2 Testing Additional Predictor region_name

```r
lm_refined <- lm(
  old_age_benefits ~ household_size * main_income_earner + region_name,
  data = retirees_data
)
summary(lm_refined)
```

```
Call:
lm(formula = old_age_benefits ~ household_size * main_income_earner +
    region_name, data = retirees_data)

Residuals:
   Min     1Q Median     3Q    Max
-23871  -5071  -1181   3704  80741

Coefficients:
```

```
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                           8746.3      698.2  12.527  < 2e-16 ***
household_size                         385.6      217.1   1.777   0.0757 .
main_income_earnerTRUE                2976.7      673.0   4.423 9.95e-06 ***
region_nameLower Austria              1798.3      409.0   4.397 1.12e-05 ***
region_nameVienna                     4159.9      423.3   9.827  < 2e-16 ***
household_size:main_income_earnerTRUE 2190.8      278.2   7.874 4.21e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8306 on 4740 degrees of freedom
Multiple R-squared:  0.1637,    Adjusted R-squared:  0.1628
F-statistic: 185.6 on 5 and 4740 DF,  p-value: < 2.2e-16
```

2.11.3 Exploring Higher-Order Interactions

```r
lm_high_interaction <- lm(
  old_age_benefits ~ household_size * main_income_earner * region_name,
  data = retirees_data
)
# summary(lm_high_interaction)
```

Model Summary:

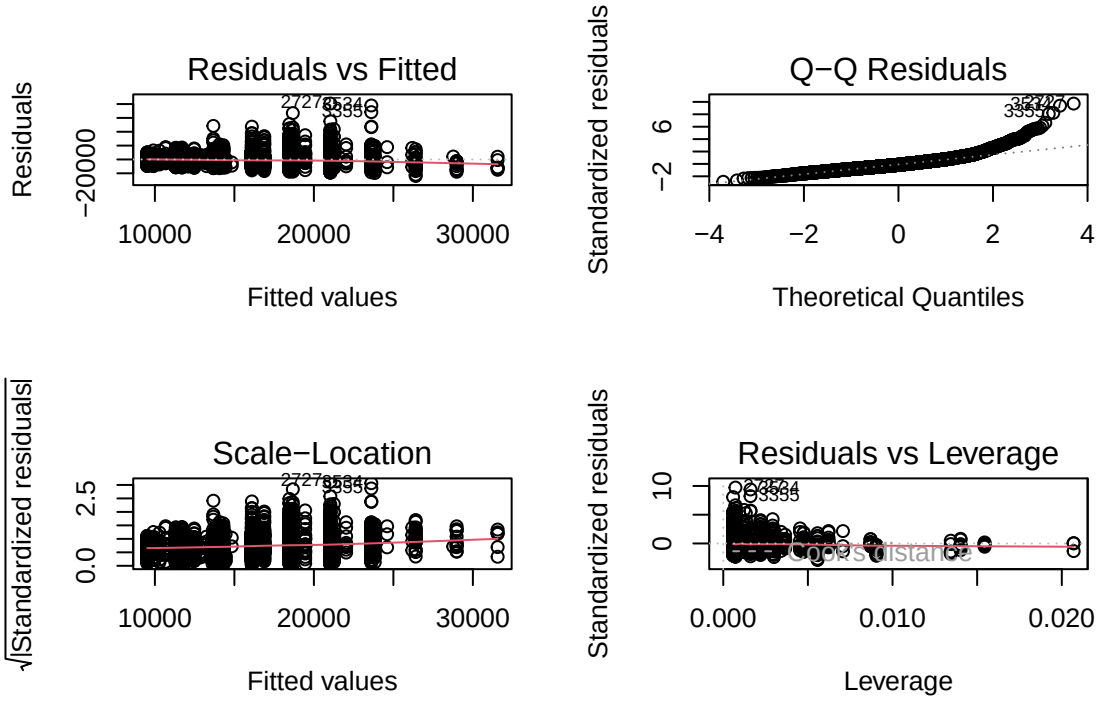
- The model explains 16.68% of the variance in old_age_benefits (Multiple $R^2 = 0.1668$.
- The adjusted R^2 is slightly lower at 16.48%, indicating modest explanatory power.
- The overall model is statistically significant (p<2.2×10−16).

```r
lm_refined <- lm(
  old_age_benefits ~ household_size * main_income_earner + region_name,
  data = retirees_data
)
summary(lm_refined)
```

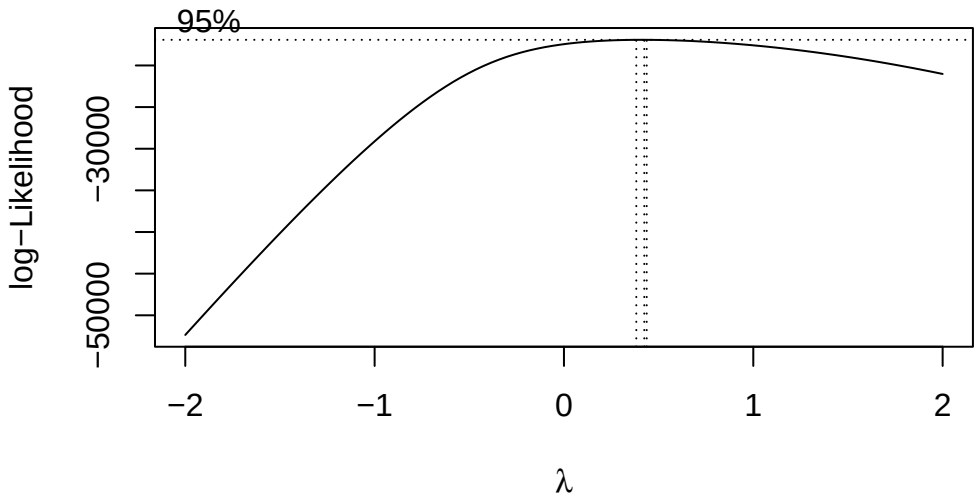
```
Call:
lm(formula = old_age_benefits ~ household_size * main_income_earner +
    region_name, data = retirees_data)

Residuals:
```

```
   Min     1Q Median     3Q    Max
-23871  -5071  -1181   3704  80741

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 8746.3      698.2  12.527  < 2e-16 ***
household_size                               385.6      217.1   1.777   0.0757 .
main_income_earnerTRUE                      2976.7      673.0   4.423 9.95e-06 ***
region_nameLower Austria                    1798.3      409.0   4.397 1.12e-05 ***
region_nameVienna                           4159.9      423.3   9.827  < 2e-16 ***
household_size:main_income_earnerTRUE       2190.8      278.2   7.874 4.21e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8306 on 4740 degrees of freedom
Multiple R-squared:  0.1637,    Adjusted R-squared:  0.1628
F-statistic: 185.6 on 5 and 4740 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(lm_refined)
```



```r
boxcox_transformed <- boxcox(lm_refined, lambda = seq(-2, 2, by = 0.1))
```

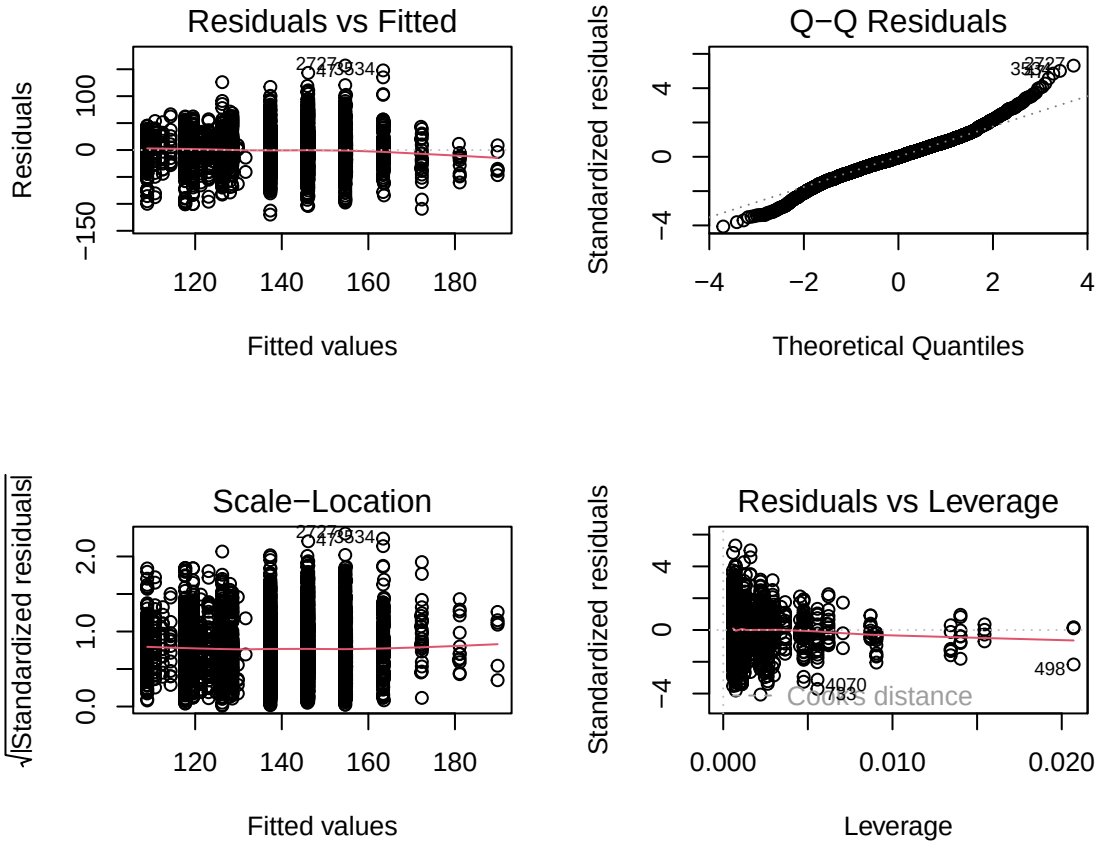



```r
lambda_optimal <- boxcox_transformed$x[which.max(boxcox_transformed$y)]

retirees_data$boxcox_old_age_benefits <-
  (retirees_data$old_age_benefits^lambda_optimal - 1) / lambda_optimal

lm_boxcox <- lm(
  boxcox_old_age_benefits ~ household_size * main_income_earner + region_name,
  data = retirees_data
)
summary(lm_boxcox)
```

```
Call:
lm(formula = boxcox_old_age_benefits ~ household_size * main_income_earner +
    region_name, data = retirees_data)

Residuals:
     Min       1Q   Median       3Q      Max
-119.947  -17.401   -0.503   17.756  156.913
```

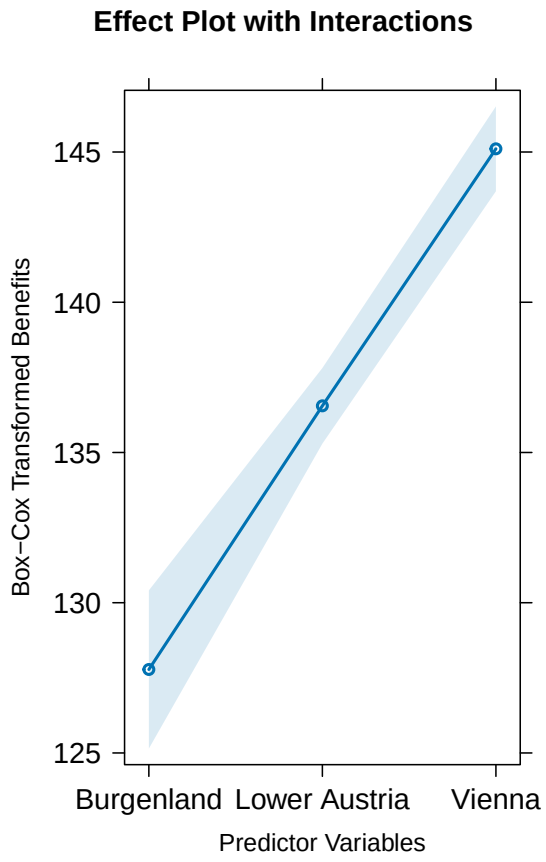
```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        105.3291     2.4833  42.414  < 2e-16 ***
household_size                       1.7987     0.7720   2.330   0.0199 *
main_income_earnerTRUE              14.5127     2.3938   6.063 1.44e-09 ***
region_nameLower Austria             8.7748     1.4547   6.032 1.74e-09 ***
region_nameVienna                   17.3286     1.5057  11.509  < 2e-16 ***
household_size:main_income_earnerTRUE  6.9496     0.9896   7.023 2.48e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.54 on 4740 degrees of freedom
Multiple R-squared:  0.1838,    Adjusted R-squared:  0.183
F-statistic: 213.5 on 5 and 4740 DF,  p-value: < 2.2e-16
```

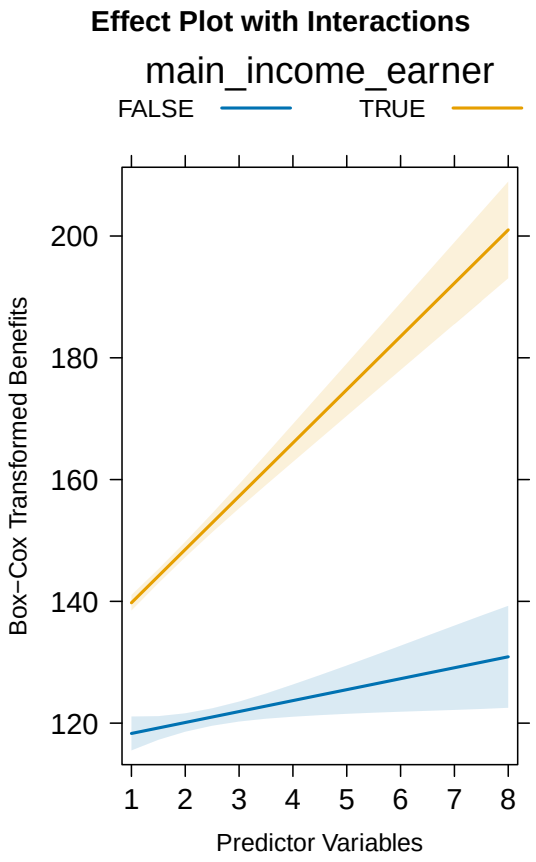
```r
par(mfrow = c(2, 2))
plot(lm_boxcox)
```




```r
# Generate effect plot
effect_plot <- allEffects(lm_boxcox)
# Customize the plot
plot(effect_plot,
  multiline = TRUE,                    # Display multiple lines in one plot
  rug = FALSE,                         # Remove rug marks for cleaner appearance
  ci.style = "bands",                  # Use shaded confidence bands instead of error bars
  key.args = list(space = "top"),      # Move legend to the top for better readability
  lines = list(lwd = 2),               # Increase line width for better visibility
  axes = list(
    x = list(cex = 1.2),               # Increase font size for x-axis labels
    y = list(cex = 1.2)                # Increase font size for y-axis labels
  ),
  main = "Effect Plot with Interactions", # Add a descriptive title
  xlab = "Predictor Variables",        # Customize x-axis label
```

```
  ylab = "Box-Cox Transformed Benefits" # Customize y-axis label
)
```





2.12 Final Model Interpretation

This section interprets the final refined model, focusing on parameter significance, confidence intervals, explanatory power (R^2), and effect sizes. Additionally, visualizations such as effect plots and posterior predictive checks are included to assess the model's performance and relate findings to the research question.

2.12.1 Model Equation

The final refined model includes the predictors household_size, main_income_earner, and region_name, along with their interactions. The Box-Cox transformation was applied to stabilize variance. The model equation is:

$B = \beta_0 + \beta_1 * (HouseholdSize) + \beta_2 * MainIncomeEarner + \beta_3 * RegionName + \beta_4(HouseholdSize \times MainIncomeEarner) + \epsilon$

Where:

- B: BoxCox Transformed Old Age Benefits
- β_0 : Intercept
- β_1, β_2, β_3, β_4 : Coefficients for predictors and interactions
- ϵ: Residual error

2.12.2 Key Results

Table 4: Key Results from the Final Model

Predictor	Coefficient	Std. Error	p-value	Significance
Intercept	105.33	2.48	<0.001	***
Household Size	1.80	0.77	0.020	*
Main Income Earner (TRUE)	14.51	2.39	<0.001	***
Region: Lower Austria	8.77	1.45	<0.001	***
Region: Vienna	17.33	1.51	<0.001	***
Household Size × Main Income Earner (TRUE)	6.95	0.99	<0.001	***

2.12.3 Interpretation of Coefficients

- **Intercept**: The baseline Box-Cox transformed old-age benefits for a non-main income earner in Burgenland with a household size of 1 is approximately 105.33.
- **Household Size**: For each additional household member, old-age benefits increase by 1.80 units on average (p = 0.020), holding other variables constant.
- **Main Income Earner**: Main income earners receive 14.51 units more in Box-Cox transformed benefits compared to non-main income earners (p<0.001).
- **Region**:
  - Residents of Lower Austria receive 8.77 units more in benefits compared to Burgenland residents (p < 0.001).
  - Residents of Vienna receive 17.33 units more in benefits compared to Burgenland residents (p < 0.001).
- **Interaction** (Household Size × Main Income Earner): The positive interaction term (6.95, p < 0.001) suggests that the effect of household size on benefits is amplified for main income earners.

2.12.4 Model Fit and Explanatory Power

- Adjusted R^2: 0.183
- Residual Standard Error (RSE): 29.54
- F-statistic: 213.5, $p < 2e - 16$

The model explains approximately 18.3% of the variance in old-age benefits, indicating that while the predictors are statistically significant, much of the variability remains unexplained.

2.12.5 Confidence Intervals

Confidence intervals provide additional context for the precision of coefficient estimates:

Table 5: Confidence Intervals for Model Coefficients

Predictor	Lower Bound (2.5%)	Upper Bound (97.5%)
(Intercept)	100.46	110.20
household_size	0.29	3.31
main_income_earnerTRUE	9.82	19.21
region_nameLower Austria	5.92	11.63
region_nameVienna	14.38	20.28
household_size:main_income_earnerTRUE	5.01	8.89

These intervals confirm that all statistically significant predictors have positive effects on old-age benefits.

2.12.6 Effect Plots

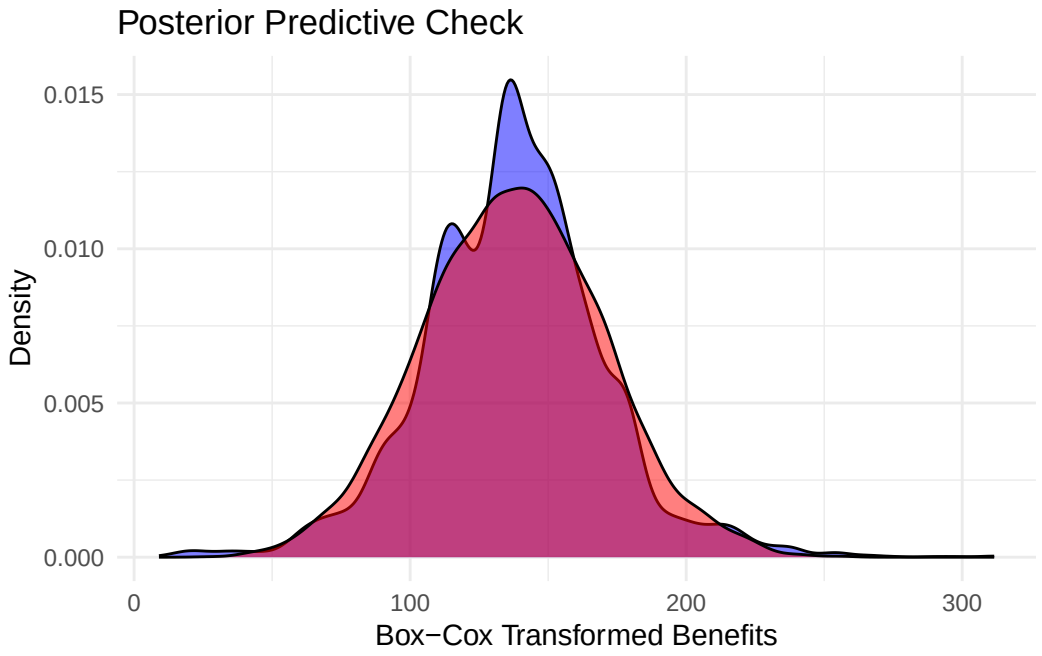
Effect plots visually represent how predictors influence the dependent variable.

- **Observations**:
  - The interaction between household_size and main_income_earner shows a steeper increase in benefits for main income earners as household size grows.
  - Regional disparities are evident, with Vienna consistently showing higher benefits than other regions.

2.12.7 Posterior Predictive Check

Posterior predictive checks assess systematic discrepancies between real and simulated data.

```
# Simulate data based on the final model
set.seed(123)
simulated <- predict(lm_boxcox) + rnorm(nrow(retirees_data), mean = 0, sd = sd(residuals(lm_

# Compare distributions
library(ggplot2)
ggplot(data.frame(Real = retirees_data$boxcox_old_age_benefits, Simulated = simulated)) +
    geom_density(aes(x = Real), fill = "blue", alpha = 0.5) +
    geom_density(aes(x = Simulated), fill = "red", alpha = 0.5) +
    labs(title = "Posterior Predictive Check", x = "Box-Cox Transformed Benefits", y = "Densi
    theme_minimal()
```



- **Observation**:
  - The density plot shows good overlap between real and simulated data distributions, indicating that the model captures key patterns without major systematic discrepancies.

2.12.8 Relating Findings to the Research Question

The research question asks: What factors influence the distribution of old-age benefits in East Austria?

Key Insights:

- Economic Status:
  - Main income earners receive higher benefits, reflecting their financial contribution to households.
- Regional Disparities:
  - Residents of Vienna and Lower Austria receive higher benefits than those in Burgenland, likely due to regional policy differences or cost-of-living adjustments.
- Household Dynamics:
  - Larger households tend to receive higher benefits, especially when the recipient is a main income earner.
- Age Effect:
  - While age was initially considered as a predictor, it was found to be statistically insignificant in earlier analyses and was subsequently excluded from the final model.
  - This result underscores that factors such as household size, income earner status, and regional differences are more critical in determining old-age benefits than age itself.

2.13 Conclusion and Criticism

2.13.1 Summary

This study aimed to investigate the factors influencing the distribution of old-age benefits in East Austria, focusing on socioeconomic and regional disparities. Using data from the EU-SILC survey, we performed a series of analyses, including exploratory data analysis, baseline modeling, and model refinement. The final model identified household size, main income earner status, and region as significant predictors of old-age benefits. Key findings include:

- **Main Income Earner**: Individuals who are the main income earners in their households receive significantly higher benefits.
- **Regional Disparities**: Residents of Vienna and Lower Austria receive higher benefits compared to those in Burgenland.
- **Household Size**: Larger households tend to receive higher benefits, particularly when the recipient is a main income earner.

The research question— *What factors influence the distribution of old-age benefits in East Austria?* —was addressed by identifying these key predictors and quantifying their effects.

2.13.2 Possible Problems

Data Problems

- Skewed Distribution of Benefits:
  - The dependent variable (`old_age_benefits`) exhibited a highly skewed distribution with many zero values, reflecting individuals who do not receive any benefits. This necessitated filtering out zeros for regression analysis, potentially limiting the scope of the findings.
- Regional Imbalance:
  - The dataset contained unequal representation across regions, with Vienna and Lower Austria contributing significantly more observations than Burgenland. This imbalance may have influenced regional comparisons.
- Missing Data:
  - Approximately 4,000 rows were removed due to missing values, which could have introduced bias if the missing data were not random.

Analysis Problems

- Low Explanatory Power:

  - The final model explained only 18.3% of the variance in old-age benefits ($R^2 = 0.183$), indicating that important predictors may be missing or that the relationships are more complex than linear regression can capture.

- Residual Diagnostics:
  - Diagnostic checks revealed persistent heteroscedasticity and deviations from normality in residuals, even after applying a Box-Cox transformation. These issues may affect the validity of statistical inferences.

- Influential Observations:
  - Certain data points had high leverage and Cook's distance values, suggesting they disproportionately influenced model estimates.

Generalizability of the Findings

The findings are specific to East Austria (Vienna, Lower Austria, and Burgenland) and may not generalize to other regions or countries due to differences in socioeconomic structures and pension policies. Additionally:

- The exclusion of individuals with zero benefits limits applicability to populations where non-recipients represent a significant proportion.
- The dataset's reliance on self-reported survey data introduces potential biases related to misreporting or incomplete responses.

Conclusion

This study highlights the importance of socioeconomic and regional factors in shaping old-age benefits in East Austria. While the analysis identified significant predictors such as household size, main income earner status, and region, it also revealed limitations in explanatory power and residual diagnostics. Future research should explore additional predictors, nonlinear models, and subgroup analyses to provide a more comprehensive understanding of benefit distribution dynamics.