Case Study: Modeling Self-Perceived Health and Longevity with SHARE Dataset

Калоян Тотев
Михаил Ангелов
Сергей Филипов

Outline

Introduction

Dataset

Aim

Charts

Methodology

Results

Conclusions

Appendix

a research infrastructure for studying the effects of health, social, economic and environmental policies over the life-course of European citizens and beyond.

9 waves from 2004 to 2021/2022

In Bulgaria – from the 7th wave (2018–)

SHARE

SURVEY OF HEALTH, AGEING AND RETIREMENT IN EUROPE

Dataset

46733
observations

3540
variables

26 European
countries + Israel

Aim

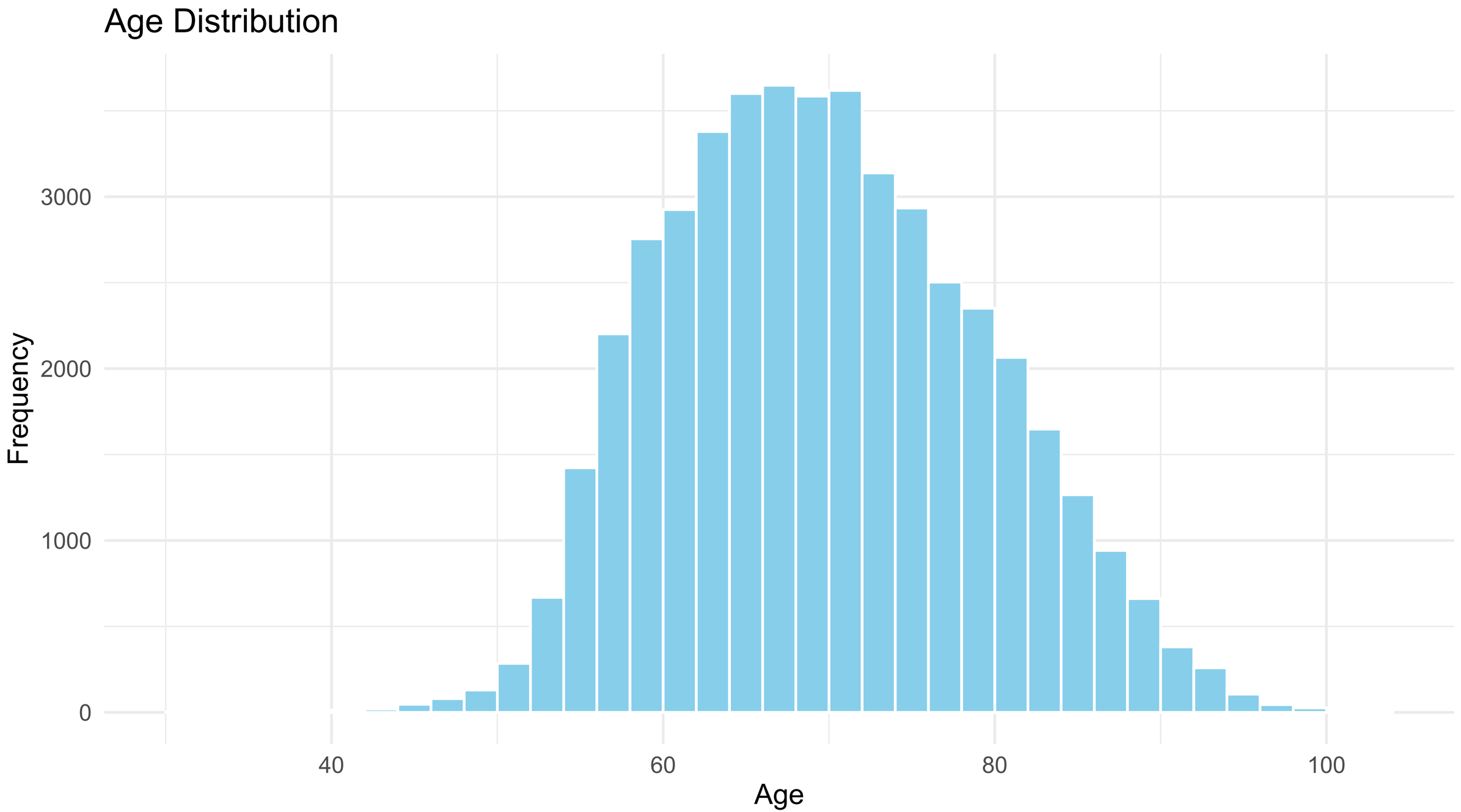
Self-Perceived Health

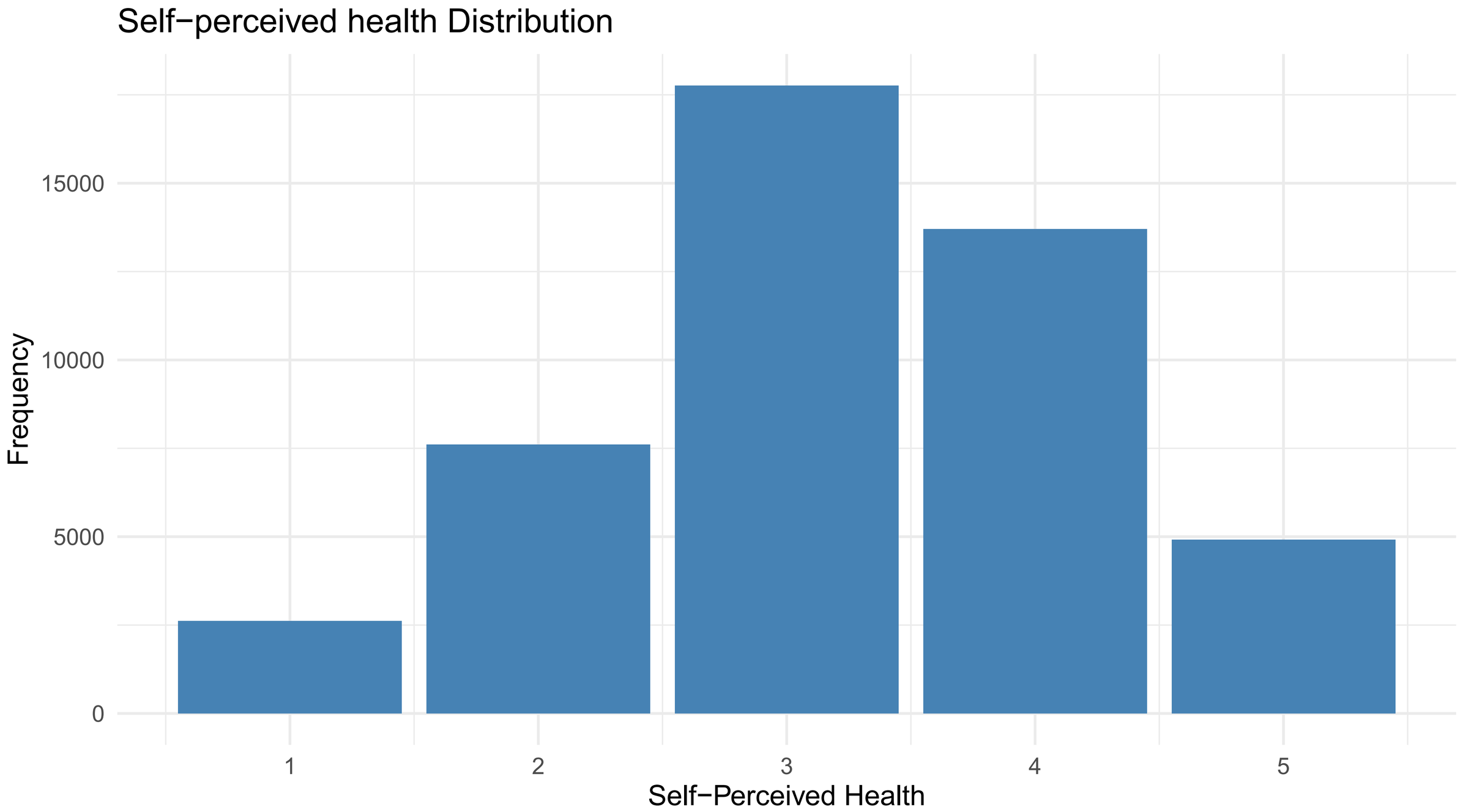
how individuals rate their health

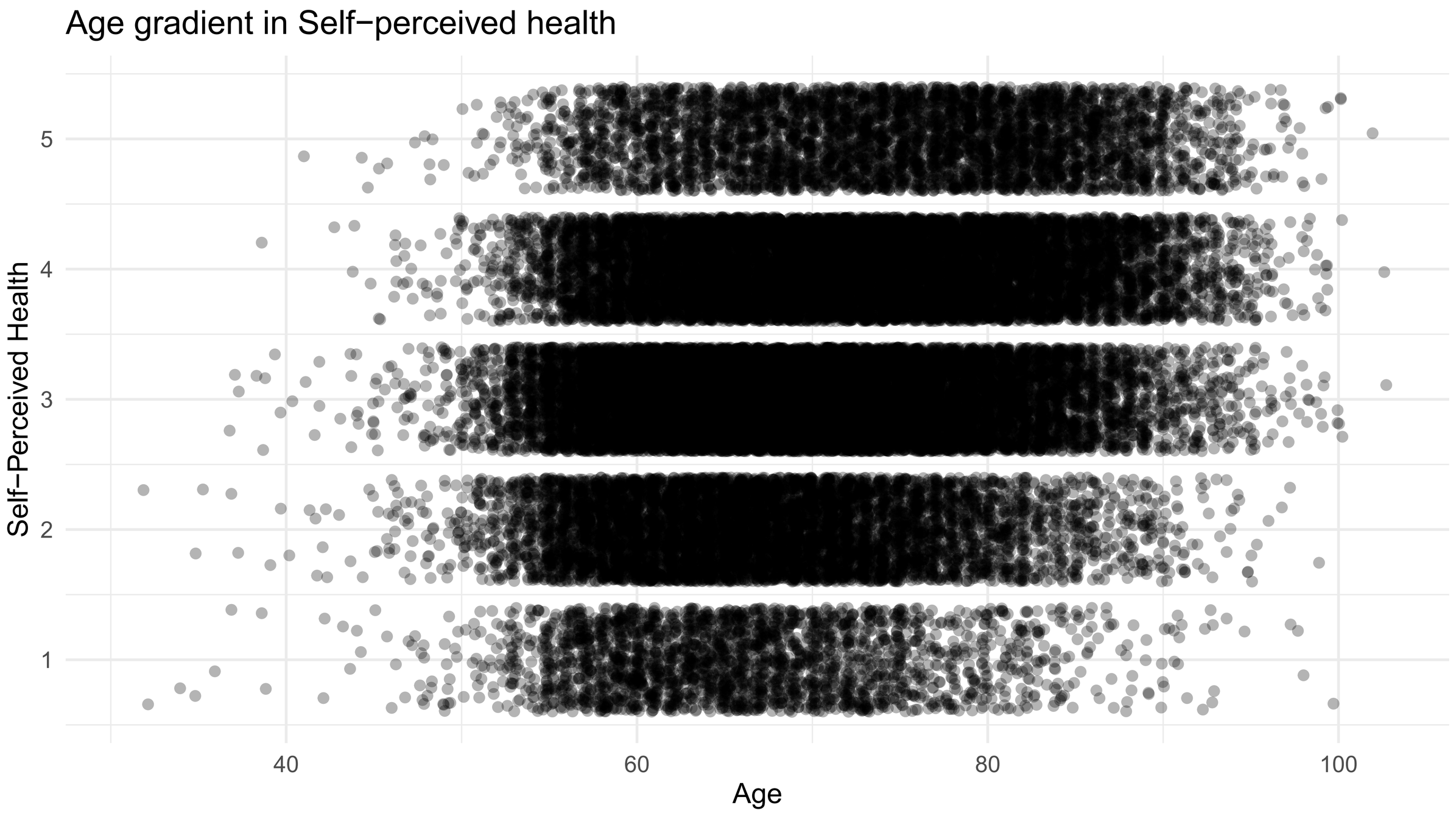
Age

used as a proxy for longevity

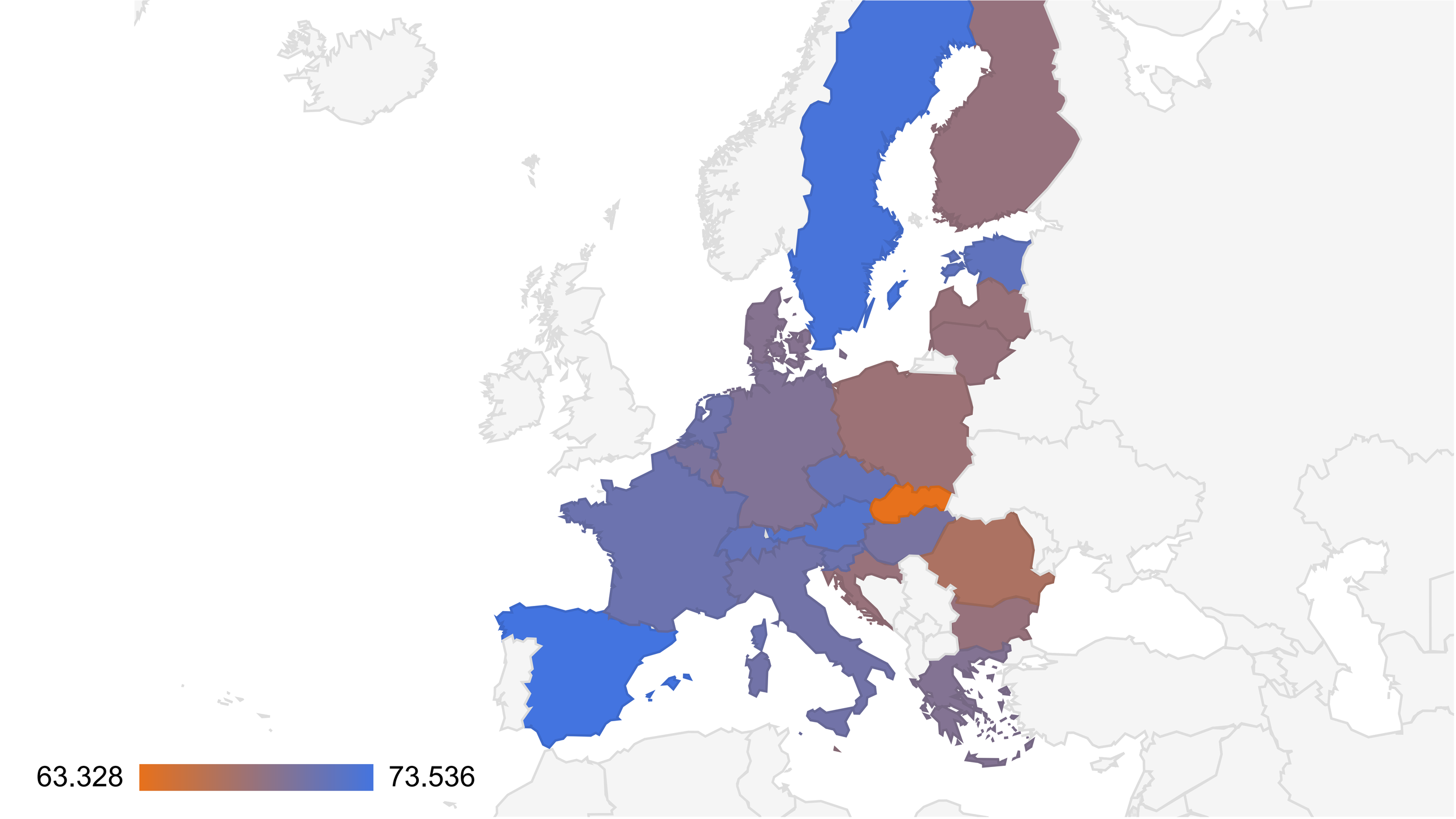
Charts

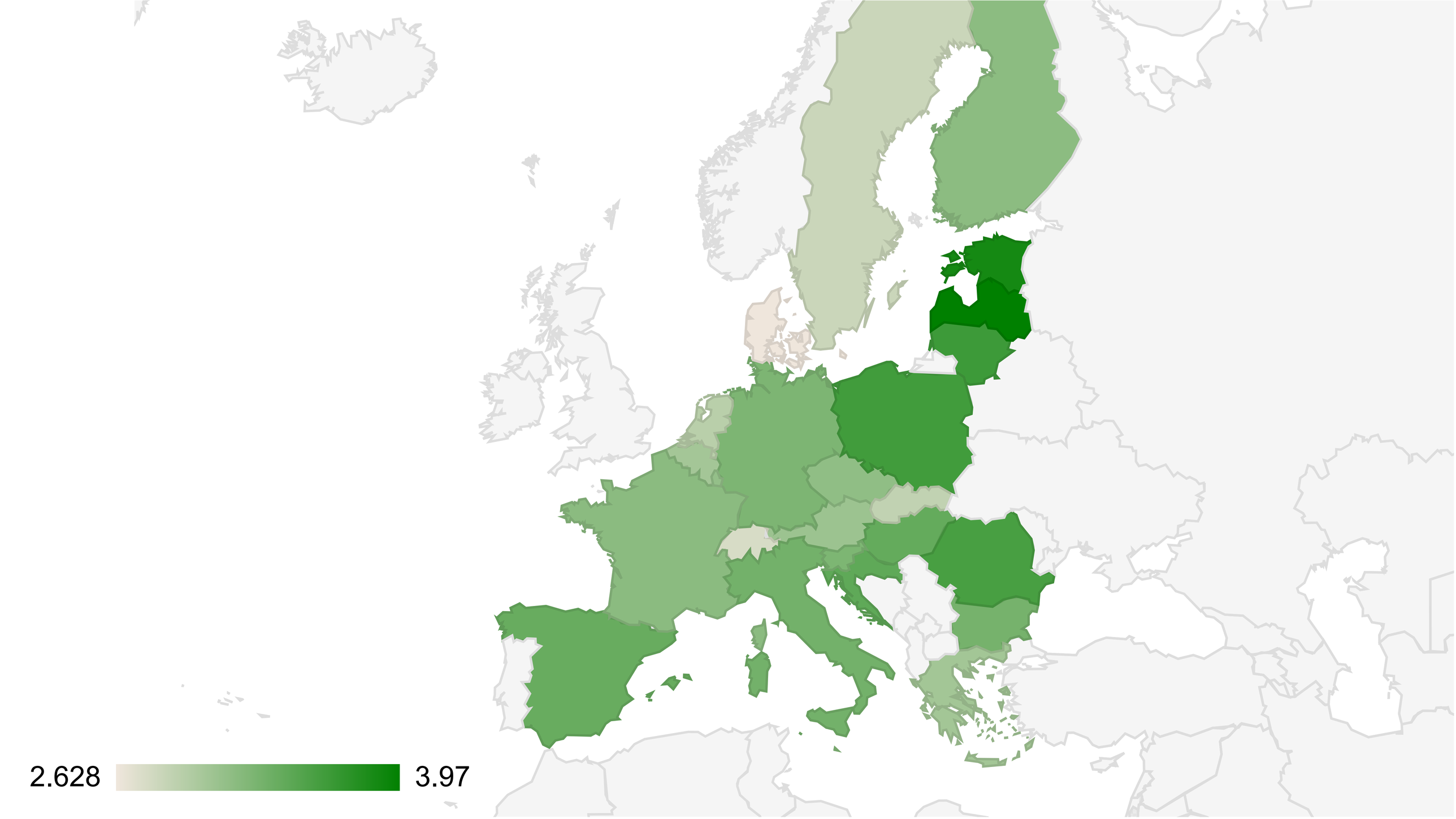

Age Distribution
Frequency
3000
2000
1000
0
40
60
80
100
Age

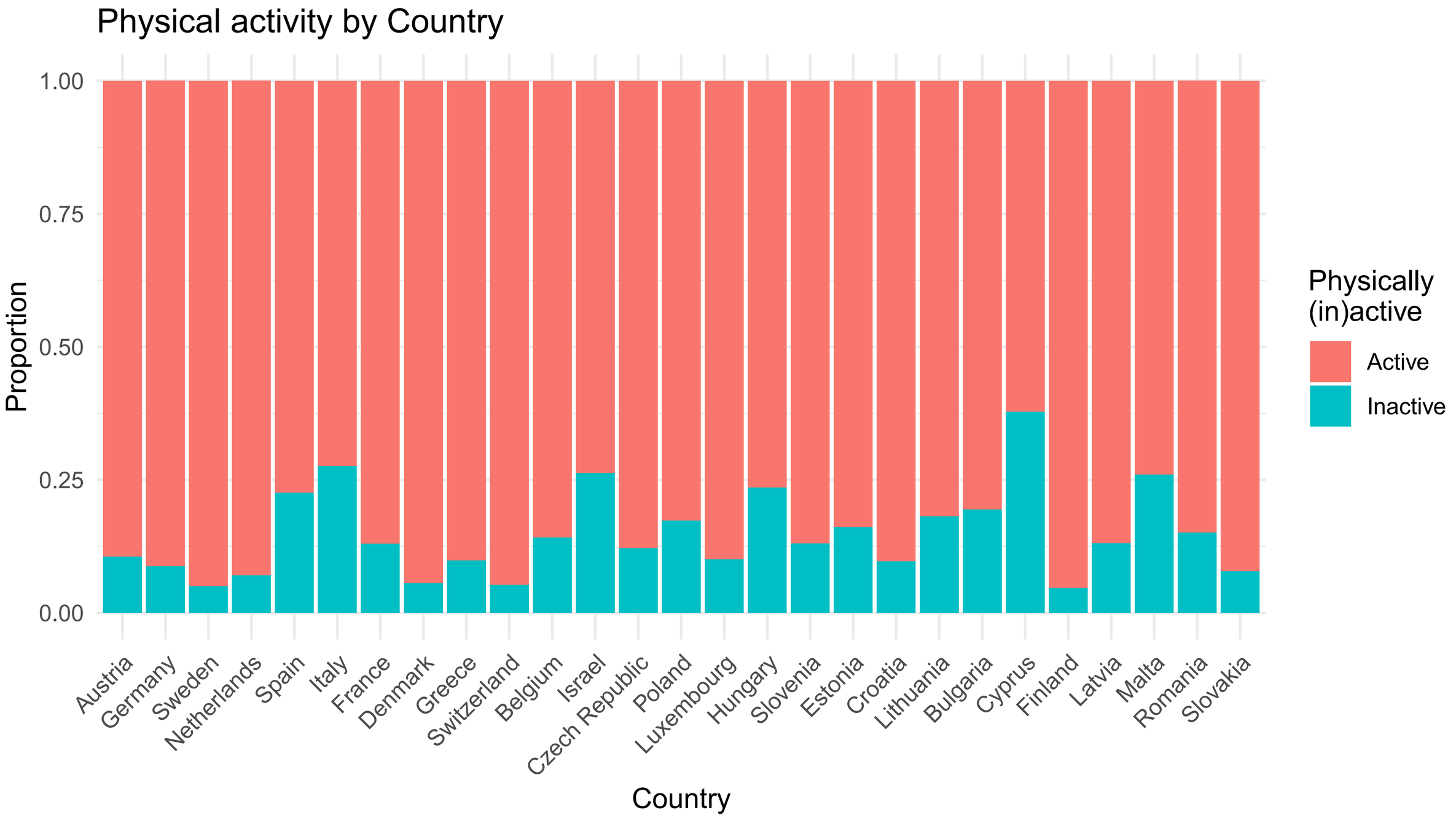

Self−perceived health Distribution
Frequency
15000
10000
5000
0
1
2
3
4
5
Self−Perceived Health

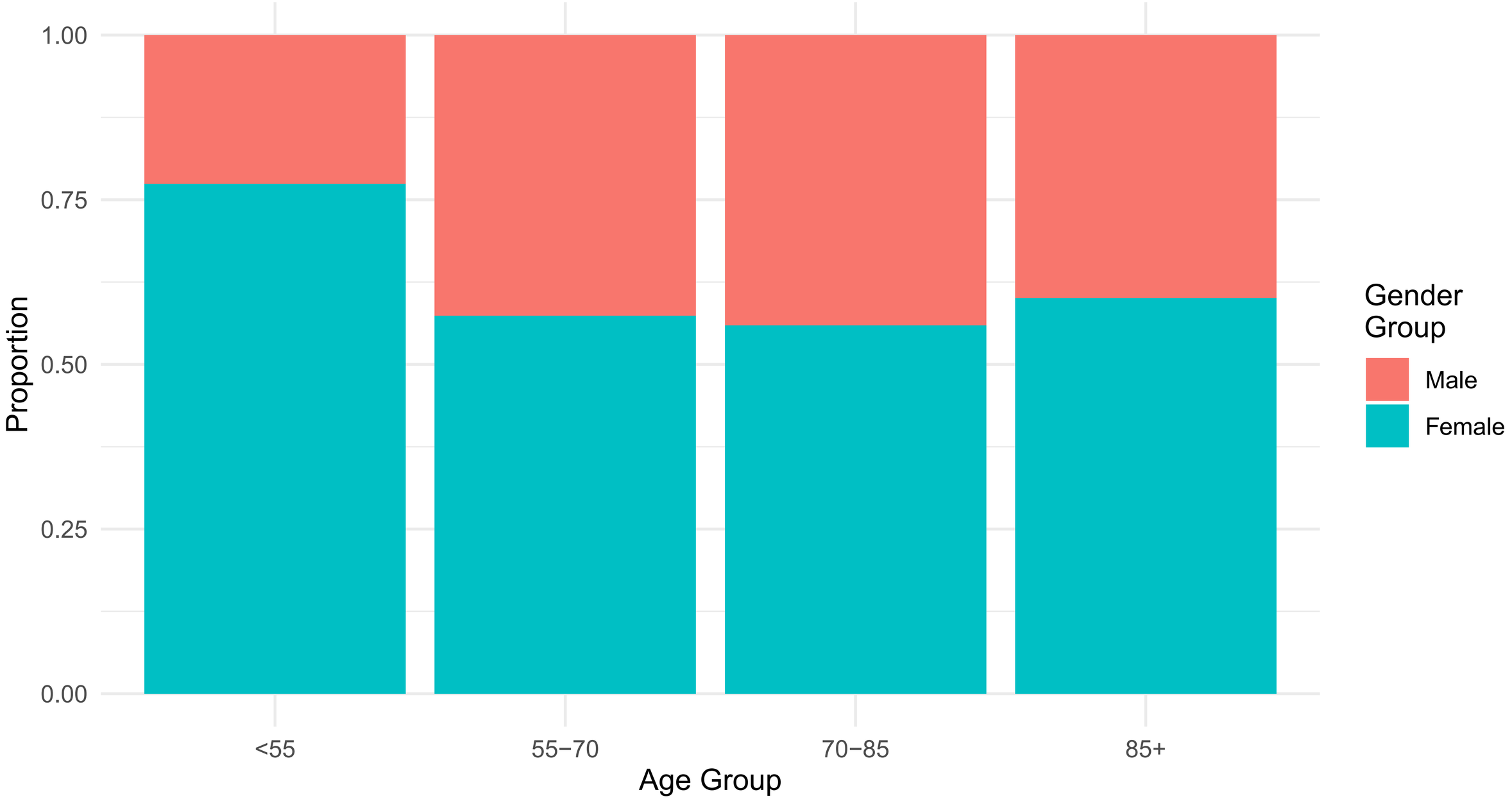

Age gradient in Self−perceived health
Self−Perceived Health
5
4
3
2
1
40
60
80
100
Age

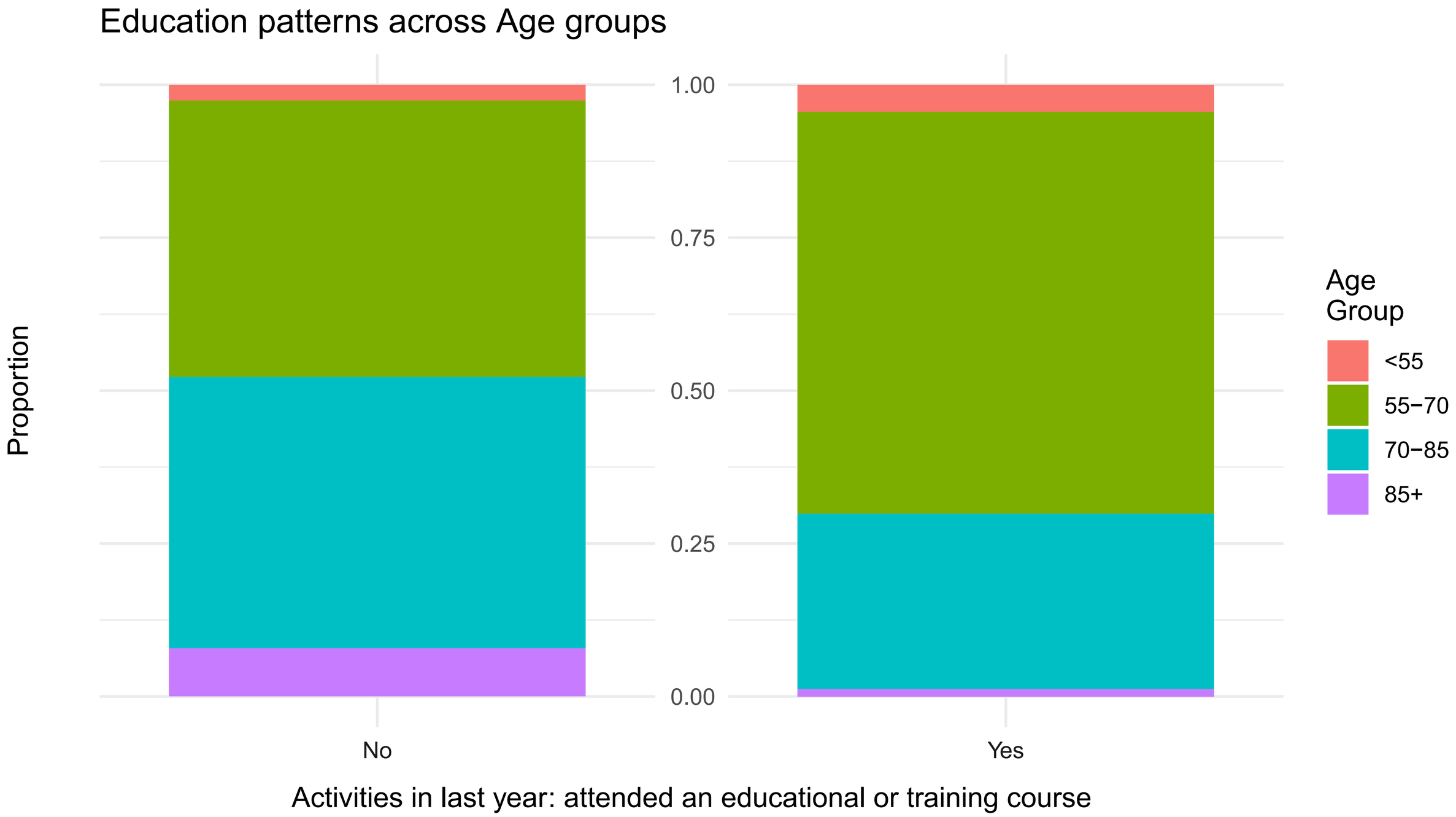
Self−perceived health distribution by Country

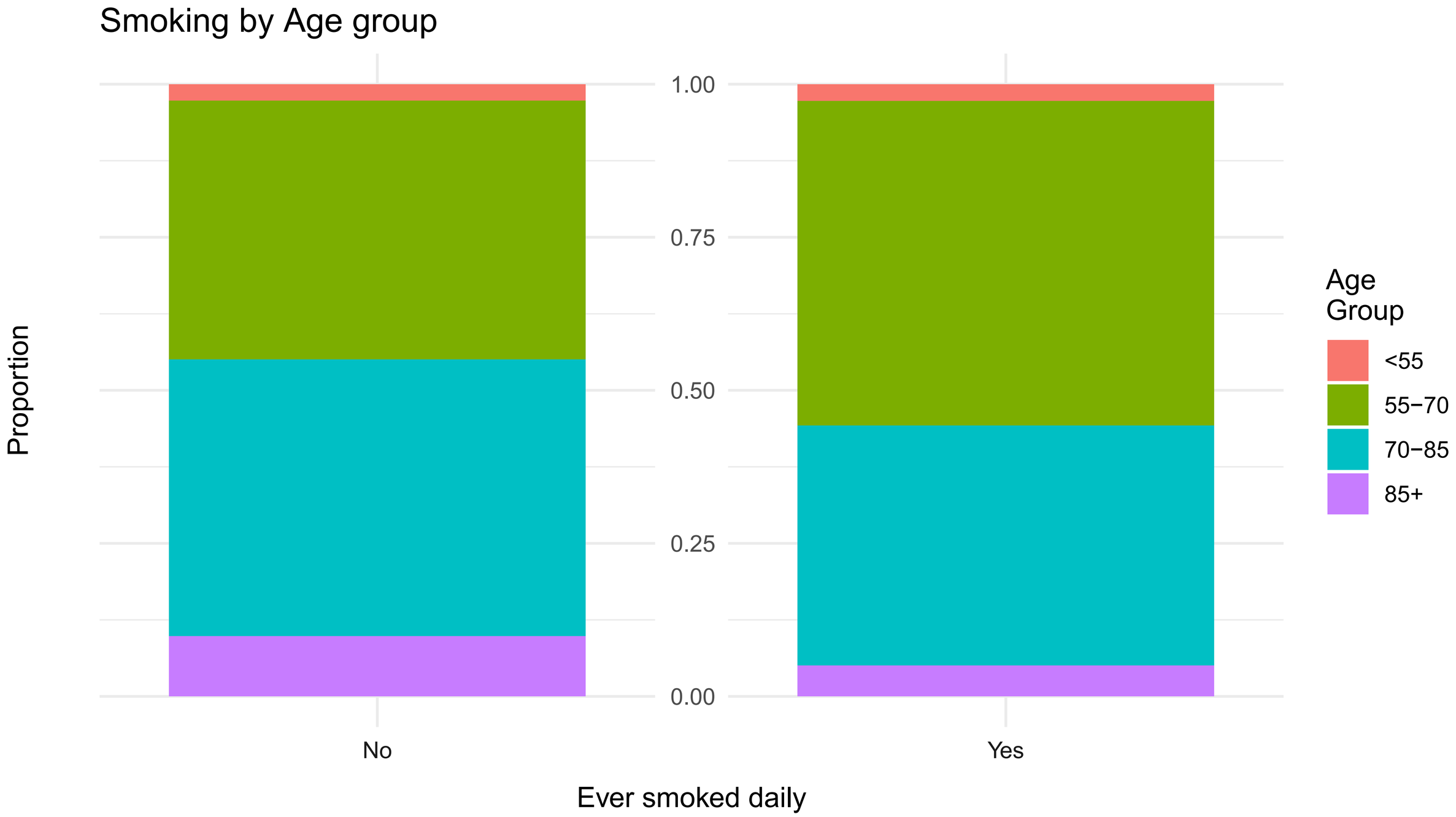

63.328
73.536


2.628
3.97


Physical activity by Country
1.00
0.75
0.50
0.25
0.00
Proportion
Austria
Germany
Sweden
Netherlands
Spain
Italy
France
Denmark
Greece
Switzerland
Belgium
Israel
Czech Republic
Poland
Luxembourg
Hungary
Slovenia
Estonia
Croatia
Lithuania
Bulgaria
Cyprus
Finland
Latvia
Malta
Romania
Slovakia
Country
Physically (in)active
Active
Inactive


Female survival advantage by Age group
1.00
0.75
0.50
0.25
0.00
Proportion
<55
55−70
70−85
85+
Age Group
Gender Group
Male
Female


Education patterns across Age groups
Proportion
1.00
0.75
0.50
0.25
0.00
No
Yes
Activities in last year: attended an educational or training course
Age Group
<55
55−70
70−85
85+


Smoking by Age group
1.00
0.75
0.50
0.25
0.00
Proportion
No
Yes
Ever smoked daily
Age Group
<55
55−70
70−85
85+

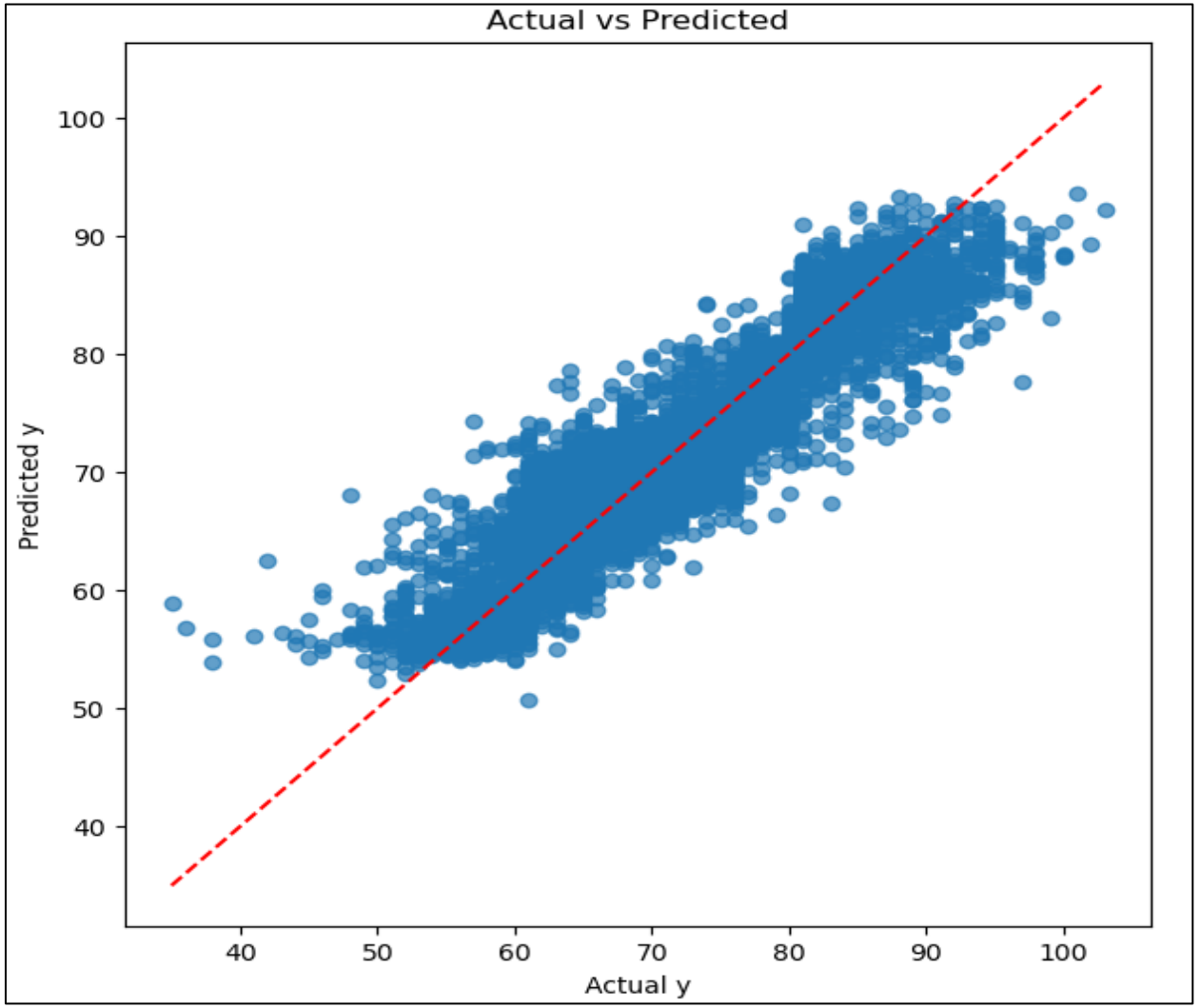
Feature Selection

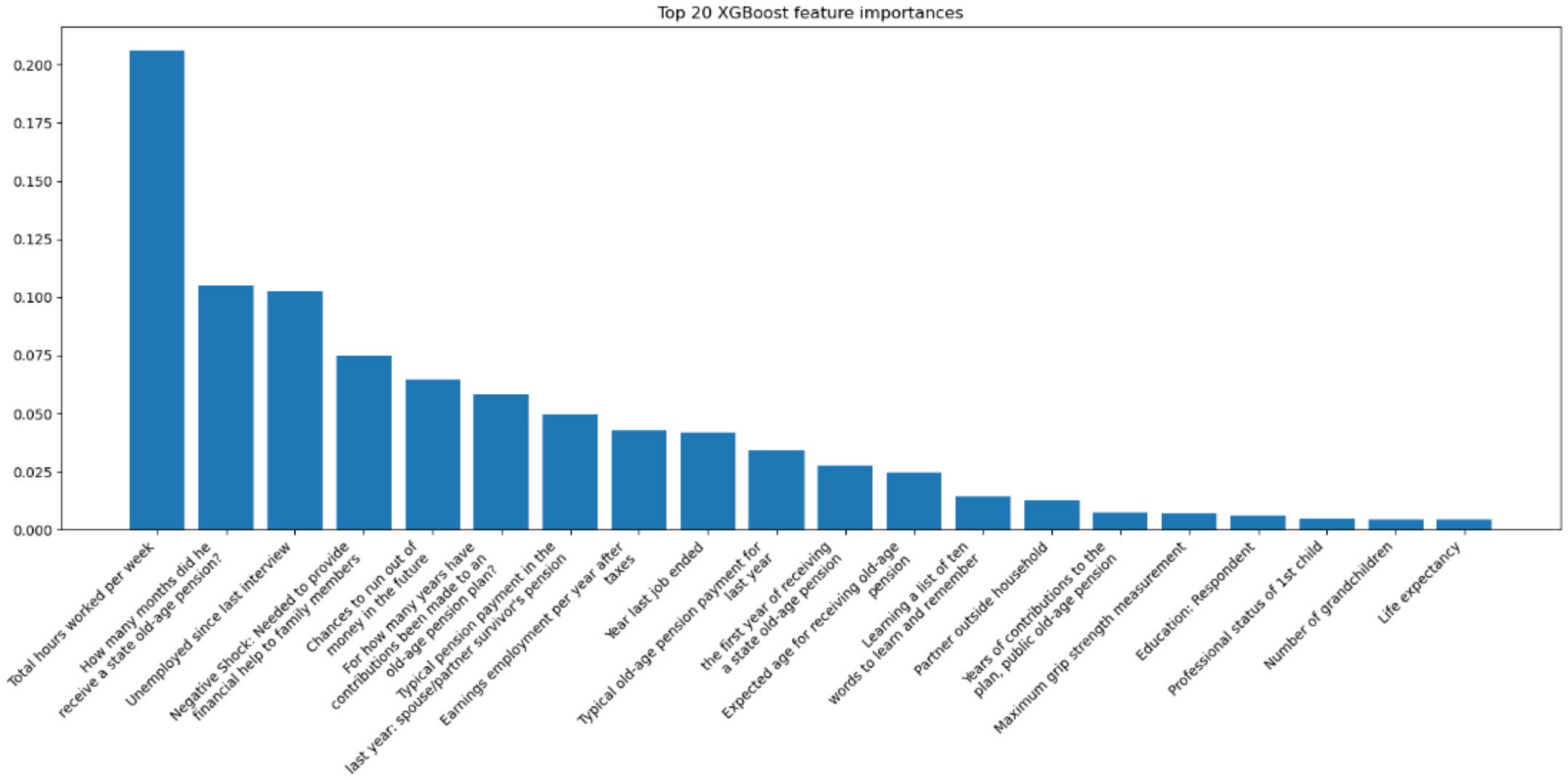
- Remove predefined redundant and multicollinear columns (from two exclusion lists)
- Replace all negative values with NaN (in numeric columns)
- Drop columns with ≥70% missing values
- Impute remaining missing numeric values with the median
- Split features into 10 chunks and compute feature importance using a Decision Tree for each chunk
- Select features

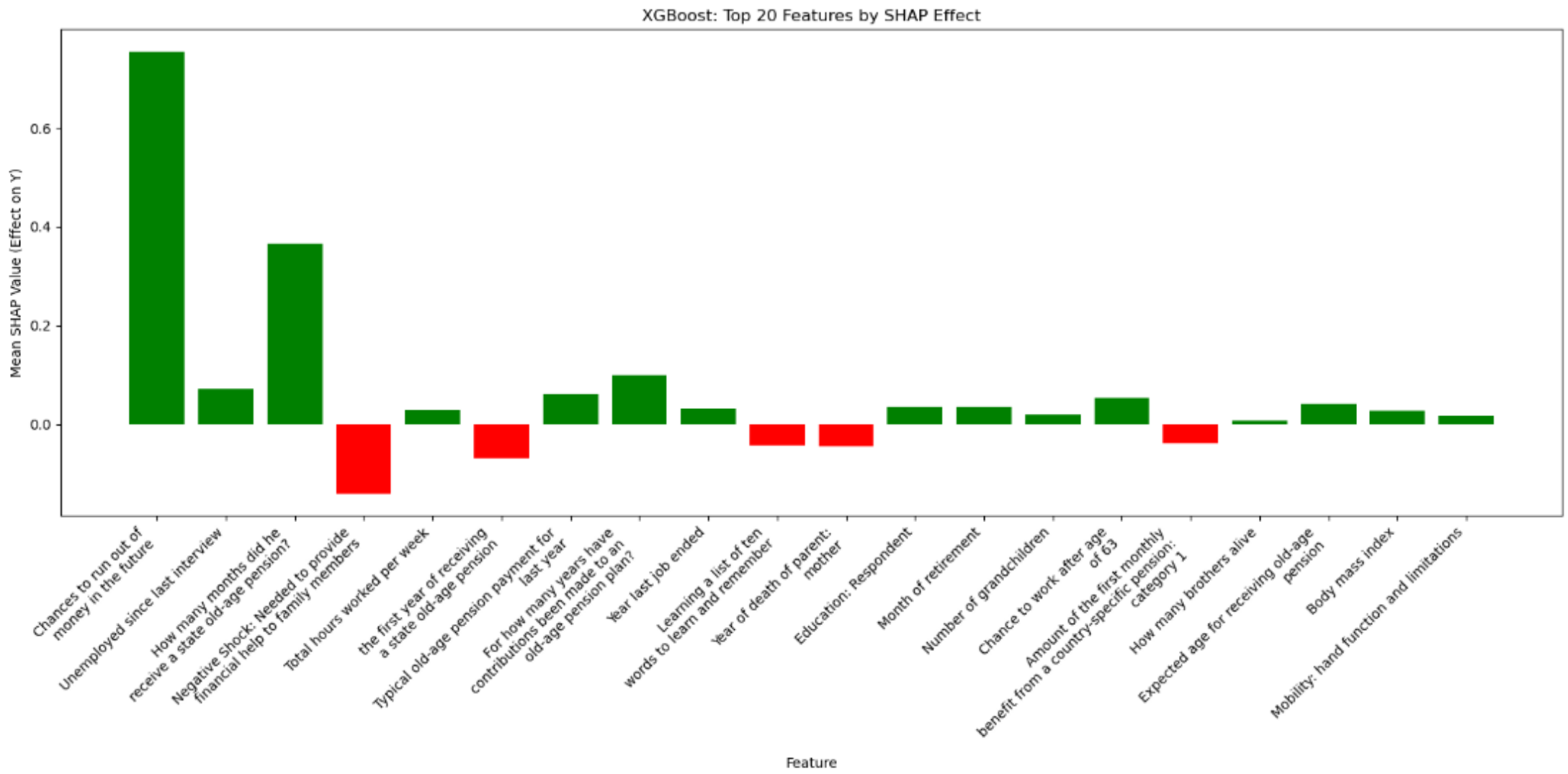
TEST set MAE: **2.534**
TEST set MAPE: **3.62%**
TEST set R^2: **0.871**
TEST set Adjusted R^2: **0.868**

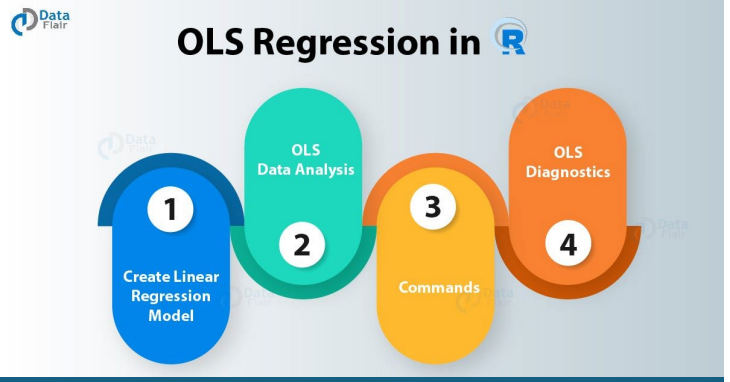
TRAIN set MAE: **2.489**
TRAIN set MAPE: **3.56%**
TRAIN set **R^2: 0.878**
TRAIN set Adjusted R^2: **0.878**

The data is split into 80% training and 20% test sets.




Top 20 XGBoost feature importances
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Total hours worked per week
How many months did he receive a state old-age pension?
Unemployed since last interview
Negative Shock: Needed to provide financial help to family members
Chances to run out of money in the future
For how many years have contributions been made to an old-age pension plan?
Typical pension payment in the last year: spouse/partner survivor's pension
Earnings employment per year after taxes
Year last job ended
Typical old-age pension payment for last year
the first year of receiving a state old-age pension
Expected age for receiving old-age pension
Learning a list of ten words to learn and remember
Partner outside household
Years of contributions to the plan, public old-age pension
Maximum grip strength measurement
Education: Respondent
Professional status of 1st child
Number of grandchildren
Life expectancy

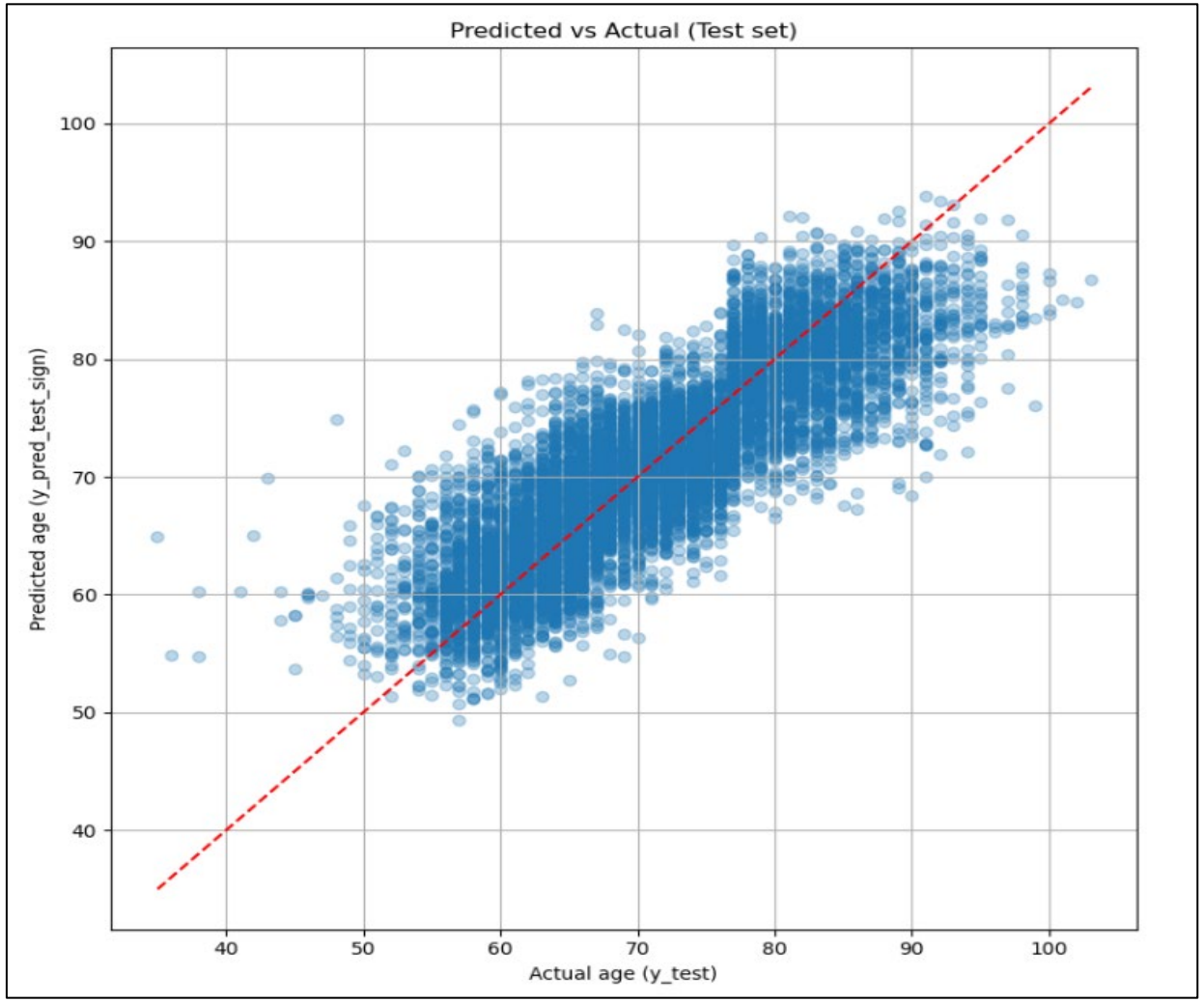

XGBoost: Top 20 Features by SHAP Effect
Mean SHAP Value (Effect on Y)
0.6
0.4
0.2
0.0
Chances to run out of money in the future
Unemployed since last interview
How many months did he receive a state old-age pension?
Negative Shock: Needed to provide financial help to family members
Total hours worked per week
the first year of receiving a state old-age pension
Typical old-age pension payment for last year
For how many years have contributions been made to an old-age pension plan?
Year last job ended
Learning a list of ten words to learn and remember
Year of death of parent: mother
Education: Respondent
Month of retirement
Number of grandchildren
Chance to work after age of 63
Amount of the first monthly benefit from a country-specific pension: category 1
How many brothers alive
Expected age for receiving old-age pension
Body mass index
Mobility: hand function and limitations
Feature



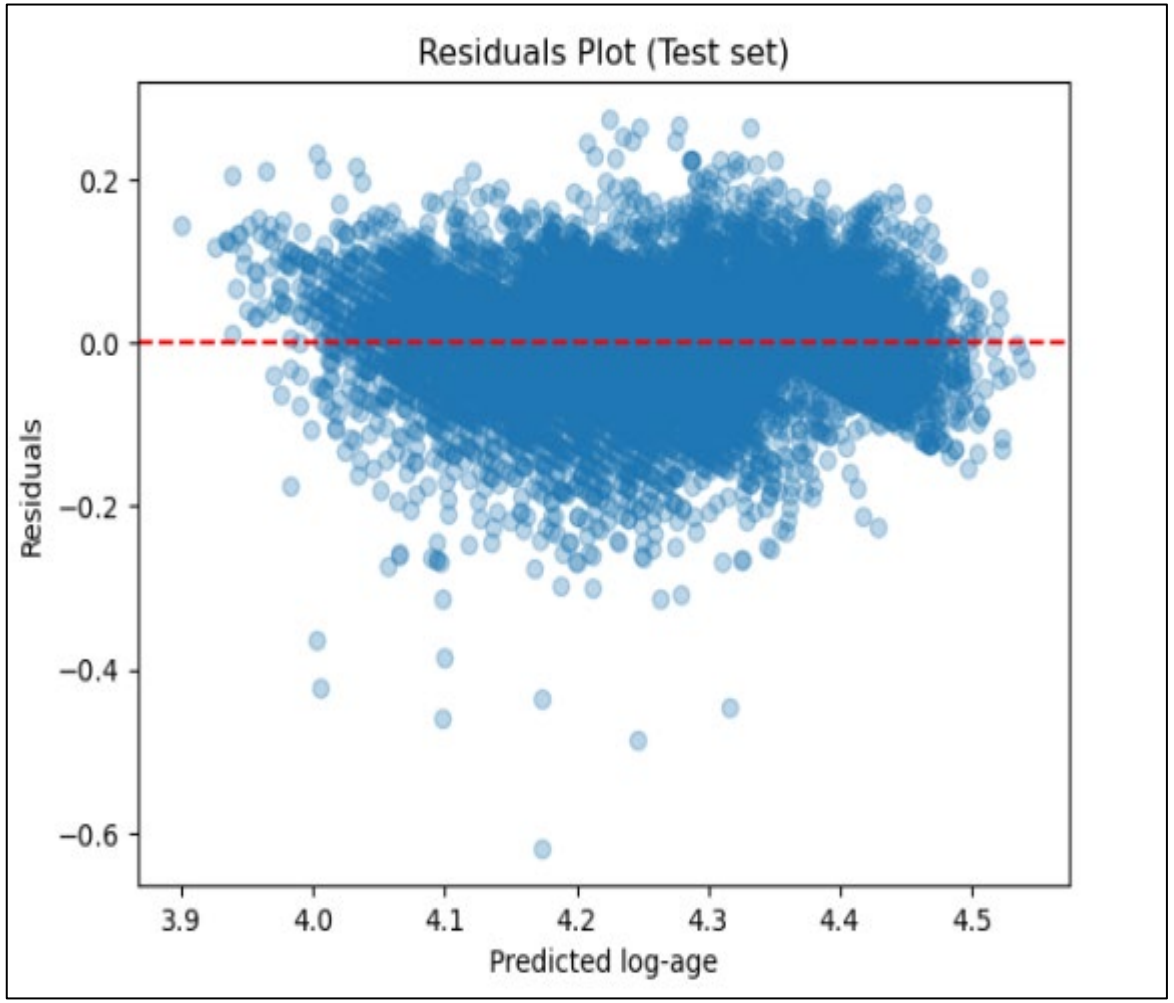
OLS Regression: Feature Selection and Collinearity Control

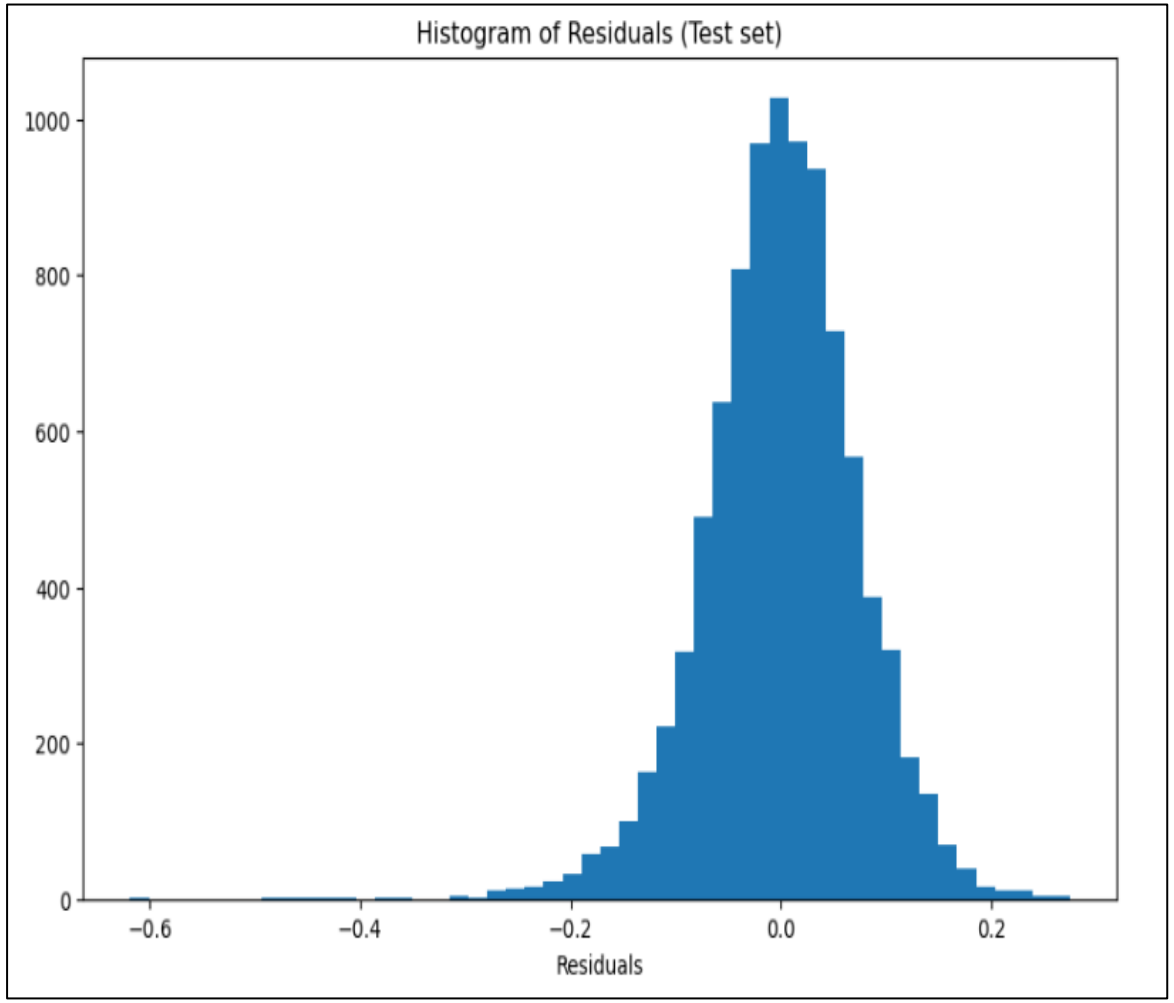
TEST set MAE: **3.916**
TEST set MAPE: **13.25%**
TEST set MSE: **25.989**
TEST set R^2: **0.708**
TEST set Adjusted R^2: **0.706**

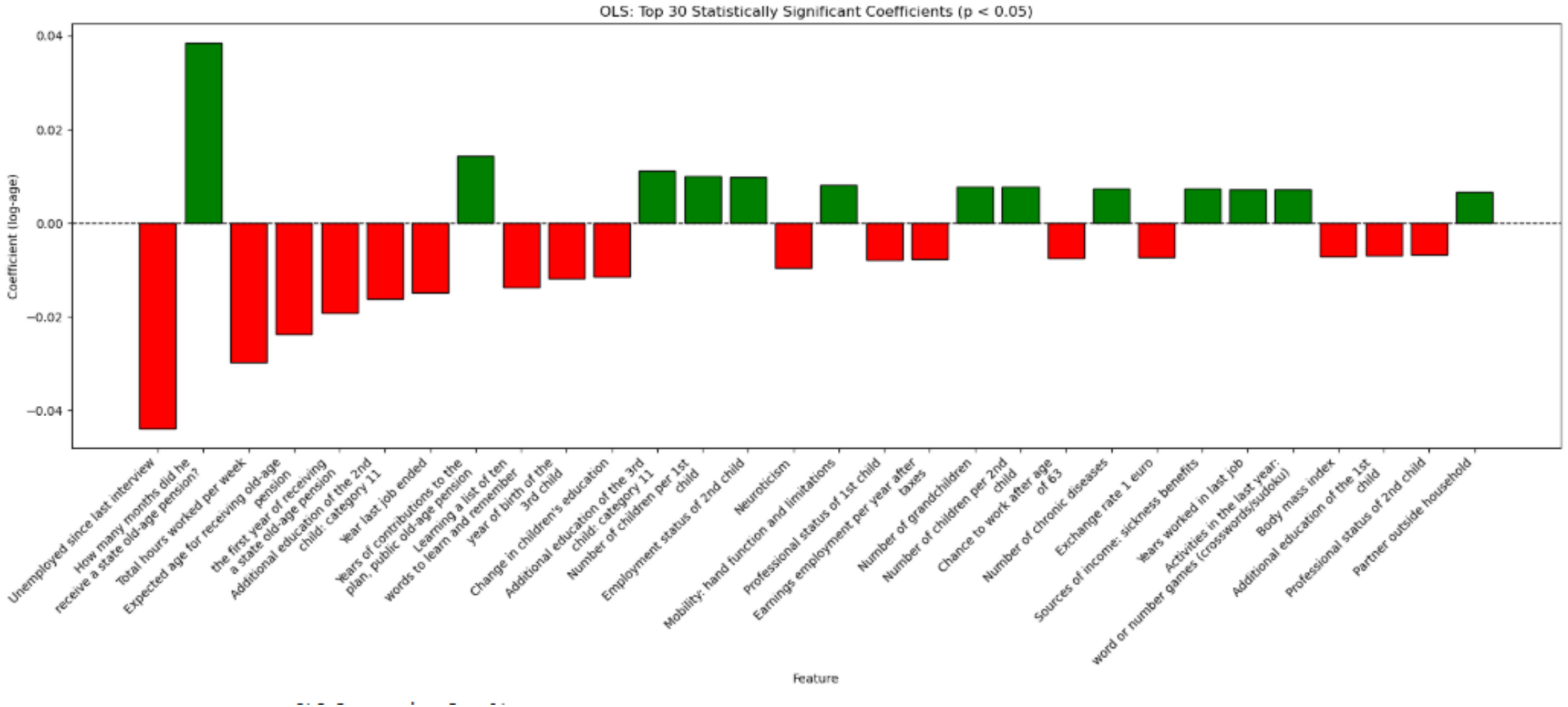
TRAIN set MAE: **3.921**
TRAIN set MAPE: **5.60%**
TRAIN set MSE: **26.027**
TRAIN set R^2: **0.711**
TRAIN set Adjusted R^2: **0.710**

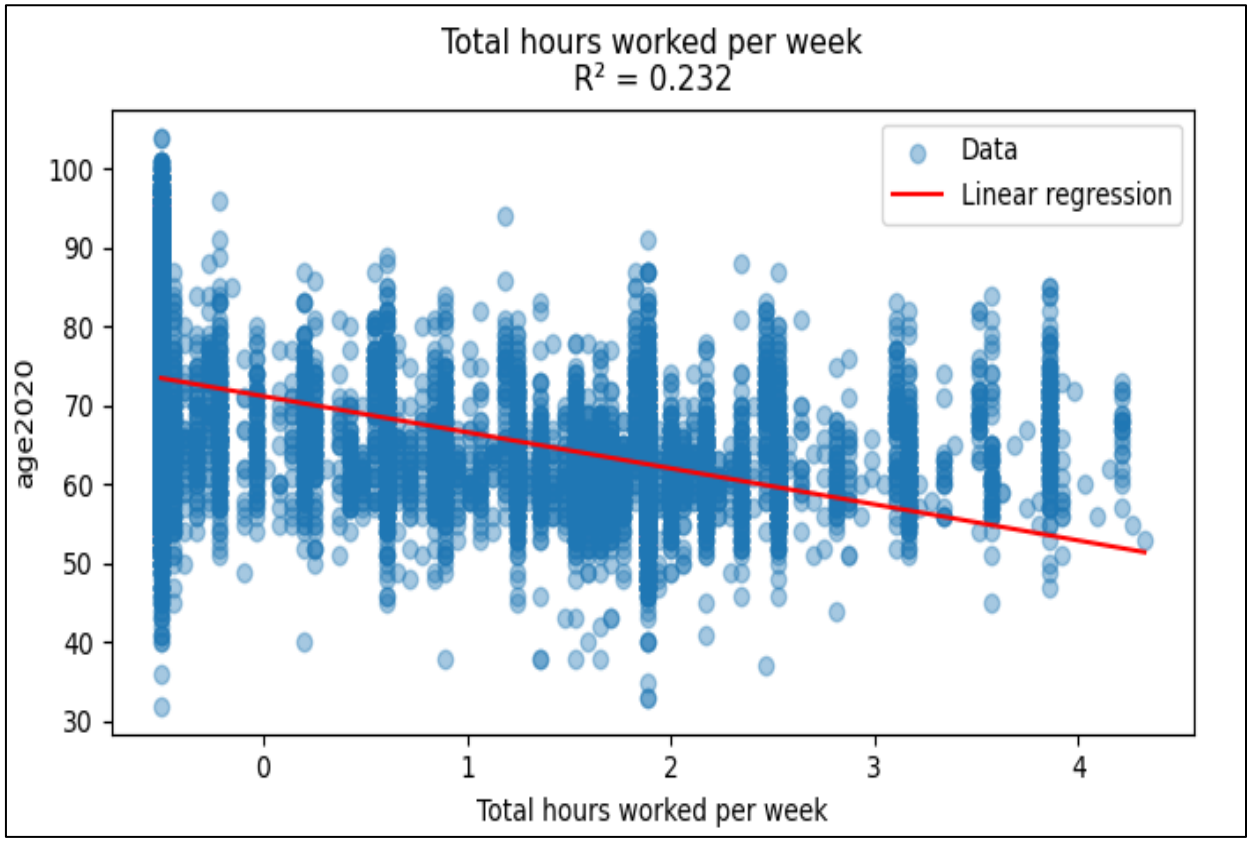


Residue diagnostics






OLS: Top 30 Statistically Significant Coefficients (p < 0.05)
Coefficient (log-age)
0.04
0.02
0.00
−0.02
−0.04
Unemployed since last interview
How many months did he receive a state old-age pension?
Total hours worked per week
Expected age for receiving old-age pension
the first year of receiving a state old-age pension
Additional education of the 2nd child: category 11
Year last job ended
Years of contributions to the plan, public old-age pension
Learning a list of ten words to learn and remember
year of birth of the 3rd child
Change in children's education
Additional education of the 3rd child: category 11
Number of children per 1st child
Employment status of 2nd child
Neuroticism
Mobility: hand function and limitations
Professional status of 1st child
Earnings employment per year after taxes
Number of grandchildren
Number of children per 2nd child
Chance to work after age of 63
Number of chronic diseases
Exchange rate 1 euro
Sources of income: sickness benefits
Years worked in last job
Activities in the last year: word or number games (crosswords/sudoku)
Body mass index
Additional education of the 1st child
Professional status of 2nd child
Partner outside household
Feature

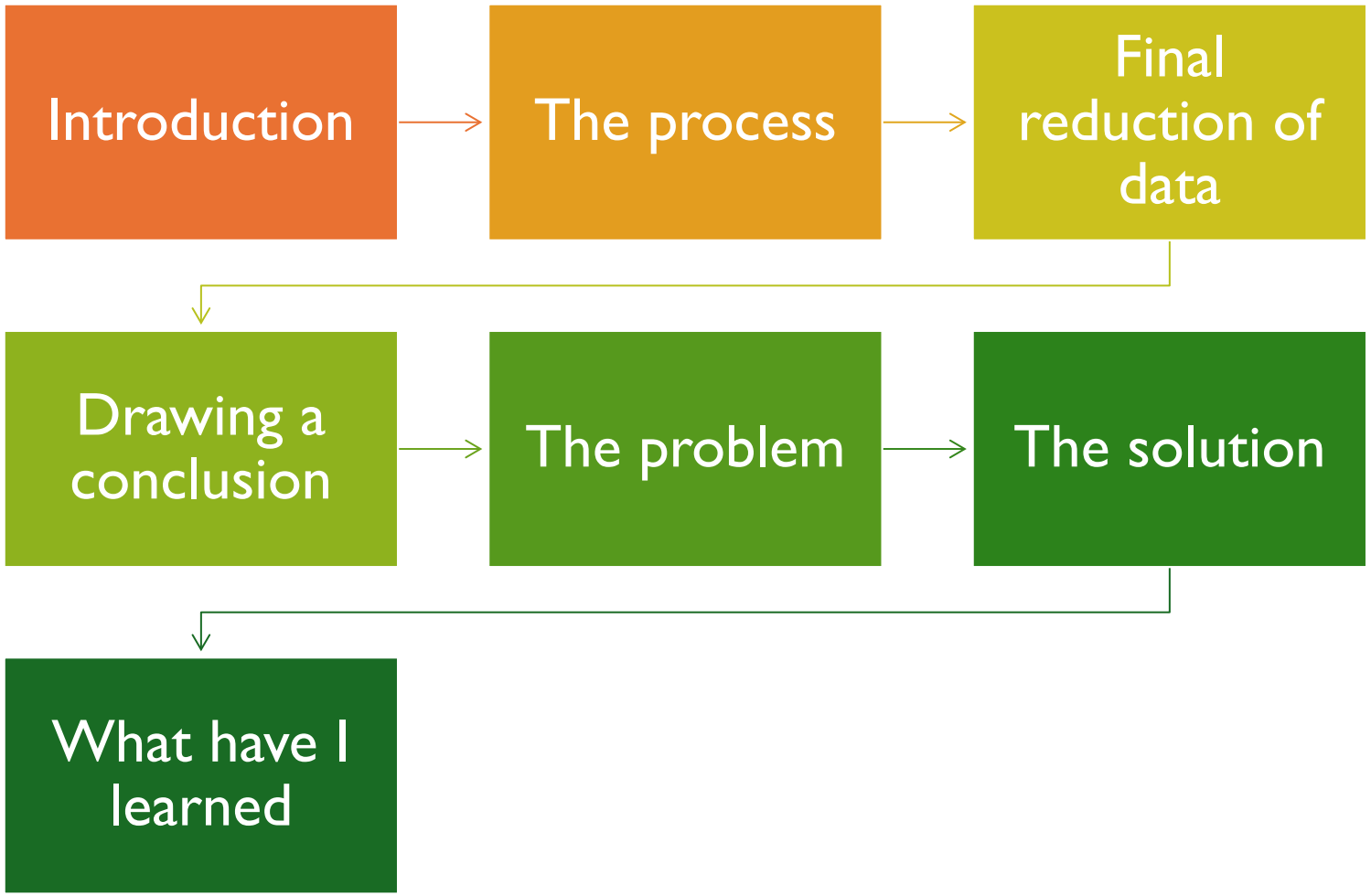


0 = 20 hours per week
1 = 30 hours per week
2 = 40 hours per week (full-time)
3 = 50 hours per week (overtime)
4 = 60 hours per week (heavy workers)

0 = 20 000
1 = 40 000
2 = 40 000
3 = 50 000
4 = 80 000

Variable_down	Effect_down	Label by SHARE	Variable_up	Effect_up	Label by SHARE3
ep325_	down	Unemployed since last interview	ep208_1	up	How many months did he receive a state old-age pension?
ep013_	down	Total hours worked per week	ch016_1	up	Years of contributions to the plan, public old-age pension
cf016tot	down	Learning a list of ten words to learn and remember	ch019_1	up	Number of children per 1st child
bfi10_neuro	down	Neuroticism	chronicw8c	up	Number of chronic diseases
bfi10_extra	down	Extraversion	co002ub	up	Amount spent on food at home
ep205e	down	Earnings employment per year after taxes	ep328_	up	Month of retirement
te035_	down	Hours spent on paid work	ac035d9	up	Activities in last year: did word or number games (crossword puzzles/Sudoku...)
ac035d5	down	Hours spent on paid work	ch019_2	up	Number of children per 2nd child
ac036_9	down	How often did word or number games the last 12 months	ch021_	up	Number of grandchildren
bmi	down	Body mass index	ac036_5	up	How often gone to a sport/social/other kind of club the last 12 months
exrate	down	Exchange rate 1 euro	sn007_2	up	Contact on social network
dn036_	down	How many brothers alive	ep671d7	up	Sources of income: sickness benefits
te005_	down	Hours spent on chores	ex801_	up	Chances to run out of money in the future
ex007_	down	Chance government reduces pension	ch_occupation_3	up	Professional status of the 3rd child
ep671d12	down	Income sources: public long-term care insurance	ep671d6	up	Sources of income: disability pension
maxgrip	down	Maximum grip strength measurement	ep649_	up	Years worked in last job
ex009_	down	Life expectancy	ch_outhh_receive_care_2	up	Outside the household: assistance received from 2nd child
ep152isco	down	ISCO code: respondent's last job	ho034_	up	Years in accommodation
dn037_	down	How many sisters alive	ch_school_education_2	up	School degree of child 2, based on ch017_2 & ch510_2
areabldgi	down	Area of building	ch_closeness_1	up	Emotional closeness with 1st child
as051e	down	Amount selling cars	ep078e_1	up	Typical old-age pension payment for last year
br003_	down	How many years smoked	ep678e	up	Amount received from all occupational pensions last year
br006_	down	Average amount of cigarettes per day	ft009_	up	Received financial gift of 250 or more
hc877_	down	Contacts specialist	ch022_	up	Has great-grandchildren
eurod	down	Depression on the EURO-D scale	isced2011_f	up	Parent's education: Father
numeracy2	down	Math result: subtraction	te047_	up	Hours spent on napping and resting during daytime
hc125_	down	Satisfaction with own coverage in basic health insurance	as070e	up	Interest or dividend income
ac036_10	down	How often played cards or games such as chess the last 12 months	ch_closeness_3	up	Emotional closeness with the 3rd child
ep616isco	down	Respondent's current main job	ep078e_2	up	Typical sickness pension payment for last year

Outline


Introduction
The process
Final reduction of data
Drawing a conclusion
The problem
The solution
What have I learned

Introduction

The process

First contact with Share

Reading the data

Sampling the data

Deviding the dataset by feature sets

Running different models

Unsatisfactory results

New approach

Dropping features with more than 50% missing values

Reducing the data by using feature importance

Reducing the data by correlation

Excluding leaky features

Training logistic regression, random forest and decision three

Results: maximum accuracy on random forest: ~ 0.6 accuracy

```
print(f"original df shape: {df.shape}")
df2 = drop_admin(df)
print(f"drop_admin df2 shape: {df2.shape}")
df2 = filter_missing(df2)
print(f"filter_missing df2 shape: {df2.shape}")
df2 = filter_variance(df2)
print(f"filter_variance df2 shape: {df2.shape}")
df2 = drop_high_card(df2)
print(f"drop_high_card df2 shape: {df2.shape}")
df2 = corr_filter(df2)
print(f"corr_filter df2 shape: {df2.shape}")
```

```
model_df shape: (46614, 4)
Accuracy: 0.4279738281668990 5
              precision    recall  f1-score   support

           1       0.00      0.00      0.00      2046
           2       0.40      0.49      0.44      3553
           3       0.46      0.60      0.52      3724

    accuracy                           0.43      9323
   macro avg       0.28      0.36      0.32      9323
weighted avg       0.33      0.43      0.37      9323
```

```python
y = df_clean['sphus']
X_feats = df_clean.drop(columns=['sphus'])
print(f"X shape before vif_filter: {X_feats.shape}")
X_feats = vif_filter(X_feats)
print(f"X afteer vif_filter: {X_feats.shape}")
print(f"df_clean after vif_filter: {df_clean.shape}")
df_clean = pd.concat([X_feats, y], axis=1)
# 3) Build X/y and train OVR
X = df_clean.select_dtypes(include='number').drop(columns=['sphus'])
y = df_clean['sphus']
print(f"fina X shape: {X.shape}")
models = train_ovr(X, y, classes=[1,2,3])
```

Model: rf
Accuracy: 0.94943579182219 94

	precision	recall	f1-score	support
not_1	0.95	0.99	0.97	36384
is_1	0.95	0.82	0.88	10230
accuracy			0.95	46614
macro avg	0.95	0.90	0.92	46614
weighted avg	0.95	0.95	0.95	46614

Model: xgb
Accuracy: 0.8150341099240571

	precision	recall	f1-score	support
not_1	0.82	0.98	0.89	36384
is_1	0.77	0.23	0.35	10230
accuracy			0.82	46614
macro avg	0.79	0.60	0.62	46614
weighted avg	0.81	0.82	0.77	46614

--- Class 2 vs Rest ---

Model: log
Accuracy: 0.5754708885742481

	precision	recall	f1-score	support
not_1	0.85	0.56	0.67	36384
is_1	0.29	0.64	0.40	10230
accuracy			0.58	46614
macro avg	0.57	0.60	0.53	46614
weighted avg	0.72	0.58	0.61	46614

Reading the cookbook

Final reduction of data

- Choosing about 1000 features, carefully evading leaky ones
- Processing these features through a pipeline
- Extracting the feature importance of the ~400 features left
- Grouping the features in 11 cathegories
- Training the model using different combinations of the groups

Drawing conclusions

1. **Mobility limitation** (`mobility` : 0.0227)
2. **Chronic conditions** (`chronicw8c` : 0.0227)
3. **Maximum grip strength** (`maxgrip` : 0.0183)
4. **Numeracy-2 test** (`numeracy2` : 0.0082)
5. **Transfer variable TE-026** (`te026_` : 0.0091)

Health and physical

Functional limitation and chronic desieses are the prime factor of chang in self perceived health

Psychological & Cognitive

- Broader well-being (CASP) matters more than raw memory scores, though numeracy and loneliness also play roles.

Less influational domains

These domains had very small individual importances

The problem

```
=== All features model ===
Accuracy: 0.6433551431942508
              precision    recall  f1-score   support

           1       0.66      0.47      0.55      2046
           2       0.55      0.65      0.59      3553
           3       0.75      0.73      0.74      3724

    accuracy                           0.64      9323
   macro avg       0.65      0.62      0.63      9323
weighted avg       0.65      0.64      0.64      9323
```


The solution

What have I learned

- Without the knowledge, but with a direction
- Keeping a journal/log is more important than I initially thought
- Avoiding leaky feature
- Feature engineering
- Dimensionality reduction

Q & A

Thank you for your attention!

Appendix

List of Countries

- Austria
- Germany
- Sweden
- Netherlands
- Spain
- Italy
- France
- Denmark
- Greece
- Switzerland
- Belgium
- Israel
- Czech Republic
- Poland
- Luxembourg
- Hungary
- Slovenia
- Estonia
- Croatia
- Lithuania
- Bulgaria
- Cyprus
- Finland
- Latvia
- Malta
- Romania
- Slovakia