Household Income Predictors based on SHARE data

Eirini Ornithopoulou

Introduction and Data Exploration

This analysis aims to identify significant predictors of household income using the Survey of Health, Ageing and Retirement in Europe (SHARE) dataset, as described here. The process involved exploratory data analysis (EDA) and multiple regression modeling to pinpoint key factors influencing household income. Additional steps were taken, to reach a more optimized solution.

In short, one of the variables, whether one lives with their partner, was eventually removed, because it impeded the analysis and did not provide enough variance to make correlations. The categorical variables were converted to numericals before the correlation matrix was tabulated, and a multiple regression model was trained. After determining the key factors, another multiple regression was performed, which had an improved predictive performance.

It is rather expected that the highest predictor was the number of people in a household, where more people meant higher income. However, less obvious was the correlation between having books at the age of 10, and the relative language literacy with the household income later in life. Perhaps this points to a confounding factor, such as that the early exposure to studies, affects the education level and ultimately the individual’s level of income. The observations are further discussed below.

Results and Discussion

Histogram of Household Income Distribution

A histogram was used to visualize the distribution of household income across the sample (Fig. 1). The histogram shows a wide variation in household income with a right-skewed distribution. A few households have exceptionally high incomes, indicating potential outliers or high-income earners in the dataset.

Regression Modeling and Results

I fitted a linear regression model to predict household income, determined the important predictors, and then retrained a model. The first model included all predictors provided such as sex, age, education, and more, except one, which lacked the necessary variance (lives_with_partner). This information is implied in part in the number of people in the household (n_household). After determining the most important factors, the second multiple linear regression model was trained on those, see Table 1.

Table 1. Regression Results

Predictor	Estimate	p-value
(Intercept)	18409	0.000104 ***
Number of Household Members	15002	< 2e-16 ***
Books at Age 10 (fewer books)	-6031	0.000840 ***
Relative Language Ability at Age 10 (worse)	-5008	0.004324 **

Significance codes: “***” 0.001 “**” 0.01

Key predictors:

Number of Household Members: This is a significant positive predictor of household income (p < 2e-16), indicating that larger households tend to have higher incomes. This is also supported in the boxplot in Fig. 3.

Books at Age 10: Having fewer books at age 10 is associated with lower household income (p < 0.001), suggesting early access to educational resources impacts future income. In Fig. 3, we can see in the related boxplot that more than 25 books at age 10 is correlated with a higher household income.

Relative Language Ability at Age 10: Worse relative language ability at age 10 is associated with lower household income (p < 0.01). And better/same relative language ability at age 10 is correlated with a lower household income, this can be observed in the related boxplot in Fig. 3.

The model can be used for predicting household income based on the significant demographic and early-life factors identified. It is applicable primarily to similar populations (older adults in Europe).

The model has an R-squared value of 0.2328, explaining about 23.28% of the variance in household income. This could be explained by the existence of unmeasured factors that also influence income. Additionally, it is possible that there are other models, better for fitting non-linear relationships, and other approaches, such as feature engineering that could improve the performance of the model. However, as mentioned before we see some correlation between the years of education and the books at age ten in the correlation matrix (Fig. 2), potentially making it a confounder for the household income.

In conclusion, future research could explore additional predictors such as occupation, geographic location, and lifestyle factors and non-linear relationships should be investigated. But it is safe to assume that proper exposure to systematic knowledge will yield more financially successful households.

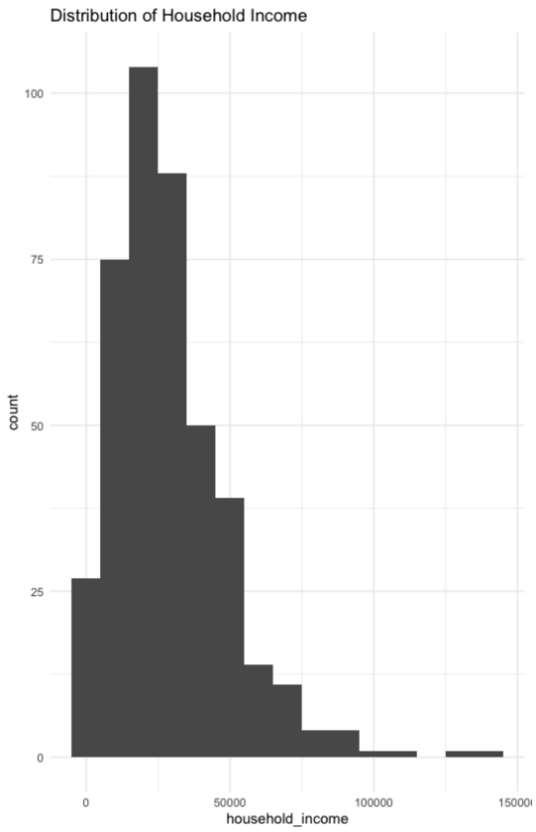

Figure 1The household income distribution. Bin size: $10000.

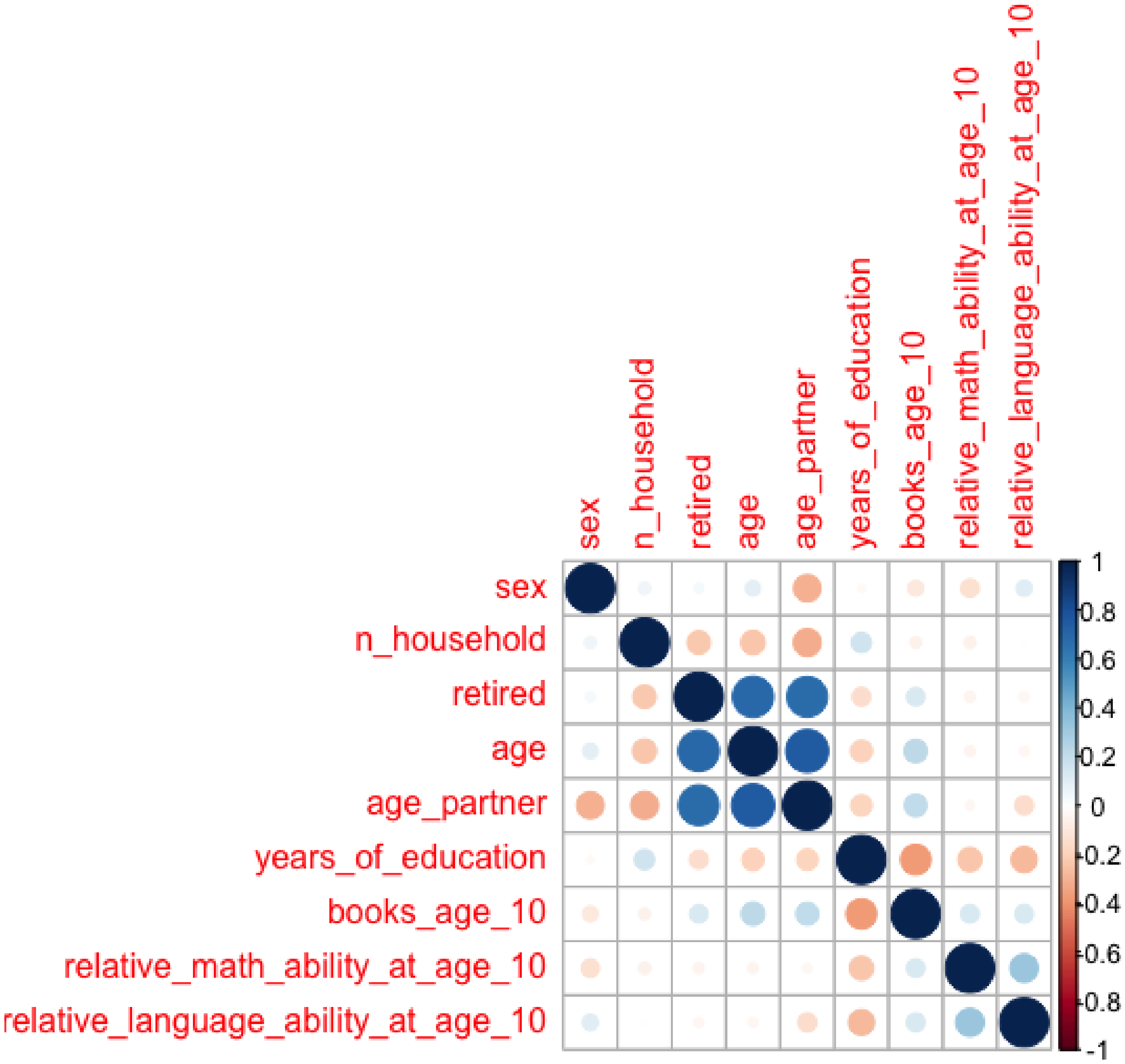

Figure 2 The correlation matrix for all the numerized variables in the dataset.

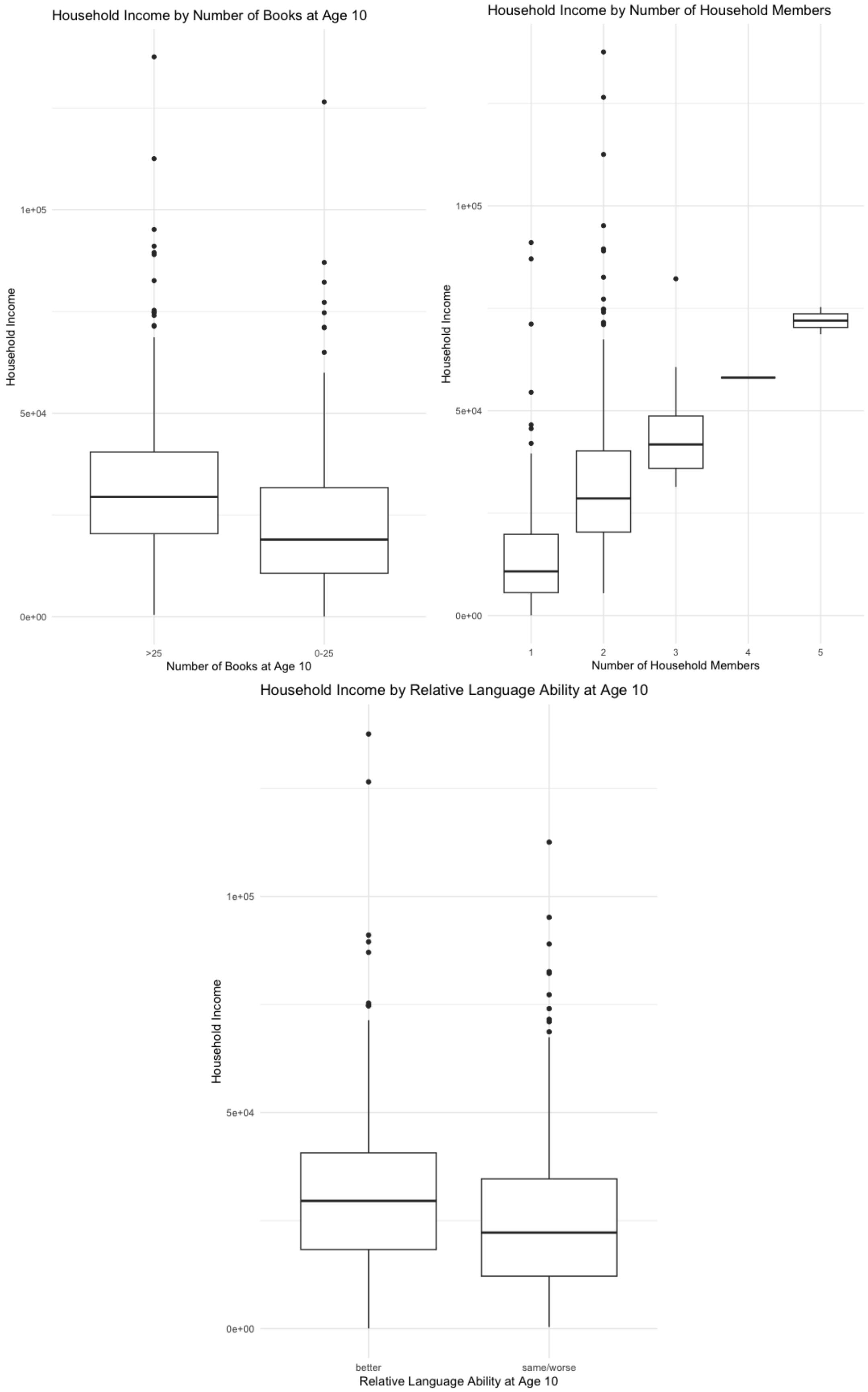

Figure 3 Top: Boxplots for the household income over the number of books at age 10 (left) and the number of household members (right). Below: A boxplot of the household income over the relative language ability at 10.