b
UNIVERSITÄT
BERN

Mathematical Institute

CAS Applied Data Science

The Influence of the Socioeconomic Status on the Health Condition

Predicting Health by Income

David von Holzen
Ziegeleiweg 1, 6048 Horw
david.vonholzen@bluewin.ch

October 4, 2024

Abstract

Poverty poses a major challenge for our societies. It does not only show itself in a lack of financial resources, but also in other life dimensions such as health. In order to get a deeper understanding of these related dimensions of poverty, the connection between the financial resources and the health status will be analyzed. To achieve this, survey data on the individual level from France and Germany will be analyzed. More specifically, a supervised machine learning approach is applied, i.e. ordinal logistic regression, to predict the health status of a person by their income. A positive finding would implicate, that political measures on lowering poverty, would have a positive impact on the health condition of the population. The presented concept will outline in a first step the aim of the project and in a second step the most important prerequisites that have to be met for a successful execution of the project. The prerequisites include the data base, the methods, the documentation and the possible risks, which could be encountered during the project.

Contents

1 Project Objectives

Poverty not only shows itself as a lack of financial resources of a private household. It is also related to other dimensions like the housing quality or the health status (Bundesamt für Sozialversicherungen 2023). These dimensions are interrelated. For instance, if there is a lack of financial resources this might lead to forego a visit by a doctor if there is a medical issue. In the longer term this can have a negative impact on the general health condition of an individual. On a societal level, income inequality, as a indicator of poverty, can be connected to a lower level of the general health condition. Wilkinson & Pickett (2009) for example show that in countries with higher income inequality, the rate of health problems is higher.

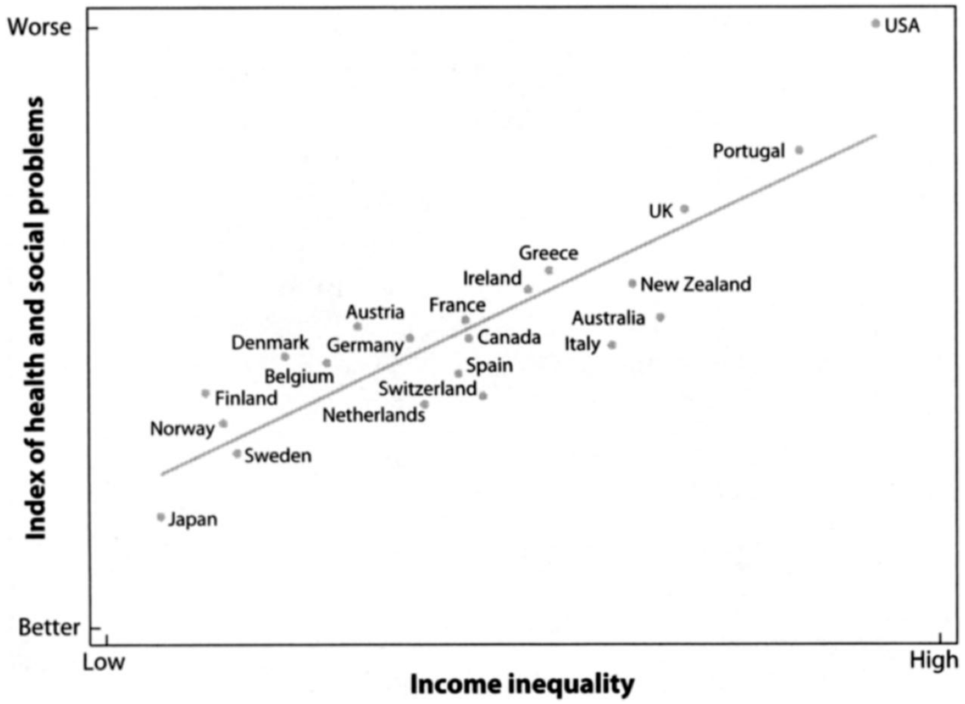


Figure 1

Index of Health and Social Problems in relation to income inequality in rich countries. Income inequality is measured by the ratio of incomes among the richest compared with the poorest 20% in each country. The index combines data for the 10 outcomes listed in Table 2. Raw scores for each variable were converted to z-scores and each country given its average z-score. Source: Wilkinson & Pickett 2009.

Therefore, the target of the project is to contribute to the understanding of the relationship between income and the health status by trying to predict the health status by the income level. This would suggest that the income inequalities can also lead to inequalities in health status. More specifically, with a supervised machine learning approach and survey data from

France and Germany, it is tried to predict the health status with the income as feature. A positive finding would implicate, that political measures on lowering the income inequality and hence poverty as well, would also have a positive impact on the health condition of the population.

2 Methods

2.1 Infrastructure & Software

Concerning the infrastructure Google Drive will be used to store the datasets (i.e. raw and final), the codebook of the used datasets and the IPython notebook.
Regarding the software (libraries) the IPython notebook will be executed on Google Colaboratory to import a number of Python libraries. The following libraries will be used for the project:

- Pandas: Data manipulation and analysis tools
- Matplotlib: Visualization tools
- Seaborn: Statistical graphic tools
- Sklearn: Machine learning tools
- Mord: Ordinal logistic regression tools

The Pandas library will be used to import and prepare the data for the analysis. The Matplotlib library will be needed to create the descriptive statistics, i.e. to visualize the distribution of the relevant variables in the dataset. The Seaborn library will be applied to create violin plots for looking at the distribution of the numeric variables. The Sklearn library serves to apply machine learning features such as splitting the dataset in a training and a test set. The Mord library supplements the Sklearn library by providing tools for ordinal logistic regressions.

2.2 Statistical Methods

Speaking about the statistical methods, the following procedure will be carried out: In a first step, descriptive statistics will be used to get an overview of the data before the analysis will be conducted. Afterwards, to predict the health status using income a supervised machine learning approach will be applied, more specifically, an ordinal logistic regression. This method is adequate since both variables, income group and health status, are ordered categorical variables. If the method does not perform well (e.g. fitting problems), it has to be checked if there are better methods like random forests or gradient-boosted trees. To improve the prediction age will be used as a second feature for the model next to the income. It can be expected that there is a relationship between the age of an individual and its health condition.

3 Data

3.1 Data Collection

The data which will be used consists of two survey datasets from Alexander-Haw et al. (2024), which is provided for use on Zenodo. The data was collected for the project called *Fundamental Decarbonisation Through Sufficiency By Lifestyle Changes* (FULFILL). However, it not only contains information about the carbon footprint of the respondents, but also detailed information about socioeconomic factors such as age, income and health status. The survey was conducted in the year 2022 from August 18th to September 5th in six countries: Denmark, France, Germany India, Italy and Latvia. The data will contain two of these six survey datasets: one from Germany (N=2028) and one from France (N=2150). These two countries will be selected and taken together due to the fact that they can be compared on an economical and on a health-related scale. Taking together these two datasets will increase the number of observations which can be analyzed and hence the quality of the analysis.

3.2 Data Cleaning

In order to prepare the data for analysis a few data cleaning measures are taken. In a first step, the variable *health status* will be recoded. In the survey the category 1 means the highest level of health and 5 the lowest. For a better understanding in comparison to the variable *income group*, where 1 is the category with the lowest income and 5 the one with the highest, the health status will be recoded accordingly. In a second step, the variable *age* will be checked for implausible outliers (1 outlier with a implausible value of 750). The implausible value(s) will be removed. Afterwards, the missing values will be checked and removed with considering important issues (e.g. proportion of missing values). In a next step, the format of the variables will be checked and if necessary changed. Subsequently, the datasets will be reduced to the necessary variables (subsetting). Finally, if the datasets of both countries are cleaned, they will be merged together.

The cleaned data set contains the variables listed in the following table.

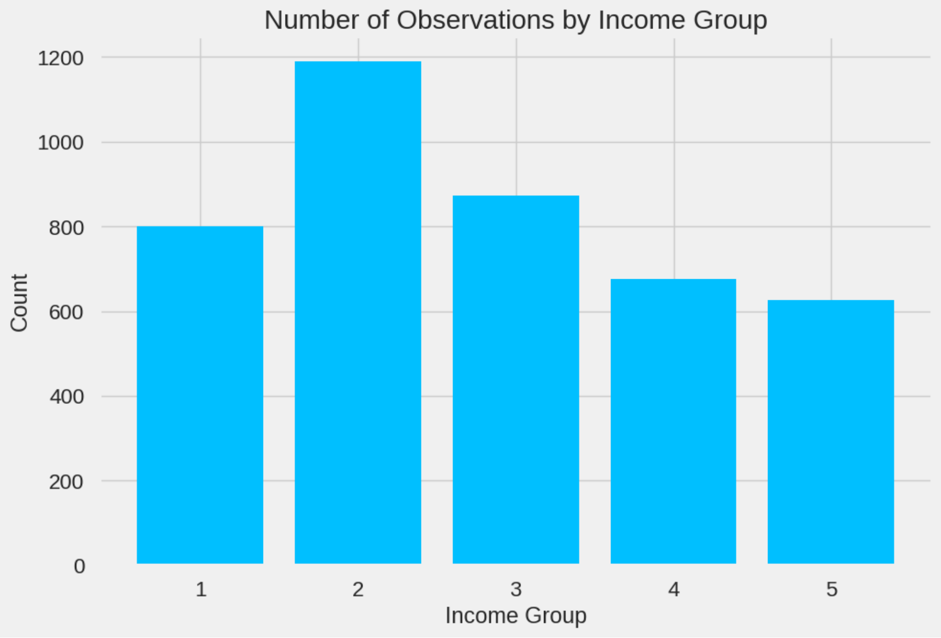
Table 1: Variables

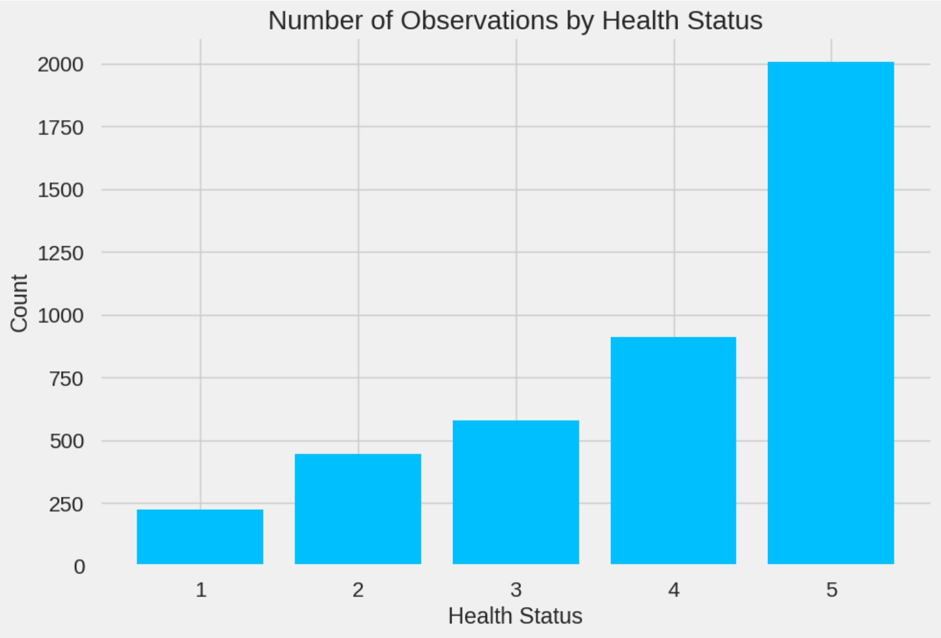
Variable Name	Data Type	Domain
Income Group	Categorical	1: < 15'600 Euro 2: 15'600 - 31'200 Euro 3: 31'200 - 43'200 Euro 4: 43'200 - 60'000 Euro 5: > 60'000 Euro
Health Status	Categorical	1: Very Dissatisfied with Health 2: Dissatisfied with Health 3: Neither nor 4: Satisfied with Health 5: Very Satisfied with Health
Age	Numeric	Min: 18 Median: 51 Max: 96

3.3 Data Overview

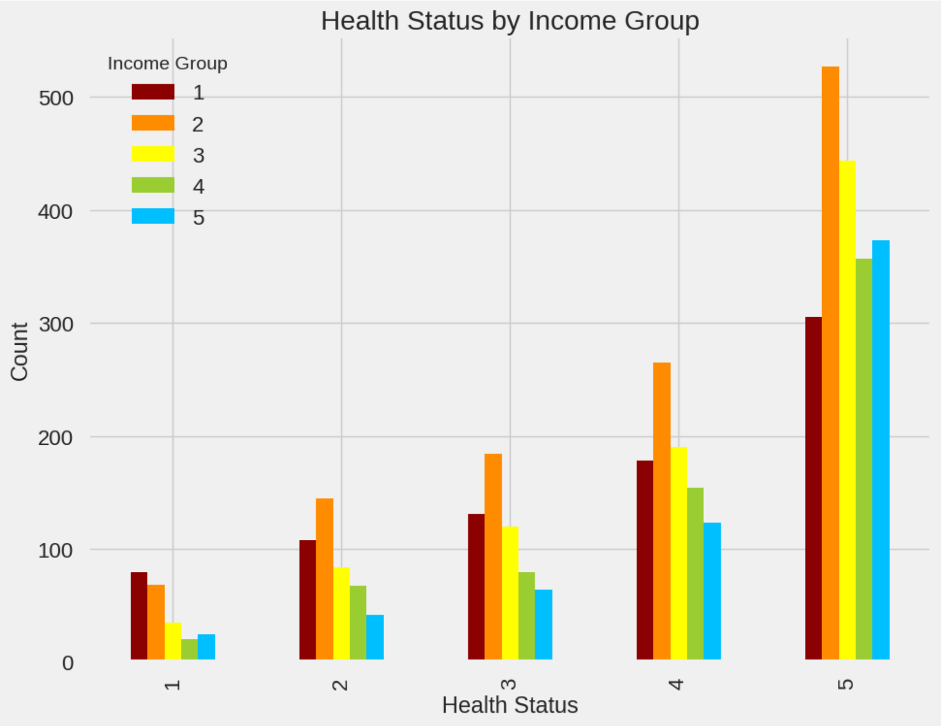
The final dataset includes the three variables listed in the table above: *income group*, *health status* and *age*. To get a better understanding of these three main variables their distribution will be described with the help of plots in the following.

The first two plot show the number of observations by income group and by health status.

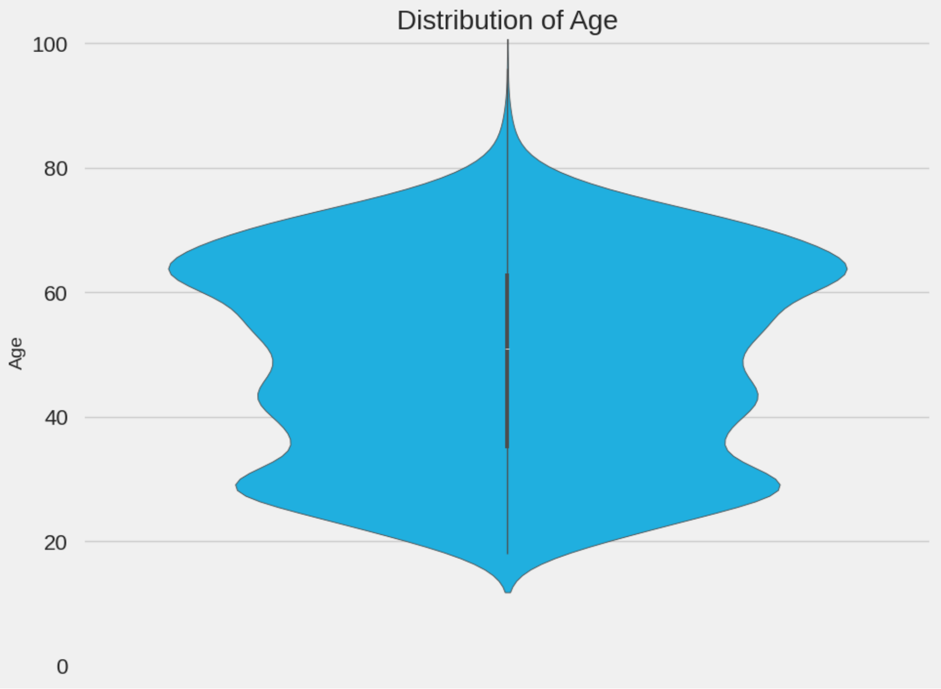


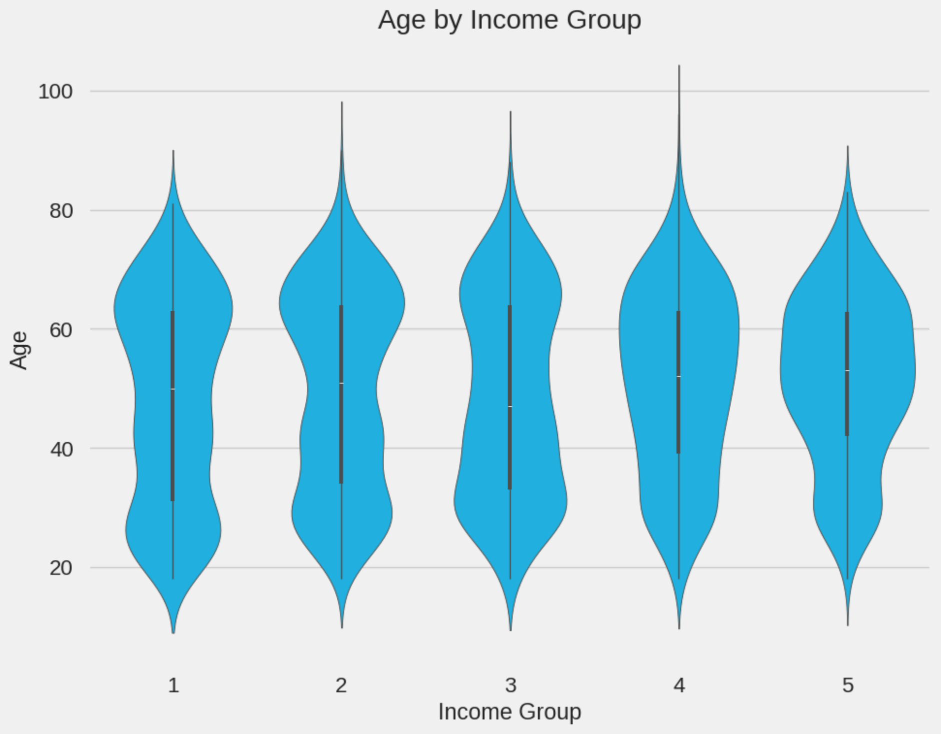


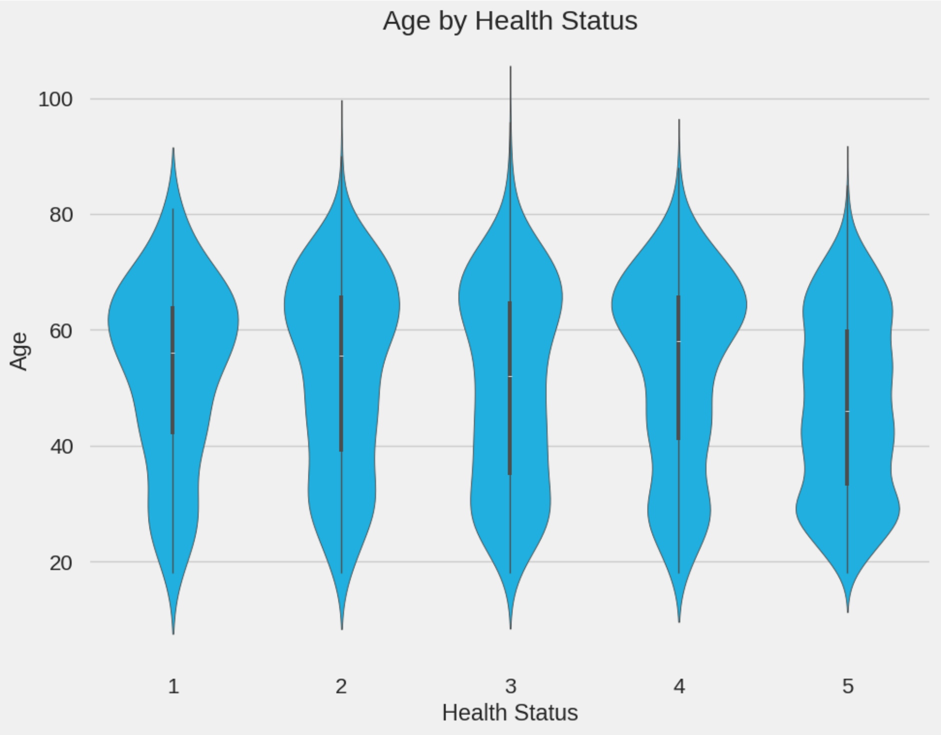
The third plot shows the number of observations per health status by income group.



The following plots show the distribution of the age, additionally by income group and health status.


Distribution of Age
100
80
60
40
20
0
Age


Age by Income Group
100
80
60
40
20
Age
1
2
3
4
5
Income Group



4 Metadata

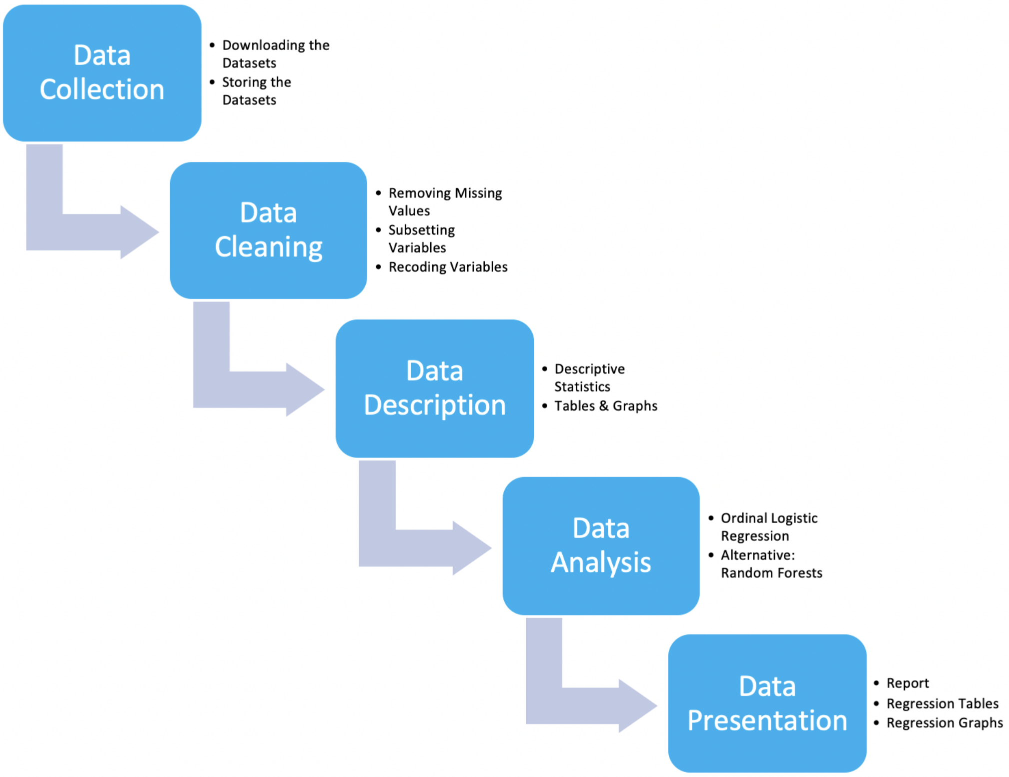
As metadata serves, among other things, the codebook from the used survey datasets. The codebook will be reduced to the used variables and also the recoding will be applied to the codebook, that it fits to the actual dataset used for the analysis. Also the IPython Notebook will be provided to reproduce the analysis. The datasets itself can be downloaded from Zenodo. The metadata will be stored in a GitHub repository. By request a link to the repository will be provided. The repository will be publicly accessible.

5 Data Quality

A main quality requirement in order to meet the project objectives is the number of observations in the dataset. In a first step, the observations from France and Germany will be used. If this number is not big enough for the model to perform well, it can be enhanced. The survey was also conducted in Denmark, India, Italy and Latvia. If the there are not enough observations by using only the France and Germany datasets, some more countries could be added. However, they differing somehow in the economic and health-related circumstances (e.g. India), which could have a negative effect on the analysis quality. Denmark and Italy could be suitable to add. India and Latvia would be probably less suitable. Another important point is the number of missing values, since both *income group* (no missing values) and *health status* (0.2%) do not show a high proportion of missing values there should be no issue with this.

6 Data Flow

Firstly, the two survey datasets will be downloaded from Zenodo and afterwards will be stored on Google Drive as a csv-file to access it with the IPython notebook on Google Colaboration. In a second step, the two datasets will be cleaned. This process step entails removing missing values, looking for outliers or implausible values, selecting the relevant variables and recode the variables if necessary. The two cleaned datasets will be merged to one final dataset. Subsequently, the relevant variables in the cleaned dataset will be described through descriptive statistics as distribution tables and graphs. As a next step, the data will be analyzed by applying an ordinal logistic regression model to examine the relationship between income and the health status. If the regression model does not fit well enough (see Methods), it has to be checked if the usage of random forests would be more adequate. Finally, the results of the analysis will be presented using regression tables and graphs. Additionally, the results will be written down in a report.



7 Data Model

On the conceptual level, the created dataset provides information to predict the health status of an individual by its income, considering the age as well. This will help to get a better understanding of the relationship between the socioeconomic status and the health condition in a society. The relevant information about the logical level of the data model (e.g. data types of the used variables) is described in section 3. On the physical level, the data is stored on Google Drive as a csv-File. Since the dataset contains a relatively low number of observations there should be no additional physical infrastructure necessary.

8 Documentation

All the relevant files will be stored in a GitHub Repository. For a better understanding of the repository, it will contain a proper description with all relevant information about the files and the project generally. Furthermore, there will be a final report as PDF-file (see also Section 6).

9 Risks

The quality of the survey data can be flawed. If this is the case, then I would have to look out for similar data, which also contains information about the income and the health status on an individual level. This would take some additional time for the project and therefore would enhance the project costs. First analyses show that the ordinal logistic regression model is biased towards the health status '5', due to the fact that a big part of the observations are located in this group. So the project may take more time to find an adequate statistical model, which would also lead to higher project costs. Therefore, these two possible cost risks have to be taken into account when the prospective project costs will be estimated.

10 Conclusions

The presented concept outlined the aim of the project and the prerequisites that have to be met for a successful execution of the planned project, such as the data basis, the statistical methods or the analytical process. Assuming that all prerequisites will be met, the presented research project could provide an important contribution to the understanding of the relationship between socioeconomic and health-related conditions of a society. The main obstacle is to control whether the data quantity and quality is high enough for the planned analysis. This has to be checked as soon as possible. Afterwards, the main task is to find the best fitting statistical model to describe the relationship between the analyzed variables. Since the data is already available, the project could be started any time soon. Moreover, the data is based on the first round of the survey. In future, next survey rounds could be added to improve the quality of the analysis.

11 Statement

”Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.”

Date: 4.10.2024 Signature:

References

Alexander-Haw, A. et al. (2024a). *FULFILL dataset round 1 France [Data set]*. Zenodo.

— (2024b). *FULFILL dataset round 1 Germany [Data set]*. Zenodo.

Bundesamt für Sozialversicherungen (2023). *Kurzfassung zum Detailkonzept des nationalen Armutsmonitorings.*

Wilkinson, Richard G. and Kate E. Pickett (2009). "Income inequality and social dysfunction". In: *Annual review of sociology* 35.1, pp. 493–511.