Income Inequality and Its Impact on Health and Life Expectancy Across Socioeconomic Groups in the Netherlands

Sarah Driessens 2854343 Luke Eising 2645467 Xander von Freytag Drabbe 2859036 \Tijs Reijling 28542

2025-06-24

Set-up your environment

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(rmarkdown)
library(yaml)
library(dplyr)
library(cbsodataR)
library(sf)
```

```
## Linking to GEOS 3.13.1, GDAL 3.11.0, PROJ 9.6.0; sf_use_s2() is TRUE
```

```
library(readr)
```

Title Page

Sarah Driessens, Luke Eising, Xander von Freytag Drabbe,
Tijs Reijling, Haris Semen, Bas van Wijk & Selim Yakali

Tutorial group 4

C. Schouwenaar

Part 1 - Identify a Social Problem

1.1 Describe the Social Problem

Topic: How does income inequality contribute to differences in health and life expectancy between high- and low-income groups in the Netherlands?

In the Netherlands, healthcare is universally granted to all citizens through mandatory health insurance. The government promotes healthcare equality through regulations that ensures equal access to services by prohibiting insurance companies to decline clients. Yet, low-income groups in the Netherlands still experience worse health outcomes than their wealthier counterparts.

Research by the Ministry of Health, Welfare and Sport shows that college or university-educated people, which correlates with a higher income, consistently score higher on health outcomes. VZinfo found that people with primary or VMBO-level education were significantly more likely to smoke, have obesity, and rate their health as "poor" compared to their higher educated peers (VZinfo, 2023).

Institutions like the CBS document imbalances between income level and life expectancy. Their analyses show that the wealthiest group in society lives, on average, eight years longer and spend 25 more years in good health than those in the lowest-income group. (Centraal Bureau voor de Statistiek [CBS], 2022).

Pharos provides us with more disparities by emphasising socio-economic differences in health. According to Pharos, receiving welfare benefits correlates with poorer health. They state that with each step up on the social ladder, their chance of good health increases (Pharos, 2022). These health differences across different income groups undermine social cohesion, which makes health inequality based on ones health a social problem that needs to be examined.

Why is this relevant? Wealth inequality in the Netherlands is growing, making these differences even more apparent. Factors like increasing housing costs, job security and education gaps contribute to the widening differences between low- and high-income individuals.

This debate about health inequality isn't new. Some policy makers have been proposing to cut deductibles (eigen risico) since they contribute to people with lower incomes delaying their care. We aim to analyse this inequality to quantify how large these differences are and identify their potential causes.

What has not been examined extensively is whether this inequality holds true across Dutch municipalities. Our aim is to analyse if this inequality hold true throughout the Netherlands, and identify regional differences in the relationship between wealth and health.

Part 2 - Data Sourcing

2.1 Load in the data

Google Drive URL with all data

```
setwd("~/GitHub/Programmeren/data")

Welzijn_Goed <- read_csv("Welzijn.goed.csv")

Ervaren_Gezondheid <- read_csv("Ervaren_Gezondheid.csv")

Levensverwachting <- read_csv("Levensverwachting.csv")

Ervaren_gezondheid_wijk <- read_delim("Ervarengezondheid_Wijk&Buurt.csv", delim = ";")
```

```r
Inkomen_per_gemeente <- read_delim("Inkomen_gemeente.csv", delim = ";")

Levensverwachting_Gemeente <- read_delim("Levensverwacht_Gemeente_Wijk&Buurt.csv", delim = ";")

ErvarenGezondheidNL <- read_csv("ErvarenGezondheidNL.csv")
```

2.2 Provide a short summary of the datasets

Welzijn dataset

```r
head(Welzijn_Goed)
```

```
## # A tibble: 6 x 69
##      ID Kenmerken Marges  Perioden ScoreGeluk_1 Ongelukkig_2
##   <dbl> <chr>     <chr>   <chr>           <dbl>        <dbl>
## 1     0 T009002   MW00000 2013JJ00          7.7          2.5
## 2     1 T009002   MW00000 2014JJ00          7.7          2.4
## 3     2 T009002   MW00000 2015JJ00          7.7          2.8
## 4     3 T009002   MW00000 2016JJ00          7.7          2.6
## 5     4 T009002   MW00000 2017JJ00          7.7          2.8
## 6     5 T009002   MW00000 2018JJ00          7.7          2.8
## # i 63 more variables: NietGelukkigNietOngelukkig_3 <dbl>, Gelukkig_4 <dbl>,
## #   ScoreTevredenheidMetHetLeven_5 <dbl>, Ontevreden_6 <dbl>,
## #   NietTevredenNietOntevreden_7 <dbl>, Tevreden_8 <dbl>,
## #   ScoreTevredenheidOpleidingskansen_9 <dbl>, Ontevreden_10 <dbl>,
## #   NietTevredenNietOntevreden_11 <dbl>, Tevreden_12 <dbl>,
## #   ScoreTevredenheidMetWerk_13 <chr>, Ontevreden_14 <chr>,
## #   NietTevredenNietOntevreden_15 <chr>, Tevreden_16 <chr>, ...
```

This data set contains and presents figures on the wellness/well-being of the population of the Netherlands aged 18 years or older. This factors in happiness and satisfaction with life, satisfaction with education, work, travel, daily activities, weight, financial situation, housing, living environment social life and total amount of time spent on leisure. In addition, concerns about financial future, feelings of insecurity and trust in others are included. This data can be categorized by gender, age, education level and origin.

URL of the data download page + Metadata

Ervaren Gezondheid dataset

The table contains estimated percentages of indicators related to health, social situation, and lifestyle at neighborhood, district, and municipal levels, based on a conducted survey.

URL of the data download page + Metadata

Levensverwachting dataset

This dataset provides us with information about the life expectancy across different periods between 1996-2022.The numbers are provided on a national and municipal level. The data at the municipal level are calculated on the basis of a 4-year period.

URL of the data download page + Metadata

Ervaren Gezondheid 'Wijken' & 'Buurten' dataset

The table contains self-reported health indicators through a questionnaire about health, social situation, and lifestyle at neighborhood, district, and municipal levels. The average of all people living in a municipality or neighbourhood is computed.

URL of the data download page + Metadata

Inkomen Gemeentes dataset

For each household within a municipality or neighbourhood, the total household income was calculated and then divided by the number of household members. This results in the average income per person, which accounts for differences in household size and provides a more accurate measure of individual economic well-being.

URL of the data download page + Metadata

2.3 Describe the type of variables included

Our data sets come from the CBS, an independent administrative body of the Dutch government. CBS is responsible for providing the public with unbiased, statistical information. Their data sets have a wide sample pool which ensure reliability and relevance for our project.

Our sources provide us with data about life expectancy, experienced health (ervaren gezondheid), and income at national, municipal, and neighbourhood levels. These datasets complement each other well because of their recent publication dates (2020+), making them very suitable for integrating them in our analysis. they also provide us with information about the health and income/wealth of Dutch citizens on different levels, allowing us to assign one indicator of health (like experienced health or life expectancy), with an income indicator (like income or net worth). Someone's health or wealth is very difficult to quantify, because there are many different factors at play. Why we decided on these datasets is because we think that they provide us with the most relevant variables when wanting to measure those things.

Despite their recent publication, they are still collected in different years. Large global or national events (like COVID-19) may have affected the data. These events could impact our data, skewing our results. Also life expectancy and Experienced health are not perfect indicators for ones health. For example, life expectancy is heavily impacted by one's genetics, so income might have little influence in this. The life expectancy could also vary due to accidents. If one group of people get in fatal accidents more often, than this could also impact their life expectancy. experienced health also has an obvious flaw, they are based of reports that are not made by an export. That means that the values gives by participants is inherently inaccurate.

Part 3 - Quantifying

3.1 Data cleaning

To prepare the data for analysis and visualization, we looked at multiple datasets retrieved from CBS, each focusing on different aspects of well-being, such as social participation, perceived health, and life satisfaction. The datasets were joined based on shared keys like `GemeenteCode` (Municipality code), `Perioden` (Year), and `Geslacht` (sex), allowing us to create a consistent panel structure. This merging approach was chosen because it preserved all available information while ensuring that each observation aligned across dimensions of time, gender, and region.

Several cleaning steps were implemented to ensure data consistency. We selected only observations where `SoortRegio_2` indicated a municipality (`"Gemeente"`) and removed unnecessary columns such as margins

("Marge") and region type indicators. New columns were created to standardize region names and construct a custom variable for gender (GeslachtCustom) to simplify filtering. These steps allowed us to make our plots more intuitive and focused.

During this process, we identified structural inconsistencies in region naming conventions and variations in column formats across datasets. We addressed these by normalizing municipality names (e.g., converting to lowercase) and aligning column names prior to merging. Missing values were handled carefully, with averages computed using na.rm = TRUE where appropriate to avoid bias in index construction.

In an ideal situation, standardized datasets with harmonized column names and formats would reduce the need for extensive preprocessing. However, given the raw CBS export format, our manual selection and transformation were necessary to produce clean and analyzable data.

Overall, these cleaning and merging operations ensured a coherent dataset ready for temporal and subpopulation visualizations, enabling meaningful insights on well-being across Dutch municipalities.

Please use a separate 'R block' of code for each type of cleaning. So, e.g. one for missing values, a new one for removing unnecessary variables etc.

3.2 Generate necessary variables

1st new variable

We sought to be able to compare different sub-populations based on their (self-reported) well-being through a questionnaire. In the datasets we used the year periods that were both from 2015 till 2018, perfect for our analysis. However, we transformed the yearly statistics from 2015 till 2018 into an average over the 4 years and gave it the title of "2016". There were plans to merge with another dataset (Ervaren_Gezondheid) who did have the year 2016 and some other keys that aligned. This merge happened based on sex but did not end up being used for any graphs or analysis that we decided to keep in the rapport. So a note has to be made that the value that is used and given the name '2016' is really the average over those 4 years.

After the averages were taken, we still had 65 columns of data. A lot of the columns were on different scales so we decided to create a new variable: Our 'WelzijnIndex'. We created a singular value from 7 of the columns of data all on the same 1 through 10 scale. This created a value that could be used to compare a lot of different sub-populations based on their well-being.
One of which we most importantly will use in 3.5, where we visualize sub-population variation based on level of education.

```
Welzijn_averages <- Welzijn_Goed %>%
  group_by(Kenmerken, Marges) %>%
  summarise(
    Perioden = "2015-2018",
    across(where(is.numeric), mean, na.rm = TRUE),
    .groups = "drop" )
```

```
Welzijn_index_2016 <- Welzijn_index_2016 %>%
  mutate(WelzijnIndex = rowMeans(select(.,
    ScoreGeluk_1,
    ScoreTevredenheidMetHetLeven_5,
    ScoreTevredenheidMetWerk_13,
    ScoreTevredenheidDagelijkseBezigheden_21,
    ScoreTevredenheidMetLichGezondheid_25,
    ScoreTevredenheidPsychischeGezondheid_29,
    ScoreTevredenheidMetSociaalLeven_57
  ), na.rm = TRUE))
```

2nd new variable

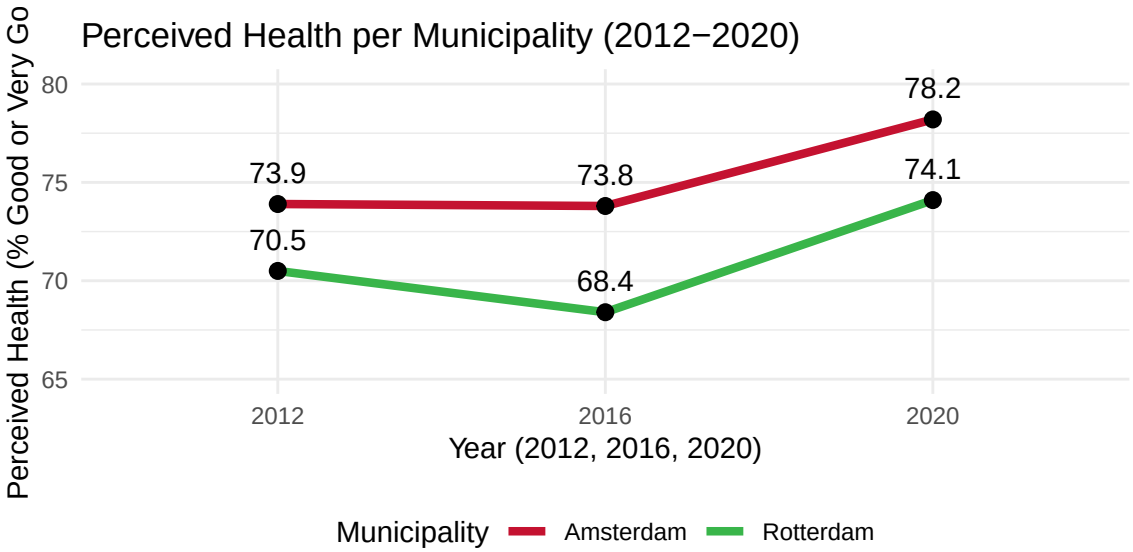
Correlation Between Income and Experienced Health:
Our second variable calculates the correlation between experienced health and average income within each municipality. The two columns selected are: ErvarenGezondheidZeerGoed_4" (the percentage of people in a neighbourhood who grade their own health as good, or very good), and "Gemiddeld" (the average yearly wage in a neighbourhood divided by €1000).

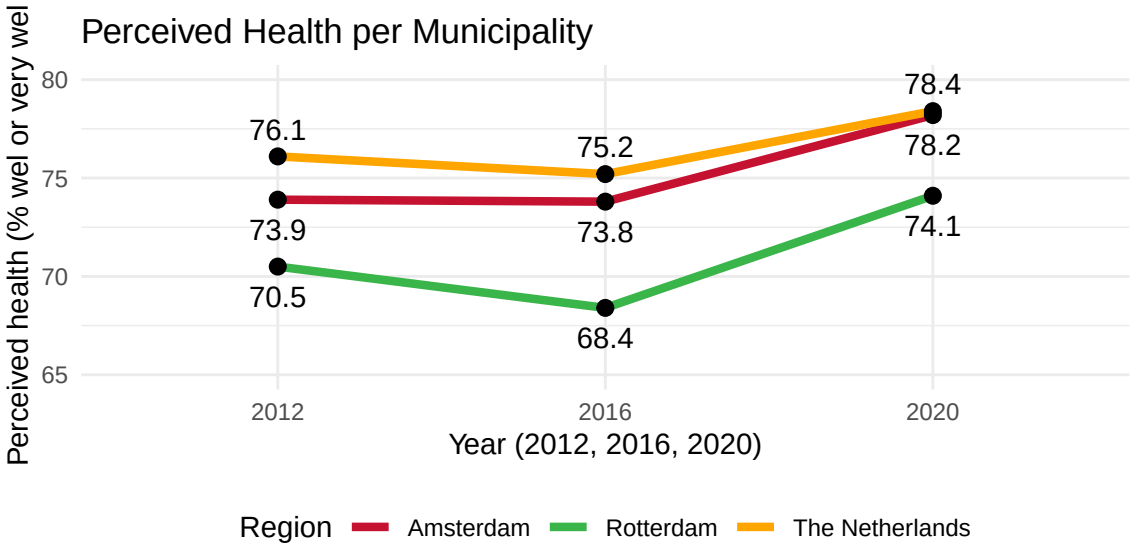
These variables were passed through the "cor()" function to calculate the correlation between these two variables for each municipality. The correlation was only computed for municipalities that had three or more neighbourhoods (with valid data).

The decision to create this variable was to map how the relationship between health and income varies across municipalities in the Netherlands. This variable can be visualised on a heat map, helping us understand where and how income inequality contributes to differences in health.

```
gemeente_cor <- gemeente_wijk_data %>%
  group_by(statnaam) %>%
  filter(n() >= 3) %>%
  summarise(
    correlation = if (sum(complete.cases(as.numeric(Gemiddeld, ErvarenGezondheidGoedZeerGoed_4))) > 1)
      {cor(as.numeric(Gemiddeld), as.numeric(ErvarenGezondheidGoedZeerGoed_4), use = "complete.obs")}
    else {NA_real_})
```

3.3 Visualize temporal variation

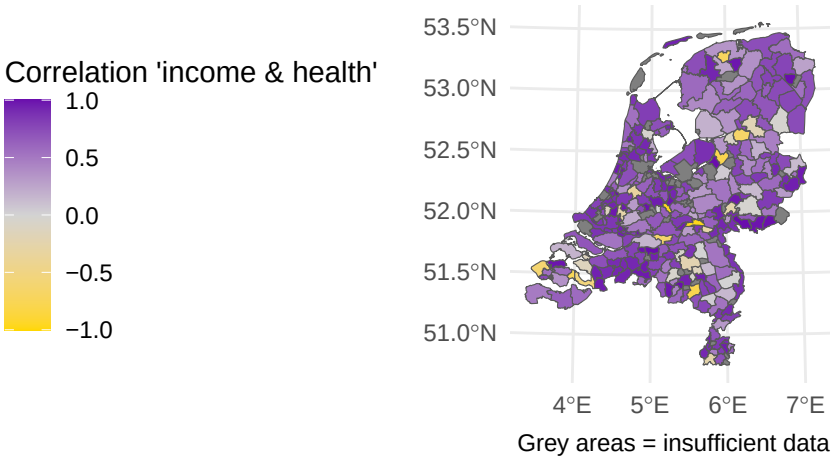




The Perceived Health per Municipality shows how the perceived health in different regions fluctuates across different time points. We see clear temporal patterns, with periods of increase and decrease indicating variability over time. This suggests that the underlying phenomenon is not static, but undergoes notable changes, highlighting the importance of considering temporal dynamics in the analysis.

Overall, we can see that the self-reported Perceived Health in the Netherlands hovered around 75-76% between the years 2012 and 2016, however, we can see a big increase between the years 2016-2020. This variation will be looked at in the "event analysis" section, to figure out why the changes can be so drastic on municipality level in relation to country-wide level.

3.4 Visualize spatial variation



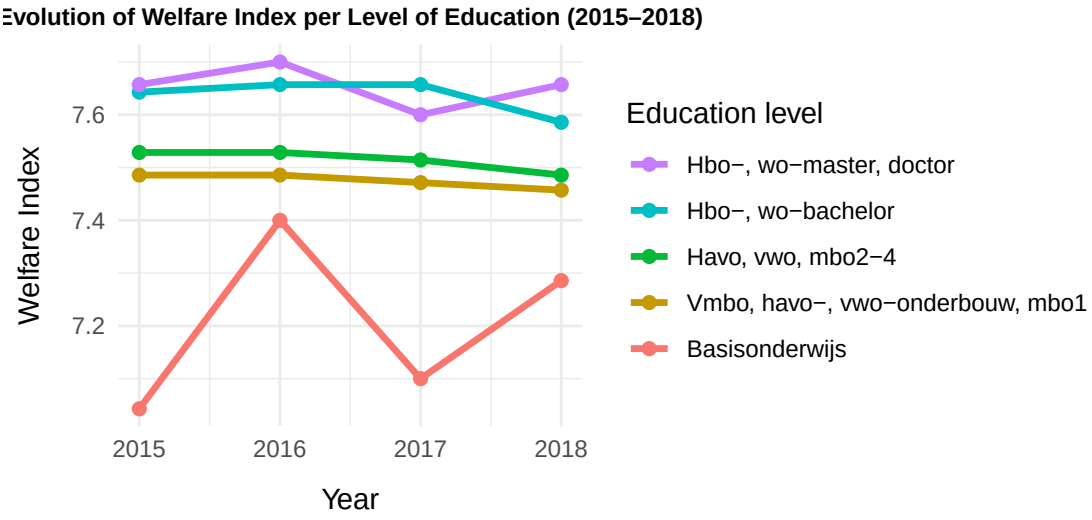
This heat map shows the strength of the correlation between the average income and experienced health across neighbourhoods within each municipality in the Netherlands. Municipalities that include fewer than three neighbourhoods with valid data are marked in grey. For each municipality with sufficient data (three or more neighbourhoods), a correlation coefficient was calculated using the "cor()" function. High correlation areas (dark purple) are mainly found in urban areas like Rotterdam and Groningen, whereas low correlation areas (yellow) are mainly found in rural areas such as Kapelle, Veere and Reimerswaal in Zeeland. These yellow areas are rare and are likely caused by data inconsistencies or an anomalies.

The prevalence of purple on the map indicates a moderate positive correlation between income and experienced health in most municipalities. Although it should be noted that correlation does not equal causation, we can confirm there exists a relation between the two.

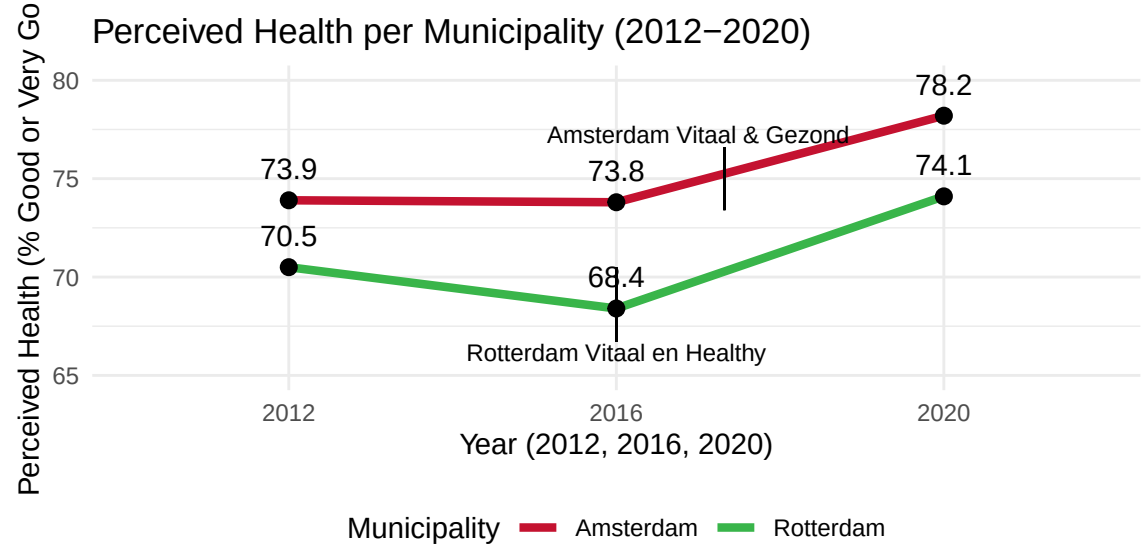
One limitation of this map is the existence of grey areas that could not be computed due to a lack of data. In general, correlations are more accurate when they are based of a larger dataset.

Therefore, we considered zooming out the map out and display it at the provincial level. However, we decided to not go with this approach, because it would lead to underrepresentation of urban areas by assigning them the same weight as less populated municipalities within each province. This issue could be solved by assigning weight to municipalities, but this would over complicate things. Additionally, keeping the map at the municipal level allows for a clearer visualisation of contrast between urban and rural areas.

3.5 Visualize sub-population variation



3.6 Event analysis



As indicated in the "Perceived Health per Municipality" diagram, before the 2020 reports, the "self-reported health" in 2016 and 2012 hovered around 74%. In 2017, the municipality of Amsterdam (in collaboration with many health insurers such as Zilveren Kruis, Sigra, Elaa, and others) started the initiative called "Amsterdam Vitaal & Gezond". This initiative aims to ensure all residents have equal access to good health by 2040, adding 10 healthy life years for vulnerable groups through integrated care, digital solutions, prevention, and community partnerships. As seen by the self-reported health in 2020, the population of Amsterdam positively responded to this initiative, bringing the perceived health up to 78.2%.

Similarly, Rotterdam's reported perceived health in 2012 and 2016, respectively, was 70.5% and 68.4%, in contrast to Amsterdam's perceived health, dropping a handful of points throughout the 4 years. The municipality of Rotterdam had started its initiative, organized by VITR in collaboration with Erasmus, Hogeschool Rotterdam, and other health-focused institutes, in 2016. Similarly to Amsterdam's initiative

results, the perceived health drastically increased from 68.4% in 2016 to 74.1% in 2020, showcasing a clear positive reaction by the general public towards the initiatives.

Part 4 - Discussion

Despite the Dutch government's commitment to universal healthcare and equal access to services, significant health disparities remain between income groups. Research consistently shows that individuals with higher income and educational attainment experience better health and live longer than those in lower-income brackets (CBS, 2022; VZinfo, 2023). These inequalities, further emphasized by Pharos (2022), demonstrate that socioeconomic status remains a key determinant of health, even within a system designed to promote equity. Growing wealth inequality, driven by rising housing costs, job insecurity, and educational disparities, continues to widen this health gap. Addressing it requires more than general policy reforms; it demands a deeper understanding of how these disparities play out locally. By examining regional differences, we can better identify where and why these inequalities are most severe and design targeted interventions to ensure that all citizens, regardless of income, have the opportunity to live a healthy life.

Part 5 - Reproducibility

5.1 Github repository link

https://github.com/TijsReijling/Programmeren

5.2 Reference list

Data set sources include links to their corresponding metadata.

#Bovenaan welvaartsladder bijna 25 jaar langer in goede gezondheid. (2022, December 20). Centraal Bureau Voor De Statistiek. https://www.cbs.nl/nl-nl/nieuws/2022/51/bovenaan-welvaartsladder-bijna-25-jaar-langer-in-goede-gezondheid?

#Pharos. (2022, July). Sociaal economische Gezondheidsverschillen (SEGV). https://www.pharos.nl/factsheets/sociaaleconomische-gezondheidsverschillen-segv/

#Sociaaleconomische gezondheidsverschillen | Volksgezondheid en Zorg. (n.d.). https://www.vzinfo.nl/sociaaleconomische-gezondheidsverschillen

Gezonde levensverwachting; inkomen en welvaart. (2025, May 23).[Data set]. CBS dataportaal. https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=85445NED&_theme=154

Welzijn; kerncijfers, persoonskenmerken. (2025, March 20).[Data set]. CBS dataportaal. https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=85542NED&_theme=178

Inkomen per gemeente en wijk, 2020. (2023, September 1).[Data set]. Centraal Bureau voor de Statisktiek. https://www.cbs.nl/nl-nl/maatwerk/2023/35/inkomen-per-gemeente-en-wijk-2020 (metadata included in dataset)

Gezondheid per wijk en buurt; 2012/2016/2020/2022. (2024, December 09). [Data set]. RIVM dataportaal. https://statline.rivm.nl/portal.html?_la=nl&_catalog=RIVM&tableId=50120NED&_theme=94

Levensverwachting op de leeftijd 0 en 65 jaar; geslacht, regio 1996-2022. (2024, December 9). [Data set]. RIVM dataportaal. https://statline.rivm.nl/portal.html?_la=nl&_catalog=RIVM&tableId=50132NED&_theme=101