Template

Studentnames and studentnumbers here

2025-06-19

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

Include your names

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

```
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 = ";")

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

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

ErvarenGezondheidNL <- read_csv("ErvarenGezondheidNL.csv")

gemeentegrenzen <- st_read("https://service.pdok.nl/cbs/gebiedsindelingen/2023/wfs/v1_0?request=GetFeatu
```

```
## Reading layer ‘gemeente_gegeneraliseerd’ from data source
##   ‘https://service.pdok.nl/cbs/gebiedsindelingen/2023/wfs/v1_0?request=GetFeature&service=WFS&versio
##   using driver ‘GeoJSON’
## Simple feature collection with 342 features and 5 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 13565.4 ymin: 306846.2 xmax: 278026.1 ymax: 619231.6
## Projected CRS: Amersfoort / RD New
```

```
gemeentegrenzen <- sf::st_read("https://service.pdok.nl/cbs/gebiedsindelingen/2023/wfs/v1_0?request=Get
```

```
## Reading layer ‘gemeente_gegeneraliseerd’ from data source
##   ‘https://service.pdok.nl/cbs/gebiedsindelingen/2023/wfs/v1_0?request=GetFeature&service=WFS&versio
##   using driver ‘GeoJSON’
## Simple feature collection with 342 features and 5 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 13565.4 ymin: 306846.2 xmax: 278026.1 ymax: 619231.6
## Projected CRS: Amersfoort / RD New
```

2.2 Provide a short summary of the datasets

Welzijn dataset

```
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'

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

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.

For the sake of this example, I will continue with the assignment...

Part 3 - Quantifying

3.1 Data cleaning

Say we want to include only larger distances (above 2) in our dataset, we can filter for this.

```
Levensverwachting <- Levensverwachting[grepl("2015G400", Levensverwachting$Perioden),]
Welzijn_Goed <- Welzijn_Goed[Welzijn_Goed$Perioden %in% c("2015JJ00", "2016JJ00", "2017JJ00", "2018JJ00
```

```
Ervaren_Gezondheid <- Ervaren_Gezondheid[
  Ervaren_Gezondheid$Gemeentenaam_1 %in% c("Amsterdam", "Rotterdam") &
    Ervaren_Gezondheid$SoortRegio_2 == "Gemeente",]
```

```
Ervaren_Gezondheid_2016 <- Ervaren_Gezondheid[Ervaren_Gezondheid$Perioden == "2016JJ00", ]
```

```
Levensverwachting_2016 <- Levensverwachting %>%
  filter(Marges == "MW00000", Perioden == "2015G400") %>%
  mutate(Perioden = "2016")
```

```
Welzijn_averages_2016 <- Welzijn_averages %>%
  filter(Marges == "MW00000", Perioden == "2015-2018") %>%
  mutate(Perioden = "2016")
```

```
Levensverwachting_2016_renamed <- Levensverwachting_2016 %>%
  rename(Kenmerken = InkomenEnWelvaart)
```

```
Welzijn_naar_opleidingsniveau <- Welzijn_naar_opleidingsniveau %>%
  select(-Kenmerken) %>%
  relocate(Opleidingsniveau, .before = everything())
```

```
Levensverwachting_geslacht <- Levensverwachting_2016_renamed %>%
  filter(Geslacht %in% c(3000, 4000))
```

```
Levensverwachting_geslacht <- Levensverwachting_geslacht %>%
  mutate(
    Geslacht = case_when(
      Geslacht == 3000 ~ "Mannen",
      Geslacht == 4000 ~ "Vrouwen"
    ))
```

```r
Levensverwachting_geslacht_gemiddeld <- Levensverwachting_geslacht %>%
  group_by(Geslacht) %>%
  summarise(
    Jaar = "2016",
    Levensverwachting = mean(Levensverwachting_1, na.rm = TRUE)
  )
```

```r
Welzijn_naar_geslacht <- Welzijn_index_2016 %>%
  filter(Kenmerken %in% c(3000, 4000)) %>%
  mutate(
    Geslacht = case_when(
      Kenmerken == 3000 ~ "Mannen",
      Kenmerken == 4000 ~ "Vrouwen"
    )
  ) %>%
  select(Geslacht, Jaar = Perioden, WelzijnIndex)
```

```r
Ultimate_dataset_of_doom_hell_and_destruction <- full_join(
  Levensverwachting_geslacht_gemiddeld,
  Welzijn_naar_geslacht,
  by = c("Geslacht", "Jaar")
)
```

```r
Welzijn_naar_opleidingsniveau <- Welzijn_naar_opleidingsniveau %>%
  filter(Marges == "MW00000") %>%
  select(-Marges) %>%
  rename(Jaar = Perioden)
```

```r
packages <- c("sf", "dplyr", "ggplot2", "readr", "tmap", "stringr")
installed <- rownames(installed.packages())
for (pkg in packages) {
  if (!pkg %in% installed) install.packages(pkg)
}
lapply(packages, library, character.only = TRUE)
```

```
## [[1]]
##  [1] "sf"        "cbsodataR" "yaml"      "rmarkdown" "lubridate" "forcats"
##  [7] "stringr"   "dplyr"     "purrr"     "readr"     "tidyr"     "tibble"
## [13] "ggplot2"   "tidyverse" "stats"     "graphics"  "grDevices" "utils"
## [19] "datasets"  "methods"   "base"
##
## [[2]]
##  [1] "sf"        "cbsodataR" "yaml"      "rmarkdown" "lubridate" "forcats"
##  [7] "stringr"   "dplyr"     "purrr"     "readr"     "tidyr"     "tibble"
## [13] "ggplot2"   "tidyverse" "stats"     "graphics"  "grDevices" "utils"
## [19] "datasets"  "methods"   "base"
##
## [[3]]
##  [1] "sf"        "cbsodataR" "yaml"      "rmarkdown" "lubridate" "forcats"
##  [7] "stringr"   "dplyr"     "purrr"     "readr"     "tidyr"     "tibble"
## [13] "ggplot2"   "tidyverse" "stats"     "graphics"  "grDevices" "utils"
## [19] "datasets"  "methods"   "base"
```

```
##
## [[4]]
##  [1] "sf"        "cbsodataR" "yaml"      "rmarkdown" "lubridate" "forcats"
##  [7] "stringr"   "dplyr"     "purrr"     "readr"     "tidyr"     "tibble"
## [13] "ggplot2"   "tidyverse" "stats"     "graphics"  "grDevices" "utils"
## [19] "datasets"  "methods"   "base"
##
## [[5]]
##  [1] "tmap"      "sf"        "cbsodataR" "yaml"      "rmarkdown" "lubridate"
##  [7] "forcats"   "stringr"   "dplyr"     "purrr"     "readr"     "tidyr"
## [13] "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics"  "grDevices"
## [19] "utils"     "datasets"  "methods"   "base"
##
## [[6]]
##  [1] "tmap"      "sf"        "cbsodataR" "yaml"      "rmarkdown" "lubridate"
##  [7] "forcats"   "stringr"   "dplyr"     "purrr"     "readr"     "tidyr"
## [13] "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics"  "grDevices"
## [19] "utils"     "datasets"  "methods"   "base"
```

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

```
gemeente_shape <- st_read(
  "https://service.pdok.nl/cbs/gebiedsindelingen/2020/wfs/v1_0?request=GetFeature&service=WFS&version=2
  quiet = TRUE)
```

```
gemeente_shape <- gemeente_shape %>%
  select(statcode, statnaam, geometry) %>%
  rename(GM_CODE = statcode, GM_NAAM = statnaam)
```

```
data <- read_csv("data/GemeentesJuist.csv", locale = locale(encoding = "UTF-8"))
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)

## Rows: 156069 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): ID, Marges, WijkenEnBuurten, Perioden, Gemeentenaam_1, SoortRegio_2
## dbl (2): Leeftijd, ErvarenGezondheidGoedZeerGoed_4
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
gemeente_data <- data %>%
  filter(str_to_lower(SoortRegio_2) == "gemeente") %>%
  select(Gemeentenaam_1, ErvarenGezondheidGoedZeerGoed_4)
```

```
gemeente_shape$GM_NAAM <- str_to_lower(gemeente_shape$GM_NAAM)
gemeente_data$Gemeentenaam_1 <- str_to_lower(gemeente_data$Gemeentenaam_1)
```

```r
kaart_data <- gemeente_shape %>%
  left_join(gemeente_data, by = c("GM_NAAM" = "Gemeentenaam_1"))
```

```r
na_count <- sum(is.na(kaart_data$ErvarenGezondheidGoedZeerGoed_4))
cat("Aantal gemeenten zonder data: ", na_count, "\n")
```

```
## Aantal gemeenten zonder data:  1
```

```r
kaart_data <- st_simplify(kaart_data, dTolerance = 100)
```

```r
tmap_mode("plot")
```

```
## i tmap mode set to "plot".
```

```r
kaart_plot <- tm_shape(kaart_data) +
  tm_fill("ErvarenGezondheidGoedZeerGoed_4",
          palette = "Blues",
          title = "Perceived Health per Municipality (%)",
          textNA = "No data") +
  tm_borders() +
  tm_layout(
    title = "Perceived Health per Municipality in 2020",
    title.position = c("center", "top"),
    inner.margins = c(0.12, 0.02, 0.10, 0.02),
    title.size = 1.5,
    title.color = "black",
    legend.outside = TRUE
  )
```

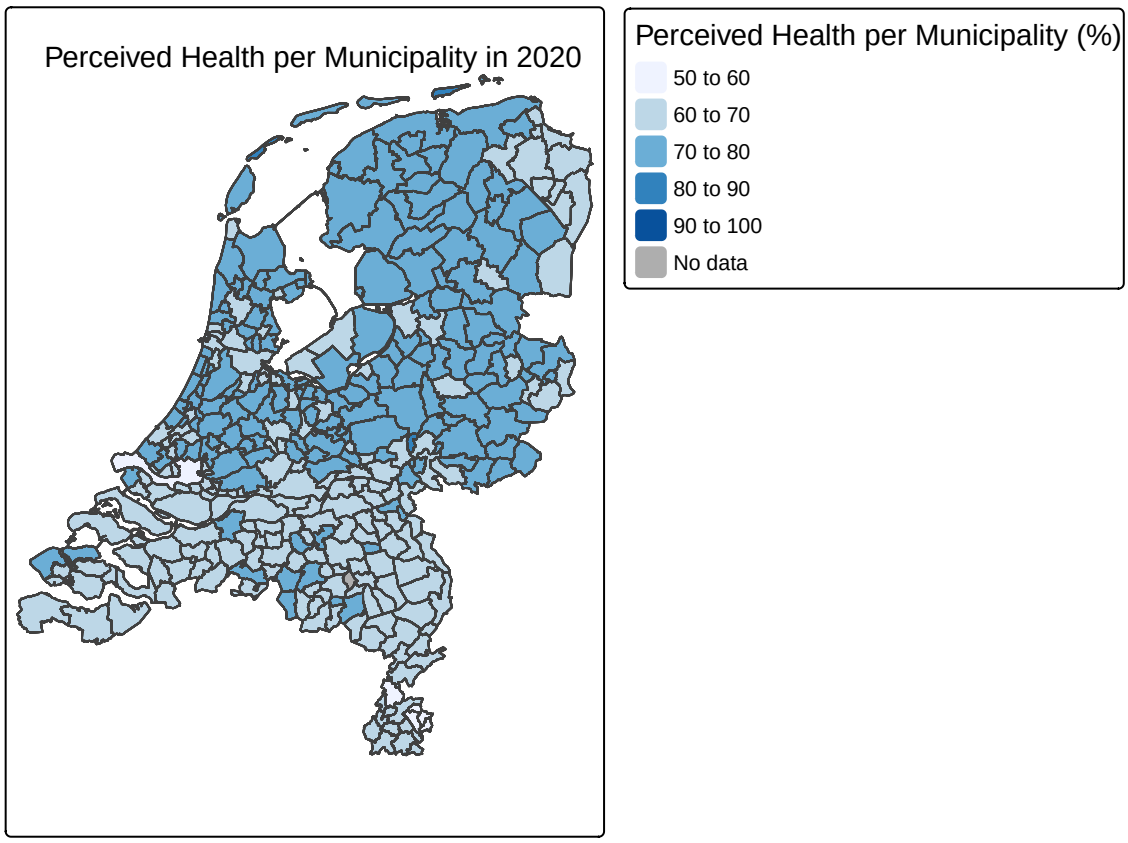
```
##
```

```
## -- tmap v3 code detected -------------------------------------------------
```

```
## [v3->v4] `tm_tm_fill()`: migrate the argument(s) related to the scale of the
## visual variable `fill` namely 'palette' (rename to 'values'), 'textNA' (rename
## to 'label.na') to fill.scale = tm_scale(<HERE>).
## i For small multiples, specify a 'tm_scale_' for each multiple, and put them in
##   a list: 'fill.scale = list(<scale1>, <scale2>, ...)'
## [v3->v4] `tm_fill()`: migrate the argument(s) related to the legend of the
## visual variable `fill` namely 'title' to 'fill.legend = tm_legend(<HERE>)'
## [v3->v4] `tm_layout()`: use `tm_title()` instead of `tm_layout(title = )`
```

```r
print(kaart_plot)
```

```
## [cols4all] color palettes: use palettes from the R package cols4all. Run
## `cols4all::c4a_gui()` to explore them. The old palette name "Blues" is named
## "brewer.blues"
## Multiple palettes called "blues" found: "brewer.blues", "matplotlib.blues". The first one, "brewer.bl
##
## [plot mode] fit legend/component: Some legend items or map compoments do not
## fit well, and are therefore rescaled.
## i Set the tmap option `component.autoscale = FALSE` to disable rescaling.
```




```
Levensverwachting_2016_renamed_baby <- Levensverwachting_2016_renamed %>%
  filter(LeeftijdOp31December == "10010", Geslacht == "T001038") %>%
  mutate(Leeftijd = 0)
```

```
Levensverwachting_2016_renamed_baby <- Levensverwachting_2016_renamed_baby %>%
  select(-Marges, -Geslacht, -LeeftijdOp31December)
```

```
Levensverwachting_2016_renamed_baby <- Levensverwachting_2016_renamed_baby %>%
  rename(Jaar = Perioden)
```

```
welvaart_codes <- c(2021770, 2021780, 2021790, 2021800, 2021810)

Levensverwachting_2016_renamed_baby <- Levensverwachting_2016_renamed_baby %>%
  filter(Kenmerken %in% welvaart_codes)
```

```
Levensverwachting_2016_renamed_baby <- Levensverwachting_2016_renamed_baby %>%
  arrange(Kenmerken) %>%
  mutate(WelvaartQuintiles = c("1st quintile", "2nd quintile", "3rd quintile", "4th quintile", "5th qui
```

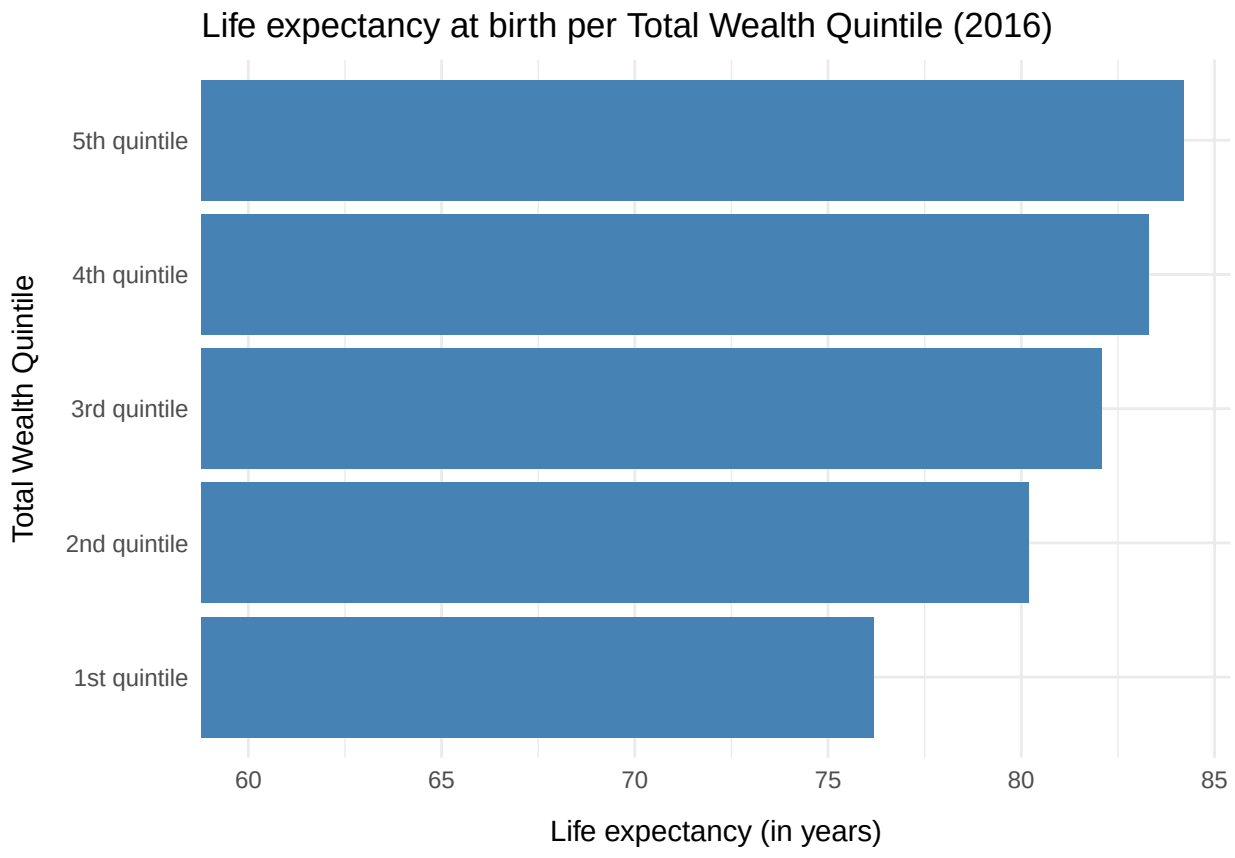
```
Levensverwachting_2016_renamed_baby <- Levensverwachting_2016_renamed_baby %>%
  select(Leeftijd, Jaar, WelvaartQuintiles, Levensverwachting_1)
```

```r
Levensverwachting_2016_renamed_baby$Levensverwachting_1 <- as.numeric(as.character(Levensverwachting_201
str(Levensverwachting_2016_renamed_baby)
```

```
## tibble [5 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Leeftijd           : num [1:5] 0 0 0 0 0
##  $ Jaar               : chr [1:5] "2016" "2016" "2016" "2016" ...
##  $ WelvaartQuintiles  : chr [1:5] "1st quintile" "2nd quintile" "3rd quintile" "4th quintile" ...
##  $ Levensverwachting_1: num [1:5] 76.2 80.2 82.1 83.3 84.2
```

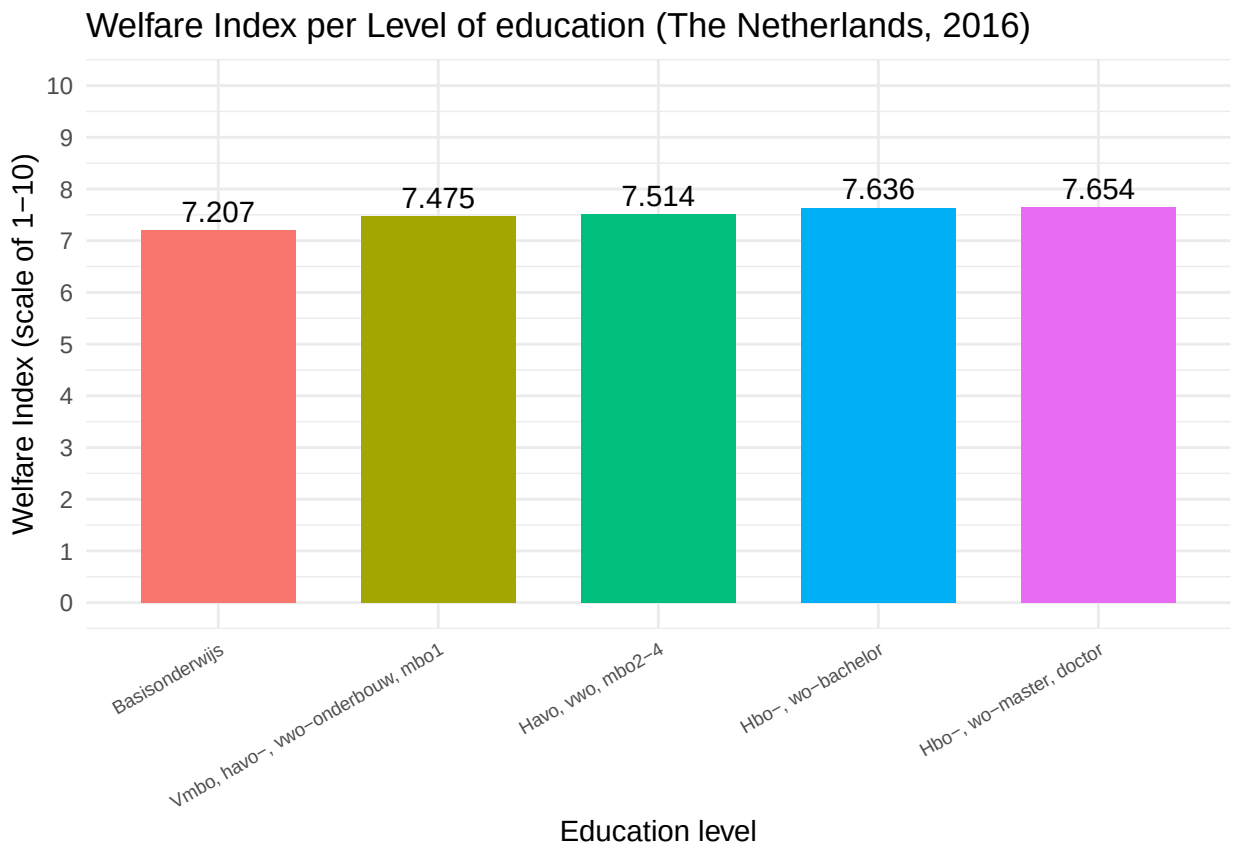
Plots:

```r
ggplot(Levensverwachting_2016_renamed_baby, aes(x = WelvaartQuintiles, y = Levensverwachting_1)) +
  geom_col(fill = "steelblue") +
  coord_flip(ylim = c(60, NA)) +  # <<--- set limit here
  labs(
    title = "Life expectancy at birth per Total Wealth Quintile (2016)",
    x = "Total Wealth Quintile",
    y = "Life expectancy (in years)"
  ) +
  theme_minimal() +
  theme(
    axis.title.y = element_text(margin = margin(r = 10)),
    axis.title.x = element_text(margin = margin(t = 10))
  )
```



```r
Welzijn_naar_opleidingsniveau$Opleidingsniveau <- factor(
  Welzijn_naar_opleidingsniveau$Opleidingsniveau,
  levels = c(
    "Basisonderwijs",
    "Vmbo, havo-, vwo-onderbouw, mbo1",
    "Havo, vwo, mbo2-4",
    "Hbo-, wo-bachelor",
    "Hbo-, wo-master, doctor"
  )
)
```

```r
ggplot(Welzijn_naar_opleidingsniveau, aes(x = Opleidingsniveau, y = WelzijnIndex, fill = Opleidingsnivea
  geom_bar(stat = "identity", width = 0.7) +
  geom_text(
    aes(label = round(WelzijnIndex, 3)),   # 3 decimals
    vjust = -0.4,
    size = 4
  ) +
  scale_y_continuous(
    limits = c(0, 10),
    breaks = seq(0, 10, 1)
  ) +
  labs(
    title = "Welfare Index per Level of education (The Netherlands, 2016)",
    x = "Education level",
    y = "Welfare Index (scale of 1-10)"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1, size = 7),
    legend.position = "none"
  )
```




```
opleidingskenmerken <- c(2018710, 2018720, 2018750, 2018800, 2018810)
```

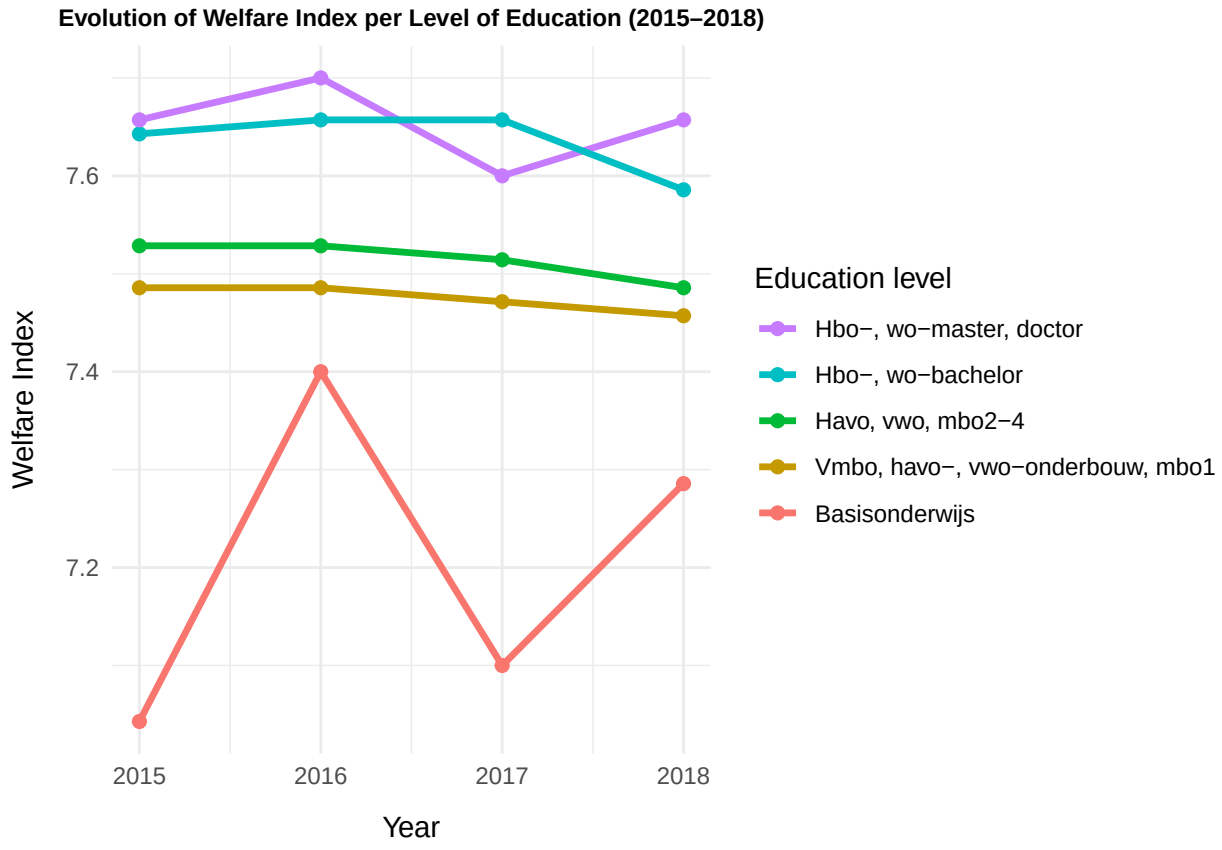
```
Welzijn_temporal_visualization <- Welzijn_Goed %>%
  filter(Kenmerken %in% opleidingskenmerken, Marges == "MW00000") %>%
  mutate(Jaar = as.numeric(substr(Perioden, 1, 4))) %>%
  rowwise() %>%
  mutate(WelzijnIndex = mean(c_across(c(
    ScoreGeluk_1,
    ScoreTevredenheidMetHetLeven_5,
    ScoreTevredenheidMetWerk_13,
    ScoreTevredenheidDagelijkseBezigheden_21,
    ScoreTevredenheidMetLichGezondheid_25,
    ScoreTevredenheidPsychischeGezondheid_29,
    ScoreTevredenheidMetSociaalLeven_57
  )), na.rm = TRUE)) %>%
  ungroup() %>%
  select(Kenmerken, Jaar, WelzijnIndex)
```

```
Welzijn_temporal_visualization <- Welzijn_temporal_visualization %>%
  mutate(
    Opleidingsniveau = case_when(
      Kenmerken == 2018710 ~ "Basisonderwijs",
      Kenmerken == 2018720 ~ "Vmbo, havo-, vwo-onderbouw, mbo1",
      Kenmerken == 2018750 ~ "Havo, vwo, mbo2-4",
      Kenmerken == 2018800 ~ "Hbo-, wo-bachelor",
```

```
      Kenmerken == 2018810 ~ "Hbo-, wo-master, doctor",
      TRUE ~ NA_character_
    )
  ) %>%
  select(Opleidingsniveau, Jaar, WelzijnIndex)
```

```
Welzijn_temporal_visualization$Opleidingsniveau <- factor(
  Welzijn_temporal_visualization$Opleidingsniveau,
  levels = c(
    "Hbo-, wo-master, doctor",
    "Hbo-, wo-bachelor",
    "Havo, vwo, mbo2-4",
    "Vmbo, havo-, vwo-onderbouw, mbo1",
    "Basisonderwijs"
  )
)
```

```
ggplot(Welzijn_temporal_visualization, aes(x = Jaar, y = WelzijnIndex, color = Opleidingsniveau, group =
  geom_line(size = 1.2) +
  geom_point(size = 2) +
  scale_x_continuous(breaks = 2015:2018) +
  scale_color_manual(
    values = c(
      "Basisonderwijs" = "#F8766D",
      "Vmbo, havo-, vwo-onderbouw, mbo1" = "#C49A00",
      "Havo, vwo, mbo2-4" = "#00BA38",
      "Hbo-, wo-bachelor" = "#00BFC4",
      "Hbo-, wo-master, doctor" = "#C77CFF"
    )
  ) +
  labs(
    title = "Evolution of Welfare Index per Level of Education (2015-2018)",
    x = "Year",
    y = "Welfare Index",
    color = "Education level"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 9, face = "bold", hjust = 0.5),
    axis.title.x = element_text(margin = margin(t = 10)),
    axis.title.y = element_text(margin = margin(r = 10))
  )
```

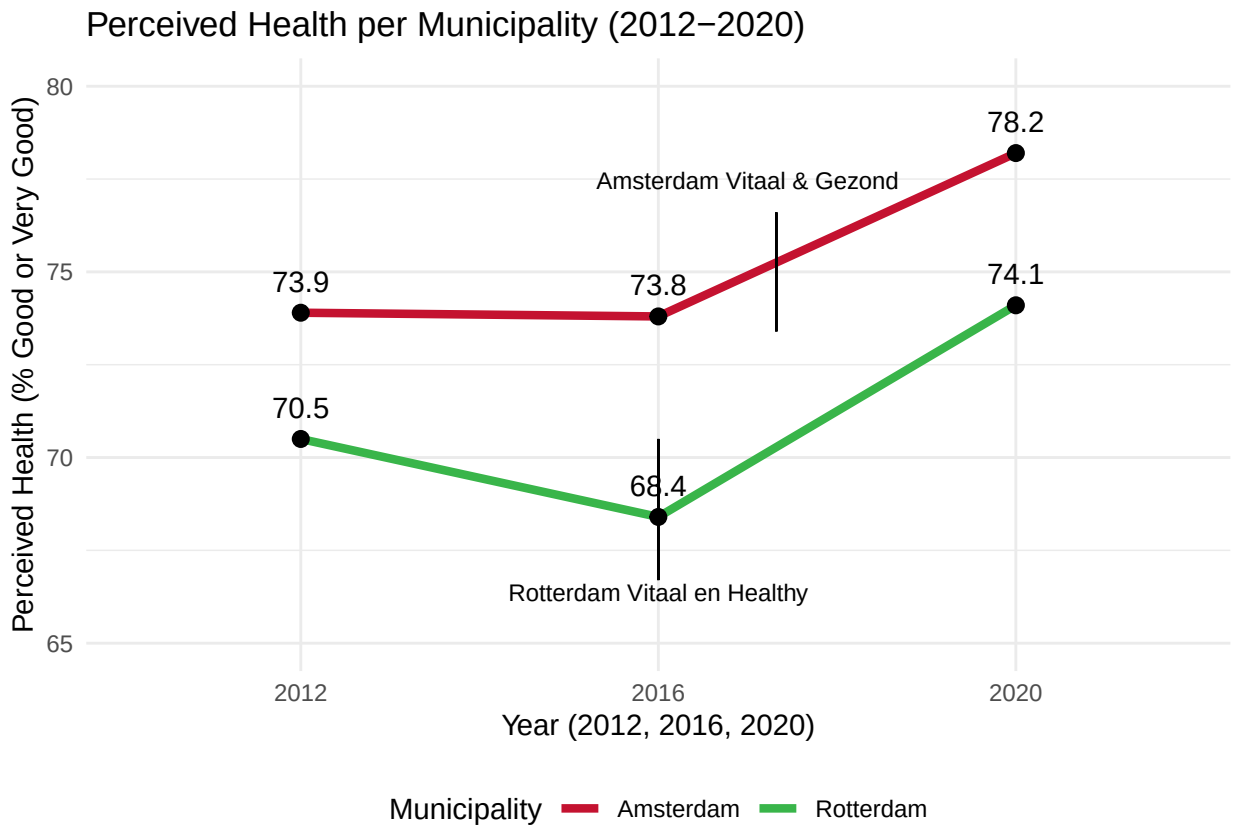



```
ggplot(Ervaren_Gezondheid,
       aes(x = Perioden, y = ErvarenGezondheidGoedZeerGoed_4,
           colour = Gemeentenaam_1, group = Gemeentenaam_1)) +

  geom_line(size = 1.5) +
  geom_text(aes(label = ErvarenGezondheidGoedZeerGoed_4),
            colour = "black", vjust = -1, show.legend = FALSE) +
  geom_point(colour = "black", size = 2.5, show.legend = FALSE) +
  geom_segment(aes(x = 2, xend = 2, y = 68.4 - 1.7, yend = 68.4 + 2.1),
               color = "black", linewidth = 0.4) +
  annotate("text", x = 2, y = 69.9 - 3.3, label = "Rotterdam Vitaal en Healthy",
           color = "black", angle = 0, vjust = 1, size = 3) +
  geom_segment(aes(x = 2.33, xend = 2.33, y = 73.8 - 0.4, yend = 73.8 + 2.8),
               color = "black", linewidth = 0.4) +
  annotate("text", x = 2.25, y = 79 - 1.3, label = "Amsterdam Vitaal & Gezond",
           color = "black", angle = 0, vjust = 1, size = 3) +
  scale_colour_manual(values = c("Amsterdam" = "#C41230", "Rotterdam" = "#39B54A")) +
  scale_x_discrete(label = c(2012, 2016, 2020)) +
  scale_y_continuous(limits = c(65, 80), breaks = seq(0, 100, 5)) +

  labs(
    x = "Year (2012, 2016, 2020)",
    y = "Perceived Health (% Good or Very Good)",
    title = "Perceived Health per Municipality (2012-2020)",
    colour = "Municipality"
  ) +
```

```
  theme_minimal() +
  theme(legend.position = "bottom")
```

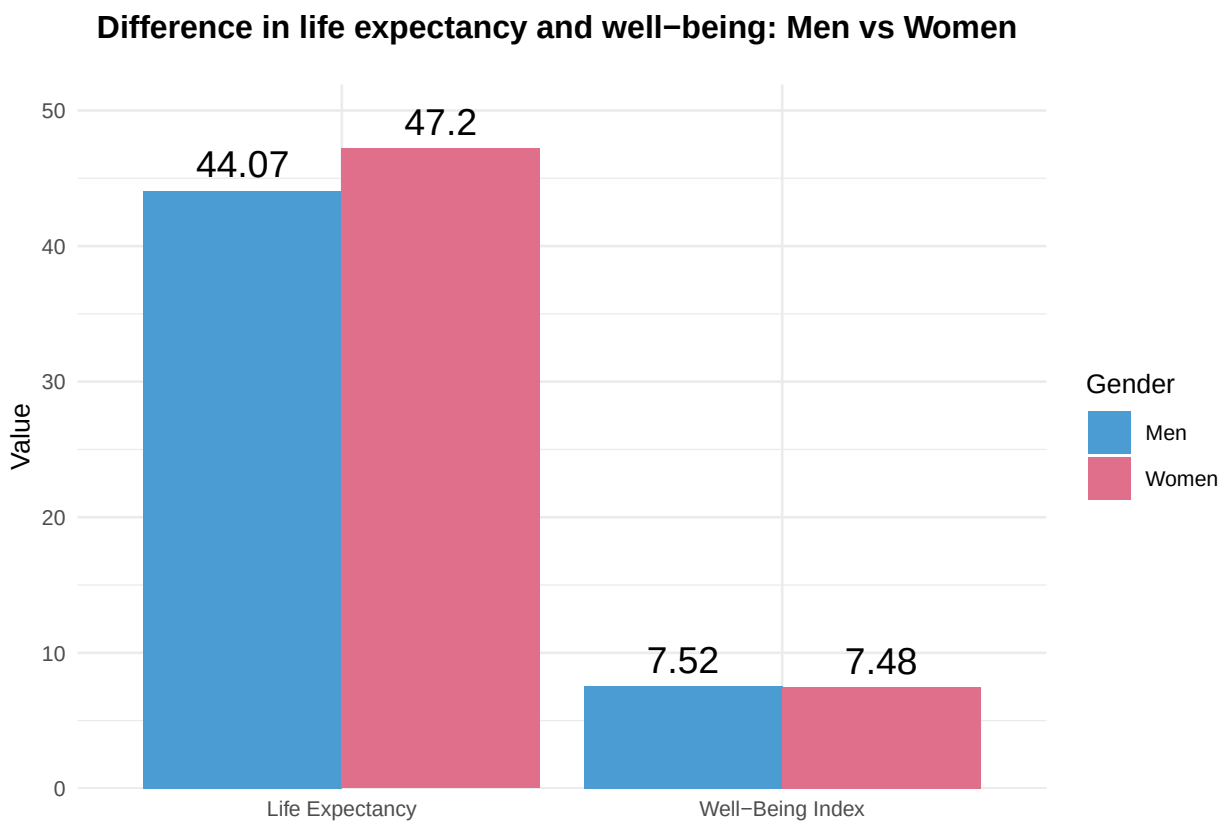



```
plot_data <- Ultimate_dataset_of_doom_hell_and_destruction %>%
  select(Geslacht, Levensverwachting, WelzijnIndex) %>%
  pivot_longer(
    cols = c("Levensverwachting", "WelzijnIndex"),
    names_to = "Metric",
    values_to = "Value"
  ) %>%
  mutate(
    Metric = recode(Metric,
                    "Levensverwachting" = "Life Expectancy",
                    "WelzijnIndex" = "Well-Being Index"
    ),
    Label = round(Value, 2)
  )
```

```
plot_data$Geslacht <- factor(plot_data$Geslacht, levels = c("Mannen", "Vrouwen"))
```

```
ggplot(plot_data, aes(x = Metric, y = Value, fill = Geslacht)) +
  geom_col(position = position_dodge(width = 0.9), width = 0.9) +
  geom_text(
    aes(label = Label),
```

```
  position = position_dodge(width = 0.9),
  vjust = -0.5,
  color = "black",
  size = 5
) +
scale_fill_manual(
  name = "Gender",
  values = c("Mannen" = "#4B9CD3", "Vrouwen" = "#E06F8B"),
  labels = c("Mannen" = "Men", "Vrouwen" = "Women")
) +
scale_y_continuous(expand = expansion(mult = c(0, 0.1))) +  # <- extra ruimte boven balken
labs(
  title = "Difference in life expectancy and well-being: Men vs Women",
  x = "",
  y = "Value"
) +
theme_minimal(base_size = 10) +
theme(
  plot.title = element_text(
    face = "bold",
    hjust = 0.4,
    margin = margin(b = 15)  # <- ruimte onder titel
  )
)
```



```r
Inkomen_per_wijk <- Inkomen_per_gemeente
Inkomen_per_wijk <- Inkomen_per_wijk %>%
  filter(Regionaam != "Totaal")
Inkomen_per_wijk <- Inkomen_per_wijk %>%
  filter(Wijkcode != "Totaal")
```

```r
Ervaren_gezondheid_wijk <- Ervaren_gezondheid_wijk %>%
  rename(Wijkcode = Codering_3, statnaam = Gemeentenaam_1)
```

```r
Ervaren_gezondheid_wijk$SoortRegio_2 <- trimws(Ervaren_gezondheid_wijk$SoortRegio_2)
Ervaren_gezondheid_wijk <- Ervaren_gezondheid_wijk %>%
  mutate(Wijkcode = trimws(toupper(Wijkcode)))
Ervaren_gezondheid_wijk <- filter(Ervaren_gezondheid_wijk, SoortRegio_2 %in% c("Gemeente", "Wijk"))
```

```r
Ervaren_gezondheid_wijk <- Ervaren_gezondheid_wijk %>%
  filter(SoortRegio_2 == "Wijk")
```

```r
gemeente_wijk_data <- inner_join(
  Inkomen_per_wijk, Ervaren_gezondheid_wijk, by = "Wijkcode") %>%
  select(statnaam, Regionaam, Wijkcode, Gemiddeld, ErvarenGezondheidGoedZeerGoed_4)
```

```r
gemeente_wijk_data$ErvarenGezondheidGoedZeerGoed_4 <- as.numeric(gsub(",", ".", gemeente_wijk_data$Ervar
```

```
## Warning: NAs introduced by coercion
```

```r
gemeente_wijk_data$Gemiddeld <- as.numeric(gsub(",", ".", gemeente_wijk_data$Gemiddeld))
```

```
## Warning: NAs introduced by coercion
```

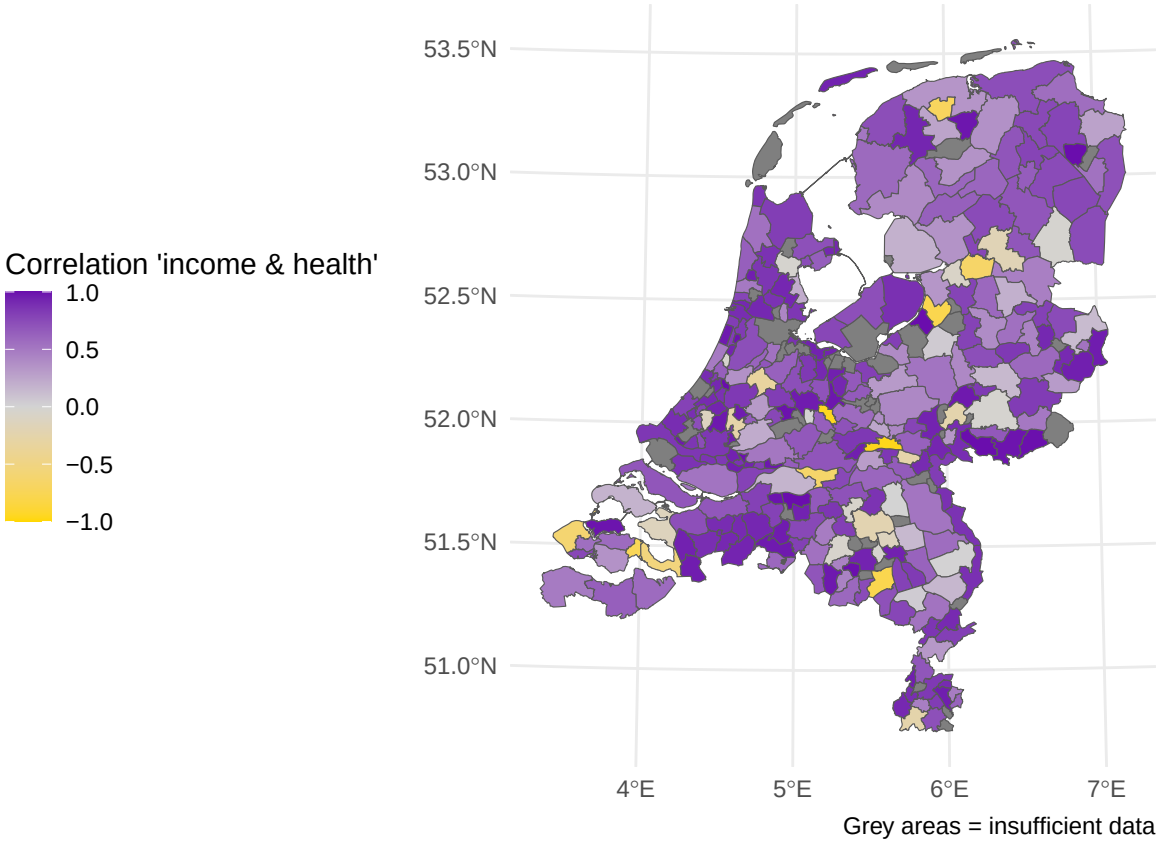
```r
gemeente_cor <- gemeente_wijk_data %>%
  group_by(statnaam) %>%
  filter(n() >= 3) %>%
  summarise(
    correlation = if (sum(complete.cases(as.numeric(Gemiddeld, ErvarenGezondheidGoedZeerGoed_4))) > 1)
    {cor(as.numeric(Gemiddeld), as.numeric(ErvarenGezondheidGoedZeerGoed_4), use = "complete.obs")}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

```r
gemeentegrenzen <- gemeentegrenzen %>%
  mutate(statnaam = tolower(trimws(statnaam)))
gemeente_cor <- gemeente_cor %>%
  mutate(statnaam = tolower(trimws(statnaam)))
```

```r
gemeente_wijk_mapdata <- gemeentegrenzen %>%
  left_join(gemeente_cor, by = "statnaam")
```

```r
ggplot(gemeente_wijk_mapdata) +
  geom_sf(aes(fill = correlation)) +
  theme_minimal() +
  scale_fill_gradient2(high = "#6A0DAD", mid = "lightgrey", low = "#FFD700",
```

```
                      limits = c(-1, 1),
                      breaks = c(-1, -0.5, 0, 0.5, 1),
                      name = "Correlation 'income & health'") +
  theme(legend.position = "left") +
  labs(caption = "Grey areas = insufficient data")
```




```
Inkomen_per_gemeente <- filter(Inkomen_per_gemeente, Wijkcode == "Totaal")
Levensverwachting_Gemeente <- filter(Levensverwachting_Gemeente, Perioden == "2019G400") #2019G400 = ge
Levensverwachting_Gemeente <- filter(Levensverwachting_Gemeente, Geslacht == "T001038") #T001038 = Man
Levensverwachting_Gemeente <- filter(Levensverwachting_Gemeente, Marges == "MW00000") #MW00000 = Waarde
Levensverwachting_Gemeente <- filter(Levensverwachting_Gemeente, Leeftijd == "10010") #10010 = 0 jarige

Levensverwachting_Gemeente <- Levensverwachting_Gemeente %>%
  filter(RegioS %in% Inkomen_per_gemeente$Gemeentecode)
```

```
gemeente_data <- left_join(Inkomen_per_gemeente, Levensverwachting_Gemeente, by = c("Gemeentecode" = "R
```

```
gemeente_mapdata <- left_join(gemeentegrenzen, gemeente_data, by = c("statcode" = "Gemeentecode"))
```

```
gemeente_mapdata$Gemiddeld <- as.numeric(gsub(",", ".", gemeente_mapdata$Gemiddeld)) #commas omzetten n
gemeente_mapdata$Levensverwachting_1 <- as.numeric(gemeente_mapdata$Levensverwachting_1)

cor(gemeente_mapdata$Levensverwachting_1, gemeente_mapdata$Gemiddeld, use = "complete.obs")
```

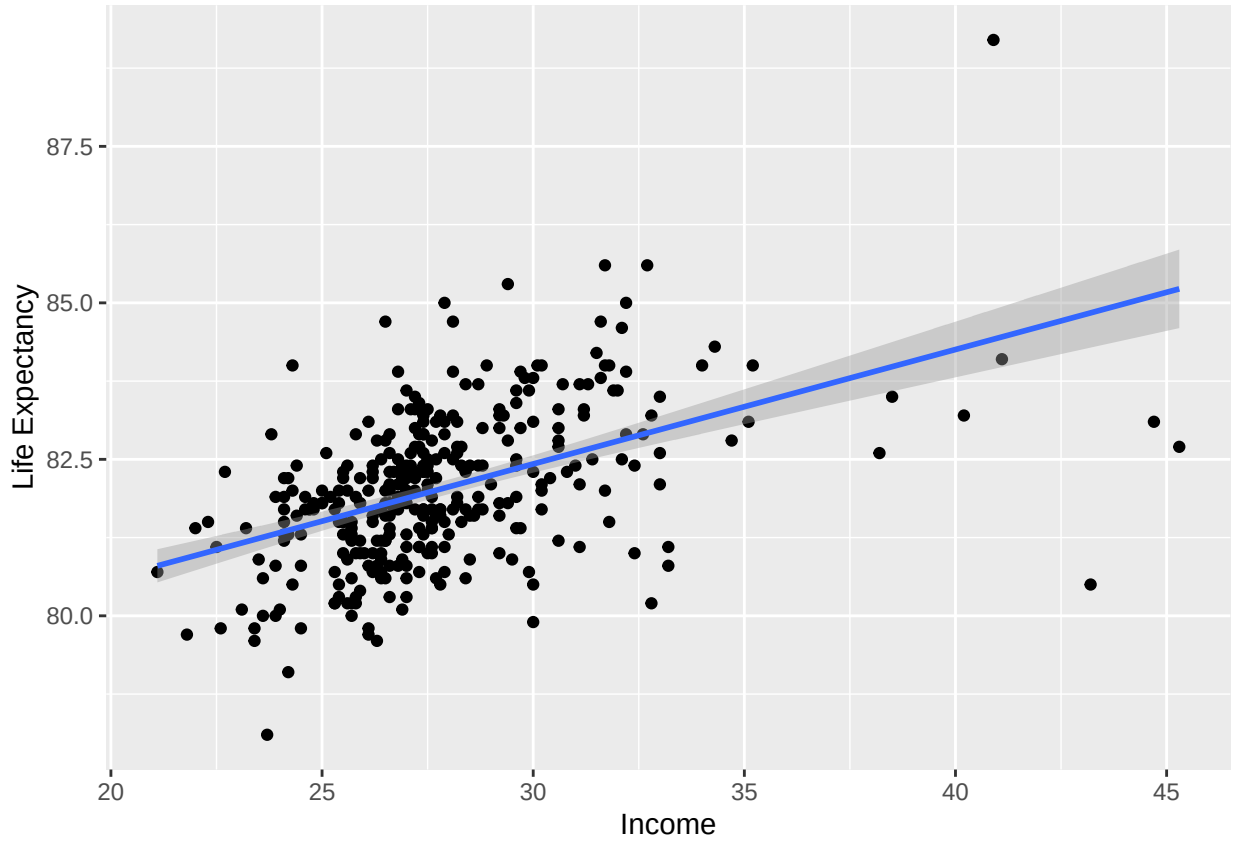
```
## [1] 0.4839739
```

```
ggplot(gemeente_mapdata, aes(x = Gemiddeld, y = Levensverwachting_1)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Income", y = "Life Expectancy")
```

```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```




```
Ervaren_Gezondheid <- Ervaren_Gezondheid[
  Ervaren_Gezondheid$Gemeentenaam_1 %in% c("Amsterdam", "Rotterdam") &
    Ervaren_Gezondheid$SoortRegio_2 == "Gemeente",]
```

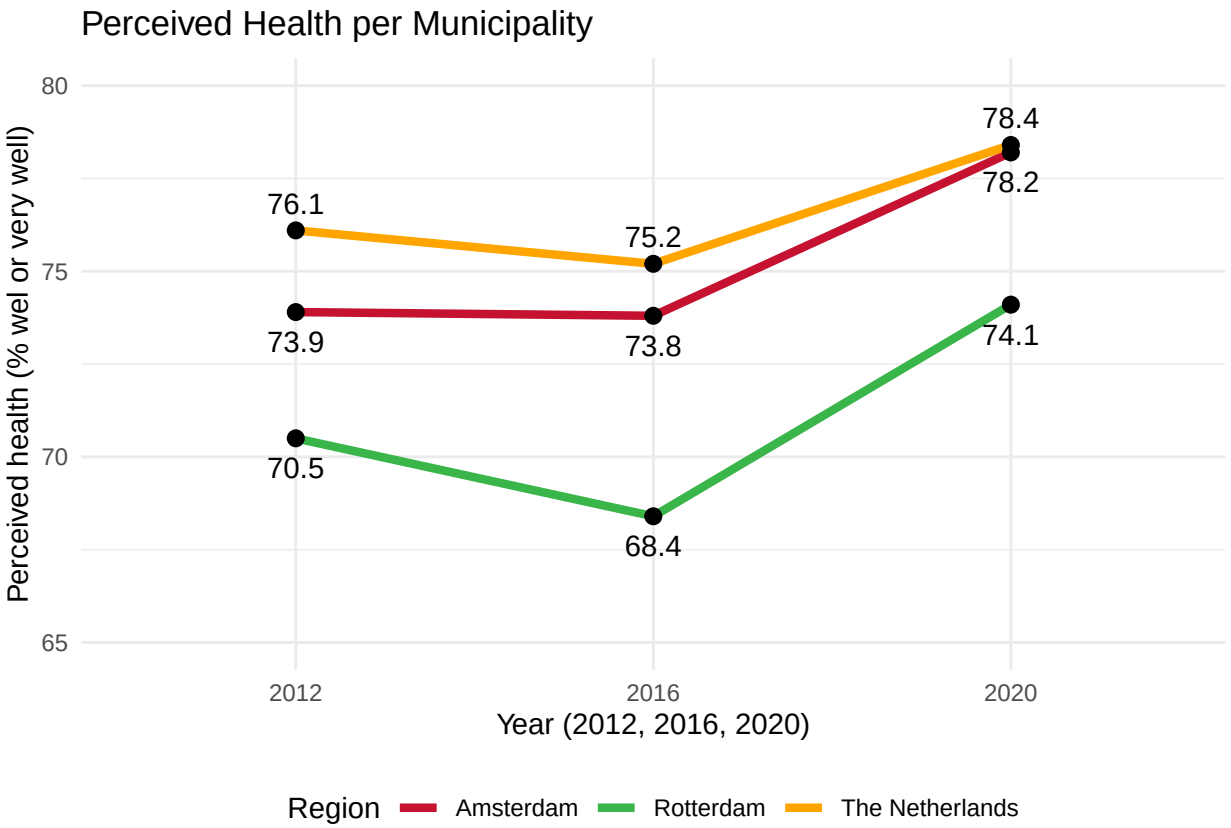
```
Ervaren_Gezondheid <- Ervaren_Gezondheid %>%
  select(-Gemeentenaam_1, -SoortRegio_2, -Codering_3)

Ervaren_Gezondheid_NL_AM_RO <- rbind(Ervaren_Gezondheid, ErvarenGezondheidNL)
```

```r
Ervaren_Gezondheid_NL_AM_RO$ErvarenGezondheidGoedZeerGoed_4 <-
  as.numeric(Ervaren_Gezondheid_NL_AM_RO$ErvarenGezondheidGoedZeerGoed_4)

Ervaren_Gezondheid_NL_AM_RO <- Ervaren_Gezondheid_NL_AM_RO %>%
  mutate(Region = case_when(
    WijkenEnBuurten == "GM0363" ~ "Amsterdam",
    WijkenEnBuurten == "GM0599" ~ "Rotterdam",
    WijkenEnBuurten == "NL01"   ~ "The Netherlands",
    TRUE ~ NA_character_
  ))
```

```r
ggplot(Ervaren_Gezondheid_NL_AM_RO,
       aes(x = Perioden, y = ErvarenGezondheidGoedZeerGoed_4,
           colour = Region, group = Region)) +
  geom_line(size = 1.5) +
  geom_point(colour = "black", size = 2.5, show.legend = FALSE) +
  geom_text(aes(label = ErvarenGezondheidGoedZeerGoed_4,
                vjust = ifelse(Region == "The Netherlands", -0.8, 1.9)),
            colour = "black",
            show.legend = FALSE) +
  scale_colour_manual(values = c(
    "Amsterdam" = "#C41230",
    "Rotterdam" = "#39B54A",
    "The Netherlands" = "orange"
  )) +
  scale_x_discrete(labels = c("2012", "2016", "2020")) +
  scale_y_continuous(limits = c(65, 80), breaks = seq(0, 100, by = 5)) +
  labs(
    x = "Year (2012, 2016, 2020)",
    y = "Perceived health (% wel or very well)",
    title = "Perceived Health per Municipality",
    colour = "Region"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



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

Variable 1

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

Variable 2

```
Welzijn_index_2016 <- Welzijn_averages_2016 %>%
  select(
    Kenmerken, Marges, Perioden,
    ScoreGeluk_1,
    ScoreTevredenheidMetHetLeven_5,
    ScoreTevredenheidMetWerk_13,
    ScoreTevredenheidDagelijkseBezigheden_21,
    ScoreTevredenheidMetLichGezondheid_25,
```

```
    ScoreTevredenheidPsychischeGezondheid_29,
    ScoreTevredenheidMetSociaalLeven_57 )
```

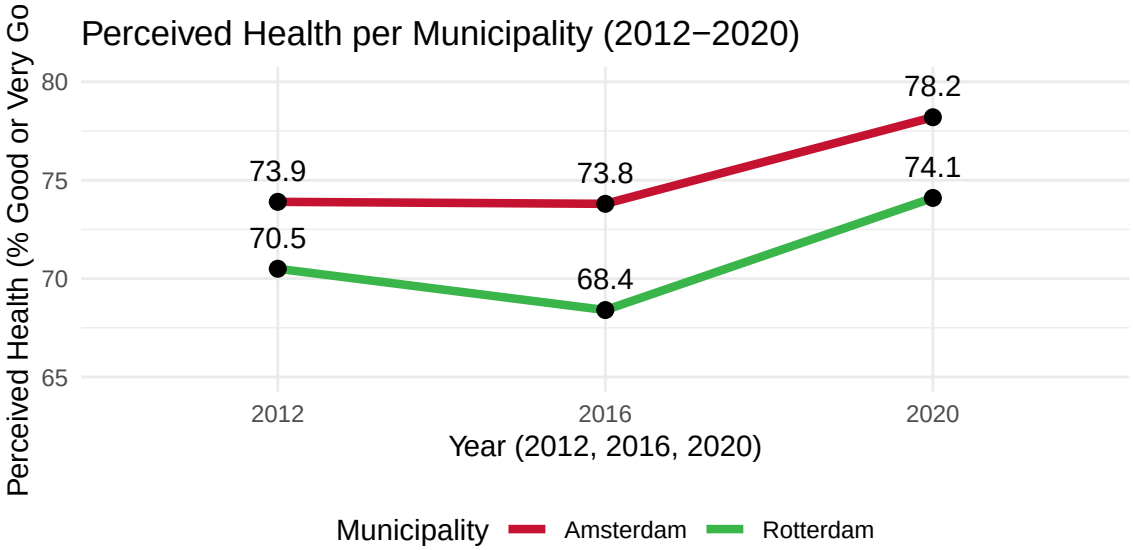
```
Welzijn_index_2016 <- Welzijn_index_2016 %>%
  mutate(WelzijnIndex = rowMeans(select(.,
                                        ScoreGeluk_1,ScoreTevredenheidMetHetLeven_5,ScoreTevredenheidMetWerk
```

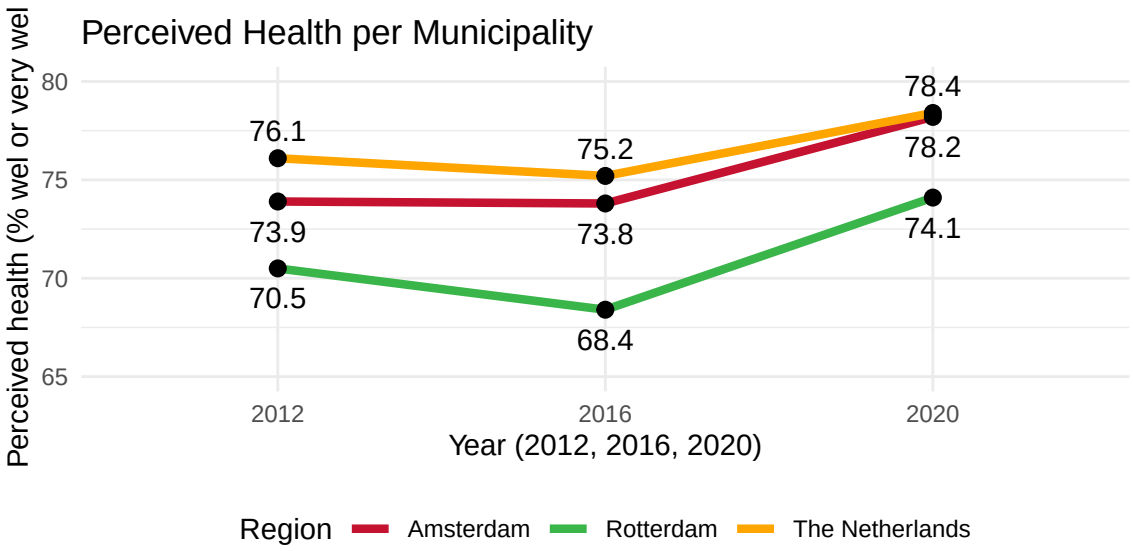
```
Welzijn_index_2016 <- Welzijn_index_2016 %>%
  select(Kenmerken, Marges, Perioden, WelzijnIndex)
```

We decided to make a new variable called "correlation" that we added

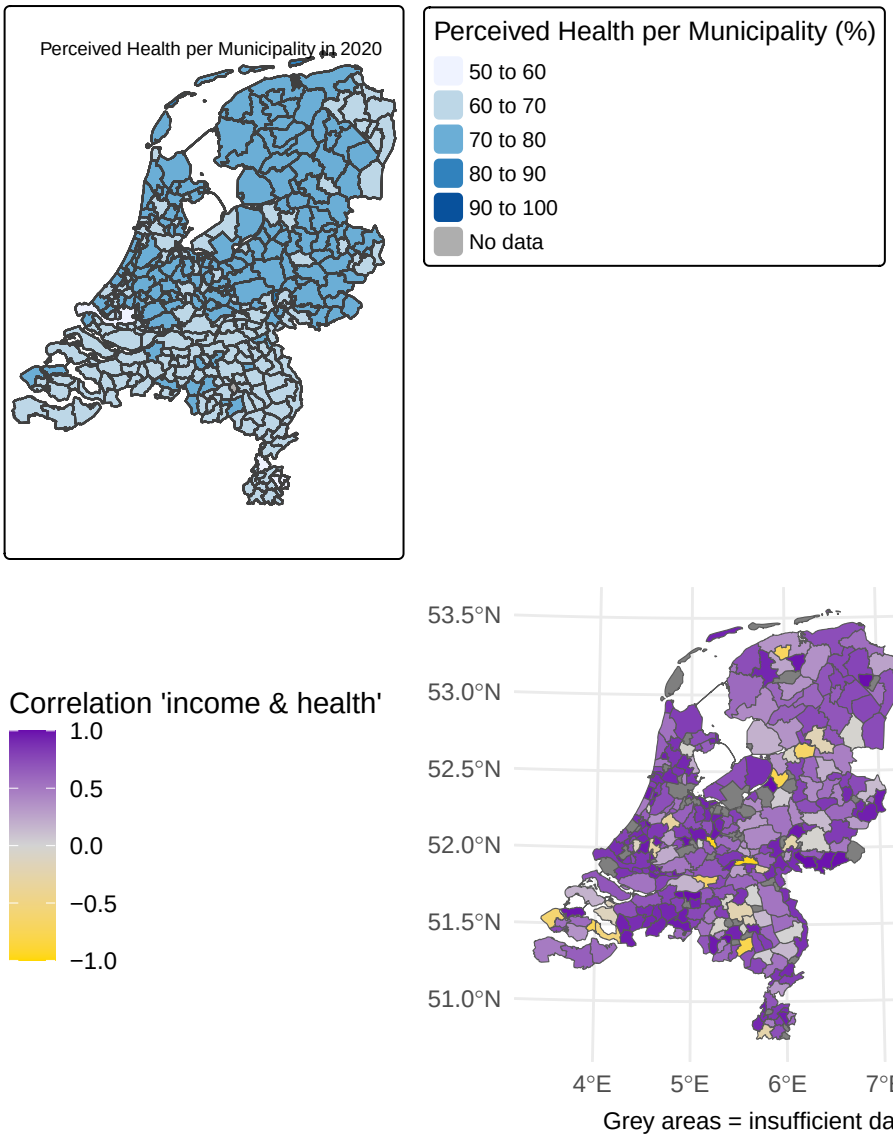
Variable 2

3.3 Visualize temporal variation





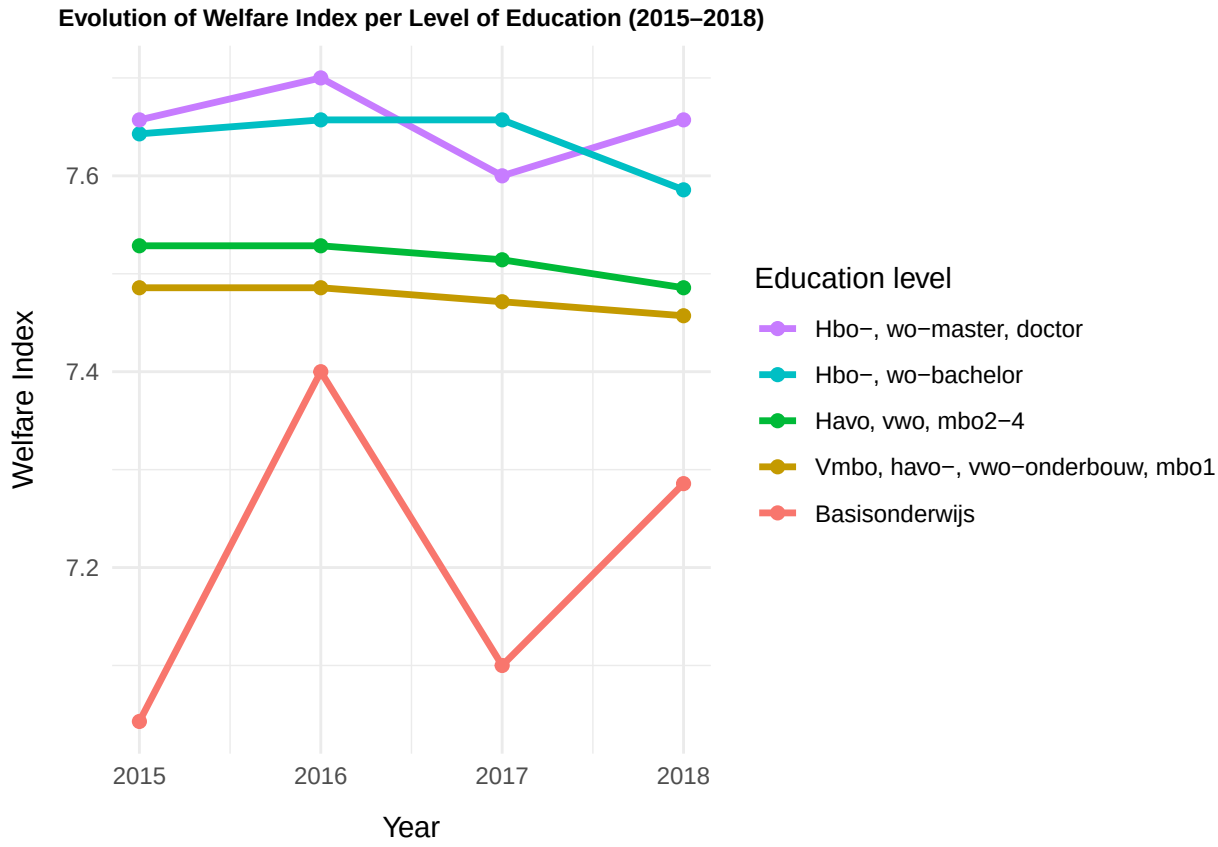
3.4 Visualize spatial variation



Here you provide a description of why the plot above is relevant to your specific social problem.

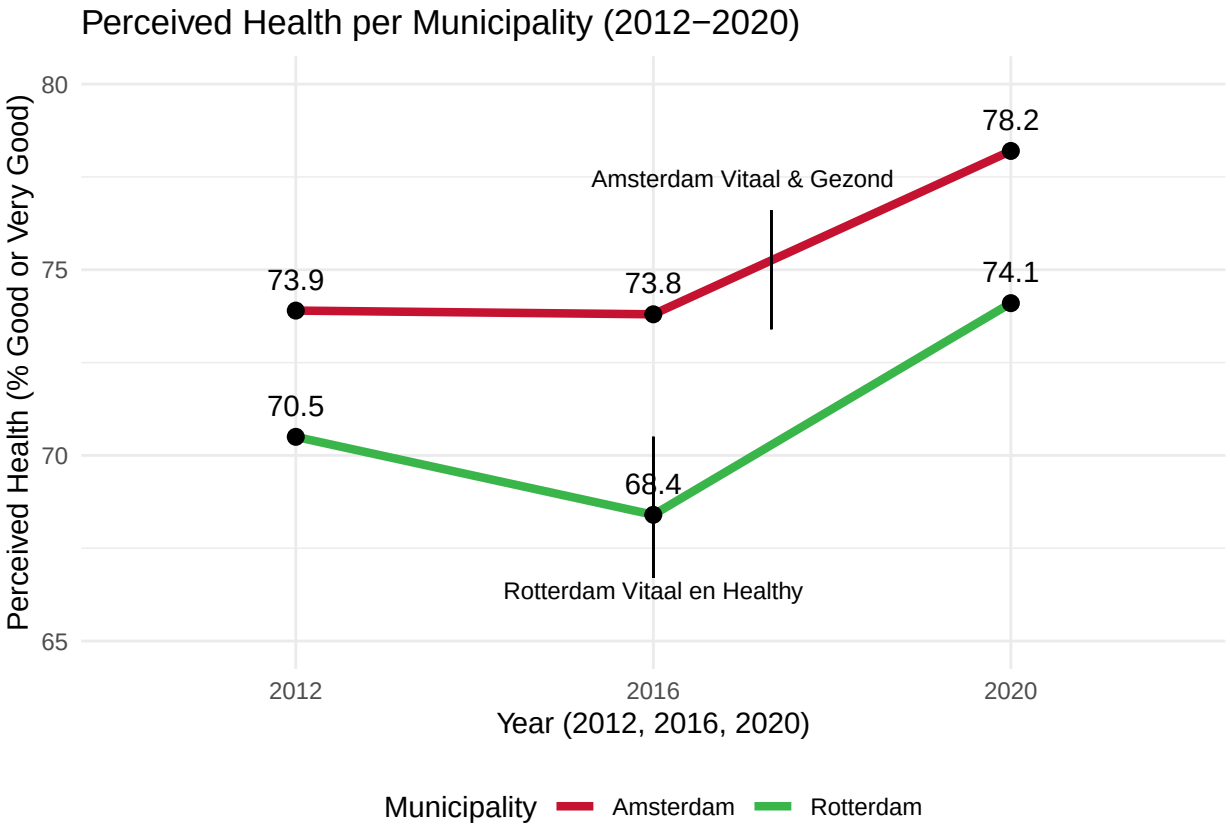
3.5 Visualize sub-population variation

What is the poverty rate by state?



Here you provide a description of why the plot above is relevant to your specific social problem.

3.6 Event analysis



Analyze the relationship between two variables.

Here you provide a description of why the plot above is relevant to your specific social problem.

Part 4 - Discussion

4.1 Discuss your findings

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