How is regional variation in life expectancy explained?

Arvid Mikkers (https://github.com/ArvidMikkers)

09 december, 2018

Contents

Foreword

For this "profielwerkstuk"(PWS) I wanted to do someting with data science and machine learning for I'm very interested in this part of computer science. I got better at statistics along the way and learned some new techniques which probably will be very useful in the future. The goal of "profielwerkstuk" for me is to do reproducible research. That's why I chose for R instead of for example Excel, because R is script based so you can see every step I made, which makes it reproducible.

I had to learn about regressions, cross-validation and random forest in this project. I did this with a book called "An introduction to Statistical learning with applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (James *et al.*, 2013).

I was already familiar with R. I did some online courses and followed a lecture in making maps in R at Tilburg University. I made some graphs for a physics project before but never before did I do a whole project in R.

I did this project under version control in a private repository on GitHub. The final version is published on a public repository (https://github.com/ArvidMikkers/PWS)

The package I used for making maps, *"thematicmaps"* is not available on Cran, you can download it, however, from my GitHub.

The data cleaning file is situated in Apendix 2.

I'm grateful for the guidance in this project of Habib Rejaibi (Cygnus Gymnasium). Next to Habib Rejaibi I got a lot of help from employees of the Dutch Healthcare Authority (NZa). I'm especially thankful to dr Gertjan Verhoeven (NZa and Tilburg University), Ramsis Croes (NZa and Erasmus University), dr Mark Klik (NZa), Annemiek van der Laan (NZa) and dr Victoria Shestalova (NZa) for their help and advice. I would like to thank Merlijn de Bruin (Cygnus Gymnasium), my grandparents Anke & Henk Mikkers and my parents Sandra Kompier & Misja Mikkers for proofreading this thesis.

1 Introduction

1.1 Research questions

Last year, 2017 around Christmas, I was reading Outliers by M.Gladwell(Gladwell, 2008). One of the chapters in the book handles of the "Roseto mystery". Roseto is a community, in which the avarage life expectancy is well above the rest of the USA. I was triggered by the question if we had such communities in the Netherlands.

In this thesis I would like to answer the following question:

What explains regional variation in life expectancy?

Several sub questions are underlying the main question:

1. Has social economic status an impact on regional variation in life expectancy (operationalized by income, education)?
2. Has ethnicity an impact on regional variation in life expectancy?
3. Has access to health care facilities an impact on regional variation in life expectancy (operationalized by distance to nearest hospital, informal care)?
4. Has health status an impact on regional variation in life expectancy (operationalized in weight, fitness and morbidity)?

I also wanted to answer the question: 2. Have environmental factors an impact on regional variation in life expectancy (operationalized by CO2 emmision)? But since I could find no complete data on this subject, I could not answer this question unfortunately.

1.2 Related literature

After reading the chapter in the book "Outliers", I went online and found some articles on regional variance in life expectancy. Altough there was already done research on the regional variance of life expectancy in the Netherlands in 1988(Poppel, 1988), it only handled of the life expectancy between 1972 and 1984. There was a study on regional variance in healthy life expectancy (a prognosis for how many years you will live in a healthy status) done by TNO in 2002(Mulder *et al.*, 2002). This article however focusses more on what a healthy life expectancy is and the regional variation in life expectancy, but does not discuss the factors which could explain this regional variation. In other countries there are only a few of those studies. Laura M Woods et al. wrote in 2005(Woods *et al.*, 2005) an article about deprivation as an explanatory factor for the geographical variation in life expectancy in Wales and England. Her study focusses on 1998. Another study, written in 2018 by Jessica Y Ho(Ho & Hendi, 2018), focusses on a decrease in expectancy between 2014-2016 in high income countries by identifying the causes of death. In an American study by LauraDwyer-Lindgren et al. from 2017 (Dwyer-Lindgren *et al.*, 2017), which is perhaps most similar to mine, they focussed on inequalities in the life expectation per county and tried to globally asses causes for that inequality. My contribution is that this study focusses on recent geographical differences in life expectancies. I will search for factors that are associated with differences in life expectancy with open data. This study is published on my GitHub account, which makes this research completely reproducible.

This paper is organized as follows: First, I will describe the data in the section "Data" and then give an overview of the data in the section "Descriptive statistics". A section about the methodology and estimation strategy follows. Then I will describe my empirical results and robustness checks in the section Analysis. Finally I will discuss my results and draw conclusions.

Appendix 1 contains some extra figures, that were not presented in the main text of this paper. Appendix 2 contains my notebook about datacleaning.In appendix 3 I present an extra regression on Health Status. In appendix 4 you can find the logbook of my activities for this project.

2 Data

To answer my research questions, I will use open-access data only. This makes my research completely reproducible.

I will use public data of CBS (Central Bureau of Statistics) for most variables and the "Ministerie van Infrastructuur en Waterstaat" (Ministry of Infrastructure and Water Management) for data on CO2-emissions. I used the public data of the RIVM for the life expectancy at birth.

On July 19th, 2018 I have downloadedthe data about life expectancy at birth from RIVM.

On July 19th, 2018 I have downloaded the data about "kerncijfers" plus metadata from CBS.

On July 19th, 2018 I have downloaded the data about "Gemeentefonds" plus metadata from CBS.

On July 19th, 2018 I have downloaded the data about the health plus metadata from CBS.

On September 27th, 2018 I have downloaded the data about "CO2 emissions" from Rijkswaterstaat.

I'm interested in the life expectancy for the total population. I have merged and cleaned the data. The notebook with the data cleaning is included in appendix 2.

3 Descriptive statistics

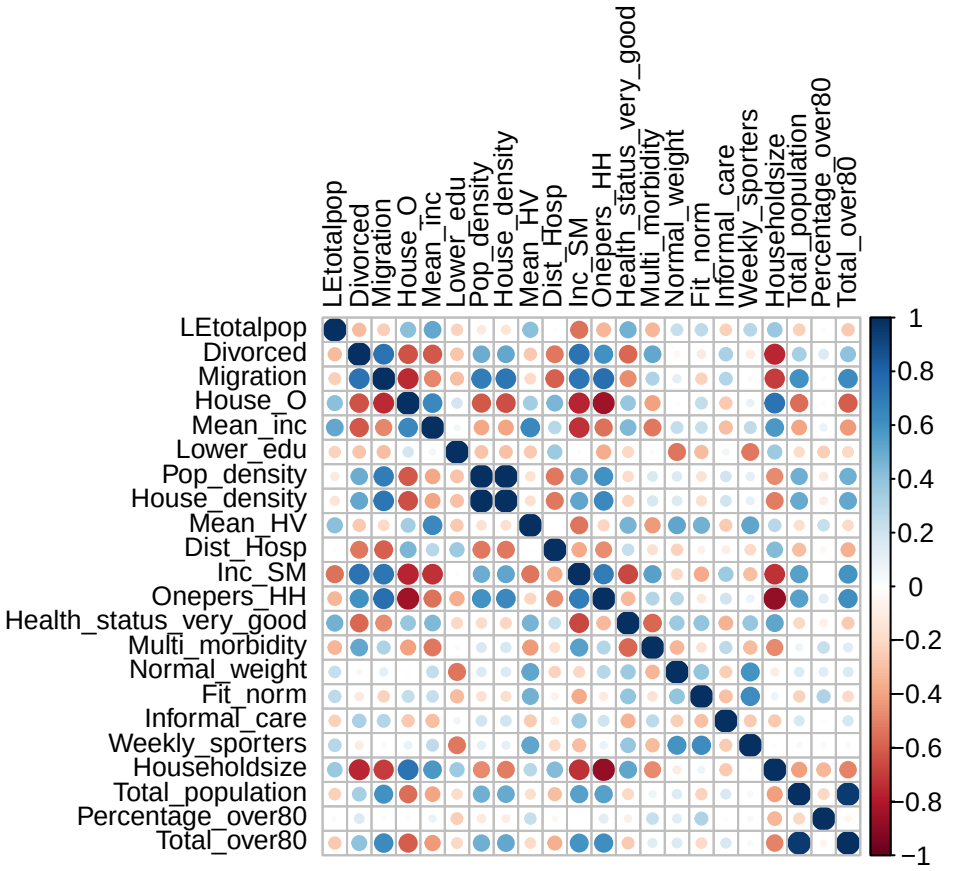
3.1 Overview

I have combined different datasets into one dataframe. I will start with some descriptive statistics of our data to be used in this paper.

Table 1: Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
LEtotalpop	374	81.910	1.242	78.500	81.200	82.600	86.400
Divorced	374	8.183	1.684	2.400	6.900	9.300	13.300
Migration	374	14.208	7.918	3.300	8.300	17.650	51.700
House_O	374	64.207	8.021	29.600	60.150	69.875	78.000
Mean_inc	364	23,197.250	1,714.319	19,500.000	21,900.000	24,300.000	28,700.000
Lower_edu	374	18.832	4.591	7	16	22	31
Pop_density	374	788.556	940.352	58	225.5	936	5,580
House_density	374	352.040	434.897	27	94.2	404.2	2,578
Mean_HV	374	221.366	52.265	122	186	251.8	554
Dist_Hosp	374	9.802	6.235	1.500	5.200	13.100	63.700
Inc_SM	374	10.202	3.783	4.100	7.600	11.775	27.800
Onepers_HH	374	30.557	6.150	18.576	26.698	32.835	64.205
Health_status_very_good	374	76.815	3.769	63.200	74.525	79.500	85.000
Multi_morbidity	374	33.440	3.788	20.100	31.000	35.900	48.900
Normal_weight	374	47.880	4.558	34.300	45.200	50.500	61.600
Fit_norm	374	27.615	4.272	13.900	25.200	30.450	42.900
Informal_care	319	10.371	2.591	5.400	8.500	11.700	19.400
Weekly_sporters	374	50.301	6.184	35.600	45.700	54.700	68.200
Householdsize	374	2.301	0.173	1.730	2.210	2.390	3.330
Total_population	374	41,501.880	61,826.470	3,611	17,964.2	43,627	833,624
Percentage_over80	374	4.742	0.985	1.800	4.100	5.300	8.400
Total_over80	374	1,855.283	2,206.888	128	838.2	2,044.8	26,160

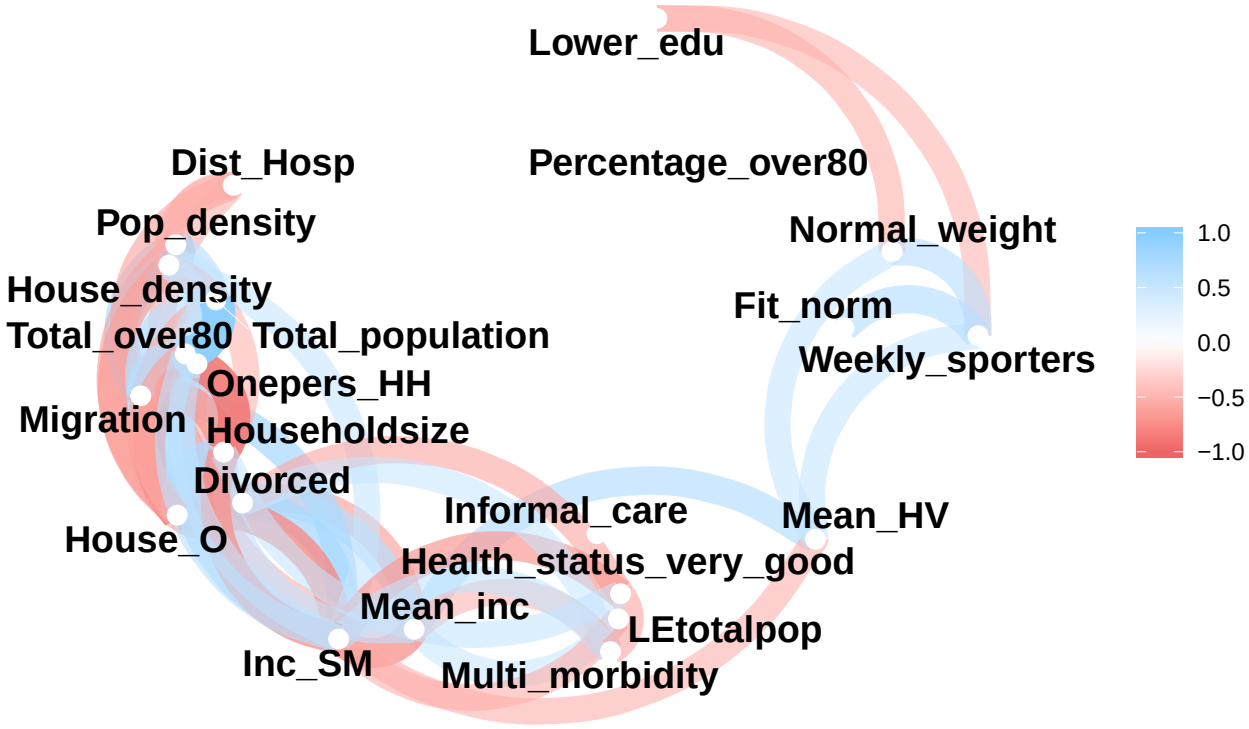
For my analysis I'm interested in the correlation between variables. Therefore, I will make a correlation plot. A correlation plot is a graphical representation of a correlation matrix. In this graph the color and size of the dot indicate the correlation coefficient. Since every variable is represented in the x-asis and the y-asis the plot is symmetrical in y=x. Obviously every correlation coeficient on the line y=x is 1, because variable on the y-asis is the same variable as on the x-asis.



In the correlation plot we can see that some variables obviously correlate with life expectancy, the most notable being, Mean_Inc (the mean income per muncipality), House_value (the meam value of the houses per municipality), Health_status_very_good (Self-reported health status good to very good), Inc_SM(percentage of people with an income upto 120% of the social minimum)

Some variables are also very strongly correlated with each other. Due to this collinearity I'm probably overstating or understating the impact some of these variables have on life expactency. I can derive from the correlation plot that, for example, the mean income per municipality (Mean_inc) correlates with the mean value of the houses per municipality (House_value) and the percentage of people with an income upto 120% of the social minimum (Inc_SM). This might be a problem for my analysis. I will discuss these problems in section 5.1.2

Due to the number of variables in my correlation plot, the plot becomes hard to read. To give a better insight in the correlation between these variables I can also present a network plot. In this plot variables closer to each other are more related. The color indicates the sign of relationship, with blue being positive and red being negative.

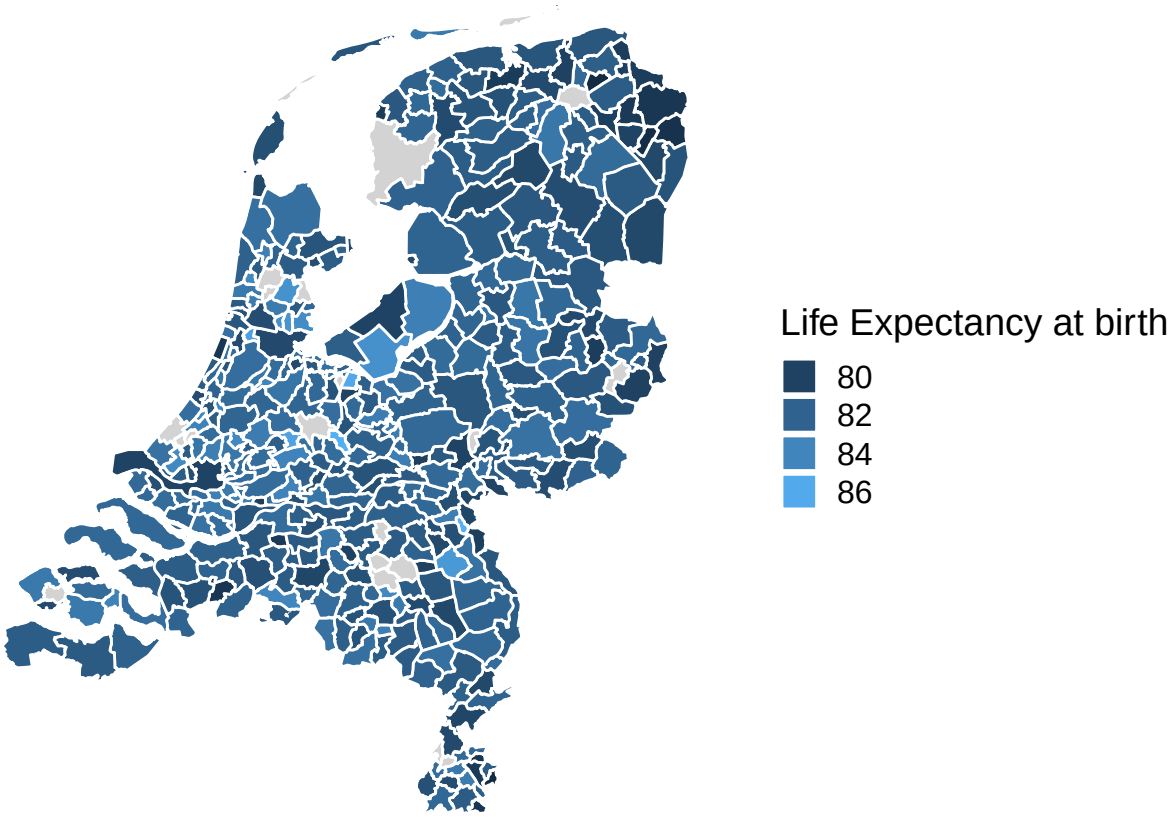


In the plot we can see numerous interesting things. Weight, the people who fullfill the fit norm and the sporting are correlated with each other. We can also see the impact of education and age on weight.

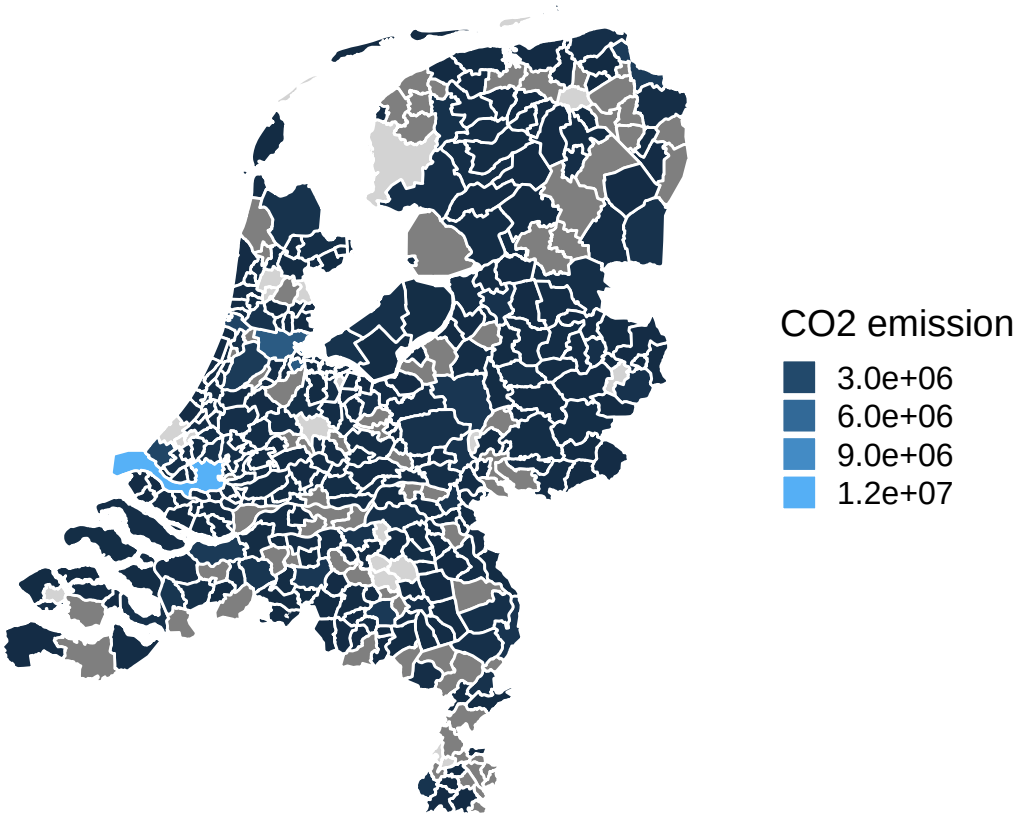
Population density, house density, total population etc. in the top left corner are also correlated with each other. And then there is the group of variables in the botom center. This group contains our response variable life expectancy of the total population. In this group there are variables with respect to income, health status and the hours of informal care given. I expect to see these relationships in our formal analysis in paragraph 5.

3.2 Regional variation

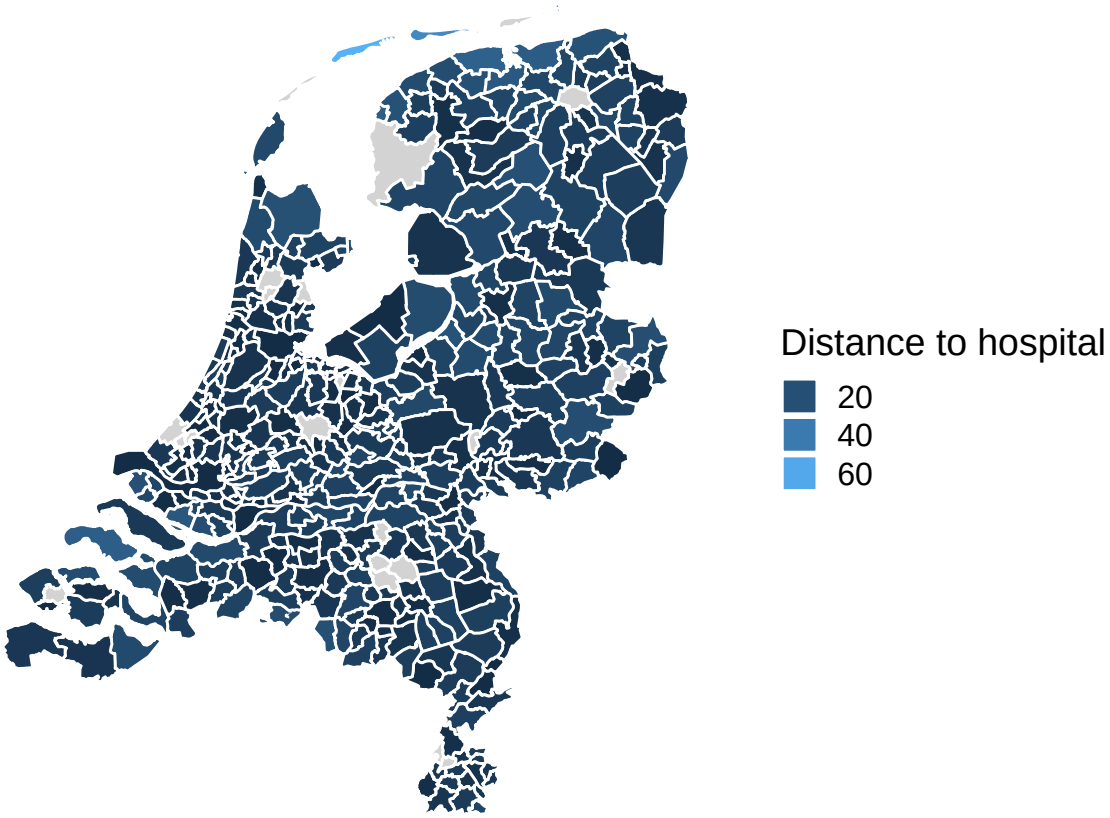
For some variables I will show the regional variation. For this purpose I created multiple maps. I think that maps offer the best insight in regional variance, and are highly readable.



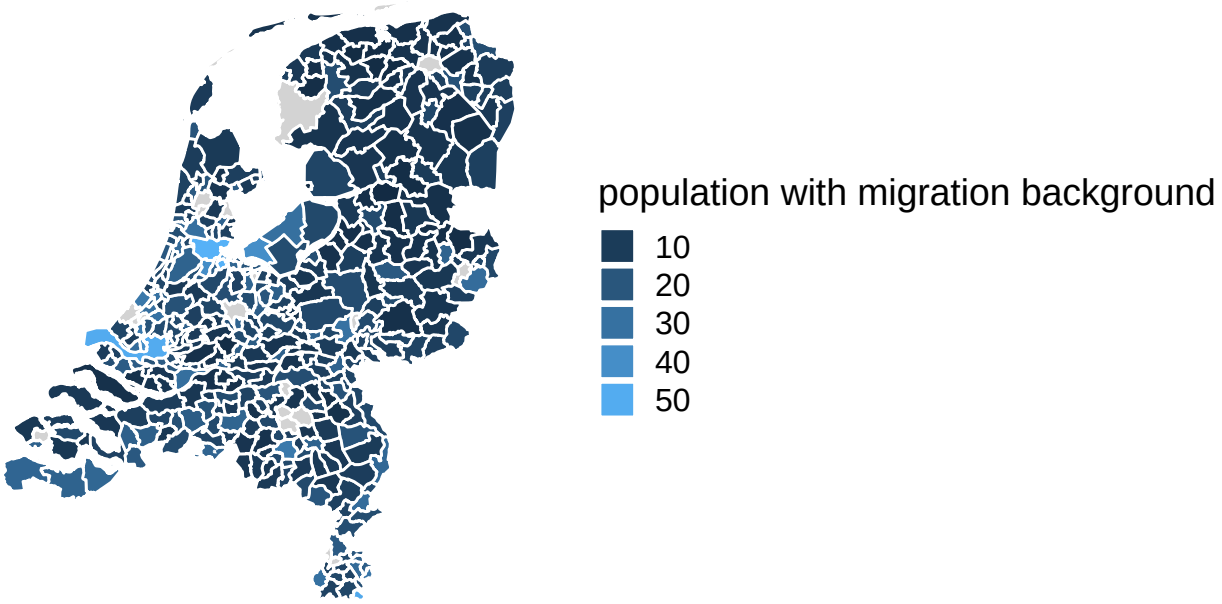
In this map we see the life expectancy per municipality, the lighter the blue the higher the life expectancy. In appendix 1 you will find a few more maps of the municipalities with the highest and lowest Life expectancy highlighted, as well as maps with the life expectancy specifically for women and specifically for men and also a map with the difference between these expectancies.



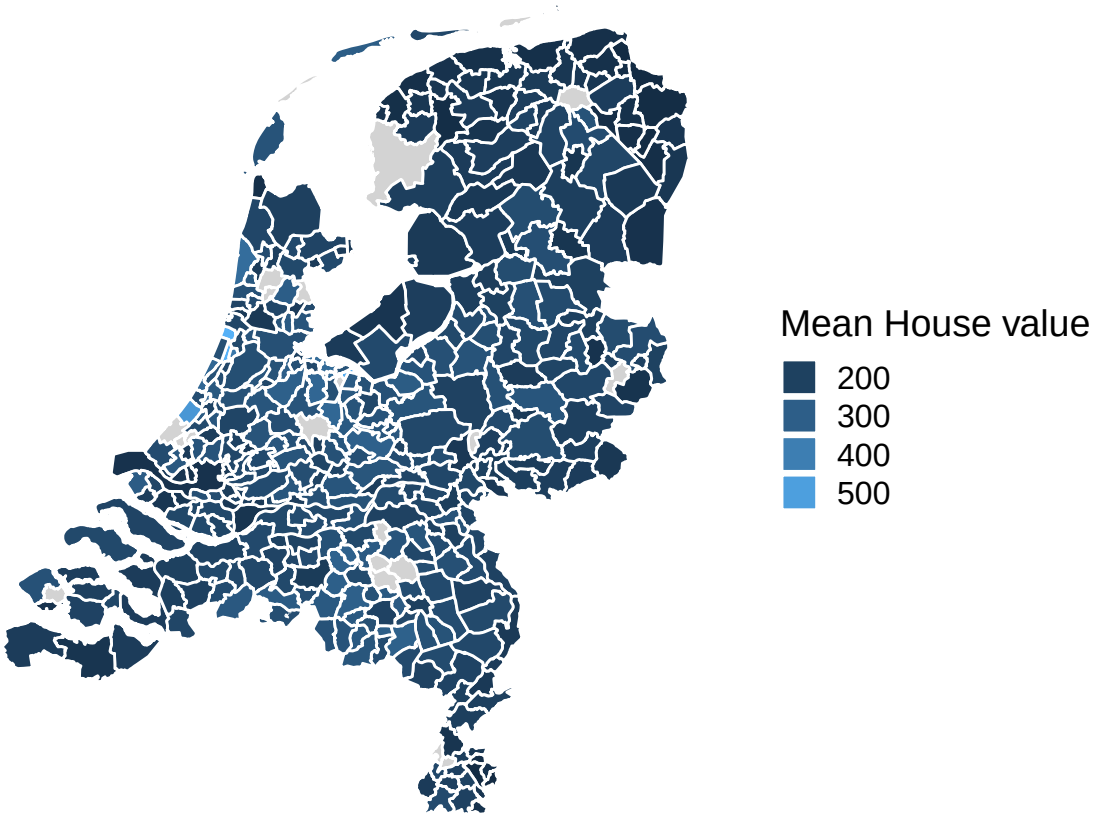
Altough this map gives an interesting insight in the CO2 emmisions in the Netherlands, the datasets unfortunately contain many gaps. This is mostly due to the very few measurement points that the RIVM has. Because of this, I cannot answer my research questions about the effect of CO2 emissions on the life expectancy. We can see clearly that in the big cities, Amsterdam and Rotterdam, the CO2 emissions are notably higher than in the rest of the country. But besides that there is more industry in those areas, there live more people as well. To correct the fact that there live more people I have made a map of the emmisions per capita as well, which you can find in appendix 1. In that map is not Rotterdam, but Delfzijl notably larger than the rest of the Netherlands.



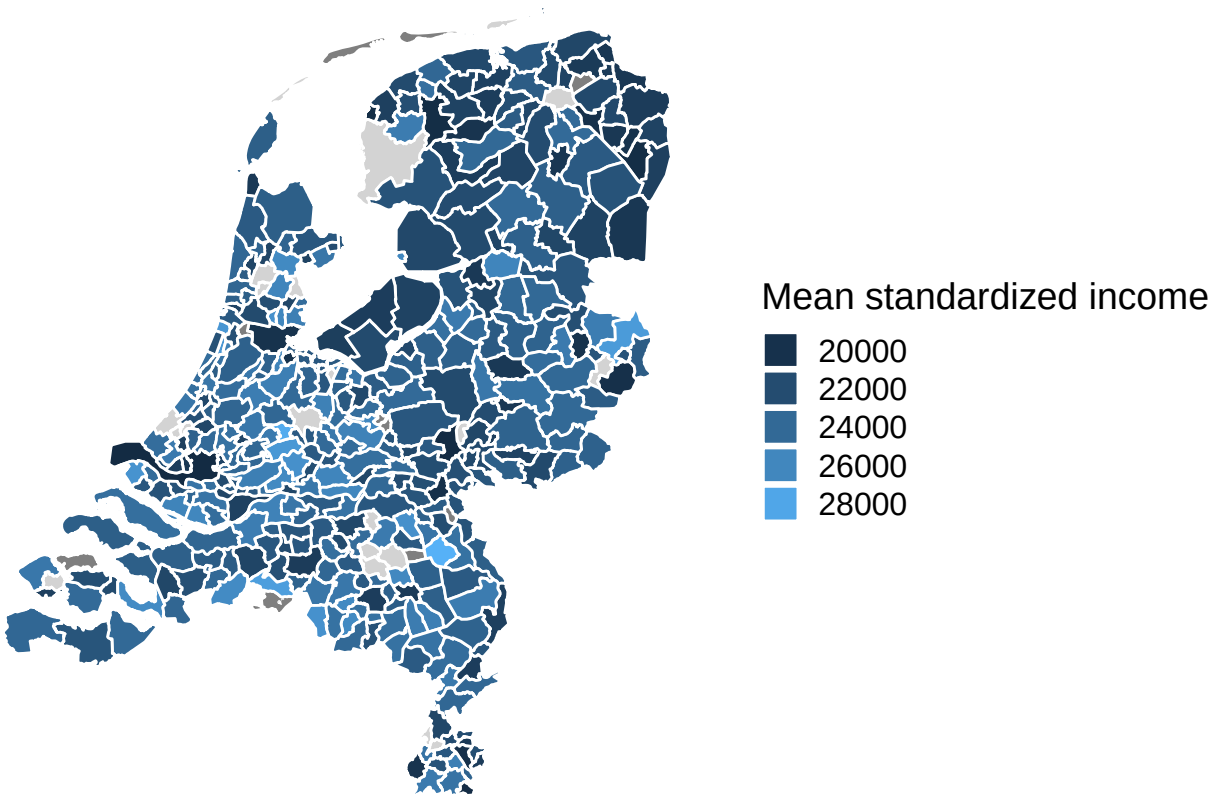
In this graph I present the regional variance in the distance to the nearest hospital. Most people only have to travel a maximum of 20 kilometers to the nearest hospital. But here we can see some interesting facts as well, because people who live at the edges of the country have to travel far further to a hospital. For example Zeeland massively under performs here. The islands in the north of the Netherlands are the obvious outliers with the nearest hospital over 60 kilometers away in some cases.



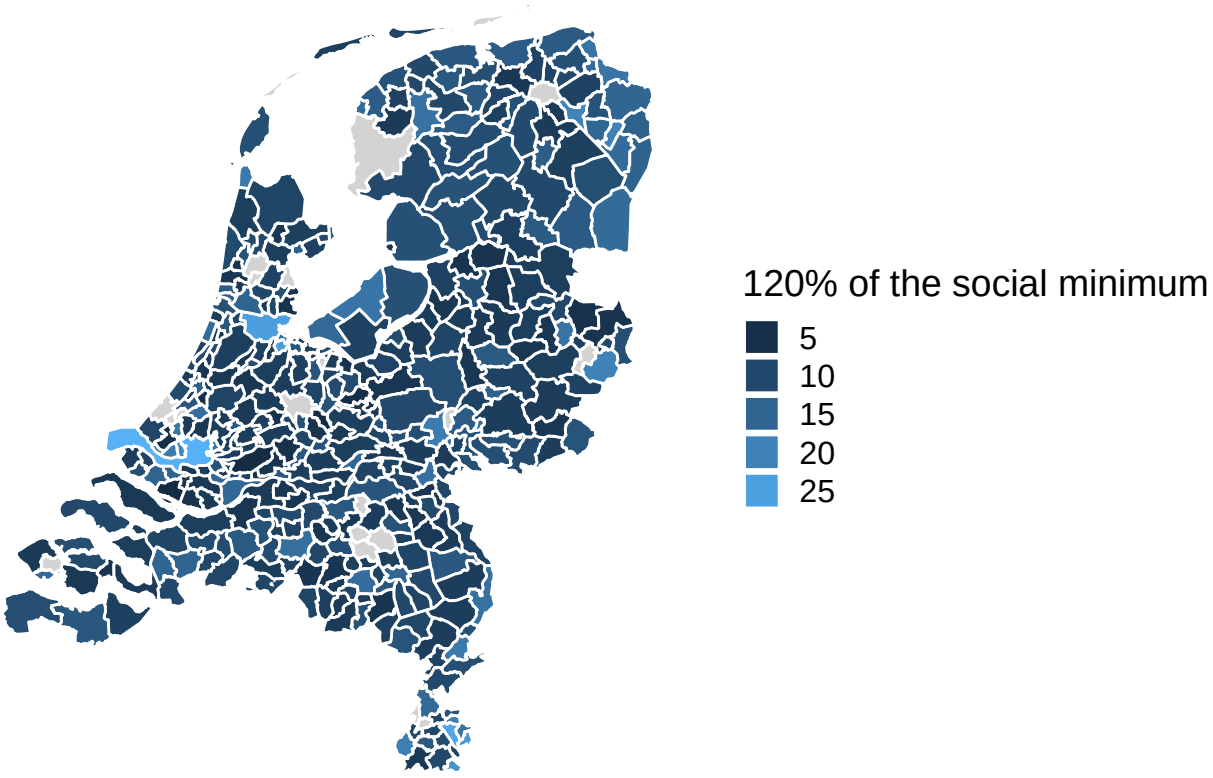
In this map I present the percentage of people with a migration background per municipality. Here you can clearly see that the urban areas in the Netherlands vastly outperform the rest of the country, with the "Randstad" 40-50% of people having a migration background. While in some other provinces, for example the north-eastern part of the Netherlands rearly exceeding more than 10%. This is interesting since there is a ungoing debate over migration in the Netherlands with many protests against immigration happening in those areas, which have in fact far less immigrants than the rest of the Netherlands.



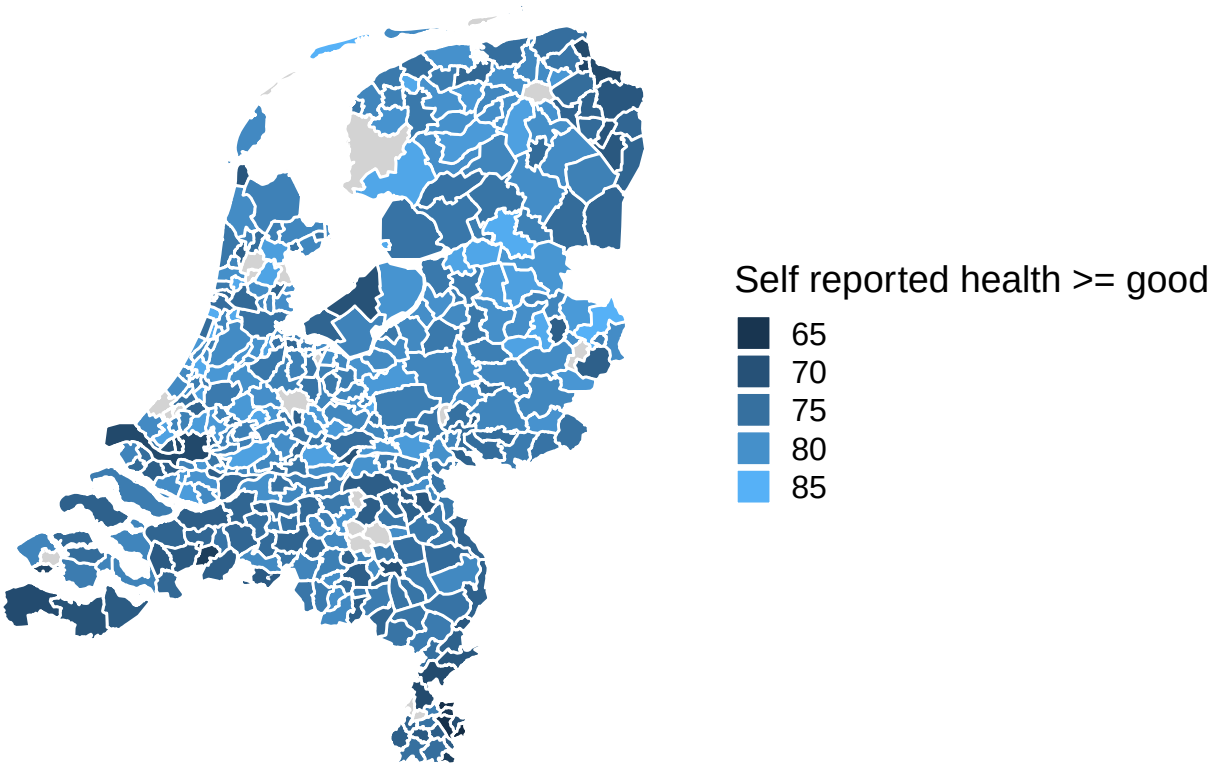
In this graph I present the mean house value per municipality in thousands of euro's. We can clearly see the outliers: the municipalities who are situated close to the North Sea, but also close to the capital of the Netherlands, Amsterdam.



This is a map which shows you the mean yearly income per municipality. The mean standardized income is the mean income of a municipality corrected for the differences in size and the structure of the average household in that municipality. By doing this we can compare the welfare of different households.



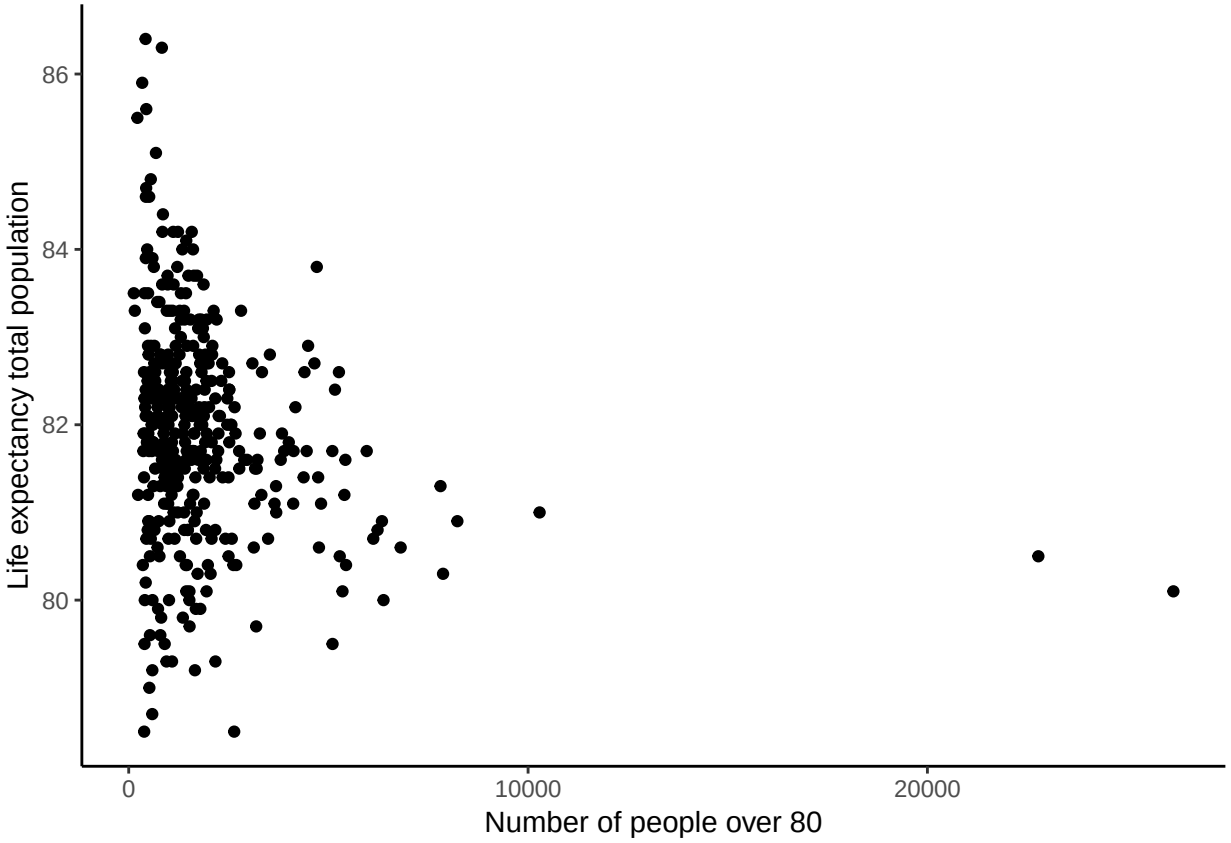
This map shows you the percentage of people with an income up to 120% of the social minimum. The social minimum represents the quantity of money you minimally need to survive. This number differs according to location, age and living conditions. If you would use medication for example, your social minimum is higher than if you don't. Because of this, the social minimum is already corrected for the specific prices in your municipality. I decided to use the variable "percentage of people with an income up to 120% of the social minimum" because if you have an income lower than that you live in absolute poverty. Clearly visible is the fact that there are far more poor people in the big cities as well as in the the northern part of the Netherlands (e.g. Groningen).



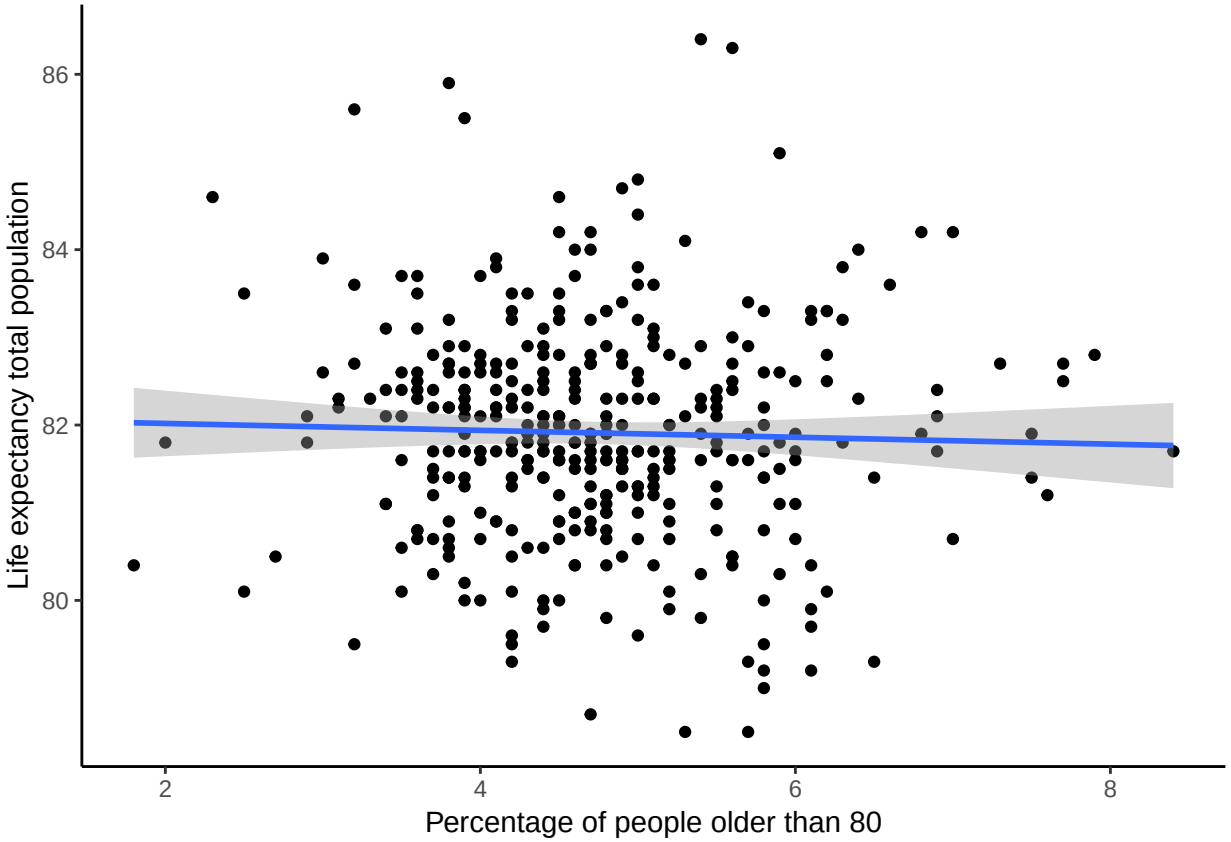
This map shows the percentage of people who self reported their health as: good/ very good. The municipalities that immediately stand out are Rotterdam, Kerkrade and Delfzijl, who are far below national avarage. This means that people in these areas feel less healthy than in the rest of the country. This is interesting considering all the other variables where Rotterdam and North East Groningen both also score worse than the national avarage.

3.3 Life expectancy and the number of elderly people

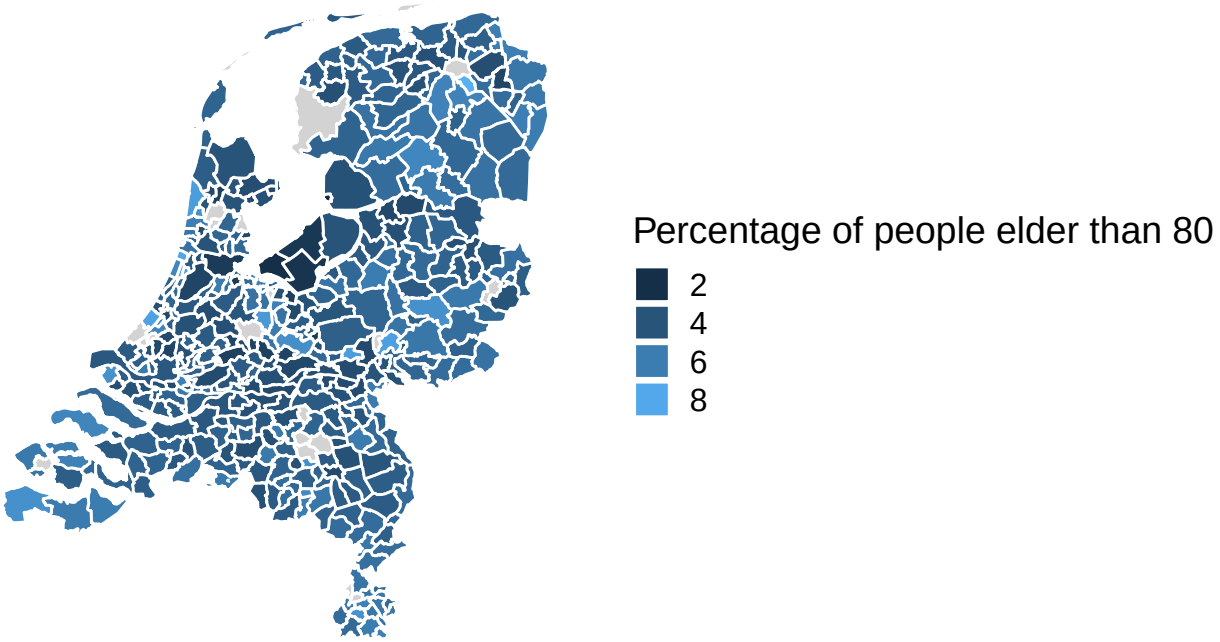
I downloaded the data about the life expectancy per municipality from the RIVM. On their website they state that this number is a prediction based on the number of deaths in every age category per municipality. But, because municipalities differ enormously in size, the prediction might not be very stable. Because of this RIVM has a model to correct for this. It is not possible for us to evaluate their system and their model to compute corrections, since it is not an open source. I therefore check if life expectancy is not correlated with the number and age of very old people in a municipality.



In this funnelplot we do not see a relationship between the number of people over 80 and the life expectancy in a municipality.



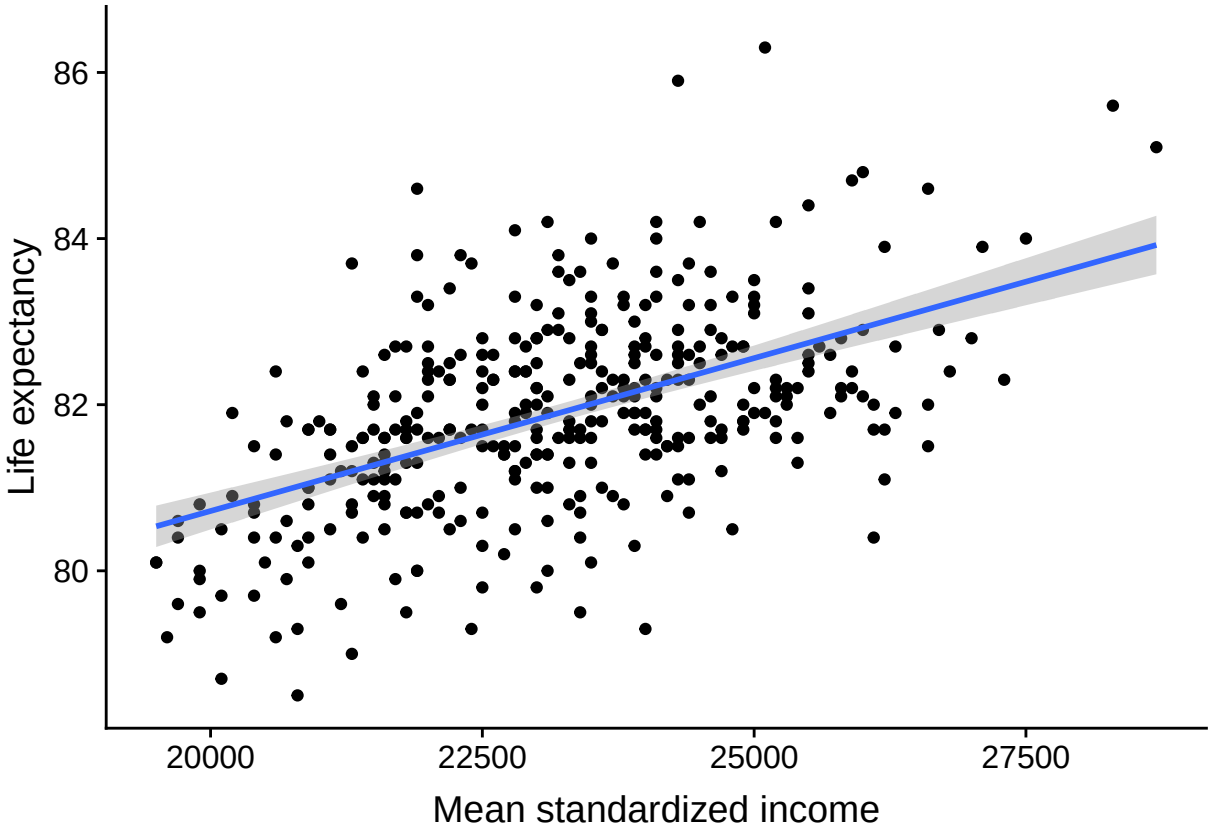
We also do not see a relationship between the percentage of people older than 80 in a municipality and the life expectancy in that municipality.



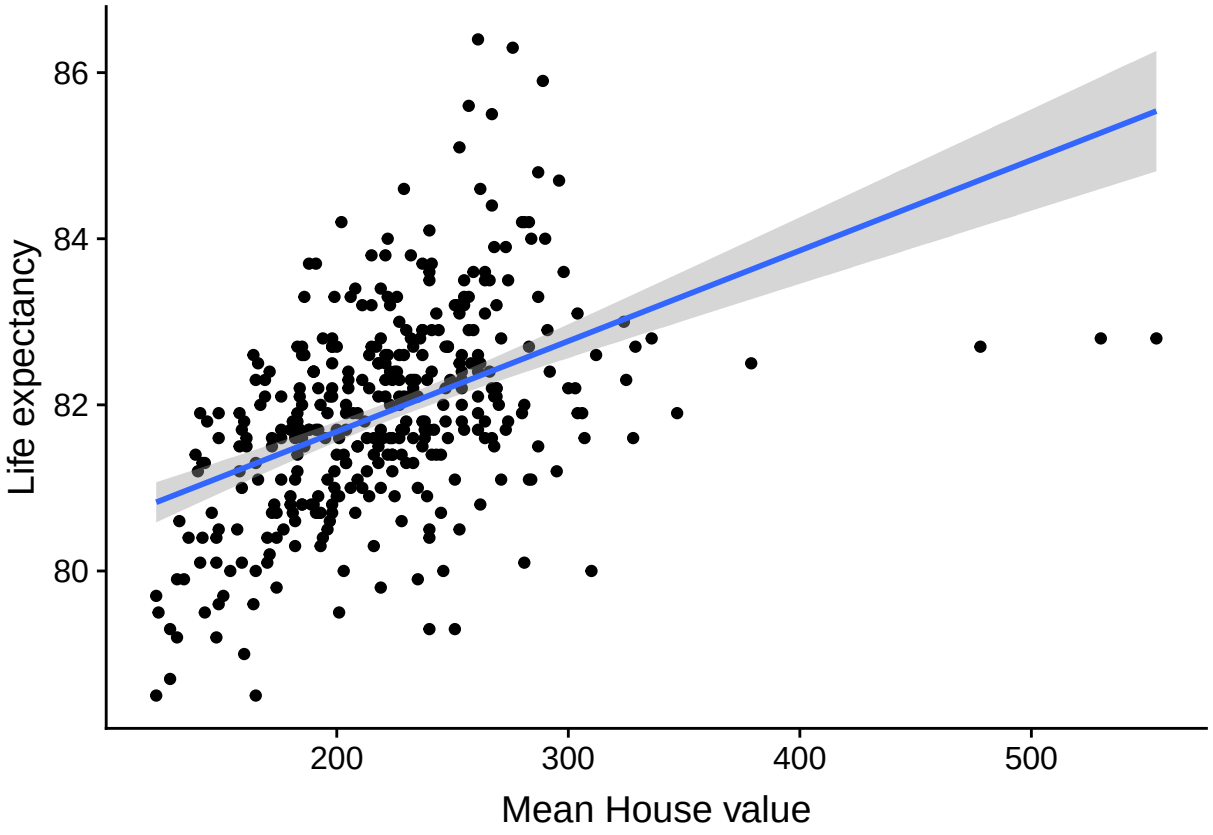
Fortunately for me, we do not see a relation between the percentage of people over 80 nor the absolute number over 80, which means these predictions of RIVM are not just based on the number of elderly in a municipality.

3.4 Relationship life expectancy with most important variables

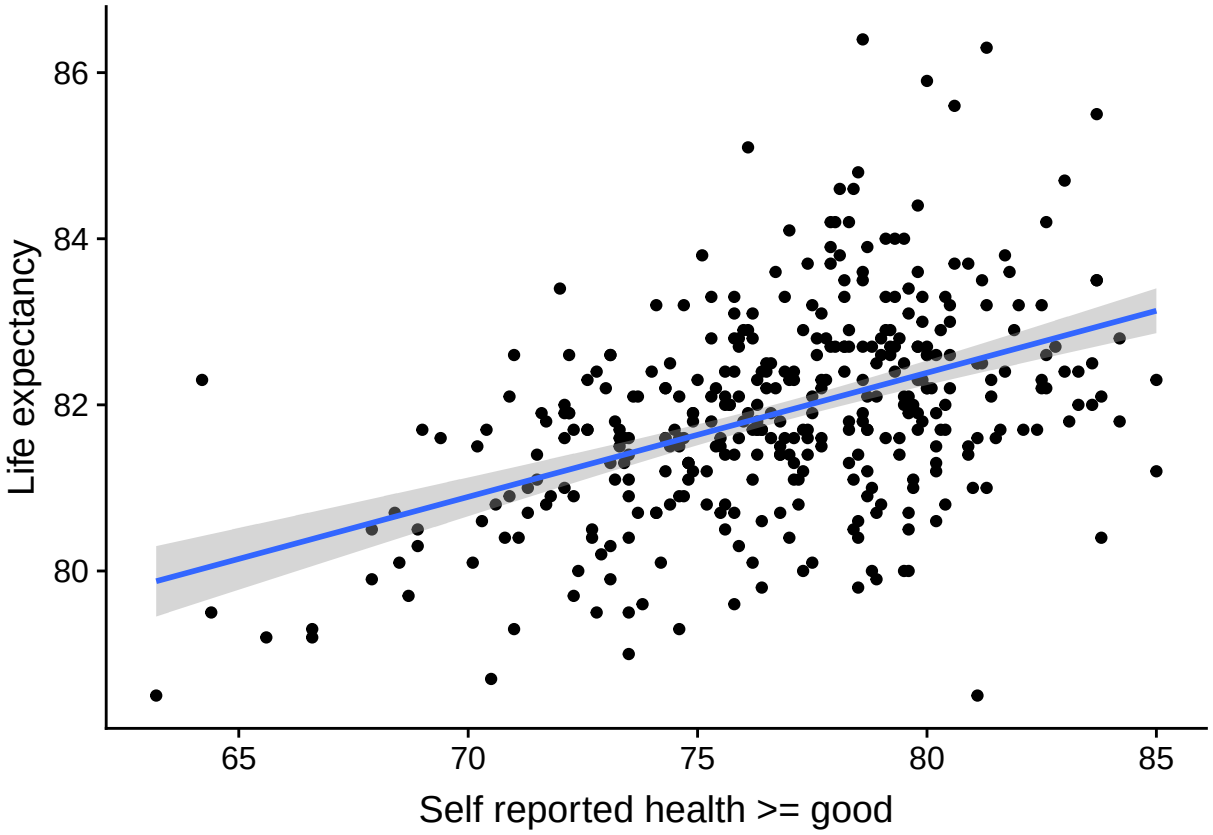
I made some simple graphs with the variables that have a big correlation according to the correlationplot. This includes the mean standardized income per municipality, the mean value of the houses of that municipality, the percentage of people who assess their own health as good to very good and the percentage of people with an income up to 120 percent of the social minimum. Because the variables are also strongly correlated (collinearity) with each other every result just indicates there might be a correlation between the life expactancy and the variable but not necessarily. R estimates the best linear regression (we will discuss that in the following paragraphs) and it will draw that as a blue line with the margin of error in grey. A larger grey area means more uncertainty. Most of the time when you have big outliers, R gives a big margin of error since it is uncertain about the α (the intercept) and β (the slope) in $y = \alpha + \beta x$



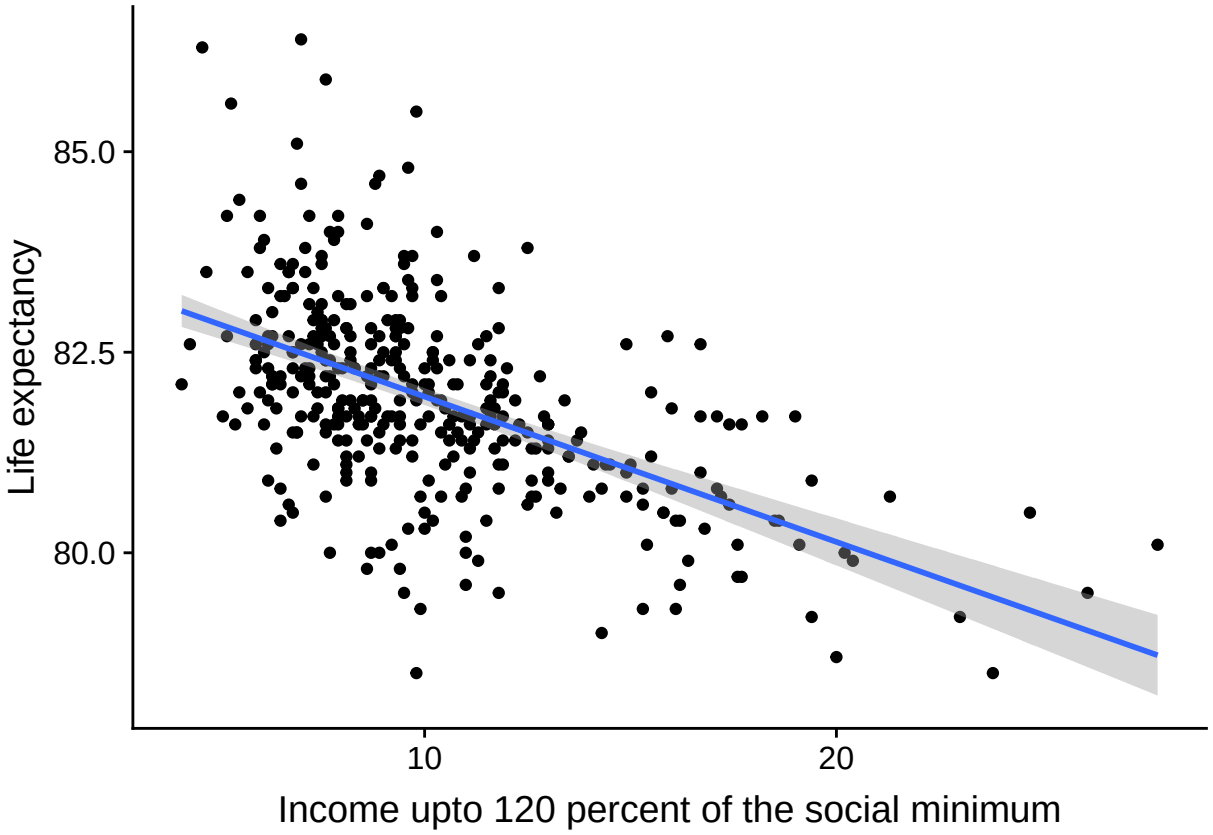
In this graph we can see the life expectancy vs the mean standardized income. We see a clear upward trend which is an indication that the life expectancy improves if the mean standardized income increases. This is not a defenitive relationship between these two since you might have other factors which improve life expectancy in the municipalities with a higher mean standardized income. This is why I will correct for other factors in the following paragraphs to truly see the effect the mean standardized income and all other variables have on the life expectancy.



In this graph we can see the life expectancy vs the mean house value in a municipality. We see a clear upward trend which is an indication that the life expectancy improves if the mean house value in a municipality increases. In this graph there are a number of very big outliers(the municipalities on the shores we talked about earlier for example) which make the estimated value of the α and β very uncertain.



In this graph we can see the life expectancy vs percentage of people who self reported their health as: good/ very good. We see a clear upward trend which is an indication that the life expectancy improves if the mean standardized income increases. In this graph there are a lot of outliers which make the estimated value of the α and β very uncertain.



In this graph we can see the life expectancy vs percentage of people with an income upto 120 percent of the social minimum. We see a clear downward trend which is an indication the life expectancy worsens if the percentage of people with an income upto 120 percent of the social minimum increases.

We can see that in all these graphs that none of the estimated variables is very predictive for life expectancy.

I would like to have a more formal method to describe the relationship between life expectancy and all variables. In the next paragraph we will introduce a methodology to test these relationships.

4 Methodology

In this paper I will use linear regression as technique to look at the relationschip between some factors (the predictors) and life expectancy of the total population (the response variable). I will use a prediction technique called random forest to assess if our model performs good, given the data. In this paragraph I will discuss linear regression and random forest. This paragraph is based on chapters 3 and 8 of James *et al.* (2013)

4.1 linear regression

To discuss linear regression, I will start with a very basic approach. I will assume that Y is a linear function of X and an error term. In mathematical form I assume:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

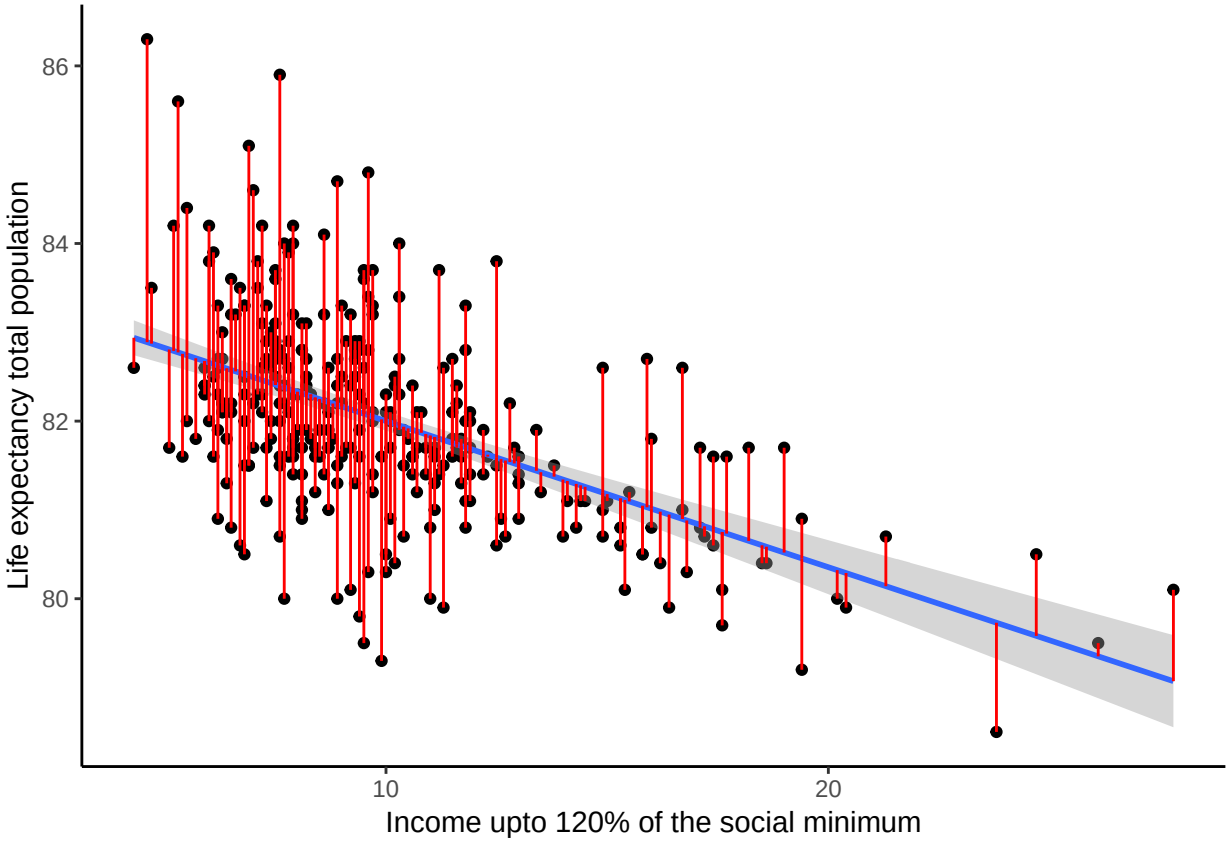
In terms of this paper I would label the life expectancy of the total population as Y, and for example the percentage of people with an income upto 120% of the social minimum in a municipality (Income upto 120% of the social minimum) as X. In a linear regression the goal is to estimate the parameters β_0 and β_1 so that the resulting regression equation is as close as possible to the data points.

With some parameters β_0 and β_1 I can predict the response ("$\hat{y_i}$") with the ith value of X by formula:

$$\hat{y_i} = \beta_0 + \beta_1 x_i$$

Then the difference between the predicted response $\hat{y_i}$ and the observed reponse in the data y_i is called the residual e_i.

In the figure below I have plotted Income upto 120% of the social minimum versus life expectancy of the total population.



In the figure I have plotted a regression line with the residuals (the red bars). As we can see, residuals can be either positive or negative. To prevent that the residuals will cancel each other out and to put more weight on very large residuals, it is common to square the residuals. The total sum of squares of the residuals is called RSS.

$$RSS = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

The best parameters β_0 and β_1 are the parameters that will minimize the total sum of squares (RSS) of the residuals. The minimizers are given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

And

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

I would also like to know if our esitmated β_0 and β_1 are accurate. Therefore I have to compute the standard error SE. I can use the following formulas:

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right]$$

and

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where $\sigma^2 = \frac{RSS}{n-k}$ for n observations and k predictors (in this example 2: β_0 and β_1).

Standard errors can be used to perform hypothesis tests on the coefficient β_1. If $\beta_1 = 0$, there is no relation between X and Y. It is common to formulate the hypotheses

$$H_0 : \beta_1 = 0$$

and

$$H_1 : \beta_1 \neq 0$$

I will test this hypothesis with the *t-statistic* t, which depends on the standard error:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta_1})}$$

This formula measures the number of standard deviations that $\hat{\beta}_1$ is away from zero. For a sufficiently large number of observations (normally more than 30), t is a normal distribution. Therefore, it is possible to compute the probability of observing any value equal $|t|$, under the assumption that $\beta_1 = 0$ ("the p-value"). With a small p-value it is possible to reject H_0. A rejection of H_0 means it is possible to conclude that there is a relation between X and Y. It is not possible to conclude that X causes Y without a causal model. Unfortunately I did not have time to study causal inference.

It is possible to extend the framework described above to a model with more predictors. In general the model with p predictors to be estimated would look like:

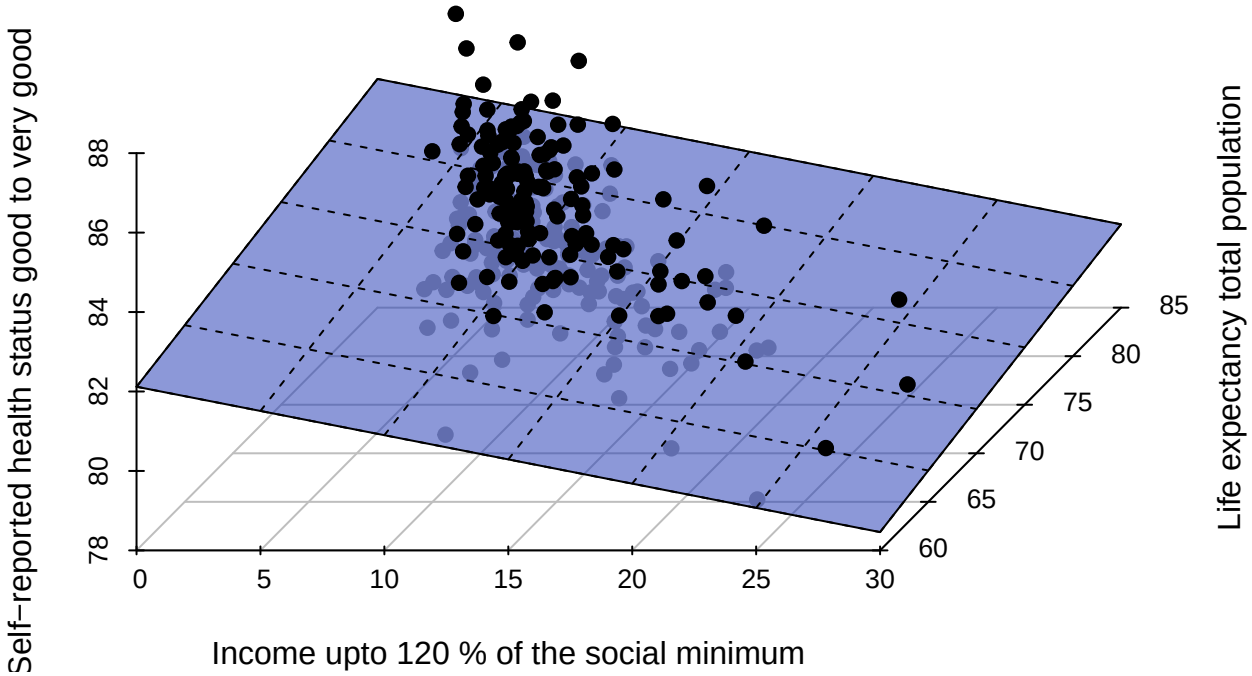
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

This formula will not lead to a "regression line", but to a "prediction plane".

It is not possible to make a plot with more than 2 predictors, but I can plot a regression plane with 2 predictors. In our example I plot the following equation:

$$LE = \beta_0 + \beta_1 * Inc\_SM + \beta_2 * Health\_status\_very\_good$$

Where LE is the life expectancy of the total population, Inc_SM income upto 120 % of the social minimum and $Health\_status\_very\_good$ is the percentage of people per municipality that score their health with "good" or "very good" (Self-reported health status good to very good).



Finally, the last statistic to be reported is the R^2. This statistic indicates which percentage of the variation in Y is explained by the model.

RSS is the *Residual Sum of Squares* and was given above by the formula

$$RSS = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

We can rewrite this formula to:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

because $e_i = y_i - \hat{y_i}$.

TSS is the *Total Sum of Squares* and given by the formula

$$TSS = \sum (y_i - \overline{y})^2$$

With the RSS and the TSS it is possible to compute the R^2:

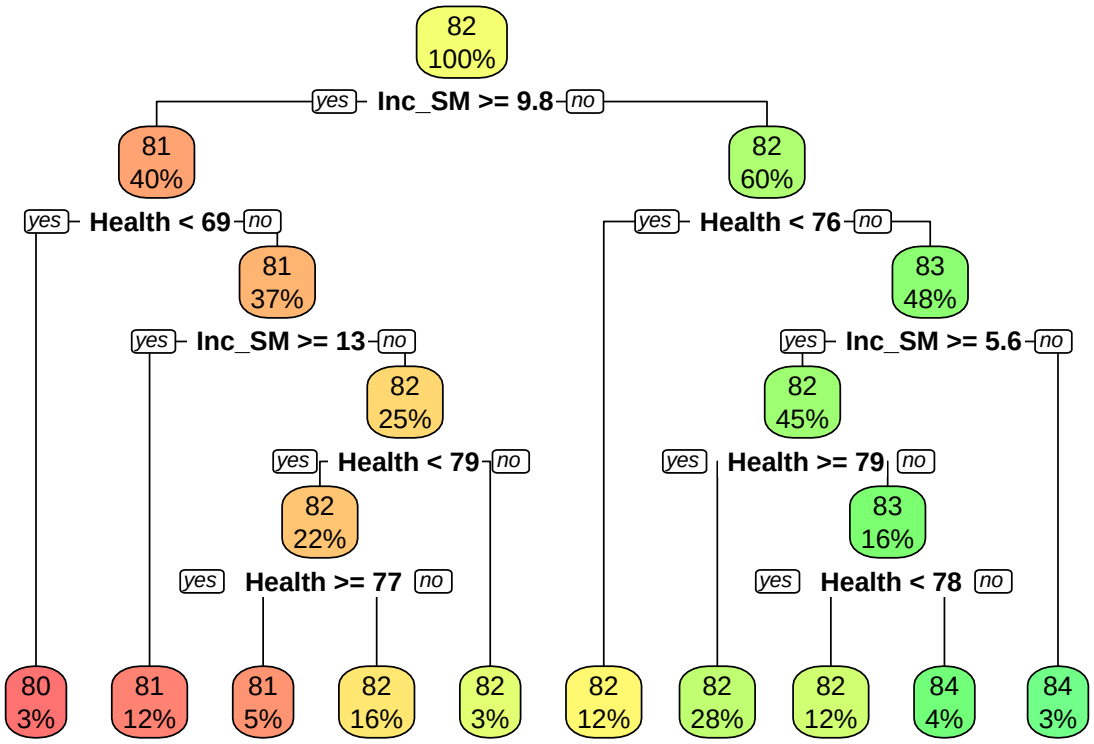
$$R^2 = 1 - \frac{RSS}{TSS}$$

4.2 Random Forest

In this paper I will use random forest to predict the life expectancy of the total population. Random forest is a prediction tool. Because it is able to pick up non-linearity and interaction terms, random forest predicts very well. However, the results are difficult to interpret.

To explain random forest, I have to explain the terms "random" and "forest". I will first turn to the term "forest".

A forest consists of trees. Random forest constructs many regression trees. I have plotted a regression tree to predict life expectancy of the total population based on the income upto 120 percent of the social minimum (Inc_SM) and the percentage of people to grade their health with good/very good (Health).



The boxes in the lowest row in the graph are called the "terminal nodes" (or in the tree anology "leaves"). The other boxes are known as the "internal nodes" The lines that connect the nodes are called the "branches".

At each node the tree splits in to 2 branches. For example, at the first split the municipalities are split according to the rule: Inc_SM >= 9.8. If a municipality fullfills this rule, it will go to the left branch. If not, it will go to the right branch. The percentage in the nodes indicates the proportion of the splitting.

The splits are done by minimizing the RSS. Since it is not possible to consider all possible splits to minimize the RSS, a "greedy" algorithm is used. In stead of looking to the best splits overall, the algorithm chooses the best split a each step.

A regression tree can be interpreted very well. Unfortunately, regression trees turn out to be very unstable. With slightly different data, regression trees will be formed totally different. Furthermore, the predictive accuracy of regression trees is pretty low. The idea of a "Forest" is to grow many trees (normally 5000) and take the average prediction of these trees. This boosts the accuracy and the robustness of the predictions.

The "random" element of random forests, refers to the characteristic of the algorithm to grow different trees. For each tree and at each split, a random sample of predictors is chosen to split the nodes. The algorithm chooses randomly a number (normally $\sqrt{number\_of\_predictors}$) of the predictors to split the node.

5 Analysis

In this paragraph I will do the formal analysis of the data and present the results of my analysis. First I will do a linear regression. Then I will check for collinearity and choose one main model. I will do some robustness checks by checking for overfitting and use a random forest model to check whether our model performs adequate, given the data. Finally I will look into predictors of random forest and marginal effects of my main model.

5.1 Regression

We saw in the paragaraph about methodology multiple demension regressions. In this case we have 21 variables so we will run a 21 demension regression. It is impossible to visualise anything with more demensions than 3, so we cannot visualise this, therefore the output is a table and not a graph.

I will estimate a linear regression of the folowing form:

$$\begin{aligned} LEtotalpop = \alpha + \beta_1 * Migration + \beta_2 * Mean\_inc + \beta_3 * pop\_density + \beta_4 * Mean\_HV + \\ \beta_5 * Inc\_SM + \beta_6 * Health\_status\_very\_good + beta_7 * Normal\_weight + \\ \beta_8 * Informal\_care + \beta_9 * House\_O + \beta_{10} * Dist\_Hosp + \beta_{11} * Fit\_norm + \\ \beta_{12} * Divorced + \beta_{13} * Lower\_edu + \beta_{14} * House\_density + \beta_{15} * Onepers_H H + \\ \beta_{16} * Multi\_morbidity + \beta_{17} * Weekly\_sporters + \beta_{18} * Householdsize + \beta_{19} * Total\_population + \\ \beta_{20} * percentage\_over80 + \beta_{21} * total\_over80 + \beta_{22} * Inc\_SM \end{aligned}$$

Which leads to the following results:

The R^2 tells us how far the data are from the fitted regression line. R^2 is given bij $\frac{explainedvariation}{totalvariation}$. In general the higher the R^2 the closer the observed data lies to the fitted line and the better your model fits your data. However a high R^2 isn't all that positive since you are probably overfitting and making a model on the datapoints you have, instead of making a model with your datapoints. If you are overfitting your model doesn't work on different data. The R^2 of the model is 44.7%. However, I will report the adjusted R^2. The adjusted R^2 takes the number of predictors into account. The adjusted R^2 increases only if an extra predictor improves the model more than would be expected by chance. My adjusted R^2 is 40.7% which is very common. In fields like economy and psychology the R^2 is most of the time between 40 and 60 percent while in physics an R^2 of 90% is pretty common.[1] Humans are just hard to predict.

The significance of the variables is indicated with the number of stars. We see a few very significant variables (indicated in the table with 3 stars). Those variables are:

- Health_status_very_good (percentage of people who self reported their Health as: good/ very good)
- Mean_inc (the mean standardized income).

I allready predicted that those variables would correlate with the life expectancy based on the correlation plot and the network plot. Still some variables that I would think to have a correlation with the life expectancy don't appear to be significant. I think this might have to do with multicollinearity: if some variables are strongly correlated with each other this will influence the estimation of the coefficients(the β's). And, therefore impact the confidence interfalls and their significance.

[1] source

Table 2: Results of the model with all predictors

	Dependent variable:
	LEtotalpop
Divorced	0.075 (0.071)
Migration	0.029* (0.015)
House_O	0.026* (0.015)
Mean_inc	0.0002*** (0.0001)
Lower_edu	−0.043** (0.018)
Pop_density	0.0003 (0.001)
House_density	−0.0004 (0.002)
Mean_HV	−0.001 (0.002)
Dist_Hosp	−0.028* (0.015)
Inc_SM	−0.086** (0.036)
Onepers_HH	−0.003 (0.032)
Health_status_very_good	0.083*** (0.025)
Multi_morbidity	0.025 (0.020)
Normal_weight	−0.013 (0.019)
Fit_norm	0.0004 (0.019)
Informal_care	−0.009 (0.022)
Weekly_sporters	−0.008 (0.015)
Householdsize	0.689 (1.152)
Total_population	−0.00000 (0.00000)
Percentage_over80	0.012 (0.088)
Total_over80	0.00000 (0.0001)
Constant	68.735*** (4.783)
Observations	313
R^2	0.447
Adjusted R^2	0.407
Residual Std. Error	0.882 (df = 291)
F Statistic	11.188*** (df = 21; 291)
Note:	*p<0.1; **p<0.05; ***p<0.01

5.1.1 Multicollinearity

As we have seen in the correlatation plots, some of the predictor variables are correlated with each other. This is problematic for linear regressions, because "collinearity" makes separtion of individual effects of the predictors difficult. In other words, "collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error [of the coefficients] to grow" (see page 101 of (James *et al.*, 2013)). We can see multicollinearity in the correlation plot. However, correlation plots do not reveal all collinearity problems in the data (e.g. between three or more variables in the data).

5.1.2 VIF

A formal check for multicollinearity is the *Variance inflation factor* (VIF). The VIF is calculated by the formula (source of the formula is James *et al.* (2013) page 102):

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

Where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j on all other predictors. So if $R^2_{X_j|X_{-j}}$ close to 1, collinearity will be present and the VIF will be large. James *et al.* (2013) advises to exclude variables with a VIF > 5 (or VIF >10) from the analysis. I have chosen to use to treshold of 5 for VIF.

It is possible to use the R package *car* to calculate the VIF. The results are:

Table 3: Multicollinearity in the first model

Divorced	5.705
Migration	5.810
House_O	5.875
Mean_inc	3.720
Lower_edu	2.522
Pop_density	365.394
House_density	390.955
Mean_HV	3.154
Dist_Hosp	1.914
Inc_SM	7.152
Onepers_HH	16.794
Health_status_very_good	3.371
Multi_morbidity	2.153
Normal_weight	2.895
Fit_norm	2.581
Informal_care	1.276
Weekly_sporters	3.173
Householdsize	14.504
Total_population	21.866
Percentage_over80	3.011
Total_over80	21.556

The following variables have a VIF larger than 10:

- Pop_density
- House_density
- Onepers_HH
- Householdsize
- Total_population
- Total over80

There are a few variables that have a VIF between 5 and 10:

- Divorced
- Migration
- House_O
- Inc_SM

To select my variables for the main model, I have done a stepwise selection procedure. Each time I deleted the variable with the highest VIF (> 5), until no variables had a VIF > 5. Since they are correlated we do not have to throw away every variable with a VIF-rating higher than five, because if we delete one of the variables (for example, House_density) the VIF-rating of the correlated variable drops because the variable that it correlated with (in the case of House_density: Pop_density (population density)) has been taken out of the dataset. I had to take the following variables out of our dataset: House_density, Total_over80, OnepersHH, Inc_SM, Householdsize and Migration. After I deleted Migration there weren't any variables with a VIF higher than five, as you can see in the table below.

Table 4: Multicollinearity in the main model

Divorced	3.457
House_O	3.053
Mean_inc	3.405
Lower_edu	2.261
Pop_density	2.139
Mean_HV	2.870
Dist_Hosp	1.804
Health_status_very_good	2.713
Multi_morbidity	2.092
Normal_weight	2.194
Fit_norm	2.481
Informal_care	1.255
Weekly_sporters	2.962
Total_population	1.758
Percentage_over80	1.645

Therefore, I can now estimate a linear model of all remaining variables. I will call this "The main model", which looks like this:

$$\begin{aligned} LEtotalpop = \alpha + \beta_1 * Mean\_inc + \beta_2 * pop\_density + \beta_3 * Mean\_HV + \beta_4 * Health\_status\_very\_good + \\ \beta_5 * Normal\_weight + \beta_6 * Informal\_care + \beta_7 * House\_O + \beta_8 * Dist\_Hosp + beta_9 * Fit\_norm + \\ \beta_{10} * Divorced + \beta_{11} * Lower\_edu + \beta_{12} * Multi\_morbidity + \beta_{13} * Weekly\_sporters + \beta_{14} * Inc\_SM + \\ \beta_{15} * percentage\_over80 + \beta_{16} * Total\_population \end{aligned}$$

Which leads to the following results:

Table 5: Results of the main model

	Dependent variable:
	LEtotalpop
Divorced	0.047
	(0.056)
House_O	0.037***
	(0.011)
Mean_inc	0.0003***
	(0.0001)
Lower_edu	−0.050***
	(0.017)
Pop_density	0.0001
	(0.0001)
Mean_HV	0.00002
	(0.002)
Dist_Hosp	−0.036**
	(0.014)
Health_status_very_good	0.098***
	(0.023)
Multi_morbidity	0.020
	(0.020)
Normal_weight	−0.018
	(0.017)
Fit_norm	−0.001
	(0.019)
Informal_care	−0.009
	(0.022)
Weekly_sporters	−0.006
	(0.015)
Total_population	−0.00000
	(0.00000)
Percentage_over80	−0.043
	(0.066)
Constant	67.801***
	(3.132)
Observations	313
R^2	0.425
Adjusted R^2	0.396
Residual Std. Error	0.890 (df = 297)
F Statistic	14.638*** (df = 15; 297)
Note:	*p<0.1; **p<0.05; ***p<0.01

In this regression, made with only the variables which passed the VIF-test, we see more significant vararіables than before the VIF. Instead of just the two significant variables before we now have four: Health_status_very_good (percentage of people who self reported their Health as: good/ very good), Mean_inc (the mean standardized income), House_O (the percentage of of households who own the house the live in) and Lower_edu (the percentage of who are lower educated[2]). As expected, multicollinearity decreases significance of the coefficients.

The adjusted R^2 for this regression is 39.6%, which is only slightly lower than the R^2 in the first model with all predictors (40.7%).

5.2 Overfitting

As described above a model might be overfitted, which means that the estimated coefficients are tailored to the data and will poorly perfom on new data. Therefore, I will check for overfitting by a technique called cross validation.

Cross validation (see chapter 5 of James *et al.* (2013)) is a technique by which we will split the data randomly in different parts, which are called "folds". I will use 10 folds. Nine of these folds will be used to estimate the model and the performance of the model is checked at the 10th fold (the "test fold"). Each fold is used as test fold, which means that the folds "rotate". Therefore the model is estimated and tested 10 times. Because I will split the data 5 times in 10 folds, I will estimate and test the model in total 50 times.

Furthermore, I will split the data in a train set and a test set. I will estimate the coefficients with cross validation on the train data and then predict the outcomes in the test data (which were not used to develop the model) with the fitted model. If the R^2 does not differ too much between the train set and the test set, the model is not overfitted.

I can only run the cross validation on complete data (we have to delete the observations with NA's). By deleting NA's we lose 61 observations.

The results of the cross validation of the main model on the data are:

Table 6: Performance of cross validated regression

RMSE	R_squared	MAE
0.9343824	0.3754164	0.7414065

I conclude that overfitting is not much of a problem : the adjusted R^2 of the main model was 39.6%, the R^2 of the cross validated model is 37.5%.

[2]lower educated are people who finished at most, primary education, the first three years of highschool, lower vocational education(VMBO) or intermediate vocational education (MBO-level 1)

5.2.1 Random forest

As mentioned in paragraph 4.2 random forest picks up non linearities and interactions. I will use random forest to check if I can get more out of the model. I expect that random forest will lead to a higher R^2, because it will use all variables (because prediction is the goal, multicollinearity is no problem) and picks up possible non linearities and interactions. I do not expect that the extra variables in the random forest will cause a much larger R^2, because we already saw in the linear regressions that omitting the variables with a VIF ≥ 5 did not change the R^2 that much.

If random forest does not lead to a much higher R^2, I assume that I don't have to look for non linearities and interactions and therefore that the main model performs reasonably well, given the data.

I will run a random forest cross validated on my train data and display the results.

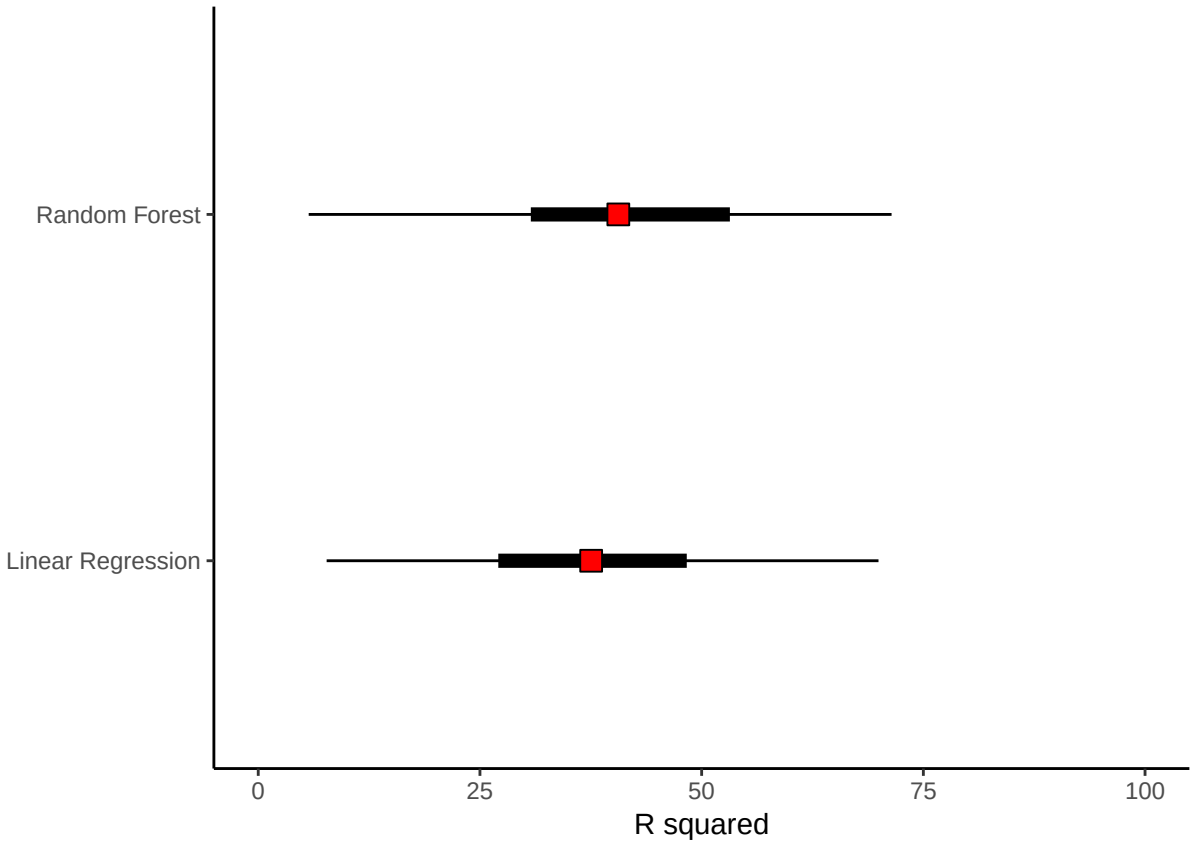
Table 7: Results random forest

mtry	splitrule	RSME	Rsquared	MAE
2	variance	0.9253983	0.3942824	0.7055045
2	extratrees	0.9263552	0.4045327	0.7095108
11	variance	0.9283274	0.3835930	0.7122703
11	extratrees	0.9118445	0.4060536	0.7021058
21	variance	0.9343824	0.3782256	0.7174154
21	extratrees	0.9131232	0.4033399	0.7022412

Tuning parameter 'min.node.size' was held constant at a value of 5. RMSE was used to select the optimal model using the smallest value. The final values used for the model were mtry = 11, splitrule = extratrees and min.node.size = 5.

The main model performs pretty well: the adjusted R^2 of our main model was 39.6% (Cross Validated was 37.5%), the R^2 of the random forest is 40.6%. This means that I don't need to look for non-linearities and interaction terms.

I can show this graphically

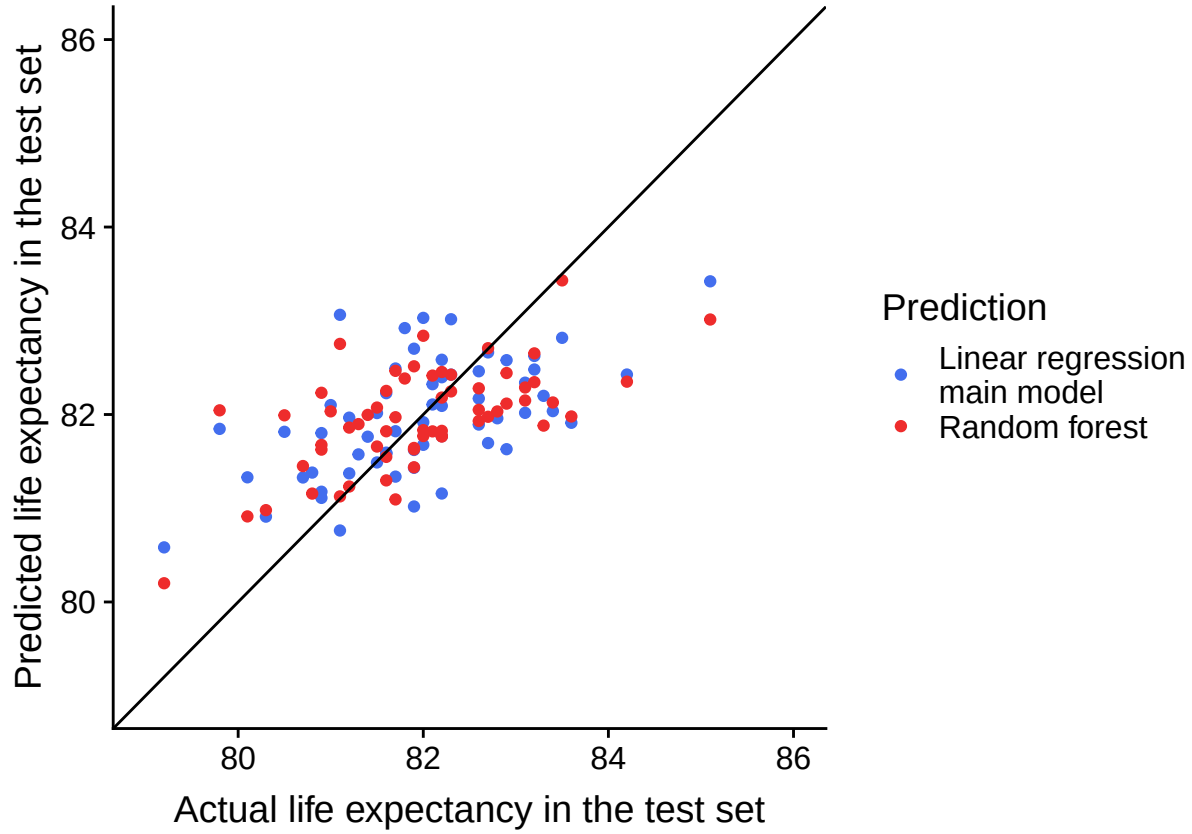


The confidence intervals are large, due to the fact that I only have 313 municipalities in the complete cases dataset.

5.3 Interpreting the results

5.3.1 How well does our model predict?

I will predict the results of the cross validated regressions and random forest on the test set. I will present a plot of real values versus predicted values.

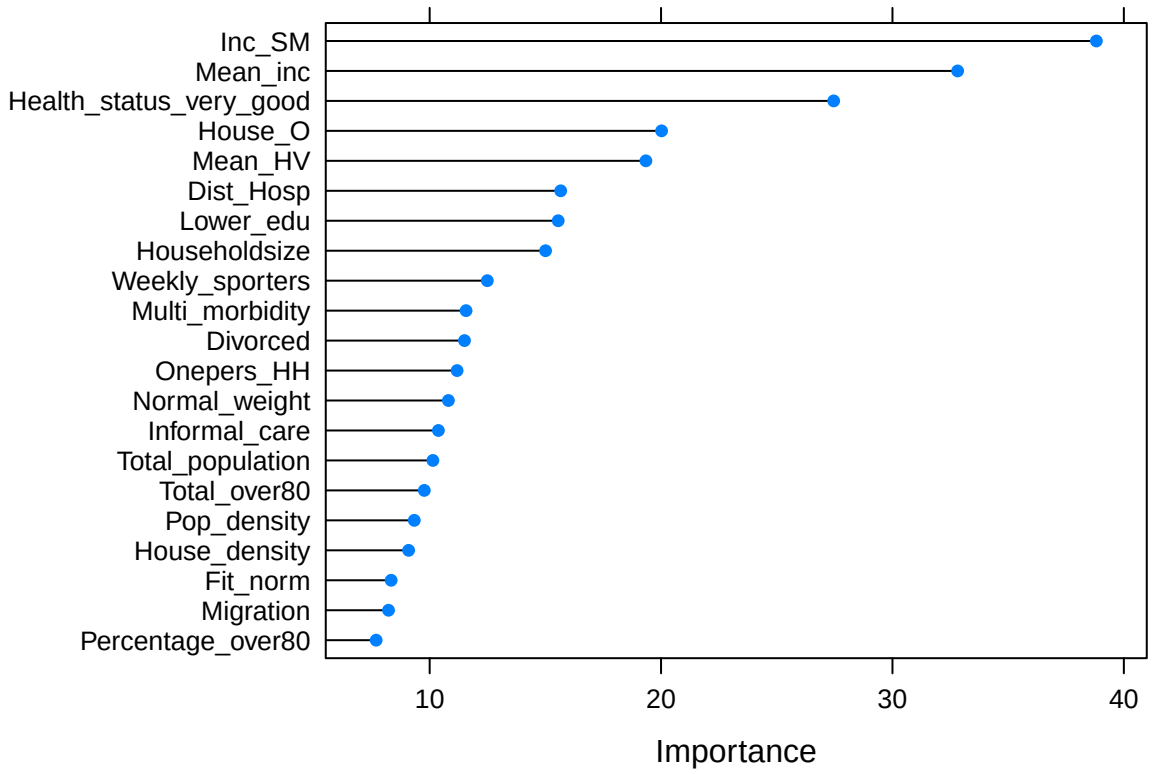


The plotted line in this graph has an angle of 45 degrees. Any prediction of life expectancy of a municipality by the models that would be equal to observed life expectancy of a municipality, would be on the line. The blue dots represent the predictions by the main regression model, the red dots represent the predictions by the random forest model.

The main model performs reasonably well, given the data. In other words, it does not seem possible to improve the analysis with assuming non-linearity and interaction between variables.

5.3.2 Variable importance of random forest

Random forest gives the possibility to plot the variable importance (a measure of how much each variable contributed to the prediction).

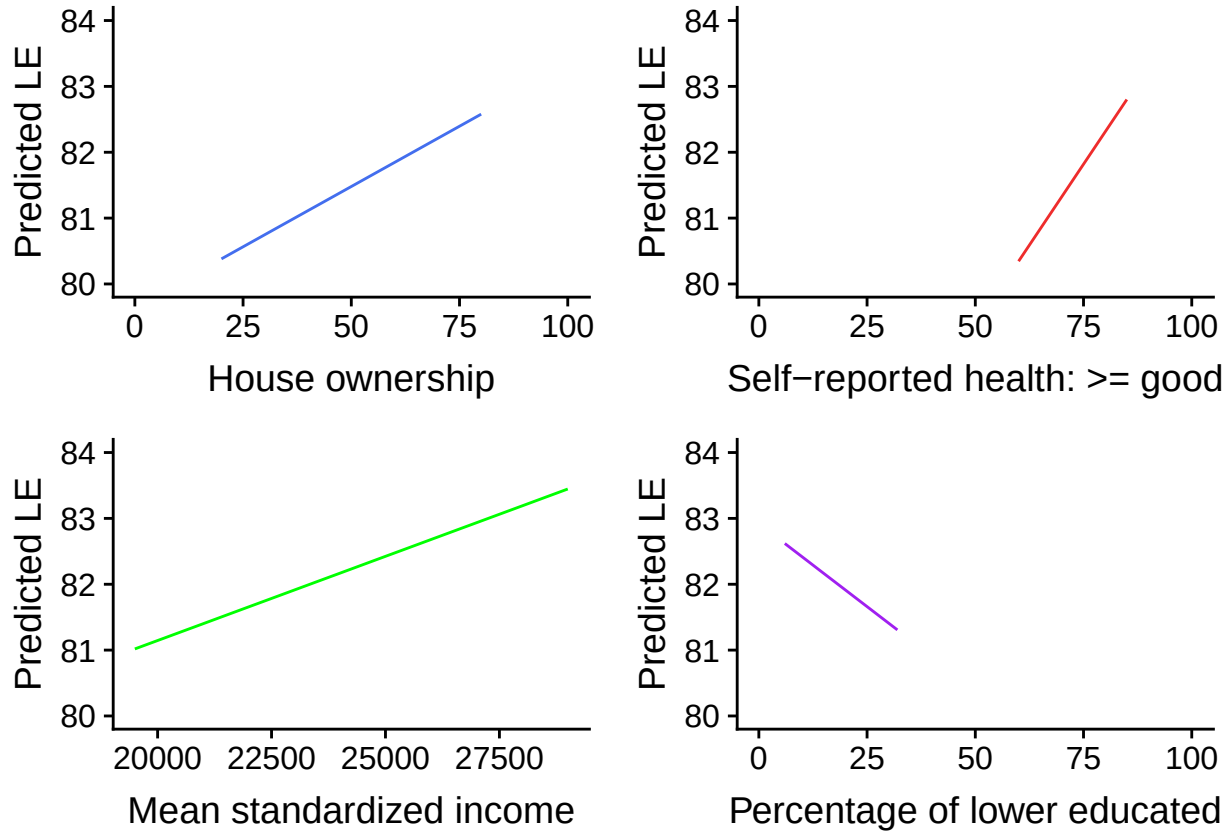


We can see that the variables that looked important in our correlation plots (Inc_SM etc), contribute most to the predictions of random forest.

5.3.3 Marginal Effects

For the regressions, it is possible to plot the marginal effects. Marginal effects are the changes in the response variable (in our case life expectancy) and a change in 1 predictor (e.g. Mean_Inc) while holding all other predictors constant. In case of a linear regression, the marginal effects are equal to the estimated coefficients (β's).

In the plot I only show the most significant variables.



The marginal effects in the main model differ from the variable importance of the random forest model. In the random forest model I used all variables as predictors, while in the main model I omitted variables due to multicollinearity.

6 Conclusions

In this study, I have examined the impact of various factors like social economic status, ethnicity, health care facilities and health status on life expectancy in Dutch municipalities. I have found that a higher social economic status is associated with a higher life expectancy. In the main model variables that are proxies for social economic status like the mean standardized income, the percentage of lower educated people, the percentage of people who own the house they live in, are highly significant with the expected signs (a higher percentage of lower educated people decreases life exectancy in the model, while the other mentioned variables increase life expectancy).

Ethnicity does not seem to have a significant impact on life expectancy. In the model with all predictors the percentage of people with a migration background is not significant. This variable is not included in the main model, because this variable is strongly correlated with other variables that are proxies for social economic status. Migration is also not an important variable in the random forest analysis. Therefore, I conclude that ethnicity is not an important factor for life expectancy. However, people with a migration background may have a lower life expectancy, because on average they have a lower social economic status.

The distance to nearest hospital has in the main model a negative significant (p-value = 0.014) impact on life expectancy. This means that the larger distance to the nearest hospital the more the life expectancy decreases. In the random forest this is a pretty important variable as well. This is probably because if you need acute medical attention the fact that there is a hospital close could save your life. The access to informal care is not a significant factor in our regressions.

I found only one significant variable which would suggest that health status is an influential factor on the life expectancy, which is the percentage of people who self reported their health as good or as very good. All other variables are not significant. But I suspect that is because those variables play a big part in if people feel healthy. So I decided to run another regression but this time not with Life expectancy as the depended variable but with the percentage of people who self reported their health as good or as very good[3]. In this regression, Divorced (the percentage of people who is divorced), Lower_edu (percentage of lower educated), Mean_HV (mean of the house value of that municipality), Multi_morbidity (the percentage of people with one or more physical defects; disabled persons), Fit_norm (the percentage of people who are able to pass the "fitnorm"(being able to preform heavy phisical acctivities for 20 minutes, three times a week)), Weakly_sporters (percentage of people who sport every week), Percentage_over80 (percentage of people older than 80) are very significant. Except Mean_HV and Fit_norm all of these variables are negatively significantly correlated with the the percentage of people who self reported their health as good or as very good. So those variables do probably impact the Life_expetancy but the system wasn't able to pick most of these up because they are already represented in the percentage of people who self reported their health as good or as very good. Notable is that a divorce negatively impacts how healthy you feel, aswell as that a lower education negatively impacts how healthy you feel.

There were some limitations to this research however; most notably that I did not use a causal model, therefore I cannot conclude that significant factors *cause* a change in life expectancy. Since I did not have enough time to study causal inference. I also could not measure environmental factors, which could have been important. It probably would have increased the explanatory value (a higher R^2) of the main model.

In the future I would like to extend this paper with a causal model. I also did not look into the differences in life expectancy between men and women. Those life expectancies seem to differ a lot and react different to different variables. I made some graphs to show this, those are situated in appendix 2. In future research more attention could be paid to the differences in life expectancy between men and women to improve my model.

Finally, I would like to conclude this paper with two policy recommandations:

Social economic status seems to be really important. The government should aim at improving the social economic status by e.g. investing in more and better eduction and circumstances that people with low SES live in (for example better houses, easier access to sport facilities).

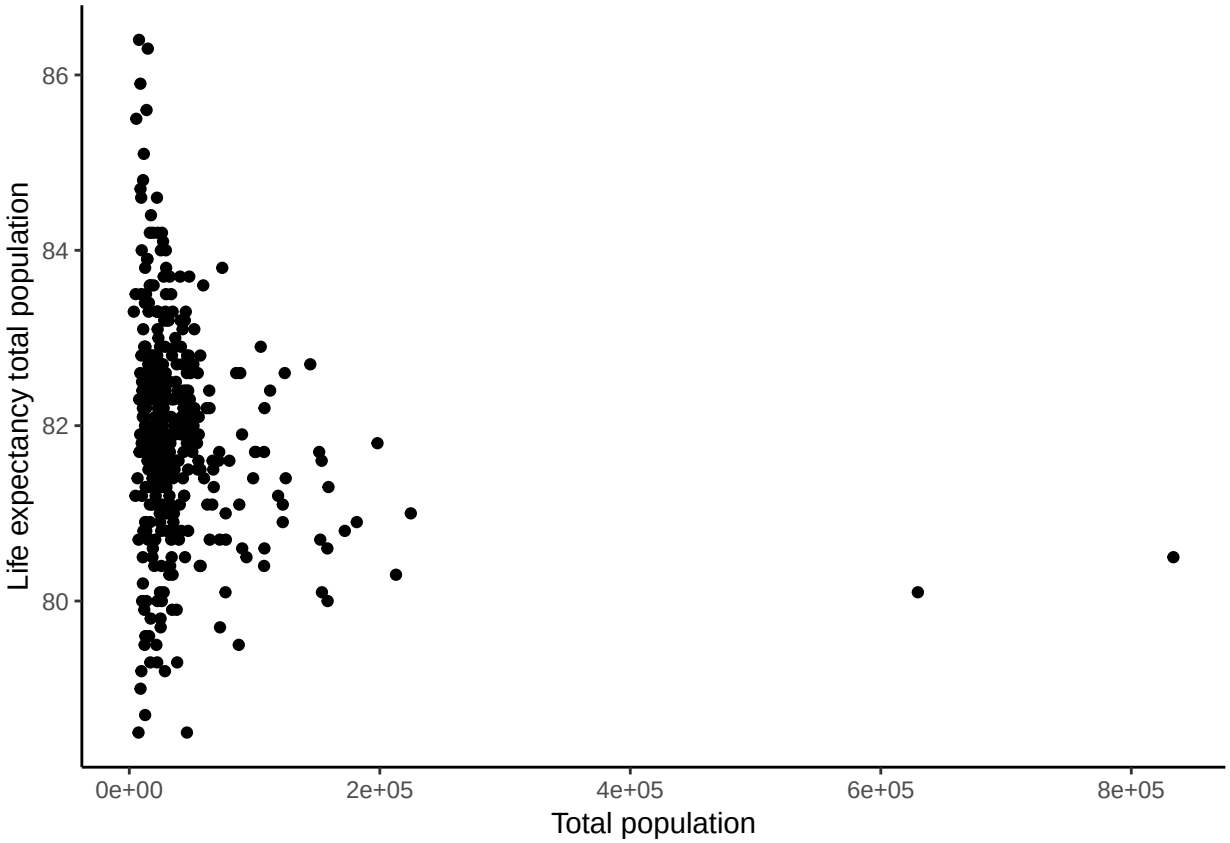
[3] appendix 3

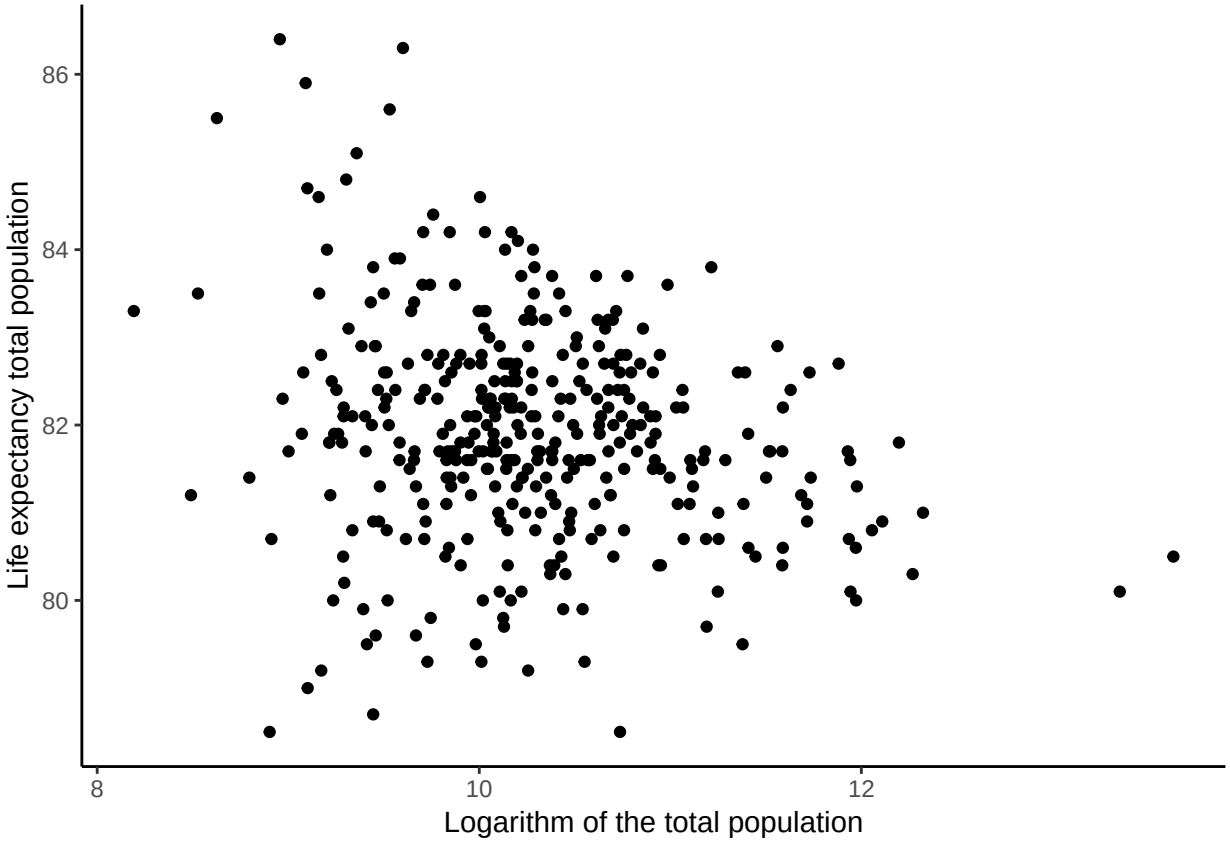
There is a lot of discussion about mergers between and bankruptcies of hospitals in the Netherlands. I would recommend that when decisions are taken about mergers and bankrupties the impact of an increased distance on life expectancy is taken into account.

Appendix 1: Extra figures

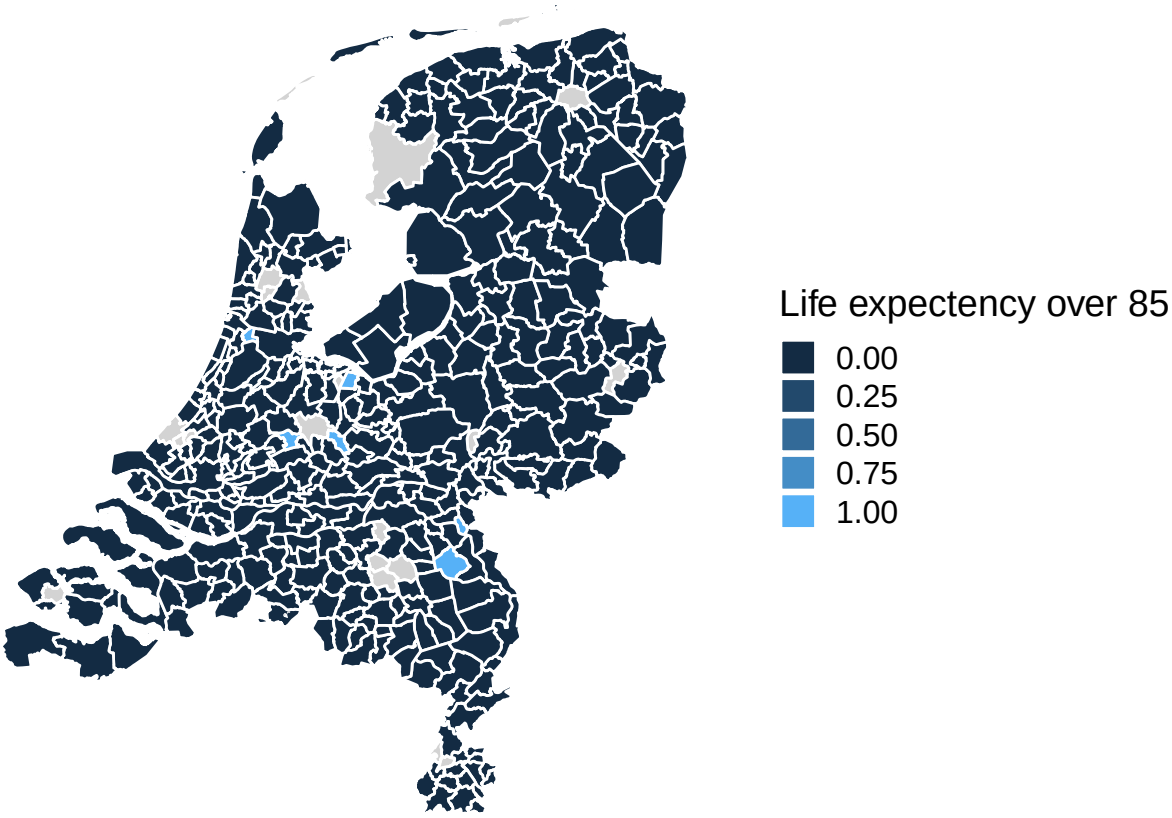
6.1 Relation of life expectancy and total population

I present a funnel plot with life expectancy as a function of the total population per municipality. To present the same information in a more accessable way, I have also taken the log of the total population.

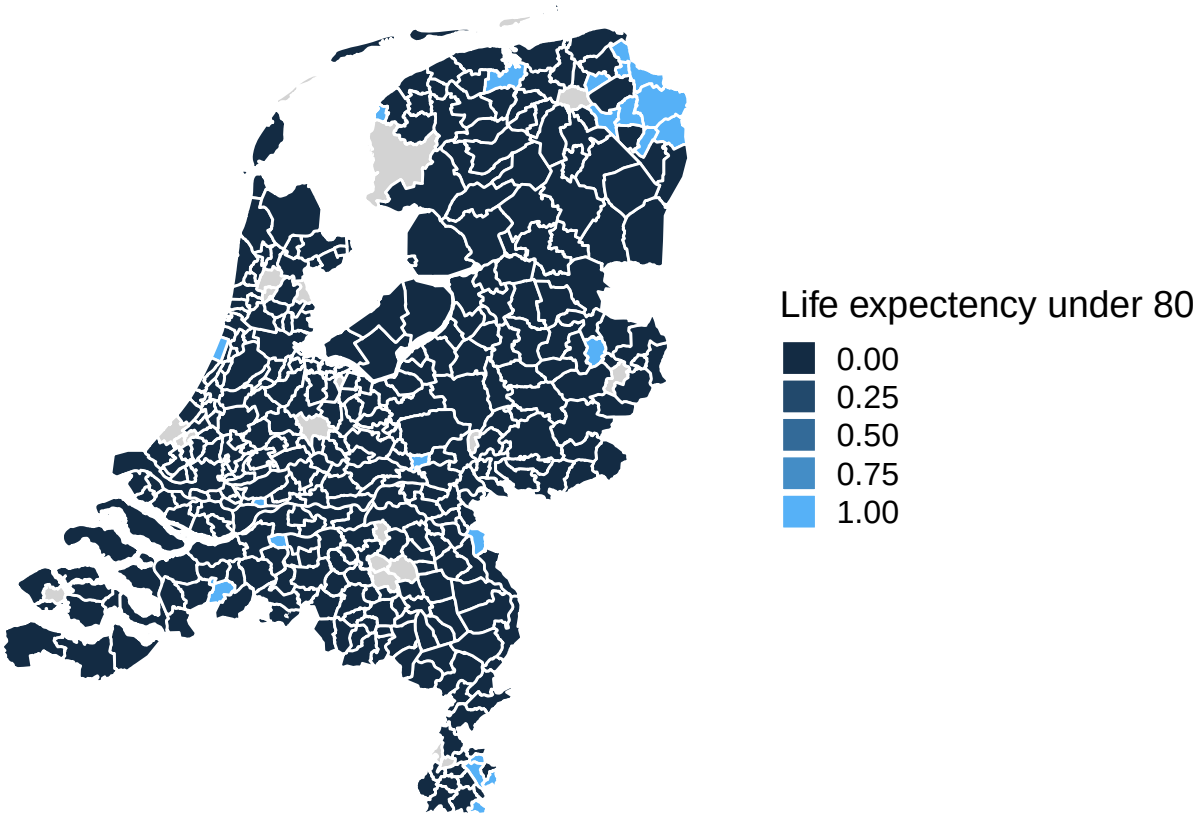



Life expectancy total population
86
84
82
80
8
10
12
Logarithm of the total population

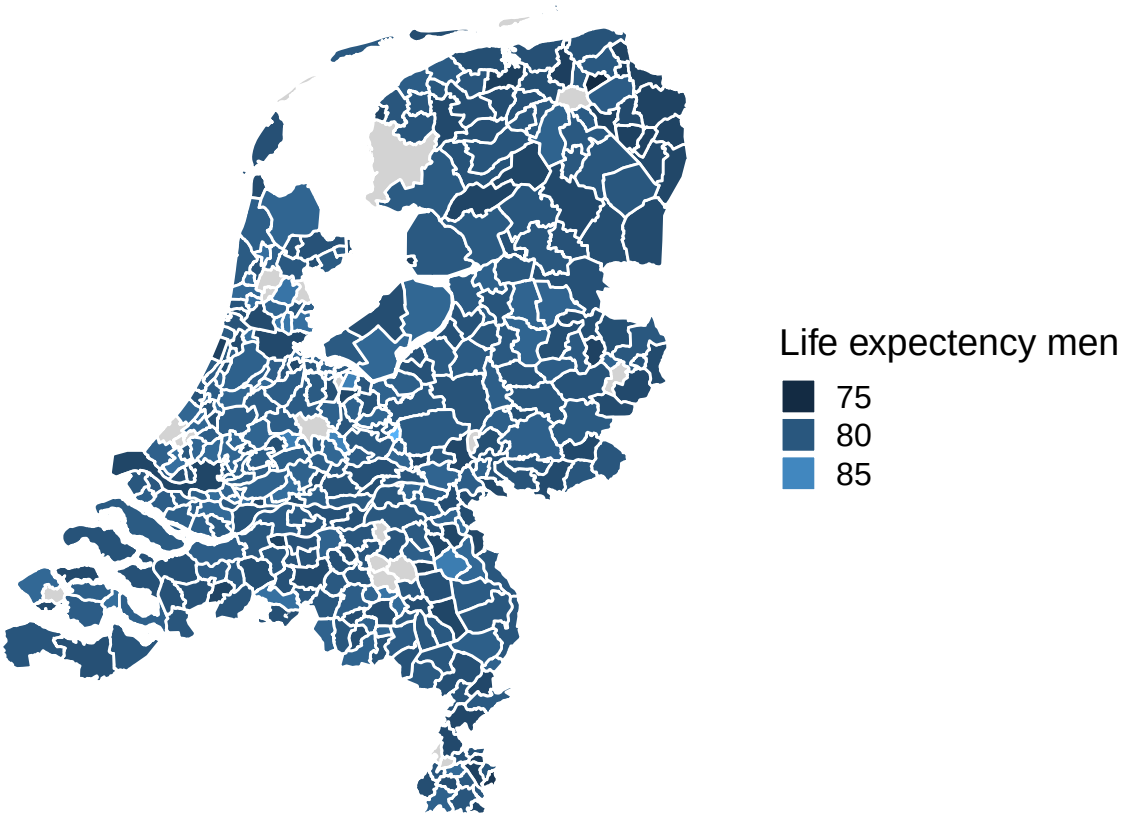
6.2 Various maps

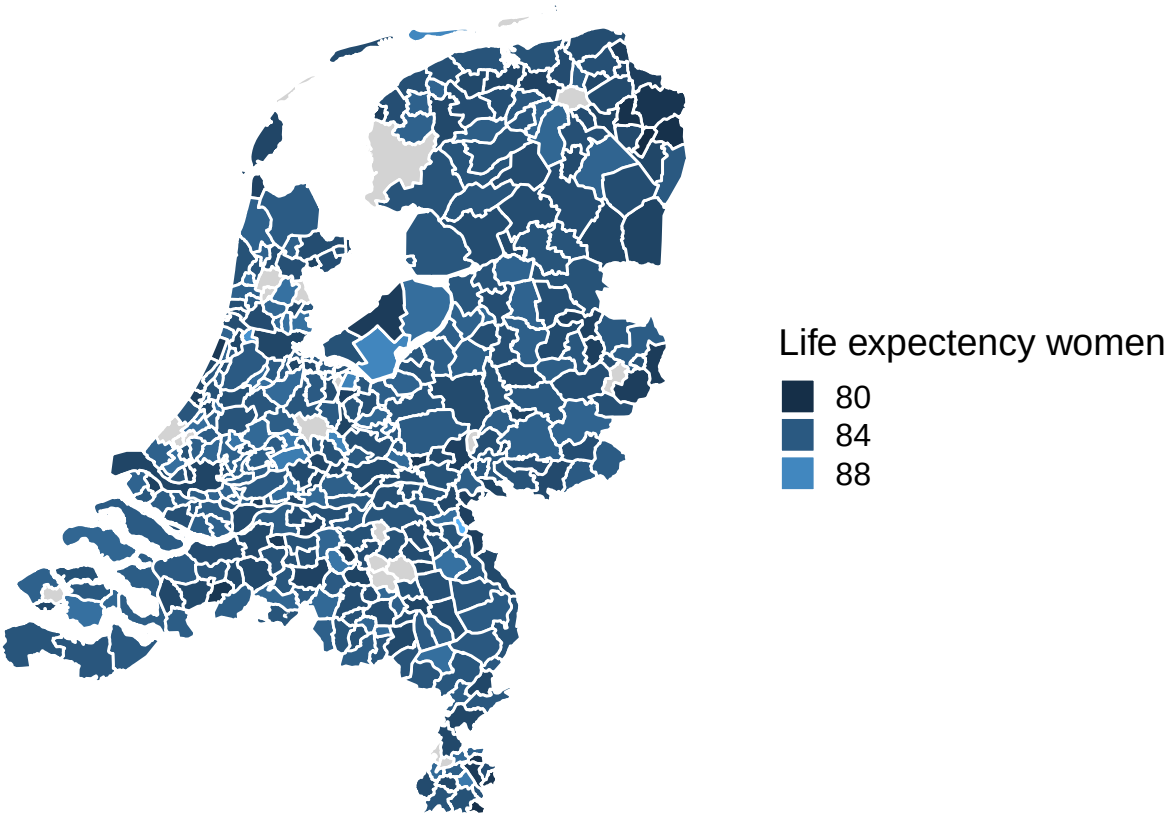


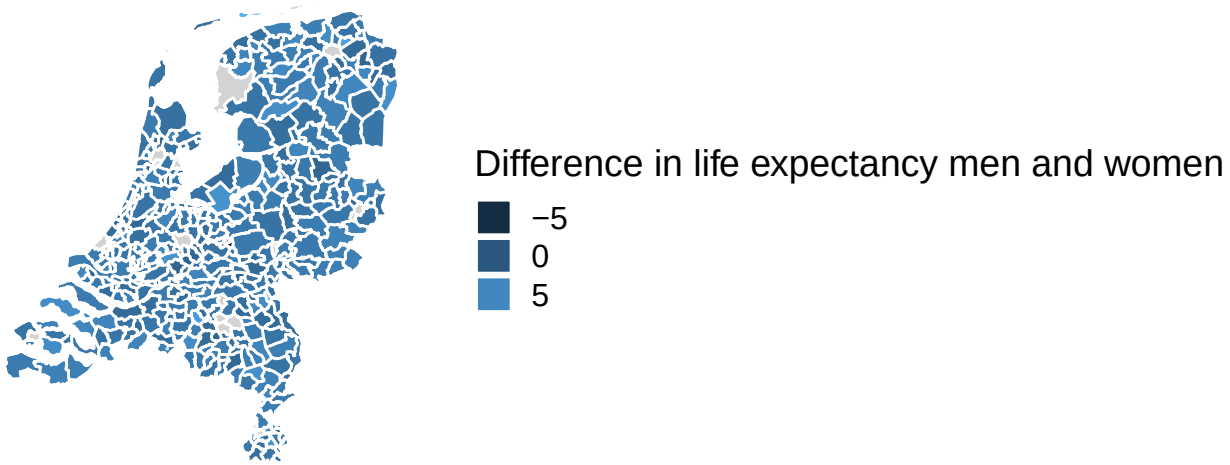
In this map you see the municipalities with the highest life expectancy of the Netherlands. Every municipality which has a life expectancy over 85 is lightblue, every municipality which has a life expectancy under 85 is darkblue.



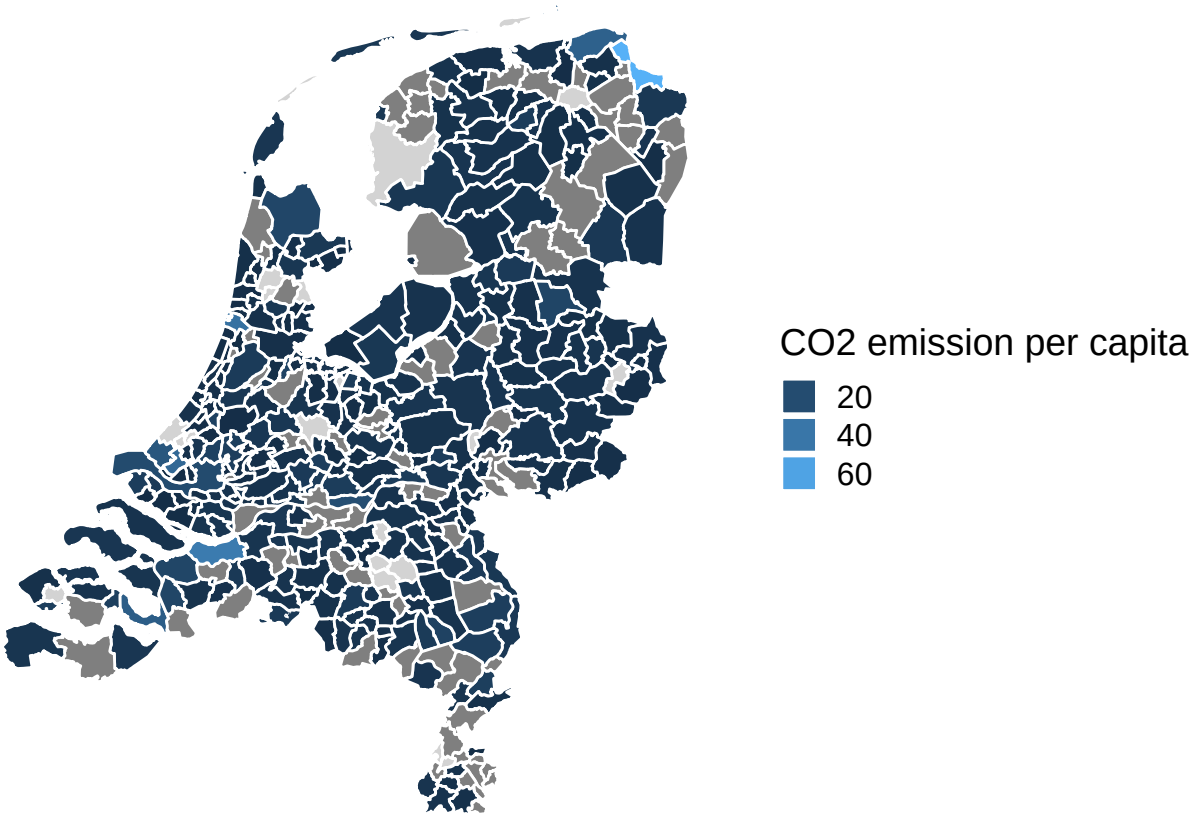
In this map you see the municipalities with the lowest life expectancy of the Netherlands. Every municipality which has a life expectancy under 80 is lightblue, every municipality which has a life expectancy over 80 is darkblue. The obvious spots which directly catch attention are the North-eastern part of Groningen and the Eastern part of Limburg.


Life expectency men
75
80
85

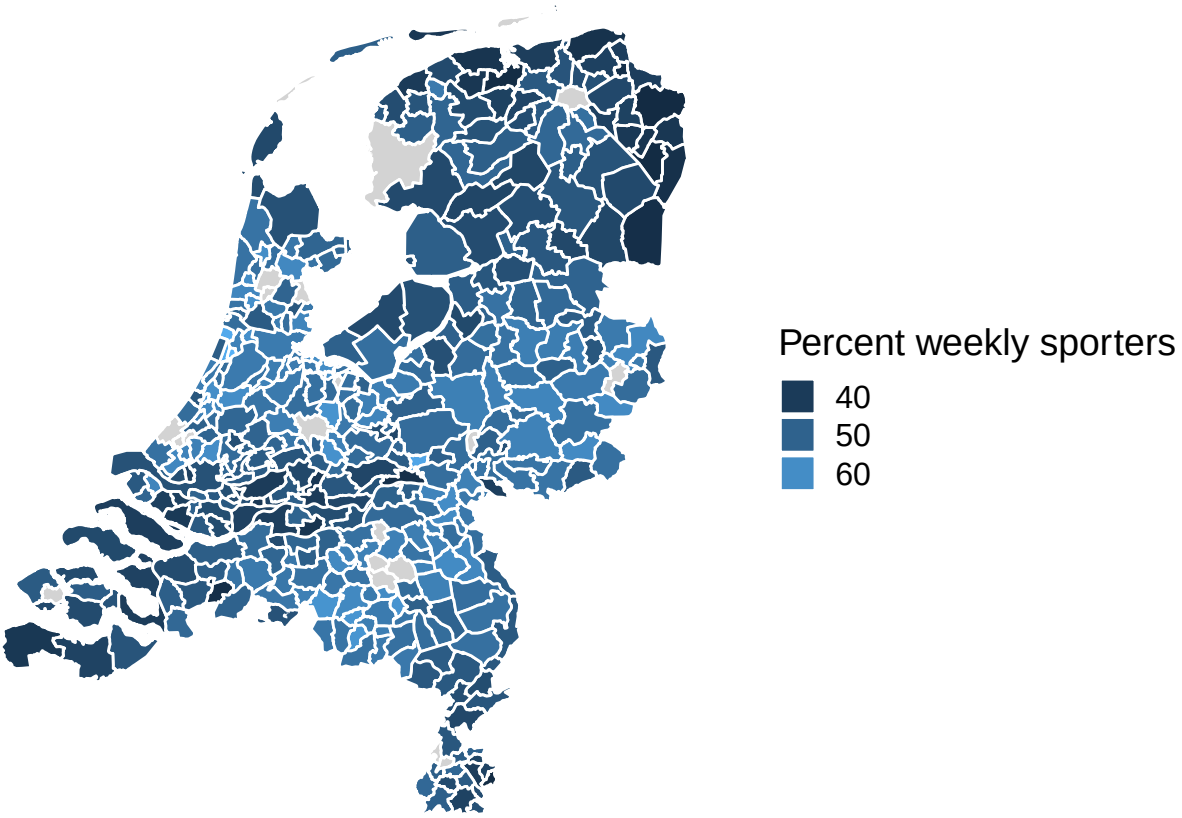

Life expectency women
80
84
88



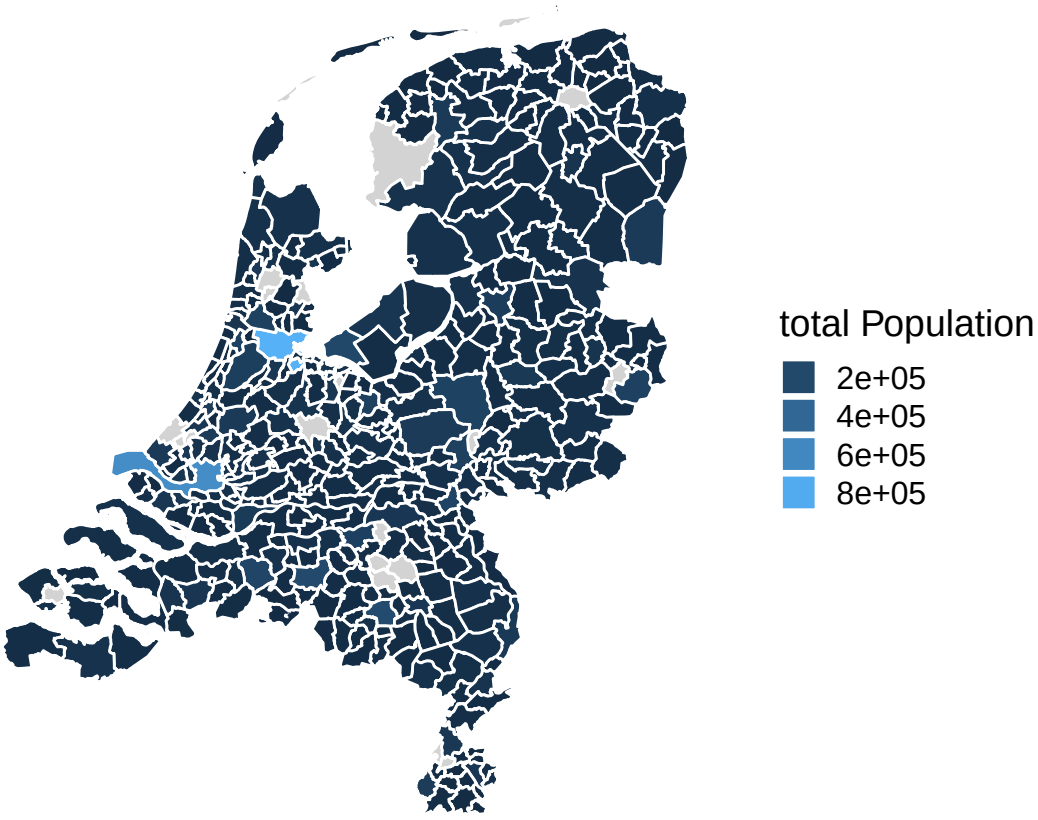
In this graph you see the life expectancy of women minus the life expectancy of men. There is only one municipality in the Netherlands where women generally die earlier than men (Renswoude), marked with a dark blue dot on the map.



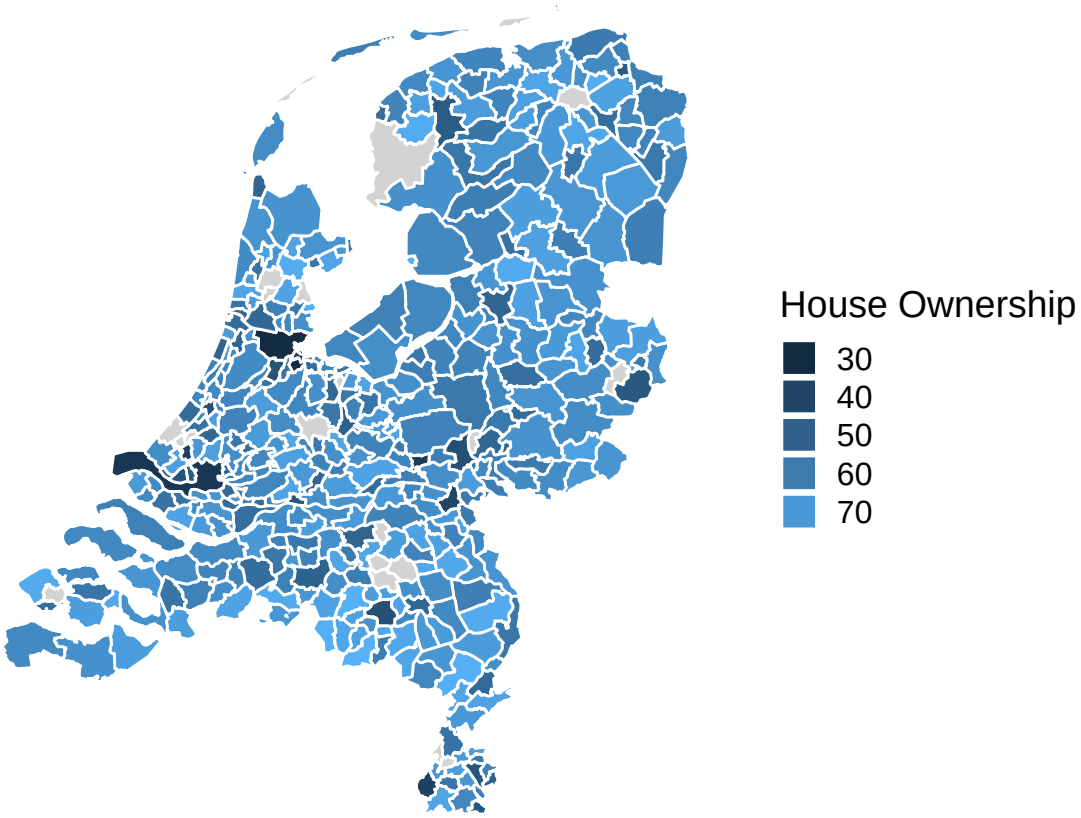
The total CO2 emmision divided by the number of people who live in that municipality. Delfzijl stands out with its enormous emissions without many people living there.



This plot presents the percentage of people who sport at least once a week. Notice the North-eastern part of Groningen where sport clearly isn't daily business so to speak.



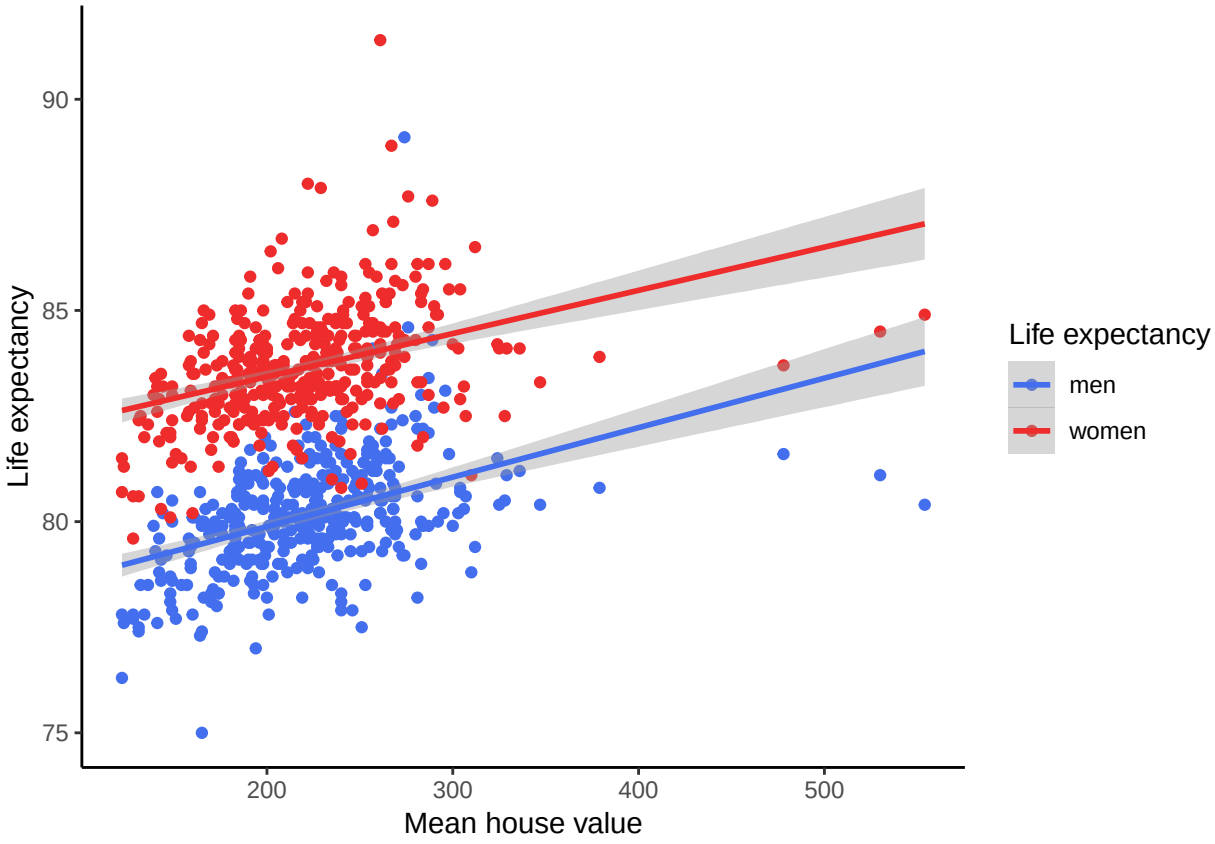
A map of the total population per municipality with Amsterdam and Rotterdam as obvious outliers.

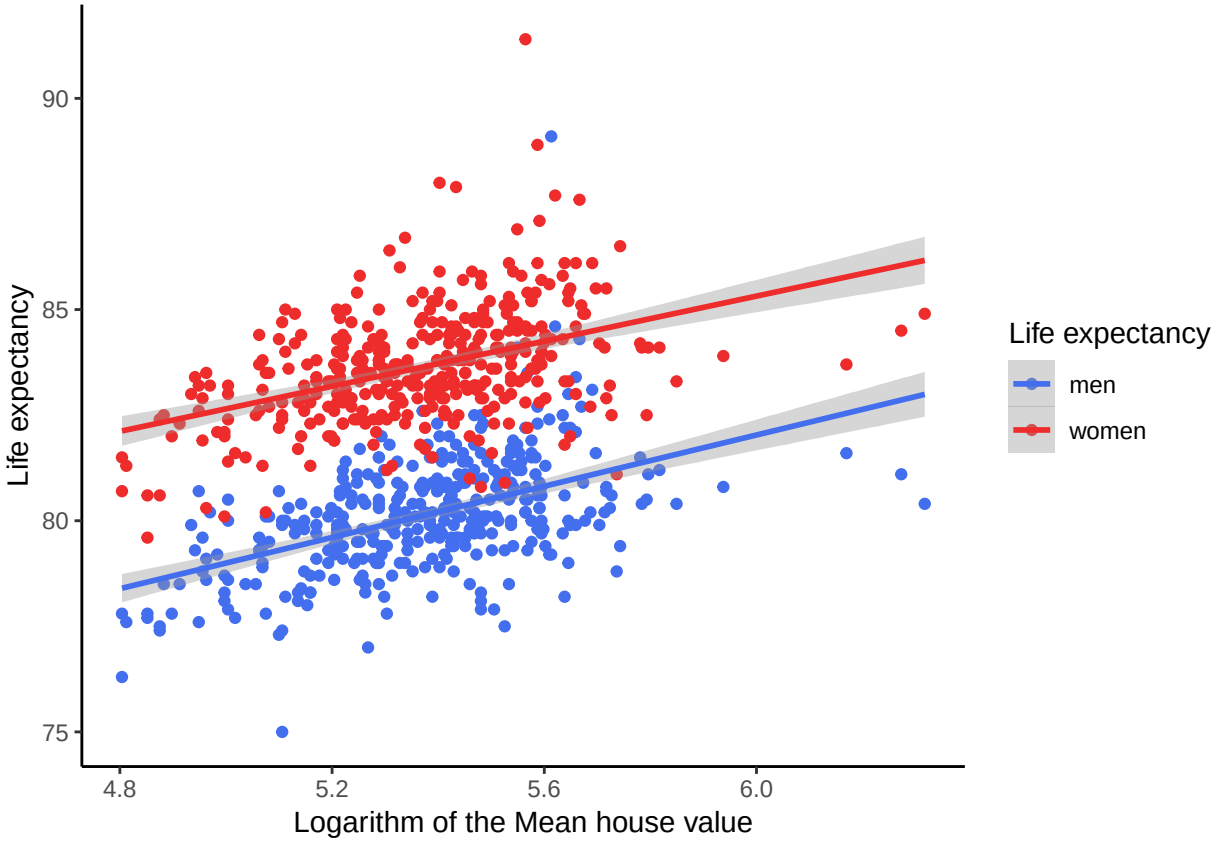


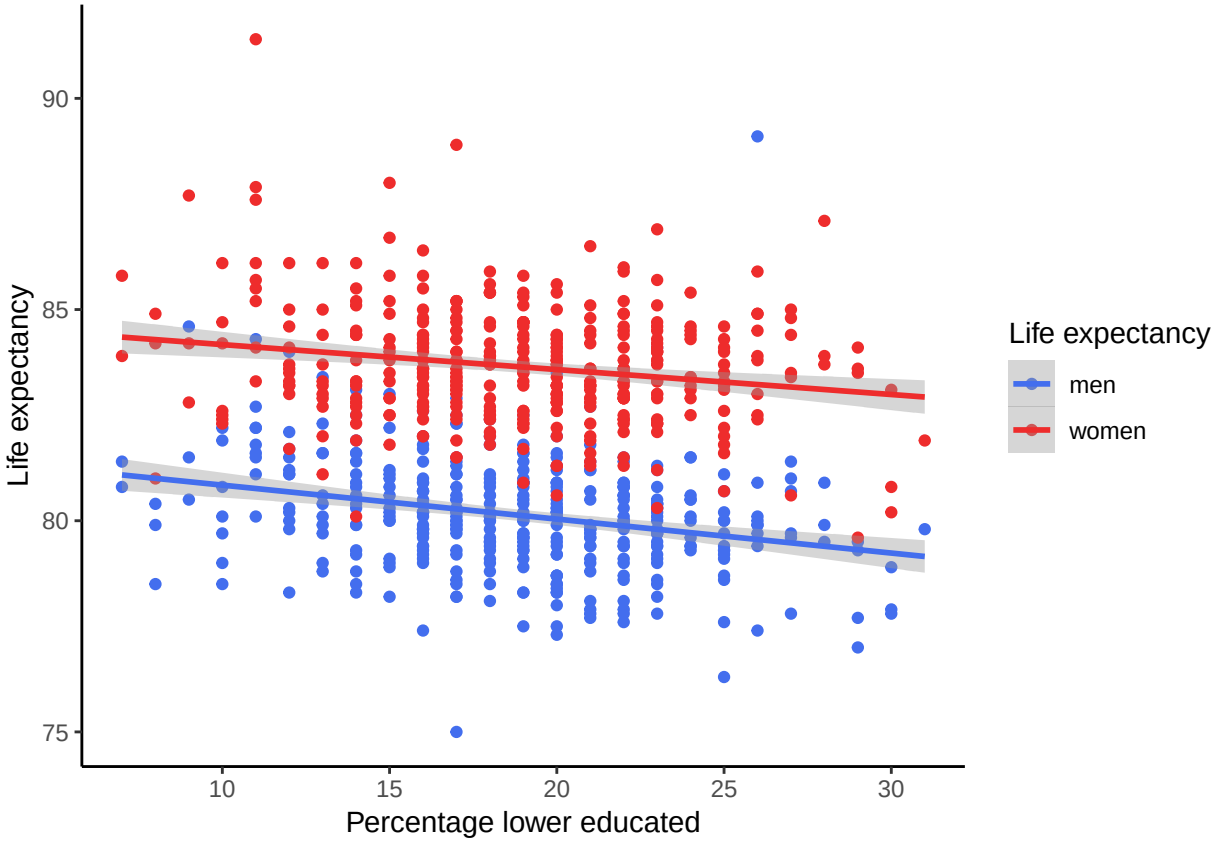
A map of the total percentage of people who owns a house. Notice Amsterdam and Rotterdam as outliers.

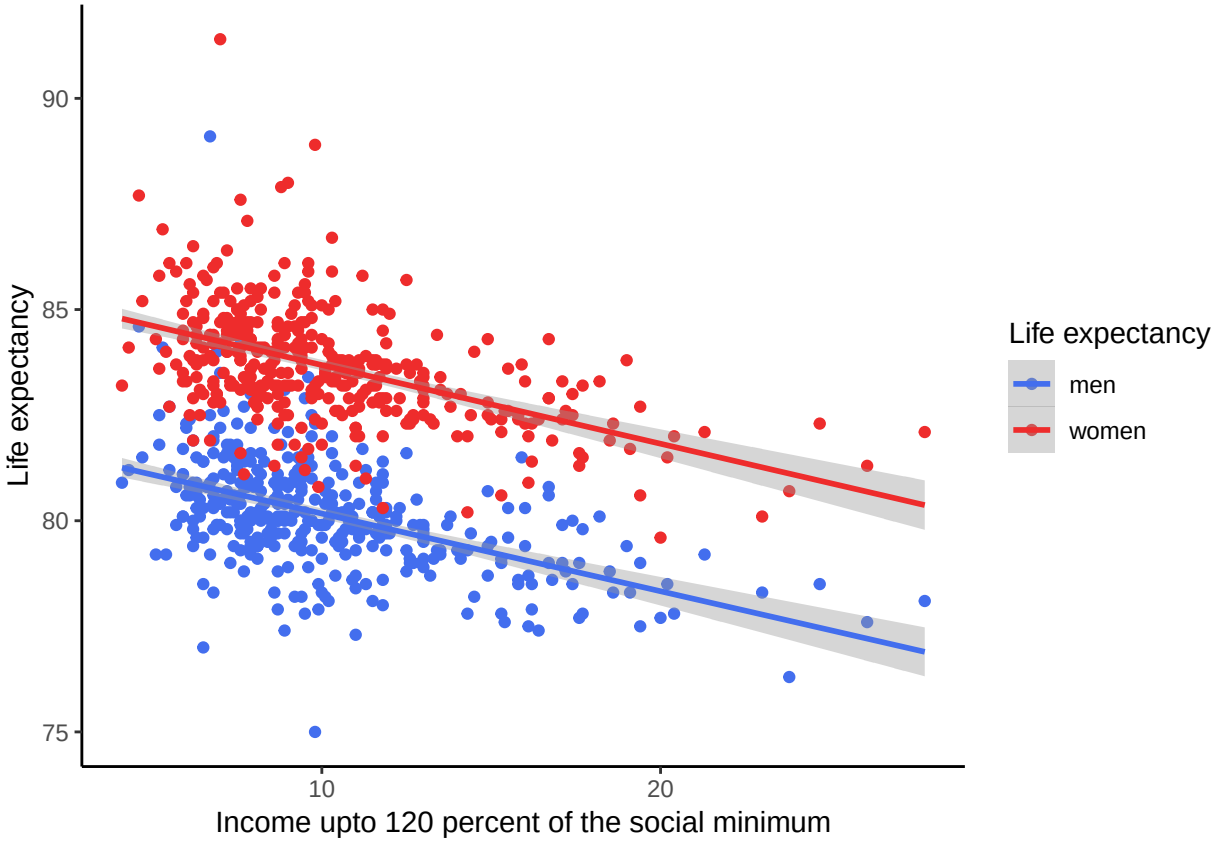
6.3 Regression graphs

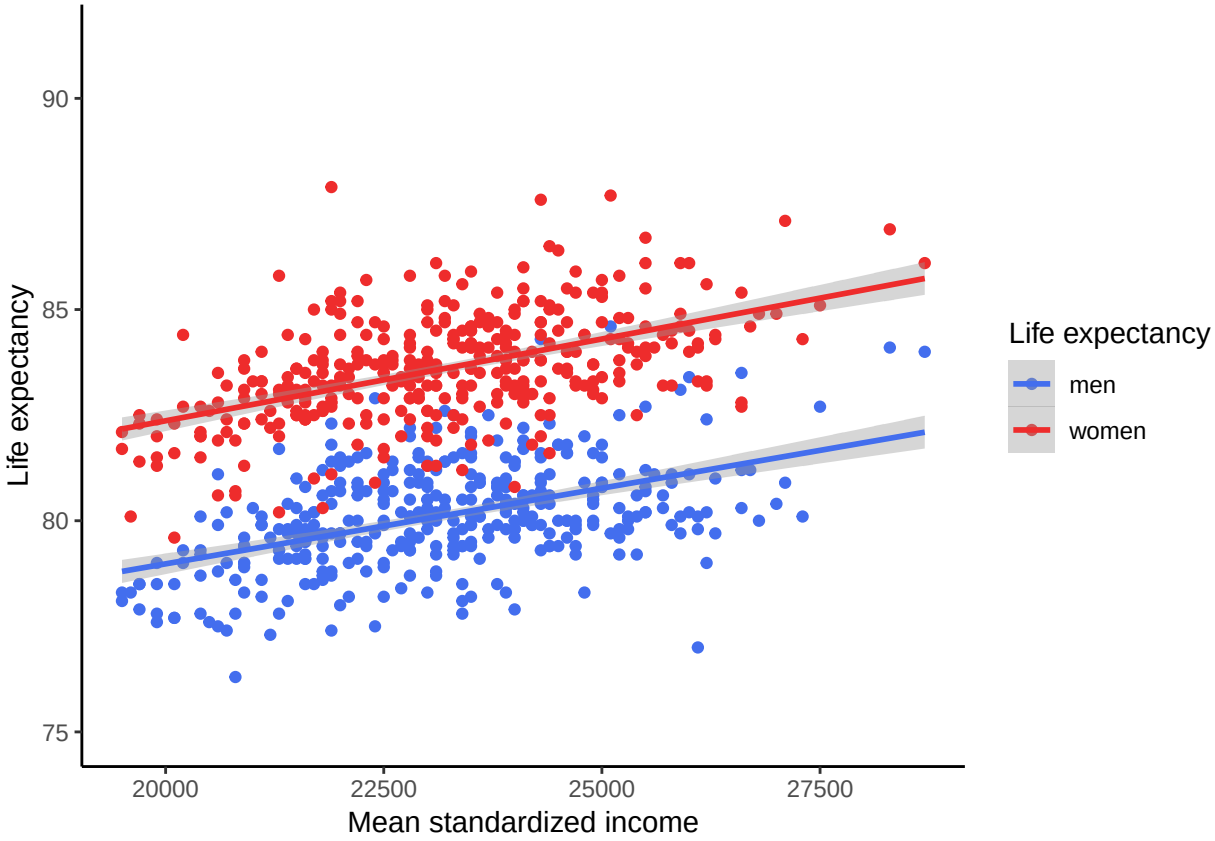
In this paragraph I will show a number of graphs with the relationship of the life expectancy of men and women and various predictors. I did not use the distinction in life expectancy between men and women in this research.

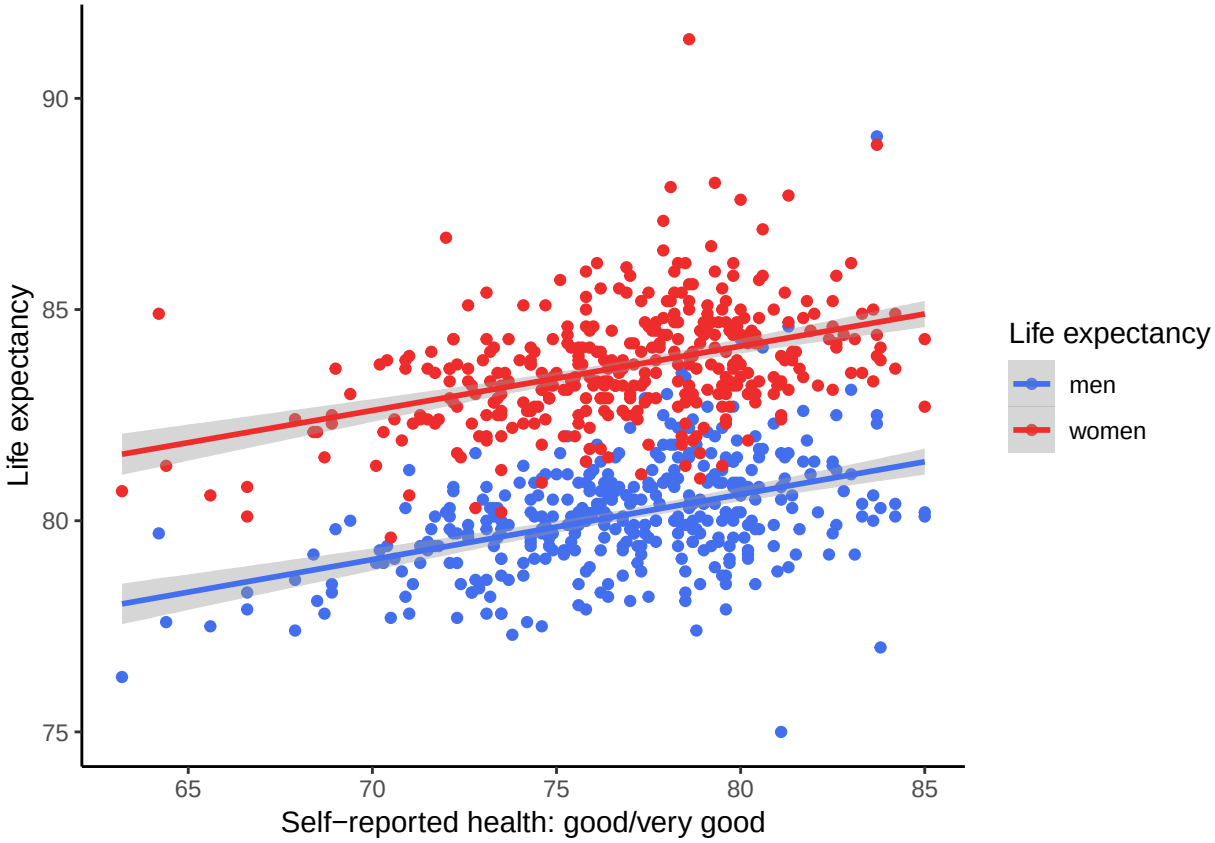


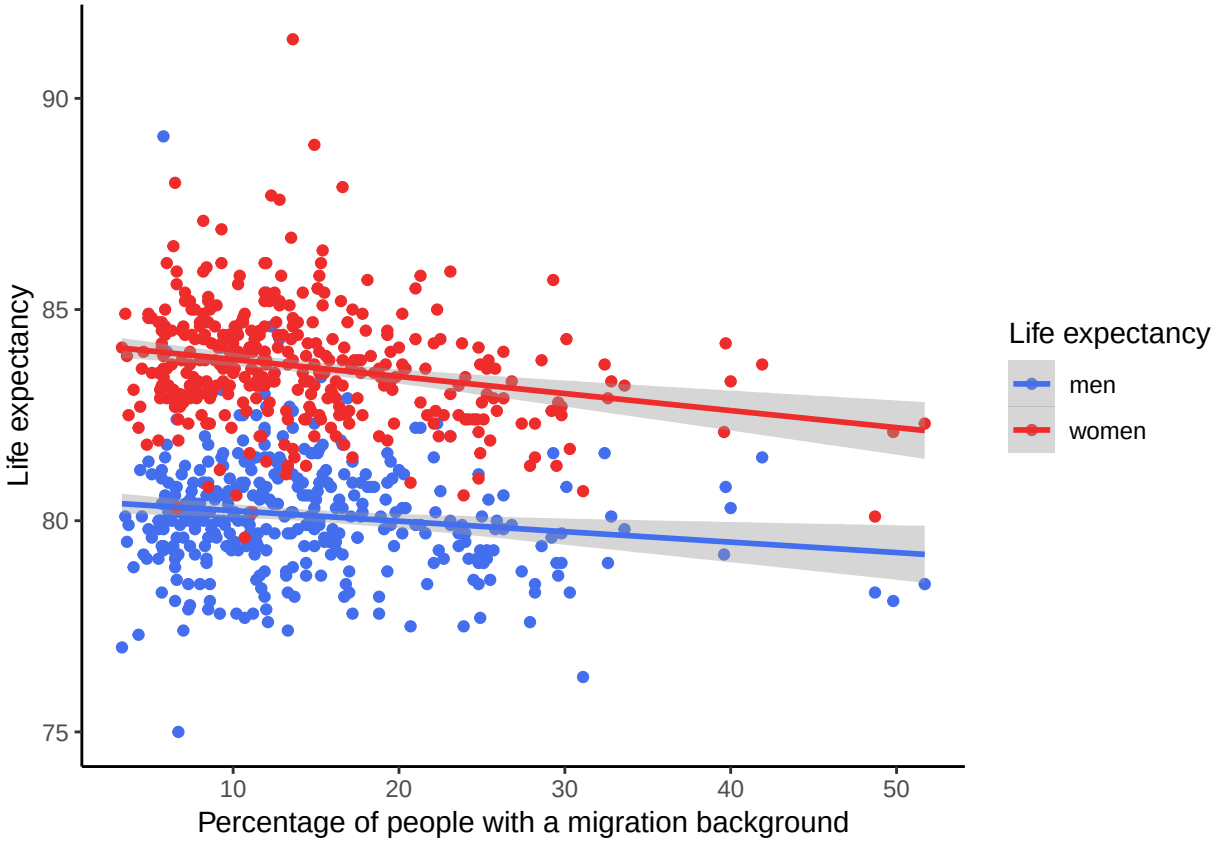

90
85
80
75
Life expectancy
4.8
5.2
5.6
6.0
Logarithm of the Mean house value
Life expectancy
men
women

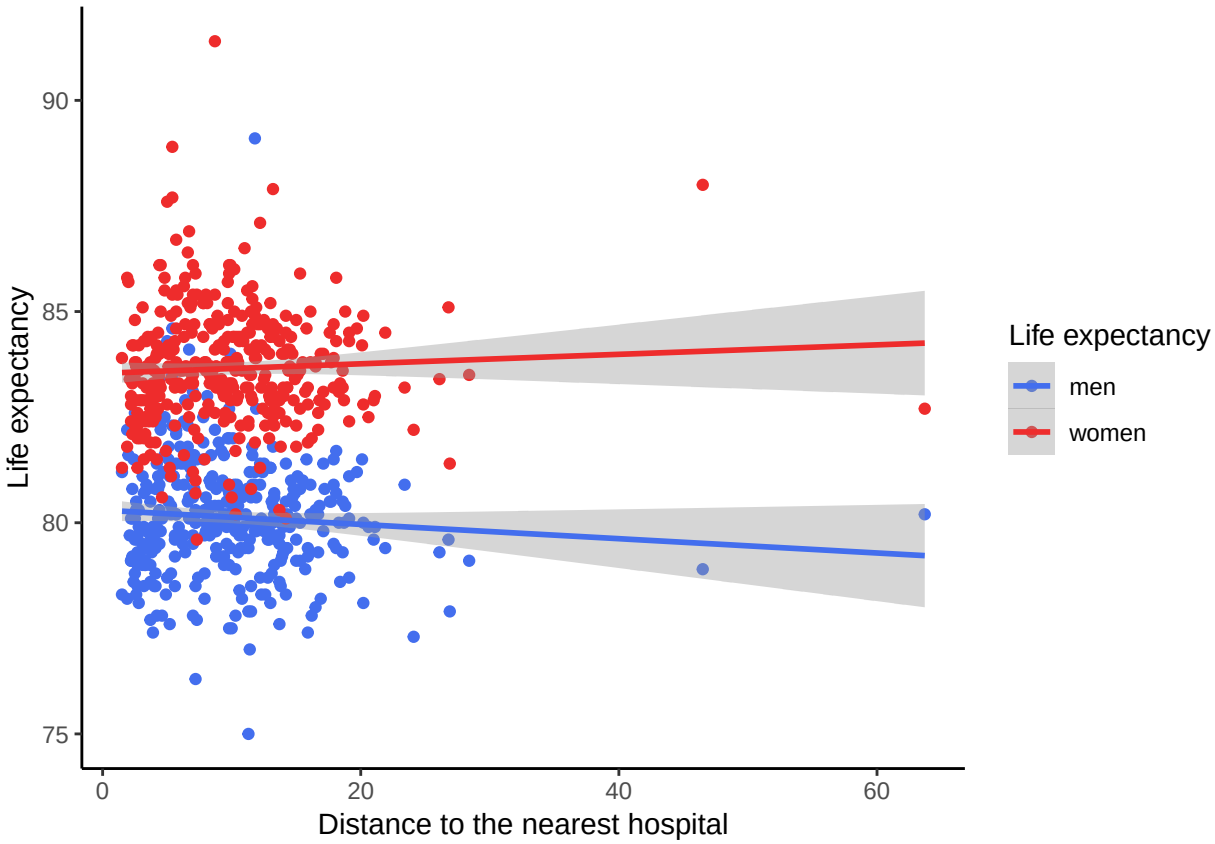

90
85
80
75
Life expectancy
10
15
20
25
30
Percentage lower educated
Life expectancy
men
women

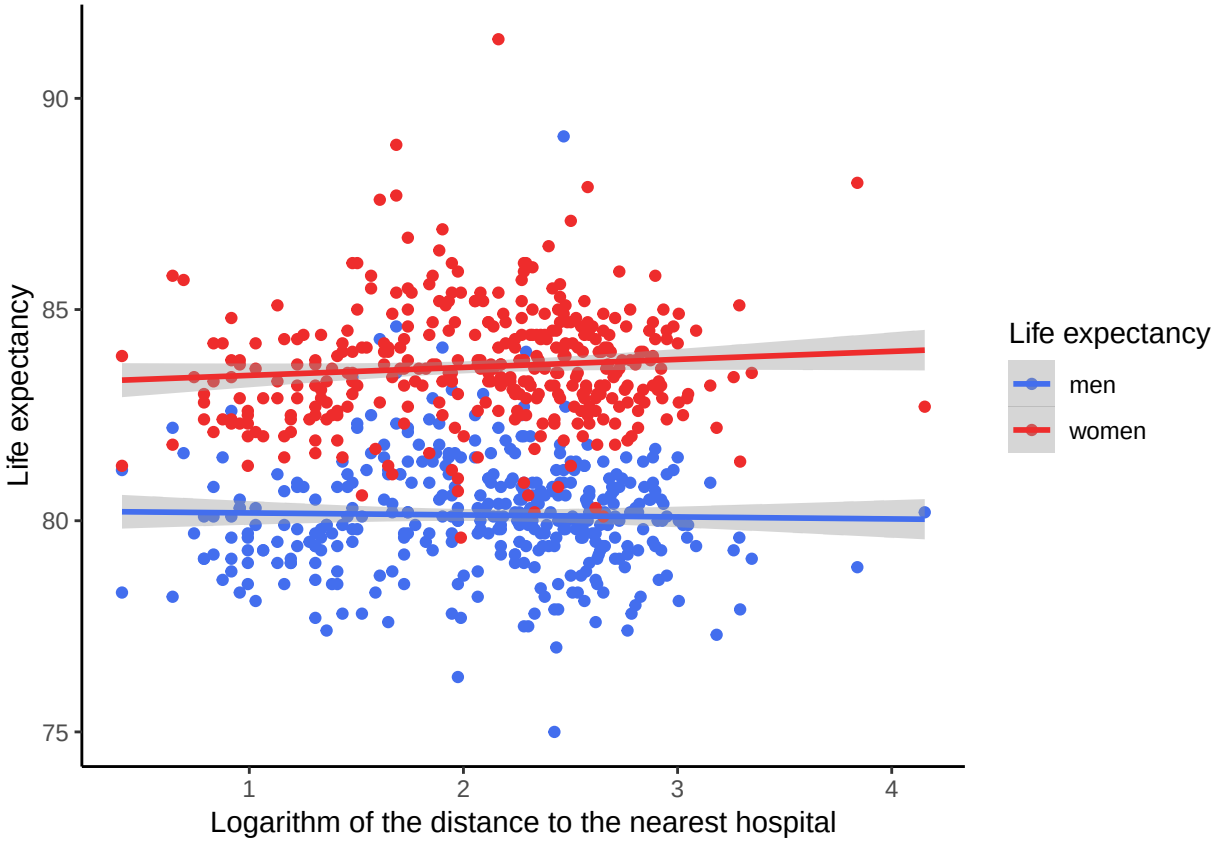

90
85
80
75
Life expectancy
10
20
Income upto 120 percent of the social minimum
Life expectancy
men
women


90
85
80
75
Life expectancy
20000
22500
25000
27500
Mean standardized income
Life expectancy
men
women


Life expectancy
90
85
80
75
65
70
75
80
85
Self−reported health: good/very good
Life expectancy
men
women


90
85
80
75
Life expectancy
10
20
30
40
50
Percentage of people with a migration background
Life expectancy
men
women


90
85
80
75
Life expectancy
0
20
40
60
Distance to the nearest hospital
Life expectancy
men
women


90
85
80
75
Life expectancy
1
2
3
4
Logarithm of the distance to the nearest hospital
Life expectancy
men
women

Appendix 2: Notebook datacleaning

6.4 Packages

We need the library Tidyverse for data manipulation.

```
library(tidyverse)
```

6.5 Data

6.5.1 RIVM

On July 19th, 2018 I have downloaded the data about life expectancy at birth from RIVM

We are interested in the life expactancy for the complete population.

```
dataLE1 <- read.csv2("../Datacleaning/sourcedata/RIVM/LE.csv") %>%
  filter(Geslacht == "Totaal")%>%
  select(Gemeente, Bij.geboorte)
```

```
  colnames(dataLE1)[colnames(dataLE1)=="Bij.geboorte"] <- "LEtotalpop"
```

```
dataLE2 <- read.csv2("../Datacleaning/sourcedata/RIVM/LE.csv") %>%
  filter(Geslacht == "Mannen")%>%
  select(Gemeente, Bij.geboorte)
```

```
  colnames(dataLE2)[colnames(dataLE2)=="Bij.geboorte"] <- "LEmen"
```

```
dataLE3 <- read.csv2("../Datacleaning/sourcedata/RIVM/LE.csv") %>%
  filter(Geslacht == "Vrouwen")%>%
  select(Gemeente, Bij.geboorte)
```

```
  colnames(dataLE3)[colnames(dataLE3)=="Bij.geboorte"] <- "LEwomen"
```

```
dataLE4 <- full_join(dataLE1, dataLE2, by = "Gemeente")
```

```
dataLE <- full_join(dataLE4, dataLE3, by = "Gemeente")
```

6.5.2 CBS

6.5.2.1 Kerncijfers

On July 19th, 2018 I have downloaded the data about "kerncijfers" plus metadata from CBS

```
kerncijfers <- read.csv2("../Datacleaning/sourcedata/CBS/kerncijfers.csv")
```

The data contain information about different aggregation levels. We are interested in the data of the municipalities. The id's for municipalities contain the string "GM". We will filter the municipalities. There are many municipalities without population, we will select municipalities with population.

```
kerncijfers2 <- kerncijfers %>%
  filter(grepl("GM", RegioS)) %>%
  filter(TotaleBevolking_1 != "NA")
```

The file `kerncijfers.csv` contains data per municipality about demographic characteristics. The data do not contain the names of the municipalities, but id's. To be able to merge these data with the RIVM data on life expecentancy, we will first produce a table with id's and names of municipalities from the metadata of the "kerncijfers". We need to rename TableInfos to RegioS and X to "Gemeente" to match the columns.

```
MD_kerncijfers <- read.csv2("../Datacleaning/sourcedata/CBS/metadata_kerncijfers.csv")
```

```
Merge_table <- MD_kerncijfers %>%
  select(RegioS = TableInfos, Gemeente = X) %>%
  filter(grepl("GM", RegioS))

Merge_table$Gemeente <- as.character(Merge_table$Gemeente)

str(Merge_table)
```

```
## 'data.frame':    713 obs. of  2 variables:
##  $ RegioS  : Factor w/ 1226 levels "0","1","10","100",..: 1136 937 744 624 807 938 708 518 745 808 .
##  $ Gemeente: chr  "Aa en Hunze" "Aalburg" "Aalsmeer" "Aalten" ...
```

Now we can merge the `kerncijfers` with the `Merge_table`. First we will change the variable `RegioS` in the dataframes to be merged in to a character.

```
Merge_table$RegioS <- as.character(Merge_table$RegioS)
kerncijfers2$RegioS <- as.character(kerncijfers2$RegioS)


kerncijfers3 <- right_join(Merge_table, kerncijfers2, by="RegioS")
```

And finally we can merge `dataLE` with `kerncijfers3` First we will change the variable `Gemeente` in to a character.

```
dataLE$Gemeente <- as.character(dataLE$Gemeente)
kerncijfers3$Gemeente <- as.character(kerncijfers3$Gemeente)
data1 <- inner_join(dataLE, kerncijfers3, by = "Gemeente")
```

6.5.2.2 Data selection

```
data1a <- data1 %>%

  select(LEtotalpop, LEwomen, LEmen, Gemeente, Gescheiden_32, TotaalMetMigratieachtergrond_44, Bevolkin



data1a$Gescheiden_32 <- as.numeric(as.character(data1a$Gescheiden_32))
data1a$TotaalMetMigratieachtergrond_44 <- as.numeric(as.character(data1a$TotaalMetMigratieachtergrond_4
data1a$GemiddeldeHuishoudensgrootte_89 <- as.numeric(as.character(data1a$GemiddeldeHuishoudensgrootte_8
data1a$Koopwoningen_94 <- as.numeric(as.character(data1a$Koopwoningen_94))
data1a$AfstandTotZiekenhuis_216 <- as.numeric(as.character(data1a$AfstandTotZiekenhuis_216))
data1a$k_80JaarOfOuder_21 <- as.numeric(as.character(data1a$k_80JaarOfOuder_21))
str(data1a)
```

```
## 'data.frame':    374 obs. of  15 variables:
##  $ LEtotalpop                     : num  82.5 80.9 83.2 82 81 82.8 82.7 81.7 79.7 81.8 ...
##  $ LEwomen                        : num  85 81.9 85.7 84.1 83.1 84.6 84.4 83.6 81.6 83.3 ...
##  $ LEmen                          : num  80 79.8 80.8 79.9 78.9 81.1 80.9 79.7 77.7 80.3 ...
##  $ Gemeente                       : chr  "Aa en Hunze" "Aalburg" "Aalsmeer" "Aalten" ...
##  $ Gescheiden_32                  : num  8.2 5.3 7.7 6.1 7.5 7.2 8.8 11 9.8 11.2 ...
##  $ TotaalMetMigratieachtergrond_44: num  5.9 5.5 18.1 11 4 13.6 19.3 21.6 24.9 40 ...
##  $ Bevolkingsdichtheid_57         : int  91 259 1555 279 274 2270 1150 974 1076 1533 ...
##  $ GemiddeldeHuishoudensgrootte_89: num  2.26 2.65 2.41 2.35 2.39 2.46 2.43 2.1 2.24 2.38 ...
##  $ Woningdichtheid_93             : int  40 100 640 122 117 935 470 456 482 617 ...
##  $ Koopwoningen_94                : num  71.1 71.7 63.1 69.4 63.3 60.1 67.3 57.7 54.6 63.8 ...
##  $ GemiddeldeWoningwaarde_99      : int  218 239 269 185 159 194 248 183 151 181 ...
##  $ AfstandTotZiekenhuis_216       : num  11.5 11.8 9.7 14.5 15.7 8.5 6.3 4 3.7 4.4 ...
##  $ TotaleBevolking_1              : int  25243 13038 31299 26912 28007 19955 24985 107615 72425 1981
##  $ k_80JaarOfOuder_21             : num  5.6 3.8 4.5 5.2 4 4.9 3.8 4.1 4.4 2 ...
##  $ k_80JaarOfOuder_12             : int  1408 490 1398 1397 1125 985 944 4457 3192 4009 ...
```

6.5.2.3 Gemeentefonds

We need to add some information to the data set about socio-economic factors (such as income, education and benefits).

On July 19th, 2018 I have downloaded the data about "Gemeentefonds" plus metadata from CBS

We need the variable `RegioS` to be able to merge. We will `filter` this variable for all observations containing ("GM")

For Income we will select the average income and the percentage of low income people. For education we will take the percentage of lower educated people We also want to select the percentage of one person households. This variable is not available, but we will construct this variable from the number of households and the number of one person households.

```
GF <- read.csv2("../Datacleaning/sourcedata/CBS/gemeentefonds.csv")

GF2 <- GF %>%
  filter(grepl("GM", RegioS)) %>%
  select(RegioS, GemiddeldGestandaardiseerdInkomen_41, InkomenTot120SociaalMinimum_13, LagerOpgeleidenP
  mutate(Percentage_eenpersoonshuishoudens = (Eenpersoonshuishoudens_44/Huishoudens_32)*100)


GF3 <- GF2 %>%
  select(-Huishoudens_32, -Eenpersoonshuishoudens_44)

GF3$GemiddeldGestandaardiseerdInkomen_41 <- as.numeric(as.character(GF3$GemiddeldGestandaardiseerdInkom
GF3$InkomenTot120SociaalMinimum_13 <- as.numeric(as.character(GF3$InkomenTot120SociaalMinimum_13))
GF3$LagerOpgeleidenPercentage_5 <- as.numeric(GF3$LagerOpgeleidenPercentage_5)

str(GF3)
```

```
## 'data.frame':    390 obs. of  5 variables:
##  $ RegioS                            : Factor w/ 391 levels "GM0003","GM0005",..: 330 234 128 65 2
##  $ GemiddeldGestandaardiseerdInkomen_41: num  23000 23700 25000 23500 20900 22500 24800 21500 20100 1
##  $ InkomenTot120SociaalMinimum_13      : num  8.2 6.2 6.6 7.6 13 8.7 6.7 11.9 17.6 16 ...
##  $ LagerOpgeleidenPercentage_5         : num  12 31 23 20 30 18 14 16 21 17 ...
##  $ Percentage_eenpersoonshuishoudens   : num  26.4 23.3 27.9 27.9 28.7 ...
```

Now we can merge `GF3` with `Merge_table`, we need to change `RegioS` in to a character. And GF4 with `data1`, we need to change `Gemeente` in to a character.

```
GF3$RegioS <- as.character(GF3$RegioS)


GF4 <-right_join(Merge_table, GF3, by="RegioS")

GF4$gemeente <- as.character(GF4$Gemeente)
data2 <- left_join(data1a, GF4, by= "Gemeente")
```

6.5.2.4 Health monitor

On July 19th, 2018 I have downloaded the data about the health plus metadata from CBS

```
HM <- read.csv2("../Datacleaning/sourcedata/CBS/health_monitor.csv")
summary(HM$Perioden)
```

```
## 2016JJ00
##     3888
```

I add the names of the municipalities to our table instead of the ID's the CBS uses.

```
HM$RegioS <- as.character(HM$RegioS)


HM2 <- full_join(Merge_table,HM , by="RegioS") %>%
 filter(grepl("GM", RegioS)) %>%
  filter(Leeftijd == 10000, Marges== "MW00000")
```

I would like to add the following data to our data set.

```
HM3 <- HM2 %>%
  select(Gemeente, ErvarenGezondheidGoedZeerGoed_1, EenOfMeerLangdurigeAandoeningen_2, NormaalGewicht_9

str(HM3)
```

```
## 'data.frame':    390 obs. of  7 variables:
##  $ Gemeente                          : chr  "Aa en Hunze" "Aalburg" "Aalsmeer" "Aalten" ...
##  $ ErvarenGezondheidGoedZeerGoed_1   : Factor w/ 421 levels "       .","    43.4",..: 300 292 310 261
##  $ EenOfMeerLangdurigeAandoeningen_2: Factor w/ 446 levels "       .","    11.0",..: 172 86 150 187 1
##  $ NormaalGewicht_9                  : Factor w/ 403 levels "       .","    19.4",..: 186 223 239 207
##  $ VoldoetAanFitnorm_14              : Factor w/ 514 levels "       .","     7.4",..: 177 103 160 208
##  $ UrenMantelzorgPerWeek_19          : Factor w/ 219 levels "       .","     2.6",..: 98 67 50 54 1 1
##  $ WekelijkseSporters_16             : Factor w/ 519 levels "       .","    14.0",..: 261 181 341 277
```

I will make the variables numeric.

```
HM3$ErvarenGezondheidGoedZeerGoed_1 <- as.numeric(as.character(HM3$ErvarenGezondheidGoedZeerGoed_1))
HM3$EenOfMeerLangdurigeAandoeningen_2 <- as.numeric(as.character(HM3$EenOfMeerLangdurigeAandoeningen_2))
HM3$NormaalGewicht_9 <- as.numeric(as.character(HM3$NormaalGewicht_9))
HM3$VoldoetAanFitnorm_14 <- as.numeric(as.character(HM3$VoldoetAanFitnorm_14))
HM3$UrenMantelzorgPerWeek_19 <- as.numeric(as.character(HM3$UrenMantelzorgPerWeek_19))
HM3$WekelijkseSporters_16 <- as.numeric(as.character(HM3$WekelijkseSporters_16))
```

Now I can merge the Health Monitor data with the rest of our data.

```
data3 <- left_join(data2, HM3, by= "Gemeente")

str(data3)
```

```
## 'data.frame':    374 obs. of  27 variables:
##  $ LEtotalpop                        : num  82.5 80.9 83.2 82 81 82.8 82.7 81.7 79.7 81.8 ...
##  $ LEwomen                           : num  85 81.9 85.7 84.1 83.1 84.6 84.4 83.6 81.6 83.3 ...
##  $ LEmen                             : num  80 79.8 80.8 79.9 78.9 81.1 80.9 79.7 77.7 80.3 ...
##  $ Gemeente                          : chr  "Aa en Hunze" "Aalburg" "Aalsmeer" "Aalten" ...
##  $ Gescheiden_32                     : num  8.2 5.3 7.7 6.1 7.5 7.2 8.8 11 9.8 11.2 ...
##  $ TotaalMetMigratieachtergrond_44   : num  5.9 5.5 18.1 11 4 13.6 19.3 21.6 24.9 40 ...
##  $ Bevolkingsdichtheid_57            : int  91 259 1555 279 274 2270 1150 974 1076 1533 ...
##  $ GemiddeldeHuishoudensgrootte_89   : num  2.26 2.65 2.41 2.35 2.39 2.46 2.43 2.1 2.24 2.38 ...
```

```
##  $ Woningdichtheid_93                 : int  40 100 640 122 117 935 470 456 482 617 ...
##  $ Koopwoningen_94                    : num  71.1 71.7 63.1 69.4 63.3 60.1 67.3 57.7 54.6 63.8 ...
##  $ GemiddeldeWoningwaarde_99          : int  218 239 269 185 159 194 248 183 151 181 ...
##  $ AfstandTotZiekenhuis_216           : num  11.5 11.8 9.7 14.5 15.7 8.5 6.3 4 3.7 4.4 ...
##  $ TotaleBevolking_1                  : int  25243 13038 31299 26912 28007 19955 24985 107615 72425
##  $ k_80JaarOfOuder_21                 : num  5.6 3.8 4.5 5.2 4 4.9 3.8 4.1 4.4 2 ...
##  $ k_80JaarOfOuder_12                 : int  1408 490 1398 1397 1125 985 944 4457 3192 4009 ...
##  $ RegioS                             : chr  "GM1680" "GM0738" "GM0358" "GM0197" ...
##  $ GemiddeldGestandaardiseerdInkomen_41: num  23000 23700 25000 23500 20900 22500 24800 21500 20100 1
##  $ InkomenTot120SociaalMinimum_13     : num  8.2 6.2 6.6 7.6 13 8.7 6.7 11.9 17.6 16 ...
##  $ LagerOpgeleidenPercentage_5        : num  12 31 23 20 30 18 14 16 21 17 ...
##  $ Percentage_eenpersoonshuishoudens  : num  26.4 23.3 27.9 27.9 28.7 ...
##  $ gemeente                           : chr  "Aa en Hunze" "Aalburg" "Aalsmeer" "Aalten" ...
##  $ ErvarenGezondheidGoedZeerGoed_1    : num  79.5 78.7 80.5 75.6 81.3 75.3 77.9 72.6 72.3 73.2 ...
##  $ EenOfMeerLangdurigeAandoeningen_2  : num  33.5 24.9 31.3 35 35.9 33.4 32.1 36.2 38.2 34.1 ...
##  $ NormaalGewicht_9                   : num  43.7 47.4 49 45.8 40.8 47.9 47.8 52.7 42.9 44 ...
##  $ VoldoetAanFitnorm_14               : num  28.6 21.2 26.9 31.7 26.1 19.9 22.3 28.2 32 18.7 ...
##  $ UrenMantelzorgPerWeek_19           : num  12.8 9.7 8 8.4 NA NA NA 8.7 10.2 16.4 ...
##  $ WekelijkseSporters_16              : num  46.3 38.3 54.3 47.9 41.7 42.3 50.1 53.3 47.1 49.7 ...
```

6.5.3 Rijkswaterstaat

On September 27th, 2018 I have downloaded the data about "CO2 emissions" from Rijkswaterstaat

```
RWS <- read.csv2("../Datacleaning/sourcedata/Rijkswaterstaat/CO2.csv", sep =",")
str(RWS)
```

```
## 'data.frame':    381 obs. of  2 variables:
##  $ Gemeenten                       : Factor w/ 381 levels "'s-Hertogenbosch",..: 2 3 4 5 6 7 8 9 10 1
##  $ Totaal.bekende.CO2.uitstoot.2016: int  NA NA NA 134297 159597 201343 117913 579053 571012 831527
```

Since in our data set the column containing the Municipalities is called "Gemeente" and in the data set from "Rijkswaterstaat" is called "Gemeenten" we have to change the name before we can merge it with our data set.

```
colnames(RWS)[colnames(RWS)=="Gemeenten"] <- "Gemeente"

rijkswaterstaat1 <- RWS
str(RWS)
```

```
## 'data.frame':    381 obs. of  2 variables:
##  $ Gemeente                        : Factor w/ 381 levels "'s-Hertogenbosch",..: 2 3 4 5 6 7 8 9 10 1
##  $ Totaal.bekende.CO2.uitstoot.2016: int  NA NA NA 134297 159597 201343 117913 579053 571012 831527
```

Now we can merge the data from "Rijkswaterstaat" with our previous made data set so that we have one final data set to work with.

```
data4 <- left_join(data3, rijkswaterstaat1, by= "Gemeente")
```

Finally we will save data4 as a csv file.

```
write.csv2(data4, "../Datacleaning/Sourcedata/Analysis/Datafile.csv")
```

7 Appendix 3: Regression Health Status

Table 8: Results of linear regression of health status on other variables

	Dependent variable:
	Health_status_very_good
Divorced	-1.089^{***} (0.129)
House_O	-0.047^{*} (0.028)
Mean_inc	-0.0003^{*} (0.0001)
Lower_edu	-0.263^{***} (0.041)
Pop_density	0.0004^{*} (0.0002)
Mean_HV	0.015^{***} (0.004)
Dist_Hosp	0.093^{**} (0.036)
Multi_morbidity	-0.278^{***} (0.050)
Normal_weight	0.024 (0.044)
Fit_norm	0.253^{***} (0.047)
Informal_care	-0.073 (0.056)
Weekly_sporters	-0.111^{***} (0.038)
Total_population	-0.00001^{**} (0.00000)
Percentage_over80	-0.591^{***} (0.165)
Constant	105.565^{***} (5.246)
Observations	313
R^2	0.631
Adjusted R^2	0.614
Residual Std. Error	2.289 (df = 298)
F Statistic	36.464^{***} (df = 14; 298)
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix 4: Logboek

Table 9: Logboek profielwerkstuk Arvid Mikkers

Datum	**Activiteit**	**Duur in uren**
14-05-2018	Overleg met dr. Gertjan Verhoeven	1
	Reistijd	2
Juni 2018	Cursussen R op Datacamp	
	Introduction to R	12
	Introduction to Tidyverse	
19-07-2018	Overleg met dr. Gerjan Verhoeven	2
	Data zoeken en downloaden	8
	Reistijd Ilpendam - Utrecht	2
20-07-2018	Datacleaning	7
09-08-2018	Datacleaning	9
01-10-2018	Data analyse en visualisatie	8
	Reistijd	2
02-10-2018	Data analyse en visualisatie	8
	Reistijd	2
03-10-2018	Data analyse en visualisatie	4
07-10-2018	Data analyse en visualisatie	4
14-10-2018	Data analyse en visualisatie	6
17-10-2018	Schrijven	2
20-10-2018	Schrijven	4
21-10-2018	Schrijven	5
22-10-2018	Schrijven	5
23-10-2018	Schrijven	9
24-10-2018	Schrijven	4
25-10-2018	Schrijven	8
26-10-2018	Tekst eerste concept controleren	3
26-10-2018	Tekst eerste concept controleren	5
26-10-2018	Tekst controleren en laatste check	2
Totaal		124

References

Dwyer-Lindgren, L., Bertozzi-Villa, A., Stubbs, R.W., Morozoff, C., Mackenbach, J.P., Lenthe, F.J. van, Mokdad, A.H. & Murray, C.J. (2017) Inequalities in life expectancy among us counties, 1980 to 2014: Temporal trends and key drivers. *JAMA internal medicine*, **177**, 1003–1011.

Gladwell, M. (2008) *Outliers: The story of success*, Hachette UK.

Ho, J.Y. & Hendi, A.S. (2018) Recent trends in life expectancy across high income countries: Retrospective observational study. *bmj*, **362**, k2562.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An introduction to statistical learning*, Springer.

Mulder, Y., Perenboom, R., Herten, L. van, Oudshoorn, K. & Hoeymans, N. (2002) Regionale verschillen in gezonde levensverwachting.

Poppel, F. van (1988) Regionale verschillen in levensverwachting in nederland in de jaren 1972-1984. *Nederlands tijdschrift voor Geneeskunde*, **132**, 571–5.

Woods, L.M., Rachet, B., Riga, M., Stone, N., Shah, A. & Coleman, M.P. (2005) Geographical variation in life expectancy at birth in england and wales is largely explained by deprivation. *Journal of Epidemiology & Community Health*, **59**, 115–120.