First analysis

Arvid Mikkers

Packages

```
library(tidyverse) # for data manipulation and plots
library(PerformanceAnalytics) # for correlation plots
library(thematicmaps)# for maps
library(corrr) # for network graphs of variables
library(ggrepel) # for names in graphs
library(glmnet) # for lasso
library(car) # to check for multicollinearity
library(caret) # for machine learning
```

Data

```
d1 <- read.csv2("../Sourcedata/Analysis/Datafile.csv")
str(d1)
```

```
## 'data.frame':    374 obs. of  29 variables:
##  $ X                                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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                           : Factor w/ 374 levels "'s-Hertogenbosch",..: 2 3 4 5 6 7 8 9
##  $ 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                             : Factor w/ 374 levels "GM0003","GM0005",..: 315 224 122 61 2
##  $ GemiddeldGestandaardiseerdInkomen_41: int  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        : int  12 31 23 20 30 18 14 16 21 17 ...
##  $ Percentage_eenpersoonshuishoudens  : num  26.4 23.3 27.9 27.9 28.7 ...
##  $ gemeente                           : Factor w/ 374 levels "'s-Hertogenbosch",..: 2 3 4 5 6 7 8 9
##  $ 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 ...
```

```
##  $ Totaal.bekende.CO2.uitstoot.2016    : int  NA NA NA 134297 159597 201343 117913 579053 571012 8315
```

Map

Since we are interested in the variation in life expectancy, we will first make a map.

To create maps we need polygon-data (data describing the polygons that constitute the map expressed in x and y coordinates). It is possible to buy this data from commercial parties, but more and more polygon data are published as open source data. For this thesis we used open source data from the CBS (central bureau of statitics in the Netherlands) and Imergis. Mark Klik (NZa) converted the polygon data to *csv* format.

For our first basic map, we need some data about the Netherlands. The polygon data from municipalities were previously saved as a *csv* in the folder "Sourcedata".

```
map_municipal1 <- read.csv2("../Sourcedata/Map/nld_municipal_map.csv", stringsAsFactors = FALSE, dec = '
head(map_municipal1)
```

```
##         name id        x        y order  hole piece group
## 1 Appingedam  0 251260.5 594393.8     1 FALSE     1   0.1
## 2 Appingedam  0 251427.8 594486.7     2 FALSE     1   0.1
## 3 Appingedam  0 251668.8 594646.8     3 FALSE     1   0.1
## 4 Appingedam  0 251713.5 594770.9     4 FALSE     1   0.1
## 5 Appingedam  0 251354.6 595461.4     5 FALSE     1   0.1
## 6 Appingedam  0 251310.3 596022.5     6 FALSE     1   0.1
```

Some data problems solved:

De Fryske Marren/De Friese Meren

In the Imergis dataset with polygons, the municipality "De Friese Meren" is called "De Fryske Marren" in the RIVM/CBS data. We will change the Imergis dataset.

Mergers

In the CBS/RIVM data the municipality "Berg en Dal" exists, this a merger between the municipalities Groesbeek, Millingen aan de Rijn and Ubbergen.

In the CBS/RIVM data the municipality "Gooise Meren" exists, this a merger between the municipalities Bussum, Naarden and Muiden.

In the CBS/RIVM data the municipality "Krimpenerwaard" exists, this a merger between the municipalities Nederlek, Ouderkerk, Vlist, Bergambacht and Schoonhoven.

In the CBS/RIVM data the municipality "Nissewaard" exists, this a merger between the municipalities Spijkenisse and Bernisse.

We will change the names of the merged municipalities in their new name.

Spelling

One municipality is spelled differently:

CBS/RIVM Nuenen, Gerwen en Nederwetten Imergis: Nuenen. Gerwen en Nederwetten

We change the name to the CBS/RIVM name

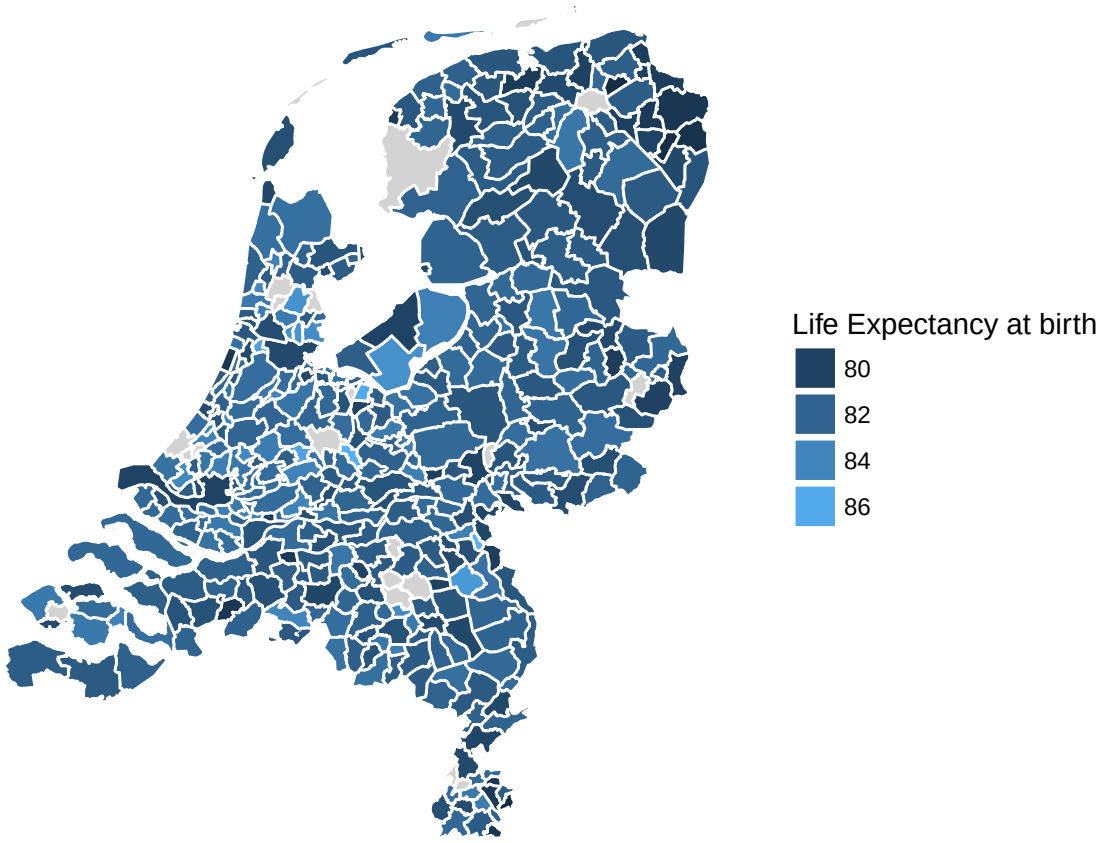
```
map_municipal <- map_municipal1 %>%
    mutate(name = replace(name, name == "De Friese Meren", "De Fryske Marren")) %>%
    mutate(name = replace(name, name == "Groesbeek"| name == "Millingen aan de Rijn" | name == "Ubberge
  mutate(name = replace(name, name == "Bussum"| name == "Naarden" | name == "Muiden", "Gooise Meren")) %
  mutate(name = replace(name, name == "Nederlek"| name == "Ouderkerk" | name == "Vlist" | name== "Bergam
  mutate(name = replace(name, name == "Spijkenisse"| name == "Bernisse", "Nissewaard")) %>%
  mutate(name = replace(name, name == "Nuenen. Gerwen en Nederwetten", "Nuenen, Gerwen en Nederwetten"))
```

Maps

```
map_info <- d1 %>%
    select(Gemeente, LEtotalpop)
```

We will use package `thematicmaps` to make a map of the Netherlands. The package `thematicmaps` is an extension to the `ggplot2` package.

```
AddMapLayer(MapPlot(), map_municipal, map_info) +
  guides(fill = guide_legend(title = "Life Expectancy at birth"))
```



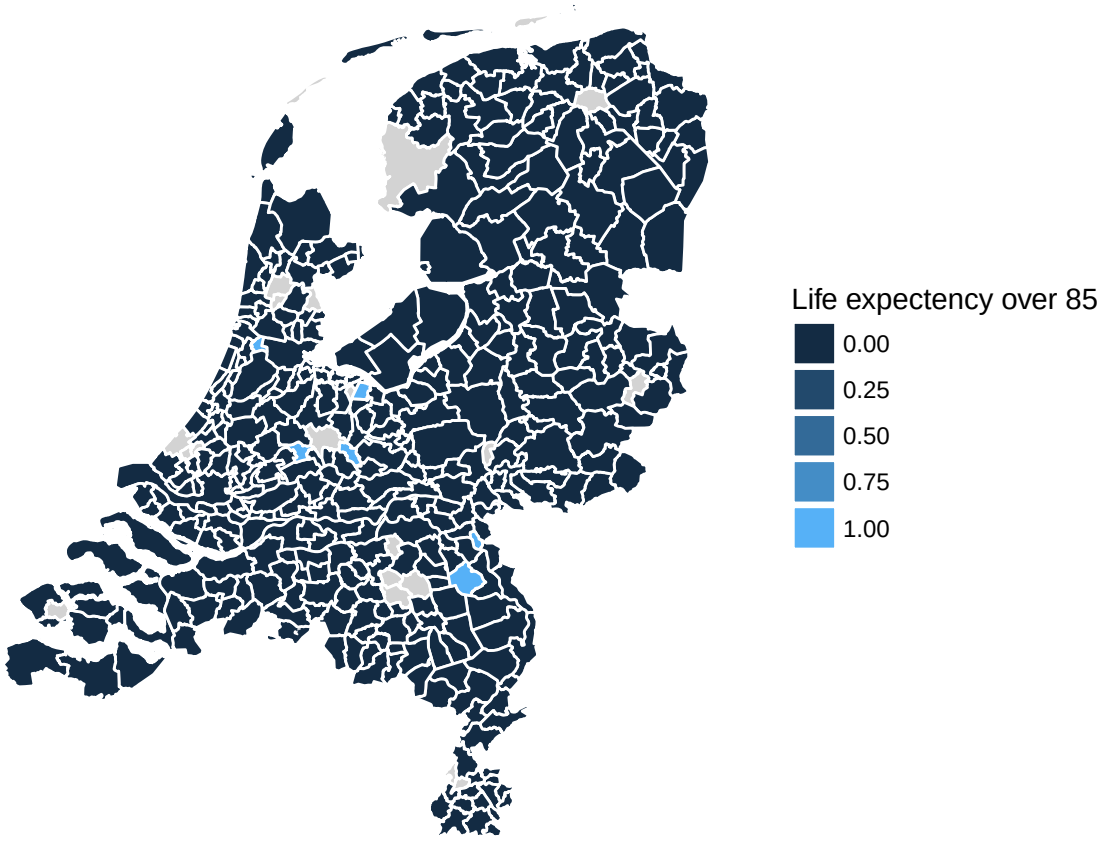

```
map_info4 <- d1 %>%
mutate( Hulp= ifelse(LEtotalpop>85,1,0))
str(map_info4)
```

```
## 'data.frame':    374 obs. of  30 variables:
##  $ X                                  : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
##  $ 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                            : Factor w/ 374 levels "'s-Hertogenbosch",..: 2 3 4 5 6 7 8 9
##  $ 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                              : Factor w/ 374 levels "GM0003","GM0005",..: 315 224 122 61 2
##  $ GemiddeldGestandaardiseerdInkomen_41: int  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         : int  12 31 23 20 30 18 14 16 21 17 ...
##  $ Percentage_eenpersoonshuishoudens   : num  26.4 23.3 27.9 27.9 28.7 ...
##  $ gemeente                            : Factor w/ 374 levels "'s-Hertogenbosch",..: 2 3 4 5 6 7 8 9
##  $ 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 ...
##  $ Totaal.bekende.CO2.uitstoot.2016    : int  NA NA NA 134297 159597 201343 117913 579053 571012 831
##  $ Hulp                                : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
map_info5 <- map_info4 %>%
  select(Gemeente, Hulp)
```

```
 AddMapLayer(MapPlot(), map_municipal, map_info5) +
  guides(fill = guide_legend(title = "Life expectency over 85"))
```



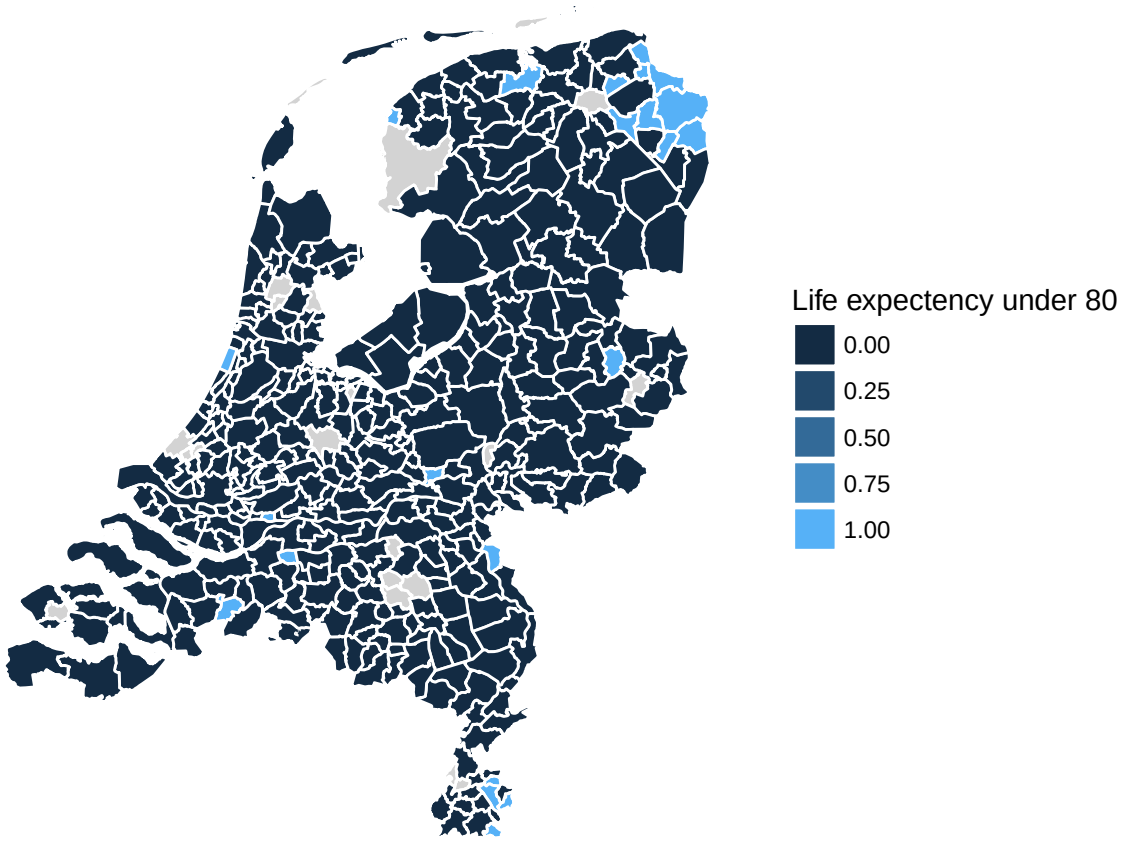

```
map_info6 <- d1 %>%
mutate( Hulp2= ifelse(LEtotalpop<80,1,0))
str(map_info4)
```

```
## 'data.frame':    374 obs. of  30 variables:
##  $ X                                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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                           : Factor w/ 374 levels "'s-Hertogenbosch",..: 2 3 4 5 6 7 8 9
##  $ 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                             : Factor w/ 374 levels "GM0003","GM0005",..: 315 224 122 61 2
##  $ GemiddeldGestandaardiseerdInkomen_41: int  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        : int  12 31 23 20 30 18 14 16 21 17 ...
##  $ Percentage_eenpersoonshuishoudens  : num  26.4 23.3 27.9 27.9 28.7 ...
```

```
##  $ gemeente                            : Factor w/ 374 levels "'s-Hertogenbosch",..: 2 3 4 5 6 7 8 9
##  $ 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 ...
##  $ Totaal.bekende.CO2.uitstoot.2016    : int  NA NA NA 134297 159597 201343 117913 579053 571012 831
##  $ Hulp                                : num  0 0 0 0 0 0 0 0 0 0 ...
```

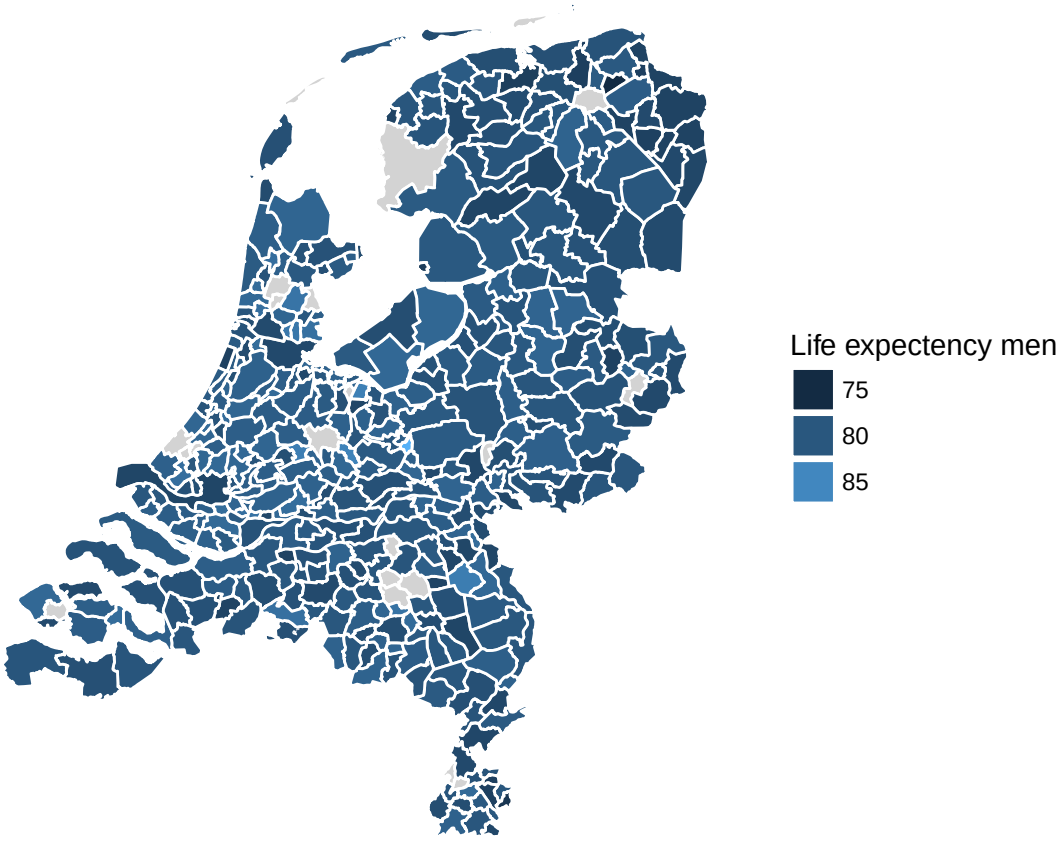
```
map_info7 <- map_info6 %>%
  select(Gemeente, Hulp2)
```

```
AddMapLayer(MapPlot(), map_municipal, map_info7) +
  guides(fill = guide_legend(title = "Life expectency under 80"))
```



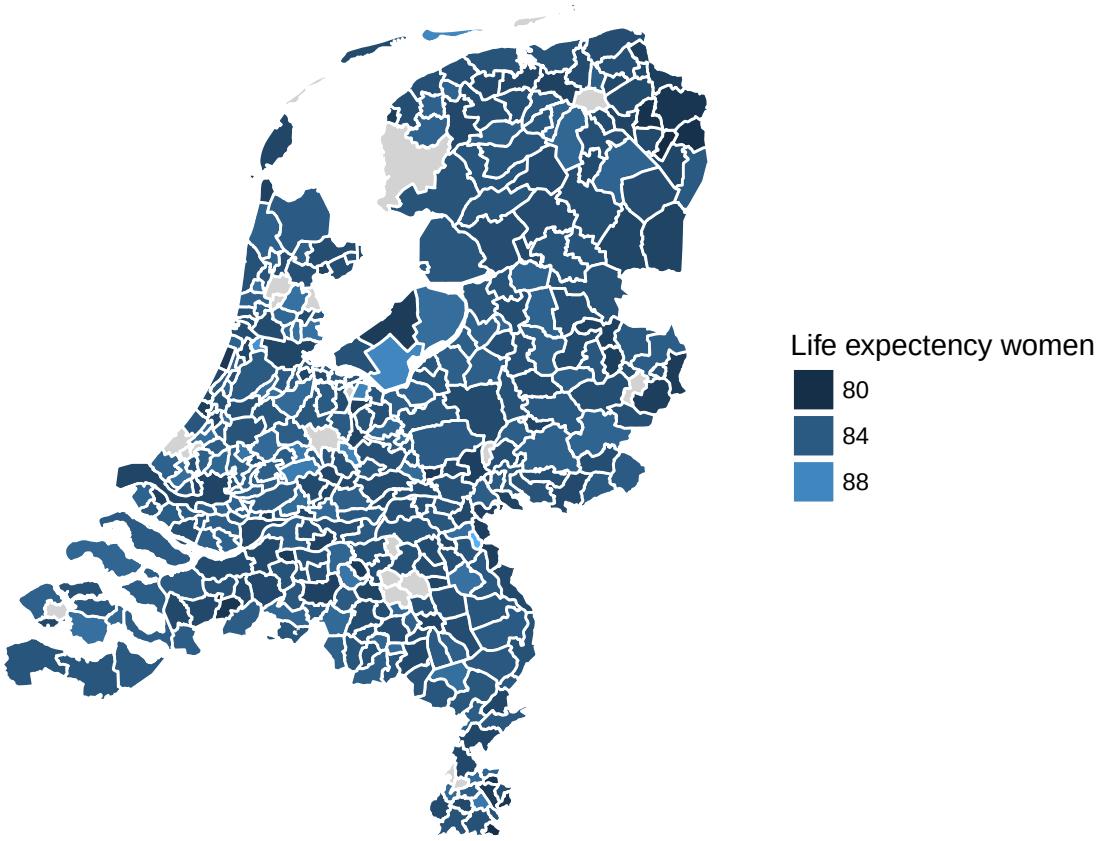

```
map_info13 <- d1 %>%
  select(Gemeente, LEmen)
```

```
AddMapLayer(MapPlot(), map_municipal, map_info13) +
  guides(fill = guide_legend(title = "Life expectency men"))
```




```
map_info14 <- d1 %>%
  select(Gemeente, LEwomen)
```

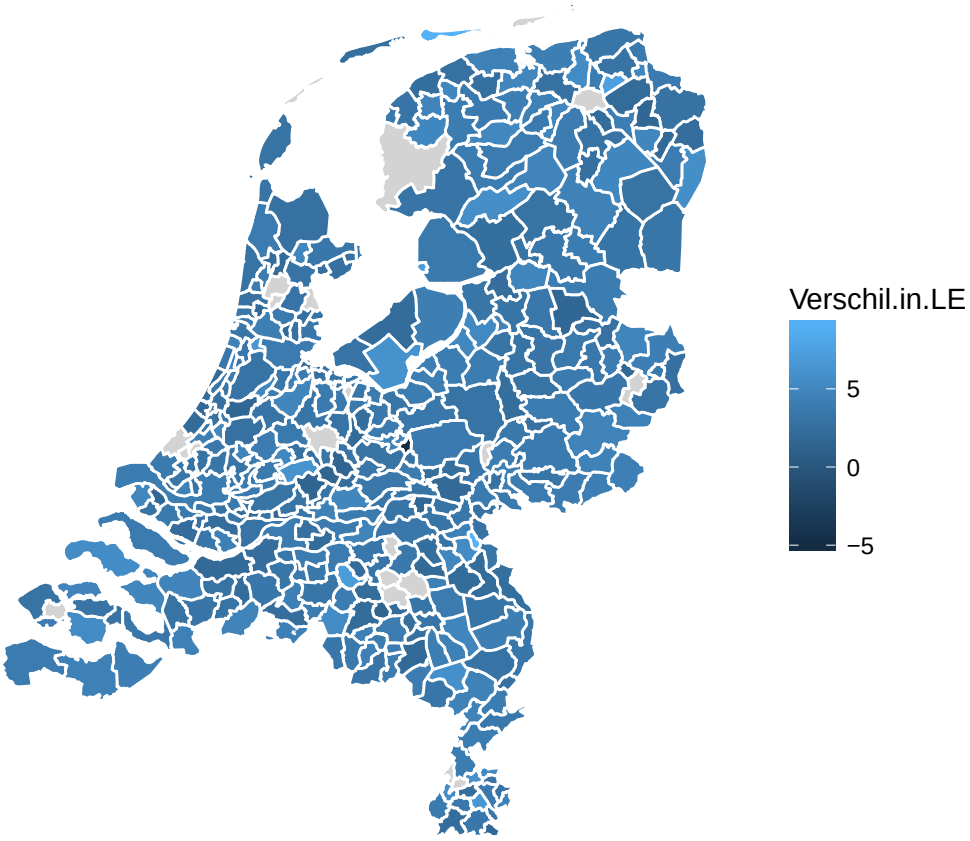
```
AddMapLayer(MapPlot(), map_municipal, map_info14) +
  guides(fill = guide_legend(title = "Life expectency women"))
```




```
hulpmapinfo15 <- d1 %>%
  mutate(Verschil.in.LE = (LEwomen - LEmen))
```

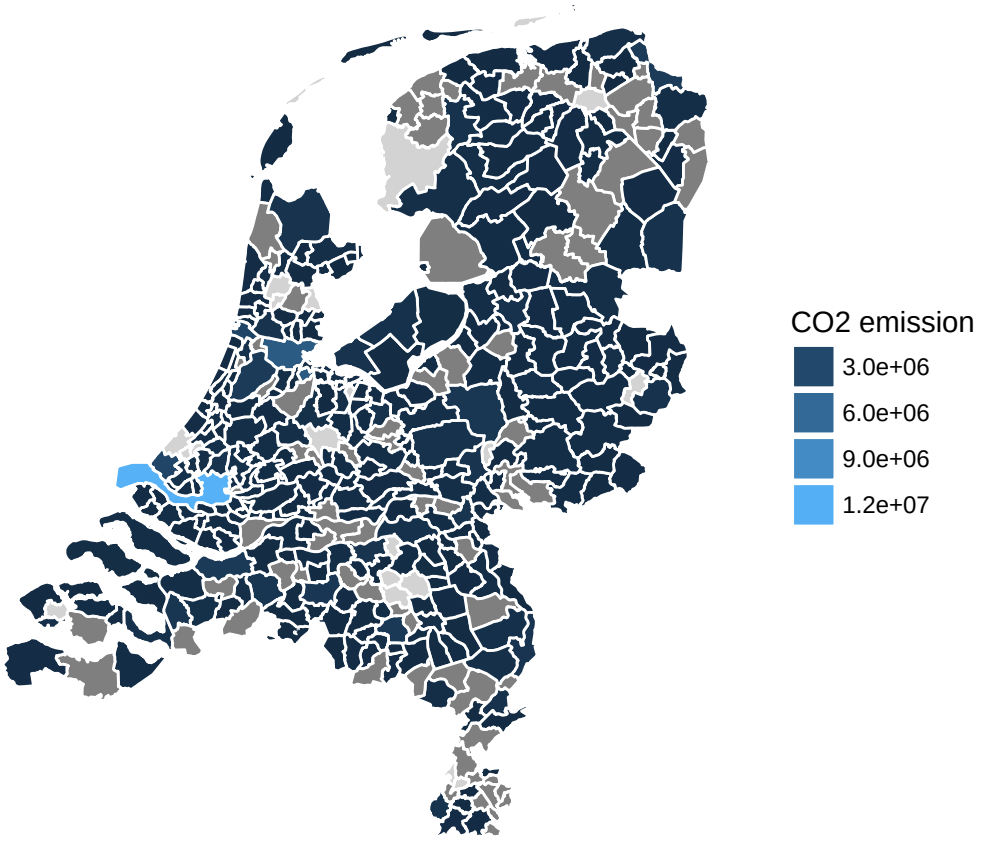
```
mapinfo15 <- hulpmapinfo15 %>%
  select(Gemeente,  Verschil.in.LE)
```

```
AddMapLayer(MapPlot(), map_municipal, mapinfo15)
```




```
map_info2 <- d1 %>%
    select(Gemeente, Totaal.bekende.CO2.uitstoot.2016)
```

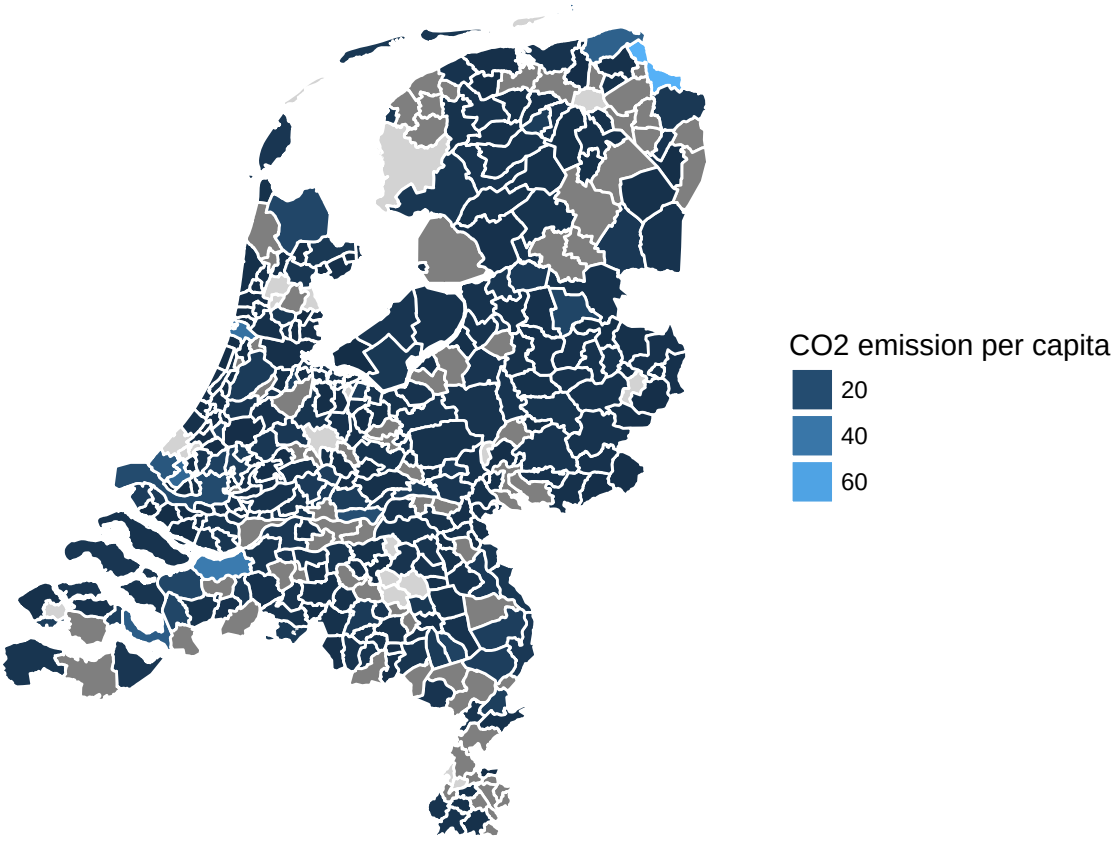
```
AddMapLayer(MapPlot(), map_municipal, map_info2) +
  guides(fill = guide_legend(title = "CO2 emission"))
```




```
d1a1 <- d1 %>%
mutate(uitstoot.per.capita = (Totaal.bekende.CO2.uitstoot.2016/TotaleBevolking_1))
```

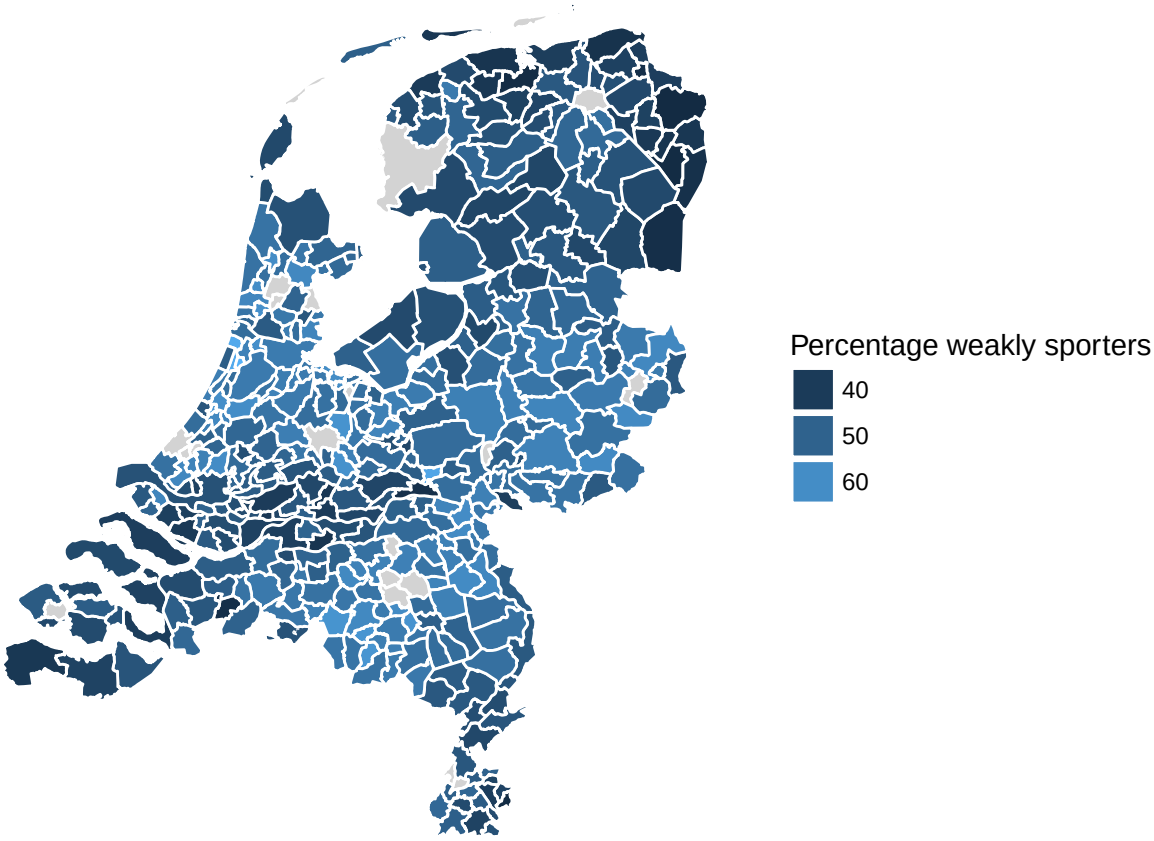
```
map_info2a <- d1a1 %>%
    select(Gemeente, uitstoot.per.capita)
```

```
AddMapLayer(MapPlot(), map_municipal, map_info2a) +
  guides(fill = guide_legend(title = "CO2 emission per capita"))
```



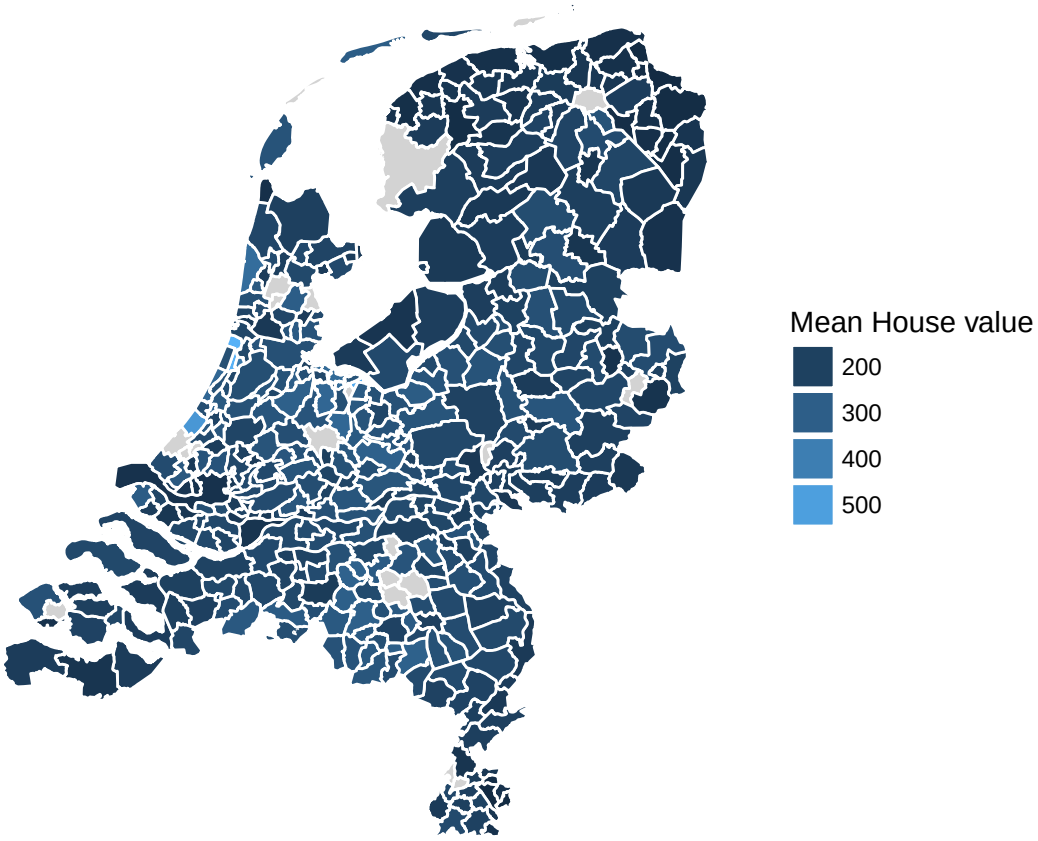

```
map_info3 <- d1 %>%
    select(Gemeente, WekelijkseSporters_16)
```

```
AddMapLayer(MapPlot(), map_municipal, map_info3) +
  guides(fill = guide_legend(title = "Percentage weakly sporters"))
```



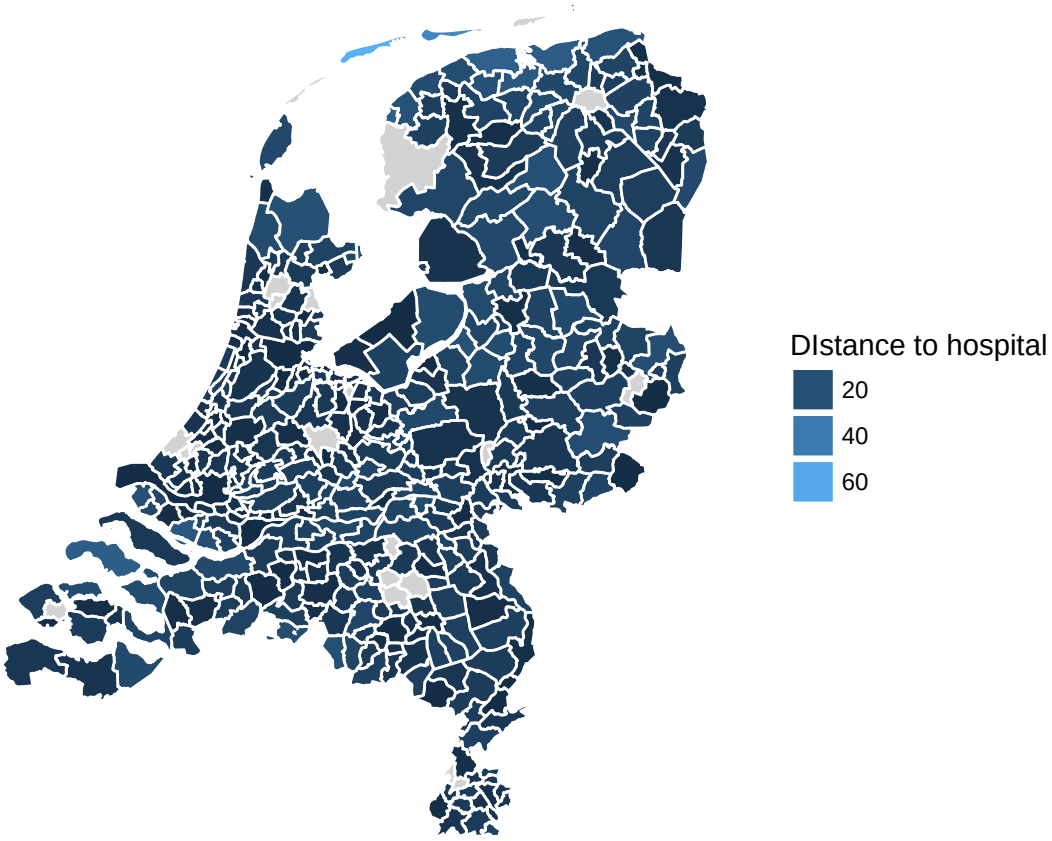

```
map_info8 <- d1 %>%
    select(Gemeente, GemiddeldeWoningwaarde_99)
```

```
AddMapLayer(MapPlot(), map_municipal, map_info8) +
  guides(fill = guide_legend(title = "Mean House value"))
```



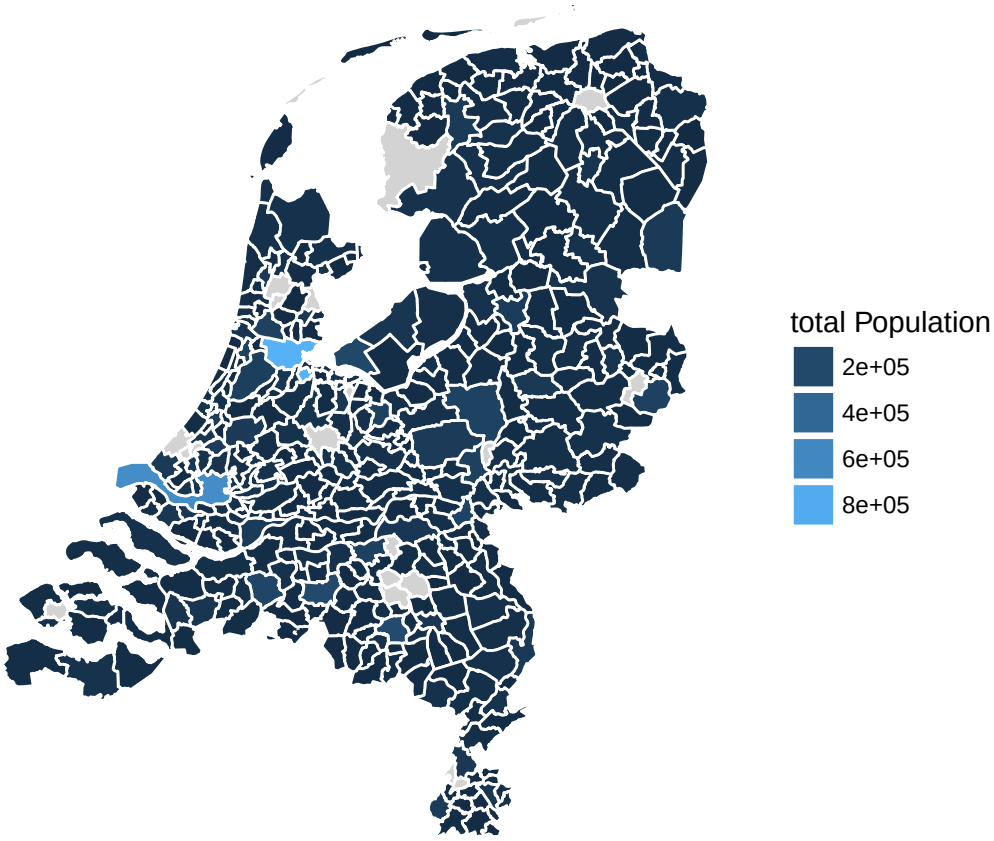

```
map_info9 <- d1 %>%
    select(Gemeente, AfstandTotZiekenhuis_216)
```

```
AddMapLayer(MapPlot(), map_municipal, map_info9) +
  guides(fill = guide_legend(title = "DIstance to hospital"))
```



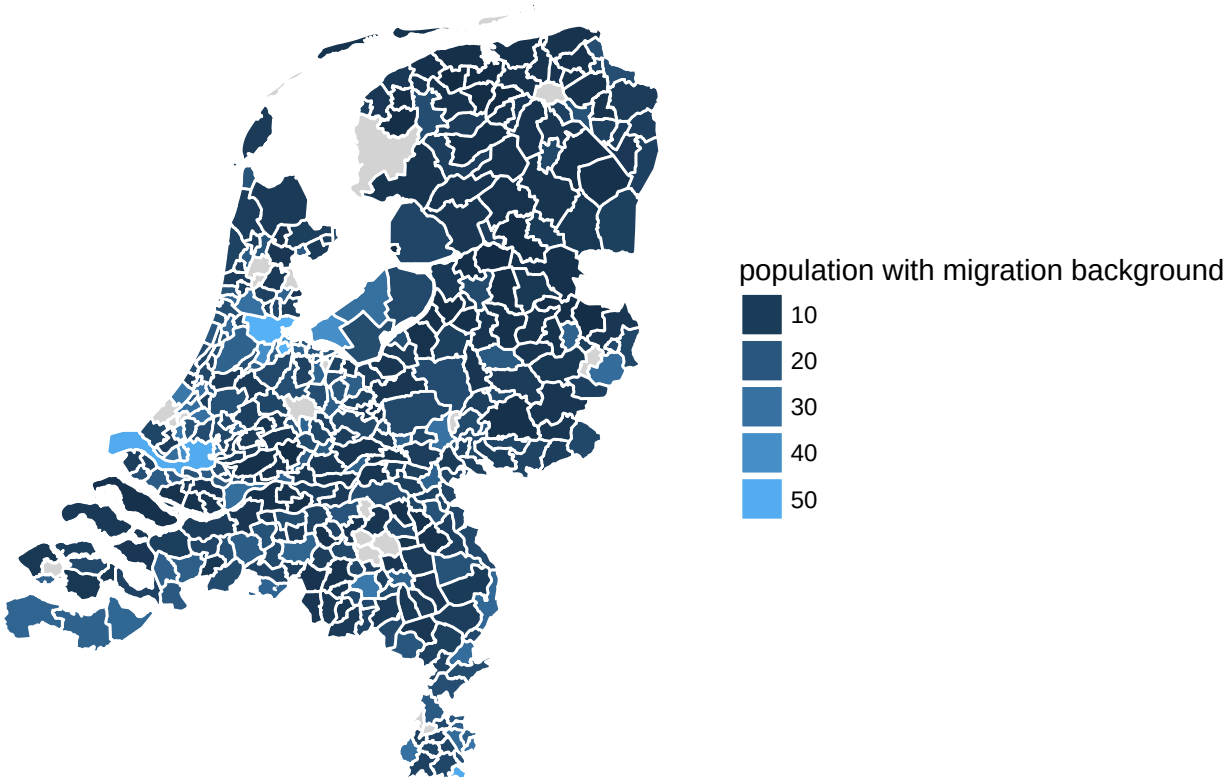

```
map_info10 <- d1 %>%
    select(Gemeente, TotaleBevolking_1)
```

```
AddMapLayer(MapPlot(), map_municipal, map_info10) +
  guides(fill = guide_legend(title = "total Population"))
```



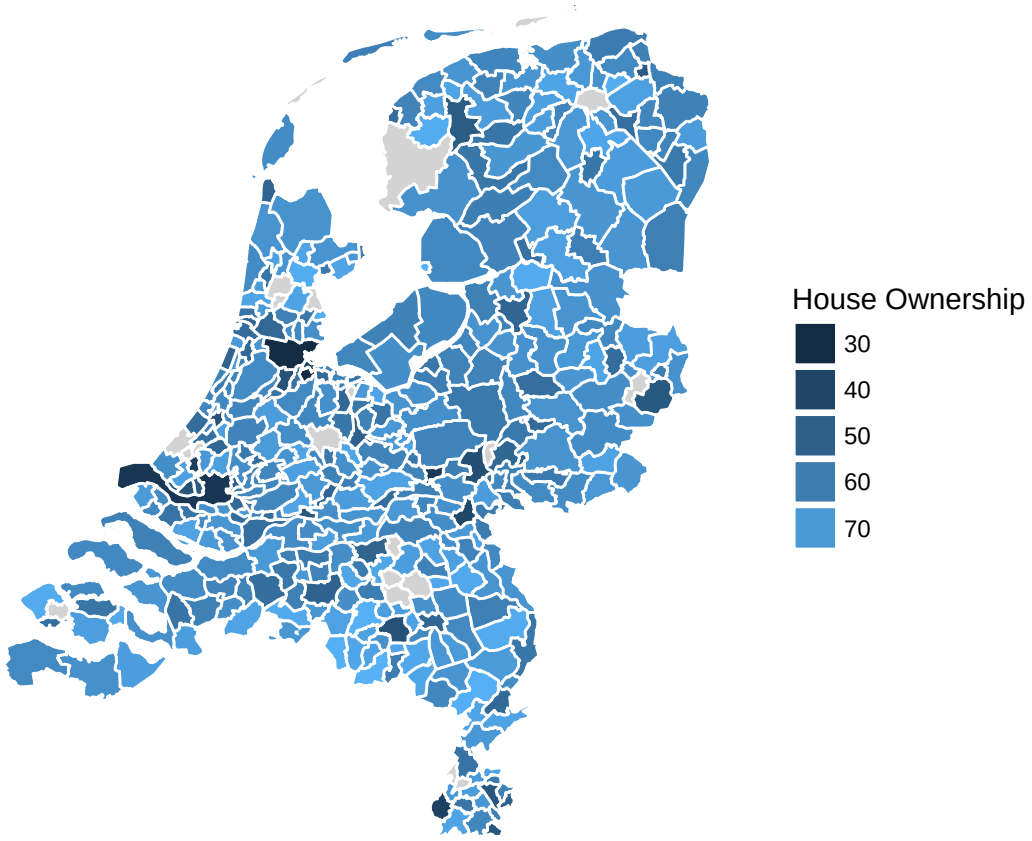

```
map_info11 <- d1 %>%
    select(Gemeente, TotaalMetMigratieachtergrond_44)
```

```
AddMapLayer(MapPlot(), map_municipal, map_info11) +
  guides(fill = guide_legend(title = "population with migration background"))
```



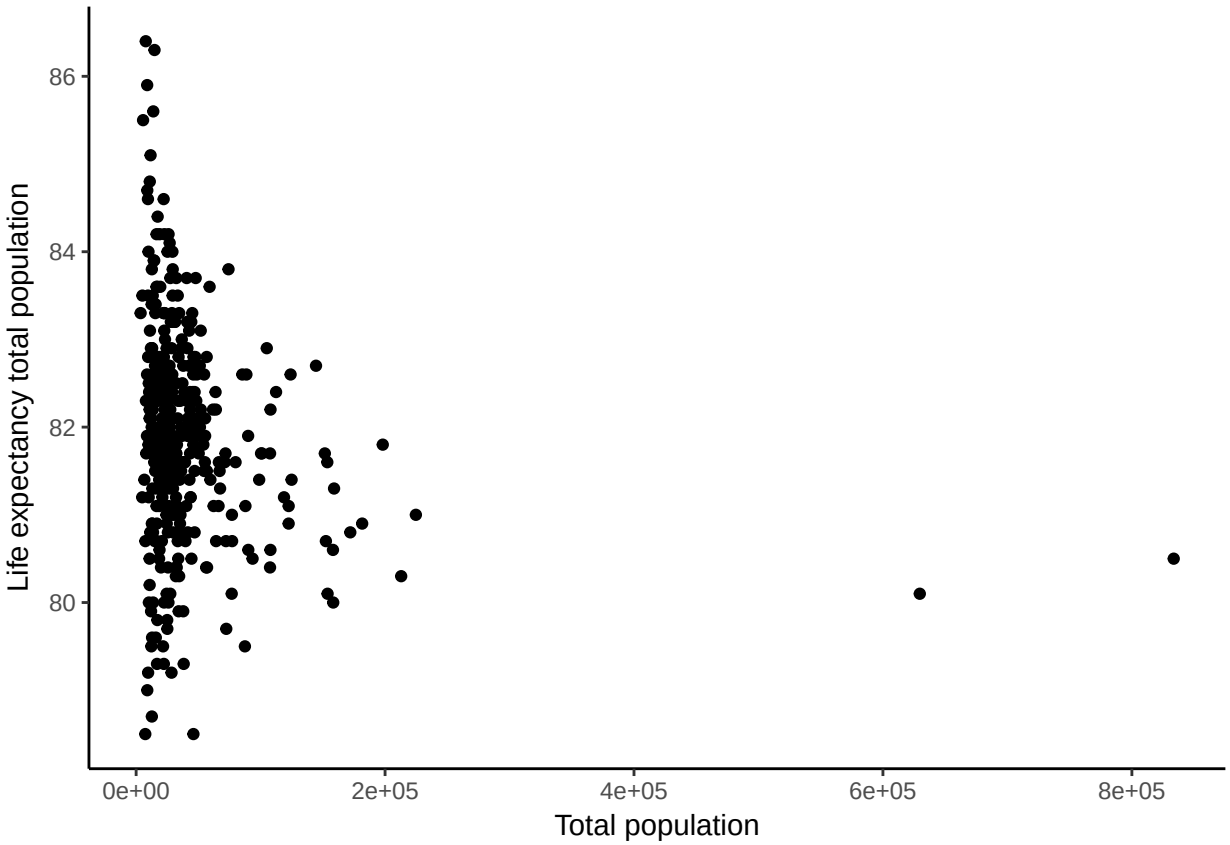

```
map_info12 <- d1 %>%
    select(Gemeente, Koopwoningen_94)
```

```
AddMapLayer(MapPlot(), map_municipal, map_info12) +
  guides(fill = guide_legend(title = "House Ownership"))
```

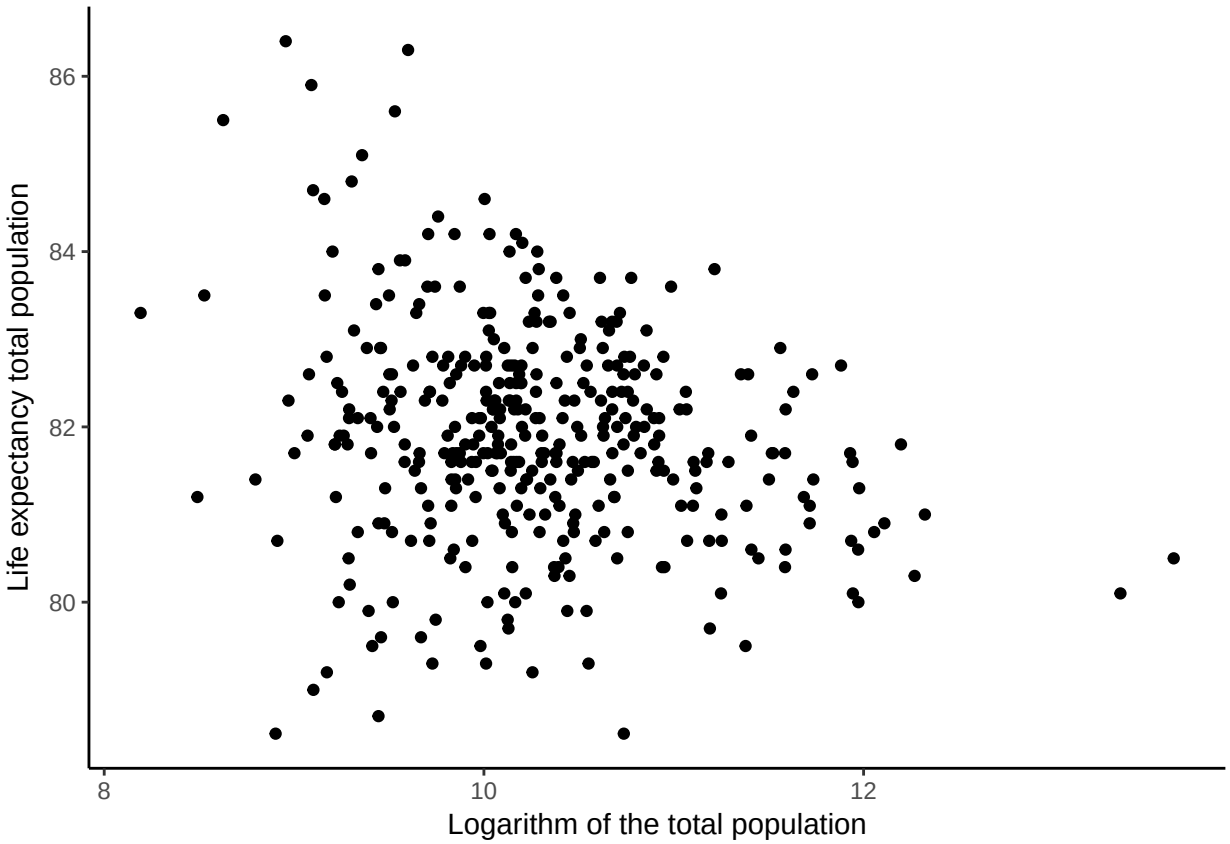


#Graphs

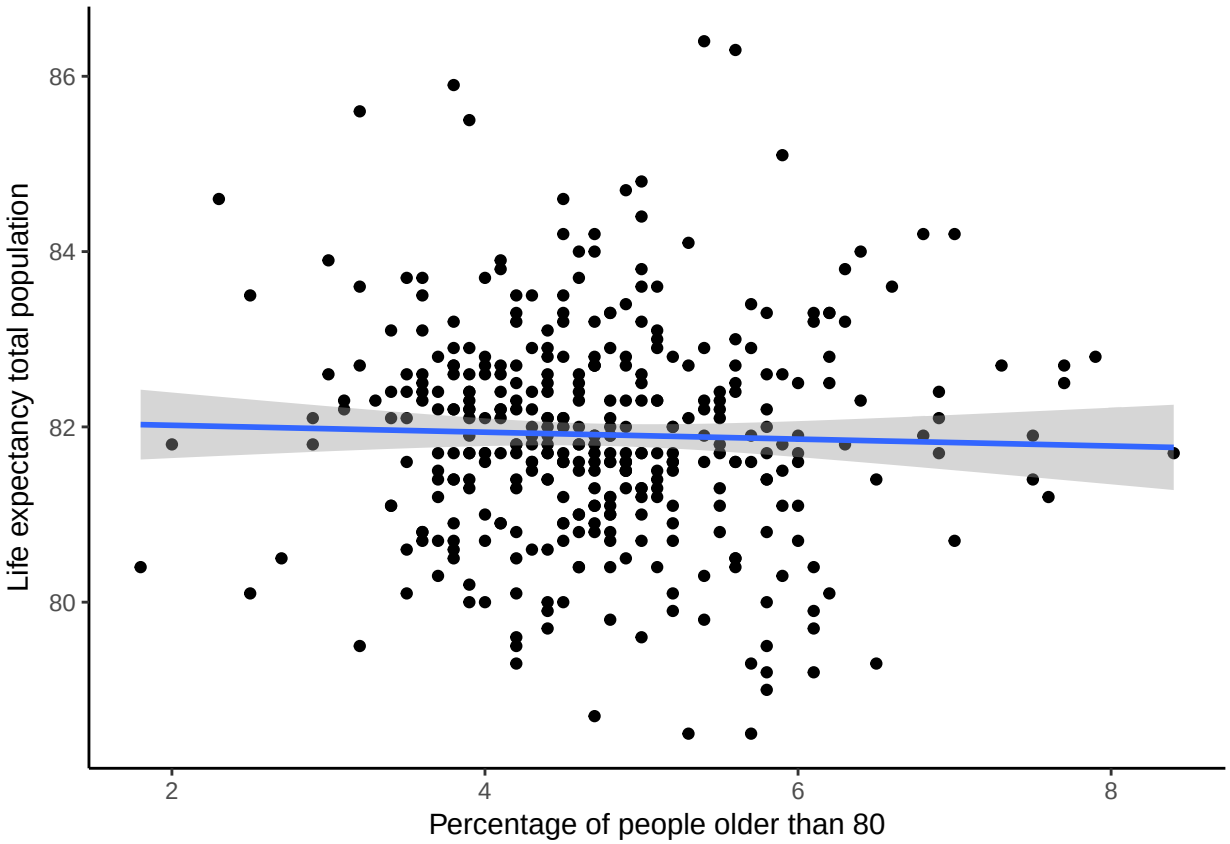
```
ggplot(d1, aes(x=TotaleBevolking_1, y=LEtotalpop)) + geom_point()+ theme_classic() + xlab("Total popula
```



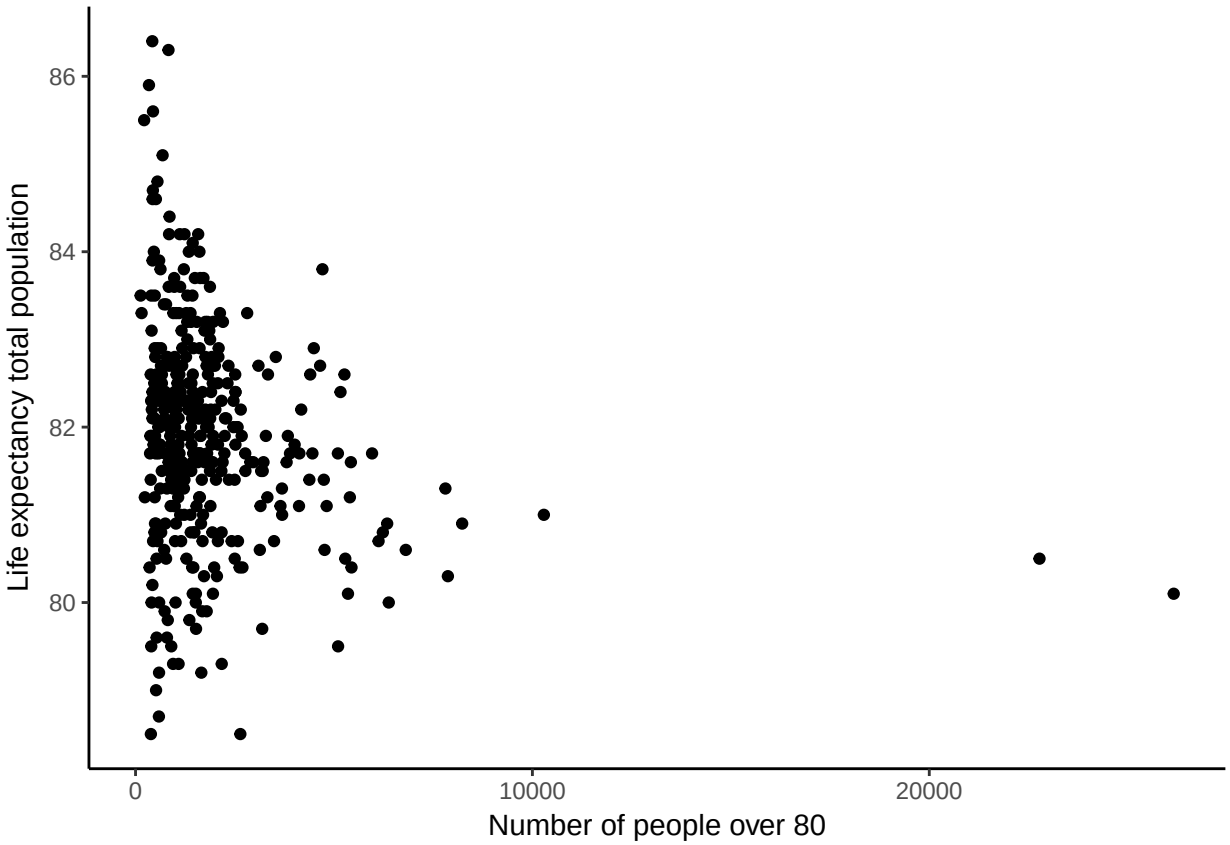

```
ggplot(d1, aes(x=log(TotaleBevolking_1), y=LEtotalpop)) + geom_point() + theme_classic()+ xlab("Logari
```



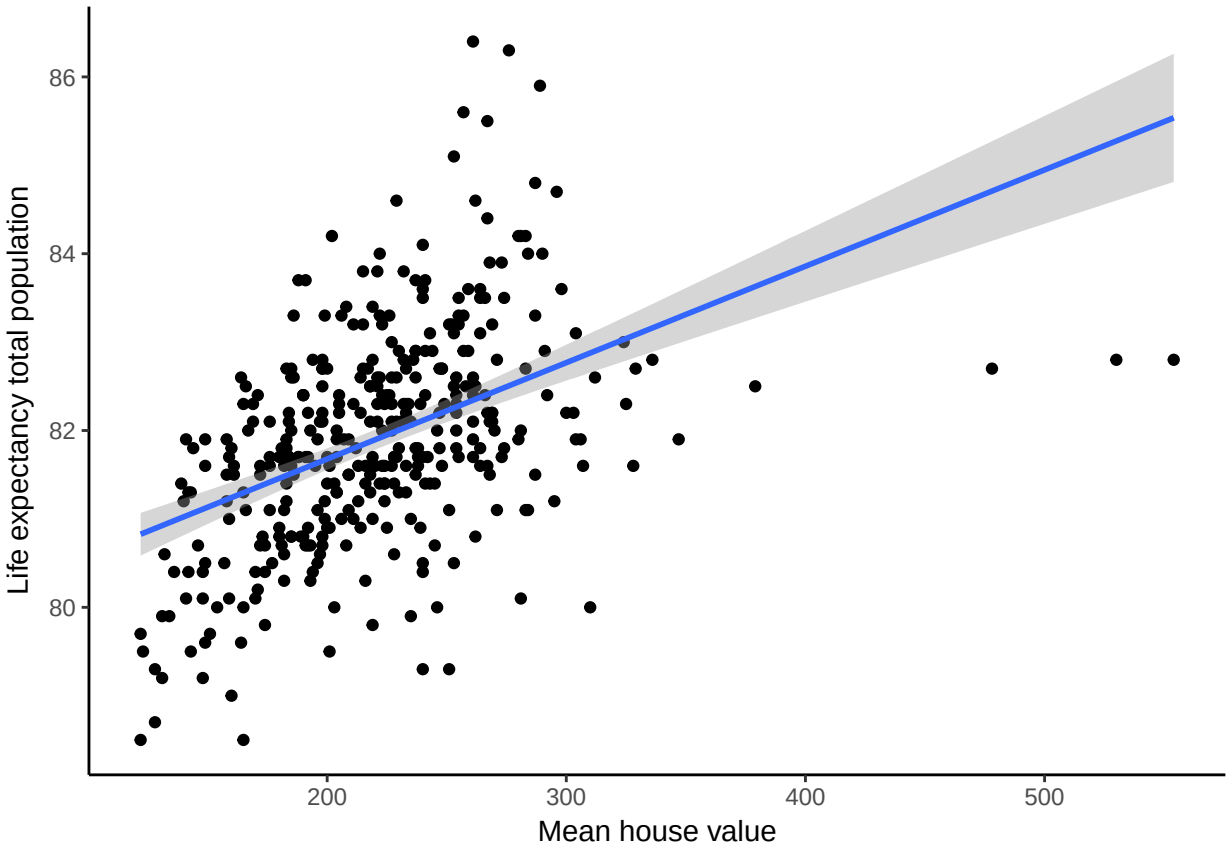

```
ggplot(d1, aes(x=k_80JaarOfOuder_21, y=LEtotalpop)) + geom_point()+geom_smooth(method = "lm") + theme_cl
```




```
ggplot(d1, aes(x=k_80JaarOfOuder_12, y=LEtotalpop)) + geom_point() + theme_classic()+ xlab("Number of p
```

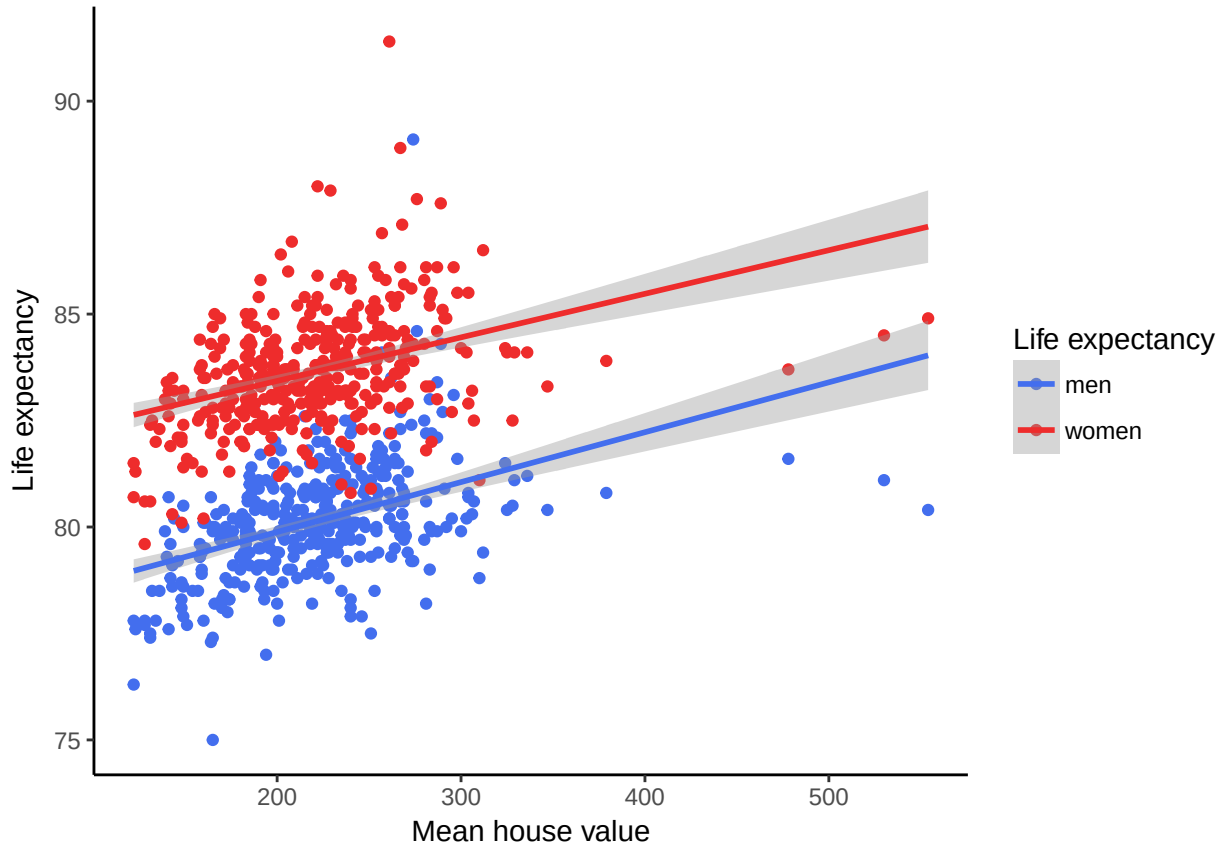



```
ggplot(d1, aes(x=GemiddeldeWoningwaarde_99, y=LEtotalpop)) + geom_point() + geom_smooth(method = "lm") +
```

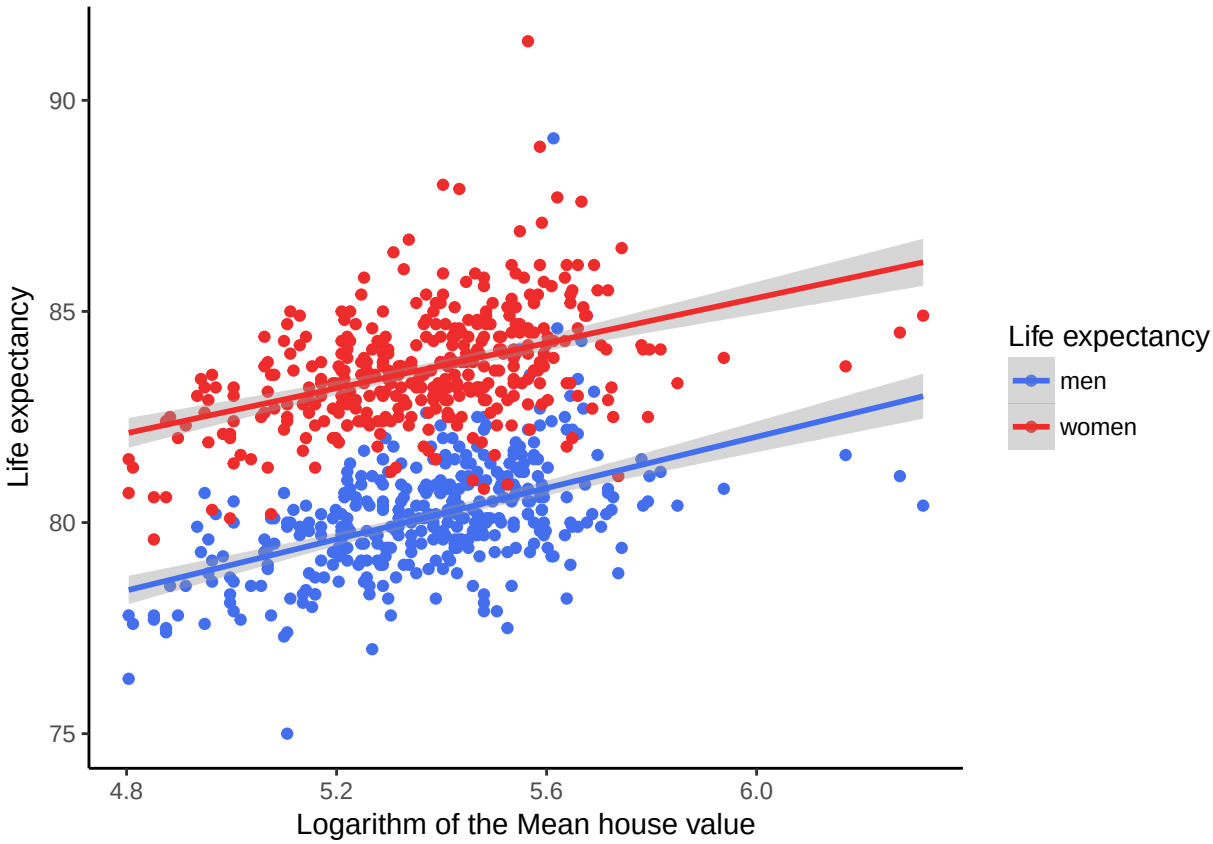



```
d_MHV <- d1 %>%
  select(Gemeente, GemiddeldeWoningwaarde_99, LEmen, LEwomen) %>%
  gather(Soort, LE, 3:4)
```

```
ggplot(d_MHV, aes(x=GemiddeldeWoningwaarde_99, y=LE, group=Soort, color= Soort)) + geom_point() + geom_s
                        values=c("royalblue2", "firebrick2"),
                        name="Life expectancy",
                        breaks=c("LEmen", "LEwomen"),
                        labels=c("men", "women"))
```

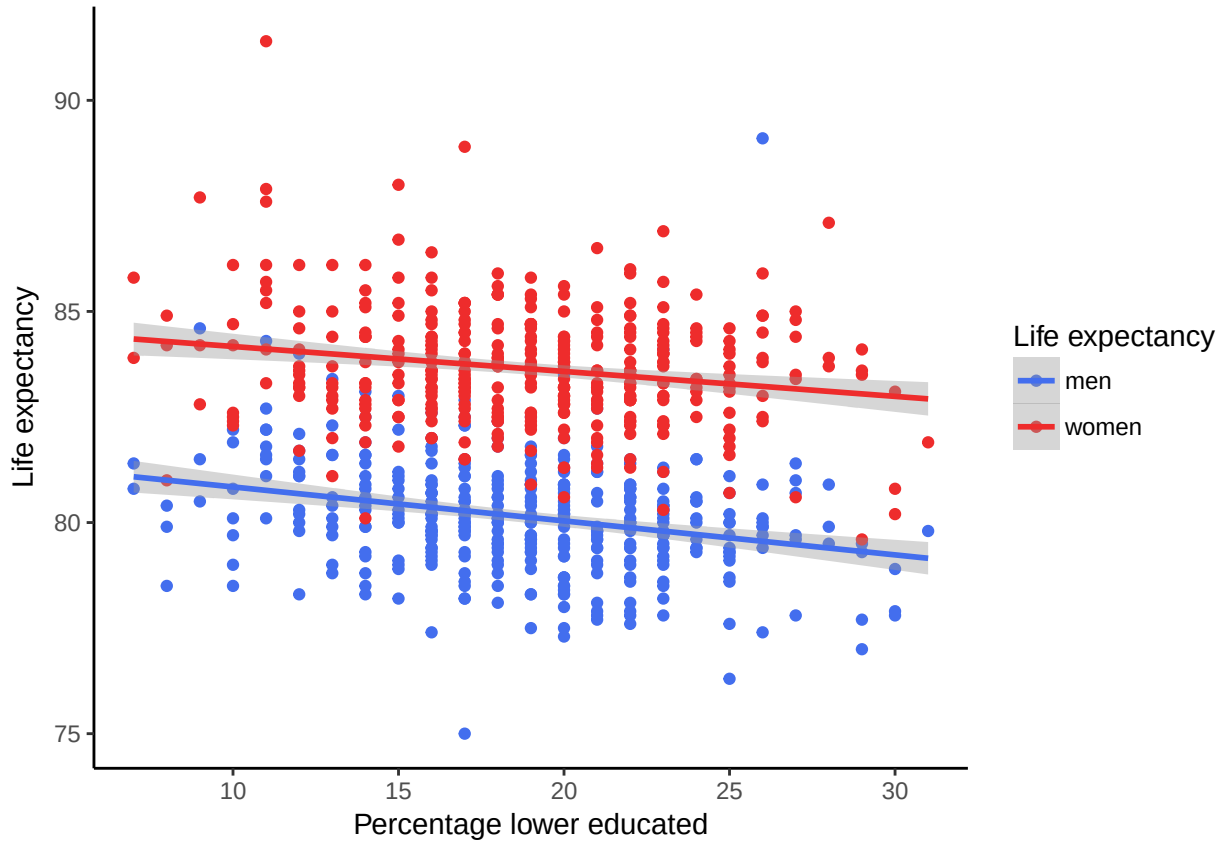



```
ggplot(d_MHV, aes(x=log(GemiddeldeWoningwaarde_99), y=LE, group=Soort, color= Soort)) + geom_point() + g
                        values=c("royalblue2", "firebrick2"),
                        name="Life expectancy",
                        breaks=c("LEmen", "LEwomen"),
                        labels=c("men", "women"))
```



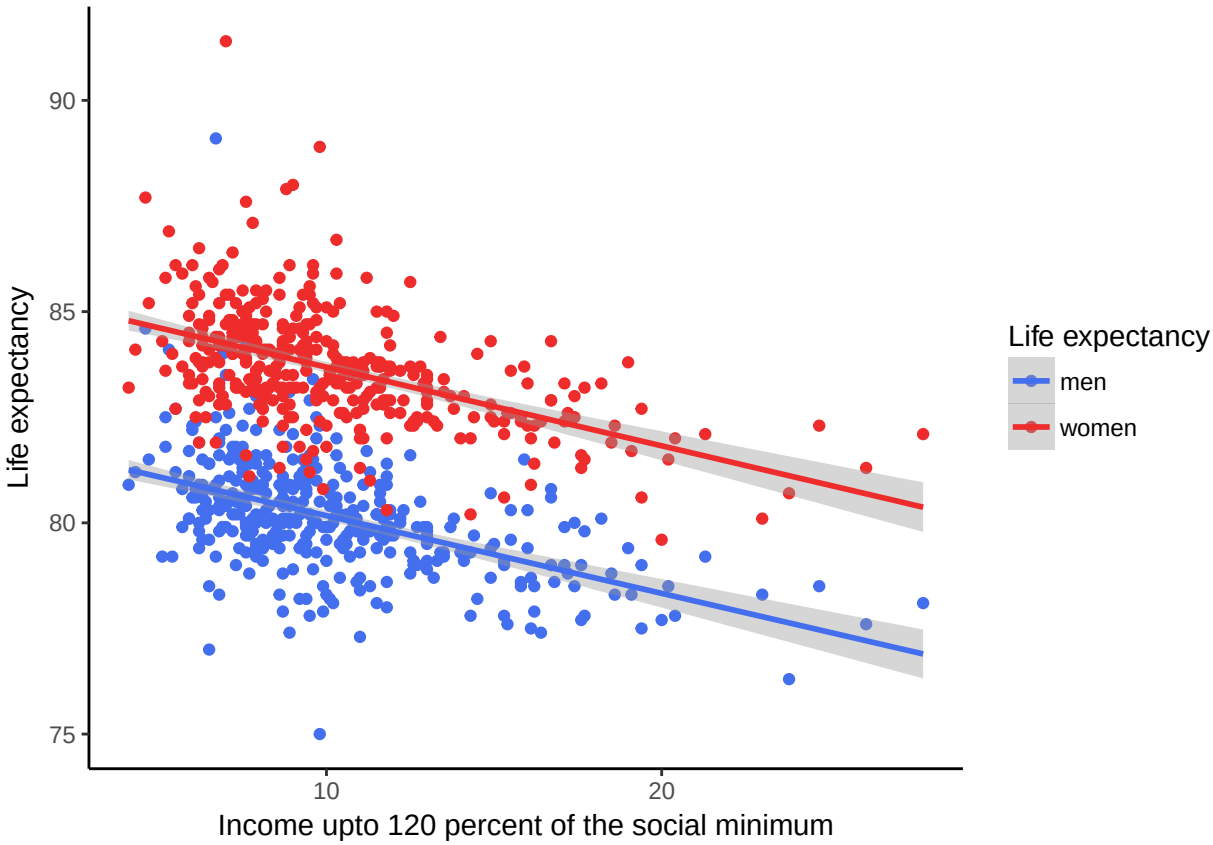

```
d_LOP <- d1 %>%
  select(Gemeente, LagerOpgeleidenPercentage_5, LEmen, LEwomen) %>%
  gather(Soort, LE, 3:4)
```

```
ggplot(d_LOP, aes(x=LagerOpgeleidenPercentage_5, y=LE, group=Soort, color= Soort)) + geom_point() + geo
                        values=c("royalblue2", "firebrick2"),
                        name="Life expectancy",
                        breaks=c("LEmen", "LEwomen"),
                        labels=c("men", "women"))
```



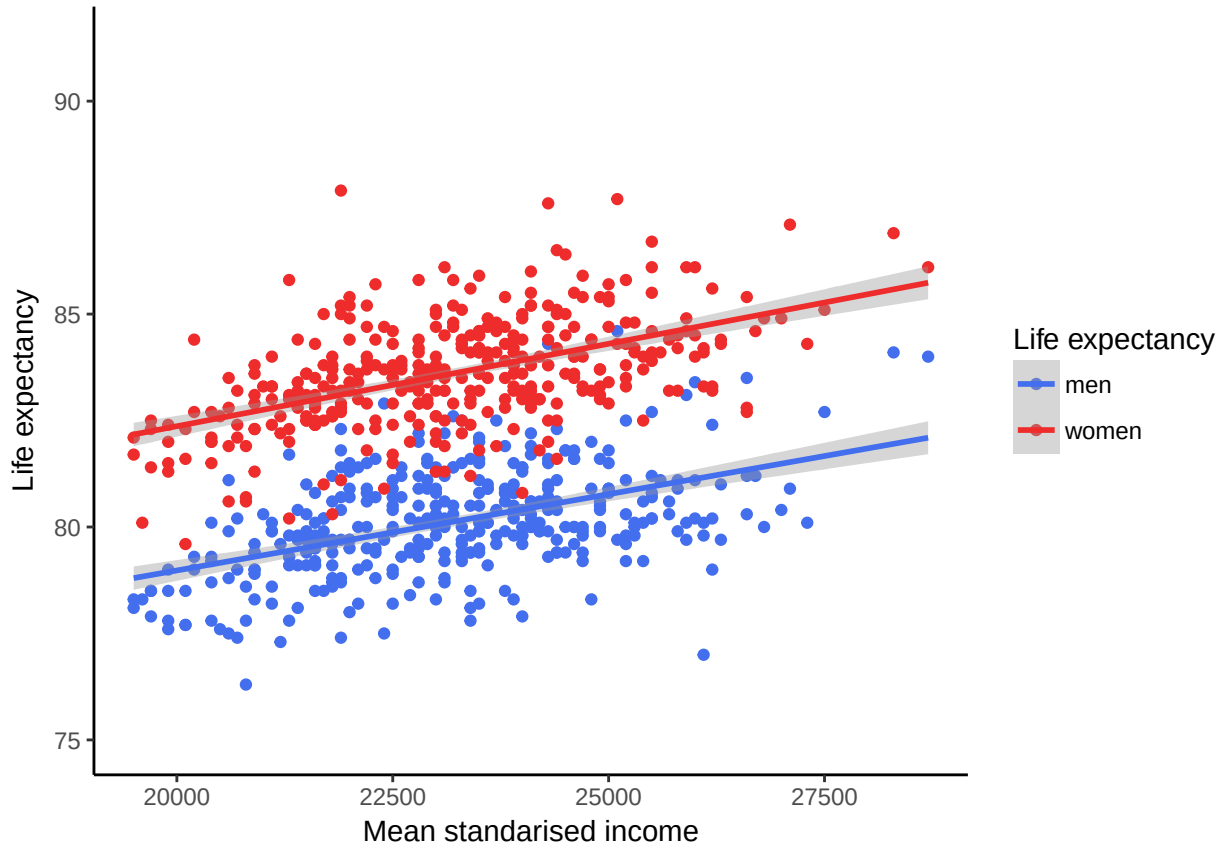

```
d_ISM <- d1 %>%
  select(Gemeente, InkomenTot120SociaalMinimum_13, LEmen, LEwomen) %>%
  gather(Soort, LE, 3:4)
```

```
ggplot(d_ISM, aes(x=InkomenTot120SociaalMinimum_13, y=LE, group=Soort, color= Soort)) + geom_point() + g
                        values=c("royalblue2", "firebrick2"),
                        name="Life expectancy",
                        breaks=c("LEmen", "LEwomen"),
                        labels=c("men", "women"))
```



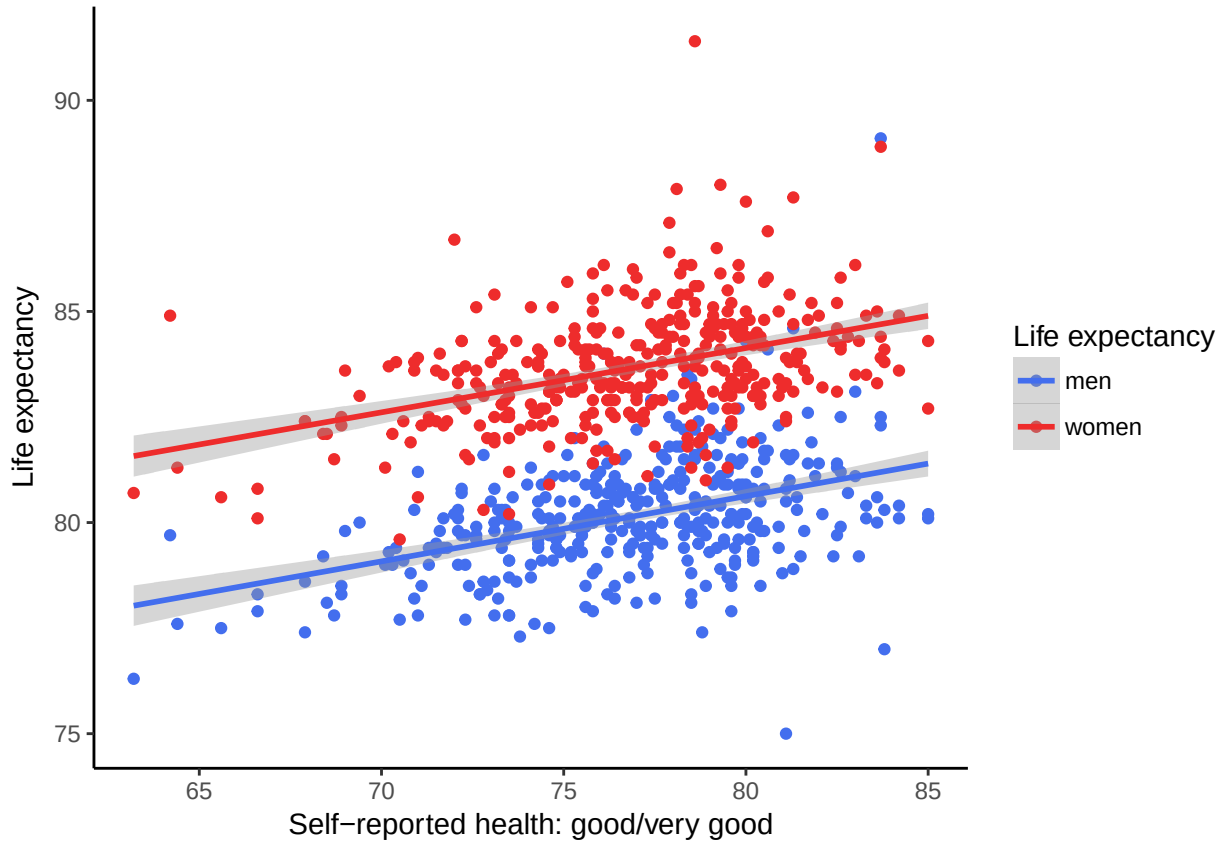

```
d_GGI <- d1 %>%
  select(Gemeente, GemiddeldGestandaardiseerdInkomen_41, LEmen, LEwomen) %>%
  gather(Soort, LE, 3:4)
```

```
ggplot(d_GGI, aes(x=GemiddeldGestandaardiseerdInkomen_41, y=LE, group=Soort, color= Soort)) + geom_poin
                        values=c("royalblue2", "firebrick2"),
                        name="Life expectancy",
                        breaks=c("LEmen", "LEwomen"),
                        labels=c("men", "women"))
```



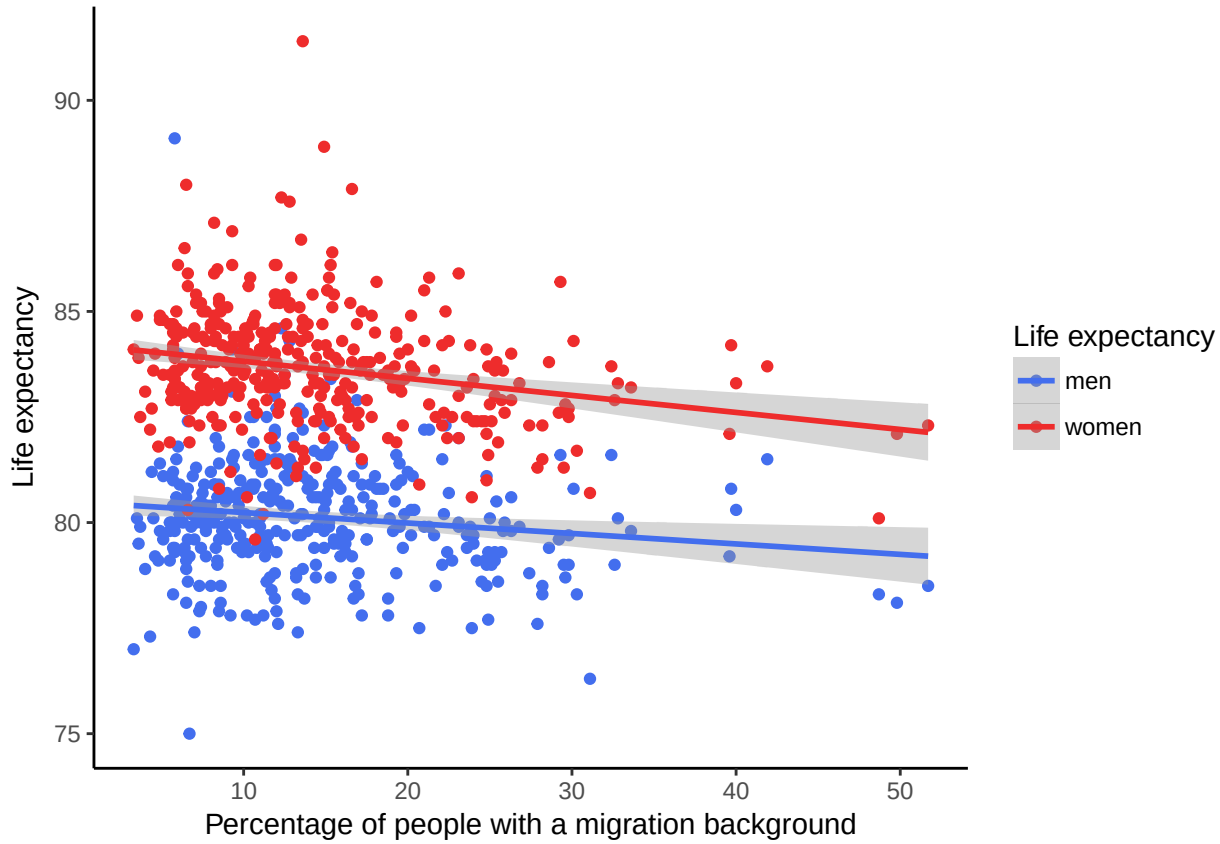

```
d_EGGZG <- d1 %>%
  select(Gemeente, ErvarenGezondheidGoedZeerGoed_1, LEmen, LEwomen) %>%
  gather(Soort, LE, 3:4)
```

```
ggplot(d_EGGZG, aes(x=ErvarenGezondheidGoedZeerGoed_1, y=LE, group=Soort, color= Soort)) + geom_point()
                        values=c("royalblue2", "firebrick2"),
                        name="Life expectancy",
                        breaks=c("LEmen", "LEwomen"),
                        labels=c("men", "women"))
```



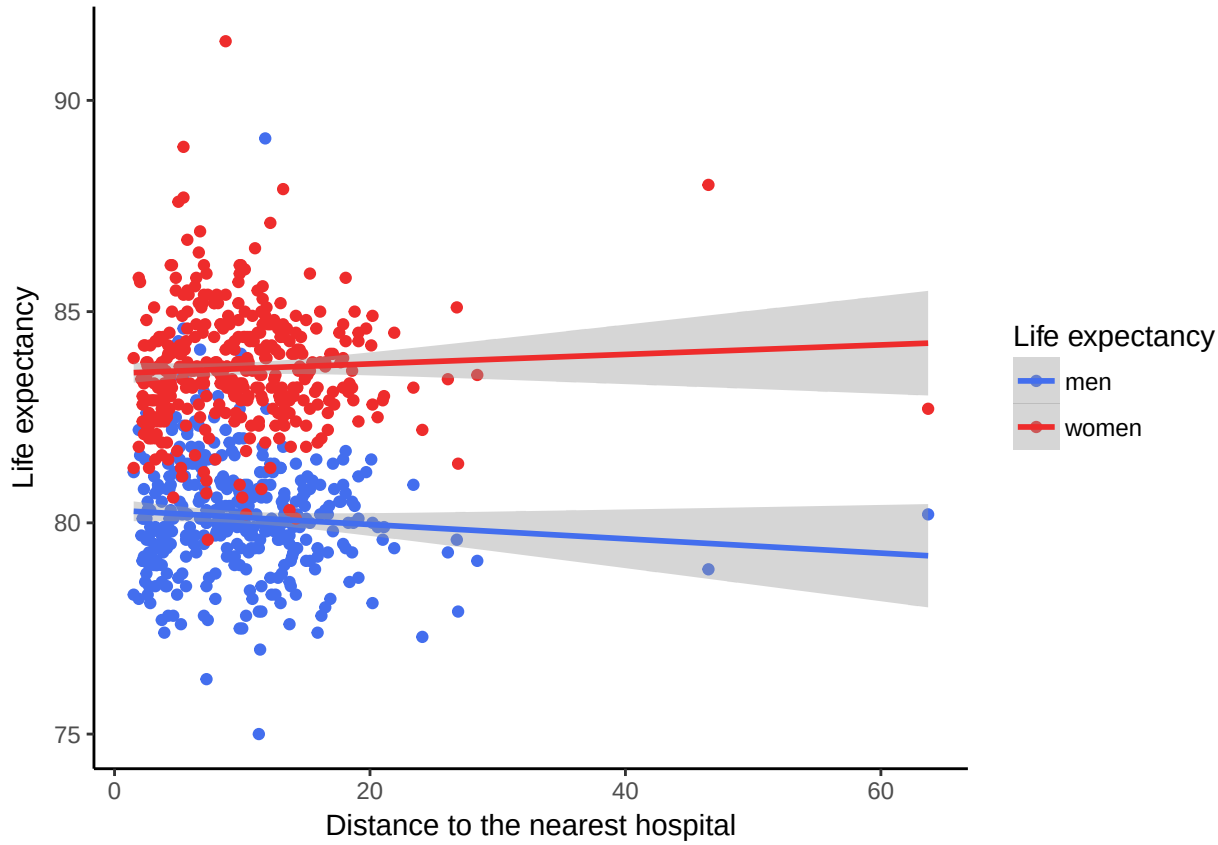

```
d_TMMA <- d1 %>%
  select(Gemeente, TotaalMetMigratieachtergrond_44, LEmen, LEwomen) %>%
  gather(Soort, LE, 3:4)
```

```
ggplot(d_TMMA, aes(x=TotaalMetMigratieachtergrond_44, y=LE, group=Soort, color= Soort)) + geom_point() +
                        values=c("royalblue2", "firebrick2"),
                        name="Life expectancy",
                        breaks=c("LEmen", "LEwomen"),
                        labels=c("men", "women"))
```

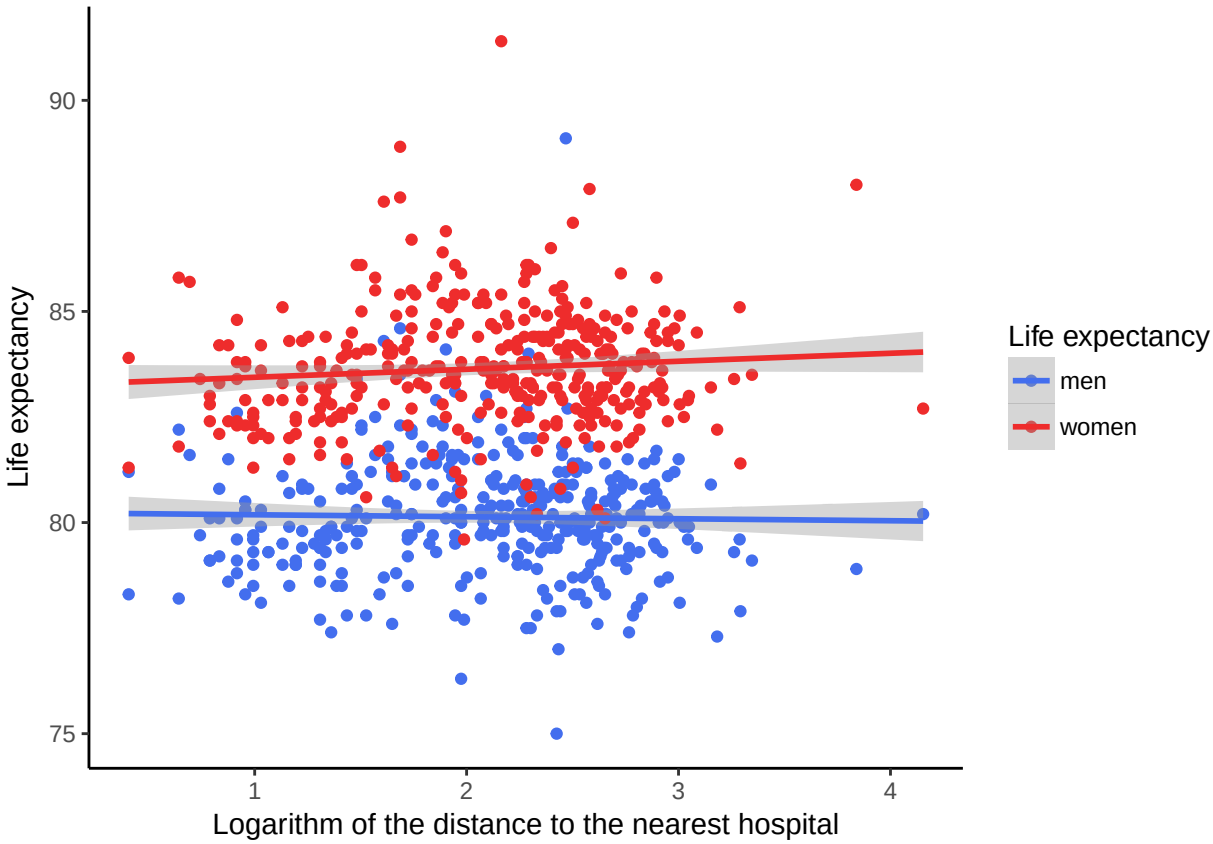



```
d_ATZ <- d1 %>%
  select(Gemeente, AfstandTotZiekenhuis_216, LEmen, LEwomen) %>%
  gather(Soort, LE, 3:4)
```

```
ggplot(d_ATZ, aes(x=AfstandTotZiekenhuis_216, y=LE, group=Soort, color= Soort)) + geom_point() + geom_sm
                        values=c("royalblue2", "firebrick2"),
                        name="Life expectancy",
                        breaks=c("LEmen", "LEwomen"),
                        labels=c("men", "women"))
```




```
ggplot(d_ATZ, aes(x=log(AfstandTotZiekenhuis_216), y=LE, group=Soort, color= Soort)) + geom_point() + g
                        values=c("royalblue2", "firebrick2"),
                        name="Life expectancy",
                        breaks=c("LEmen", "LEwomen"),
                        labels=c("men", "women"))
```



Collinearity

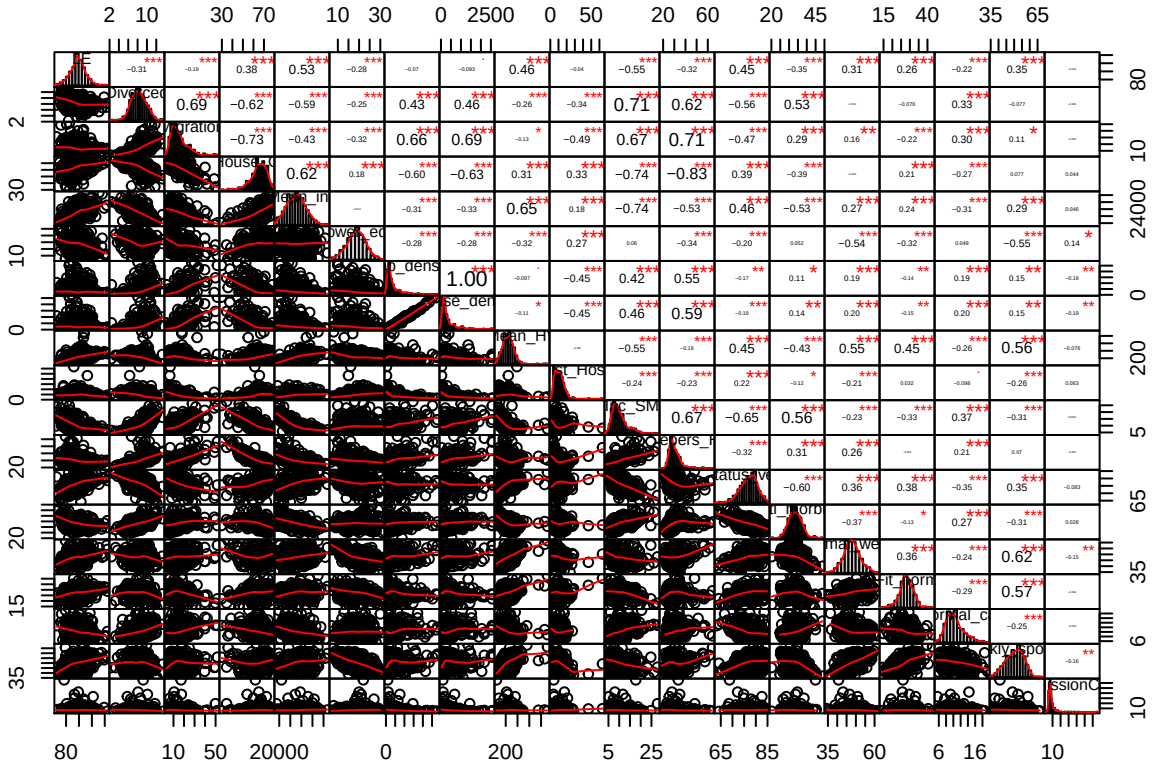
```
d1a <- d1a1 %>%
  select(-X, -Gemeente, -RegioS)


str(d1a)
```

```
## '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 ...
##  $ 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 ...
##  $ GemiddeldGestandaardiseerdInkomen_41: int  23000 23700 25000 23500 20900 22500 24800 21500 20100
```

```
##  $ InkomenTot120SociaalMinimum_13      : num  8.2 6.2 6.6 7.6 13 8.7 6.7 11.9 17.6 16 ...
##  $ LagerOpgeleidenPercentage_5         : int  12 31 23 20 30 18 14 16 21 17 ...
##  $ Percentage_eenpersoonshuishoudens   : num  26.4 23.3 27.9 27.9 28.7 ...
##  $ gemeente                            : Factor w/ 374 levels "'s-Hertogenbosch",..: 2 3 4 5 6 7 8 9
##  $ 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 ...
##  $ Totaal.bekende.CO2.uitstoot.2016    : int  NA NA NA 134297 159597 201343 117913 579053 571012 831
##  $ uitstoot.per.capita                 : num  NA NA NA 4.99 5.7 ...
```

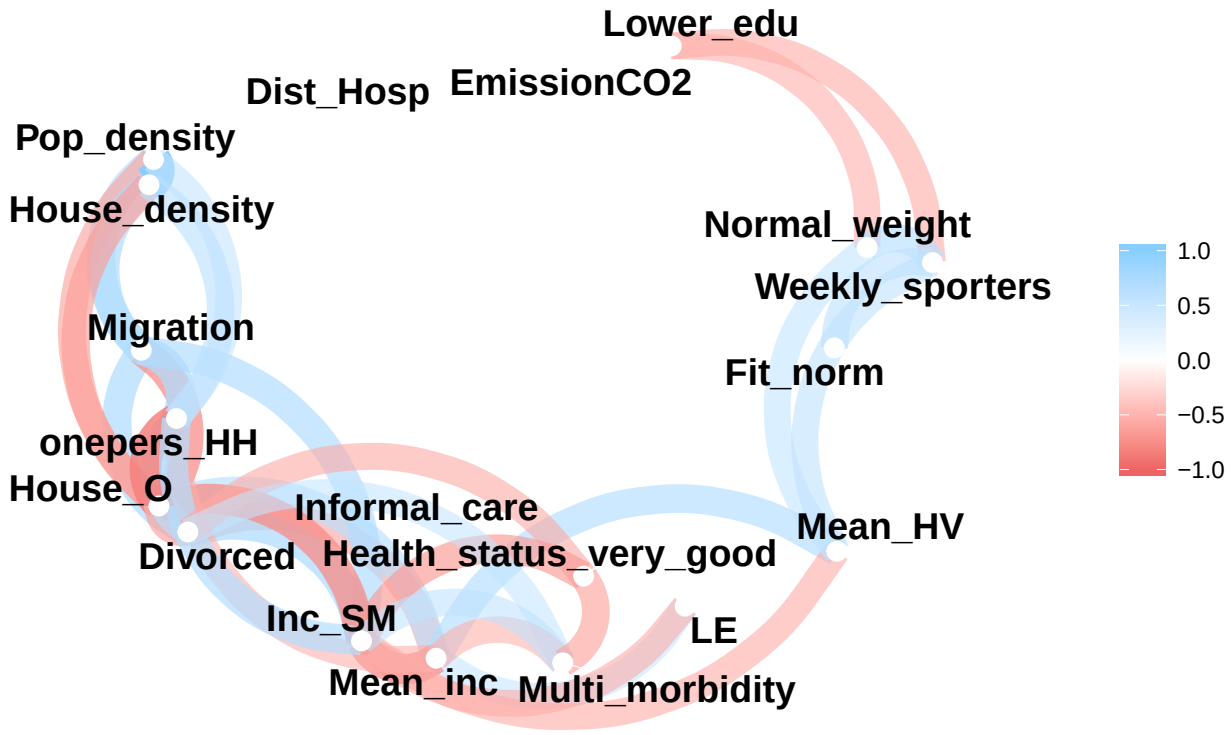
```
d1b <- d1a %>%
  select(LE = LEtotalpop, Divorced= Gescheiden_32, Migration = TotaalMetMigratieachtergrond_44, House_O
```

```
suppressWarnings(chart.Correlation(d1b))
```



```
d1b %>%
  correlate() %>%
  network_plot(min_cor = .5, repel= TRUE, curved = TRUE, legend = TRUE)
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```



Analysis

```
d2 <- d1a1 %>%
  select(-Gemeente,-gemeente, -X , -RegioS, -LEmen, -LEwomen, -uitstoot.per.capita, -Totaal.bekende.CO2
```

```
lm1 <- lm(LEtotalpop ~  ., data = d2)
summary(lm1)
```

```
##
## Call:
## lm(formula = LEtotalpop ~ ., data = d2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.24009 -0.53852 -0.01378  0.56261  2.80073
##
## Coefficients:
##                                      Estimate Std. Error t value
## (Intercept)                          6.873e+01  4.783e+00  14.371
## Gescheiden_32                        7.494e-02  7.124e-02   1.052
## TotaalMetMigratieachtergrond_44      2.881e-02  1.532e-02   1.881
## Bevolkingsdichtheid_57               2.655e-04  9.794e-04   0.271
## GemiddeldeHuishoudensgrootte_89      6.893e-01  1.152e+00   0.598
## Woningdichtheid_93                  -4.282e-04  2.182e-03  -0.196
```

```
## Koopwoningen_94                         2.576e-02  1.488e-02   1.731
## GemiddeldeWoningwaarde_99              -1.084e-03  1.704e-03  -0.636
## AfstandTotZiekenhuis_216               -2.817e-02  1.452e-02  -1.940
## TotaleBevolking_1                      -1.992e-07  3.529e-06  -0.056
## k_80JaarOfOuder_21                      1.236e-02  8.791e-02   0.141
## k_80JaarOfOuder_12                      3.735e-06  9.911e-05   0.038
## GemiddeldGestandaardiseerdInkomen_41    2.130e-04  5.702e-05   3.736
## InkomenTot120SociaalMinimum_13         -8.556e-02  3.551e-02  -2.409
## LagerOpgeleidenPercentage_5            -4.272e-02  1.779e-02  -2.401
## Percentage_eenpersoonshuishoudens      -2.761e-03  3.248e-02  -0.085
## ErvarenGezondheidGoedZeerGoed_1         8.329e-02  2.487e-02   3.348
## EenOfMeerLangdurigeAandoeningen_2       2.495e-02  2.040e-02   1.223
## NormaalGewicht_9                       -1.283e-02  1.942e-02  -0.661
## VoldoetAanFitnorm_14                    4.079e-04  1.929e-02   0.021
## UrenMantelzorgPerWeek_19               -9.331e-03  2.172e-02  -0.430
## WekelijkseSporters_16                  -7.608e-03  1.519e-02  -0.501
##                                        Pr(>|t|)
## (Intercept)                             < 2e-16 ***
## Gescheiden_32                          0.293751
## TotaalMetMigratieachtergrond_44        0.061014 .
## Bevolkingsdichtheid_57                 0.786486
## GemiddeldeHuishoudensgrootte_89        0.550113
## Woningdichtheid_93                     0.844536
## Koopwoningen_94                        0.084450 .
## GemiddeldeWoningwaarde_99              0.525031
## AfstandTotZiekenhuis_216               0.053402 .
## TotaleBevolking_1                      0.955025
## k_80JaarOfOuder_21                     0.888286
## k_80JaarOfOuder_12                     0.969964
## GemiddeldGestandaardiseerdInkomen_41 0.000225 ***
## InkomenTot120SociaalMinimum_13         0.016602 *
## LagerOpgeleidenPercentage_5            0.016973 *
## Percentage_eenpersoonshuishoudens      0.932309
## ErvarenGezondheidGoedZeerGoed_1        0.000920 ***
## EenOfMeerLangdurigeAandoeningen_2      0.222399
## NormaalGewicht_9                       0.509303
## VoldoetAanFitnorm_14                   0.983143
## UrenMantelzorgPerWeek_19               0.667793
## WekelijkseSporters_16                  0.616955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8818 on 291 degrees of freedom
##   (61 observations deleted due to missingness)
## Multiple R-squared:  0.4467, Adjusted R-squared:  0.4068
## F-statistic: 11.19 on 21 and 291 DF,  p-value: < 2.2e-16
```

Multicollinearity

VIF

```
vif(lm1)
```

```
##                         Gescheiden_32      TotaalMetMigratieachtergrond_44
##                              5.705206                              5.810465
##                Bevolkingsdichtheid_57      GemiddeldeHuishoudensgrootte_89
##                             365.393594                             14.503591
##                    Woningdichtheid_93                     Koopwoningen_94
##                             390.954875                              5.874586
##             GemiddeldeWoningwaarde_99             AfstandTotZiekenhuis_216
##                              3.153901                              1.913766
##                     TotaleBevolking_1                   k_80JaarOfOuder_21
##                             21.865585                              3.010674
##                   k_80JaarOfOuder_12 GemiddeldGestandaardiseerdInkomen_41
##                             21.555904                              3.719655
##        InkomenTot120SociaalMinimum_13          LagerOpgeleidenPercentage_5
##                              7.151678                              2.521600
##     Percentage_eenpersoonshuishoudens      ErvarenGezondheidGoedZeerGoed_1
##                             16.793665                              3.371330
##     EenOfMeerLangdurigeAandoeningen_2                    NormaalGewicht_9
##                              2.153334                              2.895083
##                  VoldoetAanFitnorm_14             UrenMantelzorgPerWeek_19
##                              2.580801                              1.275771
##                 WekelijkseSporters_16
##                              3.172771
```

We exclude the variable with the highest VIF (Which Woningdichtheid_93) and then run a VIF again.

```
lm1a <- lm(LEtotalpop ~  . -Woningdichtheid_93, data = d2)
vif(lm1a)
```

```
##                         Gescheiden_32      TotaalMetMigratieachtergrond_44
##                              5.605589                              5.807076
##                Bevolkingsdichtheid_57      GemiddeldeHuishoudensgrootte_89
##                              2.386683                             14.264664
##                     Koopwoningen_94             GemiddeldeWoningwaarde_99
##                              5.868994                              3.153779
##             AfstandTotZiekenhuis_216                     TotaleBevolking_1
##                              1.904859                             20.593716
##                   k_80JaarOfOuder_21                   k_80JaarOfOuder_12
##                              2.866814                             20.842889
## GemiddeldGestandaardiseerdInkomen_41        InkomenTot120SociaalMinimum_13
##                              3.664997                              6.688891
##          LagerOpgeleidenPercentage_5     Percentage_eenpersoonshuishoudens
##                              2.512324                             16.394575
##      ErvarenGezondheidGoedZeerGoed_1     EenOfMeerLangdurigeAandoeningen_2
##                              3.369972                              2.152358
##                     NormaalGewicht_9                  VoldoetAanFitnorm_14
##                              2.894579                              2.510979
##             UrenMantelzorgPerWeek_19                 WekelijkseSporters_16
##                              1.263101                              3.161343
```

As we see TotaleBevolking_1 now has a much smaller VIF. We will exclude k_80JaarOfOuder_12.

```
lm1b <- lm(LEtotalpop ~  . -Woningdichtheid_93 -k_80JaarOfOuder_12 , data = d2)
vif(lm1b)
```

```
##                         Gescheiden_32      TotaalMetMigratieachtergrond_44
##                              5.518427                              5.798244
##                Bevolkingsdichtheid_57      GemiddeldeHuishoudensgrootte_89
```

```
##                            2.373824                            13.820104
##                     Koopwoningen_94           GemiddeldeWoningwaarde_99
##                            5.860578                             3.056180
##            AfstandTotZiekenhuis_216                   TotaleBevolking_1
##                            1.884955                             2.145838
##                  k_80JaarOfOuder_21 GemiddeldGestandaardiseerdInkomen_41
##                            2.159121                             3.664788
##       InkomenTot120SociaalMinimum_13         LagerOpgeleidenPercentage_5
##                            6.688802                             2.494810
##    Percentage_eenpersoonshuishoudens      ErvarenGezondheidGoedZeerGoed_1
##                           15.903170                             3.361021
##    EenOfMeerLangdurigeAandoeningen_2                   NormaalGewicht_9
##                            2.139249                             2.894554
##                VoldoetAanFitnorm_14            UrenMantelzorgPerWeek_19
##                            2.508490                             1.262561
##               WekelijkseSporters_16
##                            3.129932
```

Now we have to exclude Percentage_eenpersoonshuishoudens

```
lm1c <- lm(LEtotalpop ~  . -Woningdichtheid_93 -k_80JaarOfOuder_12 -Percentage_eenpersoonshuishoudens ,
vif(lm1c)
```

```
##                        Gescheiden_32      TotaalMetMigratieachtergrond_44
##                            4.625759                             5.428687
##               Bevolkingsdichtheid_57      GemiddeldeHuishoudensgrootte_89
##                            2.344712                             5.456101
##                     Koopwoningen_94           GemiddeldeWoningwaarde_99
##                            4.307679                             3.045132
##            AfstandTotZiekenhuis_216                   TotaleBevolking_1
##                            1.882441                             2.137890
##                  k_80JaarOfOuder_21 GemiddeldGestandaardiseerdInkomen_41
##                            2.048042                             3.630407
##       InkomenTot120SociaalMinimum_13         LagerOpgeleidenPercentage_5
##                            6.615487                             2.487965
##      ErvarenGezondheidGoedZeerGoed_1    EenOfMeerLangdurigeAandoeningen_2
##                            3.232984                             2.122261
##                    NormaalGewicht_9                VoldoetAanFitnorm_14
##                            2.269688                             2.508352
##            UrenMantelzorgPerWeek_19               WekelijkseSporters_16
##                            1.262439                             3.077148
```

We have to exclude InkomenTot120SociaalMinimum_13

```
lm1d <- lm(LEtotalpop ~  . -Woningdichtheid_93 -k_80JaarOfOuder_12 -Percentage_eenpersoonshuishoudens -
vif(lm1d)
```

```
##                        Gescheiden_32      TotaalMetMigratieachtergrond_44
##                            4.440087                             5.236904
##               Bevolkingsdichtheid_57      GemiddeldeHuishoudensgrootte_89
##                            2.341934                             5.252130
##                     Koopwoningen_94           GemiddeldeWoningwaarde_99
##                            4.112214                             3.003954
##            AfstandTotZiekenhuis_216                   TotaleBevolking_1
##                            1.880795                             1.965704
##                  k_80JaarOfOuder_21 GemiddeldGestandaardiseerdInkomen_41
##                            2.047186                             3.513078
```

```
##          LagerOpgeleidenPercentage_5      ErvarenGezondheidGoedZeerGoed_1
##                             2.356467                             3.133783
##    EenOfMeerLangdurigeAandoeningen_2                    NormaalGewicht_9
##                             2.108718                             2.269499
##                VoldoetAanFitnorm_14            UrenMantelzorgPerWeek_19
##                             2.507794                             1.262068
##               WekelijkseSporters_16
##                             3.053537
```

We have to exclude GemiddeldeHuishoudensgrootte_89

```
lm1e <- lm(LEtotalpop ~  . -Woningdichtheid_93 -k_80JaarOfOuder_12 -Percentage_eenpersoonshuishoudens -
vif(lm1e)
```

```
##                        Gescheiden_32      TotaalMetMigratieachtergrond_44
##                             4.052095                             5.219084
##              Bevolkingsdichtheid_57                    Koopwoningen_94
##                             2.341277                             3.448235
##           GemiddeldeWoningwaarde_99            AfstandTotZiekenhuis_216
##                             2.870664                             1.850349
##                    TotaleBevolking_1                  k_80JaarOfOuder_21
##                             1.956895                             1.651725
## GemiddeldGestandaardiseerdInkomen_41         LagerOpgeleidenPercentage_5
##                             3.504535                             2.282380
##      ErvarenGezondheidGoedZeerGoed_1    EenOfMeerLangdurigeAandoeningen_2
##                             2.864838                             2.105555
##                    NormaalGewicht_9                VoldoetAanFitnorm_14
##                             2.195265                             2.507577
##             UrenMantelzorgPerWeek_19               WekelijkseSporters_16
##                             1.261347                             2.980982
```

We have to exclude TotaalMetMigratieachtergrond_44

```
lm1f <- lm(LEtotalpop ~  . -Woningdichtheid_93 -k_80JaarOfOuder_12 -Percentage_eenpersoonshuishoudens -
vif(lm1f)
```

```
##                        Gescheiden_32              Bevolkingsdichtheid_57
##                             3.457348                             2.138669
##                    Koopwoningen_94           GemiddeldeWoningwaarde_99
##                             3.053491                             2.869530
##            AfstandTotZiekenhuis_216                   TotaleBevolking_1
##                             1.804012                             1.757886
##                  k_80JaarOfOuder_21 GemiddeldGestandaardiseerdInkomen_41
##                             1.644660                             3.404595
##          LagerOpgeleidenPercentage_5      ErvarenGezondheidGoedZeerGoed_1
##                             2.260617                             2.713071
##    EenOfMeerLangdurigeAandoeningen_2                    NormaalGewicht_9
##                             2.091726                             2.193777
##                VoldoetAanFitnorm_14            UrenMantelzorgPerWeek_19
##                             2.481170                             1.254677
##               WekelijkseSporters_16
##                             2.961695
```

```
lm2 <- lm(LEtotalpop ~  . -Woningdichtheid_93 -k_80JaarOfOuder_12 -Percentage_eenpersoonshuishoudens -In

summary(lm2)
```

```
##
## Call:
## lm(formula = LEtotalpop ~ . - Woningdichtheid_93 - k_80JaarOfOuder_12 -
##     Percentage_eenpersoonshuishoudens - InkomenTot120SociaalMinimum_13 -
##     GemiddeldeHuishoudensgrootte_89 - TotaalMetMigratieachtergrond_44,
##     data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1724 -0.5719 -0.0007  0.5186  2.8505
##
## Coefficients:
##                                    Estimate Std. Error t value
## (Intercept)                        6.780e+01  3.132e+00  21.651
## Gescheiden_32                      4.733e-02  5.596e-02   0.846
## Bevolkingsdichtheid_57             1.133e-04  7.561e-05   1.498
## Koopwoningen_94                    3.662e-02  1.082e-02   3.383
## GemiddeldeWoningwaarde_99          1.905e-05  1.640e-03   0.012
## AfstandTotZiekenhuis_216          -3.582e-02  1.423e-02  -2.518
## TotaleBevolking_1                 -5.813e-07  1.010e-06  -0.576
## k_80JaarOfOuder_21                -4.253e-02  6.556e-02  -0.649
## GemiddeldGestandaardiseerdInkomen_41  2.556e-04  5.505e-05   4.644
## LagerOpgeleidenPercentage_5       -5.043e-02  1.700e-02  -2.967
## ErvarenGezondheidGoedZeerGoed_1    9.825e-02  2.252e-02   4.364
## EenOfMeerLangdurigeAandoeningen_2  1.987e-02  2.029e-02   0.979
## NormaalGewicht_9                  -1.810e-02  1.706e-02  -1.061
## VoldoetAanFitnorm_14              -1.238e-03  1.908e-02  -0.065
## UrenMantelzorgPerWeek_19          -8.597e-03  2.173e-02  -0.396
## WekelijkseSporters_16             -6.381e-03  1.481e-02  -0.431
##                                   Pr(>|t|)
## (Intercept)                        < 2e-16 ***
## Gescheiden_32                     0.398416
## Bevolkingsdichtheid_57            0.135136
## Koopwoningen_94                   0.000813 ***
## GemiddeldeWoningwaarde_99         0.990739
## AfstandTotZiekenhuis_216          0.012344 *
## TotaleBevolking_1                 0.565169
## k_80JaarOfOuder_21                0.517062
## GemiddeldGestandaardiseerdInkomen_41 5.15e-06 ***
## LagerOpgeleidenPercentage_5       0.003250 **
## ErvarenGezondheidGoedZeerGoed_1   1.77e-05 ***
## EenOfMeerLangdurigeAandoeningen_2 0.328195
## NormaalGewicht_9                  0.289496
## VoldoetAanFitnorm_14              0.948317
## UrenMantelzorgPerWeek_19          0.692719
## WekelijkseSporters_16             0.666936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8897 on 297 degrees of freedom
##   (61 observations deleted due to missingness)
## Multiple R-squared:  0.425,  Adjusted R-squared:  0.396
## F-statistic: 14.64 on 15 and 297 DF,  p-value: < 2.2e-16
```

Overfitting

We need complete cases

```
d2a <- d2 %>%
  na.omit()
```

Split dataset in train and test set

```
set.seed(123)
trainIndex <- createDataPartition(d2a$LEtotalpop, p=.8,
                                  list=FALSE,
                                  times =1)

df_train <- d2a[trainIndex,]
df_test <- d2a[-trainIndex,]
```

Now we can train the model

```
lmControl <- trainControl(method = "repeatedcv",
                          number = 10,
                          repeats = 5)

lmreg <- train(LEtotalpop ~  . -Woningdichtheid_93 -k_80JaarOfOuder_12 -Percentage_eenpersoonshuishouden
               data = df_train,
               method = "lm",
               trControl = lmControl)
```

```
print(lmreg)
```

```
## Linear Regression
##
## 252 samples
##  21 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 227, 228, 227, 227, 226, 228, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.9343824  0.3754164  0.7414065
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

And predict on the testset

```
df_test <-df_test %>%
mutate(predictionlm= predict(lmreg, df_test))
```

```
rfControl <- trainControl(method = "repeatedcv",
                          number = 10,
                          repeats = 5)

rf <- train(LEtotalpop ~  . ,
               data = df_train,
               method = "ranger",
               trControl = lmControl)
```

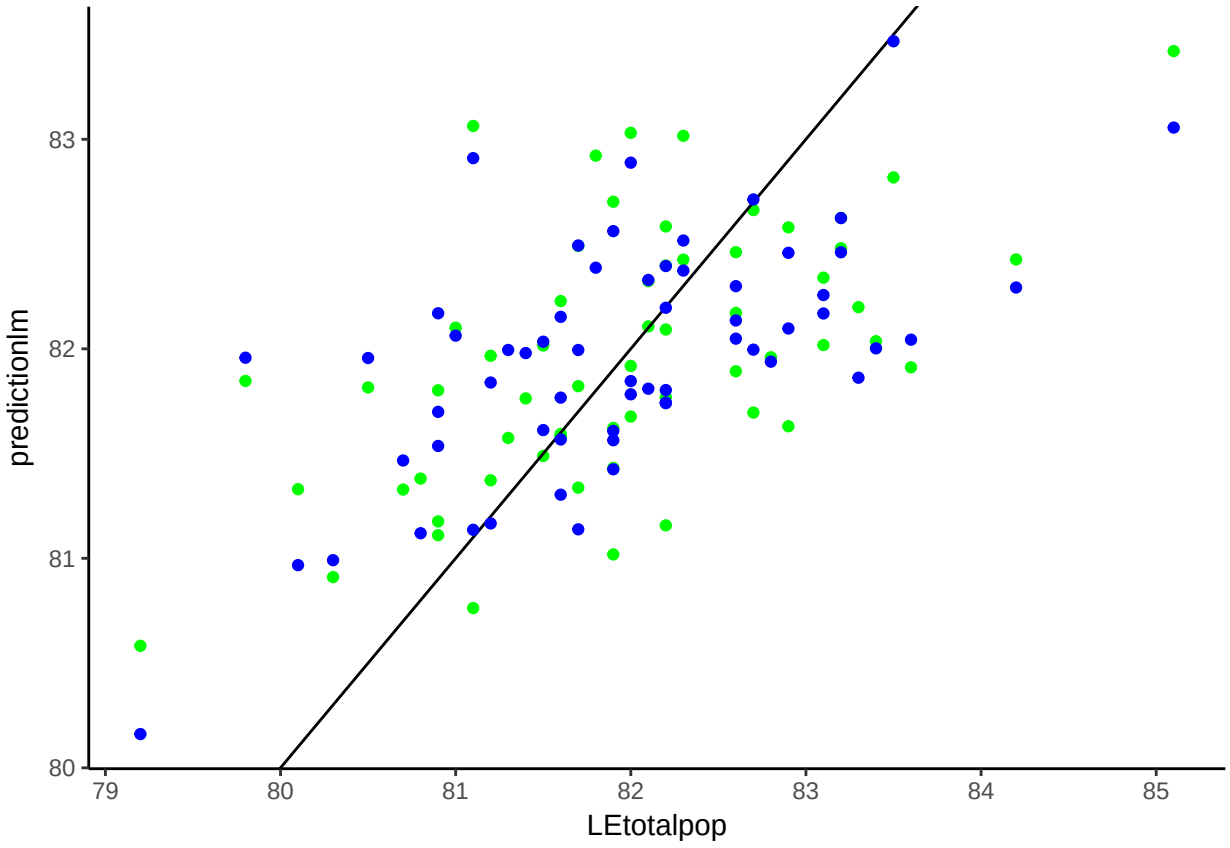
```
print(rf)
```

```
## Random Forest
##
## 252 samples
##  21 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 227, 227, 226, 226, 227, 227, ...
## Resampling results across tuning parameters:
##
##   mtry  splitrule   RMSE       Rsquared   MAE
##    2    variance    0.9272532  0.3908951  0.7086586
##    2    extratrees  0.9255243  0.4051069  0.7080013
##   11    variance    0.9282843  0.3833972  0.7109493
##   11    extratrees  0.9109605  0.4088283  0.7008000
##   21    variance    0.9346864  0.3777847  0.7182408
##   21    extratrees  0.9168971  0.3993272  0.7053522
##
## 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.
```

```
df_test <-df_test %>%
mutate(predictionrf= predict(rf, df_test))
```

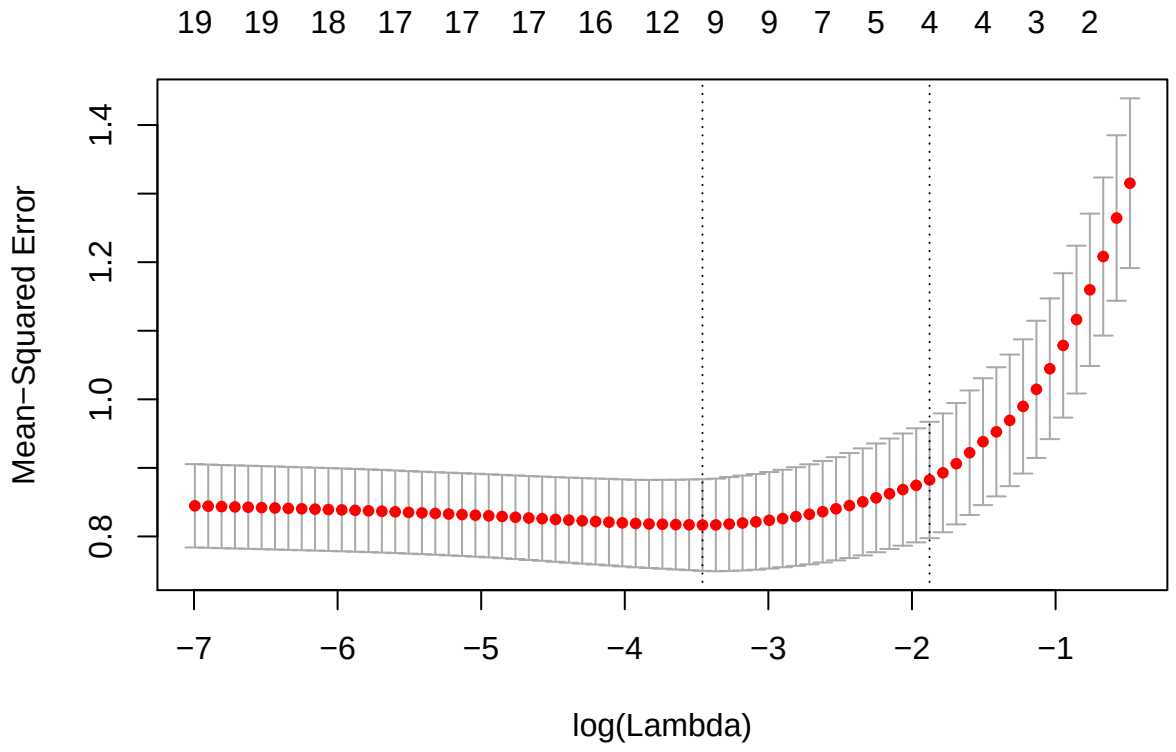
Now we can make a plot of real values versus predicted values.

```
ggplot(data= df_test, aes(x= LEtotalpop, y= predictionlm)) + geom_point(color= "green")+ geom_abline(sl
```



Lasso

```
set.seed(123)
d3 <- d2[complete.cases(d2),]  %>% select(-LEtotalpop)
d4 <- d2[complete.cases(d2),]  %>% select(LEtotalpop)


lasso <- cv.glmnet(x=as.matrix(d3),
                   y=as.matrix(d4),
                   alpha=1)

plot(lasso)
```




```
lasso$lambda.min
```

```
## [1] 0.0314621
```

```
lassobest <- glmnet(x=as.matrix(d3),
                    y=as.matrix(d4),
                    alpha=1, lambda = lasso$lambda.min)

predict(lassobest, type='coefficients', s=lasso$lambda.min)
```

```
## 22 x 1 sparse Matrix of class "dgCMatrix"
##                                                 1
## (Intercept)                          7.511830e+01
## Gescheiden_32                        .
## TotaalMetMigratieachtergrond_44      9.377194e-03
## Bevolkingsdichtheid_57               4.548143e-05
## GemiddeldeHuishoudensgrootte_89      .
## Woningdichtheid_93                   .
## Koopwoningen_94                      1.058573e-02
## GemiddeldeWoningwaarde_99            .
## AfstandTotZiekenhuis_216            -2.578973e-02
## TotaleBevolking_1                    .
## k_80JaarOfOuder_21                   .
## k_80JaarOfOuder_12                   .
## GemiddeldGestandaardiseerdInkomen_41 1.667540e-04
## InkomenTot120SociaalMinimum_13      -5.949756e-02
## LagerOpgeleidenPercentage_5         -3.906004e-02
```

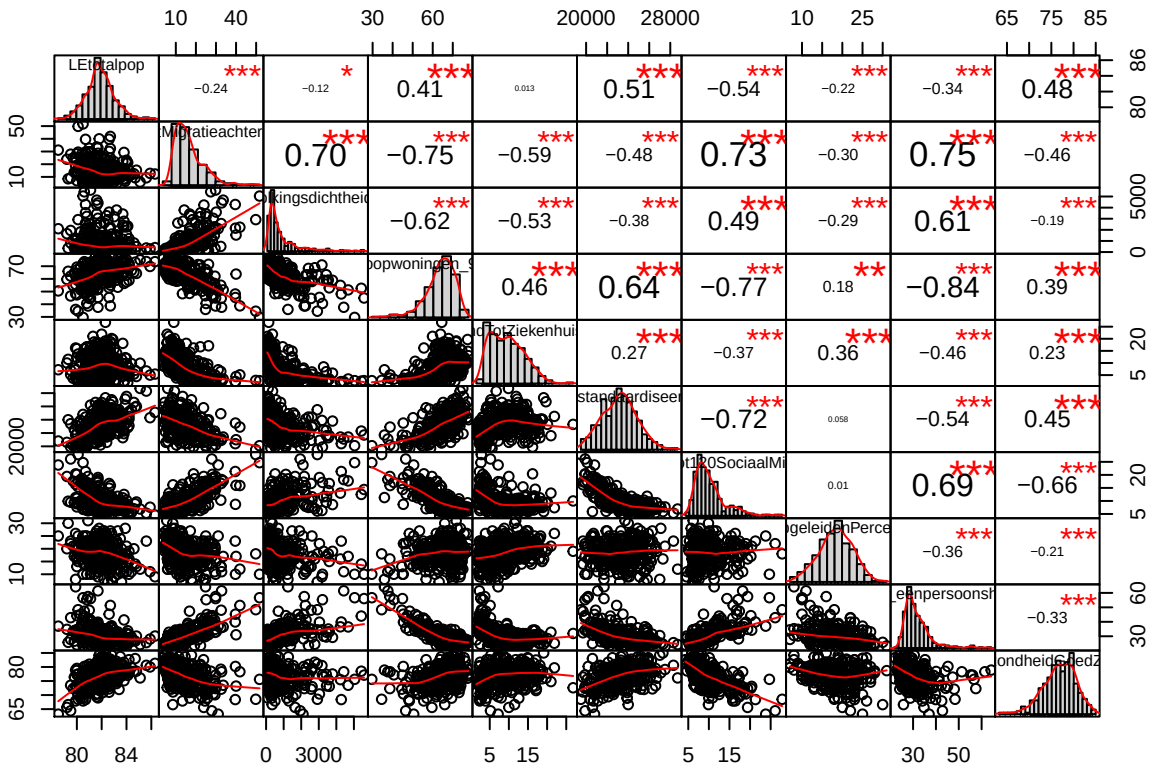
```
## Percentage_eenpersoonshuishoudens    -1.792515e-02
## ErvarenGezondheidGoedZeerGoed_1       5.508628e-02
## EenOfMeerLangdurigeAandoeningen_2     .
## NormaalGewicht_9                      .
## VoldoetAanFitnorm_14                  .
## UrenMantelzorgPerWeek_19              .
## WekelijkseSporters_16                 .
```

```
d5<- d2[complete.cases(d2),] %>%
  select( LEtotalpop, TotaalMetMigratieachtergrond_44, Bevolkingsdichtheid_57, Koopwoningen_94, Afstand
```

```
lm2 <- lm(LEtotalpop ~  ., data = d5)
summary(lm2)
```

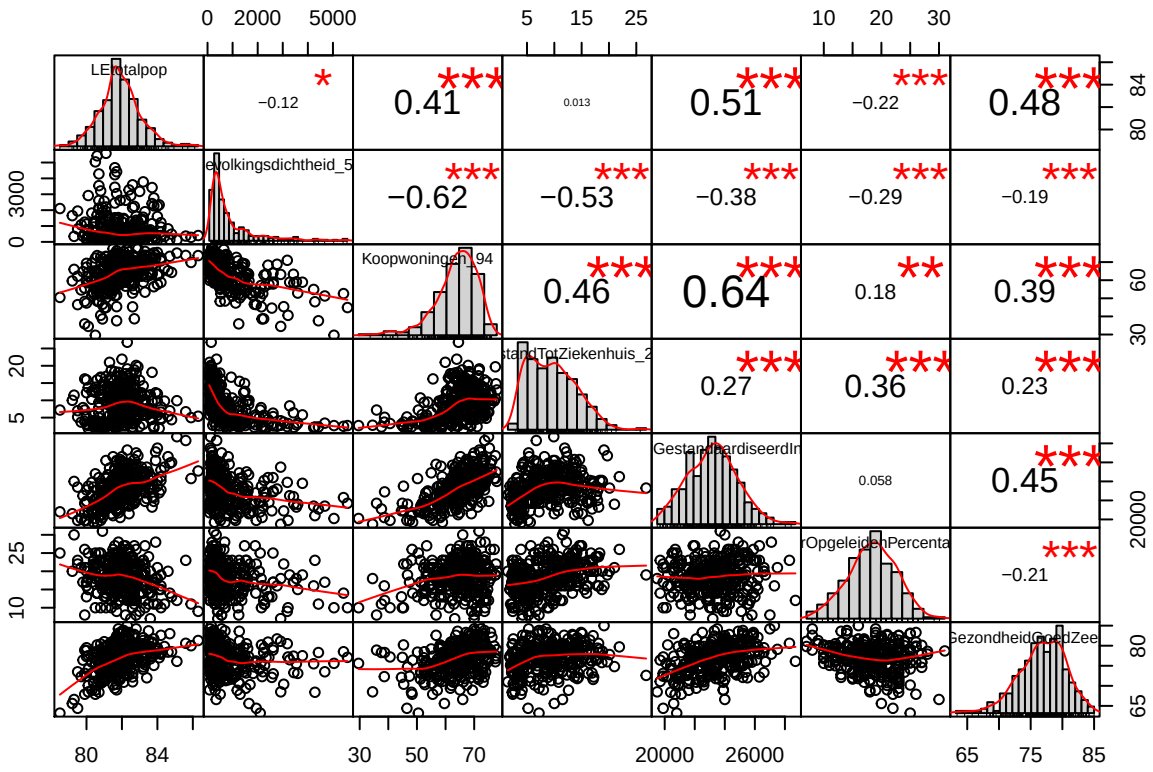
```
##
## Call:
## lm(formula = LEtotalpop ~ ., data = d5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.24707 -0.53046  0.00285  0.53156  2.86699
##
## Coefficients:
##                                   Estimate Std. Error t value
## (Intercept)                       7.431e+01  2.412e+00  30.802
## TotaalMetMigratieachtergrond_44   2.883e-02  1.330e-02   2.167
## Bevolkingsdichtheid_57            7.448e-05  7.604e-05   0.979
## Koopwoningen_94                   1.756e-02  1.376e-02   1.277
## AfstandTotZiekenhuis_216         -2.815e-02  1.360e-02  -2.069
## GemiddeldGestandaardiseerdInkomen_41  1.613e-04  4.485e-05   3.597
## InkomenTot120SociaalMinimum_13   -6.882e-02  3.091e-02  -2.226
## LagerOpgeleidenPercentage_5      -4.083e-02  1.444e-02  -2.827
## Percentage_eenpersoonshuishoudens    -3.258e-02  1.679e-02  -1.941
## ErvarenGezondheidGoedZeerGoed_1   6.518e-02  1.969e-02   3.311
##                                   Pr(>|t|)
## (Intercept)                        < 2e-16 ***
## TotaalMetMigratieachtergrond_44   0.030985 *
## Bevolkingsdichtheid_57            0.328121
## Koopwoningen_94                   0.202705
## AfstandTotZiekenhuis_216          0.039402 *
## GemiddeldGestandaardiseerdInkomen_41 0.000375 ***
## InkomenTot120SociaalMinimum_13    0.026716 *
## LagerOpgeleidenPercentage_5       0.005012 **
## Percentage_eenpersoonshuishoudens    0.053155 .
## ErvarenGezondheidGoedZeerGoed_1   0.001044 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8732 on 303 degrees of freedom
## Multiple R-squared:  0.4351, Adjusted R-squared:  0.4183
## F-statistic: 25.93 on 9 and 303 DF,  p-value: < 2.2e-16
```

```
suppressWarnings(chart.Correlation(d5))
```



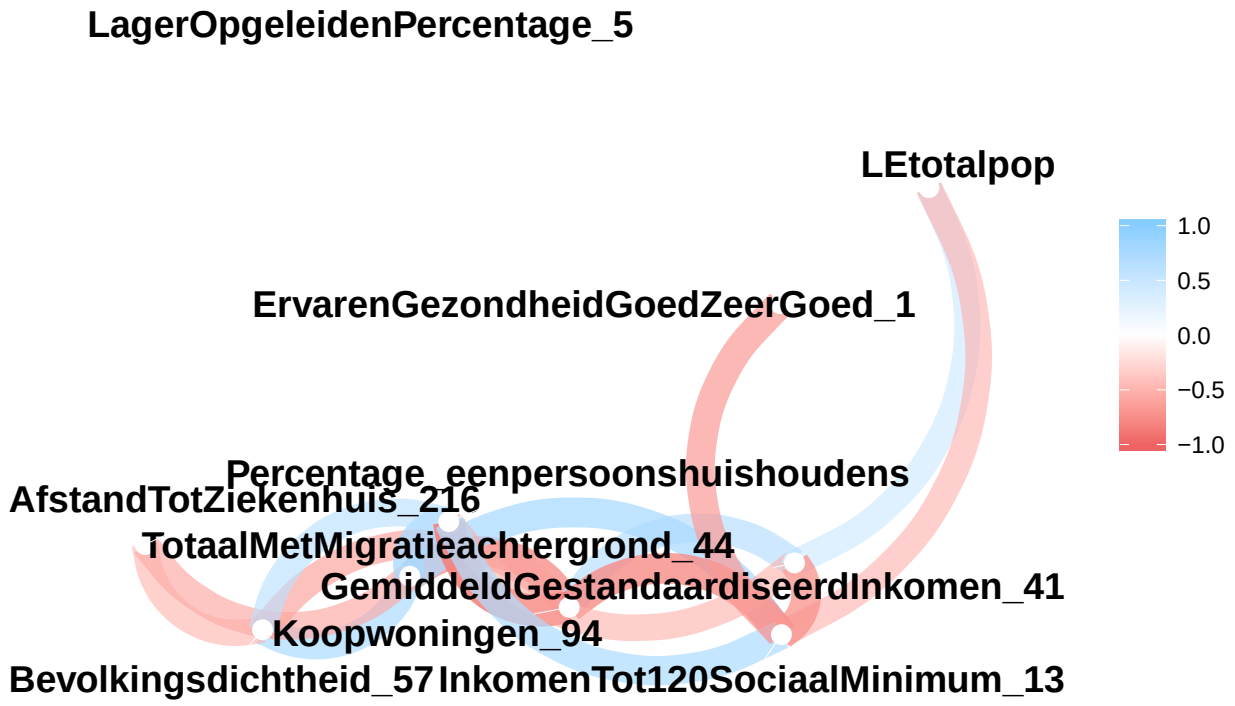

```
d6 <- d5 %>%
  select(-TotaalMetMigratieachtergrond_44,-InkomenTot120SociaalMinimum_13, -Percentage_eenpersoonshuish
```

```
suppressWarnings(chart.Correlation(d6))
```




```
d5 %>%
  correlate() %>%
  network_plot(min_cor = .5, repel= TRUE, curved = TRUE, legend = TRUE)
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```




```
lm3 <- lm(LEtotalpop ~  ., data = d6)
summary(lm3)
```

```
##
## Call:
## lm(formula = LEtotalpop ~ ., data = d6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16178 -0.58208  0.00957  0.53705  2.84655
##
## Coefficients:
##                                    Estimate Std. Error t value
## (Intercept)                       6.964e+01  1.269e+00  54.879
## Bevolkingsdichtheid_57            1.077e-04  7.068e-05   1.524
## Koopwoningen_94                   3.793e-02  9.717e-03   3.904
## AfstandTotZiekenhuis_216         -3.534e-02  1.341e-02  -2.636
## GemiddeldGestandaardiseerdInkomen_41  2.109e-04  4.056e-05   5.199
## LagerOpgeleidenPercentage_5      -4.083e-02  1.297e-02  -3.149
## ErvarenGezondheidGoedZeerGoed_1   7.790e-02  1.646e-02   4.734
##                                   Pr(>|t|)
## (Intercept)                        < 2e-16 ***
## Bevolkingsdichtheid_57            0.128508
## Koopwoningen_94                   0.000117 ***
## AfstandTotZiekenhuis_216          0.008824 **
## GemiddeldGestandaardiseerdInkomen_41 3.67e-07 ***
```

```
## LagerOpgeleidenPercentage_5          0.001802 **
## ErvarenGezondheidGoedZeerGoed_1      3.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8853 on 306 degrees of freedom
## Multiple R-squared:  0.4135, Adjusted R-squared:  0.402
## F-statistic: 35.95 on 6 and 306 DF,  p-value: < 2.2e-16
```

```
d6 %>%
  correlate() %>%
  network_plot(min_cor = .5, repel= TRUE, curved = TRUE, legend = TRUE)
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```