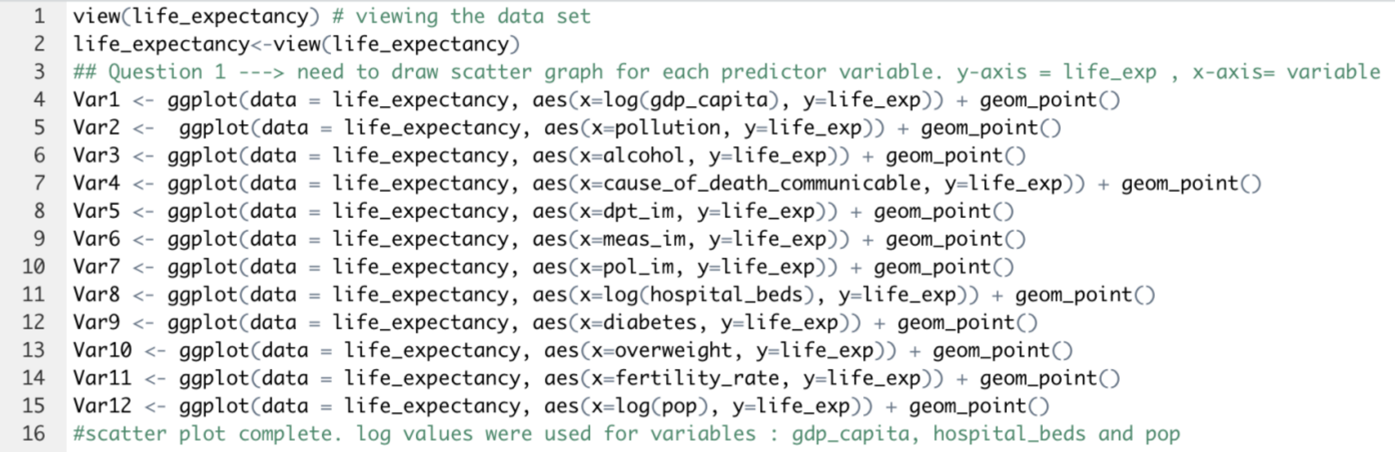
BS2004: What affects life expectancy throughout the world?

This report will look at different variables that affect life expectancy. The report analysis dataset from 157 countries. The dataset was obtained from the World Bank (life_expectancy_data.csv). In this analysis, the response variable is life expectancy(life_exp) and there are 12 predictor variables affecting the response variable.

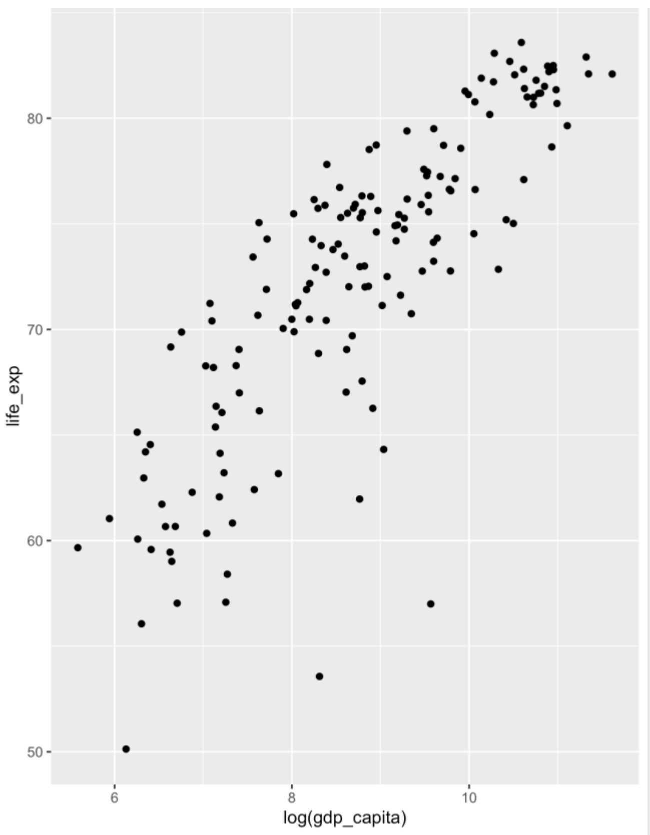
Question 1:

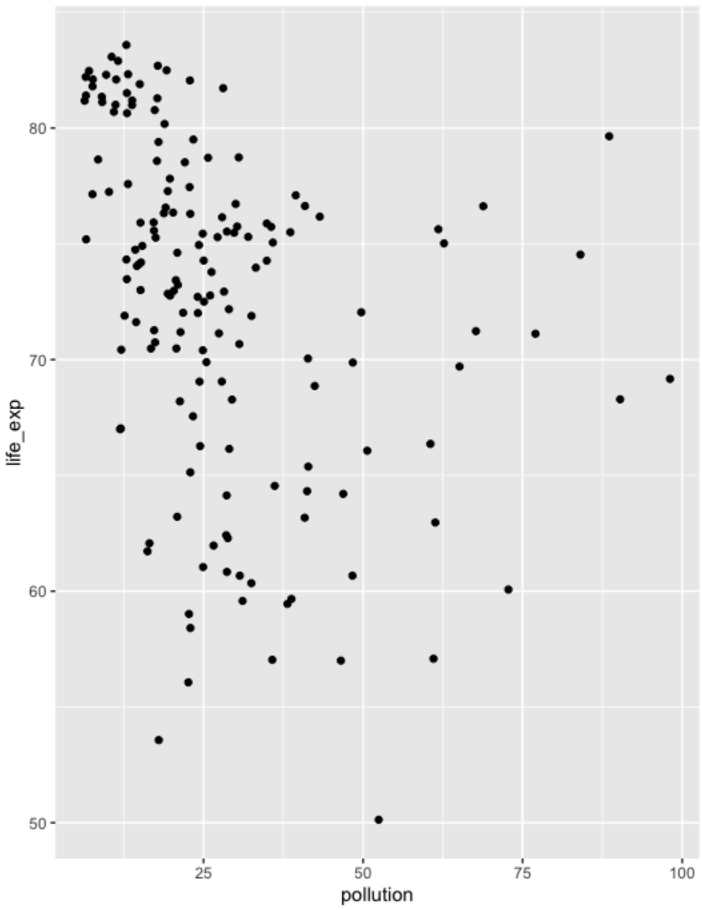


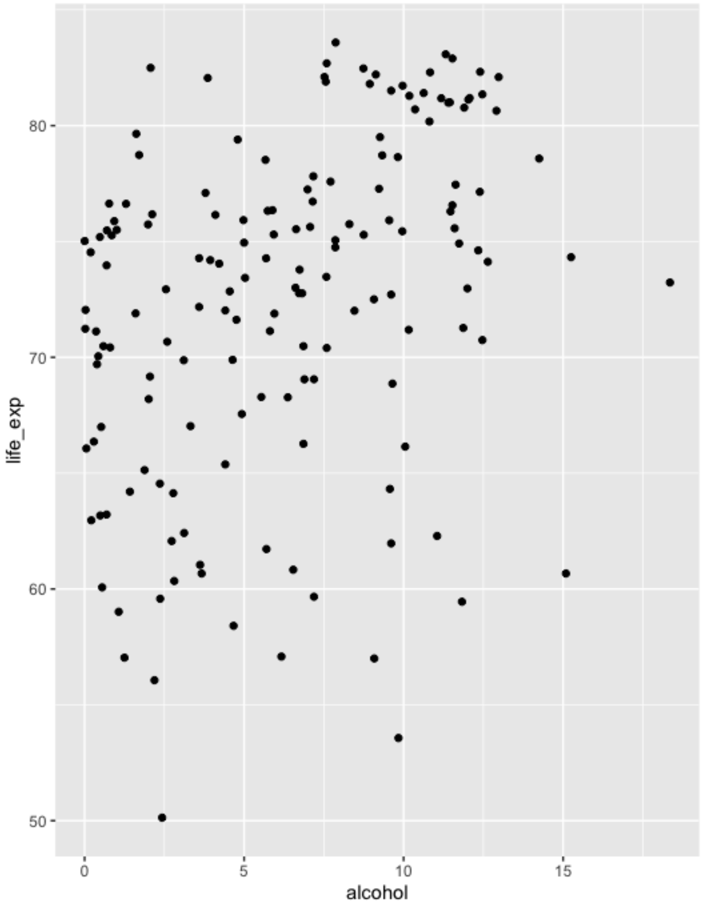
```
view(life_expectancy) # viewing the data set
life_expectancy<-view(life_expectancy)
## Question 1 ---> need to draw scatter graph for each predictor variable. y-axis = life_exp , x-axis= variable
Var1 <- ggplot(data = life_expectancy, aes(x=log(gdp_capita), y=life_exp)) + geom_point()
Var2 <-  ggplot(data = life_expectancy, aes(x=pollution, y=life_exp)) + geom_point()
Var3 <- ggplot(data = life_expectancy, aes(x=alcohol, y=life_exp)) + geom_point()
Var4 <- ggplot(data = life_expectancy, aes(x=cause_of_death_communicable, y=life_exp)) + geom_point()
Var5 <- ggplot(data = life_expectancy, aes(x=dpt_im, y=life_exp)) + geom_point()
Var6 <- ggplot(data = life_expectancy, aes(x=meas_im, y=life_exp)) + geom_point()
Var7 <- ggplot(data = life_expectancy, aes(x=pol_im, y=life_exp)) + geom_point()
Var8 <- ggplot(data = life_expectancy, aes(x=log(hospital_beds), y=life_exp)) + geom_point()
Var9 <- ggplot(data = life_expectancy, aes(x=diabetes, y=life_exp)) + geom_point()
Var10 <- ggplot(data = life_expectancy, aes(x=overweight, y=life_exp)) + geom_point()
Var11 <- ggplot(data = life_expectancy, aes(x=fertility_rate, y=life_exp)) + geom_point()
Var12 <- ggplot(data = life_expectancy, aes(x=log(pop), y=life_exp)) + geom_point()
#scatter plot complete. log values were used for variables : gdp_capita, hospital_beds and pop
```

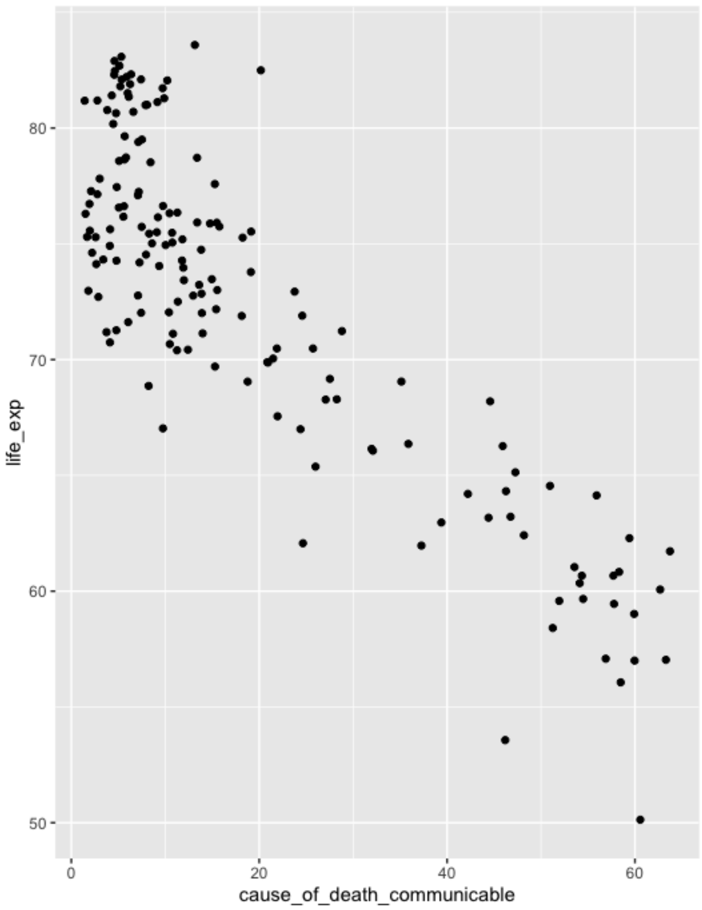
Snippet code 1. Codes used to generate different scatterplots for each predictor variable.

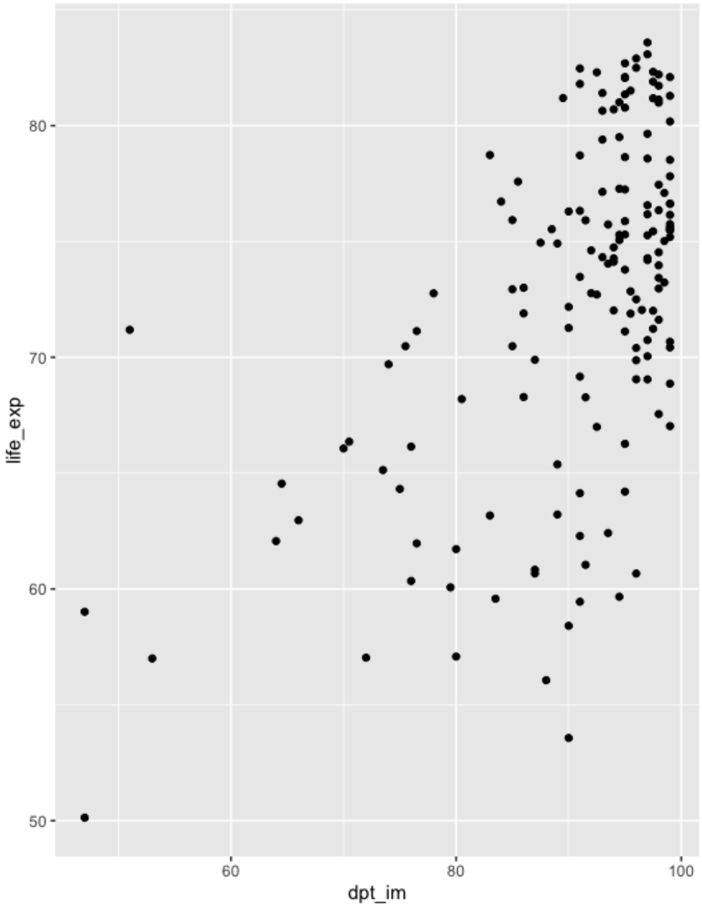
To generate scatterplots of life expectancy against each of the 12 predictor variables. I used the codes shown above in snippet code 1. Each plot was labelled individually, this is important when trying to group all the plots into one figure as will be shown later.
For the variables: gdp_capita, hospital_beds and pop, log values were used as they didn’t show a linear relationship with life_exp. The individual plots are below:

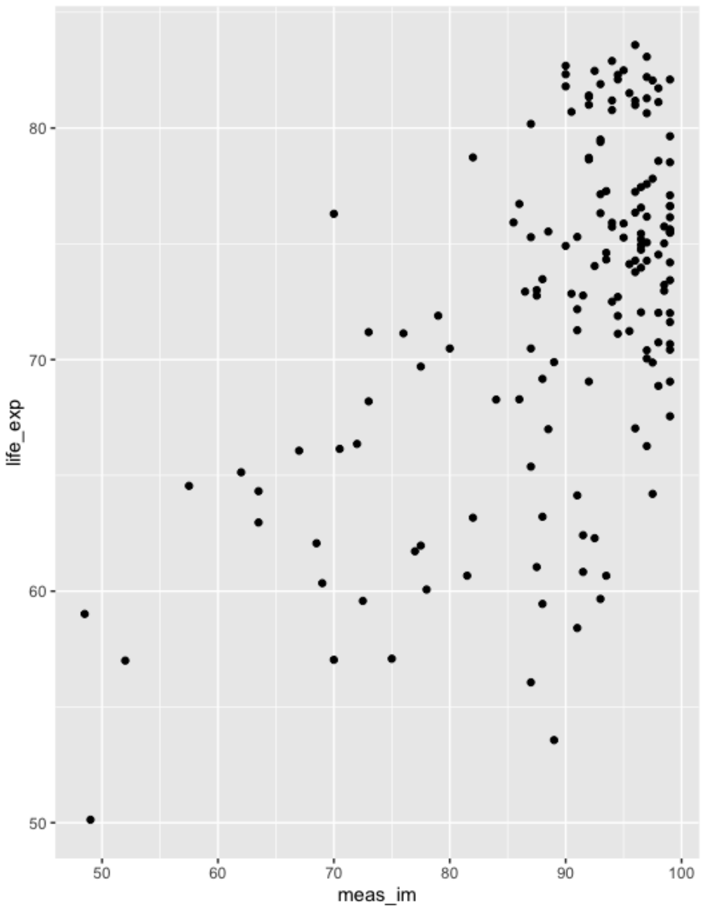

80
70
60
50
life_exp
6
8
10
log(gdp_capita)

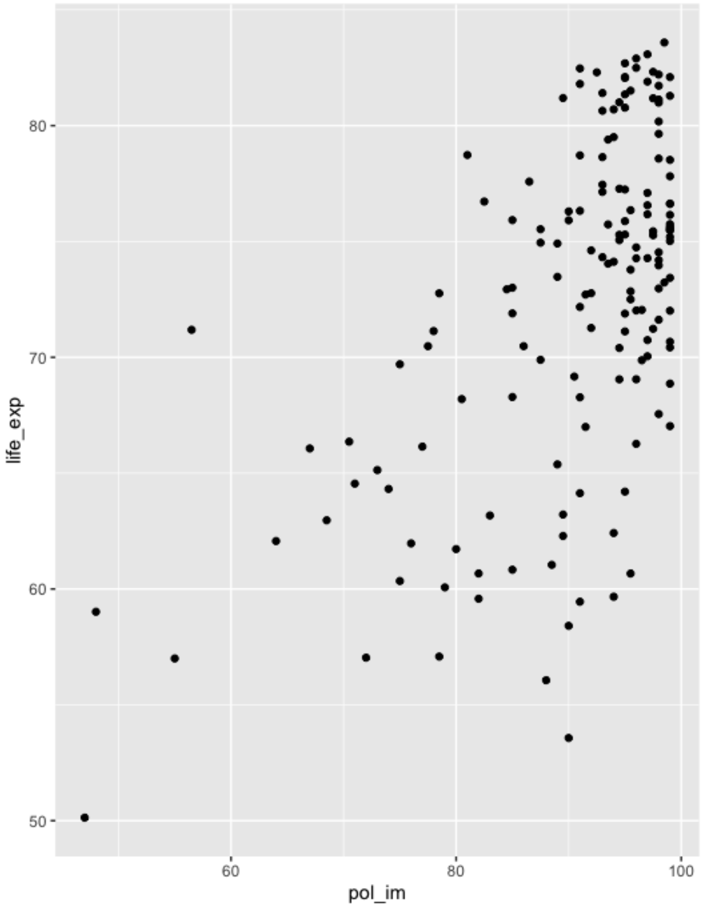

life_exp
80
70
60
50
25
50
75
100
pollution

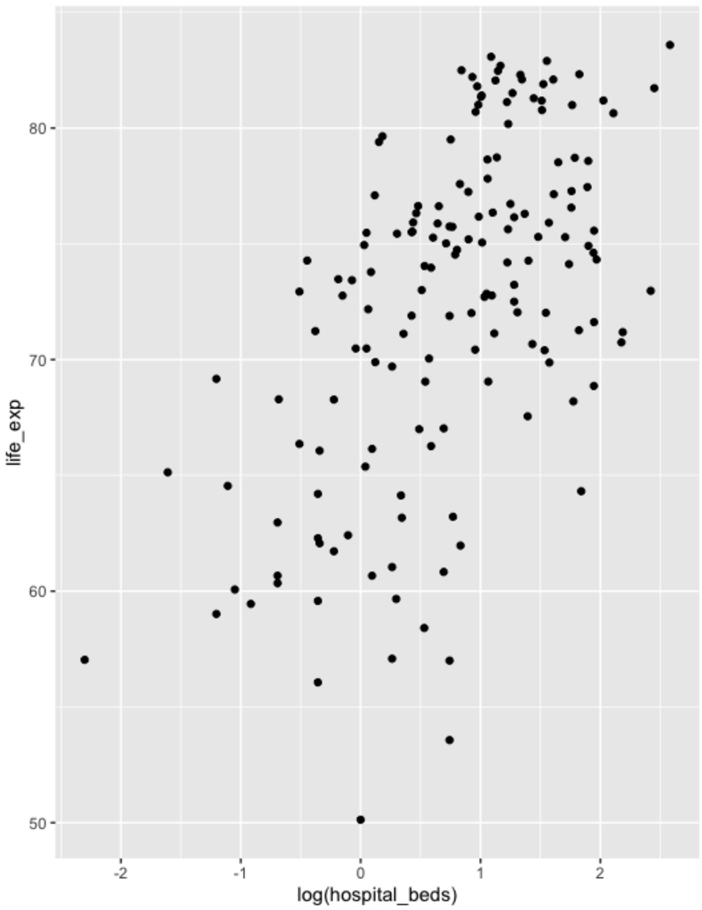

life_exp
80
70
60
50
0
5
10
15
alcohol


life_exp
80
70
60
50
0
20
40
60
cause_of_death_communicable

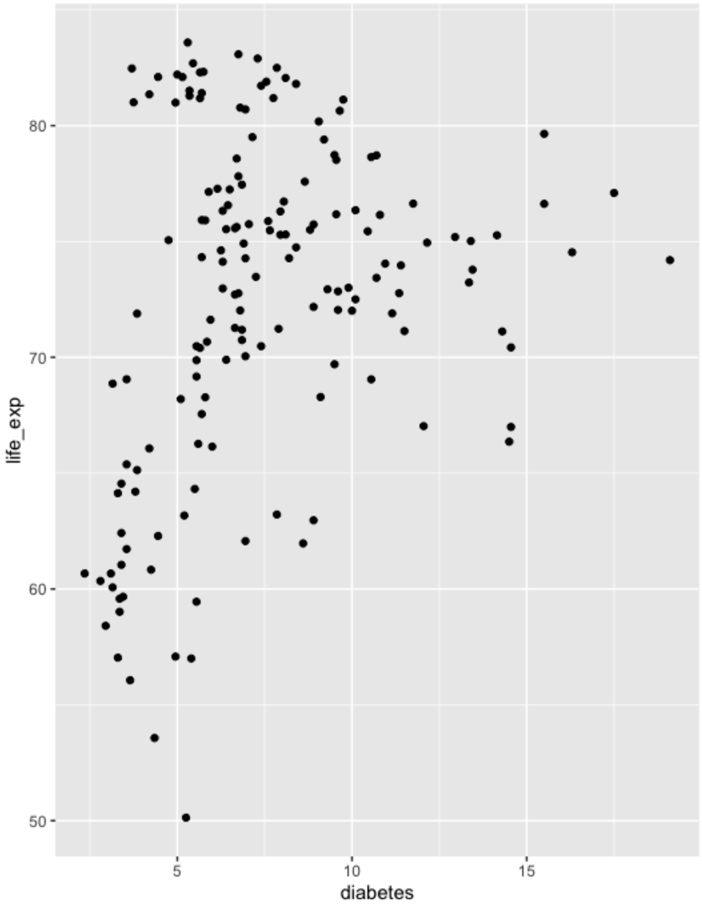

life_exp
80
70
60
50
60
80
100
dpt_im


life_exp
80
70
60
50
50
60
70
80
90
100
meas_im

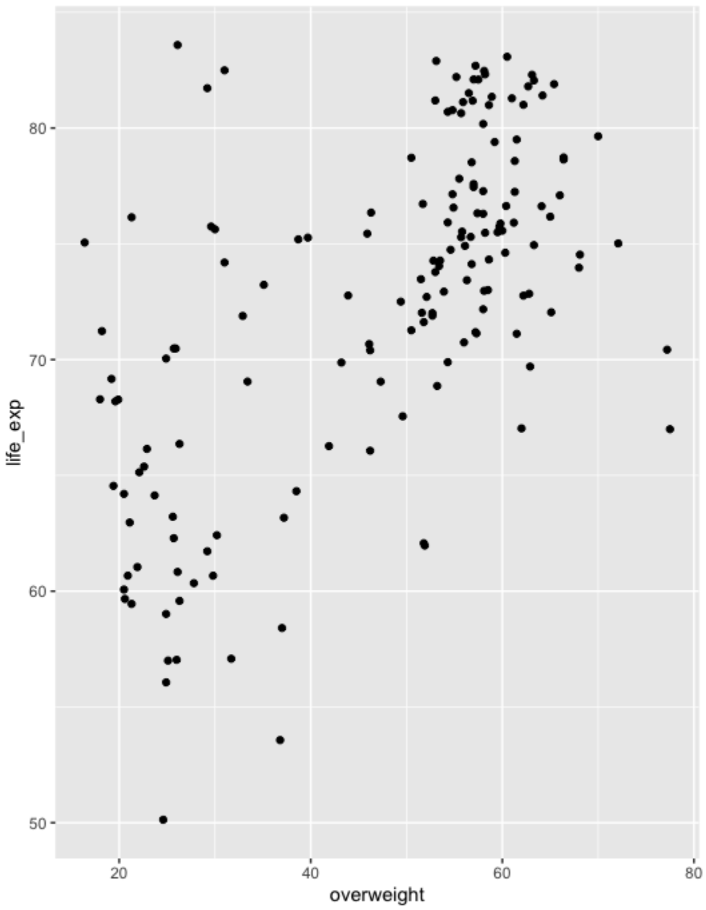

life_exp
80
70
60
50
60
80
100
pol_im



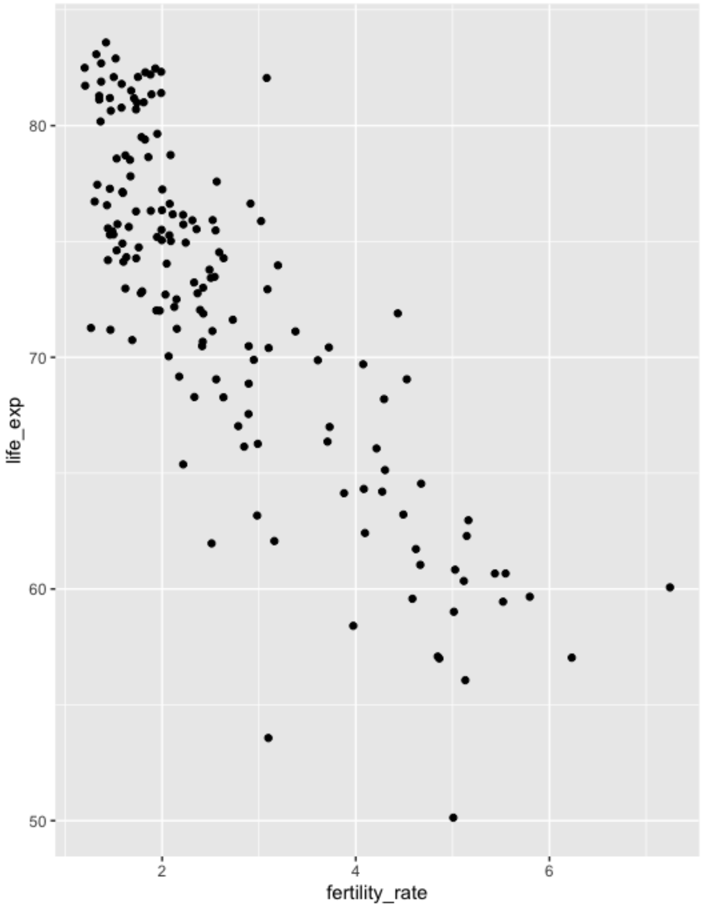
life_exp
80
70
60
50
-2
-1
0
1
2
log(hospital_beds)



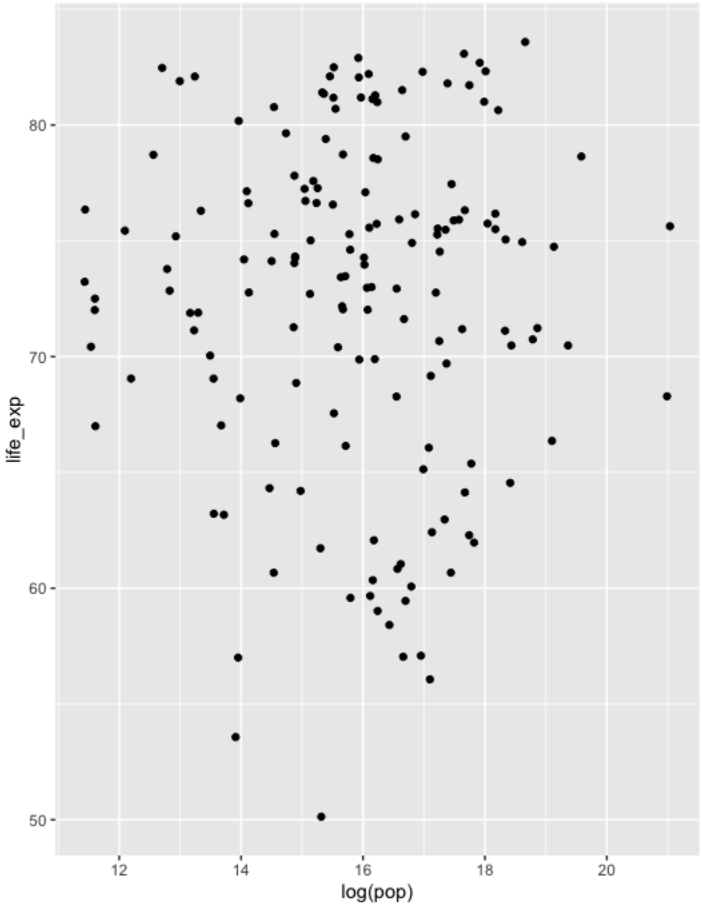
life_exp
80
70
60
50
5
10
15
diabetes

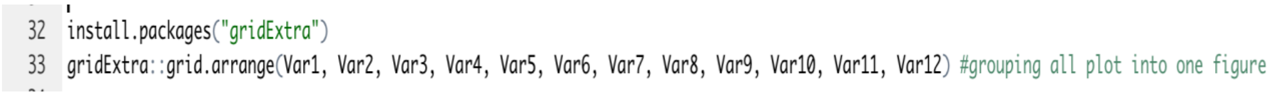


life_exp
80
70
60
50
20
40
60
80
overweight



life_exp
80
70
60
50
2
4
6
fertility_rate





```
install.packages("gridExtra")
gridExtra::grid.arrange(Var1, Var2, Var3, Var4, Var5, Var6, Var7, Var8, Var9, Var10, Var11, Var12) #grouping all plot into one figure
```

Snippet code 2. shows the ode used to arrange the plots into one figure.

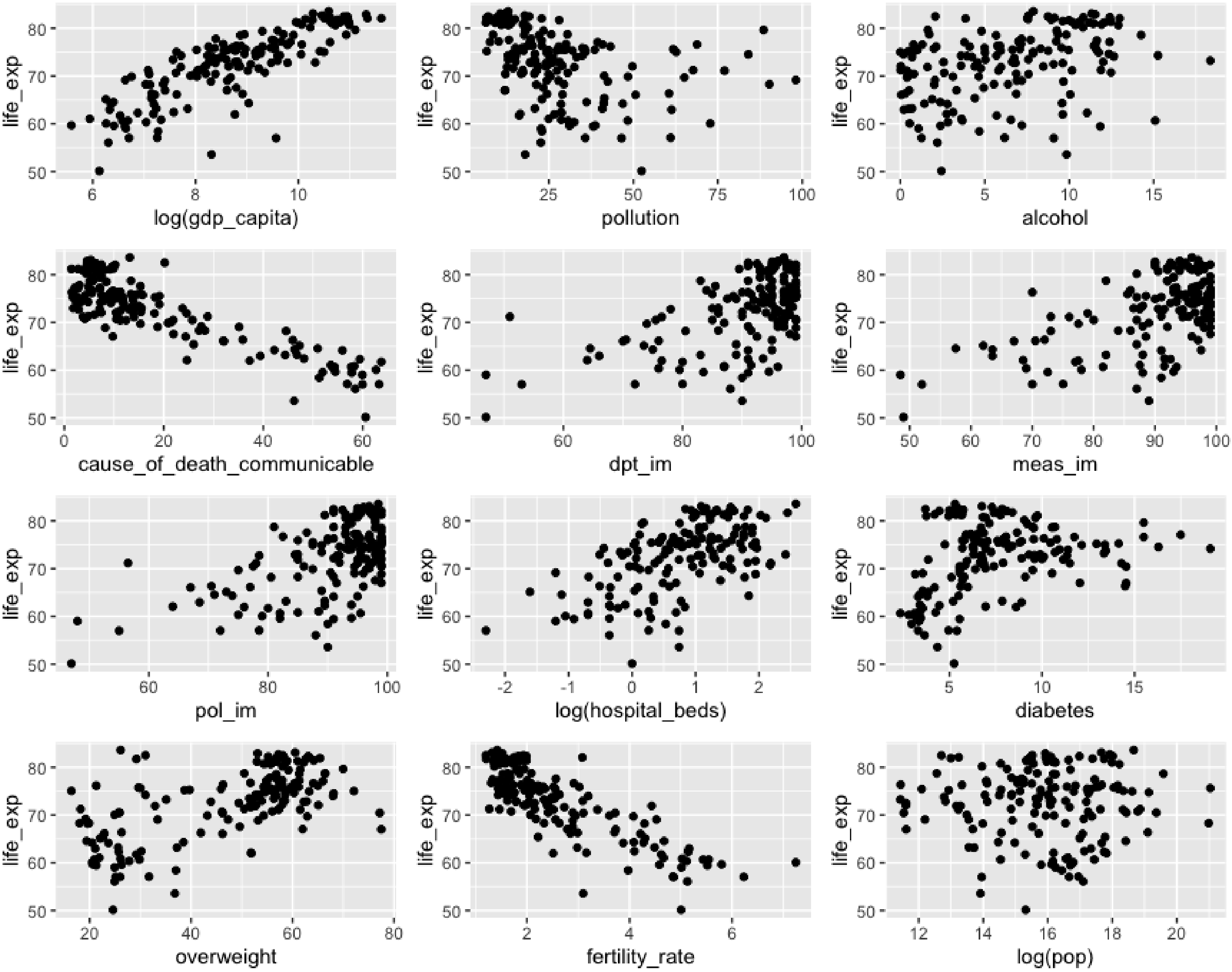


Figure 1. grouped image of individual predictor variables affecting life expectancy. Predictor variables: gdp_capita, hospital_beds and pop have been logged.

After generating the individual graphs for each predictor variable, I installed the package ‘gridExtra’and used the code shown in snippet code 2 to combine all of the individual plots into a single figure. In figure 1, Life expectancy (life_exp) is plotted on the y-axis whilst the predictor variables are plotted on the x-axis. The dataset for gdp_capita, hospital_beds and pop were logged to provide a linear relationship between the predictor variables and life expectancy.

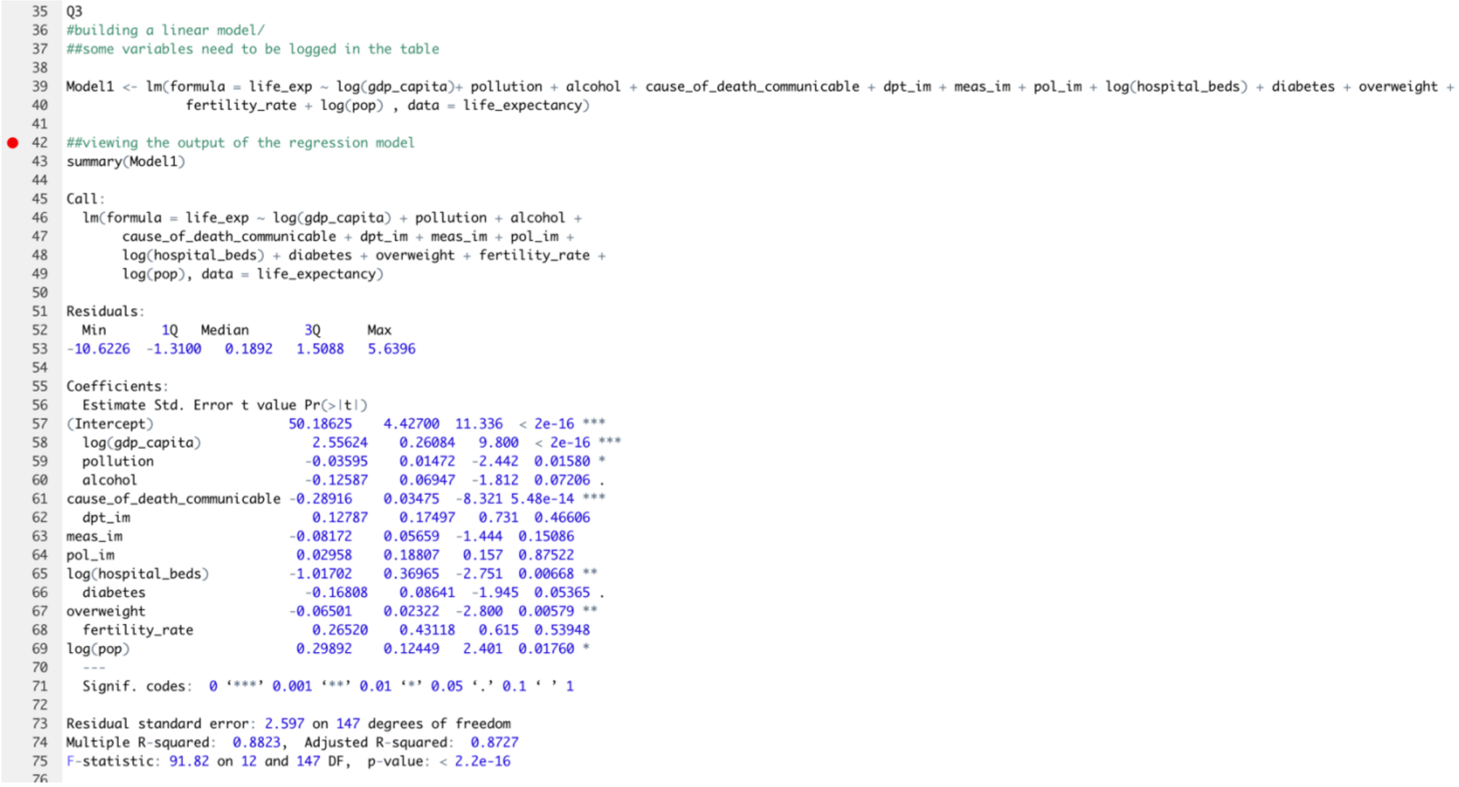
Q2.

From figure 1 we can see that there is an increasing linear relationship between life expectancy and log(gdp_capita) and a positive correlation. This could be due to the increased availability of modern sanitation and medicine in poorer countries which would increase their life expectancy.
There is a decreasing linear relationship and a negative correlation between life expectancy and fertility_rate. A reason for this could be that those countries with developed economies

such as those in Europe and Asia usually have a higher life expectancy and a lower fertility rate compared to countries with developing economies such as Africa.

The number of hospital beds per 1000 people shows an increasing linear relationship with life expectancy and a positive co-relation.
This could be that with better health care people can be admitted to hospitals and treated for their disease which will increase life expectancy.

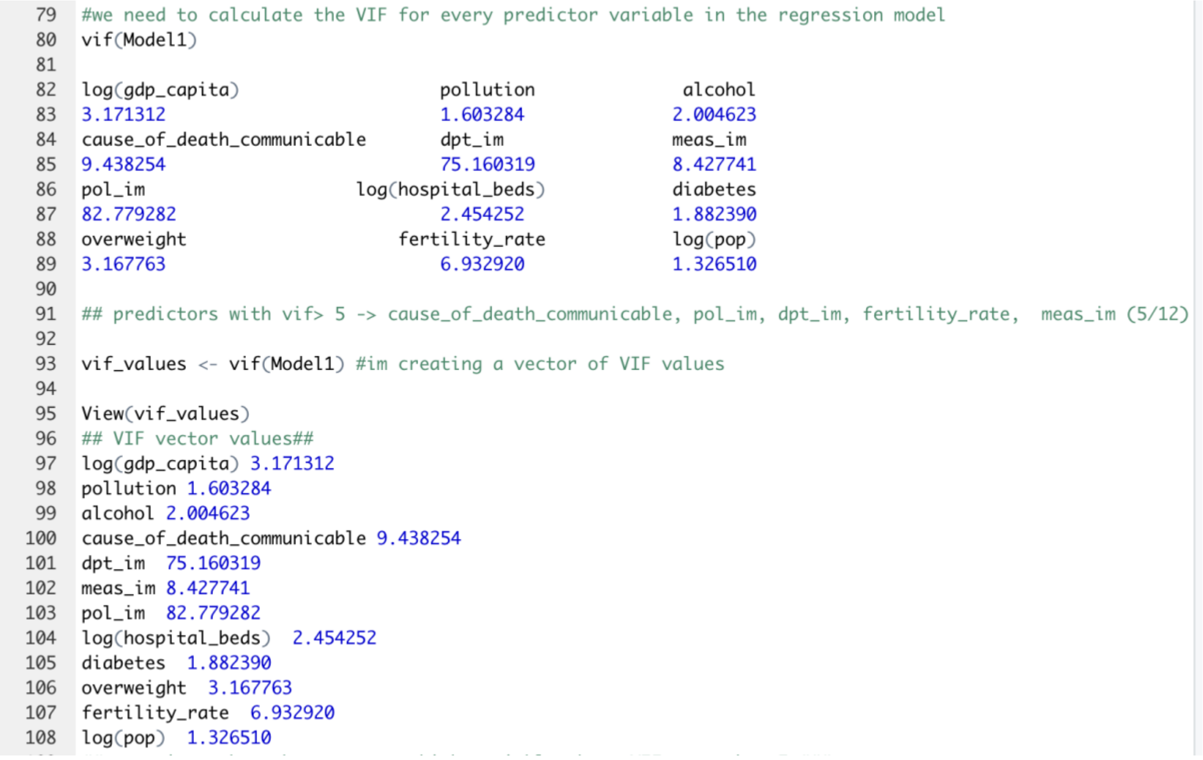
Q3.



```
35  Q3
36  #building a linear model/
37  ##some variables need to be logged in the table
38
39  Model1 <- lm(formula = life_exp ~ log(gdp_capita)+ pollution + alcohol + cause_of_death_communicable + dpt_im + meas_im + pol_im + log(hospital_beds) + diabetes + overweight +
40                fertility_rate + log(pop) , data = life_expectancy)
41
42  ##viewing the output of the regression model
43  summary(Model1)
44
45  Call:
46    lm(formula = life_exp ~ log(gdp_capita) + pollution + alcohol +
47         cause_of_death_communicable + dpt_im + meas_im + pol_im +
48         log(hospital_beds) + diabetes + overweight + fertility_rate +
49         log(pop), data = life_expectancy)
50
51  Residuals:
52    Min       1Q   Median       3Q      Max
53  -10.6226  -1.3100   0.1892   1.5088   5.6396
54
55  Coefficients:
56    Estimate Std. Error t value Pr(>|t|)
57  (Intercept)                 50.18625    4.42700  11.336  < 2e-16 ***
58    log(gdp_capita)              2.55624    0.26084   9.800  < 2e-16 ***
59    pollution                   -0.03595    0.01472  -2.442  0.01580 *
60    alcohol                     -0.12587    0.06947  -1.812  0.07206 .
61  cause_of_death_communicable -0.28916    0.03475  -8.321 5.48e-14 ***
62    dpt_im                       0.12787    0.17497   0.731  0.46606
63  meas_im                     -0.08172    0.05659  -1.444  0.15086
64  pol_im                       0.02958    0.18807   0.157  0.87522
65  log(hospital_beds)          -1.01702    0.36965  -2.751  0.00668 **
66    diabetes                    -0.16808    0.08641  -1.945  0.05365 .
67  overweight                  -0.06501    0.02322  -2.800  0.00579 **
68    fertility_rate               0.26520    0.43118   0.615  0.53948
69  log(pop)                     0.29892    0.12449   2.401  0.01760 *
70    ---
71    Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
72
73  Residual standard error: 2.597 on 147 degrees of freedom
74  Multiple R-squared:  0.8823,  Adjusted R-squared:  0.8727
75  F-statistic: 91.82 on 12 and 147 DF,  p-value: < 2.2e-16
76
```

Snippet code 3. A linear model was built with all 12 predicted variables.

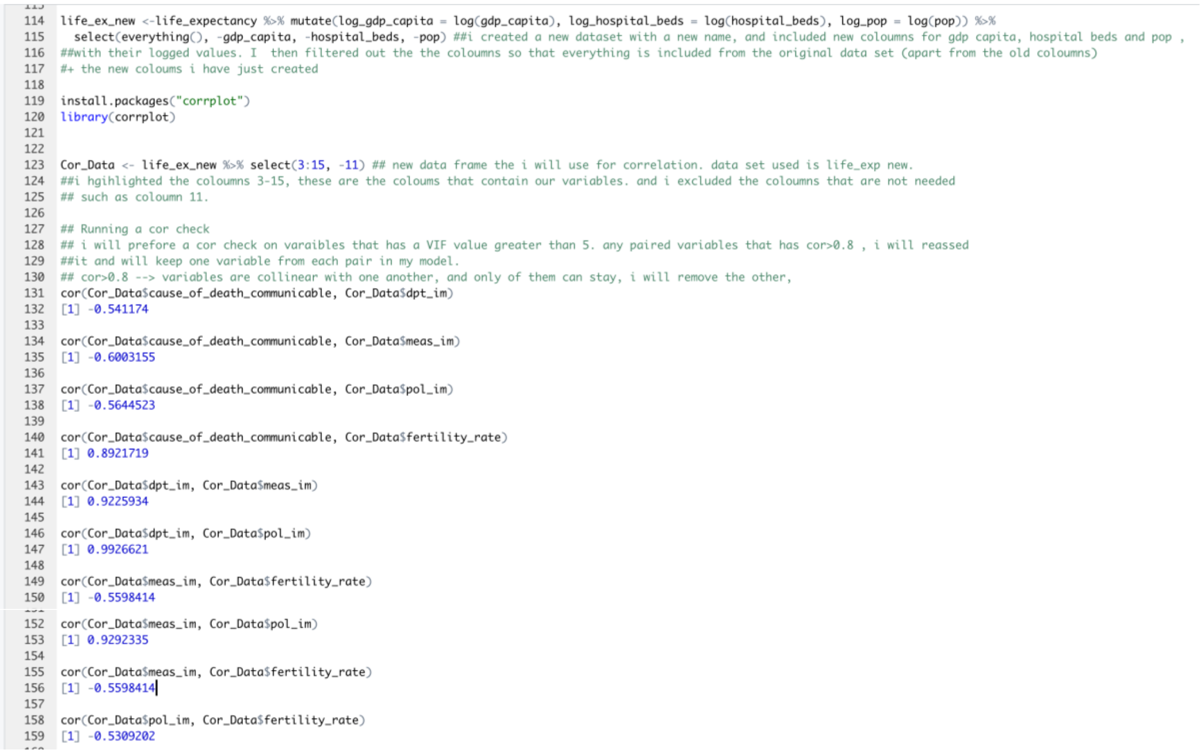
We need to check on what affects life expectancy throughout the world by creating a full linear model (Model1). This regression model will check for collinearity between life expectancy and the 12 variable predictors. The residuals of the model are shown in the figure on lines 55-75. We want to check if there is multicollinearity between the 12 variables and the strength of that correlation using the variance inflation factor (VIF). The VIF will measure the correlation between all 12 predictor variables in the model. If the VIF value is:
VIF=1, there is no collinearity
1<VIF<5 , there is moderate collinearity,
VIF≥ 5, then there is considerable collinearity.



```
#we need to calculate the VIF for every predictor variable in the regression model
vif(Model1)

log(gdp_capita)                    pollution              alcohol
3.171312                           1.603284              2.004623
cause_of_death_communicable        dpt_im                meas_im
9.438254                           75.160319             8.427741
pol_im                     log(hospital_beds)            diabetes
82.779282                          2.454252              1.882390
overweight                     fertility_rate            log(pop)
3.167763                           6.932920              1.326510

## predictors with vif> 5 -> cause_of_death_communicable, pol_im, dpt_im, fertility_rate,  meas_im (5/12)

vif_values <- vif(Model1) #im creating a vector of VIF values

View(vif_values)
## VIF vector values##
log(gdp_capita) 3.171312
pollution 1.603284
alcohol 2.004623
cause_of_death_communicable 9.438254
dpt_im  75.160319
meas_im 8.427741
pol_im  82.779282
log(hospital_beds)  2.454252
diabetes  1.882390
overweight  3.167763
fertility_rate  6.932920
log(pop)  1.326510
```

Snippet code 4.

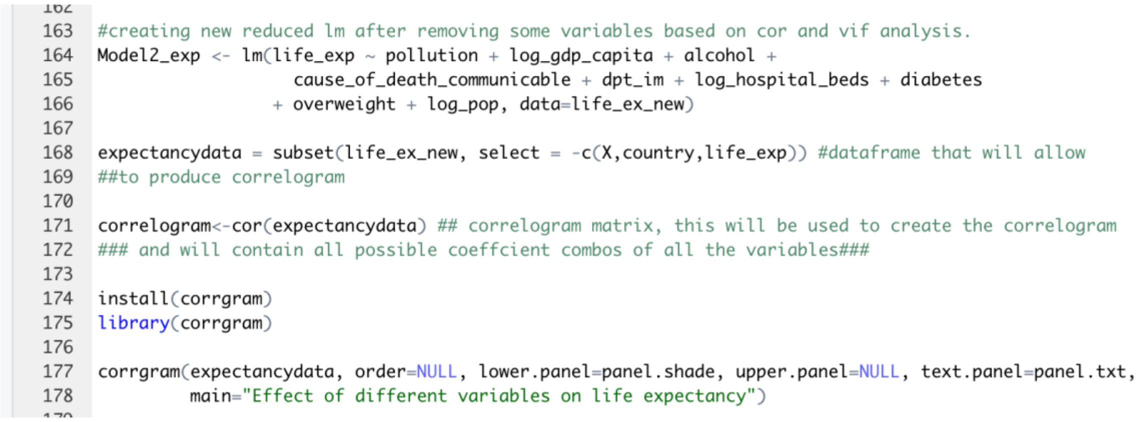
The vif() function was used for each of the 12 individual variables. Those that showed vif>5 were: cause_of_death_communicable, pol_im , dpt_im , fertility_rate, and meas_im.



```
life_ex_new <-life_expectancy %>% mutate(log_gdp_capita = log(gdp_capita), log_hospital_beds = log(hospital_beds), log_pop = log(pop)) %>%
  select(everything(), -gdp_capita, -hospital_beds, -pop) ##i created a new dataset with a new name, and included new coloumns for gdp capita, hospital beds and pop ,
##with their logged values. I  then filtered out the the coloumns so that everything is included from the original data set (apart from the old coloumns)
#+ the new coloums i have just created

install.packages("corrplot")
library(corrplot)


Cor_Data <- life_ex_new %>% select(3:15, -11) ## new data frame the i will use for correlation. data set used is life_exp new.
##i hgihlighted the coloumns 3-15, these are the coloums that contain our variables. and i excluded the coloumns that are not needed
## such as coloumn 11.

## Running a cor check
## i will prefore a cor check on varaibles that has a VIF value greater than 5. any paired variables that has cor>0.8 , i will reassed
##it and will keep one variable from each pair in my model.
## cor>0.8 --> variables are collinear with one another, and only of them can stay, i will remove the other,
cor(Cor_Data$cause_of_death_communicable, Cor_Data$dpt_im)
[1] -0.541174

cor(Cor_Data$cause_of_death_communicable, Cor_Data$meas_im)
[1] -0.6003155

cor(Cor_Data$cause_of_death_communicable, Cor_Data$pol_im)
[1] -0.5644523

cor(Cor_Data$cause_of_death_communicable, Cor_Data$fertility_rate)
[1] 0.8921719

cor(Cor_Data$dpt_im, Cor_Data$meas_im)
[1] 0.9225934

cor(Cor_Data$dpt_im, Cor_Data$pol_im)
[1] 0.9926621

cor(Cor_Data$meas_im, Cor_Data$fertility_rate)
[1] -0.5598414

cor(Cor_Data$meas_im, Cor_Data$pol_im)
[1] 0.9292335

cor(Cor_Data$meas_im, Cor_Data$fertility_rate)
[1] -0.5598414

cor(Cor_Data$pol_im, Cor_Data$fertility_rate)
[1] -0.5309202
```

Snippet code 5.

A new data set with logged columns called ‘Cor_Data’ was created, this is to not cause any confusion as well as to ease the process of collinear predictors. A Cor check was performed and variables with Cor> 0.8 were identified. Based on the collinear results, a new reduced data model was created with removed variables.



```
#creating new reduced lm after removing some variables based on cor and vif analysis.
Model2_exp <- lm(life_exp ~ pollution + log_gdp_capita + alcohol +
                   cause_of_death_communicable + dpt_im + log_hospital_beds + diabetes
                 + overweight + log_pop, data=life_ex_new)

expectancydata = subset(life_ex_new, select = -c(X,country,life_exp)) #dataframe that will allow
##to produce correlogram

correlogram<-cor(expectancydata) ## correlogram matrix, this will be used to create the correlogram
### and will contain all possible coeffcient combos of all the variables###

install(corrgram)
library(corrgram)

corrgram(expectancydata, order=NULL, lower.panel=panel.shade, upper.panel=NULL, text.panel=panel.txt,
         main="Effect of different variables on life expectancy")
```

Snippet 6.

A reduced linear model was created with removed variables based on the analysis from the VIF and Cor. A correlogram was created using the ‘corrgram’ package.

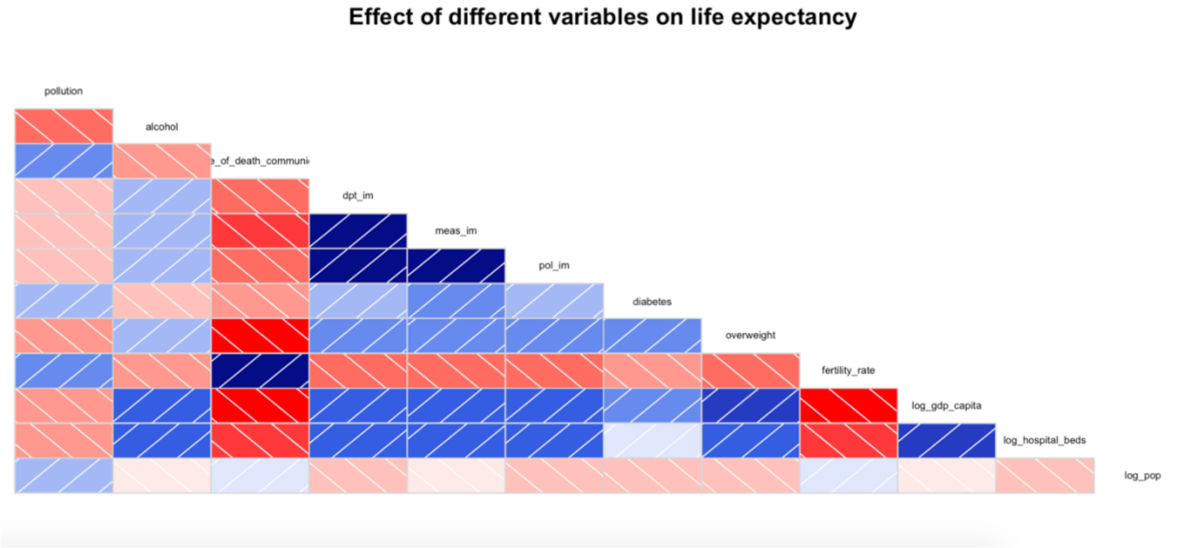
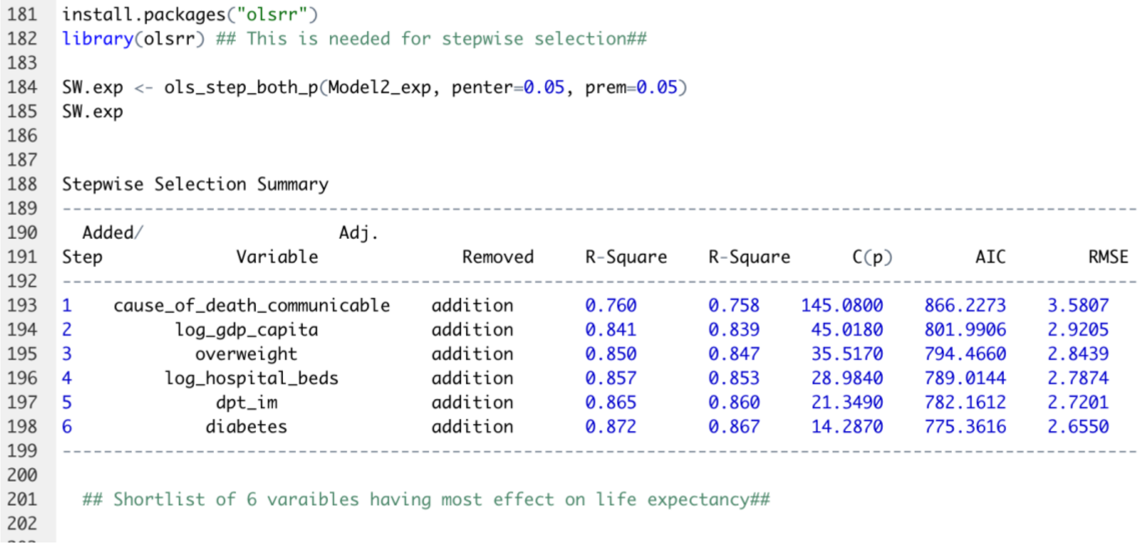


Figure 2: A correlogram showing 12 variable predictors.

Note: I generated the correlogram with the 12 variables, but I struggled with the labels as some were too long and I didn’t know how to resolve this.

Q4.

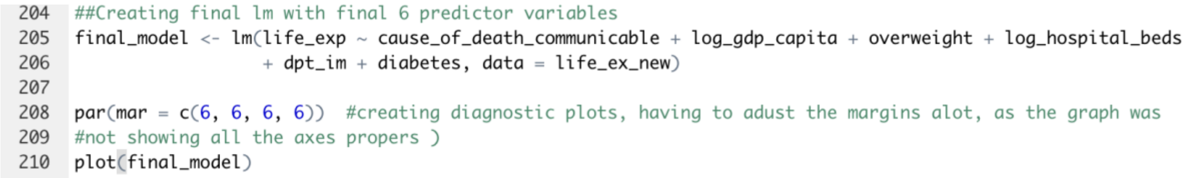


```
install.packages("olsrr")
library(olsrr) ## This is needed for stepwise selection##

SW.exp <- ols_step_both_p(Model2_exp, penter=0.05, prem=0.05)
SW.exp


Stepwise Selection Summary
-----------------------------------------------------------------------------------------------------------
  Added/                   Adj.
Step              Variable              Removed     R-Square     R-Square       C(p)        AIC         RMSE
-----------------------------------------------------------------------------------------------------------
1     cause_of_death_communicable    addition       0.760        0.758       145.0800    866.2273    3.5807
2           log_gdp_capita           addition       0.841        0.839        45.0180    801.9906    2.9205
3             overweight             addition       0.850        0.847        35.5170    794.4660    2.8439
4          log_hospital_beds         addition       0.857        0.853        28.9840    789.0144    2.7874
5               dpt_im               addition       0.865        0.860        21.3490    782.1612    2.7201
6              diabetes              addition       0.872        0.867        14.2870    775.3616    2.6550
-----------------------------------------------------------------------------------------------------------

  ## Shortlist of 6 varaibles having most effect on life expectancy##

```

Snippet 7.

Model selection for the new reduced model with removed variables. The stepwise selection was used based on p (0.05) which produce a table of the summary.

Q5.



```
##Creating final lm with final 6 predictor variables
final_model <- lm(life_exp ~ cause_of_death_communicable + log_gdp_capita + overweight + log_hospital_beds
                  + dpt_im + diabetes, data = life_ex_new)

par(mar = c(6, 6, 6, 6))  #creating diagnostic plots, having to adust the margins alot, as the graph was
#not showing all the axes propers )
plot(final_model)
```

Snippet 8.

Creating a diagnostic plot of the final model.

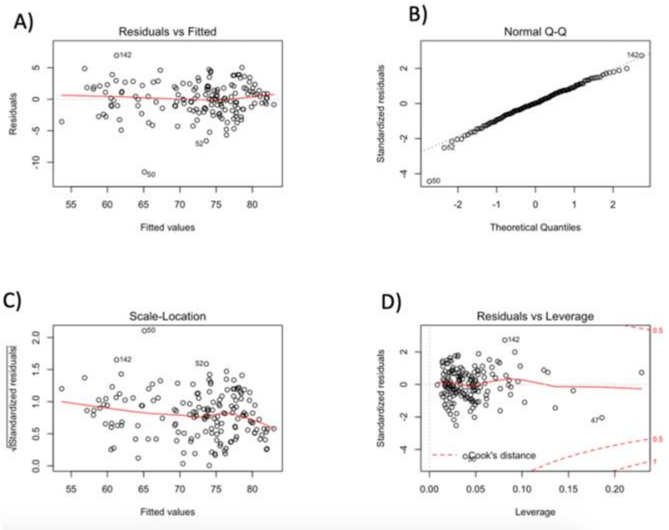


Figure 3. Diagnostic plots for the final model.

Observations of the diagnostic plots are as follow:

A) There is a spread of data across the horizontal line, however, the data is showing a negative skewness. The data is concentrated more on the left side. Thus, the data doesn't fit normality or linearity and doesn't fit the assumption.
B) Data is normally distributed showing good linearity. However, there are 2 outliers at residue 50 and residue 52. The data fit the assumption.
C) There is a spread of residues across the range, however, the data shows a slight negative regression line. With residuals more concentrated after the fitted value 70.
D) There are no data points outside of Cook's distance, thus, no outliers that can be of effect laying on the regression line. The data shows that no outliers are affecting the regression line and it does not fit the assumption.

Q6.

The final linear model showed an intercept co-efficient (mean of life expectancy) at 54.11 years and a standard error of 3.26 years. Below is an overall observation of the significant predictors starting from the one which has the most significant effect on life expectancy to the least effect (but still effective).

Significant predictor 1: Percentage of death caused by communicable diseases. It showed an estimated coefficient of -0.28, a standard error of 0.02 and a P-value of <2e-16. The results suggest that as a predictor, it is statistically significant. For every 1% increase in the total percentage of total deaths caused by communicable diseases, there is a decrease of -0.28 seen in life expectancy. We expect this; as a lot of people die because of disease, especially in third world countries where there is a poor health care system, poor sanitation and living standards.

Significant predictor 2: Log gdp_capita which represents the gross domestic product and how rich a country is. It showed a coefficient estimate of 2.51, a standard error of 0.24 and a P-value of <2e-16. This suggests that for every 1% increase in the total percentage of total death caused by the Log gdp_capita, there is an increase in life expectancy by 2.51 years. From the P-value, we can determine that it had a statistical significance. This also means that the higher the log GBP capita value is, the richer the country is, which means that the country will be able to offer better health care systems and sanitations which would impact life expectancy greatly.

Significant predictor 3: percentage of adults who are overweight. The results showed a co-efficient of -0.06, a standard error of 0.02 and a P-value of <0.005. This suggests that life expectancy decreases by 0.06 years for every 1% increase in adults who are overweight. We expect this from the results as overweight people suffer from underlying health complications which could cause death in some cases.

Significant predictor 4: log value of hospital beds. The results showed an estimated co-efficient of -1.34, a standard error of 0.35 and a P-value of <0.00017. This means that for every 1% increase in the log value of available hospital beds per 1000 people then there is a decrease in life expectancy by 1.34 years. This is an odd finding which will be discussed in question 7 of this report.

Significant predictor 5: percentage of babies that were vaccinated against DPT. The results showed a significant P-value of < 0.0016. The estimate co-efficient was 0.08 and the standard error 0.02. The results suggest that for every 1% of babies vaccinated against DPT, there is an increase in the average life expectancy by 0.08 years. This is most likely, as the vaccine protects against 3 fatal diseases and having a high percentage of children immunised against DPT will result in an increase in average life expectancy.

Significant predictor 6: Diabetes and the percentage of the population who has it. The results showed an estimated co-efficient of -0.22, standard error of 0.08 and P-value <0.0038. This means that for every 1% increase in the percentage of adults suffering from diabetes, the average life expectancy decreases by 0.22 years. Diabetes can cause serious health problems as well as death, and if not addressed accordingly then type 2 diabetes can develop into type 1 whereby the body becomes insulin resistant. It is therefore expected that with a high percentage of diabetes amongst adults in the population, the average life expectancy will decrease.

Question 7.

It is logical to think that with the availability of more hospital beds, then better treatments are available which could potentially save more lives leading to an increase in the average life expectancy of a population. But, if we assume that a certain country suffers from fatal diseases and health problems, then it will need much more beds per 1000 people and not all the people treated (also occupying a bed) will survive. In other words, just because a bed is available doesn’t necessarily mean that the person will survive an illness, especially if the entire country is suffering from a fatal disease and health complications. Perhaps we could instead use the number of hospital beds available to 1000 people as an indicator of the levels of disease and how it can implicate the healthcare system. It's important to see it from this way as well, just because a bed is available it does not mean that the level of care provided is up to a good standard, this is especially applied to poorer countries where the health system is not as strong and as a result, it will cause more deaths which will lower the average life expectancy.