An Analysis of Life Expectancy

By: Alexander Fou

5.23.2023

My goal was to analyze the effects of different factors on life expectancy. Going into this project, I had preconceived notions of what I thought would affect life expectancy: I thought that a higher Human Development Index (a score of how advanced/developed a country is) would correlate with higher life expectancies, and that a higher Air Quality Index (more particulates in the air) would correlate with a lower life expectancy, to name a few. To investigate, I found and grouped the following categories by country: life expectancy, human development index (HDI), life expectancy at birth, years of expected schooling, mean years of schooling, percent of population with access to drinking water, percent of population with access to basic sanitation, happiness score (from 0 to 10), GDP per capita, social support score , healthy life expectancy, freedom of life choices (the ability to do what you want in life), generosity, perception of government corruption, air quality index, alcohol consumption per capita (in liters/year), and percent of people who smoke. We can expect some multicollinearity, or categories correlated with each other; for example, the freedom to do what one wants in life is likely associated with an increase in happiness, or the percent of people with access to basic sanitation is likely associated with the human development index.

For this project, I chose a significance level of 0.1, as life expectancy is affected by too many variables, and thus, is difficult to predict. The next step was to figure out which factors were correlated. To do so, I ran a variance inflation factor test (VIF) to measure how correlated variables were with each other, by regressing each variable against all other variables. Some of the results were unsurprising: the HDI, life expectancy at birth (LEBirth), healthy life expectancy, and GDP per capita had high VIF scores. This makes sense, as the HDI is calculated from those categories. The VIF score for percent of population with sanitation was 10.95, just over the threshold of correlation; this also makes sense, as more developed countries have better access to sanitation. I expected the VIF score for percent of population with access to water to be extremely similar, and it was, with a VIF score of 8.44; however, this score puts it below the threshold, and so I included it for future testing.

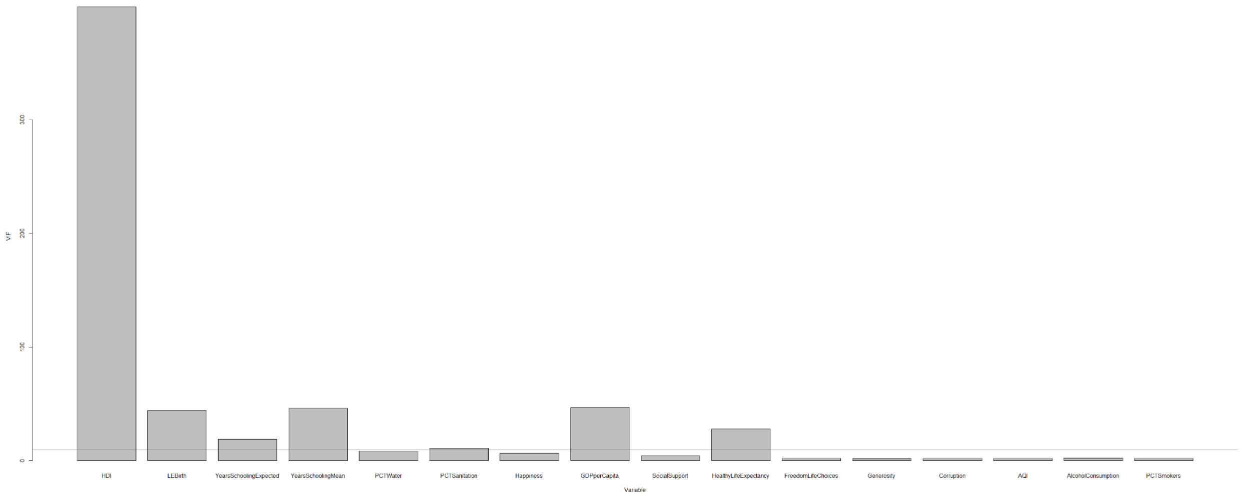


Fig 1. VIF scores of each category. Red bar marks a VIF score of 10. PCT stands for percent of population with access to ….

When making the model, I decided to include the HDI and remove the correlated categories because the HDI is derived from said categories. I then created a linear model with every category, which gave me a p-value for the F statistic of 2.2E^{-16}. In other words, there is incredibly strong evidence to suggest that there exists a relationship between life expectancy and these categories. The model had an R^2 value of 0.9626 and an adjusted R^2 value of 0.9578, meaning that it fit the data very well. However, many of the categories were insignificant: At the 0.1 significance level, only the HDI, healthy life expectancy, and % of population who smoked were significant, with respective values of 0.009, 2E^{-16}, and 0.048. The HDI value being significant makes sense: more development means more science and society, and thus, a higher life expectancy. The life expectancy of a healthy individual being very significant is expected; after all, most individuals are healthy[1], and so the life expectancies should be similar. The significance of the percent of population who smoke was surprising. However, it had a coefficient of 0.0455, meaning that it didn't affect the model that much. To revise the model, I searched for the optimal model using the olsrr library's step_all_possible() function, which brute-force checks every possible model for the best results. I sorted the results by adjusted R^2, which measures how well the model fits the data, adjusting for the amount of variables. Before creating the model, I compared life expectancy to each variable individually. Naturally, the life expectancy of a healthy individual was most closely correlated with the overall life expectancy; it follows an obvious positive linear relationship, as an increase in healthy life expectancy is very likely to correspond to an increase in overall life expectancy (r-value 0.9758). Happiness score, HDI, and social support all had strong positive linear relationships, with r-values of 0.8, 0.9332, and 0.7075. The other categories had weak linear relationships, with scores of -0.0084, -0.1368, and 0.385, respectively. Looking at the graphs, there is no clear relationship between those

[1] Citation needed

categories and life expectancy; because of this, I decided to drop them and only look at happiness, HDI, and social support, and I decided to use a linear model, as a linear relationship is easily seen in the data.

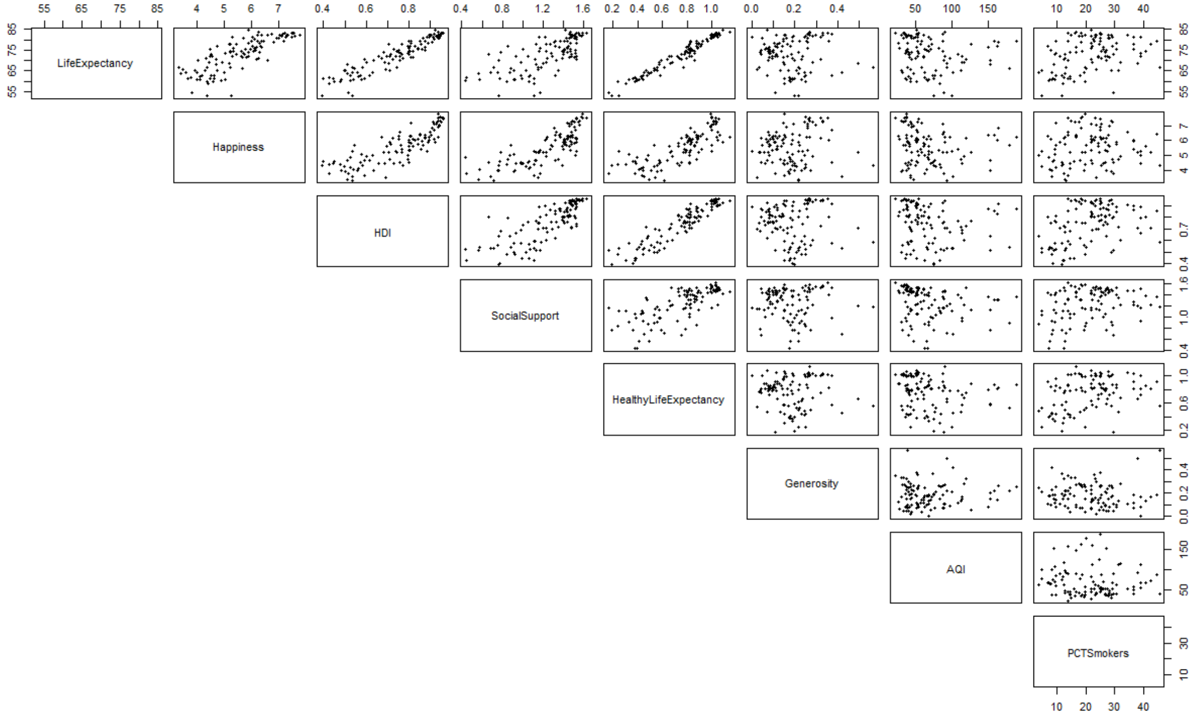


Fig 2. Life expectancy vs happiness, HDI, social support, healthy life expectancy, generosity, AQI, and % of population who smoke.

With this new model, every category was significant. As a final check, I looked at the Q-Q plot to investigate the distribution of our model: if most points lie along a straight line, then our dataset follows a normal distribution, and our model is good.

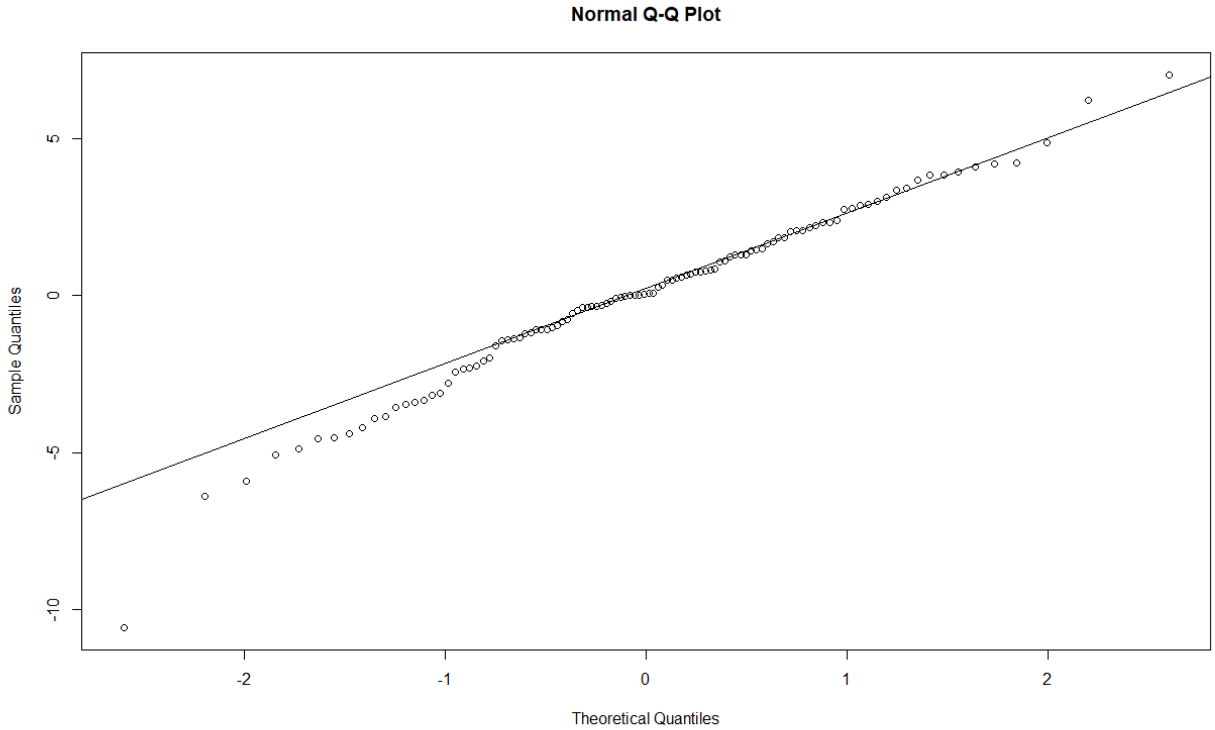


Fig 3. Q-Q Plot. Most points lie very close to the straight line with some outliers on the lower end.

```
Residuals:
     Min       1Q   Median       3Q      Max 
-10.5661  -1.3746   0.0558   1.8467   7.0151 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          38.0296     1.4036  27.094   <2e-16 ***
dfNew$Happiness       0.8546     0.4745   1.801   0.0746 .  
dfNew$HDI            45.1664     3.3170  13.617   <2e-16 ***
dfNew$SocialSupport  -2.9164     1.6705  -1.746   0.0838 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.839 on 104 degrees of freedom
Multiple R-squared:  0.8764,   Adjusted R-squared:  0.8728 
F-statistic: 245.8 on 3 and 104 DF,  p-value: < 2.2e-16
```

Fig 4. Final model: 38.0296 + 0.8546x_1 + 45.1664x_2 - 2.9164x_3 . Expected life expectancy: 72.7819 years (based on average values of 0.7443, 5.5453, and 1.2358, respectively).

So what did we learn from this experiment, and what are some limitations of our model? Well, we expect the average life expectancy to be 72.7819 years, which is 0.0017 years off the average life expectancy (72.7836). We expect that, for every 0.1 point increase in HDI, that the life expectancy increases by 4.5 years. We expect that for every 1 point increase in happiness score that life expectancy increases by 0.85 years, and that for every 1 point increase in social support score that life expectancy decreases by 2.9 years. This last part seems counterintuitive, and I could not think of any explanation for why this is. Our model is limited to the range 0-10 for happiness score, 0-1 for HDI, and 0-10 for social support scores; anything outside those ranges is meaningless. This means that our model can only predict life expectancies between 8.8656 and 91.742 years of age; the higher bound is reasonable, but the lower bound is not. A more reasonable lower bound would be 52.8571; this corresponds to two standard deviations below the mean for each variable, and is within 0.801 of the lowest recorded life expectancy, 52.777 years of age (from Chad). Our model also has a slightly worse R-squared value, (0.8728) but still fits the data well.

Lastly, it is important to note that this model is vastly oversimplified: it only contains three inputs, and is not a great predictor of life expectancy, which is affected by many variables. We cannot conclude that a higher HDI causes a higher life expectancy, but it is reasonable to conclude that they are correlated, based on common sense. And finally, the data set I used only had 108 data points, and so likely has more variance (and thus less accuracy) than the real values for each category. Because of this, this model is a reasonable starting point to predict life expectancy, but should **not** be relied on to accurately predict life expectancy.

Presentation: https://youtu.be/GIGZWnfB5I8

Appendix

https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset
Air pollution dataset, for the AQI. From user Hasib Al Muzdadid on Kaggle. 2023.
This dataset had 175 countries with many different air quality indexes from different cities. I took the average AQI per country and used that. However, AQI was insignificant in our model, so this doesn't *really* matter. The data itself is from eLichens, a company that revolves around air quality.

https://www.kaggle.com/datasets/unsdsn/world-happiness?select=2019.csv
World Happiness report, including happiness score, GDP per capita, social support score, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption. From user Sustainable Development Solutions Network on Kaggle. 2019.
This dataset has 156 entries, one per country. The data comes from the Gallup World Poll. Many of the variables were correlated with the HDI, and so were not included in the final model.

https://www.kaggle.com/datasets/utkarshxy/who-worldhealth-statistics-2020-complete
Datasets on alcohol consumption per capita, access to drinking water, access to basic sanitation, and tobacco usage. From user Zeus on Kaggle. 2021.
The alcohol abuse dataset had 2788 points, with many points per country per year. Data was split into three categories: male, female, and both sexes. Since the male data plus the female data was equal to the total data, and because the life expectancy data I worked with did not account for sex, I chose to only look at data regarding both sexes. I also chose the most recent datapoints per country, as these more accurately reflect the true values of life expectancy by country (obviously, life expectancy has gone up, and so data from 30 years ago is very likely to be less accurate).
The drinking water dataset was similar in that it had many datapoints per country, based on year (3456 total). Again, I looked only at the most recent data per country. With the sanitization dataset, they included data for urban, rural, and total percentages of the population, as well as year (9369 datapoints total). I took the most recent data for the entire population.

https://hdr.undp.org/data-center/human-development-index#/indicies/HDI
Human Development Index, including life expectancy at birth, expected years of schooling, mean years of schooling, GNI per capita, GNI minus HDI ranking, and HDI ranking. From the United Nations Development Programme. 2021. This dataset had 191 datapoints, one per country.

Source Code (+ Comments)

```
> #importing all the data and cleaning it. read.csv(“my file here”) and then
grouping data with the following parameters, in order: by country, then by most
recent year, then by total (ignoring sex or urban/rural or different cities). I
either used the average or used the total value. Merged into one huge dataframe
with everything titled df. Cleaned column names to be consistent.

#First we want to see if any datapoints are outliers: any datapoints are > 3 SD
away from mean.
> mean(df$LifeExpectancy)
[1] 72.78359
> avgLifeExpectancy = mean(df$LifeExpectancy)
> sdLifeExpectancy = sd(df$LifeExpectancy)
> outliers = which(df$LifeExpectancy > avgLifeExpectancy + 3 * sdLifeExpectancy
| df$LifeExpectancy < avgLifeExpectancy - 3 * sdLifeExpectancy)
> outliers
integer(0)

> model = lm(LifeExpectancy ~ HDI + LEBirth + YearsSchoolingExpected +
YearsSchoolingMean + PCTWater + PCTSanitation + Happiness + GDPperCapita +
SocialSupport + HealthyLifeExpectancy + FreedomLifeChoices + Generosity +
Corruption + AQI + AlcoholConsumption + PCTSmokers, data = df)#creating the
first model

>summary(model)


Residuals:
    Min      1Q  Median      3Q     Max 
-3.7806 -0.5295 -0.0161  0.4544  2.3303 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)           13.3296219  2.5873285   5.152 1.49e-06 ***
HDI                    1.9623592 11.2888071   0.174  0.86238    
LEBirth                0.7300511  0.0729224  10.011 2.42e-16 ***
YearsSchoolingExpected  0.0372133  0.1317039   0.283  0.77816    
YearsSchoolingMean    -0.0711102  0.1862694  -0.382  0.70353    
PCTWater               0.0046893  0.0127611   0.367  0.71413    
PCTSanitation         -0.0000237  0.0100922  -0.002  0.99813    
Happiness             -0.3845242  0.2055950  -1.870  0.06466 .  
GDPperCapita           0.5804624  1.5385697   0.377  0.70685    
SocialSupport          1.0260727  0.6445335   1.592  0.11486    
HealthyLifeExpectancy  6.0074279  2.0434576   2.940  0.00416 ** 
FreedomLifeChoices     1.8839962  0.9338896   2.017  0.04660 *  
Generosity            -0.2918362  1.1508524  -0.254  0.80039    
```

```
Corruption              -1.1638211  1.1985060  -0.971  0.33409
AQI                     -0.0027584  0.0035862  -0.769  0.44378
AlcoholConsumption      -0.0117649  0.0331675  -0.355  0.72363
PCTSmokers               0.0219763  0.0134165   1.638  0.10487
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9425 on 91 degrees of freedom
Multiple R-squared:  0.9881,  Adjusted R-squared:  0.986
F-statistic: 471.5 on 16 and 91 DF,  p-value: < 2.2e-16
> vifValues = vif(model)
> vifValues
HDI LEBirth YearsSchoolingExpected YearsSchoolingMean PCTWater
PCTSanitation Happiness GDPperCapita SocialSupport HealthyLifeExpectancy
FreedomLifeChoices
399.139862 44.151706 18.910710 46.082309 8.436001 10.950509 6.500960
46.629923 4.124012 27.763925 1.916417
Generosity Corruption AQI PCTAlcohol PCTSmokers
1.630963  1.864249 1.975933  2.141889  1.981753
# make a bar plot of VIF values to see
> barplot(vifValues, horiz = FALSE, col = "gray",
+        main = "VIF Scores", xlab = "Variables",
+        ylab = "VIF Score", cex.names = 0.8, las = 2)
> abline(h = 10, col = "red")

#HDI is measured from life expectancy at birth, years of expected and average
schooling, GDP per capita, so we remove.
> model = lm(LifeExpectancy ~ HDI + PCTWater + PCTSanitation + Happiness +
SocialSupport + HealthyLifeExpectancy + FreedomLifeChoices + Generosity +
Corruption + AQI + AlcoholConsumption + PCTSmokers, data=df)
> summary(model)

Call:
lm(formula = LifeExpectancy ~ HDI + PCTWater + PCTSanitation +
    Happiness + SocialSupport + HealthyLifeExpectancy + FreedomLifeChoices +
    Generosity + Corruption + AQI + AlcoholConsumption + PCTSmokers,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9084 -0.6646  0.0771  0.9842  3.6469

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            43.7152892  1.2946735  33.765  < 2e-16 ***
HDI                     9.8358573  3.6956456   2.661  0.00914 **
PCTWater               -0.0074632  0.0212315  -0.352  0.72598
PCTSanitation          -0.0009255  0.0163341  -0.057  0.95494
Happiness               0.2769457  0.3365609   0.823  0.41264
```

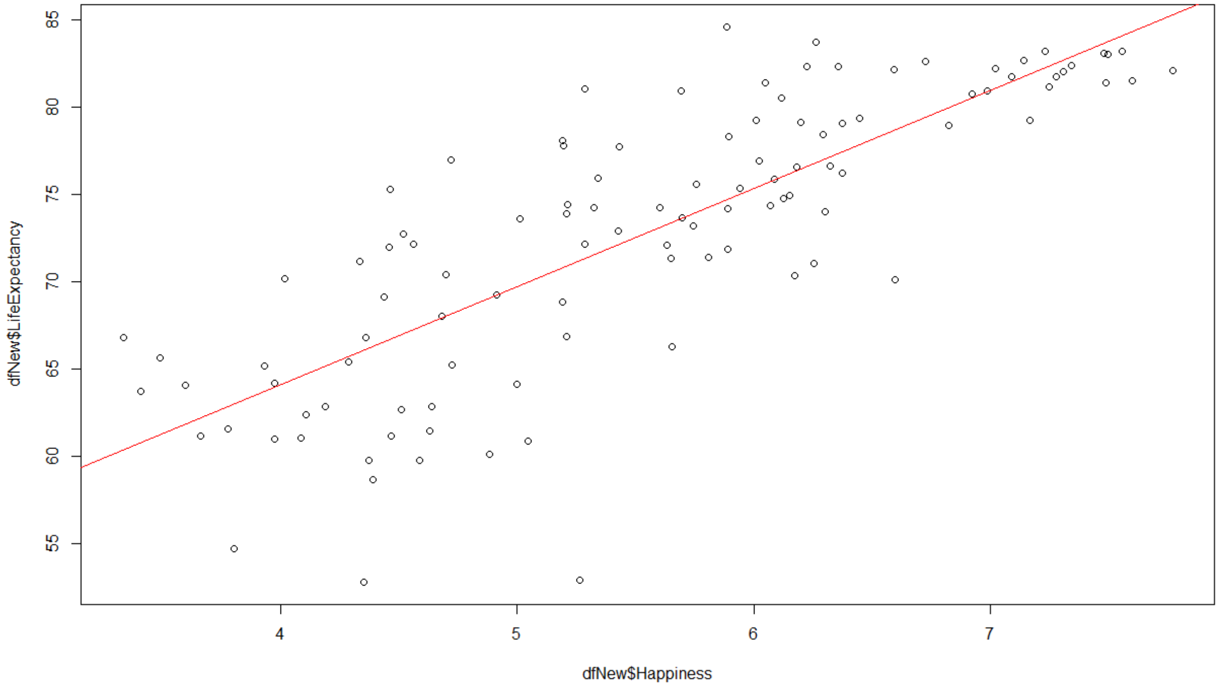
```
SocialSupport         -1.2725300  1.0636284  -1.196  0.23452
HealthyLifeExpectancy 27.0358595  1.9900884  13.585  < 2e-16 ***
FreedomLifeChoices     0.6395451  1.6019506   0.399  0.69062
Generosity             2.6668925  1.9006531   1.403  0.16383
Corruption             0.3741286  1.9900490   0.188  0.85128
AQI                    0.0078273  0.0054688   1.431  0.15563
AlcoholConsumption    -0.0188952  0.0563042  -0.336  0.73792
PCTSmokers             0.0455070  0.0227270   2.002  0.04810 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.635 on 95 degrees of freedom
Multiple R-squared:  0.9626,   Adjusted R-squared:  0.9578
F-statistic: 203.6 on 12 and 95 DF,  p-value: < 2.2e-16

#HDI significant which makes sense…
#Get all possible regression coefficients
> library(olsrr)
> potentialCoefficients = ols_step_all_possible(model)
> potentialCoefficients[order(-potentialCoefficients$adjr), ] #sort by highest
adjusted R-squared!

> potentialCoefficients[1,] #get the one with highest R^2 adjusted value

#the following I moved when writing the report for easier readability
Predictors
HDI Happiness SocialSupport HealthyLifeExpectancy Generosity AQI PCTSmokers
R-Square Adj. R-Square Mallow's Cp
0.9623841     0.9597510    3.458395

> predictors = c("Happiness", "HDI", "SocialSupport", "HealthyLifeExpectancy",
"Generosity", "AQI", "PCTSmokers")
> dfNew = df[, predictors] #to remove all other data

#create new model
> newModel = lm(dfNew$LifeExpectancy ~ predictors, data=dfNew)
> summary(newModel)

Call:
lm(formula = LifeExpectancy ~ Happiness + HDI + SocialSupport +
    HealthyLifeExpectancy + Generosity + AQI + PCTSmokers, data = dfNew)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7998 -0.6804  0.0657  0.9846  3.4780

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           43.733367   1.060982  41.220   <2e-16 ***
```

```
Happiness              0.291811   0.315749   0.924   0.3576
HDI                    9.191568   3.193938   2.878   0.0049 **
SocialSupport         -1.317755   0.963217  -1.368   0.1744
HealthyLifeExpectancy 26.963957   1.821629  14.802   <2e-16 ***
Generosity             3.274485   1.599154   2.048   0.0432 *
AQI                    0.007983   0.004422   1.805   0.0740 .
PCTSmokers             0.038746   0.019794   1.958   0.0531 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.597 on 100 degrees of freedom
Multiple R-squared:  0.9624,   Adjusted R-squared:  0.9598
F-statistic: 365.5 on 7 and 100 DF,  p-value: < 2.2e-16
```

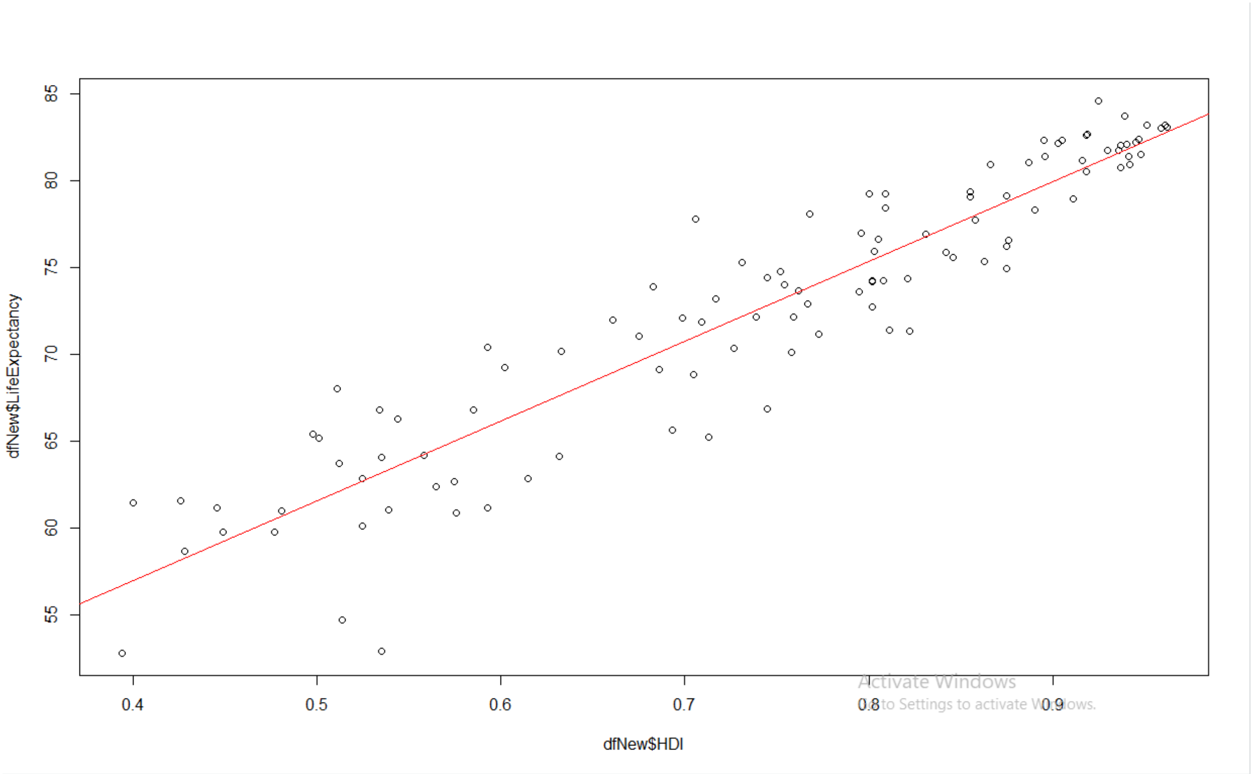
```
#Refer to Figure 2. Noticed a linear relationship between life expectancy and
happiness, HDI, social support, and healthy life expectancy. Obviously healthy
life expectancy corresponds to actual life expectancy so we remove this, it’s
like cheating.

> happy = lm(dfNew$LifeExpectancy ~ dfNew$Happiness, data=dfNew)
> summary(happy)
```

```
Call:
lm(formula = LifeExpectancy ~ Happiness, data = dfNew)

Residuals:
    Min      1Q  Median      3Q     Max
-18.317  -2.401   0.054   3.082   9.912

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.5402     2.3248   17.87   <2e-16 ***
Happiness     5.6342     0.4109   13.71   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.802 on 106 degrees of freedom
Multiple R-squared:  0.6395,   Adjusted R-squared:  0.6361
F-statistic:   188 on 1 and 106 DF,  p-value: < 2.2e-16
```

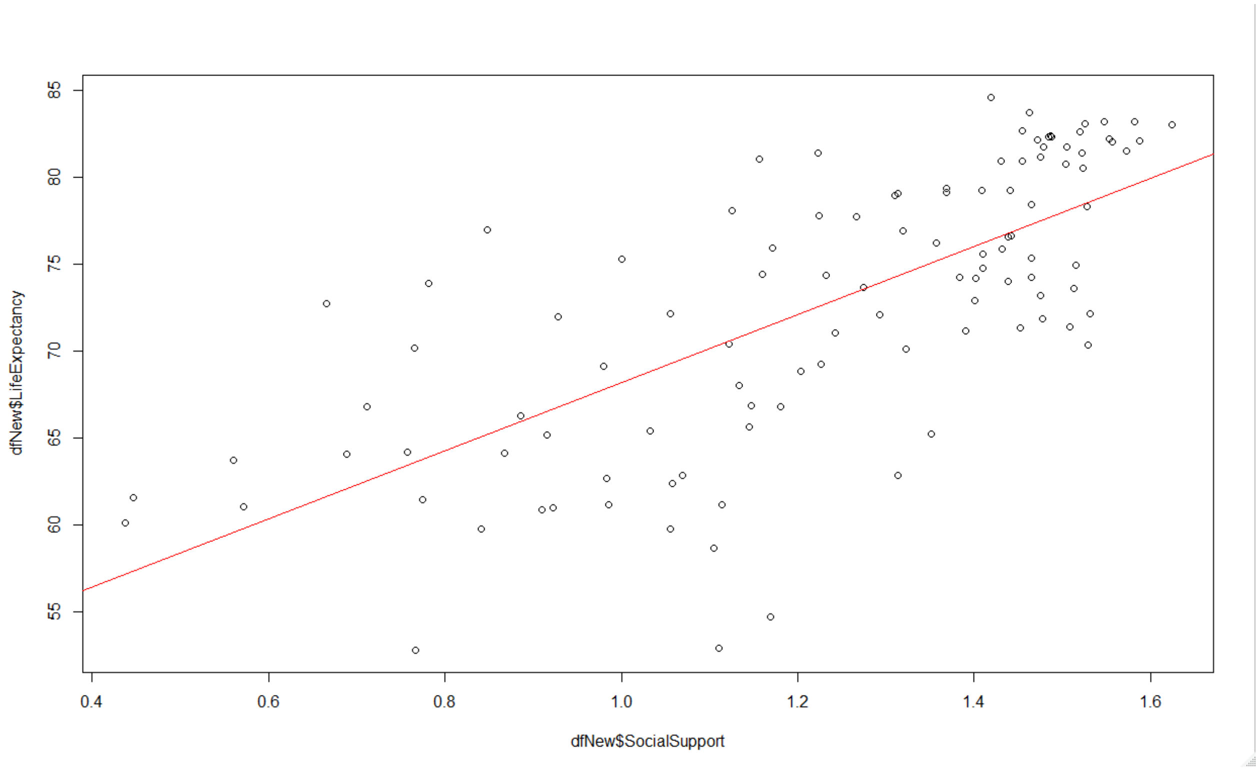
```
#looking at relationships between every category
```



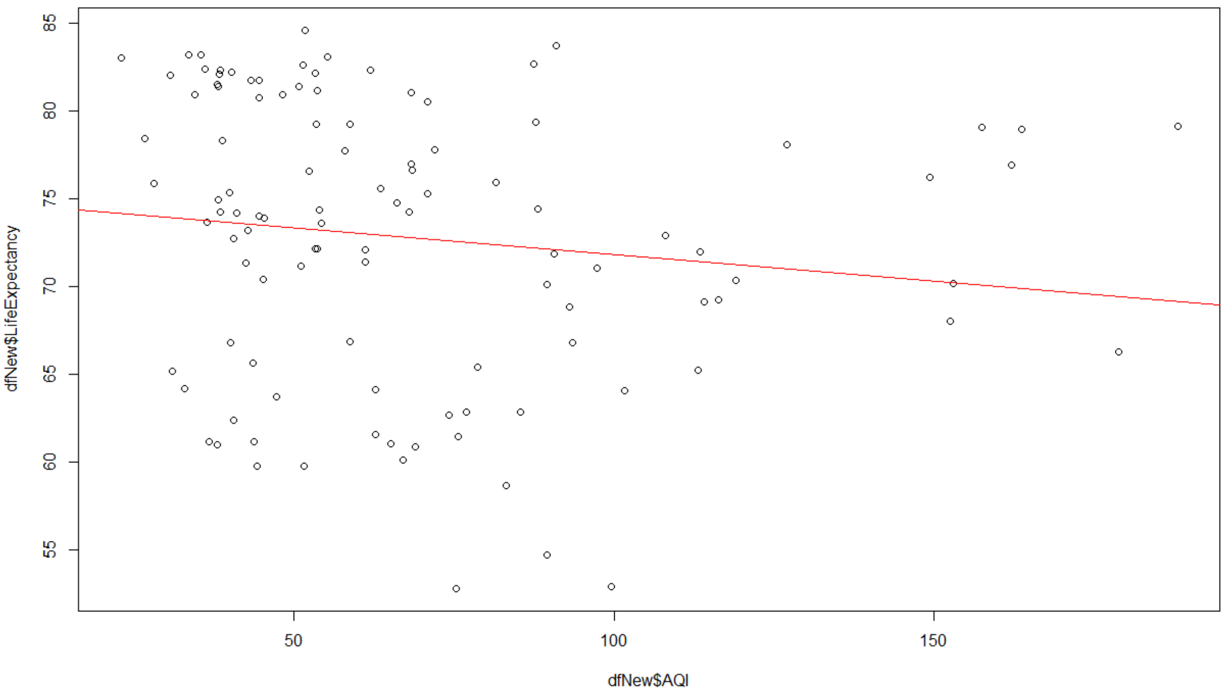
```
> plot(dfNew$Happiness, dfNew$LifeExpectancy)
> abline(happy, col=”red”)

> hdiVs = lm(LifeExpectancy ~ HDI, data=dfNew)
> summary(hdiVs)

Call:
lm(formula = LifeExpectancy ~ HDI, data = dfNew)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2513  -1.5512   0.3152   1.8356   6.7867

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.487      1.312   29.35   <2e-16 ***
HDI           46.076      1.722   26.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.873 on 106 degrees of freedom
Multiple R-squared:  0.871,    Adjusted R-squared:  0.8698
F-statistic: 715.6 on 1 and 106 DF,  p-value: < 2.2e-16
```



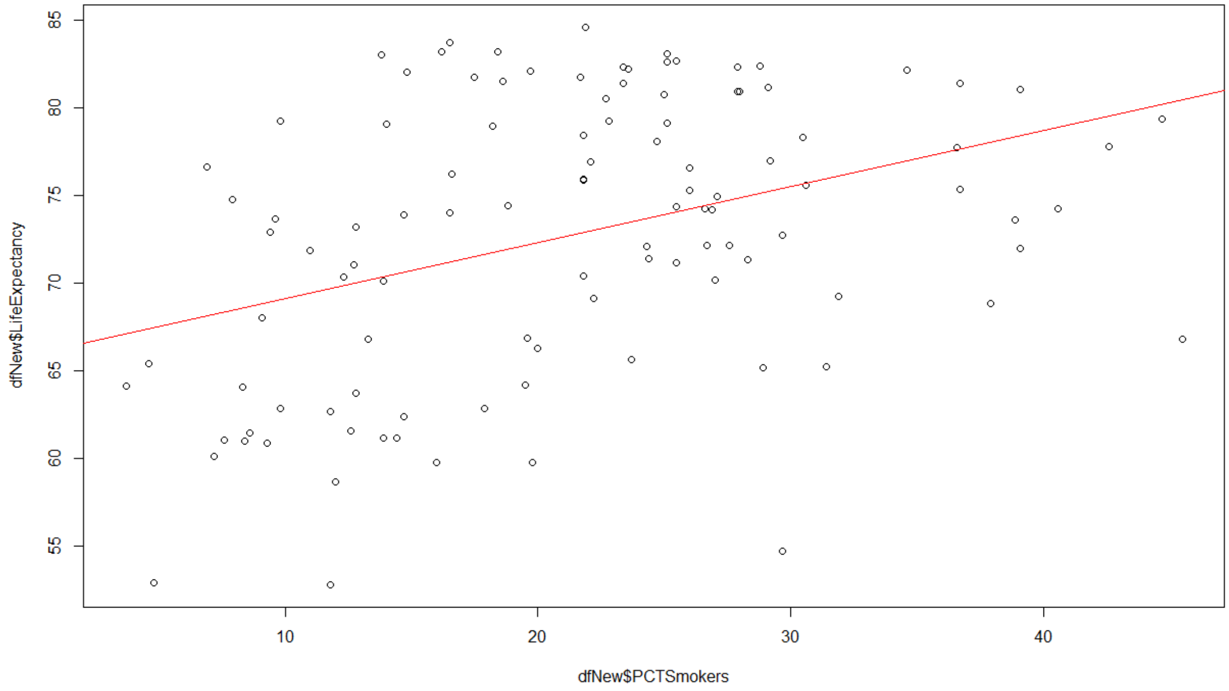
```
> plot(dfNew$HDI, dfNew$LifeExpectancy)
> abline(hdiVs, col=”red”)

> social = lm(LifeExpectancy ~ SocialSupport, data=dfNew)
> summary(social)

Call:
lm(formula = LifeExpectancy ~ SocialSupport, data = dfNew)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.449  -3.619   1.113   4.209  11.814 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     48.537      2.414   20.10   <2e-16 ***
SocialSupport   19.621      1.903   10.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.652 on 106 degrees of freedom
Multiple R-squared:  0.5006,	Adjusted R-squared:  0.4959 
F-statistic: 106.3 on 1 and 106 DF,  p-value: < 2.2e-16

> plot(dfNew$SocialSupport, dfNew$LifeExpectancy)
> abline(social, col="red")
```



```
> air = lm(LifeExpectancy ~ AQI, data=dfNew)
> plot(dfNew$AQI, dfNew$LifeExpectancy)
> abline(air, col="red")
> summary(air)

Call:
lm(formula = LifeExpectancy ~ AQI, data = dfNew)

Residuals:
     Min       1Q   Median       3Q      Max
-19.8028  -6.1879   0.5506   7.3424  11.6417

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.87624    1.65804  45.160   <2e-16 ***
AQI         -0.03048    0.02145  -1.421    0.158
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.923 on 106 degrees of freedom
Multiple R-squared:  0.0187,   Adjusted R-squared:  0.009444
F-statistic:  2.02 on 1 and 106 DF,  p-value: 0.1582
```

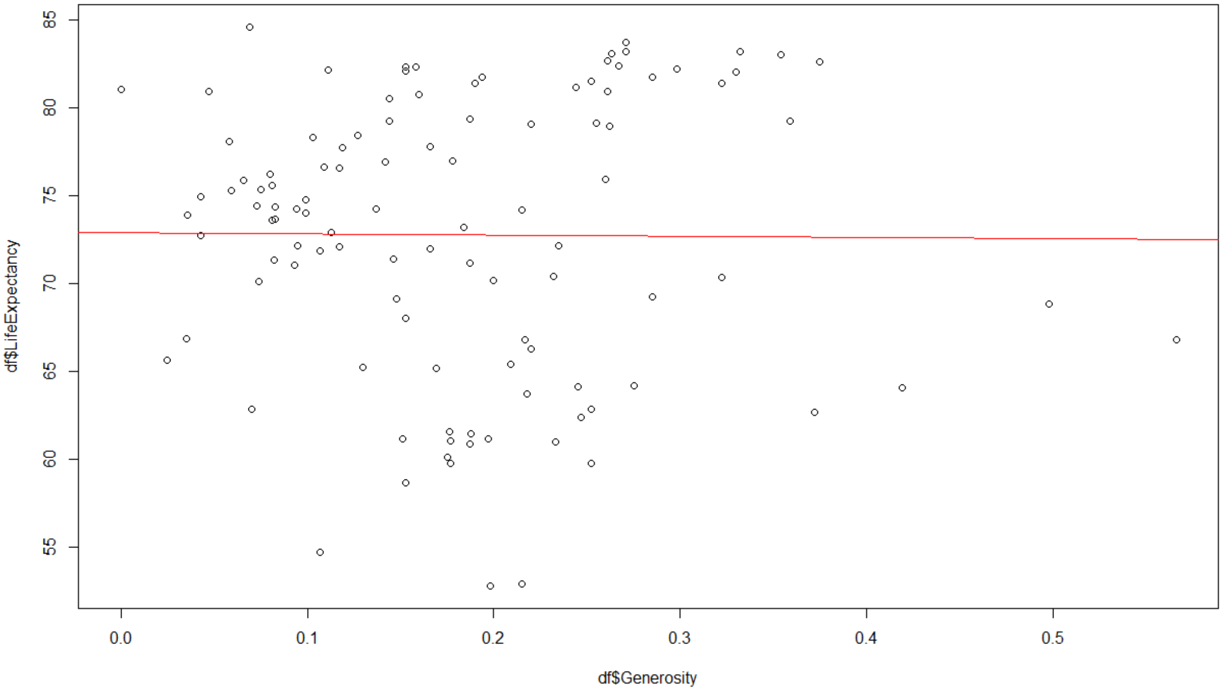


```
#this makes sense: higher AQI index means lower quality air and thus a decrease
in life expectancy.

> smoke = lm(LifeExpectancy ~ PCTSmokers, data=dfNew)
> plot(dfNew$PCTSmokers, dfNew$LifeExpectancy,)
> abline(smoke, col="red")
> summary(smoke)

Call:
lm(formula = LifeExpectancy ~ PCTSmokers, data = dfNew)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.7379  -6.0891   0.1753   6.0553  12.7325 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 65.90912    1.75116  37.637  < 2e-16 ***
PCTSmokers   0.32060    0.07465   4.295 3.88e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.382 on 106 degrees of freedom
Multiple R-squared:  0.1482,	Adjusted R-squared:  0.1402 
F-statistic: 18.45 on 1 and 106 DF,  p-value: 3.885e-05
```

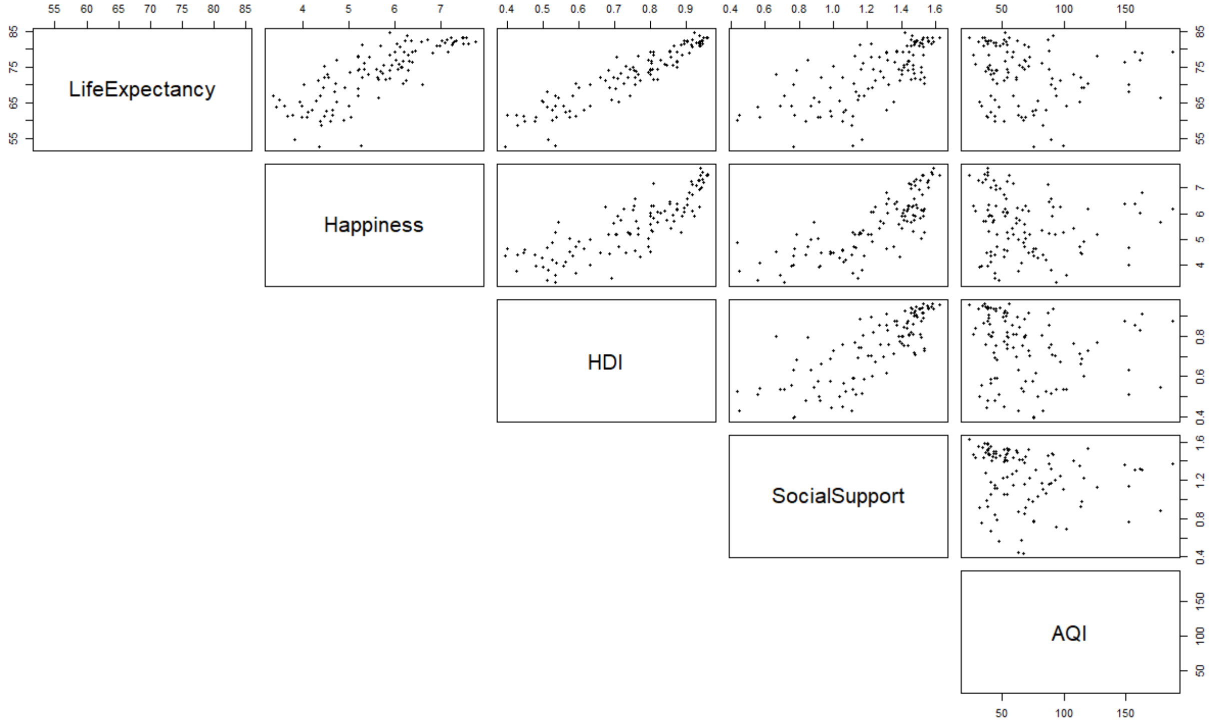


```
#this is interesting: as rate of smoking increases, life expectancy also
increases. This is counterintuitive and is probably caused by other things.
Maybe higher smoking rate means more access to money (to spend on
smoking-related things)???

> generous = lm(LifeExpectancy ~ Generosity, data=dfNew)
> plot(df$Generosity, df$LifeExpectancy)
> abline(generous, col="red")
> summary(generous)

Call:
lm(formula = LifeExpectancy ~ Generosity, data = dfNew)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.996  -6.132   1.277   6.599  11.757 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  72.9044     1.5938  45.742   <2e-16 ***
Generosity   -0.6619     7.6474  -0.087    0.931    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.998 on 106 degrees of freedom
Multiple R-squared:  7.067e-05,	Adjusted R-squared:  -0.009363
```

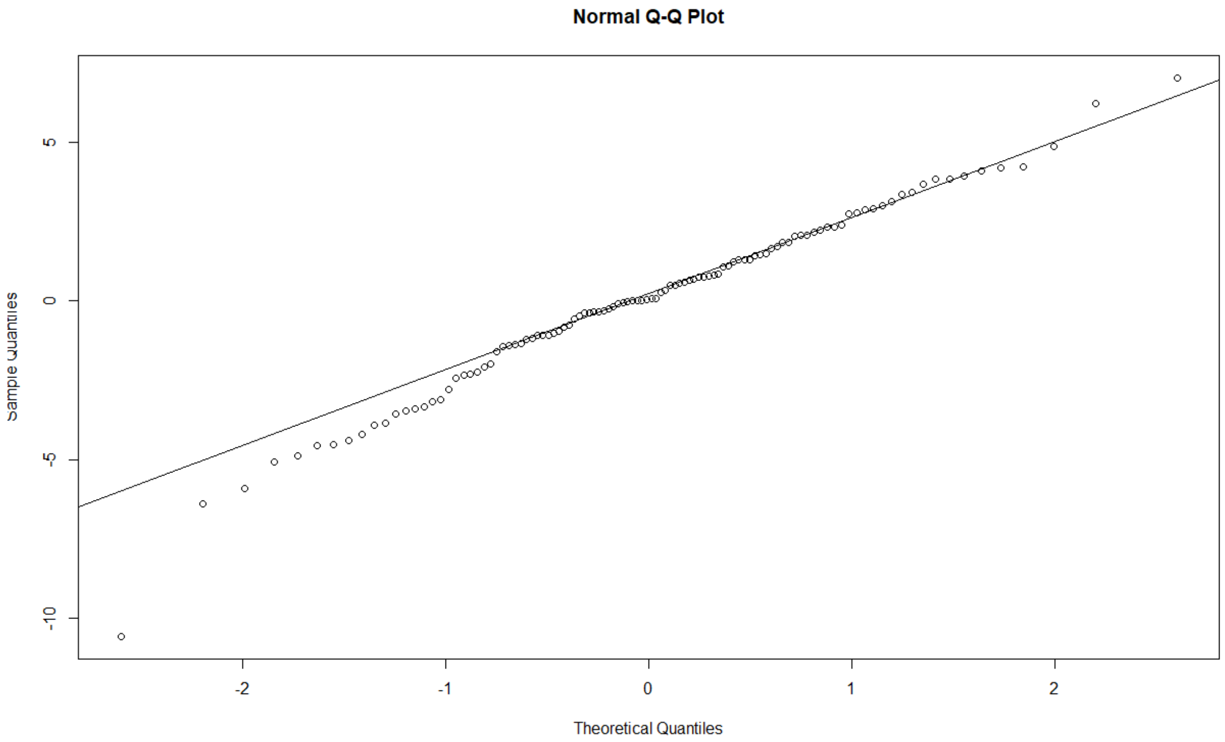
```
F-statistic: 0.007491 on 1 and 106 DF,  p-value: 0.9312
```



```
#F statistic too low, and visually checking there seems to be no relationship
between generosity and life expectancy. Can remove
```



```
#Everything seems to have a linear relationship EXCEPT AQI, which (at a glance)
seems to be random?

> final = lm(LifeExpectancy ~ dfNew$Happiness + dfNew$HDI +
dfNew$SocialSupport)
> summary(final)

Call:
lm(formula = LifeExpectancy ~ dfNew$Happiness + dfNew$HDI +
dfNew$SocialSupport,
    data = dfNew)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.5661  -1.3746   0.0558   1.8467   7.0151 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          38.0296     1.4036  27.094   <2e-16 ***
dfNew$Happiness       0.8546     0.4745   1.801   0.0746 .  
dfNew$HDI            45.1664     3.3170  13.617   <2e-16 ***
dfNew$SocialSupport  -2.9164     1.6705  -1.746   0.0838 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.839 on 104 degrees of freedom
Multiple R-squared:  0.8764,  Adjusted R-squared:  0.8728 
F-statistic: 245.8 on 3 and 104 DF,  p-value: < 2.2e-16

> residuals = residuals(model4)
> qqnorm(residuals)
> qqline(residuals)
```



```
#Most data follows the line, so seems normally distributed. On the lower end
there are low outliers: maybe those countries lag really far behind? Could be
due to active conflict in those countries?

#To make the bar chart for the presentation…
> data = round(df$LifeExpectancy)
> stripchart(data, method = "stack", offset = .5, at = 0, pch = 19,
+            col = "steelblue", main = "Life Expectancies (Rounded)", xlab =
"Age")
```