Social & Economic Welfare.md

Aneesh Didwania, Dorcas Cheung, Dorothy Zhu, Lanjing Wang, Pranav Prabhas, Zhengan Lyu

12/11/2019

Introduction

- In the 1970s, psychologist Paul Eckman suggested that happiness, along with sadness, disgust, fear, surprise, and anger are emotions that are universally experienced by all humans. Among these six emotions, happiness is one of the basic emotions and is the one that humans like to strive for. Although what happiness means differ from person to person, there are traditional believes that happiness is related to health. Some also identify that happiness and government policies are connected, and conflicts around you can play a role in determining one's happiness. Therefore, we are motivated to find what brings us pleasure and joy in our daily lives.
- We have a chance to analyze data from the World Happiness Report which is an annual survey that ranks the state of global happiness based on respondents rating of their own lives. Happiness score represented the "Ladder," is the national average response to the question of well-being evaluations. Respondents were asked to express which step of the ladder, from number 0 to 10, did they feel they stand on at the time of the survey. A higher number means that the individual is more satisfied in their current state.
- Also, this report uses 12 variables, including social support, healthy life expectancy, freedom, corruption, positive and negative effect, confidence in government, democratic quality, delivery quality, generosity, GDP and Gini of household income, to examine 160 countries for both 2017 and 2018 years.
    - The data on social support is based on the national mean of the "binary responses to the Gallup World Poll" in which individuals answered the question: "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"
    - In contrast, the data on healthy life expectancy at birth was based on the World Health Organization's Global Health Observatory, and it computes the number of years a newborn can expect to live.
    - For the freedom variable, it is the national average of binary responses to the Gallup World Poll's question "are you satisfied or dissatisfied with your freedom to choose what you do with your life?".
    - The corruption perception is the average response of the overall perception at the individual level.
    - Positive affect and negative affect define as the average of previous-day affect measure of laugh and sadness respectively.
    - Gini measures the distribution of income among individuals or households.
    - Generosity is the "residual of regressing the national average of Gallup World Poll responses to the question 'have you donated money to a charity in the past month'.

- Lastly, the data on GDP per capita was founded from the World Development Indicators, and it aimed to gauge a country's purchasing power or economic well being.

Data Background

```
#Read the data frame "the world's happiness" and name it as "whr_alternate"
#whr_alternate=read.csv("/Users/zhenganlyu/Library/Mobile Documents/com~apple~CloudDocs/whr_alternate.csv")

whr_alternate=read.csv("/Users/zhenganlyu/Library/Mobile Documents/com~apple~CloudDocs/whr_alternate.csv")

#summarize and see structure of the data frame
whr_alternate$X = NULL #X shows the list number and it is not important
summary(whr_alternate) #We observed that there are 283 observations, 17 variables and some missing values
```

```
     Country.name      Year       Life.Ladder    Log.GDP.per.capita
 Afghanistan:  2   Min.   :2017   Min.   :2.662   Min.   : 6.494
 Albania    :  2   1st Qu.:2017   1st Qu.:4.650   1st Qu.: 8.426
 Algeria    :  2   Median :2017   Median :5.481   Median : 9.460
 Argentina  :  2   Mean   :2017   Mean   :5.480   Mean   : 9.276
 Armenia    :  2   3rd Qu.:2018   3rd Qu.:6.276   3rd Qu.:10.237
 Australia  :  2   Max.   :2018   Max.   :7.858   Max.   :11.454
 (Other)    :271                                  NA's   :13
 Social.support   Healthy.life.expectancy.at.birth
 Min.   :0.3196   Min.   :45.20
 1st Qu.:0.7387   1st Qu.:59.00
 Median :0.8304   Median :66.10
 Mean   :0.8078   Mean   :64.42
 3rd Qu.:0.9053   3rd Qu.:68.90
 Max.   :0.9845   Max.   :76.80
 NA's   :1        NA's   :6
 Freedom.to.make.life.choices   Generosity       Perceptions.of.corruption
 Min.   :0.3735               Min.   :-0.33638   Min.   :0.09656
 1st Qu.:0.7146               1st Qu.:-0.14592   1st Qu.:0.68211
 Median :0.8038               Median :-0.03316   Median :0.79097
 Mean   :0.7825               Mean   :-0.01741   Mean   :0.73044
 3rd Qu.:0.8793               3rd Qu.: 0.08673   3rd Qu.:0.85146
 Max.   :0.9852               Max.   : 0.62958   Max.   :0.95439
 NA's   :1                    NA's   :15         NA's   :18
 Positive.affect  Negative.affect  Confidence.in.national.government
 Min.   :0.4210   Min.   :0.0927   Min.   :0.07971
 1st Qu.:0.6233   1st Qu.:0.2210   1st Qu.:0.33474
 Median :0.7250   Median :0.2827   Median :0.47367
```

```
 Mean   :0.7072   Mean   :0.2924   Mean   :0.49879
 3rd Qu.:0.7947   3rd Qu.:0.3556   3rd Qu.:0.63786
 Max.   :0.9028   Max.   :0.5993   Max.   :0.98812
 NA's   :2        NA's   :2        NA's   :26
 Democratic.Quality Delivery.Quality
 Min.   :-2.3266    Min.   :-2.01850
 1st Qu.:-0.7043    1st Qu.:-0.67840
 Median :-0.1644    Median :-0.17384
 Mean   :-0.1225    Mean   :-0.00347
 3rd Qu.: 0.6412    3rd Qu.: 0.68622
 Max.   : 1.5750    Max.   : 2.06917
 NA's   :136        NA's   :136
 Standard.deviation.of.ladder.by.country.year
 Min.   :1.198
 1st Qu.:1.858
 Median :2.215
 Mean   :2.243
 3rd Qu.:2.596
 Max.   :3.719

 Standard.deviation.Mean.of.ladder.by.country.year
 Min.   :0.1654
 1st Qu.:0.3123
 Median :0.4076
 Mean   :0.4371
 3rd Qu.:0.5474
 Max.   :0.9717

 gini.of.household.income.reported.in.Gallup..by.wp5.year
 Min.   :0.2010
 1st Qu.:0.3761
 Median :0.4406
 Mean   :0.4628
 3rd Qu.:0.5615
 Max.   :0.8520
 NA's   :4
```

```
str(whr_alternate)
```

```
'data.frame':   283 obs. of  17 variables:
 $ Country.name                                          : Factor w/ 152 levels "Af
ghanistan",..: 1 1 2 2 3 3 4 4 5 5 ...
 $ Year                                                  : int  2017 2018 2017 2018
2017 2018 2017 2018 2017 2018 ...
 $ Life.Ladder                                           : num  2.66 2.69 4.64 5 5.
25 ...
 $ Log.GDP.per.capita                                    : num  7.5 7.49 9.38 9.41
9.54 ...
 $ Social.support                                        : num  0.491 0.508 0.638 0
.684 0.807 ...
 $ Healthy.life.expectancy.at.birth                      : num  52.8 52.6 68.4 68.7
65.7 ...
 $ Freedom.to.make.life.choices                          : num  0.427 0.374 0.75 0.
824 0.437 ...
 $ Generosity                                            : num  -0.1122 -0.08489 -0
.03264 0.00538 -0.19152 ...
 $ Perceptions.of.corruption                             : num  0.954 0.928 0.876 0
.899 0.7 ...
 $ Positive.affect                                       : num  0.496 0.424 0.669 0
.713 0.642 ...
 $ Negative.affect                                       : num  0.371 0.405 0.334 0
.319 0.289 ...
 $ Confidence.in.national.government                     : num  0.261 0.365 0.458 0
.435 NA ...
 $ Democratic.Quality                                    : num  -1.887 NA 0.3 NA -0
.928 ...
 $ Delivery.Quality                                      : num  -1.438 NA -0.13 NA
-0.817 ...
 $ Standard.deviation.of.ladder.by.country.year          : num  1.45 1.41 2.68 2.64
2.04 ...
 $ Standard.deviation.Mean.of.ladder.by.country.year     : num  0.546 0.523 0.578 0
.528 0.389 ...
 $ gini.of.household.income.reported.in.Gallup..by.wp5.year: num  0.287 0.291 0.41 0.
456 0.528 ...
```

```
whr_alternate1=na.omit(whr_alternate)

#Incorporated another data frame and name it as new data
#data <- data.frame(read_csv("/Users/zhenganlyu/Downloads/GDP-Data-Set.csv"))
data <- data.frame(read_csv("/Users/zhenganlyu/Downloads/GDP-Data-Set.csv"))
newdata = data[order(data$GDP.PPP.Per.Capita),]

#summarize and see structure of the new included data frame
summary(newdata)
```

```
       X.          Country.name             Year      GDP.PPP.Per.Capita
 Min.   :  1.00   Length:140         Min.   :2017   Min.   :   727
 1st Qu.: 35.75   Class :character   1st Qu.:2017   1st Qu.:  4873
 Median : 70.50   Mode  :character   Median :2017   Median : 14903
 Mean   : 70.50                      Mean   :2017   Mean   : 21688
 3rd Qu.:105.25                      3rd Qu.:2017   3rd Qu.: 32729
 Max.   :140.00                      Max.   :2017   Max.   :107641
  Life.Ladder    GDP.Per.Capita.Nominal
 Min.   :2.662   Min.   :   357
 1st Qu.:4.608   1st Qu.:  1562
 Median :5.587   Median :  5769
 Mean   :5.481   Mean   : 14830
 3rd Qu.:6.278   3rd Qu.: 18825
 Max.   :7.788   Max.   :105280
```

```
str(newdata)
```

```
'data.frame':   140 obs. of  6 variables:
 $ X.                    : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Country.name          : chr  "Central African Republic" "Congo (Kinshasa)" "Niger"
"Malawi" ...
 $ Year                  : num  2017 2017 2017 2017 2017 ...
 $ GDP.PPP.Per.Capita    : num  727 889 1019 1205 1250 ...
 $ Life.Ladder           : num  3.48 4.31 4.62 3.42 4.28 ...
 $ GDP.Per.Capita.Nominal: num  424 462 376 357 441 699 504 450 618 673 ...
```

Exploratory Data Analysis

- Looking at variables, we found that our group is most interested in studying how happiness is related to the social and economic welfare of a country because the quality of life of an individual is of paramount importance to how happy they are. We assume that the variables of social and economic welfare are external and should be considered as public conditions that affect one's private wellbeing.

Histogram of Life Ladder

Pranav Prabhas

```
whr_alternate$Year <- as.factor(whr_alternate$Year)

ggplot(whr_alternate, aes(x=Life.Ladder, color = Year, fill =Year, y=..density..)) +
  geom_histogram(position="identity", bins = 25, alpha = 0.5) + labs(title="The Happi
ness Distribution",x="Life Ladder Score", y = "Density") + theme_minimal() + theme(le
gend.position="top") + geom_density(alpha=0, fill="#FB6656") +
  theme(plot.title = element_text(hjust = 0.5))
```

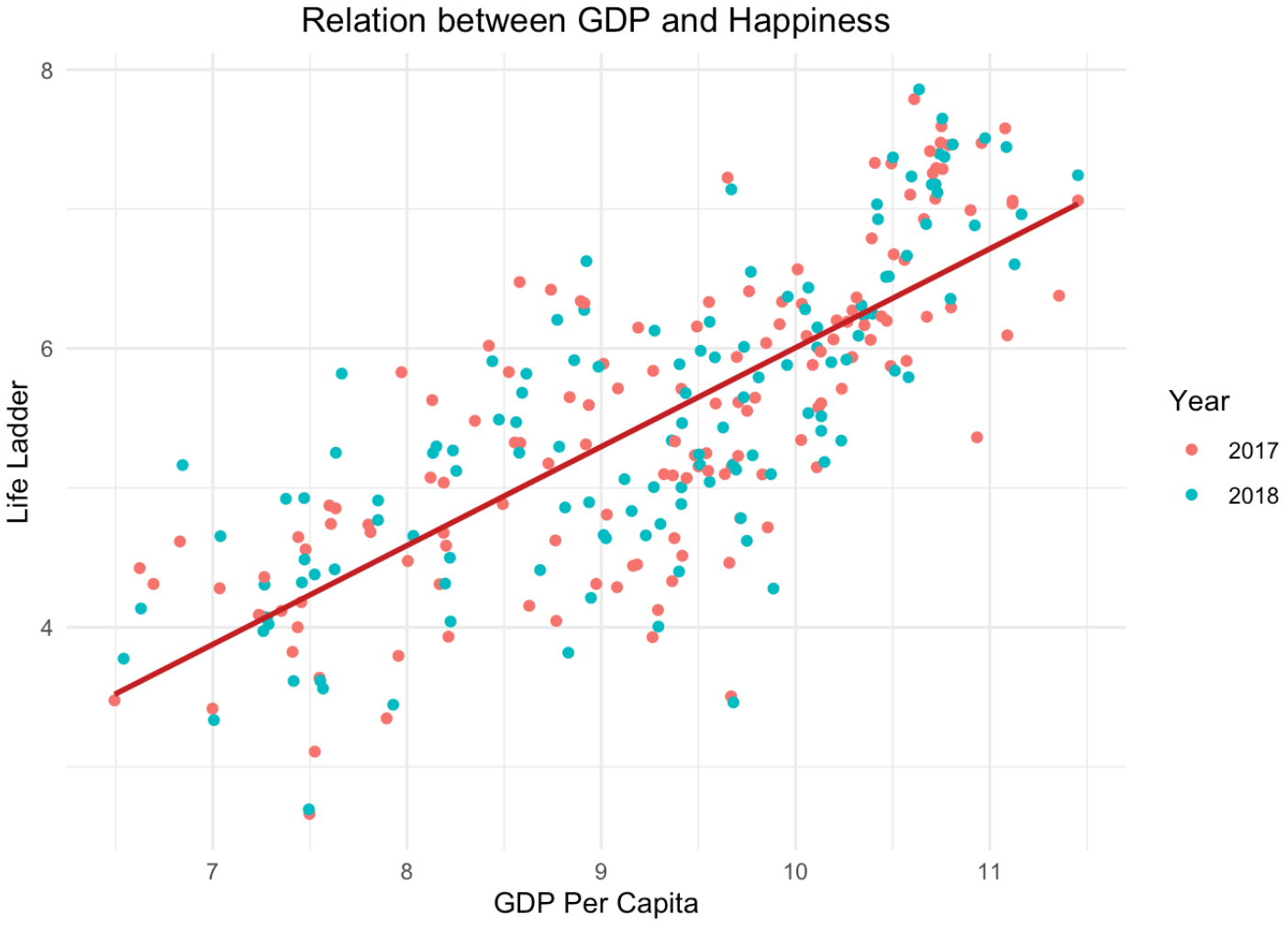
Individual Interpretation

- The graph below is a Density Histogram, which shows how the Life Ladder Scores are distributed. As seen in the Legend, the Pink bars are used to show the distribution for 2017, and the blue bars show the same for 2018. For each year, I have also drawn a Density Estimation, which can act as an alternative to the histogram, for future analysis involving Normal Curves. This graph gives us a few key statistics.
    - 1. The Mean Life Ladder Score in 2017 is 5.46, while the mean in 2018 is 5.50.
    - 2. The Median Life Ladder Score in 2017 is 5.55, while the Median in 2018 is 5.46.
    - 3. The SD of Scores in 2017 is 1.14, while the SD in 2018 is 1.10.
- I was particularly surprised to see an SD of 1.15, given the data is only plotted on an 8-point scale. After discussing with the team, we decided to investigate the question: “What Drives Happiness?”

GDP Per Capita & Life Ladder Scatter Plot

Aneesh Didwania

```
my_graph <-ggplot(whr_alternate, aes(x = (Log.GDP.per.capita), y = Life.Ladder)) +
  geom_point(aes(color = factor(Year))) +
  stat_smooth(method = "lm",
              col = "#C42126",
              se = FALSE,
              size = 1)
my_graph +
  labs(
    x = "GDP Per Capita",
    y = "Life Ladder",
    color = "Year",
    title = "Relation between GDP and Happiness"
  ) + theme_minimal() + theme(plot.title = element_text(hjust = 0.5))
```



Individual Interpretation

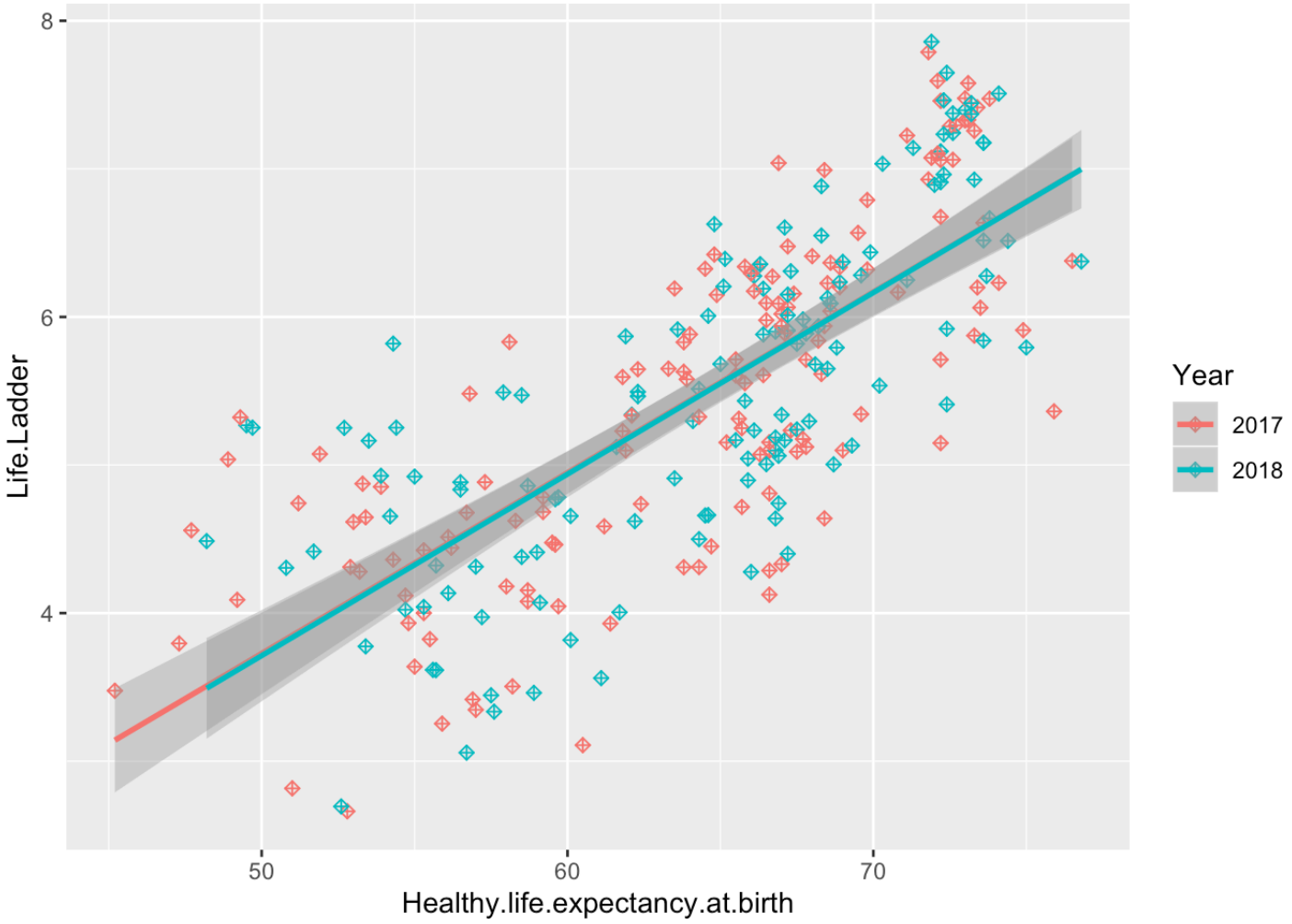
- This plot seeks to show the relationship between the Happiness Coefficient and the GDP of a Country. I sought to find this idea because I wanted to explore if a country's economic status might influence their quality of living and raise the general level of happiness within people. As the graph shows, there is a

strong correlation between the GDP of a country and its overall Happiness score. I used both Years of Data to plot this graph and show the regression line.

Healthy Life Expectancy At Birth & Life Ladder Scatter Plot

Dorothy Zhu

```
ggplot(whr_alternate, aes(x = Healthy.life.expectancy.at.birth, y = Life.Ladder,color
= Year)) +
  geom_point(size=1.5,shape=9,) +
  geom_smooth(method="lm",se=T)
```



Individual Interpretation:

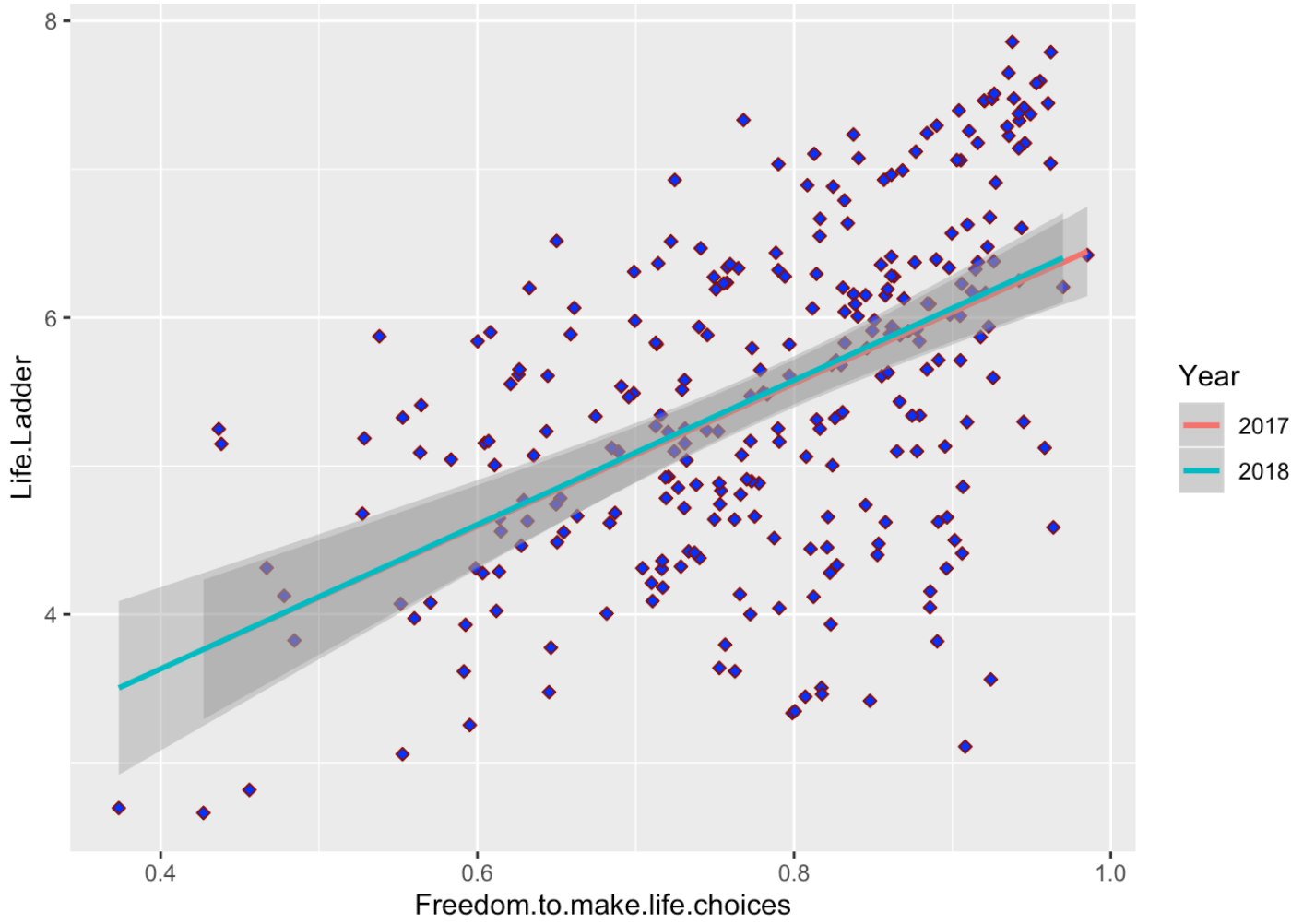
- The goal of the healthy life expectancy at birth (HALE) is to apply “disability weights” to countries to compute the expected number of years newborns can expect to live. Given the data, the HALE at birth ranges from approximately 45 years to 76 years for 2018 and 2017.
- In the scatterplot above, the darker points consist of data from 2017, and the lighter points consist of data from 2018.The regression line with the confidence interval is shown above, and the line goes through the point of averages. Notably, there exists a positive association between the life ladder and

the healthy life expectancy at birth.

Freedom & Life Ladder Scatter Plot

Dorcas Cheung

```
ggplot(whr_alternate, aes(x=Freedom.to.make.life.choices,y=Life.Ladder,color= Year))
+
  geom_point(shape=23, fill="blue", color="darkred") +
  geom_smooth(method="lm")
```




```
r=cor(whr_alternate1$Freedom.to.make.life.choices,whr_alternate1$Life.Ladder)
r
```

```
[1] 0.4955001
```

```
SDy=sd(whr_alternate1$Life.Ladder)
SDx=sd(whr_alternate1$Freedom.to.make.life.choices)
SDy
```

```
[1] 1.148219
```

```
SDx
```

```
[1] 0.1218002
```

```
m=r*(SDy/SDx)
m
```

```
[1] 4.671113
```

```
lm.out =lm(Life.Ladder ~ Freedom.to.make.life.choices, data=whr_alternate1)
summary(lm.out)
```

```

Call:
lm(formula = Life.Ladder ~ Freedom.to.make.life.choices, data = whr_alternate1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.95151 -0.55260  0.05589  0.87558  1.92529 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   1.8180     0.5817   3.125  0.00221 ** 
Freedom.to.make.life.choices   4.6711     0.7295   6.403 2.74e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001 on 126 degrees of freedom
Multiple R-squared:  0.2455,    Adjusted R-squared:  0.2395 
F-statistic:    41 on 1 and 126 DF,  p-value: 2.74e-09
```

Individual Analysis:

- The scatter plot above shows that there is a relatively weak relationship between Freedom to make life choices and Life ladder (which is a measure of happiness).
- The blue line represents the regression line and can be represented by the equation: Predicted Life

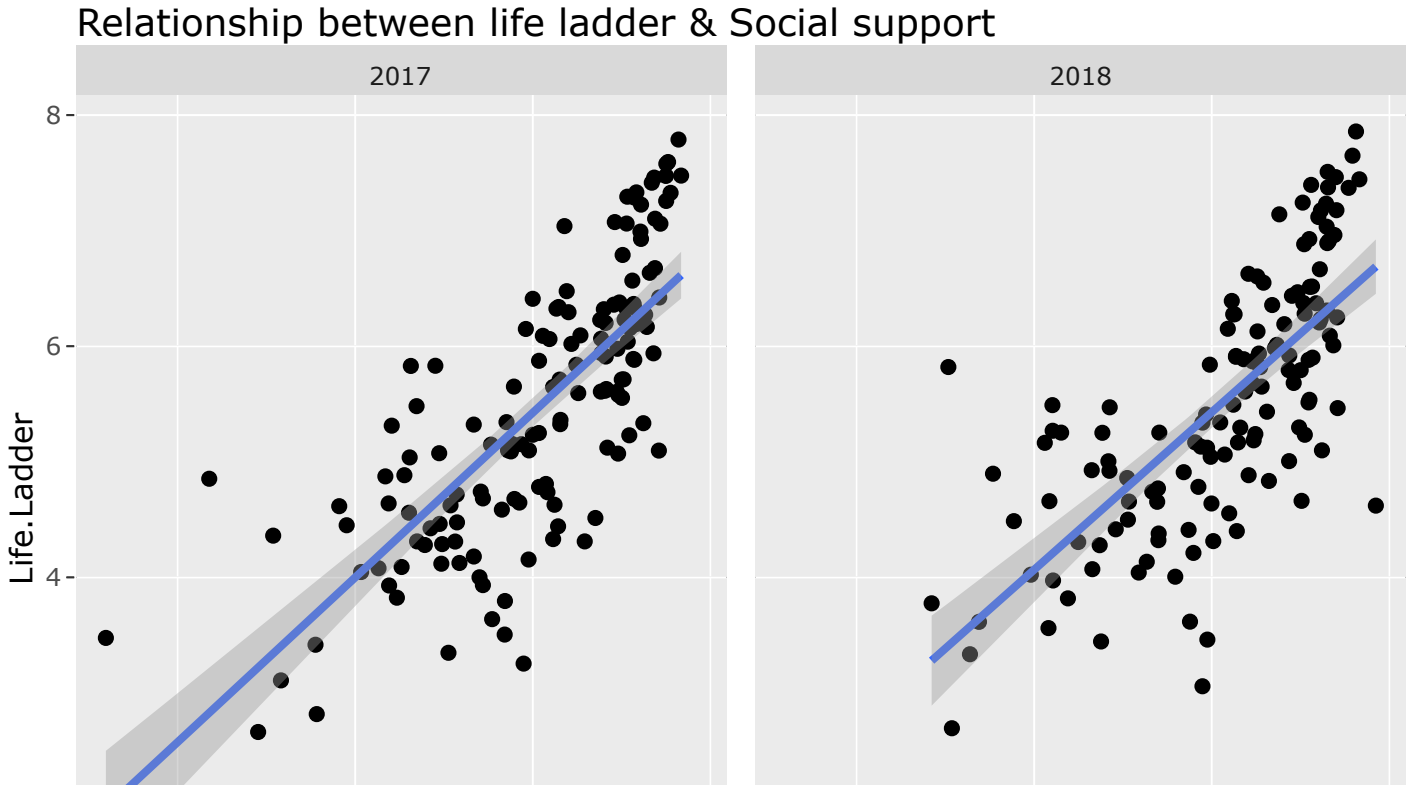
Ladder = 4.671*Freedom to make life choices + 1.818. The regression line has a slope of 4.677 and a y-intercept of 1.81.

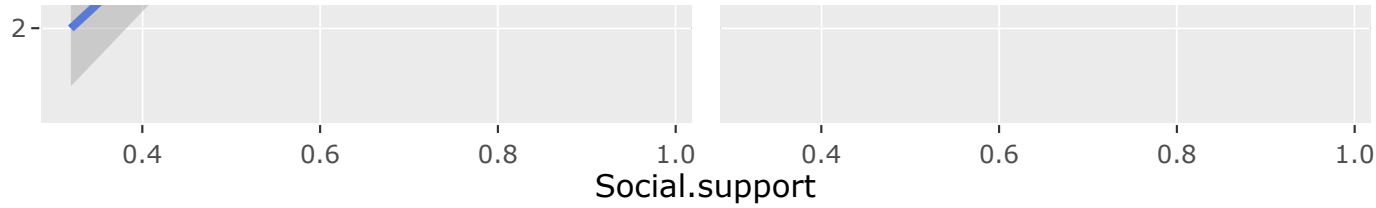
- I observed that, in general, the greater the freedom to make life choices, the higher the happiness score. In order words, countries whose citizens are highly satisfied with their freedom to choose what they do with their life are likely to have a higher national average value of ladder. This positive linear relationship is represented by the positive slope of the regression line. Considering the relatively low correlation coefficient, r=0.496, the relationship between the two variables is not strong. Having one data of variable is not that helpful to predict the other variable. The gray shading shows the confidence interval of the regression line.

Social Support & Life Ladder Scatter Plot

Zhengan Lyu

```
#visualize correlation by making a scatter plot
#Make an internative scatter plot for the year 2017 & 2018 separately; set "social su
pport" as x-axis and "life ladder" as y-axis; apply regression line and label the sca
tter plot
scatterplot1= whr_alternate %>%
  ggplot(aes(x=Social.support, y=Life.Ladder))+
  geom_point()+
  geom_smooth(method="lm", se=TRUE)+
  facet_wrap(~Year)+
  labs(title="Relationship between life ladder & Social support")
ggplotly(scatterplot1)
```





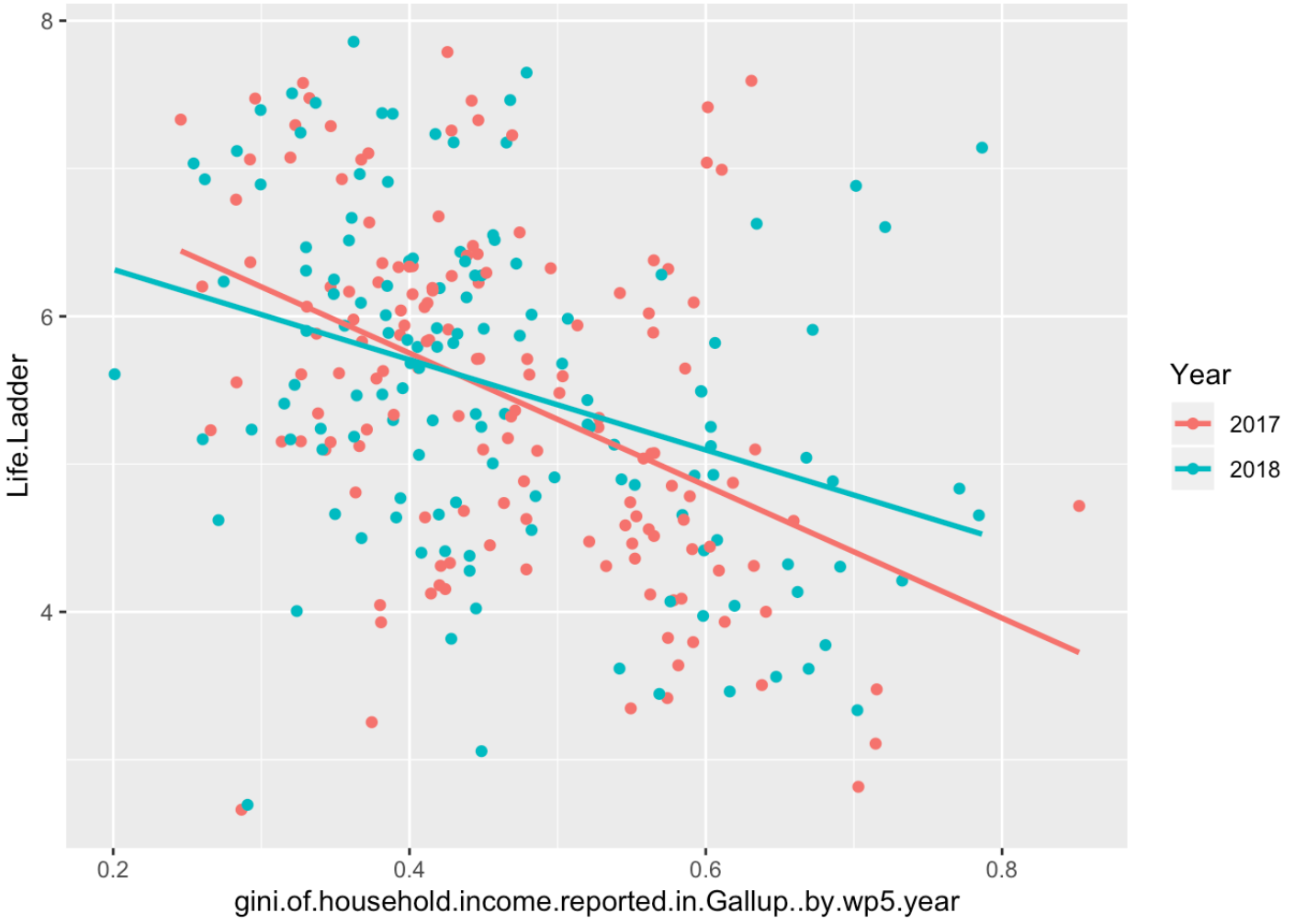
Individual Interpretation

- This scatter plot shows an upward trend and a tightly clustered football-shaped plot, meaning there is a positive strong correlation between social support and life ladder.
- Football shaped plot means x and y values are normally distributed. It also means the plot is linear and it is necessary for doing the regression method. There is the same separation in vertical strips.
- For confidence interval of regression, when there are more data concentrated, we are more certain about predicted value. It means the prediction interval is very narrow. When there are fewer observations in the area, we have less confidence in our predicted value.

GINI and Life Ladder Scatter Plot

Lanjing Wang

```
ggplot(whr_alternate, aes (x= gini.of.household.income.reported.in.Gallup..by.wp5.yea
r, y= Life.Ladder, color=Year))+
  geom_point()+
  geom_smooth(method = "lm", se= FALSE)
```




```
Year2018= whr_alternate %>%
  filter(Year=="2018") %>%
  select(-Democratic.Quality,-Delivery.Quality)
Year2018_=na.omit(Year2018)

r4=cor(Year2018_$Life.Ladder,Year2018_$gini.of.household.income.reported.in.Gallup..b
y.wp5.year)
r4
```

```
[1] -0.3839868
```

```
Year2017_=filter(whr_alternate, Year=="2017")
Year2017__=na.omit(Year2017_)
r5=cor(Year2017__$Life.Ladder,Year2017__$gini.of.household.income.reported.in.Gallup.
.by.wp5.year)
r5
```

```
[1] -0.4946095
```

Individual Interpretation

- The plot shows a negative association between life ladder and Gini index from the regression line. As a rule, the higher life ladder has smaller Gini index, representing that higher happiness score in a country usually have perfect equality income in a country.
- The correlation coefficient r which is -0.4946095 in 2017 and -0.3839868 in 2018. The r indicates that there is a weak association between Gini and the life ladder. Knowing one variable does not help much in guessing the other. Therefore, the Gini index may not be a better variable to test if Gini index has affected on a life ladder.

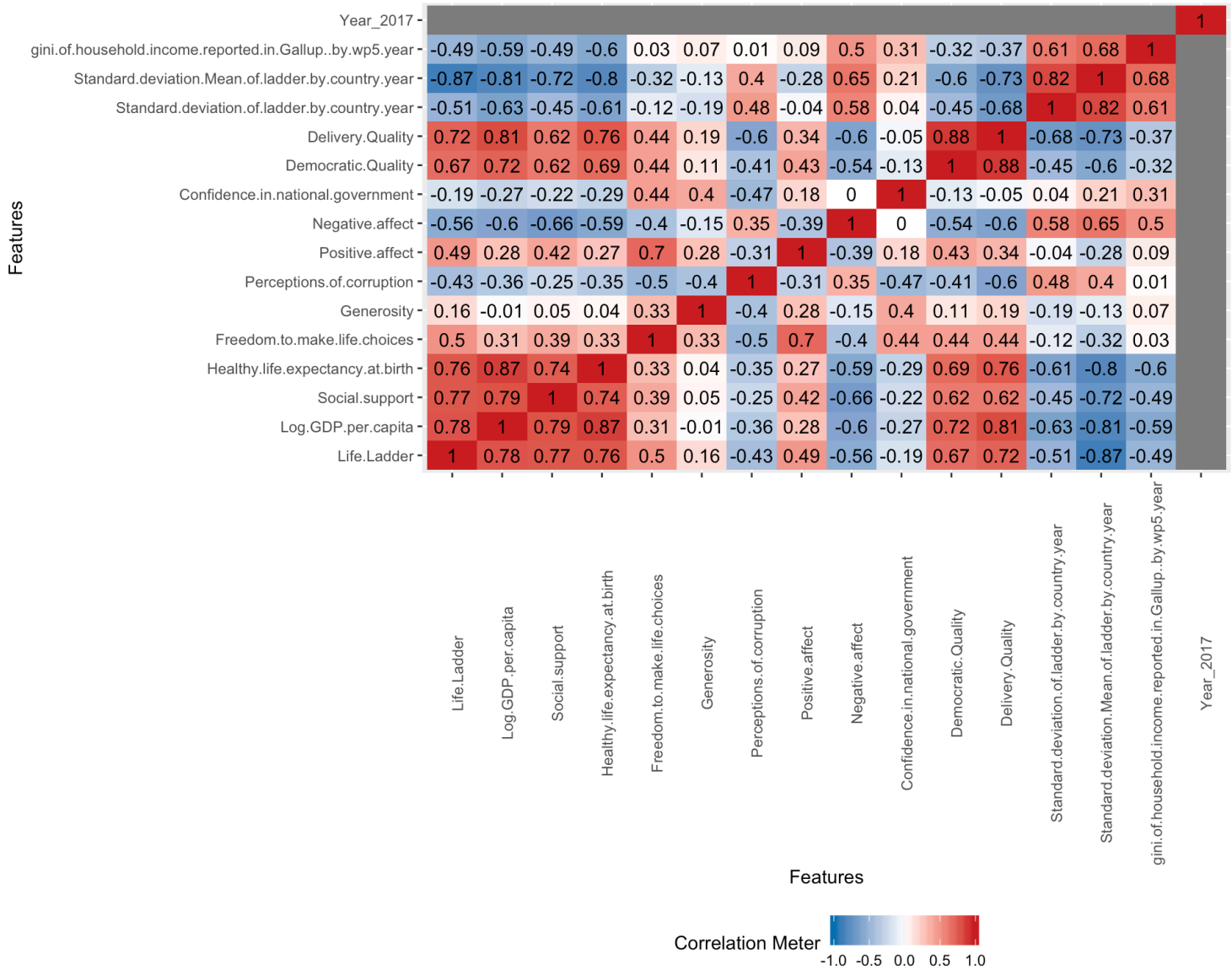
Correlation Heatmaps

Zhengan Lyu

```
# %>% take lefthand side as input for command on the righthand side
# Take rows that are only related to the year 2017 in the data frame whr_alternate, t
hen name it as Year2017
Year2017= whr_alternate %>%
  filter(Year=="2017")

## Take rows that are only related to the year 2018 in the data frame whr_alternate,
then name it as Year2018. Since there is no data in these two columns (Democratic Qua
lity and Delivery Quality), R uses select() to delete these two columns
Year2018= whr_alternate %>%
  filter(Year=="2018") %>%
  select(-Democratic.Quality,-Delivery.Quality)
```

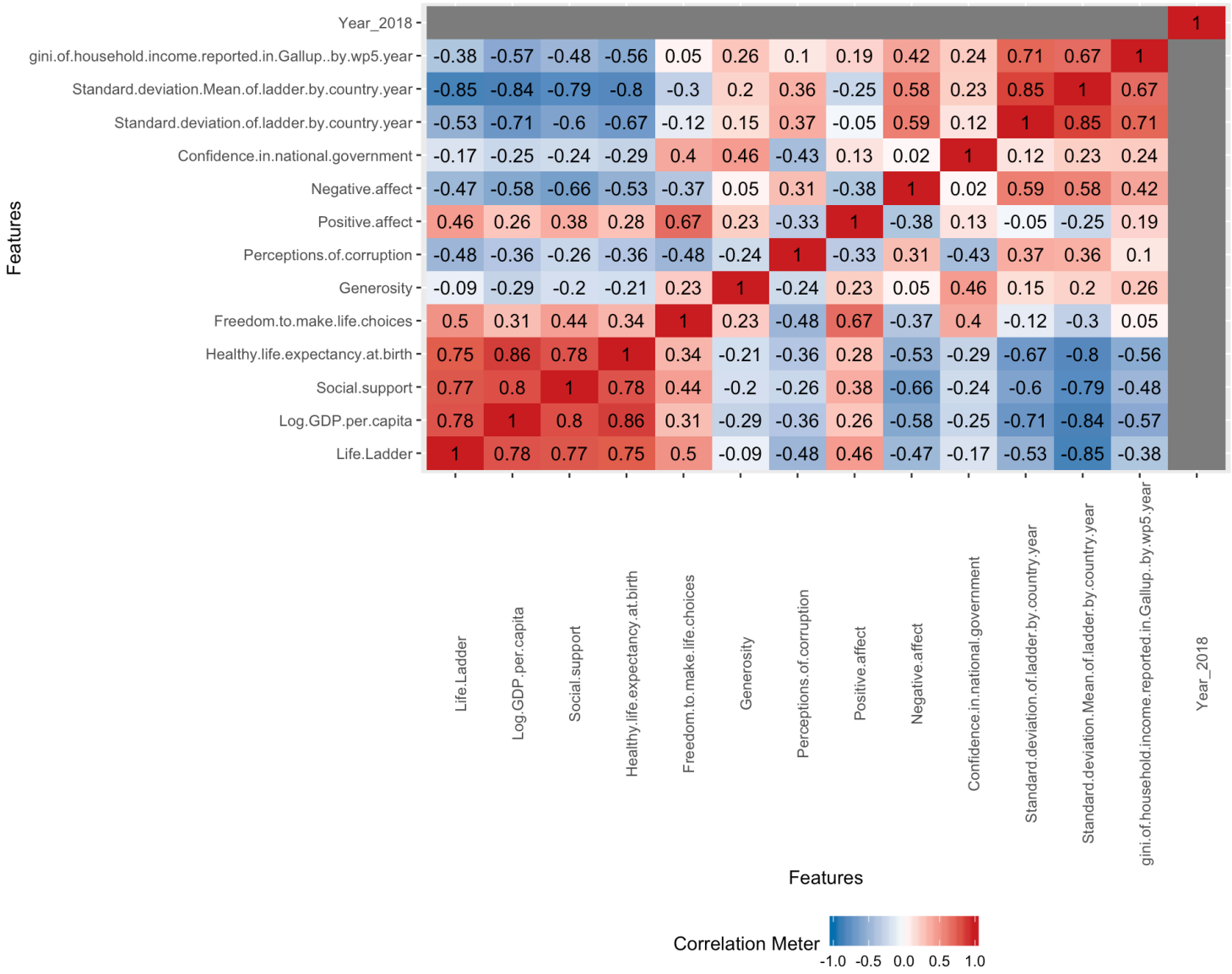
a. Correlation heatmap for the year 2017

```
#create correlation heatmap to directly see correlations between any two variables in
year 2017
#remove missing values by using na.omit()
plot_correlation(Year2017 %>% na.omit())
```



b. Correlation heatmap for the year 2018

```
#create correlation heatmap to directly see correlations between any two variables in year 2018
#remove missing values using na.omit()
plot_correlation(Year2018 %>% na.omit())
```



Individual Intepretation

- According to correlation heatmap, we observed that GDP per capita, healthy life, social life and delivery quality have stronger positive correlations with life ladder in the year 2017.
- For the year 2018, GDP per capita, healthy life, social life have stronger positive correlations with life ladder.
- Overall,GDP per capita, healthy life, social life have stronger positive correlations with life ladder in both years. Perception of corruption and standard deviation mean of ladder by country negatively affected life ladder for both years too.

Data Analysis

- After reviewing individual exploratory data analysis, we notice GDP per capita, healthy life expectancy and social support have stronger correlation with life ladder. All these three factors provide inferences on a country's economic and social welfare. Most importantly, scatter plot about these factors all show

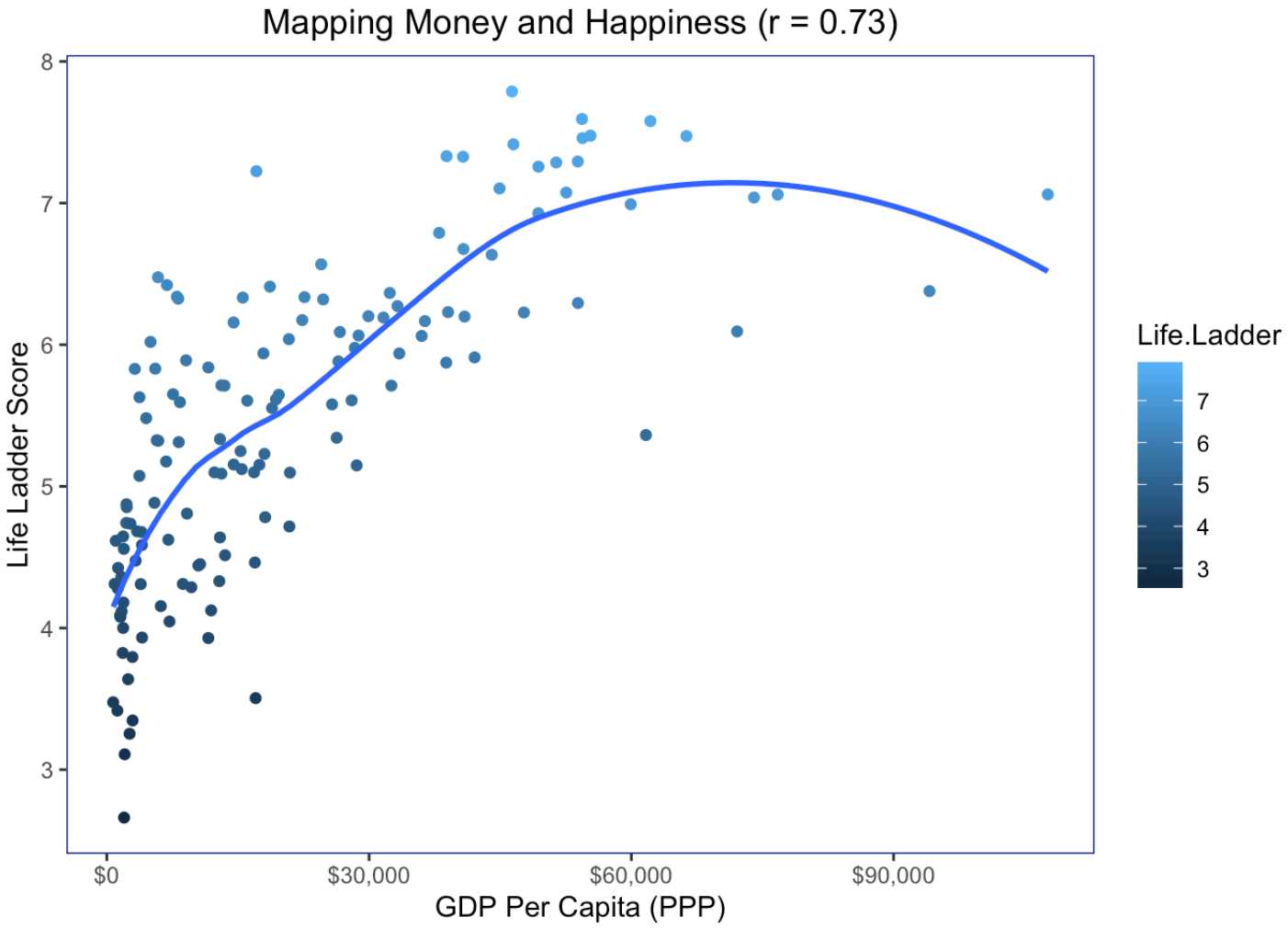
the football shaped plot, meaning values are normally distributed. Each plot is linear, so it is necessary for doing regression method. Therefore, we decide to do hypothesis tests to figure out whether these factors have an impact on life ladder, or it can be explained by chance variation.

1. GDP Per Capita Hypothesis Test

Member: Aneesh Didwania and Pranav Prabhas

- After looking at the initial data set, we saw that its GDP was measured in a log format and was quite difficult to interpret. So, we pulled GDP (PPP) Per Capita data from the World Bank, and merged it with the Data Set provided in class. Thus, it is easier to understand the slopes and correlation coefficients. Also, GDP (PPP) is a more accurate estimator of living standards, because it adjusts for the purchasing power of the local currency.

```
ggplot(newdata, aes(x = GDP.PPP.Per.Capita, y = Life.Ladder, color=Life.Ladder))+ geo
m_point() + labs(title="Mapping Money and Happiness (r = 0.73)",
        x ="GDP Per Capita (PPP)", y = "Life Ladder Score") + geom_smooth(method="loe
ss", se=FALSE) + scale_x_continuous(labels = dollar) + theme(plot.title = element_tex
t(hjust = 0.5), panel.border = element_blank(),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(), panel.background = element_rect(fill =
"white",
                                colour = "darkblue"))
```




```
cor(newdata$GDP.PPP.Per.Capita, newdata$Life.Ladder)
```

```
[1] 0.7324451
```

- The plot above seeks to find a relationship between the GDP of a country and its Happiness. After plotting the scatter, we conducted a polynomial regression to see how strong the association was. We arrived at an r value of 0.73. While this is indicative of a moderately strong relationship, it is not definitive. To be sure, that there was a relationship between the two variables, for the entire population, and not just our sample, we performed a Hypothesis Test.

- **Null Hypothesis**:
    - Results can be explained by chance
    - The relationship between the life ladder of a country and its GDP per Capita (PPP) is due to chance
- **Alternative Hypothesis**:
    - Results cannot be explained by chance
    - GDP per capita has strong efect on Life ladder

a. Statistics

```
rows = length(newdata)
happymean = mean(newdata$Life.Ladder)
happymean
```

```
[1] 5.48133
```

```
gdpmean = mean(newdata$GDP.PPP.Per.Capita)
gdpmean
```

```
[1] 21687.94
```

```
happysd = sd(newdata$Life.Ladder) * sqrt((rows-1)/rows)
happysd
```

```
[1] 1.040583
```

```
gdpsd = sd(newdata$GDP.PPP.Per.Capita) * sqrt((rows-1)/rows)
gdpsd
```

```
[1] 19306.93
```

```
r = cor(newdata$Life.Ladder, newdata$GDP.PPP.Per.Capita)
r
```

```
[1] 0.7324451
```

```
slope = r * happysd/gdpsd
slope
```

```
[1] 3.947649e-05
```

```
se = (sqrt(1-(r*r))*happysd)/(sqrt(rows - 2)*gdpsd)
se
```

```
[1] 1.834719e-05
```

```
z = slope/se
z
```

```
[1] 2.151636
```

```
pvalue=2*pnorm(-abs(z))
pvalue
```

```
[1] 0.031426
```

Interpretation for Statistics

- We use the z-test to calculate the test statistics. The observed value is the slope, and the expected value is zero. We multiplied the SD by the correction factor to find the SD of the population, not the sample.We use the data found before to calculate the P-value.

b. Linear Model Method

```
summary(lm(newdata$Life.Ladder ~ newdata$GDP.PPP.Per.Capita))
```

```

Call:
lm(formula = newdata$Life.Ladder ~ newdata$GDP.PPP.Per.Capita)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.04145 -0.51128  0.00285  0.55838  1.92457 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                4.625e+00  9.446e-02   48.96   <2e-16 ***
newdata$GDP.PPP.Per.Capita 3.948e-05  3.124e-06   12.64   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7789 on 138 degrees of freedom
Multiple R-squared:  0.5365,    Adjusted R-squared:  0.5331 
F-statistic: 159.7 on 1 and 138 DF,  p-value: < 2.2e-16
```

Interpretation for Linear Model

- We compare P-value found by doing statsitics and p-value obtained by linear model.
- Since the P-value is smaller than the significance level of the 0.01, we reject the null hypothesis in favor of the alternative. We conclude “there is sufficient evidence to conclude that there is a correlation in the

population between the predictor GDP per Capita (PPP) and the Life Ladder."

2. Healthy Life Expectancy At Birth & life Ladder Hypothesis Test

Member: Dorcas Cheung & Dorothy Zhu

- **Null Hypothesis**:
  - Results can be explained by chance
  - Correlation between life expectancy at birth and the life ladder is zero
- **Alternative Hypothesis**:
  - Results are not due to chance
  - Correlation between life expectancy at birth and the life ladder is greater than zero

a. Statistics

```
Year2018_=na.omit(Year2018)
r1=cor(Year2018_$Healthy.life.expectancy.at.birth,Year2018_$Life.Ladder)
r1
```

```
[1] 0.7543433
```

```
SDofhh=sd(Year2018_$Healthy.life.expectancy.at.birth)
SDofhh
```

```
[1] 6.811616
```

```
SDofll=sd(Year2018_$Life.Ladder)
SDofll
```

```
[1] 1.125497
```

```
m1=r1*SDofll/SDofhh
m1
```

```
[1] 0.1246416
```

```
se1=(sqrt(1-(r^2))*SDofll)/(sqrt(110 - 2)*SDofhh)
se1
```

```
[1] 0.01082476
```

```
z1=(m1-0)/se1
z1
```

```
[1] 11.51449
```

```
1-pnorm(z1)
```

```
[1] 0
```

Interpretation For Statistics

- Because there are more than 25 draws, we use z-test to analyze correlation of healthy life expectancy and life ladder in 2018. We acknowledge that both years are dependent.
- SD of healthy life expectancy (x) is small the estimation of the slope is fairly accurate as x varies a lot.
- n-2 takes account of the fact that with more data, the estimate is more accurate
- Since data of the life ladder is normal, we will use normal approximation. There is no need to use continuity correction because it is continuous.
- P-value is 0.

b. Linear Model Method

```
summary(lm(Year2018_$Life.Ladder~Year2018_$Healthy.life.expectancy.at.birth))
```

```
Call:
lm(formula = Year2018_$Life.Ladder ~ Year2018_$Healthy.life.expectancy.at.birth)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.54896 -0.50480  0.04867  0.53385  1.60421 

Coefficients:
                                          Estimate Std. Error t value
(Intercept)                               -2.50559    0.68033  -3.683
Year2018_$Healthy.life.expectancy.at.birth  0.12464    0.01044  11.942
                                          Pr(>|t|)    
(Intercept)                               0.000362 ***
Year2018_$Healthy.life.expectancy.at.birth  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7423 on 108 degrees of freedom
Multiple R-squared:  0.569, Adjusted R-squared:  0.565 
F-statistic: 142.6 on 1 and 108 DF,  p-value: < 2.2e-16
```

Interpretation For Linear Model Method

- P-value is the same for both statsics and linear model. P-value is close to 0.
- Reject null at the 1% significant level. The results are not likely due to chance. There is a positive association between the healthy life expectancy at birth and the life ladder.

3. Social Support & Life Ladder Hyoothesis Test

Member: Lanjing Wang & Zhengan Lyu

- **Null Hypothesis**:
  - Differences are due to chance
  - Social support has no effect on life ladder =0
- **Alternative Hypothesis**:
  - Differences are not due to chance
  - The correlation between social support and life ladder is >0

a. Statistics

```
r2=cor(whr_alternate1$Social.support,whr_alternate1$Life.Ladder)
r2
```

```
[1] 0.7703844
```

```
SDy=sd(whr_alternate1$Life.Ladder)
SDy
```

```
[1] 1.148219
```

```
SDx=sd(whr_alternate1$Social.support)
SDx
```

```
[1] 0.1255978
```

```
m=r2*SDy/SDx
m
```

```
[1] 7.042874
```

```
numerator=sqrt(1-r2^2)*SDy
deminator=sqrt(128-2)*SDx
SE=numerator/deminator

z2=(m-0)/SE
z2
```

```
[1] 13.56308
```

```
1-pnorm(z2)
```

```
[1] 0
```

Interpretation for statistical steps

- Z test helps us determine whether this is by chance. After calculating z, P-value is essentially 0. This conclusion is that this is not just random, and there is a correlation between social support and life ladder.
- In this hypothesis test, it does not help us decide whether there are confounding factors or other explaination, only whether the results can be explainated by chance.

b. Linear Model Method

```
#create linear model about predicting life ladder from social support
lm.out1=lm(Life.Ladder ~ Social.support, data=whr_alternate1)
summary(lm.out1)
```

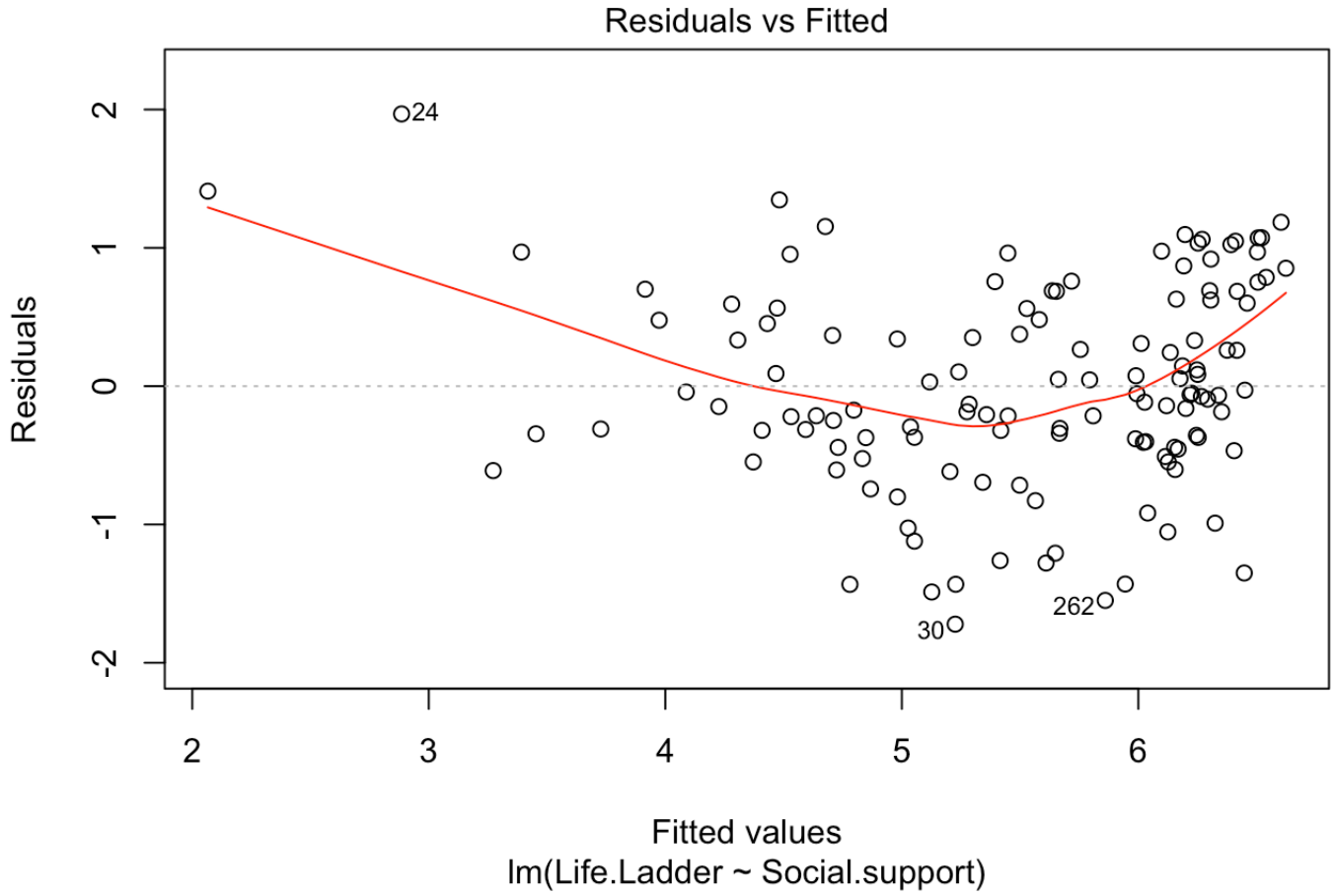
```

Call:
lm(formula = Life.Ladder ~ Social.support, data = whr_alternate1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.72102 -0.41543 -0.06384  0.59495  1.96819 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -0.1848     0.4241  -0.436    0.664    
Social.support   7.0429     0.5193  13.563   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.735 on 126 degrees of freedom
Multiple R-squared:  0.5935,    Adjusted R-squared:  0.5903 
F-statistic:   184 on 1 and 126 DF,  p-value: < 2.2e-16
```

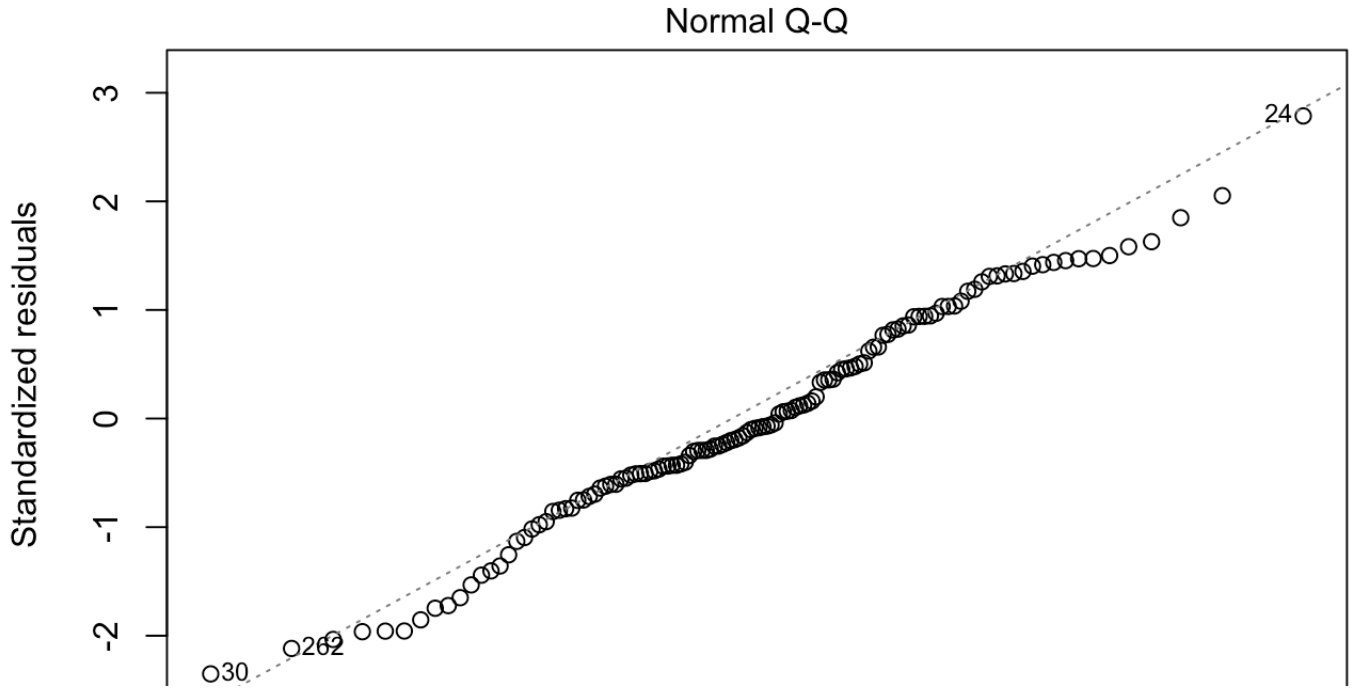
Interpretation for Linear Model

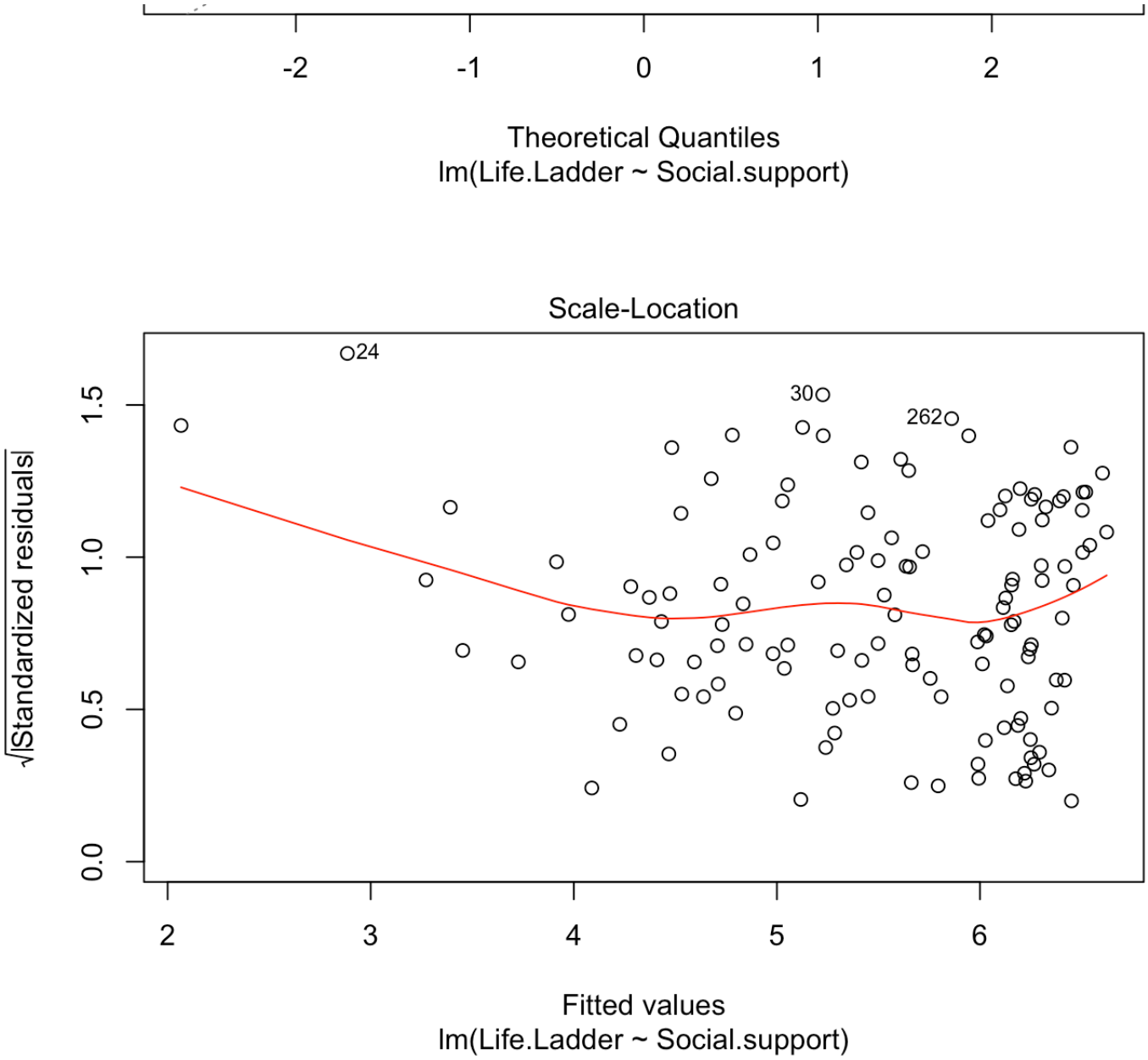
- P-value is 2e-16 < 0.01, indicating that results are not due to chance, null hypothesis seems to be failed. Social support has effect on life ladder. For every % increases in social support, life ladder increases by 7.04%
- From multiple R-squared, 55% of the variations in life ladder can be explained by social support according to linear model 2. 55% is expalained, 45% is unexplained. We should include more predictors in the model.
- Recall the correlation heatmap, GDP and healthy life are stronger correlations with life lader. Therefore, we should use these additional predictors for life ladder.

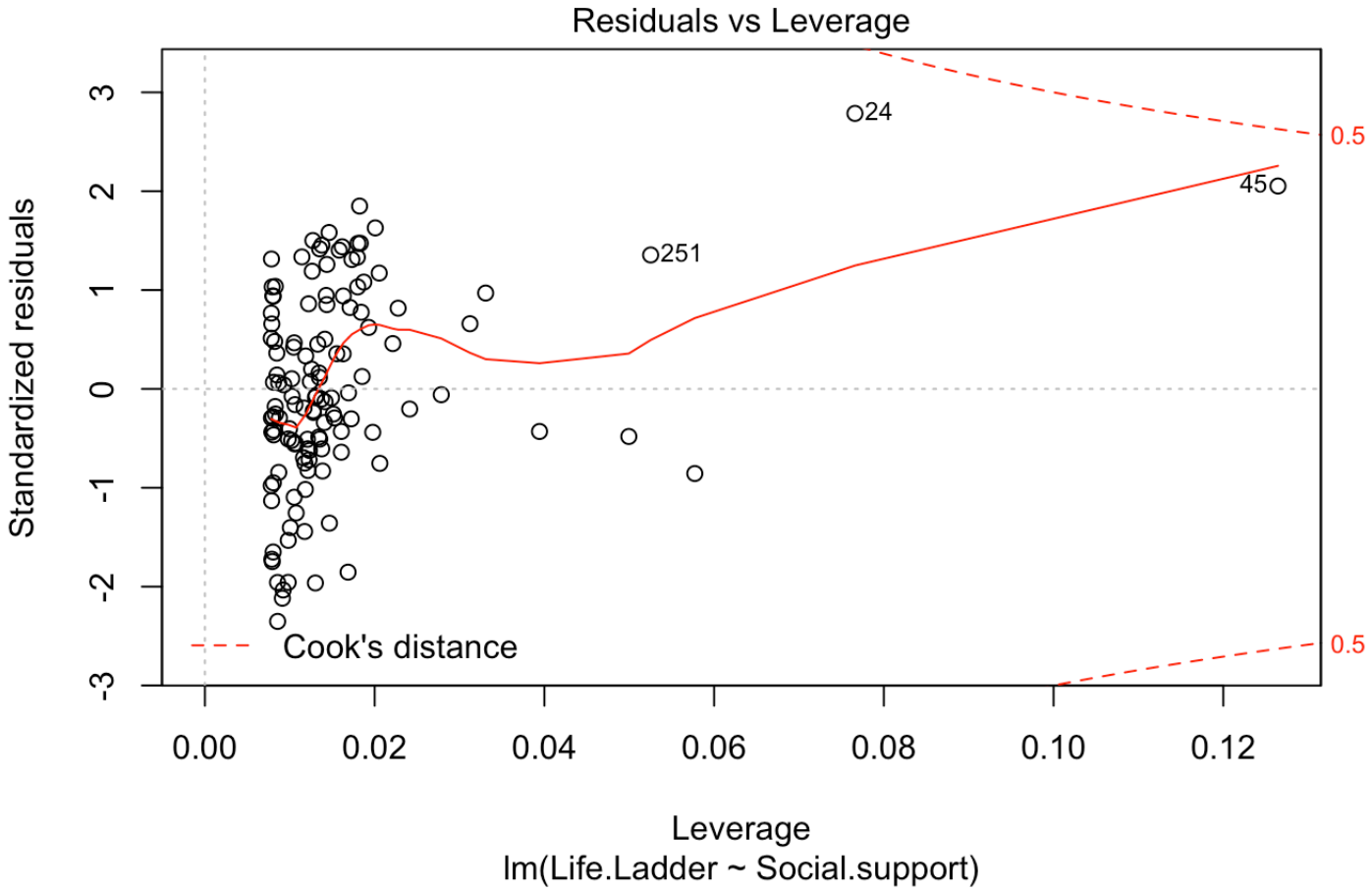
c. Check for model performance

```
plot(lm.out1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Life.Ladder ~ Social.support)
24
30
262
-2
-1
0
1
2
2
3
4
5
6


Normal Q-Q
Standardized residuals
24
30
262
-2
-1
0
1
2
3


-2
-1
0
1
2
Theoretical Quantiles
lm(Life.Ladder ~ Social.support)
Scale-Location
24
30
262
√|Standardized residuals|
0.0
0.5
1.0
1.5
2
3
4
5
6
Fitted values
lm(Life.Ladder ~ Social.support)



Interpretation for plots

- The first plot is about the residuals and fitted value. Residuals are errors which are differences between observed life ladder and predictted life ladder. Fitted values are predicted values of life ladder from social support. It shows random scatter pattern (even though the red line seem to be little curved). It suggests that the model is relatively a good fit for the data.
- The normal Q-Q plot displays a straight line pattern with small departures from normality. It supports the conclusion that the model is a good fit. Life ladder variable can be considered normally distributed.
- Standardized residuals use standard unit, so it is easy to compare. This plot shows that fitted value and standardlized residuals are correlated as displayed in the random scatter plot.
- Leverage is a measure of how far away the independent variable (social support) values an observation are from the other observations. This plot shows a fewer observations with high leverage values. For example, observation 24, 25 and 45. The predicted life ladder for these countries are lower than real observations.

Conclusion

- In conclusion, we found out that GDP, healthy life expectancy and social support have an effect on life

ladder respectively. We take a look at multiple R-squared value for all these three hypothesis tests, we noticed that they are all around 0.5 to 0.6. For the multiple R-squared value of 53%, it says that 53% of the variations in the life ladder can be explained by GDP according to a linear model. For the multiple R-squared of 56%, it says that 56% of the variations in the life ladder can be explained by healthy life expectancy at birth according to the linear mode. For the multiple R-squared 59%, it says that 59% of the variations in the life ladder can be explained by social support. Therefore, knowing only one factor of social support, healthy life expectancy or GDP cannot make 100% prediction for life ladder.

Reference

- https://data.worldbank.org/indicator/NY.GDP.PCAP.CD (https://data.worldbank.org/indicator/NY.GDP.PCAP.CD)