Implementation of Correlation in R

Asst. Prof. Ashwini Mathur

For Example

This is the idea that two things that we measure can be somehow **related to one another.**

For example, your personal happiness, **which we could try to measure say with a questionnaire**, might be **related to other thing**s in your life that we could also measure, such as number of close friends, yearly salary, how much chocolate you have in your bedroom, or how many times you have said the word Nintendo in your life.

Some of the relationships that we can **measure are meaningful**, and might reflect a causal relationship where one thing causes a change in another thing. Some of the relationships are spurious, and do not reflect a causal relationship.

Goals

1. Compute Correlation between two variables using software.
2. Discuss the possible meaning of correlations that you observe.

Note: use data from the World Happiness Report. A .csv of the data can be found here: WHR2018.csv

R

In this session, we use explore to **explore correlations** between any two variables, and also show **how to do a regression line**.

There will be three main parts.

Getting R to compute the correlation, and looking at the data using scatter plots. We'll look at some correlations from the World Happiness Report.

Then correlations using data we collect from ourselves

For Correlation (Command : cor)

R has the `cor` function for computing Pearson's r between any two variables. In fact this same function computes other versions of correlation, but we'll skip those here. To use the function you just need two variables with numbers in them like this:

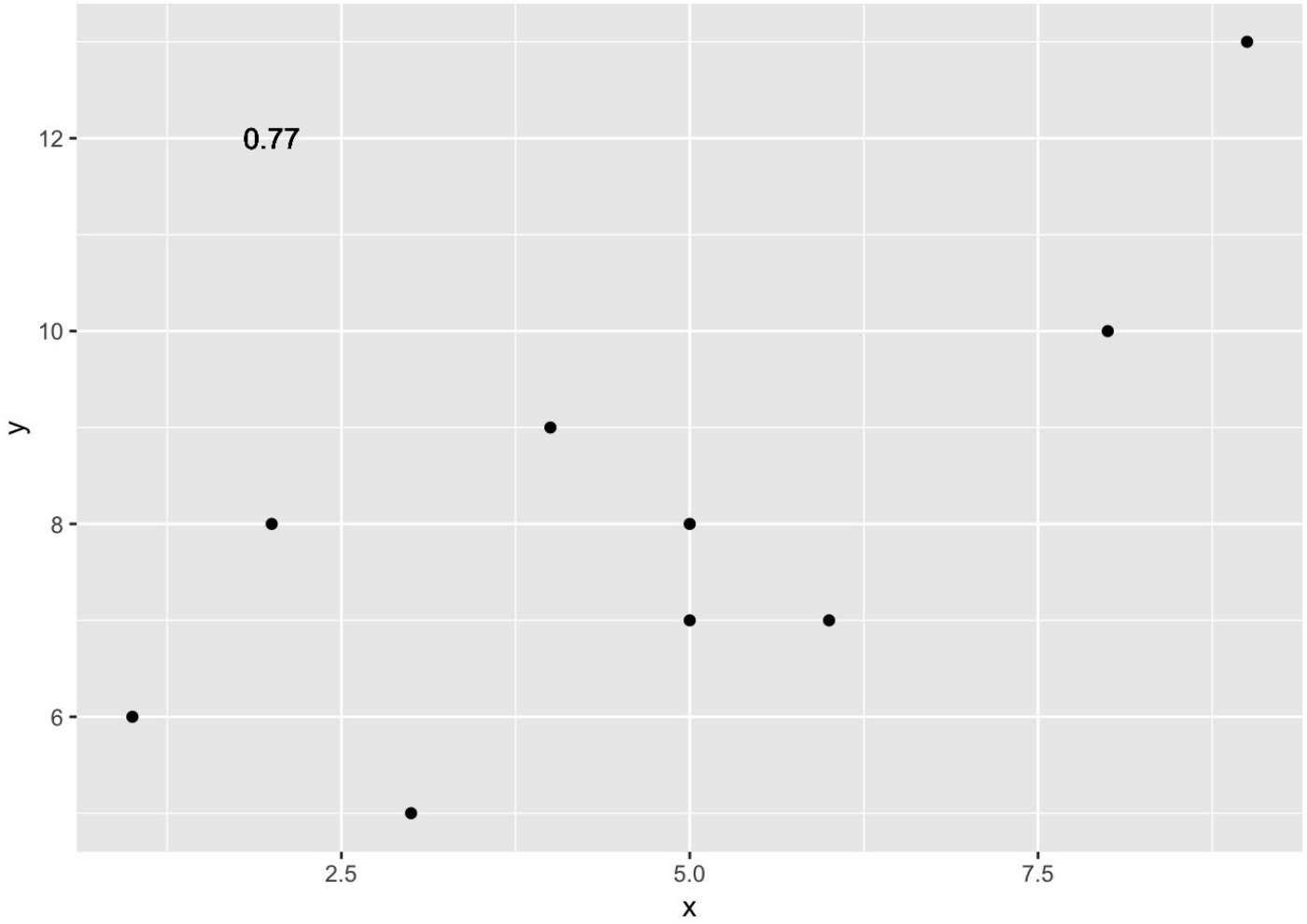
```
x  <- c(1,3,2,5,4,6,5,8,9)

y  <- c(6,5,8,7,9,7,8,10,13)

 cor(x,y)

 ## [1] 0.76539
```

Scatter Plots

Plot the data in a scatter plot using ggplot2, and let’s also return the correlation and print it on the scatter plot. Remember, ggplot2 wants the data in a data.frame, so we first put our x and y variables in a data frame.

Program in R

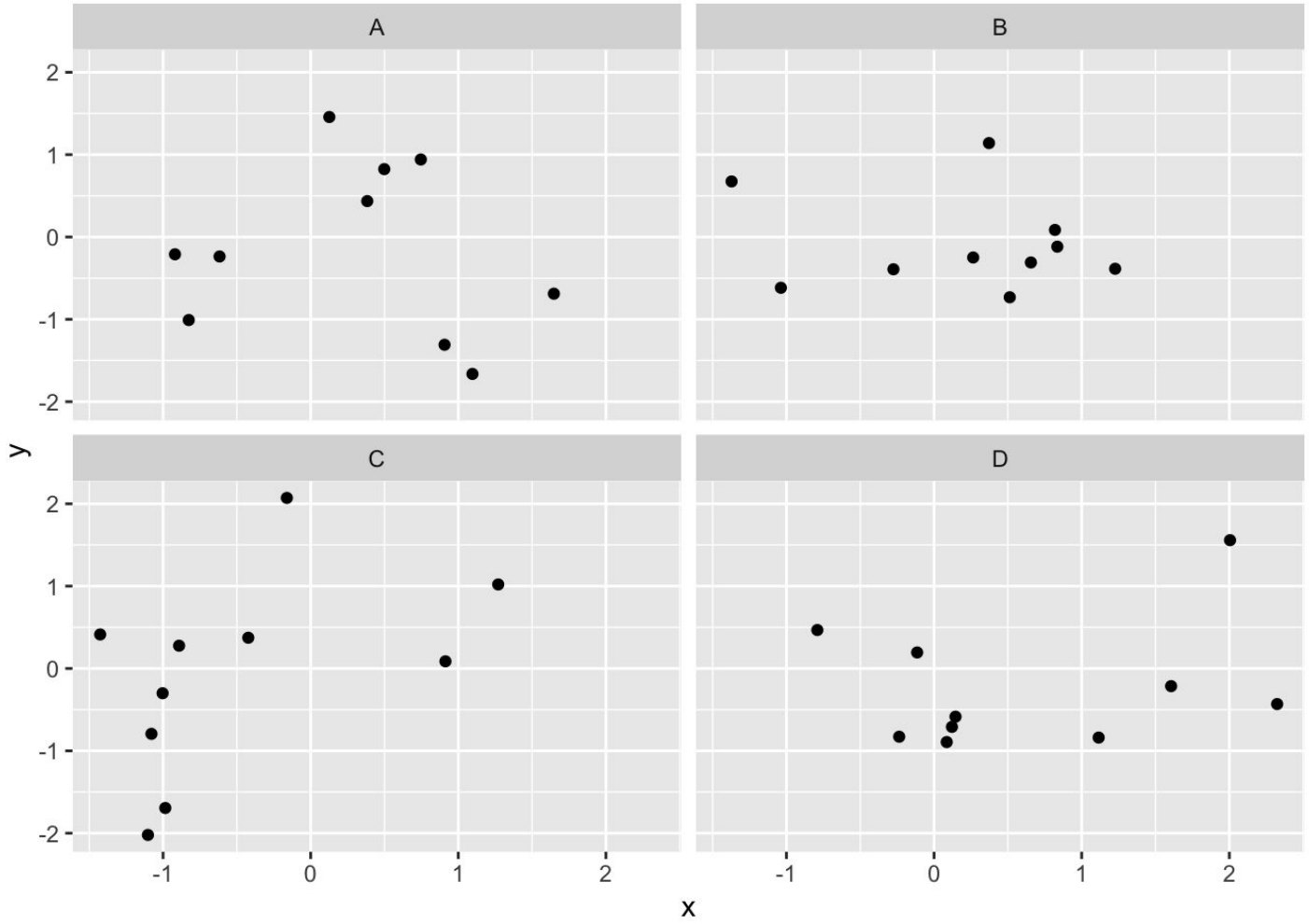
```
library(ggplot2)

# create data frame for plotting

my_df <- data.frame(x,y)

# plot it

ggplot(my_df, aes(x=x,y=y))+

  geom_point()+

  geom_text(aes(label = round(cor(x,y), digits=2), y=12, x=2 ))
```


0.77
12
10
8
6
y
2.5
5.0
7.5
x

Lots of Scatterplots

Before we move on to real data, let’s look at some fake data first. Often we will have many measures of X and Y, split between a few different conditions, for example, A, B, C, and D. Let’s make some fake data for X and Y, for each condition A, B, C, and D, and then use facet_wrapping to look at four scatter plots all at once

```r
x<-rnorm(40,0,1)

y<-rnorm(40,0,1)

conditions<-rep(c("A","B","C","D"), each=10)

all_df <- data.frame(conditions, x, y)

ggplot(all_df, aes(x=x,y=y))+

  geom_point()+

  facet_wrap(~conditions)
```


A
B
C
D
y
x
2
1
0
-1
-2
-1
0
1
2

Let's Start with the Real Data-set

Overview

Let’s take a look at some correlations in real data. We are going to look at responses to a questionnaire about happiness that was sent around the world, from the world happiness report

Load the data

```
library(data.table)
```

```
whr_data <- fread('data/WHR2018.csv')
```

Look at the data

```
library(summarytools)
```

```
view(dfSummary(whr_data))
```

You should be able to see that there is **data for many different countries, across a few different years**. There are lots of different kinds of measures, and each are given a name.

Problem to Solve with the help of Data-Science

Lets formulate the Problem First

My Question

For the year 2017 only, does a countries measure for “freedom to make life choices” correlate with that countries measure for " Confidence in national government"?

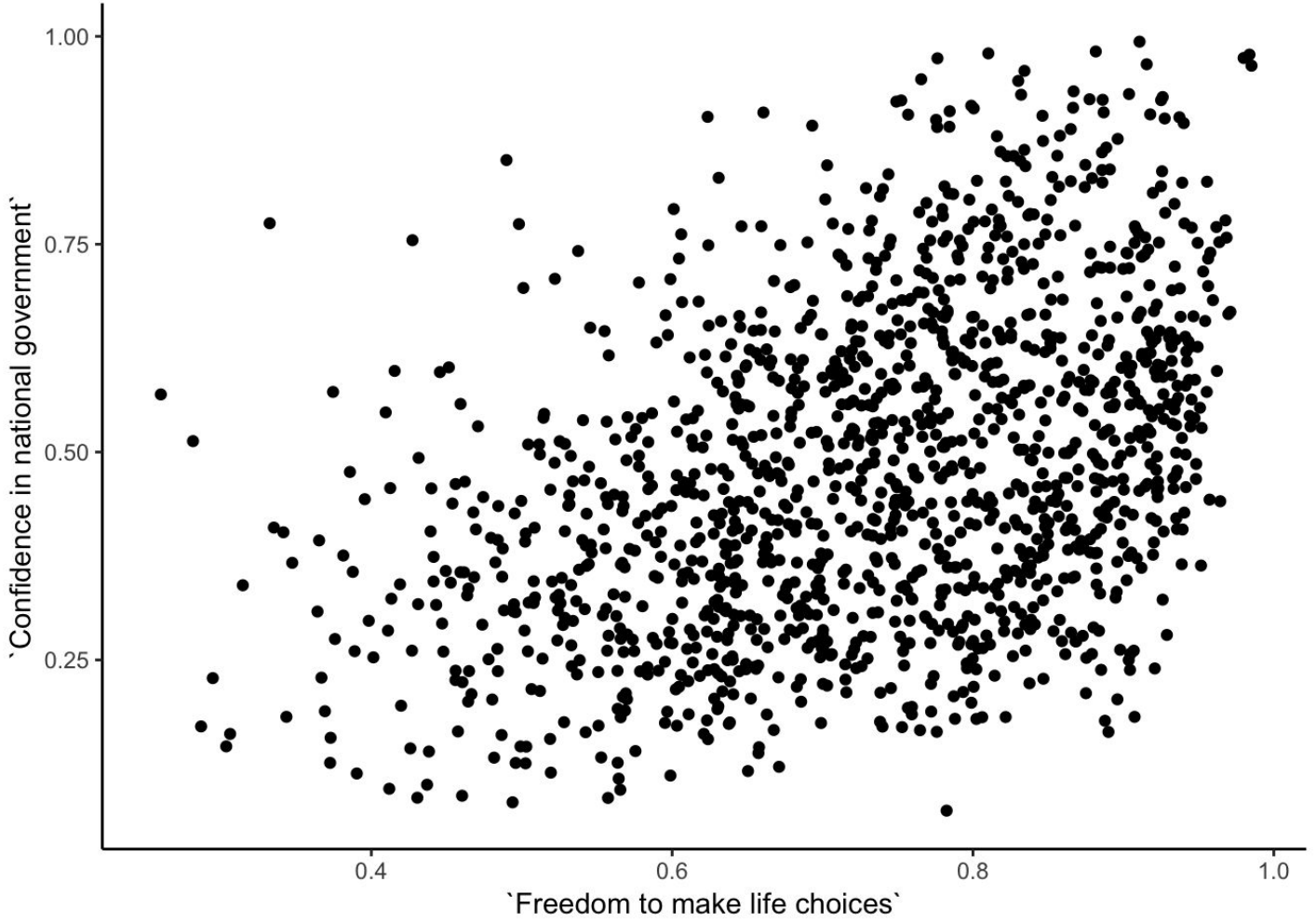
Let's find out. We calculate the correlation, and then we make the scatter plot.

```
cor(whr_data$`Freedom to make life choices`,
    whr_data$`Confidence in national government`)
```

```
## [1] NA
```

Lets try with scatter plot

```
ggplot(whr_data, aes(x=`Freedom to make life choices`,
                     y=`Confidence in national government`))+
  geom_point()+
   theme_classic()
```


1.00
0.75
0.50
0.25
`Confidence in national government`
0.4
0.6
0.8
1.0
`Freedom to make life choices`

Interesting, what happened here? We can see some dots, but the correlation was NA (meaning undefined).

This occurred because there are some missing data points in the data.

We can remove all the rows with missing data first, then do the correlation. We will do this a couple steps, first creating our own data.frame with only the numbers we want to analyse. We can select the columns we want to keep using `select`. Then we use `filter` to remove the rows with NAs.

```r
library(dplyr)

smaller_df <- whr_data %>%
                  select(country,
                         `Freedom to make life choices`,
                         `Confidence in national government`) %>%
                  filter(!is.na(`Freedom to make life choices`),
                         !is.na(`Confidence in national government`))

cor(smaller_df$`Freedom to make life choices`,
    smaller_df$`Confidence in national government`
```

Answer

```
## [1] 0.4080963
```

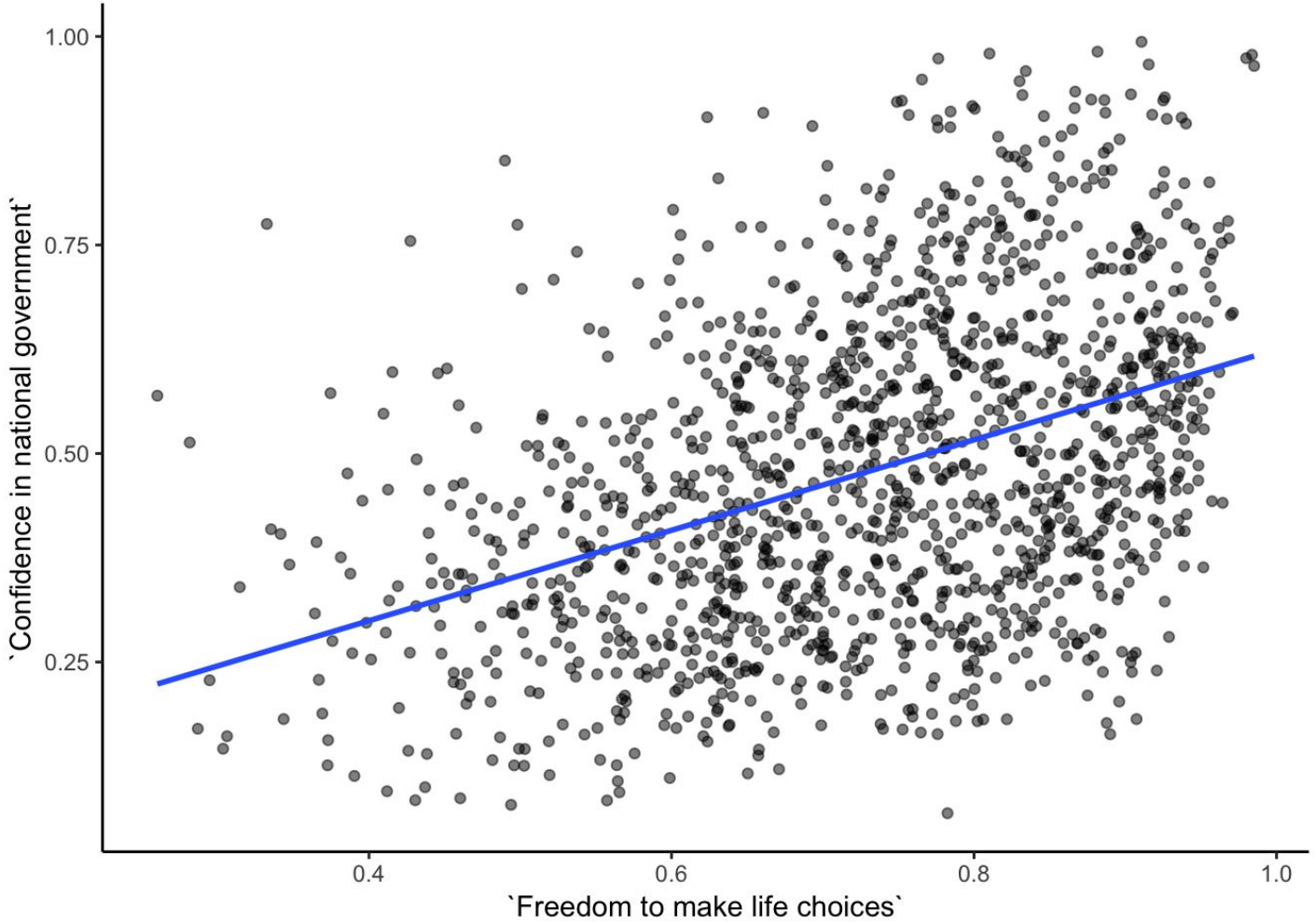
Now we see the correlation is .408.

Although the scatter plot shows the dots are everywhere, **it generally shows that as Freedom to make life choices increases in a country, that countries confidence in their national government also increase**. **This is a positive correlation**.

Let's do this again and add the best fit line, so the trend is more clear, we use `geom_smooth(method=lm, se=FALSE)`. I also change the `alpha` value of the dots so they blend it bit, and you can see more of them

Plot the data with best fit line.

```
ggplot(smaller_df, aes(x=`Freedom to make life choices`,
                       y=`Confidence in national
government`))+
  geom_point(alpha=.5)+
  geom_smooth(method=lm, se=FALSE)+
   theme_classic()
```


1.00
0.75
0.50
0.25
`Confidence in national government`
0.4
0.6
0.8
1.0
`Freedom to make life choices`

Example 2.

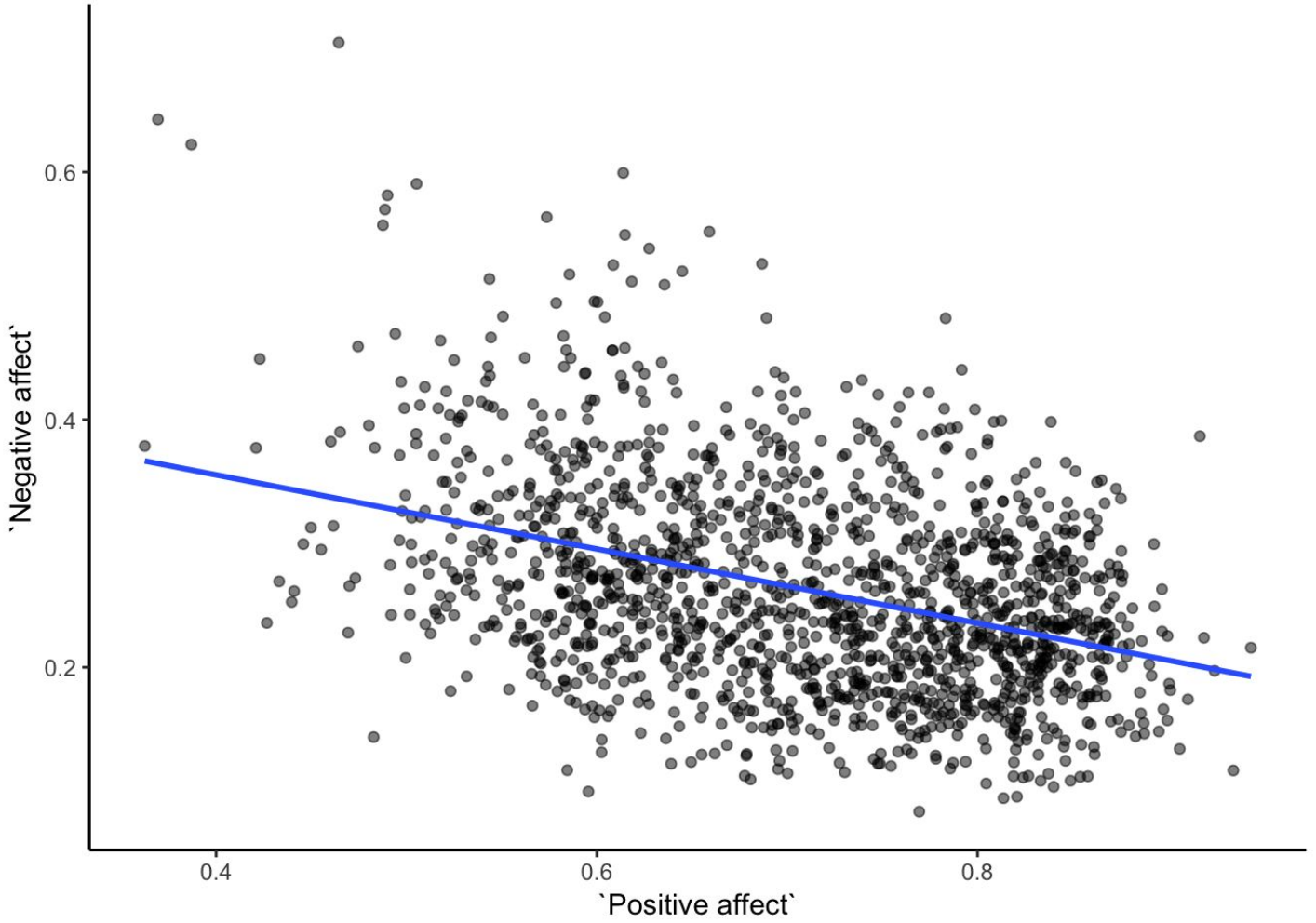
what is the relationship between positive affect in a country and negative affect in a country. I wouldn’t be surprised if there was a negative correlation here: when positive feelings generally go up, shouldn’t negative feelings generally go down?

```r
# select DVs and filter for NAs

smaller_df <- whr_data %>%

                select(country,

                       `Positive affect`,

                       `Negative affect`) %>%

                filter(!is.na(`Positive affect`),

                       !is.na(`Negative affect`))

# calcualte correlation

cor(smaller_df$`Positive affect`,

    smaller_df$`Negative affect`)
```

Answer

```
## [1] -0.3841123
```

```r
# plot the data with best fit line


ggplot(smaller_df, aes(x=`Positive affect`,
                       y=`Negative affect`))+
  geom_point(alpha=.5)+
  geom_smooth(method=lm, se=FALSE)+
   theme_classic()
```


`Negative affect`
0.6
0.4
0.2
0.4
0.6
0.8
`Positive affect`