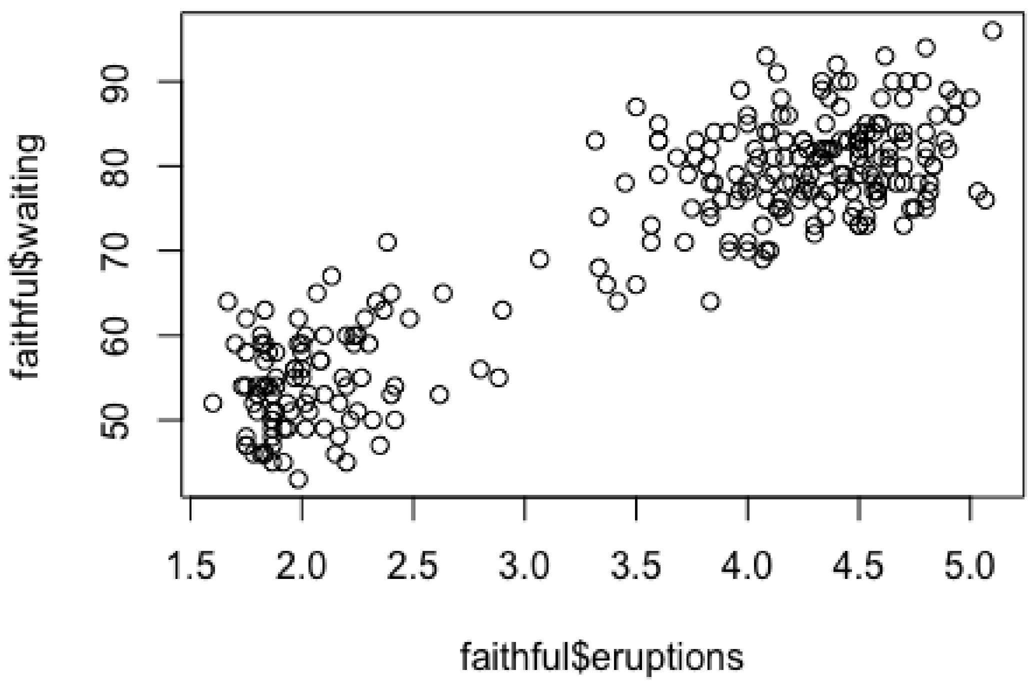
chapter 4.cor.and.plots

kim

September 6, 2023

What is the relationship between two variables? Examples: exercise and metabolism literacy and under5rate life expectancy and literacy Give some of your own examples from your field or your dataset. From chapter 0 what is one extra thing we need to consider when looking at the relationship between two variables? (hint:lu.va.)

```
plot(faithful$eruptions,faithful$waiting)
```



```
cfth=cor(faithful$eruptions,faithful$waiting)
cfth
```

```
## [1] 0.9008112
```

The above graph shows the faithful data plotted with waiting as a function of eruption length. Do you think there is a strong tendency? Does the graph indicate a relationship between the two? cor stands for correlation coefficient.

What does the correlation coefficient tell you-in general and in this case?

Relationship between correlation coefficient and linear association? +0.30. A weak uphill (positive) linear relationship

+0.50. A moderate uphill (positive) relationship

+0.70. A strong uphill (positive) linear relationship

Exactly +1. A perfect uphill (positive) linear relationship (From stats for dummies)

Similarly for the negative correlation coefficients

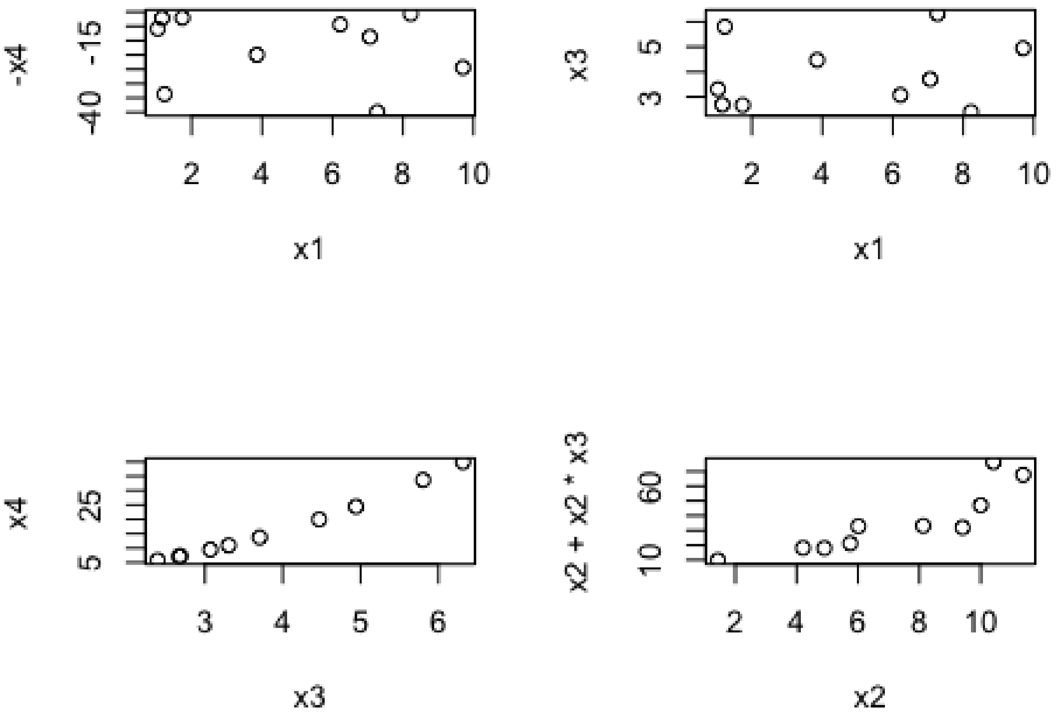
Formula: $cor\ = \frac{1}{(n-1)}\sum\frac{\left(x_i-\bar{x}\right)}{s_x}\frac{\left(y_i-\bar{y}\right)}{s_y}$ (Picture showing the center of the dots as the center of axis and what effect points in different quadrants have on linearity and the correlation coefficient)

Does correlation depend on the order of the variables?

What unit does correlation have?

More graphs

```r
#just to get some points
x1=rchisq(10,6)
x2=x1+2*rnorm(10,1,1)
x3=rnorm(10,3,2)
x4=x3^2

par(mfrow=c(2,2))
plot(x1,-x4)
plot(x1,x3)
plot(x3,x4)
plot(x2,x2+x2*x3)
```



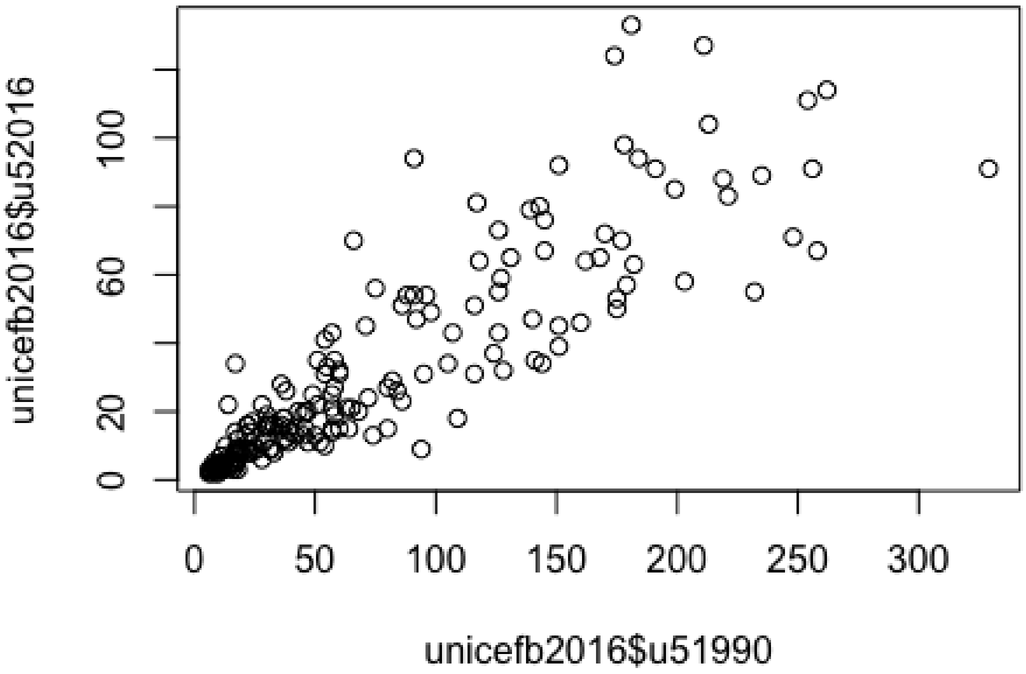
Describe the association.

Below are the associated correlations

```
cor(x1,-x4)

## [1] -0.1751277

cor(x1,x3)

## [1] 0.1794724

cor(x3,x4)

## [1] 0.9930205

cor(x2,x2+x2*x3)

## [1] 0.8734483
```

Look at Unicef data

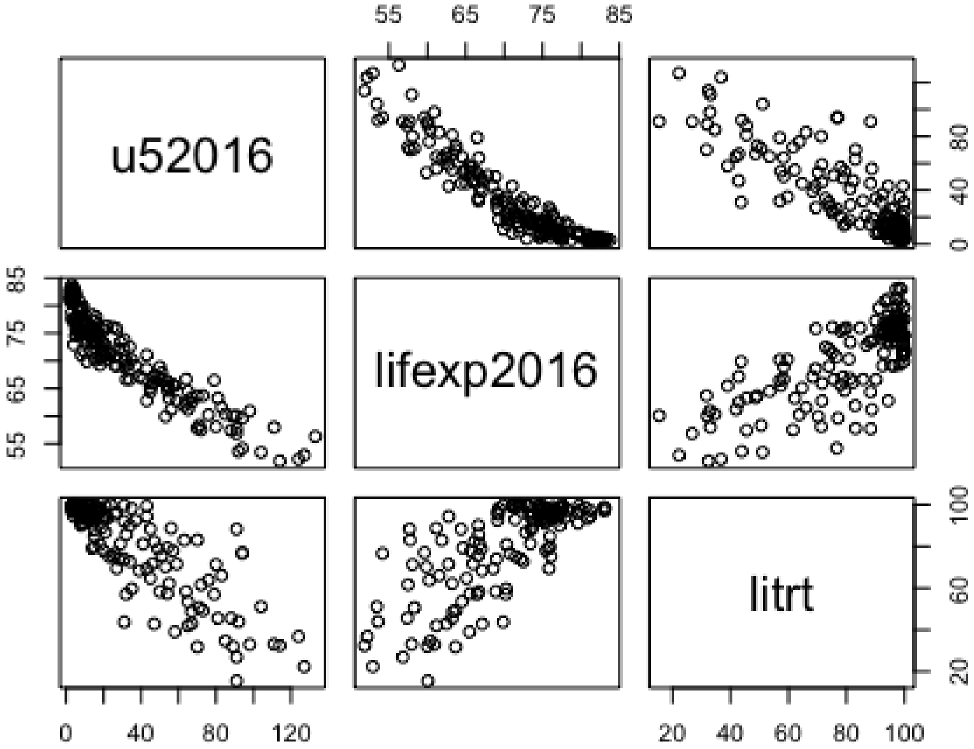
```
plot(unicefb2016$u51990,unicefb2016$u52016)
```



```
cor(unicefb2016$u51990,unicefb2016$u52016,use= "pairwise.complete.obs")  # to
deal with NA's

## [1] 0.8827976
```

Check on a point: This code works only in the console, not in an Rmarkdown-at least not that I know of: plot(unicefb2016$u51990, unicefb$2016u52016) identify(unicefb2016 $u51990, unicefb$2016u52016,labels=unicefb2016\$countries.and.areas) This will tell you which country your mouse is pointing to. Try it!

```
plot(unicefb2016[c("u52016","lifexp2016","litrt")])
```



```
#this creates a grid of scatter plots for the three listed variables
```

Explanation of each graph.

Tendency? Strong association? Positive or negative? Linear or not?

And below we have the correlations.

```
cor(unicefb2016[c("u52016","lifexp2016","litrt")],use="pairwise.complete.obs"
)

##                u52016 lifexp2016      litrt
## u52016      1.0000000 -0.9293870 -0.8372457
## lifexp2016 -0.9293870  1.0000000  0.7478669
## litrt      -0.8372457  0.7478669  1.0000000
```

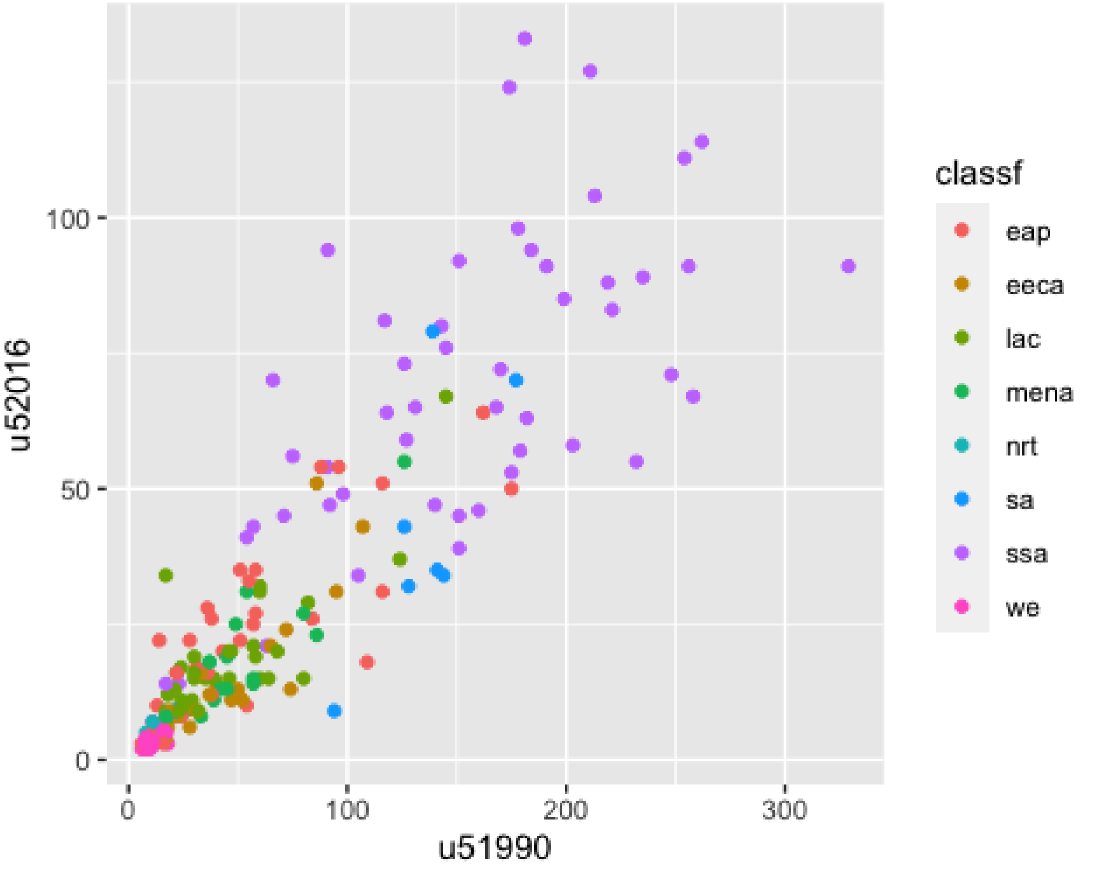
Try some other combinations of variables.

Try a couple of variables from your own data

Splitting up the unicef graph according to regions. We will use ggplot here as it looks much better and is easier to use.

```
ggplot(unicefb2016,aes(u51990,u52016,col=classf)) +geom_point()
```

```
## Warning: Removed 7 rows containing missing values (`geom_point()`).
```



What can you determine from the above graph? Is there strong linear association? Does it vary by region?

You will just want to copy and modify the above code if you use it. The dataset we are using is unicefb2016, the variables we are plotting are u51990, u52016, the points are being colored by classf. We are plotting points.

```
#You would just want to copy and modify this code.
by(unicefb2016[,c("u51990","u52016")],unicefb2016$classf,function(x){cor(x$u5
1990,x$u52016,use="pairwise.complete.obs")})
```

```
## unicefb2016$classf: eap
## [1] 0.8210564
## ------------------------------------------------------------
## unicefb2016$classf: eeca
## [1] 0.8698767
## ------------------------------------------------------------
## unicefb2016$classf: lac
## [1] 0.7858059
## ------------------------------------------------------------
## unicefb2016$classf: mena
## [1] 0.8918164
## ------------------------------------------------------------
## unicefb2016$classf: nrt
## [1] 1
## ------------------------------------------------------------
## unicefb2016$classf: sa
## [1] 0.7334938
## ------------------------------------------------------------
## unicefb2016$classf: ssa
## [1] 0.6342725
## ------------------------------------------------------------
## unicefb2016$classf: we
## [1] 0.5055257
```

```
cor(unicefb2016$u51990,unicefb2016$u52016,use="pairwise.complete.obs")
```

```
## [1] 0.8827976
```

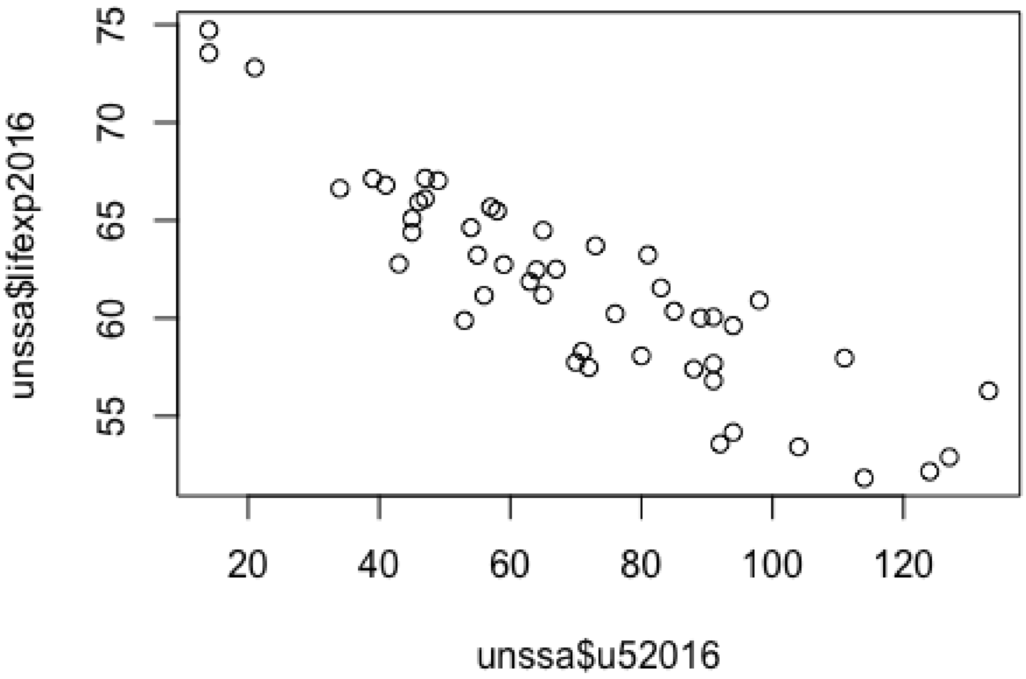
```
#another way to code to get correlation by group:
unicefb2016%>%group_by(classf)%>%summarize(cr=cor(u52016,lifexp2016,use="pairwise.complete.obs"))
```

```
## # A tibble: 8 × 2
##   classf     cr
##   <fct>   <dbl>
## 1 eap    -0.859
## 2 eeca   -0.630
## 3 lac    -0.824
## 4 mena   -0.804
## 5 nrt    -1
## 6 sa     -0.929
## 7 ssa    -0.890
## 8 we     -0.487
```

Modify the above code for graphs and correlation for u52016 and lifexp2016. What do you notice?

Next restrict to one region.

```
unssa=unicefb2016%>%filter(classf=='ssa') # create new dataset from old by just getting the ones with classf ssa, ie the countries in subsaharan Africa.
plot(unssa$u52016,unssa$lifexp2016)
```



```
cor(unssa$u52016,unssa$lifexp2016)

## [1] -0.8904534

#What would happen to the correlation if the three points in the upper left
were removed? Compute the correlation in this case.
```

Try other regions and variables.

Homework type questions: 4.28 ebola and gorillas 4.32 poverty and life expectancy