# Exercise 1 – Sharon Hadar 201023991 – 8.8.17

## Q1

## mean\_and\_sd <- function(x,use.sum=T)

{

if(is.matrix(x))

{

x<- as.vector(x)

}

if(use.sum)

{

xmean = sum(x)/length(x)

xsd = sqrt(sum((x-xmean)^2)/(length(x)-1))

}

else

{

xlength = 0

xsum = 0

for(cell in x)

{

xlength<-xlength+1

xsum <- xsum+cell

}

xmean = xsum/xlength

xsqrt=0

for(cell in x)

{

xsqrt <- xsqrt + (cell-xmean)^2

}

xsd = sqrt(xsqrt/(xlength-1))

}

return (list(mean = xmean, sd = xsd))

}

x=matrix(1:1000, nrow = 1000, ncol = 1000)

system.time(mean\_and\_sd(x,use.sum=F))["elapsed"]

system.time(mean\_and\_sd(x,use.sum=T))["elapsed"]

> system.time(mean\_and\_sd(x,use.sum=F))["elapsed"]

elapsed

0.13

> system.time(mean\_and\_sd(x,use.sum=T))["elapsed"]

elapsed

0.05

Q2:

x=matrix(rnorm(1000, mean = 5, sd = 10), nrow = 100, ncol = 10)

y=matrix(rnorm(100000, mean = 5, sd = 10), nrow = 10000, ncol = 10)

apply(x,2,mean\_and\_sd)

apply(y,2,mean\_and\_sd)

|  |  |
| --- | --- |
| 1. Is more far from mean 5 and sd 10 | B – is more close to mean 5 and sd 10 |
| > apply(x,2,mean\_and\_sd)  [[1]]  [[1]]$mean  [1] 7.138201  [[1]]$sd  [1] 8.561584  [[2]]  [[2]]$mean  [1] 5.034403  [[2]]$sd  [1] 10.63847  [[3]]  [[3]]$mean  [1] 6.896464  [[3]]$sd  [1] 9.088593  [[4]]  [[4]]$mean  [1] 4.708208  [[4]]$sd  [1] 10.5532  [[5]]  [[5]]$mean  [1] 3.85591  [[5]]$sd  [1] 9.664721  [[6]]  [[6]]$mean  [1] 5.018624  [[6]]$sd  [1] 9.606202  [[7]]  [[7]]$mean  [1] 4.545362  [[7]]$sd  [1] 10.12088  [[8]]  [[8]]$mean  [1] 5.213506  [[8]]$sd  [1] 8.504945  [[9]]  [[9]]$mean  [1] 5.406072  [[9]]$sd  [1] 8.964576  [[10]]  [[10]]$mean  [1] 5.793002  [[10]]$sd  [1] 8.284713 | > apply(y,2,mean\_and\_sd)  [[1]]  [[1]]$mean  [1] 4.95689  [[1]]$sd  [1] 10.01139  [[2]]  [[2]]$mean  [1] 4.90849  [[2]]$sd  [1] 9.937356  [[3]]  [[3]]$mean  [1] 4.881652  [[3]]$sd  [1] 10.02572  [[4]]  [[4]]$mean  [1] 5.026586  [[4]]$sd  [1] 10.00993  [[5]]  [[5]]$mean  [1] 4.940451  [[5]]$sd  [1] 9.944951  [[6]]  [[6]]$mean  [1] 5.098901  [[6]]$sd  [1] 9.963286  [[7]]  [[7]]$mean  [1] 5.228968  [[7]]$sd  [1] 10.07528  [[8]]  [[8]]$mean  [1] 5.090714  [[8]]$sd  [1] 10.34855  [[9]]  [[9]]$mean  [1] 5.158159  [[9]]$sd  [1] 10.02645  [[10]]  [[10]]$mean  [1] 4.973096  [[10]]$sd  [1] 9.99944 |

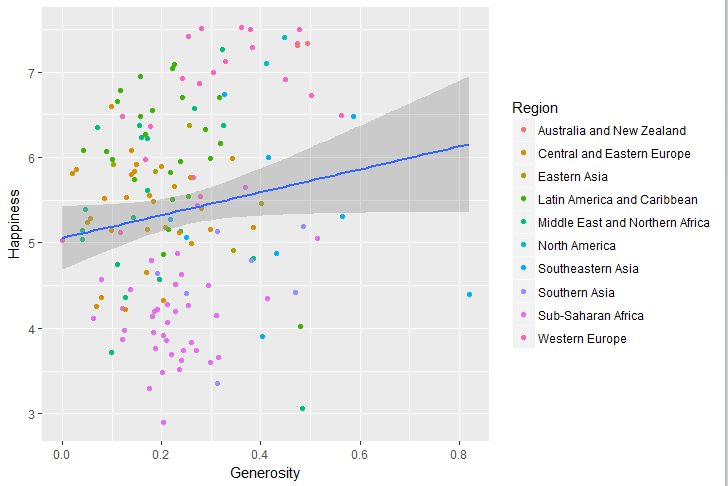
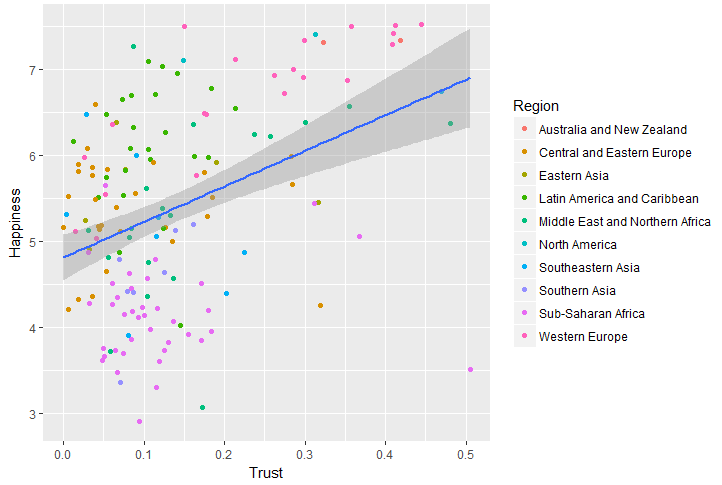
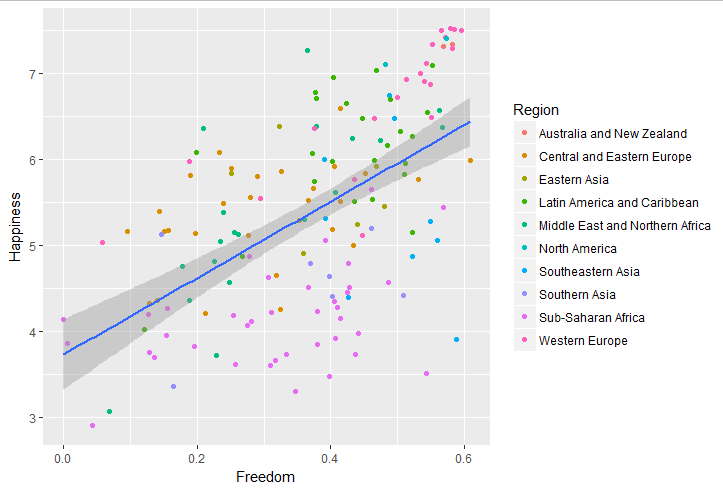
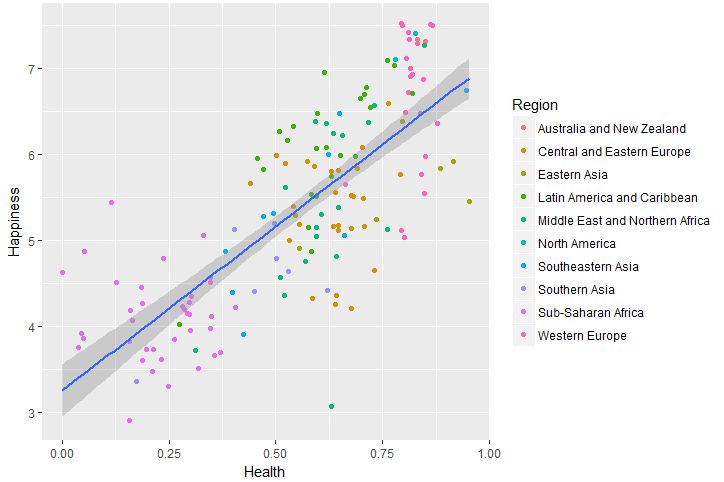
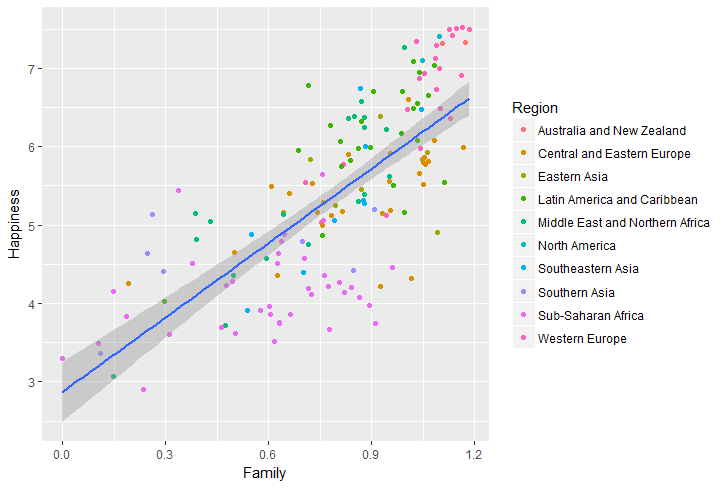
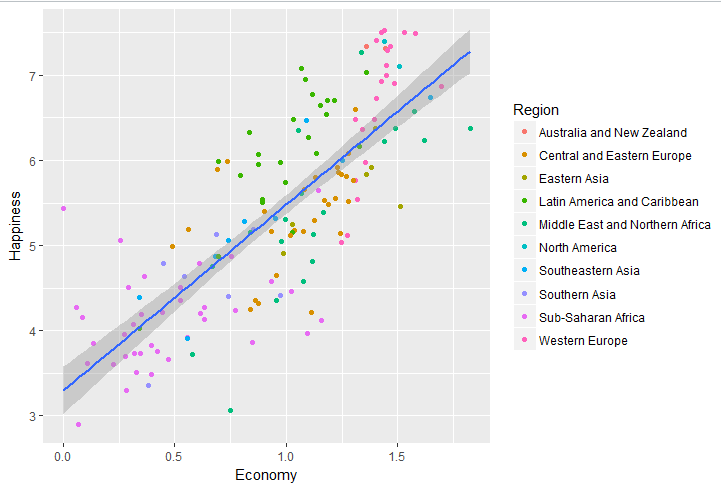
c) What can you do in order to be able to compare your results for this question with the results of your classmate? =

ill average my means and sds to get two number and compare these with others.

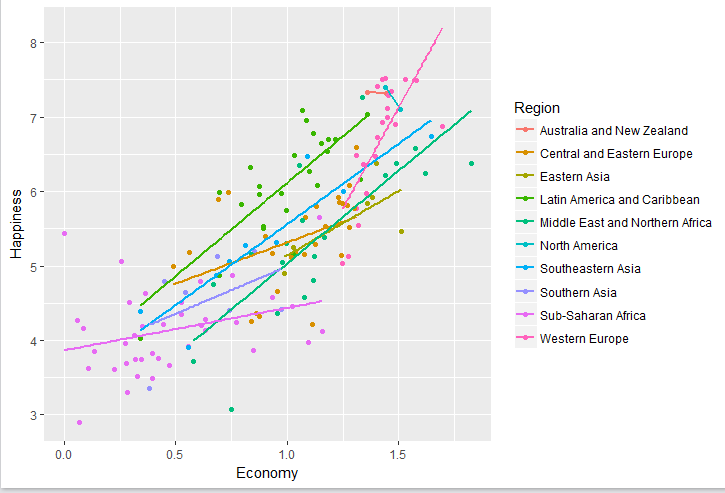
Q3. a)

|  |  |
| --- | --- |
| How many observations are in the data? | > nrow(word)  [1] 157 |
| What are the variables and what is the type of each one? | > sapply(word, class)  Country "factor"  Region "factor"  Happiness "numeric"  Economy "numeric"  Family "numeric"  Health "numeric"  Freedom "numeric"  Trust "numeric"  Generosity "numeric" |
| How many different Regions are in the data? | > length(unique(Region))  [1] 10 |

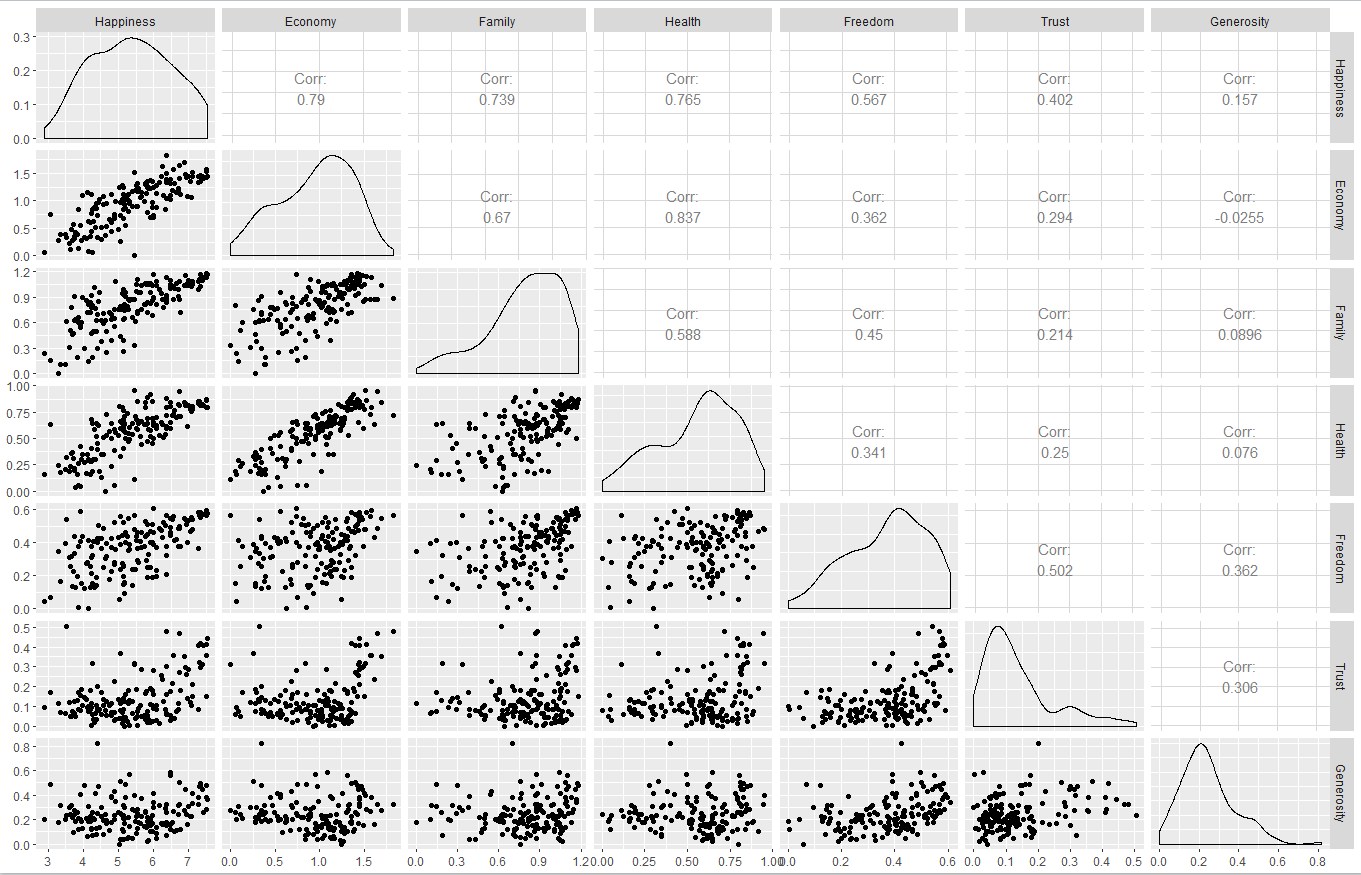
b)

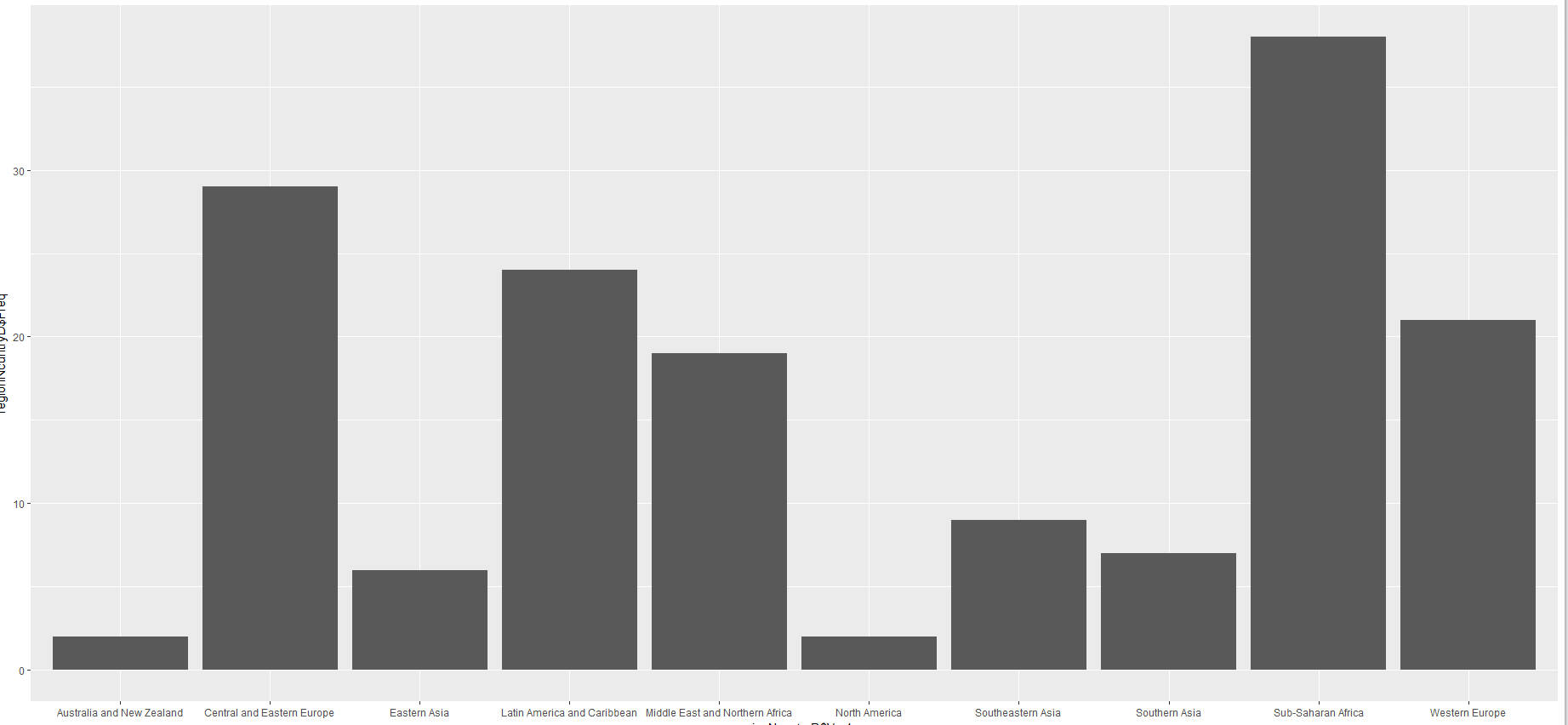
C)

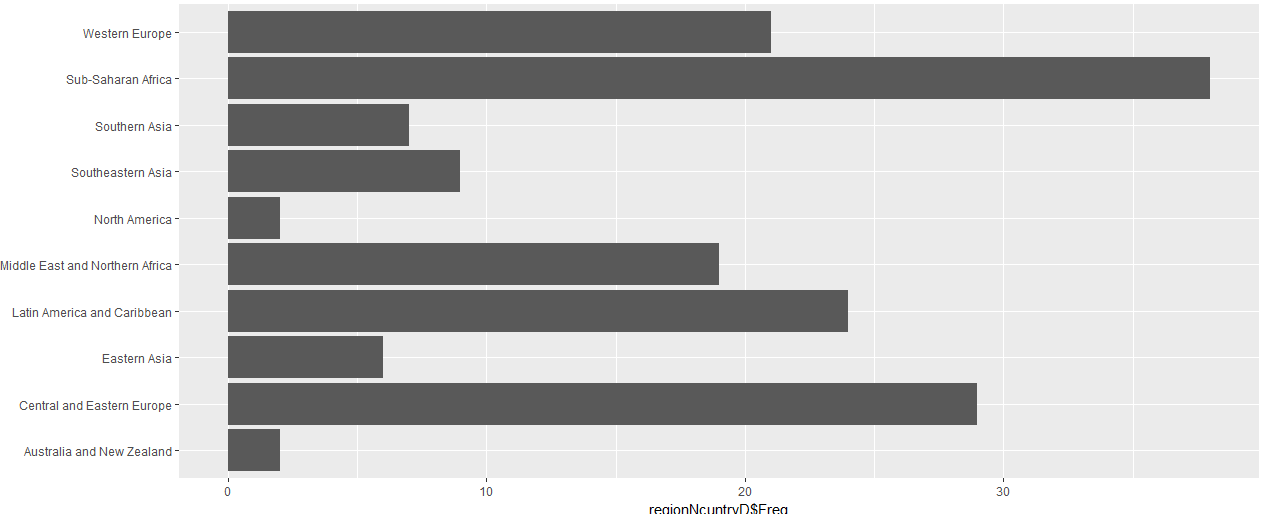


D)Install and load the package GGaly. Call the ggpairs() function with a subset of the data frame that includes only the last 7 variables.

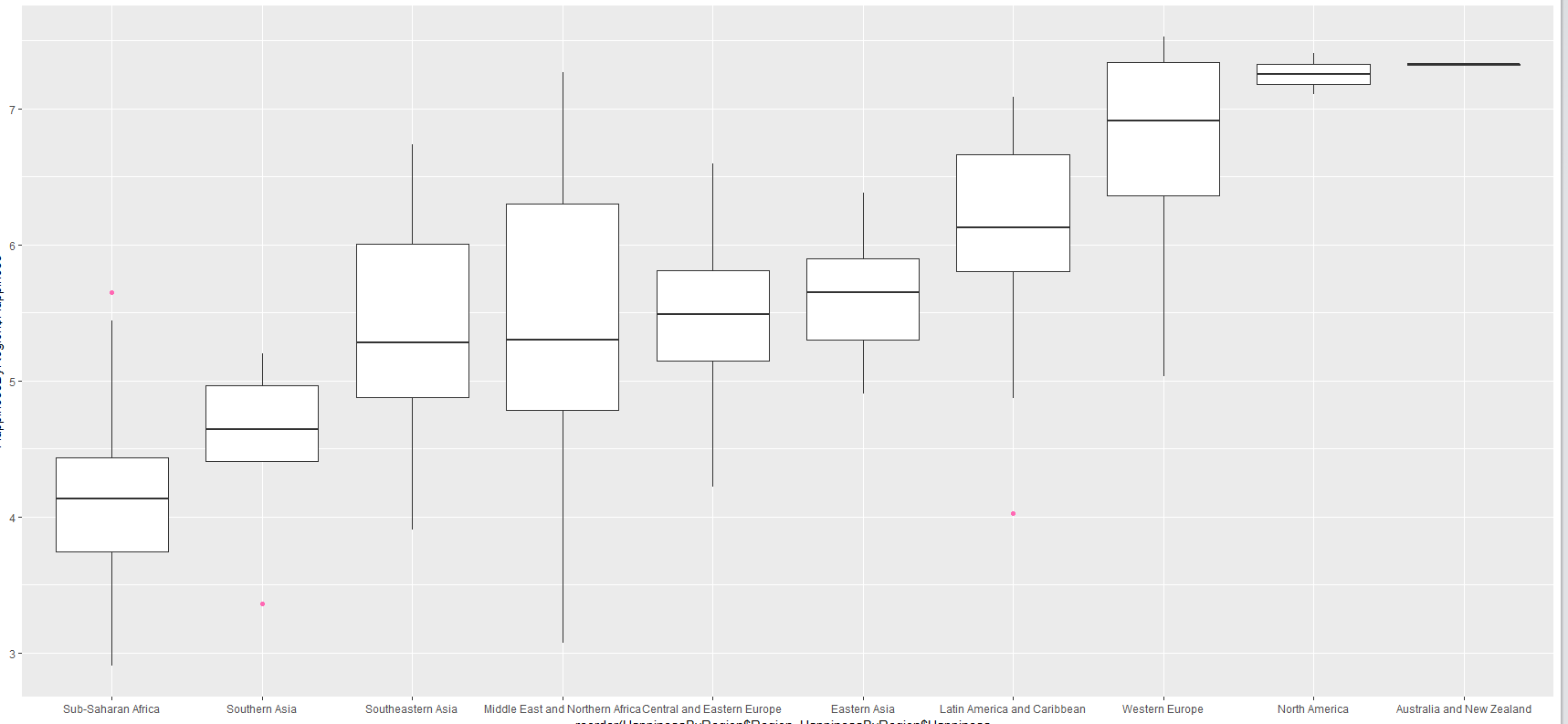


|  |  |
| --- | --- |
| What does this function plot? | Make a matrix of plots with a given data set |
| What does it show on the diagonal? | Histogram of the x axis parameter value frequency |
| Above the diagonal? | Correlation values between the parameter on the x axis and the parameter at the y axis |
| Below the diagonal? | Scatter plots between the parameter on the x axis and the parameter at the y axis |
| What can you learn from this plot that you couldnt tell from the plots you created in section (b)? | Which of the parameters defined the data better with the KPI of Correlation |

E 



|  |  |
| --- | --- |
| Why does flipping the coordinates improves the plot in this case? | We can see the attributes names better |

F 

|  |  |
| --- | --- |
| Which Region has the most variability in Happiness? | Middle East and Northern Africa |
| Which Region has the least variability? | Australia and NZ |
| What is the problem with the answer to this last question? | There are only 2 countries in this region |