**CS 6313.501**

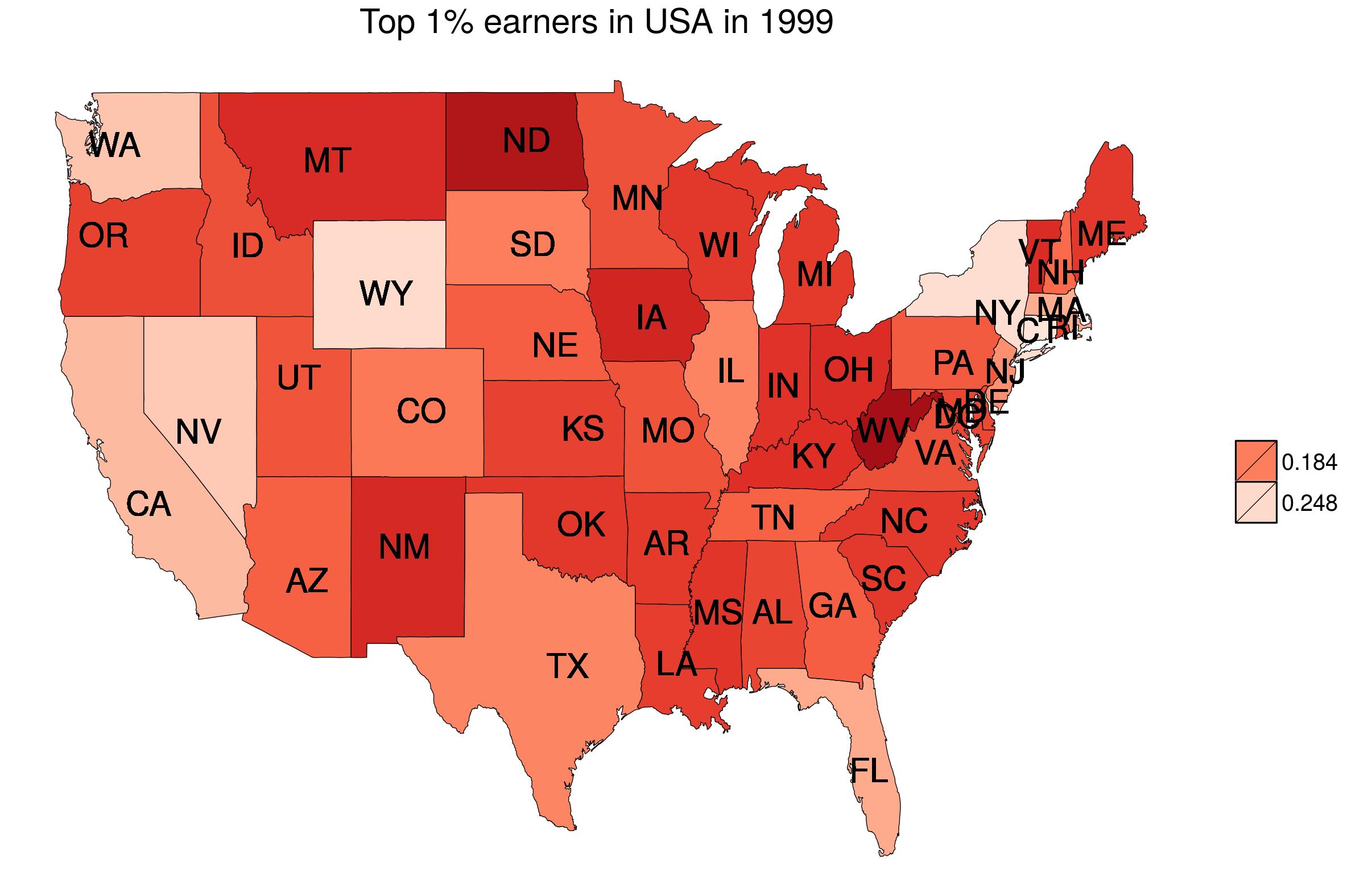
**Mini Project 2**

**Group members:** Arnav Sharma (axs144130), Divyanshu Paliwal (dxp151630)

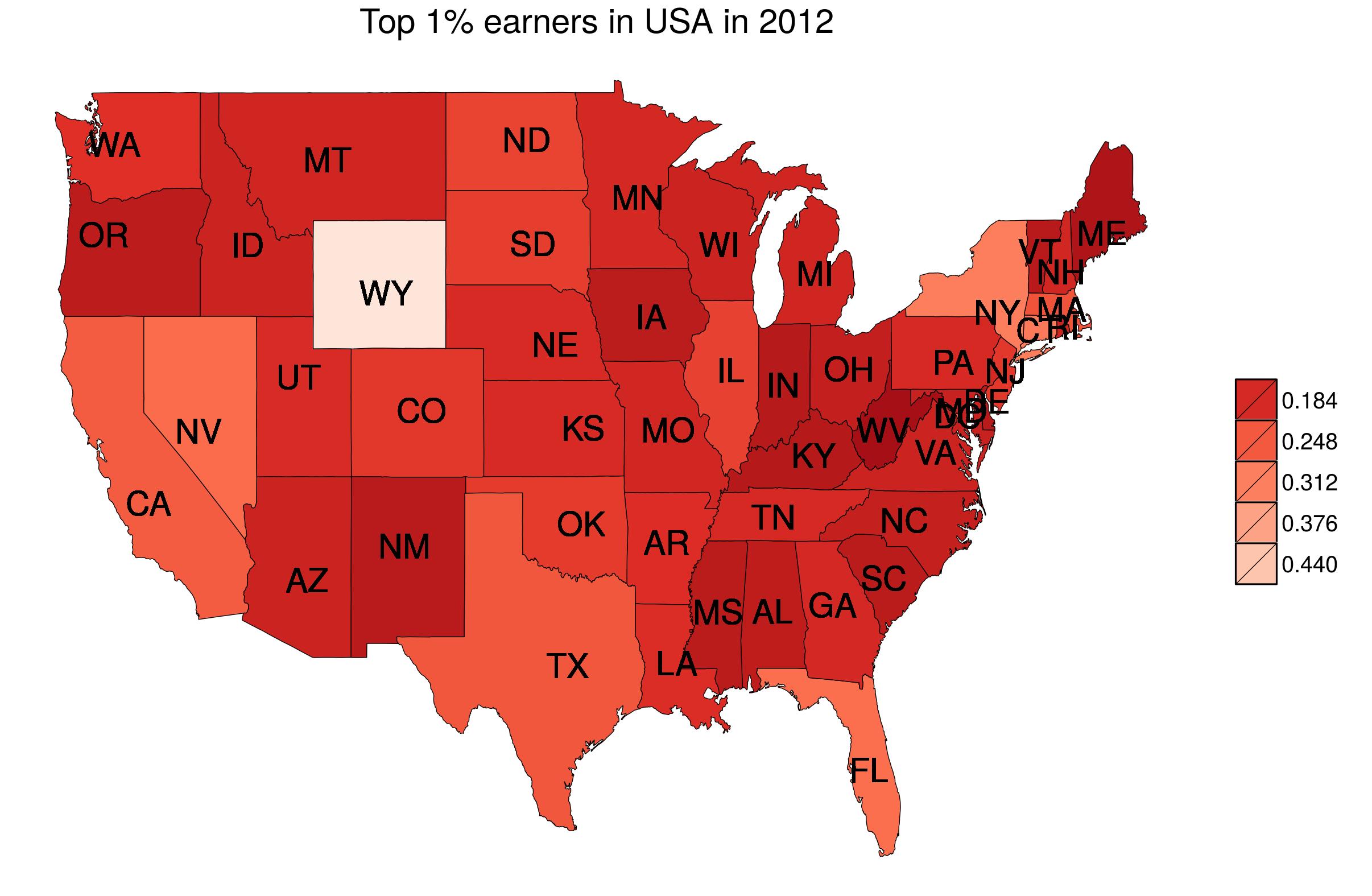
**Contribution of each member:** Both members contributed equally to the following tasks:

* Writing the R code
* Generating graphs
* Deriving conclusions

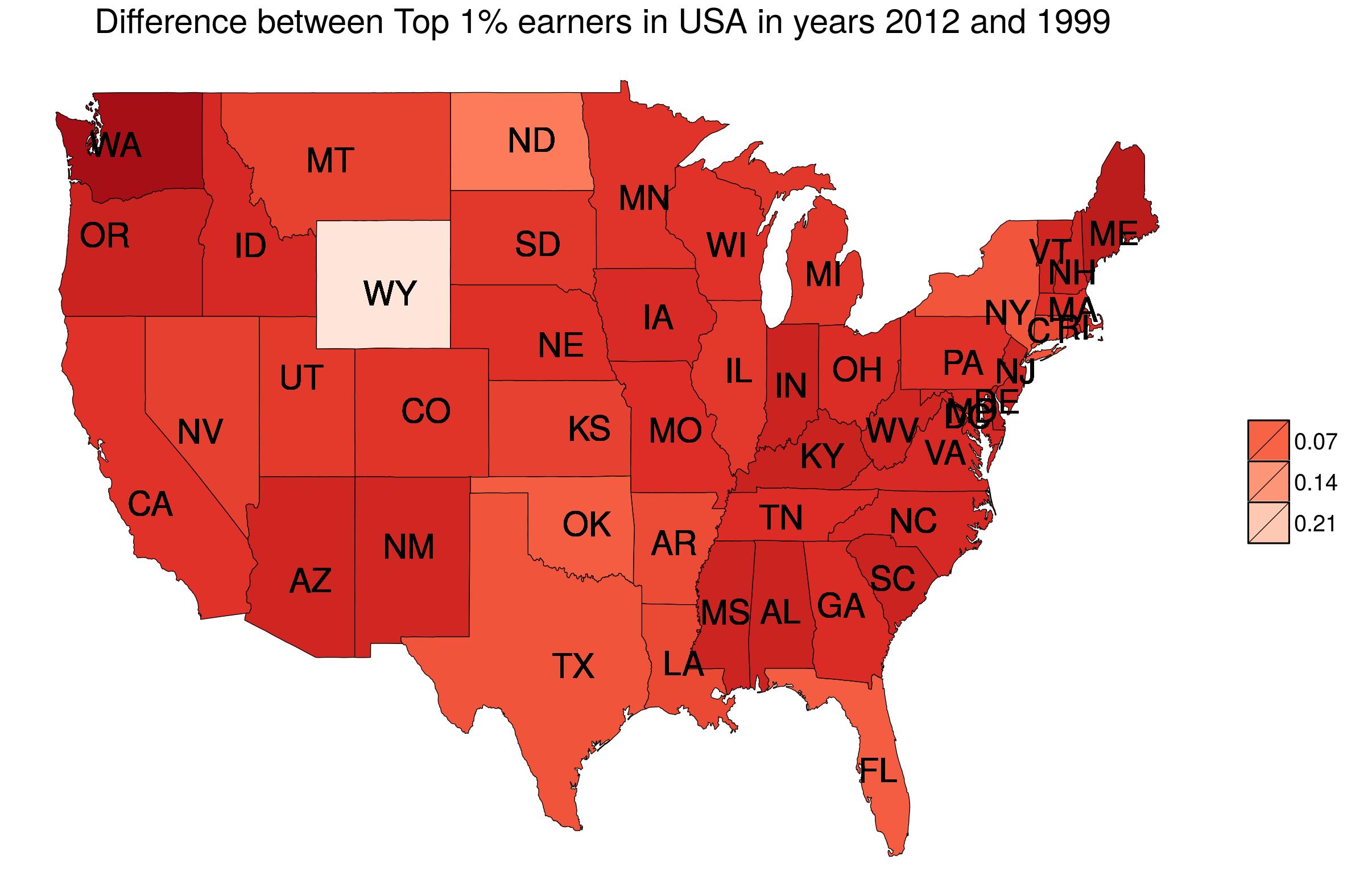
**Exercise 1**



Map 1: shows the state level income share of top 1% of income earners in 1999.



Map 2: shows the state level income share of top 1% of income earners in 2012.



Map 3: shows the difference from 1999 to 2012 in state level income share of top 1% of income earners of the USA.

**B-** Map 1 shows that states like New York followed by Wyoming and Nevada have clearly the largest share in 1999.

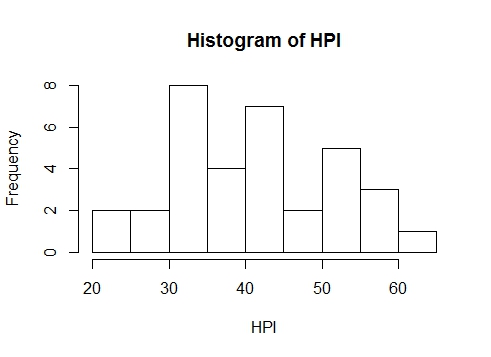
Map 2 shows that now Wyoming has by far the largest income share of top 1% of income earners in 2012. Also, the average income of top 1% income earners from all states has increased which can be noted from the increase in the scale of break points.

Map 3 validates the claim made from Map 2 as Wyoming is seen as the state having the largest difference from year 1999 to 2012.

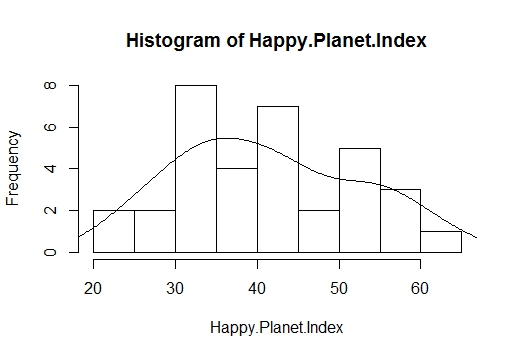
**Exercise 2**

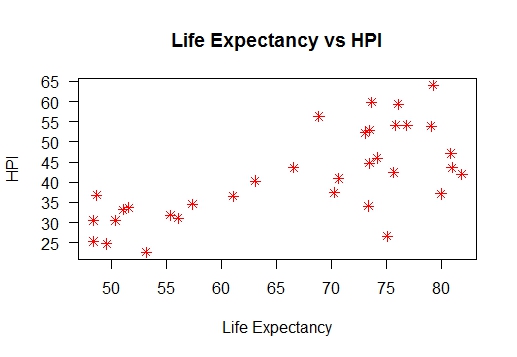
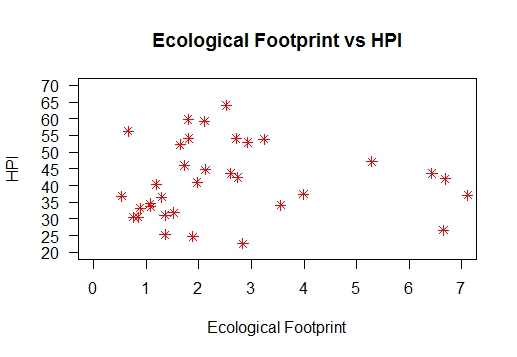
**B-** By examining the histogram for HPI variable it is clear that the distribution is almost symmetric. This is also evident from the fact that mean (HPI) is approximately equal to median(HPI) ( mean=41, median=40.6). As the distribution for HPI is normal, the appropriate measures of center and variability for this distribution are **(mean, standard deviation).**

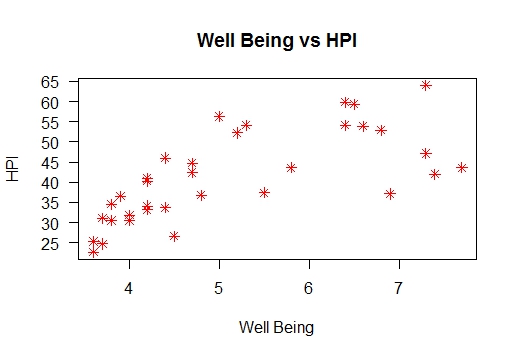
**Histogram of Happy Planet Index**

****

**Histogram of Happy Planet Index along with the density curve**

****

**C-** The scatterplots of HPI against each of the three variables is shown below. Clearly the plot between Well Being and HPI is the best.



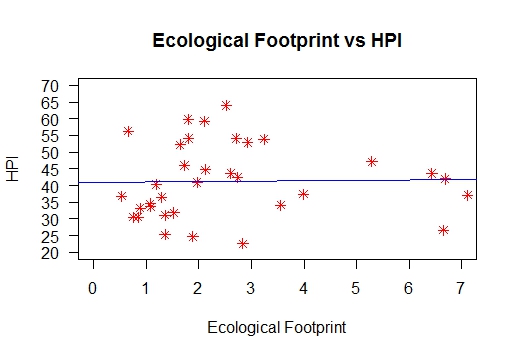
**Yes** it will be appropriate to use correlation to summarize the relationship of HPI with other three variables, because correlation measures the degree (strength) of relationship between two variables. Thus by looking at the correlations of HPI with the three variables we will be able to tell which variable is more strongly related to HPI. The correlations are as follows:

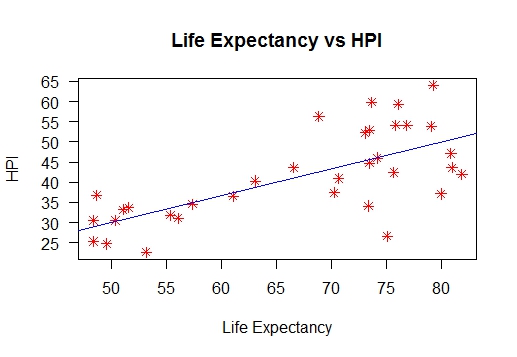
Correlation between Ecological Footprint and HPI = **0.01594187**

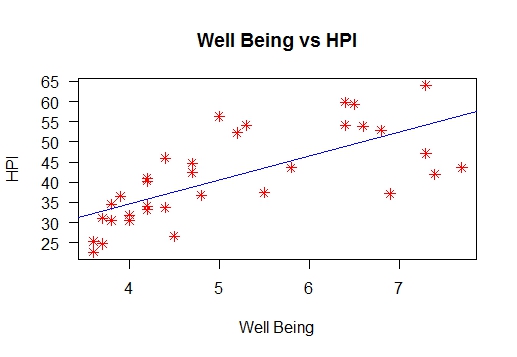
Correlation between Life Expectancy and HPI = **0.6964823**

Correlation between Well-being and HPI = **0.6995526**

**The below graphs show a linear correlation function drawn over the scatterplots of HPI against the three variables**. It is clear from these graphs that Well-being is most strongly related with HPI and Ecological footprint is the least related.







**Appendix**

R code for Exercise-1

#Mini-Project-2

#Question-1

#Generating Maps of the States in the USA for top 1% income earners

#-------------------------------------------------------------------------------------------

#map\_data() function outputs the states in the USA as regions along with their latitudinal and longitudinal values in usa.df data file

usa.df<-map\_data("state")

#Changed the name of the Column 5 from "region" to "state" in usa.df file

colnames(usa.df)[5]<-"state"

#as.factor() function factors the values in the "state" column.

usa.df$state<-as.factor(usa.df$state)

#read.table() function reads the input file showing varied state level income share for multiple years for each state and stores in usa.dat file

usa.dat <- read.table("Frank\_top\_2012.csv", header = T, sep = ",")

#join() fuction joins usa.df and usa.dat files by the column "state" using the type inner join

usa.df <- join(usa.df, usa.dat, by = "state", type = "inner")

#read.table() reads "state\_latlon.csv" file for abbreviations of each state and their positions and stores the result in a new "centers" file

centers <- read.table("state\_latlon.csv", header = T, sep = ",")

#names() function updates the name of the columns "longitude" to "long\_c" and "latitude" to "lat\_c"

names(centers)[3:4] <- c('long\_c','lat\_c')

#join() function joins usa.df and centers file by states using type inner join

usa.df <- join(usa.df, centers, by = "state", type = "inner")

#Map 1 for income share of top 1% earners by state in 2012

#-------------------------------------------------------------------------------------------

#"usa\_2012.df" file stores the value of "usa.df" file for each row where year is 2012

usa\_2012.df = usa.df[usa.df$year==2012,]

#brks stores the value of breakpoints to be used in the map

brks<-c(0.184, 0.248, 0.312, 0.376, 0.44, 0.5)

#ggplot() and geom\_polygon() are used for plotting the map with borders, scale\_fill\_distiller is used and shades of red are used for plotting which go from dark to light as income share increases

p <- ggplot() +

geom\_polygon(data = usa\_2012.df, aes(x = long, y = lat, group = group, fill = top1), color = "black", size = 0.15) +

scale\_fill\_distiller(palette = "Reds", breaks = brks, trans = "reverse") + theme\_nothing(legend = TRUE) +

labs(title = "Top 1% earners in USA in 2012 ", fill = "") +

geom\_text(aes(x = lat\_c, y = long\_c, group = group, label = label),

data= usa\_2012.df,

alpha = 1,

color = "black")

#The result obtained from previous line of code is saved as a jpeg image file

ggsave(p, file = "usa\_top1\_2012\_A.jpeg")

#Map 2 for income share of top 1% earners by state in 1999

#--------------------------------------------------------------------------------------

#"usa\_1999.df" file stores the value of "usa.df" file for each row where year is 1999

usa\_1999.df = usa.df[usa.df$year==1999,]

#brks stores the value of breakpoints to be used in the map

brks<-c(0.184, 0.248, 0.312, 0.376, 0.44, 0.5)

#ggplot() and geom\_polygon() are used for plotting the map with borders, scale\_fill\_distiller is used and shades of red are used for plotting which go from dark to light as income share increases

p <- ggplot() +

geom\_polygon(data = usa\_1999.df, aes(x = long, y = lat, group = group, fill = top1), color = "black", size = 0.15) +

scale\_fill\_distiller(palette = "Reds", breaks = brks, trans = "reverse") +

theme\_nothing(legend = TRUE) +

labs(title = "Top 1% earners in USA in 1999 ", fill = "") +

geom\_text(aes(x = lat\_c, y = long\_c, group = group, label = label),

data= usa\_1999.df,

alpha = 1,

color = "black")

#The result obtained from previous line of code is saved as a jpeg image file

ggsave(p, file = "usa\_top1\_1999.jpeg")

#Map 3 for diffrence in income share of top 1% earners by state from 1999 to 2012

#---------------------------------------------------------------------------------------

#A new column "Top1\_2012" is created

usa\_2012.df["Top1\_2012"] <- NA

#A new column "Top1\_1999" is created

usa\_2012.df["Top1\_1999"] <- NA

#Column "Top1\_2012" is given the values of column "top1" from usa\_2012.df. Therefore it now has a column with state wise income share of top 1% earners for year 2012

usa\_2012.df["Top1\_2012"] <- usa\_2012.df[,c("top1")]

#Column "Top1\_1999" is given the values of column "top1" from usa\_1999.df. Therefore it now has a column with state wise income share of top 1% earners for year 1999

#both the columns Top1\_2012 and Top1\_1999 are stored in usa\_2012.df file

usa\_2012.df["Top1\_1999"] <- usa\_1999.df[,c("top1")]

#The difference of the values in each row for the columns "Top1\_2012" and "Top1\_1999" is stored in a new column "difference" of the usa\_2012.df file

usa\_2012.df$difference <- usa\_2012.df$Top1\_2012 - usa\_2012.df$Top1\_1999

#brks stores the breakpoint values of the map

brks<-c(-0.04, 0.07, 0.14, 0.21)

#ggplot() and geom\_polygon() are used for plotting the map with borders, scale\_fill\_distiller is used and shades of red are used for plotting which go from dark to light as income share increases

p <- ggplot() +

geom\_polygon(data = usa\_2012.df, aes(x = long, y = lat, group = group, fill = difference), color = "black", size = 0.15) +

scale\_fill\_distiller(palette = "Reds", breaks = brks, trans = "reverse") +

theme\_nothing(legend = TRUE) +

labs(title = "Difference between Top 1% earners in USA in years 2012 and 1999 ", fill = "") +

geom\_text(aes(x = lat\_c, y = long\_c, group = group, label = label),

data= usa.df,

alpha = 1,

color = "black")

#The result obtained from previous line of code is saved as a jpeg image file

ggsave(p, file = "Top1\_difference.jpeg")

#----------------------------------------------------------------------------------------

R code for Exercise-2

#read the data

hpi.dat <- read.table("hpi-data.csv", header = T, sep = ",")

str(hpi.dat)

attach(hpi.dat)

#draw the histogram of HPI

myhist<-hist(Happy.Planet.Index, main = "Histogram of HPI", xlab = "HPI")

#draw the density function over the histogram to get an idea of the distribution of data

multiplier <- myhist$counts / myhist$density

mydensity <- density(Happy.Planet.Index)

mydensity$y <- mydensity$y \* multiplier[1]

lines(mydensity)

#draw the scatterplot between Well Being and HPI

plot(Well.being..0.10., Happy.Planet.Index, main="Well Being vs HPI", xlab="Well Being", ylab="HPI", las=1, pch=8, col=2, yaxp=c(0,80,16))

hpi.dat <- na.omit(hpi.dat)

#draw a linear correlation function over the scatterplot to get an idea of correlation between

#Well Being and HPI

abline(lm(Happy.Planet.Index~Well.being..0.10.), col=4)

#draw the scatterplot between Life Expectancy and HPI

plot(Life..Expectancy, Happy.Planet.Index, main="Life Expectancy vs HPI", xlab="Life Expectancy", ylab="HPI", las=1, pch=8, col=2, yaxp=c(0,80,16))

#draw a linear correlation function over the scatterplot to get an idea of correlation between Life #Expectancy and HPI

abline(lm(Happy.Planet.Index ~ Life..Expectancy), col=4)

#draw the scatterplot between Ecological Footprint and HPI

plot(Footprint..gha.capita., Happy.Planet.Index, main="Ecological Footprint vs HPI", xlab="Ecological Footprint", ylab="HPI", las=1, pch=8, col=2, xlim=(c(0,7)), ylim=(c(20,70)), yaxp=c(20,80,12))

#draw a linear correlation function over the scatterplot to get an idea of correlation between #Ecological Footprint and HPI

abline(lm(Happy.Planet.Index ~ Footprint..gha.capita.), col=4)

#calculate correlation between HPI and each of the three variables

cor(Footprint..gha.capita., Happy.Planet.Index, use= "complete")

cor(Life..Expectancy, Happy.Planet.Index, use= "complete")

cor(Well.being..0.10., Happy.Planet.Index, use= "complete")