MINI PROJECT 2 Section: **CS 6313.501**

Name: Raghunandan Nuggehalli Ramesha netid : rxn150230
Name:Panchami Rudrakshi netid : pgr150030

CONTRIBUTION :

Both the team members contributed equally towards understanding,analysis,problem solving,coding and making the report.

EXERCISE 1

PROBLEM STATEMENT:

Use R to make three maps of the states in the USA. The first map should plot state level income share of the top 1% of income earners in 2012. The second map should plot the same variable but for 1999. The third map should plot the 2012-1999 difference of the variable.

OVERVIEW:

The project involves plotting map using inbuilt library functions. Three maps are plotted for state level income share of the top 1% of income earners for year 1999, 2012 and the difference between the two is analysed. Further the states are labelled.

FUNCTIONS USED:

library(raster) - to get map shape file
library(ggplot2) - for plotting and miscellaneous things
library(ggmap) - for plotting
library(plyr) - for merging datasets
library(scales) - Graphical scales map data to aesthetics, and provide methods for automatically determining breaks and labels for axes and legends.
library(maps) - Display of maps.
map_data(map)- name of map provided by the maps package which takes values county, state, world e.t.c.
str(object)- Compactly display the internal structure of an R object. factor {base}- Encode a vector as a factor. is.factor, is.ordered, as.factor and as.ordered are the membership and coercion functions for these classes.
levels()- levels provides access to the levels attribute of a variable. The first form returns the value of the levels of its argument and the second sets the attribute.
join(x, y, by='', type='')- Merge two data frames by common columns or row names, or do other versions of database *join* operations.
type: inner- Return only the rows in which the left table have matching keys in the right table.
range()- range returns a vector containing the minimum and maximum of all the given arguments.
brks – breaks the given data into specified intervals.
**ggplot(**)-ggplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. It takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics.
geom_polygon(mapping=NULL,data=NULL,stat="identify", position="identity",..)-The aesthetic mapping, usually constructed with **aes.** Gives textual annotations.
geom_text(mapping = NULL, data = NULL, stat = "identity", position = "identity", parse = FALSE, ...)- Gives Polygon, a filled path.
labs(title = "")- The text for the axis or plot title.
ggsave(p, file = "")- ggsave is a convenient function for saving a plot. It defaults to saving the last plot that you displayed, and for a default size uses the size of the current graphics device. It also guesses the type of graphics device from the extension. This means the only argument you need to supply is the filename.

OBSERVATIONS & COMMENTS:

The maps show the distribution of top 1% of earners in USA. The map is plotted using the **ggplot()** function. The colour shades indicate the % of the income earners in that state. Lighter shades are indicators of the more % of income earners and as the shades get darker, the % gets lesser. The darker brown colour indicates the lowest contributors towards the US income for top 1 % earners.

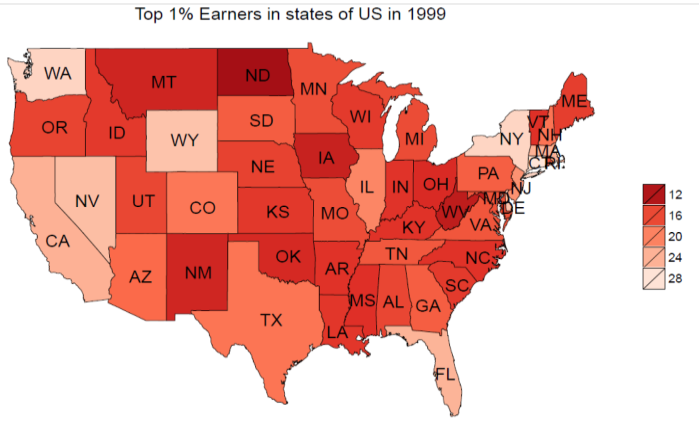
We plot three graphs for the given problem firstly for 1999. Then we filter the data for the year 2012 and plot the second graph. In the third case we calculate the difference in top 1% of data for years 2012-1999 and plot that graph.

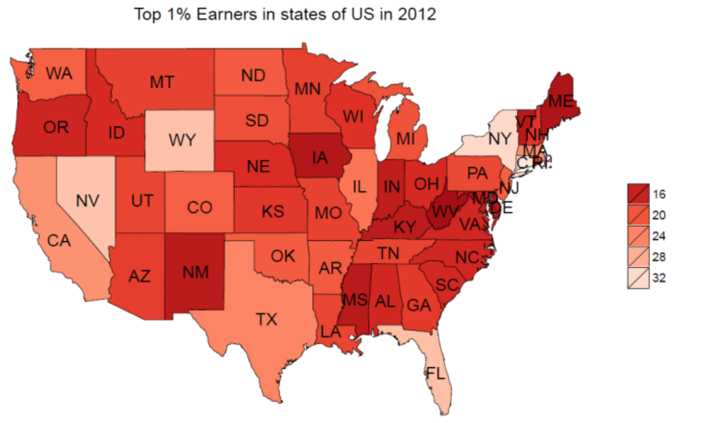
1999:

In 1999, there were 6 prominent states namely New York, Washington, Delaware, Wyoming, Nevada, Connecticut marked by the lighter colour in the graph lying between 25 and 28 . This indicates that they are amongst the top contributors of the USA's top 1 % earners in 1999.

Also we can see from the graph that there are a few states that fall in the middle range of 24 to 20 which are indicated by a slightly darker shade of brown namely Nevada, florida, Massachusetts, California, New Jersey, Illinois .

The shade gets darker as the top 1% earners get reduced. And it is dark brown for those which have less than 12.5%. States like Alaska, North Dakota, West Virginia, Montana, Iowa and New Mexico have the darker shade indicating they have lesser number of top 1% income earners.
The range of the % in the year 1999 is: 10.74956 28.15289





2012:

There were 5 states namely Connecticut, New York, Nevada, Wyoming, Florida marked with lightest shade in the graph. This indicates that they are among the top contributors of the USA top 1% income.

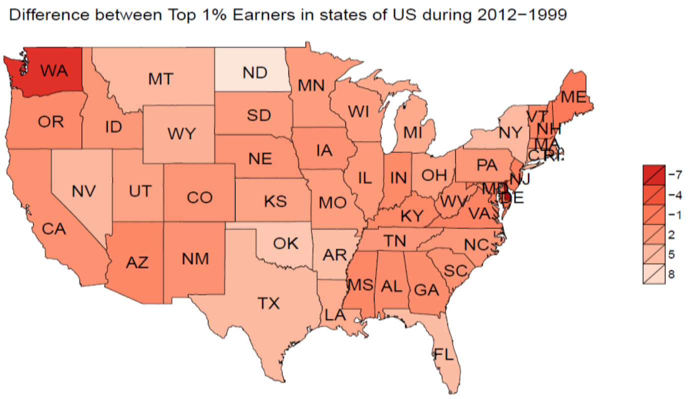
The shade gets darker as the top 1 % earners get reduced. And it is dark brown for those which have less than 15%. States like Alaska, West Virginia, Maine, Iowa, Vermont, Mississippi and New Mexico have the darker shade indicating they have lesser number of top 1 % income earners in 2012.

The range of the % in the year 2012 is: 12.50678 33.00785

2012-1999

The third map basically represents the states whose top 1% Income Share has increased/decreased from 1999 to 2012. We can clearly see that the highest increase in income share by top 1% income earners was in North Dakota having increase of more that 8% and highest decrease was in Delaware and Washington, with a decrease of more than 6%.

The maximum increase is for North Dakota with the value 9.02726.The highest decrease is for Delaware with the value -10.169.



CODE :

```
library(raster)                                         # to get map shape file
library(ggplot2)                                        # for plotting and miscellaneous things
library(ggmap)                                          # for plotting the map
library(plyr)                                           # for merging datasets
library(scales)                                         # Graphical scales map data to aesthetics
library(maps)                                           # for plotting the graphs
usa.df <- map_data("state")
str(usa.df)                                             # display the internal structure of an R object.
colnames(usa.df) [5] <- "state"
usa.df$state <- as.factor(usa.df$state)
str(usa.df)
USA.dat <- read.csv(file.choose(), header = T, sep = ",")     #reading the csv file with data for year 1999
levels(usa.df$state)                            # levels provides access to the levels attribute of a variable.
usa.df <- join(usa.df, USA.dat, by = "state", type = "inner")
range(usa.df$top1)                            # returns a vector containing the minimum and maximum
brks <- c(12,16,20,24,28)                     # breaks the given data into specified intervals
p <- ggplot() +
geom_polygon(data = usa.df, aes(x = long, y = lat, group = group, fill = top1), color = "black", size = 0.25)
+         geom_text(aes(x=state.center$x,y=state.center$y,label=state.abb)) +
scale_fill_distiller(palette = "Reds", breaks = brks, trans = "reverse") +
theme_nothing(legend = TRUE) +
labs(title = " Top 1% Earners in states of US in 1999", fill = "")
ggsave(p, file = "1999_top earners.pdf")
#Different breaks are taken for different years depending on the range of values of income for different
years.
# brks <- c(16,20,24,28,32)                                 # for 2012 data
# brks <- c(-7,-4,-1,2,5,8)                               # for difference between 2012-1999 data
```

Note : BONUS POINTS

For the labelling of the states with their observations we have used a function geom_text()

```
#geom_text(aes(x=state.center$x,y=state.center$y,label=state.abb))
```

EXERCISE 2

The Happy Planet Index measures the extent to which it delivers a long and happy life to the people who live in a particular country based on:
Life Expectancy : This is measure of life expectancy data generated for each country.
Experienced Well-Being :This is measure by asking the people themselves how great their life is going. This is done by asking the people to imagine a ladder, 0 being the worst possible life and 10, the best possible life, and report the step of the ladder they feel they are at.
Ecological Footprint :This is a per capita measure of the amount of land required to sustain a country's consumption in terms of global hectares.

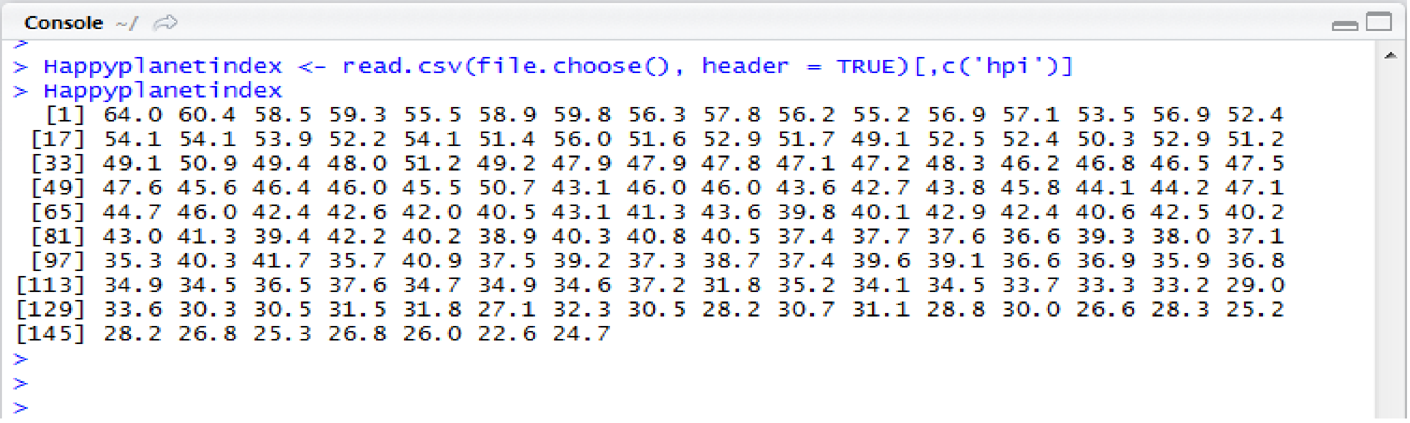
Happy Planet Index ≈ (Experienced Well-Being * Life Expectancy) / Ecological Footprint

CRITICISM:

The criticism is due to commentators incorrectly understanding HPI to be a measure of personal happiness rather than a measure of the "happiness" of the planet. The HPI completely ignores issues such as political freedom, human rights and labour rights. The World Values Survey covers only a minority of the world's nations and is only carried out every five years. As a result, much of the data for the index must comes from other sources or is estimated using regressions. The subjective measures of well-being are suspect. The ecological footprint is a controversial and much criticized concept.

The index has been criticised for weighting the carbon footprint too heavily, to the point that the U.S. would have had to be universally happy and would have had to have a life expectancy of 439 years to equal Vanuatu's score in the 2006 index.

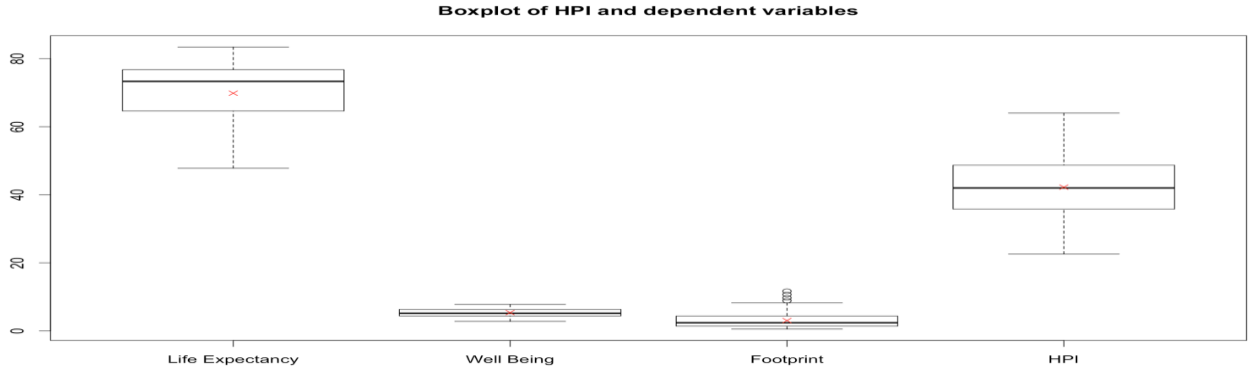
From Happy Planet Index on Wikipedia (http://www.happyplanetindex.org/data/) , the data obtained is :

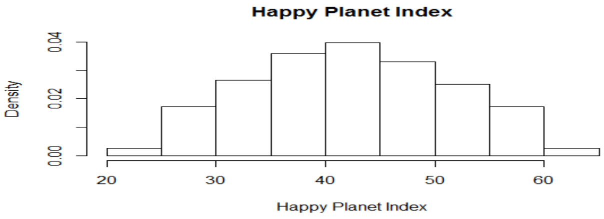


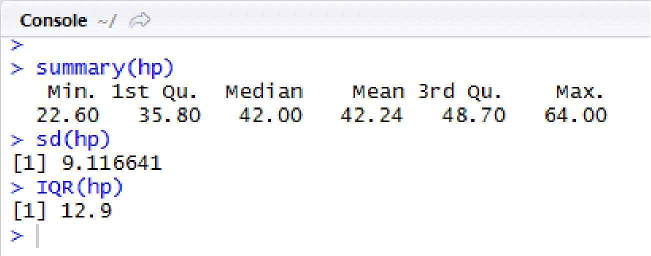
```
>
> Happyplanetindex <- read.csv(file.choose(), header = TRUE)[,c('hpi')]
> Happyplanetindex
  [1] 64.0 60.4 58.5 59.3 55.5 58.9 59.8 56.3 57.8 56.2 55.2 56.9 57.1 53.5 56.9 52.4
 [17] 54.1 54.1 53.9 52.2 54.1 51.4 56.0 51.6 52.9 51.7 49.1 52.5 52.4 50.3 52.9 51.2
 [33] 49.1 50.9 49.4 48.0 51.2 49.2 47.9 47.9 47.8 47.1 47.2 48.3 46.2 46.8 46.5 47.5
 [49] 47.6 45.6 46.4 46.0 45.5 50.7 43.1 46.0 46.0 43.6 42.7 43.8 45.8 44.1 44.2 47.1
 [65] 44.7 46.0 42.4 42.6 42.0 40.5 43.1 41.3 43.6 39.8 40.1 42.9 42.4 40.6 42.5 40.2
 [81] 43.0 41.3 39.4 42.2 40.2 38.9 40.3 40.8 40.5 37.4 37.7 37.6 36.6 39.3 38.0 37.1
 [97] 35.3 40.3 41.7 35.7 40.9 37.5 39.2 37.3 38.7 37.4 39.6 39.1 36.6 36.9 35.9 36.8
[113] 34.9 34.5 36.5 37.6 34.7 34.9 34.6 37.2 31.8 35.2 34.1 34.5 33.7 33.3 33.2 29.0
[129] 33.6 30.3 30.5 31.5 31.8 27.1 32.3 30.5 28.2 30.7 31.1 28.8 30.0 26.6 28.3 25.2
[145] 28.2 26.8 25.3 26.8 26.0 22.6 24.7
>
>
>
```

(b) Examine the distribution of the HPI variable graphically. What would be appropriate measures of centre and spread of this distribution --------- (mean, SD) or (median, IQR). Justify your answers.

Distribution of is graphically analysed using Boxplot. Boxplot gives the summary of distribution (min, median, Q1, Q3 and max)







```
>
> summary(hp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.60   35.80   42.00   42.24   48.70   64.00 
> sd(hp)
[1] 9.116641
> IQR(hp)
[1] 12.9
> 
```

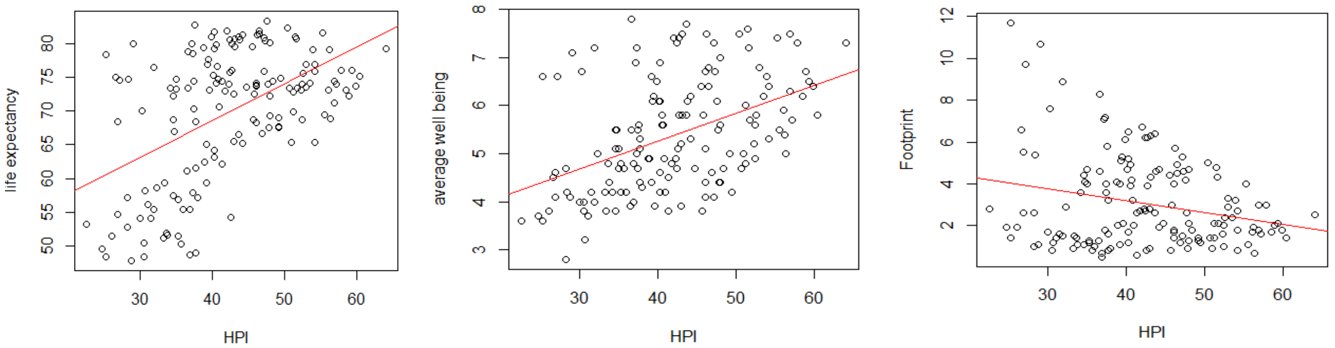
The distribution of the HPI is approximately Normal as seen from the histogram.

Justification for using mean and sd:

- From the mean and median values generated using R, its seen that HPI is slightly right skewed.
- IQR and median would be a better option when the data is skewed and/or have outliers.
- The distribution tells us that 68% of our data falls within the one standard deviation of the mean and 95% are within the two standard deviations.

- We can clearly see that Life Expectancy is left skewed and that Footprint is right skewed and contains outliers too.
- For measure of central tendency and spread of Life expectancy and Footprint, IQR and Median would be a better option but for HPI, as it does not contain and outliers, and that it is approximately normally distributed, mean and standard deviation is a better fit.

(c) Make scatter plots of HPI against each of the three variables on which the index is best. Comment on what you see. Will it be appropriate to use correlation to summarize the relationship of HPI with the other three variables? If yes, provide the correlations. Explain your answers.



Scatter plot of HPI with Life Expectancy:
It looks like a moderate *positive correlation*. It can be said, as the value of HPI increases, the life expectancy also might increase.
Cor between HPI and Life Expectancy: **0.5111565**

Scatter plot of HPI with Well Being:
It has a weak *positive correlation*. Not much can be interpreted from this scatter plot.
Cor between HPI and Experienced Well being: **0.4510568**

Scatter plot of HPI with Ecological Footprint:
It seems to be a weak *negative correlation*. As the value of HPI increases, the ecological footprint decreases. The scatter plot also shows us that there's a possibility of outliers when the HPI is low.
Cor between HPI and Ecological Footprint: **-0.2380059**

Conclusion: Correlations help us measure the strength of a linear relationship between two variables. Here, since in our scatter plot a strong linear relationship does not exists, calculating the correlation does not make sense. The only scatter plot that could make a little sense would be the one with Life Expectancy, as the Cor value is above 0.5 (Range of Cor is between -1 to 1).

CODE:

```
hp<-read.xls("hpi_dat.xls")                                  # Read the data from Excel Sheeet
range(hp$HPI)                                                # Finds the range
mean(hp$Life.Expectancy)                                     # Finds the mean
sd(hp$HPI)                                                   # Finds the standard Deviation
median(hp$Life.Expectancy)                                   # Finds the median

x<-hist(hp$HPI,main="Happy Planet Index", xlab="HPI", border="black")
                                                       # Plots histogram of Happy Planet Index

# Plots boxplot of HPI, Life Expectancy, Footprint and Well Being with the indication of outliers.

boxplot(hp$Life.Expectancy,hp$Well.Being,hp$Footprint..gha.capita. ,hp$HPI,
        names=c("Life Expectancy","WellBeing","Footprint","HPI"),
        main="Boxplot of HPI and dependent variables")
 points(mean(hp$Life.Expectancy),x=1,col="red",pch=4)
 points(mean(hp$Well.Being),x=2,col="red",pch=4)
 points(mean(hp$Footprint..gha.capita.),x=3,col="red",pch=4)
 points(mean(hp$HPI),x=4,col="red",pch=4)
# Plots scatter plot of HPI against each identifying factor.

 plot(hp$HPI,hp$Life.Expectancy,xlab="Happy Planet Index",ylab="Life Expectancy")
 abline(lm(hp$life~hp$hpi), col="red")                       # regression line (y~x)
 plot(hp$HPI,hp$Well.Being,xlab="Happy Planet Index",ylab="Experienced Well Being")
 abline(lm(hp$ Well.Being ~hp$hpi), col="red")               # regression line (y~x)
 plot(hp$HPI,hp$Footprint..gha.capita.,xlab="Happy Planet Index",ylab="Ecological Footprint")
 abline(lm(hp$ Footprint..gha.capita.~hp$hpi), col="red")  # regression line (y~x)

# Plots scatter plot of HPI against each identifying factor.
  cor(hp$HPI,hp$Life.Expectancy)                            # Correlation
  cor(hp$HPI,hp$Well.Being)
  cor(hp$HPI,hp$Footprint..gha.capita.)
```