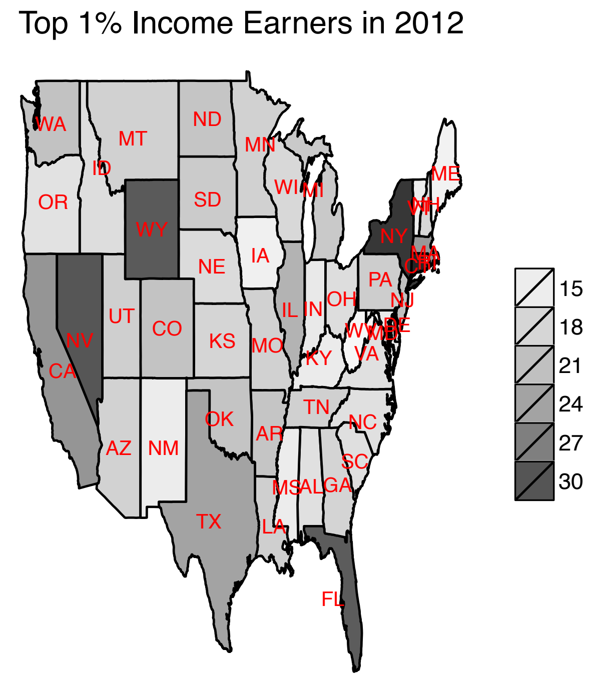
Statistical Methods for Data Science
Mini Project 2

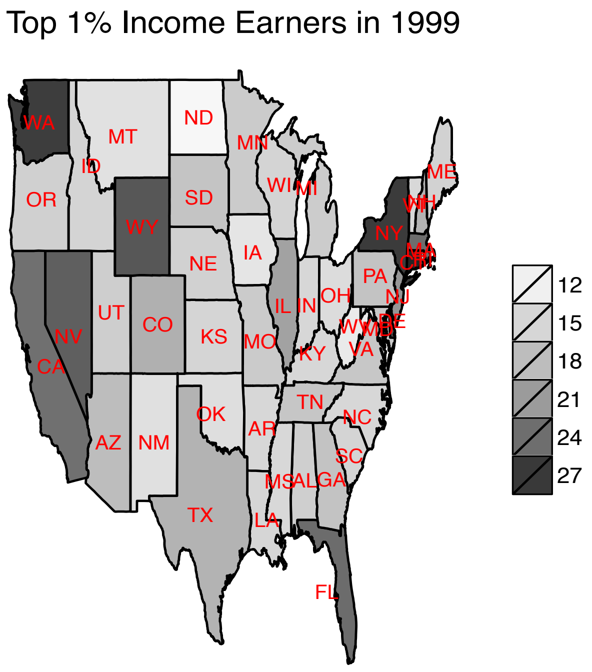
Charu Arora
cxa150730

Exercise 1

a) Plot three maps of the states in USA.



(1) State level income share of the top 1% of income earners in 2012.



(2) State level income share of the top 1% of income earners in 1999.

b) The maps show the top 1% Income Share by each State in the United States in the year 2012, 1999 and the difference in the top 1% Income Share by each State in the United States between 2012 and 1999.

The range of the percentage in the year 2012 is:

[1] 12.50678 33.00785

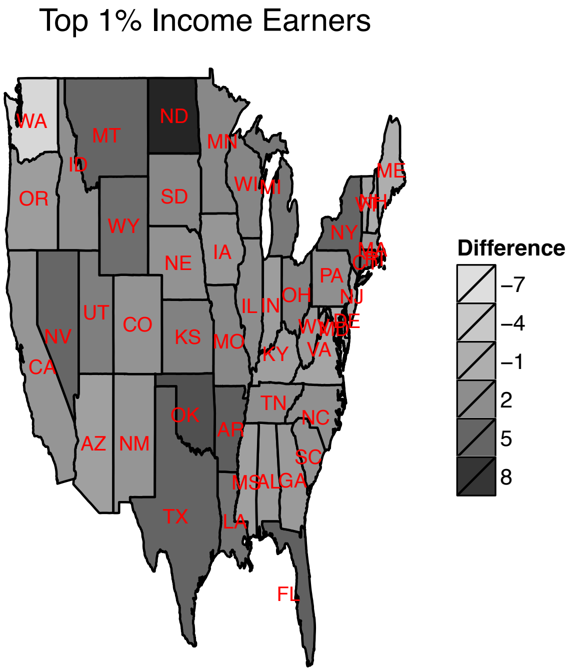
From the map, we can see that in the year 2012, the highest state level income share by top 1% income earners was in Connecticut, New York, Nevada, Wyoming and Florida, all of which having a percentage in the range of 30-34 and the lowest state level income share was in New Mexico, Iowa, Mississippi, Maine and West Virginia, all of which fall in the range 12-15.

In the year 1999, we see that the range of top 1% Income share was comparatively less.

The range of the percentage in the year 1999 is:

[1] 10.74956 28.15289

But the states with the highest state level income share by top 1% was in same states, i.e., Connecticut, New York, Nevada, Wyoming and Florida.



(3) Difference in state level income share of the top 1% of income earners between 2012 and 1999.

North Dakota, West Virginia, Iowa, Montana and New Mexico are the states that had the lowest income share in the year 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%.

Exercise 2

a) The Happy Planet Index measures the extend to which it delivers a long and happy life to the people that live in a particular country based on:

- Life Expectancy

This is measured by the life expectancy data generated for each country.

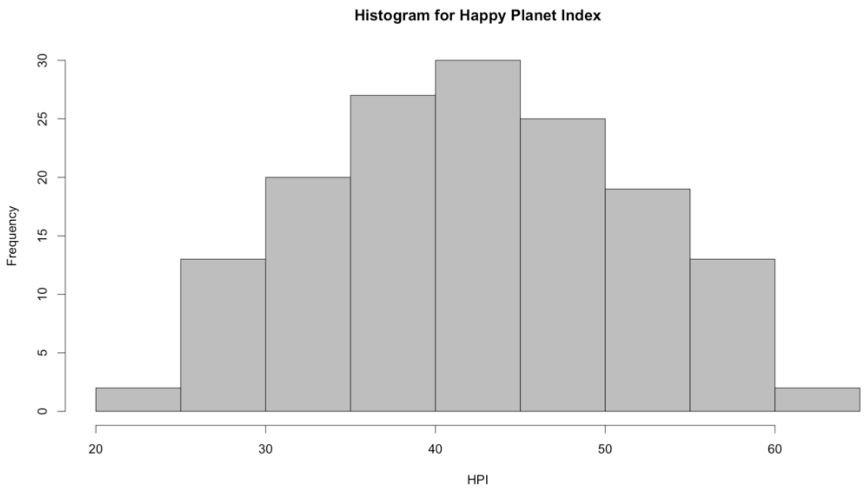
- Experienced Well-Being

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



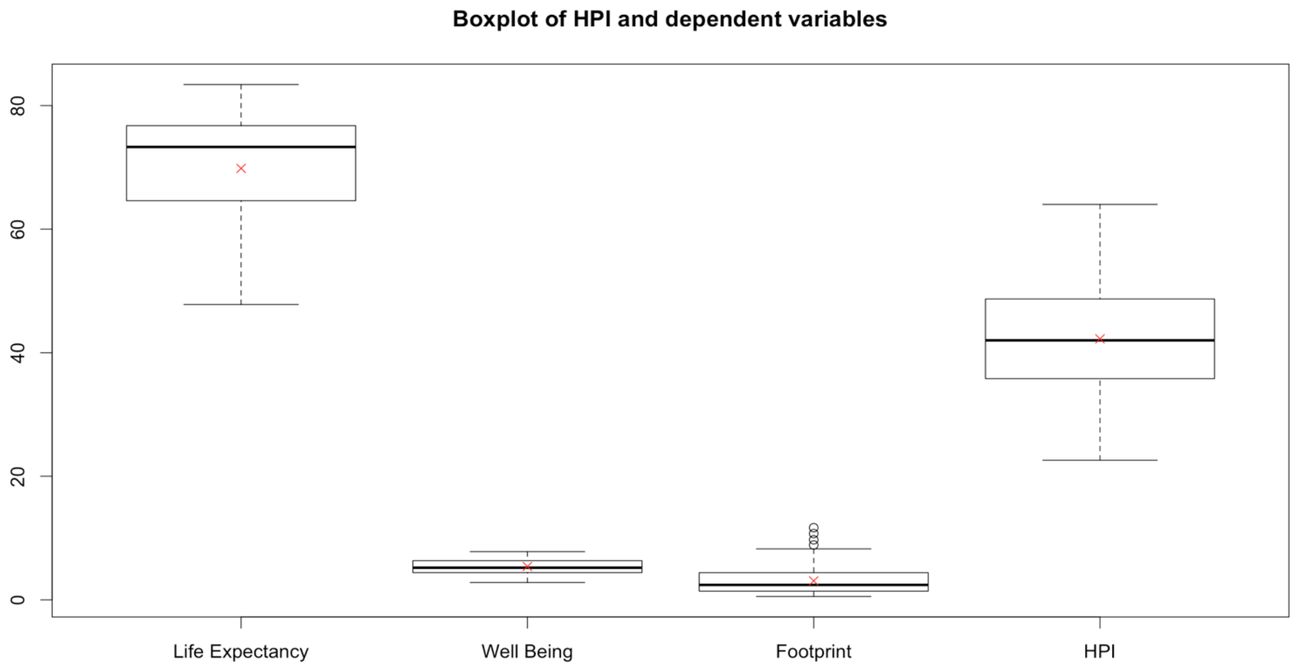
b) The distribution of the HPI is approximately Normal as seen from the histogram. From the mean and median values generated using R, its seen that HPI is very slightly right skewed.

```
                >Range
         [1] 22.6 64.0
                > Mean
          [1] 42.24371
  > Standard Deviation
          [1] 9.116641
              > Median
                [1] 42
```

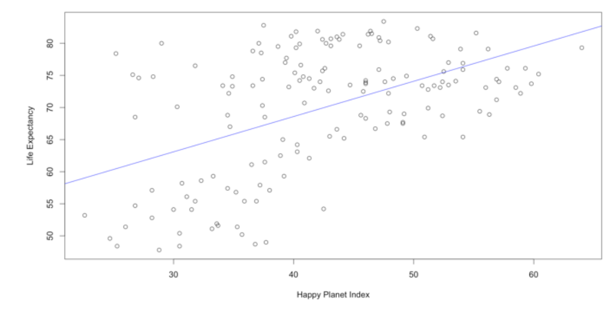
For measure of central tendency and spread of the happy planet index, our best option here is mean and standard deviation as our data is approximately normally distributed. IQR and median would be a better option when the data is skewed and/or have outliers. The distribution tells us that 68% of our data falls within the one standard deviation of the mean and 95% are within the two standard deviations.

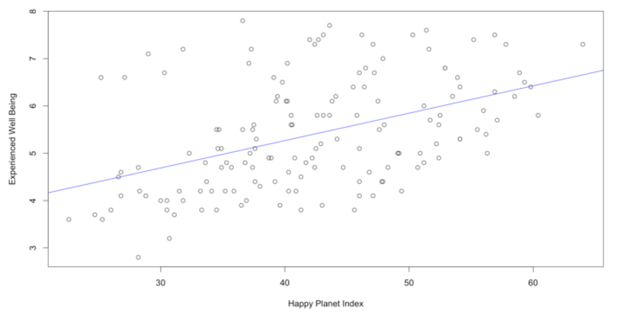
Below is a boxplot representation of the HPI variable along side the variables used to calculate HPI.

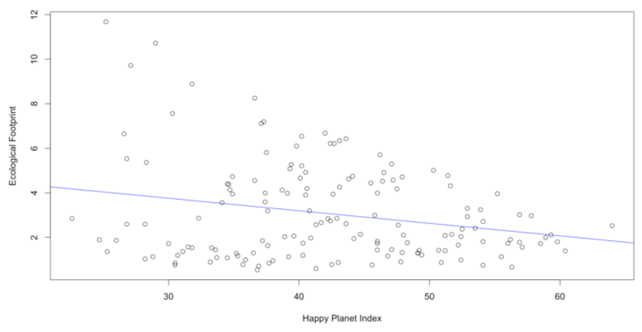
We can clearly see that Life Expectancy is left skewed and that Footprint is right skewed and contains outliers too. For measure of central tendency and spread of Life expectancy and Footprint, IQR and Median would be a better option but for HPI, as it does not contain and outliers, and that it is approximately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



c) Scatterplots of Happy Planet Index with Life Expectancy, Experienced Well-Being and Ecological Footprint.

<table>
<tr>
<td>


</td>
<td>
The scatterplot of HPI with Life Expectancy 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
[1] 0.5111565
</td>
</tr>
<tr>
<td>


</td>
<td>
The scatterplot of HPI with Well Being has a weak positive correlation. Not much can be interpreted from this scatterplot.
> Cor between HPI and Experienced Well being
[1] 0.4510568
</td>
</tr>
<tr>
<td>


</td>
<td>
The scatterplot of HPI with Ecological Footprint seems to be a weak negative correlation. As the value of HPI increases, the ecological footprint decreases. The scatterplot also shows us that there’s a possibility of outliers when the HPI is low.
> Cor between HPI and Ecological Footprint
[1] -0.2380059
</td>
</tr>
</table>

Correlations help us measure the strength of a linear relationship between two variables. Here, since in our scatterplot a strong linear relationship does not exists, calculating the correlation does not make sense. The only scatterplot 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).

R Programming Code
Exercise 1

```
>library(maps)                                # Required to create maps using R
>library(ggplot2)                             # Required to plot the map
>library(plyr)                                # Required to join two data frames



>usa_Dt<-map_data("state")                    # Contains the state map of USA
>str(usa_Dt)
>colnames(usa_Dt)[5]<-"State"                 # to join two data frames, the column name needs to be the same
>usa_Dt$State <-as.factor(usa_Dt$State)
>str(usa_Dt$State)
>levels(usa_Dt$State)
>usaData<-read.xls("usstates.xls");             # reads the top 1% data into data frame



>usa_Dat<-subset(subset(usaData,select = c("State","Top1_adj","Year")),Year==2012)
                                    # selects only those attributes which are required for the year 2012
>usa_Dat$State =tolower(usa_Dat$State)

>usa_Dat1<-subset(subset(usaData,select = c("State","Top1_adj","Year")),Year==1999)
                                    # selects only those attributes which are required for the year 1999
>usa_Dat1$State =tolower(usa_Dat1$State)



>usa_Dat2<-join(usa_Dat,usa_Dat1,by="State",type="inner")
>colnames(usa_Dat2)[2]<-"Top1_2012"
>colnames(usa_Dat2)[4]<-"Top1_1999"
>usa_Dat2$Difference<-usa_Dat2$Top1_2012-usa_Dat2$Top1_1999
                                    # takes the difference of the values in the year 2012 and 1999



>abb<-read.csv("us_states.csv")     # table with the state names and their abbreviations
>abb<-subset(abb,select=c("State","Abb"))
>abb$State =tolower(abb$State)


>usa.df<-join(usa_Dt,usa_Dat,by="State",type="inner")      # join data and map values for 2012
>usa.df<-join(usa.df,abb,by="State",type="inner")          # join values with abbreviation data for 2012
>usa.abb <- aggregate(cbind(long, lat, group, as.numeric(Top1_adj)) ~ Abb,
          data = usa.df,FUN=function(x)mean(range(x)))
```

```
        #finding the mean of the range of latitude and longitude for each state to display the the abbreviation in map
>range(usa_Dat$Top1_adj)
>brks <- c(15,18,21,24,27,30,34)            # deciding breaks in map based on the range to fill color
>p<-ggplot() +
geom_polygon(data = usa.df, aes(x = long, y = lat, group = group, fill = Top1_adj),
color = "black",size=0.5) +
  geom_text(data=usa.abb,aes(x = long, y = lat,label = Abb, fill = NULL),color = "red", size=3) +
scale_fill_distiller(palette = "Greys", breaks = brks) +
theme_nothing(legend=TRUE) +
labs(title="Top 1% Income Earners in 2012",fill="")                #plottting the map for 2012
>ggsave(p, file = "usa_map.pdf")
```

```
>usa.df1<-join(usa_Dt,usa_Dat1,by="State",type="inner")     # join data and map values for 1999
>usa.df1<-join(usa.df1,abb,by="State",type="inner")          # join values with abbreviation data for 1999
>usa.abb <- aggregate(cbind(long, lat, group, as.numeric(Top1_adj)) ~ Abb,
                  data = usa.df1,FUN=function(x)mean(range(x)))
>range(usa.df1$Top1_adj)
>brks <- c(12,15,18,21,24,27,30)            # deciding breaks in map based on the range to fill color
>p<-ggplot() +
  geom_polygon(data = usa.df1, aes(x = long, y = lat, group = group, fill = Top1_adj),
             color = "black",size=0.5) +
  geom_text(data=usa.abb,aes(x = long, y = lat,label = Abb, fill = NULL),color = "red", size=3) +
  scale_fill_distiller(palette = "Greys",breaks = brks) +
  theme_nothing(legend=TRUE) +
  labs(title="Top 1% Income Earners in 1999",fill="")          #plottting the map for 1999
>ggsave(p, file = "usa_map1.pdf")
```

```
>usa.df2<-join(usa_Dt,usa_Dat2,by="State",type="inner")     # join data and map values for 2012-1999
>usa.df2<-join(usa.df2,abb,by="State",type="inner")  # join values with abbreviation data for 2012-1999
>usa.abb <- aggregate(cbind(long, lat, group, as.numeric(Difference)) ~ Abb,
                  data = usa.df2,FUN=function(x)mean(range(x)))
>range(usa_Dat2$Difference)
>brks <- c(-7,-4,-1,2,5,8,11)               # deciding breaks in map based on the range to fill color
>p<-ggplot() +
  geom_polygon(data = usa.df2, aes(x = long, y = lat, group = group, fill = Difference),
             color = "black",size=0.5) +
  geom_text(data=usa.abb,aes(x = long, y = lat,label = Abb, fill = NULL),color = "red", size=3) +
  scale_fill_distiller(palette = "Greys", breaks = brks) +
  theme_nothing(legend=TRUE) +
  labs(title="Top 1% Income Earners")                        #plotting the map for the difference (2012-1999)

>ggsave(p, file = "usa_map2.pdf")
```

R Programming Code
Exercise 2

```
>hp<-read.xls("hpi_dat.xls")                                         #read the HPI data into R data frame


>x<-hist(hp$HPI,
    main="Histogram for Happy Planet Index",
    xlab="HPI",
    border="black",
    col="grey")
                                        #plotting the data into a histogram to find how the data is distributed


>boxplot(hp$Life.Expectancy,hp$Well.Being,hp$Footprint..gha.capita. ,hp$HPI,
      names=c("Life Expectancy","Well Being","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)
                                  #Plotting the boxplot for HPI and all the variables that are used to calculate the HPI
                                  #Plotting the mean of each variable along side its boxplot


>plot(hp$HPI,hp$Life.Expectancy,xlab="Happy Planet Index",ylab="Life Expectancy")
                                                                         #scatterplot of HPI with life expectancy
>abline(lm(hp$Life.Expectancy~hp$HPI),col="blue")            #regression line of HPI with life expectancy


>plot(hp$HPI,hp$Well.Being,xlab="Happy Planet Index",ylab="Experienced Well Being")
                                                                             #scatterplot of HPI with well being
>abline(lm(hp$Well.Being~hp$HPI),col="blue")           #regression line of HPI with well being


>plot(hp$HPI,hp$Footprint..gha.capita.,xlab="Happy Planet Index",ylab="Ecological Footprint")
                                                                              #scatterplot of HPI with footprint
>abline(lm(hp$Footprint..gha.capita.~hp$HPI),col="blue")     #regression line of HPI with footprint



>cor(hp$HPI,hp$Life.Expectancy)                              #correlation value of HPI with life expectancy
>cor(hp$HPI,hp$Well.Being)                                   #correlation value of HPI with well being
>cor(hp$HPI,hp$Footprint..gha.capita.)                       #correlation value of HPI with footprint
```