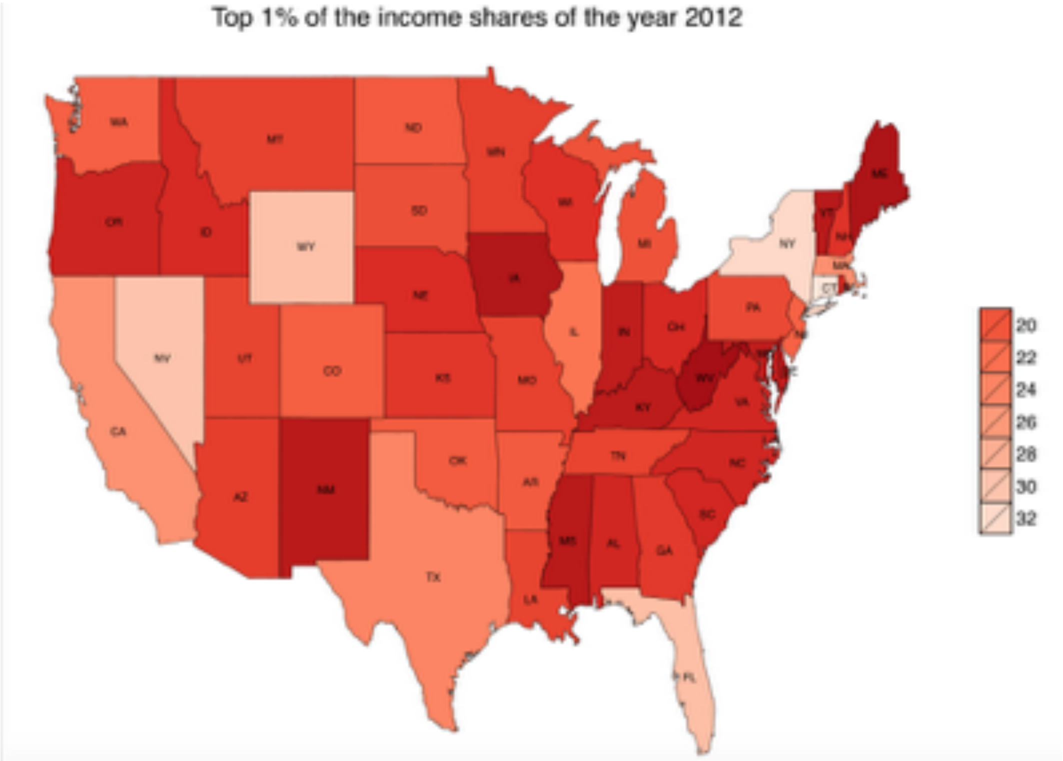
Mini Project: 2
Name: Ajay Vembu

Exercise - 1:

Part - B:

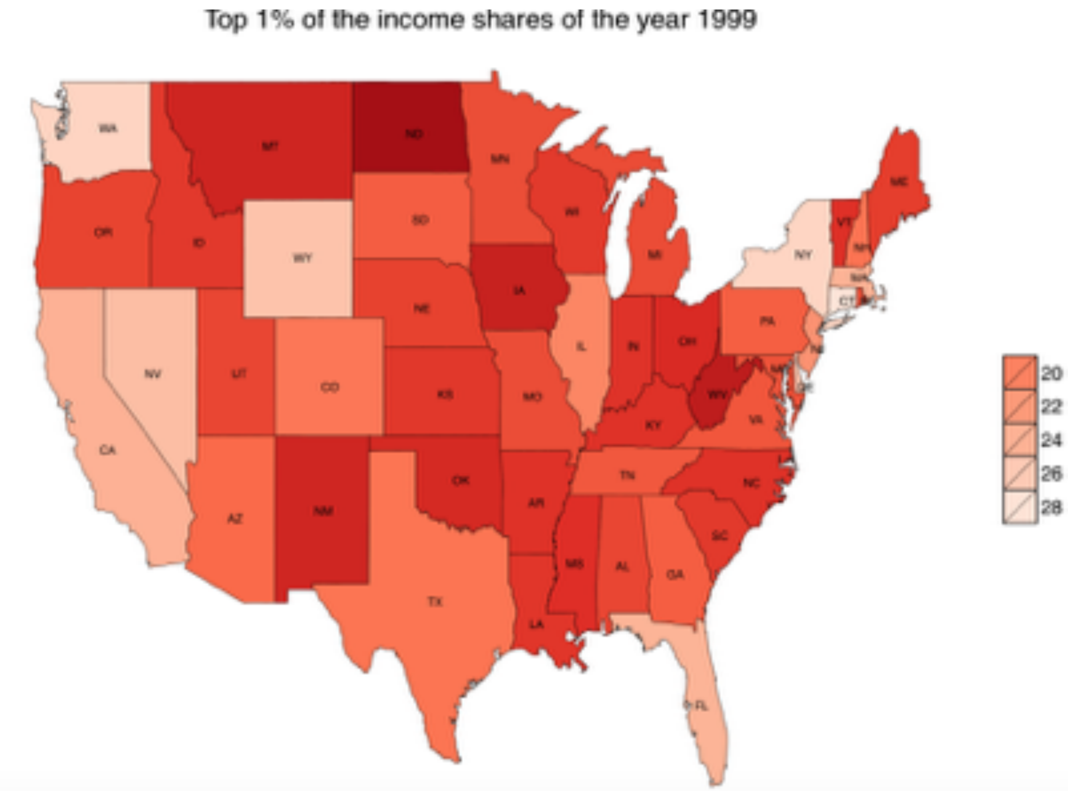
The first map which shows the state level income share of the top 1% of income earners in 2012.



The below can be inferred from the above map,

- The dark colored states like ME, NM, IA, KY, WV etc. are the states with the lowest income share that is less than 20%.
- The light color states like NV, WY, NY are the states with highest income shares that is around 32%
- The state WA which was among the high income share in the year 1999 is not among the high income share in the year 2012.
- The state ND income share has been improved from the year 1999 which was in the low income share.
- In a similar fashion the other states can be related to their respective colors to get the income share.

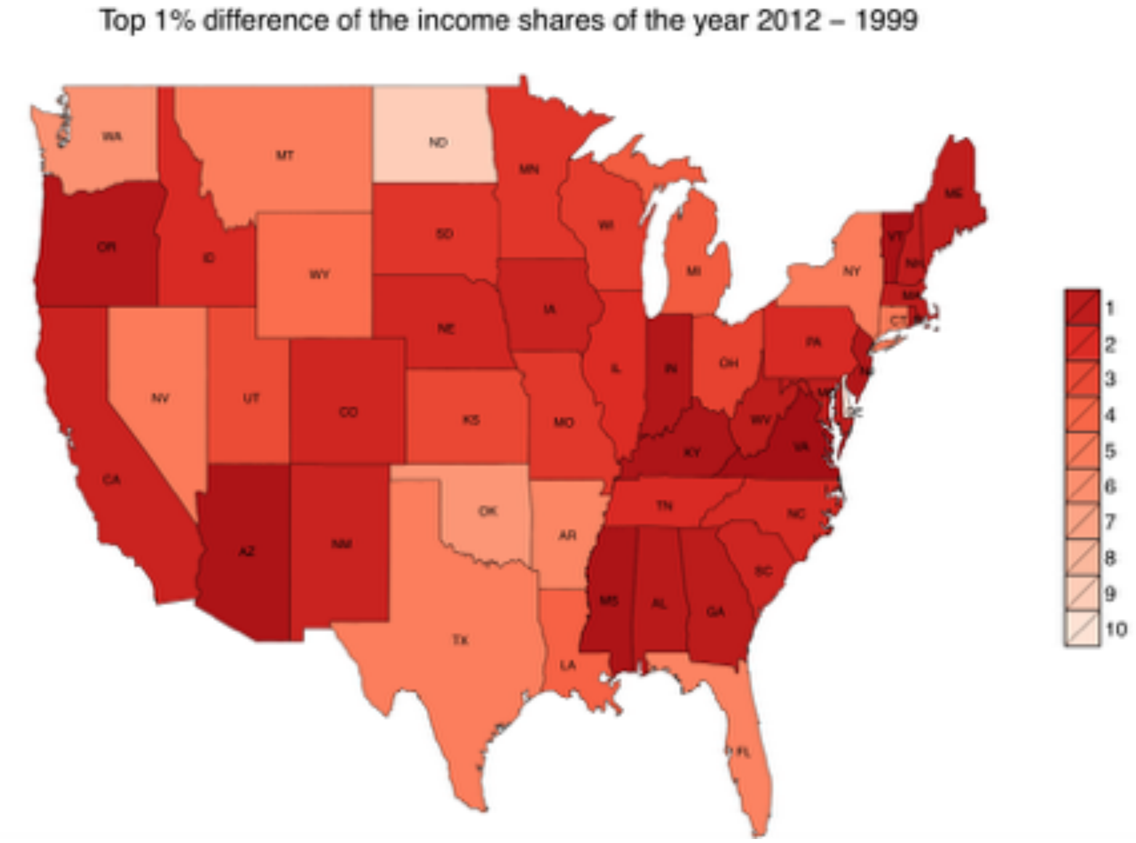
The second map which shows the state level income share of the top 1% of income earners in 1999.


Top 1% of the income shares of the year 1999

The below can be inferred from the above map,

- The dark colored states like ND, WV etc. are the states with the lowest income share that is less than 20%.
- The light color states like NV, WY, NY, WA are the states with highest income shares that is around 32%
- In a similar fashion the other states can be related to their respective colors to get the income share.

The third map which shows the difference in state level income share of the top 1% of income earners between the year 2012 and 1999.



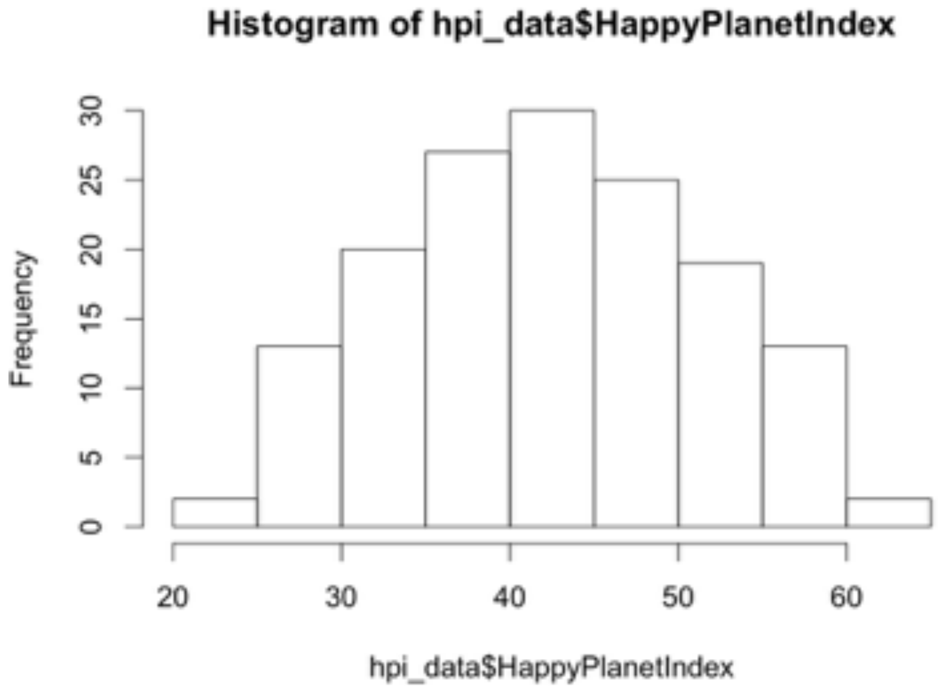
The below can be inferred from the above map,

- The dark colored states like OR, AZ, IN, KY, VA, MS, VT etc. are the states which didn't have any change in the income share between the year 1999 and 2012 and remains around 1%.
- On the other hand light colored state ND showed a significant change in the income share which is around 10%.
- The states like OK, NY CT, FL has shown some changes between the year 1999 and 2012.
- In a similar fashion the other states can be related to their respective colors to get the income share.

Exercise - 1:

Part-B:

Below is the histogram of the plot with the variables Happy Planet Index



The appropriate measure of the center and the spread of this distribution would be mean and standard deviation.

Below are the reasons,

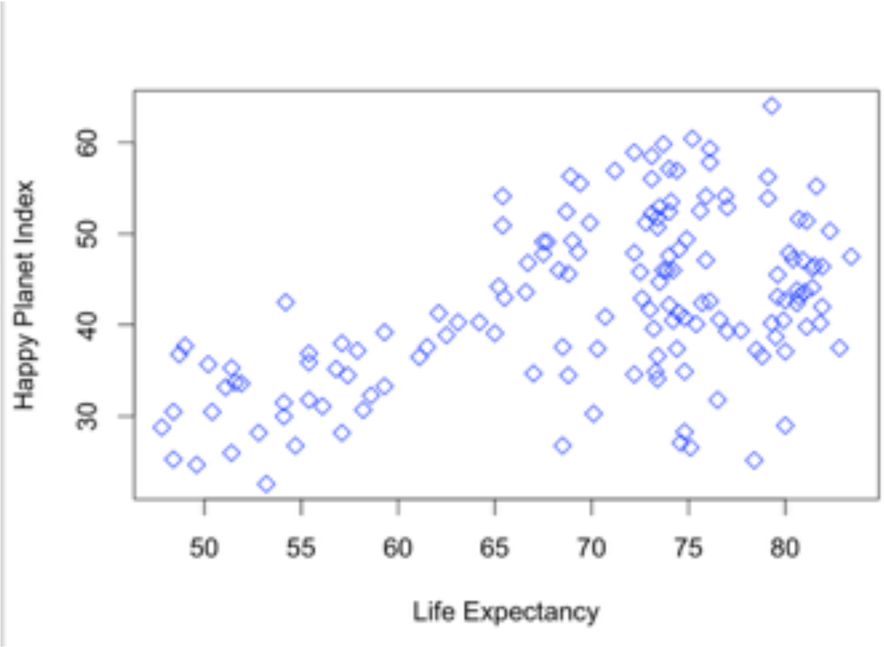
- The above histogram shows that the distribution is approximately normal with no skew either in the left or in the right.
- To use median and IQR as the center and the spread of the distribution, the distribution has to be either right or left skewed.

Part - C:

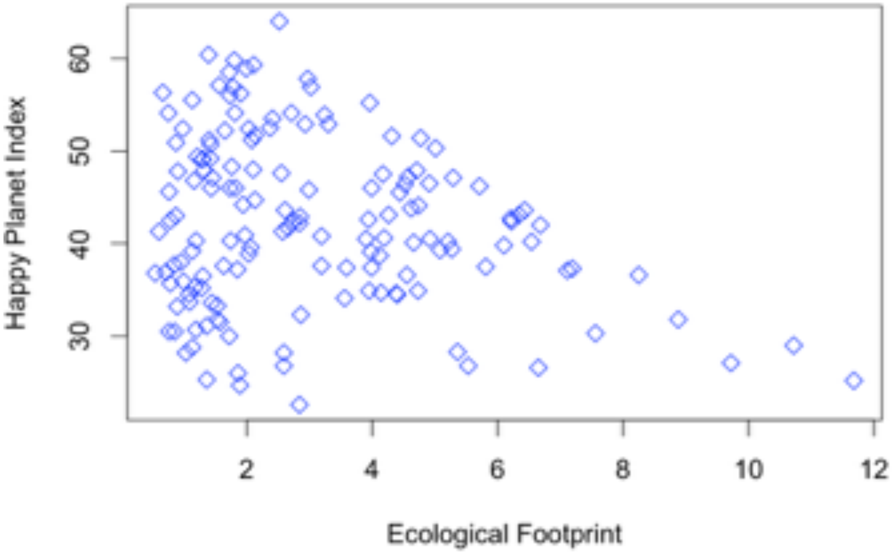
The best three variables which describe the HPI are the below,

- Life Expectancy,
- Ecological Footprint, and
- Well Being.

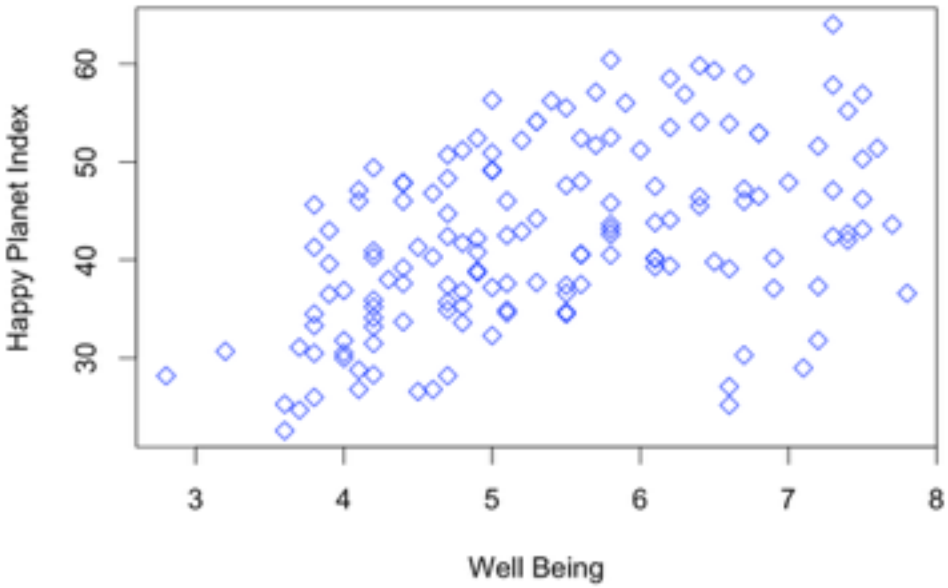
Below are the scatter plots for the HPI against the above mentioned three variables, Scatter plot for HPI against Life Expectancy,



Scatter plot for HPI against Ecological Footprint,



Scatter plot for HPI against Well Being,



As we see a linear relation between the HPI and the corresponding three best variables correlation can be used as a measure between HPI and the variables.

Below are the correlations found in R and their corresponding inferences,

The correlation between Life expectancy and HPI,

0.5111565

This shows that Life expectancy and the HPI has moderate positive relationship.

The correlation between Ecological Footprint and HPI,

-0.2380059

This shows that Ecological Footprint and the HPI has weak negative relationship.

The correlation between Well Being and HPI,

0.4510568

This shows that Well Being and the HPI has moderate positive relationship.

R code for exercise 1:

As a step of convenience in MAC book all the excel files has been translated to CSV files,

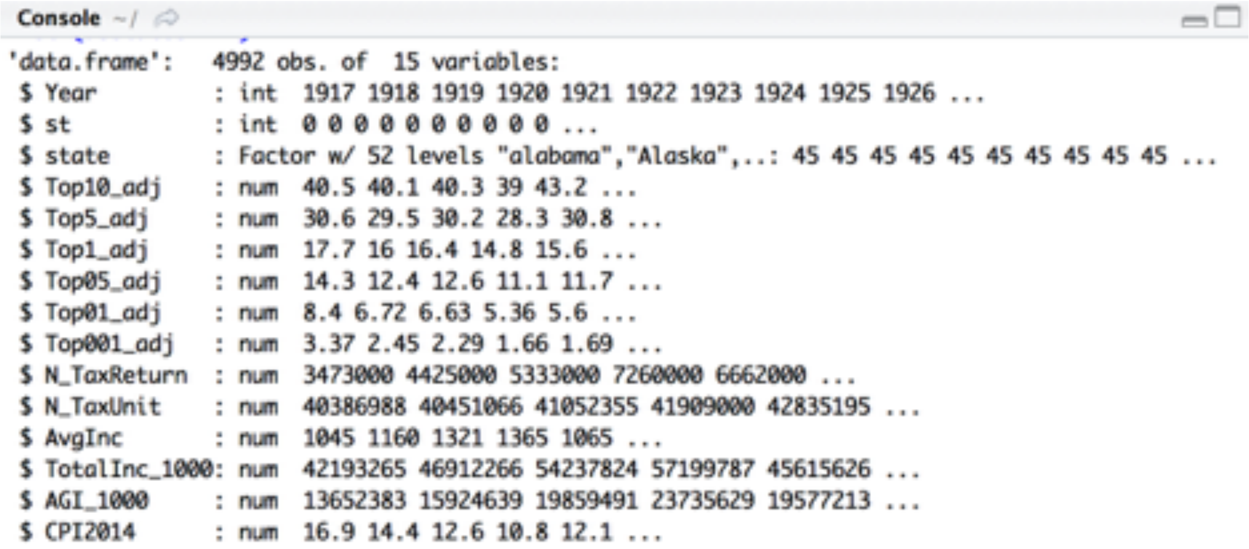
First map:

# the file has been imported,

usstatesWTID <- read.csv("~/Documents/Statistics/projects/miniproject2/usstatesWTID.csv", header=TRUE)

#structure of the file for reference

str(usstatesWTID)



```
'data.frame':   4992 obs. of  15 variables:
 $ Year         : int  1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 ...
 $ st           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ state        : Factor w/ 52 levels "alabama","Alaska",..: 45 45 45 45 45 45 45 45 45 45 ...
 $ Top10_adj    : num  40.5 40.1 40.3 39 43.2 ...
 $ Top5_adj     : num  30.6 29.5 30.2 28.3 30.8 ...
 $ Top1_adj     : num  17.7 16 16.4 14.8 15.6 ...
 $ Top05_adj    : num  14.3 12.4 12.6 11.1 11.7 ...
 $ Top01_adj    : num  8.4 6.72 6.63 5.36 5.6 ...
 $ Top001_adj   : num  3.37 2.45 2.29 1.66 1.69 ...
 $ N_TaxReturn  : num  3473000 4425000 5333000 7260000 6662000 ...
 $ N_TaxUnit    : num  40386988 40451066 41052355 41909000 42835195 ...
 $ AvgInc       : num  1045 1160 1321 1365 1065 ...
 $ TotalInc_1000: num  42193265 46912266 54237824 57199787 45615626 ...
 $ AGI_1000     : num  13652383 15924639 19859491 23735629 19577213 ...
 $ CPI2014      : num  16.9 14.4 12.6 10.8 12.1 ...
```

# the subset data of the year 2012,

dataSetOfYear2012 <- subset(usstatesWTID, Year == 2011)

# the shape file of the USA,

usaDfForTop1Percent.df <- map_data("state")

colnames(usaDfForTop1Percent.df) [5] <- “state"

usaDfForTop1Percent.df$state <- as.factor(usaDfForTop1Percent.df$state)

# the shape file merged with the data file,

```
usaDfForTop1Percent.df <- join(usaDfForTop1Percent.df, dataSetOfYear2012, by = "state", type = “inner")

# the below code has been used as the breaks for both the 1999 and 2012

brks = c(18,20,22,24,26,28,30,32)

# the below code for the plot,

plotForTop1Percent <- ggplot() + geom_polygon(data = usaDfForTop1Percent.df, aes(x = long, y = lat, group = group, fill = Top1_adj),color = "black", size = 0.15) + scale_fill_distiller(palette = "Reds", breaks = brks, trans = "reverse") + theme_nothing(legend = TRUE) + labs(title = "Top 1% of the income shares of the year 2012", fill = "")

# the below code for the state acronyms

states <- data.frame(state.center, state.abb)

states <- states[!(states$state.abb %in% c("AK", “HI")),] # as these states do not belong to the shape file

plotForTop1Percent <- plotForTop1Percent + geom_text(data=states, aes(x=x, y=y, label=state.abb, group=NULL), size=2)

# to save

ggsave(plotForTop1Percent, file = “plotForTop1Percent2012.pdf")
```

Second Map:

```
# the data subset of the year 1999

dataSetOfYear1999 <- subset(usstatesWTID, Year == 1999)
# the shape file of the USA,

usaDfForTop1Percent1999.df <- map_data("state")

colnames(usaDfForTop1Percent1999.df ) [5] <- "state"

 usaDfForTop1Percent1999.df$state <- as.factor(usaDfForTop1Percent1999.df$state)

# the shape file merged with the data file,

usaDfForTop1Percent1999.df <- join(usaDfForTop1Percent1999.df, dataSetOfYear1999, by = "state", type = “inner")

# the below code for the plot,
plotForTop1Percent1999 <- ggplot() + geom_polygon(data = usaDfForTop1Percent1999.df,
aes(x = long, y = lat, group = group, fill = Top1_adj),color = "black", size = 0.15) +
scale_fill_distiller(palette = "Reds", breaks = brks, trans = "reverse") + theme_nothing(legend =
TRUE) + labs(title = "Top 1% of the income shares of the year 1999", fill = “")

# the below code for the state acronyms

states <- data.frame(state.center, state.abb)

states <- states[!(states$state.abb %in% c("AK", “HI")),] # as these states do not belong
to the shape file

plotForTop1Percent1999 <- plotForTop1Percent1999 + geom_text(data=states, aes(x=x,
y=y, label=state.abb, group=NULL), size=2)

# to save

ggsave(plotForTop1Percent1999, file = “plotForTop1Percent1999.pdf")
```

Third Map:

```
# changed the column name of the data set of the year 1999 for the convenience of find the
difference,

colnames(dataSetOfYear1999)[6] <- "Top1_adj_1999"

# the below code for breaks

brksForDifference <- c(1,2,3,4,5,6,7,8,9,10,11,12)

# the two datasets are joined temporarily,

 tempJoinTwoFrames <- join(dataSetOfYear2012,dataSetOfYear1999, by = "state", type =
“inner")

# the below code to find the difference in the income share,

tempJoinTwoFrames$difference <- abs(tempJoinTwoFrames$Top1_adj - tempJoinTwoFrames
$Top1_adj_1999)

# code for shape file

usaDfForTop1PercentDiff.df <- map_data(“state")
colnames(usaDfForTop1PercentDiff.df ) [5] <- “state"
usaDfForTop1PercentDiff.df$state <- as.factor(usaDfForTop1PercentDiff.df$state)

# below code to merge the dataset and the shape file
```

```
usaDfForTop1PercentDiff.df <- join(usaDfForTop1PercentDiff.df, tempJoinTwoFrames, by =
"state", type = “inner")

# plot

plotForTop1PercentDiff <- ggplot() + geom_polygon(data = usaDfForTop1PercentDiff.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 = "Top 1% difference of the income shares of the year 2012 - 1999", fill = "")
> ggsave(plotForTop1PercentDiff, file = “plotForTop1PercentDiff.pdf")

# get the state acronyms

states <- data.frame(state.center, state.abb)

states <- states[!(states$state.abb %in% c("AK", “HI")),] # as these states do not belong
to the shape file

plotForTop1PercentDiff <- plotForTop1PercentDiff + geom_text(data=states, aes(x=x, y=y,
label=state.abb, group=NULL), size=2)

# to save

ggsave(plotForTop1PercentDiff, file = “plotForTop1PercentDiff.pdf")
```

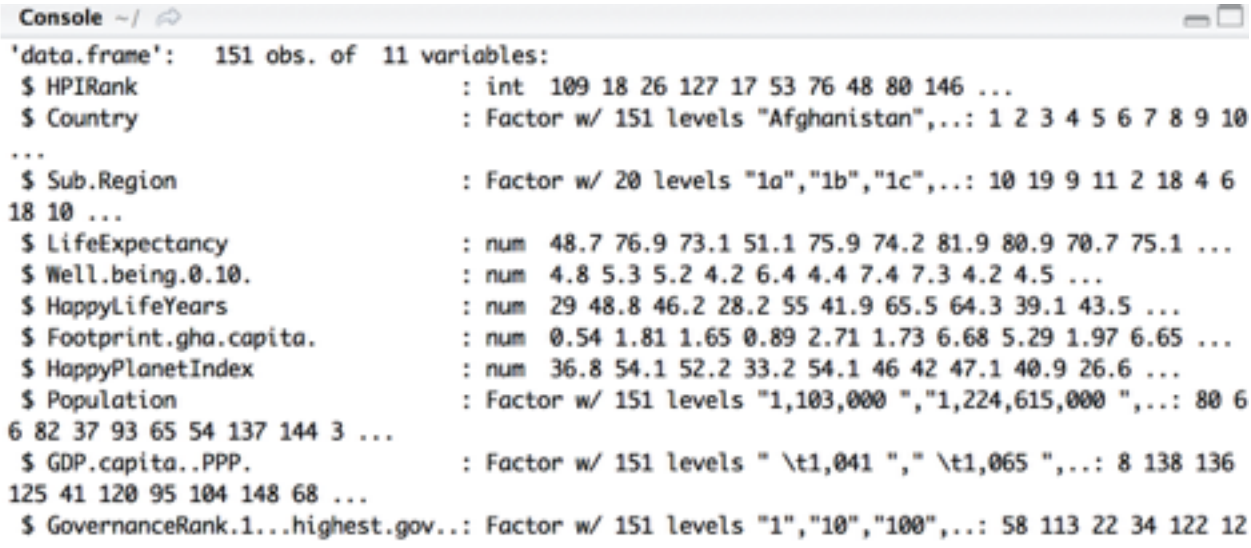
Exercise - 2:

```
# import the cvs file

hpi_data <- read.csv("~/Documents/Statistics/projects/miniproject2/hpi_data.csv",
header=FALSE)

# structure of the file for reference

str(hpi_data)
```



```
'data.frame':   151 obs. of  11 variables:
 $ HPIRank                         : int  109 18 26 127 17 53 76 48 80 146 ...
 $ Country                         : Factor w/ 151 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10
...
 $ Sub.Region                      : Factor w/ 20 levels "1a","1b","1c",..: 10 19 9 11 2 18 4 6
18 10 ...
 $ LifeExpectancy                  : num  48.7 76.9 73.1 51.1 75.9 74.2 81.9 80.9 70.7 75.1 ...
 $ Well.being.0.10.                : num  4.8 5.3 5.2 4.2 6.4 4.4 7.4 7.3 4.2 4.5 ...
 $ HappyLifeYears                  : num  29 48.8 46.2 28.2 55 41.9 65.5 64.3 39.1 43.5 ...
 $ Footprint.gha.capita.           : num  0.54 1.81 1.65 0.89 2.71 1.73 6.68 5.29 1.97 6.65 ...
 $ HappyPlanetIndex                : num  36.8 54.1 52.2 33.2 54.1 46 42 47.1 40.9 26.6 ...
 $ Population                      : Factor w/ 151 levels "1,103,000 ","1,224,615,000 ",..: 80 6
6 82 37 93 65 54 137 144 3 ...
 $ GDP.capita..PPP.                : Factor w/ 151 levels " \t1,041 "," \t1,065 ",..: 8 138 136
125 41 120 95 104 148 68 ...
 $ GovernanceRank.1...highest.gov..: Factor w/ 151 levels "1","10","100",..: 58 113 22 34 122 12
```

```
# histogram of the index

hist(hpi_data$HappyPlanetIndex)

# code for the scatter plots

# life expectancy and the HPI

plot(x = hpi_data$LifeExpectancy, y = hpi_data$HappyPlanetIndex, xlab = "Life Expectancy",
ylab = "Happy Planet Index", col = "blue", pch = 5)

# footprint and the HPI

plot(x = hpi_data$Footprint.gha.capita., y = hpi_data$HappyPlanetIndex, xlab = "Ecological
Footprint", ylab = "Happy Planet Index", col = "blue", pch = 5)

# well being and the HPI

plot(x = hpi_data$Well.being.0.10., y = hpi_data$HappyPlanetIndex, xlab = "Well Being", ylab =
"Happy Planet Index", col = "blue", pch = 5)

# below code to find the correlation

# life expectancy and the HPI

correlationBwLeAndHpi <- cor(hpi_data$LifeExpectancy,hpi_data$HappyPlanetIndex)
correlationBwLeAndHpi

# footprint and the HPI

correlationBwFpAndHpi <- cor(hpi_data$Footprint.gha.capita.,hpi_data$HappyPlanetIndex)
correlationBwFpAndHpi

# well being and the HPI

correlationWbFpAndHpi <- cor(hpi_data$Well.being.0.10.,hpi_data$HappyPlanetIndex)
correlationWbFpAndHpi
```