Golas, Ekal

UNIVERSITY OF TEXAS AT DALLAS | 800 W CaMPBELL RD RICHARDSON TX 75080

MINI PROJECT 2

Statistical methods for data science



Table of Contents

[Methodology 2](#_Toc431904931)

[Exercise 1 2](#_Toc431904932)

[Exercise 2 2](#_Toc431904933)

[Answers 3](#_Toc431904934)

[Exercise 1 3](#_Toc431904935)

[Exercise 2 4](#_Toc431904936)

[R Code 7](#_Toc431904937)

[Exercise 1 7](#_Toc431904938)

[Exercise 2 8](#_Toc431904939)

[References 9](#_Toc431904940)

# Methodology

As discussed in the problem statement (Choudhary, Mini Project 2, 2015) , the algorithms for both the exercises are as follows:-

## Exercise 1

As stated in (Choudhary, 05\_map\_making.R, 2015), the following packages were downloaded:-

1. Maps - to get map shape file
2. ggplot2 - for plotting and miscellaneous things
3. ggmap - for plotting
4. plyr - for merging datasets
5. scales - to get nice looking legends
6. XLConnect - to read xls file

Then the algorithm to implement the solution in R is:-

1. Get a dataframe of the map for the states of USA
2. Make state as the factor variable to be joined to other dataframes
3. Read the top 1% income data from the excel file
4. Change column names of the read data to match it with the map dataframe so that these dataframes can be joined
5. Extract the data for 2012, 1999 and the difference between 2012 and 1999
6. Join each extracted data set with the map dataframe
7. Aggregate the latitude and longitude values within each state to get the center point where we will display the state abbreviation
8. Plot the maps for all 3 datasets and save them in different files

## Exercise 2

The algorithm to implement the solution in R is:-

1. Load the HPI Complete Dataset from the excel file
2. Display a box plot for the Happy Planet Index
3. Display scatter plots of HPI with life expectancy, ecological footprint and well-being and observe their plots
4. Calculate correlations between HPI and these variables and output the values

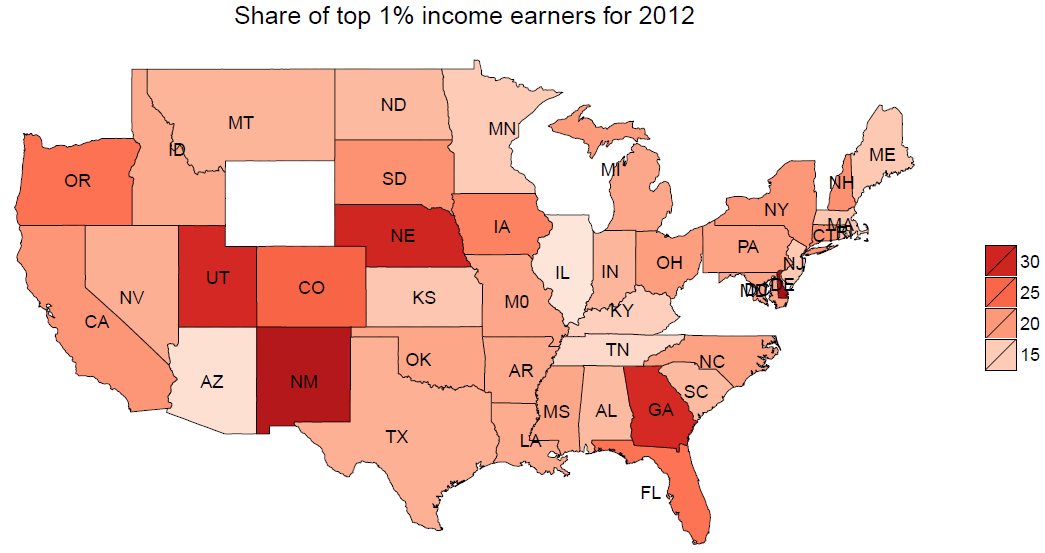
# Answers

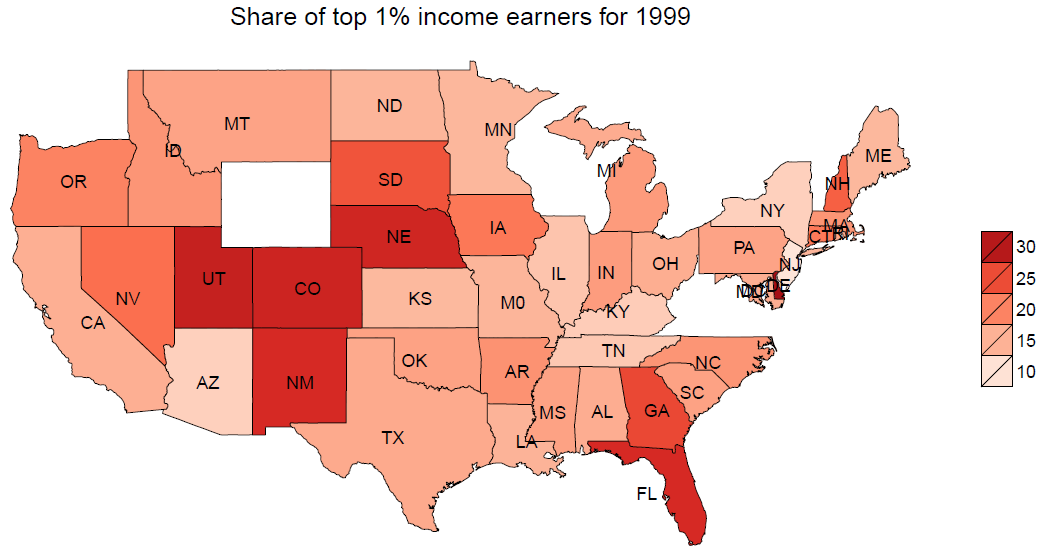
## Exercise 1

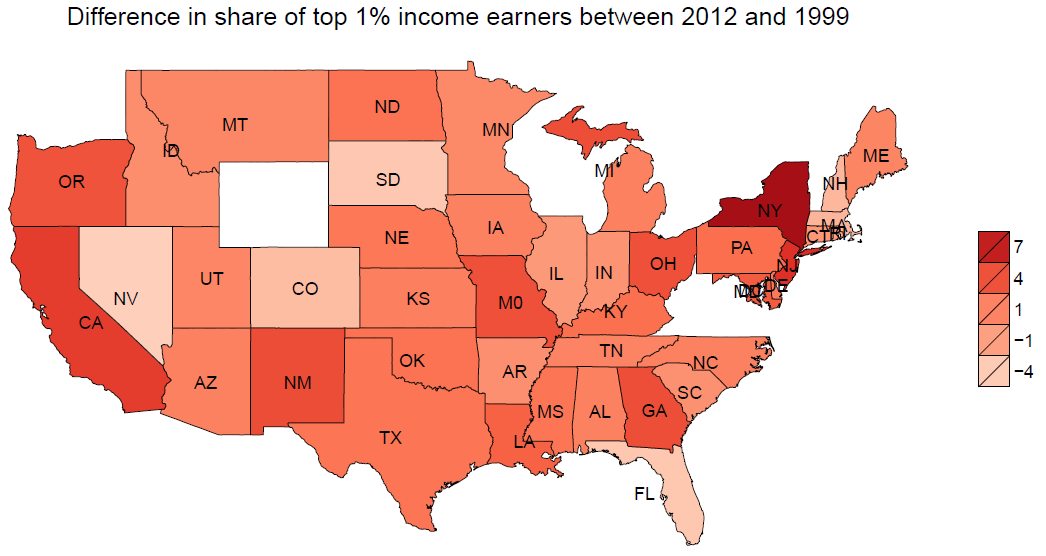
The answers to the questions stated in (Choudhary, Mini Project 2, 2015) are as follows:-

1. **What do the maps show? Justify your conclusions.**

The maps are displayed as follows:-







From these maps we can observe that:-

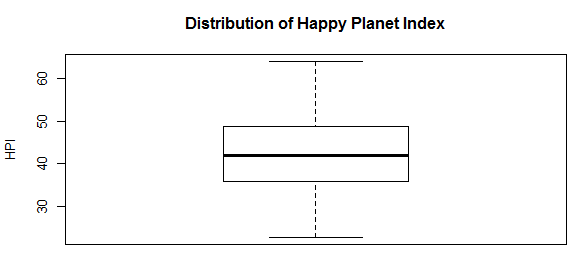
* The range of distribution has decreased from [10, 30] in 1999 to [15, 30] in 2012. So the share of income is distributed more evenly in 2012
* Some states like NY have seen an increase in share while other states like FL have seen a decrease, suggesting that some states might provide better conditions and right environment for businesses to grow and generate income than the other states
* Most of the income share is spread in the middle states of UT, CO, NE and NM

## Exercise 2

The answers to the questions are stated as follows:-

1. **Examine the distribution of the HPI variable graphically.**

The graphical denotation of the variable is as:-



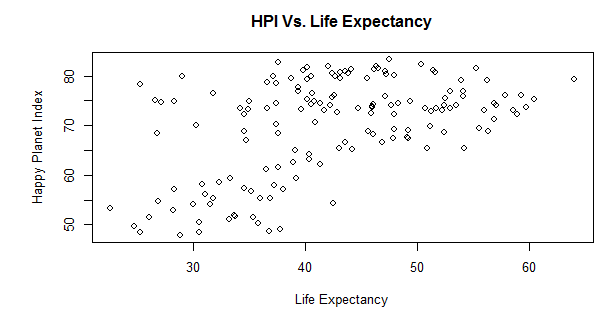
1. **What would be appropriate measures of center and spread of this distribution: (mean, SD) or (median, IQR). Justify your answers**

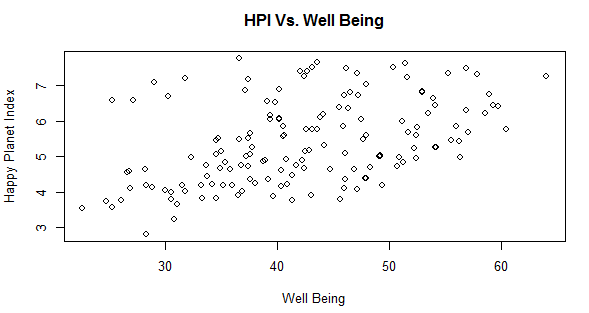
Mean, SD can be an appropriate measure because:-

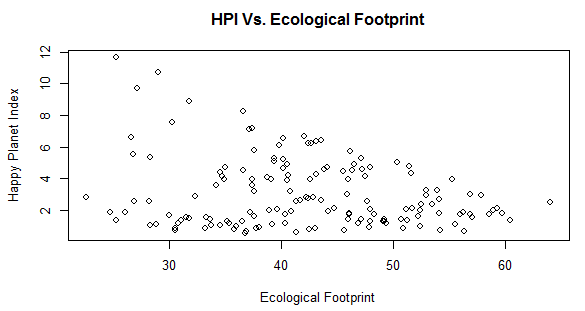
* The distribution seems normal and mean is the center of that distribution. This mean gives the average HPI of countries in the world which takes in account of all the values. Median would take just the middle value of the distribution which is not a good indicator of overall HPI in the world
* Standard deviation is a more robust measure as it takes into account how every value varies from the mean. If we use IQR we ignore the outliers in the dataset but here the dataset hardly contains outliers

1. **Make scatterplots of HPI against each of the three variables on which the index is best. Comment on what you see.**

The scatterplots are as follows:-







From the plots it is observed that:-

* The HPI variable can be expressed a function of these three variables
* The relation of HPI with these variables is not exactly linear but somewhat more complex
* Ecological footprint has a negative relation but the other two variables have a positive relation

1. **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**

Yes, we can use correlation to summarize the relationship of HPI with the other three variables because:-

* Correlation can tell that whether the relation is linear or not
* Correlation can tell how much a relation tends to be linear
* Correlation can tell the alignment of a relation between variables – negative or positive

Correlations as per output of the [R Code](#_Exercise_2) are as follows:-

Correlation of HPI with life expectancy: 0.5109242

Correlation of HPI with well being: 0.4530088

Correlation of HPI with footprint: -0.2382588

This explains that:-

* Life expectancy has the most linear relation with HPI
* Footprint has a negative relation with HPI
* None of the variable have an exact linear relation with HPI, which means that HPI is a complex function of these variables

# R Code

## Exercise 1

library(maps) # to get map shape file

library(ggplot2) # for plotting and miscellaneuous things

library(ggmap) # for plotting

library(plyr) # for merging datasets

library(scales) # to get nice looking legends

library(XLConnect) # to read xls file

usa.df = map\_data("state") # Get a dataframe of states of USA

# Let's rename 'region' as 'state' and make it a factor variable

colnames(usa.df)[5] = "state"

usa.df$state = as.factor(usa.df$state)

# Read input file

wb = loadWorkbook("usstatesWTID.xls")

data.table = readWorksheet(wb, sheet = "Sheet1", header = TRUE, keep = c("st", "Top1\_adj"))

# Change column names

colnames(data.table) = c("group", "percent")

# Filter the data to read rows related to year 2012, 1999 and 2012 - 1999 difference

# First occurence of 2012 is in line number 96 in data

# First occurence of 1999 is in line number 84 in data

# Construct a sequence to get only the rows related to those years

data.2012.table = data.table[seq(96, 4992, 96), ]

data.1999.table = data.table[seq(84, 4992, 96), ]

data.diff.table = data.2012.table

data.diff.table$percent = data.2012.table$percent - data.1999.table$percent

# Get data frames

usa.2012.dat = as.data.frame.matrix(data.2012.table)

usa.1999.dat = as.data.frame.matrix(data.1999.table)

usa.diff.dat = as.data.frame.matrix(data.diff.table)

# Merge the shape data with the income data by state name

usa.2012.df = join(usa.df, usa.2012.dat, by = "group", type = "inner")

usa.1999.df = join(usa.df, usa.1999.dat, by = "group", type = "inner")

usa.diff.df = join(usa.df, usa.diff.dat, by = "group", type = "inner")

# Map states

cnames = aggregate(cbind(lat, long) ~ state, data=usa.2012.df, FUN = function(x) mean(range(x)))

# Rename states to abbreviations

# Using setNames gives incorrect results, have to do this manually

cnames$state = c("AL", "AZ", "AR", "CA" , "CO", "CT", "DE", "DC", "FL", "GA", "ID", "IL", "IN", "IA", "KS", "KY", "LA", "ME", "MD", "MA", "MI", "MN", "MS", "M0", "MT", "NE", "NV", "NH", "NJ", "NM", "NY", "NC", "ND", "OH", "OK", "OR", "PA", "RI", "SC", "SD", "TN", "TX", "UT")

brks = c(0, 10, 15, 20, 25, 30, 35, 99) # Define breaks

# Plot map

p.2012 = ggplot(usa.2012.df, aes(long, lat)) +

geom\_polygon(aes(group = group, fill = percent), color = "black", size = .1) + # with borders (slower)

geom\_text(aes(long, lat, label = state), data = cnames, size = 3.5) + # Add labels

coord\_map() + # Remove distortion

scale\_fill\_distiller(palette = "Reds", breaks = brks) + # Change color and break

guides(fill = guide\_legend(reverse = TRUE)) + # Reverse the legend

theme\_nothing(legend = TRUE) + # Get rid of background

# Provide a title and remove the guide legend heading

labs(title = "Share of top 1% income earners for 2012", fill = "")

#Plot map for others

p.1999 = ggplot(usa.1999.df, aes(long, lat)) + geom\_polygon(aes(group = group, fill = percent), color = "black", size = .1) + geom\_text(aes(long, lat, label = state), data = cnames, size = 3.5) + coord\_map() +

scale\_fill\_distiller(palette = "Reds", breaks = brks) + guides(fill = guide\_legend(reverse = TRUE)) +

theme\_nothing(legend = TRUE) + labs(title = "Share of top 1% income earners for 1999", fill = "")

brks = c(-10, -7, -4, -1, 1, 4, 7, 10) # Define breaks for difference

p.diff = ggplot(usa.diff.df, aes(long, lat)) + geom\_polygon(aes(group = group, fill = percent), color = "black", size = .1) + geom\_text(aes(long, lat, label = state), data = cnames, size = 3.5) + coord\_map() + scale\_fill\_distiller(palette = "Reds", breaks = brks) + guides(fill = guide\_legend(reverse = TRUE)) + theme\_nothing(legend = TRUE) + labs(title = "Difference in share of top 1% income earners between 2012 and 1999", fill = "")

# Save the map to a file for viewing

ggsave(p.2012, file = "USA\_2012.pdf")

ggsave(p.1999, file = "USA\_1999.pdf")

ggsave(p.diff, file = "USA\_diff.pdf")

## Exercise 2

library(XLConnect) # to read xls file

# Read input file

wb = loadWorkbook("hpi-data.xlsx")

data.table = readWorksheet(wb, sheet = "Complete HPI Dataset", header = TRUE, startRow = 6, startCol = 2)

par(mfrow = c(2, 2)) # Plot 4 figures

boxplot(data.table$Happy.Planet.Index, data = data.table, main = "Distribution of Happy Planet Index", ylab = "HPI") # Display the distribution of HPI variable graphically

# Make scatterplots

plot(data.table$Happy.Planet.Index, data.table$Life..Expectancy, main = "HPI Vs. Life Expectancy", xlab = "Life Expectancy", ylab = "Happy Planet Index")

plot(data.table$Happy.Planet.Index, data.table$Well.being..0.10., main = "HPI Vs. Well Being", xlab = "Well Being", ylab = "Happy Planet Index")

plot(data.table$Happy.Planet.Index, data.table$Footprint..gha.capita., main = "HPI Vs. Ecological Footprint", xlab = "Ecological Footprint", ylab = "Happy Planet Index")

# Calculate correlations

cat("Correlation of HPI with life expectancy:", cor(data.table$Happy.Planet.Index, data.table$Life..Expectancy))

cat("Correlation of HPI with well being:", cor(data.table$Happy.Planet.Index, data.table$Well.being..0.10.))

cat("Correlation of HPI with footprint:", cor(data.table$Happy.Planet.Index, data.table$Footprint..gha.capita.))

# References

Choudhary, P. (2015). 05\_map\_making.R. *E-learning*.

Choudhary, P. (2015). *Mini Project 2.* Richardson, USA.