Group 1 Lab 2 exercises and Assignment 2- Part 1

Thilanka Munasinghe

Data Analytics

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

Group 1, Lab 2/Assignment2 – Part1, February 6th, 2020

Lab2- Part1: 2a, 2b

Do the BOTH (Lab2a, Lab2b)

- Lab2a. Measures of Central Tendency/Histograms/ Data Manipulation:

- Lab2b. Regression
  - using EPI dataset

The Dataset(s)

- http://aquarius.tw.rpi.edu/html/DA

- See slides: Last week slides and in-class work as a reference.
- Code fragments, i.e. they **will not** run as-is, on the following slides as.

Remember a few useful commands

```
head(<object>)
tail(<object>)
summary(<object>)
```

Lab2a

Measures of Central Tendency:

- Generate Central Tendency values for EPI variable
- Generate Central Tendency values for DALY variable

Generate the Histogram for EPI and DALY variables

- Generate the Histogram for EPI variable
- Generate the Histogram for DALY variable

Dplyr exercises

Lab2a:

Using sample_n() function in dplyr, get 5 random data points
From EPI, DALY

Using sample_frac() function in dplyr, get 10% random data points
From EPI, DALY

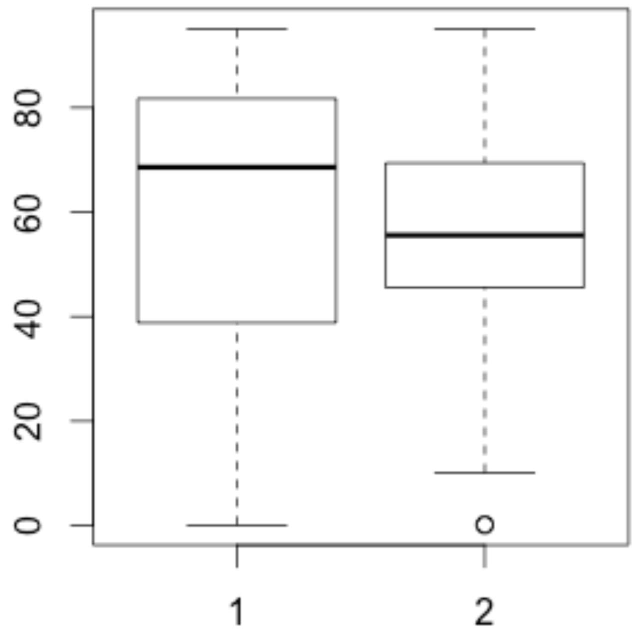
Use the arrange() and desc() functions to arrange values in the descending order in the EPI and DALY and assign them to new variables: *new_decs_EPI* and *new_decs_DALY*

Using the mutate() function, create new columns: double_EPI and double_DALY where multiplying the values in EPI and DALY by 2

Using the summarise() function along with the mean() function to find the mean for EPI and DALY

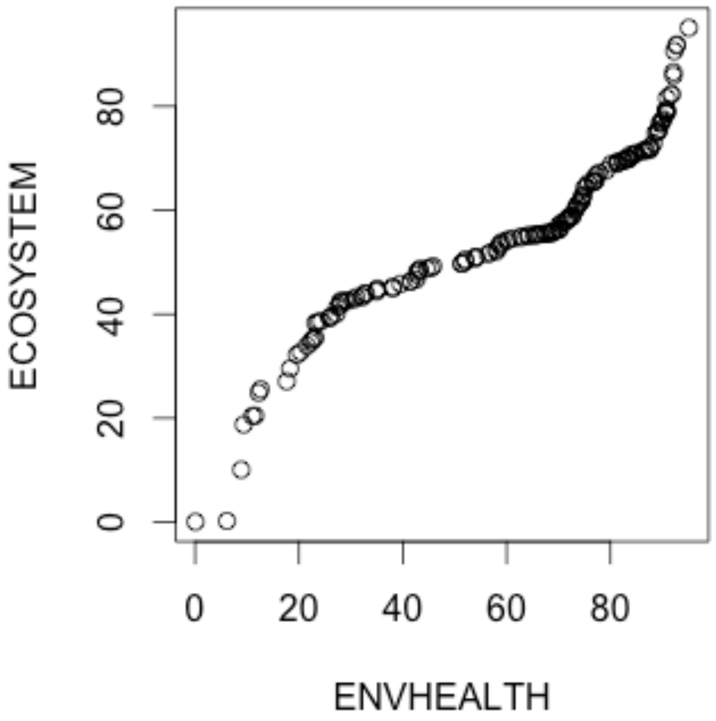
boxplot(ENVHEALTH,ECOSYSTEM)

(Generate the box plot)



qqplot(ENVHEALTH,ECOSYSTEM)

(generate the Q-Q plot)…



2(b):Regression Exercises

- Using the EPI (under /EPI on web) dataset find the single most important factor in increasing the EPI in a *given region*

Linear and least-squares

```
> EPI_data <- read.csv(”EPI_data.csv")
> attach(EPI_data);
> boxplot(ENVHEALTH,DALY,AIR_H,WATER_H)
> lmENVH<-
lm(ENVHEALTH~DALY+AIR_H+WATER_H)


> lmENVH


> summary(lmENVH)


> cENVH<-coef(lmENVH)
```

Predict

```
> DALYNEW<-c(seq(5,95,5))
> AIR_HNEW<-c(seq(5,95,5))
> WATER_HNEW<-c(seq(5,95,5))
> NEW<-data.frame(DALYNEW,AIR_HNEW,WATER_HNEW)
> pENV<-predict(lmENVH,NEW,interval=“prediction”)
> cENV<-predict(lmENVH,NEW,interval=“confidence”)
```

NOTE: Read the documentation for the predict() function in R : https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/predict.lm

Repeat for

AIR_E

CLIMATE

Due Dates:

- Part 1 of the Assignment2 (Lab2 – Part 1) – February 6th
- Part 2 will be given on February 13th, 2020 during the class.
- Due Date: (Both Lab2-Part 1 & Lab2-Part 2 submit together): **17th February, 2020, Monday by 11:59pm.** Submit on LMS.