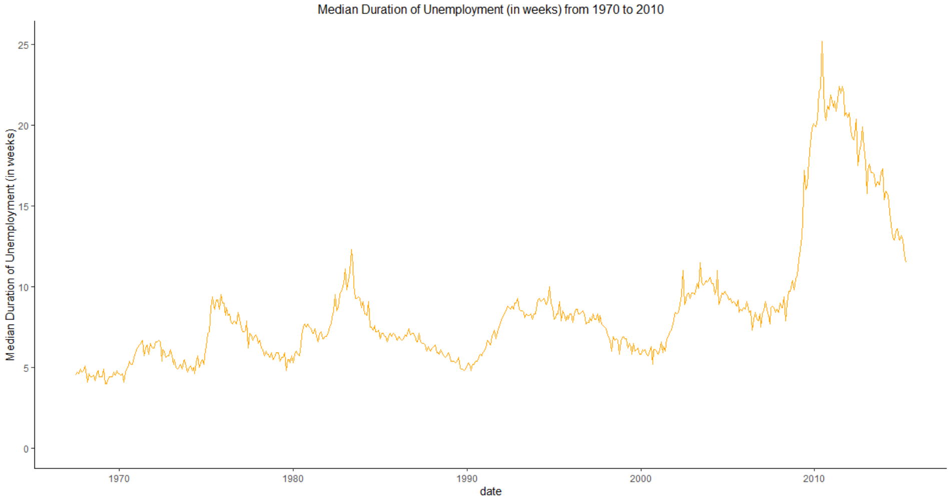
Assignment 1

(i) You are allowed to form a group of up to three individuals for the joint submission of your group work in hardcopy during the lecture on September 14.

(ii) Additionally, email both your R script and your work in PDF format.

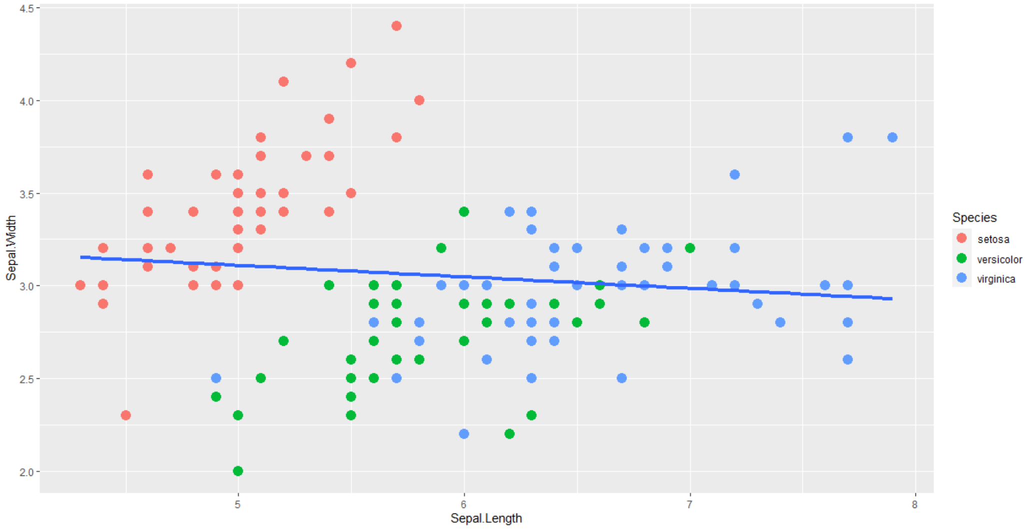
1. [ggplot2]

(a) using economics data set in tidyverse to replicate the time series graphic below. Write down your ggplot2 function.

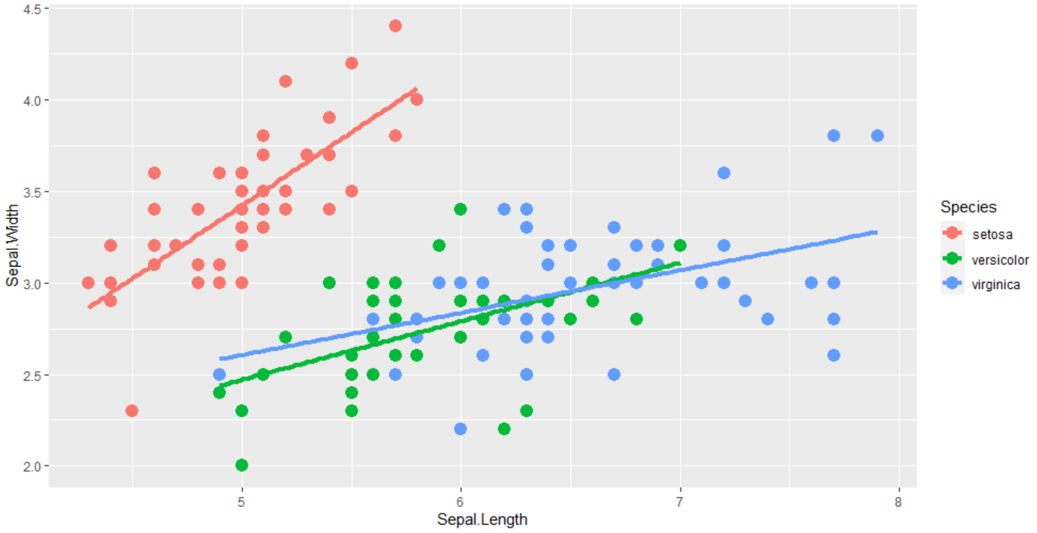


(b) Use the iris data set to generate the following graphs. Write down your codes.

(b1) a regression line fit all the data.



(b2) Three regression lines fit three groups separately.



2. [**loop and statistics**]

Part I (Using loops to answer the following questions)

(a) Write down the function to calculate (1+3+5+7+9+…99). What is the answer to this math question?

(b) Write down the R function to calculate $\sum_{i=1}^{100}(i + 2i^2 + \sqrt{i^3})$. What is the answer to this math question?

Part II

Suppose μ is the population parameter for a population, where $E(X) = \mu$. In addition, $X \sim N(165, 20)$.

(a) Randomly sample 2,000 observations from the population. What is your sample mean $\bar{x}$?

(b) Construct a 95% confidence interval for μ based on the sample from (a). Report this confidence interval. Does it include the true value $\mu = 165$? Does it surprise you? Why?

(c) Now, repeat the process in (a) for 10,000 times, and create a histogram based on these 10,000 sample means, $\bar{x}$s. Label the average of the sample means in your histogram. Does the result surprise you? Why or why not? (Note: you have ten thousand samples. Therefore, you have ten thousand sample means. Find out $\frac{1}{10000}\sum \bar{x}$)

(d) Construct 95% confidence intervals for μ based on each of the 10000 samples in (c). Report the number of confidence intervals that contain the true value $\mu = 165$. Does the answer make sense? Explain. (Note: There are ten thousand samples. Therefore, you can construct ten thousand confidence intervals)

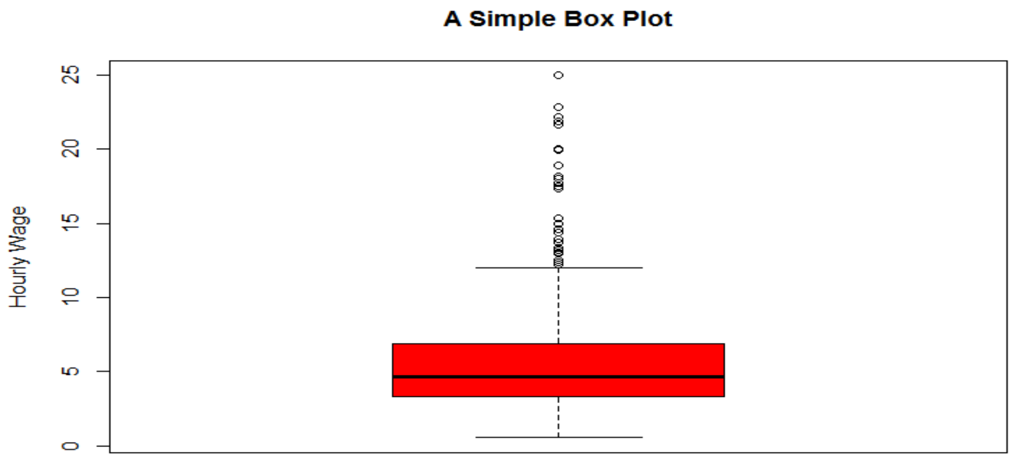
4.

(a) Import WAGE1.dta and rename the dataset to “mydata”.

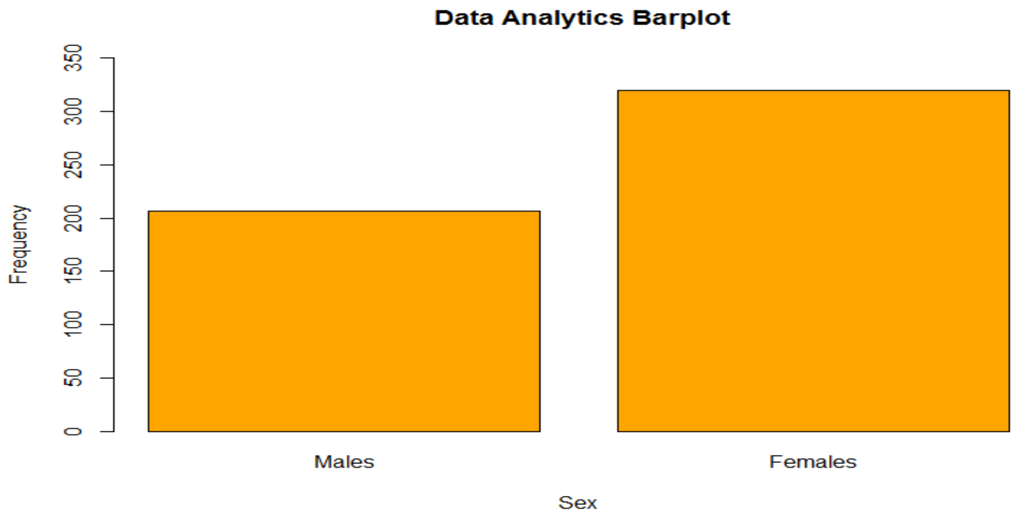
(b) Use the head( ) function to print out first 10 observations.

(c) Calculate the average, max, min and median for (hourly) wage.

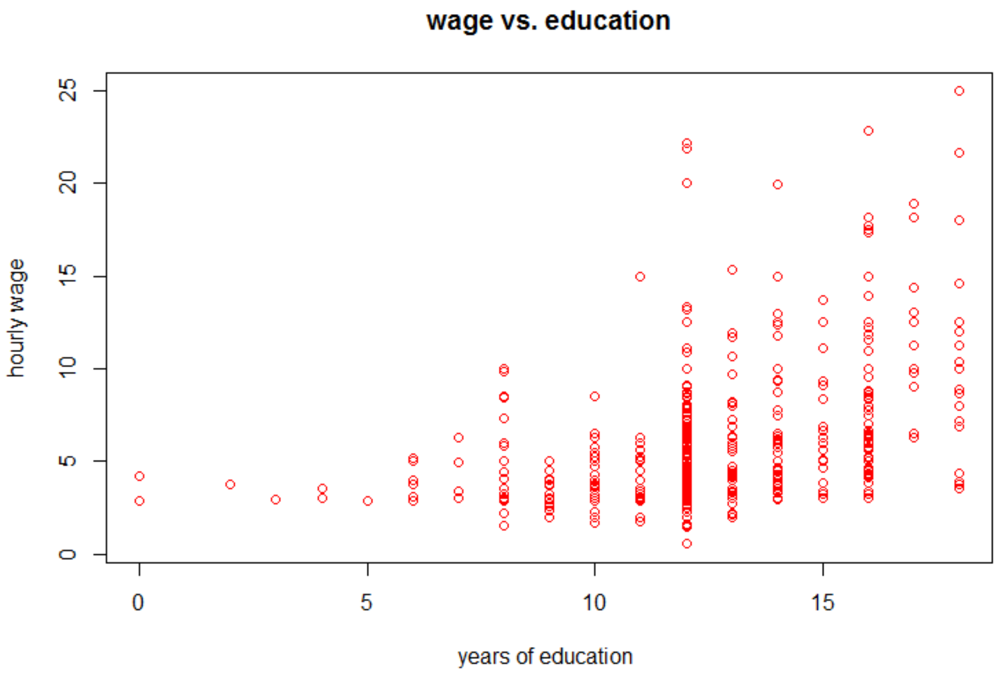
(d) Create a graphic similar to the following box plot for (hourly) wage and change the red color to any other colors you like. Explain your finding.



(e) Create a graphic similar to the following graph and change the orange color to any other colors you like.



(f) Create a scatter plot similar to the example below to show how years of education are related to hourly wage. Change the red color to any other color you like.



5. What are the differences between supervised learning and unsupervised learning? Explain in great details. [You can find the answer from the textbook or online, but you need to use your own words and examples.]