Data Analysis and Visualization - Assignment 1

Ma Jingchun, 2020111235

Due Date: In class on Thursday Oct 9th, 2022.

Please print out your assignment in pages. DO NOT SEND ELECTRONIC COPIES TO MY EMAIL.

Notes:

1. The first thing you need to do is change "Name" and "Student No." of this template. You can modify those in the 3rd line of this Rmd file (author:....). Use *pinyin* in order of last name plus first name, instead of Mandarin character in case that there will be compling errors. For instance, use "Zhang Sansan, 20XXXXXXXX" to replace "Name, Student No.".

2. All you have to do for this assignment is to write R codes in the chunks in this .Rmd file. You can find in each question the words "Please write your R code in this chunk". Just follow this instruction.

3. For questions that require outputs of figures, such as boxplots, please just show them in the R chunk, instead of producing them in R firstly, export them out then insert the plot in Rmd. Just produce graphs in the R chunk in this Rmd file.

4. For questions that involve short answers, put them in "Your comments if needed:" at the end of each question. PLEASE WRITE IN ENGLISH in case of any compiling error. Your language skill will not count for marks.

5. For questions that involve calculation, "PRINT" outputs in the R chunk. For example, if you are required to find the mean of a variable, then in the R chunk, use "mean(variable)" to show the output. Do not write them in words in "Your comments if needed.." section. You have to show that the result is calculated by your own R code rather than anywhere else.

6. DO NOT CHANGE ANYTHING ELSE IN THIS RMD FILE EXCEPT FOR THE R CODE YOU WRITE, ESPECIALLY THE SETUP COMMAND FOR R CHUNKS. Otherwise your R output may not appear.

7. DO NOT COPY CODES FROM OTHERS. STRICT PUNISHMENT WILL FOLLOW IF FOUND.

1. Write a R program to create a sequence of numbers from 20 to 50 and find the mean of numbers from 20 to 60 and sum of numbers from 51 to 91.

```
######### Please write your R code in this chunk #########
### Solution to Q1
20:50
```

```
##  [1] 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
## [26] 45 46 47 48 49 50
```

```
mean(20:60)
```

```
## [1] 40
```

```
sum(51:91)
```

```
## [1] 2911
```

2. Write a R program to get the first 10 Fibonacci numbers with initial two terms as $a_1 = 1$ and $a_2 = 1$.

```
######### Please write your R code in this chunk #########
### Solution to Q2
a <- 1
b <- 1
for (i in 1:10){
  print(a)
  c = a + b
  a = b
  b = c
}
```

```
## [1] 1
## [1] 1
## [1] 2
## [1] 3
## [1] 5
## [1] 8
## [1] 13
## [1] 21
## [1] 34
## [1] 55
```

3. Write a R program to print the numbers from 1 to 100 and print "Fizz" for multiples of 3, print "Buzz" for multiples of 5, and print "FizzBuzz" for multiples of both.

```r
######### Please write your R code in this chunk #########
### Solution to Q3
for (i in 1:100){
  if (i %% (3*5) == 0){
    print("FizzBuzz")
  } else if(i %% 3 == 0){
    print("Fizz")
  } else if(i %% 5 == 0){
    print("Buzz")
  } else{
    print(i)
  }
}
```

```
## [1] 1
## [1] 2
## [1] "Fizz"
## [1] 4
## [1] "Buzz"
## [1] "Fizz"
## [1] 7
## [1] 8
## [1] "Fizz"
## [1] "Buzz"
## [1] 11
## [1] "Fizz"
## [1] 13
## [1] 14
## [1] "FizzBuzz"
## [1] 16
## [1] 17
## [1] "Fizz"
## [1] 19
## [1] "Buzz"
## [1] "Fizz"
## [1] 22
## [1] 23
## [1] "Fizz"
## [1] "Buzz"
## [1] 26
## [1] "Fizz"
## [1] 28
## [1] 29
## [1] "FizzBuzz"
## [1] 31
## [1] 32
## [1] "Fizz"
## [1] 34
## [1] "Buzz"
## [1] "Fizz"
## [1] 37
## [1] 38
## [1] "Fizz"
## [1] "Buzz"
```

```
## [1] 41
## [1] "Fizz"
## [1] 43
## [1] 44
## [1] "FizzBuzz"
## [1] 46
## [1] 47
## [1] "Fizz"
## [1] 49
## [1] "Buzz"
## [1] "Fizz"
## [1] 52
## [1] 53
## [1] "Fizz"
## [1] "Buzz"
## [1] 56
## [1] "Fizz"
## [1] 58
## [1] 59
## [1] "FizzBuzz"
## [1] 61
## [1] 62
## [1] "Fizz"
## [1] 64
## [1] "Buzz"
## [1] "Fizz"
## [1] 67
## [1] 68
## [1] "Fizz"
## [1] "Buzz"
## [1] 71
## [1] "Fizz"
## [1] 73
## [1] 74
## [1] "FizzBuzz"
## [1] 76
## [1] 77
## [1] "Fizz"
## [1] 79
## [1] "Buzz"
## [1] "Fizz"
## [1] 82
## [1] 83
## [1] "Fizz"
## [1] "Buzz"
## [1] 86
## [1] "Fizz"
## [1] 88
## [1] 89
## [1] "FizzBuzz"
## [1] 91
## [1] 92
## [1] "Fizz"
## [1] 94
```

```
## [1] "Buzz"
## [1] "Fizz"
## [1] 97
## [1] 98
## [1] "Fizz"
## [1] "Buzz"
```

4. Write a R program to create three vectors a, b and c with 3 arbitrary integers. Combine the three vectors to become a 3×3 matrix A, where each column represents a vector. Print the content of the matrix A.

```
######### Please write your R code in this chunk #########
### Solution to Q4
a <- sample(1:10, 3, replace = FALSE)
b <- sample(1:10, 3, replace = FALSE)
c <- sample(1:10, 3, replace = FALSE)
A <- cbind(a,b,c)
print(A)
```

```
##      a b c
## [1,] 9 6 2
## [2,] 3 1 4
## [3,] 1 7 5
```

5. Write a R program to find row and column index of maximum and minimum value in a given matrix. Check your code using the matrix A below (do not change A).

```
######### Please write your R code in this chunk #########
### Solution to Q5
set.seed(123)
A = matrix(rnorm(20,0,1), nrow=4, ncol=5, byrow=T)
apply(A,1,min) # row min
```

```
## [1] -0.5604756 -1.2650612 -0.5558411 -1.9666172
```

```
apply(A,1,max) # row max
```

```
## [1] 1.558708 1.715065 1.224082 1.786913
```

```r
apply(A,2,min) # column min
```

```
## [1] -0.5604756 -0.2301775 -1.9666172 -0.6868529 -0.5558411
```

```r
apply(A,2,max) # column max
```

```
## [1] 1.7869131 0.4978505 1.5587083 0.7013559 0.1292877
```

6. Generate a sample of size $n = 100$ from $(\mathbf{X}, Y)$, using a linear model $Y = \mathbf{X}\boldsymbol{\beta} + \mathbf{0.1} \times \varepsilon$, where $\boldsymbol{\beta} = (1, 2, 3)^\top$ and X_1, X_2, X_3 and $\varepsilon \sim N(0, 1)$ independently.

(a) Use OLS to find the estimated β based on the generated sample.

(b) Find the residual vector using $Y - \hat{Y}$, where $\hat{Y} = \boldsymbol{X}\hat{\beta}$, and report the mean squared error (MSE).

```r
######### Please write your R code in this chunk #########
### Solution to Q6
# (a)
e = rnorm(100,0,1)
beta = c(1,2,3)
X = matrix(rnorm(200,1,1),
           nrow=100, ncol=2, byrow=T)
X = cbind(rep(1,100), X)
Y = X %*% beta + 0.1 * e # your model: y=xB+0.1*e
bethat = solve(t(X) %*% X) %*% t(X) %*% Y
rownames(bethat) = c('intercept', 'beta1', 'beta2')
bethat
```

```
##                [,1]
## intercept 1.009222
## beta1     1.998454
## beta2     2.991106
```

```r
# (b)
Yhat = X %*% bethat
MSE = apply((Y - Yhat) ^ 2 , 2, sum) * ( 1 / 100 )
MSE
```

```
## [1] 0.007609865
```