University of Colombo School of Computing

SCS2211: Laboratory II

Assignment – R Programming

Instructions

Execute the R expressions, record the outputs, and write the answers to the questions using Markdown section if required. Indicate the question number in each section.

Save the notebook as a ".ipynb" file. File name should be <Index number>.ipynb (Eg: 2000000.ipynb)

Upload both ".ipynb" and ".csv" files to the given link.

Question 1

Create an artificial dataset using R for employee data of a company for 50 employees using following guidelines. (Give suitable variable names.)

Note: Make sure the assignment will be reproducible when using random values (Use your index number as seed value).

I. Generate Employee IDs starting from ~~E1001 to E1020.~~ E1001 to E1050

II. Randomly assign gender to employees, ensuring a male-to-female ratio of 6:2 (use "Male" and "Female" as category names).

III. Assign an age to each employee, ensuring the age is a random value between 25 & 40.

IV. Assign each employee a job position, randomly selecting from the following categories:

"Software Engineer", "Data Analyst", "System Administrator", "Project Manager"

V. Assign a base salary to each employee, ensuring it is a random value between 100,000 and 200,000 in multiples of 5,000 (e.g., 105,000, 110,000, etc.).

VI. Randomly assign a department from the following list:

"IT", "Finance", "HR", "Marketing"

VII. Each employee receives a performance-based bonus, randomly assigned between 5,000 and 20,000 in multiples of 1,000.

VIII. Randomly generate years of experience between 1 and 15 years for each employee.

IX. Each employee is given a performance rating on a scale of 1 to 5, where:

1 = Poor
2 = Below Average
3 = Average
4 = Good
5 = Excellent

X. Assign Remote Work Eligibility, each employee is either eligible or not eligible for remote work. The probability of eligibility should be 60%.

a. Display the structure of the dataset.

b. Generate summary statistics (e.g., mean, median, min, max) for all numeric variables (such as Age, Salary, Bonus, Years of Experience, Performance Rating)

c. Create a new variable called TotalCompensation that is the sum of Salary and Bonus.

d. Convert the Employee ID from a character string (e.g., "E1001") to a factor variable.

e. Extract a subset of employees who belong to the "IT" department and are eligible for remote work.

f. Create another subset containing only those employees with a Performance Rating of 4 or higher.

g. Calculate the average Salary for each Department (use either the aggregate function or dplyr’s group_by and summarise).

h. Determine the proportion of Male and Female employees in the dataset.

i. Create a histogram to show the distribution of employee Ages.

j. Generate a boxplot comparing Salary distributions across different Job Positions.

k. Create a scatter plot of Years of Experience versus Salary, and add a regression line to this plot.

l. Using group-by operations, compute the average Bonus and average Performance Rating for each Job Position.

m. Compare the average Years of Experience between male and female employees.

n. Construct a 95% confidence interval for the mean Salary of employees in the "Finance" department and interpret what this confidence interval means in the context of the dataset.

o. Write a function that accepts a Department name as its input and returns a list containing:

   The number of employees in that department.
   The average Salary.
   The average Performance Rating.

   Test your function using the "HR" department.

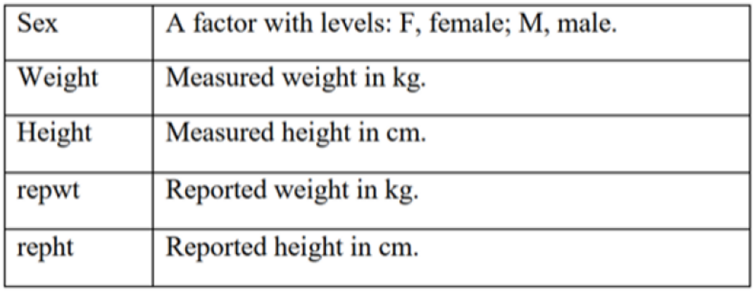
p. Calculate the correlation between Years of Experience and Salary.

q. Fit a linear regression model with Salary as the response variable and Years of Experience as the predictor. Provide a summary of the model and interpret the coefficients.

r. Save the final modified dataset (including any new variables you created, such as TotalCompensation) to a CSV file named "employee_data.csv".

Question 2

The data set named “Davis” contains 200 rows and 5 columns. The subjects were men and women engaged in regular exercise. Variables in the data set are as follows.



Sex	A factor with levels: F, female; M, male.
Weight	Measured weight in kg.
Height	Measured height in cm.
repwt	Reported weight in kg.
repht	Reported height in cm.

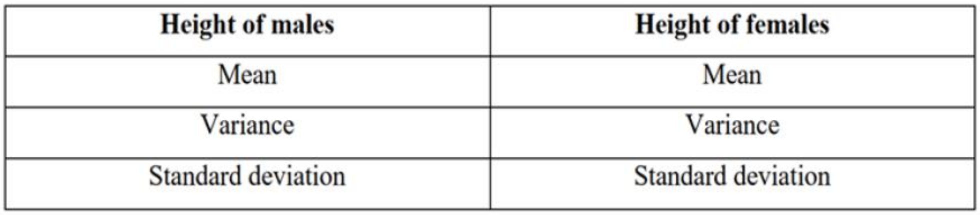
a. Load the data set in the package “carData”.

b. Carry out a descriptive analysis for the above variables and comment on your findings.

c. Find the male proportion in this sample and then construct a 99% confidence interval for the population proportion of males.

d. Create a new data frame named “males” by extracting only the records corresponding to males from the data set “Davis”. Similarly, create a data frame named “females”.

e. Consider the variable “Height” and find the following measures for males and females separately.



Height of males	**Height of females**
Mean	Mean
Variance	Variance
Standard deviation	Standard deviation

f. Calculate the pooled sample standard deviation of height considering males and females as two samples drawn from two populations. [Assume that the two population variances are equal]

g. Obtain a point estimate for the difference between the mean heights of males and females.

h. Construct a 95% confidence interval for the difference between mean heights of males and females.