Cochin University of Science and Technology

Re-accredited by NAAC with ' A+ ' Grade

കൊച്ചി ശാസ്ത്ര സാങ്കേതിക സർവ്വകലാശാല

21-805-0506 R FOR DATA SCIENCE

Dr. Remya K Sasi

ASSISTANT PROFESSOR, DCS, CUSAT

VISION

- To globally excel in innovative research, teaching, and technology development inspired by social obligation.

MISSION

- To contribute to knowledge development and dissemination.
- To facilitate learning and innovative research in frontier areas of computer science.
- To drive students for technology development to solve problems of interest.
- To create socially responsible professionals.

Welcome to the R for Data Science Laboratory!

This document outlines the steps to follow for each lab session, including setting up your environment, working with data, and submitting your work on GitHub.

Installation and software setup

- **Install R:** Follow the installation instructions on the R Project website for your operating system. You can download it for free from https://cran.r-project.org/bin/windows/base/.
- **Install RStudio:** Download and install RStudio following the instructions on their website. https://www.rstudio.com/categories/integrated-development-environment/.

Labcycle submission on GitHub:

Create a Repository: Create a new repository on your GitHub account for this lab session.

Additional Resources:

- R for Dummies by Andrie de Vries & Jennifer Ogle
- R Programming for Data Science by Roger D. Peng & Andrew Schlichting
- RStudio Cheat Sheets: https://posit.co/resources/cheatsheets/

Submission deadlines

Labcycle 1 :July 30, 2024

Labcycle 2 :August 30, 2024

Labcycle 3 :September 30, 2024

Project Presentation deadline : October 30, 2024

Pattern of Evaluation

Continuous Assessment	30 marks (10 marks for each cycle)
Project completion	20 marks
External Examination	50 marks
Total	100 marks

By following these instructions, you'll be well-equipped to successfully complete your R for Data Science Laboratory sessions and showcase your work effectively!

LAB CYCLE 1

1. Develop a program to read a paragraph of text and perform the following tasks: a. Count the total number of words. b. Calculate the average word length. c. Identify and print the longest word. d. Replace all occurrences of a specific word with another word of your choice.

2. String Encryption Write a program that reads a sentence from the user and encrypts it using a simple Caesar cipher. The user can specify the shift value. Implement the encryption such that only alphabetic characters are shifted, while maintaining their case.

3. Data Validation and User Input Develop a program to read student records with their names, ages, and grades. Implement validation checks: a. Ensure age is a positive integer. b. Ensure grade is a valid letter grade (A, B, C, D, F). c. Calculate and display the average age of students with valid records.

4. Password Generator Write a program to generate a random password for a user. The password should include a mix of uppercase letters, lowercase letters, digits, and special characters. Allow the user to specify the length of the password.

5. Series Summation Develop a program to calculate the sum of the series: 1 - 2/3 + 3/5 - 4/7 + ... up to a specified number of terms. Allow the user to input the number of terms in the series.

6. Prime Number Checker Write a program to check whether a given number is prime or not. Implement this using both loops and functions. Additionally, allow the user to input a range and identify all prime numbers within that range.

7. Fibonacci Series with a Twist Develop a program to generate the Fibonacci series, but with a twist. Allow the user to input the number of terms and generate the series where each term is the sum of the last three terms.
8. Palindrome Checker Write a program that reads a string and checks if it's a palindrome. A palindrome is a string that reads the same forwards and backwards, ignoring spaces and punctuation.

9. Data Compression Design a program to read a string and compress it using run-length encoding. In run-length encoding, consecutive repeated characters are replaced with a single character followed by the count of occurrences.
10. Data Reversal Write a program to reverse the order of elements in a given list. Implement this without using any built-in functions or loops

Lab Cycle 2

1. Create a scatterplot of the Sepal.Length and Petal.Length variables in the iris dataset using the plot function? Add appropriate labels and title to the plot. Save the plot as a high-resolution image file.
2. Create a scatterplot of the mpg and disp variables in the mtcars dataset. Use different colors to represent the cyl variable and add a smooth line to show the trend. Add appropriate labels, title, and legend to the plot
3. Create a bar plot of the number of cylinders (cyl) in the mtcars dataset. Use different colors to represent the transmission type (am). Add appropriate title, labels, and legend to the plot.
4. Create a histogram of the miles per gallon (mpg) in the mtcars dataset. Use different shades of blue to represent the frequency of each bin. Add appropriate title and labels to the plot. Calculate and display the mean and standard deviation of mpg on the plot.
5. Create a box plot of the horsepower (hp) in the mtcars dataset. Use different shapes to represent the number of gears (gear). Add appropriate title, labels, and legend to the plot. Identify and label any outliers on the plot.
6. Create a scatter plot of the displacement (disp) versus the weight (wt) in the mtcars dataset. Use different colors and sizes to represent the number of carburetors (carb). Add appropriate title, labels, and legend to the plot. Add a smooth line to show the trend of the relationship.

7. Develop an R program to create a time series plot using real-world data. (https://www.kaggle.com/datasets/niketchauhan/covid-19-time-series-data)
8. Perform EDA on "Titanic Dataset". You are given the Titanic dataset, which contains information about passengers on the Titanic, including their survival status, age, class, and gender. a)plot the histogram of Number of parents and children of the passenger aboard(parch)**.**

   b)Perform a detailed EDA, including advanced statistical analysis, to explore factors influencing survival rates.

   c)Create a customized box plot to visualize the age distribution of survivors and non-survivors.
9. EDA on "Iris Dataset"
   a) For the Iris dataset, which contains measurements of various iris flowers, conduct an EDA.
   a. Determine if there are statistically significant differences in sepal lengths between different species using a suitable statistical test.
   b. Create a pair plot to visualize the relationships between all variables.

Lab Cycle 3

1. Suppose you have a dataset containing information about house prices (dependent variable, denoted as price) and the size of the houses (in square feet, independent variable, denoted as size). You want to build a linear regression model to predict house prices based on their size.

Write an **R** code snippet to perform the following steps:

a) Load the dataset https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

b)

c) Fit a simple linear regression model with price as the dependent variable and size as the independent variable.

d) Calculate the regression coefficients (slope and intercept).

e) Plot the regression line along with the scatter plot of the data points.

2. Graph

a) Create an adjacency list representation for a given undirected graph

b) Implement a function to add an edge between two vertices in the graph.

c) Write an **R** function to perform DFS traversal on a graph starting from a specific vertex.

References

https://r.igraph.org/index.html

3. Suppose we have a dataset of motor trend car road tests (mtcars). The dataset contains information about 32 car brands and 11 attributes. We want to investigate the correlation between the horsepower (hp) and miles per gallon (mpg). Perform a Pearson correlation test to analyze this relationship.
4. Suppose we have a dataset of motor trend car road tests (mtcars). The dataset contains information about 32 car brands and 11 attributes. We want to investigate whether there are any significant variations in the average displacement (disp) across different gear types (gear). Perform a one-way ANOVA test to analyze this

5. We want to investigate the behavior of the total positive COVID-19 cases weekly from 22 January 2020 to 15 December 2020 in India. Perform the following tasks:

Data set link https://raw.githubusercontent.com/datasets/covid-19/master/data/time-series-19-covid-combined.csv

a) Univariate Time Series Analysis:
    i. Create a time series object for the total positive COVID-19 cases
    ii. Visualize the time series data using a line chart.
b) Multivariate Time Series Analysis:
    i. Also, consider the **total deaths** from COVID-19 during the same period.
    ii. Create a multivariate time series object that includes both the total positive cases and total deaths.
    iii. Plot both series on a single chart.

c) Time Series Forecasting:

i. Use the **auto.arima()** function from the **forecast** library to fit an ARIMA model to the total positive cases.

ii. Forecast the next 5 data points.

iii. Plot the forecasted values.

6. The Boston data set comes from the real estate industry in Boston (US). This is a regression problem. The data has 506 rows and 14 columns.

a. Perform data exploration and visualization using R programing.

b. Perform Regression analysis on the dataset.

c. Predict the median value of owner occupied homes.

https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset