VISION

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

Our 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.

Course code 21-805-0506: Lab 9 - R for Data Science Lab

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. 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.
   b) Determine if there are statistically significant differences in sepal lengths between different species using a suitable statistical test.
   c)Create a pair plot to visualize the relationships between all variables.
10. 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:

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

2. Fit a simple linear regression model with `price` as the dependent variable and `size` as the independent variable.
3. Calculate the regression coefficients (slope and intercept).
4. Plot the regression line along with the scatter plot of the data points.

11.

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.

Reference

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

12 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.

13. 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

14.

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

1. Univariate Time Series Analysis:
    - Create a time series object for the total positive COVID-19 cases
    - Visualize the time series data using a line chart.
2. Multivariate Time Series Analysis:
    - Also, consider the **total deaths** from COVID-19 during the same period.

    - Create a multivariate time series object that includes both the total positive cases and total deaths.
    - Plot both series on a single chart.
3. Time Series Forecasting:
    - Use the **auto.arima()** function from the **forecast** library to fit an ARIMA model to the total positive cases.
    - Forecast the next 5 data points.
    - Plot the forecasted values.