

Lab 02: Inferential Statistics and Data Exploration

Due date: Wednesday, Feb 2, 2025 submitted as Word document to Canvas **Lab02** link

This lab counts 9 % toward your total grade.

Objectives: In this lab, you will practice your skills in

- a) Inferential statistics
- b) Confidence interval
- c) Statistical graph
- d) Data exploration

Format of answer: Submit your answers as a **Word document** with graphs and verbal descriptions, properly labeled in the task sequence, with answers in **red text** and only relevant content included

Task 1: Confidence Interval (3 pts)

A survey was conducted on a sample of 1,000 university students to determine the proportion of students who regularly use public transportation. Out of the 1,000 students surveyed, 72.4% reported that they regularly use public transportation.

- a) Calculate the 95% confidence interval for the proportion of students who use public transportation. (1 pts)
- b) Calculate the 99% confidence interval for the same proportion. (1 pts)
- c) Interpret the results. (1 pts)

Task 2. Statistical Graph (3 pts)

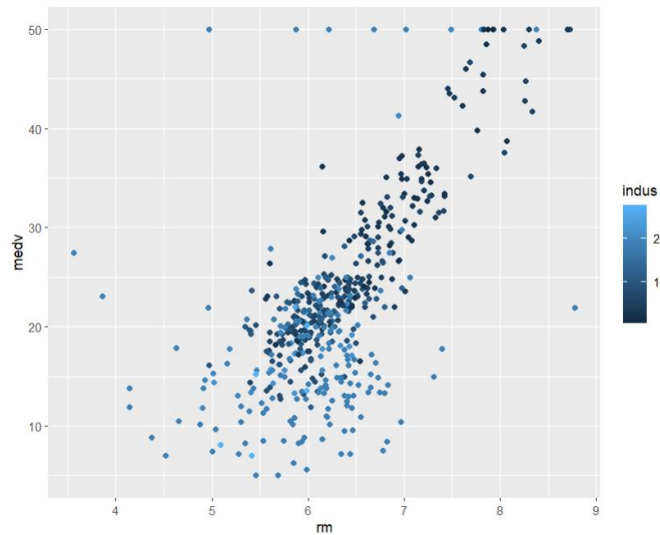
The following graph is reproduction from the [R Graphics Cookbook](#). R has extensive documentation to introduce its functions. You should be able to learn how to understand the documentation and apply the function to your study. Study the code in the [link](#) and reproduce the graph with a different data (**Boston**).

Boston data in MASS package (**MASS::Boston**) contain variables about the housing value in suburbs of Boston.

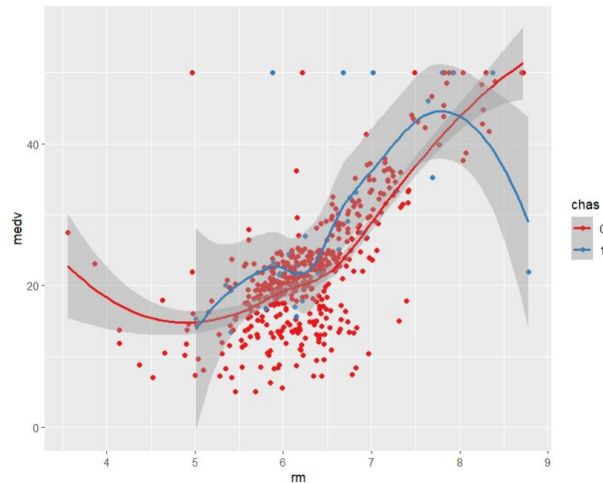
In this task, you will work with ggplot2 to create a scatter plot with additional layers to enhance the visualization. The output can show relationship between two variables, using

different colors to distinguish the third variable. **Show your R code and graphs for this calculation.**

- a) Reproduce the graph below using **Boston** data in **MASS** package. In **ggplot** function, set up parameters as: **aes(x = rm, y = medv, colour = indus)** . (1pts)



- b) Explain the data distribution pattern based on **Boston** data. (0.5pts)
- c) Based on the scatterplot, a smoothing line is added to the plot using the **geom_smooth()** function. Your task is to reproduce the below graph using Boston data in **MASS** package. Before you produce the graph, convert the **chas** variable in **Boston** data to factor:
Boston\$chas = as.factor(Boston\$chas).
Set up the parameters in **ggplot()** function as follows:
aes(x = rm, y = medv, colour = chas). (1.5 pts)



Task 3: Data Exploration (3pts)

The **MplsDemo** Demographic Data 2015 in **carData** package include the demographic data from the 2015 American Community Survey. **Show your R code for this calculation.**

- Import the **MplsDemo** data use the function **data()**. (1 pts)
- Examine the histogram and pairwise relationships between variables using **scatterplotMatrix()**. Identify any skewed pattern visually and provide a description. Please use following variables: **~population + white + black + hhIncome**. (1 pts)
- Evaluate the skewness of the variable identified as having a skewed pattern using **e1071::skewness ()**. (1 pts)