Practice 01: Introduction to R for Statistics

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

1. Import data into R
2. Understand the basics of working with data frames.
3. Learn basic R commands for data manipulation and exploration.
4. Perform summary statistics
5. Create basic statistical graphs.

# Task 1: Setting Up Your Environment

1. Open RStudio.
2. Create a new R script (File > New File > R Script).
3. Use function **setwd()** to setup working directory. Show your R code for this calculation. (0.5 pts)
4. Click  to save your R document

# Task 2: Importing Data

Import **economic\_indicators.csv** and **free\_wifi\_locations.xls** file using function in R. show your R code for this calculation.

1. Use function **read.csv()** toimport **economic\_indicators.csv** file and assign it to an object named **economics**.
2. Use function **read\_excel()** from library **readxl** to import **free\_wifi\_locations.xls** file and assign it to an object named **wifi**.
3. Make a screenshot of **GLOBAL ENVIRONMENT** to display all 2 data-frames.

economics = read.csv('economic\_indicators.csv')

wifi = readxl::read\_excel('free\_wifi\_locations.xls')

A screenshot of a computer

Description automatically generated

# Task 3: Data-Frame Basics

Economic indicators data include values related to topics such employment, housing and real estate development, covering the period from Jan 2013 and Dec 2019. Show your R code for this calculation.

1. Access unemp\_rate and labor\_force\_part\_rate columns.

economics$unemp\_rate

economics$labor\_force\_part\_rate

A table of numbers with numbers on it

Description automatically generated

1. Use **labor\_force\_part\_rate** to minus **unemp\_rate** to calculate the difference between these two values and add the new variable **diff\_unemp\_labor** to the **economics** data-frame. (0.5 pts)

economics$diff\_unemp\_labor = economics$labor\_force\_part\_rate - economics$unemp\_rate

A table of numbers with numbers on it

Description automatically generated

1. Apply the statement

**economics[order(economics$diff\_unemp\_labor, decreasing = TRUE),c('Year','Month')]**

What is this statement doing?

It sorts the rows of the economics data frame based on the values in the diff\_unemp\_labor column in descending order (largest to smallest).

After sorting, it returns a subset of the data frame that includes only the Year and Month columns, arranged according to the sorted order of diff\_unemp\_labor. Based on the result, December 2019 has the highest diff\_unemp\_labor value within the time period.

A computer screen shot of a computer code

Description automatically generated

1. Use **summary()** to see the summary information of the **wifi** data-frame.

summary(wifi)

A screenshot of a computer

Description automatically generated

1. Describe the summary information for **OID\_** and **neightborhood\_id**, and explain why they are different?

summary(wifi$OID\_)

summary(wifi$neighborhood\_id)

A close-up of a white background

Description automatically generated

The OID\_ variable is numerical, so its summary provides detailed statistics such as the minimum value, first quartile, median, mean, third quartile, and maximum value.

On the other hand, the neighborhood\_id variable is categorical (character type). As a result, its summary only displays the length, class, and mode. The length indicates that there are 309 values for this variable, which corresponds to the number of rows in the data frame.

# Task 4: Plot basics

Boxplot analysis

1. Make a boxplot based on column **logan\_intl\_flights** in economics data-frame. (hint: using boxplot() and input variable is **logan\_intl\_flights** from **economics** data-frame)
2. Apply below statement:

**boxplot(logan\_intl\_flights ~ Month, data = economics)**

what insights can we gather about seasonal trends in international flights from grouped boxplot?

A diagram of a diagram

Description automatically generated

Regarding seasonal trend, we can Identify variability and outliers in data.

The thick horizontal line inside the box represents the median (3960) number of international flights. (The median is the central value of the data and divides it into two halves.)

The edges of the box represent the 1st Quartile (Q1, 25th percentile= 3408) and the 3rd Quartile (Q3, 75th percentile= 4516).Additionally, range contains the middle 50% of the data, indicating the most typical range for the number of flights.

Regression line analysis

1. Apply below statements:

**plot(logan\_intl\_flights~Time, data = economics, type = 'l')**

**abline(lm(logan\_intl\_flights~Time, data=economics))**

how does the trend of **logan\_intl\_flights** change over **Time** based on the first plotted line?

**cyclical pattern** of international flights with seasonal fluctuations showing positive long term trend

What does the regression line added to the plot tell us about the relationship between **logan\_intl\_flights** and **Time**?  
strong linear relationship, positive slope, international flights increasing with time

A graph with lines and numbers

Description automatically generated