Practice 01: Introduction to R for Statistics

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

a) Import data into R
b) Understand the basics of working with data frames.
c) Learn basic R commands for data manipulation and exploration.
d) Perform summary statistics
e) Create basic statistical graphs.

Task 1: Setting Up Your Environment

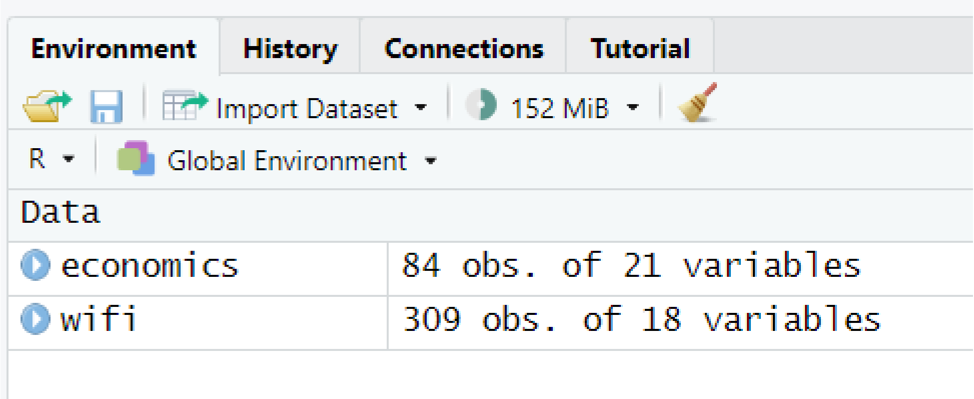
a) Open RStudio.
b) Create a new R script (File > New File > R Script).
c) Use function **setwd()** to setup working directory. Show your R code for this calculation. (0.5 pts)
d) Click 💾 to save your R document

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.

a) Use function **read.csv()** to import **economic_indicators.csv** file and assign it to an object named **economics**.
b) Use function **read_excel()** from library **readxl** to import **free_wifi_locations.xls** file and assign it to an object named **wifi**.
c) 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')
```



economics	84 obs. of 21 variables
wifi	309 obs. of 18 variables

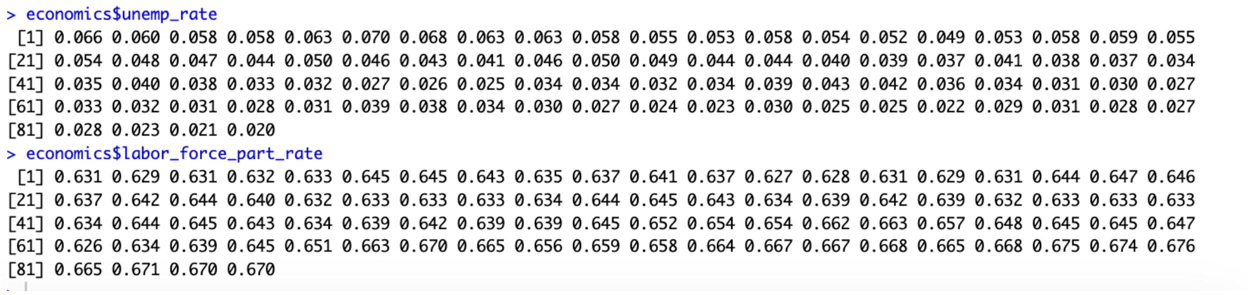
Task 3: Data-Frame Basics

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.

a) Access unemp_rate and labor_force_part_rate columns.

economics$unemp_rate

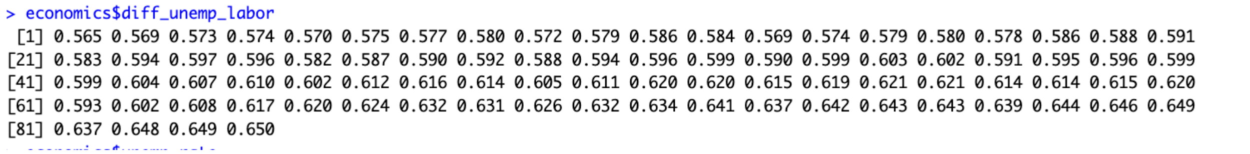
economics$labor_force_part_rate



```
> economics$unemp_rate
 [1] 0.066 0.060 0.058 0.058 0.063 0.070 0.068 0.063 0.063 0.058 0.055 0.053 0.058 0.054 0.052 0.049 0.053 0.058 0.059 0.055
[21] 0.054 0.048 0.047 0.044 0.050 0.046 0.043 0.041 0.046 0.050 0.049 0.044 0.044 0.040 0.039 0.037 0.041 0.038 0.037 0.034
[41] 0.035 0.040 0.038 0.033 0.032 0.027 0.026 0.025 0.034 0.034 0.032 0.034 0.039 0.043 0.042 0.036 0.034 0.031 0.030 0.027
[61] 0.033 0.032 0.031 0.028 0.031 0.039 0.038 0.034 0.030 0.027 0.024 0.023 0.030 0.025 0.025 0.022 0.029 0.031 0.028 0.027
[81] 0.028 0.023 0.021 0.020
> economics$labor_force_part_rate
 [1] 0.631 0.629 0.631 0.632 0.633 0.645 0.645 0.643 0.635 0.637 0.641 0.637 0.627 0.628 0.631 0.629 0.631 0.644 0.647 0.646
[21] 0.637 0.642 0.644 0.640 0.632 0.633 0.633 0.633 0.634 0.644 0.645 0.643 0.634 0.639 0.642 0.639 0.632 0.633 0.633 0.633
[41] 0.634 0.644 0.645 0.643 0.634 0.639 0.642 0.639 0.639 0.645 0.652 0.654 0.654 0.662 0.663 0.657 0.648 0.645 0.645 0.647
[61] 0.626 0.634 0.639 0.645 0.651 0.663 0.670 0.665 0.656 0.659 0.658 0.664 0.667 0.667 0.668 0.665 0.668 0.675 0.674 0.676
[81] 0.665 0.671 0.670 0.670
```

b) 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



```
> economics$diff_unemp_labor
 [1] 0.565 0.569 0.573 0.574 0.570 0.575 0.577 0.580 0.572 0.579 0.586 0.584 0.569 0.574 0.579 0.580 0.578 0.586 0.588 0.591
[21] 0.583 0.594 0.597 0.596 0.582 0.587 0.590 0.592 0.588 0.594 0.596 0.599 0.590 0.599 0.603 0.602 0.591 0.595 0.596 0.599
[41] 0.599 0.604 0.607 0.610 0.602 0.612 0.616 0.614 0.605 0.611 0.620 0.620 0.615 0.619 0.621 0.621 0.614 0.614 0.615 0.620
[61] 0.593 0.602 0.608 0.617 0.620 0.624 0.632 0.631 0.626 0.632 0.634 0.641 0.637 0.642 0.643 0.643 0.639 0.644 0.646 0.649
[81] 0.637 0.648 0.649 0.650
```

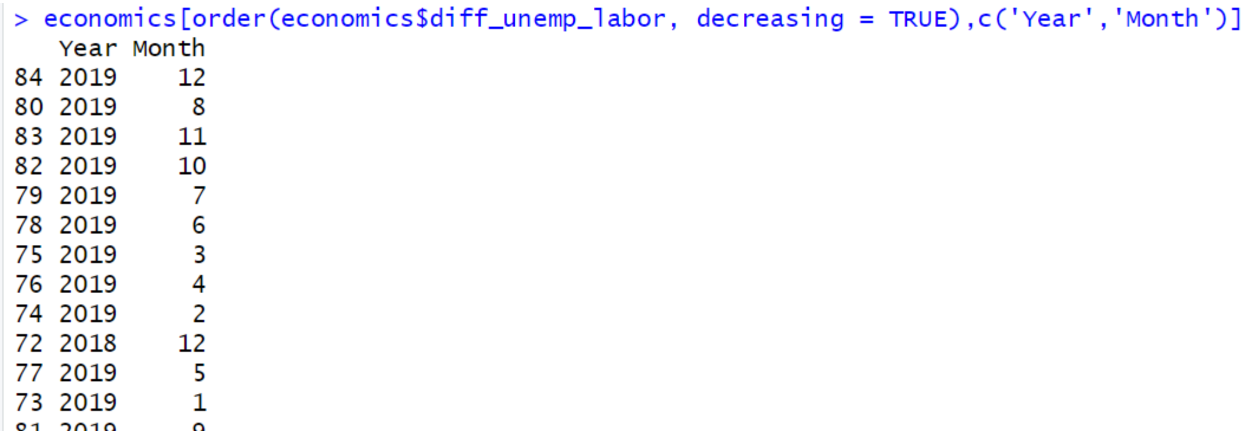
c) Apply the statement

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).

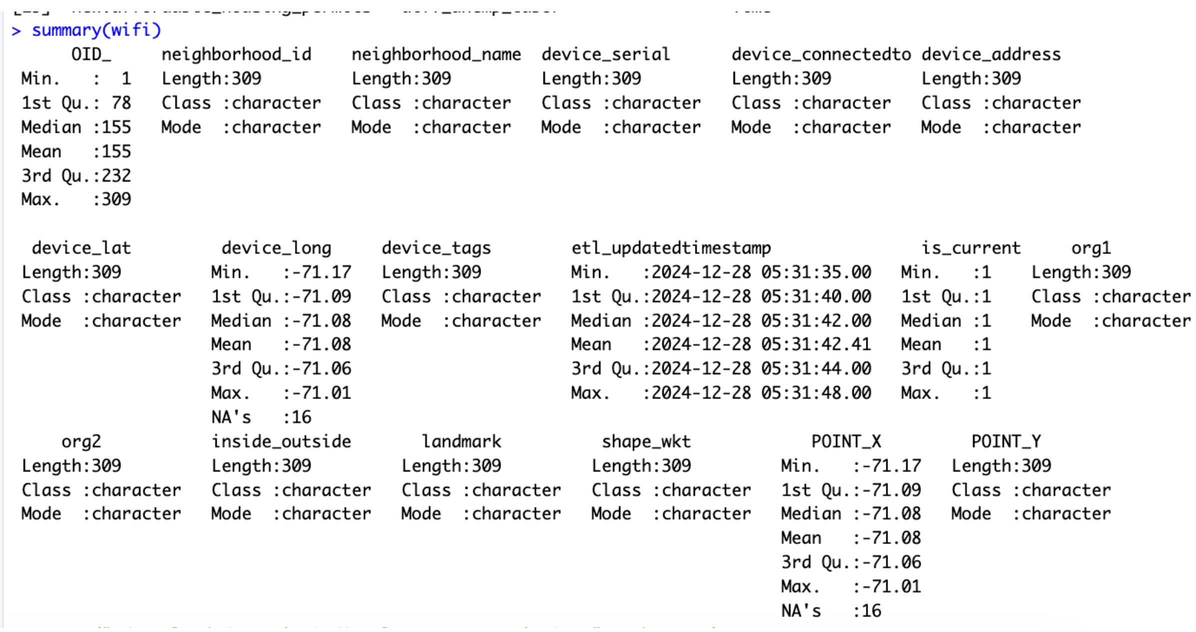
After sorting, it returns a subset of the data frame that includes only the Year and Month columns, arranged according to the sorted order of diff_unemp_labor. Based on the result, December 2019 has the highest diff_unemp_labor value within the time period.



```
> economics[order(economics$diff_unemp_labor, decreasing = TRUE),c('Year','Month')]
   Year Month
84 2019    12
80 2019     8
83 2019    11
82 2019    10
79 2019     7
78 2019     6
75 2019     3
76 2019     4
74 2019     2
72 2018    12
77 2019     5
73 2019     1
81 2019     9
```

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

summary(wifi)

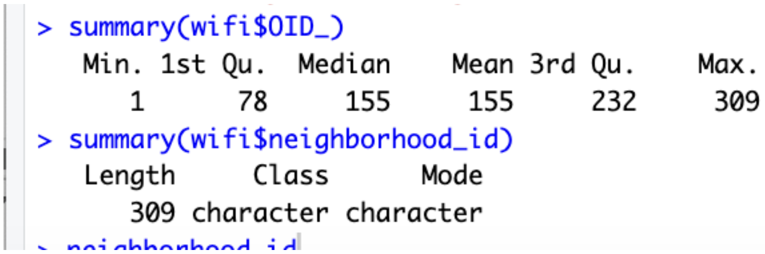


```
> summary(wifi)
      OID_        neighborhood_id    neighborhood_name  device_serial      device_connectedto device_address    
 Min.   :  1   Length:309         Length:309         Length:309         Length:309         Length:309        
 1st Qu.: 78   Class :character   Class :character   Class :character   Class :character   Class :character  
 Median :155   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :155                                                                                                 
 3rd Qu.:232                                                                                                 
 Max.   :309                                                                                                 
                                                                                                             
  device_lat         device_long      device_tags        etl_updatedtimestamp            is_current     org1          
 Length:309         Min.   :-71.17   Length:309         Min.   :2024-12-28 05:31:35.00   Min.   :1   Length:309        
 Class :character   1st Qu.:-71.09   Class :character   1st Qu.:2024-12-28 05:31:40.00   1st Qu.:1   Class :character  
 Mode  :character   Median :-71.08   Mode  :character   Median :2024-12-28 05:31:42.00   Median :1   Mode  :character  
                    Mean   :-71.08                      Mean   :2024-12-28 05:31:42.41   Mean   :1                     
                    3rd Qu.:-71.06                      3rd Qu.:2024-12-28 05:31:44.00   3rd Qu.:1                     
                    Max.   :-71.01                      Max.   :2024-12-28 05:31:48.00   Max.   :1                     
                    NA's   :16                                                                                         
     org2           inside_outside       landmark          shape_wkt            POINT_X          POINT_Y         
 Length:309         Length:309         Length:309         Length:309         Min.   :-71.17   Length:309        
 Class :character   Class :character   Class :character   Class :character   1st Qu.:-71.09   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :-71.08   Mode  :character  
                                                                             Mean   :-71.08                     
                                                                             3rd Qu.:-71.06                     
                                                                             Max.   :-71.01                     
                                                                             NA's   :16                         
```

e) Describe the summary information for **OID_** and **neightborhood_id**, and explain why they are different?

summary(wifi$OID_)

summary(wifi$neighborhood_id)



```
> summary(wifi$OID_)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      1      78     155     155     232     309 
> summary(wifi$neighborhood_id)
   Length     Class      Mode 
      309 character character 
```

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

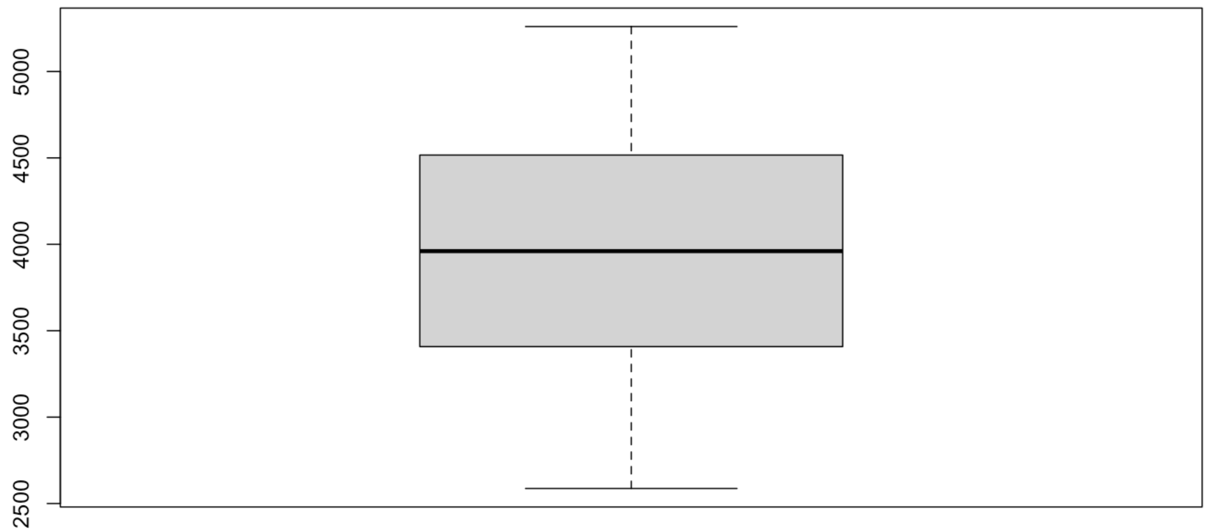
Boxplot analysis

a) 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)

b) Apply below statement:

boxplot(logan_intl_flights ~ Month, data = economics)

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



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

c) Apply below statements:

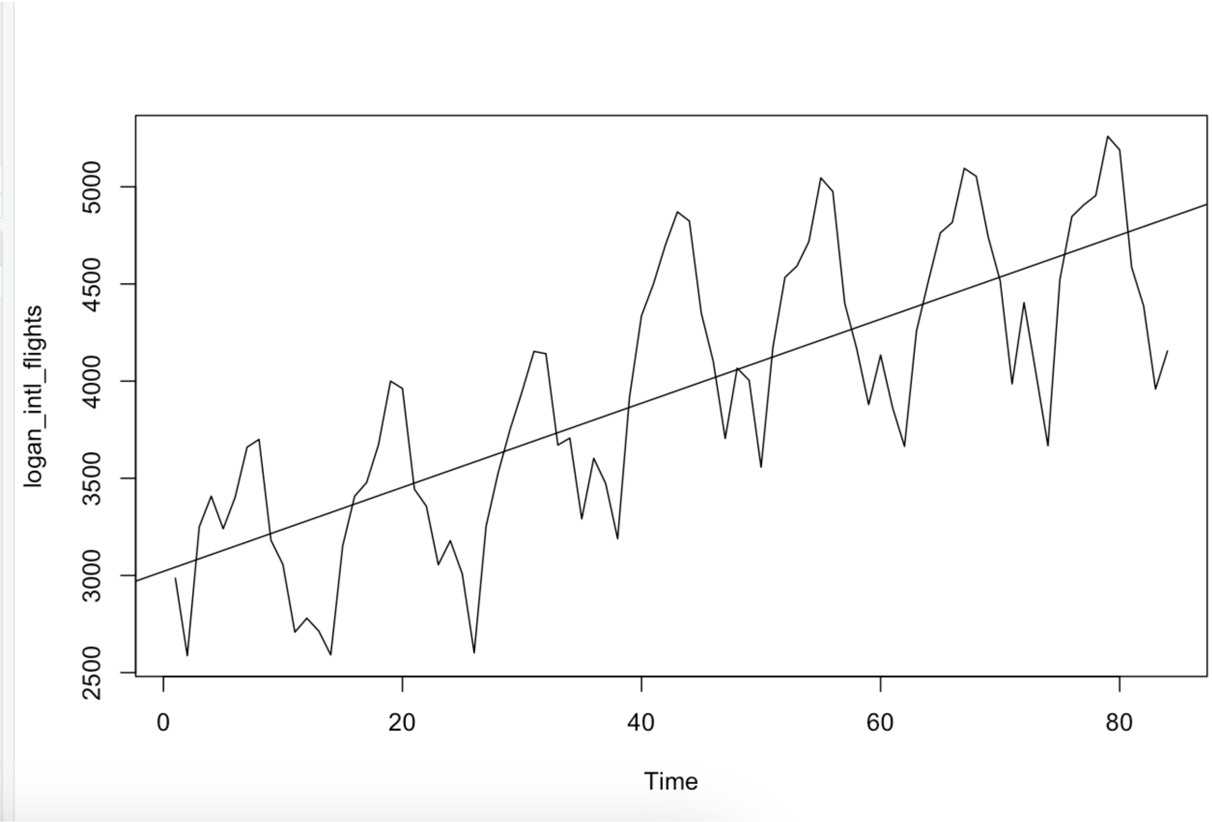
```
plot(logan_intl_flights~Time, data = economics, type = 'l')
abline(lm(logan_intl_flights~Time, data=economics))
```

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


logan_intl_flights
Time