Assignment01

2024-09-22

Assignment 01

Name: Jose E. Encarnacion Satana

Student No: 8982860

Big Data Solution Architecture, Conestoga College

Data Analysis Mathematics, Algorithms and Modeling

PROG8435 - Fall 2024 - Section 1

Yun Qian Miao, PhD

Setting Enviroment

```
knitr::opts_chunk$set(echo = TRUE)
# init env
if(!is.null(dev.list())) dev.off()
```

```
## null device
##           1
```

```
cat("\014")
```

```
rm(list = ls())
options(scipen = 9)

# setting the current directory
setwd('C:/Users/RXPCOMPUTER/Source/RProjects/dataanalysisr')
```

Assignment Tasks

Task 1:

You are working for a human resources consulting firm. The following statement is made by your manager.

Wages are going up but productivity is going down.

Use MS CoPilot to try to transform this in to a question that can be answered with data analytics. Include the prompt and the response in your answer. Now, based on the examples and discussion in Week 1, transform it in to a question that can be answered with data analytics. Make sure you discuss the logic and reasoning you use to transform it and what questions you might ask.

Discuss how your answer is different from the Gen AI answer and why you think your answer is better.

Prompt: "Rewrite the statement 'Wages are going up but productivity is going down' into a question that can be answered through data analytics."

MS CoPilot Response: "How are changes in wages over time correlated with changes in productivity, and what factors are influencing the observed trends?"

Answer: What variables are impacting the observed patterns, and how are changes in productivity and pay associated over time?

The difference between the answer IA and the answer generated by my logic and reason is that IA uses an algorithm for appearance by a human. The answer IA is too structured and sound robotic. Humans write in a form natural and make mistakes in the spelling word. I think that my question is more natural. I used the method of paraphraser to transform the statement into a question that can be understood by a manager in human resources.

Task 2:

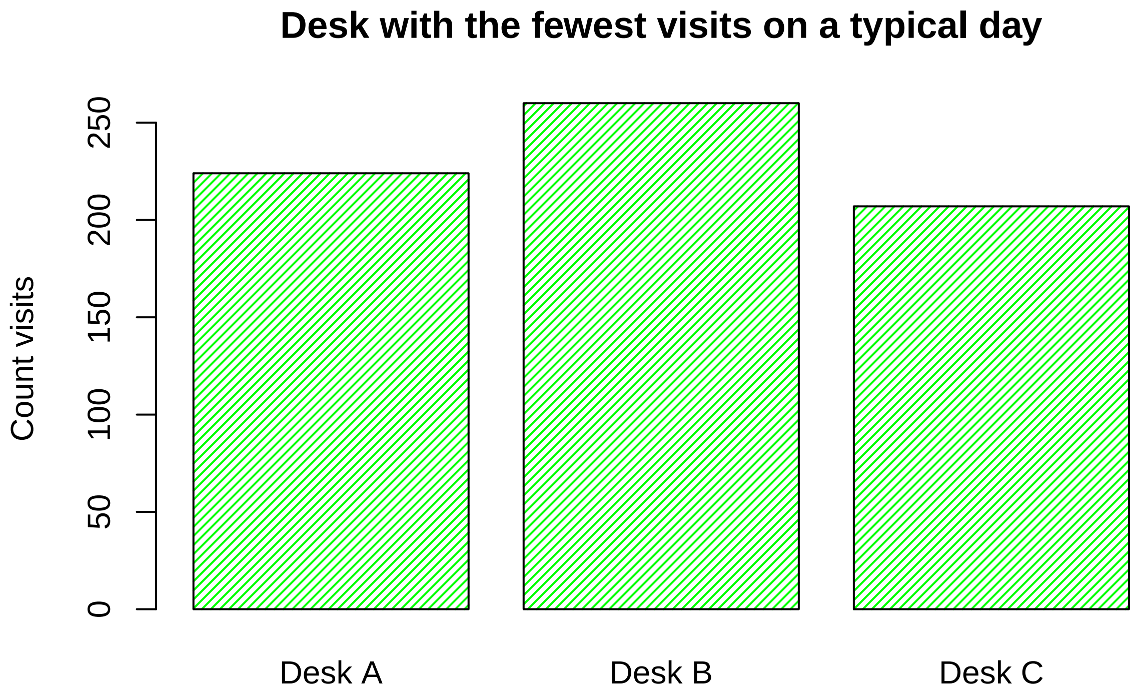
Consider the following three arrays of data. Each array is data for one in- person help desk. The numbers in the array represent the number of unique

visitors to each help desk in a day (for example, Desk A had 250 visitors on the first day, 255 on the second and so on). Desk A: (250 255 230 257 237 224 232 240 229 242 246) Desk B: (264 273 265 269 269 270 271 260 268 275 276) Desk C: (255 257 250 229 255 261 272 237 207 233 243) Based on the data provided, and using the skills learned in this class, answer the following questions. Make sure to provide evidence for your answers. a) Which Desk has the fewest visits on a typical day? b) Which Desk has the most consistent usage?

a) Which Desk has the fewest visits on a typical day?

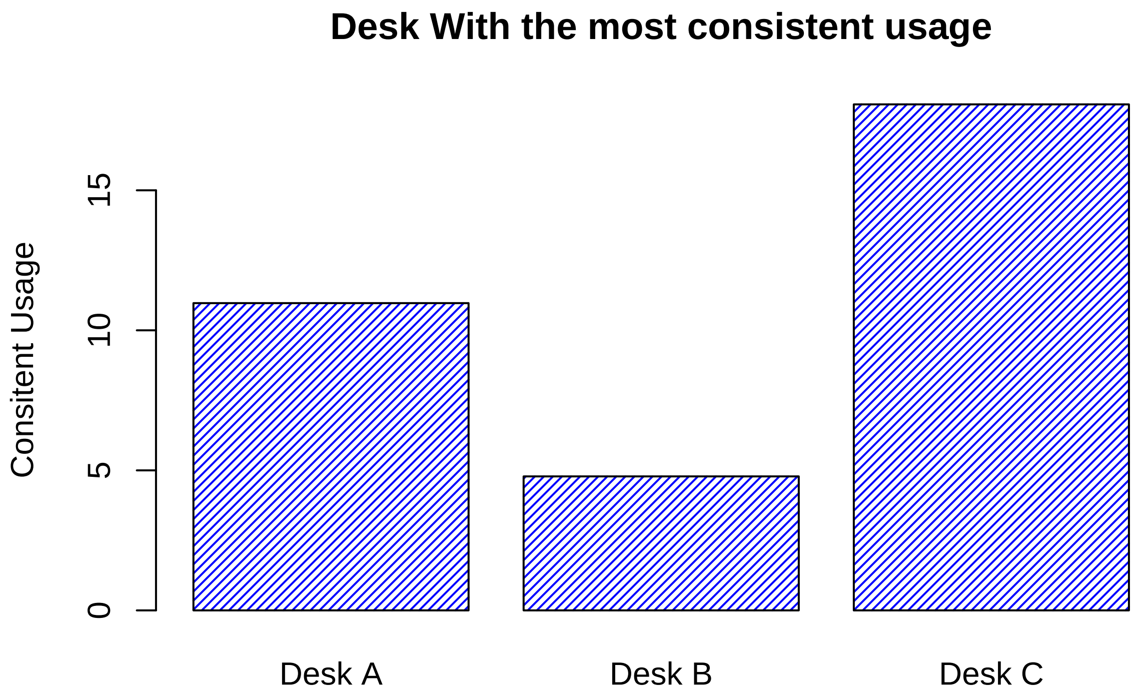
```
# Create arrays for the three help desks
desk_a_je <- c(250, 255, 230, 257, 237, 224, 232, 240, 229, 242, 246)
desk_b_je <- c(264, 273, 265, 269, 269, 270, 271, 260, 268, 275, 276)
desk_c_je <- c(255, 257, 250, 229, 255, 261, 272, 237, 207, 233, 243)
```

```r
list_je <- list(desk_A= desk_a_je, desk_B = desk_b_je, desk_C = desk_c_je)

df_je = as.data.frame(list_je)

# summary(df_je)

# a) Which Desk has the fewest visits on a typical day?
#
#     We calculate the mean (average) number of visits for each desk.
#     The desk with the lowest mean is the one that typically has the fewest visits.

min_desk_a_je <- min(df_je$desk_A)
min_desk_b_je <- min(df_je$desk_B)
min_desk_c_je <- min(df_je$desk_C)

min_desk_day <- c(min_desk_a_je, min_desk_b_je, min_desk_c_je)

barplot(min_desk_day,
        main = "Desk with the fewest visits on a typical day",
        ylab = "Count visits",
        names.arg = c("Desk A","Desk B","Desk C"),
        col="green", density = 30, angle = 45)
```



b) Which Desk has the most consistent usage?

```r
# b) Which Desk has the most consistent usage?
#
#     We calculate the standard deviation of the visits for each desk.
#     The desk with the lowest standard deviation is the most consistent
#     (has the least variation day-to-day).

std_desk_a_je <- sd(df_je$desk_A)
std_desk_b_je <- sd(df_je$desk_B)
std_desk_c_je <- sd(df_je$desk_C)

std_desk_usage <- c(std_desk_a_je, std_desk_b_je, std_desk_c_je)

bar_usage_je <- barplot(std_desk_usage,
        main = "Desk With the most consistent usage",
        ylab = "Consitent Usage",
        names.arg = c("Desk A","Desk B","Desk C"),
        col="blue", density = 30, angle = 45)
```



Task 2

PART 2 : Every question in Part 2 should be answered using the dataset provided. The dataset is a subset of employment data gathered and published by Statistics Canada. The following tasks will seek to describe and explore some of the employment data which has been gathered by Statistics Canada.

Basic Manipulation

1. Read in the text file and change to a data frame

```
# 1. Read in the text file and change to a data frame

df_st_canada_je <- read.table('PROG8435-24F-A1_data.txt', sep = ',', header = TRUE);
head(df_st_canada_je)
```

```
##                     Province Year UnEmp Part  Emp
## 1 Newfoundland and Labrador 1976  13.4 49.4 42.8
## 2 Newfoundland and Labrador 1977  15.4 50.6 42.8
## 3 Newfoundland and Labrador 1978  15.9 51.7 43.5
## 4 Newfoundland and Labrador 1979  14.8 53.4 45.5
## 5 Newfoundland and Labrador 1980  13.3 53.2 46.2
## 6 Newfoundland and Labrador 1981  13.5 53.5 46.3
```

2. Append your initials to all variables in the data frame (Note – you will need to do this in all your subsequent assignments).

```
# 2. Append your initials to all variables in the data frame (Note - you will
# need to do this in all your subsequent assignments).
colnames(df_st_canada_je) <- paste(colnames(df_st_canada_je), "je", sep="_")
head(df_st_canada_je)
```

```
##                 Province_je Year_je UnEmp_je Part_je Emp_je
## 1 Newfoundland and Labrador    1976     13.4    49.4   42.8
## 2 Newfoundland and Labrador    1977     15.4    50.6   42.8
## 3 Newfoundland and Labrador    1978     15.9    51.7   43.5
## 4 Newfoundland and Labrador    1979     14.8    53.4   45.5
## 5 Newfoundland and Labrador    1980     13.3    53.2   46.2
## 6 Newfoundland and Labrador    1981     13.5    53.5   46.3
```

3. Change each character variable to a factor variable

```
# 3. Change each character variable to a factor variable

df_st_canada_je$Province_je <- as.factor(df_st_canada_je$Province_je)
str(df_st_canada_je$Province_je)
```

```
##  Factor w/ 10 levels "Alberta","British Columbia",..: 5 5 5 5 5 5 5 5 5 5 ...
```

4. Create a new variable showing the level of unemployment. This new variable should be discrete with three levels: "L" if Unemployment $<$ 6.5 "M" if Unemployment between 6.5 and 11.5 (inclusive) "H" if Unemployment $>$ 11.5

```
# 4. Create a new variable showing the level of unemployment. This new
# variable should be discrete with three levels:
# "L" if Unemployment < 6.5
# "M" if Unemployment between 6.5 and 11.5 (inclusive)
# "H" if Unemployment > 11.5
```

```r
UnEmp_lvl_je <- cut(df_st_canada_je$UnEmp_je,
                                breaks = c(-Inf, 6.5, 11.5, Inf),
                                labels = c("L", "M", "H"))
table(UnEmp_lvl_je)
```

```
## UnEmp_lvl_je
##   L   M   H
## 126 237 117
```

5. What are the dimensions of the dataset (rows and columns)?

```r
# 5. What are the dimensions of the dataset (rows and columns)?

dm_je <- dim(df_st_canada_je)
dm_je
```

```
## [1] 480   5
```

Summarizing Data

1. Means and Standard Deviations
2. Calculate the mean and standard deviation for Unemployment.

```r
mean_unemp_je <- mean(df_st_canada_je$UnEmp_je)
mean_unemp_je
```

```
## [1] 9.220417
```

```r
st_unemp_je <- sd(df_st_canada_je$UnEmp_je)
st_unemp_je
```

```
## [1] 3.553075
```

b. Use the results above to calculate the coefficient of variation (rounded to 3 decimal places).

```r
cv_unemp_je <-(st_unemp_je / mean_unemp_je) * 100
cv_unemp_je <- round(cv_unemp_je, 3)
print(cv_unemp_je)
```

```
## [1] 38.535
```

c. Calculate the mean and standard deviation for Participation Rate.

```r
mean_part_je <- mean(df_st_canada_je$Part_je)
mean_part_je
```

```
## [1] 64.42875
```

```
sd_part_je <- sd(df_st_canada_je$Part_je)
sd_part_je
```

```
## [1] 4.790083
```

d. Also calculate the coefficient of variation (rounded to 3 decimal places).

```
cv_part_je <- (sd_part_je/mean_part_je) * 100
cv_part_je <- round(cv_part_je, 3)
cv_part_je
```

```
## [1] 7.435
```

e. Does the Unemployment or Participation have more variation?

```
vr_result_je <- ""
if(cv_unemp_je > cv_part_je){
   vr_result_je <- "Unemployment has more variation."
}else if (cv_unemp_je < cv_part_je){
   vr_result_je <- "Participation Rate has more variation."
}else{
  vr_result_je <- "Both have the same variation."
}
vr_result_je
```

```
## [1] "Unemployment has more variation."
```

2. Calculate the 74th percentile of the number of Employment Rate. This calculation should be rounded to the nearest whole number (no decimal places).

```
per_unemp_74_je <- quantile(df_st_canada_je$Emp_je, 0.74, na.rm = TRUE)
per_unemp_74_je <- round(per_unemp_74_je)
per_unemp_74_je
```

```
## 74%
##  62
```

Organizing Data

1. Summary Table

a. Create a table showing the average unemployment rate by province. This should be rounded to two decimal places.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
avg_unemp_by_prov_je <- df_st_canada_je %>%
  group_by(df_st_canada_je$Province_je) %>%
  summarise(avg_unemp_by_prov_je = round(mean(df_st_canada_je$UnEmp_je, na.rm=TRUE), 2))
avg_unemp_by_prov_je
```

```
## # A tibble: 10 x 2
##    `df_st_canada_je$Province_je` avg_unemp_by_prov_je
##    <fct>                                        <dbl>
##  1 Alberta                                       9.22
##  2 British Columbia                              9.22
##  3 Manitoba                                      9.22
##  4 New Brunswick                                 9.22
##  5 Newfoundland and Labrador                     9.22
##  6 Nova Scotia                                   9.22
##  7 Ontario                                       9.22
##  8 Prince Edward Island                          9.22
##  9 Quebec                                        9.22
## 10 Saskatchewan                                  9.22
```

b. Which province has, on average, the highest unemployment rate?

```
prov_hight_unemp_je <- avg_unemp_by_prov_je %>%
  filter(avg_unemp_by_prov_je == max(avg_unemp_by_prov_je))
prov_hight_unemp_je
```

```
## # A tibble: 10 x 2
##    `df_st_canada_je$Province_je` avg_unemp_by_prov_je
##    <fct>                                        <dbl>
##  1 Alberta                                       9.22
##  2 British Columbia                              9.22
##  3 Manitoba                                      9.22
##  4 New Brunswick                                 9.22
##  5 Newfoundland and Labrador                     9.22
##  6 Nova Scotia                                   9.22
##  7 Ontario                                       9.22
##  8 Prince Edward Island                          9.22
##  9 Quebec                                        9.22
## 10 Saskatchewan                                  9.22
```

2. Cross Tabulation

a. Create a table counting all levels of unemployment (the variable you created in Part 2: Q1.4) by province.

```r
unemp_level_by_provice_je <- table(df_st_canada_je$Province_je,UnEmp_lvl_je)
unemp_level_by_provice_je
```

```
##                            UnEmp_lvl_je
##                              L  M  H
##   Alberta                   26 22  0
##   British Columbia          12 30  6
##   Manitoba                  32 16  0
##   New Brunswick              0 29 19
##   Newfoundland and Labrador  0  2 46
##   Nova Scotia                2 33 13
##   Ontario                   16 32  0
##   Prince Edward Island       0 24 24
##   Quebec                     6 33  9
##   Saskatchewan              32 16  0
```

b. Change the table to show the percentage of each Unemployment level in each Province. This should be rounded to three decimal places.

```r
unemp_per_by_provice_je <- prop.table(unemp_level_by_provice_je, 1) * 100
unemp_per_by_provice_round_je <- round(unemp_per_by_provice_je,3)

unemp_per_by_provice_round_je
```

```
##                            UnEmp_lvl_je
##                                  L      M      H
##   Alberta                   54.167 45.833  0.000
##   British Columbia          25.000 62.500 12.500
##   Manitoba                  66.667 33.333  0.000
##   New Brunswick              0.000 60.417 39.583
##   Newfoundland and Labrador  0.000  4.167 95.833
##   Nova Scotia                4.167 68.750 27.083
##   Ontario                   33.333 66.667  0.000
##   Prince Edward Island       0.000 50.000 50.000
##   Quebec                    12.500 68.750 18.750
##   Saskatchewan              66.667 33.333  0.000
```

c. What percentage of high unemployment levels were in Prince Edward Island?

```r
pei_high_unemp_percentage_je <- unemp_per_by_provice_round_je["Prince Edward Island", "H"]
pei_high_unemp_percentage_je
```

```
## [1] 50
```

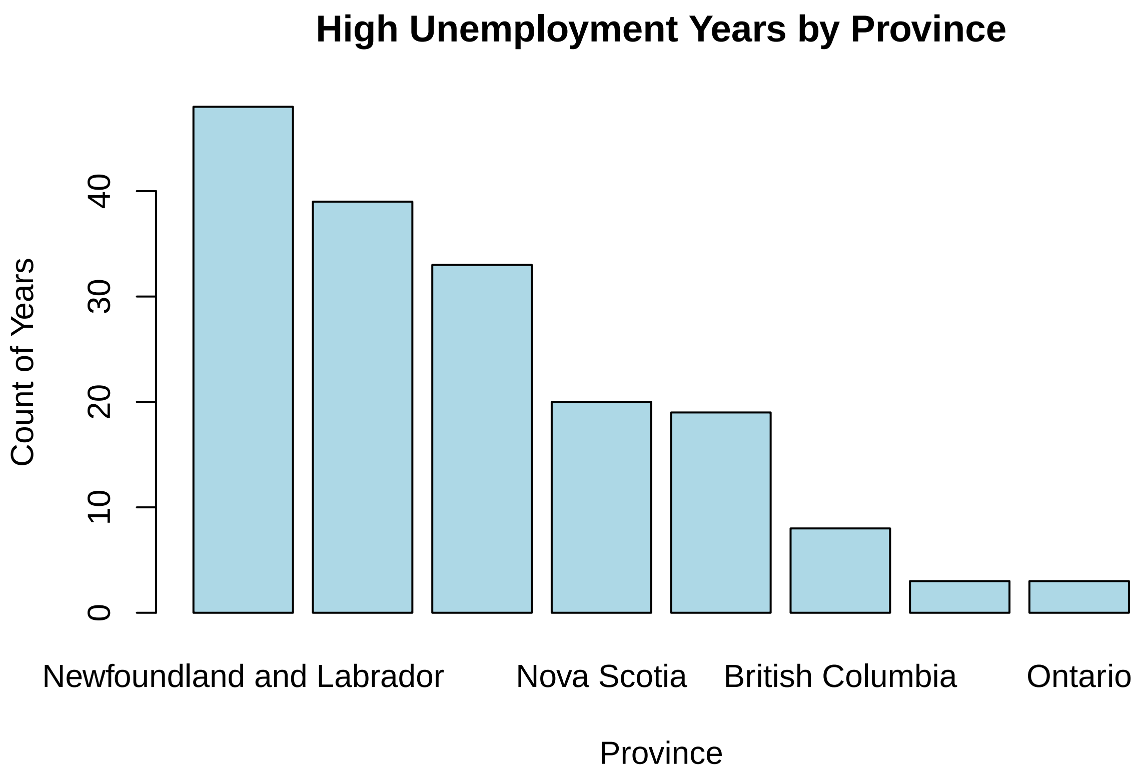
3. Bar Plot

a. Create a column plot of years of high unemployment in each province.
b. The plot should be:
c. Rank ordered by highest count of high unemployment.

ii. Properly labeled (title, x-axis, etc)

iii. The bars should have a different colour than the oneshown in class.

```
hight_unemp_je <- 10

hight_unemp_year_je <- df_st_canada_je %>%
  filter(df_st_canada_je$UnEmp_je >= hight_unemp_je)
# hight_unemp_year_je

hight_unemp_year_count_je <- hight_unemp_year_je %>%
  group_by(hight_unemp_year_je$Province_je) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))


# Bar plot
barplot(hight_unemp_year_count_je$Count,
        names.arg = hight_unemp_year_count_je$`hight_unemp_year_je$Province_je`,
        col = "lightblue", main = "High Unemployment Years by Province",
        xlab = "Province", ylab = "Count of Years")
```

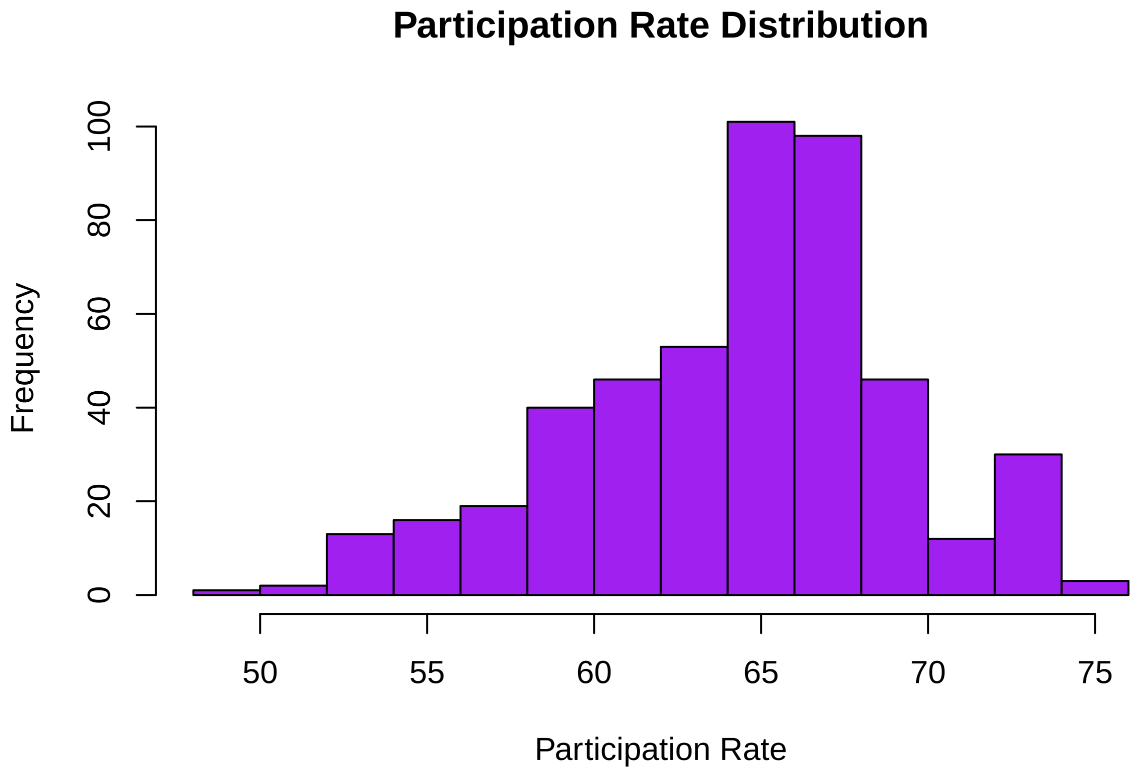


c. Based on the bar plot, (approximately) how many of years did Nova Scotia experience high unemployment?

4. Histogram

a. Create a histogram of Participation Rate.
b. The plot should be properly labeled and a unique colour and have 10 breaks.
c. Which range of Participation Rate is the most common?

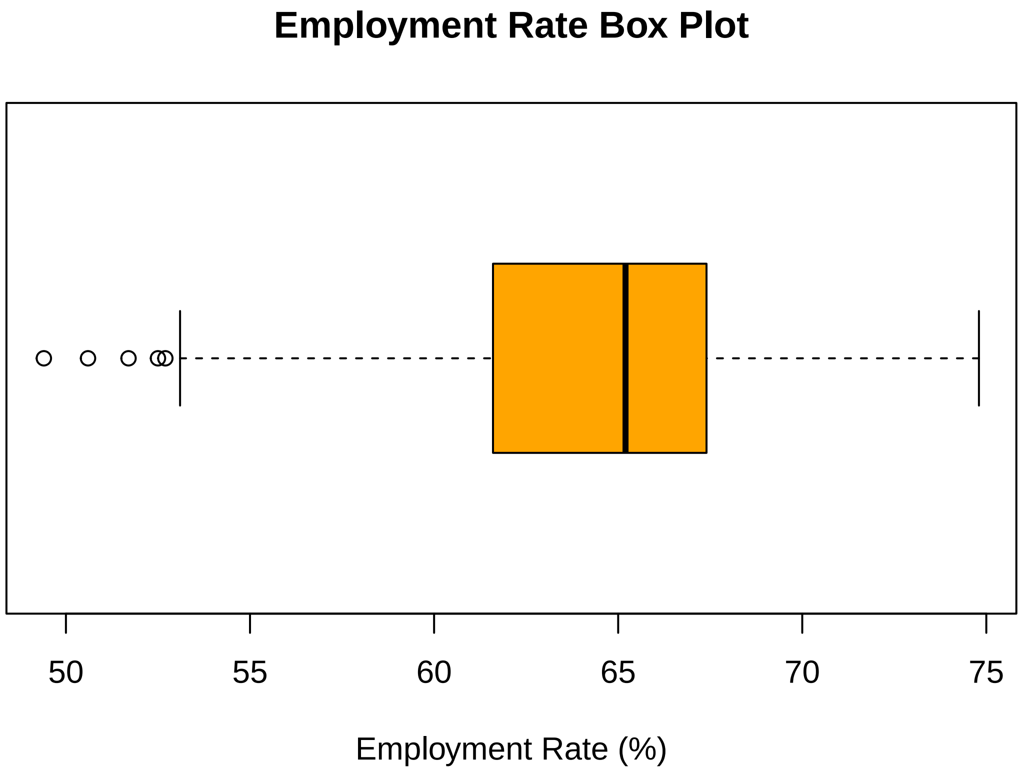
```
hist(df_st_canada_je$Part_je,
     breaks = 10,
     col = "purple",
     main = "Participation Rate Distribution",
     xlab = "Participation Rate",
     ylab = "Frequency")
```



5. Box plot

a. Create a horizontal box plot of number of Employment Rate.
b. The plot should be properly labeled and a unique colour.
c. Based on the box plot, approximately how many years had an Employment Rate less than ~ 60%?

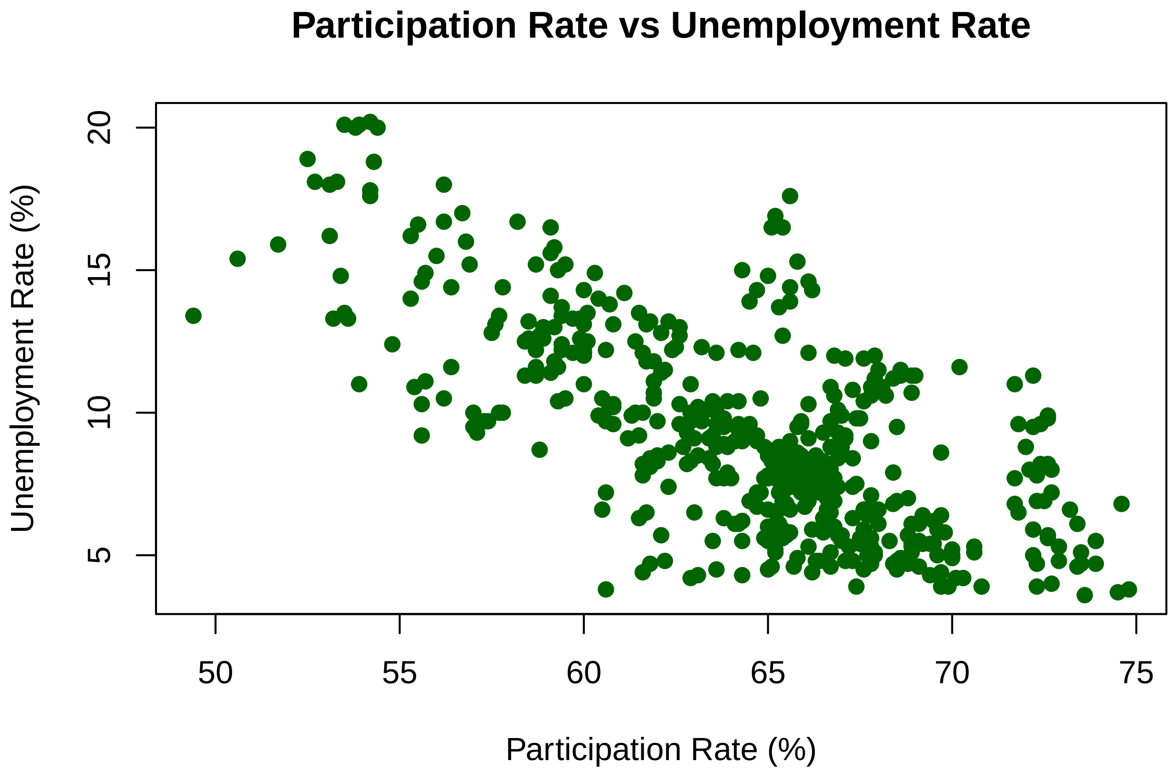
```
boxplot(df_st_canada_je$Part_je,
        horizontal = TRUE,
        col = "orange",
        main = "Employment Rate Box Plot",
        xlab = "Employment Rate (%)")
```



6. Scatter Plot

a. Create a scatter plot comparing Participation Rate and Unemployment Rate.
b. The plot should be properly labeled with a marker type different than the one demonstrated in class.
c. Does there appear to be an association between Participation Rate and Unemployment Rate?

```
plot(df_st_canada_je$Part_je, df_st_canada_je$UnEmp_je,
     main = "Participation Rate vs Unemployment Rate",
     xlab = "Participation Rate (%)",
     ylab = "Unemployment Rate (%)",
     pch = 19,  # Change marker type
     col = "darkgreen")
```



Appendix one: Study file data

Each row of the dataset represents information for a particular year and province. Variable Description Province Name of Canadian province Year Full Calendar Year

UnEmp

The unemployment rate is the number of unemployed persons expressed as a percentage of the labour force. The unemployment rate for a particular group (age, sex, marital status, etc.) is the number unemployed in that group expressed as a percentage of the labour force for that group. Estimates are percentages, rounded to the nearest tenth.

Part

The participation rate is the number of labour force participants expressed as a percentage of the population 15 years of age and over. The participation rate for a particular group (age, sex, marital status, etc.) is the number of labour force participants in that group expressed as a percentage of the population for that group. Estimates are percentages, rounded to the nearest tenth.

Emp

The employment rate is the number of persons employed expressed as a percentage of the population 15 years of age and over. The employment rate for a particular group (age, sex, marital status, etc.) is the number employed in that group expressed as a percentage of the population for that group. Estimates are percentages, rounded to the nearest tenth.

Source: Statistics Canada. Table 14-10-0327-02 Unemployment rate, participation rate and employment rate by sex, annual https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410032702