Homework 3

The COVID-19 pandemic disrupted human society, and governments worldwide implemented diverse policies to curb its spread, including non-pharmaceutical interventions like Japan's early school closures. This assignment draws from a study focused on quantifying the effects of such policies on reducing infection rates.

This exercise is based on the following study:

Fukumoto, K., McClean, C.T. & Nakagawa, K. No causal effect of school closures in Japan on the spread of COVID-19 in spring 2020. *Nature Medicine 27*, 2111–2119 (2021)

To address the potential confounding bias caused by the absence of randomized treatment assignment, the authors of this observational study used a statistical method called "matching." Matching methods utilize an algorithm to match each treated unit with a control unit that has the most similar pre-treatment characteristics. Then, one computes the causal effects using matched units alone while dropping the unmatched observations from the data set. The question of how exactly matching can be done is beyond the scope of this exercise. Instead, we take the matched control units as given and analyze the resulting data set.

This exercise focuses on the school closure of April 6 as the only treatment variable. If this treatment variable equals 1, it means that in a given municipality all elementary and junior high schools are closed as of the survey date, and 0 if they are open.

The names and descriptions of variables in both data sets are:

Variable	Description
`municipality_code`	Municipality
`labor`	Labor Force (2015)
`elder.pop`	Elderly Population (65 and older) (2015)
`hospitals`	Density Index of Hospitals (2017)
`elementary`	Population Density Index of Elementary School Students (2018)
`junior`	Population Density Index of Junior High School Students (2018)
`prec_mean`	Average Precipitation (average between 1981–2010)
`log.number`	Number of Bordering Municipalities (2020)
`age.0406`	Mayor's Age (as of 2020.04.06)
`shutdown.0406`	Treatment Variable (treatment status on the 6th April)
`X2020.X.XX`	Number of Infections per 100,000 Municipal Residents on Day XX Month X Year 2020

Loading the data

```
covid <- read.csv("covid.csv")
```

Round all answers to two decimal places

Question 1

1.1. Compute the differences between the means of the treatment and control groups for each of the the following variables: `labor`, `elder.pop`, and `hospitals`. In all cases subtract the mean of the control group from the mean of the treatment group.

```
# Difference in means for 'labor'
mean_labor_treatment <- mean(covid$labor[covid$shutdown.0406 ==
    1], na.rm = TRUE)
mean_labor_control <- mean(covid$labor[covid$shutdown.0406 ==
    0], na.rm = TRUE)
diff_labor <- mean_labor_treatment - mean_labor_control
round(diff_labor, digits = 2)
```

```
## [1] 0
```

```
# Difference in means for 'elder.pop'
mean_elder_pop_treatment <- mean(covid$elder.pop[covid$shutdown.0406 ==
    1], na.rm = TRUE)
mean_elder_pop_control <- mean(covid$elder.pop[covid$shutdown.0406 ==
    0], na.rm = TRUE)
diff_elder_pop <- mean_elder_pop_treatment - mean_elder_pop_control
round(diff_elder_pop, digits = 2)
```

```
## [1] -0.05
```

```
# Difference in means for 'hospitals'
mean_hospitals_treatment <- mean(covid$hospitals[covid$shutdown.0406 ==
    1], na.rm = TRUE)
mean_hospitals_control <- mean(covid$hospitals[covid$shutdown.0406 ==
    0], na.rm = TRUE)
diff_hospitals <- mean_hospitals_treatment - mean_hospitals_control
round(diff_hospitals, digits = 2)
```

```
## [1] 0
```

Answer (labor): 0
Answer (elder.pop): -0.05
Answer (hospitals): 0

1.2. Compute the differences between the means of the treatment and control groups for `elementary`, `junior`, `prec_mean`, `log.number` and `age.0406`.

```
# Difference in means for 'elementary'
mean_elementary_treatment <- mean(covid$elementary[covid$shutdown.0406 ==
    1], na.rm = TRUE)
mean_elementary_control <- mean(covid$elementary[covid$shutdown.0406 ==
    0], na.rm = TRUE)
diff_elementary <- mean_elementary_treatment - mean_elementary_control
round(diff_elementary, digits = 2)
```

```
## [1] 0
```

```r
# Difference in means for 'junior'
mean_junior_treatment <- mean(covid$junior[covid$shutdown.0406 ==
    1], na.rm = TRUE)
mean_junior_control <- mean(covid$junior[covid$shutdown.0406 ==
    0], na.rm = TRUE)
diff_junior <- mean_junior_treatment - mean_junior_control
round(diff_junior, digits = 2)
```

```
## [1] 0
```

```r
# Difference in means for 'prec_mean'
mean_prec_mean_treatment <- mean(covid$prec_mean[covid$shutdown.0406 ==
    1], na.rm = TRUE)
mean_prec_mean_control <- mean(covid$prec_mean[covid$shutdown.0406 ==
    0], na.rm = TRUE)
diff_prec_mean <- mean_prec_mean_treatment - mean_prec_mean_control
round(diff_prec_mean, digits = 2)
```

```
## [1] -189.61
```

```r
# Difference in means for 'log.number'
mean_log_number_treatment <- mean(covid$log.number[covid$shutdown.0406 ==
    1], na.rm = TRUE)
mean_log_number_control <- mean(covid$log.number[covid$shutdown.0406 ==
    0], na.rm = TRUE)
diff_log_number <- mean_log_number_treatment - mean_log_number_control
round(diff_log_number, digits = 2)
```

```
## [1] 0.14
```

```r
# Difference in means for 'age.0406'
mean_age_0406_treatment <- mean(covid$age.0406[covid$shutdown.0406 ==
    1], na.rm = TRUE)
mean_age_0406_control <- mean(covid$age.0406[covid$shutdown.0406 ==
    0], na.rm = TRUE)
diff_age_0406 <- mean_age_0406_treatment - mean_age_0406_control
round(diff_age_0406, digits = 2)
```

```
## [1] -1.91
```

Answer (elementary): 0
Answer (junior): 0
Answer (prec_mean): -189.61
Answer (log.number): .14
Answer (age.0406): -1.91

1.3. Interpret the results for `hospitals`, `elementary` and `junior`. Are there any consistent patterns of differences between the two groups of municipalities?

Answer: The mean differences for all three variables are essentially zero. This suggests that there are no consistent patterns of differences between the two groups of municipalities for these variables.

1.4. Create a new variable called `students` that is the sum of `elementary` and `junior`. Compute the differences between the means of the treatment and control groups for `students`.

```r
# Creating a new variable 'students' as the sum of
# 'elementary' and 'junior'
covid$students <- covid$elementary + covid$junior

# Compute differences between the means of the treatment
# (shutdown.0406) and control groups for 'students'
mean_students_treatment <- mean(covid$students[covid$shutdown.0406 ==
    1], na.rm = TRUE)
mean_students_control <- mean(covid$students[covid$shutdown.0406 ==
    0], na.rm = TRUE)
diff_students <- mean_students_treatment - mean_students_control

round(diff_students, digits = 2)
```

```
## [1] 0.01
```

Answer: 0.01

Question 2

2.1. Compute a single variable `total_infections_april` that records the number of COVID-19 infections per 100,000 residents for April 1, 2020 to April 30, 2020. To do this you will need to sum the values of `X2020.4.01` to `X2020.4.30`. What is the total number of infections for this period across all locations?

Hint: there are many ways to do this. You can optionally use the `rowSums` function in R.

```r
# Directly specifying the column names for April 1 to April
# 30, 2020
covid$total_infections_april <- covid$X2020.4.1 + covid$X2020.4.2 +
    covid$X2020.4.3 + covid$X2020.4.4 + covid$X2020.4.5 + covid$X2020.4.6 +
    covid$X2020.4.7 + covid$X2020.4.8 + covid$X2020.4.9 + covid$X2020.4.10 +
    covid$X2020.4.11 + covid$X2020.4.12 + covid$X2020.4.13 +
    covid$X2020.4.14 + covid$X2020.4.15 + covid$X2020.4.16 +
    covid$X2020.4.17 + covid$X2020.4.18 + covid$X2020.4.19 +
    covid$X2020.4.20 + covid$X2020.4.21 + covid$X2020.4.22 +
    covid$X2020.4.23 + covid$X2020.4.24 + covid$X2020.4.25 +
    covid$X2020.4.26 + covid$X2020.4.27 + covid$X2020.4.28 +
    covid$X2020.4.29 + covid$X2020.4.30

# The easy way to do this is to use the `rowSums` function
# in R. This function calculates the sum of the values in
# each row of a matrix or data frame.

april_columns <- c("X2020.4.1", "X2020.4.2", "X2020.4.3", "X2020.4.4",
    "X2020.4.5", "X2020.4.6", "X2020.4.7", "X2020.4.8", "X2020.4.9",
    "X2020.4.10", "X2020.4.11", "X2020.4.12", "X2020.4.13", "X2020.4.14",
    "X2020.4.15", "X2020.4.16", "X2020.4.17", "X2020.4.18", "X2020.4.19",
    "X2020.4.20", "X2020.4.21", "X2020.4.22", "X2020.4.23", "X2020.4.24",
    "X2020.4.25", "X2020.4.26", "X2020.4.27", "X2020.4.28", "X2020.4.29",
    "X2020.4.30")
covid$total_infections_april <- rowSums(covid[, april_columns],
    na.rm = TRUE)
```

```r
# The REALLY easy way is to use grep
covid$total_infections_april <- rowSums(covid[, grep("X2020.4",
    names(covid))], na.rm = TRUE)

# Compute the total number of infections for this period
# across all locations
sum(covid$total_infections_april)
```

```
## [1] 3753
```

Answer: 3753

2.2. Compute the average treatment effect (ATE) for school closure on the number of COVID-19 infections per 100,000 residents in April 2020. The ATE is defined as the difference in the average number of infections between the treatment and control groups.

```r
mean_infections_treatment <- mean(covid$total_infections_april[covid$shutdown.0406 ==
    1], na.rm = TRUE)
mean_infections_control <- mean(covid$total_infections_april[covid$shutdown.0406 ==
    0], na.rm = TRUE)

# Compute the Average Treatment Effect (ATE) for school
# closure
ate_school_closure <- mean_infections_treatment - mean_infections_control

round(ate_school_closure, 2)
```

```
## [1] 10.24
```

Answer: 10.24

2.3. Interpret the ATE. What does this value tell us about the effect of school closures on the number of COVID-19 infections per 100,000 residents in April 2020?

Answer: The ATE of 10.24 indicates that school closures led to an average increase of 10.24 COVID-19 infections per 100,000 residents in April 2020.

2.4. Compute the 25th and 75th percentiles of the `students` variable. **Round to four decimal places.**

```r
# Calculate the 25th percentile
percentile_25 <- quantile(covid$students, 0.25, na.rm = TRUE)

# Calculate the 75th percentile
percentile_75 <- quantile(covid$students, 0.75, na.rm = TRUE)

round(percentile_25, 4)
```

```
##    25%
## 0.0668
```

```r
round(percentile_75, 4)
```

```
##     75%
## 0.0817
```

Answer (25%): .0668
Answer (75%): .0817

2.5. Using the values you computed above, divide the municipalities that introduced the school closure interventions into three groups based on the `students`. Among these municipalities, the 'High exposure' group represents the group of municipalities whose student population density index was greater than or equal to the 75 percentile. The 'Low exposure' group represents the group of municipalities whose student population density index was less than or equal to the 25 percentile. The 'Medium exposure' group represents the remaining municipalities. Store these three labels to a new column called `student_density_level`. How many municipalities are in each group?

```
# Assign 'Medium exposure' to all rows and then update to
# high or low.

covid$student_density_level[covid$shutdown.0406 == 1] <- "Medium exposure"

# Assign 'High exposure' for student count >= 75th
# percentile among municipalities with school closures
covid$student_density_level[covid$shutdown.0406 == 1 & covid$students >=
    percentile_75] <- "High exposure"

# Assign 'Low exposure' for student count <= 25th
# percentile among municipalities with school closures
covid$student_density_level[covid$shutdown.0406 == 1 & covid$students <=
    percentile_25] <- "Low exposure"

table(covid$student_density_level)
```

```
##
##   High exposure    Low exposure Medium exposure
##              81              39             136
```

```
# or

# Assign 'Medium exposure' to all rows and then update to
# high or low.

covid$student_density_level[covid$shutdown.0406 == 1] <- "Medium exposure"

# Assign 'High exposure' for student count >= 75th
# percentile among municipalities with school closures
covid$student_density_level[covid$shutdown.0406 == 1 & covid$students >=
    0.0817] <- "High exposure"

# Assign 'Low exposure' for student count <= 25th
# percentile among municipalities with school closures
covid$student_density_level[covid$shutdown.0406 == 1 & covid$students <=
    0.0668] <- "Low exposure"

table(covid$student_density_level)
```

```
##
##   High exposure    Low exposure Medium exposure
##              80              39             137
```

Answer (High exposure): 80 or 81
Answer (Low exposure): 39
Answer (Medium exposure): 137 or 136

2.6. Why is the number of municipalities in the 'High exposure' group different from the number of municipalities in the 'Low exposure' group?

Answer: We computed the quartiles based on the entire dataset, not just the municipalities with school closures. As a result, the number of municipalities in the 'High exposure' group is different from the number of municipalities in the 'Low exposure' group.

2.7. What *percent* of municipalities had a proportion of elderly residents over 25% and a mayor who was over 65 years old?

```
# Number of municipalities with a proportion of elderly
# residents over 25% and a mayor who was over 65 years old
old_places <- nrow(covid[covid$age.0406 > 60 & covid$elder.pop >
    0.25, ])

round(old_places/nrow(covid) * 100, 2)
```

```
## [1] 52.5
```

Answer: 52.5%