API - 209 | Problem Set 1

Prof. Dan Levy

Due on Tuesday, September 6 2022 at 10:00 am.

INSTRUCTIONS

To successfully complete this problem set, please follow these steps:

1. **Download this RMarkdown document file into your computer.**
2. **Insert all your answers into this document.** Guidance **here** on how to insert objects such as handwritten work or screenshot images in your answers.
3. **SAVE your work frequently**.
4. To make things easier to visualize in RStudio, you can set the view mode as "Visual" instead of as "Source" in the top left of your screen (just below the Save button).
5. Once your document is complete, please save it as a PDF by clicking the **KNIT** button.
6. Please submit an electronic copy of the PDF (and any separate requested files) to the Canvas course page.

   6.a) If you want to check a PDF version of this problem set before starting to work on it, you can always knit it. In fact, you can knit the document at any point.

   6.b) If you cannot Knit and it's time to submit the problem set, submit the RMarkdown file and make an appointment with a member of the teaching team
7. Remember to consult the R resources from math camp, particularly the HKS R cheat sheet (available **here**, which contains many of the commands needed to answer the questions in this problem set.

IDENTIFICATION

1. Your information

```
Last Name:  Chaturvedi
First Name: Shreya
```

2. Group Members (please list below the classmates you worked with on this problem set):

```
Group members: Manisha Jha, Kelly Jiang, Alice Zhang, Rushabh Sanghvi, Vardan
```

3. Compliance with Harvard Kennedy School Academic Code: Do you certify that my work in this problem set complies with the Harvard Kennedy School Academic Code[1] (mark with an X below)?

[1]We abide by the Harvard Kennedy School Academic code (available here) for all aspects of the course. In terms of problem sets, unless explicitly written otherwise, the norms are the following: You are free (and encouraged) to discuss problem sets with your classmates. However, you must hand in your own unique written work and code in all cases. Any copy/paste of another's work is plagiarism. In other words, you can work with your classmate(s), sitting side-by-side and going through the problem set question-by-question, but you must each type your own answers and your own code. For more details, please see syllabus.

[X] YES [ ] NO

QUESTION 1 – COVID-19 TESTING

The goal of this question is to help you develop your ability to use statistics to understand COVID-19 testing. Imagine this is mid 2020 and you are trying to understand how to accurately interpret COVID test results. First a quick primer on COVID tests.

A quick primer on COVID tests[2]

There are two types of tests: **diagnostic tests** and antibody tests. A diagnostic test aims at detecting whether the person currently has an active coronavirus infection. Currently there are two types of diagnostic tests which detect the virus – molecular tests, such as RT-PCR tests, that detect the virus's genetic material, and antigen tests that detect specific proteins on the surface of the virus.

An **antibody (serology) test** aims at detecting whether the person had an infection, by assessing whether the person has developed antibodies against the virus. If test results show that the person has antibodies, it indicates that the person was likely infected with COVID-19 at some time in the past. It may also mean that the person has some immunity. But there is no definitive evidence on whether having antibodies means the person is protected against reinfection with COVID-19. The level of immunity and how long immunity lasts are not yet known.

Our question

We will start with antigen tests, which are diagnostic tests that are rapid to administer and get results. We will focus on one of the tests that have been approved by the FDA. This test has a specificity rate (percent not infected correctly identified as negative) of 100%. The sensitivity rate (percent of infected correctly identified as positive) is 97%[3]. Assume that the prevalence of COVID-19 in your population of interest is 1.2%.

(a) Calculate the probability that a person who tests positive is infected (i.e. P(COVID|+)). This number is usually referred to as the positive predictive value of the test.

Answer: 1 - Everyone who tested positive does have COVID

(b) Calculate the probability that a person who tests negative is not infected (i.e. P(NO COVID|-)). This number is usually referred to as the negative predictive value of the test.

Answer: 0.9996

(c) Is the result in (a) substantially different than the ones we got for mammograms in class? If so, explain why this is the case. If not, explain why not. [2-3 sentences]

Answer: Yes, the result is significantly different from the one we derived in class. This is because there are no people who do not have COVID and still test positive. This results in the second term in the denominator being zero and the overall expression being significantly higher.

+:==========+ +——————————————————————–+

[2] Sources: FDA, Mayo clinic.

[3] Note that the sensitivity rate of various COVID antigen tests vary. This figure is from Color, the provider of tests for Harvard. Source: https://www.color.com/covid19-details.

(d) The FDA indicates that for antigen tests "positive results are usually highly accurate but negative results may need to be confirmed with a molecular test."

(d.1) Do you agree with the FDA statement? Explain why or why not.

Answer: I think that people who do not test positive for COVID do not need to undergo a molecular test, as the odds of a person having COVID given they tested positive is very low (less than 4 in 10000).

+:==========+ +——————————————————————+

(d.2) Suppose you are in charge of COVID testing in your city, and are trying to evaluate whether to use this antigen test for your citizens. Suppose you find it unacceptable for the probability of having COVID, conditional on a negative test, to be more than 5%. For what range of values of the COVID prevalence rate would you expect this conditional probability to be greater than 5%?

Answer: For a prevalence rate that is greater than 63.6%, this conditional probability would be greater than 5%.

+:==========+ +——————————————————————+

Now let's explore *antibody tests*, which aim at detecting whether the person had the disease at some point in the past. We will focus our attention on the Wadsworth New York SARS-CoV Microsphere Immunoassay for Antibody Detection test, which is reported to have a sensitivity of 88% and a specificity of 98.8 [4]%.

(e) A random person in your country gets this test and has a positive result. Calculate the probability that this person had COVID-19 at some point in the recent past. To do this, explain:

(e.1) How you got the prevalence rate for your country. Note: Given COVID is still rapidly evolving, you won't find this number easily. Take your best guess of this number based on data that you find on sites such as this one. Do not spend more than 10 minutes trying to estimate this number. The goal is to get you to confront the imperfections of data in the real world, not to turn this into a research project.

Answer: The prevalence rate in India is the proportion of the cumulative cases in the whole population. The website shows a total of 44.46 million cases in India for a total population of 1.38 billion. This results in a prevalance rate of 3.2%.

+:==========+ +——————————————————————+

(e.2) What is the probability that this person had COVID-19 at some point in the recent past?

Answer: P(COVID | +ve) is 0.7079. This means that there is a 71% chance this person had COVID in the past.

+:==========+ +——————————————————————+

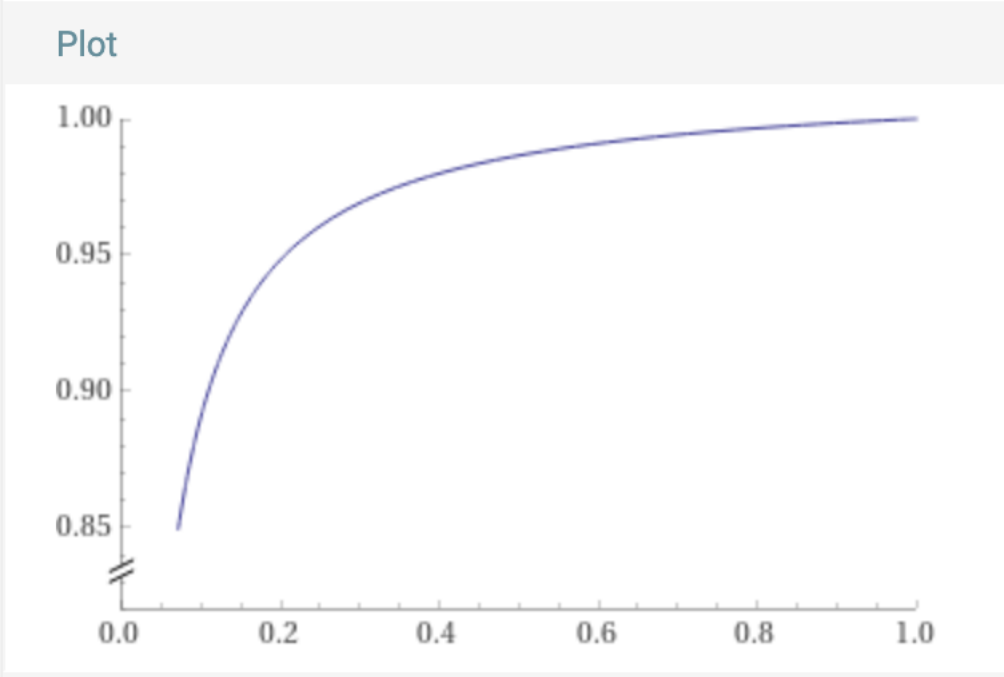
(f) Draw a graph (by hand, Excel, R or **here**) of the positive predictive value of the antibody test as a function of the prevalence of COVID-19. Describe the graph and discuss the implications for interpreting the results of antibody tests.

[4] Source: The Wall Street Journal.

Answer: The curve for the positive prediction rate grows steeply till the prevalance rate is about 25%, post which it somewhat flattens. This implies that the positive prediction value of a antibody test is high and reliable for a prevalance rate that is higher than roughly 25%.

+:==========+ +——————————————————————+

Please insert your graph here: |



+:==========+ +——————————+

(g) Suppose a random member of the population tests positive on a first antibody test. He goes to take another antibody test with the same sensitivity and specificity as the one described above, and the result is also positive. What is the probability that the person has the disease after the second test? Assume the two tests are independent.

Write your calculations here (or insert a screenshot of your work):

(g) $P(covid \mid ++) = \frac{P(+|covid)^2 \times P(covid)}{P(+|covid)^2 P(covid) + P(+|\neg covid)^2 \cdot P(\neg covid)}$

$= \frac{0.88^2 \times 0.032}{0.88^2 \times 0.032 + 0.012^2 \times 0.968}$

$= 0.9944$

QUESTION 2 – LEARNING ABOUT THE WORLD ECONOMY

The purpose of this question is to help you learn about the world economy while developing your R skills to analyze data. This will be the first of several problem sets in which you will be developing your R skills. There are some hints in the Appendix below to help you with the questions below. My advice is that you refer to the appendix only after having tried to figure out how to do it by yourself; this will better allow you to develop your R skills.

We strongly recommend you set up your environment as provided in day 5 of math camp. Specifically, you should create a subdirectory for problem set 1, keep all your data, code, and work in that directory, and make a RStudio Project specifically for that project directory.

The dataset for this question is an extract from World Bank's World Development Indicators (WDI). Please download the dataset "PS1_dataset.csv" from our Canvas website. The dataset includes three tabs: one with variable definitions and source information; one with the data extract that you will use for this question; and one with the full time series of the key variables from 1960 through 2019 for those that may be interested.

First, familiarize yourself with the worksheet "Data Extract (1993 and 2019)". Then import these data into R and answer the questions below. We will focus a lot of attention on the gross domestic product (GDP), a concept you dealt with in your Macro class with Prof. Frankel. Note that the WDI dataset reports GDP values in 2010 U.S. dollars, so that so you can directly compare values in different years. Also note that GDP and population data are not available in some of the years in the spreadsheet. For this problem set, use only those observations for which data are available. (A question dealing with missing data is further below).

GENERAL GUIDANCE FOR QUESTIONS INVOLVING R

(1) Explore the data set: An essential practice before doing any data analysis is to explore the data set. Here are some questions to ask:

- What is the unit of observation (i.e. country, year, country/year, etc.)?
- How many observations are in the data set?
- What are the key variables in the data set?
- For the key variables, how are they coded, what is the extent of missing data, and how will I deal with this missing data?

Once you have done this (no need to type answers to these questions, but do answer them), create an analysis data set in which:

- You transform the population variables so that they are expressed in millions of people (for example, 158,000,000 should become 158).
- You transform the gdp variables so that they are expressed in millions of dollars
- You keep only observations that have non-missing data for both population 2019 and gdp 2019.

This will be the data set you will you for the remainder of this problem set, so assign it to an object you can use. Now report the mean and the number of observations for gdp 2019 for this analysis data set.

```
## [1] 450988.6
```

```
## [1] 185
```

(2) Totals: Please calculate and report: a. Total World GDP (expressed in trillions of 2010 dollars) and world population in 2019 (expressed in billions of people)

```
## # A tibble: 1 x 2
##   totalgdp totalpop
##      <dbl>    <dbl>
## 1     83.4     7.53
```

b. Top 5 countries in terms of GDP in 2019 and their respective GDPs and top 5 countries in terms of population in 2019 and their respective populations. You may round your values.

```
## # A tibble: 5 x 2
##   country       gdp_2019_mn
##   <chr>               <dbl>
## 1 United States    18300000
## 2 China            11500000
## 3 Japan             6210000
## 4 Germany           3940000
## 5 France            2970000
```

```
## # A tibble: 5 x 2
##   country       pop_2019_mn
##   <chr>               <dbl>
## 1 China               1398.
## 2 India               1366.
## 3 United States        328.
## 4 Indonesia            271.
## 5 Pakistan             217.
```

(3) Central Tendencies: Calculate GDP per capita (which is equal to total GDP divided over population, but beware of units) for each country in the database for 1993 and 2019. Then report the following statistics. To help you code more efficiently, feel free to write code chunks that answer several of the questions below at the same time: a. GDP per capita for the average country in 2019.

```
## # A tibble: 1 x 1
##   mean_gdp_per_capita_2019
##                      <dbl>
## 1                   15011.
```

b. World GDP per capita in 2019 (equal to: Total world GPD / Total world population).

```
## # A tibble: 1 x 3
##   totalgdp   totalpop total_gdp_per_capita
##      <dbl>      <dbl>                <dbl>
## 1  8.34e13 7528454316               11082.
```

c. Explain what is the difference between (a) and (b) in a language that someone not well-versed in statistics can understand.

Answer: The first answer is the gotten by computing the average value of the GDP per capita for all countries in our dataset. The second value is a more distributed measure - that is the total GDP of the world divided by the total population of the world.

+:===+ +———————————————————————+

d. What do you think drives the difference between (a) and (b)?

Answer: The average value of GDP per capita for all countries is higher than the GDP per capita of the whole world (that is, total GDP by total population). This is because the average is skewed or biased because of certain countries in the data set having really high GDP per capita values.

+:===+ +———————————————————————+

e. Median GDP per capita in 2019.

```
## # A tibble: 1 x 1
##   median_gdp_per_capita_2019
##                        <dbl>
## 1                      5922.
```

f. What do you think drives the difference between (a) and (e)?

Answer: The median is less sensitive to outliers (that is, the median does not jump because of the presence of some countries with really high GDP per capita values).

+:===+ +———————————————————————+

g. The total population of all countries in 2019 with GDP per capita below the world mean calculated in part (a).

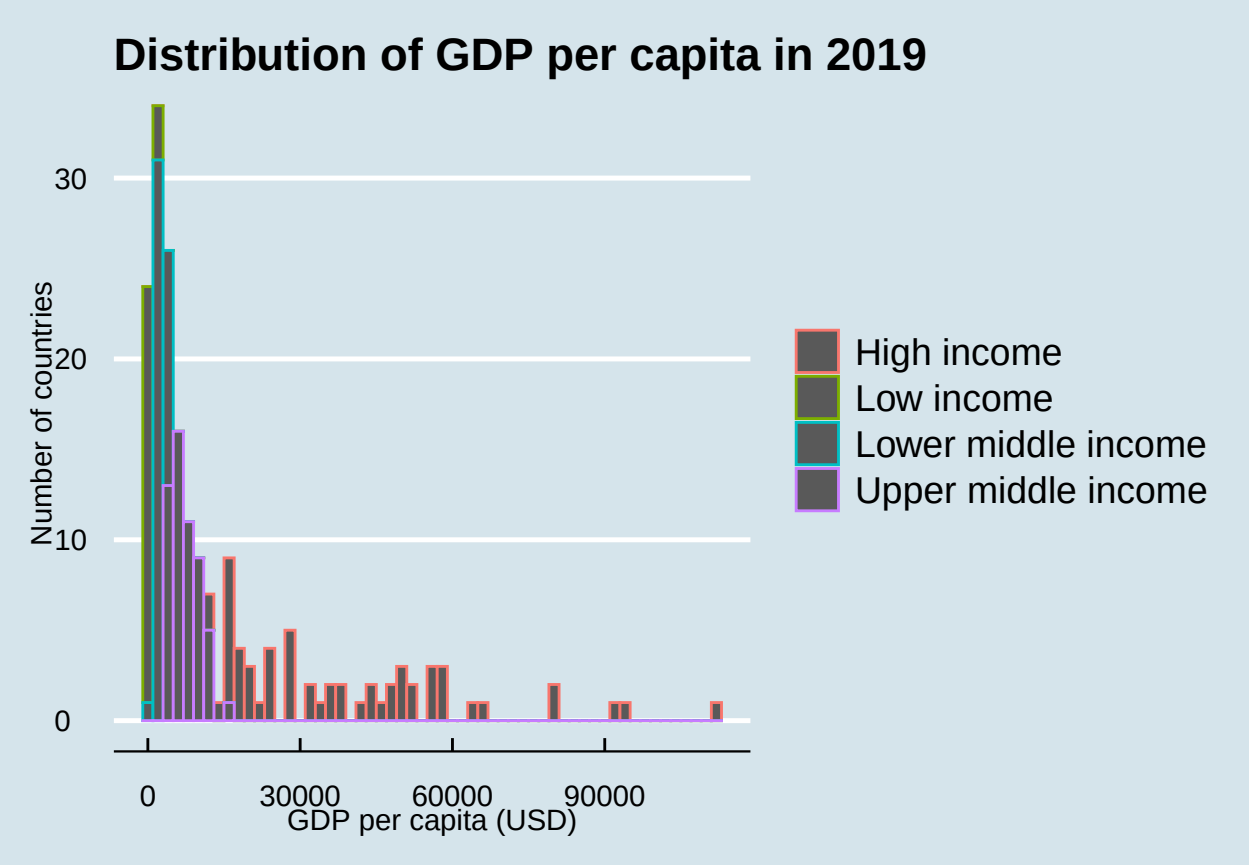
```
## # A tibble: 1 x 1
##     totalpop
##        <dbl>
## 1 6262875551
```

h. What is your answer to (g) telling you in terms of how appropriate mean world GDP per capita is in characterizing the economic well-being of the average person in the world? [2-3 sentences]

Answer: This number shows that a majority of the people (more than 80%) live in countries that have less than the mean GDP per capita in the year 2019. In other words, the GDP "overestimates" how well off the average person is in reality. Instead, only a few people (less than 20%) in the world are in a country that has more than the mean GDP per capita .

+:===+ +———————————————————————+

i. Draw a histogram showing the distribution of GDP per capita in 2019, using bins that are $2,000 wide. Imagine this histogram were to appear in The Economist, i.e. make it well-labeled and professionally looking. Feel free to tweet it using #api209.



(4) Regional Variation: The data set also contains information for each country regarding the World Bank's classification of region of the world and income group. Answer the following questions:

a. Produce a table that contains the percent of countries in each income group category. Make sure that the table is sorted from low income to high income (rather than alphabetically).

Income-wise Categorization of Countries

Income Group	Percentage (%)
Low income	14.29
Lower middle income	21.66
Upper middle income	27.65
High income	36.41

b. What percent of countries in Sub-Saharan Africa are low income?

```
## [1] 0.5
```

c. Assess whether the following statement is supported by the evidence: The majority of low-income countries are in sub-Saharan Africa.

Answer: 24 of the 31 low income countries are in sub saharan africa. Therefore, this statement is true.

+:==========+ +——————————————————+

(5) Missing data: Missing data is, unfortunately, a fact of life, and we face it here. While we cannot always fix the problem of missing data, it is important to consider its effects.

a. The database does not have a 1993 GDP figure for how many countries?

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 2.38e+07 3.83e+09 1.36e+10 2.34e+11 1.01e+11 9.56e+12
##     NA's
##       18
```

b. Calculate the mean GDP per capita in 1993 ignoring missing values.

```
## # A tibble: 1 x 1
##   mean_gdp_per_capita_1993
##                      <dbl>
## 1                    9998.
```

c. Divide countries into two groups: those who have missing GDP per capita data in 1993 and those that don't. Calculate the mean GDP per capita in 2019 for these two groups.

```
## # A tibble: 2 x 2
##   group   mean
##   <dbl>  <dbl>
## 1     0 15396.
## 2     1 11648.
```

d. Which countries seem more likely to have missing data in 1993? Feel free to use your answer in (c) and/or do additional calculations/explorations of the data.

(6) Putting it all together: In one crisp paragraph, summarize your findings from your analyses in parts (1)-(4)?

Answer: This dataset shows the GDP and population statistics for certain countries in 1993 and 2019. Using this dataset, we can understand that the GDP per capita in 2019 distribution is right skewed. In fact, only less than 20% of the world's population lives in countries with GDP per capita higher than the mean value. About 14% of the countries (by count) are classified as low income as per the World Bank classification, and most of these fall in Sub Saharan Africa.

+:===+ +———————————————————————+

QUESTION 3 – MEXICAN PENSIONS FOR THE POOR

Read the case study "Providing Pensions for the Poor: Targeting Cash Transfers for the Elderly in Mexico." (linked in class #4 readings in home page of our Canvas site).

(1) Calculate the leakage and undercoverage rates for each of the three options. Show your calculations and report the final results below:

OPTION 1

$$\text{leakage} = \frac{0.1871 \times 4592726 - 0.3407 \times 187233}{0.1871 \times 4592726}$$

$$= 0.2577$$

$$\text{undercoverage} = 1 - 0.3407$$

$$= 0.6593$$

OPTION 2

$$\text{leakage} = \frac{0.3077 \times 4592726 - 0.4281 \times 187233}{0.3077 \times 4592726}$$

$$= 0.4328$$

$$\text{undercoverage} = 1 - 0.4281$$

$$= 0.5719$$

OPTION 3

$$\text{leakage} = \frac{0.6621 \times 4592726 - 0.5597 \times 187233}{0.6621 \times 4592726}$$

$$= 0.656$$

$$\text{undercoverage} = 1 - 0.5597$$

$$= 0.4403$$

(2)

Answer:

Option 1 Leakage = 0.2577
Option 2 Leakage = 0.4328
Option 3 Leakage = 0.656
Option 1 Undercoverage = 0.6593
Option 2 Undercoverage = 0.5719
Option 3 Undercoverage = 0.4403

+==========+ +——————————————————+

(2) Assess the advantages and disadvantages of each of the options. In doing so, consider the targeting effectiveness as well as other criteria (including political, logistical, and financial). Summarize your findings in the table below.

Option 1 Advantages: Ease of implementation (built on existing infrastructure of Oportunidades), low leakage
Option 1 Disadvantages: Highest undercoverage of all options
Option 2 Advantages: Focus on rural communities that have been previously underserved
Option 2 Disadvantages: Implementation challenges due to a lack of set up in sparse communities
Option 3 Advantages: Prioritization of marginalized individuals through targeting
Option 3 Disadvantages: Highest leakage of all options, highest financial cost of all options

(3) Write one crisp paragraph to the Secretary of Social Development (Sedesol) recommending which option should be selected to target the pension program and why. Justify your recommendation using the advantages and disadvantages you identified above. The paragraph should be written in a language that the head of Sedesol (who you can assume is intelligent and well-educated, but not well-versed in statistics) can understand. Enter your answer **here**.

Enter your answers in the link above. Once completed, type "DONE" into this field.

QUESTION 4 – ONLINE UNIT ON DECISION ANALYSIS

The goal of this problem set question is to help you prepare you for the class on **Decision Analysis** that will be held on **Tuesday, September 6** (the day this problem set is due). The idea is to get everyone familiar with the basics of decision trees so that we can delve deeper in class on this topic than we would be able to do if we had to go through the basics in class.

You will be asked to engage with a short module (total running time is about 10 minutes) and answer some questions in a quiz. The quiz results will give me information about overall performance of the class that I will use to prepare class; your individual performance in the quiz will be registered in the system but will not count towards your grade in any way.

To get full credit for this question, you need to watch the module and complete the quiz. The module is available **here**.

Please enter "Done" in this field once you have completed the quiz.
Done!

TIME USE

Please enter in **this** form the time you spent on each question. This information will only be used for teaching improvements; **please be candid**.

Please enter "Done" in this field once you have completed the quiz.

This is a copy of your code.

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(options(width = 60))
knitr::opts_chunk$set(class.output = "bg-warning")

packages <- c('haven','dplyr', 'ggplot2', 'reshape2', 'tidyverse', 'pracma',
              'lubridate', 'scales', 'ggthemes')
to_install <- packages[!(packages %in% installed.packages()[,"Package"])]
if(length(to_install)>0) install.packages(to_install)
lapply(packages, require, character.only=TRUE)


Last Name:  Chaturvedi
First Name: Shreya
Group members: Manisha Jha, Kelly Jiang, Alice Zhang, Rushabh Sanghvi, Vardan
                              [ X ] YES                [   ] NO
# You can use this code chunk to upload the dataset.
rawdata <- read_csv("WDI Data Extract API-209 - PS 1 - 2022.csv")

data <- rawdata %>%
  mutate(pop_2019_mn = pop_2019/1000000, pop_1993_mn = pop_1993/1000000, gdp_1993_mn = gdp_1993/1000000,
  drop_na(gdp_2019) %>%
  drop_na(pop_2019)

mean_gdp_2019 <- mean(data$gdp_2019_mn)
count_gdp_2019 <- nrow(data)
mean_gdp_2019
count_gdp_2019
# Insert your code here.
data %>%
  summarise(totalgdp = sum(gdp_2019_mn), totalpop = sum(pop_2019_mn)) %>%
#converting to trillions of USD and billions of people
  mutate(totalgdp = totalgdp/1000000, totalpop = totalpop/1000)

# Insert your code here.
top5gdp <- data %>%
  arrange(desc(gdp_2019_mn)) %>%
  slice(1:5) %>%
  select(country, gdp_2019_mn)



top5pop <- data %>%
  arrange(desc(pop_2019_mn)) %>%
  slice(1:5) %>%
  select(country, pop_2019_mn)

top5gdp
top5pop

# Insert your code here.
data %>%
  mutate(gdp_per_capita_1993 = gdp_1993/pop_1993, gdp_per_capita_2019 = gdp_2019/pop_2019) %>%
  summarize(mean_gdp_per_capita_2019 = mean(gdp_per_capita_2019))

# Insert your code here.
data%>%
  summarise(totalgdp = sum(gdp_2019), totalpop = sum(pop_2019)) %>%
  mutate(total_gdp_per_capita = totalgdp/totalpop)

# Insert your code here.
```

```r
data %>%
  mutate(gdp_per_capita_1993 = gdp_1993/pop_1993, gdp_per_capita_2019 = gdp_2019/pop_2019) %>%
  summarize(median_gdp_per_capita_2019 = median(gdp_per_capita_2019))

# Insert your code here.
data %>%
  mutate(gdp_per_capita_1993 = gdp_1993/pop_1993, gdp_per_capita_2019 = gdp_2019/pop_2019) %>%
  filter(gdp_per_capita_2019 < mean(gdp_per_capita_2019)) %>%
  summarise(totalpop = sum(pop_2019))



# Insert your code here.


plot <- data %>%
  mutate(gdp_per_capita_1993 = gdp_1993/pop_1993, gdp_per_capita_2019 = gdp_2019/pop_2019) %>%
  ggplot(aes(x = gdp_per_capita_2019, color = income_group)) +
  geom_histogram(binwidth = 2000) +
  labs(title = "Distribution of GDP per capita in 2019",
       x = "GDP per capita (USD)",
       y = "Number of countries")  +
  theme_economist() +
  theme(legend.position = "right",legend.title = element_blank())


plot
# Insert your code here.
library(gt)
options(digits = 4)
rawdata %>%
  group_by(income_group) %>%
  summarise(count = n()) %>%
  mutate(percent = count/sum(count) * 100) %>%
  mutate(income_cat = case_when(income_group == "Low income" ~ 1,
                                income_group == "Lower middle income" ~ 2,
                                income_group == "Upper middle income" ~ 3,
                                income_group == "High income" ~ 4)) %>%
  arrange(income_cat) %>%
  select(income_group, percent) %>%
  gt() %>%
  tab_header(
    title = "Income-wise Categorization of Countries",
  ) %>%
  cols_label(
    income_group = "Income Group",
    percent = "Percentage (%)"
  )
# Insert your code here.
ssa <- sum(rawdata$region == "Sub-Saharan Africa")

ssali <- sum(rawdata$income_group == "Low income" & rawdata$region == "Sub-Saharan Africa")



ssali/ssa

#50% of the countries in sub saharan africa are low income.

li <- sum(rawdata$income_group == "Low income")
```

```r
# Insert your code here.
summary(data$gdp_1993)

#The dataset does not have GDP values for 18 countries in 1993.
# Insert your code here.

data %>%
  mutate(gdp_per_capita_1993 = gdp_1993/pop_1993) %>%
  summarize(mean_gdp_per_capita_1993 =  mean(gdp_per_capita_1993, na.rm = TRUE))

# Insert your code here
data %>%
  mutate(gdp_per_capita_1993 = gdp_1993/pop_1993, gdp_per_capita_2019 = gdp_2019/pop_2019) %>%
  mutate(group = ifelse(is.na(gdp_per_capita_1993),1,0)) %>%
  group_by(group) %>%
  summarise(mean = mean(gdp_per_capita_2019))




# Insert your code here.

#Countries with really small populations or political instability are more likely to have missing data in
```