PROBLEM SET 1 – Suggested Answers

**Due on Tuesday, Sept 8, 2020, 8:30 am.**

I - INSTRUCTIONS

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

1. **Download this Word document file into your computer**
2. **Insert all your answers into this Word document**. Guidance [here](https://www.dropbox.com/s/ox9fhmbpvy2viw5/How%20to%20incorporate%20handwritten%20work%2C%20Stata%20output%2C%20and%20screenshot%20images.pdf?dl=0) on how to insert non-Word objects such as handwritten work or screenshot images in your answers.
3. **Once your document is complete, please save it as a PDF**. This is important to make sure all your work is preserved in the process of submission to Canvas.
4. **Please submit an electronic copy of the PDF (and any separate requested files) to the Canvas course page**.

II - IDENTIFICATION

1. **Your information**

|  |  |
| --- | --- |
| Your Last Name: |  |
| Your First Name: |  |

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

|  |
| --- |
|  |

1. **Compliance with Harvard Kennedy School Academic Code****[[1]](#footnote-1) (mark with an X below)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | **Yes** | **No** | | I certify that my work in this problem set complies with the Harvard Kennedy School Academic Code |  |  | |

**Table of Contents**

[QUESTION 1 – COVID-19 TESTING 2](#_Toc49920119)

[QUESTION 2 – ONLINE UNIT ON DECISION ANALYSIS 4](#_Toc49920120)

[QUESTION 3 – MEXICAN PENSIONS FOR THE POOR 5](#_Toc49920121)

[QUESTION 4 – LEARNING ABOUT THE WORLD ECONOMY 7](#_Toc49920122)

# 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. First a quick primer on COVID tests.

|  |
| --- |
| **A quick primer on COVID tests**[[2]](#footnote-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 a lack of 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. |

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 two 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 84%.[[3]](#footnote-3) Assume that the prevalence of COVID-19 in your population of interest is 1%.

1. 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.

|  |
| --- |
| *P(COVID|+)===*  Note, in this case, the only way to receive a positive test result is by having the disease, because the specificity of the test is perfectly accurate. |
|  |

1. 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.[[4]](#footnote-4)

|  |
| --- |
| *P(No COVID|-)===* |

1. 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 sentence]

|  |
| --- |
| *The answer in part (a) is very different from the result obtained in class. This is driven by specificity of the test. With perfect specificity, there is no chance of a false positive, and therefore the probability of having COVID conditional on a positive test result is 100%.* |

1. 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.”
   1. Do you agree with the FDA statement? Explain why or why not.
   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 to have a false negative rate of more than 5%. For what range of values of the COVID prevalence rate would you expect the rate of false negatives to be greater than 5%?

|  |
| --- |
| *The sensitivity of the test is not perfect, so there is a chance we have false negatives. However, based on our calculation from part (b), given the low prevalence of COVID and the perfect specificity of these tests, conditional on having a negative test result, we are already 99.84% sure that person does not have COVID. While molecular tests could increase this probability closer to 100%, the gains would be marginal.*  ***Note, we should have been more precise in our definition of ‘false negative rate’ in this problem. Here, we meant for false negative rate to refer to P(COVID|-), which would equal the proportion of the total negative test results which are false negatives. Other sources correctly point out the false negative rate is defined as P(-|COVID), which is directly related to the sensitivity of the test. P(-|COVID) would not however vary by the underlying prevalence P(COVID), which is the concept this question intended to focus on. We will award full credit for answers which instead focused on P(-|COVID). By Bayes’ Rule, P(COVID|-) is equal to:***  *We want to solve for the prevalence of COVID such that this value is greater than 0.05. Defining the unknown COVID prevalence as x, we therefore require:*  *Solving for x, we have: 0.2475 < x < 1.* |

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%.[[5]](#footnote-5)

1. 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:
2. How you got the prevalence rate for your country. Note: Given COVID is 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](https://ourworldindata.org/coronavirus). 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.
3. What is the probability that this person had COVID-19 at some point in the recent past?

|  |
| --- |
| *Answers will vary. There are many possible arguments for the answer to part (a). Perhaps the simplest way to arrive at this number is by comparing the total confirmed cases in a country divided by the country’s population. This would give a crude lower bound of the prevalence during the period. This number is likely a lower bound because not all cases in the population will be confirmed with tests.*  *Alternatively, you might think it is most relevant to calculate this prevalence in a window around when the person was tested. A very high prevalence in the early months of the pandemic may not be indicative of current prevalence or probability of acquiring the disease.*  *Call whatever estimate you obtain from this first part ‘x’. We can then apply Bayes’ Rule in the usual way to solve for part (b), where ‘+’ now refers to a positive antibody test.*    *P(COVID|+)==* |

1. Draw a graph (by hand, Excel, R or [here](http://www.wolframalpha.com/examples/mathematics/plotting-and-graphics/)) 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.

|  |
| --- |
| *Generated with the following code:*  *library(tidyverse)*  *# create a dataframe showing the range of x*  *x\_range <- tibble(x = c(0, 0.4))*  *# draw a continuous function using stat\_function based on Bayes' Rule calculation*  *ggplot(x\_range, aes(x = x)) +*  *stat\_function(fun = function(p) (0.88\*p)/(0.88\*p + (1-0.988)\*(1-p))) +*  *scale\_x\_continuous(expand = c(0, 0)) +*  *labs(x = "Prevalence of COVID",*  *y = "Positive Predicted Value of Antibody Test")+*  *ylim(c(0,1))+*  *theme\_light()*  *ggsave("COVID\_pred-value-curve.pdf", width = 4, height = 4)*  *ggsave("COVID\_pred-value-curve.png", width = 4, height = 4)* |

|  |
| --- |
| *There is a positive relationship between baseline prevalence and positive predictive value of the test. As the baseline prevalence increases, we become more likely in our predicted positive result. This makes sense because the number of false negatives will decrease as the prevalence increases. In the extreme case, if the prevalence were 100%, every positive test result would correspond to a positive case (because there are only positive cases) and this predictive value would be 100%.* |

1. 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.

|  |
| --- |
| ***Approach #1 – Single Event***  *Perhaps the easiest way to solve this problem is by treating the person getting two positive test results as a single event. By the independence of the tests, the probability of two tests showing the same result, conditional on disease status, will just be the product of their individual probabilities. Recalling that the prevalence of COVID was denoted by ‘x’, we then have:*  *This will increase our assessment of the probability the person has COVID, as it is extremely unlikely to get two false positives on two tests given the specificity.*  *If x=0.01, then P(COVID|++) = 0.9819*  ***Approach #2 - Iterative Approach***  *You could also update iteratively, treating your first calculation of P(COVID|+) as the new ‘prior’ probability of COVID going into the calculation for the second test. The two approaches are equivalent and will yield the same answer, but the above involves less algebra.*  *Assuming P(COVID)=0.01, after the person takes the first test:*  *P(COVID|+)===*  *Now before the person takes the second test, the probability that the person has COVID has been updated from 0.01 to 0.4255. So to calculate the probability that the person has COVID given a second test result, we employ Bayes rule again but with this updated probability of COVID, i.e. P(COVID)=0.4255 is now the prevalence for this second test.*  *P(COVID|+)===*  *As you can see, the two methods yield identical results!*  *The reason it is helpful to think about the second iterative approach is that it helps you understand what “Bayesian Updating” means. Prior to the first test, you have a probability of disease. You get a positive result, and update this probability. When you go into the second test, your probability of disease has been updated.* |
|  |

# QUESTION 2 – ONLINE UNIT ON DECISION ANALYSIS[[6]](#footnote-6)

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 8** (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](https://canvas.harvard.edu/courses/78089/pages/0-introduction-2?module_item_id=705071).

|  |
| --- |
| *Please enter “Done” in this field once you have completed the quiz.* |

# 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](https://canvas.harvard.edu/courses/78089) 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 in the table below:

|  |
| --- |
| *Leakage = = 1 –*  *Undercoverage= = 1 –*  *For Option 1,*  *Leakage =*  *Undercoverage =*  *For Option 2,*  *Leakage =*  *Undercoverage =*  *For Option 3,*  *Leakage =*  *Undercoverage =* |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Report the final results for this question in the below table:*   |  |  |  | | --- | --- | --- | |  | Leakage | Undercoverage | | Option 1 | 25.77% | 65.93% | | Option 2 | 43.28% | 57.19% | | Option 3 | 65.54% | 44.03% | |

1. 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 (next page)**.
2. 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.

|  |
| --- |
| *See your notes from class discussion.* |

1. Based on your assessment above, which of the three options would you recommend to the Secretary of Sedesol? Enter your answer [here](https://forms.gle/jEmW32A3L4o3zjbF6).

|  |
| --- |
| *Enter your answers in the link above. Once completed, type “Done” into this field.* |

# QUESTION 4 – 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 workbook “WDI Data Extract API-209 PS1” from our Canvas website. The workbook 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 2018 for those that may be interested.

First, familiarize yourself with the worksheet “Data Extract”. 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).

**INSTRUCTIONS ON WHAT TO SUBMIT**

* A replicable .R file with proper syntax (please copy and paste in yellow box at the end of this question). The code should follow the style guide from math camp, be able to run from beginning to end (following the restart + re-run all procedure discussed in math camp), and reproduce all the results you submit in this problem set question.
* Your answers to all sub-questions below. Please paste R output (graphs, tables, etc.) into the yellow boxes below when appropriate

**GENERAL GUIDANCE FOR QUESTIONS INVOLVING R**

* Remember to consult the R resources from math camp, particularly the HKS R cheat sheet (available at [bit.ly/HKS-R](https://docs.google.com/spreadsheets/d/1_CvPHcUlL-GJCy9hkIBwAjrIJU6KukRi4ctthHe1ulQ/edit#gid=0)), which contains many of the commands needed to answer the questions in this problem set.
* Following the style guide for your code and routinely checking your intermediate code passes the “restart + re-run” test is not only a requirement for the final code, but also an effective way to save time. Cleanly-written and well-tested code is easier to debug and build up from.
* When pasting output or code into the yellow boxes, use a monotype font like Courier New, to ensure that it will be legible to the CA grading your work.

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 2018 and gdp 2018.

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 2018 for this analysis data set.

|  |
| --- |
| *Mean: $438,901mm*  *Number of observations: 184* |

1. **Totals:** Please calculate and report:
2. Total World GDP (expressed in trillions of 2010 dollars) and world population in 2018 (expressed in billions of people)

|  |
| --- |
| 1. *Total world GDP: $80,757,820 million* 2. *Total world population: 7,373 million* |

1. Top 5 countries in terms of GDP in 2018 and their respective GDPs and top 5 countries in terms of population in 2018 and their respective populations. You may round your values.

|  |
| --- |
|  |

1. **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 2018. 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:
2. GDP per capita for the average country in 2018. [[7]](#footnote-7)

|  |
| --- |
| *$14,183* |

1. World GDP per capita in 2018 (equal to: Total world GPD / Total world population). [[8]](#footnote-8)

|  |
| --- |
| *$10,953* |

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

|  |
| --- |
| *All countries receive the same weighting for (a) regardless of size.* |

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

|  |
| --- |
| *Since countries with larger populations tend to have lower GDP per capita, the weighted average is lower than the simple average. India and China have very large populations and are probably driving a lot of the difference.* |

1. Median GDP per capita in 2018.

|  |
| --- |
| *$5,640* |

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

|  |
| --- |
| *The distribution of World GDP per capita is right skewed, which means that a few countries have a very high per capita GDP while most are clustered at moderate/low levels. Those few countries with very high per capita GDP push the mean up, far above the per capita GDP of the country that is at the center of the distribution.* |

1. The total population of all countries in 2018 with GDP per capita below the world mean calculated in part (a).

|  |
| --- |
| *6,111 million* |

1. 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]

|  |
| --- |
| *Country average GDP per capita will disproportionally represent countries that account for about 1/7 of the world population. In other words, it tells us that average per capita GDP is not a good measurement of wellbeing for the average person in the world since most people live in countries with low GDP per capita.* |

1. Draw a histogram showing the distribution of GDP per capita in 2018, 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.

|  |
| --- |
| *Yes: the data suggest that the GDP per capita of Luxembourg is around 256 times that of the DRC.* |

1. **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:
   1. 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).

|  |
| --- |
|  |

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

|  |
| --- |
| *46.7%, as shown by the following table:* |

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

|  |
| --- |
| *21/26 in Sub-Saharan Africa.* |

1. **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.
   1. The database does not have a 1993 GDP figure for how many countries?

|  |
| --- |
| *19* |

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

|  |
| --- |
| *$9,366* |

* 1. 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 2018 for these two groups.

|  |
| --- |
| *Missing data: $10,414*  *Not missing data: $14,592* |

* 1. 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.

|  |
| --- |
| *Poorer countries. Probably less developed statistical agencies in 1993.* |

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

|  |
| --- |
| *Discussion based on results above. 6/7 of the World population live in countries with average yearly incomes below $14k. Long term growth varies geographically. Half of countries in Africa are low income. Etc…* |

1. Please insert your full reproducible R script below (see example in yellow box below using Courier New font; if you discover a workflow that works well for submitting R code, please contact us and/or share with your classmates in the discussion forum)

|  |
| --- |
| 1. ibrary(tidyverse) 2. library(readxl) 3. library(scales) 4. ## for optional questions 5. library(sf) 6. library(maps) 7. library(maptools) 8. library(rgeos) 9. library(mapproj) 10. # data ----- 11. wdi\_raw <- read\_excel("WDI Data Extract API-209 - PS 1.xlsx") 12. # transform data to analysis ---- 13. wdi <- wdi\_raw %>% 14. mutate( 15. gdp\_1993 = gdp\_1993 / 1e6, 16. gdp\_2018 = gdp\_2018 / 1e6, 17. pop\_1993 = pop\_1993 / 1e6, 18. pop\_2018 = pop\_2018 / 1e6, 19. ) %>% 20. mutate(gdp\_pc\_1993 = gdp\_1993 / pop\_1993, 21. gdp\_pc\_2018 = gdp\_2018 / pop\_2018) %>% 22. filter(!is.na(gdp\_pc\_2018)) 23. # (1) ---- 24. mean(wdi$gdp\_2018, na.rm = TRUE) 25. nrow(wdi) 26. # (2) ---- 27. # a 28. sum(wdi$gdp\_2018, na.rm = TRUE) 29. sum(wdi$pop\_2018, na.rm = TRUE) 30. # b 31. arrange(wdi, desc(gdp\_2018)) %>% 32. select(country, gdp\_2018) 33. arrange(wdi, desc(pop\_2018)) %>% 34. select(country, pop\_2018) 35. # (3) ---- 36. wdi %>% 37. summarize(mean\_gdp\_pc\_2018 = mean(gdp\_pc\_2018, na.rm = TRUE), 38. tot\_gdp\_pc\_2018 = sum(gdp\_2018, na.rm = TRUE) / sum(pop\_2018, na.rm = TRUE), 39. med\_gdp\_pc\_2018 = median(gdp\_pc\_2018, na.rm = TRUE), 40. pop\_belo\_mean = sum(pop\_2018\*(scale(gdp\_pc\_2018) < 0), na.rm = TRUE)) 41. wdi %>% 42. ggplot(aes(x = gdp\_pc\_2018), aes(x = gdp\_pc\_2018)) + 43. geom\_histogram(binwidth = 2000) + 44. scale\_x\_continuous(labels = dollar) + 45. labs(x = "GDP per capita", 46. y = "Number of Countries") 47. # (4) ---- 48. # (a) Top 5 49. wdi %>% 50. mutate(growth = (gdp\_pc\_2018 / gdp\_pc\_1993) ^ (1/25) - 1) %>% 51. arrange(desc(growth)) %>% 52. select(country, growth) %>% 53. slice(1:5) 54. # (b) Bottom 5 55. wdi %>% 56. mutate(growth = (gdp\_pc\_2018 / gdp\_pc\_1993) ^ (1/25) - 1) %>% 57. arrange(growth) %>% 58. select(country, growth) %>% 59. slice(1:5) 60. # (5) ---- 61. wdi\_fct <- wdi %>% 62. add\_tally(name = "total\_n") %>% 63. mutate(income\_group = fct\_relevel(income\_group, 64. "Low income", 65. "Lower middle income", 66. "Upper middle income", 67. "High income" 68. )) 69. wdi\_fct %>% 70. group\_by(income\_group, total\_n) %>% 71. summarize(n = n()) %>% 72. mutate(prop = n/total\_n) %>% 73. select(Income = income\_group, Proportion = prop) 74. wdi\_fct %>% 75. filter(region == "Sub-Saharan Africa") %>% 76. count(income\_group) 77. # (6) ------- 78. sum(is.na(wdi$gdp\_pc\_1993)) 79. mean(wdi$gdp\_pc\_1993, na.rm = TRUE) 80. wdi %>% 81. mutate(is\_na = is.na(gdp\_1993)) %>% 82. group\_by(is\_na) %>% 83. summarize( 84. mean\_2018 = mean(gdp\_pc\_2018, na.rm = TRUE) 85. ) |

APPENDIX – HINTS FOR R QUESTION

This appendix provides hints (in green) for some of the parts of the R question that require R commands. The goal is to scaffold your learning for those of you with little or no experience with R. As the semester progresses and your mastery of R improves, less scaffolding will be provided to ensure that you have internalized the skills you learned.

1. **Explore the data set:** Create an analysis data set in which:
   * You recode the population variables so that they are expressed in millions of people
   * You recode the gdp variables so that they are expressed in millions of dollars
   * You keep only observations that have non-missing data for both population 2018 and gdp 2018.

(hint: use the mutate functions to transform the variables, and the filter function – along with is.na – to keep the observations you need. Also, note that in R, you can refer to large numbers like a million by shorthand by using the “e” (for exponent) character between two numbers. Typing in 1e6 will be interpreted as 1 times 10 to the sixth power (a million). This is often clearer than trying to type in six zeros and re-counting that you have the right number of zeros (1000000)).

Now report the mean and the number of observations for gdp 2018 for this analysis data set. (hint: use summarize command and read off the results from the Console)

1. **Totals:** Please calculate and report:
2. Total World GDP and world population in 2018 (hint: use the summarize command)
3. Top 5 countries in terms of GDP in 2018 and their respective GDPs and top 5 countries in terms of population in 2018 and their respective populations (hint: use the arrange function to create a variable that ranks countries by gdp or by population, the select function to select the variables you would like to display, and the slice function to select the observations you want to display.)
4. **Central Tendencies:** Calculate GDP per capita for each country in the database for 1993 and 2018. Then report the following statistics:
5. GDP per capita for the average country in 2018. (hint: use mutate to create new per capita variables. Then use summarize).
6. World GDP per capita in 2018 (hint: take the ratio between the total (sum) GDP and the total (sum) population.)
7. Explain what is the difference between (a) and (b) in a language that someone not well-versed in statistics can understand.
8. What do you think drives the difference between (a) and (b)?
9. Median GDP per capita in 2017. (hint: use summarize command combined with median function).
10. What do you think drives the difference between (a) and (e)?
11. The total population of all countries with GDP per capita below the world mean calculated in part (a) (hint: you can use summarize after filtering the data).
12. 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]
13. A histogram showing the distribution of GDP per capita in 2017, using bins that are $2,000 wide (hint: use the ggplot() function with the histogram geom, specifying its binwidth argument).
14. **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:
    1. 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). (hint: use the fct\_relevel command (in the “Work with vectors” section of the cheat sheet) to relevel the income\_group variable so you can list income categories in the right order.

Then group by the income\_group variable, obtain counts, and transform that to a proportion by dividing by the total number of countries. Or, instead of using group\_by, you can also make a one-way table by combining prop.table and xtabs (as suggested in the Descriptive Statistics version of the cheat sheet, i.e. prop.table(xtabs(~ race, gss\_cat)).

* 1. What percent of countries in Sub-Saharan Africa are low income? (hint: filter to get the region you want, then use hint from (a) above)

1. **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:
   1. The database does not have a 1993 GDP figure for how many countries? (hint: use sum and is.na)
   2. Calculate the mean GDP per capita in 1993 ignoring missing values. (hint: use summarize, and na.rm = TRUE)
   3. 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 2018 for these two groups. (hint: use mutate and group\_by)

1. We abide by the Harvard Kennedy School Academic code (available [here](https://www.hks.harvard.edu/educational-programs/academic-calendars-policies/student-handbook/general-regulations-and-1)) 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. [↑](#footnote-ref-1)
2. Sources: [FDA](https://www.fda.gov/consumers/consumer-updates/coronavirus-testing-basics), [Mayo clinic](https://www.mayoclinic.org/covid-antibody-tests/expert-answers/faq-20484429), [↑](#footnote-ref-2)
3. Source: [CDC](https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/antigen-tests-guidelines.html). [↑](#footnote-ref-3)
4. You can verify your results in (a) and (b) by using this nice [tool](https://graphics.wsj.com/dynamic-inset-iframer/?url=https://asset.wsj.net/wsjnewsgraphics/dice/antibody-test-reliability-calculator-433fb3e7-60e4-4bb9-aaa1-1f268cd14a0a/inset.json) that also illustrates the intuition of the test results. [↑](#footnote-ref-4)
5. Source: [The Wall Street Journal.](https://graphics.wsj.com/dynamic-inset-iframer/?url=https://asset.wsj.net/wsjnewsgraphics/dice/antibody-test-reliability-calculator-433fb3e7-60e4-4bb9-aaa1-1f268cd14a0a/inset.json) [↑](#footnote-ref-5)
6. There is one online unit in this problem set. There will be several more during the course of the semester. [↑](#footnote-ref-6)
7. This is the simple average of the GDP per capita of all countries in the database. [↑](#footnote-ref-7)
8. Another way of thinking about this statistic is that this is the mean weighted World GDP per capita in 2018, where the weight is the population of each country in 2018. [↑](#footnote-ref-8)