Problem Set 1

Due Monday September 11, 2023, EOD

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

1. Download the contents of the assignment Dropbox link in a folder (directory) on your computer designated for this problem set only. Follow the directory structure, e.g., putting the `data` folder containing the datasets as a subdirectory of the project directory.

2. Insert your answers in the yellow boxes using Microsoft Word, and prepare a single `.R` script for what you produce. Save the word document as a`.PDF`.

3. Please submit the PDF to the designated `PS-XX: pdf` link and your R Script to the `PS-XX: R` link.

(1) Your name:

Nikhilla Bhuvana Sundar

(2) Group members, if any:

N/A

(3) Compliance with the Academic Code on problem set[1] (sign with an X below)

X

[1] You may use the same code from classmates, Ed Discussion Board, instructors, and generative AI. 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, or use generative AI to provide potential code for you, but you must each type your own answers and your own code.

In this problem set, we will work with visualization and tables for a study we will read for the next 1-2 weeks:

> Banerjee, A., Duflo, E., Karlan, D. et al., 2015. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*. (Canvas reading link)

Gradescope will have pre-installed the following packages:

```
tidyverse
scales
giscoR
sf
ggthemes
```

Code Style

Throughout the course's assignments, we will be enforcing a code style. Deviation from the code style can result in point deductions. Please read through part 1 of the tidyverse style guide, chapters 1-5, focusing on the Good vs. Bad examples and the brief logic.[2] Then to summarize, *restyle the following pipelines following the style guide and paste in the re-styled version*. `flights` is a dataset of domestic flights.

You are not asked to run the code on the console – for now, focus on the form. You do not need to understand what the coding is doing exactly, although good code with good style is usually self-explanatory. Do this in your RStudio script editor so that you can easily indent the code, e.g. with the `Cmd/Ctrl` + `I` shortcut.

```
flights|>filter(carrier=="UA",dest%in%c("IAH","HOU"),sched_dep_ti
me>0900,sched_arr_time<2000)|>group_by(flight)|>summarize(delay=m
ean(arr_delay,na.rm=TRUE),cancelled=sum(is.na(arr_delay)),n=n())|
>filter(n>10)
```

```
flights |>
  filter(
    carrier == "UA",
    dest %in% c("IAH", "HOU"),
    sched_dep_time > 0900,
    sched_arr_time < 2000
    ) |>
  group_by(flight) |>
  summarize(
    delay = mean(arr_delay, na.rm = TRUE),
    cancelled = sum(is.na(arr_delay)),
    n = n()
    ) |>
  filter(n > 10)
```

[2] The most important parts of this style guide are summarized in R for Data Science (R4DS) chapter 7, but it's worth reading the actual style guide. The exercise comes from the R4DS chapter.

Problem 1

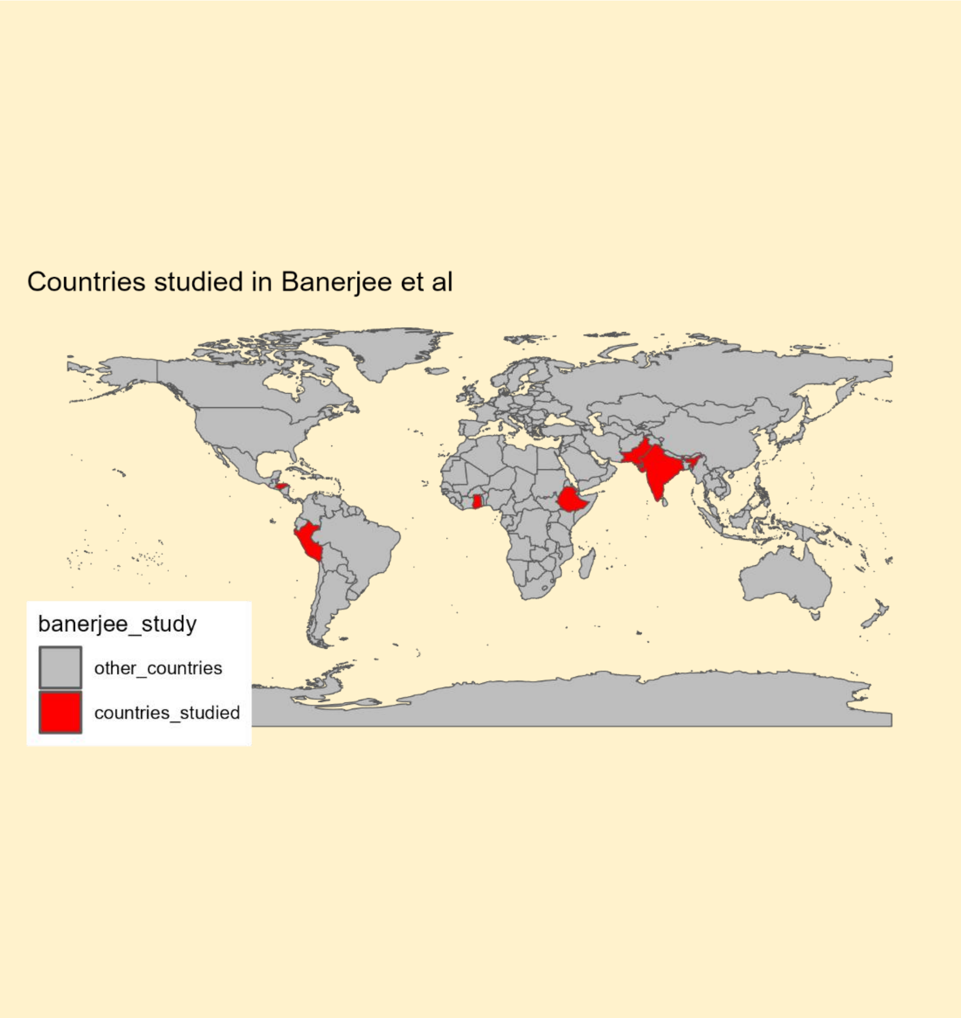
Using a shapefile obtained from `giscoR::gisco_get_countries()`, make a map of the world that highlights the six countries studied in the Banerjee et al. article: Honduras, Ghana, Pakistan, India, Ethiopia, and Peru.

In this map, manually assign colors with `scale_fill_manual()` to specify that the six countries are shown in a shade of red, and the remaining countries in gray. That is, do not leave the ggplot default as-is.

Put the png file or the screenshot of the image here.

(In this problem and subsequent problems, there will be a "Hints" section at the end with more details and tips).

You might want to use `ggthemes::theme_map()` for formatting changes that are typically good choices for maps.



Problem 2

We have provided you with a dataset from the paper in the `data` folder: `graduation_honduras.csv`. Read in the dataset by the `read_csv` function that will be loaded in `tidyverse`.[3] Assign the object name `dat` for your dataset.

This dataset has 28 variables, including the following:

Variable name	Description
`id`	Household ID
`country`	Country in which Survey took place
`hh_size`	Size of household

[3] More precisely, `read_csv` is a function from the `readr` package. That package is loaded when you load `tidyverse`, with a set of other tidyverse packages.

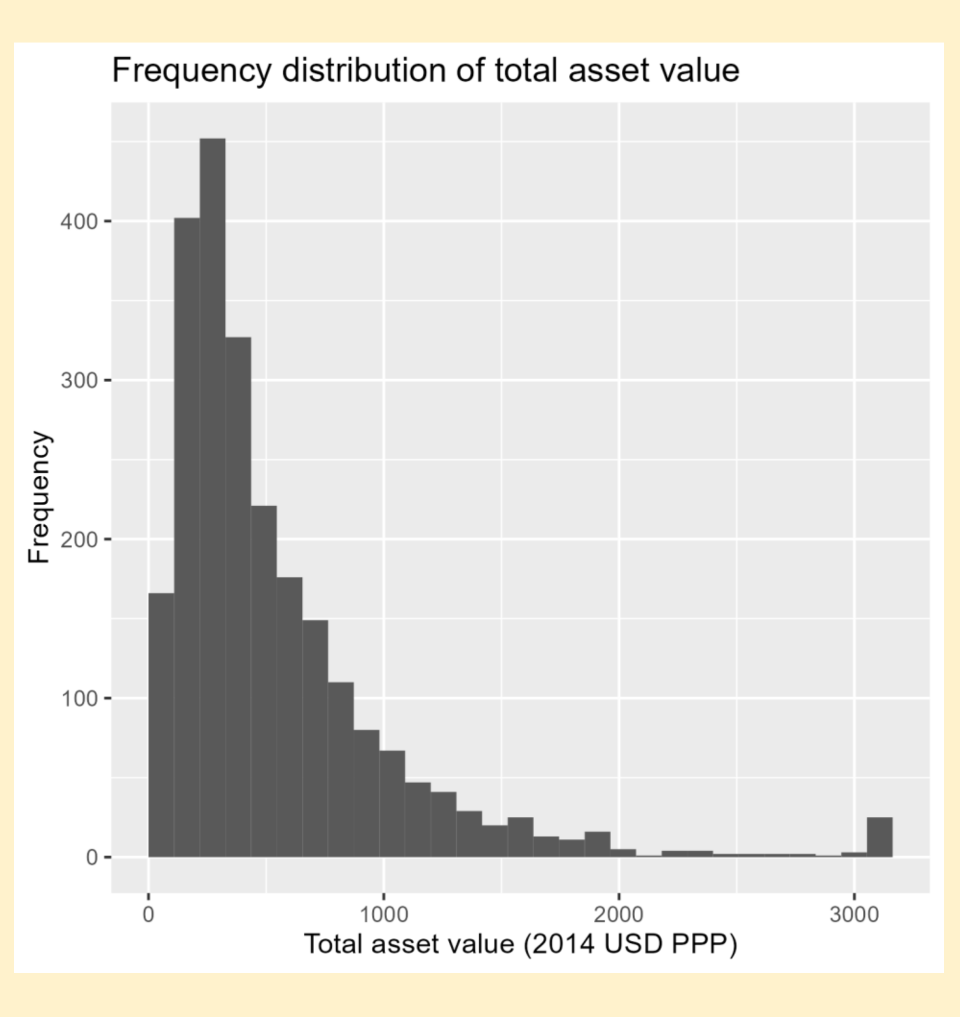
Variable name	Description
`hh_head_female`	Whether or not the household is headed by a female
`hh_education`	Whether or not the household head has a formal education
`asset_tot_value_bsl`	Total asset value (2014 USD PPP)
`asset_prod_value_bsl`	Productive asset value (2014 USD PPP)
`asset_hh_value_bsl`	Household asset value (2014 USD PPP)
`asset_index_bsl`	Total asset index
`asset_prod_index_bsl`	Productive asset index
`asset_hh_index_bsl`	Household asset index
`iagri_month_bsl`	Household Income from Agriculture (2014 USD PPP)
`ibusiness_month_bsl`	Household Income from self-operated microenterprise (2014 USD PPP)
`ipaidlabor_month_bsl`	Household Income from Paid Labor (2014 USD PPP)
`ranimals_month_bsl`	Monthly household revenue from livestock (2014 USD PPP)
`ctotal_pcmonth_bsl`	Total monthly per capita consumption (2014 USD PPP)
`cnonfood_pcmonth_bsl`	Monthly per capita non-food expenditures (2014 USD PPP)

Each row in the spreadsheet is a survey respondent representing a household in Honduras. The survey asked a series of questions about the household's finances, labor situation, and well-being, and some of them are listed here.

Create a **histogram** of the total asset value and paste a screenshot or the png in the box. Modify two things in this graph before considering it done:

(a) Label the x-axis title of your histogram as `Total asset value (2014 USD PPP)`, and

(b) In the geom, shift the location of the bars by setting an additional argument, `boundary = 0`. This makes sure that the bins of data are left-aligned rather than centered, and avoids the false visual impression that there are respondents with negative asset values.


Frequency distribution of total asset value
Frequency
400
300
200
100
0
0
1000
2000
3000
Total asset value (2014 USD PPP)

Problem 3: Summaries

3.1

The variable hh_head_female is 1 if the household is headed by a woman and 0 otherwise. What proportion of households in the sample are headed by a woman? Use the mean function, since a mean of a binary variable is a proportion.

17. 04% of the households are headed by a woman

3.2

Using the functions group_by and summarize, compute the following summary statistics for households headed by women and other households separately:

- The *proportion* of households whose household head has a formal education
- The *median* asset value
- The *mean* monthly income from business
- The *total* household income from paid labor across all households.

Summarize the difference you find in a sentence or two. Here and in all future problem sets, round your numbers when reporting results like this, as a journal article or newspaper would do.

Even though only 17. 04% of the households are headed by women, a significant proportion (read: ~69%) of the females heading these households have received a formal education. Additionally, the figure for median value of assets shows that the value is only 79.32 USD PPP higher for households headed by a male when compared to a female.

Interestingly, the average loss from business income for households headed by females is only 0.03 USD PPP and lower than losses faced by the counterpart. But when it comes to household income from paid labor, male headed households receive a total of 166918.28 USD PPP which is substantially higher than female headed households suggesting possible gender based disparities in economic opportunities which require further analysis.

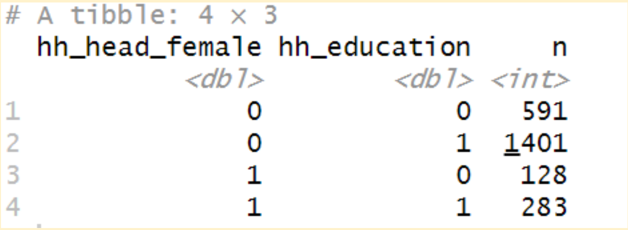
3.3

Another common statistic is simply the number of observations, or rows. dplyr has a function n() that counts the number of rows of the group it is operating in (it is a rare function that does not have any argument, but it reads from the dataset it is called in). So, summarize(n = n()) will count the number of rows in each group defined by group_by().

The `dplyr::count()` function is a shortcut for this procedure. Go to the help page of the function by `?count` and you will read that:

> "`df |> count(a, b)` is roughly equivalent to `df |> group_by(a, b) |> summarise(n = n())`"

Here, use the `count` function to count the number of observations in four groups, defined by two variables: whether the household is headed by a woman, and whether the head of the household has a formal education. Show the R tibble output as a screenshot (you do not need to format it into a clean table).

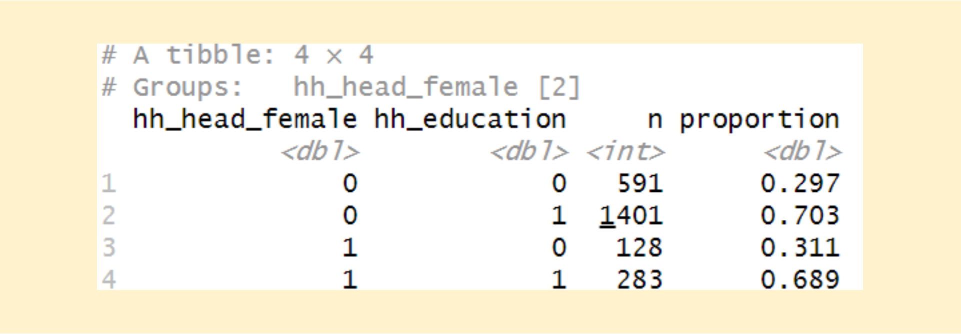


```
# A tibble: 4 × 3
  hh_head_female hh_education     n
           <dbl>        <dbl> <int>
1              0            0   591
2              0            1  1401
3              1            0   128
4              1            1   283
```

3.4

Finally, we will use `group_by` with `mutate` instead of with `summarize`. For each of the four rows, translate the count `n` into a *fraction* of the total number of households in each group of `hh_head_female`. In other words, if row 1 shows that there are X households headed by a woman with a formal education and row 2 shows that there are Y households headed by a woman without a formal education, the values of the fractions should be X/(X + Y) in row 1 and Y/(X + Y) in row 2. Verify that your answers are consistent with the output in 3.2.

Show the R tibble output as a screenshot (you do not need to format it into a clean table).



```
# A tibble: 4 × 4
# Groups:   hh_head_female [2]
  hh_head_female hh_education     n proportion
           <dbl>        <dbl> <int>      <dbl>
1              0            0   591      0.297
2              0            1  1401      0.703
3              1            0   128      0.311
4              1            1   283      0.689
```

Problem 4

Before you submit, please complete this brief survey on Canvas. Your response is important for us to adjust the course.

https://yale.instructure.com/courses/89156/quizzes/62866

Also, on Wednesday September 13th, we will be discussing the Banerjee et al. paper. Please read it, come with questions, and answer the following Canvas pre-class exercise by 10am on that Wednesday. The link includes a reading guide and a video to accompany the paper.

https://yale.instructure.com/courses/89156/quizzes/62867

Reminders Before Submitting

Remember that your R script, and not only your PDF will be graded. The autograder will check whether your code can run independently, and we will also check for compliance with the tidyverse style. Here's a checklist:

- Explicitly load the necessary libraries at the beginning of your script.
- Delete any failed attempts or duplicative code.
- Label the relevant question number by comment (e.g., `## Problem 1.1 -------`. Follow the style guide for the exact format).
- The results should be "reproducible" from the script. This means that an instructor who receives your script should be able to run it and reproduce correct answers. To preview roughly what will happen on gradescope, try restarting R (Toolbar `Session` > `Restart`) and running your entire code at once (e.g., Select All Text and Run, or `Run All` by the hot-key `option` + `command` + `R`.
- Follow other guidelines from the style guide, such as breaking up long lines and properly using spaces.

Hints

Problem 1

- Make sure you have installed the `giscoR` package first (but do not put that installation code on your script).
- Many modern packages have websites that mirror their help page, e.g. https://ropengov.github.io/giscoR/. You can take a look here first to see if any code snippets are useful. Each function also has its help page, with examples.
- Review the syntax of `%in%`. This checks, for each element of the left-hand side of the operator, whether the element is in any of the elements of the right-hand side. See the help page for `?%in%` and its examples. For example, you will see the R code `1:10 %in% c(1, 3, 5, 9)`. The single colon in `1:10` is shorthand for "all integers starting from 1 to 10", and is unrelated to the double colon used for invoking a package.
- For `scale_fill_manual()`, the help page of the function is helpful, especially the example section. Again, there is a website equivalent: https://ggplot2.tidyverse.org/reference/scale_manual.html#ref-usage. You will want to use both the `values` and `labels` option.

Problem 2

- For the histogram, follow the ggplot rstudio but use a histogram geometry, `geom_histogram()`. You can also explore all the functions through the `ggplot2` website https://ggplot2.tidyverse.org/. That website includes a more comprehensive cheatsheet.
- There are multiple ways to add a axis title. You can add `labs(x = "X axis title", y = "y axis title", title = "Figure title")` as a layer, for example.
- To specify an argument for a layer like a boundary, make sure to do it within parentheses. See the examples by executing `?geom_histogram` on your Console.

Problem 3

- R4DS (2e) 4.5 https://r4ds.hadley.nz/data-transform.html#sec-summarize gives some good examples on how to compute summary statistics by groups.

3.4

- Continue to use the output from 3.3.
- The difference between `mutate` and `summarize` in this context is that `summarize` should always produce one row per group, while it is ok for `mutate` to produce more than one row per group. The groupings here should be defined by the `hh_head_female` variable.