MPA/ID Math Camp 2022: Day #4

2022-06-27

MPA/ID Math Camp 2022: Day #4

Goals:

1. Gain additional practice **reading and manipulating datasets** (`read_csv()` , `dplyr` functions, etc.).
2. Practice **data exploration** functions (e.g. `unique()`, `table()`, `mean()`, etc.).
3. Calculate **group-level statistics** (`group_by()` and `summarise()`).
4. Practice **data visualization** with `ggplot2`.

International Education

Today, we are going to explore data on international education levels. Our source will be data from the **Unicef Multiple Indicator Cluster Surveys** (MICS), a source of internationally comparable surveys on children and women conducted in over 115 countries. A core use of MICS data is to measure progress towards the UN's Sustainable Development Goals, with more information available here. The MICS data is also regularly used for academic and policy work, including this recent paper by two MPA/ID alums! We thank Marla Spivack and Jason Silberstein for sharing their cleaned data with all of us. You can read more about the data source at the Unicef website here, or even look through the codebook for the MICS data here.

Question 1 First, load the `tidyverse` library and read the dataset, called "misc6.csv" in the `data` folder. Assign this data to an object with a name of your choice. The MICS data is a survey administered to children living in each country, so the rows in this dataset are individuals living in that country. Of course, these rows are not exhaustive of every child living in the country, but they are the "sample" of respondents. We will learn much more about sampling in API-209.

In the RStudio console (not the .Rmd file), use the `View()` function to look around the dataset. `View()` is an extremely useful function for looking at data. However, in general we suggest **not** putting `View()` directly in your .Rmd file, because R will try to open a window to visualize the dataset every time you knit the file.

```
library(tidyverse)

df <- read_csv("data/mics6.csv")
```

Question 2 We are working with a subset of the MICS6 data from a few randomly selected countries. How many different countries are in the `country` column? Remember, outside of a tidyverse "pipe" (`%>%`), you can access the values in a particular column from a dataset called `df` with `df$country`.

Hint: like with most programming tasks, there are many possible ways to do this. Feel free to be creative and / or come up with multiple answers.

```
unique(df$country)
```

```
## [1] "Ghana"        "Kosovo"       "Mongolia"     "Nepal"        "Turkmenistan"
## [6] "Zimbabwe"
```

```
table(df$country)
```

```
##
##        Ghana       Kosovo     Mongolia        Nepal Turkmenistan     Zimbabwe
##         5153         1006         4100         4077         2019         3848
```

Question 3 The MICS survey studies many health and education outcomes. One skill they measure is called `numeracy`, which determines a respondent's ability to, for example, read numbers aloud, compare magnitudes, and perform calculations like addition. In this survey, the `numeracy` column is a binary measure(either 1 or 0) indicating whether the respondent passed a skill-based test of numeracy.

Find the proportion of respondents in the full dataset who are "numerate" (i.e. have an `numeracy` value equal to 1). Write a sentence in text to interpret your result.

Hint: with a binary measure, you can calculate a proportion (i.e. the fraction of respondents who reported "yes") by taking the average.

Hint #2: your initial answer might not be very informative. Remember the summer assignment and try to remove NA values from your calculation. If you are stuck, try reading the documentation for the function you're trying to use by running `?function_name` in the console.

```
mean(df$numeracy, na.rm = TRUE)
```

```
## [1] 0.2975283
```

Question 4 Create a new dataset by filtering the data to observations from one country of your choice from Question 2. Then, calculate the average `numeracy` only within that one country.

```
# Option 1
mongolia <- df %>% filter(country == "mongolia")
mean(mongolia$numeracy, na.rm = TRUE)
```

```
## [1] NaN
```

```
# Option 2
df %>%
  filter(country == "mongolia") %>%
  summarise(avg = mean(numeracy, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##     avg
##   <dbl>
## 1   NaN
```

Question 5 Using `group_by()` and `summarise()` together, calculate the average `numeracy` in each country from the MISC6 dataset. For a reminder on how these functions work, feel free to consult Notebook #4 from the Summer Assignment, use Google, or ask a group member / teaching staff.

```
df %>%
  group_by(country) %>%
  summarise(avg = mean(numeracy, na.rm = TRUE))
```

```
## # A tibble: 6 x 2
##   country        avg
##   <chr>        <dbl>
## 1 Ghana        0.119
```

```
## 2 Kosovo       0.416
## 3 Mongolia     0.465
## 4 Nepal        0.253
## 5 Turkmenistan 0.661
## 6 Zimbabwe     0.179
```

Question 6 Finally, use the averages in each country from Question 5 to design a visualization showing the average `numeracy` in each country among the MISC6 surveyed children. Feel free to be creative, but we recommend a barplot (`geom_col()`) as a place to start. We encourage you to make your visualization appealing by practicing the various tools from the Summer Assignment. (e.g. labels and themes from Notebook #3).

```
df %>%
  group_by(country) %>%
  summarise(avg = mean(numeracy, na.rm = TRUE)) %>%
  ggplot(aes(x = country, y = avg)) +
    geom_col() +
    labs(title = "Proportion of Children with Foundational Numeracy",
         caption = "Source: MISC6 Data",
         x = "Country",
         y = "Average Numeracy") +
    theme_bw()
```