CHSE Code Challenge

Philip Nguyen

2022-08-30

Background and Data

We are conducting a comparative study of the health care workforce in two states: **Idaho and Hawaii**. As part of this study, you are comparing the availability of providers with a primary specialty of **Emergency Medicine (EM) or Obstetrics and Gynecology (OBGYN)** in the two states. To do so, you have obtained a database from a nationwide provider directory.

Details about the database are provided in the attached data dictionary. Note that providers can declare multiple specialties, but each provider can choose only one primary specialty. For example, a provider might declare both Family Medicine and OBGYN as specializations, but indicate that Family Medicine is their primary specialization. Please provide the code you use to calculate the summary statistics and extract outlined below along with any assumptions you made.

Instructions

Your role is to validate the NPI (National Provider Identifier) data and prepare an extract for analysis. For this study, your team is interested in **individual physicians only**; do not include organizations in results. Furthermore, exclude all records known to be updated prior to **January 1, 2008**.

---

0) Data Exploration, Cleaning, and Formatting

Before computing any summary statistics, I will explore the data to understand how it is structured and to address any outstanding issues that may affect the computed statistics such as data missingness. First let's load in the data and look at the first few rows using R.

```
#|
# load in relevant libraries
library(tidyverse)
library(lubridate)
library(knitr)
library(kableExtra)

# read in data
npi_data <- vroom::vroom("../dataset.csv")

# look at first few rows of data
npi_data %>%
  head() %>%
  kable(align = "l",
        booktabs = T) %>%
  kable_styling(bootstrap_options = c("striped"),
                position = "left",
                full_width = F, font_size = 12)
```

NPI	provider_state	HealthcareProviderTaxonomyCode_1	HealthcareProviderTaxonomyC
1396749347	HI	207N00000X	207N00000X
1124022801	ID	207N00000X	207N00000X
1801891049	VT	207N00000X	NA
1801891569	ID	174400000X	207V00000X
1598761306	VT	207V00000X	NA
1215933940	ID	207N00000X	NA

Right away we notice that the data are stored in a wide format. Each provider can have multiple medical specialties – up to 15 in this dataset, each separated by an individual medical specialty code in its own column. Additionally, each medical specialty code has a corresponding column that determines whether or not a specialty is the provider's primary specialty. It will be easier to compute summary statistics later if the data are in a longer format based on these columns, which can be done relatively easily.

Prior to doing so, let's get a sense of which columns are potentially missing values using the `skim` function from the `skimr` package.

```
npi_data %>%
  skimr::skim() %>%
  select(skim_type, skim_variable, n_missing) %>%
  arrange(skim_variable) %>%
  kable(align = "l", booktabs = T) %>%
  kable_styling(bootstrap_options = c("striped"),
                position = "left",
                full_width = F,  font_size = 12)
```

skim_type	skim_variable	n_missing
character	entity_type	0
character	HealthcareProviderPrimaryTaxonomySwitch_1	0
logical	HealthcareProviderPrimaryTaxonomySwitch_10	1966
logical	HealthcareProviderPrimaryTaxonomySwitch_11	1966
logical	HealthcareProviderPrimaryTaxonomySwitch_12	1966
logical	HealthcareProviderPrimaryTaxonomySwitch_13	1966
logical	HealthcareProviderPrimaryTaxonomySwitch_14	1966
logical	HealthcareProviderPrimaryTaxonomySwitch_15	1966
character	HealthcareProviderPrimaryTaxonomySwitch_2	1548
character	HealthcareProviderPrimaryTaxonomySwitch_3	1867
character	HealthcareProviderPrimaryTaxonomySwitch_4	1945
logical	HealthcareProviderPrimaryTaxonomySwitch_5	1966
logical	HealthcareProviderPrimaryTaxonomySwitch_6	1966
logical	HealthcareProviderPrimaryTaxonomySwitch_7	1966
logical	HealthcareProviderPrimaryTaxonomySwitch_8	1966
logical	HealthcareProviderPrimaryTaxonomySwitch_9	1966
character	HealthcareProviderTaxonomyCode_1	0
logical	HealthcareProviderTaxonomyCode_10	1966
logical	HealthcareProviderTaxonomyCode_11	1966
logical	HealthcareProviderTaxonomyCode_12	1966
logical	HealthcareProviderTaxonomyCode_13	1966
logical	HealthcareProviderTaxonomyCode_14	1966
logical	HealthcareProviderTaxonomyCode_15	1966
character	HealthcareProviderTaxonomyCode_2	1548
character	HealthcareProviderTaxonomyCode_3	1867
character	HealthcareProviderTaxonomyCode_4	1945
logical	HealthcareProviderTaxonomyCode_5	1966
logical	HealthcareProviderTaxonomyCode_6	1966
logical	HealthcareProviderTaxonomyCode_7	1966
logical	HealthcareProviderTaxonomyCode_8	1966
logical	HealthcareProviderTaxonomyCode_9	1966
Date	last_updated	0
numeric	NPI	0
numeric	provider_age	311
character	provider_gender	311
character	provider_state	0

The bulk of the missing values comes from the columns that indicate the additional specialties a provider might have. No provider IDs (NPI) or last update dates are missing. 311 values are missing for provider age and gender, which may correspond to entities that are organizations rather than individual providers. This may not be an issue since we are only concerned with the latter. Overall there doesn't appear to be any huge red flags with the data as it stands.

Let's now filter the data to include only **individual** providers from **Hawaii or Idaho** that have updated their information since the start of **2008**.

```
# filter by state, entity type, and date of last update

HI_ID_wide <- npi_data %>%
  filter(provider_state == "ID" |
           provider_state == "HI") %>%
  filter(entity_type == "individual") %>%
  filter(year(last_updated) >= 2008)

# skim again to check for missingness in date, gender, age

HI_ID_wide %>%
  select(NPI, last_updated, provider_age, provider_gender) %>%
  skimr::skim() %>%
  select(skim_type, skim_variable, n_missing) %>%
  arrange(skim_variable) %>%
  kable(align = "l", booktabs = T) %>%
  kable_styling(bootstrap_options = c("striped"),
                position = "left",
                full_width = F,  font_size = 12)
```

skim_type	skim_variable	n_missing
Date	last_updated	0
numeric	NPI	0
numeric	provider_age	0
character	provider_gender	0

Now there are no missing values in age and gender. Let's now transpose the data into a longer format so that medical specialty and whether or not they are the provider's primary specialty are contained within two columns rather than split across 30 columns. Then we will filter based on specialty code (include only emergency medicine and OBGYN), and select providers that have either specialty as their primary specialty. This should yield the desired dataset since providers can only list one primary specialty.

```r
# pivot table
HI_ID_long <- HI_ID_wide %>%
  pivot_longer(cols = contains("HealthcareProvider"),
               names_to = c(".value", "pair_number"),
               names_sep = "_") %>%
  filter(HealthcareProviderTaxonomyCode == "207P00000X" |
           HealthcareProviderTaxonomyCode == "207V00000X") %>%
  filter(HealthcareProviderPrimaryTaxonomySwitch == "Y")

# look at first few rows
HI_ID_long %>%
  select(-pair_number) %>%
  head() %>%
  kable(align = "l", booktabs = T) %>%
  kable_styling(bootstrap_options = c("striped"),
                position = "left",
                full_width = F,  font_size = 12)
```

NPI	provider_state	last_updated	entity_type	provider_gender	provider_age	Health
1801891569	ID	2008-09-02	individual	female	47	207V0
1710987664	HI	2009-07-27	individual	male	61	207P0
1275533184	ID	2014-08-07	individual	male	49	207V0
1386644318	ID	2012-03-01	individual	male	51	207V0
1497757223	HI	2011-09-15	individual	male	62	207V0
1427041730	ID	2016-01-28	individual	male	55	207P0

We're left with 505 providers after subsetting the data based on the desired specifications. A quick check also shows that there are no duplicate providers, meaning the pivoting and filtering worked as planned. Let's now compute the summary statistics.

1) Summary Statistics

Calculate the following summary statistics for providers with a primary specialty of EM versus OBGYN, in the two states of interest (Hawaii and Idaho). Please present the results in a table.

a. Total number of providers
b. Percentage of providers who are female
c. Mean and standard deviation of age.

Describe the steps that you take to validate the data and any concerns that you find.

a) Total number of providers

```
HI_ID_long %>%
  group_by(provider_state) %>%
  summarize(n=n()) %>%
  rename(`Provider State` = provider_state) %>%
  kable(align = "l", booktabs = T) %>%
  kable_styling(bootstrap_options = c("striped"),
                position = "left",
                full_width = F, font_size = 12)
```

Provider State	n
HI	256
ID	249

In total we have 505 providers: 256 in Hawaii and 249 in Idaho.

b) Percentage of providers who are female

```
HI_ID_long %>%
  group_by(provider_gender) %>%
  summarize(n=n()) %>%
  mutate(percentage = round(n/sum(n), 2)) %>%
  rename(`Provider Gender` = provider_gender) %>%
  kable(align = "l", booktabs = T) %>%
  kable_styling(bootstrap_options = c("striped"),
                position = "left",
                full_width = F, font_size = 12)
```

Provider Gender	n	percentage
female	195	0.39
male	310	0.61

Between Idaho and Hawaii, 39% of emergency medicine or OBGYN primary providers are female. We can also look at the proportions by state:

```
HI_ID_long %>%
  group_by(provider_state, provider_gender) %>%
  summarize(n=n()) %>%
  mutate(percentage = round(n/sum(n), 2)) %>%
  rename(`Provider State` = provider_state,
         `Provider Gender` = provider_gender) %>%
  kable(align = "l", booktabs = T) %>%
  kable_styling(bootstrap_options = c("striped"),
                position = "left",
                full_width = F, font_size = 12)
```

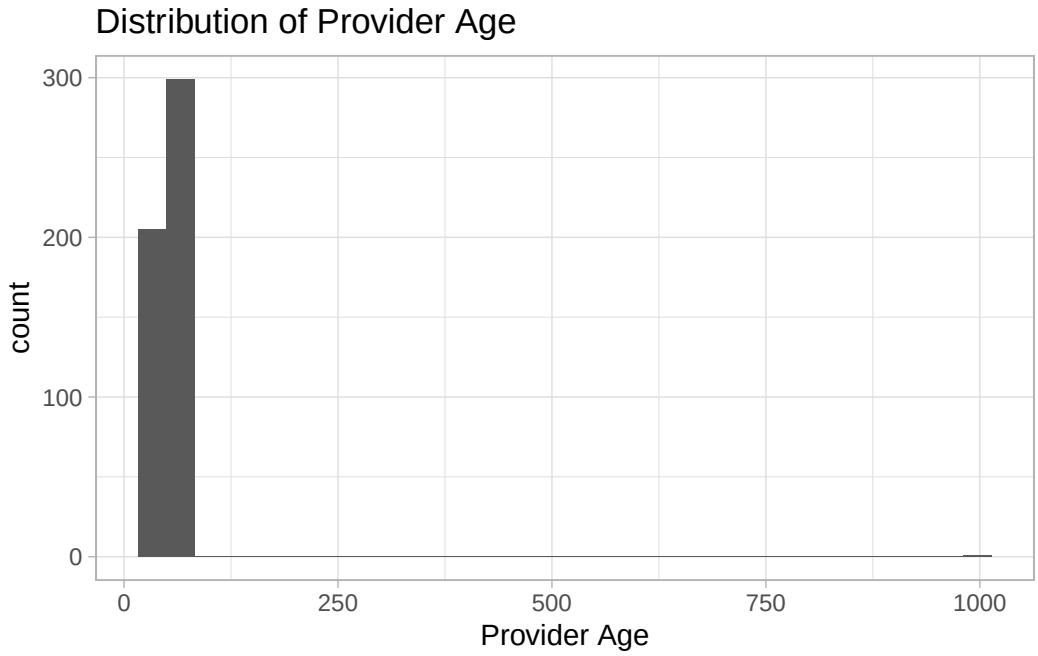
Provider State	Provider Gender	n	percentage
HI	female	115	0.45
HI	male	141	0.55
ID	female	80	0.32
ID	male	169	0.68

We see that there is a more balanced proportion of gender for emergency medicine or OBGYN primary providers in Hawaii than in Idaho.

c) Mean and standard deviation of age

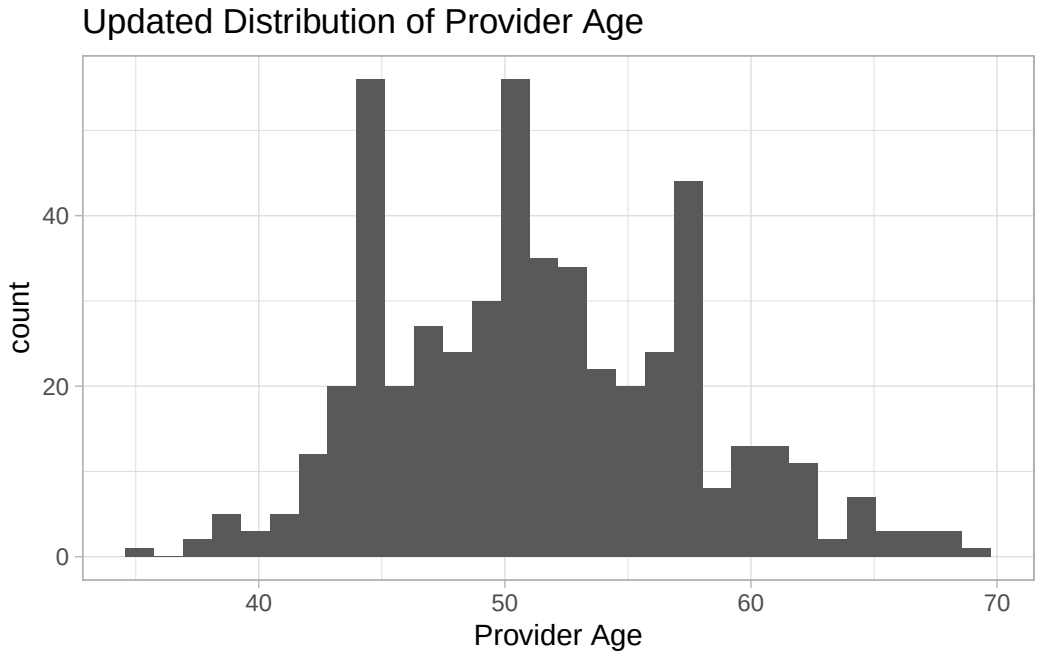
Prior to computing the mean and standard deviation, let's plot the distribution of age since the mean is sensitive to outlier values.

```
HI_ID_long %>%
  ggplot(aes(x = provider_age)) +
  geom_histogram() +
  theme_light() +
  labs(x = "Provider Age") +
  ggtitle("Distribution of Provider Age")
```



As the histogram shows, there is a provider that has listed their age to be 1000 years old. If this provider were really 1000 years old, further investigation would be warranted (What is their secret? Perhaps this is a ploy so that the provider receives more patients in the quest for everlasting longevity), but for now, we will filter out any providers over the age of 100 for the sake of computing our mean and standard deviation.

```
HI_ID_long %>%
  filter(provider_age <= 100) %>%
  ggplot(aes(x = provider_age)) +
  geom_histogram() +
  theme_light() +
  labs(x = "Provider Age") +
  ggtitle("Updated Distribution of Provider Age")
```



This looks much better. Let's now compute the mean and standard deviation.

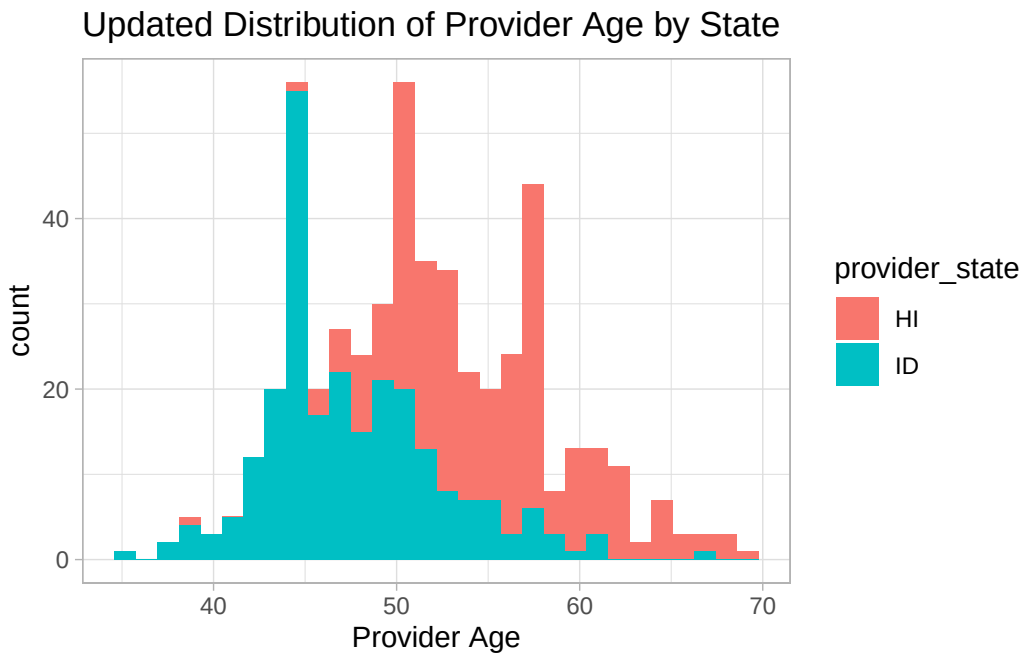
```
HI_ID_long %>%
  filter(provider_age <= 100) %>%
  summarize(`Mean Provider Age` = round(mean(provider_age), 2),
            `SD Provider Age` = round(sd(provider_age), 2)) %>%
  kable(align = "l", booktabs = T) %>%
  kable_styling(bootstrap_options = c("striped"),
                position = "left",
                full_width = F, font_size = 12)
```

Mean Provider Age	SD Provider Age
51.35	6.24

The mean provider age is approximately 51 years old, with a standard deviation of 6.24 years. We can also compute the mean and standard deviation by state:

```
HI_ID_long %>%
  filter(provider_age <= 100) %>%
  ggplot(aes(x = provider_age, fill = provider_state)) +
  geom_histogram() +
  theme_light() +
```

```
  labs(x = "Provider Age") +
  ggtitle("Updated Distribution of Provider Age by State")
```




```
HI_ID_long %>%
  filter(provider_age <= 100) %>%
  group_by(provider_state) %>%
  summarize(`Mean Provider Age` = round(mean(provider_age), 2),
            `SD Provider Age` = round(sd(provider_age), 2)) %>%
  rename(`Provider State` = provider_state) %>%
  kable(align = "l", booktabs = T) %>%
  kable_styling(bootstrap_options = c("striped"),
                position = "left",
                full_width = F, font_size = 12)
```

Provider State	Mean Provider Age	SD Provider Age
HI	55.05	5.03
ID	47.55	4.94

Idaho seems to have younger providers with a mean provider age of approximately 47 years old versus 55 in Hawaii, with about the same amount of variation in both states.

2) Data Extract

> Create an extract from the **dataset.csv** table with all of the records that meet the following criteria: - Have a primary specialty of Emergency Medicine (EM) or Obstetrics & Gynecology (OBGYN) - Practice in the states of Hawaii (HI) or Idaho (ID) - Have a record update of January 1, 2008 and after - Are individual providers (not organizations)

Since we have been working with a filtered dataset that meets the above specification, we can simply use the `readr::write_csv()` function to create the data extract.

```
# export filtered data as csv file

HI_ID_long %>%
  select(-pair_number) %>%
  write_csv("data_extract.csv")
```

That concludes the code challenge!