Problem set 4

2025-10-05

In the next problem set, we plan to explore the relationship between COVID-19 death rates and vaccination rates across US states by visually examining their correlation. This analysis will involve gathering COVID-19 related data from the CDC's API and then extensively processing it to merge the various datasets. Since the population sizes of states vary significantly, we will focus on comparing rates rather than absolute numbers. To facilitate this, we will also source population data from the US Census to accurately calculate these rates.

In this problem set we will learn how to extract and wrangle data from the data US Census and CDC APIs.

1. Get an API key from the US Census at https://api.census.gov/data/key_signup.html. You can't share this public key. But your code has to run on a TFs computer. Assume the TF will have a file in their working directory named `census-key.R` with the following one line of code:

```
census_key <- "A_CENSUS_KEY_THAT_WORKS"
```

Write a first line of code for your problem set that defines `census_key` by running the code in the file `census-key.R`.

```
source("census-key.R")
```

2. The US Census API User Guide provides details on how to leverage this valuable resource. We are interested in vintage population estimates for years 2021 and 2022. From the documentation we find that the *endpoint* is:

```
url <- "https://api.census.gov/data/2021/pep/population"
```

Use the **httr2** package to construct the following GET request.

```
https://api.census.gov/data/2021/pep/population?get=POP_2020,POP_2021,NAME&for=state:*&key=Y
```

Create an object called `request` of class `httr2_request` with this URL as an endpoint. Hint: Print out `request` to check that the URL matches what we want.

```
library(httr2)
request <- request(url) |>
  req_url_query(
    get = "POP_2020,POP_2021,NAME",
    `for` = "state:*",
    key = census_key
  )
request
```

```
<httr2_request>
GET https://api.census.gov/data/2021/pep/population?get=POP_2020%2CPOP_2021%2CNAME&for=state%
Body: empty
```

3. Make a request to the US Census API using the `request` object. Save the response to and object named `response`. Check the response status of your request and make sure it was successful. You can learn about *status codes* here.

```
response <- req_perform(request)
resp_status(response)
```

```
[1] 200
```

4. Use a function from the **httr2** package to determine the content type of your response.

```
content_type <- resp_content_type(response)
content_type
```

```
[1] "application/json"
```

5. Use just one line of code and one function to extract the data into a matrix. Hints: 1) Use the `resp_body_json` function. 2) The first row of the matrix will be the variable names and this OK as we will fix in the next exercise.

```
population <- response |> resp_body_json(simplifyVector = TRUE)
```

6. Examine the `population` matrix you just created. Notice that 1) it is not tidy, 2) the column types are not what we want, and 3) the first row is a header. Convert `population` to a tidy dataset. Remove the state ID column and change the name of the column with state names to `state_name`. Add a column with state abbreviations called `state`. Make sure you assign the abbreviations for DC and PR correctly. Hint: Use the **janitor** package to make the first row the header.

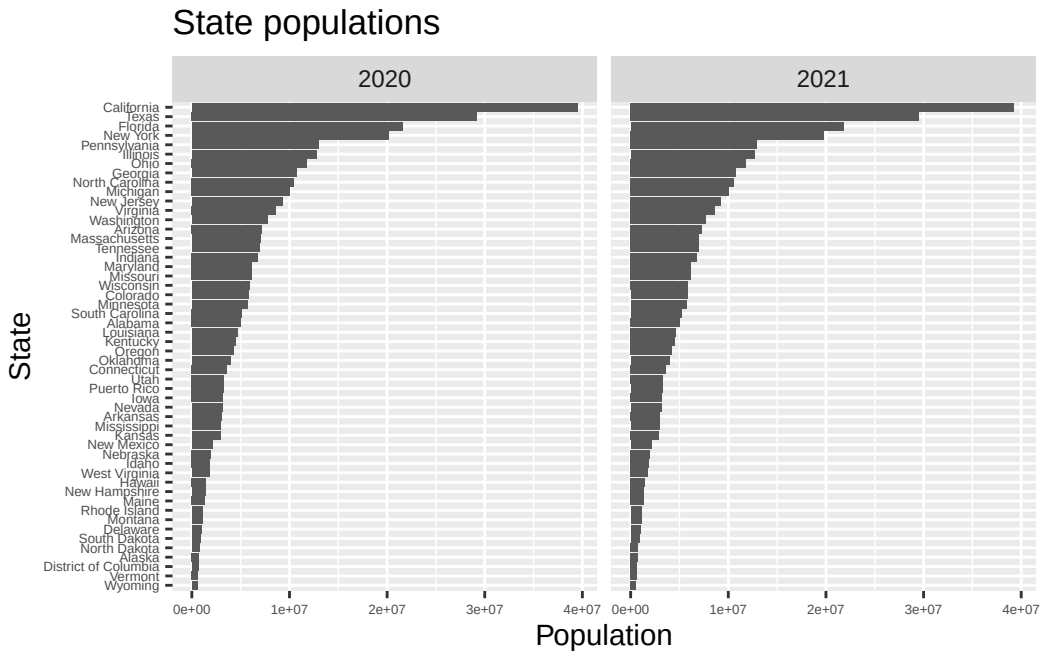
```
library(tidyverse)
library(janitor)

population <- population |>
  as_tibble(.name_repair = "minimal") |>  # convert to tibble
  row_to_names(row_number = 1) |> # Use janitor row to names function
  select(-state) |> # remove stat column
  rename(state_name = NAME)|> # rename state column to state_name
  pivot_longer(
    cols = c(POP_2020, POP_2021),   # use pivot_longer to tidy
    names_to = "year",
    values_to = "population"
  ) |>
  mutate(
    year = as.integer(sub("POP_", "", year)), # remove POP_ from year
    population = as.numeric(population) # parese all relevant columns to numeric
  ) |>
  left_join(tibble(state_name = state.name, state = state.abb),
            by = "state_name") |> # add state abbreviations using state.abb variable
  mutate(  # use case_when to add abbreviations for DC and PR
    state = case_when(
      state_name == "District of Columbia" ~ "DC",
      state_name == "Puerto Rico"          ~ "PR",
      TRUE                                 ~ state
    )
  )

population
```

```
# A tibble: 104 x 4
   state_name    year population state
   <chr>        <int>      <dbl> <chr>
 1 Oklahoma      2020    3962031 OK
 2 Oklahoma      2021    3986639 OK
 3 Nebraska      2020    1961455 NE
```

```
 4 Nebraska       2021    1963692 NE
 5 Hawaii         2020    1451911 HI
 6 Hawaii         2021    1441553 HI
 7 South Dakota   2020     887099 SD
 8 South Dakota   2021     895376 SD
 9 Tennessee      2020    6920119 TN
10 Tennessee      2021    6975218 TN
# i 94 more rows
```

7. As a check, make a barplot of states' 2020 and 2021 populations. Show the state names in the y-axis ordered by population size. Hint: You will need to use `reorder` and use `facet_wrap`.

```
population |>
  group_by(year) |>
  mutate(state_order = reorder(state_name, population)) |>
  ungroup() |>
  ggplot(aes(x = population, y = state_order)) +
  geom_col() +
  facet_wrap(~ year) +
  labs(title = "State populations", x = "Population", y = "State") +
  theme(axis.text  = element_text(size = 5))
```



8. The following URL:

```
url <- "https://github.com/datasciencelabs/2025/raw/refs/heads/main/data/regions.json"
```

points to a JSON file that lists the states in the 10 Public Health Service (PHS) defined by CDC. We want to add these regions to the `population` dataset. To facilitate this create a data frame called `regions` that has two columns `state_name`, `region`, `region_name`. One of the regions has a long name. Change it to something shorter.

```
library(jsonlite)
library(purrr)
url <- "https://github.com/datasciencelabs/2025/raw/refs/heads/main/data/regions.json"
regions_list <- fromJSON(url, simplifyVector = FALSE)

regions <- map_df(regions_list, function(region_item) {
  tibble(
    state_name = unlist(region_item$states),
    region = region_item$region[[1]],
    region_name = region_item$region_name
  )
})

regions <- regions |>
  filter(state_name %in% c(state.name, "District of Columbia", "Puerto Rico")) |>
  mutate(region_name = ifelse(region_name ==
    "New York and New Jersey, Puerto Rico, Virgin Islands",
    "NY/NJ & Territories",
    region_name))

regions
```

```
# A tibble: 52 x 3
   state_name    region region_name
   <chr>          <int> <chr>
 1 Connecticut        1 New England
 2 Maine              1 New England
 3 Massachusetts      1 New England
 4 New Hampshire      1 New England
 5 Rhode Island       1 New England
 6 Vermont            1 New England
 7 New Jersey         2 NY/NJ & Territories
 8 New York           2 NY/NJ & Territories
 9 Puerto Rico        2 NY/NJ & Territories
```

```
10 Delaware           3 Mid-Atlantic
# i 42 more rows
```

9. Add a region and region name columns to the `population` data frame.

```
population <- population |>
  left_join(regions, by = "state_name")
population
```

```
# A tibble: 104 x 6
   state_name    year population state region region_name
   <chr>        <int>      <dbl> <chr>  <int> <chr>
 1 Oklahoma      2020    3962031 OK         6 South Central
 2 Oklahoma      2021    3986639 OK         6 South Central
 3 Nebraska      2020    1961455 NE         7 Central Plains
 4 Nebraska      2021    1963692 NE         7 Central Plains
 5 Hawaii        2020    1451911 HI         9 Pacific
 6 Hawaii        2021    1441553 HI         9 Pacific
 7 South Dakota  2020     887099 SD         8 Mountain States
 8 South Dakota  2021     895376 SD         8 Mountain States
 9 Tennessee     2020    6920119 TN         4 Southeast
10 Tennessee     2021    6975218 TN         4 Southeast
# i 94 more rows
```

10. From reading https://data.cdc.gov/ we learn the endpoint `https://data.cdc.gov/resource/pwn4-m3yp.` provides state level data from SARS-COV2 cases. Use the **httr2** tools you have learned to download this into a data frame. Is all the data there? If not, comment on why.

```
api <- "https://data.cdc.gov/resource/pwn4-m3yp.json"

cases_raw <- request(api) |>
  req_perform() |>
  resp_body_json(simplifyVector = TRUE) |>
  as_tibble()

cases_raw
```

```
# A tibble: 1,000 x 10
   date_updated         state start_date end_date tot_cases new_cases tot_deaths
   <chr>                <chr> <chr>      <chr>    <chr>     <chr>     <chr>
 1 2023-02-23T00:00:00~ AZ    2023-02-1~ 2023-02~ 2434631.0 3716.0    33042.0
 2 2022-12-22T00:00:00~ LA    2022-12-1~ 2022-12~ 1507707.0 4041.0    18345.0
```

```
 3 2023-02-23T00:00:00~ GA    2023-02-1~ 2023-02~ 3061141.0 5298.0    42324.0
 4 2023-03-30T00:00:00~ LA    2023-03-2~ 2023-03~ 1588259.0 2203.0    18858.0
 5 2023-02-02T00:00:00~ LA    2023-01-2~ 2023-02~ 1548508.0 5725.0    18572.0
 6 2023-03-23T00:00:00~ LA    2023-03-1~ 2023-03~ 1580709.0 1961.0    18835.0
 7 2023-04-27T00:00:00~ LA    2023-04-2~ 2023-04~ 1597070.0 1884.0    18937.0
 8 2023-03-16T00:00:00~ NV    2023-03-0~ 2023-03~ 891702.0  1233.0    11937.0
 9 2023-05-11T00:00:00~ FL    2023-05-0~ 2023-05~ 7572282.0 6937.0    88248.0
10 2022-10-27T00:00:00~ NYC   2022-10-2~ 2022-10~ 2928439.0 14590.0   42863.0
# i 990 more rows
# i 3 more variables: new_deaths <chr>, new_historic_cases <chr>,
#   new_historic_deaths <chr>
```

It does not return all data but only gives the first 1,000 rows because of request().

We see exactly 1,000 rows. We should be seeing over 52×3 rows per state.

11. The reason you see exactly 1,000 rows is because CDC has a default limit. You can change this limit by adding `$limit=10000000000` to the request. Rewrite the previous request to ensure that you receive all the data. Then wrangle the resulting data frame to produce a data frame with columns `state`, `date` (should be the end date) and `cases`. Make sure the cases are numeric and the dates are in `Date` ISO-8601 format.

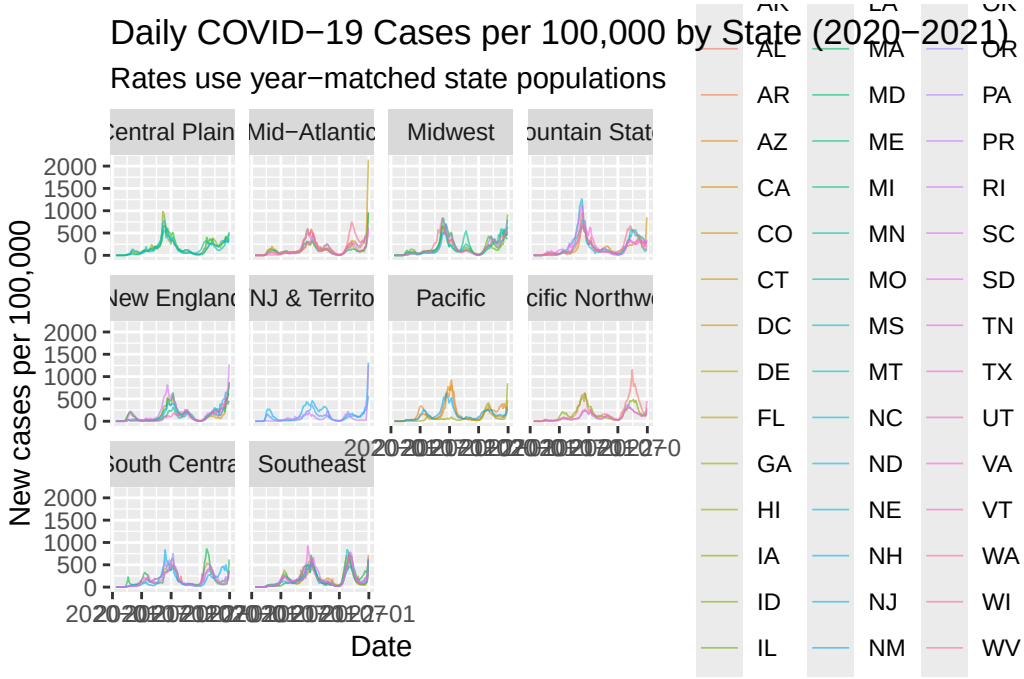
```
api <- "https://data.cdc.gov/resource/pwn4-m3yp.json"
cases <- request("https://data.cdc.gov/resource/pwn4-m3yp.json") |>
  req_url_query(`$limit` = 10000000000) |>
  req_perform() |>
  resp_body_json(simplifyVector = TRUE) |>
  as_tibble() |>
  transmute(
    state,
    date  = as.Date(end_date),
    cases = suppressWarnings(as.numeric(new_cases))
  ) |>
  arrange(state, date)
cases
```

```
# A tibble: 10,380 x 3
   state date       cases
   <chr> <date>     <dbl>
 1 AK    2020-01-22     0
 2 AK    2020-01-29     0
 3 AK    2020-02-05     0
 4 AK    2020-02-12     0
```

```
 5 AK    2020-02-19     0
 6 AK    2020-02-26     0
 7 AK    2020-03-04     0
 8 AK    2020-03-11     0
 9 AK    2020-03-18    11
10 AK    2020-03-25    52
# i 10,370 more rows
```

12. For 2020 and 2021, make a time series plot of cases per 100,000 versus time for each state. Stratify the plot by region name. Make sure to label you graph appropriately.

```
cases |>
  filter(between(date, as.Date("2020-01-01"), as.Date("2021-12-31"))) |>
  mutate(year = year(date)) |>
  left_join(population, by = c("state","year")) |>
  filter(!is.na(.data$population), !is.na(.data$region_name)) |>
  mutate(cases_per_100k = cases / .data$population * 1e5) |>
  ggplot(aes(x = date, y = cases_per_100k, color = state)) +
  geom_line(alpha = 0.6, linewidth = 0.3) +
  facet_wrap(~ region_name,) +
  labs(
    title = "Daily COVID-19 Cases per 100,000 by State (2020-2021)",
    subtitle = "Rates use year-matched state populations",
    x = "Date", y = "New cases per 100,000"
  )
```



13. The dates in the `cases` dataset are stored as character strings. Use the **lubridate** package to properly parse the `date` column, then create a summary table showing the total COVID-19 cases by month and year for 2020 and 2021. The table should have columns for year, month (as month name), and total cases across all states. Order by year and month. Use the **knitr** package and `kable()` function to display the results as a formatted table.

```
library(lubridate)
library(knitr)
cases |>
  mutate(
    date  = ymd(date),
    cases = as.numeric(cases),
    year  = year(date),
    month = month(date, label = TRUE, abbr = FALSE)
  ) |>
  filter(year %in% c(2020, 2021)) |>
  group_by(year, month) |>
  summarise(total_cases = sum(cases, na.rm = TRUE), .groups = "drop") |>
  arrange(year, month) |>
  kable()
```

year	month	total_cases
2020	January	11
2020	February	68
2020	March	68245
2020	April	974032
2020	May	650943
2020	June	654904
2020	July	1989512
2020	August	1461283
2020	September	1415438
2020	October	1628598
2020	November	3932646
2020	December	7027128
2021	January	5808063
2021	February	2667511
2021	March	2068441
2021	April	1773591
2021	May	972915
2021	June	493635
2021	July	1137440
2021	August	3572562
2021	September	5027537
2021	October	2356302
2021	November	2322814
2021	December	5615644

14. The following URL provides additional COVID-19 data from the CDC in JSON format:

```
deaths_url <- "https://data.cdc.gov/resource/9bhg-hcku.json"
```

Use **httr2** to download COVID-19 death data from this endpoint. Make sure to remove the default limit to get all available data. Create a clean dataset called `deaths` with columns `state`, `date`, and `deaths` (renamed from the original column name). Ensure dates are in proper Date format and deaths are numeric.

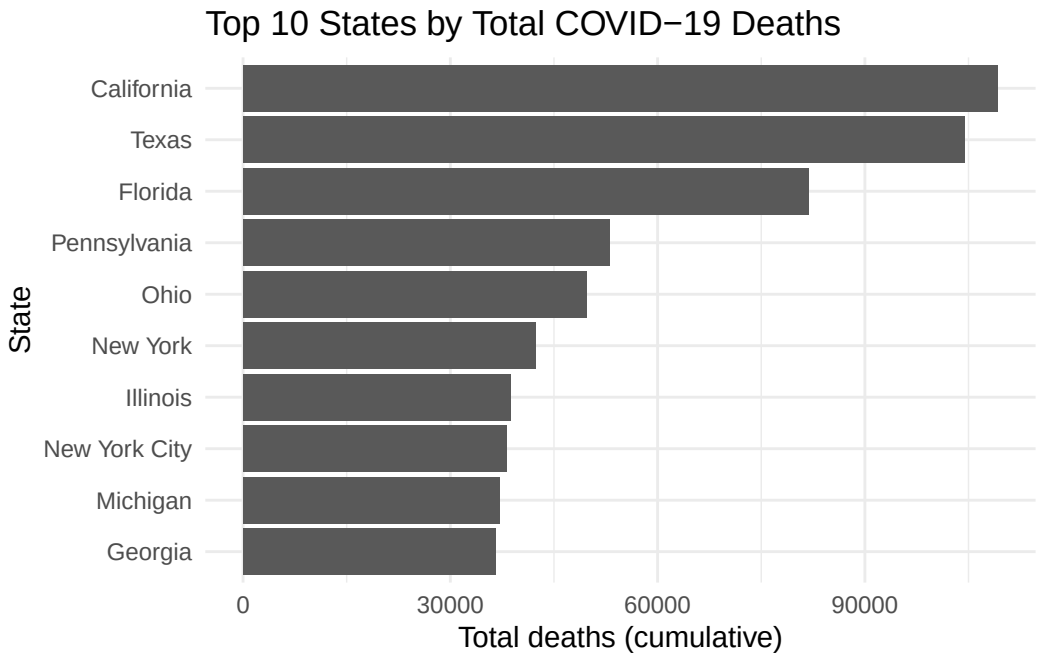
```
deaths <- request(deaths_url) |>
  req_url_query(`$limit` = 10000000000) |>
  req_perform() |>
  resp_body_json(simplifyVector = TRUE) |>
  as_tibble() |>
  (\(df) {
```

```
    df$date   <- as.Date(df$end_date)
    df$deaths <- suppressWarnings(as.numeric(df$covid_19_deaths))
    df[, c("state", "date", "deaths")]
  })()
deaths
```

```
# A tibble: 137,700 x 3
   state         date        deaths
   <chr>         <date>       <dbl>
 1 United States 2023-09-23 1146774
 2 United States 2023-09-23     519
 3 United States 2023-09-23    1696
 4 United States 2023-09-23     285
 5 United States 2023-09-23     509
 6 United States 2023-09-23    3021
 7 United States 2023-09-23    7030
 8 United States 2023-09-23   12401
 9 United States 2023-09-23   19886
10 United States 2023-09-23   30108
# i 137,690 more rows
```

15. Using the `deaths` dataset you created, make a bar plot showing the total COVID-19 deaths by state. Show only the top 10 states with the highest death counts. Order the bars from highest to lowest and use appropriate labels and title.

```
deaths |>
  filter(state != "United States") |>
  group_by(state) |>
  summarise(total_deaths = max(deaths, na.rm = TRUE), .groups = "drop") |>  # use sum(...) i
  slice_max(total_deaths, n = 10, with_ties = FALSE) |>
  ggplot(aes(x = reorder(state, total_deaths), y = total_deaths)) +
  geom_col() +
  coord_flip() +
  labs(
    title = "Top 10 States by Total COVID-19 Deaths",
    x = "State",
    y = "Total deaths (cumulative)"
  ) +
  theme_minimal()
```


Top 10 States by Total COVID−19 Deaths
California
Texas
Florida
Pennsylvania
Ohio
New York
Illinois
New York City
Michigan
Georgia
State
0
30000
60000
90000
Total deaths (cumulative)