Coding lab project

Ari Anisfeld

8/28/2020

Your final project is quite simple. You will pick a data set that speaks to you and try to uncover something interesting which you will visualize in a plot. You will also compute some summary statistics that you will show in a summary table. Your TAs will review your work and provide a non-binding suggestion as to whether you are ready for the 'accelerated' Fall Coding Lab.

Due date: September 21, 5pm upload your Rmd and pdf to gradescope. Feel free to submit your work earlier. We will not review work that is turned in after that point.

Format:

You will turn in a knitted pdf[1] that has the following sections. - `graph` in which you load your data set and provide the minimal code that produces your graph. - `table` in which you create and print a summary table with minimal code. - `appendix` (optional) in which you share code you used during data exploration, e.g. extensions of your main plot, other plots you attempted on your search for your main plot.

I have provided a sample project at the end of this document.

You are welcome to use google and stackoverflow as you procede. Please cite your sources if you borrow a solution from stackoverflow or someone's blog. To cite, just add a comment with the url. See, the last code line of the sample project where I used stackoverflow to figure out how to reformat my legend.

Your TAs will review how to use Rmds with you. But here are some quick tips.

- You make a new section in Rmd using `# section title`.
- When you read in you data, we do not want to see messages or warnings. To avoid this start the code block where you read the data with the following `{r, message = FALSE, warning = FALSE}`.

- If you have a line of code that is too long, it will be cutoff. Most R code can be split across two lines.
- Knit early and often (this is how you test if the Rmd is working how you think it is).

Datasets:

Below is a list of suitable data sources. You are welcome to and encouraged to find a data source not on this list that speaks to your policy interests. Many of these data sources have a wide range of data sets. Pick one that comes in tabular format with several variables that are interesting[2]. I recommend that once you pick a data source that is sufficient work with that so you have ample time to focus on your R skills.

Description	url
Weekly Covid data from US CDC (several datasets available, follow links)	https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/

---

[1]Occasionally students have technical issues knitting to pdfs because of some missing dependency or latex. If that's the case for you, try knitting to html instead.

[2]By which I mean there's variation.

Google maps data aggregating how visits to places, such as grocery stores and parks, are changing in each geographic region since February 15, 2020 until today (3-4 days delay), compared to the same week of the day in January, 2020.	https://www.google.com/covid19/mobility/index.html?hl=en
World Inequality Database which allows you to download a customizable dataset. You are able to choose the indicators you want (per adult gdp, top 10% income share & dozens others), countries you want and date range that you want.	https://wid.world/data/
The Humanitarian Data Exchange (HDX) is an open platform for sharing data across crises and organizations. They host thousands of datasets including development indicator data, geospatial data, damage assessments, and more.	https://data.humdata.org/
World Bank publishes hundreds of different global development related datasets including datasets on World Development Indicators, all of World Bank's lending projects and access to sample survey data etc. Also able to search data by country or indicator.	https://data.worldbank.org/
The City of Chicago publishes many different datasets, including ones on public finance, public safety, transportation, and education	https://data.cityofchicago.org/
NYC also has an open data initiative that aims to provide data from different agencies in one central platform. Data on ride-share programs can be found there as well, but of course, the public version	https://opendata.cityofnewyork.us/
FiveThirtyEight is a website focused on opinion poll analysis, politics, economics, and sports blogging. Currently, it has a unique repository of datasets about 2020 election polls and forecasts, Trump's popularity, Americans' view on COVID crisis Trump's response, NBA/NFL/MLB/Soccer predictions, among others.	https://data.fivethirtyeight.com/
National Bureau of Economic Research has a master page of data resources.	https://data.nber.org/data/
Google also has a dataset search tool that is pretty handy for finding data without having to comb through niche websites. I would not rely on this completely, though, since not all datasets follow the standards required to show up in these search results	https://datasetsearch.research.google.com/

Example project:

Introduction

I analyze weekly covid-19 data from the US Center for Disease Control. I show the extent to which racial disparities exist as measured by the percentage change in deaths in 2020 compared to 2015-2019. The plot below shows data for the United States except the tri-state area NY-NJ-CT. NYC is a large diverse city that was particularly hard hit by the coronavirus, so it is plausible that the racial disparities reported on are driven by those facts. The plot shows that NYC does not appear to drive the disparities. In the appendix, I examine the same question in states with high Latinx populations that experienced a covid-19 surge in the late summer (TX-CA-AZ-FL). And, I look at the least densely populated states.

The data description is found here: https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/ The data can be downloaded directly from here: https://data.cdc.gov/api/views/qfhf-uhaa/rows.csv?accessType=DOWNLOAD&bom=true&format=true%20target=

graph

```
library(tidyverse)

covid_data <-
  read_csv("../data/Weekly_counts_of_deaths_by_jurisdiction_and_race_and_Hispanic_origin.csv",
           col_types = cols(Suppress = col_character())) %>%
    mutate(week = `Week Ending Date`,
           race_ethnicity = `Race/Ethnicity`,
           n_deaths = `Number of Deaths`,
           diff = `Difference from 2015-2019 to 2020`,
           expected_deaths = n_deaths - diff,
           perc_diff = `Percent Difference from 2015-2019 to 2020`,
           year = MMWRYear,
           week_no = MMWRWeek,
           jurisdiction = Jurisdiction,
           state = `State Abbreviation`
           ) %>%
    filter(`Time Period` == "2020", Outcome == "All Cause", Type != "Unweighted") %>%
    select(jurisdiction, state, week, year, week_no,
           race_ethnicity, n_deaths, expected_deaths, diff, perc_diff)
```

```
data_for_plot <-
covid_data %>%
  mutate(week = lubridate::mdy(week)) %>%
  filter(race_ethnicity %in%
         c("Hispanic", "Non-Hispanic White", "Non-Hispanic Black", "Non-Hispanic Asian")) %>%
  filter(! state %in% c("US","NY","YC", "NJ",  "CT", "PR"), week_no <= 29) %>%
  group_by(race_ethnicity, week ) %>%
  summarize(actual_deaths = sum(n_deaths, na.rm  = TRUE),
            diff_deaths = sum(diff, na.rm = TRUE),
            expected_deaths = actual_deaths - diff_deaths,
            perc_above_expected = 100 * diff_deaths / expected_deaths)
```

```
## `summarise()` regrouping output by 'race_ethnicity' (override with `.groups` argument)
```

```
data_for_plot %>%
  ggplot(aes(x = week, color = race_ethnicity)) +
    geom_line(aes(y = perc_above_expected)) +
  theme_minimal() +
  labs(y = "Percent of deaths above expected deaths\n(weekly mean 2015-2019 for given week)",
       x = "",
       title = "Racial disparities of Covid-19, USA excluding NY-NJ-CT" ,
       subtitle = "Data source: CDC",
       color = "") +
  theme(legend.position = "bottom")
```

Racial disparities of Covid−19, USA excluding NY−NJ−CT
Data source: CDC
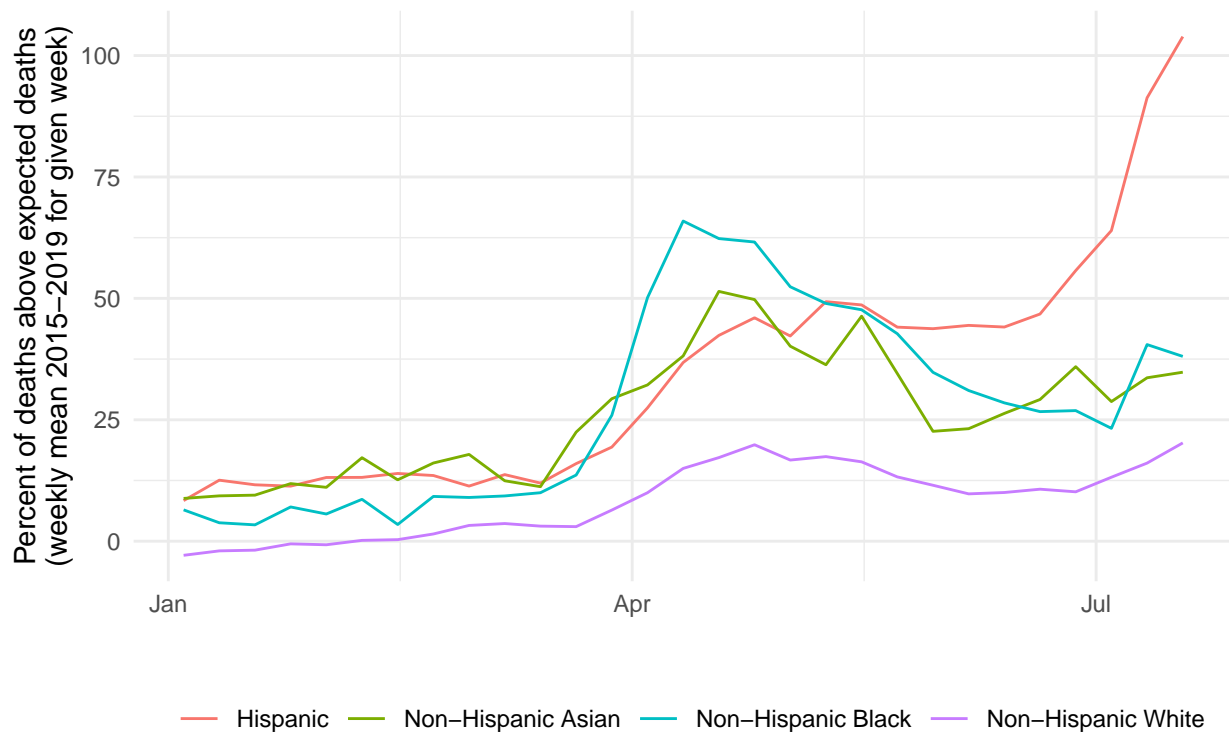
table

```
summary_table <-
covid_data %>%
  filter(state == "US") %>%
  group_by(race_ethnicity) %>%
  summarize(expected_deaths = sum(expected_deaths, na.rm = TRUE),
            total_additional_deaths = sum(diff, na.rm = TRUE),
            percent_diff = 100 * total_additional_deaths / expected_deaths
            )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
summary_table %>%
  knitr::kable()
```

race_ethnicity	expected_deaths	total_additional_deaths	percent_diff
Hispanic	121787	54977	45.141928
Non-Hispanic American Indian or Alaska Native	10749	2335	21.722951
Non-Hispanic Asian	40741	14364	35.256867
Non-Hispanic Black	204431	61969	30.312917
Non-Hispanic White	1337231	130967	9.793895
Other	15094	2613	17.311515

Appendix

```r
# We don't expect you to use functions. For this project it's acceptable to repeat code.
# As you grow as a programmer, when you find that you want to copy and paste a code block
# over and over again. It means it's time for a function or a loop. We'll discuss these
# in the fall.
data_for_plot <-
  function(states,
    ethnicities = c("Hispanic", "Non-Hispanic White", "Non-Hispanic Black", "Non-Hispanic Asian")) {

    covid_data %>%
      mutate(week = lubridate::mdy(week)) %>%
      filter(race_ethnicity %in% ethnicities) %>%
      filter(! state %in% "US", state %in% states, week_no <= 29) %>%
      group_by(race_ethnicity, week ) %>%
      summarize(actual_deaths = sum(n_deaths, na.rm  = TRUE),
                diff_deaths = sum(diff, na.rm = TRUE),
                expected_deaths = actual_deaths - diff_deaths,
                perc_above_expected = 100 * diff_deaths / expected_deaths)
}

make_plot <- function(data_for_plot, title) {
    data_for_plot %>%
      ggplot(aes(x = week, color = race_ethnicity)) +
        geom_line(aes(y = perc_above_expected)) +
      theme_minimal() +
      labs(y = "Percent of deaths above expected deaths\n(weekly mean 2015-2019 for given week)",
          x = "",
          title = glue::glue("Racial disparities of Covid-19 {title}"),
          subtitle = "Data source: CDC",
          color = "") +
      theme(legend.position = "bottom")
}

data_for_plot(c("AZ", "TX", "FL", "CA")) %>% make_plot("in TX-FL-AZ-CA")

## `summarise()` regrouping output by 'race_ethnicity' (override with `.groups` argument)
```
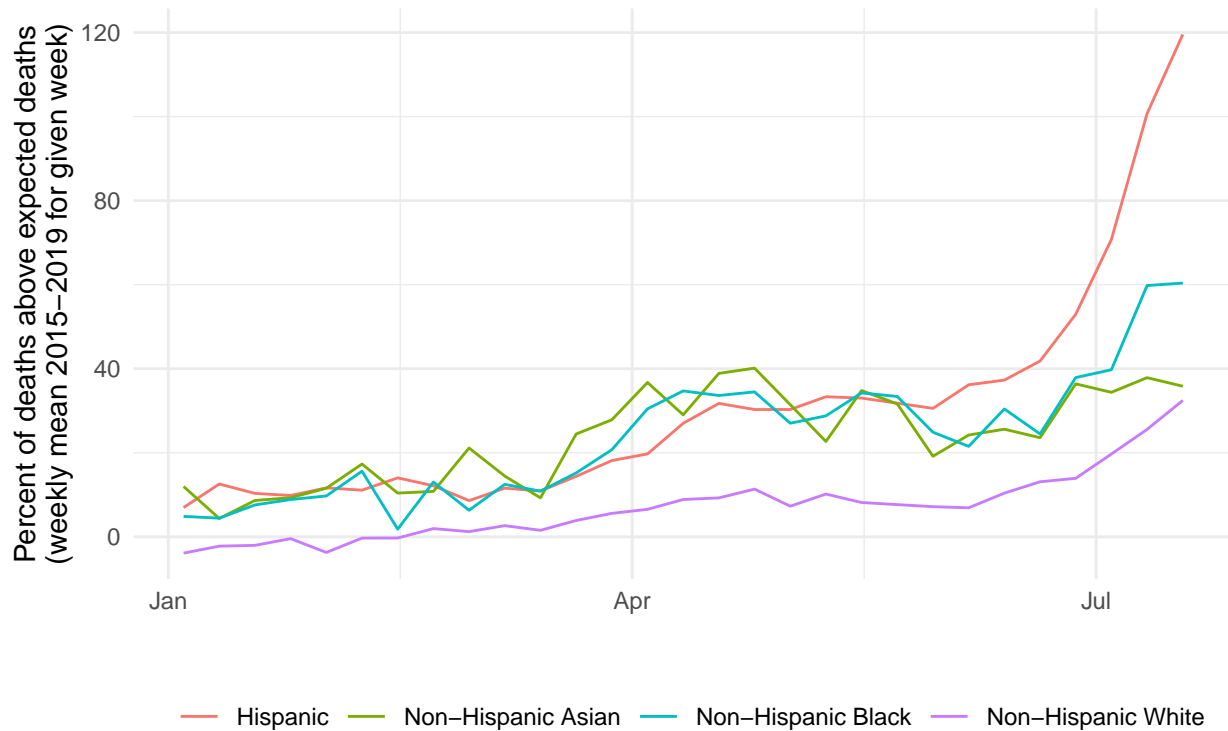
Racial disparities of Covid−19 in TX−FL−AZ−CA
Data source: CDC


```
data_for_plot(c("AK","MT", "WY", "SD", "ND", "NM", "NE", "ID"),
              ethnicities = c("Hispanic", "Non-Hispanic White",
                              "Non-Hispanic Black", "Non-Hispanic Asian",
                              "Non-Hispanic American Indian or Alaska Native")
              ) %>% make_plot("low-population density states") +
  # code from https://stackoverflow.com/questions/10332387/wrap-legend-text-in-ggplot2
  guides(colour = guide_legend(nrow = 2))
```

```
## `summarise()` regrouping output by 'race_ethnicity' (override with `.groups` argument)
```

```
## Warning: Removed 29 row(s) containing missing values (geom_path).
```

Racial disparities of Covid−19 low−population density states
Data source: CDC