Assignment 2

Nicole Blackburn and Jianing Zou

You may work in pairs or individually for this assignment. Make sure you join a group in Canvas if you are working in pairs. Turn in this assignment as an HTML or PDF file to ELMS. Make sure to include the R Markdown or Quarto file that was used to generate it.

```
library(tidyverse)
library(epidatr)
library(censusapi)
```

In this assignment, you will pull from APIs to get data from various data sources and use your data wrangling skills to use them all together. You should turn in a report in PDF or HTML format that addresses all of the questions in this assignment, and describes the data that you pulled and analyzed. You do not need to include full introduction and conclusion sections like a full report, but you should make sure to answer the questions in paragraph form, and include all relevant tables and graphics.

Whenever possible, use piping and `dplyr`. Avoid hard-coding any numbers within the report as much as possible.

Pulling from APIs

Our first data source is the Delphi COVIDcast data. You can access this using the Epidata API built by Carnegie Mellon University's Delphi Research group. Documentation for this API can be found here: https://cmu-delphi.github.io/delphi-epidata/. Here, we find the smoothed estimate of the proportion of people experiencing Covid-like symptoms by county from April 6, 2020 to April 14, 2020.

```
covid_april <- pub_covidcast('fb-survey',
                        'smoothed_wcli',
                        'county',
                        'day',
```

```
                        time_values = c(20200406:20200414))
head(covid_april)
```

```
# A tibble: 6 x 15
  geo_value signal     source geo_type time_type time_value direction issue
  <chr>     <chr>      <chr>  <fct>    <fct>     <date>         <dbl> <date>
1 01000     smoothed_~ fb-su~ county   day       2020-04-06        NA 2020-09-03
2 01073     smoothed_~ fb-su~ county   day       2020-04-06        NA 2020-09-03
3 01089     smoothed_~ fb-su~ county   day       2020-04-06        NA 2020-09-03
4 01097     smoothed_~ fb-su~ county   day       2020-04-06        NA 2020-09-03
5 02000     smoothed_~ fb-su~ county   day       2020-04-06        NA 2020-09-03
6 02020     smoothed_~ fb-su~ county   day       2020-04-06        NA 2020-09-03
# i 7 more variables: lag <dbl>, missing_value <dbl>, missing_stderr <dbl>,
#   missing_sample_size <dbl>, value <dbl>, stderr <dbl>, sample_size <dbl>
```

For more information about the data, see: https://cmu-delphi.github.io/delphi-epidata/api/covidcast_signals.html

Answer the following questions:

- Change the data from long to wide format by including the estimate of Covid-like symptoms for each day as a column. There should be a column for `geo_value` as well as a column for each of the days in the dataset.

```
covid_april |>
  pivot_wider(id_cols = geo_value,
              names_from = time_value,
              values_from = value)
```

```
# A tibble: 1,462 x 10
   geo_value `2020-04-06` `2020-04-07` `2020-04-08` `2020-04-09` `2020-04-10`
   <chr>            <dbl>        <dbl>        <dbl>        <dbl>        <dbl>
 1 01000            1.19         1.06         0.924        0.855        0.895
 2 01073            1.94         1.54         1.25         1.03         0.903
 3 01089            0.723        0.490        0.654        0.539        0.545
 4 01097            1.14         0.935        0.894        0.918        1.03
 5 02000            1.76         1.02         1.41         1.42         1.28
 6 02020            0.332        0.711        0.554        0.455        0.520
 7 04000            1.02         1.36         1.22         0.270        0
 8 04013            0.858        0.819        0.879        0.821        0.822
 9 04015            1.30         0.867        0.535        0.356        0.414
```

```
10 04019            0.966        0.828        0.948        0.972        0.940
# i 1,452 more rows
# i 4 more variables: `2020-04-11` <dbl>, `2020-04-12` <dbl>,
#   `2020-04-13` <dbl>, `2020-04-14` <dbl>
```

- Find the mean, median, and variance of the estimate on each of the days from April 6, 2020 to April 14, 2020. (Note that this is not the appropriate way of finding the overall measures in reality because we aren't using weights)

```
covid_april |>
  group_by(time_value) |>
  summarize(
    mean = mean(value),
    median = median(value),
    variance = var(value)
  )
```

```
# A tibble: 9 x 4
  time_value  mean median variance
  <date>     <dbl>  <dbl>    <dbl>
1 2020-04-06 0.955  0.849    0.454
2 2020-04-07 0.890  0.779    0.341
3 2020-04-08 0.871  0.789    0.300
4 2020-04-09 0.856  0.778    0.278
5 2020-04-10 0.850  0.777    0.283
6 2020-04-11 0.853  0.776    0.264
7 2020-04-12 0.854  0.784    0.260
8 2020-04-13 0.830  0.765    0.244
9 2020-04-14 0.796  0.719    0.255
```

- Which counties had the highest report Covid-like symptoms on each of the days within this range?

Answer

On April 6, 2020, county 36005 (Bronx County, NY) had the highest report of Covid-like symptoms. From April 7, 2020 - April 13, 2020, county 36087 (Rockland County, NY) has the highest report of Covid-like symptoms. On April 14, 2020, county 36079 (Putnam County, NY) had the highest report of Covid-like symptoms.

```
covid_april |>
  select(time_value, geo_value, value) |>
  group_by(time_value) |>
```

```
  arrange(desc(value)) |>
  slice_head(n = 1)
```

```
# A tibble: 9 x 3
# Groups:   time_value [9]
  time_value geo_value value
  <date>     <chr>     <dbl>
1 2020-04-06 36005      3.41
2 2020-04-07 36087      4.59
3 2020-04-08 36087      5.16
4 2020-04-09 36087      4.63
5 2020-04-10 36087      4.52
6 2020-04-11 36087      4.29
7 2020-04-12 36087      4.41
8 2020-04-13 36087      4.69
9 2020-04-14 36079      3.97
```

Using the API, get the actual COVID cases from the JHU Cases and Deaths (using the link above, `confirmed_7dav_incidence_prop`) from May 6, 2020 to May 14, 2020. This is the number of confirmed COVID cases per 100,000 people. Find the correlation between reported COVID-like symptoms and actual COVID cases per 100,000 people within each county a month later. Is there a relationship?

Answer

The correlation coefficient is 0.08998326, which indicates that there is a weak positive correlation between the April illness symptoms and may COVID cases by county. . The p-value of <2.2e-16 indicates the correlation is statistically significant. The 95% condifidence interval is $[0.07078256, 0.10911729]$, which does not include zero, further indicating we can reject the null hypothesis that there is no correlation.

```
covid_may <- pub_covidcast('jhu-csse',
                        'confirmed_7dav_incidence_prop',
                        'county',
                        'day',
                        time_values = c(20200506:20200514))
```

```
covid_both_apr <-
  covid_april |>
  select(geo_value, time_value, value) |>
  mutate(same_date = time_value + 30)
```

```r
covid_both_may <-
  covid_may |>
  select(geo_value, time_value, value)

covid_both <- inner_join(covid_both_apr,
                         covid_both_may,
                         join_by(geo_value, same_date == time_value))
```

```r
cor.test(covid_both$value.x, covid_both$value.y)
```

```
    Pearson's product-moment correlation

data:  covid_both$value.x and covid_both$value.y
t = 9.1633, df = 10286, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.07078256 0.10911729
sample estimates:
       cor
0.08998326
```

Covidcast API Data + ACS

Now lets add another data set. The `censusapi` package provides a nice R interface for communicating with this API. However, before running queries we need an access key. This (easy) process can be completed here:

https://api.census.gov/data/key_signup.html

Once you have an access key, save it as a text file, then read this key in the `cs_key` object. We will use this object in all following API queries. Note that I called my text file `census-key.txt` – yours might be different!

```r
cs_key <- read_file("C:/Users/npbvb/OneDrive - University of Maryland/R_Projects/727/APIs/ce
```

```r
cs_key <- read_file("census-key")
```

You can navigate through the documentation for all Census Data APIs here: https://www.census.gov/data/developers/data-sets.html Documentation for the 5-year ACS API can be found here: https://www.census.gov/data/developers/data-sets/acs-5year.html.

In the following, we request basic socio-demographic information (population, median age, median household income, income per capita) for counties in the US. The information about the variables used here can be found here: https://api.census.gov/data/2022/acs/acs5/variables.html.

```
acs <- getCensus(name = "acs/acs5",
                 vintage = 2020,
                 vars = c("NAME",
                          "B01001_001E",
                          "B06002_001E",
                          "B19013_001E",
                          "B19301_001E"),
                 region = "county",
                 key = cs_key)
head(acs)
```

```
  state county                    NAME B01001_001E B06002_001E B19013_001E
1    01    001 Autauga County, Alabama       55639        38.6       57982
2    01    003 Baldwin County, Alabama      218289        43.2       61756
3    01    005 Barbour County, Alabama       25026        40.1       34990
4    01    007    Bibb County, Alabama       22374        39.9       51721
5    01    009  Blount County, Alabama       57755        41.0       48922
6    01    011 Bullock County, Alabama       10173        39.7       33866
  B19301_001E
1       29804
2       33751
3       20074
4       22626
5       25457
6       20783
```

Now, it might be useful to rename the socio-demographic variables (`B01001_001E` etc.) in our data set and assign more meaningful names.

```
acs <-
  acs %>%
  rename(pop = B01001_001E,
         age = B06002_001E,
         hh_income = B19013_001E,
         income = B19301_001E)
```

It seems like we could try to use this location information listed above to merge this data set with the COVID data. However, we first have to clean the geography data to match the two datasets. The COVID data has a five digit geography code, with the first two digits representing the state and the last three representing the county within that state. The ACS data has this separated out. Add a new variable `location` to the ACS data that has the geography value in the same format as the COVID data.

```
acs <- acs |>
        mutate(
        location = paste0(state, county),
         .after = county)
```

Answer the following questions with the COVID data and ACS data.

- First, check how many counties aren't matched. Then, create a new data set by joining the two datasets. Keep only counties that appear in both data sets.

Answer

In the covid dataset, there were 51 counties that weren't matched to the ACS dataset. These counties ended with 000. Per the Delphi Epidata API on Geographic coding, "Megacounty estimates are reported with a FIPS code ending with 000, which is never a FIPS code for a real county". In the ACS dataset, there were 1,810 counties that didn't match to the covid dataset. After combining the dataset, there are 1,391 counties that appear in both data sets.

```
covid_both |>
  distinct(location = geo_value) |>
  anti_join(acs)

acs |>
  distinct(geo_value = location) |>
  anti_join(covid_both)

combined_full <- inner_join(acs, covid_both, join_by(location == geo_value))

combined_full |>
  count(location) |>
  filter(n > 1)
```

- Compute the mean of the proportion of people with covid-like illness symptoms on April 6, 2020 for counties that have an above average median household income and for those that have an below average median household income. When building your pipe, start with creating the grouping variable and then proceed with the remaining tasks. What conclusions might you draw from this?

Answer

On April 6th, 2020, counties with an above average median household income had a mean proportion of 1.0148242 covid-like illness symptoms, while counties with a below average median household income had a mean proportion of 0.8810505 covid-like illness symptoms. A possible explanation of this is the exposure to COVID through a job; households with two or more working adults and a higher household income have more exposure to COVID through multiple workplaces while households with only one working adult and a lower household income have exposure to COVID from only one workplace. Another conclusion could be that counties with above average household incomes reported covid-like symptoms more than counties with below average household incomes.

```
combined <- combined_full |>
  select(location, hh_income, time_value, value.x)

combined |>
  group_by(above_avg_hh = hh_income > mean(combined$hh_income)) |>
  filter(time_value == as.Date("2020-04-06")) |>
  summarize(mean_prop = mean(value.x))
```

```
# A tibble: 2 x 2
  above_avg_hh mean_prop
  <lgl>            <dbl>
1 FALSE            0.881
2 TRUE             1.01
```

- Is there a relationship between the median household income and the proportion of people reporting Covid-like illness symptoms? Describe the relationship and use a scatterplot.

Answer

The correlation coefficient of -0.06415457 indicates a weak, negative relationship between median household income and the proportion of people reporting covid-like symptoms. The p-value of 1.763e-10 is less than 0.05, showing the relationship to be statistically significant. The 95% confidence interval $(-0.08377330, -0.04448613)$ does not include zero, further indicating that we can reject the null hypothesis of having no relationship.

```
cor.test(combined$hh_income, combined$value.x)
```

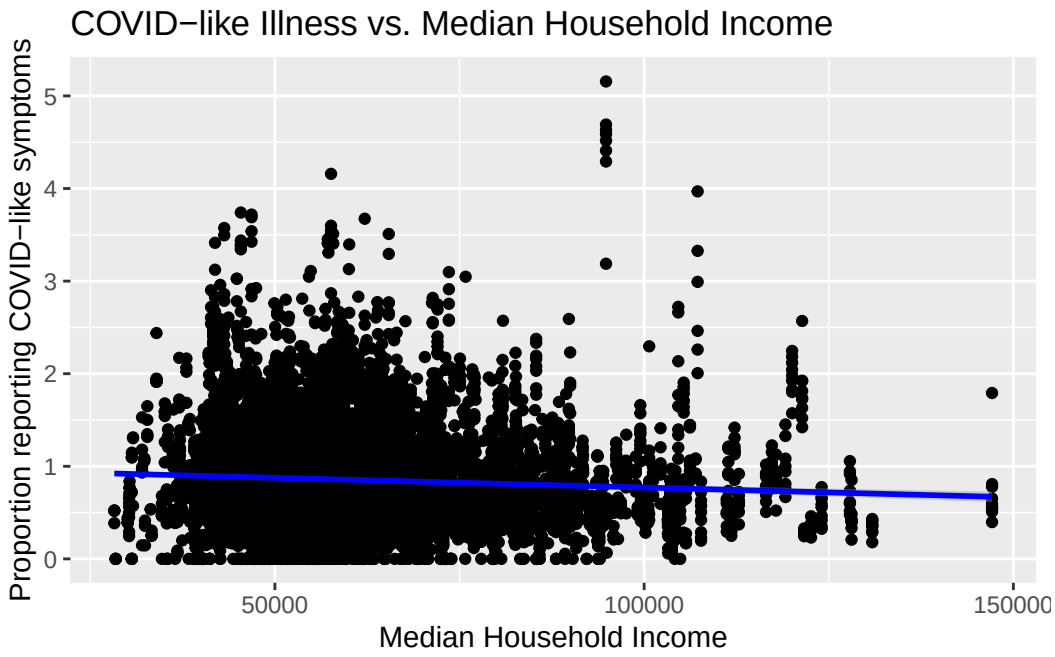
```
    Pearson's product-moment correlation

data:  combined$hh_income and combined$value.x
```

```
t = -6.3874, df = 9872, p-value = 1.763e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08377330 -0.04448613
sample estimates:
        cor
-0.06415457
```

```r
ggplot(combined, aes(x = hh_income, y = value.x)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(
    x = "Median Household Income",
    y = "Proportion reporting COVID-like symptoms",
    title = "COVID-like Illness vs. Median Household Income"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Using Other Census Data

Suppose we wanted to use the 2020 1-year ACS instead of the 5-year ACS. Why would we be unable to do this?

Hint: Read the documentation for the 1-year ACS

Answer

We can't do this because The Census Bureau does not release its standard 2020 ACS 1-year estimates because of the impact of the COVID-19 pandemic on data collection. Experimental estimates, developed from 2020 ACS 1-year data are available on the ACS Experimental Data page. They will not be available on data.census.gov or the Application Programming Interface (API).

Instead, repeat the steps above to merge the Delphi COVIDcast data to the 1-year ACS from 2021 (rather than the 5-year ACS). Do the same analysis as above.

```
acs_2021 <- getCensus(name = "acs/acs1",
                      vintage = 2021,
                      vars = c("NAME",
                               "B01001_001E",
                               "B06002_001E",
                               "B19013_001E",
                               "B19301_001E"),
                      region = "county",
                      key = cs_key)
head(acs_2021)
```

```
  state county                    NAME B01001_001E B06002_001E B19013_001E
1    01    003 Baldwin County, Alabama      239294        43.9       63866
2    01    015 Calhoun County, Alabama      115972        40.2       46524
3    01    043 Cullman County, Alabama       89496        40.5       55517
4    01    049  DeKalb County, Alabama       71813        40.2       41800
5    01    051  Elmore County, Alabama       89304        40.0       59032
6    01    055  Etowah County, Alabama      103162        41.5       45298
  B19301_001E
1       35824
2       24804
3       28024
4       25633
5       28515
6       23768
```

```r
acs_2021 <-
  acs_2021 %>%
  rename(pop = B01001_001E,
         age = B06002_001E,
         hh_income = B19013_001E,
         income = B19301_001E)

acs_2021 <- acs_2021 |>
        mutate(
        location = paste0(state, county),
         .after = county)
```

- First, check how many counties aren't matched. Then, create a new data set by joining the two datasets. Keep only counties that appear in both data sets.

Answer

Unmatched counties:

647 counties in the covid data set that aren't in acs_2021.

26 counties in acs_2021 dataset that aren't in covid data.

Matched:

6,486 matched pairs at total. There are 807 counties in the combined data set.

```r
covid_both |>
  distinct(location = geo_value) |>
  anti_join(acs_2021)

acs_2021 |>
  distinct(geo_value = location) |>
  anti_join(covid_both)

combined2_full <- inner_join(acs_2021, covid_both, join_by(location == geo_value))

combined2_full |>
  count(location) |>
  filter(n > 1)
```

- Compute the mean of the proportion of people with covid-like illness symptoms on April 6, 2020 for counties that have an above average median household income and for those that have an below average median household income. When building your pipe, start

with creating the grouping variable and then proceed with the remaining tasks. What conclusions might you draw from this?

Answer

Then mean of the proportion of people with covid-like illness symptoms on April 6, 2020 of the above average income group is 0.9624972, and the mean of the below average income group is 0.9893342, which is higher than the above one.

This difference indicates that the above average income group is less likely to report illness symptoms than the below average income group.

```
combined_2021 <- combined2_full |>
  select(location, hh_income, time_value, value.x)

combined_2021 |>
  group_by(above_avg_hh = hh_income > mean(combined_2021$hh_income)) |>
  filter(time_value == as.Date("2020-04-06")) |>
  summarize(mean_prop = mean(value.x))
```

```
# A tibble: 2 x 2
  above_avg_hh mean_prop
  <lgl>            <dbl>
1 FALSE            0.989
2 TRUE             0.962
```

- Is there a relationship between the median household income and the proportion of people reporting Covid-like illness symptoms? Describe the relationship and use a scatterplot.

Answer

The Correlation coefficient is -0.08024641, indicating a weak, negative correlation. However, the p-value is 9.678e-11, which is lower than 0.05, so the correlation is statistically significant. The 95% CI is $[-0.10437997, -0.05601841]$ which does not include zero, further indicating the null hypothesis of no correlation can be rejected.

The scatterplot suggests a weak negative relationship between median household income and the proportion of residents reporting illness symptoms. While there may be a slight trend, the wide scatter and presence of outliers indicate the relationship is inconsistent across counties.

In sum, there is evidence of a significant but weak negative correlation between median household income and COVID-like symptom reporting by county.

```
cor.test(combined_2021$hh_income, combined_2021$value.x)
```

```
    Pearson's product-moment correlation

data:  combined_2021$hh_income and combined_2021$value.x
t = -6.4826, df = 6484, p-value = 9.678e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.10437997 -0.05601841
sample estimates:
        cor
-0.08024641
```

```r
ggplot(combined_2021, aes(x = hh_income, y = value.x)) +
  geom_point() +
  geom_smooth(method = "lm",
              se = TRUE,
              color = "blue") +
  labs(x = "Median Household Income",
       y = "Proportion reporting COVID-like symptoms",
       title = "COVID-like Illness vs. Median Household Income")
```

```
`geom_smooth()` using formula = 'y ~ x'
```