Assignment 2

Due at 11:59pm on October 3.

Leng Seong Che; Bozhou (Peter) Tan

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.

```r
library(tidyverse)
library(gtrendsR)
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.

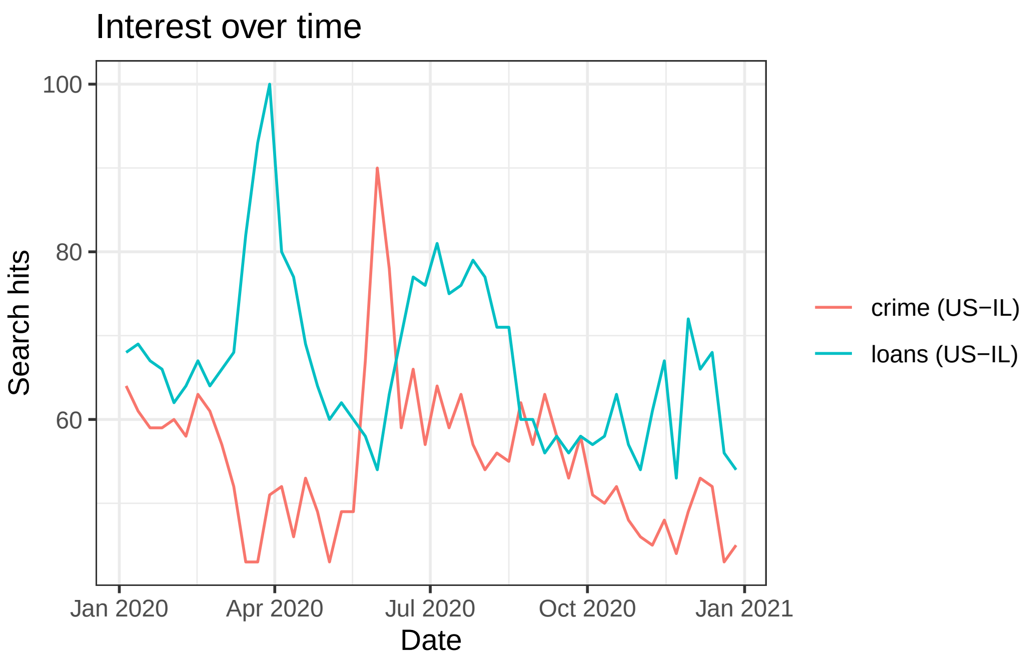
Whenever possible, use piping and `dplyr`. Avoid hard-coding any numbers within the report as much as possible.

Pulling from APIs

`crime` and `loans`

Our first data source is the Google Trends API. Suppose we are interested in the search trends for `crime` and `loans` in Illinois in the year 2020. We could find this using the following code:

```r
res = gtrends(c("crime", "loans"),
              geo = "US-IL",
              time = "2020-01-01 2020-12-31",
              low_search_volume = TRUE)
plot(res)
```



Answer the following questions for the keywords “crime” and “loans”.

- Find the mean, median and variance of the search hits for the keywords.

```
# transfer the data into tibble
rest = as_tibble(res$interest_over_time)

# find the mean, median and variance of the search hits
library(dplyr)
library(tidyr)
library(knitr)

descriptive = rest %>%
  group_by(keyword) %>%
  summarise(n = n(),
            mean = mean(hits),
            median = median(hits),
            variance = var(hits))
kable(descriptive, caption = "Descriptive Statistics of Keywords")
```

Table 1: Descriptive Statistics of Keywords

keyword	n	mean	median	variance
crime	52	55.26923	54.5	79.02413
loans	52	66.73077	66.0	98.67119

According to Table 1, we can find that the keyword `crime` has a mean of 55.00000, a median of 54 and a variance of 86.43137. The keyword `loans` has a mean of 66.48077, a median of 65 and a variance of 95.39178.

- Which cities (locations) have the highest search frequency for `loans`? Note that there might be multiple rows for each city if there were hits for both "crime" and "loans" in that city. It might be easier to answer this question if we had the search hits info for both search terms in two separate variables. That is, each row would represent a unique city.

```
rescity = as_tibble(res$interest_by_city) %>%
  pivot_wider(., names_from = keyword, values_from = hits) %>%
  arrange(., desc(loans))
kable(head(rescity), caption = "Highest Search Frequency for Loans")
```

Table 2: Highest Search Frequency for Loans

location	geo	gprop	crime	loans
White City	US-IL	web	NA	100
Alorton	US-IL	web	NA	78
Oakwood	US-IL	web	NA	62
Rosemont	US-IL	web	28	60
Roseville	US-IL	web	NA	59
Washington Park	US-IL	web	NA	55

According to Table 2, Midlothia has the highest search frequency for `loans` with the value of 100.

- Is there a relationship between the search intensities between the two keywords we used?

```
crime = rest %>%
  filter(keyword == "crime") %>%
  select(date, hits) %>%
  rename(., crimehits = hits)
```

```
loan = rest %>%
  filter(keyword == "loans") %>%
  select(date, hits) %>%
  rename(., loanshits = hits)

crimloan = left_join(crime, loan, by = "date")
cor.test(crimloan$crimehits, crimloan$loanshits)
```

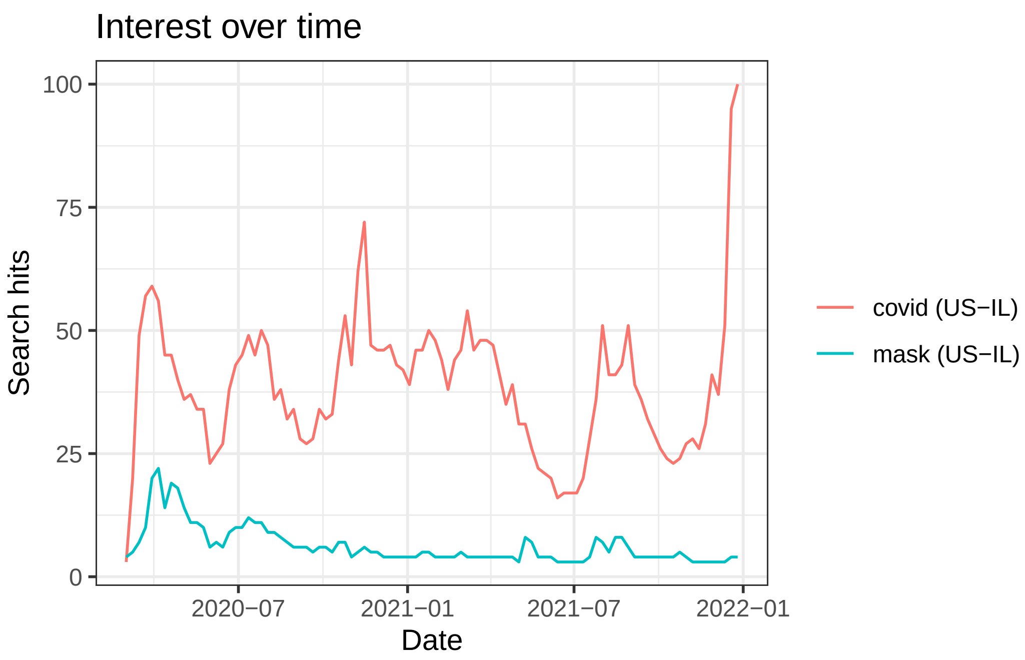
```

   Pearson's product-moment correlation

data:  crimloan$crimehits and crimloan$loanshits
t = -0.53574, df = 50, p-value = 0.5945
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3414103  0.2015059
sample estimates:
        cor
-0.07554894
```

From the plot above, it seems like there is a negative correlation between crime and loans. However, if we use the quantitative method to compute the t-statistic and corresponding p-value, we can see that the p-value is bigger than 0.05, which means there is no significant negative relationship between crime and loans.

`covid` and `mask`

Repeat the above for keywords related to covid. Make sure you use multiple keywords like we did above. Try several different combinations and think carefully about words that might make sense within this context.

```
res2 = gtrends(c("covid", "mask"),
               geo = "US-IL",
               time = "2020-03-01 2021-12-31",
               low_search_volume = TRUE)
plot(res2)
```



```
# transfer the data into tibble
rest2 = as_tibble(res2$interest_over_time)

# find the mean, median and variance of the search hits
descriptive2 = rest2 %>%
  group_by(keyword) %>%
  summarise(n = n(),
            mean = mean(hits),
            median = median(hits),
            variance = var(hits))
kable(descriptive2, caption = "Descriptive Statistics of Keywords")
```

Table 3: Descriptive Statistics of Keywords

keyword	n	mean	median	variance
covid	96	38.86458	39	211.31831
mask	96	6.25000	5	14.69474

```
rescity2 = as_tibble(res2$interest_by_city) %>%
  pivot_wider(., names_from = keyword, values_from = hits) %>%
```

```r
  arrange(., desc(mask))
kable(head(rescity2), caption = "Highest Search Frequency for mask")
```

Table 4: Highest Search Frequency for mask

location	geo	gprop	covid	mask
Cerro Gordo	US-IL	web	NA	100
Raymond	US-IL	web	NA	95
Divernon	US-IL	web	74	87
Farmersville	US-IL	web	NA	86
Winnetka	US-IL	web	NA	84
Prairie Grove	US-IL	web	83	81

```r
mask = rest2 %>%
  filter(keyword == "mask") %>%
  select(date, hits) %>%
  rename(., maskhits = hits)

covid = rest2 %>%
  filter(keyword == "covid") %>%
  select(date, hits) %>%
  rename(., covidhits = hits)

maskcovid = left_join(mask, covid, by = "date")
cor.test(maskcovid$maskhits, maskcovid$covidhits)
```

```

    Pearson's product-moment correlation

data:  maskcovid$maskhits and maskcovid$covidhits
t = 2.253, df = 94, p-value = 0.02659
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.02709019 0.40830417
sample estimates:
      cor
0.2263468
```

Google Trends + 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, store this key in the `cs_key` object. We will use this object in all following API queries.

```r
cs_key <- "c0fd12402e23b7a95923e694f046015d624c91c5"
```

In the following, we request basic socio-demographic information (population, median age, median household income, income per capita) for cities and villages in the state of Illinois.

```r
acs_il <- getCensus(name = "acs/acs5",
                    vintage = 2020,
                    vars = c("NAME",
                             "B01001_001E",
                             "B06002_001E",
                             "B19013_001E",
                             "B19301_001E"),
                    region = "place:*",
                    regionin = "state:17",
                    key = cs_key)
head(acs_il)
```

```
  state place                        NAME B01001_001E B06002_001E B19013_001E
1    17 15261 Coatsburg village, Illinois         180        35.6       55714
2    17 15300    Cobden village, Illinois        1018        44.2       38750
3    17 15352      Coffeen city, Illinois         640        33.4       35781
4    17 15378   Colchester city, Illinois        1347        42.2       43942
5    17 15469    Coleta village, Illinois         230        27.7       56875
6    17 15495    Colfax village, Illinois        1088        32.5       58889
  B19301_001E
1       27821
2       19979
3       26697
4       24095
5       23749
6       24861
```

Convert values that represent missings to NAs.

```
acs_il[acs_il == -666666666] <- NA
```

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

```
acs_il <-
  acs_il %>%
  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 Google Trends data. However, we first have to clean NAME so that it has the same structure as location in the search interest by city data. Add a new variable location to the ACS data that only includes city names.

```
library(stringr)
pattern = c("St." = "Saint")

acs_il = acs_il %>%
  mutate(location = str_remove_all(NAME, c(" town,| city,| village,| Illinois"))) %>%
  mutate(location = str_replace_all(location, coll(pattern)))
```

Answer the following questions with the “crime” and “loans” Google trends data and the ACS data.

- First, check how many cities don’t appear in both data sets, i.e. cannot be matched. Then, create a new data set by joining the Google Trends and the ACS data. Keep only cities that appear in both data sets.

```
check = left_join(acs_il, rescity, by = "location")
check %>% filter(is.na(geo)) %>% count()
```

```
     n
1 1128
```

```
joint = inner_join(rescity, acs_il, by = "location")
```

- Compute the mean of the search popularity for both keywords for cities 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?

```
group1 = joint %>%
  mutate(mean = mean(hh_income, na.rm = TRUE)) %>%
  mutate(group = ifelse(hh_income > mean, "high", "low")) %>%
  group_by(group) %>%
  summarise(crime = mean(crime, na.rm = TRUE),
            loans = mean(loans, na.rm = TRUE)) %>%
  filter(!is.na(group))
kable(group1, caption = "Search Popularity by Household Income")
```

Table 5: Search Popularity by Household Income

group	crime	loans
high	25.82979	26.82000
low	27.75000	32.18681

- Is there a relationship between the median household income and the search popularity of the Google trends terms? Describe the relationship and use a scatterplot with `qplot()`.

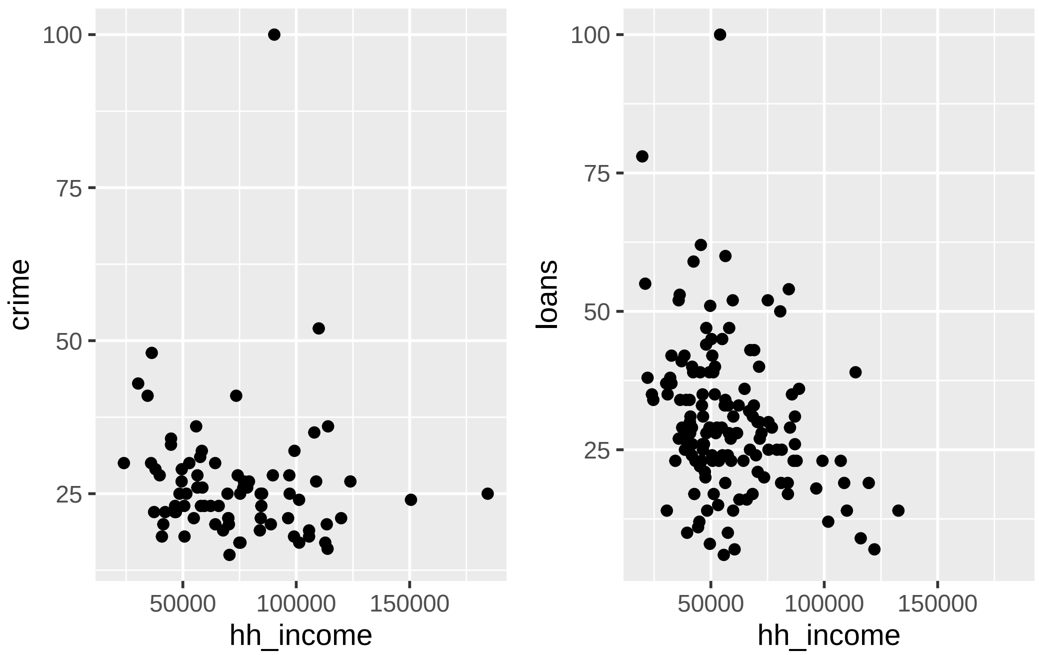
```
library(ggpubr)

p1 = qplot(x = hh_income, y = crime, data = joint)
```

```
Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```

```
p2 = qplot(x = hh_income, y = loans, data = joint)
ggarrange(p1, p2, ncol = 2, nrow = 1)
```

```
Warning: Removed 259 rows containing missing values (`geom_point()`).
```

```
Warning: Removed 197 rows containing missing values (`geom_point()`).
```



Repeat the above steps using the covid data and the ACS data.