Assignment 2

Due at 11:59pm on September 30.

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.

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 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 2021-12-31",
               low_search_volume = TRUE)
names(res)
```

```
[1] "interest_over_time"  "interest_by_country" "interest_by_region"
[4] "interest_by_dma"     "interest_by_city"    "related_topics"
[7] "related_queries"
```

```r
glimpse(res)
```

```
List of 7
 $ interest_over_time :'data.frame':    210 obs. of  7 variables:
  ..$ date    : POSIXct[1:210], format: "2019-12-29" "2020-01-05" ...
  ..$ hits    : int [1:210] 61 62 61 58 57 59 58 59 63 57 ...
  ..$ keyword : chr [1:210] "crime" "crime" "crime" "crime" ...
  ..$ geo     : chr [1:210] "US-IL" "US-IL" "US-IL" "US-IL" ...
  ..$ time    : chr [1:210] "2020-01-01 2021-12-31" "2020-01-01 2021-12-31" "2020-01-01 2021-
  ..$ gprop   : chr [1:210] "web" "web" "web" "web" ...
  ..$ category: int [1:210] 0 0 0 0 0 0 0 0 0 0 ...
 $ interest_by_country: NULL
 $ interest_by_region : NULL
 $ interest_by_dma    :'data.frame':    20 obs. of  5 variables:
  ..$ location: chr [1:20] "Rockford IL" "Chicago IL" "St. Louis MO" "Champaign & Springfiel
  ..$ hits    : int [1:20] 100 94 93 89 85 83 75 75 70 66 ...
  ..$ keyword : chr [1:20] "crime" "crime" "crime" "crime" ...
  ..$ geo     : chr [1:20] "US-IL" "US-IL" "US-IL" "US-IL" ...
  ..$ gprop   : chr [1:20] "web" "web" "web" "web" ...
 $ interest_by_city   :'data.frame':    400 obs. of  5 variables:
  ..$ location: chr [1:400] "Morton Grove" "Byron" "Macomb" "Cahokia" ...
  ..$ hits    : int [1:400] 100 94 86 76 75 74 72 71 71 71 ...
  ..$ keyword : chr [1:400] "crime" "crime" "crime" "crime" ...
  ..$ geo     : chr [1:400] "US-IL" "US-IL" "US-IL" "US-IL" ...
  ..$ gprop   : chr [1:400] "web" "web" "web" "web" ...
 $ related_topics     : NULL
 $ related_queries    :'data.frame':    100 obs. of  6 variables:
  ..$ subject       : chr [1:100] "100" "48" "39" "30" ...
  ..$ related_topics: chr [1:100] "top" "top" "top" "top" ...
  ..$ value         : chr [1:100] "crime chicago" "true crime" "crime news" "american crime"
  ..$ geo           : chr [1:100] "US-IL" "US-IL" "US-IL" "US-IL" ...
  ..$ keyword       : chr [1:100] "crime" "crime" "crime" "crime" ...
  ..$ category      : int [1:100] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "reshapeLong")=List of 4
  .. ..$ varying:List of 1
  .. .. ..- attr(*, "v.names")= chr "value"
  .. .. ..- attr(*, "times")= chr "top"
  .. ..$ v.names: chr "value"
  .. ..$ idvar  : chr "id"
  .. ..$ timevar: chr "related_topics"
 - attr(*, "class")= chr [1:2] "gtrends" "list"
```

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

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

```r
stats <- res$interest_over_time %>%
  group_by(keyword) %>%
  summarise(
    mean_hits   = mean(hits, na.rm = TRUE),
    median_hits = median(hits, na.rm = TRUE),
    var_hits    = var(hits, na.rm = TRUE)
  )

stats
```

```
# A tibble: 2 x 4
  keyword mean_hits median_hits var_hits
  <chr>       <dbl>       <int>    <dbl>
1 crime        57.7          57     63.8
2 loans        66.9          66     68.1
```

The mean number of crime hits was 57.7, with a median of 57 and a variance of 63.8. The mean number of loans hits was 66.9 with a median of 66 and a variance of 68.1.

- 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. - The 10 cities below had the highest search frequency for “loans” during our query time period.

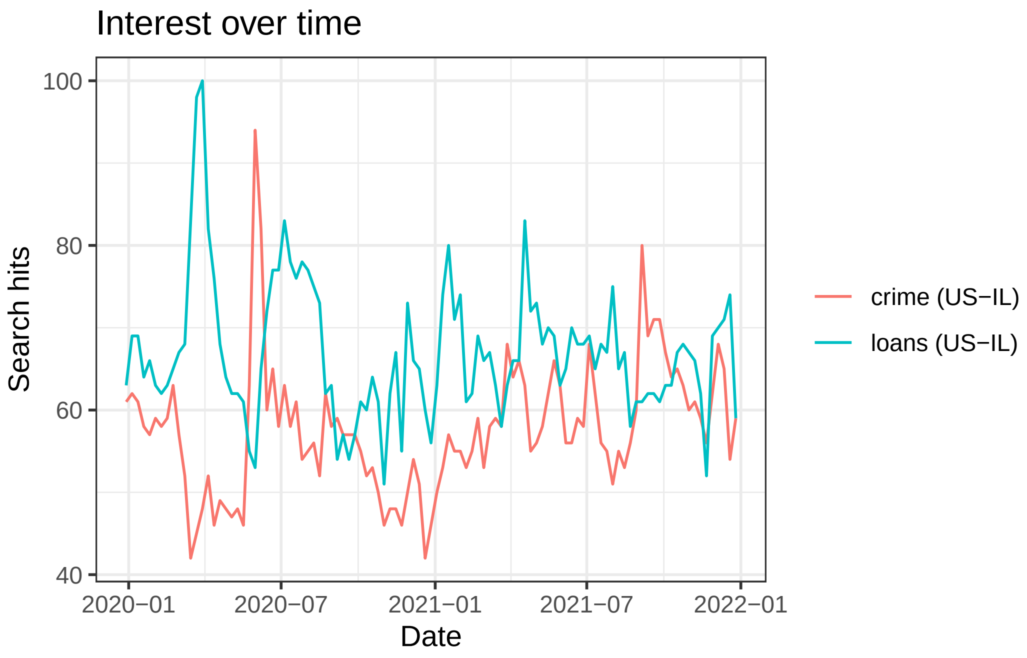
```r
cities <- res$interest_by_city %>%
  filter(keyword == "loans") %>%
  select(location, keyword, hits) %>%
  pivot_wider(names_from = keyword, values_from = hits)

# Top cities by "loans"
top_loans <- cities %>%
  arrange(desc(loans))

head (top_loans, 10) # top 10 cities with the highest search frequency for "loans" below
```

```
# A tibble: 10 x 2
   location           loans
   <chr>              <int>
 1 Robbins              100
 2 Riverdale             91
 3 Ford Heights          85
 4 Rosemont              84
 5 Cahokia               83
 6 Dolton                83
 7 Madison               81
 8 Hazel Crest           79
 9 University Park       76
10 Country Club Hills    75
```

Riverdale had the highest search frequency of loans with 100 hits, with Ford Heights and Rosemont have the second highest search frequency with 93 each.

- Is there a relationship between the search intensities between the two keywords we used?
  - From the time series plot below, there does not seem to be an apparent association between "crime" and "loans" as they show different trends over time. It indicates that interest in loans surged sharply and peaked around April 2020, during the early months of the COVID-19 pandemic. This was likely due to searches related to financial support, unemployment, and loans. Conversely, crime-related searches surged around June 2020, a time characterized by increased media coverage of crime and protests. Afterwards, searches declined, with moderate activity observed. Loan searches were generally higher than crime searches throughout most of 2020, except in June when crime search spiked.

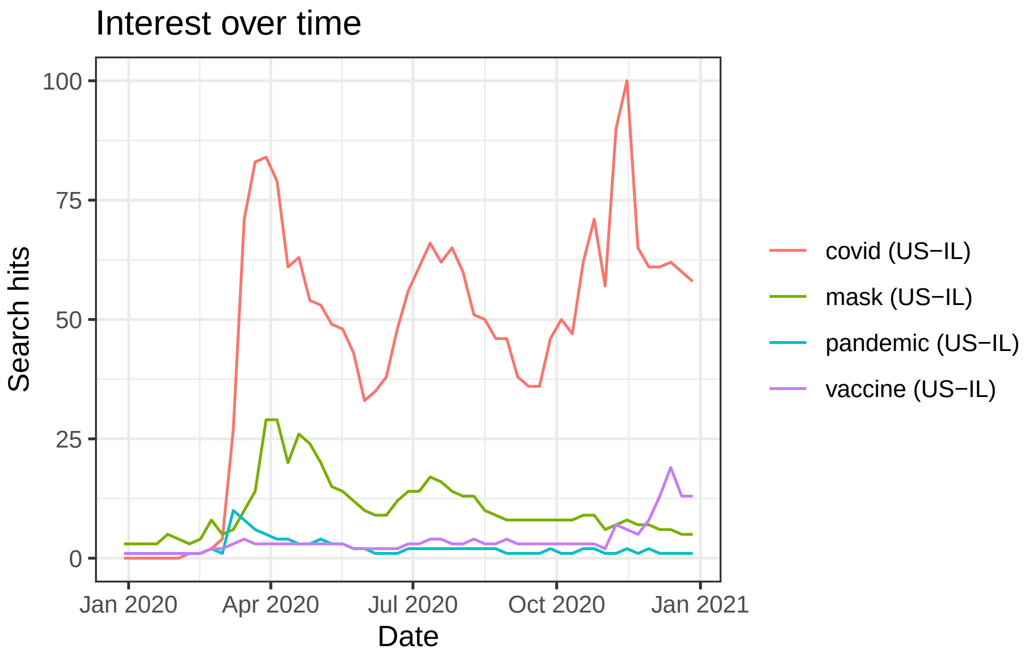
  ```r
  plot(res)
  ```



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.

- Initial search keywords include: covid, mask, pandemic, and vaccine.

```
covid <- gtrends(c("covid", "mask",
                   "pandemic", "vaccine"), # multiple keywords entered
               geo = "US-IL",
               time = "2020-01-01 2020-12-31",
               low_search_volume = TRUE)

plot(covid)
```



- The two keywords of relevance used are: “covid” and “vaccine”.

```
covid_keywords <- gtrends(c("covid",
                            "vaccine"), # two keyword of relevance picked
                          geo = "US-IL",
                          time = "2020-01-01 2020-12-31",
                          low_search_volume = TRUE)


names(covid_keywords)
```

```
[1] "interest_over_time"  "interest_by_country" "interest_by_region"
[4] "interest_by_dma"     "interest_by_city"    "related_topics"
[7] "related_queries"
```

```
glimpse(covid_keywords)
```

```
List of 7
 $ interest_over_time :'data.frame':	106 obs. of  7 variables:
  ..$ date    : POSIXct[1:106], format: "2019-12-29" "2020-01-05" ...
  ..$ hits    : chr [1:106] "0" "0" "0" "0" ...
  ..$ keyword : chr [1:106] "covid" "covid" "covid" "covid" ...
  ..$ geo     : chr [1:106] "US-IL" "US-IL" "US-IL" "US-IL" ...
  ..$ time    : chr [1:106] "2020-01-01 2020-12-31" "2020-01-01 2020-12-31" "2020-01-01 2020-
  ..$ gprop   : chr [1:106] "web" "web" "web" "web" ...
  ..$ category: int [1:106] 0 0 0 0 0 0 0 0 0 0 ...
 $ interest_by_country: NULL
 $ interest_by_region : NULL
 $ interest_by_dma    :'data.frame':    20 obs. of  5 variables:
  ..$ location: chr [1:20] "Peoria-Bloomington IL" "Chicago IL" "Champaign & Springfield-Deca
  ..$ hits    : int [1:20] 100 98 97 96 92 89 87 84 82 82 ...
  ..$ keyword : chr [1:20] "covid" "covid" "covid" "covid" ...
  ..$ geo     : chr [1:20] "US-IL" "US-IL" "US-IL" "US-IL" ...
  ..$ gprop   : chr [1:20] "web" "web" "web" "web" ...
 $ interest_by_city   :'data.frame':    400 obs. of  5 variables:
  ..$ location: chr [1:400] "Wheeler" "Evergreen Park" "Beckemeyer" "Clarendon Hills" ...
  ..$ hits    : int [1:400] 100 81 81 80 79 78 77 76 76 75 ...
  ..$ keyword : chr [1:400] "covid" "covid" "covid" "covid" ...
  ..$ geo     : chr [1:400] "US-IL" "US-IL" "US-IL" "US-IL" ...
  ..$ gprop   : chr [1:400] "web" "web" "web" "web" ...
 $ related_topics     : NULL
 $ related_queries    :'data.frame':    100 obs. of  6 variables:
  ..$ subject       : chr [1:100] "100" "51" "32" "30" ...
  ..$ related_topics: chr [1:100] "top" "top" "top" "top" ...
  ..$ value         : chr [1:100] "covid 19" "covid illinois" "covid testing" "covid cases"
  ..$ geo           : chr [1:100] "US-IL" "US-IL" "US-IL" "US-IL" ...
  ..$ keyword       : chr [1:100] "covid" "covid" "covid" "covid" ...
  ..$ category      : int [1:100] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "reshapeLong")=List of 4
  .. ..$ varying:List of 1
  .. .. ..- attr(*, "v.names")= chr "value"
  .. .. ..- attr(*, "times")= chr "top"
  .. ..$ v.names: chr "value"
  .. ..$ idvar  : chr "id"
  .. ..$ timevar: chr "related_topics"
 - attr(*, "class")= chr [1:2] "gtrends" "list"
```

Answer the following questions for the keywords “covid” and “vaccine”.

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

```r
stats <- covid_keywords$interest_over_time %>%
  group_by(keyword) %>%
  mutate(hits = ifelse(hits == "<1", .5, hits),
         hits = as.numeric(hits)) %>%
  summarise(
    mean_hits   = mean(hits, na.rm = TRUE),
    median_hits = median(hits, na.rm = TRUE),
    var_hits    = var(hits, na.rm = TRUE)
  )

stats
```

```
# A tibble: 2 x 4
  keyword mean_hits median_hits var_hits
  <chr>       <dbl>       <dbl>    <dbl>
1 covid       46.0           50    684.
2 vaccine      3.77           3     12.0
```

The mean number of covid hits was 46.0 with a median of 50 and a variance of 683.7. The mean number of vaccine hits was 3.8 with a median of 3 and a variance of 12.0.

- Which cities (locations) have the highest search frequency for **covid**? 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.
  - The 10 cities below had the highest search frequency for "covid" during our query time period.

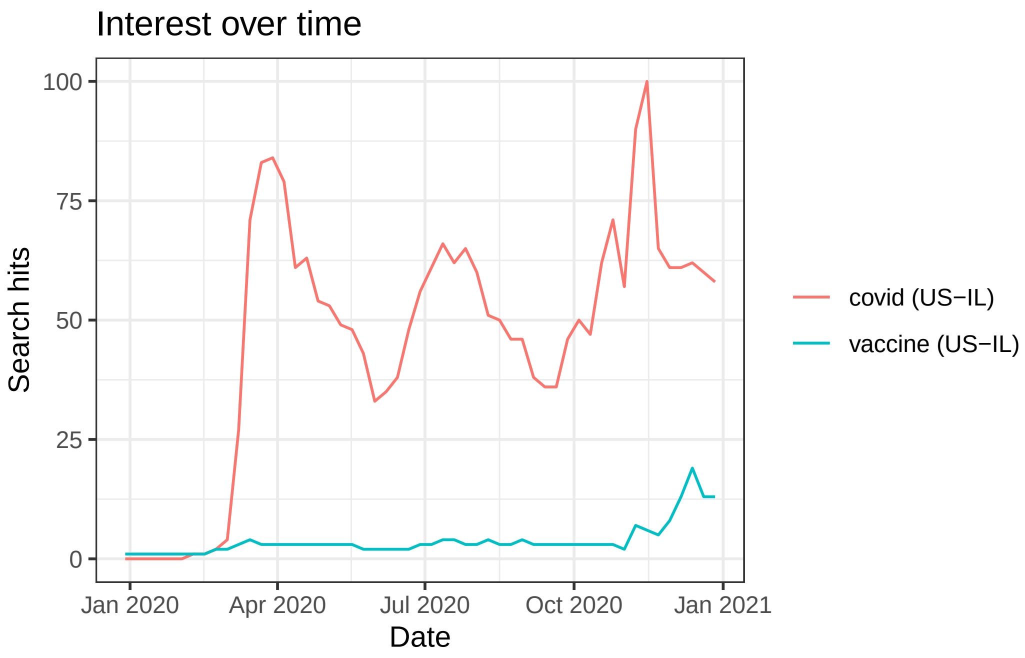
```
cities_covid <- covid_keywords$interest_by_city %>%
  filter(keyword == "covid") %>%
  select(location, keyword, hits) %>%
  pivot_wider(names_from = keyword, values_from = hits)

# Top cities by "covid"
top_covid <- cities_covid %>%
  arrange(desc(covid))

head (top_covid, 10)
```

```
# A tibble: 10 x 2
   location        covid
   <chr>           <int>
 1 Wheeler           100
 2 Evergreen Park     81
 3 Beckemeyer         81
 4 Clarendon Hills    80
 5 Geneva             79
 6 Oak Lawn           78
 7 Westfield          77
 8 Albany             76
 9 New Burnside       76
10 Winnetka           75
```

Wheeler had the highest search frequency of covid with 100 hits, with Evergreen Park and Beckemeyer have the second highest search frequency with 81 each.

- Is there a relationship between the search intensities between the two keywords we used?
    - Again, from the time series plot below, there does not seem to be an apparent association between “covid” and “vaccine” as they show different trends over time. It looks like “covid” searches surged the most and sharply twice - the outbreak in spring 2020 (around February/March) and again towards the end of 2020 (November/December) which might be attributable to the onset of vaccine approvals and distribution.

```r
plot(covid_keywords)
```



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!

```
cs_key <- read_rtf("~/Desktop/SURV 727/sg_key.rtf")
```

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. Documentation for the 5-year ACS API can be found here: https://www.census.gov/data/developers/data-sets/acs-5year.html. The information about the variables used here can be found here: https://api.census.gov/data/2022/acs/acs5/variables.html.

```
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) # to view first few rows
```

```
  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
```

```
# view(acs_il) # review entire data sets
```

Convert values that represent missings to NAs.

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

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
```

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)

head(acs_il)
```

```
  state place                        NAME  pop  age hh_income income
1    17 15261 Coatsburg village, Illinois  180 35.6     55714  27821
2    17 15300    Cobden village, Illinois 1018 44.2     38750  19979
3    17 15352      Coffeen city, Illinois  640 33.4     35781  26697
```

```
4     17 15378   Colchester city, Illinois 1347 42.2     43942  24095
5     17 15469    Coleta village, Illinois  230 27.7     56875  23749
6     17 15495    Colfax village, Illinois 1088 32.5     58889  24861
```

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.

- Adding a new variable “location”

```r
acs_il <- acs_il |>  # adding a new variable "location"
  mutate (location = str_extract(NAME, "[^,]+"),
        location = str_remove(location, "village|city|\\stown|CDP"),
        location = str_trim(location))
```

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.

8 cities that do not appear in both data sets

```r
asc_crime_loans <- res$interest_by_city |> as_tibble() |>
  as_tibble() |>
  mutate(location = str_replace(location, "Saint\\s", "St. "),
         location = str_replace(location, "Sainte\\s", "Ste. "),
         location = str_remove(location, "Fort ") ) |>
  # joining crime/loans google trends data with ACS
  left_join(acs_il, relationship = "many-to-many")
```

```
Joining with `by = join_by(location)`
```

```r
# cities that do not appear in both data sets
asc_crime_loans |>
  filter(is.na(NAME)) |>
  select(location)  # unmatches between both data sets
```

```
# A tibble: 7 x 1
  location
  <chr>
1 Barnhill
2 Massbach
3 Rosamond
4 Colusa
5 Berwick
6 Ledbetter
7 Ancona
```

```
count(asc_crime_loans)
```

```
# A tibble: 1 x 1
      n
  <int>
1   403
```

- Joining Google Trends and ACS data sets

  ```
  asc_crime_loans <- res$interest_by_city |> as_tibble() |>
    as_tibble() |>
    mutate(location = str_replace(location, "Saint\\s", "St. "),
           location = str_replace(location, "Sainte\\s", "Ste. "),
           location = str_remove(location, "Fort ") ) |>
    # join with ACS
    left_join(acs_il, relationship = "many-to-many")
  ```

  ```
  Joining with `by = join_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?

  Income and Crime: Cities with higher median household income have a slightly higher mean search popularity (66.5) compared to cities with lower income (66.1) – indicating that as household income increases, interest in crime search popularity increases. This could be due to higher income areas being more vigilant about crimes.

  Income and Loans: Cities with higher median income have a slightly lower mean search popularity for loans (73.8) than cities with lower income (76.0) – an indication that higher-income cities have better access to loans or financial opportunities.

```r
stats_asc_crime_loans <- asc_crime_loans |>
  mutate(hh_income_median = ifelse(hh_income > median(hh_income, na.rm=TRUE),
                                   "Higher", "Lower")) |>
  filter(!is.na(hh_income_median)) |>
  group_by(hh_income_median, keyword) |>
  summarise(mean_hits = mean(hits, na.rm=TRUE)) |>
  pivot_wider(names_from = keyword, values_from = mean_hits)
```

```
`summarise()` has grouped output by 'hh_income_median'. You can override using
the `.groups` argument.
```

```r
stats_asc_crime_loans
```

```
# A tibble: 2 x 3
# Groups:   hh_income_median [2]
  hh_income_median crime loans
  <chr>            <dbl> <dbl>
1 Higher            66.5  66.1
2 Lower             66.1  70.2
```

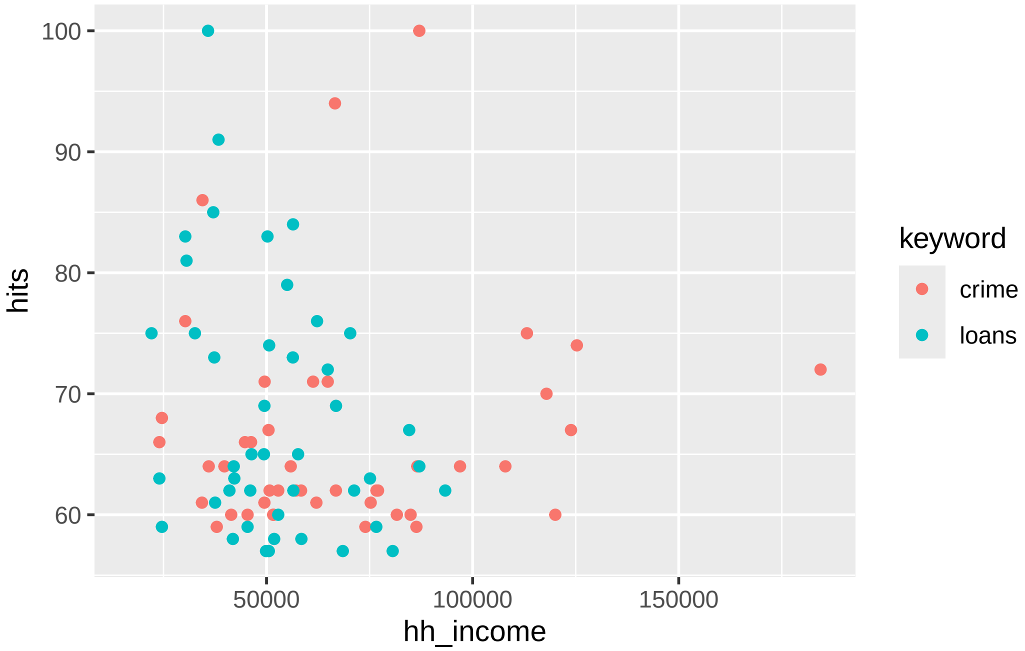
- Is there a relationship between the median household income and the search popularity of the Google trends terms? Describe the relationship and use a scatterplot with `qplot()`.

  There seems to be linear relationships between median household income and the search popularity of crime and loans. Similar to the observation in the previous question, as median household income increases, crime searches seem to increase. Conversely, as median household income increases, there appears to be a slight decrease in loan searches.

  - Median household income and income_loans plot

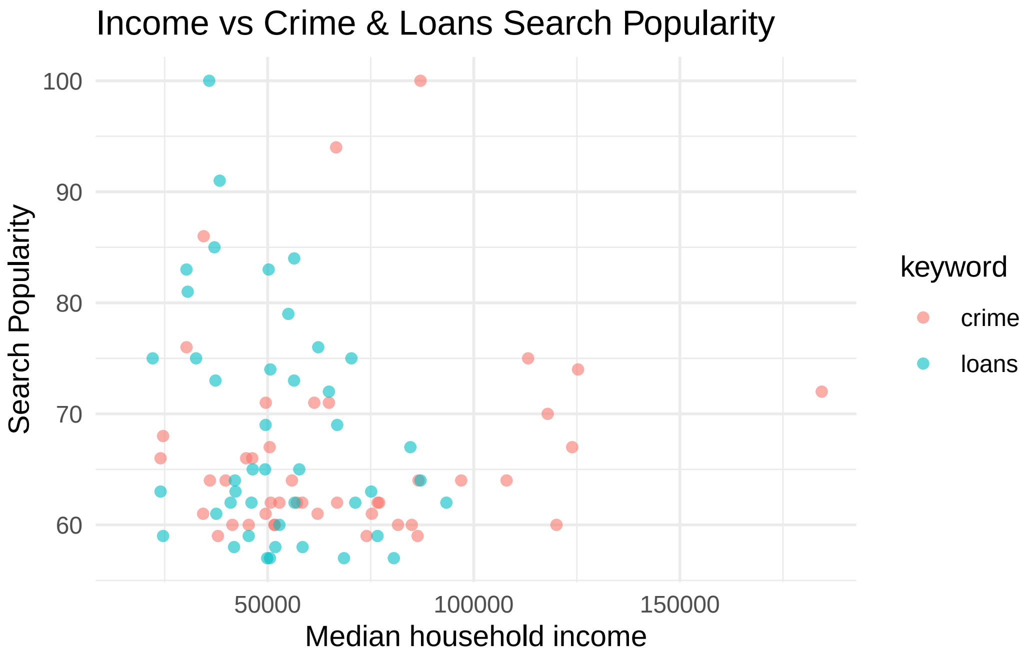
```r
asc_crime_loans |>
  ggplot(aes(x=hh_income, y=hits, color=keyword)) +
  geom_point()
```

```
Warning: Removed 315 rows containing missing values or values outside the scale range
(`geom_point()`).
```



```
ggplot(asc_crime_loans, aes(x = hh_income, y = hits, color = keyword)) +
  geom_point(alpha = 0.6) +
  labs(
    title = "Income vs Crime & Loans Search Popularity",
    x = "Median household income",
    y = "Search Popularity"
  ) +
  theme_minimal()
```

```
Warning: Removed 315 rows containing missing values or values outside the scale range
(`geom_point()`).
```

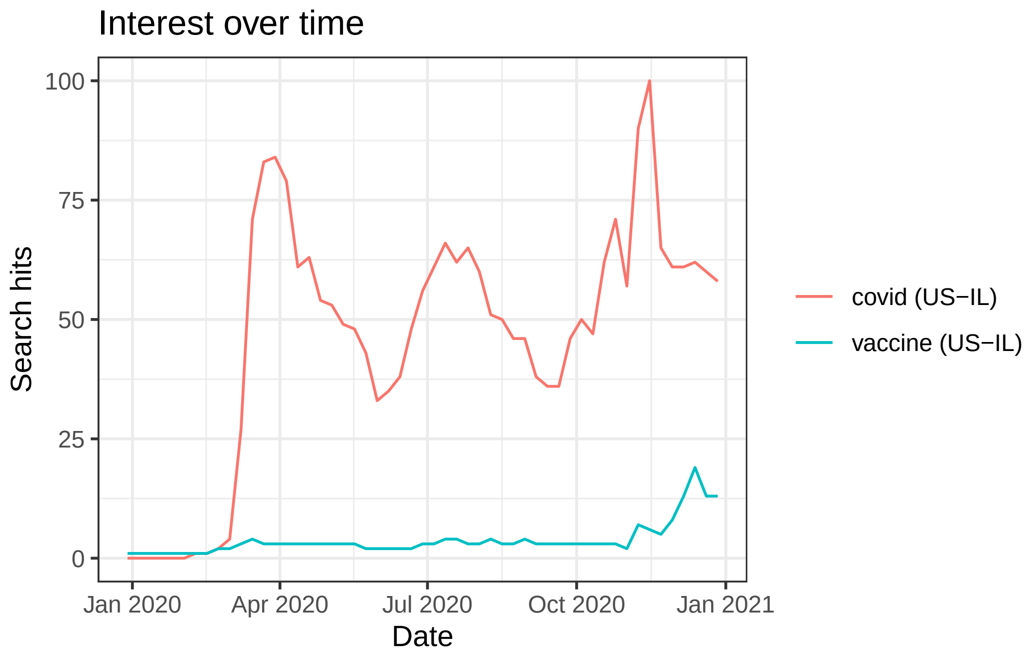


Repeat the above steps using the covid data and the ACS data.

- 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.

Repeating the same steps for COVID keywords

```
# Pull covid keywords
plot(covid_keywords)
```



9 cities that do not appear in both data sets

```r
asc_covid_vaccine <- covid_keywords$interest_by_city |> as_tibble() |>
  as_tibble() |>
  mutate(location = str_replace(location, "Saint\\s", "St. "),
         location = str_replace(location, "Sainte\\s", "Ste. "),
         location = str_remove(location, "Fort ") ) |>
  # joining covid/vaccine google trends data with ACS
  left_join(acs_il, relationship = "many-to-many")
```

```
Joining with `by = join_by(location)`
```

```r
# cities that do not appear in both data sets
asc_covid_vaccine |>
  filter(is.na(NAME)) |>
  select(location)  # unmatches between both data sets
```

```
# A tibble: 9 x 1
  location
  <chr>
1 Graymont
2 Plato Center
3 Village of Lakewood
```

```
4 Gulf Port
5 Chana
6 Village of Lakewood
7 Chana
8 Colusa
9 Graymont
```

```
count(asc_covid_vaccine )
```

```
# A tibble: 1 x 1
      n
  <int>
1   404
```

- 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?

```
stats_asc_covid_vaccine <-
  asc_covid_vaccine |>
  mutate(hh_income_median =
           ifelse(hh_income > median(hh_income, na.rm=TRUE),
                  "Higher", "Lower")) |>
  filter(!is.na(hh_income_median)) |>
  group_by(hh_income_median, keyword) |>
  reframe(mean_hits = mean(hits, na.rm=TRUE)) |>
  pivot_wider(names_from = keyword, values_from = mean_hits)

stats_asc_covid_vaccine
```

```
# A tibble: 2 x 3
  hh_income_median covid vaccine
  <chr>            <dbl>   <dbl>
1 Higher            64.9    33.7
2 Lower             64.7    36.4
```

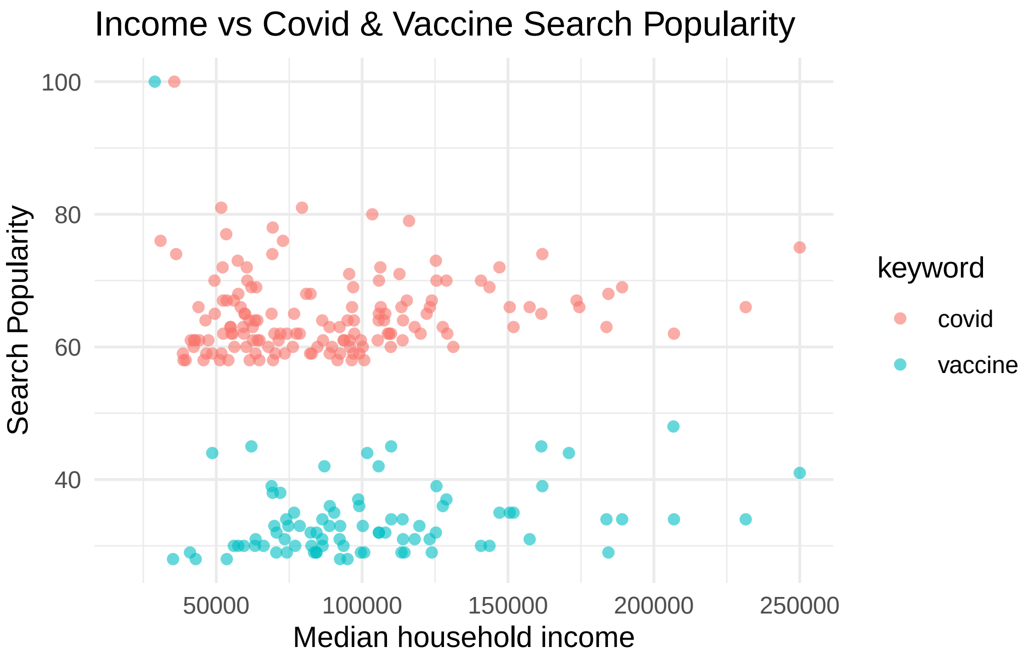
Income and Covid: Cities with higher median household income have a slightly higher mean search popularity (64.9) compared to cities with lower income (64.7) – indicating that as household income increases, interest in covid search popularity increases. This could be due to higher education levels of people in wealthier cities feeling confident to research covid on their own time.

Income and Loans: Cities with higher median income have a lower mean search popularity for vaccine (33.8) than cities with lower income (36.9) – an indication that higher-income cities have more remote workers, while in lower income cities workers may be required to be in person, and would have needed to search vaccine requirements so they could return to work.

- Median household income and covid_vaccine plot

```
ggplot(asc_covid_vaccine, aes(x = hh_income, y = hits, color = keyword)) +
  geom_point(alpha = 0.6) +
  labs(
    title = "Income vs Covid & Vaccine Search Popularity",
    x = "Median household income",
    y = "Search Popularity"
  ) +
  theme_minimal()
```

```
Warning: Removed 172 rows containing missing values or values outside the scale range
(`geom_point()`).
```



There appear to be slight positive linear trends with both covid and vaccine, where searches increase as income increases.