Assignment 2

Sungjoo Cho, Catherine (Kate) Lamoreaux

2023-10-03

Packages

Pulling from APIs

```r
# pulling from APIs
res <- gtrends(c("crime", "loans"),
               geo = "US-IL",
               time = "2020-01-01 2020-12-31",
               low_search_volume = TRUE)

# transforming the `data.frame` into a `tibble`
str(res)
```

```
## List of 7
##  $ interest_over_time :'data.frame': 104 obs. of  7 variables:
##   ..$ date    : POSIXct[1:104], format: "2020-01-05" "2020-01-12" ...
##   ..$ hits    : int [1:104] 63 61 59 60 59 59 62 60 57 51 ...
##   ..$ keyword : chr [1:104] "crime" "crime" "crime" "crime" ...
##   ..$ geo     : chr [1:104] "US-IL" "US-IL" "US-IL" "US-IL" ...
##   ..$ time    : chr [1:104] "2020-01-01 2020-12-31" "2020-01-01 2020-12-31" "2020-01-01 2020-12-31"
##   ..$ gprop   : chr [1:104] "web" "web" "web" "web" ...
##   ..$ category: int [1:104] 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" "St. Louis MO" "Chicago IL" "Quincy IL-Hannibal MO-Keokuk I
##   ..$ hits    : int [1:20] 100 96 95 90 81 81 80 80 75 75 ...
##   ..$ 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] "Braceville" "Hampshire" "Anna" "South Jacksonville" ...
##   ..$ hits    : int [1:400] 100 74 71 62 60 60 59 55 54 52 ...
##   ..$ 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" "89" "46" "37" ...
##   ..$ related_queries: chr [1:100] "top" "top" "top" "top" ...
##   ..$ value          : chr [1:100] "chicago crime" "crime rate" "true crime" "crime news" ...
##   ..$ 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
##   .. .. ..$ value: chr "top"
##   .. .. ..- attr(*, "v.names")= chr "value"
##   .. .. ..- attr(*, "times")= chr "top"
##   .. ..$ v.names: chr "value"
##   .. ..$ idvar  : chr "id"
##   .. ..$ timevar: chr "related_queries"
##  - attr(*, "class")= chr [1:2] "gtrends" "list"
```

```r
res_time <- as_tibble(res$interest_over_time)

glimpse(res_time)
```

```
## Rows: 104
## Columns: 7
## $ date     <dttm> 2020-01-05, 2020-01-12, 2020-01-19, 2020-01-26, 2020-02-02, ~
## $ hits     <int> 63, 61, 59, 60, 59, 59, 62, 60, 57, 51, 40, 42, 50, 52, 47, 5~
## $ keyword  <chr> "crime", "crime", "crime", "crime", "crime", "crime", "crime"~
## $ geo      <chr> "US-IL", "US-IL", "US-IL", "US-IL", "US-IL", "US-IL", "US-IL"~
## $ time     <chr> "2020-01-01 2020-12-31", "2020-01-01 2020-12-31", "2020-01-01~
## $ gprop    <chr> "web", "web", "web", "web", "web", "web", "web", "web", "web"~
## $ category <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

Answer the following questions for the keywords "crime" and "loans".

- Find the mean, median and variance of the search hits for the keywords.
    - The Table 1 below shows the mean, median, and variance of the search hits for the both keywords 'crime' and 'loans'.

```r
# mean, median and variance of the search hits for the keywords
stat_keywords <- res_time %>%
  group_by(keyword) %>%
  summarize(mean = mean(hits),
            median = median(hits),
            variance = var(hits))

stat_keywords
```

```
## # A tibble: 2 x 4
##   keyword  mean median variance
##   <chr>   <dbl>  <dbl>    <dbl>
## 1 crime    54.8     55     69.7
## 2 loans    66.2     65    104.
```

```r
knitr::kable(stat_keywords, caption =
               "Search-hits Statistics: 'crime' and 'loans'")
```

Table 1: Search-hits Statistics: 'crime' and 'loans'

keyword	mean	median	variance
crime	54.76923	55	69.67119
loans	66.17308	65	103.87142

- 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.
  - Table 2 below shows the top six Illinois cities with the highest search frequency for `loans`.

```r
# using pivot_wider
res_city <- res$interest_by_city %>%
  pivot_wider(names_from = keyword,
              values_from = hits)

# changing NA values to 0 for loans and crime
res_city['loans'][is.na(res_city['loans'])] <- 0
res_city['crime'][is.na(res_city['crime'])] <- 0

# sorting
res_city <- res_city[order(-res_city$loans), ]

knitr::kable(head(res_city), caption =
               "IL Cities with the Highest Search Frequency for Loans")
```

Table 2: IL Cities with the Highest Search Frequency for Loans

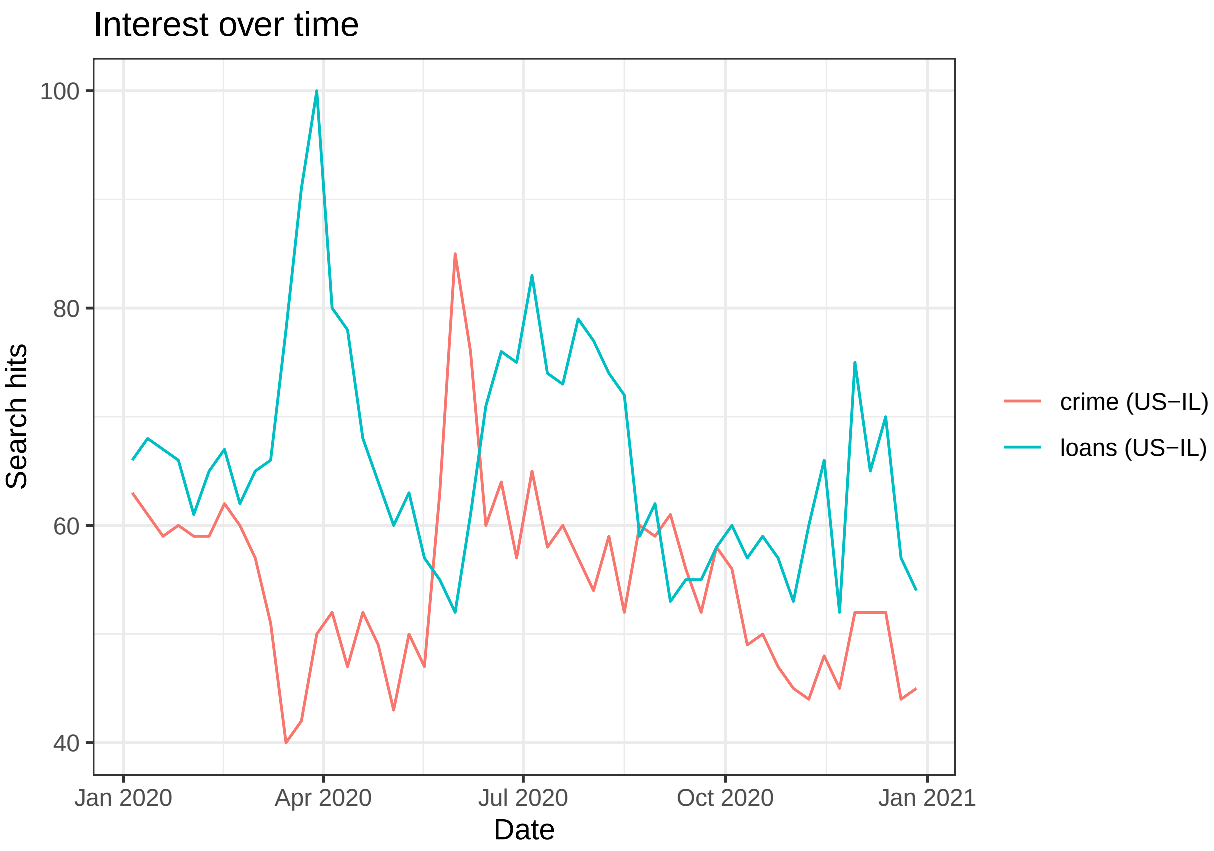
location	geo	gprop	crime	loans
Alorton	US-IL	web	0	100
Braceville	US-IL	web	100	97
Long Lake	US-IL	web	0	95
New Athens	US-IL	web	0	90
Jonesboro	US-IL	web	0	87
Rosemont	US-IL	web	0	85

- Is there a relationship between the search intensities between the two keywords we used?
  - The correlation between the search intensities of the two keywords is low and negative. This means that there is not a strong linear relationship between searches for 'crime' and 'loans' over time, but generally, as the search intensity for 'loans' increases, the search intensity for 'crime' tends to decline.
  - The plot below illustrates this inverse relationship, but with more complexity. There are points in the year 2020 where the search teams appear to be inversely related, but there are other points of the year where the trendlines intersect (late May and September) or appear to run parallel to one another (after October). This change in this relationship between keywords over time suggests exogenous variables influencing the nature of these trends.

```r
# correlation
cor(res_city$crime, res_city$loans)
```

```
## [1] -0.1159915
```

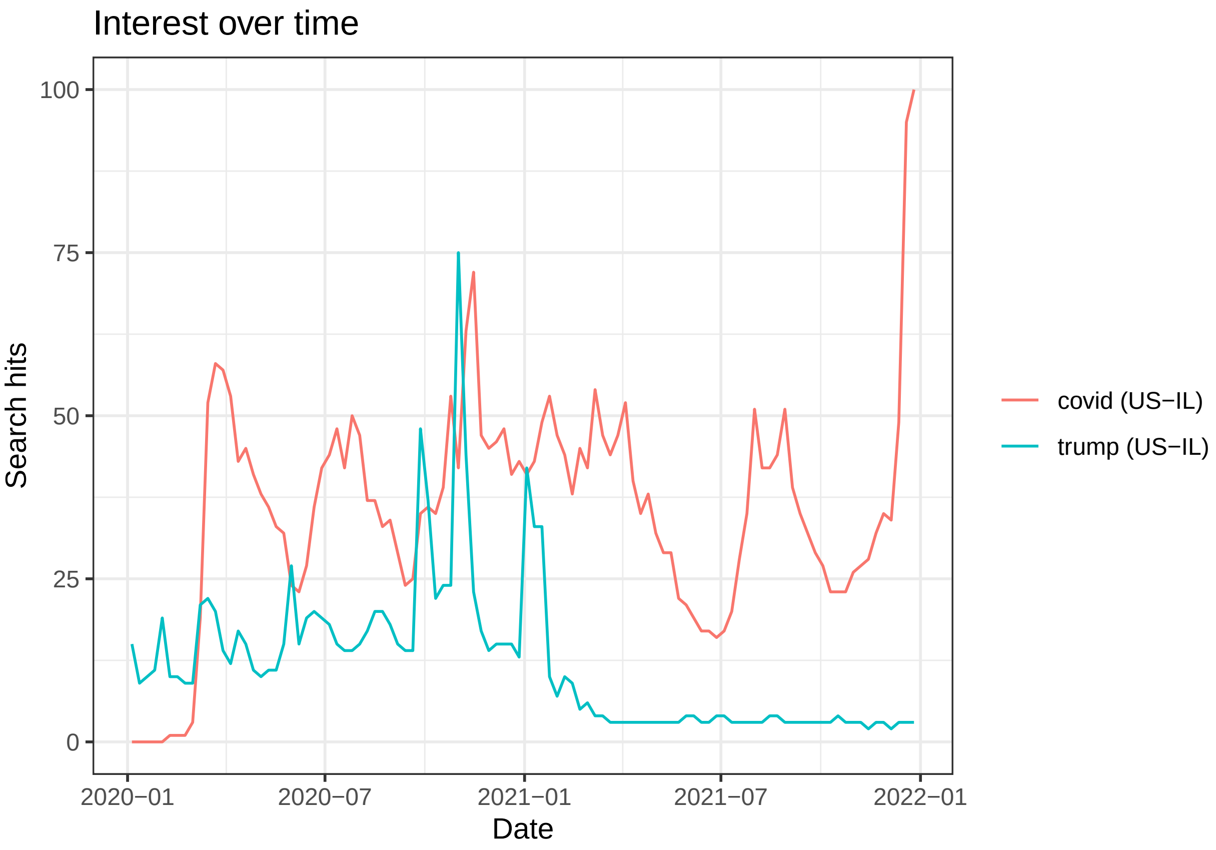
```r
# plot of the number of search hits changes over time
plot(res)
```



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.

- We tried several different combinations of keywords, including “trump”, “death”, “mask”, and “virus.” We found that “death” and “mask” were not searched nearly as frequently as “covid.” “Virus” searches peaked early, during the onset of the U.S. lockdowns, but rapidly decreased, stabilizing around June 2020. In contrast, searches for “trump” remained high throughout 2020.

```r
#Commenting out code above to see if multiple keywords break the code.
res_covid <- gtrends(c("covid", "trump"),
               geo = "US-IL",
               time = "2020-01-01 2021-12-31",
               low_search_volume = TRUE)
plot(res_covid)
```



```r
str(res_covid)
```

```
## List of 7
##  $ interest_over_time :’data.frame’: 208 obs. of  7 variables:
##   ..$ date    : POSIXct[1:208], format: "2020-01-05" "2020-01-12" ...
##   ..$ hits    : chr [1:208] "0" "0" "0" "0" ...
##   ..$ keyword : chr [1:208] "covid" "covid" "covid" "covid" ...
##   ..$ geo     : chr [1:208] "US-IL" "US-IL" "US-IL" "US-IL" ...
##   ..$ time    : chr [1:208] "2020-01-01 2021-12-31" "2020-01-01 2021-12-31" "2020-01-01 2021-12-31"
##   ..$ gprop   : chr [1:208] "web" "web" "web" "web" ...
##   ..$ category: int [1:208] 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] "Chicago IL" "Peoria-Bloomington IL" "Davenport IA-Rock Island-Moline IL"
##   ..$ hits    : int [1:20] 100 94 91 90 89 87 82 82 77 71 ...
##   ..$ 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] "Oak Lawn" "Northbrook" "Wheaton" "Highland Park" ...
##   ..$ hits    : int [1:400] 100 98 92 90 89 89 88 88 87 86 ...
##   ..$ 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" "71" "69" "63" ...
##   ..$ related_queries: chr [1:100] "top" "top" "top" "top" ...
##   ..$ value          : chr [1:100] "covid 19" "covid vaccine" "vaccine" "illinois covid" ...
##   ..$ 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
##   .. .. ..$ value: chr "top"
##   .. .. ..- attr(*, "v.names")= chr "value"
##   .. .. ..- attr(*, "times")= chr "top"
##   .. ..$ v.names: chr "value"
##   .. ..$ idvar  : chr "id"
##   .. ..$ timevar: chr "related_queries"
##  - attr(*, "class")= chr [1:2] "gtrends" "list"
```

```
# transforming the `data.frame` into a `tibble`
res_covid_time <- as_tibble(res_covid$interest_over_time)
head(res_covid_time)
```

```
## # A tibble: 6 x 7
##   date                hits  keyword geo   time                  gprop category
##   <dttm>              <chr> <chr>   <chr> <chr>                 <chr>    <int>
## 1 2020-01-05 00:00:00 0     covid   US-IL 2020-01-01 2021-12-31 web          0
## 2 2020-01-12 00:00:00 0     covid   US-IL 2020-01-01 2021-12-31 web          0
## 3 2020-01-19 00:00:00 0     covid   US-IL 2020-01-01 2021-12-31 web          0
## 4 2020-01-26 00:00:00 0     covid   US-IL 2020-01-01 2021-12-31 web          0
## 5 2020-02-02 00:00:00 0     covid   US-IL 2020-01-01 2021-12-31 web          0
## 6 2020-02-09 00:00:00 <1    covid   US-IL 2020-01-01 2021-12-31 web          0
```

```
# changing '<1' values to 0 for hits values
res_covid_time$hits <- ifelse(res_covid_time$hits == '<1', 0,
                              res_covid_time$hits)
res_covid_time$hits <- as.integer(res_covid_time$hits)
str(res_covid_time)
```

```
## tibble [208 x 7] (S3: tbl_df/tbl/data.frame)
##  $ date    : POSIXct[1:208], format: "2020-01-05" "2020-01-12" ...
##  $ hits    : int [1:208] 0 0 0 0 0 0 0 1 3 19 ...
##  $ keyword : chr [1:208] "covid" "covid" "covid" "covid" ...
##  $ geo     : chr [1:208] "US-IL" "US-IL" "US-IL" "US-IL" ...
##  $ time    : chr [1:208] "2020-01-01 2021-12-31" "2020-01-01 2021-12-31" "2020-01-01 2021-12-31" "20
##  $ gprop   : chr [1:208] "web" "web" "web" "web" ...
##  $ category: int [1:208] 0 0 0 0 0 0 0 0 0 0 ...
```

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

- Find the mean, median and variance of the search hits for the keywords.
  - The Table 3 below shows the mean, median, and variance of the search hits for the keywords “covid” and “trump”.

```r
# mean, median and variance of the search hits for the keywords
stat_covid_keywords <- res_covid_time %>%
  group_by(keyword) %>%
  summarize(mean = mean(hits),
            median = median(hits),
            variance = var(hits))

knitr::kable(stat_covid_keywords, caption = "Statistics of the search hits for the keywords")
```

Table 3: Statistics of the search hits for the keywords

keyword	mean	median	variance
covid	35.63462	36.5	302.4283
trump	12.03846	10.0	134.1538

- 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 keywords 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.
  - Table 4 below shows the top six Illinois cities with the highest search frequency for `covid`.

```r
res_covid_city<- as_tibble(res_covid$interest_by_city)

# changing NA values to 0 for hits
res_covid_city['hits'][is.na(res_covid_city['hits'])] <- 0

# We found in running this analysis that Google Trends often gave us two cities
# named Windsor for covid searches. Illinois has two cities called Windsor,
# with the larger of these also known as New Windsor. Based on the higher number of searches, we rename
res_covid_city[res_covid_city$location == "Windsor" & res_covid_city$hits ==63, "location"] <- "New Win

# using pivot_wider and changing NA values to 0
res_covid_city<- res_covid_city %>%
 pivot_wider(names_from = keyword,
            values_from = hits,
            values_fill = 0)

# sorting
res_covid_city <- res_covid_city[order(-res_covid_city$covid), ]

knitr::kable(head(res_covid_city), caption = "IL Cities with the Highest Search Frequency for Covid")
```

Table 4: IL Cities with the Highest Search Frequency for Covid

location	geo	gprop	covid	trump
Oak Lawn	US-IL	web	100	0
Northbrook	US-IL	web	98	65
Wheaton	US-IL	web	92	53
Highland Park	US-IL	web	90	0

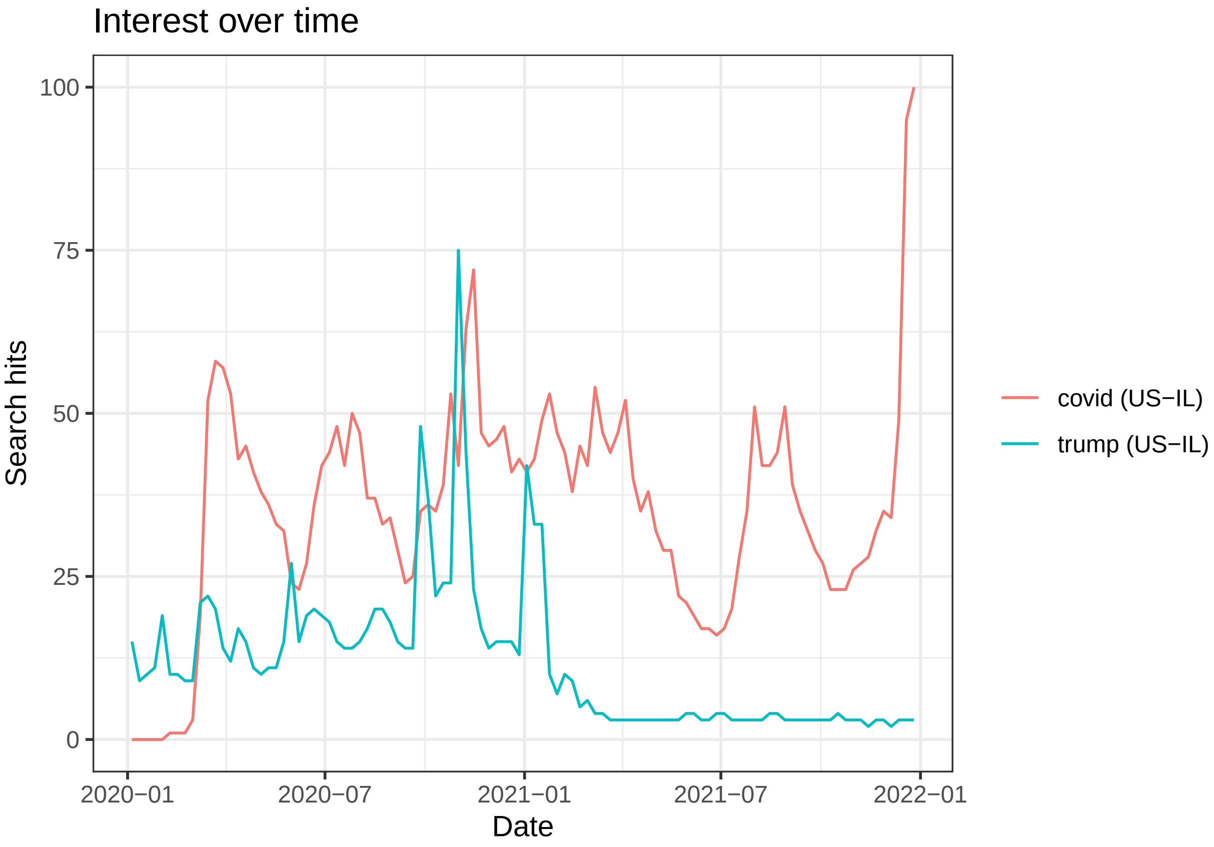
location	geo	gprop	covid	trump
Lake Forest	US-IL	web	89	0
Western Springs	US-IL	web	89	0

- Is there a relationship between the search intensities between the two keywords we used?
    - The correlation between the search intensities of the two keywords is negative and of moderate strength. This means that generally, as the search frequency for 'covid' increases, the search frequency for 'trump' decreases.
    - Like the plot of crime and loans searches, the plot below demonstrates a less striaghtforward and more complex relationship between the two searches, suggesting external factors influence both trends. There are times in 2020 when the search trends inversely mirror one another (April, July-August, November-December), times when they intersect (October-November), and times when they appear to be following the same course (September).

```
# correlation
cor(res_covid_city$covid, res_covid_city$trump)
```

```
## [1] -0.4625865
```

```
# plot of the number of search hits changes over time
plot(res_covid)
```



Google Trends + ACS

Pulling Data

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.

```
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
```

```
# rename the socio-demographic variables
acs_il <- acs_il %>%
  rename(pop = B01001_001E,
         age = B06002_001E,
         hh_income = B19013_001E,
         income = B19301_001E)
```

Cleaning NAME variable in ACS data

- We added a new variable ‘location’ to the ACS data that only includes city names in order to merge this data set with the Google Trends data.

```
# Cleaning NAME in ACS data by adding location variable to ACS

acs_il$location <-  gsub(", .*", "", acs_il$NAME)
```

```
acs_il$location <- gsub("(city|village|CDP|town)", "", acs_il$location)
acs_il$location <- trimws(acs_il$location, "right")
head(acs_il)
```

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

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.
  - In the below tables, we are able to see how many cities appear or don't appear in both datasets. Those categorized as `FALSE` under `Google Trends IL Cities in ACS Data` are the number of cities that appear in Google Trends data, but not in the ACS data. Those categorized as `TRUE` under `Google Trends IL Cities in ACS Data` are the number of cities that appear in both the Google Trends data and the ACS data. Similarly, those categorized as `FALSE` under `ACS IL Cities in Google Trends Data` are the number of cities that appear in the ACS data, but not in the Google Trends data. Those categorized as `TRUE` under `ACS IL Cities in Google Trends Data` are the number of cities that appear in both the ACS data and the Google Trends data .

```
# Are any of the locations in our search data also in our ACS data?
# If yes, will print TRUE.
any(res_city$location %in% acs_il$location)
```

```
## [1] TRUE
```

```
# Printing how many cities don't appear in both data sets
paste("Google Trends IL Cities in ACS Data")
```

```
## [1] "Google Trends IL Cities in ACS Data"
```

```
(summary(res_city$location %in% acs_il$location))
```

```
##    Mode   FALSE    TRUE
## logical      13     335
```

```
paste("ACS IL Cities in Google Trends Data")
```

```
## [1] "ACS IL Cities in Google Trends Data"
```

```
(summary(acs_il$location %in% res_city$location))
```

```
##    Mode   FALSE    TRUE
## logical    1127     339
```

```
# Doing an inner join, only keeping variables common to both datasets
res_city_acs <- inner_join(res_city, acs_il,
                           by = join_by("location" == "location"))

# Printing the number of rows in each dataset to QC our merge matches
# the numbers in our logical table above
nrow(res_city)
```

```
## [1] 348
```

```
nrow(acs_il)
```

```
## [1] 1466
```

```
nrow(res_city_acs)
```

```
## [1] 339
```

- 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?
  - Table 5 below shows 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.

```
# Compute the mean of the search popularity for both keywords by income group
popsearchmean <- res_city_acs %>%
  mutate(high_hh_income = ifelse(hh_income > mean(hh_income, na.rm = TRUE),
                                 "Above", "Below")) %>%
  group_by(high_hh_income) %>%
  summarize(mean_pop_crime = mean(crime),
            mean_pop_loans = mean(loans))

knitr::kable(popsearchmean, caption = "Popularity of Crime and Loans Searches in IL Cities Above and Bel
```

Table 5: Popularity of Crime and Loans Searches in IL Cities Above and Below the Average Median Household Income

high_hh_income	mean_pop_crime	mean_pop_loans
Above	12.046875	17.11719
Below	6.219048	19.76667

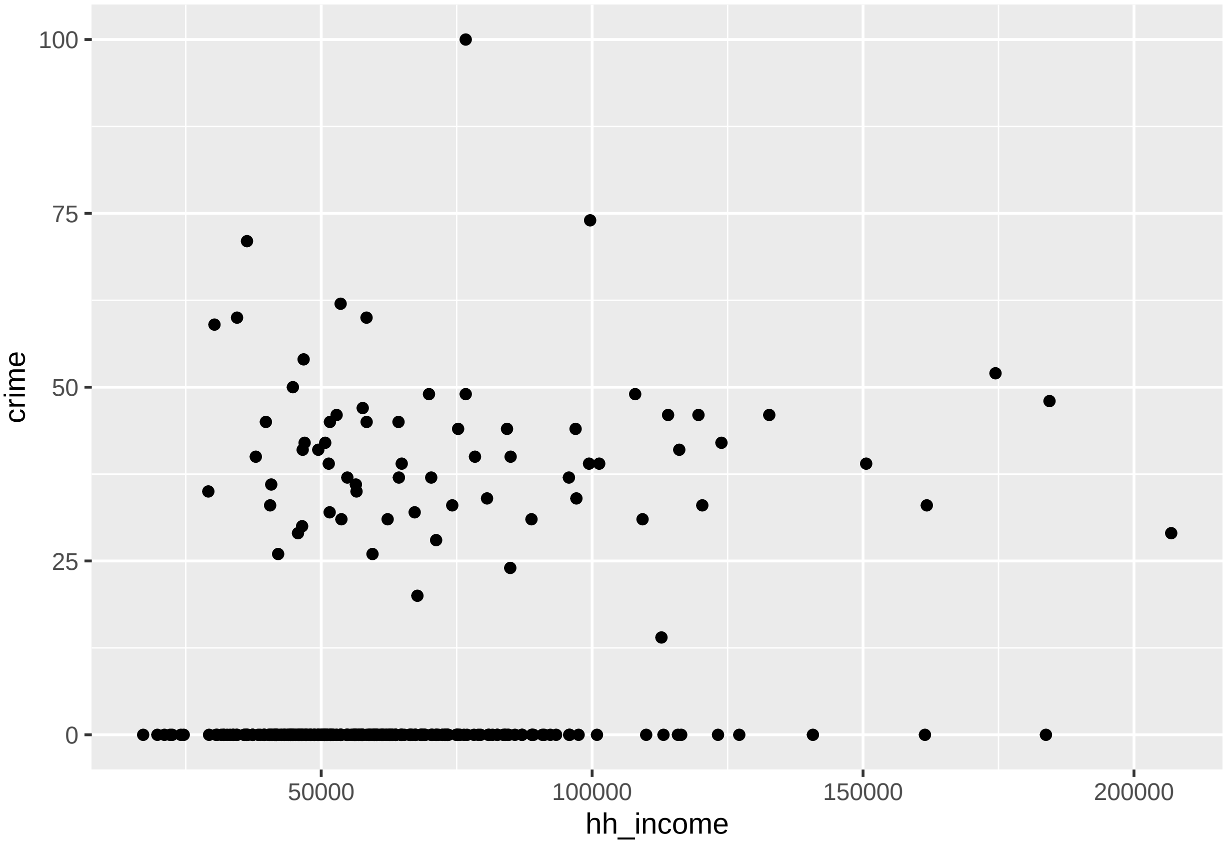
high_hh_income	mean_pop_crime	mean_pop_loans
NA	0.000000	0.00000

- 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()`.
  - Most of the searches for crime appear to be clustered around lower income levels. There appears to be a slightly negative relationship between searches for crime and median household income.
  - Similarly, most of the searches of loans appear to be clustered around lower income levels. However, in this case there is a clearer, stronger downward trajectory of loan searches as income increases.
  - The searches seen together with income in different shades of blue paint a clearer picture: there are few high income households searching for loans, but the same isn't true for crime. There appears to be more variability in crime searches across income levels. Meanwhile, there are a few cities with high search frequencies of both crime and loans, which tend to be low or middle income.

```r
# plot for a relationship between hh_income and crime
res_city_acs %>%
qplot(x = hh_income, y = crime, data = .,
      geom = "auto")
```

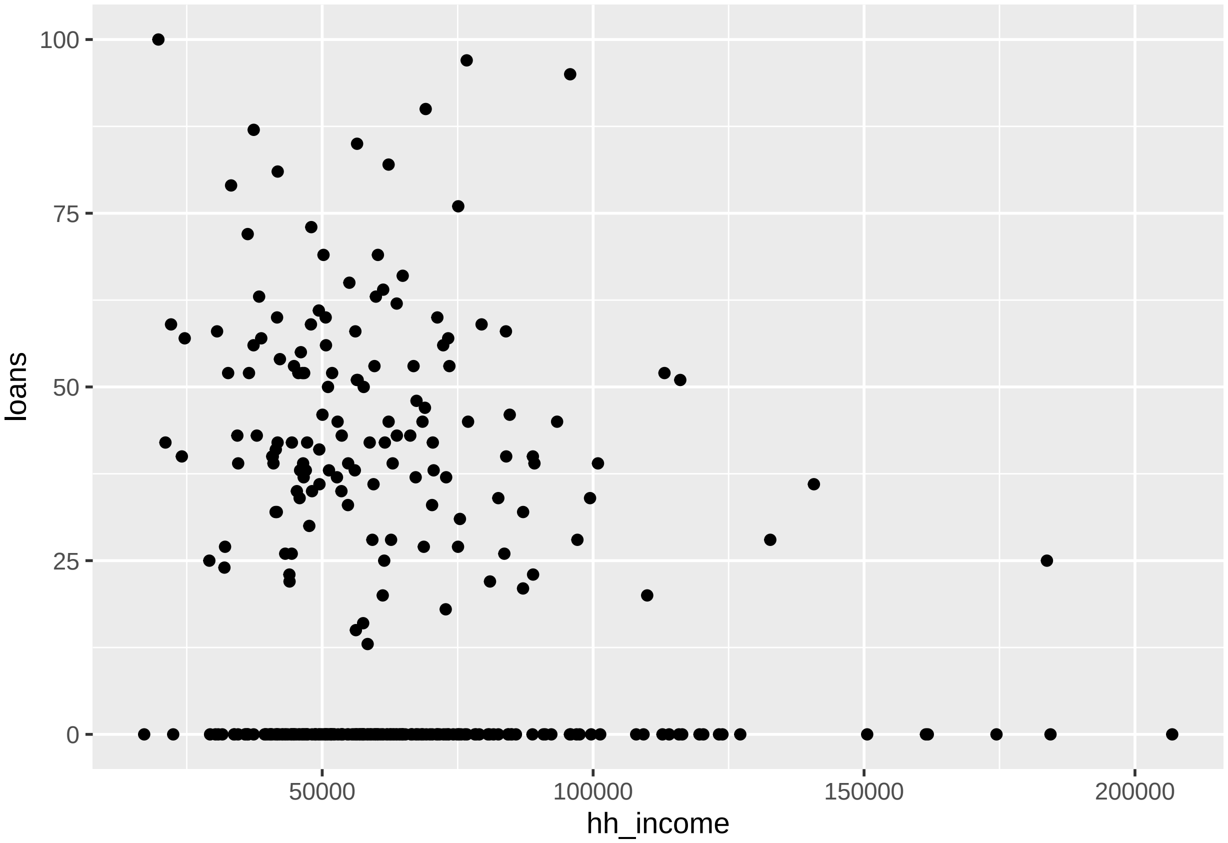
```
## Warning: ‘qplot()‘ was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```

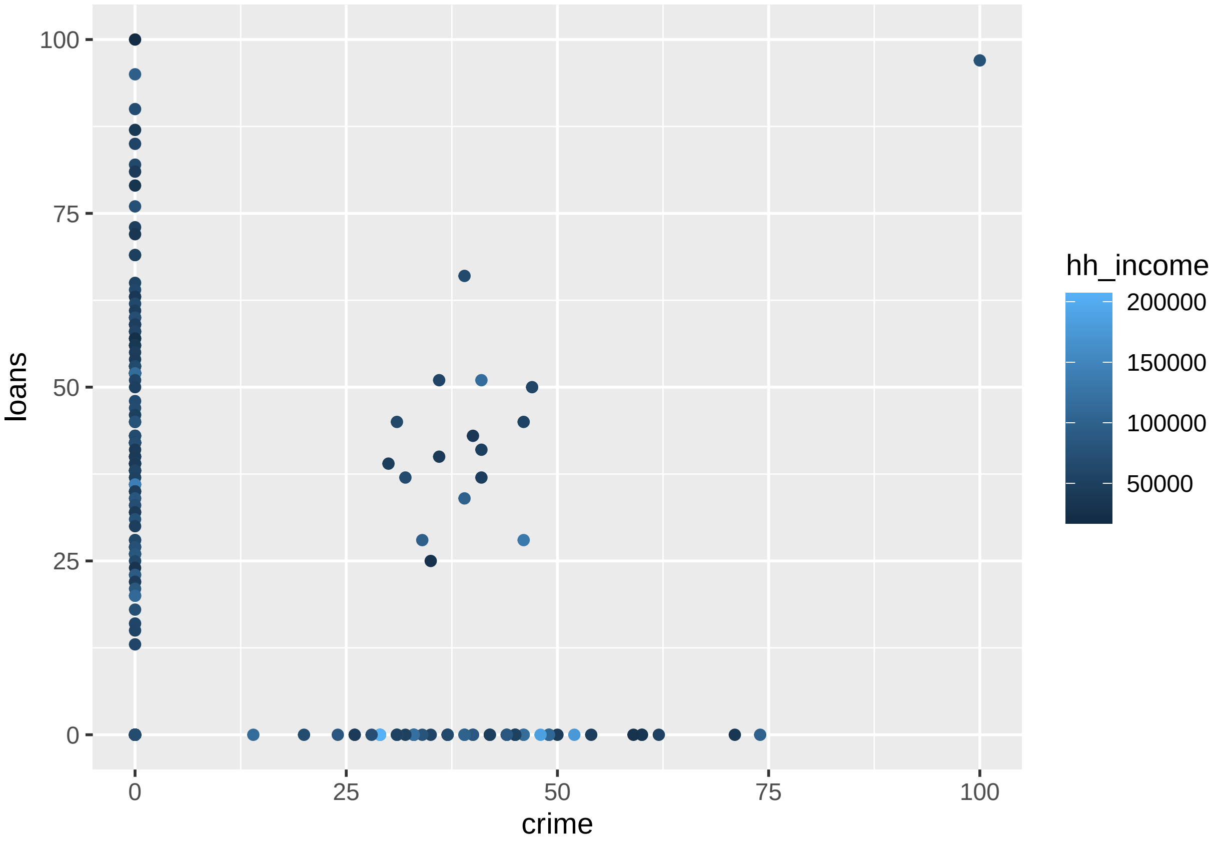


```
# plot for a relationship between hh_income and loans
res_city_acs %>%
  qplot(x = hh_income, y = loans, data = .,
        geom = "auto")
```

```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```



```
# Plotting
ggplot(res_city_acs, aes(crime, loans, colour = hh_income)) +
  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.
    - In the below tables, we are able to see how many cities appear or don't appear in both datasets.

```
# Are any of the locations in our search data also in our ACS data?
# If yes, will print TRUE.
any(res_covid_city$location %in% acs_il$location)
```

```
## [1] TRUE
```

```
# Printing how many cities don't appear in both data sets
paste("Google Trends IL Cities in ACS Data")
```

```
## [1] "Google Trends IL Cities in ACS Data"
```

```
summary(res_covid_city$location %in% acs_il$location)
```

```
##    Mode   FALSE    TRUE
## logical      12     333
```

```r
paste("ACS IL Cities in Google Trends Data")
```

```
## [1] "ACS IL Cities in Google Trends Data"
```

```r
summary(acs_il$location %in% res_covid_city$location)
```

```
##    Mode   FALSE    TRUE
## logical    1130     336
```

```r
# Doing an inner join, only keeping variables common to both datasets
res_covid_city_acs <- inner_join(res_covid_city, acs_il,
                                 by = join_by("location" == "location"))

# Printing the number of rows in each dataset to QC our merge matches
# the numbers in our logical table above
nrow(res_covid_city)
```

```
## [1] 345
```

```r
nrow(acs_il)
```

```
## [1] 1466
```

```r
nrow(res_covid_city_acs)
```

```
## [1] 336
```

- 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?
    - Table 6 below shows the mean of the search popularity for both keywords (‘covid’ and ‘trump’) for cities that have an average median household income and for those that have an below average median household income.

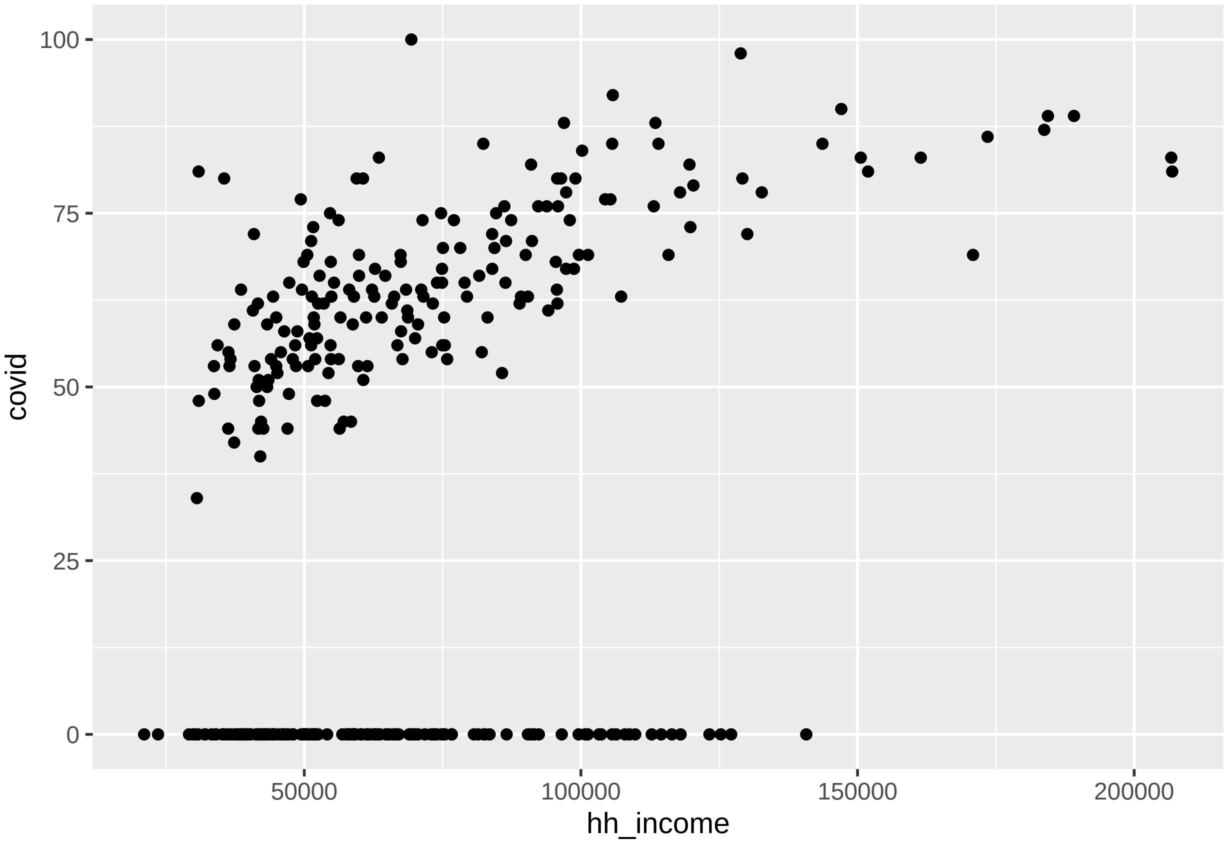
```r
# removing NA values in the household income variable
res_covid_city_acs <- res_covid_city_acs %>%
  drop_na(hh_income)

popsearchmean_covid <- res_covid_city_acs%>%
  mutate(high_hh_income = ifelse(hh_income > mean(hh_income, na.rm = TRUE),
                                 "Above", "Below")) %>%
  group_by(high_hh_income) %>%
  summarize(mean_pop_covid = mean(covid),
            mean_pop_trump = mean(trump))

knitr::kable(popsearchmean_covid,
  caption = "Popularity of COVID and Trump Searches Across Low and High-income Illinois Cities")
```

Table 6: Popularity of COVID and Trump Searches Across Low and High-income Illinois Cities

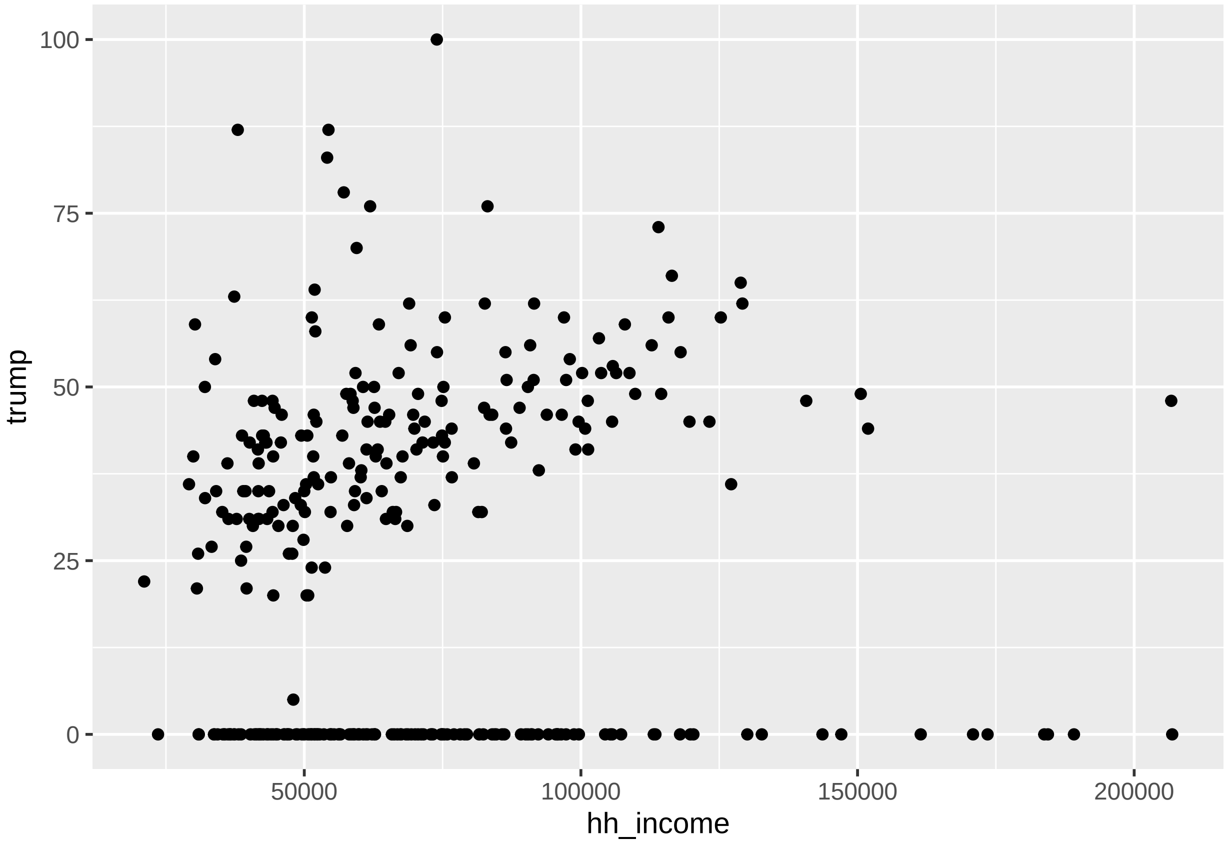
high_hh_income	mean_pop_covid	mean_pop_trump
Above	48.06923	27.26923
Below	29.48039	22.79902

- 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()`.
  - The scatter plots below depict a positive relationship between median household income and the search popularity of both the keywords 'covid' and 'trump'. They show that cities with higher median household incomes tend to exhibit elevated search popularity for these Google Trends keywords.
  - There appears to be a strong relationship between high frequency searches for "covid" and median household income. Above median household incomes of $125,000, there are few cities with low frequency "covid" searches.
  - There is a slight increase of searches for "trump" as median household income rises. However most of the high frequency searches of "trump" appear to be clustered around median household incomes below $125,000.
  - Seen together with income in different shades of blue, it appears there are several cities with medium to high incomes that search for both "trump" and "covid" frequently. However, it appears that high frequency searches for "trump" only occur mostly within cities with lower to medium household incomes. Meanwhile, high frequency searches for "covid" only tend to be in cities with much higher median household incomes overall.

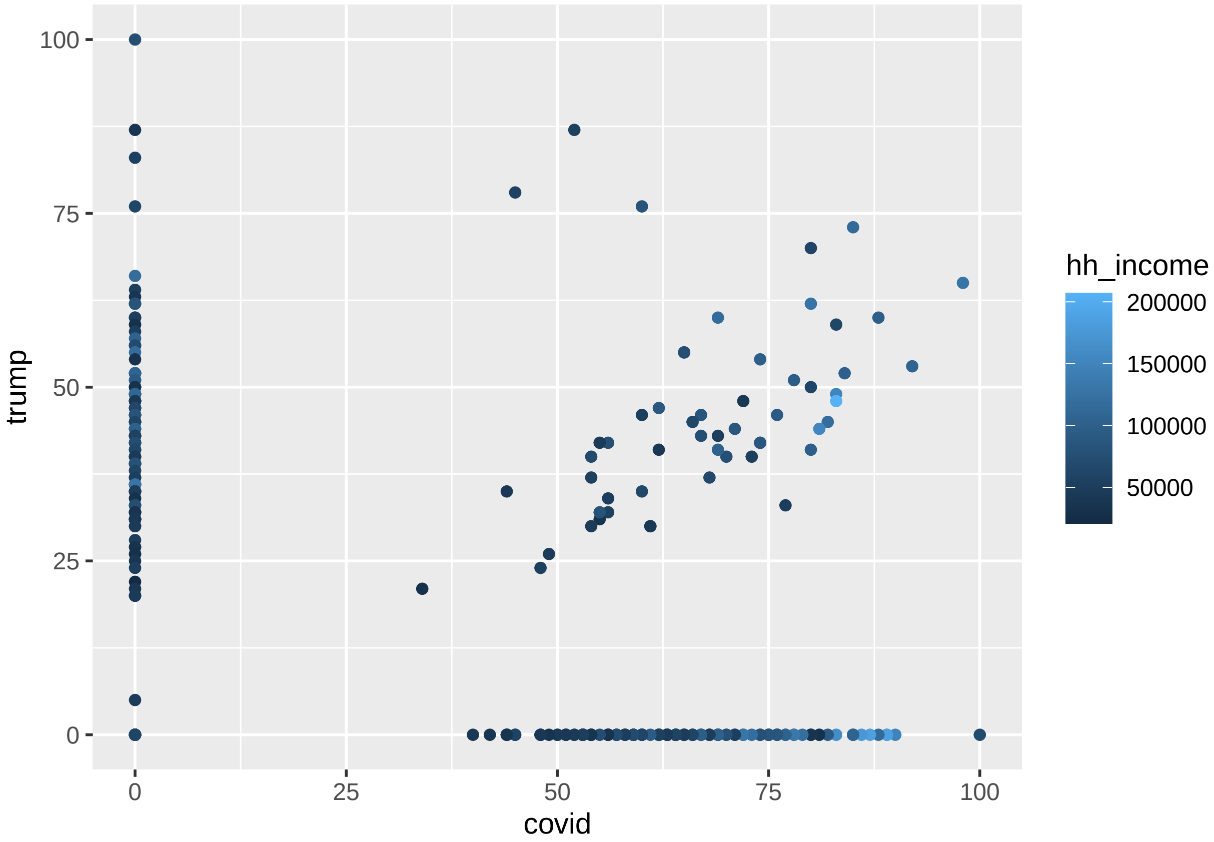
```r
# Doing qplots of each
res_covid_city_acs %>%
  qplot(x = hh_income, y = covid, data = .,
        geom = "auto")
```



```
res_covid_city_acs %>%
  qplot(x = hh_income, y = trump, data = .,
        geom = "auto")
```



```r
# Using ggplot
ggplot(res_covid_city_acs, aes(covid, trump, colour = hh_income)) +
geom_point()
```



trump
100
75
50
25
0
covid
0
25
50
75
100
hh_income
200000
150000
100000
50000