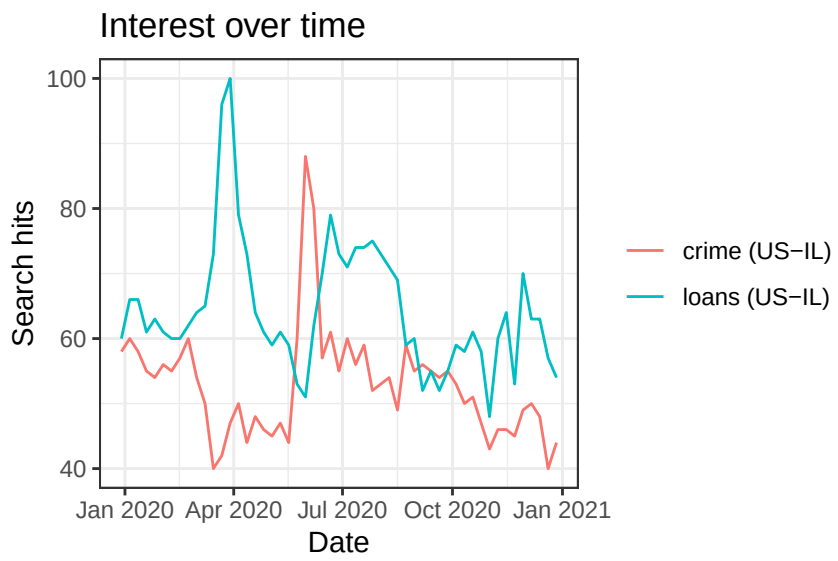
Assignment 2

Angelina Lu & Yuchen Ding

2024-09-28

GitHub Repository (https://github.com/angelinalu61/HW2_727_0918)

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




```
str(res)
```

```
## 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    : int [1:106] 58 60 58 55 54 56 55 57 60 54 ...
##   ..$ keyword : chr [1:106] "crime" "crime" "crime" "crime" ...
##   ..$ 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 2020-12-31" "
##   ..$ 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] "Rockford IL" "Chicago IL" "St. Louis MO" "Quincy IL-Hannibal MO-Keokuk IA
##   ..$ hits    : int [1:20] 100 100 97 92 90 89 83 79 78 76 ...
##   ..$ 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] "Anna" "Hampshire" "Streamwood" "East Saint Louis" ...
##   ..$ hits    : int [1:400] 100 90 85 85 84 82 80 76 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    : NULL
##  - attr(*, "class")= chr [1:2] "gtrends" "list"
```

```
head(res$interest_over_time)
```

```
##         date hits keyword   geo                  time gprop category
## 1 2019-12-29   58   crime US-IL 2020-01-01 2020-12-31   web        0
## 2 2020-01-05   60   crime US-IL 2020-01-01 2020-12-31   web        0
## 3 2020-01-12   58   crime US-IL 2020-01-01 2020-12-31   web        0
## 4 2020-01-19   55   crime US-IL 2020-01-01 2020-12-31   web        0
## 5 2020-01-26   54   crime US-IL 2020-01-01 2020-12-31   web        0
## 6 2020-02-02   56   crime US-IL 2020-01-01 2020-12-31   web        0
```

```
library(tidyverse)
trend_data <- res$interest_over_time
glimpse(trend_data)
```

```
## Rows: 106
## Columns: 7
## $ date     <dttm> 2019-12-29, 2020-01-05, 2020-01-12, 2020-01-19, 2020-01-26, ~
## $ hits     <int> 58, 60, 58, 55, 54, 56, 55, 57, 60, 54, 50, 40, 42, 47, 50, 4~
## $ 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~
```

1A. Find the mean, median and variance of the search hits for the keywords.

```
crime_data <- trend_data %>% filter(keyword == "crime")
mean_crime <- mean(crime_data$hits)
median_crime <- median(crime_data$hits)
var_crime <- var(crime_data$hits)

loans_data <- trend_data %>% filter(keyword == "loans")
mean_loans <- mean(loans_data$hits)
median_loans <- median(loans_data$hits)
var_loans <- var(loans_data$hits)
```

```
mean_crime
```

```
## [1] 52.83019
```

```
median_crime
```

```
## [1] 53
```

```
var_crime
```

```
## [1] 71.72061
```

```
mean_loans
```

```
## [1] 64.32075
```

```
median_loans
```

```
## [1] 62
```

```
var_loans
```

```
## [1] 99.76052
```

1B. 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.

```
trend_data_city <- res$interest_by_city

trend_data_city <- trend_data_city %>%
  mutate(hits = as.numeric(hits))

loans_by_city <- trend_data_city %>%
  filter(keyword == "loans") %>%
  group_by(location) %>%
  summarize(total_hits = sum(hits, na.rm = TRUE)) %>%
  arrange(desc(total_hits))

head(loans_by_city)
```

```
## # A tibble: 6 x 2
##   location         total_hits
##   <chr>                 <dbl>
## 1 Long Lake               100
## 2 Rosemont                 81
## 3 East Saint Louis         80
## 4 Coal City                79
## 5 Dolton                   78
## 6 Ford Heights             78
```

1C. Is there a relationship between the search intensities between the two keywords we used? Is there a relationship between the search intensities between the two keywords we used?

```
merged_data <- trend_data %>%
  select(date, keyword, hits) %>%
  pivot_wider(names_from = keyword, values_from = hits)

correlation <- cor(merged_data$crime, merged_data$loans, use = "complete.obs")
correlation
```

```
## [1] -0.1516683
```

The correlation coefficient indicates a slight negative correlation between the search popularity of "crime" and "loans." However, when the correlation coefficient is close to 0, it suggests almost no linear relationship. Therefore, there is no significant linear correlation between the search popularity of these two keywords, and their changes are likely influenced by different factors.

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.

```
library(gtrendsR)
library(tidyverse)

covid_res <- gtrends(c("death", "lockdown", "hospital"),
                     geo = "US-IL",
                     time = "2020-01-01 2020-12-31",
                     low_search_volume = TRUE)

covid_data <- covid_res$interest_over_time
head(covid_data)
```

```
##         date hits keyword   geo                  time gprop category
## 1 2019-12-29   42   death US-IL 2020-01-01 2020-12-31   web        0
## 2 2020-01-05   40   death US-IL 2020-01-01 2020-12-31   web        0
## 3 2020-01-12   42   death US-IL 2020-01-01 2020-12-31   web        0
## 4 2020-01-19   42   death US-IL 2020-01-01 2020-12-31   web        0
## 5 2020-01-26   77   death US-IL 2020-01-01 2020-12-31   web        0
## 6 2020-02-02   44   death US-IL 2020-01-01 2020-12-31   web        0
```

```
str(covid_data)
```

```
## 'data.frame':    159 obs. of  7 variables:
##  $ date    : POSIXct, format: "2019-12-29" "2020-01-05" ...
##  $ hits    : chr  "42" "40" "42" "42" ...
##  $ keyword : chr  "death" "death" "death" "death" ...
##  $ geo     : chr  "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 2020-12-31" "2020-01-0
##  $ gprop   : chr  "web" "web" "web" "web" ...
##  $ category: int  0 0 0 0 0 0 0 0 0 0 ...
```

```r
class(covid_data$hits)
```

```
## [1] "character"
```

```r
covid_data <- covid_data %>%
  mutate(hits = as.numeric(hits))
str(covid_data)
```

```
## 'data.frame':    159 obs. of  7 variables:
##  $ date    : POSIXct, format: "2019-12-29" "2020-01-05" ...
##  $ hits    : num  42 40 42 42 77 44 42 43 46 54 ...
##  $ keyword : chr  "death" "death" "death" "death" ...
##  $ geo     : chr  "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 2020-12-31" "2020-01-0
##  $ gprop   : chr  "web" "web" "web" "web" ...
##  $ category: int  0 0 0 0 0 0 0 0 0 0 ...
```

```r
covid_data <- covid_data %>%
  mutate(hits = gsub("<1", "1", hits)) %>%
  mutate(hits = as.numeric(hits))
str(covid_data)
```

```
## 'data.frame':    159 obs. of  7 variables:
##  $ date    : POSIXct, format: "2019-12-29" "2020-01-05" ...
##  $ hits    : num  42 40 42 42 77 44 42 43 46 54 ...
##  $ keyword : chr  "death" "death" "death" "death" ...
##  $ geo     : chr  "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 2020-12-31" "2020-01-0
##  $ gprop   : chr  "web" "web" "web" "web" ...
##  $ category: int  0 0 0 0 0 0 0 0 0 0 ...
```

```r
death_data <- covid_data %>% filter(keyword == "death")
mean_death <- mean(death_data$hits, na.rm = TRUE)
var_death <- var(death_data$hits, na.rm = TRUE)

lockdown_data <- covid_data %>% filter(keyword == "lockdown")
mean_lockdown <- mean(lockdown_data$hits, na.rm = TRUE)
var_lockdown <- var(lockdown_data$hits, na.rm = TRUE)

hospital_data <- covid_data %>% filter(keyword == "hospital")
mean_hospital <- mean(hospital_data$hits, na.rm = TRUE)
var_hospital <- var(hospital_data$hits, na.rm = TRUE)

mean_death
```

```
## [1] 46.88679
```

```r
var_death
```

```
## [1] 85.60232
```

```
mean_lockdown
```

```
## [1] 7.08
```

```
var_lockdown
```

```
## [1] 212.4833
```

```
mean_hospital
```

```
## [1] 56.64151
```

```
var_hospital
```

```
## [1] 28.58055
```

From the data structure, it shows that the hits column is of the character type, so I couldn't directly perform numerical operations. Therefore, I changed its structure and also addressed the issue with NA values.

2A. 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.

```
library(censusapi)
library(readr)
library(dplyr)
library(stringr)
library(gtrendsR)

cs_key <- read_file("census-key.txt")

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
```

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

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

```
acs_il <- acs_il %>%
  mutate(location = str_replace_all(NAME, c(" village" = "", " city" = "","," CDP" = "", ", Illinois" = "

head(acs_il$location)
```

```
## [1] "Coatsburg"  "Cobden"     "Coffeen"    "Colchester" "Coleta"
## [6] "Colfax"
```

```
trend_cities <- res$interest_by_city$location
acs_cities <- acs_il$location

unmatched_in_trend <- setdiff(trend_cities, acs_cities)

unmatched_in_acs <- setdiff(acs_cities, trend_cities)

num_unmatched_in_trend <- length(unmatched_in_trend)
num_unmatched_in_acs <- length(unmatched_in_acs)

num_unmatched_in_trend + num_unmatched_in_acs
```

```
## [1] 1148
```

There are 1134 cities that don't appear in both datasets and therefore cannot be merged.

2B. 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?

```
library(gtrendsR)
avg_income <- mean(acs_il$hh_income, na.rm = TRUE)
```

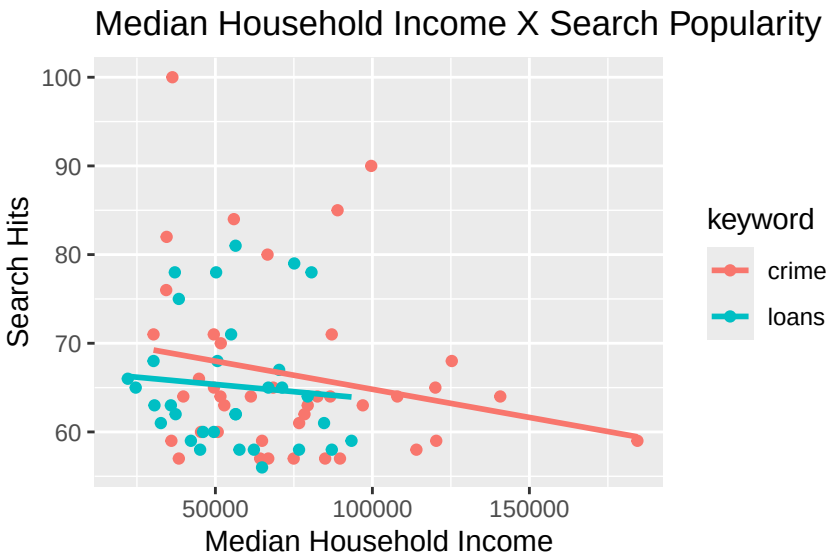
```r
acs_il <- acs_il %>%
  mutate(income_group = ifelse(hh_income > avg_income, "Above Average", "Below Average"))
merged_data <- inner_join(res$interest_by_city, acs_il, by = "location")
merged_data <- merged_data %>%
  filter(!is.na(income_group))

grouped_data <- merged_data %>%
  group_by(income_group, keyword) %>%
  summarize(mean_hits = mean(hits, na.rm = TRUE))

grouped_data
```

```
## # A tibble: 4 x 3
## # Groups:   income_group [2]
##   income_group  keyword mean_hits
##   <chr>         <chr>       <dbl>
## 1 Above Average crime        64.5
## 2 Above Average loans        64.5
## 3 Below Average crime        69.2
## 4 Below Average loans        65.5
```

In both higher and lower income cities, searches for "loans" were more popular than searches for crime, and the gap between the two was small. This may suggest that demand for and attention to loans is relatively consistent regardless of income. As for the search popularity of crime in higher income or lower income cities, it is more similar and almost the same, so the attention to crime does not show a significant difference between high income and low income groups.

2C. 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().

```r
library(ggplot2)

qplot(x = hh_income,
      y = hits,
      data = merged_data,
      color = keyword,
      geom = c("point", "smooth"),
      method = "lm",
      se = FALSE,
      main = "Median Household Income X Search Popularity",
      xlab = "Median Household Income",
      ylab = "Search Hits")
```




```
#ggplot(merged_data, aes(x = hh_income, y = hits, color = keyword)) +
  #geom_point() +
  #geom_smooth(method = "lm", se = FALSE) +
  #labs(title = "Median Household Income X Search Popularity",
       #x = "Median Household Income",
       #y = "Search Hits")
```

The overall distribution of the number of "loans" search hits shows vertical divergence, with a concentration of search hits in low- and middle-income households.

Meanwhile, the number of search hits for "crime" is negatively correlated with median household income, indicating that people in lower-income cities are more concerned with or search for "crime" more frequently, while this attention decreases in higher-income cities.

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

```
covid_cities <- covid_res$interest_by_city$location
acs_cities <- acs_il$location

unmatched_in_covid <- setdiff(covid_cities, acs_cities)

unmatched_in_acs <- setdiff(acs_cities, covid_cities)

num_unmatched_in_covid <- length(unmatched_in_covid)
num_unmatched_in_acs <- length(unmatched_in_acs)

total_unmatched <- num_unmatched_in_covid + num_unmatched_in_acs
total_unmatched
```

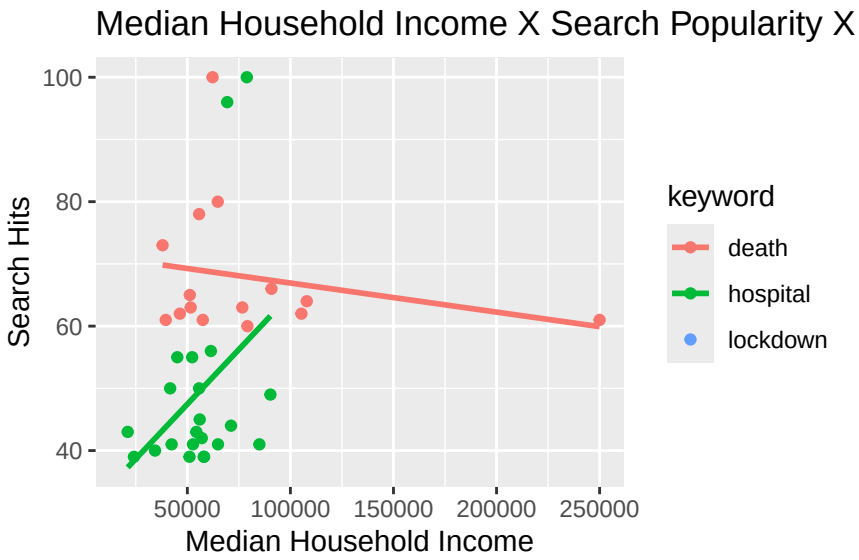
```
## [1] 1041
```

There are 1006 cities that don't appear in both datasets and therefore cannot be merged.

```r
avg_income <- mean(acs_il$hh_income, na.rm = TRUE)
acs_il <- acs_il %>%
  mutate(income_group = ifelse(hh_income > avg_income, "Above Average", "Below Average"))

merged_data_covid <- inner_join(covid_res$interest_by_city, acs_il, by = "location")

merged_data_covid <- merged_data_covid %>%
  filter(!is.na(income_group))

grouped_data_covid <- merged_data_covid %>%
  group_by(income_group, keyword) %>%
  summarize(mean_hits = mean(hits, na.rm = TRUE))

grouped_data_covid
```

```
## # A tibble: 6 x 3
## # Groups:   income_group [2]
##   income_group  keyword  mean_hits
##   <chr>         <chr>        <dbl>
## 1 Above Average death         65.1
## 2 Above Average hospital      61.8
## 3 Above Average lockdown     NaN
## 4 Below Average death         70.4
## 5 Below Average hospital      44.8
## 6 Below Average lockdown     NaN
```

```r
qplot(x = hh_income,
      y = hits,
      data = merged_data_covid,
      color = keyword,
      geom = c("point", "smooth"),
      method = "lm",
      se = FALSE,
      main = "Median Household Income X Search Popularity X COVID",
      xlab = "Median Household Income",
      ylab = "Search Hits")
```



The distribution of the keyword "death" is concentrated, particularly among low-income groups (household income below approximately 50,000).

There is a positive correlation between the number of searches for "hospital" and household income, indicating that higher-income households tend to search for "hospital" more frequently.

Finally, the keyword "lockdown" has fewer data points, which are mostly distributed in areas with high search frequency. However, the majority of these searches occur within the low-income range, with most search hits happening in households with an income below 12,500.