727-HW2

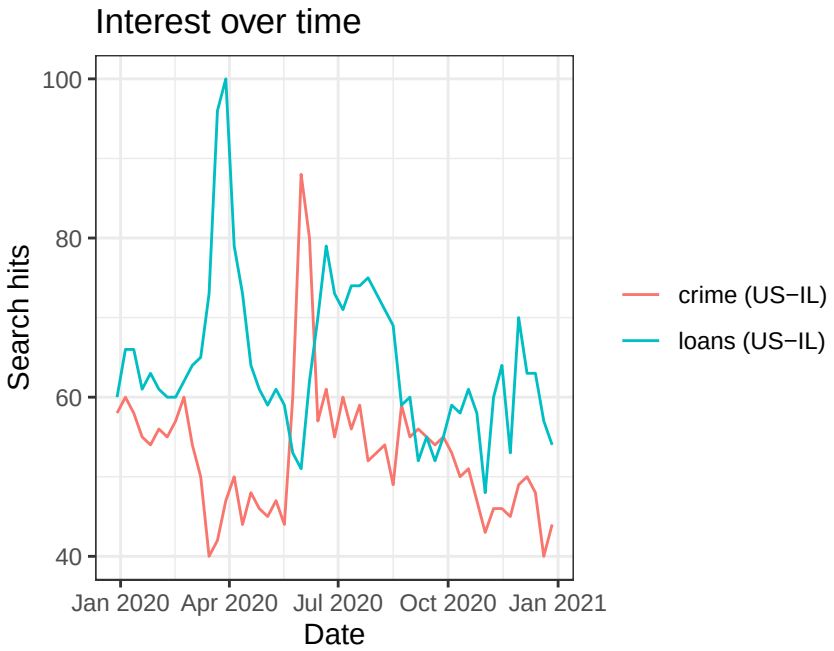
Yuchen Ding

2024-09-28

```
# Packages
if (!requireNamespace("httr", quietly = TRUE)) {
  install.packages("httr")
}
if (!requireNamespace("jsonlite", quietly = TRUE)) {
  install.packages("jsonlite")
}

library(httr)
library(jsonlite)
library(tidyverse)
```

Pulling from APIs

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



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

```
library(dplyr)
trend_data <- res$interest_over_time

crime_stats <- trend_data %>%
  filter(keyword == "crime") %>%
  summarise(
    mean = mean(hits, na.rm = TRUE),
    median = median(hits, na.rm = TRUE),
    variance = var(hits, na.rm = TRUE)
  )
crime_stats
```

```
##       mean median variance
## 1 52.83019     53 71.72061
```

```
loans_stats <- trend_data %>%
  filter(keyword == "loans") %>%
  summarise(
    mean_hits = mean(hits, na.rm = TRUE),
    median_hits = median(hits, na.rm = TRUE),
    variance_hits = var(hits, na.rm = TRUE)
  )
loans_stats
```

```
##   mean_hits median_hits variance_hits
## 1  64.32075          62      99.76052
```

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.

```
city_data <- res$interest_by_city

loans_data <- city_data %>%
  filter(keyword == "loans")

crime_data <- city_data %>%
  filter(keyword == "crime")

combined_data <- merge(crime_data, loans_data, by = "location", suffixes = c("_crime", "_loans"))

#head(combined_data)

top_loans_cities <- loans_data %>%
  arrange(desc(hits)) %>%
  select(location, hits)

head(top_loans_cities)
```

```
##           location hits
```

```
## 1        Long Lake  100
## 2         Rosemont   81
## 3 East Saint Louis   80
## 4        Coal City   79
## 5          Peotone   78
## 6           Dolton   78
```

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
```

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.

```
res_covid <- gtrends(c("covid vaccine", "covid deaths"),
                     geo = "US-IL",
                     time = "2020-01-01 2020-12-31",
                     low_search_volume = TRUE)

covid_trend_data <- res_covid$interest_over_time

covid_vaccine_data <- covid_trend_data %>%
  filter(keyword == "covid vaccine")

covid_deaths_data <- covid_trend_data %>%
  filter(keyword == "covid deaths")

combined_covid_data <- merge(covid_vaccine_data, covid_deaths_data, by = "date", suffixes = c("_deaths"

combined_covid_data$hits_deaths <- as.numeric(combined_covid_data$hits_deaths)
combined_covid_data$hits_vaccine <- as.numeric(combined_covid_data$hits_vaccine)

deaths_stats <- covid_trend_data %>%
  filter(keyword == "deaths") %>%
  summarise(
    mean_deaths = mean(hits, na.rm = TRUE),
    median_deaths = median(hits, na.rm = TRUE),
    variance_deaths = var(hits, na.rm = TRUE)
  )
deaths_stats
```

```
##   mean_deaths median_deaths variance_deaths
## 1          NA          <NA>              NA
```

```r
vaccine_stats <- covid_trend_data %>%
  filter(keyword == "vaccine") %>%
  summarise(
    mean_vaccine = mean(hits, na.rm = TRUE),
    median_vaccine = median(hits, na.rm = TRUE),
    variance_vaccine = var(hits, na.rm = TRUE)
  )
vaccine_stats
```

```
##   mean_vaccine median_vaccine variance_vaccine
## 1           NA           <NA>                NA
```

```r
cor_deaths_vaccine <- cor(combined_covid_data$hits_deaths, combined_covid_data$hits_vaccine, use = "comp

cor_deaths_vaccine
```

```
## [1] 0.02508857
```

Google Trends + ACS

```r
cs_key <- read_file("/Users/asuka/Library/Mobile Documents/com~apple~TextEdit/Documents/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
```

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

acs_il <-
acs_il %>%
rename(pop = B01001_001E,
age = B06002_001E,
```

```
hh_income = B19013_001E,
income = B19301_001E)
```

```
library(dplyr)
library(stringr)
acs_il <- acs_il %>%
  mutate(location = str_extract(NAME, "^[^,]+"))%>%
  mutate(location = str_replace_all(location, "\\s*(city|village|CDP)\\s*", ""))%>%
 select(-NAME)


head(acs_il)
```

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

```
head(res$interest_by_city)
```

```
##           location hits keyword   geo gprop
## 1             Anna  100   crime US-IL   web
## 2        Hampshire   90   crime US-IL   web
## 3       Streamwood   85   crime US-IL   web
## 4 East Saint Louis   85   crime US-IL   web
## 5  North Riverside   84   crime US-IL   web
## 6           Macomb   82   crime US-IL   web
```

```
summary(acs_il)
```

```
##     state              place                pop                 age
##  Length:1466        Length:1466        Min.   :      0.0   Min.   : 9.40
##  Class :character   Class :character   1st Qu.:    314.5   1st Qu.:36.70
##  Mode  :character   Mode  :character   Median :    944.0   Median :40.60
##                                        Mean   :   7674.1   Mean   :41.57
##                                        3rd Qu.:   4159.0   3rd Qu.:45.98
##                                        Max.   :2699347.0   Max.   :90.10
##                                                            NA's   :16
##    hh_income          income         location
##  Min.   : 11016   Min.   :  4800   Length:1466
##  1st Qu.: 45848   1st Qu.: 23356   Class :character
##  Median : 56466   Median : 27534   Mode  :character
##  Mean   : 63215   Mean   : 30715
##  3rd Qu.: 73030   3rd Qu.: 34017
##  Max.   :250001   Max.   :134596
##  NA's   :72       NA's   :23
```

```
merged_data <- merge(acs_il, combined_data, by = "location")

missing_in_acs <- anti_join(res$interest_by_city, acs_il, by = "location")

missing_in_trends <- anti_join(acs_il, res$interest_by_city, by = "location")
```

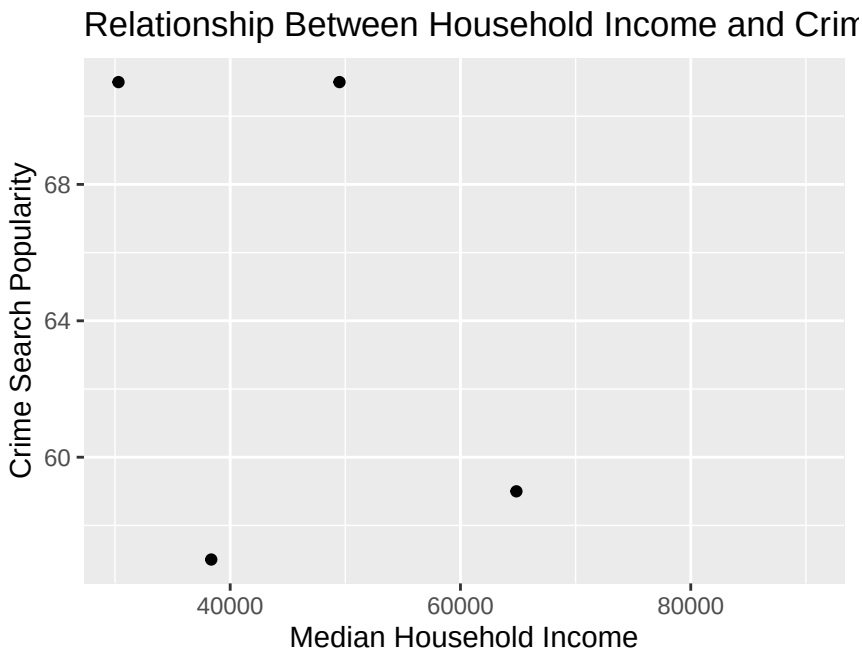
```r
n_missing_in_acs <- nrow(missing_in_acs)
n_missing_in_trends <- nrow(missing_in_trends)
n_total <- n_missing_in_acs + n_missing_in_trends

n_total
```

```
## [1] 1138
```

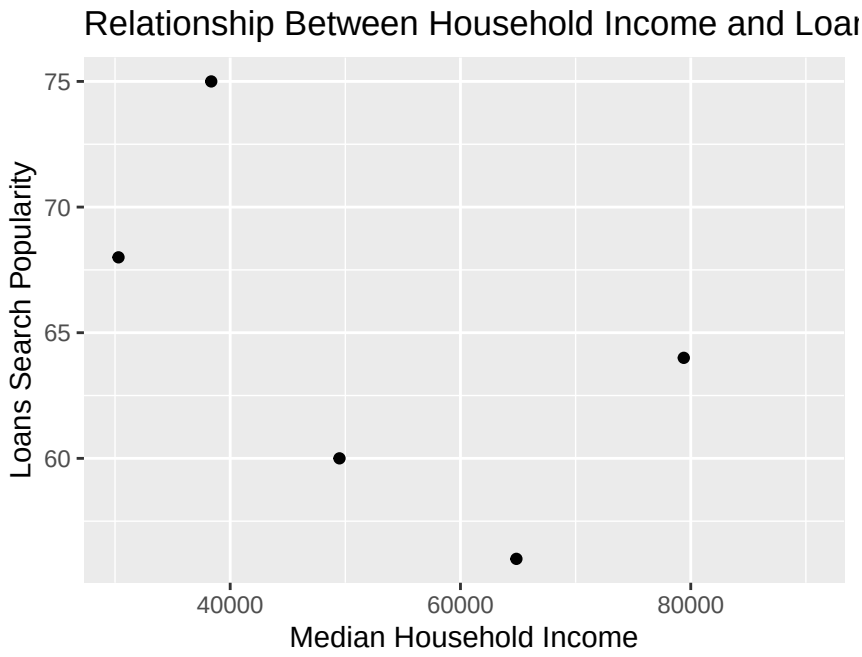
```r
average_income <- mean(merged_data$hh_income, na.rm = TRUE)

merged_data <- merged_data %>%
  mutate(income_group = ifelse(hh_income > average_income, "above_average", "below_average"))

search_means <- merged_data %>%
  group_by(income_group) %>%
  summarise(
    mean_crime_hits = mean(hits_crime, na.rm = TRUE),
    mean_loans_hits = mean(hits_loans, na.rm = TRUE)
  )

search_means
```

```
## # A tibble: 2 x 3
##   income_group  mean_crime_hits mean_loans_hits
##   <chr>                   <dbl>           <dbl>
## 1 above_average            59              60
## 2 below_average            66.3            67.7
```

```r
library(ggplot2)
qplot(hh_income, hits_crime, data = merged_data,
      xlab = "Median Household Income",
      ylab = "Crime Search Popularity",
      main = "Relationship Between Household Income and Crime Searches")
```



```
qplot(hh_income, hits_loans, data = merged_data,
      xlab = "Median Household Income",
      ylab = "Loans Search Popularity",
      main = "Relationship Between Household Income and Loans Searches")
```




```
merged_covid_data <- merge(acs_il, res_covid$interest_by_city, by = "location")

merged_covid_data <- merged_covid_data %>%
  mutate(income_group = ifelse(hh_income > average_income, "above_average", "below_average"))

merged_vaccine_data <- merged_covid_data %>%
  filter(keyword == "covid vaccine")

merged_deaths_data <- merged_covid_data %>%
  filter(keyword == "covid deaths")

combined_covid_data <- merge(merged_deaths_data, merged_vaccine_data, by = "location", suffixes = c("_d

missing_in_acs_covid <- anti_join(res_covid$interest_by_city, acs_il, by = "location")

missing_in_trends_covid <- anti_join(acs_il, res_covid$interest_by_city, by = "location")

n_missing_in_acs_covid <- nrow(missing_in_acs_covid)
n_missing_in_trends_covid <- nrow(missing_in_trends_covid)
n_total_covid <- n_missing_in_acs_covid + n_missing_in_trends_covid

n_total_covid
```

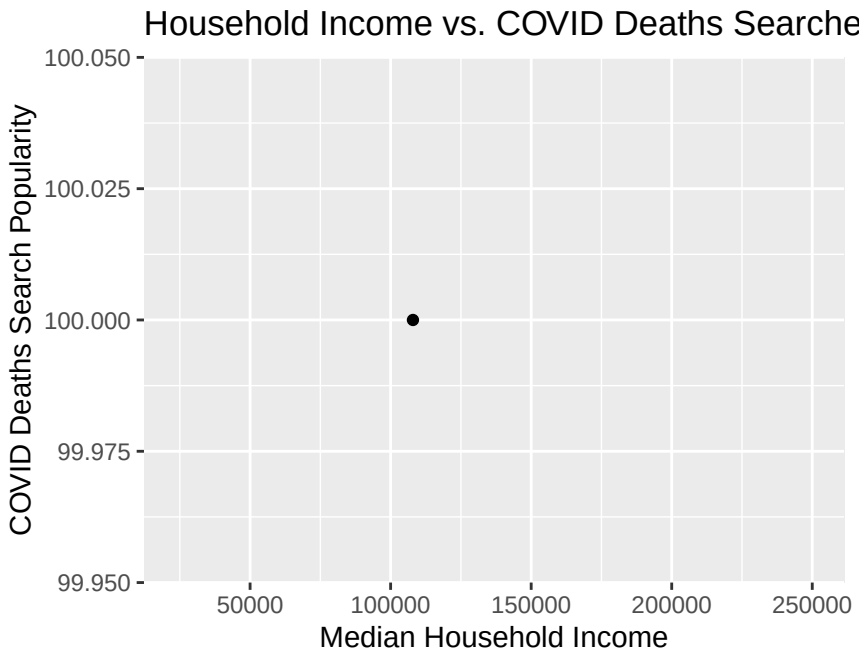
```
## [1] 1148
```

```
average_income_covid <- mean(combined_covid_data$hh_income, na.rm = TRUE)

search_means_covid <- combined_covid_data %>%
  group_by(income_group_deaths) %>%
```

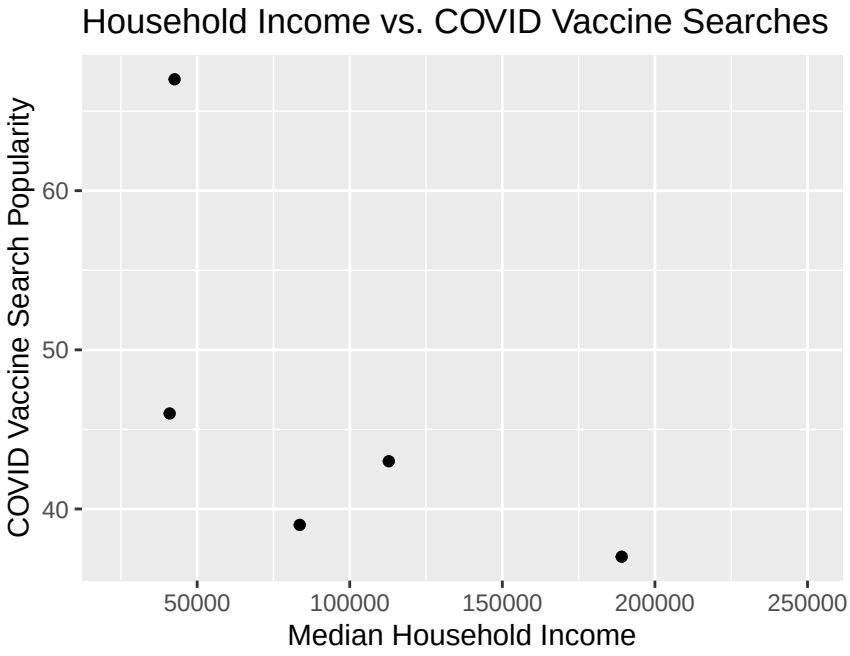
```
  summarise(
    mean_deaths_hits = mean(hits_deaths, na.rm = TRUE),
    mean_vaccine_hits = mean(hits_vaccine, na.rm = TRUE)
  )

search_means
```

```
## # A tibble: 2 x 3
##   income_group  mean_crime_hits mean_loans_hits
##   <chr>                   <dbl>           <dbl>
## 1 above_average            59              60
## 2 below_average            66.3            67.7
```

```
qplot(hh_income_deaths, hits_deaths, data = combined_covid_data,
      xlab = "Median Household Income",
      ylab = "COVID Deaths Search Popularity",
      main = "Household Income vs. COVID Deaths Searches")
```




```
qplot(hh_income_deaths, hits_vaccine, data = combined_covid_data,
      xlab = "Median Household Income",
      ylab = "COVID Vaccine Search Popularity",
      main = "Household Income vs. COVID Vaccine Searches")
```


Household Income vs. COVID Vaccine Searches
COVID Vaccine Search Popularity
60
50
40
50000
100000
150000
200000
250000
Median Household Income