Pulling from APIs - crime and loans

Isabel O'Malley

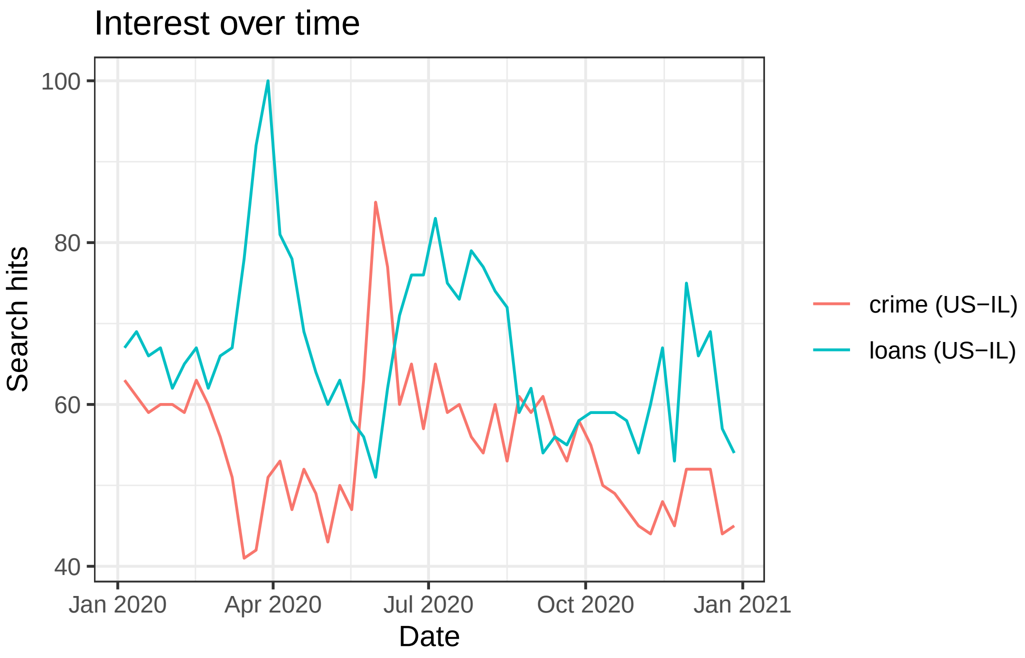
Link to github repo: https://github.com/isabelshaheen/JPSM727-assignment2.git

```
library(tidyverse)
library(gtrendsR)
library(censusapi)
```

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:

```
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.

First, we transform the `data.frame` into a `tibble`.

```
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, 60, 59, 63, 60, 56, 51, 41, 42, 51, 53, 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~
```

Then, we use the group_by function and we find mean, SD, median, and variance of hits for the two keywords.

```
res_time %>%
  group_by(keyword) %>%
  summarize(mean_hits = mean(hits),
            sd_hits = sd(hits),
            median_hits = median(hits),
            var_hits = var(hits))
```

```
# A tibble: 2 x 5
  keyword mean_hits sd_hits median_hits var_hits
  <chr>       <dbl>   <dbl>       <dbl>    <dbl>
1 crime        54.9    8.41        54.5     70.8
2 loans        66.5   10.1         66      103.
```

- **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.

Note that the original results object `res` contains some additional information, such as the search interest by city/ region.

```
res$interest_by_city
```

Make res$interest_by_city into a tibble and shorten name to res_city

Pivot wider to split the hits column into two variables: one for crime and one for loans

Plot the search hits for each keyword by city, using res_city_w

```
library(ggplot2)
ggplot (res_city_w, aes(x = location, y = loans)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Search hits for loan by city", x = 'city', y = 'hits')
```

Plot only the 10 observations with the highest # of hits on loans

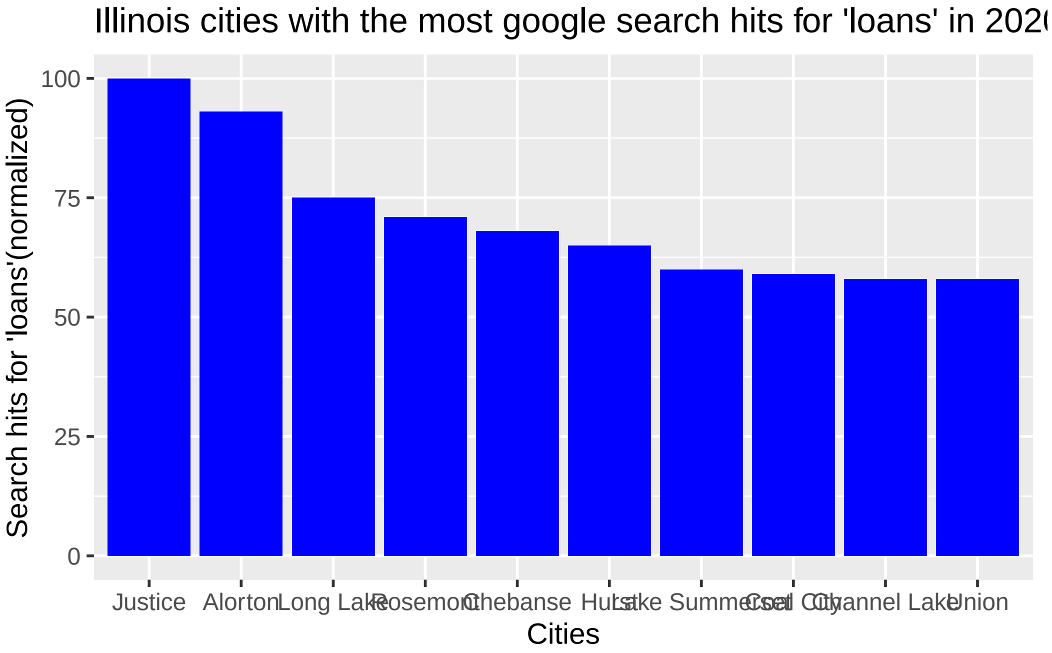
```
# Arrange the dataframe in descending order of the loans variable
res_city_w <- res_city_w %>%
  arrange(desc(loans))

# Select the top 10 observations
top_10 <- head(res_city_w, 10)
```

```
top_10
```

```
# A tibble: 10 x 5
   location       geo   gprop crime loans
   <chr>          <chr> <chr> <int> <int>
 1 Justice        US-IL web      NA   100
 2 Alorton        US-IL web      NA    93
 3 Long Lake      US-IL web      NA    75
 4 Rosemont       US-IL web      NA    71
 5 Chebanse       US-IL web      NA    68
 6 Hurst          US-IL web      NA    65
 7 Lake Summerset US-IL web      NA    60
 8 Coal City      US-IL web      NA    59
 9 Union          US-IL web      NA    58
10 Channel Lake   US-IL web      NA    58
```

```
# Create a bar plot using ggplot2
ggplot(data = top_10, aes(x = reorder(location, -loans), y = loans)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Illinois cities with the most google search hits for 'loans' in 2020", x =
```



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

Convert NAs to 0

Find the correlation between crime and loans hits

```
cor_test_result <- cor.test(res_city_w$crime, res_city_w$loans)

cor_test_result
```

```

    Pearson's product-moment correlation

data:  res_city_w$crime and res_city_w$loans
t = -2.0897, df = 344, p-value = 0.03738
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.214858472 -0.006603125
sample estimates:
     cor
-0.11196
```

Answer: The p-value is $< .001$ and the t-value is -4.23 indicating a significant negative correlation between the number of google searches for “crime” and the number of searches for “loans” in Illinois in 2020.