Assignment 2

Due at 11:59pm on October 3.

Leng Seong Che; Bozhou (Peter) Tan

The results are based on the data at 3:30pm on Oct 1st.

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.

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

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.

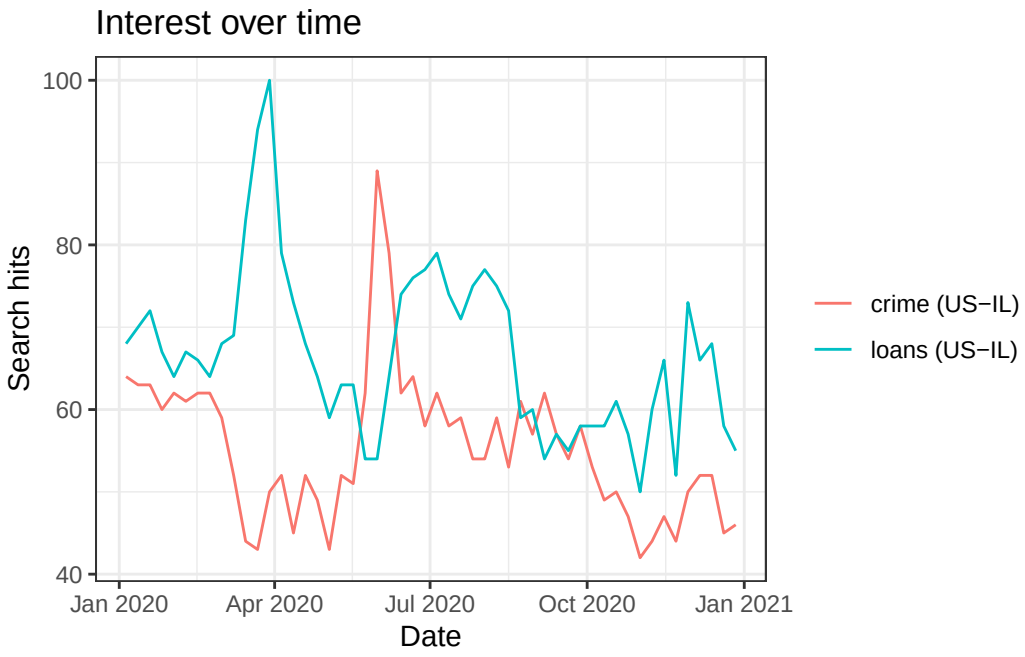
Pulling from APIs

`crime` and `loans`

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.

```
# transfer the data into tibble
rest = as_tibble(res$interest_over_time)

# find the mean, median and variance of the search hits
library(dplyr)
library(tidyr)
library(knitr)

descriptive = rest %>%
  group_by(keyword) %>%
  summarise(n = n(),
            mean = mean(hits),
            median = median(hits),
            variance = var(hits))
```

```r
kable(descriptive, caption = "Descriptive Statistics of Keywords")
```

Table 1: Descriptive Statistics of Keywords

keyword	n	mean	median	variance
crime	52	55.23077	54	78.41629
loans	52	66.69231	66	101.66818

According to Table 1, we can find that the keyword crime has a mean of 55.2307692307692, a median of 54 and a variance of 78.4162895927602. The keyword loans has a mean of 66.6923076923077, a median of 66 and a variance of 101.668174962293.

- 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.

```r
rescity = as_tibble(res$interest_by_city) %>%
  pivot_wider(., names_from = keyword, values_from = hits) %>%
  arrange(., desc(loans))
kable(head(rescity), caption = "Highest Search Frequency for Loans")
```

Table 2: Highest Search Frequency for Loans

location	geo	gprop	crime	loans
Alorton	US-IL	web	NA	100
Roseville	US-IL	web	NA	98
Henry	US-IL	web	NA	77
Cerro Gordo	US-IL	web	NA	74
Long Lake	US-IL	web	NA	71
Rosemont	US-IL	web	62	66

According to Table 2, Alorton has the highest search frequency for `loans` with the value of 100, followed by Roseville and Henry.

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

```r
crime = rest %>%
  filter(keyword == "crime") %>%
  select(date, hits) %>%
  rename(., crimehits = hits)

loan = rest %>%
  filter(keyword == "loans") %>%
  select(date, hits) %>%
  rename(., loanshits = hits)

crimloan = left_join(crime, loan, by = "date")
cor.test(crimloan$crimehits, crimloan$loanshits)
```

```
    Pearson's product-moment correlation

data:  crimloan$crimehits and crimloan$loanshits
t = -0.44078, df = 50, p-value = 0.6613
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3295202  0.2143246
sample estimates:
        cor
-0.06221463
```

According to the graph above, search frequencies for "crime" and "loans" have similar trends at the beginning of 2020, where they both went up and down from January to around February 2020. From March to April, search frequency for "loans" increased drastically from approximately 65 to 100, while search frequency for "crime" decreased before it increased again. In other words, the two keywords have a similar trend between January and February and most time between July 2020 and January 2021. However, from March to June 2020, they seem to have a inverse relationship.

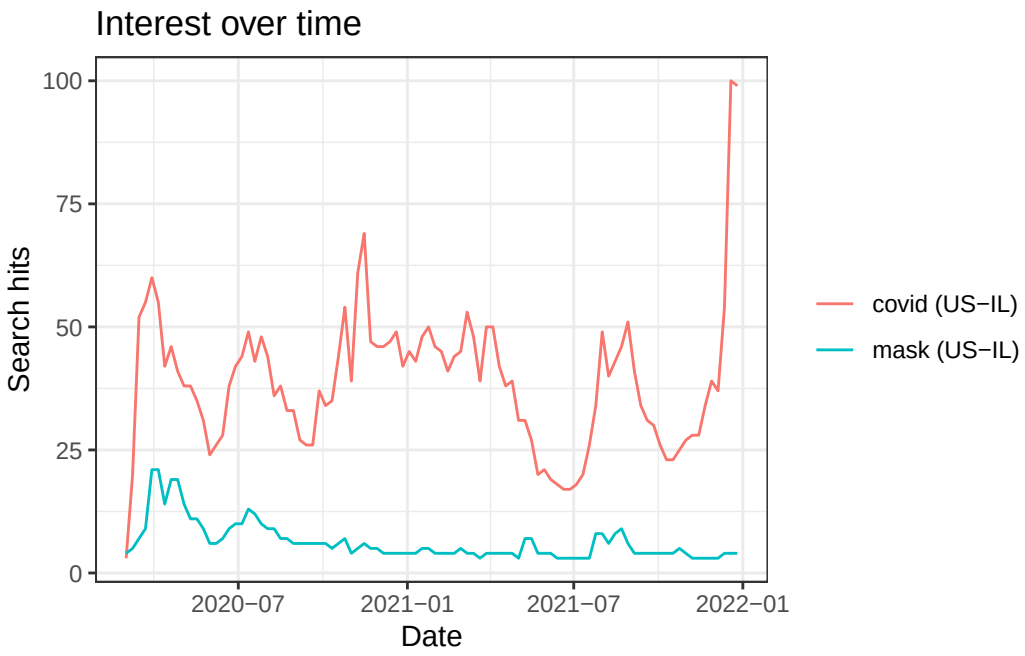
If we use the quantitative method to compute the t-statistic and corresponding p-value, we can see that the p-value is bigger than 0.05, which means there is no statistically significant negative relationship between crime and loans.

`covid` and `mask`

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 choose `covid` and `mask` as our keywords for analysis.

```
res2 = gtrends(c("covid", "mask"),
               geo = "US-IL",
               time = "2020-03-01 2021-12-31",
               low_search_volume = TRUE)
plot(res2)
```




```
# transfer the data into tibble
rest2 = as_tibble(res2$interest_over_time)

# find the mean, median and variance of the search hits
descriptive2 = rest2 %>%
  group_by(keyword) %>%
  summarise(n = n(),
            mean = mean(hits),
            median = median(hits),
            variance = var(hits))
kable(descriptive2, caption = "Descriptive Statistics of Keywords")
```

Table 3: Descriptive Statistics of Keywords

keyword	n	mean	median	variance
covid	96	39.00	39	213.57895
mask	96	6.25	5	14.98947

From the table, we can find that the keyword covid has a mean of 39, a median of 39 and a variance of 213.578947368421. The keyword mask has a mean of 6.25, a median of 5 and a variance of 14.9894736842105.

```
rescity2 = as_tibble(res2$interest_by_city) %>%
  pivot_wider(., names_from = keyword, values_from = hits) %>%
  arrange(., desc(covid))
kable(head(rescity2), caption = "Highest Search Frequency for covid")
```

Table 4: Highest Search Frequency for covid

location	geo	gprop	covid	mask
Barrington	US-IL	web	100	66
Evergreen Park	US-IL	web	97	NA
Oak Lawn	US-IL	web	96	44
Evanston	US-IL	web	96	NA
Wheaton	US-IL	web	96	NA
Lake Forest	US-IL	web	96	70

From the table, we can see that Barrington has the highest search frequency for `covid` with the value of 100, followed by Evergreen Park and Oak Lawn.

```
mask = rest2 %>%
  filter(keyword == "mask") %>%
  select(date, hits) %>%
  rename(., maskhits = hits)

covid = rest2 %>%
  filter(keyword == "covid") %>%
  select(date, hits) %>%
  rename(., covidhits = hits)

maskcovid = left_join(mask, covid, by = "date")
cor.test(maskcovid$maskhits, maskcovid$covidhits)
```

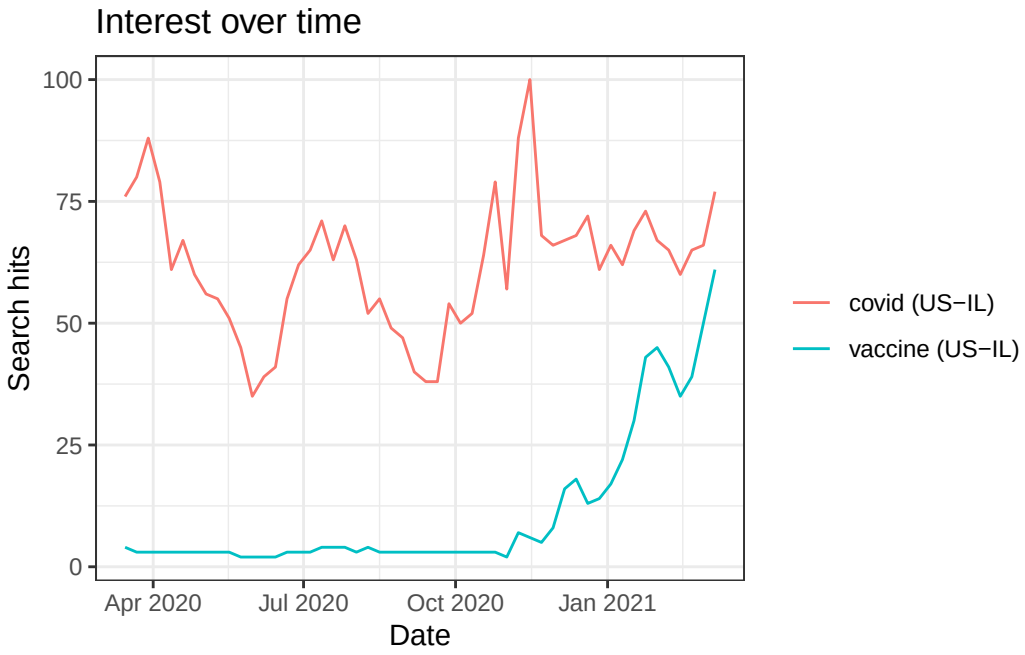
```
    Pearson's product-moment correlation

data:  maskcovid$maskhits and maskcovid$covidhits
t = 2.1777, df = 94, p-value = 0.03193
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01952591 0.40197806
sample estimates:
      cor
0.2191539
```

From the correlation test, we can see that `covid` has a significantly positive correlation with `mask` at 0.05 level. The correlation probably means that people will search for mask when Covid-19 is severe in one place.

`covid` and `vaccine`

We choose `covid` and `vaccine` as our keywords for analysis.

```
res3 = gtrends(c("covid", "vaccine"),
               geo = "US-IL",
               time = "2020-03-11 2021-3-11",
               low_search_volume = TRUE)
plot(res3)
```




```
# transfer the data into tibble
rest3 = as_tibble(res3$interest_over_time)

# find the mean, median and variance of the search hits
descriptive3 = rest3 %>%
  group_by(keyword) %>%
  summarise(n = n(),
            mean = mean(hits),
            median = median(hits),
            variance = var(hits))
kable(descriptive3, caption = "Descriptive Statistics of Keywords")
```

Table 5: Descriptive Statistics of Keywords

keyword	n	mean	median	variance
covid	52	61.86538	63	186.7854
vaccine	52	11.00000	3	222.3137

From the table, we can find that the keyword covid has a mean of 61.8653846153846, a median of 63 and a variance of 186.78544494721. The keyword vaccine has a mean of 11, a median of 3 and a variance of 222.313725490196.

```
rescity3 = as_tibble(res3$interest_by_city) %>%
  pivot_wider(., names_from = keyword, values_from = hits) %>%
  arrange(., desc(vaccine))
kable(head(rescity3), caption = "Highest Search Frequency for vaccine")
```

Table 6: Highest Search Frequency for vaccine

location	geo	gprop	covid	vaccine
Hurst	US-IL	web	68	100
Northbrook	US-IL	web	79	86
Highland Park	US-IL	web	92	82
Hinsdale	US-IL	web	NA	82
Deer Park	US-IL	web	NA	79
Riverwoods	US-IL	web	NA	77

From the table, we can see that Hurst has the highest search frequency for `vaccine` with the value of 100, followed by Northbrook and Highland Park.

```
vaccine = rest3 %>%
  filter(keyword == "vaccine") %>%
  select(date, hits) %>%
  rename(., vaccinehits = hits)

covid = rest3 %>%
  filter(keyword == "covid") %>%
  select(date, hits) %>%
  rename(., covidhits = hits)

vaccinecovid = left_join(vaccine, covid, by = "date")
cor.test(vaccinecovid$vaccinehits, vaccinecovid$covidhits)
```

```
    Pearson's product-moment correlation

data:  vaccinecovid$vaccinehits and vaccinecovid$covidhits
t = 1.96, df = 50, p-value = 0.05558
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.006242617  0.503323250
sample estimates:
```

```
      cor
0.2671128
```

The search popularities for covid and vaccine seem to have a positive relationship starting from November 2020. As the vaccine became available around this time, the search for vaccine increases drastically and follows a similar pattern of covid.

From the correlation test, we can see that `covid` has a significantly positive correlation with `vaccine` at 0.1 level.

Google Trends + ACS

`crime` and `loans`

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, store this key in the `cs_key` object. We will use this object in all following API queries.

```
cs_key <- "c0fd12402e23b7a95923e694f046015d624c91c5"
```

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
```

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)
```

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.

```
library(stringr)
pattern = c("St." = "Saint")

acs_il = acs_il %>%
  mutate(location = str_remove_all(NAME, c(" town,| city,| village,| Illinois"))) %>%
  mutate(location = str_replace_all(location, coll(pattern)))
```

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.

```
joint = inner_join(rescity, acs_il, by = "location")
nrow(joint)
```

```
[1] 333
```

```
# check how many cities do not appear in both datasets
n = (nrow(acs_il) - nrow(joint) ) + (nrow(rescity) - nrow(joint))
n
```

```
[1] 1150
```

- 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?

```
group1 = joint %>%
  mutate(mean = mean(hh_income, na.rm = TRUE)) %>%
  mutate(group = ifelse(hh_income > mean, "high", "low")) %>%
  group_by(group) %>%
  summarise(crime = mean(crime, na.rm = TRUE),
            loans = mean(loans, na.rm = TRUE)) %>%
  filter(!is.na(group))
kable(group1, caption = "Search Popularity by Household Income")
```

Table 7: Search Popularity by Household Income

group	crime	loans
high	50.92683	29.15556
low	55.14634	35.80952

From the table, cities that have an above average median household income have lower crime hits and lower loans hits, which means crime and loans may correlate with income. The reason for higher mean search popularity of "crime" can be that those with lower average

median household income live in some neighborhoods with a relatively higher number of crimes. Houses in areas with more crimes can be more affordable. The reason for higher mean search popularity of "loans" can be these households need more loans for various living expenses such as education. Also, the low search popularity might be due to less access to internet for lower-income.

- 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()`.

```
cor.test(joint$hh_income, joint$crime, method = "pearson")
```

```
    Pearson's product-moment correlation

data:  joint$hh_income and joint$crime
t = -1.2801, df = 80, p-value = 0.2042
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.34798920  0.07771525
sample estimates:
       cor
-0.1416813
```

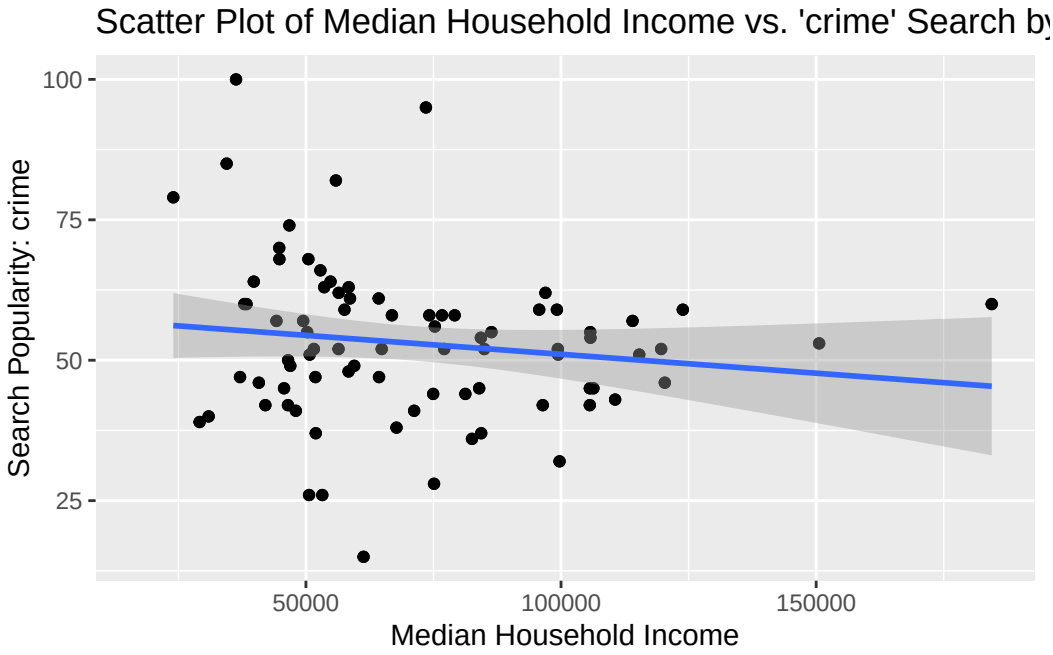
```
cor.test(joint$hh_income, joint$loans, method = "pearson")
```

```
    Pearson's product-moment correlation

data:  joint$hh_income and joint$loans
t = -3.6064, df = 127, p-value = 0.0004447
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4537407 -0.1392747
sample estimates:
       cor
-0.3047914
```

```
p1 = qplot(x = hh_income, y = crime, data = joint) +
  geom_point() +
  geom_smooth(method = lm) +
```

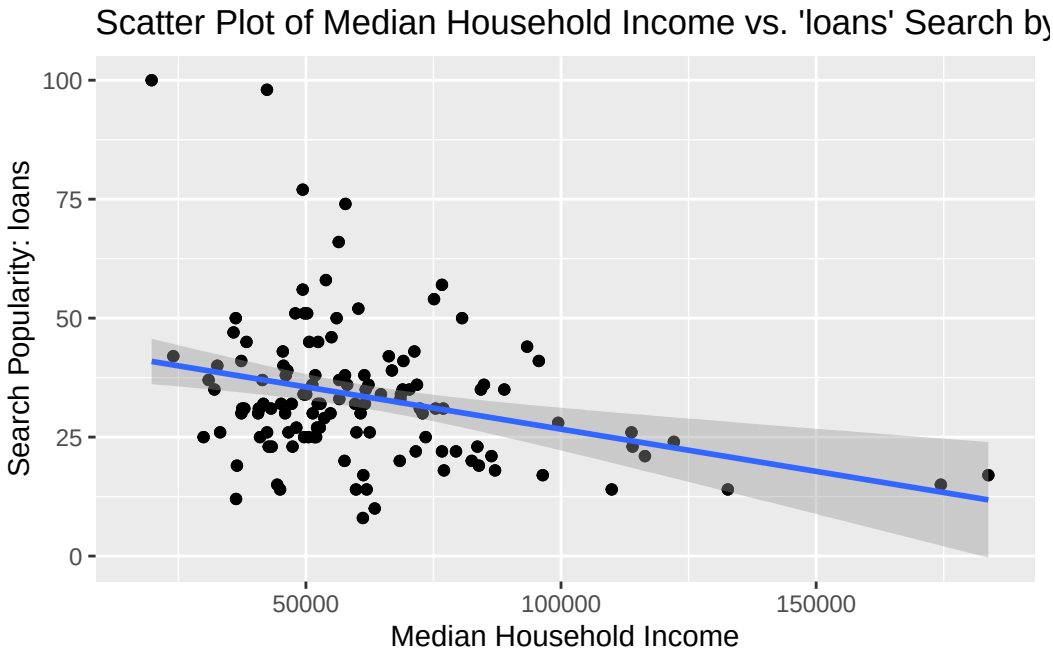
```
  labs(
    title = "Scatter Plot of Median Household Income vs. 'crime' Search by City",
    x = "Median Household Income",
    y = "Search Popularity: crime"
  )

p1
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
p2 = qplot(x = hh_income, y = loans, data = joint) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(
    title = "Scatter Plot of Median Household Income vs. 'loans' Search by City",
    x = "Median Household Income",
    y = "Search Popularity: loans"
  )

p2
```

```
`geom_smooth()` using formula = 'y ~ x'
```



The results from the Pearson correlation test suggest a negative statistically significant correlation between the median household income and the search popularity for "loans" and a statistically non-significant correlation between the median household income and "crime". These can also be observed in the scatter plots. For "crime", the majority of the cities have search popularity below 40 regardless of median household income. A slightly decreasing trend according to the the regression line, however, the correlation test suggests an absence of statistically significant relationship between them. For "loans", a decreasing trend is suggested based on the regression line. As the median household income increases, the search popularity for "loans" decrease. For those with median household income higher than $100,000, the searches are mostly lower than 25. Those median household income higher than lower than $100,000 have a wider range of search numbers.

`covid` and `mask`

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

```
joint2 = inner_join(rescity2, acs_il, by = "location")

group2 = joint2 %>%
```

```
  mutate(mean = mean(hh_income, na.rm = TRUE)) %>%
  mutate(group = ifelse(hh_income > mean, "high", "low")) %>%
  group_by(group) %>%
  summarise(covid = mean(covid, na.rm = TRUE),
            mask = mean(mask, na.rm = TRUE)) %>%
  filter(!is.na(group))
kable(group2, caption = "Search Popularity by Household Income")
```

Table 8: Search Popularity by Household Income

group	covid	mask
high	75.90909	65.25714
low	62.31579	64.62245

From the table, we can see cities that have an above average median household income have higher covid hits and higher mask hits, which means search hits of covid and mask may correlate with income positively.

```
cor.test(joint2$hh_income, joint2$covid, method = "pearson")
```

```
    Pearson's product-moment correlation

data:  joint2$hh_income and joint2$covid
t = 10.933, df = 189, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5270419 0.7023537
sample estimates:
      cor
0.6224443
```

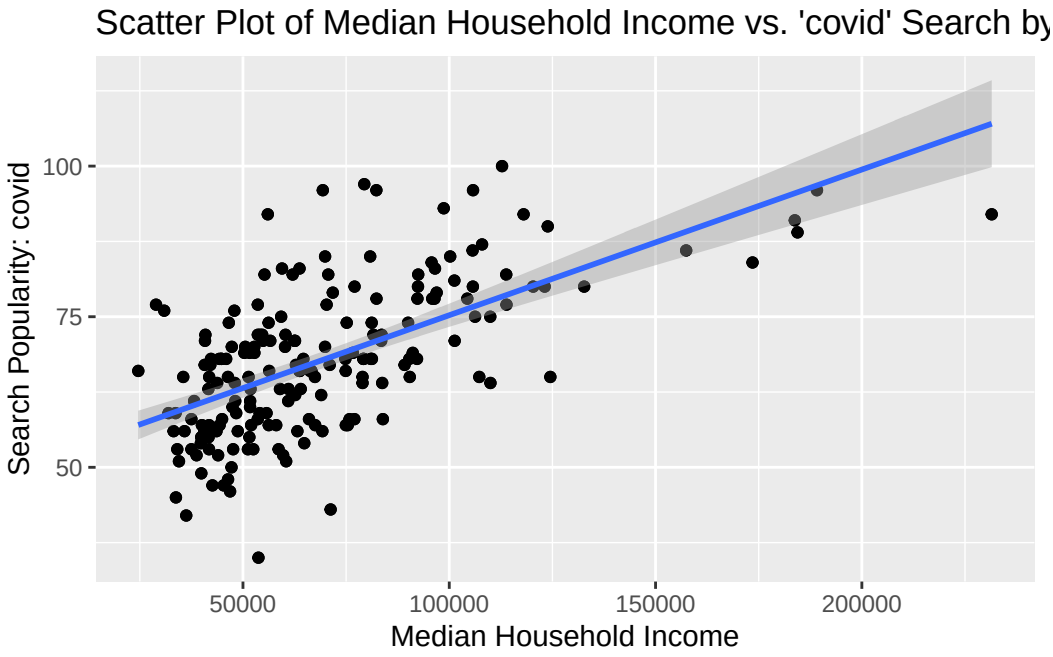
```
cor.test(joint2$hh_income, joint2$mask, method = "pearson")
```

```
    Pearson's product-moment correlation

data:  joint2$hh_income and joint2$mask
t = 2.253, df = 166, p-value = 0.02557
```

```
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.02139895 0.31543379
sample estimates:
      cor
0.1722507
```

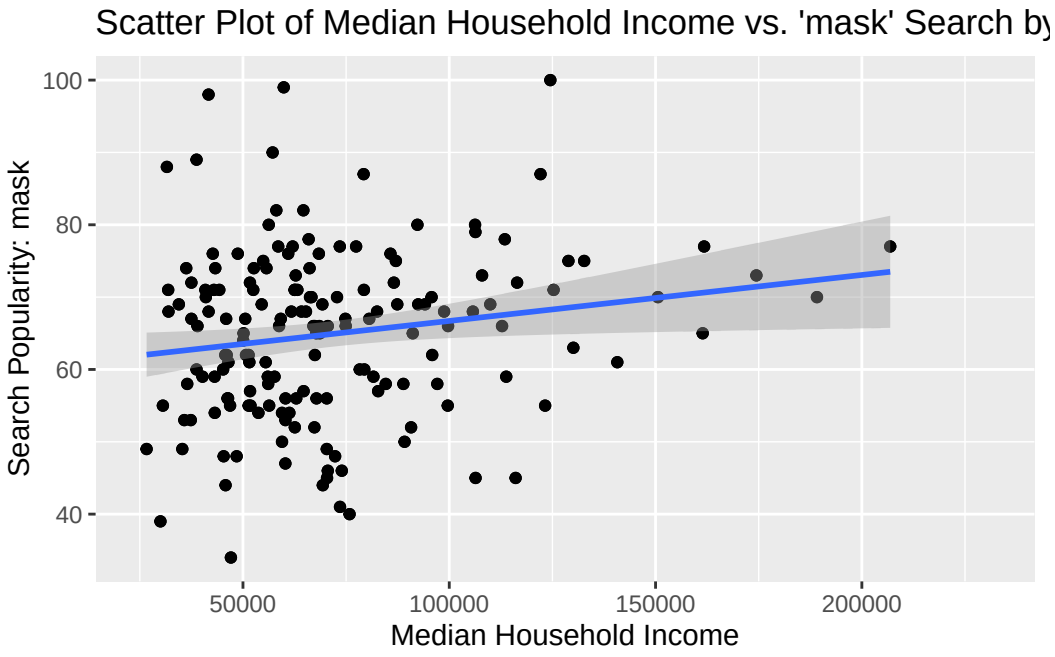
```r
p3 = qplot(x = hh_income, y = covid, data = joint2) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(
    title = "Scatter Plot of Median Household Income vs. 'covid' Search by City",
    x = "Median Household Income",
    y = "Search Popularity: covid"
  )

p3
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```r
p4 = qplot(x = hh_income, y = mask, data = joint2) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(
    title = "Scatter Plot of Median Household Income vs. 'mask' Search by City",
    x = "Median Household Income",
    y = "Search Popularity: mask"
  )

p4
```

```
`geom_smooth()` using formula = 'y ~ x'
```



According to the scatterplots, we can see that the income is both positively correlated with covid and mask. This indicates that people in rich areas may pay attention to covid and its protection more, showing one kind of social inequality. From the correlation test, we can see that the p-value of both tests are both less than 0.05, indicating that income has a statistically significant positive relation with both covid and mask.

covid and vaccine

```
joint3 = inner_join(rescity3, acs_il, by = "location")

group3 = joint3 %>%
  mutate(mean = mean(hh_income, na.rm = TRUE)) %>%
  mutate(group = ifelse(hh_income > mean, "high", "low")) %>%
  group_by(group) %>%
  summarise(covid = mean(covid, na.rm = TRUE),
            vaccine = mean(vaccine, na.rm = TRUE)) %>%
  filter(!is.na(group))
kable(group3, caption = "Search Popularity by Household Income")
```

Table 9: Search Popularity by Household Income

group	covid	vaccine
high	69.16176	53.95946
low	58.73950	35.93617

From the table, we can see cities that have an above average median household income have higher covid hits and higher mask hits, which means search hits of covid and mask may correlate with income positively.

```
cor.test(joint3$hh_income, joint3$covid, method = "pearson")
```

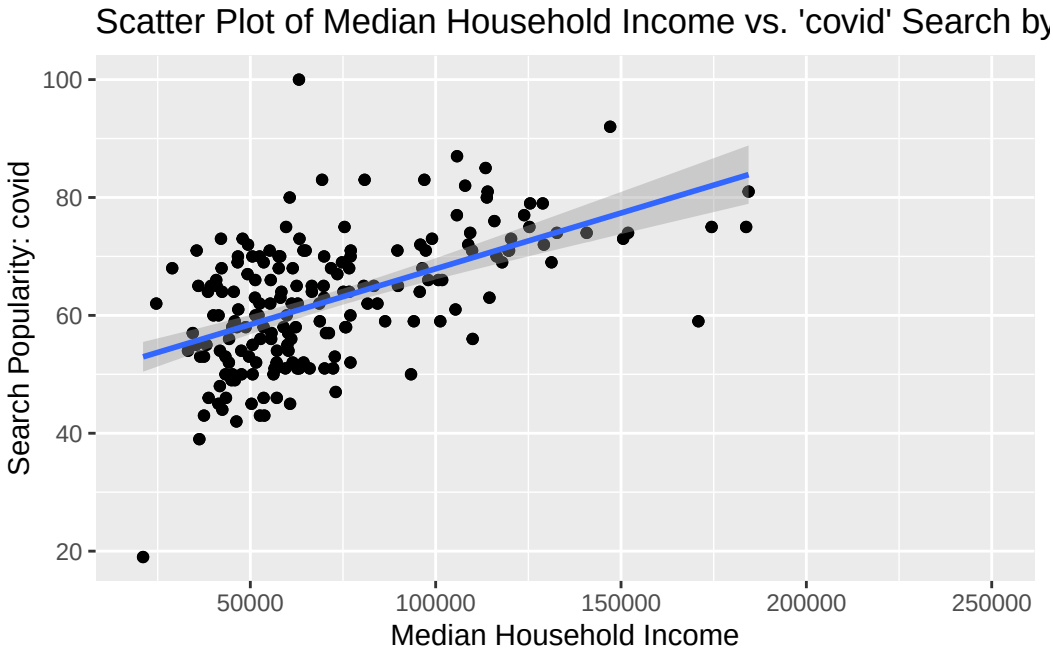
```
    Pearson's product-moment correlation

data:  joint3$hh_income and joint3$covid
t = 8.8467, df = 185, p-value = 7.13e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4358424 0.6387548
sample estimates:
      cor
0.5452363
```

```
cor.test(joint3$hh_income, joint3$vaccine, method = "pearson")
```

```
    Pearson's product-moment correlation

data:  joint3$hh_income and joint3$vaccine
t = 10.572, df = 166, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5342122 0.7168754
sample estimates:
      cor
0.6343155
```

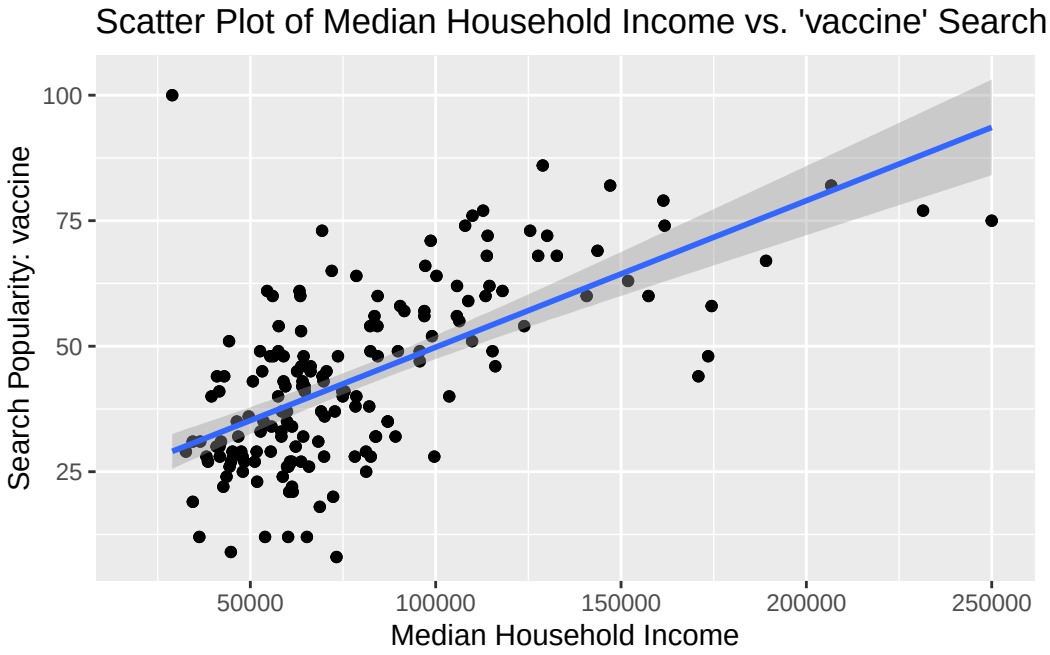
```r
p5 = qplot(x = hh_income, y = covid, data = joint3) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(
    title = "Scatter Plot of Median Household Income vs. 'covid' Search by City",
    x = "Median Household Income",
    y = "Search Popularity: covid"
  )

p5
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
p6 = qplot(x = hh_income, y = vaccine, data = joint3) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(
    title = "Scatter Plot of Median Household Income vs. 'vaccine' Search by City",
    x = "Median Household Income",
    y = "Search Popularity: vaccine"
  )

p6
```

```
`geom_smooth()` using formula = 'y ~ x'
```



The results from the Pearson correlation test suggest a positive statistically significant correlation between the median household income and both keywords "covid" and "vaccine". The scatter plot results are consistent with the correlation tests. For "covid", the majority of the cities with median household income lower than $10,000 have search popularity centered around 40. They generally have a wider range of search popularity than those with median household income higher than $10,000. The latter mostly have over 70 searches for "covid". Based on the plot of median household income and "loans", as the median household income increases, the search popularity for "loans" seems to increase as well. About half of the cities with median household income lower than $125,000 have search popularity below 60, while the majority of those with median household income higher than $125,000 have search popularity above 60.