# Problem Set #1

### Insert Name

### September 28th, 2022

*Note:* Please prepare your answers using Rmarkdown and submit a pdf via Canvas. Each submission has to include all code and R output used to answer the questions. I encourage you to work on the assignments together, but each of you have to type up their responses individually. Identical submissions will not be accepted. Late submissions may only receive partial credit at my discretion.

Penney (2016) explored whether the widespread publicity about NSA/PRISM surveillance (i.e., the Snowden revelations) in June 2013 was associated with a sharp and sudden decrease in traffic to Wikipedia articles on topics that raise privacy concerns. If so, this change in behavior would be consistent with a chilling effect resulting from mass surveillance. The approach of Penney (2016) is sometimes called an interrupted time series design, and it is related to the approaches described in section 2.4.3. of Salganik (2019).

To choose the topic keywords, Penney referred to the list used by the US Department of Homeland Security for tracking and monitoring social media. The DHS list categorizes certain search terms into a range of issues, i.e., "Health Concern," "Infrastructure Security," and "Terrorism." For the study group, Penney used the 48 keywords related to "Terrorism" (see appendix table 8). He then aggregated Wikipedia article view counts on a monthly basis for the corresponding 48 Wikipedia articles over a 32-month period from the beginning of January 2012 to the end of August 2014. To strengthen his argument, he also created several comparison groups by tracking article views on other topics. Now, you are going to replicate Penney (2016). All the raw data that you will need for this activity is available from Wikipedia (https://dumps.wikimedia.org/other/pagecounts-raw/ (Links to an external site.)). Or you can get it from the R-package wikipediatrend. When you write up your responses, please note which data source you used. This activity will give you practice in data wrangling and thinking about discovering natural experiments in big data sources. It will also get you up and running with a potentially interesting data source for future projects.

```
library(lubridate)
library(dplyr)
library(ggplot2)
library(tidyverse)
library(wikipediatrend)
library(rvest)
```

   a) Read Penney (2016) and replicate his figure 2, which shows the page views for "Terrorism"-related pages before and after the Snowden revelations. Interpret the findings.

```
data1 <- wp_trend(page = c("Al-Qaeda",
                           "Terrorism",
                           "Terror",
                           "Attack",
                           "Iraq",
                           "Afghanistan",
                           "Iran",
                           "Pakistan",
                           "Agro",
                           "Environmental terrorism",
```

```
                            "Eco-terrorism",
                            "Conventional weapon",
                            "Weapons grade",
                            "Dirty bomb",
                            "Nuclear enrichment",
                            "Nuclear",
                            "Chemical weapon",
                            "Biological weapon",
                            "Ammonium nitrate",
                            "Improvised_explosive_device",
                            "Abu Sayyaf",
                            "Hamas",
                            "FARC",
                            "Irish_Republican_Army",
                            "Euskadi_ta_Askatasuna",
                            "Hezbollah",
                            "Tamil Tigers",
                            "PLO",
                            "Palestine_Liberation_Front",
                            "Car bomb",
                            "Jihad",
                            "Taliban",
                            "Suicide bomber",
                            "Suicide attack",
                            "Al-Qaeda_in_the_Arabian_Peninsula",
                            "Al-Qaeda_in_the_Islamic_Maghreb",
                            "Tehrik-i_Taliban_Pakistan",
                            "Yemen",
                            "Pirates",
                            "Extremism",
                            "Somalia",
                            "Nigeria",
                            "Political radicalism",
                            "Al-Shabaab",
                            "Nationalism",
                            "Recruitment",
                            "Fundamentalism",
                            "Islamist"),
                    from = "2012-01-01",
                    to = "2014-08-31",
                    lang = "en") %>%
    as_tibble(.) %>%
    mutate(.,
           month = format(date, "%m"),
           year = format(date, "%y")) %>%
    group_by(., month, year) %>%
    summarise(., total = sum(views)) %>%
    arrange(., year) %>%
    rowid_to_column(var = "totalmonth") %>%
    mutate(., group = factor(ifelse(totalmonth < 18, 0, 1)))

# Alternative way to go about things using the "rvest" package and with a loop.

library(tidyverse)
```

```r
library(rvest)

# Doing it for a single file.

file <- read_html("https://dumps.wikimedia.org/other/pagecounts-raw/2012/2012-01/") %>%
  html_nodes("li") %>%
  html_node("a") %>%
  html_attr("href") %>%
  str_subset("pagecounts")

temp <- tempfile()
# The next line of code will download from the internet. May take some time depending on your download
download.file(paste0("https://dumps.wikimedia.org/other/pagecounts-raw/2012/2012-01/", file[1]), temp)
gzfile(temp, 'rt')
```

```
## A connection with
## description "/var/folders/vt/yy7p3snn4r35wrhsc39ym1z80000gn/T//Rtmp4VGRtw/file8af82f5d18ac"
## class       "gzfile"
## mode        "rt"
## text        "text"
## opened      "opened"
## can read    "yes"
## can write   "no"
```

```r
mydata1 <- read.table(temp, sep = " ") %>%
  as_tibble() %>%
  filter(V1 == "en",
         V2 == "Al-Qaeda")

# What the above did was to isolate the page counts of the page "Al-Qaeda" in the English language. If
#
# for (i in 2:length(file)) {
#
#   # Doing the same deal, but within the loop.
#   download.file(paste0("https://dumps.wikimedia.org/other/pagecounts-raw/2012/2012-01/", file[i]), te
#   gzfile(temp, 'rt')
#   df <- read.table(temp, sep = "", fill = T) %>%
#     as_tibble() %>%
#     filter(V1 == "en",
#            V2 == "Al-Qaeda")
#
#   # Adding the created 1x4 data frame to the end of the main data.
#   mydata1 <- rbind(mydata1, df)
#
# }
# unlink(temp)
```

## Interpretation for A

```r
# This ought to work (to some extent; obviously it's not identical). However, the total view numbers-as

ggplot(data1, aes(x = totalmonth,
                  y = total,
```

```
                group = group)) +
 geom_point() +
 geom_smooth(method = 'lm',
             se = FALSE) +
 geom_vline(xintercept = 17.75) +
 geom_text(aes(x = data1$totalmonth[20], y = 4250000,
               label = "Mid-June 2013"), colour = "black",
               vjust = 1, size = 2.75) +
 scale_y_continuous(name = "Total Views (All 48 Wikipedia Articles)",
                    labels = scales::comma,
                    limits = c(1750000, NA),
                    breaks = seq(1500000, 4000000, 500000)) +
 scale_x_continuous(name = "Time (Months)",
                    breaks = seq(0, 32, 2))
```

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

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## Warning: Removed 2 rows containing missing values (geom_point).
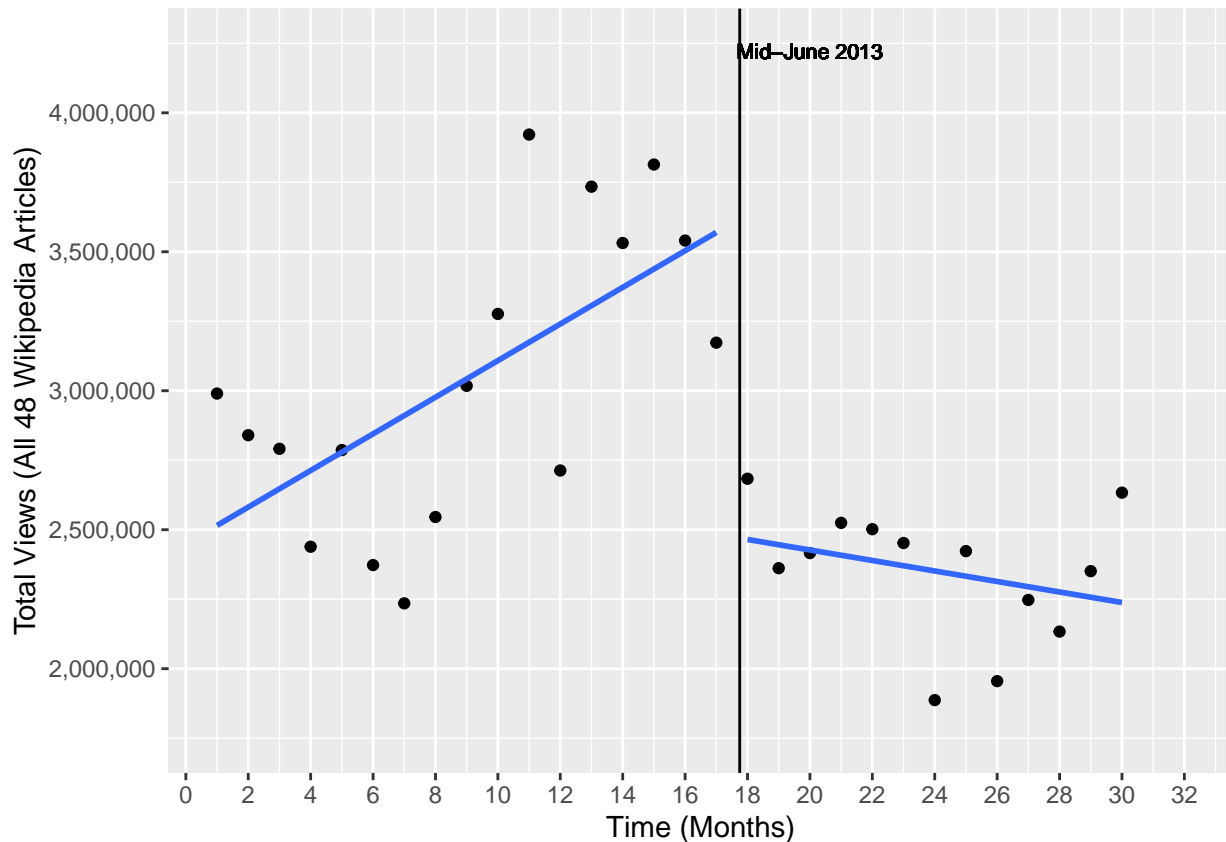


b) Next, replicate figure 4A, which compares the study group ("Terrorism"- related articles) with a comparator group using keywords categorized under "DHS & Other Agencies" from the DHS list (see appendix table 10 and footnote 139). Interpret the findings.

## Interpretation for B

```r
# Creating a data frame of the DHS and other agencies table, Table 10, p. 180.
data2 <- wp_trend(page = c("Department_of_Homeland_Security",
                           "Federal_Emergency_Management_Agency",
                           "Coast_Guard",
                           "Customs_and_Border_Protection",
                           "Border_patrol",
                           "Secret_Service",
                           "Bureau_of_Land_Management",
                           "Homeland_defense",
                           "Espionage",
                           "Task_Force_88",
                           "Central_Intelligence_Agency",
                           "Fusion_center",
                           "DEA",
                           "Secure_Border_Initiative",
                           "Federal_Bureau_of_Investigation",
                           "Alcohol_and_Tobacco_Tax_and_Trade_Bureau",
                           "U.S._Citizenship_and_Immigration_Services",
                           "Federal_Air_Marshal_Service",
                           "Transportation_Security_Administration",
                           "Air_Marshal",
                           "Federal_Aviation_Administration",
                           "National_Guard",
                           "Emergency_Management",
                           "U.S._Immigration_and_Customs_Enforcement",
                           "United_Nations"),
                  from = "2012-01-01",
                  to = "2014-08-31",
                  lang = "en") %>%
  as_tibble(.) %>%
  mutate(.,
         month = format(date, "%m"),
         year = format(date, "%y")) %>%
  group_by(., month, year) %>%
  summarise(., total = sum(views)) %>%
  arrange(., year) %>%
  rowid_to_column(var = "totalmonth") %>%
  mutate(., group = factor(ifelse(totalmonth < 18, 0, 1)))

# Replicating Figure 4A.
ggplot(data1, aes(x = totalmonth,
                  y = total,
                  group = group)) +
  geom_point(data = data1, color = "blue") +
  geom_point(data = data2, color = "red") +
  geom_smooth(data = data1,
              method = 'lm') +
  geom_smooth(data = data2,
              method = 'lm',
              color = "red") +
  geom_vline(xintercept = 17.75) +
  geom_text(aes(x = data1$totalmonth[20], y = 4250000,
```
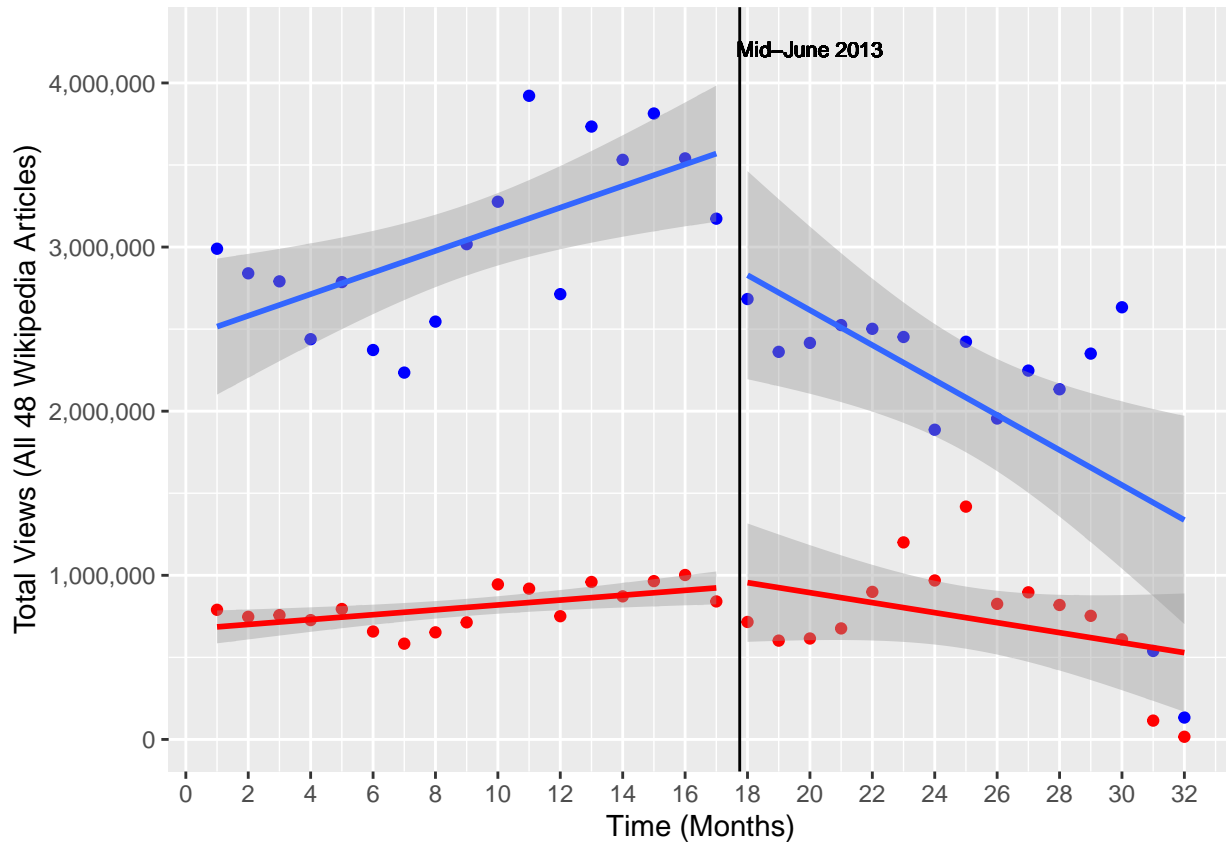
```
                    label = "Mid-June 2013"), colour = "black",
                    vjust = 1, size = 2.75) +
    scale_y_continuous(name = "Total Views (All 48 Wikipedia Articles)",
                       labels = scales::comma) +
    scale_x_continuous(name = "Time (Months)",
                       breaks = seq(0, 32, 2))
```



c) In part (b) you compared the study group with one comparator group. Penney also compared with two other comparator groups: "Infrastructure Security"–related articles (appendix table 11) and popular Wikipedia pages (appendix table 12). Come up with an alternative comparator group, and test whether the findings from part (b) are sensitive to your choice of comparator group. Which choice makes most sense? Why?

## Interpretation for C

```
# Use table 9(data1) and table 11 (data3);
# Uses the infrastructure and total views

data3 <- wp_trend(page = c("Infrastructure_security",
                           "Airport",
                           "Airplane",
                           "Chemical_burn",
                           "CIKR",
                           "Amtrak",
                           "Collapse",
                           "Information_infrastructure",
```

6

```
                               "Telecommunications_network",
                               "Telecommunication",
                               "Critical_Infrastructure",
                               "National_Information_Infrastructure",
                               "Metro_station",
                               "Washington_Metropolitan_Area_Transit_Authority",
                               "Subway",
                               "Bay_Area_Rapid_Transit",
                               "Metropolitan_Atlanta_Rapid_Transit_Authority",
                               "Port_authority",
                               "NBIC",
                               "Electrical_grid",
                               "Power",
                               "Smart",
                               "Full_body_scanner",
                               "Electric_power",
                               "Failure",
                               "Power_outage",
                               "Blackout",
                               "Brownout",
                               "Port",
                               "Dock_(maritime)",
                               "Bridge",
                               "Flight_cancellation_and_delay",
                               "Delay",
                               "Electric_power_transmission"),
                     from = "2012-01-01",
                     to = "2014-08-31",
                     lang = "en") %>%
  as_tibble(.) %>%
  mutate(.,
         month = format(date, "%m"),
         year = format(date, "%y")) %>%
  group_by(., month, year) %>%
  summarise(., total = sum(views)) %>%
  arrange(., year) %>%
  rowid_to_column(var = "totalmonth") %>%
  mutate(., group = factor(ifelse(totalmonth < 18, 0, 1)))
```

```
## `summarise()` has grouped output by 'month'. You can override using the
## `.groups` argument.
# Plot
ggplot(data1, aes(x = totalmonth,
                  y = total,
                  group = group)) +
  geom_point(data = data1, color = "black") +
  geom_point(data = data3, color = "black") +
  geom_smooth(data = data1,
              method = "lm",
              color = "orange") +
  geom_smooth(data = data3,
              method = "lm",
              color = "blue") +
```

```
    geom_vline(xintercept = 17.75) +
    geom_text(aes(x = data1$totalmonth[20], y = 4250000,
                  label = "Mid-June 2013"), colour = "black",
                  vjust = 1, size = 2.75) +
    scale_y_continuous(name = "Total Views (All Articles)",
                       labels = scales::comma) +
    scale_x_continuous(name = "Time (Months)",
                       breaks = seq(0, 32, 2)) +
  ggtitle("Terrorism and Infrastructure Security Related Articles") +
  theme_minimal()
```

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