BST 270 Individual Project

Reproducible Data Science: Joe Biden's 2019 Media Coverage (FiveThirtyEight)

Tzu-Hsi Jen

2025-01-22

Contents

1. Introduction

The purpose of this notebook is to meet the individual project criterion for the course BST270: Reproducible Data Science, taken Winter 2025.

2. Motivation and Reproducibility

The aim of this project is to accurately reproduce one figure and one table from FiveThirtyEight's The Media Frenzy Around Biden Is Fading. It utilises datasets capturing media mentions from both online and cable news sources. The online news data, drawn from the Media Cloud database, includes U.S. Top Online News and U.S. Top Digital Native News and is stored at `../data/online_weekly.csv`. Cable news data, collected through the GDELT Television API from the TV News Archive, can be found at `../data/cable_weekly.csv`.

3. Set Up

3.1 Install Required Packages and Load Libraries

First, we install and load the necessary packages for data manipulation and visualisation. This includes `lubridate` for date conversions, `dplyr` and `tidyr` for data organization, and `ggplot2` for visualizations. We also use `knitr` for generating table.

```r
# Please uncomment the following lines to install packages if they are not already installed:
#install.packages("dplyr")
#install.packages("lubridate")
#install.packages("ggplot2")
#install.packages("tidyr")
#install.packages("knitr")

# For data wrangling purposes:
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)

# For generating the plot and the table:
library(ggplot2)
library(knitr)
```

3.2 Import Data

Next, we import the datasets for analysis.

```r
online_data <- read.csv("../data/online_weekly.csv") # online news weekly data
cable_data <- read.csv("../data/cable_weekly.csv") # cable news weekly data
```

4. Data Wrangling

4.1 Figure 1

4.1.1 Online News

We start by processing `online_data` to compute the weekly percentage of online news mentions for each candidate starting from 2019-05-01. This start date aligns with the timeline presented in Figure 1 in the original article.

```r
online_news <- online_data|>
  mutate(date = ymd(date))|> # convert date into ymd format
  filter(date >= "2019-05-01")|> # select data starting from 2019-05-01
  mutate(week = floor_date(date, unit = "week", week_start = 7))|>  # week starts from Sunday
  group_by(name, week)|> # aggregate story count by week for each candidate
  summarise(matched_stores = sum(matched_stories, na.rm = TRUE),
            all_candidate_stories = sum(all_candidate_stories, na.rm = TRUE),
      # calculate percentage of story (online news) mentions for each candidate
            pct_stories = (matched_stores/all_candidate_stories)*100,
            .groups = 'drop')|>
  arrange(desc(pct_stories))|> # arrange pct_stories in descending order
  select(name, week, pct_stories)
```

We then determine the candidates with the highest and the second-highest percentage of online news mentions each week, focusing particularly on Joe Biden, as he consistently ranks highest. From here, we can compute how many times higher Joe Biden's mentions are compared to the second highest candidate, which will be used for Figure 1.

```
online_rank <- online_news|>
  group_by(week)|> # aggregate the data by week
  arrange(week, desc(pct_stories))|> # arrange by week and by pct_stories in descending order
  mutate(rank = rank(-pct_stories, ties.method = "first"))|> # rank by pct_stories
  filter(rank %in% c(1, 2))|> # select the candidates with the top two highest pct_stories
  summarise(biden_pct_stories = first(pct_stories), # Biden's pct_stories
            second_highest_pct_stories = last(pct_stories), # second highest pct_stories

            # ratio of Biden's pct_stories/ second rank's pct_stories
            times_pct_stories = (biden_pct_stories/second_highest_pct_stories),
            .groups = 'drop')
```

4.1.2 Cable News

We then proceed to process `cable_data` to compute the weekly percentage of cable news mentions for each candidate starting from 2019-05-01.

```
cable_news <- cable_data|>
  mutate(date = ymd(date))|> # convert date into date format
  filter(date >= "2019-05-01")|> # select data starting from 2019-05-01
  mutate(week = floor_date(date, unit = "week", week_start = 7))|> # week starts from Sunday
  group_by(name, week)|> # aggregate clip count by week for each candidate
  summarise(matched_clips = sum(matched_clips, na.rm = TRUE),
            all_candidate_clips = sum(all_candidate_clips, na.rm = TRUE),
            total_clips = sum(total_clips, na.rm = TRUE),

            # calculate percentage of clip (cable news) mentions for each candidate
            pct_clips = (matched_clips/all_candidate_clips)*100,
            .groups = 'drop')|>
  arrange(desc(pct_clips))|> # arrange pct_clips in descening order
  select(name, week, pct_clips)
```

Similar to the online news data, we determine the candidates with the highest and the second-highest percentages of cable news mentions each week, again focusing on Joe Biden when calculating how much more frequently he is mentioned compared to the second-highest candidate. This calculation will also contribute to Figure 1.

```
cable_rank <- cable_news|>
  group_by(week)|> # aggregate the data by week
  arrange(week, desc(pct_clips))|> # arrange by week and by pct_clips in descening order
  mutate(rank = rank(-pct_clips, ties.method = "first"))|> # rank by pct_clips
  filter(rank %in% c(1, 2))|> # show only candidates with the top two highest pct_clips
  summarise(biden_pct_clips = first(pct_clips),
            second_highest_pct_clips = last(pct_clips),

            # ratios of Biden's pct_clips/ second rank's pct_clips
            times_pct_clips = (biden_pct_clips/second_highest_pct_clips),
            .groups = 'drop')
```

After calculating the weekly mention ratios for both online and cable news, we combine these results to create a comprehensive view of Joe Biden's media coverage relative to the second-ranking candidate across both platforms. This step involves joining the datasets by week and selecting the relevant columns.

```
ratio_table <- online_rank|>
  left_join(cable_rank, by = "week")|> # join the two datasets by week
  select(week, times_pct_stories, times_pct_clips)
```

4.2 Table 1

To prepare the data for Table 1, we compute the differences in media mentions for each candidate between two specific weeks. We analyse these differences separately for cable news and online news to observe any shifts in media coverage over this one-week interval.

4.2.1 Online News

```
online_diff <- online_news|>
  filter(week %in% c("2019-10-13",  "2019-10-20"))|> # as stated in the original Table 1
  mutate(pct_stories = round(pct_stories, 1))|> # round the data to one decimal place
  select(name, week, pct_stories)|>
  # convert data frame for calculating the difference
  pivot_wider(names_from = week, values_from = pct_stories)|>
  mutate(diff = `2019-10-20` - `2019-10-13`) # calculate difference in online news mentioned
```

4.2.2 Cable News

```
cable_diff <- cable_news|>
  filter(week %in% c("2019-10-13",  "2019-10-20"))|> # as stated in the original Table
  mutate(pct_clips = round(pct_clips, 1))|> # round the data to one decimal place
  select(name, week, pct_clips)|>
  # convert data frame for calculating the difference
  pivot_wider(names_from = "week", values_from = "pct_clips")|>
  mutate(diff = `2019-10-20` - `2019-10-13`) # calculate the difference in cable news mentioned
```

5. Reproducing Plot and Table

Figure 1. Trends in Media Coverage: Joe Biden's Relative Share of Mentions Across Platforms

Figure 1 illustrates the weekly trend in media mentions for Joe Biden compared to the second most-mentioned candidate across both online and cable platforms from 2019 May to 2019 October.

```
figure1 <- ratio_table|>
  ggplot(aes(x=week))+
  geom_line(aes(y=times_pct_clips, color = "Cable News"), linewidth = 1.2)+
  geom_line(aes(y=times_pct_stories, color = "Online News"), linewidth = 1.2)+
  labs(
    title =
      "Biden's share of media mentions on each medium relative to \nthe next most-mentioned candidate ea
       x = "Month",
       y = "News Coverage")+
  scale_color_manual(values = c("Cable News" = "cadetblue3",
                                "Online News" = "darkorange2"))

figure1
```

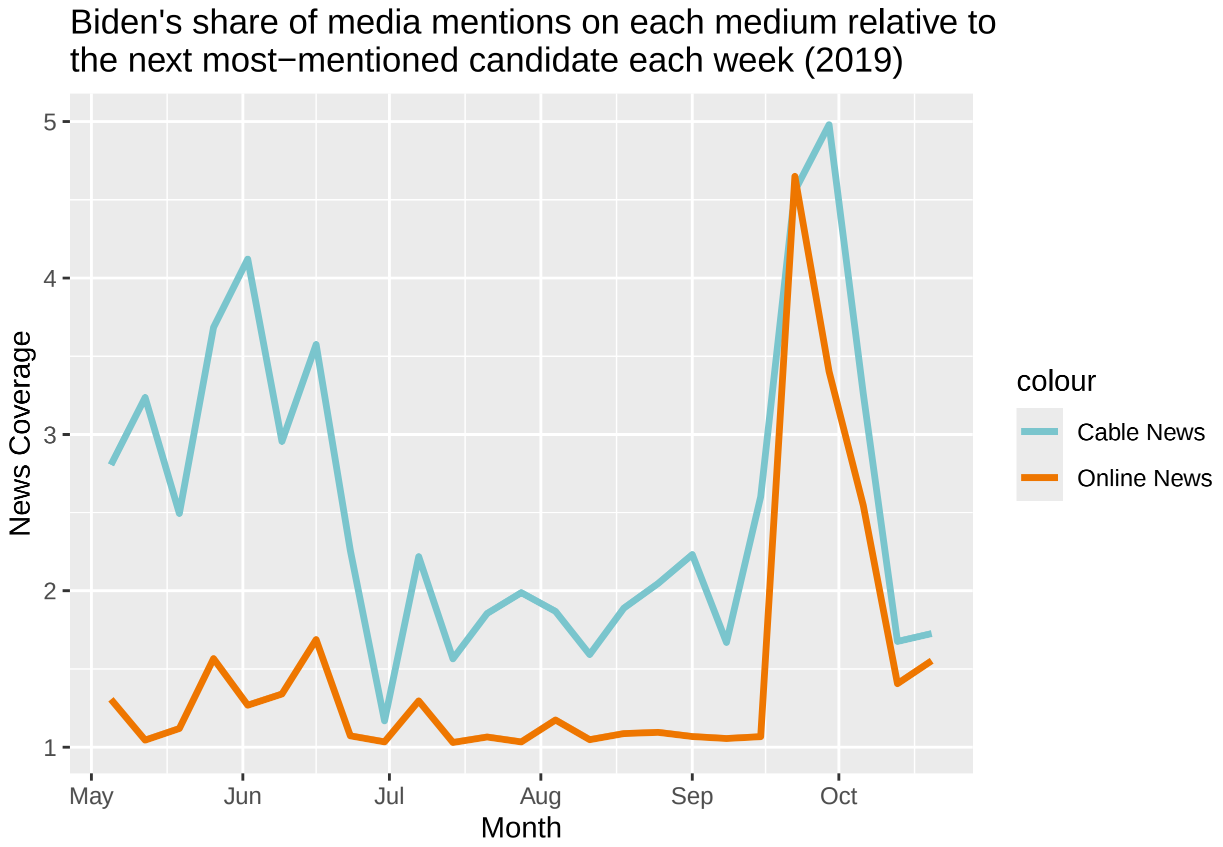

Table 1. Comparative Weekly Shifts in Media Coverage Across Platforms

Table 1 displays changes in media mentions across both online and cable platforms between the weeks of 2019-10-13 and 2019-10-20. We generate the table by combining `cable_diff` and `online_diff`.

```
# Combine the two datasets for the cable and the online news differences for each candidate
table1 <- cable_diff|>
  full_join(online_diff, by = "name")|>
  rename("Candidate" = name,             # rename the column names for clarity
         "10/13/19 Cable (%)" = `2019-10-13.x`,
         "10/20/19 Cable (%)" = `2019-10-20.x`,
         "Diff Cable (%)" = `diff.x`,
         "10/13/19 Online (%)" = `2019-10-13.y`,
         "10/20/19 Online (%)" = `2019-10-20.y`,
         "Diff Online (%)" = `diff.y`)

kable(table1)
```

Candidate	10/13/19 Cable (%)	10/20/19 Cable (%)	Diff Cable (%)	10/13/19 Online (%)	10/20/19 Online (%)	Diff Online (%)
Joe Biden	44.1	39.7	-4.4	59.2	59.0	-0.2
Elizabeth Warren	26.3	23.0	-3.3	42.1	38.0	-4.1
Bernie Sanders	16.5	17.0	0.5	34.9	30.0	-4.9
Tulsi Gabbard	7.6	15.4	7.8	12.9	13.0	0.1

Candidate	10/13/19 Cable (%)	10/20/19 Cable (%)	Diff Cable (%)	10/13/19 Online (%)	10/20/19 Online (%)	Diff Online (%)
Pete Buttigieg	5.9	6.0	0.1	19.9	18.2	-1.7
Amy Klobuchar	3.8	3.3	-0.5	13.7	8.8	-4.9
Kamala Harris	3.8	3.8	0.0	15.6	12.5	-3.1
Beto O'Rourke	3.6	2.0	-1.6	5.6	4.0	-1.6
Tom Steyer	2.4	1.8	-0.6	9.6	3.4	-6.2
Cory Booker	2.3	1.5	-0.8	11.4	7.6	-3.8
Andrew Yang	1.4	0.5	-0.9	9.9	6.0	-3.9
Tim Ryan	0.0	1.1	1.1	0.9	2.4	1.5
Julian Castro	0.7	0.9	0.2	3.2	1.9	-1.3
John Delaney	0.1	0.2	0.1	1.2	1.2	0.0
Marianne Williamson	0.2	0.2	0.0	2.0	1.6	-0.4
Michael Bennet	0.0	0.1	0.1	1.5	2.1	0.6
Steve Bullock	0.1	0.1	0.0	1.1	0.9	-0.2
Joe Sestak	0.0	0.0	0.0	0.8	0.3	-0.5

6. Limitations and Biases

The limitation of this project primary involves the lack of transparency of the data retrieval process and the difficulties in accessing the data sources. The GitHub of the original article mentions that the cable news data was sourced through the GDELT Television API, but the specific process on how this API was utilised was not clearly detailed. Although it does not affect our process in data wrangling and reproducing the figure and the table, as the `.csv` files for the data are already provided on the article's GitHub, it would still impact our attempt in trying to understand how the data was originally generated and extracted.

Furthermore, accessing online news data from Media Cloud database (collected from U.S. Top Online News and U.S. Top Digital Native News) led to security and privacy warnings. Similarly, the warnings do not affect the purpose of this project, they hinder the process of verifying how the data was originally extracted.

To improve this project, future documentation should provide clearer guidelines on data retrieval and ensure the data sources are secure and transparent about their data handling practices, thus enhancing the project's credibility and usability.