

# Problem Set 2: Data Wrangling

Thomas-J-Griffiths

## Background

Political advertising has traditionally been focused on the medium of television, but in recent cycles, online advertising has become much more popular. In this problem set, you will explore a dataset that has information on Facebook ad spending and impressions by candidates in the 2018 election cycle in the United States. The variables in this data are described below.

Name	Description
<code>cand_id</code>	unique identifier code for candidate
<code>cand_name</code>	full name of the candidate
<code>cand_name_last</code>	last name of the candidate
<code>party</code>	party affiliation of the candidate (R = Republican, D = Democrat)
<code>office</code>	office being sought by candidate
<code>state</code>	state in which the candidate is running
<code>incumbency</code>	incumbency status of candidate (incumbent, challenger, or open seat)
<code>spend</code>	estimated total spending on Facebook ads by candidate
<code>impressions</code>	estimated total impressions of Facebook ads
<code>ad_tone_attack</code>	proportion of FB ads that mention candidate's opponent only
<code>ad_tone_promote</code>	proportion of FB ads that mention candidate only
<code>ad_tone_contrast</code>	proportion of FB ads that mention candidate and candidate's opponent

## Question 1 (8 points)

Load the data using the `read_csv` function and save it as `fb_ads` (using this will automatically make `fb_ads` a tibble). In the text, describe how many candidates there are in the dataset.

Use `dplyr` functions to create a table with the number of candidates in each type of incumbency status in the data set. Save this table output as `incumbency_table` (for the autograder). Use the function `knitr::kable()` on this table to have a nicely formatted table produced in the knitted output.

**Rubric:** 2pt for loading the data (autograder); 1pt for describing the number of candidates (PDF); 3pts for creating the table (autograder); 2pt for using `kable()` to nicely format the output (PDF)

## Answer 1

```
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

library(readr)
library(knitr)
library(ggplot2)
fb_ads <-
  read_csv("data/fb_ads.csv")

## Rows: 7014 Columns: 12

## -- Column specification -----
## Delimiter: ","
## chr (7): cand_id, cand_name, cand_name_last, party, office, state, incumbency
## dbl (5): spend, impressions, ad_tone_attack, ad_tone_promote, ad_tone_contrast
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.

## There are 7014 candidates in the data set

incumbency_table <- fb_ads |>
  group_by(incumbency) |>
  summarize(candidate_count = n())

knitr::kable(incumbency_table)
```

incumbency	candiate_count
Challenger	2510
Incumbent	2022
Open Seat	2482

## Question 2 (7 points)

Filter the data to just US House and US Senate races and use this to create a tibble called `party_incumbent_promote` that has 6 rows that summarizes the average of `ad_tone_promote` for each combination of `party` and `incumbency`. Call the variable summarizing the promote variable as `promote_prop` and be sure to remove any missing values when computing the averages.

Use `knitr::kable()` to produce a nicely formatted table. In this call, set the `digits` arguments to 3 and use the `col.names` argument to pass a nicer set of names. You can use the following as a template:

```
knitr::kable(my_table, col.names = c("Variable 1", "Variable 2", ...))
```

In the writeup, describe which type of candidate sponsored the most promoting ads on average.

**Rubric:** 3pts for creating `party_incumbent_promote` correctly (autograder); 2pt for a nicely formatted table (PDF); 1pt for changing the column names of the output table (PDF); 1pt for correctly identifying the type of candidate with highest average (PDF)

## Answer 2

```
party_incumbent_promote <- fb_ads |>
  filter(office == "US Senate" | office == "US House") |>
  mutate(promote_prop = mean(ad_tone_promote)) |>
  group_by(incumbency, party) |>
  summarize(promote_prop = mean(ad_tone_promote, na.rm = TRUE))
```

## 'summarise()' has grouped output by 'incumbency'. You can override using the '.groups' argument.

```
party_incumbent_promote
```

```
## # A tibble: 6 x 3
## # Groups:   incumbency [3]
##   incumbency party promote_prop
##   <chr>      <chr>      <dbl>
## 1 Challenger D          0.833
## 2 Challenger R          0.813
## 3 Incumbent  D          0.854
## 4 Incumbent  R          0.792
## 5 Open Seat  D          0.845
## 6 Open Seat  R          0.828
```

```
knitr::kable(digits = 3, party_incumbent_promote, col.names = c("Incumbency", "Party", "Promote Ad Proportion"))
```

Incumbency	Party	Promote Ad Proportion
Challenger	D	0.833
Challenger	R	0.813
Incumbent	D	0.854
Incumbent	R	0.792
Open Seat	D	0.845

Incumbency	Party	Promote Ad Proportion
Open Seat	R	0.828

*## Type of candidate with highest average ad\_tone\_promote are Incumbent Democrats at 85.4%*

### Question 3 (7 points)

Create a new variable called `impressions_millions` that is the total Facebook ad impressions in millions (as opposed to single impressions). Make sure to save the resulting dataset back as `fb_ads`.

Create a histogram of this variable for just the US House races. Save the ggplot output as `plot_q3` and also print it to produce a plot in the output. In the text, describe the shape of the histogram and tell the reader if most of the House candidates had more than 10 million ads impressions on Facebook.

**Rubric:** 2pt for creating the new variable (autograder); 3pts for creating the histogram object (autograder); 2pts for answering the question about the histogram (PDF)

### Answer 3

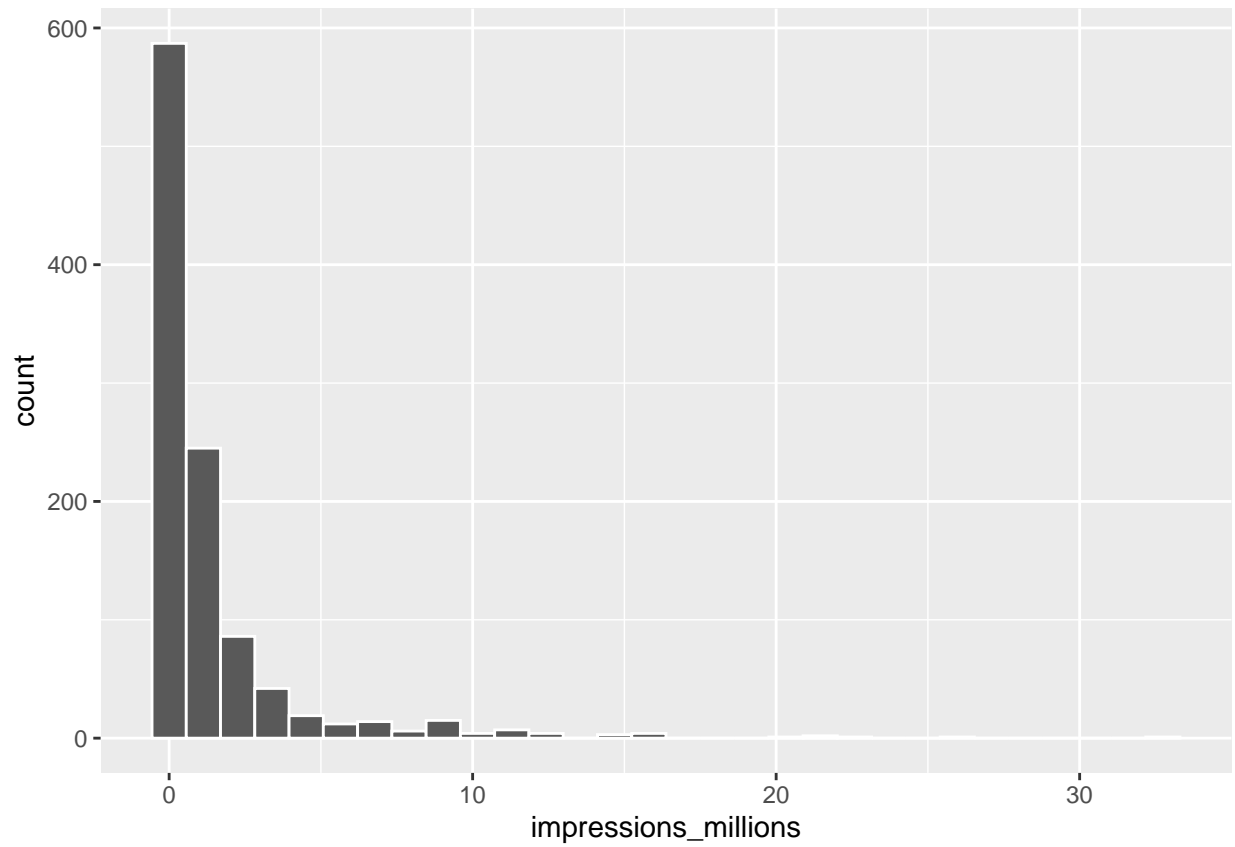
```
impressions_millions <- fb_ads |>
  mutate(impressions_millions = impressions / 1e6)
  impressions_millions
```

```
## # A tibble: 7,014 x 13
##   cand_id cand_name cand_name_last party office state incumbency spend impressions ad_tone_attack
##   <chr>    <chr>      <chr>      <chr> <chr>  <chr> <chr>      <dbl>      <dbl>      <dbl>
## 1 ftm_223~ RAUNER, ~ RAUNER      R    Gover~ IL    Incumbent 2.88e5 18066608 0.413
## 2 ftm_223~ PRITZKER~ PRITZKER    D    Gover~ IL    Challenger 3.84e6 97795310 0.0210
## 3 ftm_223~ RAOUL, K~ RAOUL      D    Other~ IL    Open Seat 1.67e5 7246472. 0.0583
## 4 ftm_223~ HAROLD, ~ HAROLD    R    Other~ IL    Open Seat 3.85e4 4275971 0.0309
## 5 ftm_223~ WHITE, J~ WHITE      D    Other~ IL    Incumbent 7.07e3 1768474. 0.0367
## 6 ftm_223~ HELLAND,~ HELLAND    R    Other~ IL    Challenger 2.54e3 262490. 0.169
## 7 ftm_223~ FRERICHS~ FRERICHS    D    Other~ IL    Incumbent 1.38e5 9123175 0.0848
## 8 ftm_223~ SALAZAR,~ SALAZAR    D    Other~ CO    Open Seat 4.95e1 3000. 0.0171
## 9 ftm_223~ WEISER, ~ WEISER    D    Other~ CO    Open Seat 7.53e4 4722822. 0.0151
## 10 ftm_223~ HILL, ST~ HILL      R    Other~ CO    Open Seat 1.39e3 46994. 0.0180
## # i 7,004 more rows
## # i 3 more variables: ad_tone_promote <dbl>, ad_tone_contrast <dbl>, impressions_millions <dbl>
```

```
fb_UShouse <- impressions_millions |>
  filter(office == "US House")

plot_q3 <- ggplot(data = fb_UShouse, mapping = aes(x = impressions_millions)) +
  geom_histogram(color = "white")
print(plot_q3)
```

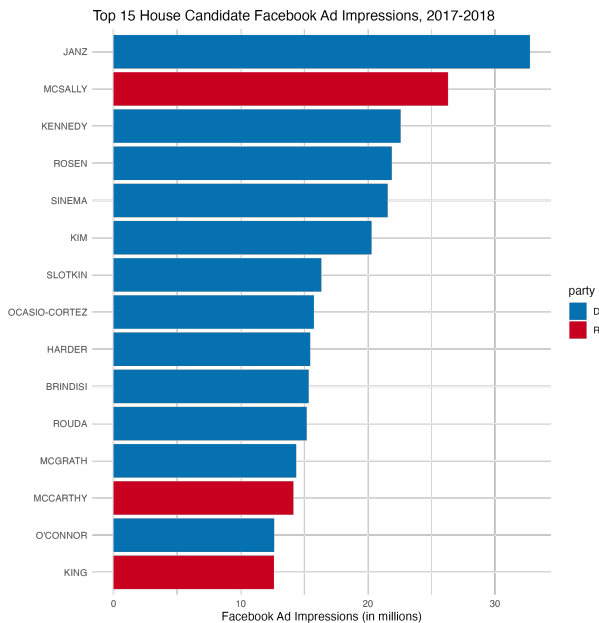
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



*## Histogram heavily tilted towards lower amount of impressions, most candidates didn't have more than*

## Question 4 (13 points)

Let's now recreate the following plot that shows the top 15 House candidates in terms of Facebook ad impressions.



You should save the `ggplot` output as `fb_top_plot`. You should also write `fb_top_plot` on its own line in the chunk to produce the actual plot. The key features of this graph that you should replicate for the autograder are:

- The barplot should have candidate last names on the y-axis and the `impressions_millions` variable from question 3 on the x-axis.
- The data feeding into the `ggplot` call should only have US House candidates and only the candidates with the highest 15 `impressions_millions` values.
- The y-axis should be ordered in ascending values of `impressions_millions` so that the lowest values are at the bottom. You may want to manipulate `cand_name_last` to achieve this.
- The fill color of the bar plot should be mapped to the `party` variable (but not globally!).

You do not need to exactly match the labels, but you should have informative labels. The color does not need to match, but if you want to change the fill colors, you can use the `scale_fill_manual(values = c(R = "red", D = "blue"))` function (where you can change the red and blue to whatever you want).

**Rubric:** 3pts for correct axes (autograder); 3pts for correct data fed into `ggplot` (autograder); 3pts for the correct ordering of the y-axis (PDF); 3pts for fill being mapped to `party` (autograder); 1pt for plot being in knitted output and having informative labels (PDF)

## Answer 4

```
house_top15 <- fb_UShouse |>
  arrange(desc(impressions)) |>
  select(cand_name_last, impressions, party) |>
  slice_head(n = 15)

house_top15
```

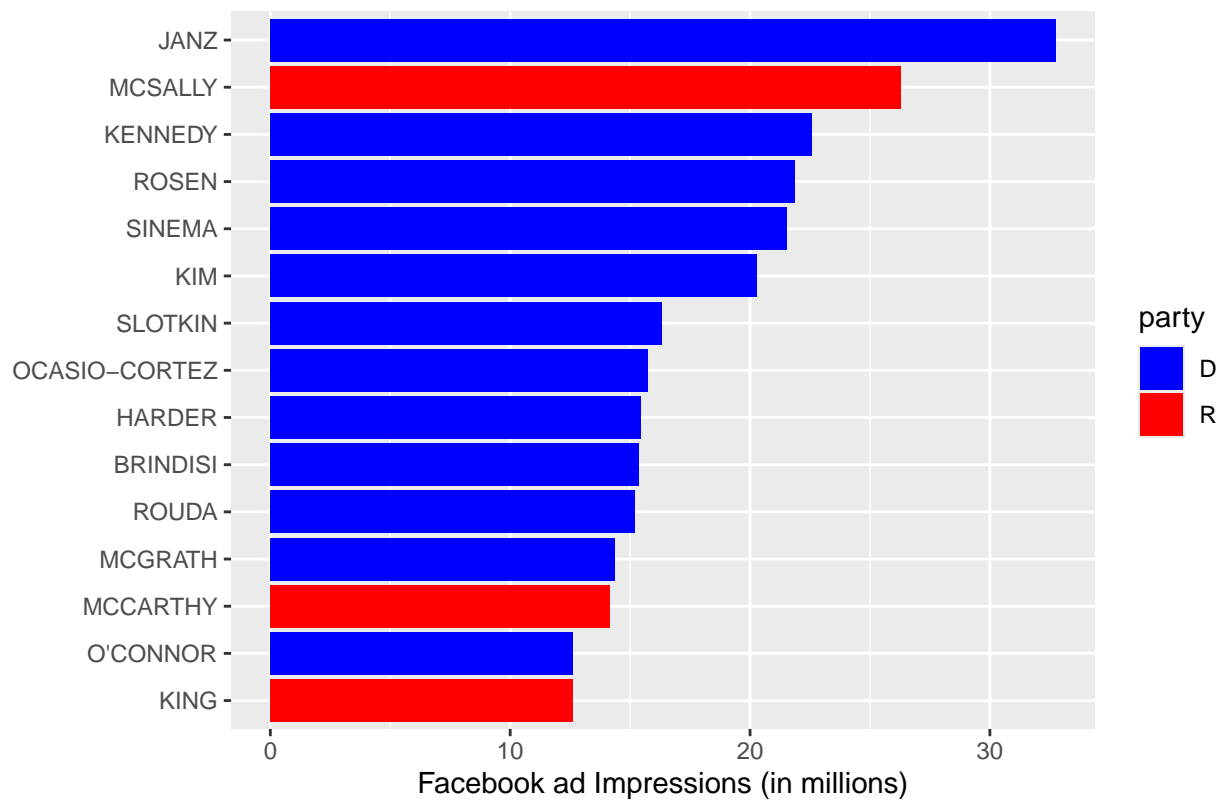


```
## # A tibble: 15 x 3
##   cand_name_last impressions party
##   <chr>           <dbl> <chr>
## 1 JANZ             32741906 D
## 2 MCSALLY          26300804 R
## 3 KENNEDY          22569831 D
## 4 ROSEN            21876048. D
## 5 SINEMA           21550630. D
## 6 KIM              20286520. D
## 7 SLOTKIN          16337279 D
## 8 OCASIO-CORTEZ    15757652. D
## 9 HARDER           15463694. D
## 10 BRINDISI         15345345 D
## 11 ROUDA            15186202. D
## 12 MCGRATH          14363196. D
## 13 MCCARTHY         14142056 R
## 14 O'CONNOR         12631748 D
## 15 KING            12615799 R
```

```
fb_top_plot <- ggplot(data = house_top15, mapping = aes(y = reorder(cand_name_last, impressions), x = 
  geom_bar(stat = "identity") +
  labs(title = "Top 15 House Canidate Facebook Ad Impressions, 2017-2018",
        x = "Facebook ad Impressions (in millions)",
        y = NULL) +
  scale_fill_manual(values = c(R = "red", D = "blue"))
```

```
fb_top_plot
```

Top 15 House Canidate Facebook Ad Impressions, 2017–2018



## Code

```
options(width = 100)
library(dplyr)
library(readr)
library(knitr)
library(ggplot2)
fb_ads <-
  read_csv("data/fb_ads.csv")
## There are 7014 candidates in the data set

incumbency_table <- fb_ads |>
  group_by(incumbency) |>
  summarize(candidate_count = n())

knitr::kable(incumbency_table)
knitr::kable(my_table, col.names = c("Variable 1", "Variable 2", ...))
party_incumbent_promote <- fb_ads |>
  filter(office == "US Senate" | office == "US House") |>
  mutate(promote_prop = mean(ad_tone_promote)) |>
  group_by(incumbency, party) |>
  summarize(promote_prop = mean(ad_tone_promote, na.rm = TRUE))
party_incumbent_promote

knitr::kable(digits = 3, party_incumbent_promote, col.names = c("Incumbency", "Party", "Promote Ad Prop"))

## Type of candidate with highest average ad_tone_promote are Incumbent Democrats at 85.4%
impressions_millions <- fb_ads |>
  mutate(impressions_millions = impressions / 1e6)
impressions_millions

fb_UShouse <- impressions_millions |>
  filter(office == "US House")

plot_q3 <- ggplot(data = fb_UShouse, mapping = aes(x = impressions_millions)) +
  geom_histogram(color = "white")
print(plot_q3)

## Histogram heavily tilted towards lower amount of impressions, most candidates didn't have more than
house_top15 <- fb_UShouse |>
  arrange(desc(impressions)) |>
  select(cand_name_last, impressions, party) |>
  slice_head(n = 15)

house_top15

fb_top_plot <- ggplot(data = house_top15, mapping = aes(y = reorder(cand_name_last, impressions), x =
  geom_bar(stat = "identity") +
  labs(title = "Top 15 House Candiate Facebook Ad Impressions, 2017-2018",
    x = "Facebook ad Impressions (in millions)",
    y = NULL) +
  scale_fill_manual(values = c(R = "red", D = "blue"))
```

fb\_top\_plot