Political Analysis of Social Media Data

R Basics 2

Instructor: Gregory Eady
Office: 18.2.10
Office hours: Fridays 13-15

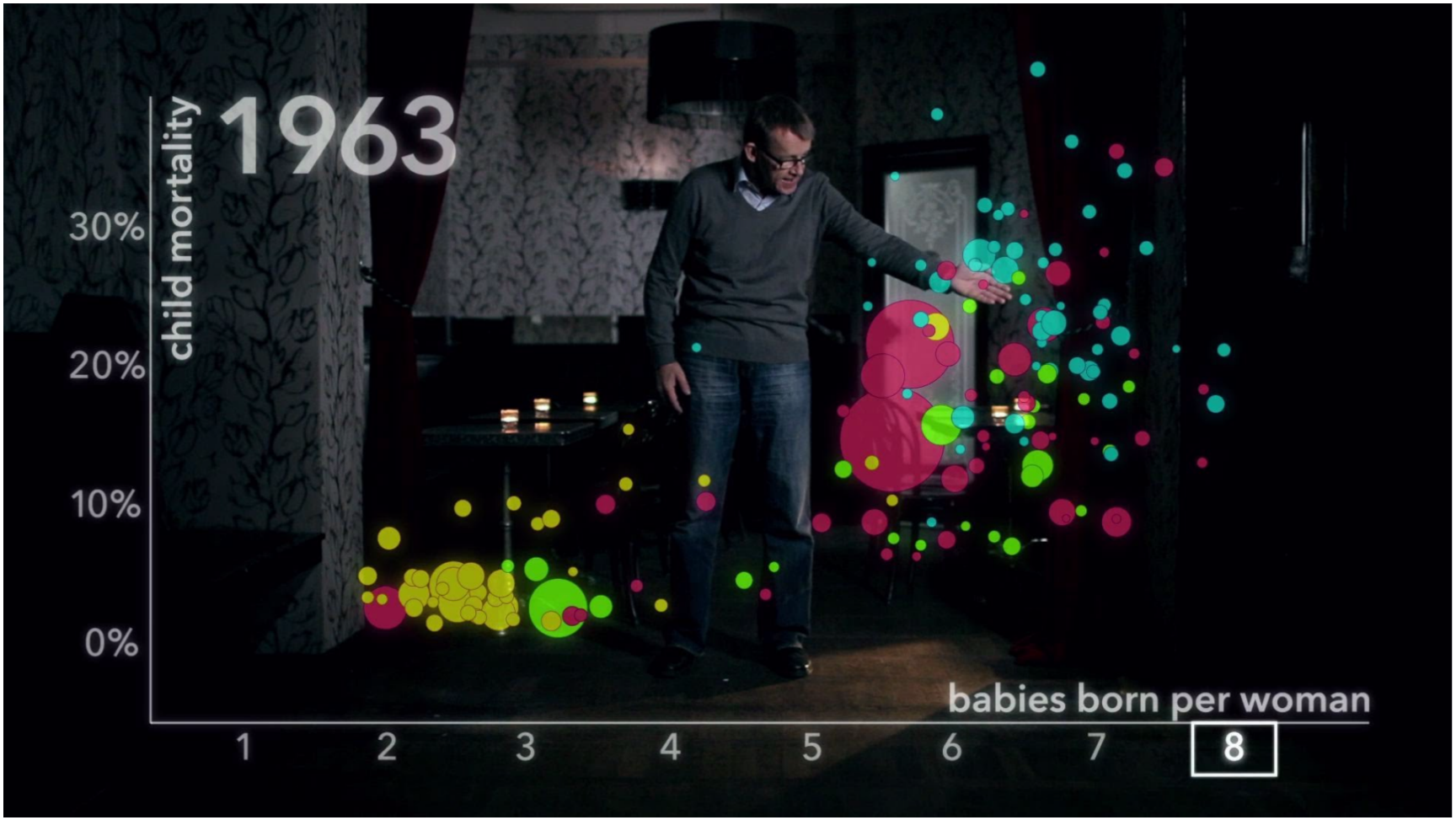
Today

- Tidyverse exercise
- Data visualization

If you haven't already, download the following file from the course website

`GM.csv`

Gapminder data



Loading data in R

```
# Load the tidyverse library
library(tidyverse)

# Set your working directory (where you saved your file)
setwd("~/Downloads/")

# Load GapMinder data
GM <- read_csv("GM.csv")
```

Note: The path (or file name) will depend on where you save the .csv file.

Gapminder variables

Variable	Description
`country`	Country
`continent`	Continent that country is located in
`year`	Year
`lifeExp`	Life expectancy at birth
`pop`	Total population size
`gdpPercap`	GDP (per capita)

Creating variables within a data.frame

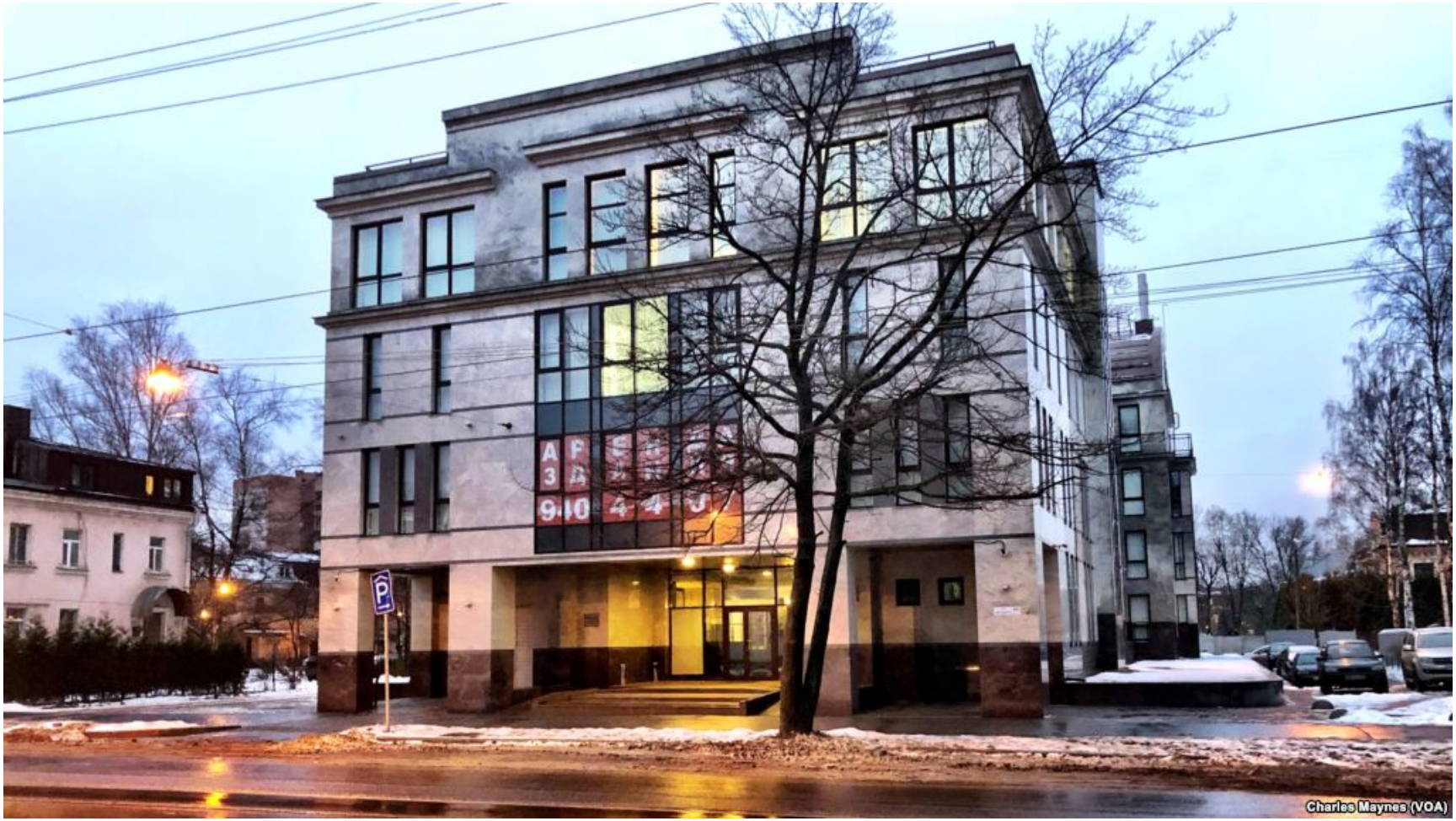
```
##### Create variable as the log of population
Result_2 <- Result_2 %>%
            mutate(log_median_pop = log(median_pop))

##### Or similarly without mutate()
Result_2$log_median_pop <- log(Result_2$median_pop)


##### Recode values in a variable
Result_2 <- Result_2 %>%
            mutate(country = recode(country,
                                    "United States" = "USA",
                                    "United Kingdom" = "UK"))

# Or like this
Result_2 <- Result_2 %>%
            mutate(country = case_when(country == "United States"
                                          ~ "USA",
                                       country == "United Kingdom"
                                          ~ "UK"))
```

Russia's Internet Research Agency



3 million Russian troll tweets

This directory contains data on nearly 3 million tweets sent from Twitter handles connected to the Internet Research Agency, a Russian "troll factory" and a defendant in an indictment filed by the Justice Department in February 2018, as part of special counsel Robert Mueller's Russia investigation. The tweets in this database were sent between February 2012 and May 2018, with the vast majority posted from 2015 through 2017.

Source:

`https://github.com/fivethirtyeight/russian-troll-tweets`

Exercise

1. Download & load a Russian troll dataset from `https://github.com/fivethirtyeight/russian-troll-tweets`
2. What are the most frequent and second-most frequent languages?
3. What region had the most tweets received by followers?
4. On average, how many followers did each tweet reach in each region?
5. How many tweets are retweets in each language?
6. How many tweets are *not* retweets in each language?
7. How frequently are Trump and Clinton mentioned in the tweets?

Exercise 1 (solution)

```
# EXERCISE 1
# 1. Download and load a Russian troll dataset from
# https://github.com/fivethirtyeight/russian-troll-tweets
# Note: The file and path will depend on what file you
# download and where you saved it on your computer
library(tidyverse)

# Set working directory
setwd("~/Downloads/")

# Load data
IRA <- read_csv("IRAhandle_tweets_1.csv.bz2")
```

Exercise 2 (solution)

```
# EXERCISE 2
# 2. What are the most frequent and second-most
# frequent languages?
IRA %>%
  count(language) %>%
  arrange(desc(n))
```

Exercise 3 (solution)

```
# EXERCISE 3
# 3. What region had the most tweets received by followers?
# arrange() is not necessary here, but it is often useful
# to use for looking at the resulting data
IRA %>%
  group_by(region) %>%
  summarize(followers = sum(followers)) %>%
  arrange(desc(followers))
```

Exercise 4 (solution)

```
# EXERCISE 4
# 4. On average, how many followers did each tweet
# reach in each region?
IRA %>%
  group_by(region) %>%
  summarize(followers = mean(followers)) %>%
  arrange(desc(followers))
```

Exercise 5 (solution)

```
# EXERCISE 5
# 5. How many tweets are retweets in each language?
IRA %>%
  group_by(language) %>%
  summarize(num_retweets = sum(retweet)) %>%
  arrange(desc(num_retweets))
```

Exercise 6 (solution)

```
# EXERCISE 6
# 6. How many tweets are _not_ retweets in each language?
# There are various ways to test if a tweet is a retweet
# E.g. sum(retweet != 1), sum(!retweet), sum(retweet == 0)
IRA %>%
  group_by(language) %>%
  summarize(num_retweets = sum(retweet == 0)) %>%
  arrange(desc(num_retweets))
```

Exercise 7 (solution)

```
# EXERCISE 7
# 7. How frequently are Trump and Clinton mentioned in the tweets?
# This exercise is more involved.
# str_count(): tells you the number of pattern matches in a string
# e.g. str_count("Trump Trumps Clinton", pattern = "Trump") is 2
# e.g. str_count("Trump Trumps Clinton", pattern = "Clinton") is 1
# Then we sum up all of these counts with sum()
# Notice, however, the use of tolower(), which makes a string lowercase.
# Why do this?
# Because str_count("Trump trumps Clinton", pattern = "trump") return 1 not 2
# If we turn all characters to lower case first, however, we don't need to
# worry about case matching
# i.e. tolower("Trump trumps Clinton") returns "trump trumps clinton"
# Thus str_count(tolower("Trump trumps Clinton"), "trump") returns 2
IRA %>%
  summarize(trump_mentions = sum(str_count(tolower(content),
                                           pattern = "trump")),
            clinton_mentions = sum(str_count(tolower(content),
                                             pattern = "clinton")))
```