Lecture 7: Obtaining data for RStudio

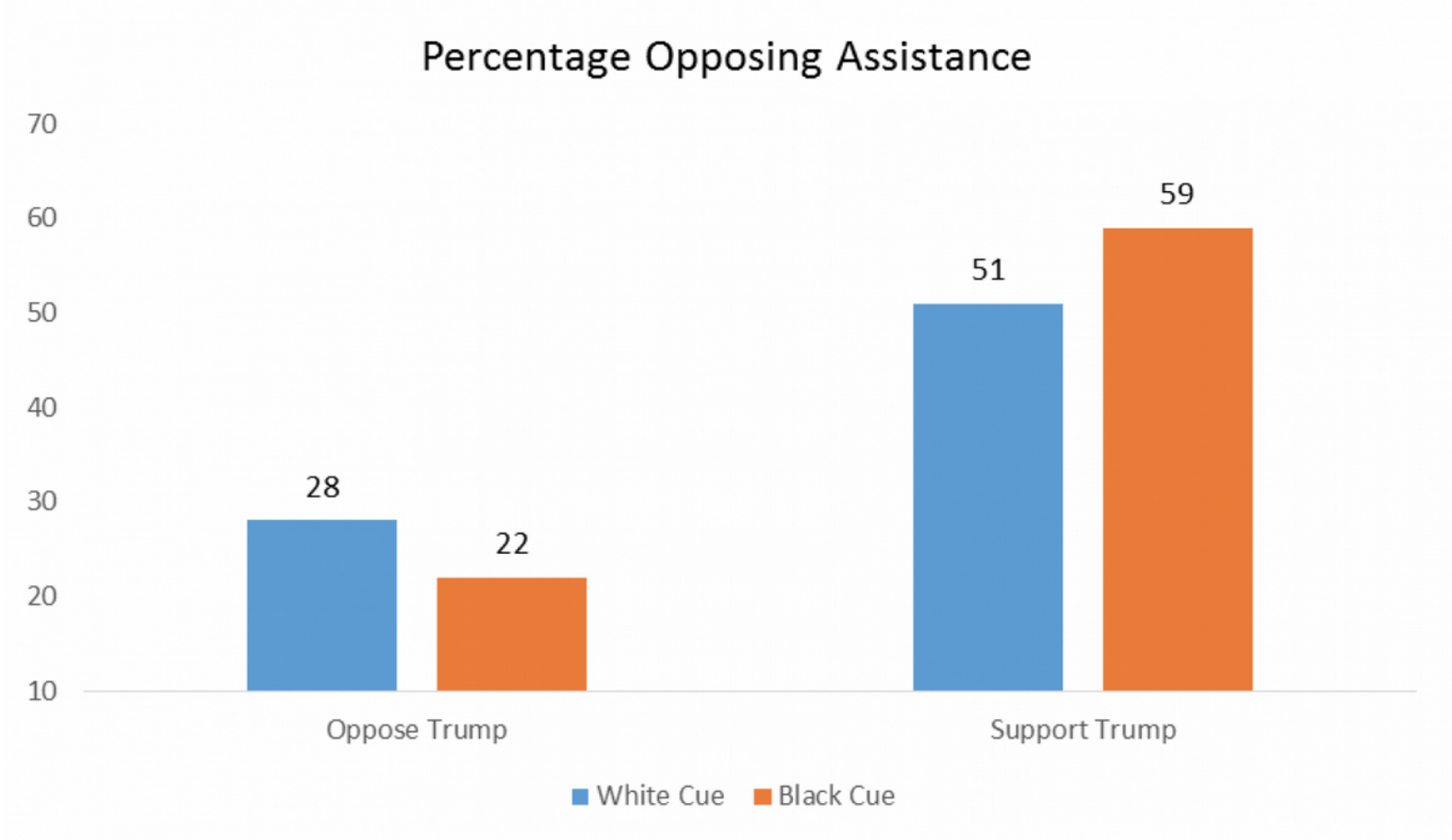
Michael Lopez, Skidmore College

Data viz's in the news

We showed Trump voters photos of black and white Americans. Here's how it affected their views., (Wired.com)

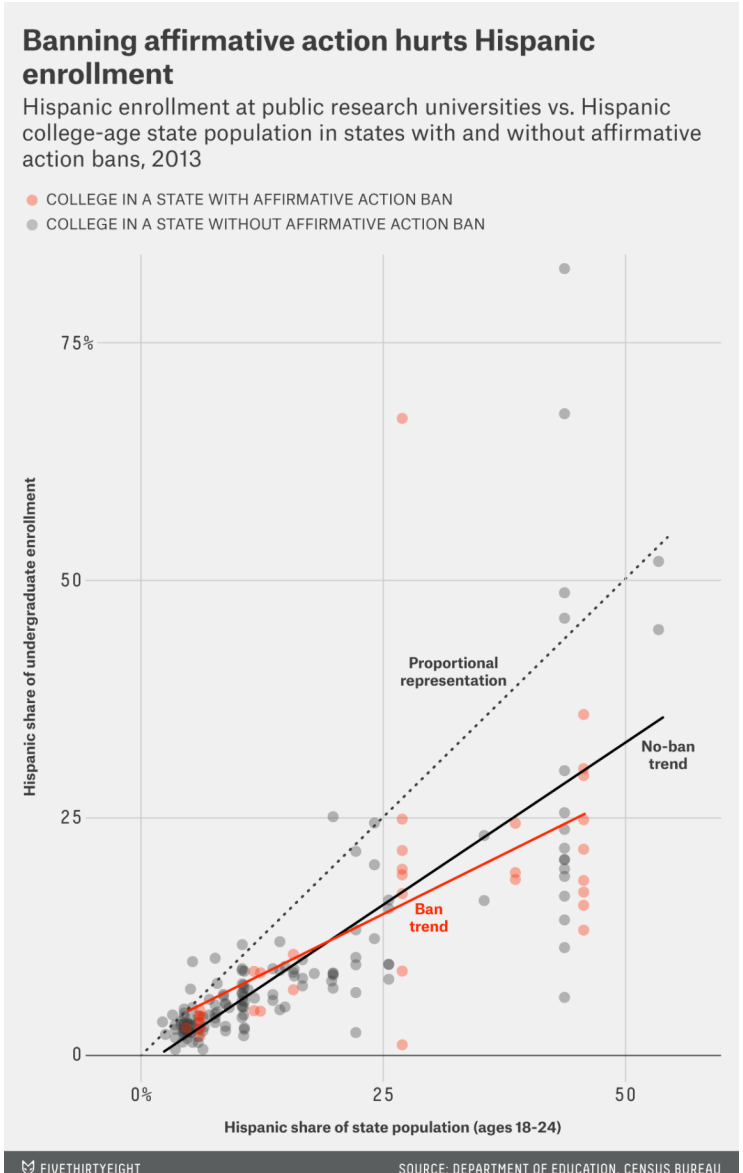
How we know race shapes Trump supporters' political attitudes

In the graph below, we present the percentage of respondents who said that they opposed the mortgage relief program, split by Trump support and which experimental condition they were assigned to.



Data viz's in the news

Here's What Happens When You Ban Affirmative Action In College Admissions, (fivethirtyeight)



Data viz's in the news

Most of Trump's charts skew the data. And not always in his favor., (Washington Post)

https://www.washingtonpost.com/graphics/politics/2016-election/trump-charts/

Today's goals

0. What does your data look like?
1. Accessing stored data
2. Data stored on the Internet
3. Data stored on Googlesheets
4. Scraping an HTML table from the Internet

What does your data look like?

- First row is a header
- First column is subject/unit identifier
- Avoid names, values or fields with blank spaces, otherwise each word will be interpreted as a separate variable
- If you want to concatenate words, inserting a . in between to words instead of a space
- Short names are prefered over longer names;
- Avoid using names that contain symbols such as ?, $,%, ^, &, *, (,),-,#, ?,,,<,>, /, |, , [,] ,{, and };
- Delete any comments that you have made in your Excel file to avoid extra columns
- Missing values in your data set are indicated with NA.

Accessing stored data, table

```
df <- read.table("<FileName>.txt")
```

`<FileName>` example, Dropbox: `~/Dropbox/DataViz/Name.txt`
`<FileName>` example, Windows: `C:/My Documents/DataViz/Name.txt`

Hint: Right click to find file location

Accessing stored data, csv

```
df <- read.table("<FileName>.csv")
```

Contents of .csv file

```
Col1,Col2,Col3
1,2,3
4,5,6
7,8,9
a,b,c
```

Hint: Save as a .csv using Excel

Hint: csv's easier to use and deal with than xlsx's or xls's.

Hint: the `readxl` package is useful for important Excel files

Hint: the `foreign` package is useful for importing Stata or SPSS files

Googlesheets

Here's a public link to a sample data set, stored on google sheets (link)

Googlesheets

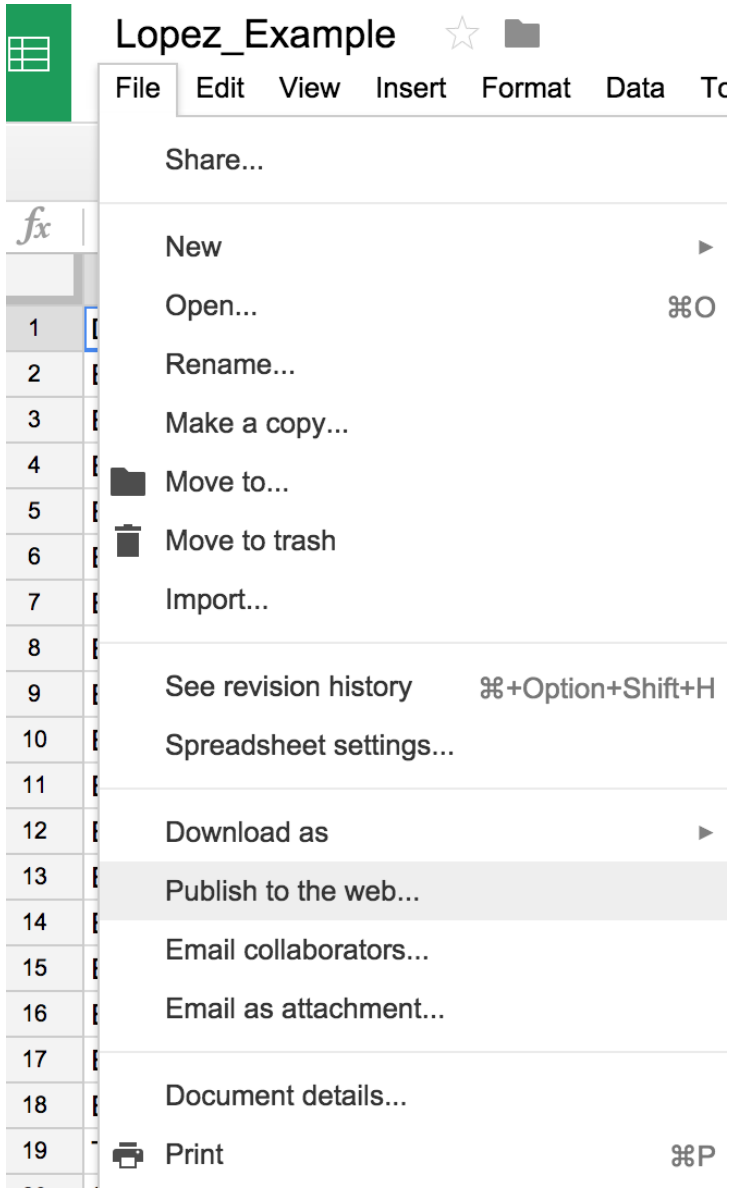


Figure 3:Step 1

Googlesheets

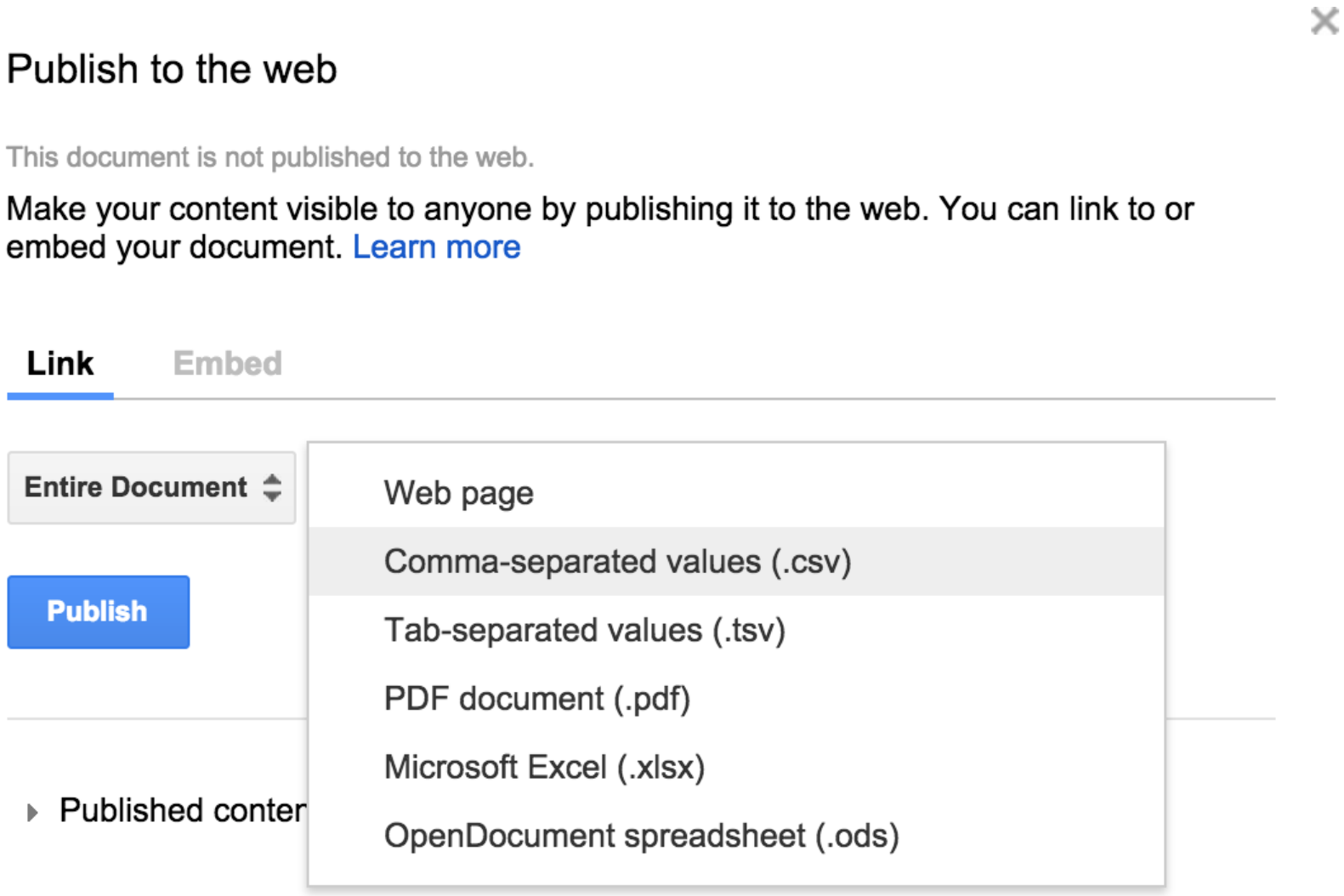


Figure 4:Step 2

Googlesheets

```
library(dplyr)
url <- "https://docs.google.com/spreadsheets/d/1wRAIt7W2mgaTqbdvutkZm_gOzJEuLaMcD7bIJj1UAg(
lopez.example <- read.csv(url)
lopez.example %>% head()
```

```
##   DraftGroup WinShares      DraftLocation
## 1   Expected      -1.4 Pick No. 1 - No. 5
## 2   Expected      33.2 Pick No. 1 - No. 5
## 3   Expected      11.4 Pick No. 1 - No. 5
## 4   Expected      17.0 Pick No. 1 - No. 5
## 5   Expected      21.6 Pick No. 1 - No. 5
## 6   Expected      41.9 Pick No. 1 - No. 5
```

Hint: You'll need an Internet connection, and it may take a few seconds to run

Scraping data, an introduction

Here's the webpage we are going to scrape

`http://www.hockey-reference.com/leagues/NHL_2015_games.html`

```
library(XML) #You'll need to install this
library(RCurl) #You'll need to install this

url <- "http://www.hockey-reference.com/leagues/NHL_2015_games.html"
tables.web <- readHTMLTable(url)
```

Scraping data, an introduction

```
names(tables.web)
```

```
## [1] "games"          "games_playoffs"
```

```
tables.web$games %>% head(3)
```

```
##         Date              Visitor G              Home G  Att. LOG Notes
## 1 2014-10-08 Philadelphia Flyers 1     Boston Bruins 2
## 2 2014-10-08   Vancouver Canucks 4    Calgary Flames 2
## 3 2014-10-08     San Jose Sharks 4 Los Angeles Kings 0
```

```
tables.web$games_playoffs %>% head(3)
```

```
##         Date             Visitor G                 Home G     Att. LOG Notes
## 1 2015-04-15    Ottawa Senators 3  Montreal Canadiens 4
## 2 2015-04-15 Chicago Blackhawks 4 Nashville Predators 3 2OT
## 3 2015-04-15     Calgary Flames 2   Vancouver Canucks 1
```

Data manipulation

```
reg.season <- tables.web$games
names(reg.season)
```

```
## [1] "Date"    "Visitor" "G"       "Home"    "G"       ""        "Att."
## [8] "LOG"     "Notes"
```

```
names(reg.season) <- c("Date", "Visitor", "Vis.Goals", "Home", "Home.Goals", "OT",
                       "Blank1", "Blank2", "Blank3")
reg.season %>% head(3)
```

```
##         Date             Visitor Vis.Goals             Home Home.Goals OT
## 1 2014-10-08 Philadelphia Flyers         1     Boston Bruins          2
## 2 2014-10-08   Vancouver Canucks         4    Calgary Flames          2
## 3 2014-10-08     San Jose Sharks         4 Los Angeles Kings          0
##   Blank1 Blank2 Blank3
## 1
## 2
## 3
```

Data manipulation

```
reg.season1 <- reg.season %>%
  select(Date:OT) %>%
  mutate(OT = ifelse(!OT == "", "Yes", "No"))

reg.season1 %>% head(3)
```

```
##         Date              Visitor Vis.Goals              Home Home.Goals OT
## 1 2014-10-08 Philadelphia Flyers         1     Boston Bruins          2 No
## 2 2014-10-08   Vancouver Canucks         4    Calgary Flames          2 No
## 3 2014-10-08     San Jose Sharks         4 Los Angeles Kings          0 No
```

Data manipulation

```
str(reg.season1)
```

```
## 'data.frame':    1230 obs. of  6 variables:
##  $ Date      : Factor w/ 178 levels "2014-10-08","2014-10-09",..: 1 1 1 1 2 2 2 2 2 2 .
##  $ Visitor   : Factor w/ 30 levels "Anaheim Ducks",..: 22 28 24 16 30 9 7 3 5 8 ...
##  $ Vis.Goals : Factor w/ 9 levels "0","1","2","3",..: 2 5 5 5 7 4 4 2 6 1 ...
##  $ Home      : Factor w/ 30 levels "Anaheim Ducks",..: 3 5 14 27 2 4 10 11 12 15 ...
##  $ Home.Goals: Factor w/ 9 levels "0","1","2","3",..: 3 3 1 4 3 2 3 3 3 6 ...
##  $ OT        : chr  "No" "No" "No" "No" ...
```

```
reg.season2 <- reg.season1 %>%
  mutate(Vis.Goals = as.numeric(as.character(Vis.Goals)),
         Home.Goals = as.numeric(as.character(Home.Goals)),
         Total.Goals = Vis.Goals + Home.Goals)
reg.season2 %>% head(3) %>% print.data.frame()
```

```
##         Date             Visitor Vis.Goals             Home Home.Goals OT
## 1 2014-10-08 Philadelphia Flyers         1     Boston Bruins          2 No
## 2 2014-10-08   Vancouver Canucks         4    Calgary Flames          2 No
## 3 2014-10-08     San Jose Sharks         4 Los Angeles Kings          0 No
##   Total.Goals
## 1           3
## 2           6
## 3           4
```

Data manipulation

Additional notes

- `lubridate` package for dealing with dates
- Additional webscraping via `rvest` package
- `googlesheets` package for direct editing
- Advantages to scraping