More Web Scraping

Stat 220

Bastola

February 16 2022

Get Links

```r
bow(url = "https://www.imdb.com/search/title/?groups=best_
  scrape() %>%
  html_elements(css = ".lister-item-header a") %>%
  html_attr(name = "href") %>%
  url_absolute(base = "https://www.imdb.com")
```

```
 [1] "https://www.imdb.com/title/tt9770150/?ref_=adv_li_tt
 [2] "https://www.imdb.com/title/tt6751668/?ref_=adv_li_tt
 [3] "https://www.imdb.com/title/tt6966692/?ref_=adv_li_tt
 [4] "https://www.imdb.com/title/tt5580390/?ref_=adv_li_tt
 [5] "https://www.imdb.com/title/tt4975722/?ref_=adv_li_tt
 [6] "https://www.imdb.com/title/tt1895587/?ref_=adv_li_tt
 [7] "https://www.imdb.com/title/tt2562232/?ref_=adv_li_tt
 [8] "https://www.imdb.com/title/tt2024544/?ref_=adv_li_tt
 [9] "https://www.imdb.com/title/tt1024648/?ref_=adv_li_tt
[10] "https://www.imdb.com/title/tt1655442/?ref_=adv_li_tt
[11] "https://www.imdb.com/title/tt1504320/?ref_=adv_li_tt
[12] "https://www.imdb.com/title/tt1010048/?ref_=adv_li_tt
[13] "https://www.imdb.com/title/tt0887912/?ref_=adv_li_tt
[14] "https://www.imdb.com/title/tt0477348/?ref_=adv_li_tt
[15] "https://www.imdb.com/title/tt0407887/?ref_=adv_li_tt
[16] "https://www.imdb.com/title/tt0405159/?ref_=adv_li_tt
[17] "https://www.imdb.com/title/tt0375679/?ref_=adv_li_tt
[18] "https://www.imdb.com/title/tt0167260/?ref_=adv_li_tt
[19] "https://www.imdb.com/title/tt0299658/?ref_=adv_li_tt
[20] "https://www.imdb.com/title/tt0268978/?ref_=adv_li_tt
[21] "https://www.imdb.com/title/tt0172495/?ref_=adv_li_tt
[22] "https://www.imdb.com/title/tt0169547/?ref_=adv_li_tt
[23] "https://www.imdb.com/title/tt0138097/?ref_=adv_li_tt
[24] "https://www.imdb.com/title/tt0120338/?ref_=adv_li_tt
[25] "https://www.imdb.com/title/tt0116209/?ref_=adv_li_tt
[26] "https://www.imdb.com/title/tt0112573/?ref_=adv_li_tt
[27] "https://www.imdb.com/title/tt0109830/?ref_=adv_li_tt
[28] "https://www.imdb.com/title/tt0108052/?ref_=adv_li_tt
[29] "https://www.imdb.com/title/tt0105695/?ref_=adv_li_tt
[30] "https://www.imdb.com/title/tt0102926/?ref_=adv_li_tt
[31] "https://www.imdb.com/title/tt0099348/?ref_=adv_li_tt
[32] "https://www.imdb.com/title/tt0097239/?ref_=adv_li_tt
[33] "https://www.imdb.com/title/tt0095953/?ref_=adv_li_tt
[34] "https://www.imdb.com/title/tt0093389/?ref_=adv_li_tt
```

Scrape Table

```r
table_usafacts <- bow(url = "https://usafacts.org/visualizations/covid-vaccine-tracker-states/state
  scrape() %>% html_elements(css = "table") %>% html_table()
```

```r
knitr::kable(table_usafacts[[3]], format = "html")
```

State	% of population with at least one dose	% fully vaccinated	% with booster or additional dose
Alabama	61.5%	49.8%	16.7%
Alaska	68.1%	59.9%	24.2%
Arizona	70.8%	59.4%	22.8%
Arkansas	65.3%	53%	19.6%
California	81.1%	69.6%	32.5%
Colorado	77.8%	68.7%	33.6%
Connecticut	93.2%	77.2%	37.1%
Delaware	81%	66.8%	28.6%

Click here to take a look at the webpage

Your Turn 1

05:00

Please clone the repository on advanced web scraping and visualization to your local folder. Go to this webpage and scrape the table that has the latest county-level coronavirus stats for the state of Minnesota.

County	7-day avg. cases	7-day avg. deaths	Cases	Deaths
Aitkin County	7	0	2,836	59
Anoka County	244	2	96,275	757
Becker County	34	0	8,386	86
Beltrami County	35	0	11,038	118
Benton County	53	0	13,555	164
Big Stone County	4	0	1,335	8
Blue Earth County	60	0	17,117	89
Brown County	16	0	6,391	72
Carlton County	57	0	8,500	88
Carver County	71	1	25,772	107

What are the top 10 counties in Minnesota by the number of COVID cases?

Scraping multiple tables

```
all_url <- "https://finance.yahoo.com/losers?count=25&offset="
```

```
idx <- seq(0, 250, by = 25)

table_new <-data.frame()
df <- data.frame()

for (i in seq_along(idx)) {
  new_webpage <- read_html(str_glue(all_url, {idx[i]}))
  table_new <- html_table(new_webpage)[[1]] %>%
    as_tibble(.name_repair = "unique")
  df <- rbind(df, table_new)
}
```

Click here to take a look at the webpage

Scraping multiple tables

```
all_url <- "https://finance.yahoo.com/losers?count=25&offset="
```

```
idx <- seq(0, 250, by = 25)

table_new <-data.frame()
df <- data.frame()

for (i in seq_along(idx)) {
  new_webpage <- read_html(str_glue(all_url, {idx[i]}))
  table_new <- html_table(new_webpage)[[1]] %>%
    as_tibble(.name_repair = "unique")
  df <- rbind(df, table_new)
}
```

Click here to take a look at the webpage

Scraping multiple tables

```
all_url <- "https://finance.yahoo.com/losers?count=25&offset="
```

```
idx <- seq(0, 250, by = 25)

table_new <-data.frame()
df <- data.frame()

for (i in seq_along(idx)) {
  new_webpage <- read_html(str_glue(all_url, {idx[i]}))
  table_new <- html_table(new_webpage)[[1]] %>%
    as_tibble(.name_repair = "unique")
  df <- rbind(df, table_new)
}
```

Click here to take a look at the webpage

Scraping multiple tables

```r
all_url <- "https://finance.yahoo.com/losers?count=25&offset="
```

```r
idx <- seq(0, 250, by = 25)

table_new <-data.frame()
df <- data.frame()

for (i in seq_along(idx)) {
  new_webpage <- read_html(str_glue(all_url, {idx[i]}))
  table_new <- html_table(new_webpage)[[1]] %>%
    as_tibble(.name_repair = "unique")
  df <- rbind(df, table_new)
}
```

Click here to take a look at the webpage

Scraping multiple tables

```
all_url <- "https://finance.yahoo.com/losers?count=25&offset="
```

```
idx <- seq(0, 250, by = 25)

table_new <-data.frame()
df <- data.frame()

for (i in seq_along(idx)) {
  new_webpage <- read_html(str_glue(all_url, {idx[i]}))
  table_new <- html_table(new_webpage)[[1]] %>%
    as_tibble(.name_repair = "unique")
  df <- rbind(df, table_new)
}
```

Click here to take a look at the webpage

Scraping multiple tables

```r
all_url <- "https://finance.yahoo.com/losers?count=25&offset="
```

```r
idx <- seq(0, 250, by = 25)

table_new <-data.frame()
df <- data.frame()

for (i in seq_along(idx)) {
  new_webpage <- read_html(str_glue(all_url, {idx[i]}))
  table_new <- html_table(new_webpage)[[1]] %>%
    as_tibble(.name_repair = "unique")
  df <- rbind(df, table_new)
}
```

Click here to take a look at the webpage

Scraping multiple tables

```
all_url <- "https://finance.yahoo.com/losers?count=25&offset="
```

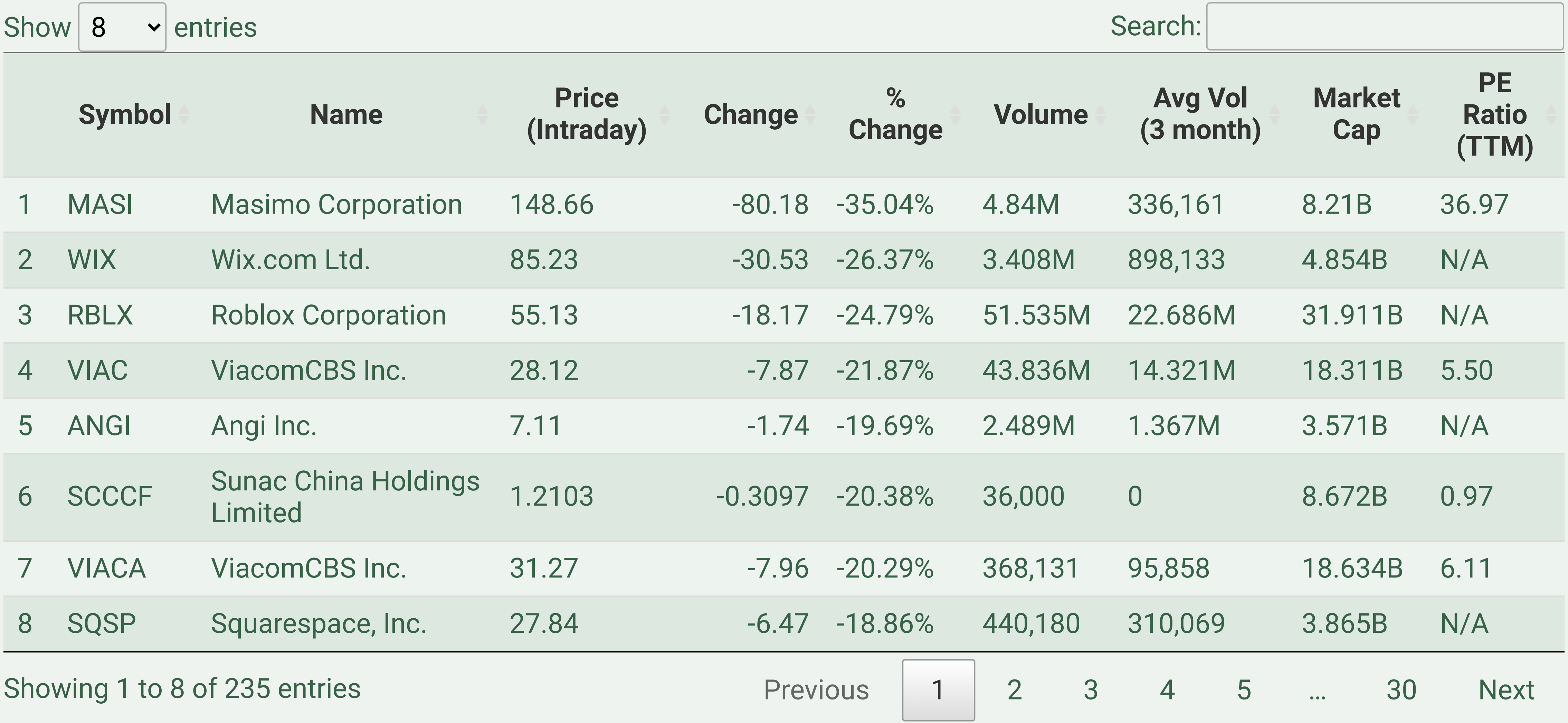
```
idx <- seq(0, 250, by = 25)

table_new <-data.frame()
df <- data.frame()

for (i in seq_along(idx)) {
  new_webpage <- read_html(str_glue(all_url, {idx[i]}))
  table_new <- html_table(new_webpage)[[1]] %>% as_tibble(.name_repair = "unique")
  df <- rbind(df, table_new)
}
```

Click here to take a look at the webpage

Multiple tables combined



Show 8 entries

Search:

	Symbol	Name	Price (Intraday)	Change	% Change	Volume	Avg Vol (3 month)	Market Cap	PE Ratio (TTM)
1	MASI	Masimo Corporation	148.66	-80.18	-35.04%	4.84M	336,161	8.21B	36.97
2	WIX	Wix.com Ltd.	85.23	-30.53	-26.37%	3.408M	898,133	4.854B	N/A
3	RBLX	Roblox Corporation	55.13	-18.17	-24.79%	51.535M	22.686M	31.911B	N/A
4	VIAC	ViacomCBS Inc.	28.12	-7.87	-21.87%	43.836M	14.321M	18.311B	5.50
5	ANGI	Angi Inc.	7.11	-1.74	-19.69%	2.489M	1.367M	3.571B	N/A
6	SCCCF	Sunac China Holdings Limited	1.2103	-0.3097	-20.38%	36,000	0	8.672B	0.97
7	VIACA	ViacomCBS Inc.	31.27	-7.96	-20.29%	368,131	95,858	18.634B	6.11
8	SQSP	Squarespace, Inc.	27.84	-6.47	-18.86%	440,180	310,069	3.865B	N/A

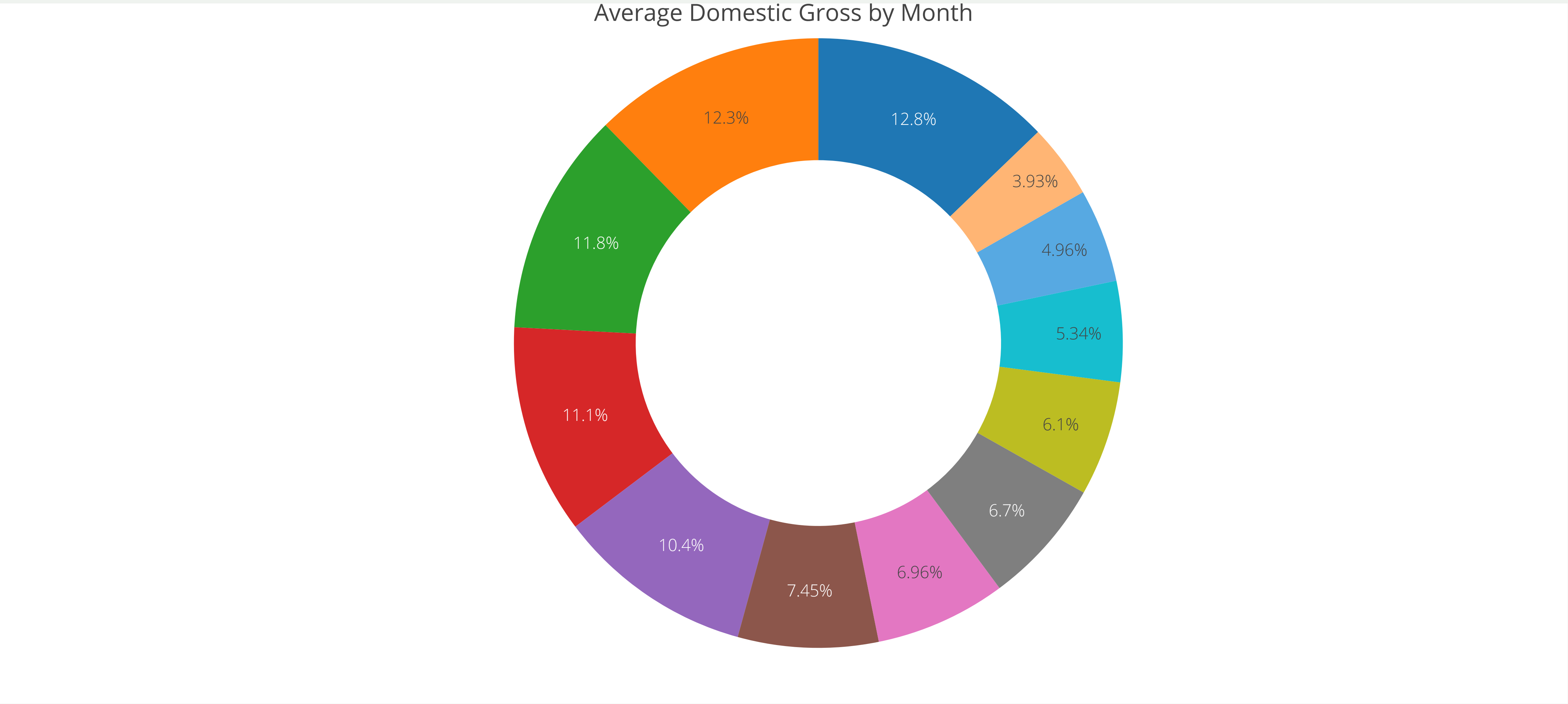
Showing 1 to 8 of 235 entries

Tidy further

```
df_movies %>%
  rename(ID = `...1`) %>%
  mutate(ProductionBudget = parse_number(ProductionBudget))
  mutate(DomesticGross = parse_number(DomesticGross))  %>%
  mutate(WorldwideGross = parse_number(WorldwideGross)) %>%
  mutate(ReleaseDate = mdy(ReleaseDate)) %>%
  mutate(ReleaseDate = replace_na(ReleaseDate, make_date())
  mutate(MonthOfRelease = month(ReleaseDate, label = TRUE))
  mutate(YearOfRelease = year(ReleaseDate)) %>%
  select(MonthOfRelease, DomesticGross) %>%
  group_by(MonthOfRelease) %>%
  summarize(AverageByMonth = mean(DomesticGross))
```

```
# A tibble: 12 × 2
   MonthOfRelease AverageByMonth
   <ord>                   <dbl>
 1 Jan                 19586029.
 2 Feb                 34671367.
 3 Mar                 37145452.
 4 Apr                 33383875.
 5 May                 61400715.
 6 Jun                 63916536.
 7 Jul                 55392411.
 8 Aug                 30375324.
 9 Sep                 24712010.
10 Oct                 26629138.
11 Nov                 52033198.
12 Dec                 59045166.
```

Interactive Donut Plot



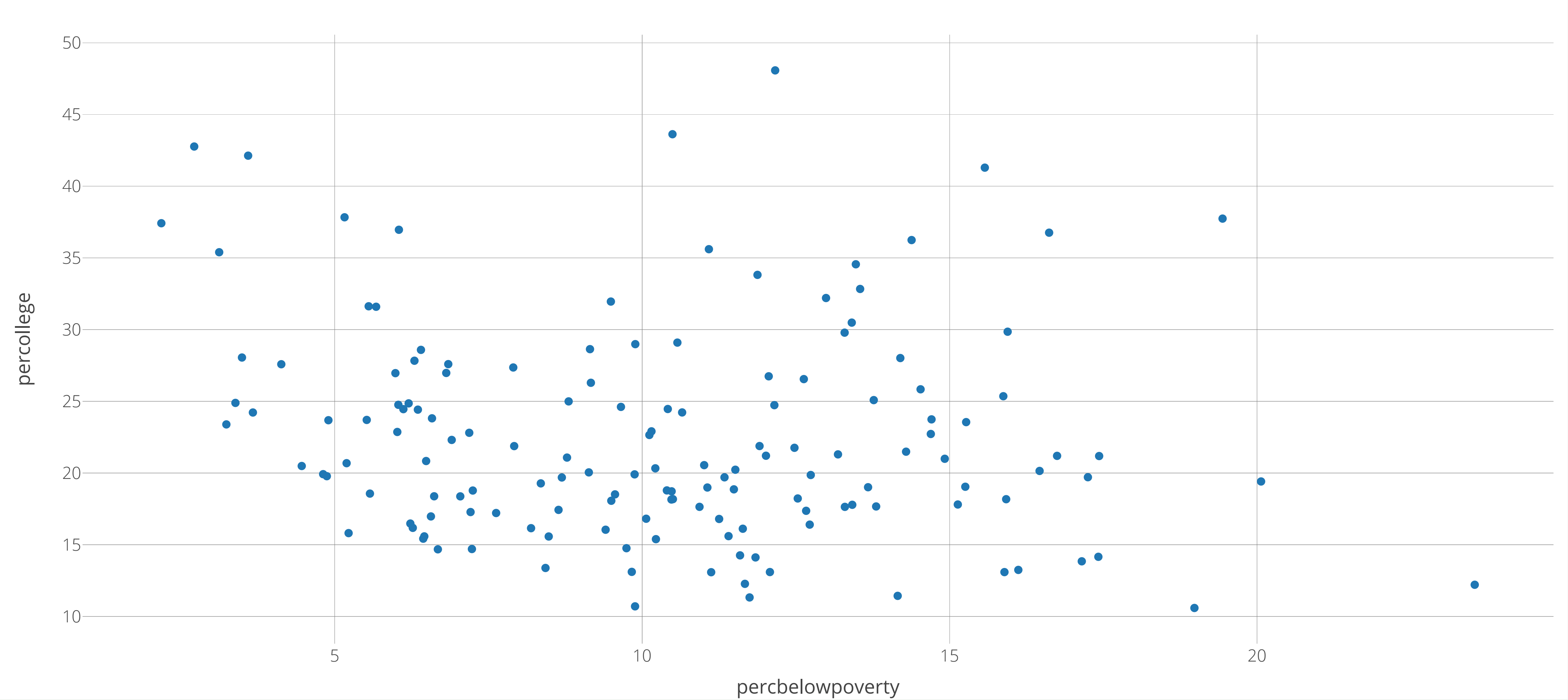
Interactive visualizations using Plotly

```
midwest %>% as_tibble()
# A tibble: 437 × 28
     PID county   state  area poptotal popdensity popwhite popblack popamerindian
   <int> <chr>    <chr> <dbl>    <int>      <dbl>    <int>    <int>         <int>
 1   561 ADAMS    IL    0.052    66090      1271.    63917     1702            98
 2   562 ALEXAN… IL    0.014    10626       759      7054     3496            19
 3   563 BOND     IL    0.022    14991       681.    14477      429            35
 4   564 BOONE    IL    0.017    30806      1812.    29344      127            46
 5   565 BROWN    IL    0.018     5836       324.     5264      547            14
 6   566 BUREAU   IL    0.05     35688       714.    35157       50            65
 7   567 CALHOUN  IL    0.017     5322       313.     5298        1             8
 8   568 CARROLL  IL    0.027    16805       622.    16519      111            30
 9   569 CASS     IL    0.024    13437       560.    13384       16             8
10   570 CHAMPA… IL    0.058   173025      2983.   146506    16559           331
```

```
library(plotly)

midwest %>%
  filter(inmetro == T) %>%
  plot_ly(x =  ~ percbelowpoverty, y =  ~ percollege) %>%
  add_markers()
```

Interactive visualizations using `Plotly`



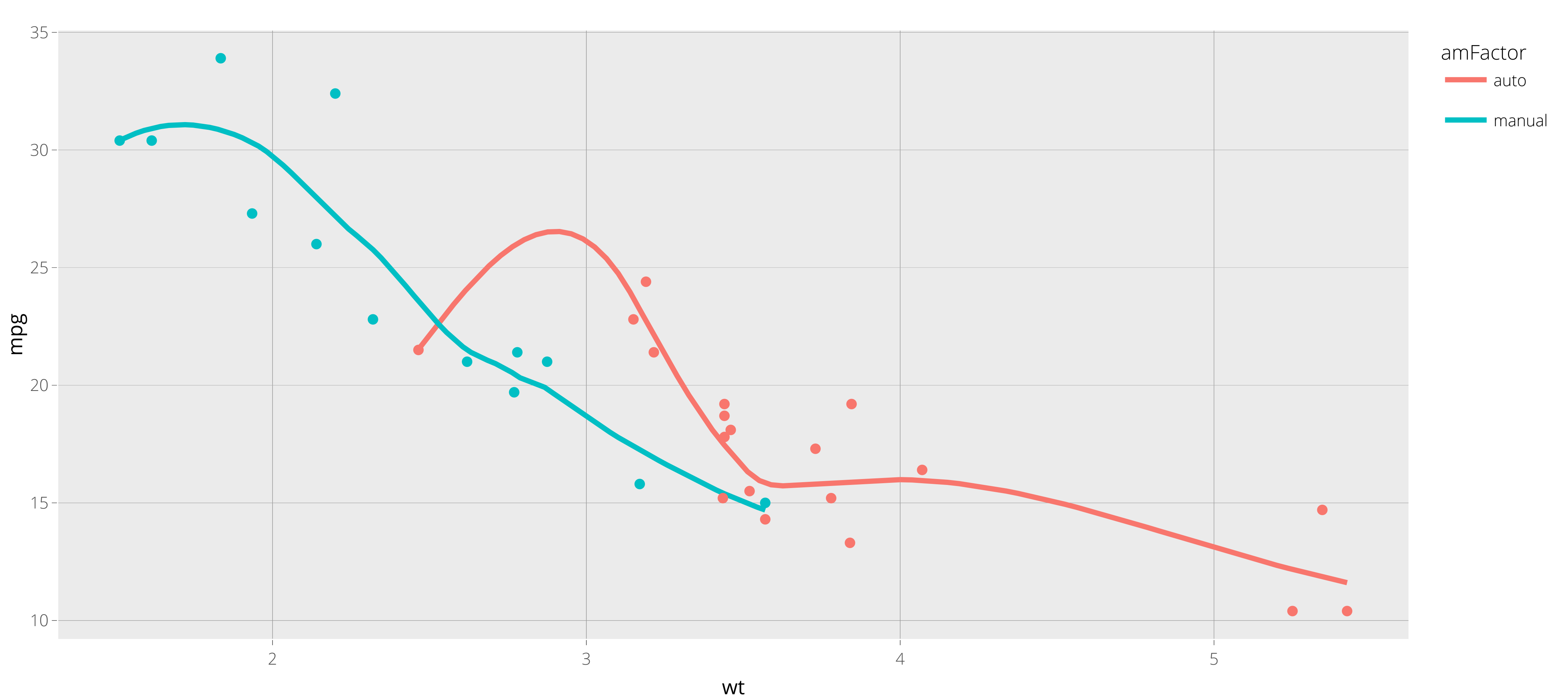
Interactive visualizations using ggplotly

```
mtcars %>% as_tibble() %>% head()
# A tibble: 6 × 11
    mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  21       6   160   110  3.9   2.62  16.5     0     1     4     4
2  21       6   160   110  3.9   2.88  17.0     0     1     4     4
3  22.8     4   108    93  3.85  2.32  18.6     1     1     4     1
4  21.4     6   258   110  3.08  3.22  19.4     1     0     3     1
5  18.7     8   360   175  3.15  3.44  17.0     0     0     3     2
6  18.1     6   225   105  2.76  3.46  20.2     1     0     3     1
```

```
gp = mtcars %>%
  mutate(amFactor = factor(am, labels = c('auto', 'manual')),
         hovertext = paste(wt, mpg, amFactor)) %>%
  arrange(wt) %>%
  ggplot(aes(x = wt, y = mpg, color = amFactor)) +
  geom_smooth(se = F) +
  geom_point(aes(color = amFactor))
```

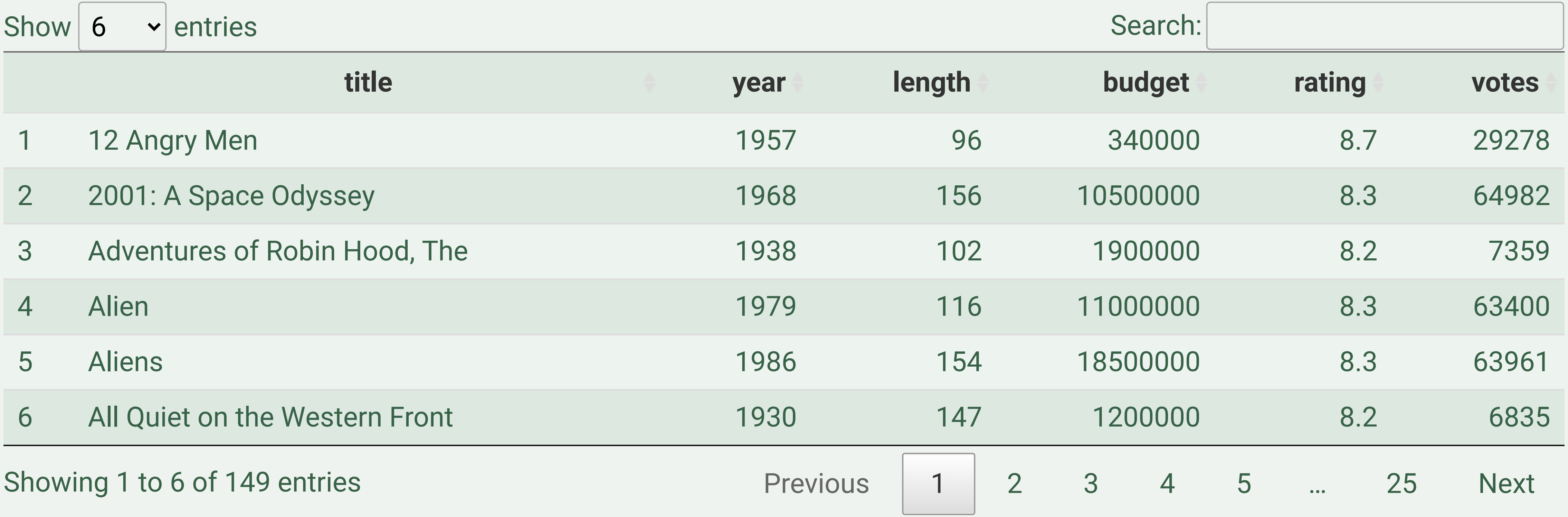
```
ggplotly()
```

Interactive visualizations using `ggplotly`



DT: Interactive Data Tables

```
library(ggplot2movies)
movies %>%
  select(1:6) %>%
  filter(rating > 8, !is.na(budget), votes > 1000) %>%
  datatable(fillContainer = FALSE, options = list(pageLength = 6))
```

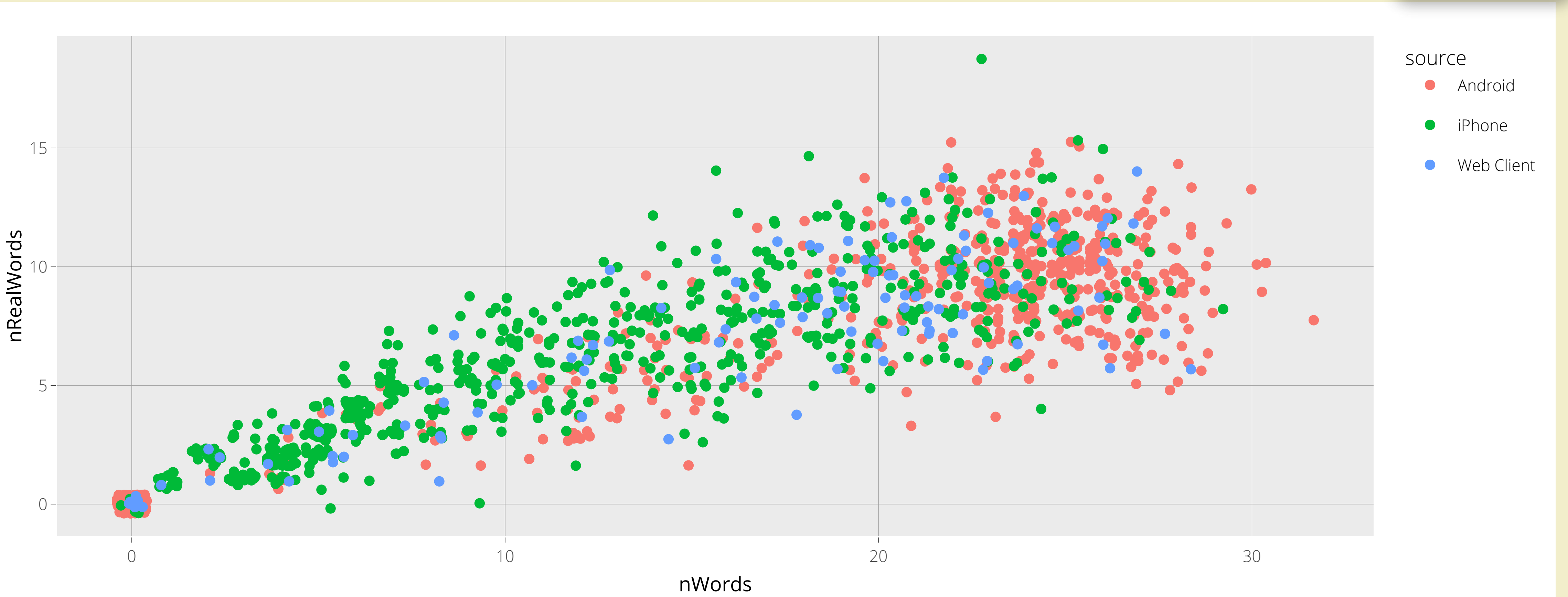


	title	year	length	budget	rating	votes
1	12 Angry Men	1957	96	340000	8.7	29278
2	2001: A Space Odyssey	1968	156	10500000	8.3	64982
3	Adventures of Robin Hood, The	1938	102	1900000	8.2	7359
4	Alien	1979	116	11000000	8.3	63400
5	Aliens	1986	154	18500000	8.3	63961
6	All Quiet on the Western Front	1930	147	1200000	8.2	6835

Showing 1 to 6 of 149 entries

Your Turn 3

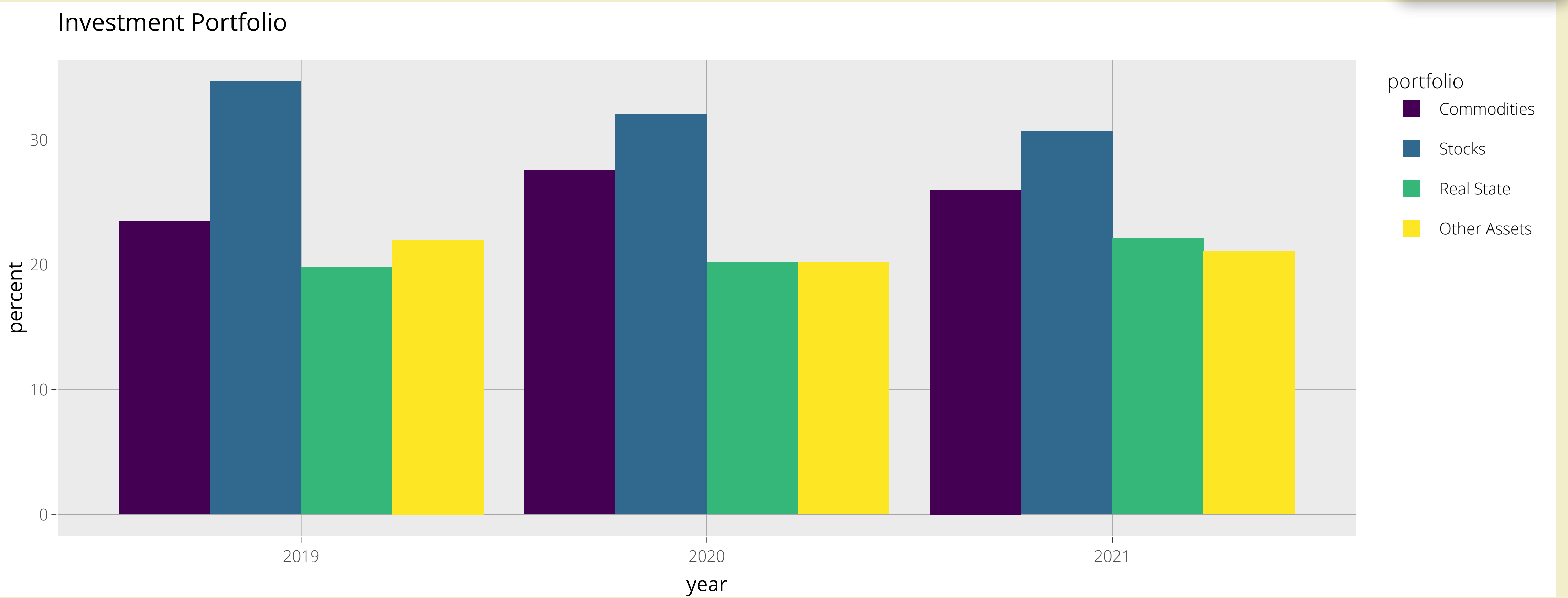
05:00



Reproduce the plot using `plotly`

Your Turn 4

05:00



Reproduce the plot using `plotly`