

Great Datasets: A Story


Alex Cookson

 [@alexcookson](https://twitter.com/alexcookson)

 [tacookson](https://github.com/tacookson)

Bored? Busy?


- Watch the 5-minute version of this talk



DOWNLOADSUPPORTDOCSCOMMUNITY

Products Solutions CustomersResourcesAboutPricing

WebinarsRStudio EssentialsData Science EssentialsAdvanced Data ScienceWorking with SparkRStudio Pro AdministrationMateriales en EspañolAdditional TalksConferences by year



RSTUDIO::GLOBAL 2021PACKAGE DEV

The Power of Great Datasets

How do love stories usually go?

Non-binary person
meets Girl

Non-binary person
meets Boy

Girl meets Boy

Girl meets Girl

Boy meets Girl

Boy meets Boy

Non-binary person meets
Non-binary person

Non-binary person
meets Girl

Non-binary person
meets Boy

Girl meets Boy

Girl meets Girl

Boy meets Girl

Boy meets Boy

Non-binary person meets
Non-binary person

Learner
meets Data

Ever seen a romantic comedy?

The Meeting



Sparks Fly



Tension / Crisis



Falling In Love



The Meeting: *Hamilton*



- *Hamilton* tickets for Christmas!
- My brother the Broadway fan
- Weekly grosses: tickets sales, ticket prices, percent of seats sold – all the way back to 1985

Sparks Fly: Weekly Grosses on Playbill



PLAYBILL									
BROADWAY GROSSES									
All data provided by The Broadway League									
BROADWAY GROSSES WEEK ENDING									
10/16/16									
RANK	THEATRE	WEEKEND GROSS	WEEKS	GROSS TO DATE	PERFORMANCES	WEEKLY AVERAGE	WEEKLY PERFORMANCES	WEEKLY AVERAGE	WEEKLY PERFORMANCES
1	THEATRE	\$1,000,000	1	\$1,000,000	1	\$1,000,000	1	\$1,000,000	1
2	THEATRE	\$800,000	1	\$800,000	1	\$800,000	1	\$800,000	1
3	THEATRE	\$600,000	1	\$600,000	1	\$600,000	1	\$600,000	1
4	THEATRE	\$400,000	1	\$400,000	1	\$400,000	1	\$400,000	1
5	THEATRE	\$200,000	1	\$200,000	1	\$200,000	1	\$200,000	1
6	THEATRE	\$100,000	1	\$100,000	1	\$100,000	1	\$100,000	1
7	THEATRE	\$50,000	1	\$50,000	1	\$50,000	1	\$50,000	1
8	THEATRE	\$25,000	1	\$25,000	1	\$25,000	1	\$25,000	1
9	THEATRE	\$12,500	1	\$12,500	1	\$12,500	1	\$12,500	1
10	THEATRE	\$6,250	1	\$6,250	1	\$6,250	1	\$6,250	1

BROADWAY GROSSES

All data provided by The Broadway League

BROADWAY GROSSES WEEK ENDING

WEEK'S TOTAL
\$26,700,955.53

2020-03-08

SHOW	THIS WEEK GROSS POTENTIAL GROSS	DIFF \$	AVG TICKET TOP TICKET ▲	SEATS SOLD SEATS IN THEATRE	PERFS PREVIEWS	% CAP	DIFF % CAP
HAMILTON RICHARD RODGERS THEATRE	\$2,688,721.00 \$2,605,608.00	-\$7,468.00	\$250.07 \$847.00	10752 1324	8 0	101.51%	-0.04%
THE LEHMAN TRILOGY NEDERLANDER THEATRE	\$188,126.00 \$171,298.00	\$0.00	\$163.02 \$397.00	1154 1154	0 1	100.00%	0.00%
MOULIN ROUGE! THE MUSICAL! AL HIRSCHFELD THEATRE	\$1,514,716.50 \$1,754,340.00	-\$56,397.25	\$147.92 \$399.00	10240 1302	8 0	98.31%	-1.69%
HADESTOWN WALTER KERR THEATRE	\$1,086,477.75 \$1,094,012.00	\$69,803.00	\$147.10 \$0.00	7386 918	8 0	100.57%	-0.04%
TINA: THE TINA TURNER MUSICAL LUNT-FONTANNE THEATRE	\$1,225,999.00 \$1,566,688.00	-\$94,767.00	\$129.68 \$297.00	9454 1478	8 0	79.96%	-4.65%
DEAR EVAN HANSEN MUSIC BOX THEATRE	\$981,271.30 \$1,234,034.00	\$15,538.20	\$126.62 \$297.00	7750 984	8 0	98.45%	1.42%
COMPANY BERNARD B. JACOBS THEATRE	\$779,588.10 \$875,460.00	\$0.00	\$124.16 \$327.00	6279 1030	0 6	101.60%	0.00%
TO KILL A MOCKINGBIRD SAM S. SHUBERT THEATRE	\$1,237,497.00 \$1,751,250.00	\$105,218.46	\$118.06 \$423.00	10482 1435	8 0	91.31%	-6.36%
THE BOOK OF MORMON EUGENE O'NEILL THEATRE	\$929,168.30 \$1,190,502.00	\$32,168.80	\$112.67 \$477.50	8247 1047	8 0	98.46%	3.43%
SIX: THE MUSICAL BROOKS ATKINSON THEATRE	\$884,878.20 \$1,057,206.00	-\$12,006.80	\$108.79 \$297.00	8134 1031	0 8	98.62%	0.83%

BROADWAY GROSSES

All data provided by The Broadway League

BROADWAY GROSSES WEEK ENDING

2020-03-08

WEEK'S TOTAL
\$26,700,955.53

SHOW	THIS WEEK GROSS POTENTIAL GROSS	DIFF \$	AVG TICKET TOP TICKET ▲	SEATS SOLD SEATS IN THEATRE	PERFS PREVIEWS	% CAP	DIFF % CAP
HAMILTON RICHARD RODGERS THEATRE	\$2,688,721.00 \$2,605,608.00	-\$7,468.00	\$250.07 \$847.00	10752 1324	8 0	101.51%	-0.04%
THE LEHMAN TRILOGY NEDERLANDER THEATRE	\$188,126.00 \$171,298.00	\$0.00	\$163.02 \$397.00	1154 1154	0 1	100.00%	0.00%
MOULIN ROUGE! THE MUSICAL! AL HIRSCHFELD THEATRE	\$1,514,716.50 \$1,754,340.00	-\$56,397.25	\$147.92 \$399.00	10240 1302	8 0	98.31%	-1.69%
HADESTOWN WALTER KERR THEATRE	\$1,082,477.75 \$1,094,012.00	\$69,803.88	\$147.10 \$0.00	7386 918	8 0	100.57%	-0.04%
TIM & ERIC'S BACKSTAGE MUSICAL THE MONTAGNE THEATRE	\$1,259,981.00 \$1,259,981.00	-\$24,767.00	\$129.68 \$297.00	9454 1478	8 0	79.96%	-4.65%
DEAR EVAN HANSEN MUSIC BOX THEATRE	\$981,271.30 \$1,234,034.00	\$15,538.20	\$126.62 \$297.00	7750 984	8 0	98.45%	1.42%
COMPANY BERNARD B. JACOBS THEATRE	\$779,588.10 \$875,460.00	\$0.00	\$124.16 \$327.00	6279 1030	0 6	101.60%	0.00%
TO KILL A MOCKINGBIRD SAM S. SHUBERT THEATRE	\$1,237,497.00 \$1,751,250.00	\$105,218.46	\$118.06 \$423.00	10482 1435	8 0	91.31%	-6.36%
THE BOOK OF MORMON EUGENE O'NEILL THEATRE	\$929,168.30 \$1,190,502.00	\$32,168.80	\$112.67 \$477.50	8247 1047	8 0	98.46%	3.43%
SIX: THE MUSICAL BROOKS ATKINSON THEATRE	\$884,878.20 \$1,057,206.00	-\$12,006.80	\$108.79 \$297.00	8134 1031	0 8	98.62%	0.83%

Are Hamilton tickets on
Broadway really that expensive?

BROADWAY GROSSES

All data provided by The Broadway League

BROADWAY GROSSES WEEK ENDING

WEEK'S TOTAL
\$26,700,955.53

2020-03-08

SHOW	THIS WEEK GROSS POTENTIAL GROSS	DIFF \$	AVG TICKET TOP TICKET ▲	SEATS SOLD SEATS IN THEATRE	PERFS PREVIEWS	% CAP	DIFF % CAP
HAMILTON RICHARD RODGERS THEATRE	\$2,688,721.00 \$2,605,608.00	-\$7,468.00	\$250.07 \$847.00	10752 1324	8 0	101.51%	-0.04%
THE LEHMAN TRILOGY NEDERLANDER THEATRE	\$188,126.00 \$171,298.00	\$0.00	\$163.02 \$397.00	1154 1154	0 1	100.00%	0.00%
MOULIN ROUGE! THE MUSICAL! AL HIRSCHFELD THEATRE	\$1,514,716.50 \$1,754,340.00	-\$56,397.25	\$147.92 \$399.00	10240 1302	8 0	99.27%	-0.69%
HADESTOWN WALTER KERR THEATRE	\$1,086,477.75 \$1,094,012.00	\$69,803.00	\$147.10 \$0.00	7386 918	8 0	100.00%	-0.04%
TINA: THE TINA TURNER MUSICAL LUNT-FONTANNE THEATRE	\$1,225,999.00 \$1,566,688.00	-\$94,767.00	\$129.68 \$297.00	9454 1478	8 0	79.96%	-4.65%
DEAR EVAN HANSEN MUSIC BOX THEATRE	\$981,271.30 \$1,234,034.00	\$15,538.20	\$126.62 \$297.00	7750 984	8 0	98.45%	1.42%
COMPANY BERNARD B. JACOBS THEATRE	\$779,588.10 \$875,460.00	\$0.00	\$124.16 \$327.00	6279 1030	0 6	101.60%	0.00%
TO KILL A MOCKINGBIRD SAM S. SHUBERT THEATRE	\$1,237,497.00 \$1,751,250.00	\$105,218.46	\$118.06 \$423.00	10482 1435	8 0	91.31%	-6.36%
THE BOOK OF MORMON EUGENE O'NEILL THEATRE	\$929,168.30 \$1,190,502.00	\$32,168.80	\$112.67 \$477.50	8247 1047	8 0	98.46%	3.43%
SIX: THE MUSICAL BROOKS ATKINSON THEATRE	\$884,878.20 \$1,057,206.00	-\$12,006.80	\$108.79 \$297.00	8134 1031	0 8	98.62%	0.83%

Yes! \$250 (USD) a pop!
(\$850 if you're fancy)

BROADWAY GROSSES

All data provided by The Broadway League

BROADWAY GROSSES WEEK ENDING

1988-11-20

WEEK
25

WEEK'S TOTAL
\$5,093,602.00

SHOW	THIS WEEK GROSS POTENTIAL GROSS ▲	DIFF \$	AVG TICKET TOP TICKET	SEATS SOLD SEATS IN THEATRE	PERFS PREVIEWS	% CAP	DIFF % CAP
THE PHANTOM OF THE OPERA MAJESTIC THEATRE	\$536,239.00 \$0.00	-\$2,833.00	\$40.95 \$0.00	13096 1609	8 0	101.74%	0.00%
LES MISÉRABLES IMPERIAL THEATRE	\$497,755.00 \$0.00	-\$2,873.00	\$40.59 \$0.00	12264 1752	7 0	100.00%	0.00%
CATS WINTER GARDEN THEATRE	\$447,568.00 \$0.00	-\$5,715.00	\$39.16 \$0.00	11430 1482	8 0	96.41%	0.15%
LEGS DIAMOND MARK HELLINGER THEATRE	\$428,961.00 \$0.00	-\$18,561.00	\$42.94 \$0.00	9989 1603	0 8	77.89%	9.62%
ME AND MY GIRL MOROSOFF THEATRE	\$425,317.00 \$0.00	-\$34,279.00	\$38.27 \$0.00	11113 1570	8 0	88.48%	0.94%
ANYTHING GOES VIVIAN BEAUMONT THEATRE	\$333,373.00 \$0.00	\$13,843.00	\$40.11 \$0.00	8311 1050	8 0	98.94%	0.44%
INTO THE WOODS MARTIN BECK THEATRE	\$302,008.00 \$0.00	-\$31,319.00	\$35.76 \$0.00	8445 1282	8 0	82.34%	-7.89%
M. BUTTERFLY EUGENE O'NEILL THEATRE	\$289,186.00 \$0.00	\$6,765.00	\$33.81 \$0.00	8553 1059	8 0	100.96%	1.11%
STARLIGHT EXPRESS GERSHWIN THEATRE	\$261,729.00 \$0.00	-\$57,498.00	\$34.08 \$0.00	7680 1803	8 0	53.24%	-9.76%
AIN'T MISBEHAVIN' AMBASSADOR THEATRE	\$221,248.00 \$0.00	-\$31,191.00	\$33.75 \$0.00	6556 1108	8 0	73.96%	-8.31%

What was popular the week I was born?

BROADWAY GROSSES

All data provided by The Broadway League

BROADWAY GROSSES WEEK ENDING

1988-11-20

WEEK
25

WEEK'S TOTAL
\$5,093,602.00

SHOW	THIS WEEK GROSS POTENTIAL GROSS ▲	DIFF \$	AVG TICKET TOP TICKET	SEATS SOLD SEATS IN THEATRE	PERFS PREVIEWS	% CAP	DIFF % CAP
THE PHANTOM OF THE OPERA MAJESTIC THEATRE	\$536,239.00 \$0.00	-\$2,833.00	\$40.95 \$0.00	13096 1609	8 0	101.74%	0.00%
LES MISÉRABLES IMPERIAL THEATRE	\$497,755.00 \$0.00	-\$2,873.00	\$40.59 \$0.00	12264 1752	7	100.00%	0.00%
CATS WINTER GARDEN THEATRE	\$447,568.00 \$0.00	-\$5,715.00	\$39.16 \$0.00	11430 1482	8 0	90.41%	-0.15%
LEGS DIAMOND MARK HELLINGER THEATRE	\$428,961.00 \$0.00	-\$18,561.00	\$42.94 \$0.00	9989 1603	0 8	77.02%	-9.62%
ME AND MY GIRL MARQUIS THEATRE	\$425,317.00 \$0.00	-\$34,279.00	\$38.27 \$0.00	11113 1570	8 0	88.48%	0.94%
ANYTHING GOES VIVIAN BEAUMONT THEATER	\$333,373.00 \$0.00	\$13,843.00	\$40.11 \$0.00	8311 1050	8 0	98.94%	0.44%
INTO THE WOODS MARTIN BECK THEATRE	\$302,008.00 \$0.00	-\$31,319.00	\$35.76 \$0.00	8445 1282	8 0	82.34%	-7.89%
M. BUTTERFLY EUGENE O'NEILL THEATRE	\$289,186.00 \$0.00	\$6,765.00	\$33.81 \$0.00	8553 1059	8 0	100.96%	1.11%
STARLIGHT EXPRESS GERSHWIN THEATRE	\$261,729.00 \$0.00	-\$57,498.00	\$34.08 \$0.00	7680 1803	8 0	53.24%	-9.76%
AIN'T MISBEHAVIN' AMBASSADOR THEATRE	\$221,248.00 \$0.00	-\$31,191.00	\$33.75 \$0.00	6556 1108	8 0	73.96%	-8.31%

Phantom, Les Mis,
and Cats!

BROADWAY GROSSES

All data provided by The Broadway League

BROADWAY GROSSES WEEK ENDING

1988-11-20

Is there any seasonality to ticket prices?

Is Hamilton's success unprecedented?

What's the highest-grossing show EVER?

Do revivals perform as well as their originals?

Have average ticket prices gotten more expensive?

SHOW	THIS WEEK GROSS POTENTIAL GROSS	DIFF \$	AVG TICKET TOP TICKET	SEATS SOLD SEATS IN THEATRE	PERFS PREVIEWS	% CAP	DIFF % CAP
THE PHOENIX OF THE OPERA MAJESTIC THEATRE	\$536,239.00 \$0.00	\$2,833.00 \$0.00	\$40.11 \$0.00	11011 1200	8 0	91.74%	0.00%
LES MISÉRABLES IMPERIAL THEATRE	\$497,755.00 \$0.00	\$7,755.00 \$0.00	\$40.58 \$0.00	12264 1200	7 0	100.00%	0.00%
CATS WINTER GARDEN THEATRE	\$447,568.00 \$0.00	-\$5,715.00 \$0.00	\$39.16 \$0.00	11430 1482	8 0	91.12%	-0.12%
LEGS DIAMOND MARK HELLINGER THEATRE	\$428,961.00 \$0.00	-\$18,561.00 \$0.00	\$42.94 \$0.00	9989 1603	0 8	77.89%	9.62%
MY GIRL MARQUIS THEATRE	\$425,317.00 \$0.00	-\$34,279.00 \$0.00	\$35.76 \$0.00	11113 1270	8 0	88.48%	0.94%
ANYTHING'S POSSIBLE VIVIAN BEAUMONT THEATRE	\$333,373.00 \$0.00	\$13,843.00 \$0.00	\$40.11 \$0.00	8311 1050	8 0	98.94%	0.44%
INTO THE WOODS MARTIN BECK THEATRE	\$302,008.00 \$0.00	-\$31,319.00 \$0.00	\$35.76 \$0.00	8415 1222	8 0	82.31%	-7.89%
M. BUTTERFLY EUGENE O'NEILL THEATRE	\$289,186.00 \$0.00	\$6,765.00 \$0.00	\$33.81 \$0.00	8553 1059	0 0	100.96%	1.11%
STARLIGHT EXPRESS GERSHWIN THEATRE	\$261,729.00 \$0.00	-\$57,498.00 \$0.00	\$34.08 \$0.00	7680 1803	8 0	53.24%	-1.11%
AIN'T MISBEHAVIN' AMBASSADOR THEATRE	\$221,248.00 \$0.00	-\$31,191.00 \$0.00	\$33.75 \$0.00	6556 1108	8 0	73.96%	-8.31%

BROADWAY GROSSES

All data provided by The Broadway League

BROADWAY GROSSES WEEK ENDING

1988-11-20

Is there any seasonality to ticket prices?

Is Hamilton's success unprecedented?

Do revivals perform as well as their originals?

What's the highest-grossing show EVER?

Have average ticket prices gotten more expensive?

SHOW	THIS WEEK'S GROSS	WEEKS IN PRODUCTION	AVERAGE TICKET PRICE	TOTAL TICKETS SOLD	PERFORMANCES	% CAPACITY	DIFF % CAP
THE PHOENIX OF THE OPERA MAJESTIC THEATRE	\$19,000	1	\$0.00	0	8	1.74%	0.00%
LES MISÉRABLES IMPERIAL THEATRE	\$1755.00	1	\$0.00	0	7	100.00%	0.00%
CATS WINTER GARDEN THEATRE	\$447,568.00	1	-\$5,715.00	\$39.16	11430	9.12%	0.12%
LEGS DIAMOND MARK HELINGER THEATRE	\$0.00	1	-\$18,000.00	\$42.94	9989	77.89%	9.62%
MY GIRL MARQUIS THEATRE	\$425,317.00	1	-\$34,279.00	\$39.16	11430	88.48%	0.94%
ANYTHING GOES VIVIAN BEAUMONT THEATRE	\$333,373.00	1	\$13,843.00	\$40.11	8311	98.94%	0.44%
INTO THE WOODS MARTIN BECK THEATRE	\$302,008.00	1	-\$31,319.00	\$35.76	8415	82.31%	-7.69%
M. BUTTERFLY EUGENE O'NEILL THEATRE	\$289,186.00	1	\$6,765.00	\$33.81	8553	100.96%	1.11%
STARLIGHT EXPRESS GERSHWIN THEATRE	\$261,729.00	1	-\$57,498.00	\$34.08	7680	53.24%	-46.76%
AIN'T MISBEHAVIN' AMBASSADOR THEATRE	\$221,248.00	1	-\$31,191.00	\$33.75	6556	73.96%	-8.31%

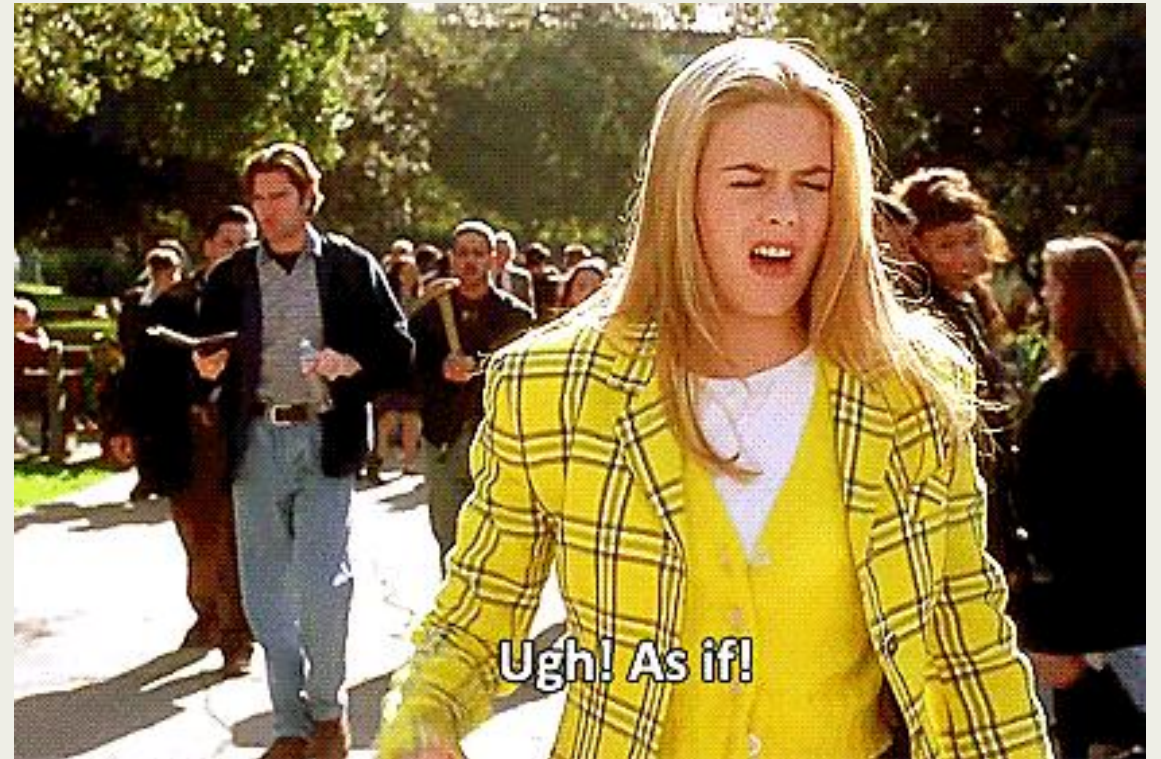
Tension / Crisis: Stuck!

- I didn't want data (one week at a time) – I wanted DATA (full dataset)
- The data was right there, but I couldn't get to it

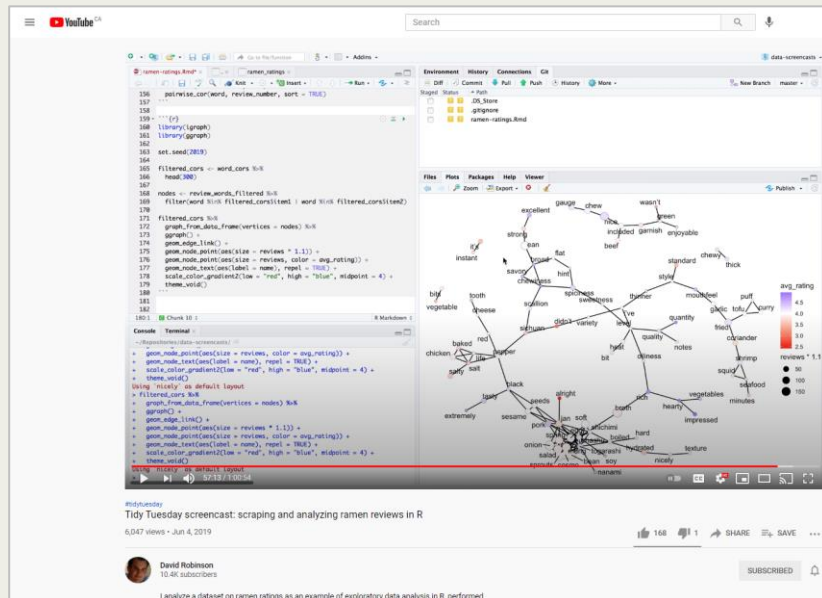


Tension / Crisis: Stuck!

Should I
learn how to
web scrape?



Tension / Crisis: Stuck!



David Robinson
Screencast



{rvest} package

```
1 # Load packages
2 library(tidyverse)
3 library(lubridate)
4 library(rvest)
5
6
7
8 # Weekly grosses
9 ## Create function to scrape grosses table
10 get_playbill_data = function(url) {
11   message(url)
12
13   website <- read_html(url)
14
15   show_stats <- list(
16     week_number = html_nodes(website, ".week-count .accent") %>% html_text(trim = TRUE),
17     weekly_gross_overall = html_nodes(website, ".week-total .accent") %>% html_text(trim = TRUE),
18     show = html_nodes(website, ".col-0 .data-value") %>% html_text(trim = TRUE),
19     theatre = html_nodes(website, ".col-0 .subtext") %>% html_text(trim = TRUE),
20     weekly_gross = html_nodes(website, ".col-1 .data-value") %>% html_text(trim = TRUE),
21     potential_gross = html_nodes(website, ".td.col-1 .subtext") %>% html_text(trim = TRUE),
22     avg_ticket_price = html_nodes(website, ".col-3 .data-value") %>% html_text(trim = TRUE),
23     top_ticket_price = html_nodes(website, ".td.col-3 .subtext") %>% html_text(trim = TRUE),
24     seats_sold = html_nodes(website, ".col-4 .data-value") %>% html_text(trim = TRUE),
25     seats_in_theatre = html_nodes(website, ".td.col-4 .subtext") %>% html_text(trim = TRUE),
26     pct_capacity = html_nodes(website, ".col-6 .data-value") %>% html_text(trim = TRUE),
27     performances = html_nodes(website, ".col-5 .data-value") %>% html_text(trim = TRUE),
28     previews = html_nodes(website, ".td.col-5 .subtext") %>% html_text(trim = TRUE)
29   )
30
31   tibble(show_stats = show_stats) %>%
32     mutate(variable_name = names(show_stats)) %>%
33     pivot_wider(names_from = variable_name, values_from = show_stats) %>%
34     unnest(cols = everything())
35 }
36
37
38 ## Create tibble with list of URLs and scrape data
39 ## TAKES A LONG TIME (~10 HOURS)
40 broadway_grosses_raw <-
41   tibble(week_ending = seq(ymd("1985-06-09"), ymd("2020-03-01"), by = "1 week")) %>%
42     mutate(grosses_url = paste0("https://www.playbill.com/grosses/week=", week_ending)) %>%
43     mutate(week_data = map(grosses_url, possibly(get_playbill_data, NULL, quiet = FALSE)))
44
45 ## Clean grosses data
46 broadway_grosses <- broadway_grosses_raw %>%
47   unnest(week_data, keep_empty = TRUE) %>%
48   mutate_at(vars(week_number:weekly_gross_overall, weekly_gross:previews),
49     parse_number) %>%
50   mutate(
51     pct_capacity = pct_capacity / 100,
52     show = stringi::stri_trans_general(show, "Latin-ASCII")
53   ) %>%
54   mutate_at(vars(potential_gross, top_ticket_price), ~ ifelse(. == 0, NA, .)) %>%
55   select(-grosses_url)
56
57 ## Write to CSV
58 broadway_grosses %>%
59   write_csv("../broadway-grosses/grosses.csv")
60
61
62
63
64 # Show synopses
65 ## Create function to scrape show synopses
66 get_synopsis <- function(url) {
67   message(url)
68
69   read_html(url) %>%
70     html_nodes(".spotlight-search-result .bsp-list-promo-desc") %>%
71     html_text(trim = TRUE)
72 }
73
74 sync
75 di
76 m
77
78
79
80
81
82 synopsis = map(
83   synopsis_url,
84   possibly(get_synopsis, NA_character_, quiet = FALSE)
85 )
86
87
88 # Clean synopsis data
89 synopsis <- synopsis_raw %>%
90   select(-synopsis_url) %>%
91   unnest(cols = c(synopsis), keep_empty = TRUE)
92
93 ## Write to CSV
94 synopsis %>%
95   write_csv("../broadway-grosses/synopses.csv")
```

Web scraping
code

Falling In Love

What's the most successful Broadway show of all time?

Analyzing Broadway box office grosses

Apr 23, 2020
data exploration / data cleaning / time series data

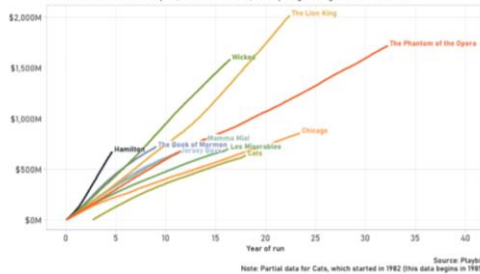
I love musicals! Who doesn't? That feeling when the lights dim at the beginning of the show. The intermission conversation (post-bathroom!) of which songs you enjoyed the most. Spending the rest of the week (maybe month?) humming your favourites to the annoyance of everyone around you.

What's that? *Les Misérables* is obviously the best musical? I know, I know. I mean, *Hamilton* is good and all that, and it deserves praise, but it's no *Les Mis* (don't @ me).

Speaking of *Hamilton*, have you ever wondered how much money it and other Broadway shows have made? Or whether any other shows have come even close to *Hamilton*'s record-breaking ticket prices? We're going to investigate exactly those questions today.

- What are the most successful Broadway shows?
- How have ticket prices changed over time?

Could *Hamilton* overtake *The Lion King* in...20 years?
Cumulative box office receipts (Jan. 2020 dollars) for top 10 grossing shows since 1985



Blog post



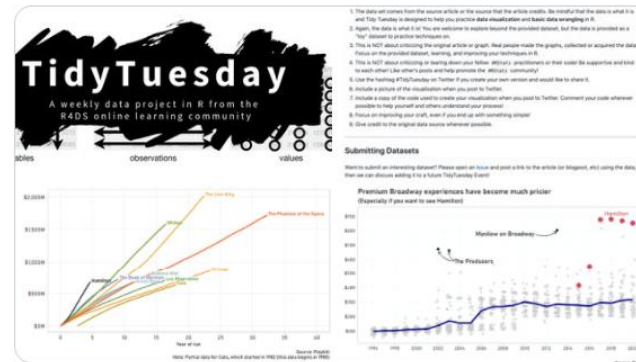
Tom Mock
@thomas_mock

The @R4DScommunity welcomes you to week 18 of #TidyTuesday! We're exploring Broadway shows!!

bit.ly/tidyreadme

bit.ly/2Sa4yWk

#r4ds #tidyverse #rstats #dataviz



11:30 AM · Apr 27, 2020 · r_tweet bot



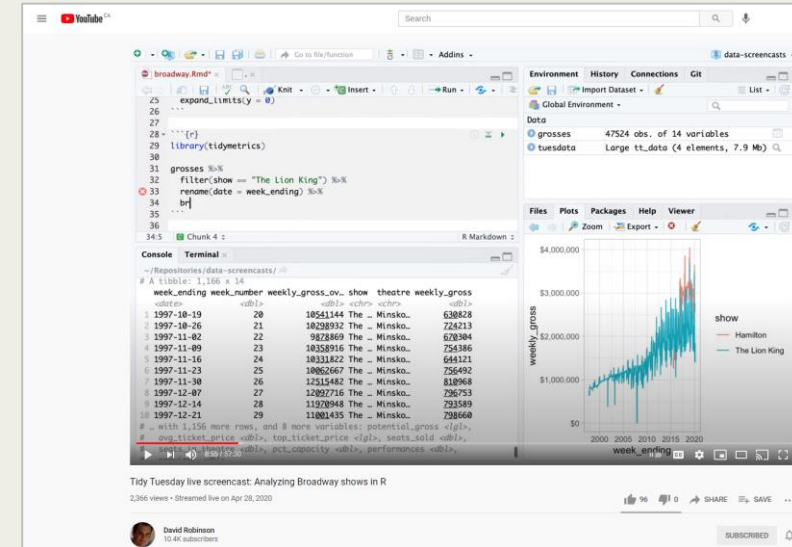
Tom Mock @thomas_mock · Apr 27, 2020

Replying to @thomas_mock

Major shoutout to @alexcookson who provided this week's data, cleaning script and readme!



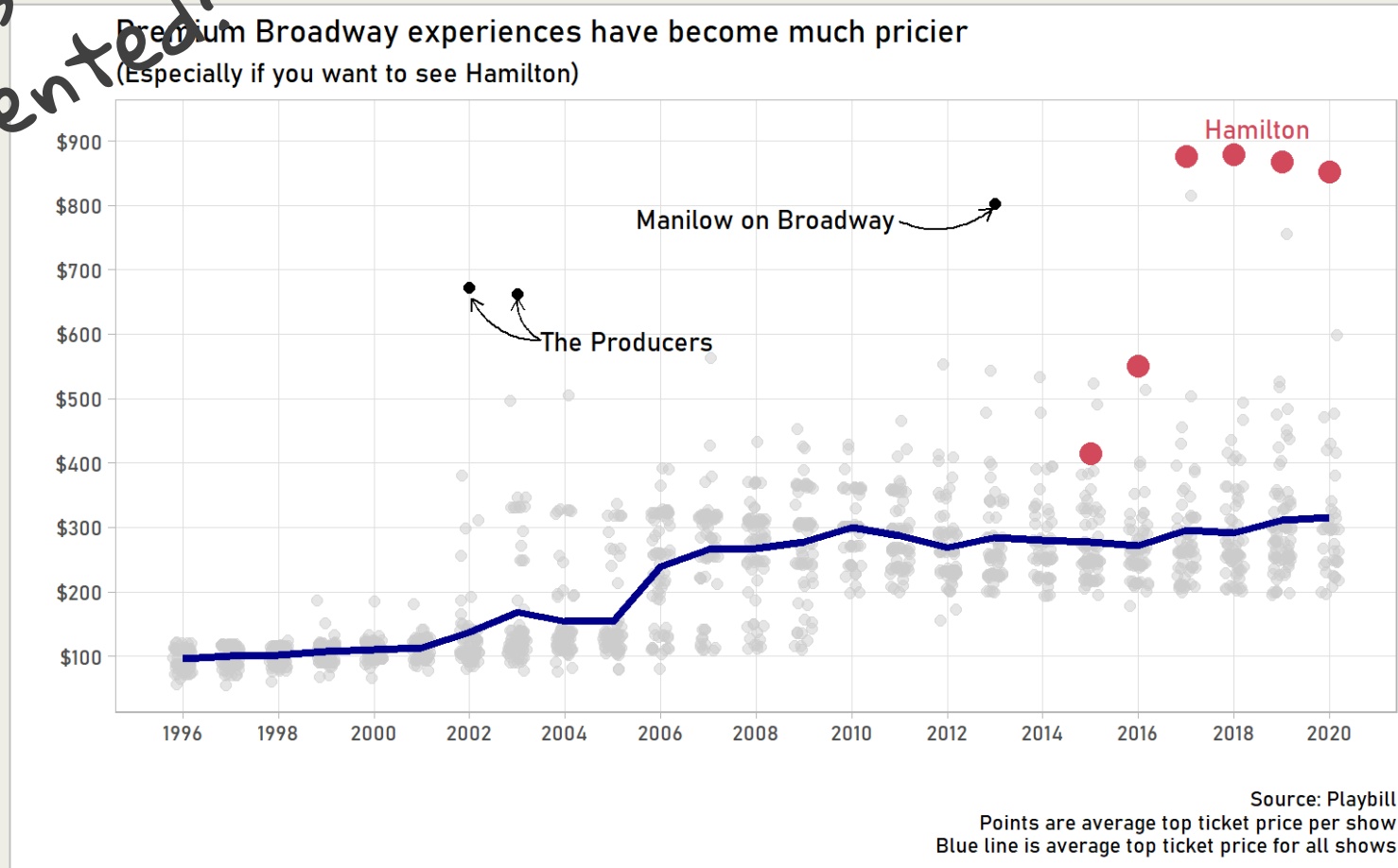
Dataset featured on
Tidy Tuesday



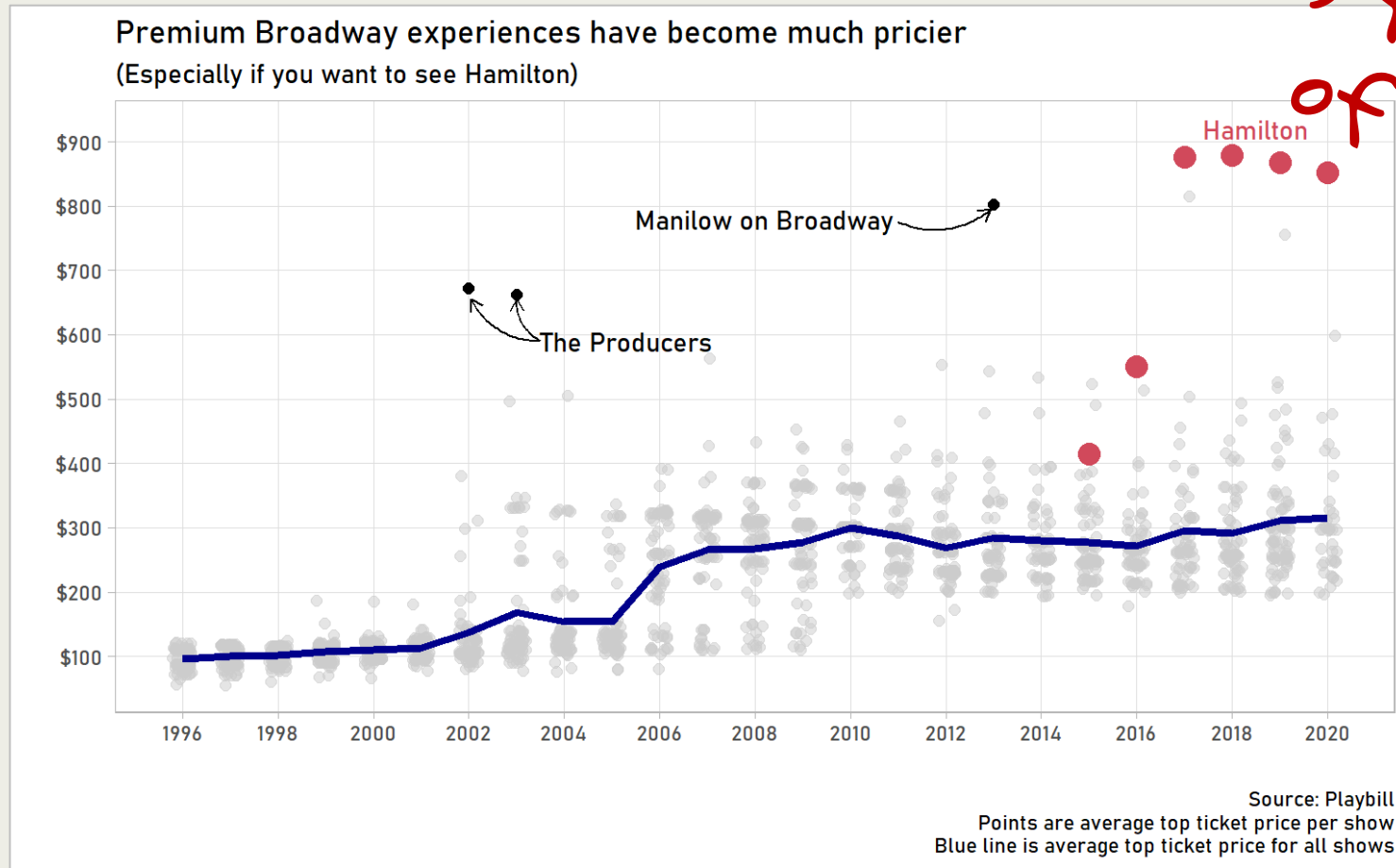
Another David
Robinson screencast

Falling In Love

Is Hamilton's success
unprecedented?

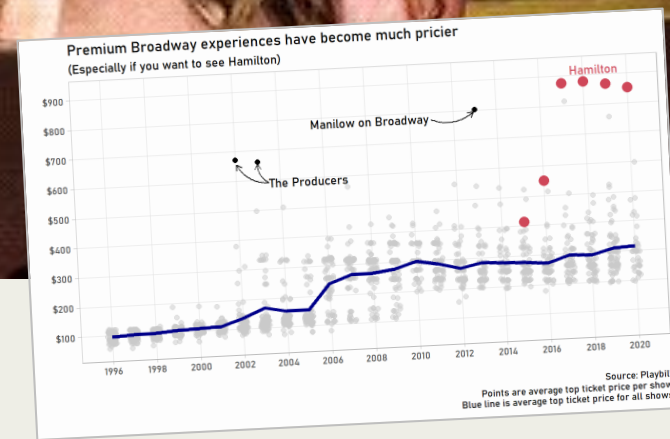
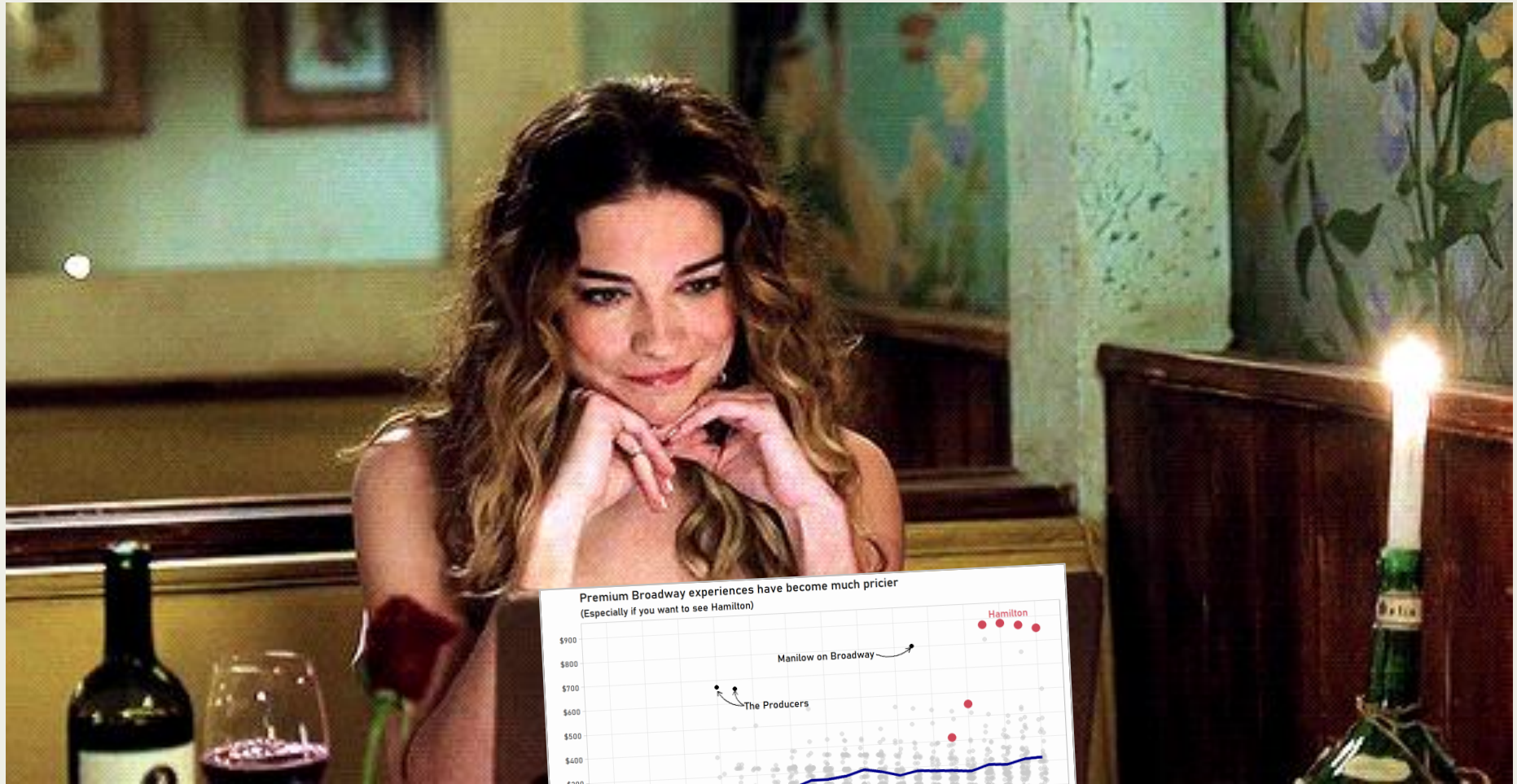


Falling In Love




No! The Producers
was the Hamilton
of its day

Falling In Love



Great Datasets: A Story

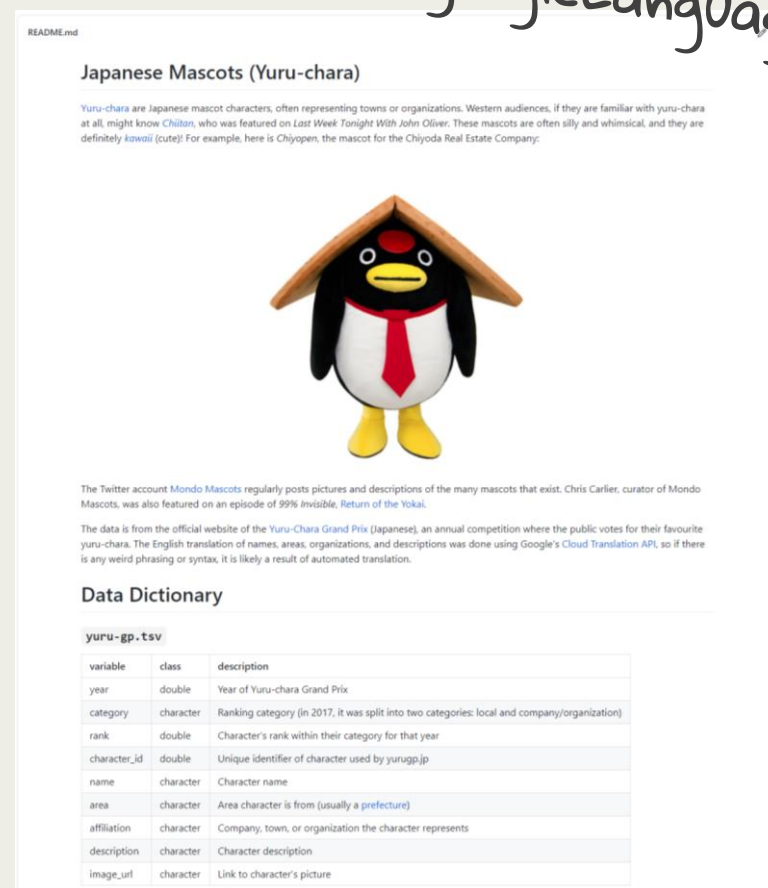
Great Datasets:

~~A~~  Story^{ies}
Many

This Love Story kept repeating itself

More web scraping!

Using Google Language API
with {googleLanguageR}

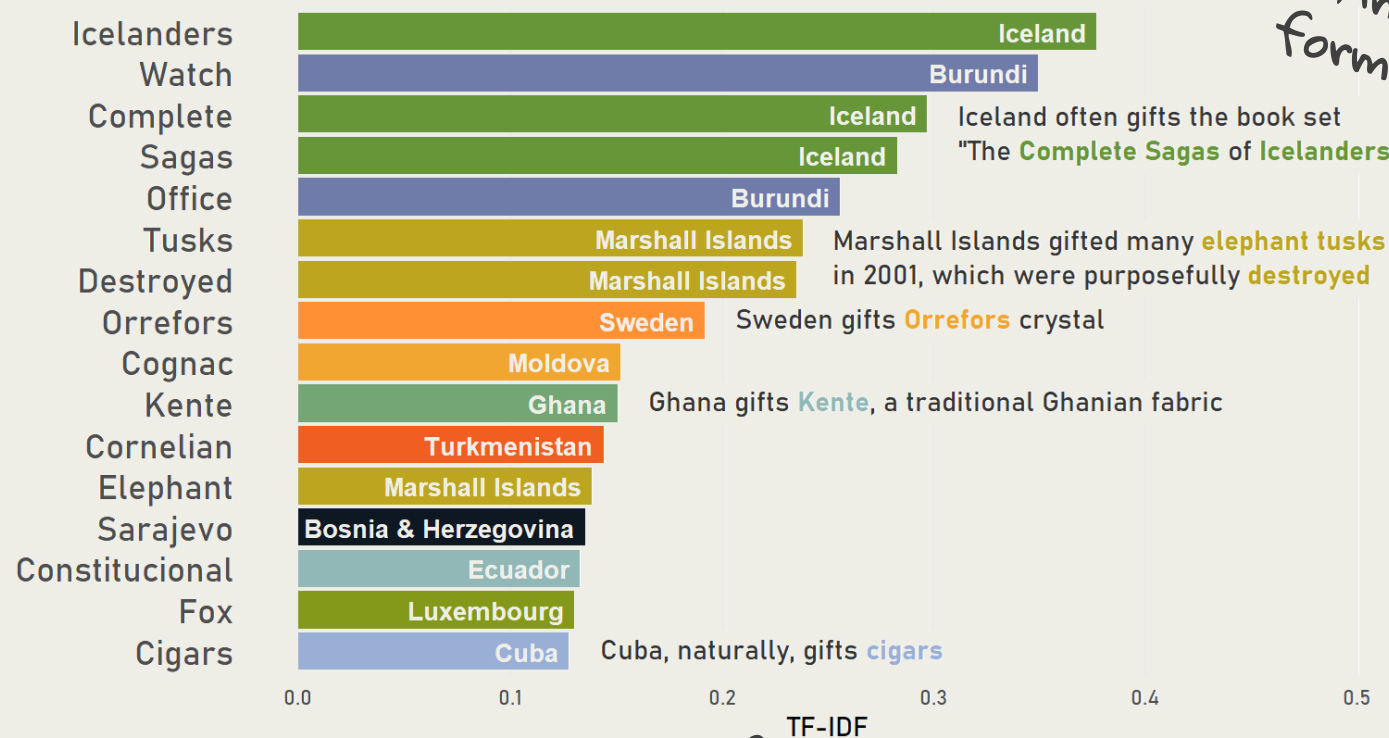


This Love Story kept repeating itself

Parsing dirty PDF data

What are the most "distinctive" words associated with diplomatic gifts?

Based on TF-IDF of gift descriptions from reports in the Federal Register



Annotations and text formatting with `{ggtext}`

Basic Natural Language Processing (TF-IDF)

Visualization: @alexcookson

This Love Story kept repeating itself

More text formatting
with {ggtext}

Creating beautiful
graphs with {ggplot2}

Principal Component
Analysis (PCA)

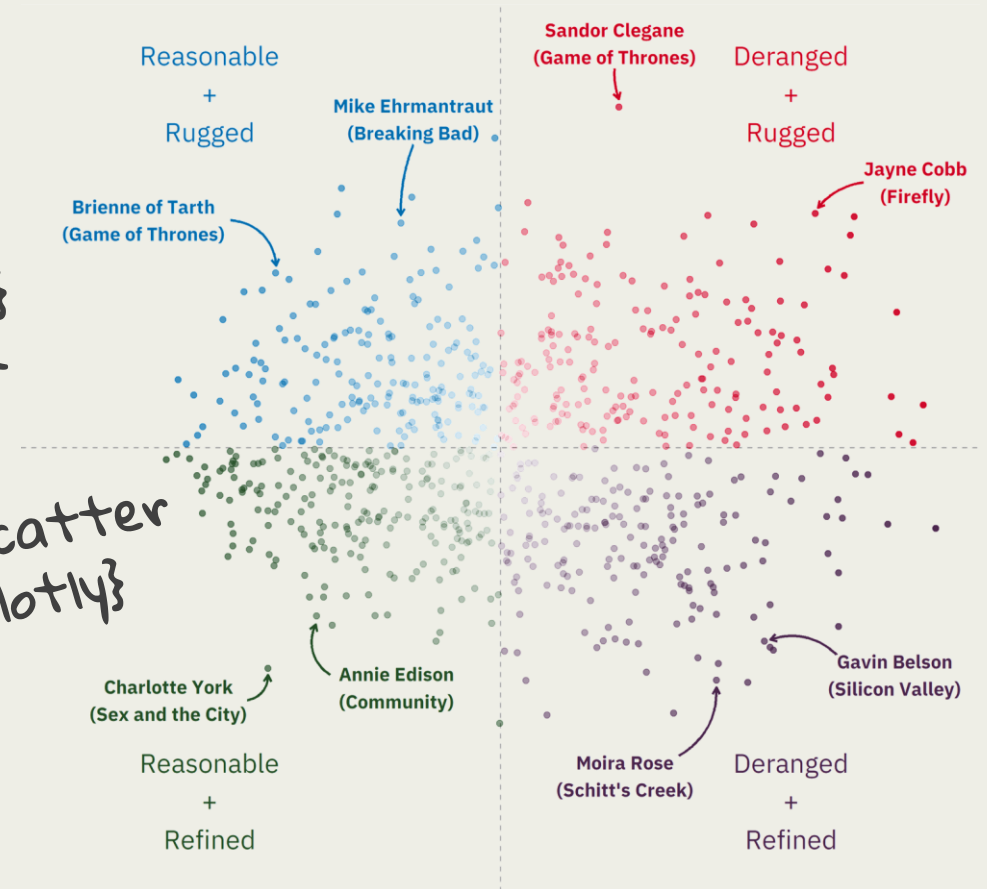
Pride and Prejudice



Visualization: @alexcookson | Data: Open Source Psychometrics Project

{tidymodels}
framework

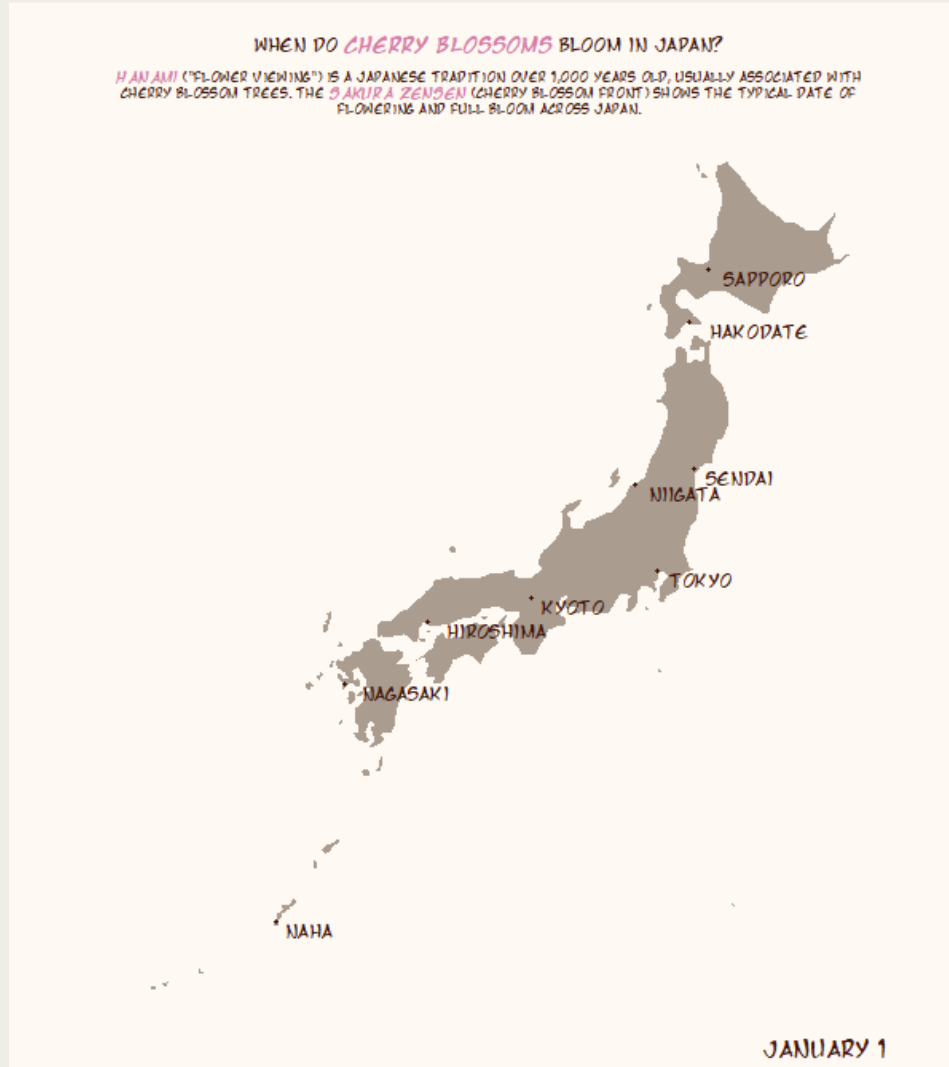
Interactive scatter
plot with {plotly}



This Love Story kept repeating itself

So. Much. {egganimate}.

Custom icons with
{eggimage}



Text formatting with
{eggtext}. Again.

Installing and using
custom fonts

The One Big Benefit

(and Two Small Benefits)

of Great Datasets

Great Datasets are...
a Learning Path

Great Datasets are...a Learning Path

1. Decide what you want to learn

"I want to learn
logistic regression"

Great Datasets are...a Learning Path

1. Decide what you want
to learn

2. Find a dataset to
practice on

"I want to learn
logistic regression"



Titanic Passengers

Great Datasets are...a Learning Path

1. Decide what you want to learn

"I want to learn logistic regression"

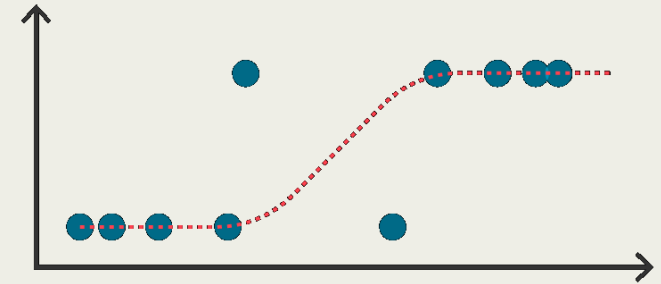


Titanic Passengers

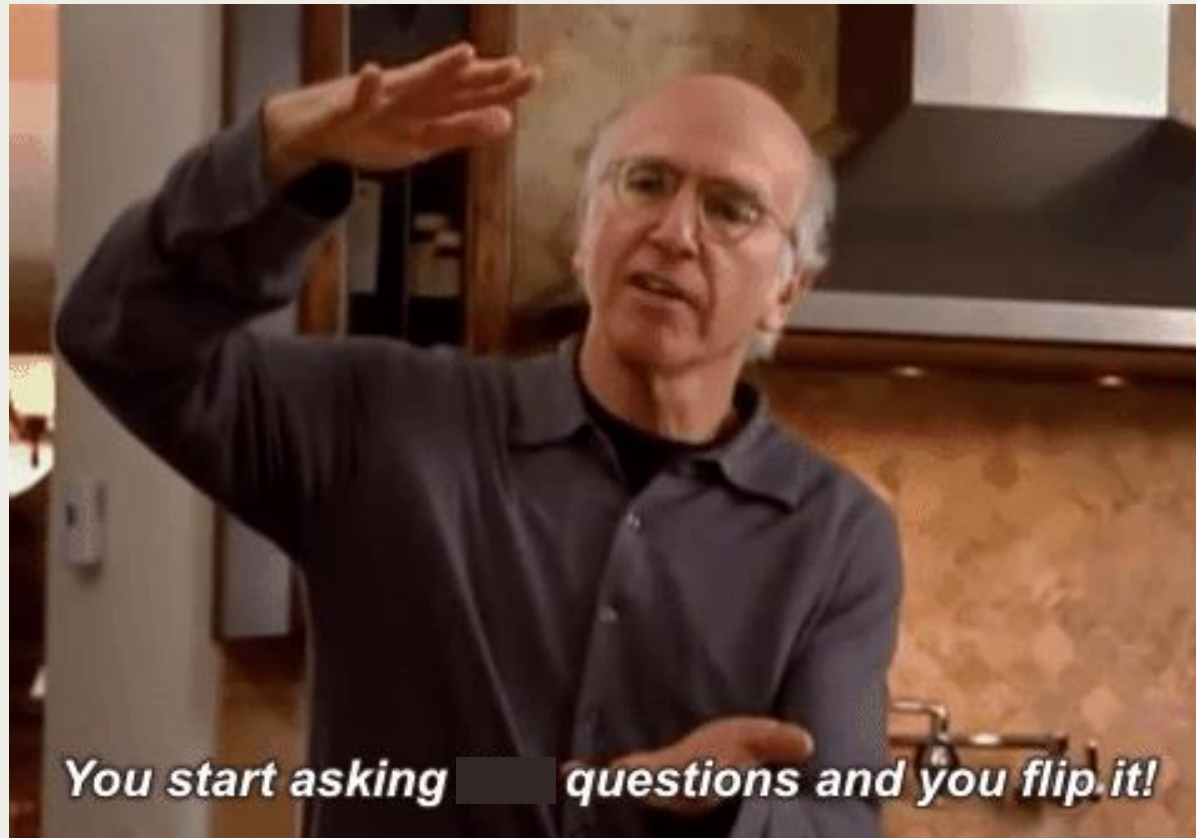
2. Find a dataset to practice on



3. Achieve the outcome



Flip it!



Great Datasets are...a Learning Path

1. Find a dataset that piques your curiosity

Broadway Grosses

Great Datasets are...a Learning Path

1. Find a dataset that
piques your curiosity

2. Let your curiosity
guide you

Broadway Grosses



What questions do I
have for this data?

Great Datasets are...a Learning Path

1. Find a dataset that piques your curiosity

2. Let your curiosity guide you

3. Learn something as a by-product

Broadway Grosses



What questions do I have for this data?



- Web scraping?
- Time-series analysis?
- Forecasting?
- NLP?
- Causal inference?

Great Datasets are...a Learning Path

Focus on this...



1. Find a dataset that piques your **curiosity**

2. Let your curiosity guide you

3. Learn something as a by-product

Broadway Grosses



What questions do I have for this data?



- Web scraping?
- Time-series analysis?
- Forecasting?
- NLP?
- Causal inference?

Great Datasets are...a Learning Path

Focus on this...



1. Find a dataset that piques your **curiosity**

2. Let your curiosity guide you

...not this



3. **Learn something** as a by-product

Broadway Grosses



What questions do I have for this data?



- Web scraping?
- Time-series analysis?
- Forecasting?
- NLP?
- Causal inference?

Great Datasets are...
Sharing who you are

Great Datasets are...Sharing who you are

AMIT LEVINSON

BLOG TALKS

Distances to Golda Ice-Cream Locations in Israel

Feb 10, 2021 · 8 min read · R

What's this all about

This past November I participated several times in the [#30DaysMapChallenge](#), a daily mapping visualization challenge. While I was satisfied with what I came up with, my main outcome was that I have no idea how to work with maps. Well, to say it differently, I was able to fiddle around and hit home eventually, but my knowledge of Coordinate Referencing Systems (CRS) and other important features was limited. For that purpose, I knew I'll be back to explore some additional geographic data, leading to the following blog post.

Static maps

Version 1

We can visualize it on a static map enabling us to easily share it as an image:

Distance from Golda ice-cream locations

Black dots represent Golda ice-cream locations, and the filled color indicates distances from the center of a 25m² area

Distance (km)

- 0-5
- 5-10
- 10-20
- 20-30
- 30-40
- 40-50

Static map of distances to Golda ice-cream locations

**Jesse Mostipak**
@kierisi

TFW you realize [@DukeLemurCenter](#) has released decades worth of lemur data and you have some free time this weekend: [lemur.duke.edu/duke-lemur-cen...](#)



GIF

5:36 PM · May 6, 2020 · Twitter Web App

1 Retweet 31 Likes

**Jesse Mostipak** @kierisi · May 6, 2020
Replying to @kierisi
y'all I can't even 🤔🤔🤔🤔

"Hiddleston is a blue-eyed black lemur, which makes him one of the 25 most endangered primates in the world. Since he was born in March 2013, staff at the Duke Lemur Center have catalogued minute details of his life in their daily logbooks...

SPORT DATA SCIENCE
Join the R Stats Adventure Here!

Home About Contact cricket football F1 Machine Learning

F1 2020 Season Review

JANUARY 4, 2021 BY PART TIME ANALYST

Hello readers, its Monday the 4th January and this is the first of my hopefully weekly blogs in 2021. We will see how long that lasts! Today I'm going to be looking at the data underpinning the 2020 F1 season. The story of the season is clear Lewis Hamilton dominated to win his 7th world title. In the process he has now won the most races of any F1 driver ever. I'm going to delve behind the headlines and really look at the best and worst performing drivers and teams.

2020 Season xPos

Driver

Mean xPos

Subscribe to Blog via Email

Enter your email address to subscribe to this blog and receive notifications of new posts by email.

Join 278 other followers

Enter your email address

Subscribe

Follow Sport Data Science

Recent Posts

- [Building a Model in R to Predict FPL Points March 9, 2021](#)
- [Win Probability Added - Batsman Review January 12, 2021](#)
- [F1 2020 Season Review January 4, 2021](#)
- [F1 2020 -Season So Far and Why Racing Point's Method of](#)

Great Datasets are...
Community-builders

Great Datasets are...Community-builders

Me when someone likes a tweet
about a dataset I shared

- If you like a dataset...so does someone else
- More (and more diverse) datasets enrich the community
- Curating and sharing datasets is a way for you to engage with the community



Okay Alex, I'm convinced!

How do I start?

Getting Going with Great Datasets

1. Start with existing datasets ([TidyTuesday](#), [Kaggle](#), my [data](#) repository)
2. Don't force it – know your interests and be watchful
3. Know that curating your own datasets will probably involve web scraping
4. Share! Encouragement from others is a great motivator

Great Datasets: A Story

Alex Cookson

 [@alexcookson](https://twitter.com/alexcookson)

 [tacookson](https://github.com/tacookson)