# Some messy work from our TidyTuesday hangout

Liza Bolton

2020-05-05

Thanks to everyone who came to hangout for this informal session. It gave me some good ideas about what you might be interested in next.

**Was this recorded?** No, as this was just an informal session I didn't record it, but if people are interested in more structured future tutorials we can record those for sure.

**Is this going to be a regular thing?** I'd be keen to run a little something each week, though probably not a full-blown tutorial. Let me know what you'd be most interested in. I'm picturing a mix of social check-in and a stats version of a 'let's play', hopefully with some tips related to the week's portfolio building theme (this week's theme was TidyTuesday itself + GitHub). Next week's theme is data viz and data wrangling, so definitely some fun stuff to do there.

**I couldn't make this time because I had class!** Please keep an eye out for our survey in the next 6*sigma* Sunday, we'll check in on a few things, and one of them will be what times are good for the most people for running webinars or social events.

## We talked about:

- What TidyTuesday is
- the 'lazy' way to get data and that you can use GitHub too
- took a quick look at the data and played around
- briefly looked at sentiment analysis in R
- talked about how many employers appreciate text processing skills and database skills (things like joins)
- why Twitter can be a great way to build your professional presence (and how people have said they got jobs from sharing their TidyTuesday work)

Here is an example:

My #rstats friends, can you believe that I got a job interview and possibly an offer because I made this treemap for #TidyTuesday? You never know who is watching you till you get calls from those who would like hire you. Keep up the good work and share with us! https://t.co/bdI0zJ5vjJ

— Zhi Yang, PhD (@zhiiiyang) October 18, 2019

## Twitter

On this note, some folks I'd recommend following on Twitter

- @hadleywickham (Hadley Wickham) - if R users have a rock star. . . it is Hadley. My fried got him to sign his laptop. He is also a genuinely good bloke. Check out his FREE textbook R for Data Science https://r4ds.had.co.nz/

- @juliasilge (Julia Silge) - talented in a range of areas, but I know her most for her work with text in R, see here FREE book on it: https://www.tidytextmining.com/
- @drob (David Robinson) - he's actually doing some TidyTuesday screencasts
- @djnavarro (Danielle Navarro) - she has also put up some great YouTube mini courses, including one about generative aRt (you'll see beautiful examples on her feed)
- @thomas_mock (Tom Mock) - one of the people behind TidyTuesday
- @minebocek (Mine CetinkayaRundel) - she is a stats education MVP and runs Duke's ASA DataFest chapter

And soooo many more. I'll keep a list as I think of them and share again. You should also check out the DoSS Twitter account: https://twitter.com/UofTStatSci, and there are lists of people you might want to follow, like our faculty, or their bigger data science or [statistics])(https://twitter.com/i/lists/1004778877691015169/members) lists.

Check out what is happening in the #rstats and #TidyTuesday hashtags too.

### A 'gathering data' webinar

Keep an eye out for a webinar Rohan Alexander (@RohanALexander) is planning to run on 'gathering data'. Looks like it will be May 22 at 11 am and cover things like API, web scraping and ethics but all TBC at this point.

## TidyTuesday: Animal Crossing

So the real purpose of today was a chill intro to TidyTuesday and to take a peek at this week's dataset on Animal Crossing.

### Loading the data

I did this the very lazy way, just by copying from the TidyTuesday data page. This chunk loads the packages I used (you may need to install them, the commented out part) and the datasets.

*Installing packages is a pretty core foundational professional skill for an R user (at least in my opinion), so if you still have some discomfort with it, let me know! Might be a topic for a mini-tutorial.*

```
#install.packages("tidyverse")
library("tidyverse")
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
#install.packages("tidytext")
library(tidytext)

critic <- readr::read_tsv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/20
```

```
## Parsed with column specification:
## cols(
##   grade = col_double(),
##   publication = col_character(),
##   text = col_character(),
##   date = col_date(format = "")
## )
```

```r
user_reviews <- readr::read_tsv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da
```

```
## Parsed with column specification:
## cols(
##   grade = col_double(),
##   user_name = col_character(),
##   text = col_character(),
##   date = col_date(format = "")
## )
```

```r
items <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020
```

```
## Parsed with column specification:
## cols(
##   num_id = col_double(),
##   id = col_character(),
##   name = col_character(),
##   category = col_character(),
##   orderable = col_logical(),
##   sell_value = col_double(),
##   sell_currency = col_character(),
##   buy_value = col_double(),
##   buy_currency = col_character(),
##   sources = col_character(),
##   customizable = col_logical(),
##   recipe = col_double(),
##   recipe_id = col_character(),
##   games_id = col_character(),
##   id_full = col_character(),
##   image_url = col_character()
## )

## Warning: 2 parsing failures.
##  row          col           expected actual
## 4472 customizable 1/0/T/F/TRUE/FALSE    Yes 'https://raw.githubusercontent.com/rfordatascience/tidytu
## 4473 customizable 1/0/T/F/TRUE/FALSE    Yes 'https://raw.githubusercontent.com/rfordatascience/tidytu
```

```
villagers <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/
```

```
## Parsed with column specification:
## cols(
##   row_n = col_double(),
##   id = col_character(),
##   name = col_character(),
##   gender = col_character(),
##   species = col_character(),
##   birthday = col_character(),
##   personality = col_character(),
##   song = col_character(),
##   phrase = col_character(),
##   full_id = col_character(),
##   url = col_character()
## )
```

## Playing around

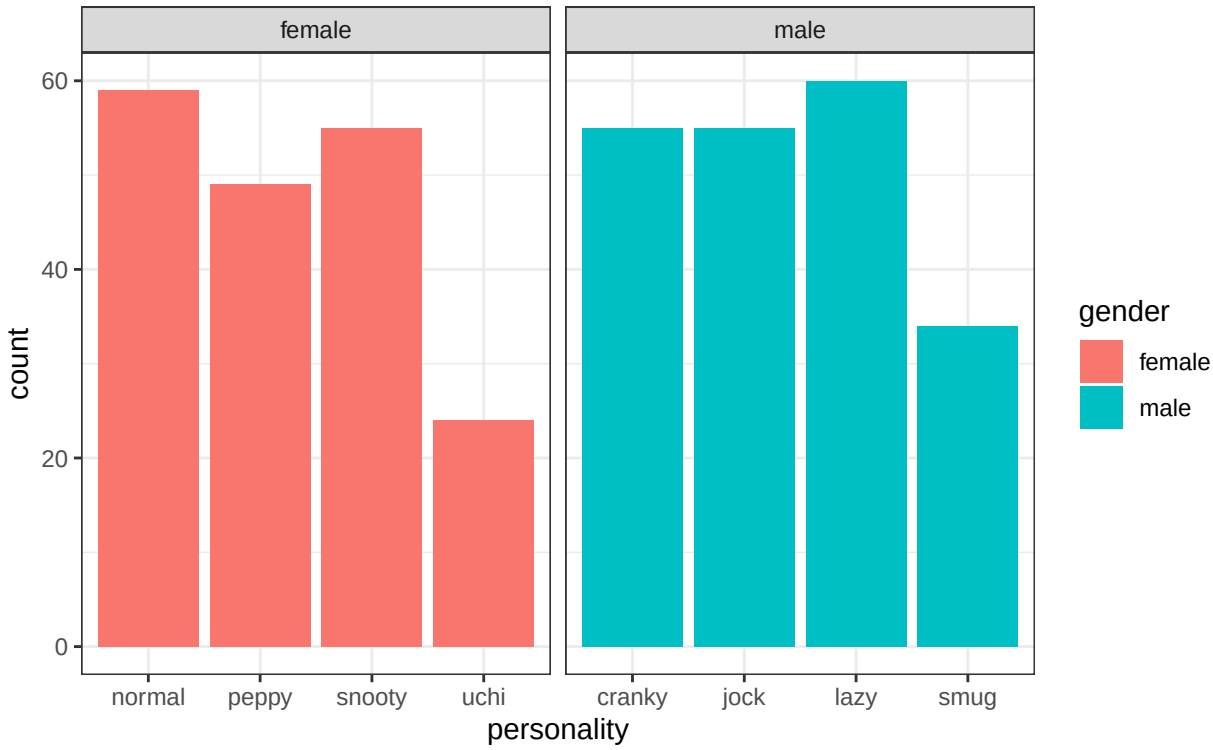
### Gender and personality

Who knew? Only men can be cranky and only women can be snooty! At least, among villagers...

```
villagers %>%
  group_by(personality, gender) %>%
  summarise(n = n()) %>%
  arrange(desc(n))
```

```
## # A tibble: 8 x 3
## # Groups:   personality [8]
##   personality gender     n
##   <chr>       <chr>  <int>
## 1 lazy        male      60
## 2 normal      female    59
## 3 cranky      male      55
## 4 jock        male      55
## 5 snooty      female    55
## 6 peppy       female    49
## 7 smug        male      34
## 8 uchi        female    24
```

```
villagers %>%
  group_by(personality, gender) %>%
  ggplot(aes(x = personality, fill = gender)) +
  facet_wrap(~gender, scales = "free_x") + #great function for splitting out plots on a categorical var
  geom_bar() +
  theme_bw() #change the look of a plot really quickly with different theme options
```

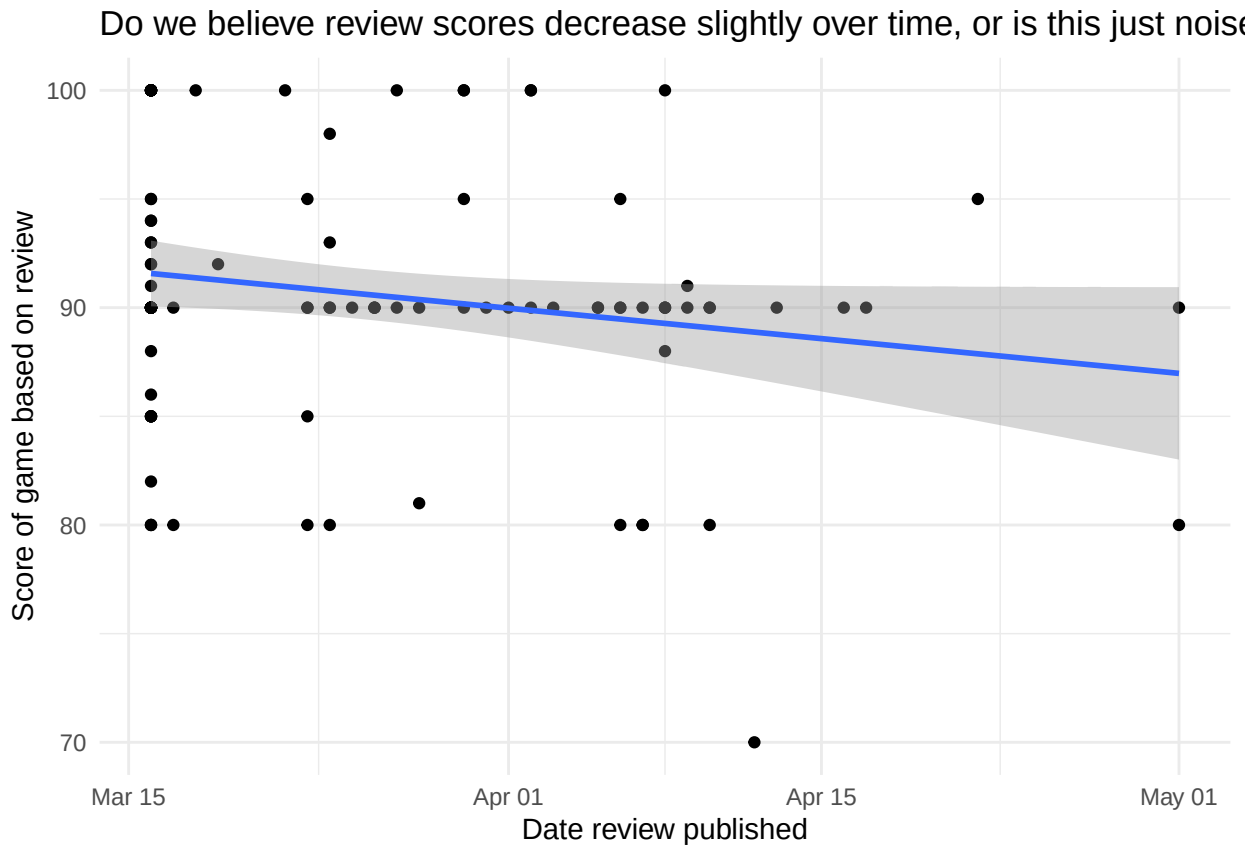



**Everyone's a critic: reviews over time**

Anything interesting about the reviews overtime? Sometimes with movies the prescreen reviews are biased to be more positive than the movie "deserves"....

```
critic %>%
  ggplot(aes(x = date, y = grade)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_minimal() +
  ggtitle("Do we believe review scores decrease slightly over time, or is this just noise?") +
  xlab("Date review published") +
  ylab("Score of game based on review")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




I guess there is a slight downward curve... but I don't know if i *trust* it as a finding.

```
summary(lm(grade ~ date, data = critic))
```

```
##
## Call:
## lm(formula = grade ~ date, data = critic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8750  -1.5732  -0.0742   2.7253  10.7253
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1924.07418  958.94999   2.006   0.0474 *
## date          -0.09993    0.05227  -1.912   0.0586 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.039 on 105 degrees of freedom
## Multiple R-squared:  0.03364,    Adjusted R-squared:  0.02444
## F-statistic: 3.655 on 1 and 105 DF,  p-value: 0.05861
```

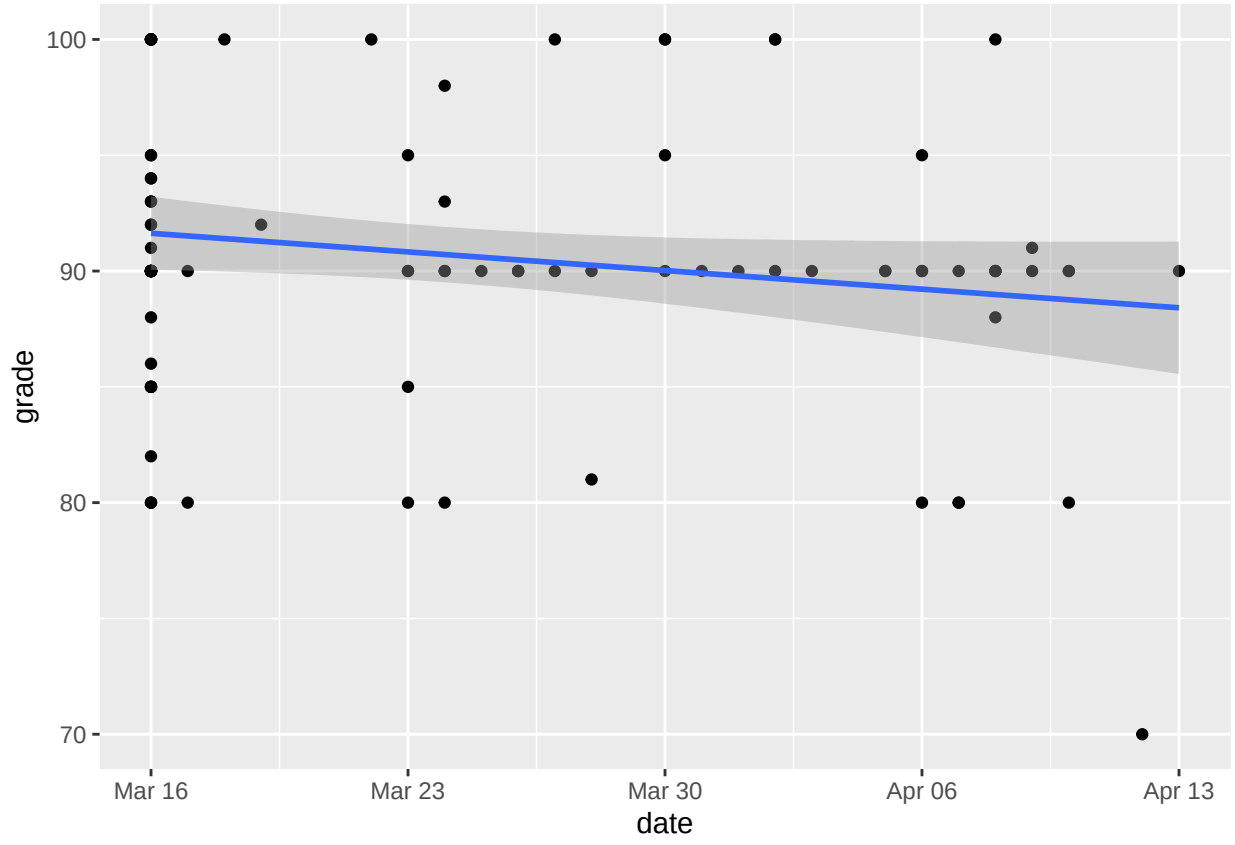
But there is only some evidence against the slope being 0 (i.e. a flat line)...

Any different if we restrict to the first month reviews (March 15 to April 15)?

```r
critic_restricted <- critic %>%
  filter(date < "2020-04-16")

critic_restricted %>%
  ggplot(aes(x = date, y = grade)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
summary(lm(grade ~ date, data = critic_restricted))
```

```
##
## Call:
## lm(formula = grade ~ date, data = critic_restricted)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.5276  -1.6306  -0.1365   2.3694  11.0127
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2199.02983 1197.17899   1.837   0.0692 .
## date          -0.11493    0.06526  -1.761   0.0813 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.094 on 100 degrees of freedom
## Multiple R-squared:  0.03008,    Adjusted R-squared:  0.02038
## F-statistic: 3.101 on 1 and 100 DF,  p-value: 0.08128
```

Not really.

This isn't enough to make me claim that there is a 'hype' effect at the beginning, but I'd be curious if the gamers in the ISSC had any thoughts.

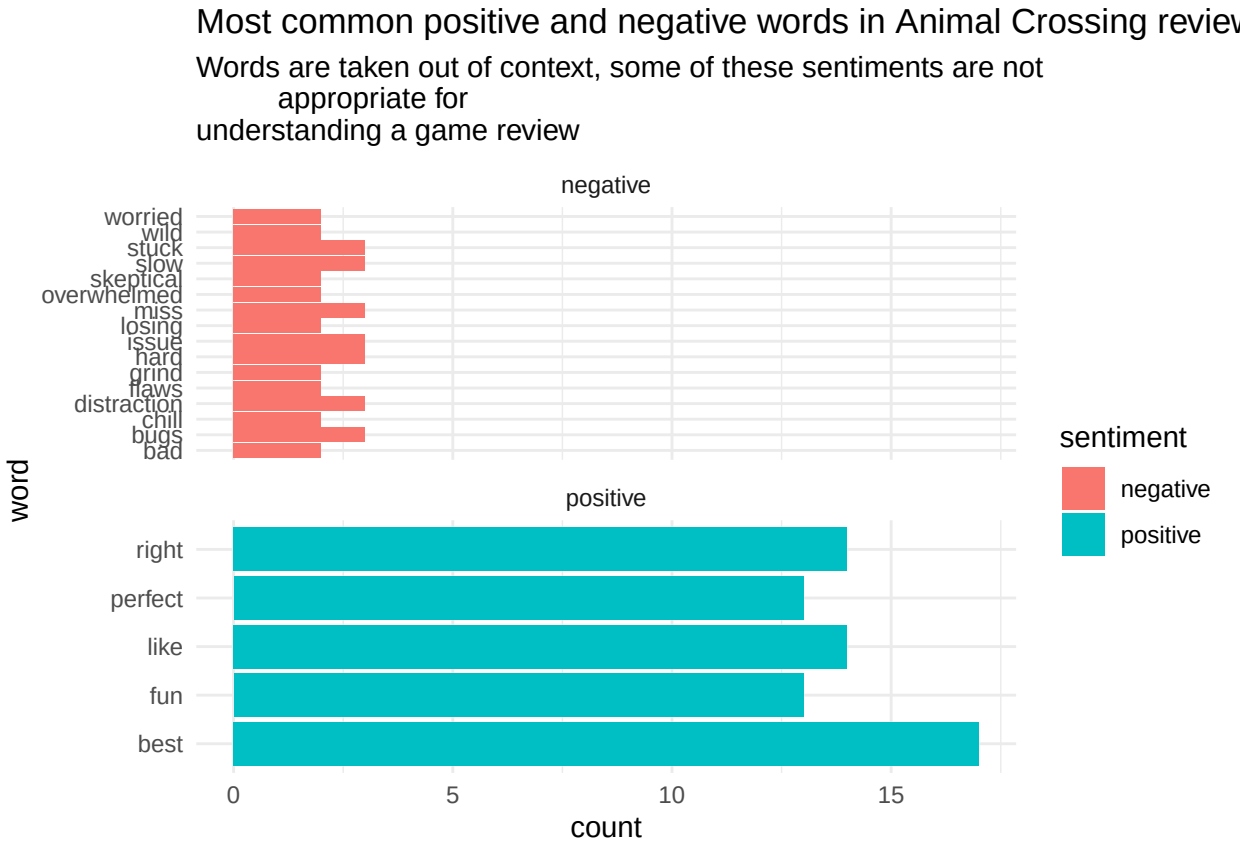
### Everyone's a critic: sentiment analysis quick example

This little block is an example of some basic text analysis AND using a join to put together two datasets. Both are skills that some employers really appreciate (at least based on panels and workshops I've been to. Joins are especially useful for the consultants I was training back in NZ).

Take a look at the TidyText book for more: https://www.tidytextmining.com/

```
# Get the sentiments of words from a preset library
bing <- get_sentiments("bing")

# Join the sentiments on the teh words from the reviews
critic %>%
  select(text) %>%
  unnest_tokens(word, text) %>%
  #group_by(word) %>%
  #summarise(count = n()) %>%
  left_join(bing, by="word") %>%
  filter(!is.na(sentiment)) %>%
  group_by(word, sentiment) %>%
  summarise(count = n()) %>%
  filter(count>1) %>%  # filter to words appearing more than once (and a sentiment score)
  arrange(desc(count)) %>%
  group_by(sentiment) %>%
  filter(count > max(count) - 5) %>% # get the top couple words of each sentiment
  ggplot(aes(x = word, y = count, fill = sentiment)) +
    geom_bar(stat = "identity") + #to just use the count var for the height of the bars
  coord_flip() +
    facet_wrap(~sentiment, nrow = 2, scales = "free_y") + # this drops the unused levels
    theme_minimal() +
  ggtitle("Most common positive and negative words in Animal Crossing reviews",
          subtitle = "Words are taken out of context, some of these sentiments are not
          appropriate for\nunderstanding a game review")
```


Most common positive and negative words in Animal Crossing reviev
Words are taken out of context, some of these sentiments are not
appropriate for
understanding a game review
negative
positive
word
worried
wild
stuck
slow
skeptical
overwhelmed
miss
losing
issue
hard
grind
flaws
distraction
chill
bugs
bad
right
perfect
like
fun
best
0
5
10
15
count
sentiment
negative
positive