BST 270 Individual Project Analysis

2022-1-18

We seek to reproduce the figures presented in the 2017 538 article "How 'Qi' and 'Za' Changed Scrabble."

Load Data

Load scrabble game data directly from corresponding 538 GitHub repo.

```
scrabble_df = read.csv("https://media.githubusercontent.com/media/fivethirtyeight/data/master/scrabble-
```

```
head(scrabble_df)
```

```
##   gameid tourneyid   tie winnerid          winnername winnerscore
## 1      1         1 False      268 Harriette Lakernick           0
## 2      2         1 False      268 Harriette Lakernick           0
## 3      3         1 False      268 Harriette Lakernick           0
## 4      4         1 False      268 Harriette Lakernick           0
## 5      5         1 False      268 Harriette Lakernick           0
## 6      6         1 False      268 Harriette Lakernick           0
##   winneroldrating winnernewrating winnerpos loserid        losername loserscore
## 1            1568            1684         1     429 Patricia Barrett          0
## 2            1568            1684         1     435       Chris Cree          0
## 3            1568            1684         1     441 Caesar Jaramillo          0
## 4            1568            1684         1     456    Mike Chitwood          0
## 5            1568            1684         1    1334      Nancy Scott          0
## 6            1568            1684         1     454     Mary Rhoades          0
##   loseroldrating losernewrating loserpos round division       date lexicon
## 1           1915           1872        3     1        1 1998-12-06   False
## 2           1840           1798        6     2        1 1998-12-06   False
## 3           1622           1606       10     3        1 1998-12-06   False
## 4           1612           1600        9     4        1 1998-12-06   False
## 5           1537           1590        4     6        1 1998-12-06   False
## 6           1676           1647        8     8        1 1998-12-06   False
```

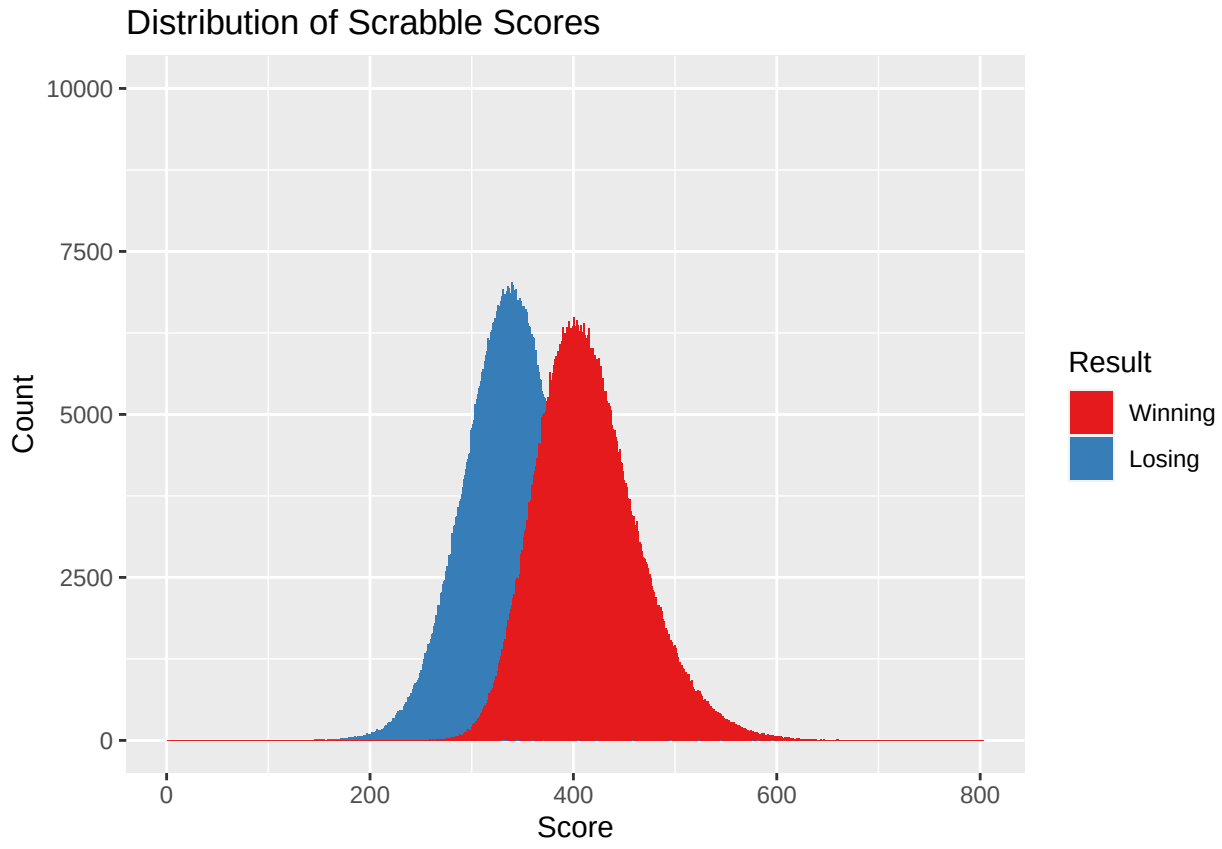
Visualize Distribution of Winning and Losing Scrabble Scores

```
# Keep winner scores and/or loser scores strictly greater than 0; ties are included
wl_df = scrabble_df[scrabble_df$winnerscore > 0 &
                    scrabble_df$loserscore > 0,]

# Assume games where winner scores are less than loser scores were recorded incorrectly
# Swap scores if winner score is less than loser score
```

```r
wl_df2 = wl_df %>% transform(
  winnerscore_new = ifelse(winnerscore < loserscore, loserscore, winnerscore),
  loserscore_new = ifelse(loserscore > winnerscore, winnerscore, loserscore))

dim(wl_df2) # This leaves us with 770653 Scrabble games
```

```
## [1] 770653     21
```

```r
# Plot histograms of winner and loser scores
p1 = ggplot(data = wl_df2) +
  geom_histogram(aes(x = loserscore_new, fill = "Losing"), binwidth = 1) +
  geom_histogram(aes(x = winnerscore_new, fill = "Winning"), binwidth = 1) +
  labs(x = "Score", y = "Count", title = "Distribution of Scrabble Scores", fill = "Result") +
  ylim(0, 10000) +
  scale_fill_manual(values = c("Winning" = "#E41A1C", "Losing" = "#377EB8"))
p1
```



Plot Average Scrabble Scores Before and After 'Qi'/'Za'

```r
# Add average score column and divide date into years, months, and days
score_df = wl_df %>% mutate(avgscore = (winnerscore + loserscore)/2) %>%
  mutate(year = as.numeric(format(as.Date(date), format = "%Y")),
```

```
         month = as.numeric(format(as.Date(date), format = "%m")),
         day = as.numeric(format(as.Date(date), format = "%d"))) %>%
  select(avgscore, year, month, day)

head(score_df)
```

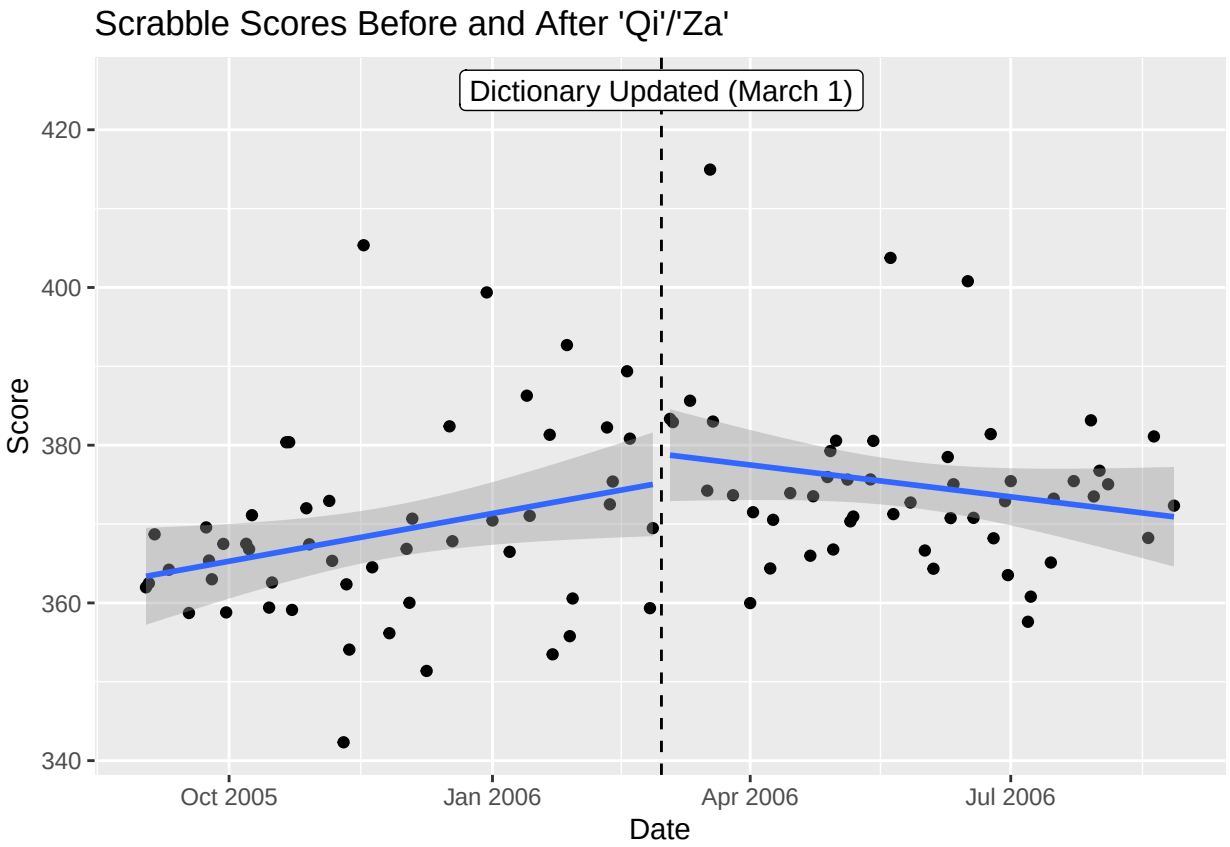
```
##      avgscore year month day
## 2720    379.0 1999     1  15
## 2721    375.0 1999     1  15
## 2722    397.5 1999     1  15
## 2723    385.5 1999     1  15
## 2724    348.0 1999     1  15
## 2725    427.5 1999     1  15
```

```
# Find average score for tournaments between September 2005 and September 2006
score_new_df = score_df[score_df$year >= 2005 & score_df$year < 2007,] %>% group_by(year, month, day) %

head(score_new_df)
```

```
## # A tibble: 6 x 5
## # Groups:   year, month [1]
##    year month   day avgscore date
##   <dbl> <dbl> <dbl>    <dbl> <date>
## 1  2005     9     2     362. 2005-09-02
## 2  2005     9     3     363. 2005-09-03
## 3  2005     9     5     369. 2005-09-05
## 4  2005     9    10     364. 2005-09-10
## 5  2005     9    17     359. 2005-09-17
## 6  2005     9    23     370. 2005-09-23
```

```
# Fit regression lines to data before and after March 1, 2006
score_new_df1 = score_new_df[(score_new_df$year == 2005) | (score_new_df$year == 2006 & score_new_df$mon
score_new_df2 = score_new_df[(score_new_df$year == 2006) & (score_new_df$month >= 3),]

# Plot average Scrabble scores between September 2005 and September 2006
p2 = ggplot(data = score_new_df, aes(x = date, y = avgscore)) +
  geom_point() +
  geom_smooth(data = score_new_df1, method = lm, se = TRUE) +
  geom_smooth(data = score_new_df2, method = lm, se = TRUE) +
  geom_vline(xintercept = as.numeric(as.Date("2006-03-01")), linetype="dashed") +
  annotate(geom="label", label="Dictionary Updated (March 1)", x = as.Date("2006-03-01"), y = 425) +
  labs(x = "Date",
       y = "Score",
       title = "Scrabble Scores Before and After 'Qi'/'Za'",
       fill = "Result")
p2
```

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


Scrabble Scores Before and After 'Qi'/'Za'
Dictionary Updated (March 1)
420
400
380
360
340
Score
Oct 2005
Jan 2006
Apr 2006
Jul 2006
Date