Project Proposal

Dancing Queens: Karianna Klassen, Madeline Waterfield, Mark Yukelis, Shrikrishna Sriram

```r
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

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

```r
library(tidytuesdayR)
```

```r
spotify_data <- read_csv("spotify_songs.csv")
```

```
## Rows: 32833 Columns: 23

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): track_id, track_name, track_artist, track_album_id, track_album_na...
## dbl (13): track_popularity, danceability, energy, key, loudness, mode, speec...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
covid_data <- read_csv("covid_approval_polls.csv")
```

```
## Rows: 3302 Columns: 13

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): pollster, sponsor, population, party, subject, text, url
## dbl  (3): sample_size, approve, disapprove
## lgl  (1): tracking
## date (2): start_date, end_date

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
mlb_data <- read_csv("mlb_elo.csv")
```

```
## Rows: 223427 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (5): playoff, team1, team2, pitcher1, pitcher2
## dbl  (20): season, neutral, elo1_pre, elo2_pre, elo_prob1, elo_prob2, elo1_p...
## date  (1): date

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Dataset 1

Data Set #1: Spotify Song Data by Musical Metrics

This data set comes from the TidyTuesday Github and was sourced using the Spotify API. The data set contains 32,833 observations (songs) from various genres. There are 23 variables, all in all, and are split between identifiers and key metrics. A few important variables are track popularity, energy, and duration_ms.

Research questions for Data Set #1: (1) To what degree does the length of a song affect its popularity? How might this vary across genres? (2) Here, are closely aligned metrics such as danceablity and energy truly independent or do certain underlying factors separate them altogether?

```
glimpse(spotify_data)
```

```
## Rows: 32,833
## Columns: 23
## $ track_id                 <chr> "6f807x0ima9a1j3VPbc7VN", "0r7CVbZTWZgbTCYdfa~
## $ track_name               <chr> "I Don't Care (with Justin Bieber) - Loud Lux~
## $ track_artist             <chr> "Ed Sheeran", "Maroon 5", "Zara Larsson", "Th~
## $ track_popularity         <dbl> 66, 67, 70, 60, 69, 67, 62, 69, 68, 67, 58, 6~
## $ track_album_id           <chr> "2oCs0DGTsRO98Gh5ZSl2Cx", "63rPSO264uRjW1X5E6~
## $ track_album_name         <chr> "I Don't Care (with Justin Bieber) [Loud Luxu~
## $ track_album_release_date <chr> "2019-06-14", "2019-12-13", "2019-07-05", "20~
## $ playlist_name            <chr> "Pop Remix", "Pop Remix", "Pop Remix", "Pop R~
## $ playlist_id              <chr> "37i9dQZF1DXcZDD7cfEKhW", "37i9dQZF1DXcZDD7cf~
## $ playlist_genre           <chr> "pop", "pop", "pop", "pop", "pop", "pop", "po~
## $ playlist_subgenre        <chr> "dance pop", "dance pop", "dance pop", "dance~
## $ danceability             <dbl> 0.748, 0.726, 0.675, 0.718, 0.650, 0.675, 0.4~
## $ energy                   <dbl> 0.916, 0.815, 0.931, 0.930, 0.833, 0.919, 0.8~
## $ key                      <dbl> 6, 11, 1, 7, 1, 8, 5, 4, 8, 2, 6, 8, 1, 5, 5,~
## $ loudness                 <dbl> -2.634, -4.969, -3.432, -3.778, -4.672, -5.38~
## $ mode                     <dbl> 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, ~
## $ speechiness              <dbl> 0.0583, 0.0373, 0.0742, 0.1020, 0.0359, 0.127~
## $ acousticness             <dbl> 0.10200, 0.07240, 0.07940, 0.02870, 0.08030, ~
## $ instrumentalness         <dbl> 0.00e+00, 4.21e-03, 2.33e-05, 9.43e-06, 0.00e~
## $ liveness                 <dbl> 0.0653, 0.3570, 0.1100, 0.2040, 0.0833, 0.143~
## $ valence                  <dbl> 0.518, 0.693, 0.613, 0.277, 0.725, 0.585, 0.1~
## $ tempo                    <dbl> 122.036, 99.972, 124.008, 121.956, 123.976, 1~
## $ duration_ms              <dbl> 194754, 162600, 176616, 169093, 189052, 16304~
```

Dataset 2

Data Set #2: How American's View Biden's Response to the Coronavirus Crisis from Five Thirty Eight.

This data is sourced from a variety of polls conducted by different pollsters across the United States who each collected their data in slightly different ways. There are 3,302 observations, or polls, from various pollsters. There are 13 variables that are split between identifiers of the polls (like the pollster) and key metrics like the approval rate. Some important variables are political party of the respondent, sponsor of the poll, and the sample size of the poll.

Research Questions for Data Set #2: (1) To what degree does political party of the respondents affect the poll's outcome? (2) Are a poll's sponsor/pollster and its outcome independent variables, or do some sponsors/pollsters have a higher probability of a certain outcome?

```
glimpse(covid_data)
```

```
## Rows: 3,302
## Columns: 13
## $ start_date  <date> 2020-02-02, 2020-02-02, 2020-02-02, 2020-02-02, 2020-02-0~
## $ end_date    <date> 2020-02-04, 2020-02-04, 2020-02-04, 2020-02-04, 2020-02-0~
## $ pollster    <chr> "YouGov", "YouGov", "YouGov", "YouGov", "Morning Consult",~
## $ sponsor     <chr> "Economist", "Economist", "Economist", "Economist", NA, NA~
## $ sample_size <dbl> 1500, 376, 523, 599, 2200, 684, 817, 700, 1996, 700, 788, ~
## $ population  <chr> "a", "a", "a", "a", "a", "a", "a", "a", "rv", "rv", "rv", ~
## $ party       <chr> "all", "R", "D", "I", "all", "R", "D", "I", "all", "R", "D~
## $ subject     <chr> "Trump", "Trump", "Trump", "Trump", "Trump", "Trump", "Tru~
## $ tracking    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA~
## $ text        <chr> "Do you approve or disapprove of Donald Trump's handling o~
## $ approve     <dbl> 42, 75, 21, 39, 57, 88, 37, 50, 39, 71, 15, 34, 39, 74, 19~
## $ disapprove  <dbl> 29, 6, 51, 25, 22, 4, 37, 22, 35, 8, 60, 33, 28, 7, 50, 25~
## $ url         <chr> "https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/doc~
```

Dataset 3

Data Set #3: MLB Elo Data from FiveThirtyEight

This data was sourced from MLB games dating back to 1871. FiveThirtyEight does not specify exactly where the game data is sourced from, but one can assume it was curated from an existing data set of games. It contains game-by-game forecasts and elo ratings calculated by FiveThirtyEight. Some interesting variables are the pitchers that started in the game, the pitchers rolling game score before the game, and the participating teams' elo ratings after the game. ELO ratings are used to rank teams based off of various metrics calculated by FiveThirtyEight.

Research Questions for Data Set #3: (1) Which starting pitchers contributed to the greatest increase in their team's elo rating on average? (2) How accurate has ELO win probability been over the years, and what factors could it use to be more accurate?

```
glimpse(mlb_data)
```

```
## Rows: 223,427
## Columns: 26
## $ date         <date> 2021-11-02, 2021-10-31, 2021-10-30, 2021-10-29, 2021-10-~
## $ season       <dbl> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 202~
## $ neutral      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ playoff      <chr> "w", "w", "w", "w", "w", "w", "l", "l", "l", "l", "l", "l~
## $ team1        <chr> "HOU", "ATL", "ATL", "ATL", "HOU", "HOU", "ATL", "HOU", "~
## $ team2        <chr> "ATL", "HOU", "HOU", "HOU", "ATL", "ATL", "LAD", "BOS", "~
## $ elo1_pre     <dbl> 1571.006, 1579.588, 1577.861, 1575.511, 1566.264, 1570.54~
## $ elo2_pre     <dbl> 1574.993, 1566.410, 1568.137, 1570.487, 1579.734, 1575.45~
## $ elo_prob1    <dbl> 0.5383260, 0.5708577, 0.5643511, 0.5554618, 0.5201953, 0.~
## $ elo_prob2    <dbl> 0.4616740, 0.4291423, 0.4356489, 0.4445382, 0.4798047, 0.~
## $ elo1_post    <dbl> 1565.052, 1574.993, 1579.588, 1577.861, 1570.487, 1566.26~
## $ elo2_post    <dbl> 1580.946, 1571.006, 1566.410, 1568.137, 1575.511, 1579.73~
## $ rating1_pre  <dbl> 1570.205, 1568.778, 1567.264, 1565.409, 1567.487, 1570.70~
## $ rating2_pre  <dbl> 1565.920, 1567.346, 1568.860, 1570.716, 1568.638, 1565.42~
## $ pitcher1     <chr> "Luis Garcia", "Tucker Davidson", "Dylan Lee", "Ian Ander~
## $ pitcher2     <chr> "Max Fried", "Framber Valdez", "Zack Greinke", "Luis Garc~
## $ pitcher1_rgs <dbl> 52.57546, 47.23319, 47.40000, 54.11173, 53.01470, 53.9926~
## $ pitcher2_rgs <dbl> 57.34148, 52.27263, 49.68246, 52.48326, 58.05077, 58.3929~
## $ pitcher1_adj <dbl> 2.7027202, -29.9992453, -29.7782514, 2.4882531, 4.3711403~
## $ pitcher2_adj <dbl> 18.32132724, 0.07780625, -11.64291017, 1.62374141, 20.540~
## $ rating_prob1 <dbl> 0.5242841, 0.4910878, 0.5081904, 0.5374569, 0.5128144, 0.~
## $ rating_prob2 <dbl> 0.4757159, 0.5089122, 0.4918096, 0.4625431, 0.4871856, 0.~
## $ rating1_post <dbl> 1565.883, 1565.920, 1568.778, 1567.264, 1570.716, 1567.48~
## $ rating2_post <dbl> 1570.242, 1570.205, 1567.346, 1568.860, 1565.409, 1568.63~
## $ score1       <dbl> 0, 5, 3, 2, 7, 2, 4, 5, 11, 2, 1, 2, 6, 12, 5, 3, 5, 5, 1~
## $ score2       <dbl> 7, 9, 2, 0, 2, 6, 2, 0, 2, 9, 9, 9, 5, 3, 4, 2, 9, 4, 2, ~
```