Project Proposal

due October 15, 2021 by 11:59 PM

Sonya Patel, Aimi Wen, Ethan, and Luca

10/8/21

Load Packages

```
library(tidyverse)
```

Load Data

```
movies <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202
raw_bechdel <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da
```

```
summary_survivor <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/maste

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

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

viewers <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/20

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

```
billboard <-
  readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-0
audio_features <-
  readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-0
```

Introduction and Data

Bechdel Test

The Bechdel Test coined by Alison Bechdel states that if a movie can satisfy three criteria — there are at least two named women in the picture, they have a conversation with each other at some point, and that conversation isn't about a male character — then it passes "The Rule," whereby female characters are allocated a bare minimum of depth.

Using the movies dataset and possibly joining it with the raw_bechdel data, we can observe trends and patterns in gender bias across different movies. The source of the data is FiveThirtyEight. The raw_bechdel.csv includes data from 1970 - 2020, for ONLY bechdel testing, while the movies.csv includes IMDB scores, budget/gross revenue, and ratings but only from 1970 - 2013. The cases are the movies. The data collected comes from the site BechdelTest.com, which is operated by committed moviegoers who analyze films and

ascertain if they pass the Bechdel test. The financial information on these films comes from The-Numbers.com, a leading site for box office and budget data.

Survivor

Survivor is a show that features a group of contestants that are given tribes and left in an isolated location, where they must provide food, water, fire, and shelter for themselves. The contestants compete in challenges for rewards and immunity from elimination. The contestants are progressively eliminated from the game as they are voted out by their fellow-contestants until only one remains to be awarded the grand prize of $1,000,000.

This data was found on TidyTuesday, and the data package is originally from the survival package, which was created by Daniel Oehm. The data has information on 40 seasons and 596 episodes, which covers about 20 years of data (2000-2020). These datasets give information on each season,including tribe set-up, information on players (personality_type, age, hometown), the types of challenges, how the players voted for the finalists, location, and the number of viewers of the show at different intervals. Across all datasets, the cases are the seasons, making it possible to join several of these datasets together through the season variable.

Billboard 100

"The Billboard Hot 100 is the music industry standard record chart in the United States for songs, published weekly by Billboard magazine. Chart rankings are based on sales (physical and digital), radio play, and online streaming in the United States."

The data was found on TidyTuesday and is from Data.World with the original data points found on Billboard.com and Spotify. The cases are songs from the certain week(s) in which they appeared on the Billboard 100 chart. It includes every weekly Hot 100 singles chart from Billboard.com. Relevant variables from Billboard include song, performer, instance (ordinal for which appearance on the chart it is for the song), previous_week_position, and more. Relevant variables from Spotify include spotify_genre, spotify_track_duration, danceability (double 0-1; factoring in tempo, rhythm stability, beat strength), energy (double 0-1; perceptual measure of intensity and activity), key, acousticness (double 0-1), and valence (double 0-1; "musical positiveness"), and more. Both include a song_id that we anticipate will be used for joining.

Research Questions

Bechdel Test

What attributes and variables correlate to movies that fail the Bechdel test?

I hypothesize that action movies and movies written by men would correlate with failing the Bechdel test.

What genre of movies fails the Bechdel Test most frequently?

I hypothesize that more action movies fail the Bechdel test because many of them have few female characters that interact with one another and are generally geared towards the male population.

Do movies that pass the Bechdel test have a lower budget than those that fail?

I hypothesize that movies with a lower budget pass the Bechdel test.

Survivor

How do different factors such as age, home state, personality type influence how well a player does in a particular season?

I hypothesize that younger people with more outgoing personality types and from home states that are famous for its nature will have a higher probability of winning.

How does the area of the world that the season was shot affect the popularity?

I hypothesize that seasons with more well-known "dangerous" areas of the world are be more popular. For example, I hypothesize that Australia is one of the most popular seasons.

Correlation between winning different immunity/reward challenges and how this affects the success of the tribe?

I hypothesize that the more wins a tribe has, the more successful the tribe would be.

Billboard 100

Which audio features best predict a songs success on the Billboard 100 charts (peak or duration)? Hypothesis: Danceability will be the best predictors of a song's duration, in weeks, on the Billboard 100 charts.

Does a song's number of appearances on the Billboard 100 charts make it more probable to reach a higher peak position?

Hypothesis: There will be no correlation between a song's number of appearances on the chart and it's peak position.

Which song keys and modes are the most popular on Billboard 100 charts?

Hypothesis: No key/mode combination will be significantly more popular than others on the Billboard 100 charts.

Glimpse

Please use `glimpse` for your datasets here.

```
glimpse(movies)
```

```
## Rows: 1,794
## Columns: 34
## $ year          <dbl> 2013, 2012, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 20~
## $ imdb          <chr> "tt1711425", "tt1343727", "tt2024544", "tt1272878", "tt0~
## $ title         <chr> "21 &amp;amp; Over", "Dredd 3D", "12 Years a Slave", "2 Guns~
## $ test          <chr> "notalk", "ok-disagree", "notalk-disagree", "notalk", "m~
## $ clean_test    <chr> "notalk", "ok", "notalk", "notalk", "men", "men", "notal~
## $ binary        <chr> "FAIL", "PASS", "FAIL", "FAIL", "FAIL", "FAIL", "FAIL", ~
## $ budget        <dbl> 1.30e+07, 4.50e+07, 2.00e+07, 6.10e+07, 4.00e+07, 2.25e+~
## $ domgross      <chr> "25682380", "13414714", "53107035", "75612460", "9502021~
## $ intgross      <chr> "42195766", "40868994", "158607035", "132493015", "95020~
## $ code          <chr> "2013FAIL", "2012PASS", "2013FAIL", "2013FAIL", "2013FAI~
## $ budget_2013   <dbl> 13000000, 45658735, 20000000, 61000000, 40000000, 225000~
## $ domgross_2013 <chr> "25682380", "13611086", "53107035", "75612460", "9502021~
## $ intgross_2013 <chr> "42195766", "41467257", "158607035", "132493015", "95020~
## $ period_code   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ decade_code   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ imdb_id       <chr> "1711425", "1343727", "2024544", "1272878", "0453562", "~
## $ plot          <chr> NA, NA, "In the antebellum United States, Solomon Northu~
## $ rated         <chr> NA, NA, "R", "R", "PG-13", "PG-13", "R", "R", "PG-13", "~
## $ response      <lgl> NA, NA, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, ~
## $ language      <chr> NA, NA, "English", "English, Spanish", "English", "Engli~
## $ country       <chr> NA, NA, "USA, UK", "USA", "USA", "USA", "USA", "UK", "US~
```

```
## $ writer        <chr> NA, NA, "John Ridley (screenplay), Solomon Northup (base~
## $ metascore     <dbl> NA, NA, 97, 55, 62, 29, 28, 55, 48, 33, 90, 58, 52, 78, ~
## $ imdb_rating   <dbl> NA, NA, 8.3, 6.8, 7.6, 6.6, 5.4, 7.8, 5.7, 5.0, 7.5, 7.4~
## $ director      <chr> NA, NA, "Steve McQueen", "Baltasar Kormákur", "Brian Hel~
## $ released      <chr> NA, NA, "08 Nov 2013", "02 Aug 2013", "12 Apr 2013", "25~
## $ actors        <chr> NA, NA, "Chiwetel Ejiofor, Dwight Henry, Dickie Gravois,~
## $ genre         <chr> NA, NA, "Biography, Drama, History", "Action, Comedy, Cr~
## $ awards        <chr> NA, NA, "Won 3 Oscars. Another 131 wins & 137 nomination~
## $ runtime       <chr> NA, NA, "134 min", "109 min", "128 min", "118 min", "98 ~
## $ type          <chr> NA, NA, "movie", "movie", "movie", "movie", "movie", "mo~
## $ poster        <chr> NA, NA, "http://ia.media-imdb.com/images/M/MV5BMjExMTEzO~
## $ imdb_votes    <dbl> NA, NA, 143446, 87301, 43608, 25735, 123837, 85871, 1897~
## $ error         <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

```
glimpse(raw_bechdel)
```

```
## Rows: 8,839
## Columns: 5
## $ year    <dbl> 1888, 1892, 1895, 1895, 1896, 1896, 1896, 1896, 1897, 1898, 18~
## $ id      <dbl> 8040, 5433, 6200, 5444, 5406, 5445, 6199, 4982, 9328, 4978, 54~
## $ imdb_id <chr> "0392728", "0000003", "0132134", "0000014", "0000131", "022334~
## $ title   <chr> "Roundhay Garden Scene", "Pauvre Pierrot", "The Execution of M~
## $ rating  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0,~
```

```
glimpse(castaways)
```

```
## Rows: 744
## Columns: 18
## $ season_name          <chr> "Survivor: Winners at War", "Survivor: Winners at~
## $ season               <dbl> 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 4~
## $ full_name            <chr> "Natalie Anderson", "Amber Mariano", "Danni Boatw~
## $ castaway             <chr> "Natalie", "Amber", "Danni", "Ethan", "Tyson", "R~
## $ age                  <dbl> 33, 40, 43, 45, 39, 43, 36, 44, 44, 35, 28, 39, 2~
## $ city                 <chr> "Edgewater", "Pensacola", "Shawnee", "Hillsboroug~
## $ state                <chr> "New Jersey", "Florida", "Kansas", "New Hampshire~
## $ personality_type     <chr> "ESTP", "ISFP", "ENFJ", "ISFP", "ESTP", "ESTJ", "~
## $ day                  <dbl> 2, 3, 6, 9, 11, 14, 16, 16, 18, 21, 23, 25, 28, 2~
## $ order                <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15~
## $ result               <chr> "1st voted out", "2nd voted out", "3rd voted out"~
## $ jury_status          <chr> NA, "1st jury member", "2nd jury member", "3rd ju~
## $ original_tribe       <chr> "Sele", "Dakal", "Sele", "Sele", "Dakal", "Sele",~
## $ swapped_tribe        <chr> NA, NA, NA, NA, NA, "Yara", "Sele", "Dakal", "Sel~
## $ swapped_tribe2       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ merged_tribe         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "Koru", "Koru~
## $ total_votes_received <dbl> 11, 6, 8, 4, 12, 4, 8, 2, 3, 14, 15, 12, 6, 9, 9,~
## $ immunity_idols_won   <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 2~
```

```
glimpse(challenges)
```

```
## Rows: 5,023
## Columns: 8
## $ season_name    <chr> "Survivor: Winners at War", "Survivor: Winners at War",~
## $ season         <dbl> 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40,~
## $ episode        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ title          <chr> "Greatest of the Greats", "Greatest of the Greats", "Gr~
## $ day            <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
```

```
## $ challenge_type <chr> "reward", "reward", "reward", "reward", "reward", "rewa~
## $ winners        <chr> "Amber", "Tyson", "Sandra", "Yul", "Wendell", "Sophie",~
## $ winning_tribe  <chr> "Dakal", "Dakal", "Dakal", "Dakal", "Dakal", "Dakal", "~
```

```
glimpse(jury_votes)
```

```
## Rows: 909
## Columns: 5
## $ season_name <chr> "Survivor: Winners at War", "Survivor: Winners at War", "S~
## $ season      <dbl> 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40~
## $ castaway    <chr> "Sarah", "Sarah", "Sarah", "Ben", "Ben", "Ben", "Denise", ~
## $ finalist    <chr> "Michele", "Natalie", "Tony", "Michele", "Natalie", "Tony"~
## $ vote        <dbl> 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0~
```

```
glimpse(summary)
```

```
## function (object, ...)
```

```
glimpse(viewers)
```

```
## Rows: 596
## Columns: 9
## $ season_name            <chr> "Survivor: Winners at War", "Survivor: Winners ~
## $ season                 <dbl> 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40,~
## $ episode_number_overall <dbl> 583, 584, 585, 586, 587, 588, 589, 590, 591, 59~
## $ episode                <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, ~
## $ title                  <chr> "Greatest of the Greats", "It's Like a Survivor~
## $ episode_date           <date> 2020-02-12, 2020-02-19, 2020-02-26, 2020-03-04~
## $ viewers                <dbl> 6.68, 7.16, 7.14, 7.08, 6.91, 7.83, 8.18, 8.23,~
## $ rating_18_49           <dbl> 1.3, 1.4, 1.4, 1.4, 1.4, 1.6, 1.7, 1.6, 1.5, 1.~
## $ share_18_49            <dbl> 7, 7, 7, 7, 6, 7, 8, 7, 6, 7, 6, 6, 5, 7, 6, 6,~
```

```
glimpse(billboard)
```

```
## Rows: 327,895
## Columns: 10
## $ url                   <chr> "http://www.billboard.com/charts/hot-100/1965-0~
## $ week_id               <chr> "7/17/1965", "7/24/1965", "7/31/1965", "8/7/196~
## $ week_position         <dbl> 34, 22, 14, 10, 8, 8, 14, 36, 97, 90, 97, 97, 9~
## $ song                  <chr> "Don't Just Stand There", "Don't Just Stand The~
## $ performer             <chr> "Patty Duke", "Patty Duke", "Patty Duke", "Patt~
## $ song_id               <chr> "Don't Just Stand TherePatty Duke", "Don't Just~
## $ instance              <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ previous_week_position <dbl> 45, 34, 22, 14, 10, 8, 8, 14, NA, 97, 90, 97, 9~
## $ peak_position         <dbl> 34, 22, 14, 10, 8, 8, 8, 8, 97, 90, 90, 90, 90,~
## $ weeks_on_chart        <dbl> 4, 5, 6, 7, 8, 9, 10, 11, 1, 2, 3, 4, 5, 6, 1, ~
```

```
glimpse(audio_features)
```

```
## Rows: 29,503
## Columns: 22
## $ song_id                  <chr> "-twistin'-White Silver SandsBill Black's Co~
## $ performer                <chr> "Bill Black's Combo", "Augie Rios", "Andy Wi~
## $ song                     <chr> "-twistin'-White Silver Sands", "¿Dònde Està~
## $ spotify_genre            <chr> "[]", "['novelty']", "['adult standards', 'b~
## $ spotify_track_id         <chr> NA, NA, "3tvqPPpXyIgKrm4PR9HCf0", "1fHHq3qHU~
## $ spotify_track_preview_url <chr> NA, NA, "https://p.scdn.co/mp3-preview/cef48~
## $ spotify_track_duration_ms <dbl> NA, NA, 166106, 172066, 211066, 208186, 2055~
```

```
## $ spotify_track_explicit    <lgl> NA, NA, FALSE, FALSE, FALSE, FALSE, TRUE, FA~
## $ spotify_track_album       <chr> NA, NA, "The Essential Andy Williams", "Comp~
## $ danceability              <dbl> NA, NA, 0.154, 0.588, 0.759, 0.613, NA, 0.64~
## $ energy                    <dbl> NA, NA, 0.185, 0.672, 0.699, 0.764, NA, 0.68~
## $ key                       <dbl> NA, NA, 5, 11, 0, 2, NA, 2, NA, NA, 7, NA, 1~
## $ loudness                  <dbl> NA, NA, -14.063, -17.278, -5.745, -6.509, NA~
## $ mode                      <dbl> NA, NA, 1, 0, 0, 1, NA, 0, NA, NA, 1, NA, 0,~
## $ speechiness               <dbl> NA, NA, 0.0315, 0.0361, 0.0307, 0.1360, NA, ~
## $ acousticness              <dbl> NA, NA, 0.91100, 0.00256, 0.20200, 0.05270, ~
## $ instrumentalness          <dbl> NA, NA, 2.67e-04, 7.45e-01, 1.31e-04, 0.00e+~
## $ liveness                  <dbl> NA, NA, 0.1120, 0.1450, 0.4430, 0.1970, NA, ~
## $ valence                   <dbl> NA, NA, 0.150, 0.801, 0.907, 0.417, NA, 0.95~
## $ tempo                     <dbl> NA, NA, 83.969, 121.962, 92.960, 160.015, NA~
## $ time_signature            <dbl> NA, NA, 4, 4, 4, 4, NA, 4, NA, NA, 4, NA, 4,~
## $ spotify_track_popularity  <dbl> NA, NA, 38, 11, 77, 73, 61, 40, NA, NA, 31, ~
```