Final Project Proposal

Abby Harris

12/3/2020

1. My Blog Link

My blog is available at https://abbyharris.netlify.app/

2. Spotify Songs

```r
library(here)
library(tidyverse)
library(ggplot2)
library(readxl)

dat1 <- read_csv(here::here("tidytuesday", "data", "2020", "2020-01-21", "spotify_songs.csv"))

glimpse(dat1)
```

```
## Observations: 32,833
## Variables: 23
## $ track_id                 <chr> "6f807x0ima9a1j3VPbc7VN", "0r7CVbZTWZgbTCYd...
## $ track_name               <chr> "I Don't Care (with Justin Bieber) - Loud L...
## $ track_artist             <chr> "Ed Sheeran", "Maroon 5", "Zara Larsson", "...
## $ track_popularity         <dbl> 66, 67, 70, 60, 69, 67, 62, 69, 68, 67, 58,...
## $ track_album_id           <chr> "2oCs0DGTsRO98Gh5ZSl2Cx", "63rPSO264uRjW1X5...
## $ track_album_name         <chr> "I Don't Care (with Justin Bieber) [Loud Lu...
## $ track_album_release_date <chr> "2019-06-14", "2019-12-13", "2019-07-05", "...
## $ playlist_name            <chr> "Pop Remix", "Pop Remix", "Pop Remix", "Pop...
## $ playlist_id              <chr> "37i9dQZF1DXcZDD7cfEKhW", "37i9dQZF1DXcZDD7...
## $ playlist_genre           <chr> "pop", "pop", "pop", "pop", "pop", "pop", "...
## $ playlist_subgenre        <chr> "dance pop", "dance pop", "dance pop", "dan...
## $ danceability             <dbl> 0.748, 0.726, 0.675, 0.718, 0.650, 0.675, 0...
## $ energy                   <dbl> 0.916, 0.815, 0.931, 0.930, 0.833, 0.919, 0...
## $ key                      <dbl> 6, 11, 1, 7, 1, 8, 5, 4, 8, 2, 6, 8, 1, 5, ...
## $ loudness                 <dbl> -2.634, -4.969, -3.432, -3.778, -4.672, -5....
## $ 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, 0.1...
## $ 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, 0.0...
## $ liveness                 <dbl> 0.0653, 0.3570, 0.1100, 0.2040, 0.0833, 0.1...
## $ valence                  <dbl> 0.518, 0.693, 0.613, 0.277, 0.725, 0.585, 0...
## $ tempo                    <dbl> 122.036, 99.972, 124.008, 121.956, 123.976,...
## $ duration_ms              <dbl> 194754, 162600, 176616, 169093, 189052, 163...
```

This data set comes from the `spotify_songs.csv` file on the Tidy Tuesday website. The data consists of 32,833 observations of 23 variables. The variables `track_id`, `track_album_id`, and `playlist_id` are all unique IDs for the track, album, and playlist, respectively. The variables `track_name`, `track_artist`, `track_album_id`, `track_album_name`, and `track_album_release_date` are all variables that give details on the track to help with identification. The variables `playlist_name`, `playlist_genre`, and `playlist_subgenre` give information on what type of track it is, in other words what type of playlists it is on. The variables, `danceability`, `energy`, `key`, `loudness`, `mode`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valence`, and `tempo` describe different characteristics of the track using a numeric scale. The last variable `duration_ms` gives the duration of the track in milliseconds.

- Question 1: Is there a positive correlation between the danceability and the popularity of the track?
  - I will explore this by creating a scatterplot of danceability and popularity of a track. I will also explore how other factors are grouped within the scatterplot using color or facets on different variables.

- Question 2: The valence of the track determines if it sounds happy or sad, does the valence change throughout the seasons of the year?
  - I will explore this by plotting the valence of a song vs. the release date of a song using a line plot to see if there is a difference in the numerical value of valence across the year. For instance, is valence higher (indicating a happier song) during summer months as opposed to songs released in the winter.
- Question 3: What is the distribution of speechiness?
  - I will explore this by creating a histogram of the speechiness value (indicates if a song is almost all words or mostly music) and how the distribution changes when it is broken down into other variables.
- Question 4: Does the duration of a track impact the popularity?
  - I will explore this by plotting the duration of the song against the popularity of the track variable to determine what legnth of track is ideal to give the highest possible popularity.

3. Tennis Grand Slams

```r
dat2 <- read_csv(here::here("tidytuesday", "data", "2019", "2019-04-09", "grand_slam_timeline.csv"))

glimpse(dat2)
```

```
## Observations: 12,605
## Variables: 5
## $ player     <chr> "Margaret Court", "Billie Jean Moffitt King", "Maria Buen...
## $ year       <dbl> 1968, 1968, 1968, 1968, 1968, 1968, 1968, 1968, 1968, 196...
## $ tournament <chr> "Australian Open", "Australian Open", "Australian Open", ...
## $ outcome    <chr> "Finalist", "Won", "Absent", "Absent", "Absent", "Semi-fi...
## $ gender     <chr> "Female", "Female", "Female", "Female", "Female", "Female...
```

```r
dat3 <- read_csv(here::here("tidytuesday", "data", "2019", "2019-04-09", "grand_slams.csv"))

glimpse(dat3)
```

```
## Observations: 416
## Variables: 6
## $ year             <dbl> 1968, 1968, 1968, 1968, 1969, 1969, 1969, 1969, 19...
## $ grand_slam       <chr> "australian_open", "french_open", "wimbledon", "us...
## $ name             <chr> "Billie Jean King", "Nancy Richey", "Billie Jean K...
## $ rolling_win_count <dbl> 1, 1, 2, 1, 1, 2, 1, 3, 4, 5, 6, 7, 8, 1, 2, 3, 2,...
## $ tournament_date  <date> 1968-01-10, 1968-06-09, 1968-07-14, 1968-09-09, 1...
## $ gender           <chr> "Female", "Female", "Female", "Female", "Female", ...
```

```
dat4 <- read_csv(here::here("tidytuesday", "data", "2019", "2019-04-09", "player_dob.csv"))

glimpse(dat4)
```

```
## Observations: 105
## Variables: 5
## $ name               <chr> "Nancy Richey", "Virginia Wade", "Billie Jean Ki...
## $ grand_slam         <chr> "French Open", "US Open", "Wimbledon", "Australi...
## $ date_of_birth      <date> 1942-08-23, 1945-07-10, 1943-11-22, 1942-07-16,...
## $ date_of_first_title <date> 1968-06-08, 1968-09-07, 1968-07-05, 1969-01-26,...
## $ age                <dbl> 9421, 8460, 8992, 9691, 7249, 7116, 7360, 10747,...
```

The first data set comes from the grand_slam_timeline.csv file on the Tidy Tuesday website. The data set has 12,605 observations of 5 variables. The variables are fairly self-explanatory, player is the name of the player, year is the year of the tournament, tournament gives the name of the tournament, outcome gives what position in the tournament the player finished, and gender gives the gender of the player. The second data set comes from the grand_slams.csv file on the Tidy Tuesday website. This data set has only 416 observations on 6 variables. This data set provides the same variables year and gender as the previous data set. Similarly, grand_slam provides the same information as tournament in the previous data set, and name provides the same information as player. This data set additionally provides the rolling_win_count which gives the total number of wins for the player at the time of that tournament, and tournament_date which gives the approximate date that the tournament took place. The third data set comes from the player_dob.csv file on the Tidy Tuesday website. This data set has 105 observations of 5 variables. The name variable is the same as in the previous data set. For this data set, grand_slam indicates what tournament the player was playing at when they won their first grand slam. date_of_first_title gives the date at which this occurs. date_of_birth gives the date of birth of the player, allowing age to be calculated by finding the difference in number of days between the date_of_first_title and date_of_birth to give the age at the time of first grand slam championship.

- Question 1: What grand slam is most common to win a first title at?
  - I will explore this by using the third data set to create a bar chart displaying how many players won their first grand slam at a given tournament. This could also be explored with facets such as age and gender.
- Question 2: Do players that are younger at the time of first grand slam win more tournaments overall?
  - I will analyze this by plotting the age at first win vs. number of rolling wins at the end of the data set. This could also be analyzed broken down into gender categories.
- Question 3: How does the age of first title won differ between males and females?
  - This can be analyzed by creating a histogram of the age at which the first title is won and faceting by gender.
- Question 4: What tournament was most frequently missed?
  - This can be found by creating a bar chart of the tournaments and the counts of the number of players that were absent from the tournament. Additionally this could be split into male and female categories.