Final Project Draft

##Instructions/Notes to Group (Delete this after) I made some changes becaues as I was writing, I realized some of the things that we talked about just aren't feasible. For example, we can't really predict which songs will end up #1 with our knowledge. So to make things simpler, I adapted some of the things we talked about to predict instance (a variable that would indicate longevity). I recommend double checking the meaning of instance. I also recommend reading through this carefully and see if this makes sense, considering I'm only one person, it's past 1am and I'm sick. No idea if my brain is working.

Since we're supposed to have one main research question, I took the liberty of adjusting things and writing the "one" question.

I think our length is ok? It's currently around 3.5 pages without code, visualizations, or discussion. But feel free to add other things we want to explore.

Introduction

"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 name in the variable "song" that we'll use for joining.

From this dataset, we're interested in examining songs from the 2000s and songs that made it to the #1 song position on the Billboard 100. Since there are so many song attributes in the dataset, we decided to focus on attributes that are easily identifiable through hearing by the general population: genre, danceability, energy, speechiness, valence, and tempo. Our main research question is: what factors play a role in a song becoming #1 on the Billboard 100 and the longevity of songs on the Billboard 100 after peaking at #1?

Someone insert the description of each variable that we'll use here, like how Professor Smith does at the top of every lab. Danceability- means this, the values from this to that, and the values mean this (i.e. the higher the value, the more danceable it is)

Methodology

```
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  3.1.5     v dplyr   1.0.7
```

```
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

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

Setting Up Our Project

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

```
## Rows: 327895 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): url, week_id, song, performer, song_id
## dbl (5): week_position, instance, previous_week_position, peak_position, wee...
```

```
##
## 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.
```

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

```
## Rows: 29503 Columns: 22
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): song_id, performer, song, spotify_genre, spotify_track_id, spotify...
## dbl (14): spotify_track_duration_ms, danceability, energy, key, loudness, mo...
## lgl  (1): spotify_track_explicit
```

```
##
## 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.
```

To conduct our work, we first need to combine our datasets, which is slightly complicated. For the billboard dataset, each entry is not a song, but rather the position of a song for each week. For example, a song is on the Billboard Top 100 for multiple weeks, it would be entered multiple times, each time with it's week's ranking. In contrast, each of the audio_feature's rows is a song. Songs are not entered in multiple times. To be able to combine the datasets together, we have to condense the Billboard 100 to having each row represent a song. Since we only care about #1 songs and songs from the 2000s, we *insert the process*

```
#this is to create the dataset that we'll end up using

billboard_top <- billboard %>%
  filter(peak_position == 1) %>%
  filter(week_position==1)

#could possibly use the variable of instance? Instance tells you the number of times that a song has ap

#this doesn't work because some songs are #1 for multiple weeks, so need to clean to have it appear onc
```

Attributes and Time

##By Year

To examine how attributes in #1 Billboard songs have changed over time, we created visualizations to show the distribution of values for each numerical factor that we're interested in: danceability, energy, speechiness, valence, and tempo. As mentioned in our introduction, in Billboard #1 songs, we predict that speechiness and tempo have decreased and increased over the 2000s due to the rising popularity of rap. We also predict that there are no significant patterns for danceability, valence, and energy; we think that the values for these factors will remain relatively constant. To keep our visualization readable, we decided that instead of keeping the time format that the current dataset has (M/D/Y), we will group by year.

```
#time in the dataset is in the form of dates, such as 7/17/2001. We need to figure out how to only get
```

```
#probably create one graph with each factor in a different color
```

Insert discussion about the graph

By Season

Besides looking by year, there are other time components that would influence what attributes are preferred over others. Seasons can influence what song attributes are preferred. For example, winter has many holidays, so that could mean that happier songs with medium tempos are preferred. To further examine attributes and #1 songs, we separated the dataset by seasons and looked at the distribution of attributes' values. We defined fall as from September to November, winter as from December to February, spring as March to May, and summer as June to August.

Genre and Longtivity Billboard #1 Songs

If this happens: Since it seems like *insert factor or factors* don't really follow a pattern, we decided to remove *insert factor or factors* in the below analyses.

Testing Pop Genre

To examine the correlation between genre and Billboard rankings, we created a null and alternative hypothees to test with simulation-based methods. Since the labels "pop genre" and "popular genre"tend to be interchangeable, we predict that the majority of songs that are #1 on the Billboard are of the pop genre.

Null Hypothesis: 50% of the #1 Billboard songs are of the pop genre. Alternative Hypothesis: Over 50% of the #1 Billboard songs are of the pop genre.

$$H_0 : p_{pop} = 0.5 \text{v.s.} H_a : p_{pop} > 0.5$$

"'{ visualizing-null-p-value}

"' *Insert Discussion*

To verify our conclusion from above, we also decided to perform a 98% Confidence Interval Test.

Insert Discussion

Testing Predictors

*NEED TO DISCUSS AS A GROUP BECAUSE IF OUR WHOLE DATASET IS SONGS THAT ARE #1, WE CAN WE DO A LINEAR REGRESSION TEST? Wouldn't linear regression not work if we're trying to find which attributes predict a song being #1 if the whole dataset is songs that are #1? I'm going to change it so the response variable is instance, thus we're examining the longetivity of songs being on the chart.

To further examine the relationship between attributes and songs on the Billboard, we decided to use linear regression. We want to see what attribute is the strongest predictor of a song remaining on the Billboard 100 after peaking at #1. We predict that energy is the strongest predictor for longevity.

To test this, we created a linear regression model with the explanatory variable as an attribute and the response variable as the instance variable. Instance in our dataset is defined as the number of times a song has appeared on the chart. We then found the R-squared value to see which attribute is the strongest predictor.

Insert Discussion

Probability of Songs Becoming #1

Now, that we've found which attribute/attributes is/are the strongest predictors, we want to delve deeper into these attributes' relationship with song's longevity. We first found the average value of these attributes. Given that the song is #1 and the attribute is at its average value, what is the probability that the song will remain on the billboard for over a month?

Things We Want Feedback on from Peer Review

Is our research question clear?

What do you think of the tests and visualizations that we're planning on performing?

How do you feel about our organization? Do you have suggestions on how we should be organizing our things?

Are there other tests/hypotheseses we should look at?