Spotify Music Project

RiSK: Rhea Tejwani, Sue Zhang, Keena Gao

10/25/20

Your written report goes here! Before you submit, make sure your code chunks are turned off with `echo = FALSE` and there are no warnings or messages with `warning = FALSE` and `message = FALSE`

Introduction and Data

Music is one of the most accessible ways to experience and communicate emotional experiences and opinions across cultural norms and language barriers. Popular music is especially able to broadcast its message, but popularity depends on the breadth of people that enjoy listening to it.

Trends in popular music are constantly changing, and these changes will affect globalization and cultural communication. For instance, hip-hop music, an aspect of the hip-hop cultural movement, was stated to be the most popular genre of music in the U.S. in 2017 in this Rolling Stone article: https://www.rollingstone.com/music/music-news/hip-hop-continued-to-dominate-the-music-business-in-2018-774422/. As students who frequently listen to music, we want to analyze the trends of popular music in our generation and the generations before us.

The data that we chose to analyze was curated by Sumat Singh (@iamsumat) on Kaggle, and contains variables that measure various characteristics of the most popular music in the world over the years 1956 to 2019 on the streaming service Spotify. https://www.kaggle.com/iamsumat/spotify-top-2000s-mega-dataset

The original data set was taken from the playlist on Spotify "Top 2000s" by the user PlaylistMachinery (@plamere) using Selenium with Python. It was scraped from http://sortyourmusic.playlistmachinery.com/. This data was uploaded 9 months ago.

Our research question: Has popular music shifted to be more diverse in characteristics such as Top Genre, Beats per Minute (BPM), Acousticness, Speechiness? The goal of our report is to observe how these characteristics changed over time and how these variables may affect one another.

This data set has 15 columns and 1994 rows. The observations in the data set describe the characteristics of the top 2000 most popular songs from 1956 to 2019 from Spotify.

Relevant Variables:

Title: title of the song

Artist: the musician/group who performed the song

Top genre: genre of the track year: year it was released

Beats per minute (BPM): tempo of the song

Energy: how energetic the song is

Danceability: how easy the song is to dance to

Loudness (dB): how loud the song is

Liveness: the likeliness of the song being a live recording

Valence: how positive a song is

Length (duration): the length of a track

Acousticness: how acoustic a song is

Speechiness: how much spoken word is in the song

Popularity: how popular a song is

Exploratory Data Analysis

There are no missing data values so we don't have to clean the data. We will create new variables using the mutate and case_when functions as follows: decade: the decade as named by "50s", "60s", etc period: "oldies" for before 1990 and "present" for after 1989 foreign_lang: categorical variable("Yes" for any song whose genre has a foreign country/language in it) broad genre: broad category "pop" or "rock" or etc

We will use dplyr functions to explore our data set in terms of summarising the variance and counting different variables and visualizing them. We are looking for unusual patterns or clusters of observations whose relationships we can further explore through hypothesis testing.

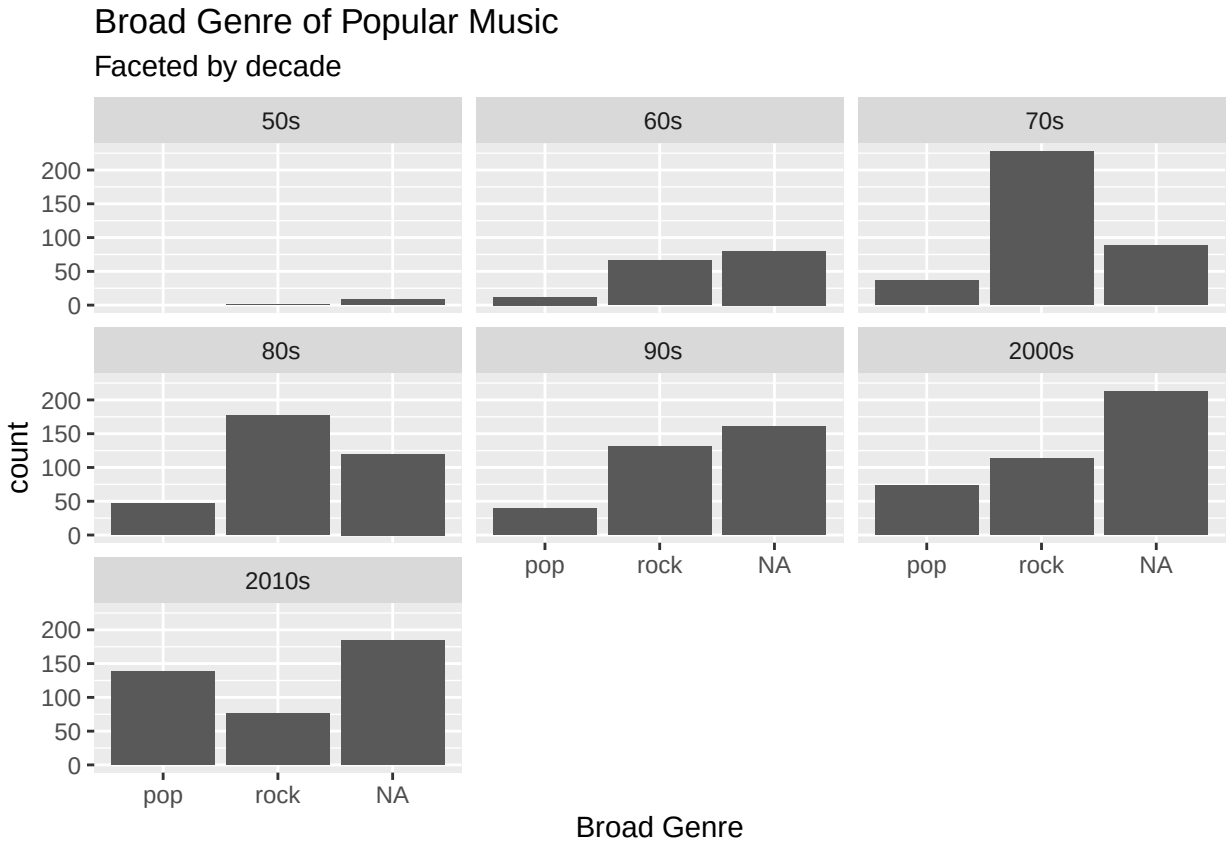
```
## # A tibble: 1 x 1
##   variance
##      <dbl>
## 1     812.
```

```
## # A tibble: 1 x 1
##   variance
##      <dbl>
## 1     849.
```

This increase in variance in acousticness suggests variation from the time period of the "Oldies" to the "Present" that may be due to error in measurement or other factors.

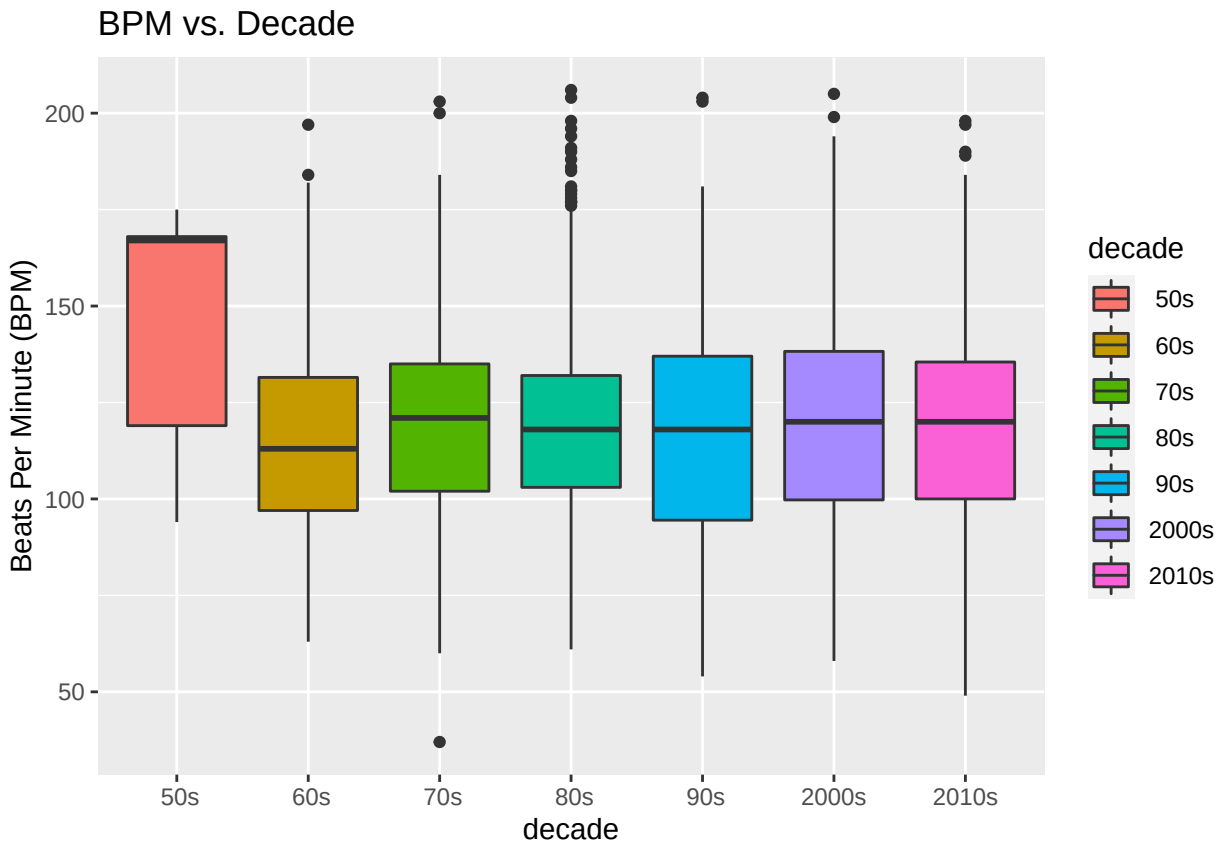
```
## # A tibble: 346 x 3
## # Groups:   decade [7]
##    decade  `Top Genre`           n
##    <chr>   <chr>             <int>
##  1 " 70s"  album rock          181
##  2 " 80s"  album rock           95
##  3 " 60s"  album rock           57
##  4 " 90s"  alternative rock     51
##  5 "2010s" dutch pop            44
##  6 " 70s"  adult standards      40
##  7 " 90s"  album rock           37
##  8 " 60s"  adult standards      34
##  9 "2010s" dutch indie          33
## 10 "2010s" dance pop            32
## # ... with 336 more rows
```

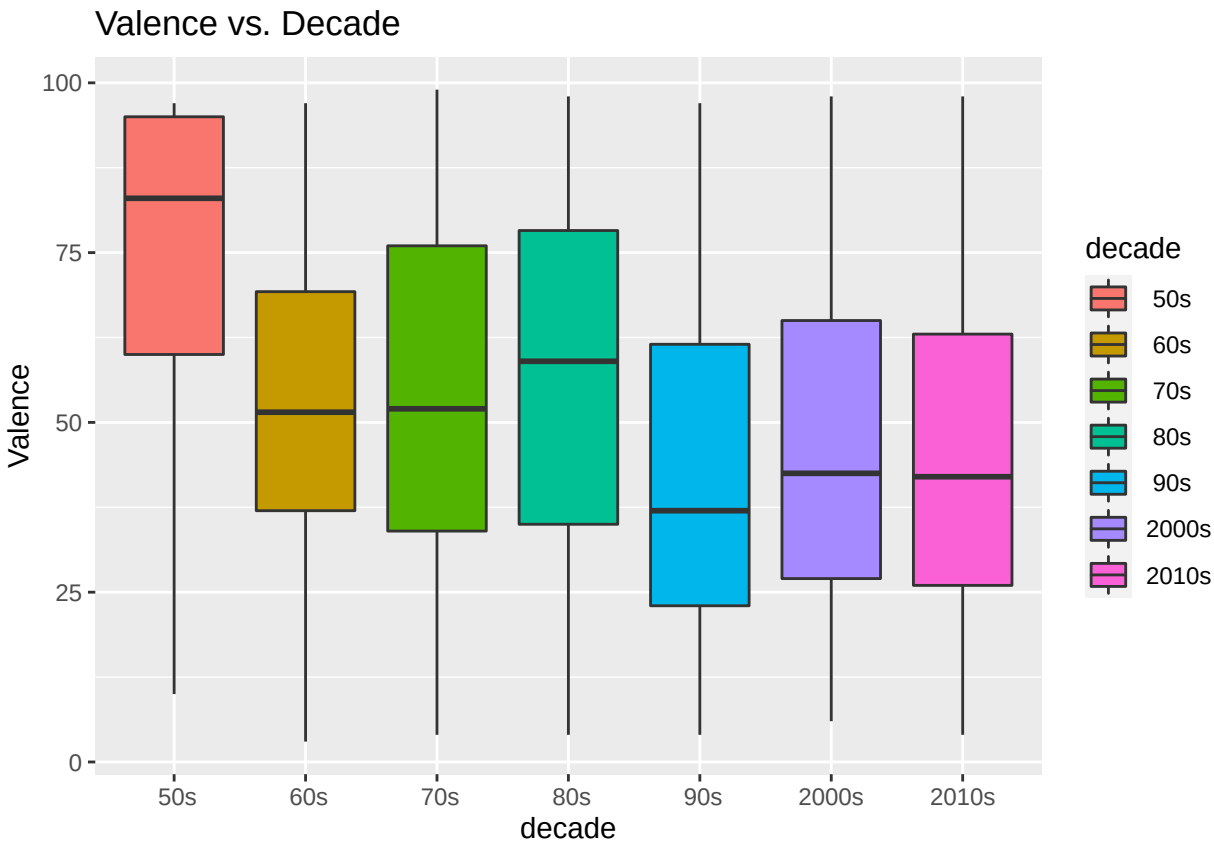
We wanted to ask what the most popular genre for each decade was. The most popular genre across these top 2000 songs by decade is album rock in the 70s. Album rock is also especially popular in the 80s and 60s.

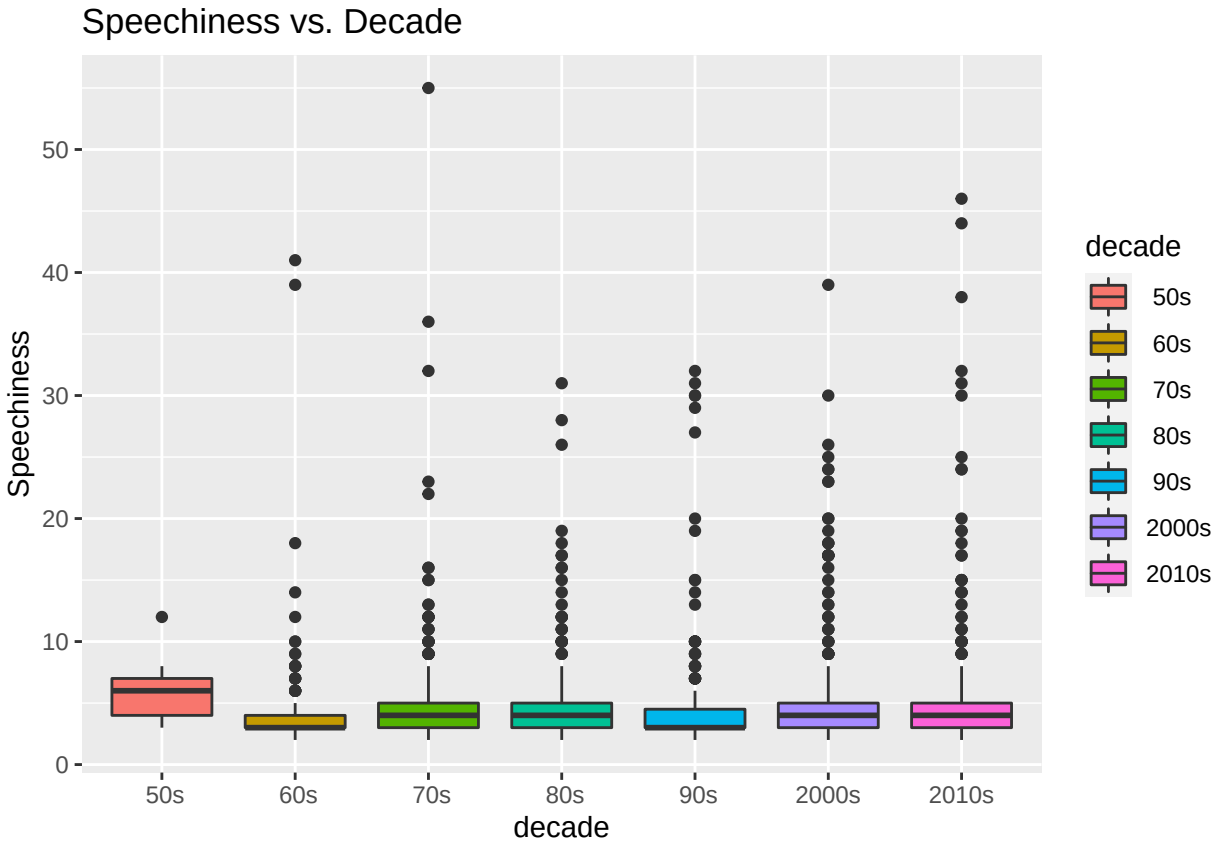


Methodology

We will look at the correlation, variance, standard deviations, and IQR of the variables for BPM, energy, valence, acousticness and speechiness. We will visualize our data using ggplot with scatterplots, boxplots and histograms. We will try to use linear models and the tools from library(broom). We can use summary statistics to find the mean, median and range of our data. We will also use the library(tidyverse) functions to explore our data set. We will find the p-value and use hypothesis tests to analyze the statistical significance of our tests. Some limitations of our analysis include other factors which are not variables or included in our modeling and the limited number of samples in our data set.


BPM vs. Decade
Beats Per Minute (BPM)
200
150
100
50
50s
60s
70s
80s
90s
2000s
2010s
decade
decade
50s
60s
70s
80s
90s
2000s
2010s


Valence vs. Decade
Valence
100
75
50
25
0
50s
60s
70s
80s
90s
2000s
2010s
decade
decade
50s
60s
70s
80s
90s
2000s
2010s



Since we are interested in discovering whether the level of acousticness has changed over time, these are our hypotheses:

H0: The true mean acousticness of hit music in the 90s, 2000s, and 2010s is equal to the true mean acousticness of hit music in the 50s, 60s, 70s, and 80s.

Ha: The true mean acousticness of hit music in the 90s, 2000s, and 2010s is less than the true mean acousticness of hit music in the 50s, 60s, 70s, and 80s.

Significance level: alpha = 0.05

```
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 -6.09
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

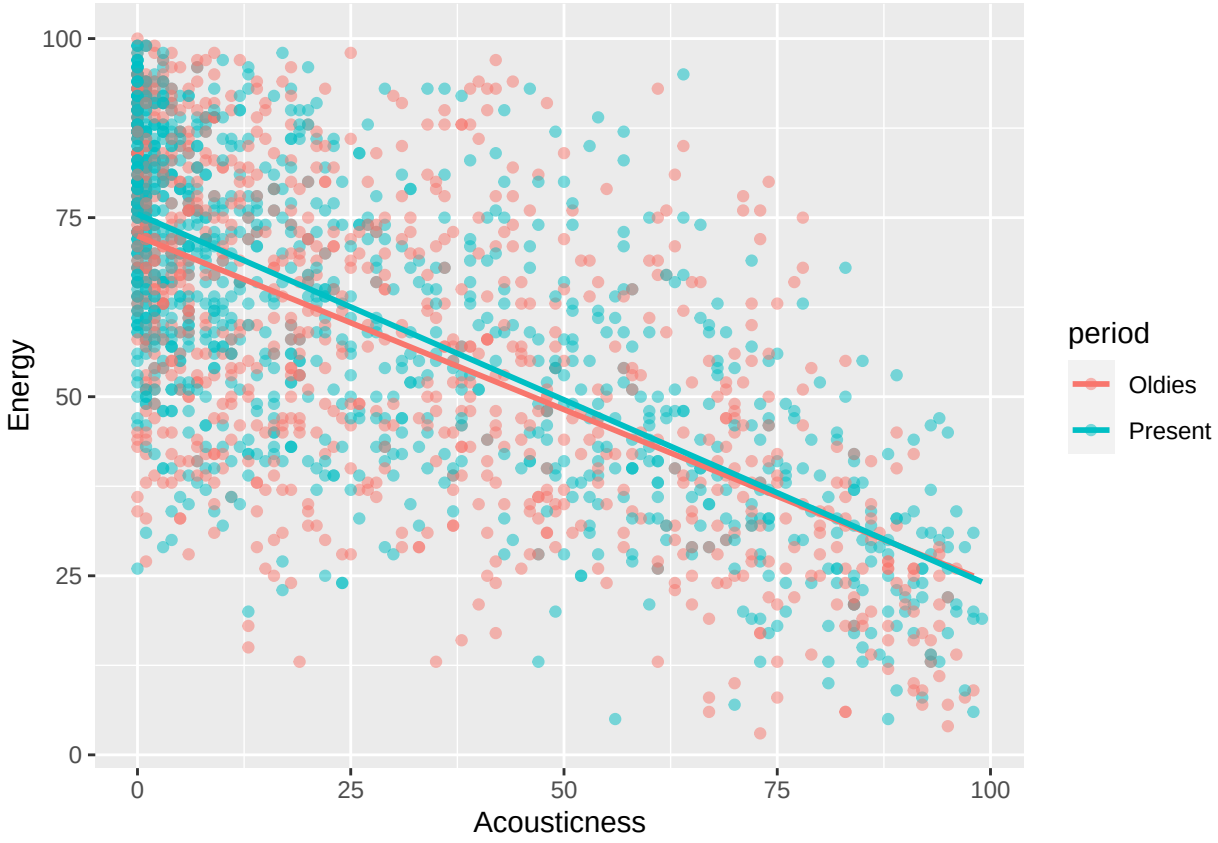
Our p-value is very little and smaller than our significance level of 0.05, so we will reject the null. We have sufficient evidence that the true mean acousticness of hit music in the 90s, 2000s, and 2010s is less than the true mean acousticness of hit music in the 50s, 60s, 70s, and 80s.

we are confused as to how to use simulation with our dataset, so we would appreciate help/advice on how to do this please! :)

Linear Regression of Energy vs. Acousticness:

```
## # A tibble: 2 x 2
##   term          estimate
##   <chr>            <dbl>
## 1 (Intercept)    74.3
## 2 Acousticness   -0.508
```

For each increase in one unit of Acousticness, the Energy is predicted to decrease by -0.508. If there is no units of Acousticness, the Energy is predicted to be 74.34.



Discussion

BPM appears to not have many discrepancies by decade. Valence appears to have decreased throughout time. The spread of speechiness appears to be increasing over time including the number of outliers, which indicates a more widespread acceptance of typically "speechy" genres in popular music, like hip hop.

Our summary statistics for variance show that the variance of acousticness is larger in the present era of music than it is in the past era of music.

No variables that we tested for in hit music have become less diverse over time, but only a few became more varied. These were acousticness, as demonstrated by our summary statistics for variance and our conclusion from our hypothesis test, and speechiness, as demonstrated by an increased spread over time by our box plots.

(Add paragraph about what we would do differently in the final draft)