

# Spotify Music Project

RiSK: Rhea Tejawani, 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** The dataset contains statistics of the most popular music in the world over the years on the streaming service Spotify. Music is constantly changing, and as students who frequently listen to music, we want to analyze the trends of popular music in our generation and the generations before us. Our research question is “how have characteristics of the most listened to music changed over time?” Our hypotheses are that popular music has shifted to be more diverse in these categories like acoustics, BPM, valence, etc. 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. 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 The original data set was taken from the playlist on spotify Top 2000s on PlaylistMachinery(@plamere) using Selenium with Python. It was scraped from <http://sortyourmusic.playlistmachinery.com/>.

## Methodology

We will look at the correlation, variance and IQR of the variables for beats per minute, 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.

## Data Analysis

Top Genres of songs across decades:

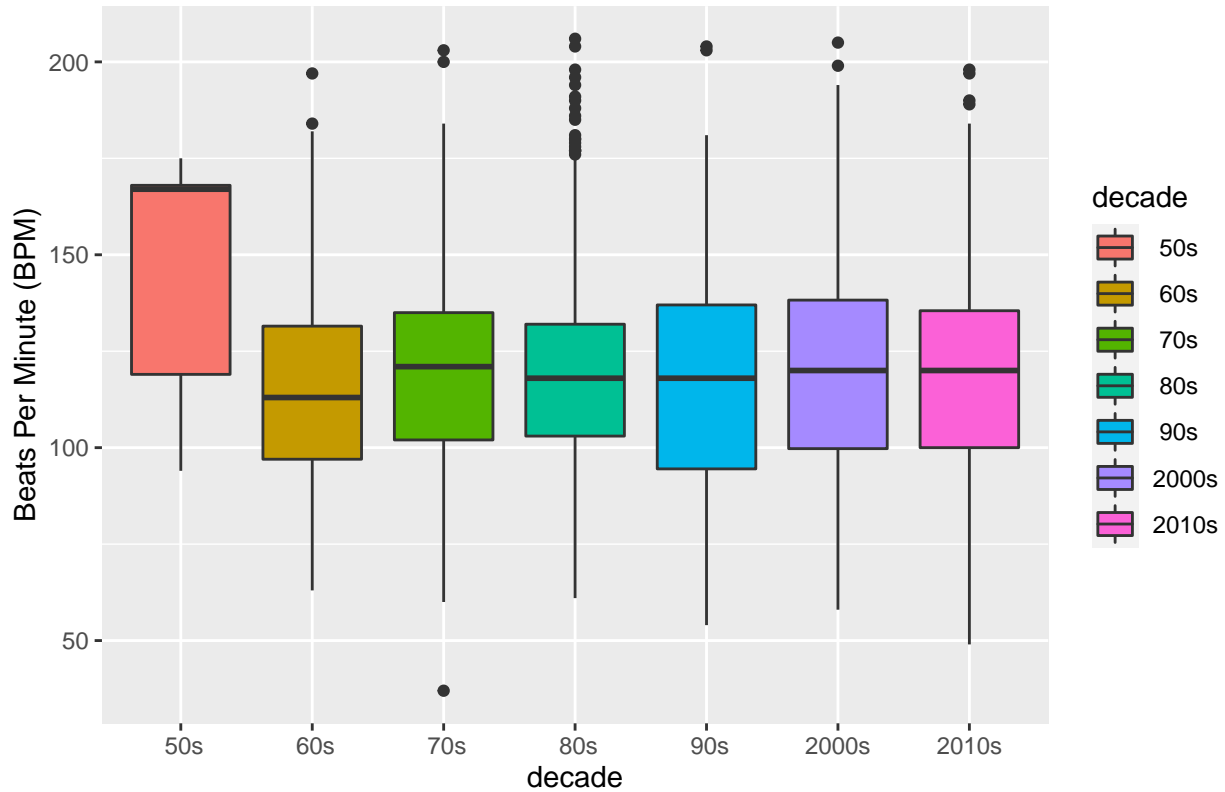
```
## # A tibble: 346 x 3
## # Groups:   decade [7]
##   decade `Top Genre`      n
##   <chr>   <chr>         <int>
## 1 " 70s" album rock      181
## 2 " 80s" album rock       95
## 3 " 60s" album rock       57
## 4 " 90s" alternative rock  51
## 5 "2010s" dutch pop       44
## 6 " 70s" adult standards  40
## 7 " 90s" album rock       37
```

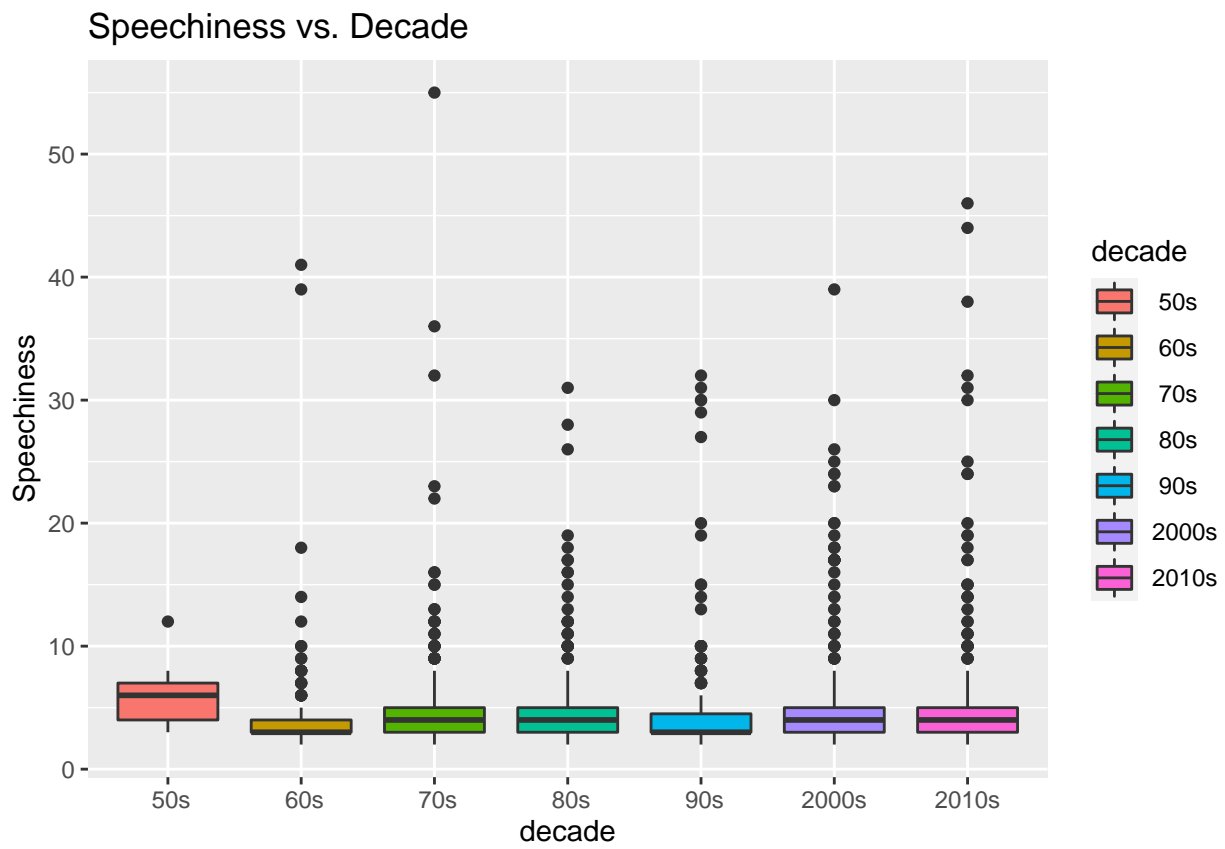
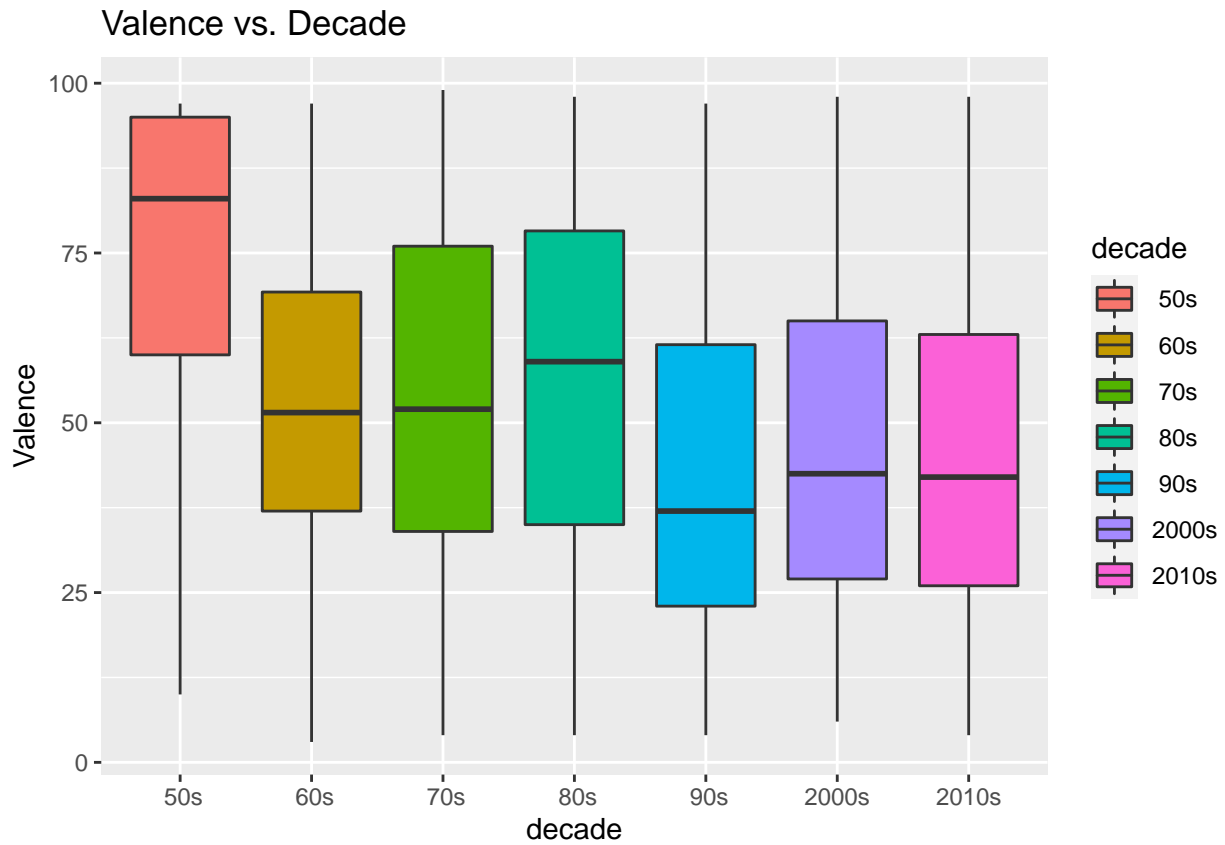
```
## 8 " 60s" adult standards      34
## 9 "2010s" dutch indie         33
## 10 "2010s" dance pop         32
## # ... with 336 more rows
```

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.

## Data Inference

BPM vs. Decade





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

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

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

## Conclusions

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, as well as the number of outliers.

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 speechiness, as demonstrated by an increased spread over time by our box plots.