Spotify Music Project

RiSK: Rhea Tejwani, Sue Zhang, Keena Gao

10/25/20

Questions: What can we improve Linear regression and interactive model graph

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.

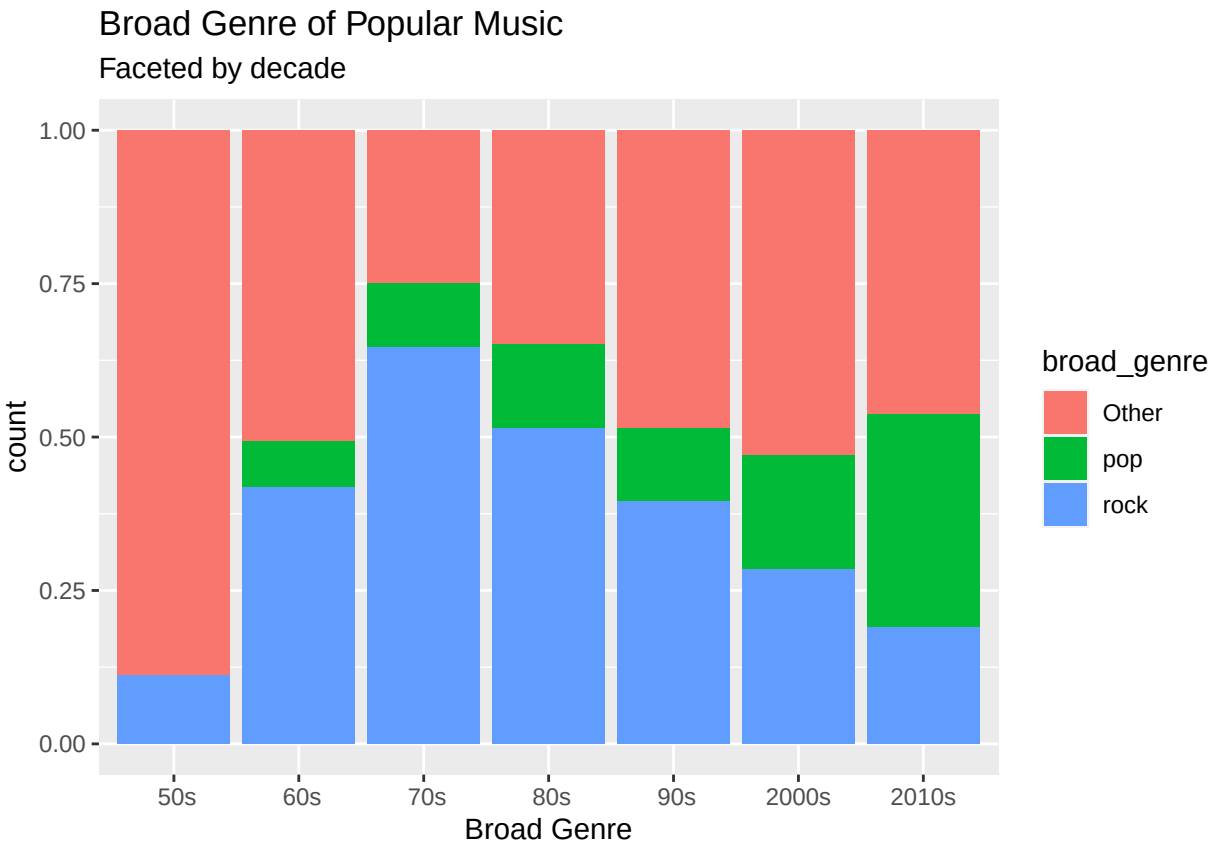
Our research question: Has popular music shifted to be more diverse in characteristics such as the genre and acousticness? The goal of our report is to observe how these characteristics changed over time and how these variables may affect one another.

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.

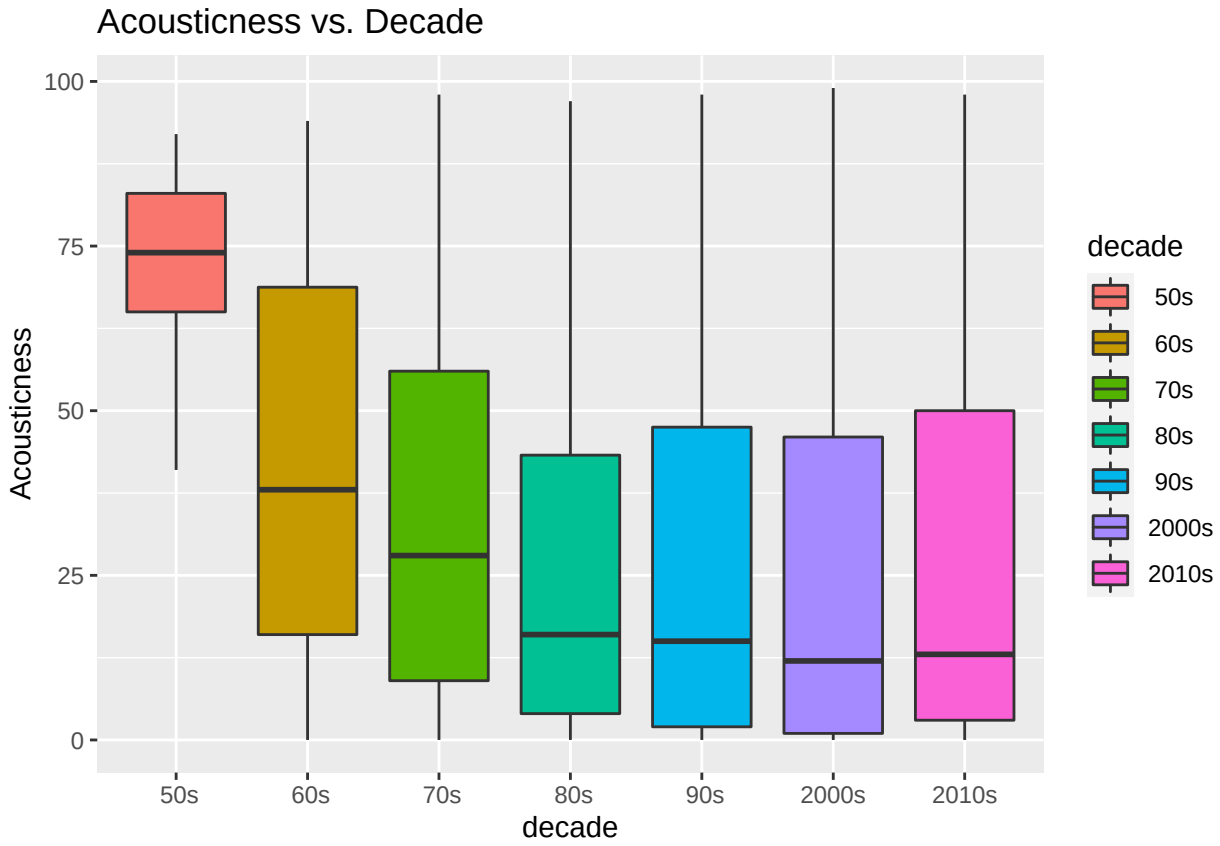
Exploratory Data Analysis

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("TRUE" for any song whose genre has a foreign country/language in it, "FALSE" otherwise) broad genre: broad category "pop" or "rock" or etc var: variance of Acousticness per year

We will use dplyr functions to explore our data set in terms of summarising the variance, 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.



The proportion of rock as a popular song genre was highest in the 70s and has decreased since then. The proportion of pop has increased slightly since the 60s. The dataset was started midway through the 50s so there aren't as many data points for genre as the other decades.



Older music tends to have a higher median value for acousticness. The median acousticness has remained relatively constant from the 80s to the 2010s. Additionally, from the 80s to the 2010s, the majority of the IQR overlaps. This could be due to an increase in technology used in music production in the modern era.

Methodology

We will look at the correlation, variance, standard deviations, and IQR of the acousticness variable. We will also analyze the changes in genre across the decade through simulation based tests. 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.

We will be using the variable `var`, which is the variance of the `Acousticness` of the songs in our dataset.

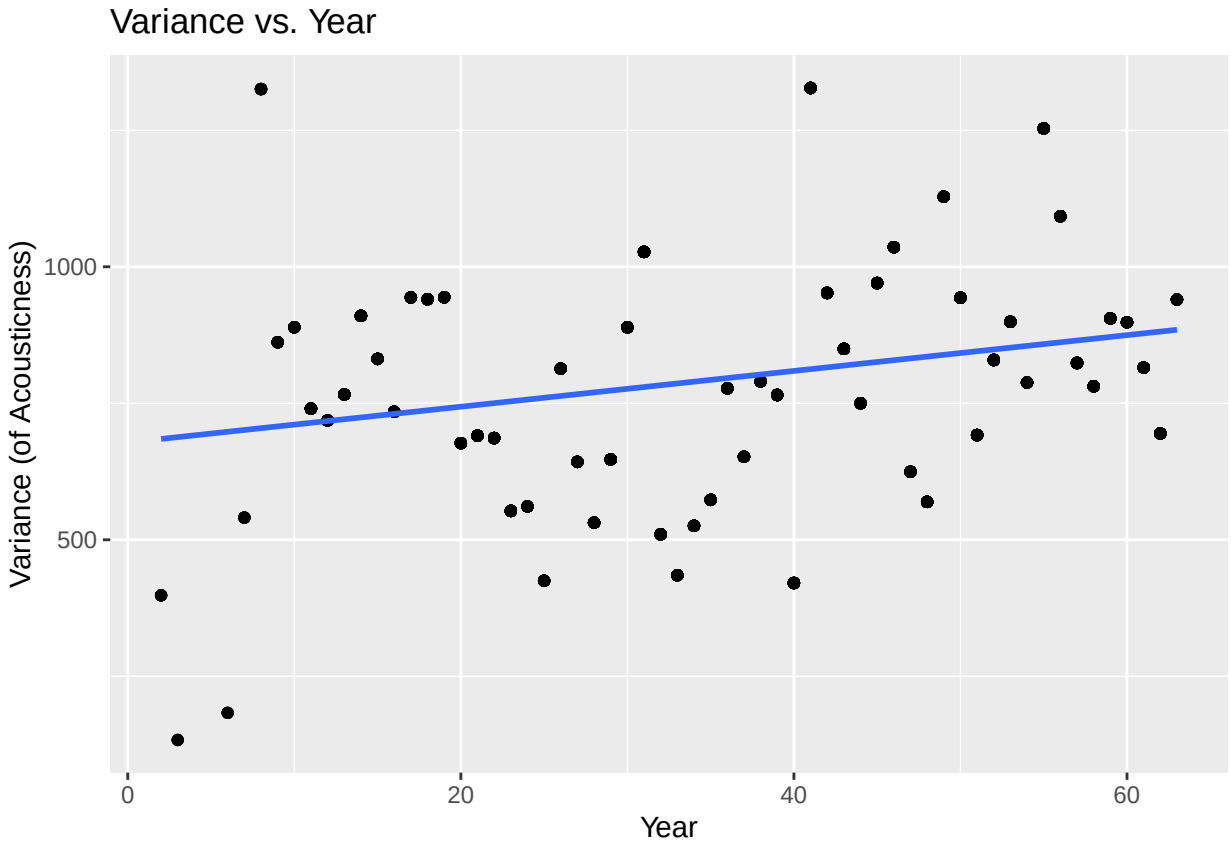
presentation answer why do I care for different models/graphs clear chronological order

###Linear model for acousticness

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   678.      10.8        62.6 0.
## 2 year            3.28     0.268      12.2 3.64e-33
```

The linear equation for the model is $\widehat{Acousticness} = 678.192 + 3.277 \times year$

intercept -when year is 0 extrapolation recode years(earliest as 0), latest this year - earliest year –> progression




```
## # A tibble: 4 x 5
##   term                    estimate std.error statistic  p.value
##   <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)               675.       11.3      60.0   0.
## 2 year                        3.36      0.290    11.6   3.45e-30
## 3 foreign_lang_factor        44.1      49.3       0.894 3.72e- 1
## 4 year:foreign_lang_factor   -0.978     1.02     -0.959 3.38e- 1
```

$\widehat{Acousticness} = 675.288 + 3.365 \times year - 44.060 \times foreign_lang_factor - 0.978 \times year : foreign_lang_factor$

R-squared values for the models:

```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1    0.0698        0.0694
```

```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1    0.0703        0.0689
```

**move this to conclusion

Our linear model where we predict the acousticness of a song with the number of years since 1956 is a better fit for the data because its adjusted r-squared value is greater than the adjusted r-squared value for the interaction model where we predict the acousticness of a song with the number of years since 1956 and whether the song is in a foreign language or not.

###Hypothesis test for genre (foreign language)

H0: The proportion of popular foreign language songs in the modern decades is equal to the proportion of popular foreign language songs in the older decades ($\rho_m = \rho_o$, where ρ_m is the proportion of popular foreign language songs in the modern decades and ρ_o the proportion of popular foreign language songs in the older decades)

Ha: The proportion of popular foreign language songs in the modern decades is greater than the proportion of popular foreign language songs in the older decades ($\rho_m > \rho_o$, where ρ_m is the proportion of popular foreign language songs in the modern decades and ρ_o the proportion of popular foreign language songs in the older decades)

Alpha: 0.05 ($\alpha = 0.05$)

```
## # A tibble: 1,000 x 2
##    replicate     stat
##        <int>    <dbl>
##  1         1 -0.0187
##  2         2  0.0283
##  3         3 -0.00438
##  4         4  0.00379
##  5         5  0.00788
##  6         6  0.00379
##  7         7  0.0222
##  8         8  0.00788
##  9         9 -0.0187
## 10        10  0.00788
## # ... with 990 more rows
```

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

The p-value is very small, which is less than the pre-determined alpha level of 0.05, so we reject the null hypothesis. There is sufficient evidence to conclude that the proportion of popular foreign language songs in the modern decades is greater than the proportion of popular foreign language songs in the older decades.

Discussion

LIMITATIONS: 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.

**say something about the 50s

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