# Spotify Music Project

RiSK: Rhea Tejwani, Sue Zhang, Keena Gao

10/25/20

**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.

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.

The variables we will be focusing on are `acousticness`, which is a measure from 0 to 100 representing the confidence measure that the song is acoustic (with 100 being high confidence and 0 being no/low confidence), `year`, which is the year the song was created, and `genre`, which is the genre of the music. We chose acousticness because we thought there trends observed in acousticness would reflect the development of technology used in music production in the modern era. Additionally, we focused on genres outside of english speaking countries because an increase in technology may also increase cross cultural communication including music.

We explored patterns and mutated our chosen variables in the methodology section of this report to better explore the change of the standard deviation of Acousticness and the prevalence of non-english languages in popular music over time.

Our research question: Does our data provide sufficient evidence that the proportion of songs in non english languages has increased from the time period of 1956-1989 to the time period of 1990 to present? Does our data provide sufficient evidence that the standard deviation in Acousticness has increased over time? The goal of our report is to observe how popular music has shifted to be more diverse in characteristics such as non english languages or Acousticness over time.

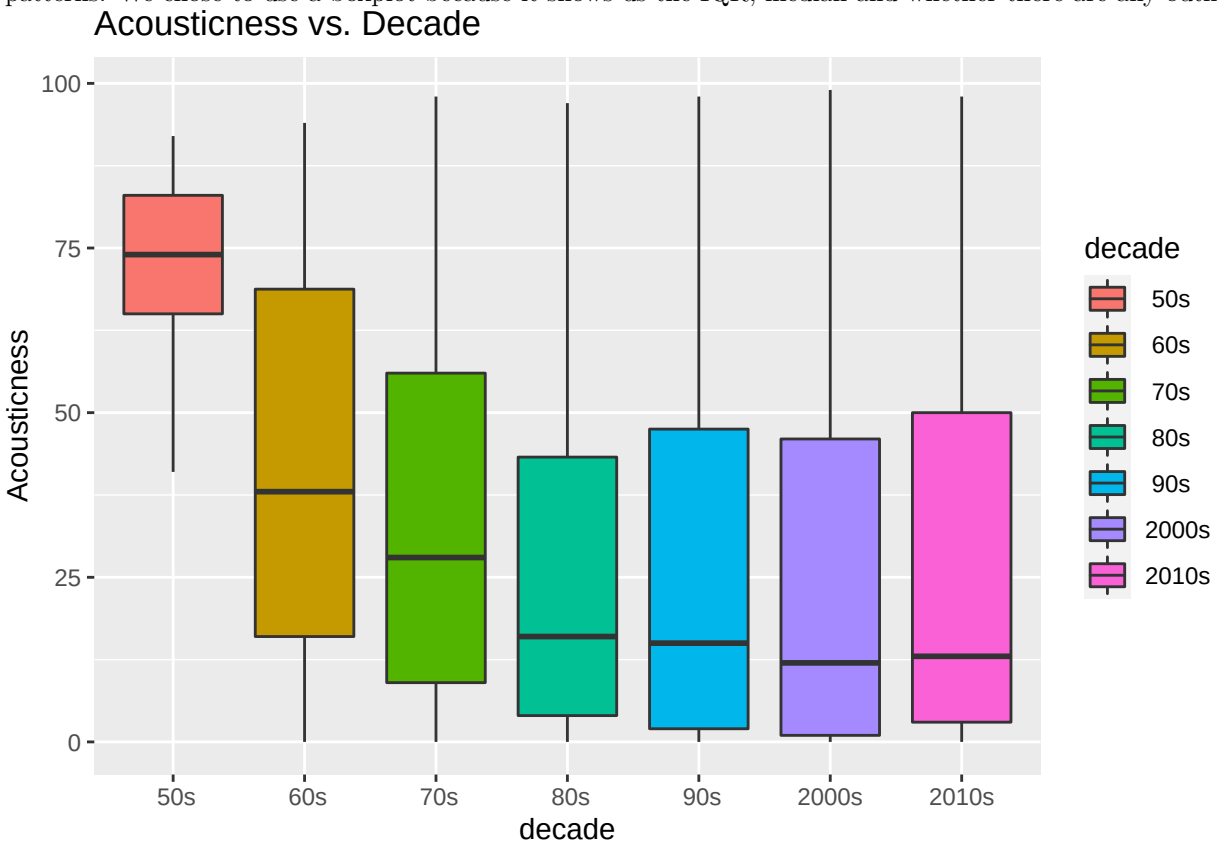
**Methodology**

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: categorical variable indicating the decade when the song was created using the categorical name of "50s", "60s", and so on period: categorical variable in which "Oldies" represents when Year is before 1990 and "Present" for when Year is after 1989 foreign_lang: categorical variable in which "TRUE" stands for any song whose Top Genre name has a non english speaking country in it, "FALSE" otherwise pop: categorical variable in which "TRUE" stands for any song whose Top Genre name has "pop" in it, "FALSE" otherwise sd: standard deviation of Acousticness per year year: numerical variable representing the years since the first Year in our data set(1956)

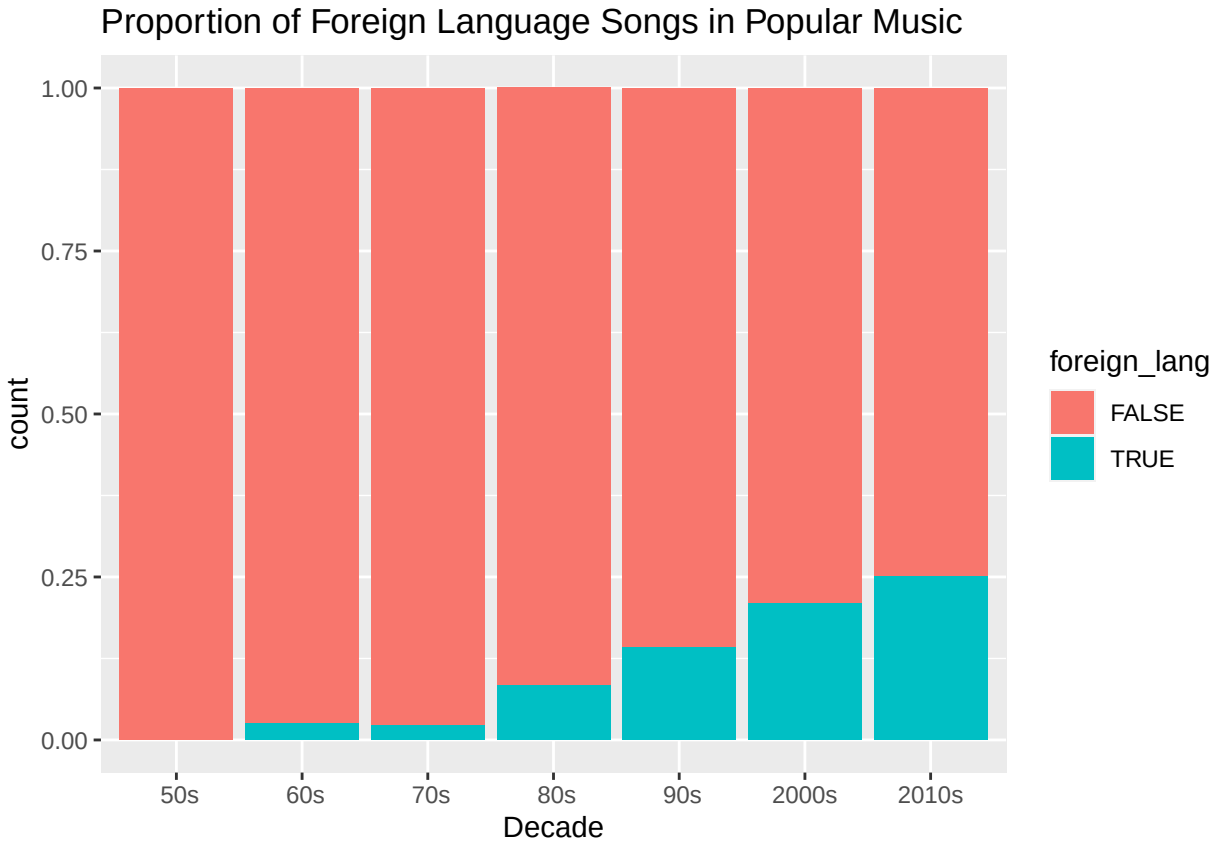
We will use dplyr functions to explore our data set in terms of summarising the variance, counting different variables and visualizing them. We are looking for unusual patterns or clusters of observations whose relationships we can further explore through hypothesis testing.

We wanted to look at the spread of Acousticness over the different decades to see if there were any identifiable patterns. We chose to use a boxplot because it shows us the IQR, median and whether there are any outliers.




Older music tends to have a higher median value for acousticness. The median acousticness has remained relatively constant from the 1990s to the 2010s and the majority of the interquartile range overlaps and is around the same size, which indicates they have similar spreads for Acousticness. There aren't any outliers in our dataset.

We wanted to look at the proportion of foreign language songs changing over different decades to see if there were any patterns we could investigate. We chose to use a bar plot since it effectively visualizes proportion.




The proportion of popular foreign language songs has increased over time. We noticed that the proportion of foreign language in the 50s is 0 or close to 0 and we think this is because we don't have as many observations in this decade since our data set starts at 1956.

We will further investigate the change in the number of foreign language songs in popular music over two time periods. We have decided to divide up the years into these time periods because the internet became widely available in the public in 1993, so people were able to discover and enjoy songs from other countries.

We used simulation-based hypothesis testing to see if the proportion of popular songs in non english languages increased from the "Oldies" period to the "Present" period. We want to be able to test our null and alternative hypotheses and generate a p-value so we can answer our research question about whether we have significant evidence to claim that the proportion of non english songs in popular music has increased from the "Oldies" time period to the "Present" time period.

We used linear and interaction effects modeling to examine how the standard deviation of Acousticness has changed over time. This will allow us to represent the relationship between year, standard deviation of Acousticness, and whether a song is in the "pop" genre with a function.

**Results**

We will look at the correlation, 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 `sd`, which is the standard deviation of the `Acousticness` of the songs in our dataset.

###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 $\rho_m$ is the proportion of popular foreign language songs in the modern decades and $\rho_o$ 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 $\rho_m$ is the proportion of popular foreign language songs in the modern decades and $\rho_o$ 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
```

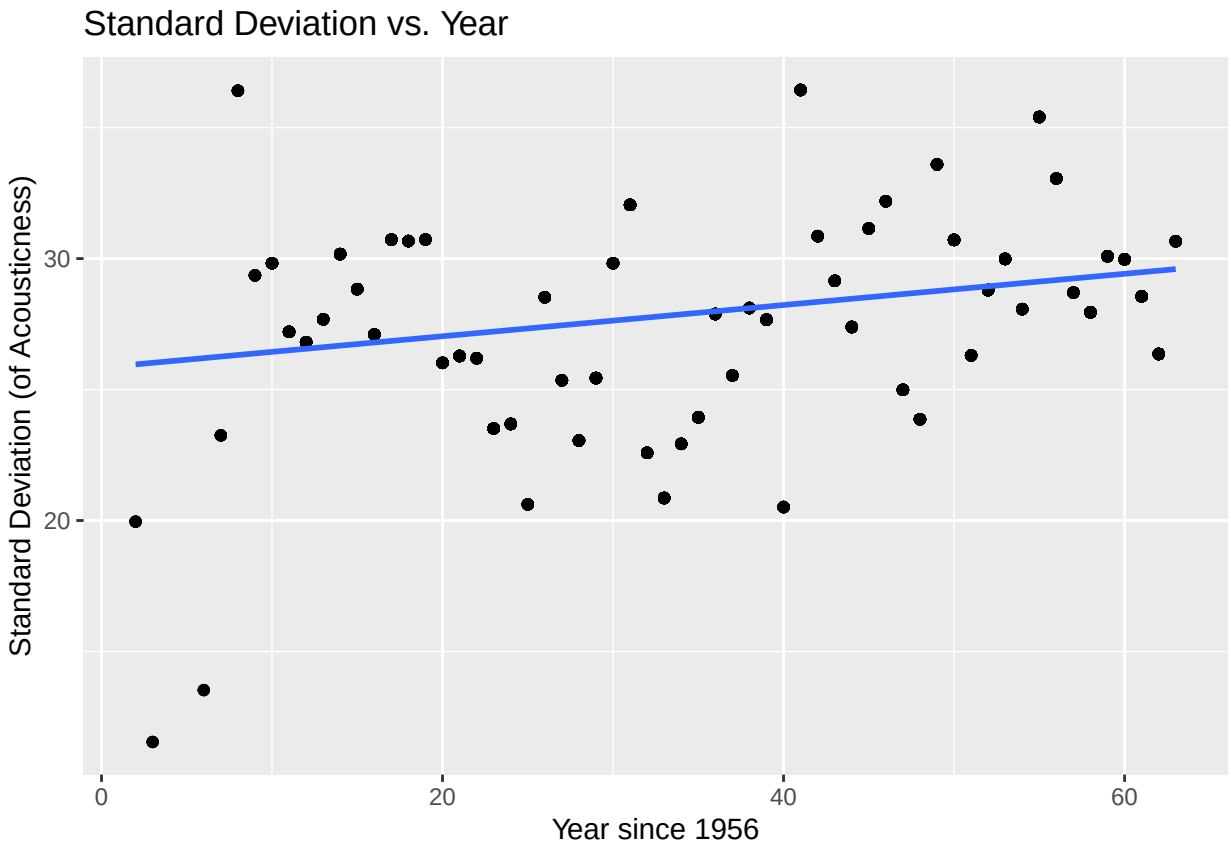
The p-value is very small, which is less than the pre-determined alpha level of 0.05, so we reject the null hypothesis. There is sufficient evidence to conclude that the proportion of popular foreign language songs in the modern decades is greater than the proportion of popular foreign language songs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

###Linear model for acousticness

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  25.8      0.194       133.  0.
## 2 year          0.0596   0.00480      12.4 3.90e-34
```

The linear equation for the model is $\widehat{Acousticness} = 25.842 + 0.0596 \times year$

The intercept of this graph is when the year is 1956, we expect the standard deviation for acousticness to be 25.842, holding all else constant. For each additional year, we expect the standard deviation of acousticness to increase by 0.0596, holding all else constant.

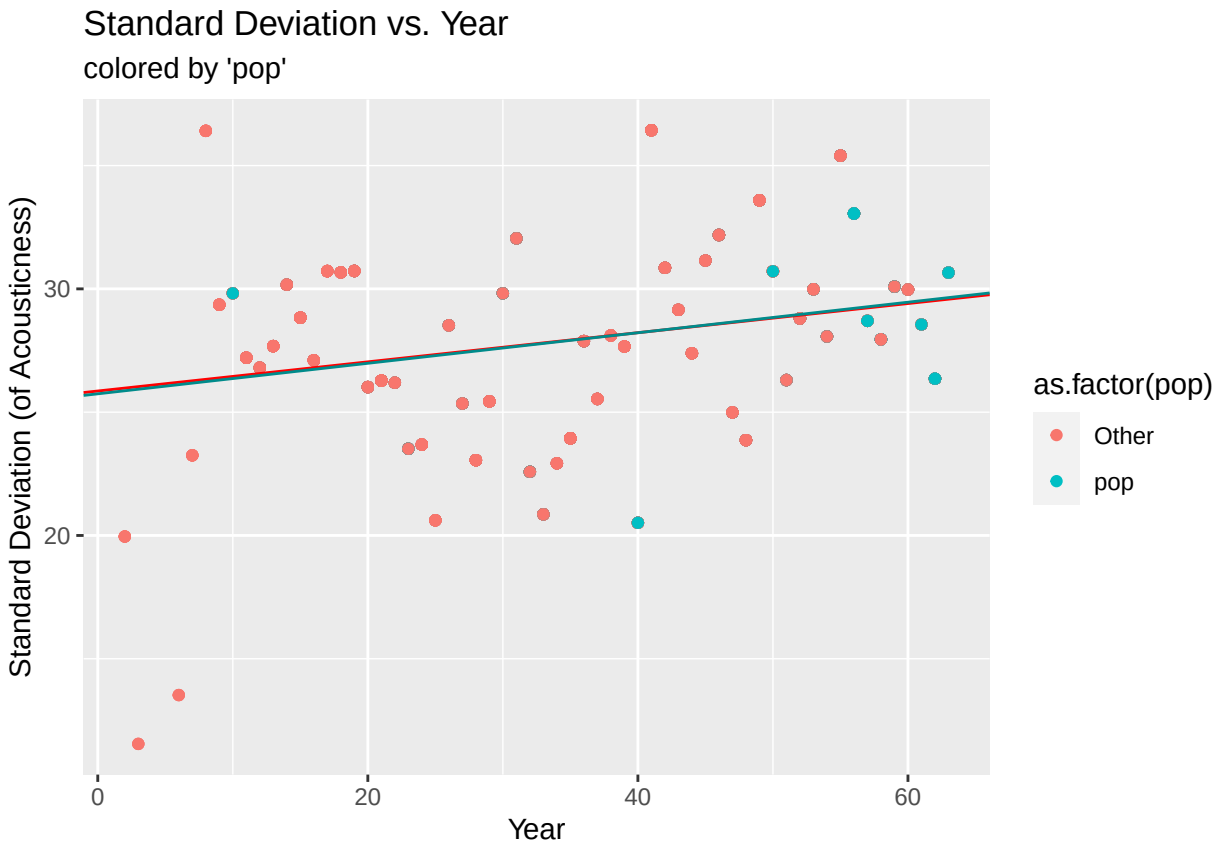



The standard deviation in the acousticness of popular songs has increased over time.

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 25.9       0.209     124.    0.
## 2 year         0.0591    0.00538    11.0   2.66e-27
## 3 poppop      -0.113     0.610      -0.186 8.53e- 1
## 4 year:poppop  0.00278   0.0133      0.209 8.35e- 1
```

$\widehat{Acousticness} = 25.858 + 0.059 \times year - 0.113 \times foreign_lang_factor - 0.003 \times year : foreign_lang_factor$

The intercept of this graph is when the year is 1956, we expect the standard deviation for acousticness to be 25.858, holding all else constant. For each additional year, we expect the standard deviation of acousticness to increase by 0.059, holding pop constant. We expect pop songs to have a standard deviation of acousticness that is 0.113 less than songs that are not pop, on average, holding year constant. **EXPLAIN INTERACTION**




R-squared values for the models:

```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1    0.0719        0.0714
```

```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1    0.0719        0.0705
```

**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)