RiSK: Analysis of the Changes in Popular Music in the World

RiSK: Rhea Tejwani, Sue Zhang, Keena Gao

November 15, 2020

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.

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 1,994 of the 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 that 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, such as the popularization of the Internet in the 1990s according to the World Wide Web Foundation, may also increase cross cultural communication including music.

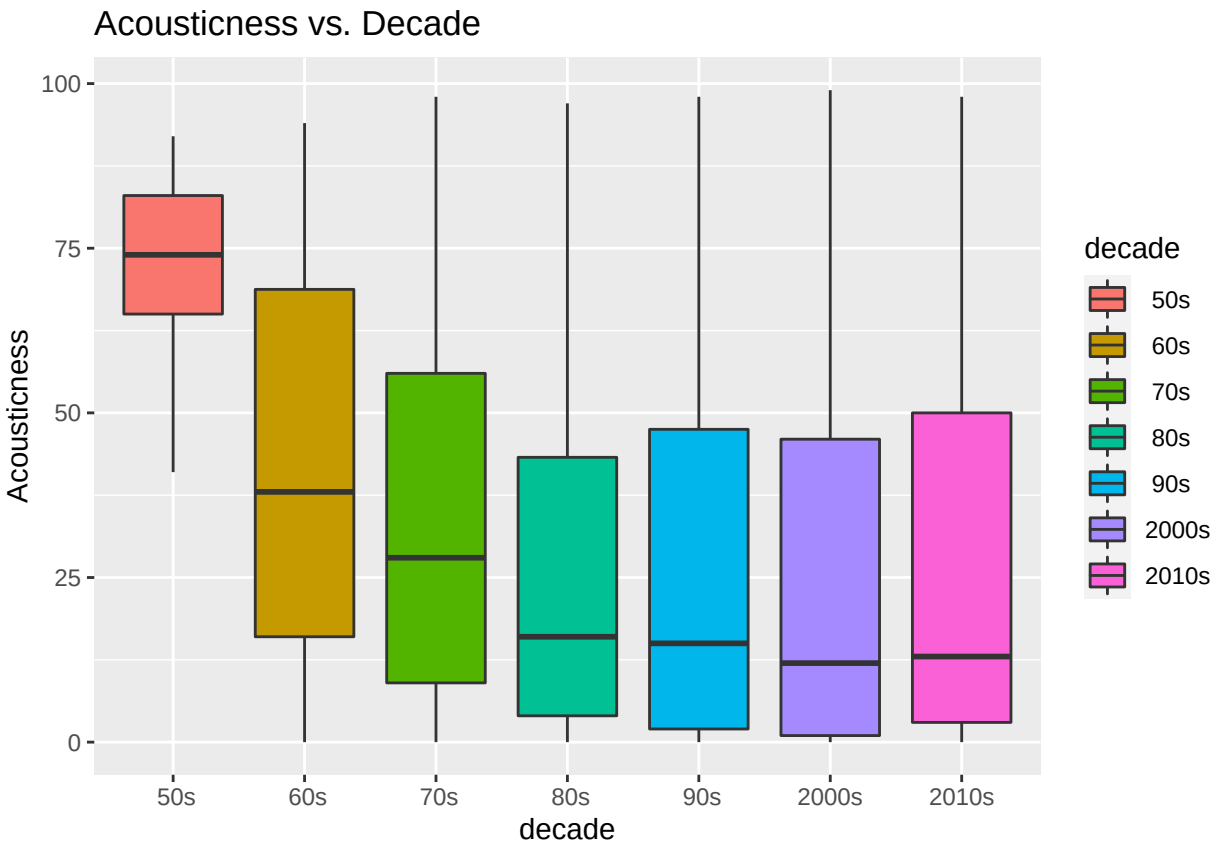
We explored patterns and mutated our chosen variables in the methodology section of this report to better explore the change of the standard deviation of Acousticness and the prevalence of non-English languages in popular music over time.

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 the characteristics of non-English genres or `Acousticness` over time.

Methodology

There are no missing values in the data set. We will create new variables using the mutate and case_when functions as follows:

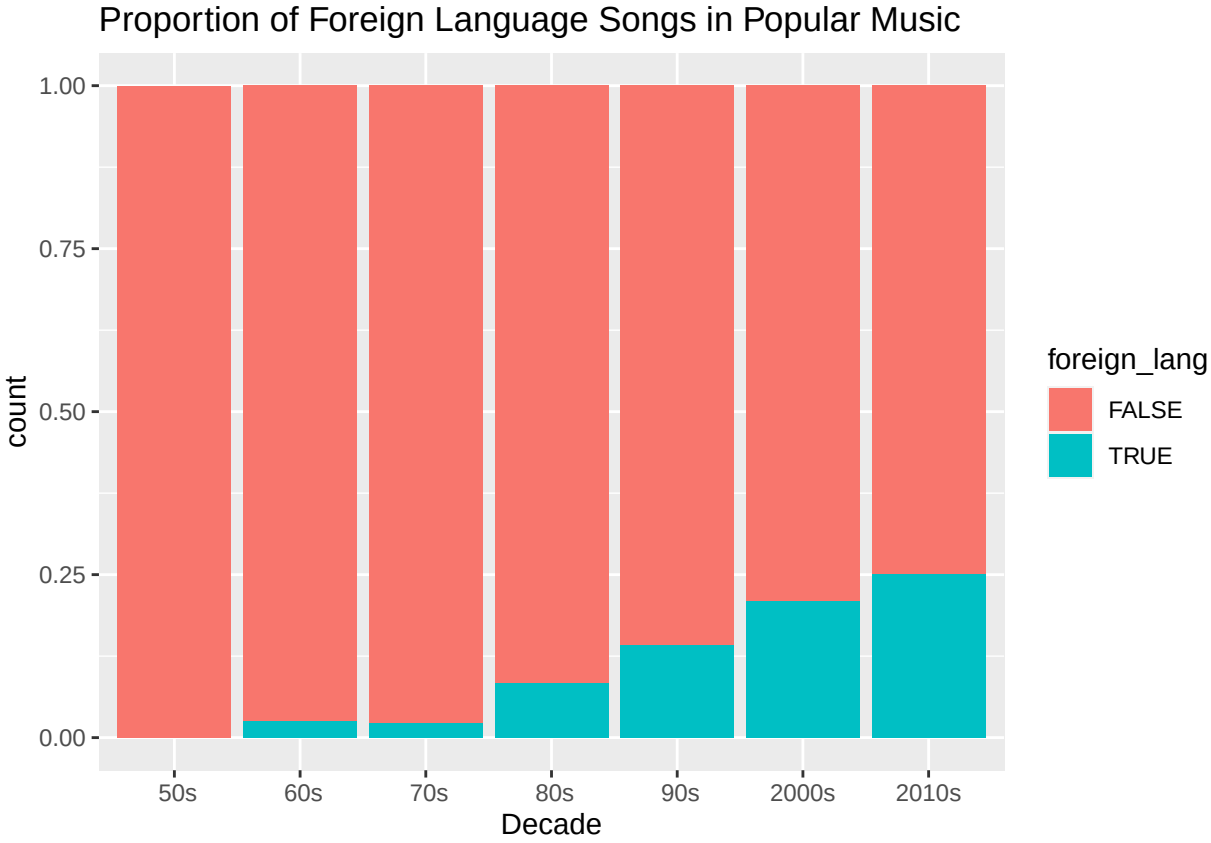
1. `decade`: categorical variable indicating the decade when the song was created using the categorical name of "50s", "60s", and so on
2. `period`: categorical variable in which "Pre 90s" represents when Year is before 1990 and "90s and After" for when Year is after 1989
3. `foreign_lang`: categorical variable in which "TRUE" stands for any song whose `Top Genre` name indicates that the song is spoken in a non-English language, "FALSE" otherwise
4. `pop`: categorical variable in which "TRUE" stands for any song whose `Top Genre` name has "pop" in it, "FALSE" otherwise
5. `sd`: standard deviation of `Acousticness` per year
6. `yearSince1956`: numerical variable representing the years since the first `Year` in our data set (1956)



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 data set.

```
## # A tibble: 6 x 3
## # Groups:   decade [6]
##   decade  pop       n
##   <chr>   <chr> <int>
## 1 " 60s"  pop      12
## 2 " 70s"  pop      37
## 3 " 80s"  pop      47
## 4 " 90s"  pop      39
## 5 "2000s" pop      74
## 6 "2010s" pop     138
```

Using the dplyr count function, we can see that the number of pop genre songs have generally increased over decades. We will further explore later whether these songs classified as the pop genre have an impact on the standard deviation of acousticness over time.



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 use simulation-based hypothesis testing to see if the proportion of `foreign_lang` songs in this data set increased from the "Pre 90s" period to the "90s and After" 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 "Pre 90s" time period to the "90s and After" time period.

We will use 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

Hypothesis test for non-English popular songs

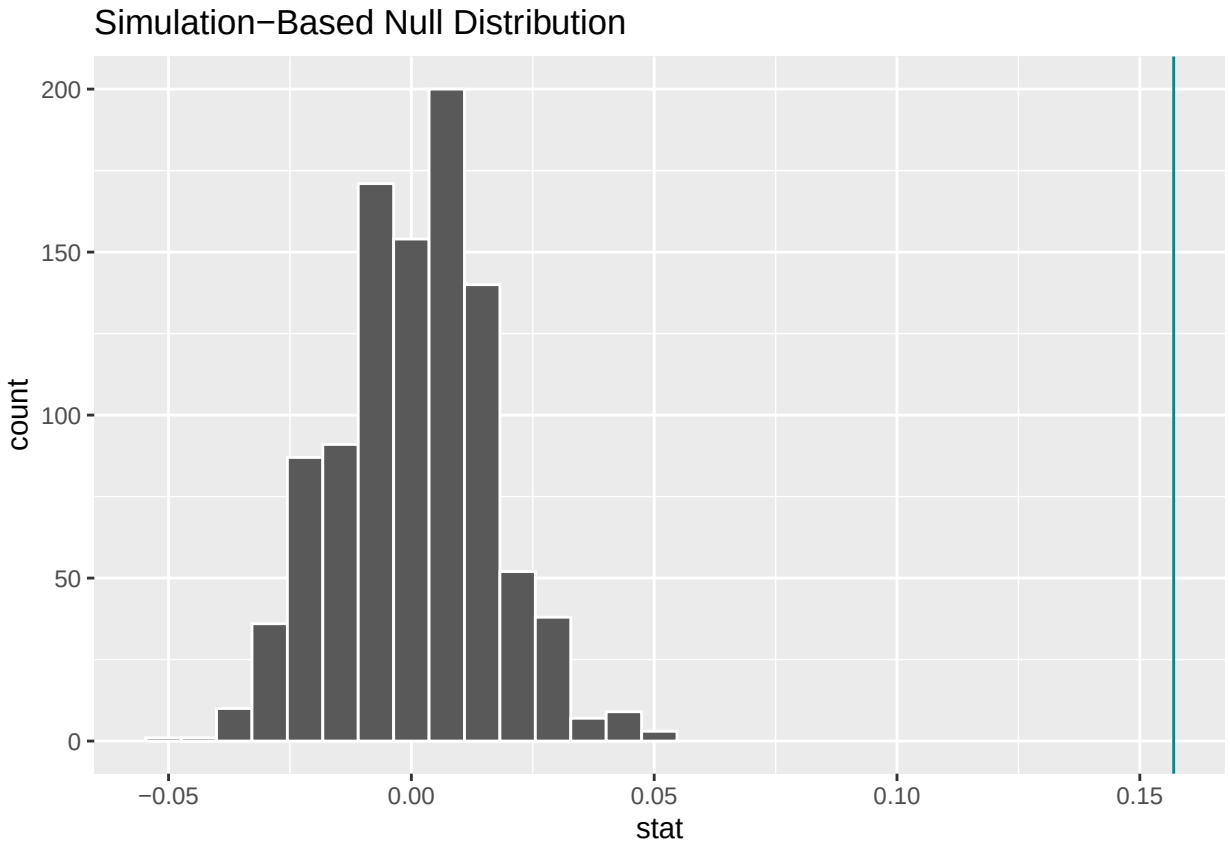
H0: The proportion of popular foreign language songs in the "90s and After" time period is equal to the proportion of popular foreign language songs in the "Pre 90s" time period ($\rho_m = \rho_o$, where ρ_m is the proportion of popular non-English language songs in the "90s and After" time period and ρ_o the proportion of non-English language songs in the "Pre 90s" time period)

Ha: The proportion of popular foreign language songs in the "90s and After" is greater than the proportion of popular foreign language songs in the "Pre 90s" time period ($\rho_m = \rho_o$, where ρ_m is the proportion of popular non-English language songs in the "90s and After" time period and ρ_o the proportion of non-English language songs in the "Pre 90s" time period)

Alpha: 0.05 ($\alpha = 0.05$)

```
## # 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 non-English songs in the time frame of 1990-2019 is greater than the proportion of popular non-English songs in the time frame of 1956-1989.



Linear model for standard deviation of acousticness

The null hypothesis is that the coefficient is equal to zero and the year has no effect on the standard deviation of `Acousticness`.

The alternative hypothesis is that the coefficient is not equal to zero and the year has an effect on the standard deviation of `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 yearSince1956   0.0596   0.00480      12.4 3.90e-34
```

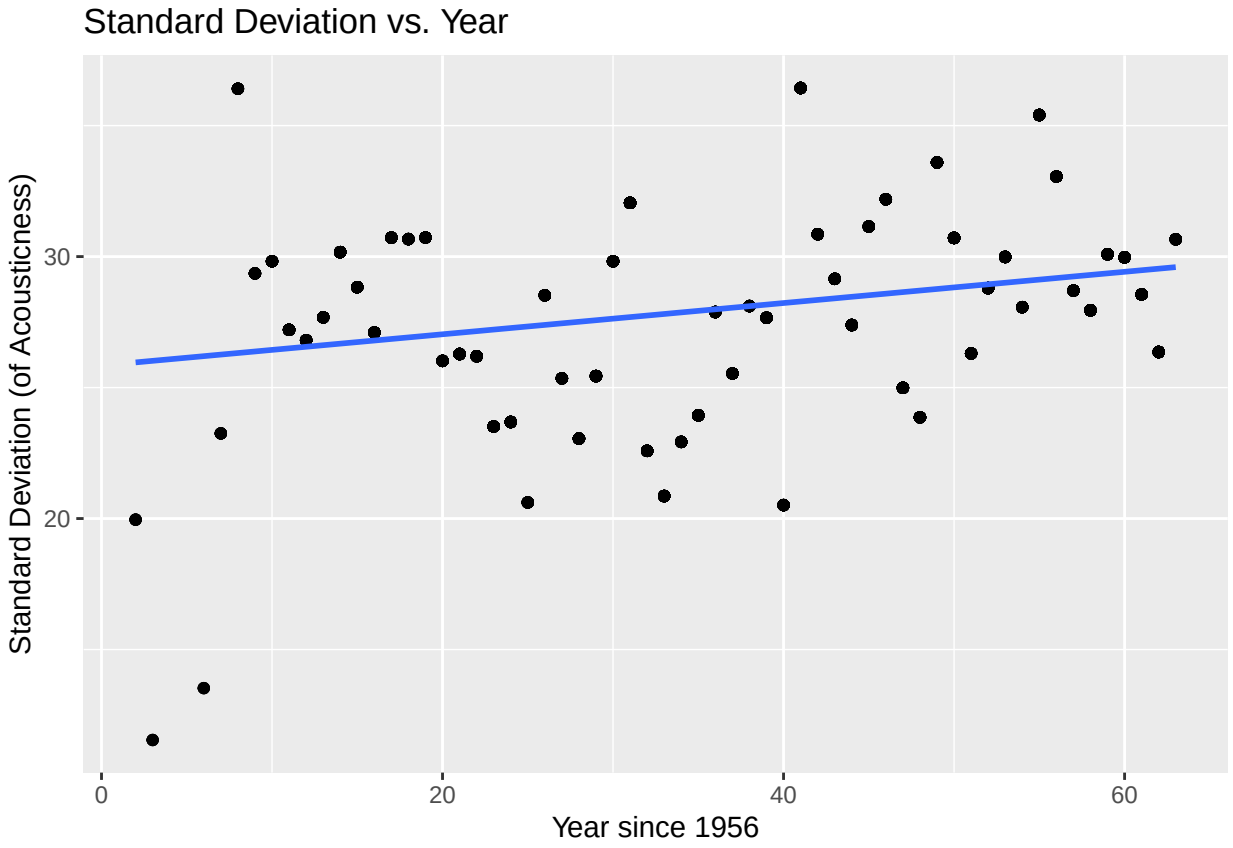
We will be using the variable `sd`, which is the standard deviation of the `Acousticness` of the songs per year.

The linear equation for the model is $\widehat{sdofAcousticness} = 25.842 + 0.0596 \times yearSince1956$

The intercept of this graph tells us that when the year is 1956, we expect the standard deviation for `Acousticness` to be 25.842, holding all else constant.

The slope tells us that for each additional year, the standard deviation of `Acousticness` is expected to increase, on average, by 0.06.

Our significance level is 0.05 and since the p-value is less than 0.05, we reject the null hypothesis and say there is sufficient evidence that the coefficient is not equal to zero and the year has an effect on the standard deviation of `Acousticness`.



This linear model suggests a slight positive linear relationship of the standard deviation of `Acousticness` of popular songs and the year the popular songs were created.

Next, we chose to do a multiple linear regression model with interaction effects to allow for different slopes and explore whether the relationship between the explanatory variable of year and standard deviation of `Acousticness` depends on the relationship between the explanatory variable of the pop genre and the standard deviation of `Acousticness`.

The null hypothesis is that the coefficients are equal to zero and the year and pop genre have no effect on the

standard deviation of `Acousticness`. The alternative hypothesis is that the coefficients are not equal to zero and the year and pop genre have an effect on the standard deviation of `Acousticness`.

```
## # 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 yearSince1956         0.0591    0.00538    11.0   2.66e-27
## 3 poppop               -0.113     0.610      -0.186 8.53e- 1
## 4 yearSince1956:poppop  0.00278   0.0133      0.209 8.35e- 1
```

$\widehat{sdAcousticness} = 25.858 + 0.059 \times yearSince1956 - 0.113 \times pop - 0.003 \times yearSince1956 : pop$

The intercept of this graph tells us that when the year is 1956, we expect the standard deviation for `Acousticness` to be 25.858. All else held constant, for every additional year since 1956, the standard deviation of `Acousticness` for pop songs is predicted to increase, on average, by 0.056.

Linear Model for pop songs: $\widehat{sdAcousticness} = 25.745 + 0.056 \times yearSince1956$

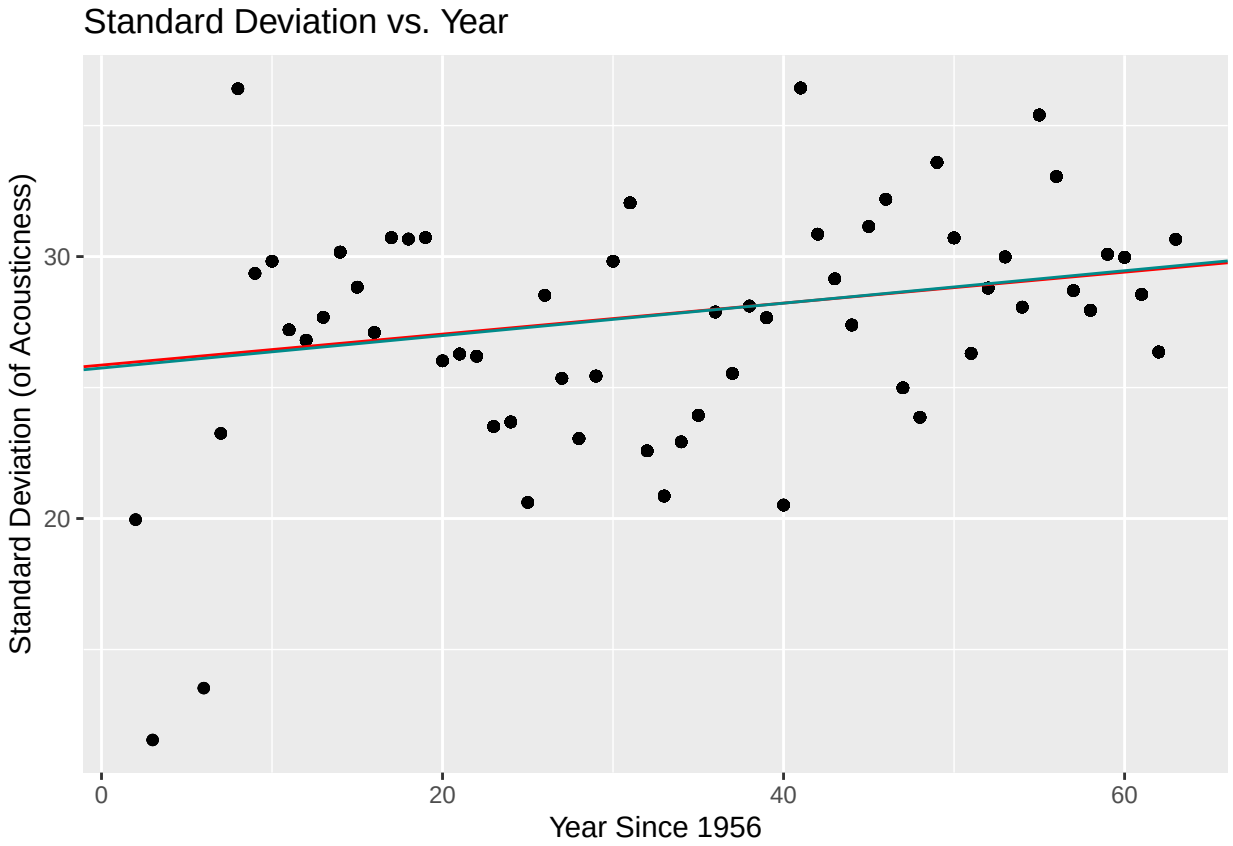
All else held constant, for every additional year since 1956, the standard deviation of `Acousticness` for pop songs is predicted to increase, on average, by 0.056.

Linear Model for non-pop songs: $\widehat{sdAcousticness} = 25.858 + 0.059 \times yearSince1956$

Our models predict:

1. Rate of change in the standard deviation of `Acousticness` as year increases depends on the genre since they have different slopes.
2. Non-pop songs consistently have a larger standard deviation of `Acousticness` than pop songs since it has a higher intercept.

Our significance level is 0.05 and since the p-value is less than 0.05, we reject the null hypothesis and say there is sufficient evidence that the coefficients are not equal to zero and the year and pop genre have an effect on the standard deviation of `Acousticness`.



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
```

Because the adjusted R^2 value is larger for the non-interactive linear model compared to the interactive linear model, the non-interactive linear model is a better fit for our data.

Discussion

Summary of Statistical Analysis

Our research question asked whether the data provides 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 and whether the data provides sufficient evidence that the standard deviation in `Acousticness` has increased over time. Overall, we wanted to observe how popular music has shifted to be more diverse in the characteristics of non-English genres or `Acousticness` over time.

To answer our research question, we created visualizations to show some general trends in `Acousticness`, the prevalence of pop as a genre, and non-English languages in popular music over time that we wanted to analyze further using hypothesis tests and regression. Using our simulation-based hypothesis test, we found that there is significant evidence that the proportion of non-English songs in popular music has increased

from the "Pre 90s" time period of before 1990 to the "90s and After" time period of after 1989. We have learned that there is enough evidence to show using linear regression with and without interaction effects that the standard deviation of `Acousticness` increased over time because our p-value for both regression models were both very very little and therefore less than our significance level so we had sufficient evidence to reject the null hypothesis in both cases.

Reliability of the Data

Kaggle, the website where we took our data from, is reputable within the statistics community but given that the data was scraped from Spotify, the top 1,994 songs only take into account Spotify listeners and not radio listeners, Youtube users, or users on other music platforms. The validity of our categorical variables like `Top Genre` and `pop` are questionable because these are subjective factors, like the labeling of one song as "alternative pop rock" or "pop rock". There are limitations to the analysis we can do with this data set, since there may be many other factors that are not variables which we cannot include in our modeling which impact the diversity of popular music. For instance, if an artist is already popular or well-known, multiple songs may become popular in the same year in which they released. an album. In addition, we have a limited number of samples of popular songs.

Critiques and Next Steps

We chose to create the variable for time period ("Pre 90s" and "90s and After") based on the fact that the Internet became widely used by the public in the 1990s but our choice to create these two time periods the way we did is somewhat arbitrary and could be misleading. The data also began in 1956 so for our decades variable, the 50s had fewer data points than the other decades. This means that the proportion of non English songs for the 50s may have impacted the "Pre 90s" time period more and skewed our data and conclusions from the simulation-based hypothesis test. There also could be more information or patterns within specific decades or years that we missed by using the larger time period so in the future, we would do more hypothesis tests comparing individual years or decades.

There was not a strong correlation for our linear regression models and our graph shows a large number of residuals from our line of best fit. Furthermore, the interaction effect was very small between the pop genre and `Acousticness` so our second model which included the interaction effect overlapped the original model. If we were to redo our regression models, we would choose different variables and explore the possibility of nonlinear models such as logistic models.

If we were to start over, we would use a data set where there is popular music in the United States so we could better examine how one country responds to cultural shifts in music instead of worldwide. It can also be seen as a form of extrapolation if we use our conclusions from the data set to talk about years outside the data set, so if we were to redo our project, we find a data set with popular songs from more years.

However, we believe our statistical analysis and testing is a good starting point for future analysis onto music data and was able to show us interesting patterns and trends in popular music.

Works Cited

https://www.rollingstone.com/music/music-news/hip-hop-continued-to-dominate-the-music-business-in-2018-774422/

https://www.kaggle.com/iamsumat/spotify-top-2000s-mega-dataset

https://webfoundation.org/about/vision/history-of-the-web/