CSIT 528 Statistics for Data Science Final Project

What is the correlation between song duration and popularity?

Marc Stern

1. Description

1.1 Motivation

In a digital world where it becomes harder to grab people's attention, it has become increasingly common for artists to release short songs. While the three-minute pop song has been an industry standard for decades, it should not be surprising to find that in general, song lengths, regardless of genre, have minimized.

1.2 Project Objectives

The goal of this project is to determine whether the above is true based on the data presented -- to determine if song duration has decreased. Another factor that will be investigated is whether there is a correlation between the length of a song and its popularity. As previously stated, pop music (the most listened to genre, hence its name "popular music") is often thought to have a standard of three minutes per song, while less popular genres such as jam band rock, or jazz, seemingly have no designated end time. Is this idea backed by the data, and does it stretch across music as a whole?

1.3 Description of the Data Set

This is a dataset containing over 160,000 songs from Spotify's API [1] released between 1921 and 2020. Included are categorical data such as song key, release date, artist name, and song name. Additionally, each song contains data quantified musical components like energy, danceability, acousticness, and many more.

Specifically, this project will be working with song duration (in ms) and popularity. The categories of data are taken from two subsets of years, 2019, which is the most recent full year's worth of data, and 1972. It is worth noting that the popularity category measures song popularity throughout the history of its existence on Spotify, as compared to the other songs in Spotify's massive catalog, and not by its popularity at the time of release by other metrics.

2. Exploratory Data Analysis

2.1 Descriptive Statistics

From the data set, a random sample of 100 samples were collected from 2019 and 1972 each. Broken down into two subgroups, duration (ms) and popularity, which is graded on a scale from 0 to 100. The sample means $\bar{x}$ and standard deviations *s* are calculated for each.

Song Duration (ms) 2019: $\bar{x}_1 = 197{,}759.21$, $s_1 = 48{,}103.65$

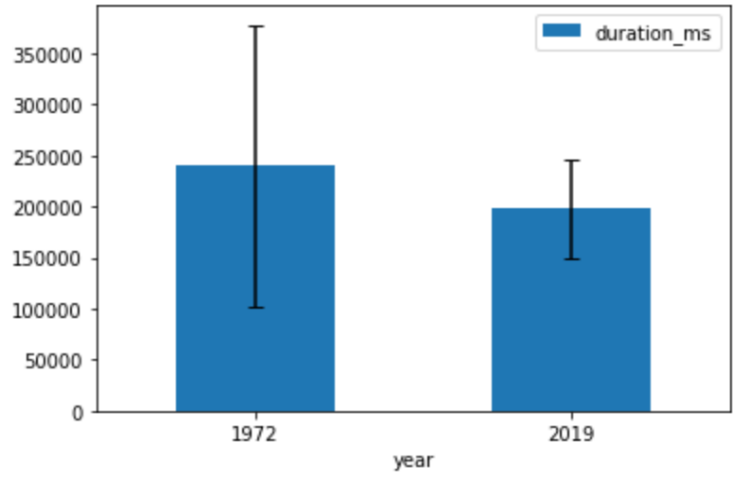
Song Duration (ms) 1972: $\bar{x}_2 = 240{,}297.74$, $s_2 = 137{,}602.31$

Song Popularity 2019: $\bar{x}_3 = 69.44$, $s_3 = 6.70$

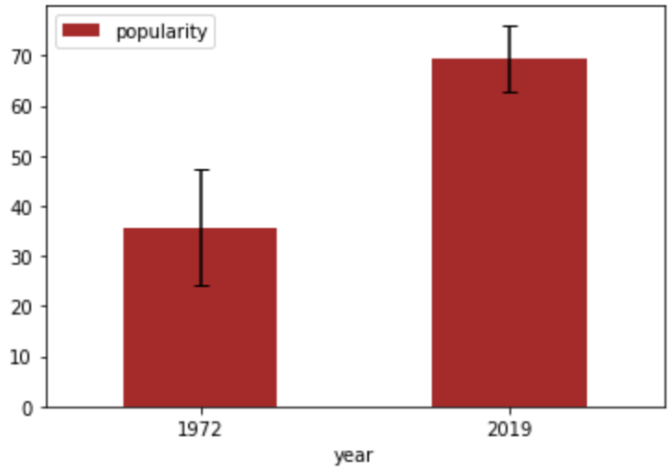
Song Popularity 1972: $\bar{x}_4 = 35.67$, $s_4 = 11.64$

The data for both categories can be seen through the visualizations below:

A)



B)



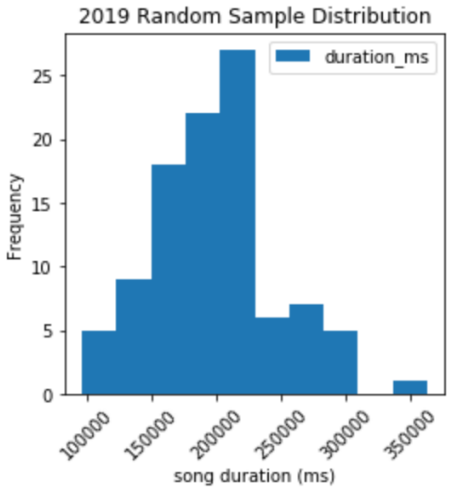
Bar chart *A* displays the sample mean song duration (ms) in 2019 and 1972, with the sample standard deviation shown by the error bars. Bar chart *B* shows the sample mean popularity rating in both years, and the sample standard deviation with error bars.

The data visualization *A* shows that average song length was shorter in 2019 than it was in 1972. Visual *B* shows that the popularity of music released in 2019 is more popular than music released in 1972.

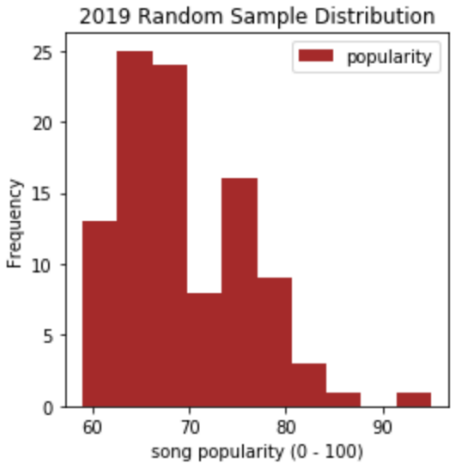
2.2 Distribution of Variables

For each variable calculated in section 2.1, there is a corresponding distribution of the data that creates the sample statistic. Creating a relative frequency table is the first step in determining the distribution of each data variable. Then, using the "square root rule" to find the number of bins for $n = 100$ is 10 bins. Finally, a histogram can display the visual representation of the distribution. All four variables' distributions are displayed below:

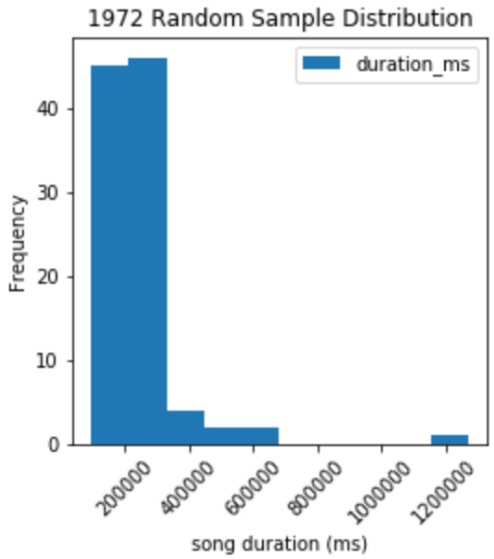
i)



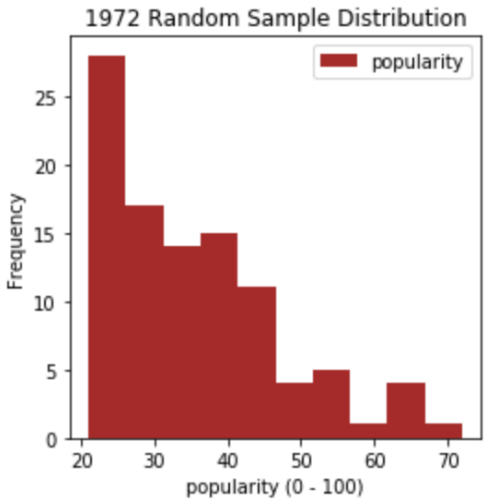
ii)



iii)



iv)



All of the variables have a right skewed distribution, except for *i*, which appears to have an approximately normal distribution. However, it is worth noting that a significant portion of the data is to the left of the center of a normal distribution, which is nearly right skewing as well.

3. Data Analysis and Results

3.1 Hypothesis Testing

For the hypothesis test performed in this section, the data, although not always of a normal distribution, is a random sample selection of $n = 100$. Since $n \geq 30$, the large sample theory applies, and a normal distribution of the data is not necessary.

The first hypothesis is that the population mean song duration is smaller for releases in 2019 compared to 1972. Using a 1% level of significance $\alpha = 0.01$, the null and alternative hypothesis can be written as:

$$H_0: \mu 1 = \mu 2$$
$$H_1: \mu 1 < \mu 2$$

With the already calculated sample standard deviations $s1$and $s2$, a sample statistic $t = -2.92$ can be computed with the following code:

```
t = (x1_bar - x2_bar) / np.sqrt(s1**2 / n1 + s2**2 / n2)
```

Because the sample statistic is negative, a left-tailed test with degree of freedom $= 99$ is used to find the corresponding P-value $= 0.0022$, with the following code:

```
p_val = stats.t.cdf(t, df)
```

$Pvalue = 0.0022 < \alpha = 0.01$

Therefore, we can reject the null hypothesis that the average song duration is not the same in 2019 as it was in 1972.

Furthermore, based on the sample data, we can predict an interval in which the population mean μ lies. To calculate an 85% confidence interval, the following code was used to first calculate the sample statistic $t_c = 1.45$ and an error tolerance $E = 21{,}136.38$:

```
c = 0.85
alpha = (1 - c) / 2
df = n1 - 1
t_c = round(abs(stats.t.ppf(q=alpha, df=df)), 2)
E = t_c * np.sqrt((s1 ** 2 /n1) + (s2 ** 2 /n2))
conf = round((x1_bar - x2_bar) - E, 2), round((x1_bar - x2_bar) + E, 2)
```

Resulting in a confidence interval:

$-63{,}674.91 < \mu < -21{,}402.15$

This interval is negative, therefore with 85% confidence we can claim that $\mu 1 < \mu 2$, and accept the alternative hypothesis that song duration in 2019 music releases is less than song duration in 1972 releases. We can also conclude that song releases in 2019 are specifically in the range of 21,402.15 to 63,674.91 ms shorter, or about 21.4 to 63.7 seconds shorter.

3.2 Correlation and Regression

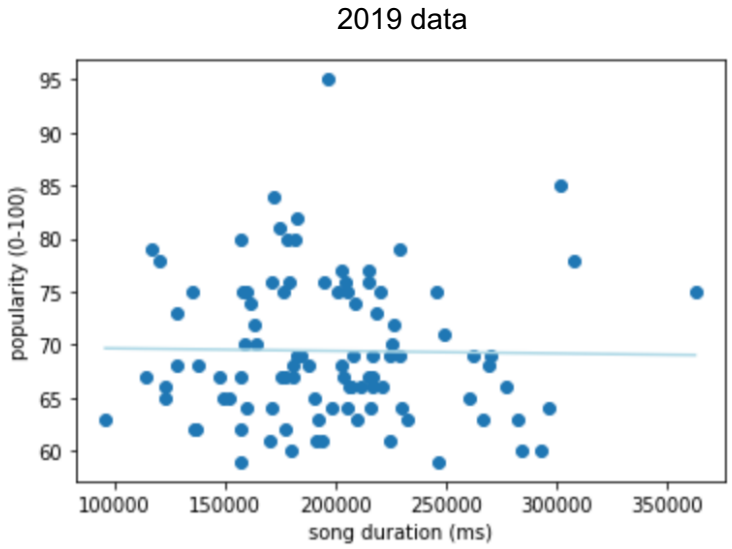
This section will look at the 2019 data, and compare variables of song duration and population in order to draw or reject correlation using regression analysis.

First, we use the following code to pass the raw data in order to return important values including the slope, intercept, the correlation coefficient r, and the P-value:

```
slope, intercept, r_val, p_val, std_err = stats.linregress(dur19, pop19)
```

Using the slope and intercept values, we can calculate the least-of-squares line, which shows correlation has the equation: $\hat{y} = 69.91 + -0.0x$

This produces a scatter plot that looks like this:



The least-of-squares line does not suggest any correlation, nor does its equation with a rounded slope of $b = 0$.

$r^2 = 0.0003$, meaning the explained variable only accounts for 0.03% of the data, while the unexplained variable accounts for the remaining 99.97% of the data. Combined with a previous result $r = -0.017$, both r values are so close 0 that it does not make sense to use the least-of-squares line.

Using the least-of-squares line can usually predict a confidence interval for a value. While this line has not shown correlation, we'll follow the steps anyway for further proof of its unreliability:

Using a confidence level of 85%, we can estimate the popularity of a song released in 2019 that is exactly 3 minutes long, or 180,000 ms:

```
def std_err (x,y):
    slope = stats.linregress(x,y).slope
    intercept = stats.linregress(x,y).intercept
    y_hat = slope * x + intercept
    SE = np.sqrt(((y-y_hat)**2).sum()/(len(y)-2))
    return SE
```

```
def margin_y(x,y,c,x_value):
    n = len(x)
    x_bar = x.mean()
    t_c = abs (stats.t.ppf(q=(1-c)/2, df=n-2))
    S_e = std_err(x,y)
    x_sum = x.sum()
    x_sqsum = (x**2).sum()
    E = t_c*S_e*np.sqrt(1 + 1/n + n*((x_value-x_bar)**2)/(n*x_sqsum-x_sum**2))
    y_hat = stats.linregress(x,y).slope *x_value + stats.linregress(x,y).intercept
    beta_lower = y_hat-E
    beta_upper = y_hat+E
    return (round(beta_lower, 4), round(beta_upper, 4))
```

```
x_value = 180000 # 3 minute song
c = 0.85
conf = margin_y(dur19, pop19, c, x_value)
```

The results are such that with 85% confidence we can conclude that a 3 minute song in 2019 has a popularity rating on Spotify between 59.65 and 79.31.

4. Conclusion

There are three main conclusions that can be drawn from these results. The first is that song duration is shorter for 2019 releases than 1972 releases. It is also noticeable that 2019 releases are more popular on Spotify than older releases from 1972. Lastly, while the song popularity was able to be estimated with 85% confidence for 2019 releases, the overall correlation between song popularity and duration is none, and the prediction is not necessarily reliable. The data does not just show a lack of correlation, but it overwhelmingly demonstrates with confidence that there is no correlation.

Citations

[1] Yamaç Eren Ay. (2020). *Spotify Dataset 1921-2020, 160k+ Tracks: Audio features of 160k+ songs released in between 1921 and 2020.* [Online]. Available: https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks