Spotify Data is being used

Reading in the dataset below and showcasing columns via str

```
library('ggplot2')
library('tidyverse')
songs <- read.csv('spotify_songs.csv')
str(songs)
```

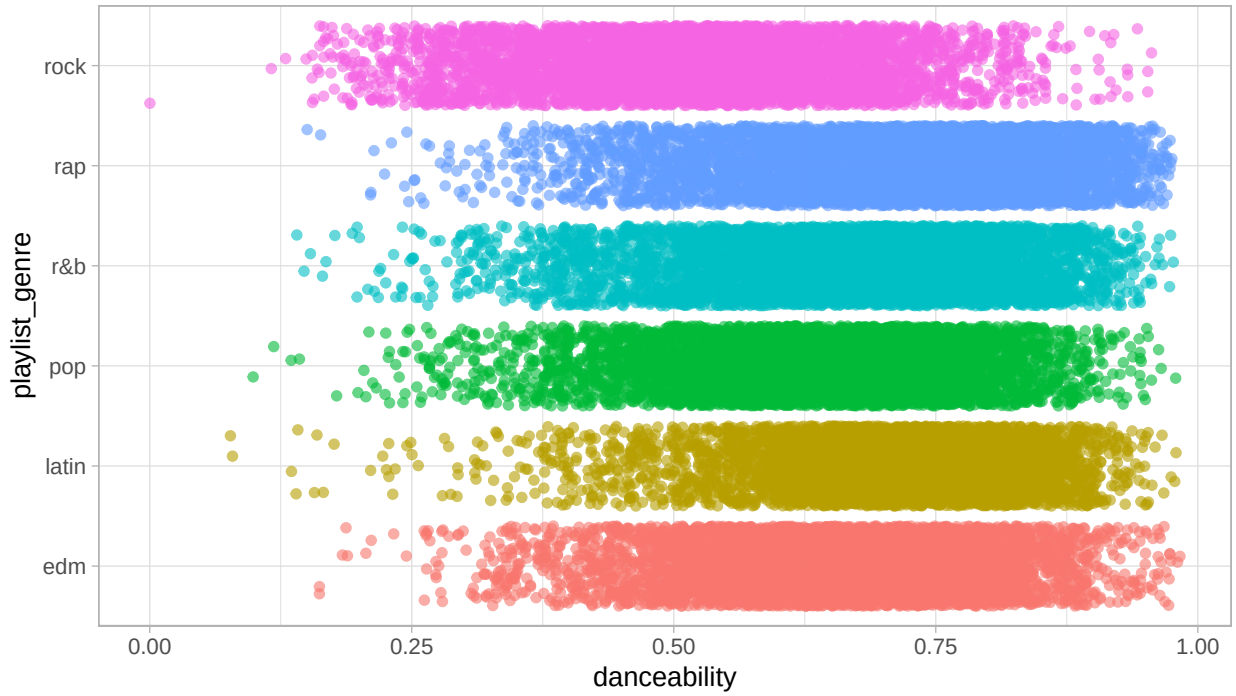
```
## 'data.frame':    32833 obs. of  23 variables:
##  $ track_id                : chr  "6f807x0ima9a1j3VPbc7VN" "0r7CVbZTWZgbTCYdfa2P31" "1z1Hg7Vb0AhHDiEm
##  $ track_name              : chr  "I Don't Care (with Justin Bieber) - Loud Luxury Remix" "Memories -
##  $ track_artist            : chr  "Ed Sheeran" "Maroon 5" "Zara Larsson" "The Chainsmokers" ...
##  $ track_popularity        : int  66 67 70 60 69 67 62 69 68 67 ...
##  $ track_album_id          : chr  "2oCs0DGTsRO98Gh5ZSl2Cx" "63rPSO264uRjW1X5E6cWv6" "1HoSmj2eLcsrR0vl
##  $ track_album_name        : chr  "I Don't Care (with Justin Bieber) [Loud Luxury Remix]" "Memories
##  $ track_album_release_date: chr  "2019-06-14" "2019-12-13" "2019-07-05" "2019-07-19" ...
##  $ playlist_name           : chr  "Pop Remix" "Pop Remix" "Pop Remix" "Pop Remix" ...
##  $ playlist_id             : chr  "37i9dQZF1DXcZDD7cfEKhW" "37i9dQZF1DXcZDD7cfEKhW" "37i9dQZF1DXcZDD
##  $ playlist_genre          : chr  "pop" "pop" "pop" "pop" ...
##  $ playlist_subgenre       : chr  "dance pop" "dance pop" "dance pop" "dance pop" ...
##  $ danceability            : num  0.748 0.726 0.675 0.718 0.65 0.675 0.449 0.542 0.594 0.642 ...
##  $ energy                  : num  0.916 0.815 0.931 0.93 0.833 0.919 0.856 0.903 0.935 0.818 ...
##  $ key                     : int  6 11 1 7 1 8 5 4 8 2 ...
##  $ loudness                : num  -2.63 -4.97 -3.43 -3.78 -4.67 ...
##  $ mode                    : int  1 1 0 1 1 1 0 0 1 1 ...
##  $ speechiness             : num  0.0583 0.0373 0.0742 0.102 0.0359 0.127 0.0623 0.0434 0.0565 0.032
##  $ acousticness            : num  0.102 0.0724 0.0794 0.0287 0.0803 0.0799 0.187 0.0335 0.0249 0.056
##  $ instrumentalness        : num  0.00 4.21e-03 2.33e-05 9.43e-06 0.00 0.00 0.00 4.83e-06 3.97e-06 0
##  $ liveness                : num  0.0653 0.357 0.11 0.204 0.0833 0.143 0.176 0.111 0.637 0.0919 ...
##  $ valence                 : num  0.518 0.693 0.613 0.277 0.725 0.585 0.152 0.367 0.366 0.59 ...
##  $ tempo                   : num  122 100 124 122 124 ...
##  $ duration_ms             : int  194754 162600 176616 169093 189052 163049 187675 207619 193187 253
```

Q2 Comparing Two Variances

Continuous variable being chosen is 'danceability'. Categorical variable being chosen is 'playlist_genre'

Below we will showcase a jitter plot with these two variables

```
ggplot(songs, aes(y=playlist_genre, x=(danceability), color=playlist_genre)) +
  geom_jitter(alpha=.6) +
  theme_light() +
  theme(legend.position = 'none')+labs(x='danceability')
```



The two values from our categorical variable selected will be 'pop' and 'rap' These can be taken as our two samples.

```
sample1 <- subset(songs, playlist_genre =='pop')$danceability
sample2 <- subset(songs, playlist_genre  =='rap')$danceability
```

Now we will list the different null hypothesis which we will test. These are the hypothesis's for comparing our population variances.

1. $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$
2. $H_0 : \mu_1 - \mu_2 \geq 0$
3. $H_0 : \mu_1 - \mu_2 \leq 0$

First Hypothesis Check

```
var.test(sample1,sample2)
```

```
##
##  F test to compare two variances
##
## data:  sample1 and sample2
## F = 0.88301, num df = 5506, denom df = 5745, p-value = 3.146e-06
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8380405 0.9304119
## sample estimates:
## ratio of variances
##          0.8830062
```

Since our p-value is very small, we will reject the hypothesis that danceability in both genres are the same.

Second hypothesis check

```
var.test(sample1,sample2, alternative = 'less')
```

```
##
##  F test to compare two variances
##
## data:  sample1 and sample2
## F = 0.88301, num df = 5506, denom df = 5745, p-value = 1.573e-06
## alternative hypothesis: true ratio of variances is less than 1
## 95 percent confidence interval:
##  0.000000 0.922621
## sample estimates:
## ratio of variances
##          0.8830062
```

This hypothesis is also rejected as our p-value is very small, we cannot say mean danceability in pop is higher than rap.

Third hypothesis check

```
var.test(sample1,sample2, alternative = 'greater')
```

```
##
##  F test to compare two variances
##
## data:  sample1 and sample2
## F = 0.88301, num df = 5506, denom df = 5745, p-value = 1
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  0.8451124       Inf
## sample estimates:
## ratio of variances
##          0.8830062
```

With the p-value equal to 1, we thus cannot reject this hypothesis that the mean danceability for rap is higher than pop.

Q3 Comparing two population means

Listing the three hypothesis:

1. $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$
2. $H_0 : \mu_1 - \mu_2 \geq 0$
3. $H_0 : \mu_1 - \mu_2 \leq 0$

for our first text we know the the variances do not equal each other

```
t.test(sample1,sample2, var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  sample1 and sample2
## t = -31.682, df = 11247, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08394199 -0.07416013
## sample estimates:
## mean of x mean of y
## 0.6393017 0.7183528
```

Given our low p-value we reject the null hypothesis of the two pop means are equal to each other.

Now for our second hypothesis, we also had to reject this when testing the variance, thus:

```
t.test(sample1,sample2, var.equal = F, alternative = 'less')
```

```
##
##  Welch Two Sample t-test
##
## data:  sample1 and sample2
## t = -31.682, df = 11247, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.07494657
## sample estimates:
## mean of x mean of y
## 0.6393017 0.7183528
```

Thus given our P-val, we reject the null hypothesis that pop has a higher danceability mean than rap.

Final hypothesis check. In our variance check we were unable to reject the null hypothesis, thus var.equal = t.

```
t.test(sample1,sample2, var.equal = T, alternative = 'greater')
```

```
##
##  Two Sample t-test
##
## data:  sample1 and sample2
## t = -31.64, df = 11251, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.08316097         Inf
## sample estimates:
## mean of x mean of y
## 0.6393017 0.7183528
```

```
# var.equal = true as thus null hypothesis when testing variance was unable to
# be rejected.
```

Thus, we cannot reject this hypothesis.