Spotify and Youtube

Nhi Le

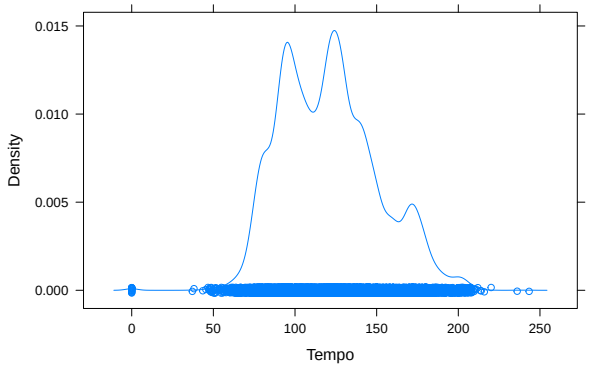
2023-05-23

Introduction: We are examining the relationships between characteristics of the top ten songs of artists on Spotify and Youtube. The research question we are investigating is whether there is a relationship between tempo and danceability, as well as whether most songs are rated danceable or not. Additionally we are looking to investigate whether there is a relationship between danceability and energy as well as valence. Additionally the relationship between danceability and views on Youtube will be explored. Investigation into these relationships can give insight into what qualities contribute to the danceability of a song, as well as how songs with different qualities are perceived.

Data: The dataset contains 26 variables. Most of the variables will be excluded from analysis. Variables that will be considered are:

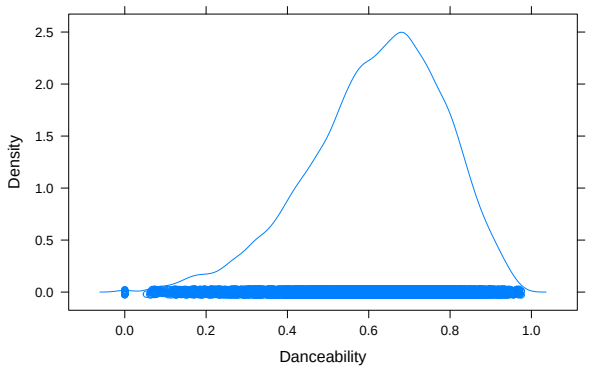
Tempo, measured in beats per minute. The density plot is bimodal, indicating a great density of tempos around 95 and around 125 beats per minute.

Figure 1: Density plot of Tempos (in beats per minute)



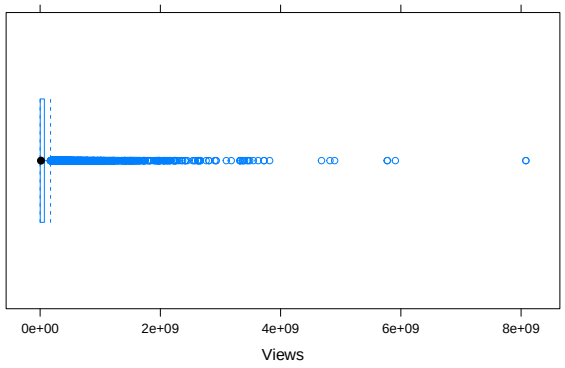
Danceability, measured as a value from 0.0 to 1.0, zero being least danceable and 1 being most danceable. This variable indicates how suitable a track is for dancing. The density plot indicates that the greatest density of songs are rated between 0.6 and 0.8. The distribution is approximately normal, though it is skewed to the left.

Figure 2: Density plot of Danceability



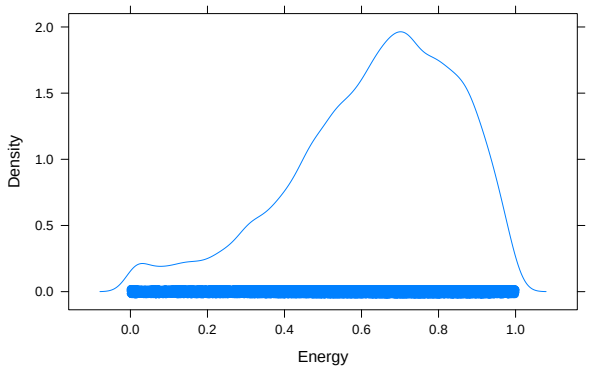
Views is number of views on Youtube. The box and whisker plot shows a large quantity of outliers above the 1.5IQR with numbers of views multiple magnitudes above the majority of the population.

Figure 3: Box & Whisker Plot of Views



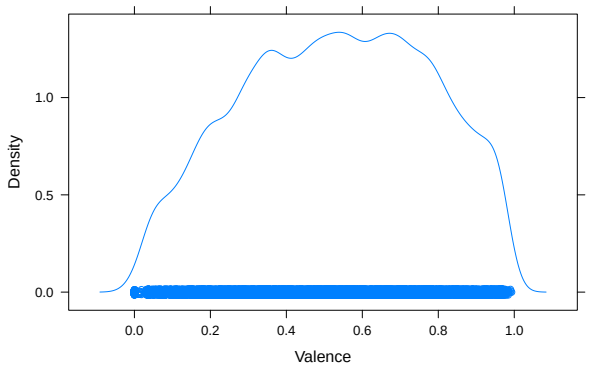
Energy, measured as a value from 0.0 to 1.0. This variable is a subjective rating of how the song feel energetically, with 0 being low in energy and 1 being highest in energy. The density plot indicates that the greatest density of songs are rated around a 0.7. The distribution looks approximately normal, though it is skewed to the left.

Figure 4: Density plot of Energy



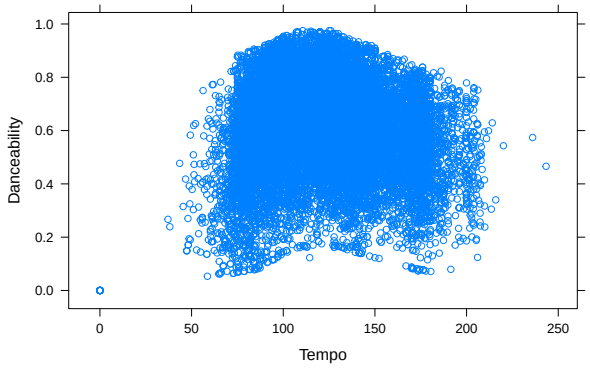
Valence is measured from 0.0 to 1.0. This variable is a subjective rating of the positiveness of the song. High valence songs, rated closer to 1, sound more happy. Low valence songs, closer to 0, sound more negative more sad or angry. The density plot indicates a relatively even spread of valence ratings between 0.4 and 0.8. Few songs are rated extremely sad or extremely happy.

Figure 5: Density plot of Valence



Analysis: Is there a linear relationship between the tempo of a song and its rated danceability?

Figure 6: Linear Regression Plot of Danceability and Tempo




```
## [1] -0.06594322
```

There is no linear relationship between tempo and danceability. The correlation coefficient of r= -0.066 is too low to even indicate a weak linear relationship.

The average tempo of the best-selling songs of 2020 was 122 bpm. Given this fact, we will perform a t-test to test the hypothesis that the average population bpm is 122. Null hypothesis: The population mean is equal to 122 bpm Alternative hypothesis: The population mean is not equal to 122 bpm

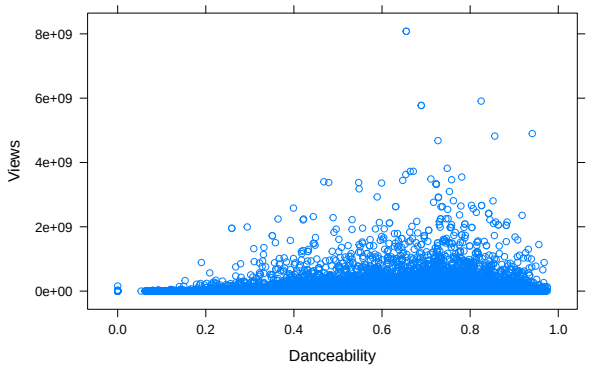
```
tstar <- qt(.975, df=20717)
t.test(~ Tempo, data=spotify, alternative="two.sided", mu=122)
```

```
##
##  One Sample t-test
##
## data:  Tempo
## t = -6.6258, df = 20715, p-value = 3.539e-11
## alternative hypothesis: true mean is not equal to 122
## 95 percent confidence interval:
##  120.2355 121.0412
## sample estimates:
## mean of x
##  120.6383
```

We are 95% confident that the population mean of the tempo is not equal to 122 beats per minute. The p-value is 0.001898 and the 95% confidence interval does not include 122.

Is there a relationship between the danceability rating of a song and the number of views on Youtube?

Figure 7: Linear Regression Plot of Danceability and Views




```
## [1] 0.08854618
```

There is no linear relationship between danceability and number of views. The correlation coefficient of r= 0.089 is too low to even indicate a weak linear relationship. It can be observed, however, that some of the highest outliers in terms of number of views also have a higher danceability rating.

The average number of views for a song video on Youtube is 26,000. Given this fact, we will perform a t-test to test the hypothesis that the average population views is 26,000. Null hypothesis: The population mean is equal to 26,000 bpm Alternative hypothesis: The population mean is not equal to 26,000 bpm

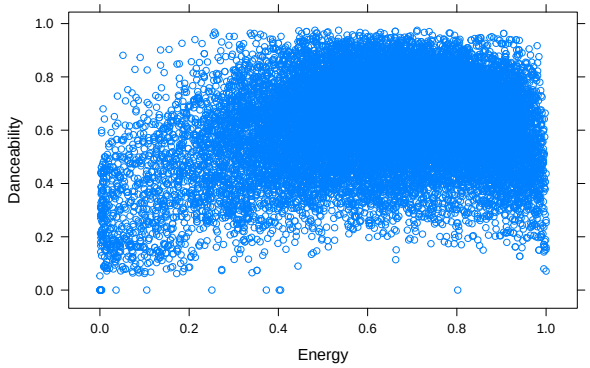
```
tstar <- qt(.975, df=20717)
t.test(~ Views, data=spotify, alternative="two.sided", mu=26000)
```

```
##
##  One Sample t-test
##
## data:  Views
## t = 48.656, df = 20247, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 26000
## 95 percent confidence interval:
##  90154671 97720972
## sample estimates:
## mean of x
##  93937821
```

We are 95% confident that the population mean of Youtube views is not 26000. The p-value is very close to zero and the 95% confidence interval does not include 26000. This is likely skewed due to some of the extreme outliers as well as the fact that the average number of views across all songs, whereas this dataset includes the top ten songs from artists.

Is there a relationship between the danceability rating and the energy rating of a song?

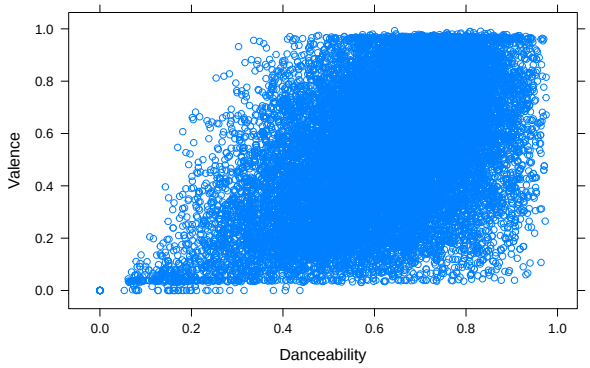
Figure 8: Linear Regression Plot of Danceability and Energy




```
## [1] 0.236596
```

There looks to be a very weak positive linear relationship between danceability and energy. The correlation coefficient of r= 0.236, however, is too low to even indicate a weak linear relationship.

Figure 9: Linear Regression Plot of Danceability and Valence




```
## [1] 0.465756
```

There is a moderate positive linear relationship between valence and danceability. The correlation coefficient is r= 0.466.

Conclusions: Overall there does not appear to be a relationship between the subjective rating of the danceability of a song and the tempo, views, or energy of the song. There is a moderate positive linear relationship between the valence and danceability of a song. Additionally, we conclude that there is sufficient evidence that the average beats per minute of a song is not equal to the 122 beats per minute average of the top selling songs of 2020. Finally, we conclude that there is sufficient evidence that the average number of views per song is not equal to the average number of views per song on Youtube, which is 26,000. One issue with this data set is how skewed the view count is. It is clear that some of the songs were immensely popular whereas a lot of them had few views. If new data was collected to address this same research question then outliers would be excluded from the analysis.