

SPRINGBOARD DATASCIENCE CAREER TRACK

# RECOMMENDATION SYSTEM

FOR THE MILLION SONG DATASET

---

Ivette M Tapia

# BASIC INFORMATION

Total Unique Songs: 905, 712

Total Unique Users: 1,019,318

Total Unique Artists: 29,559

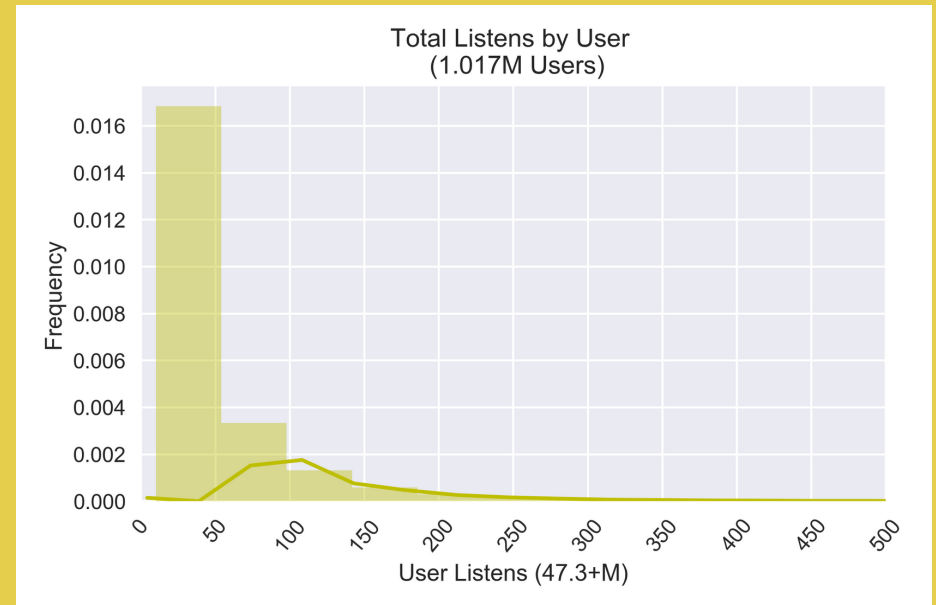
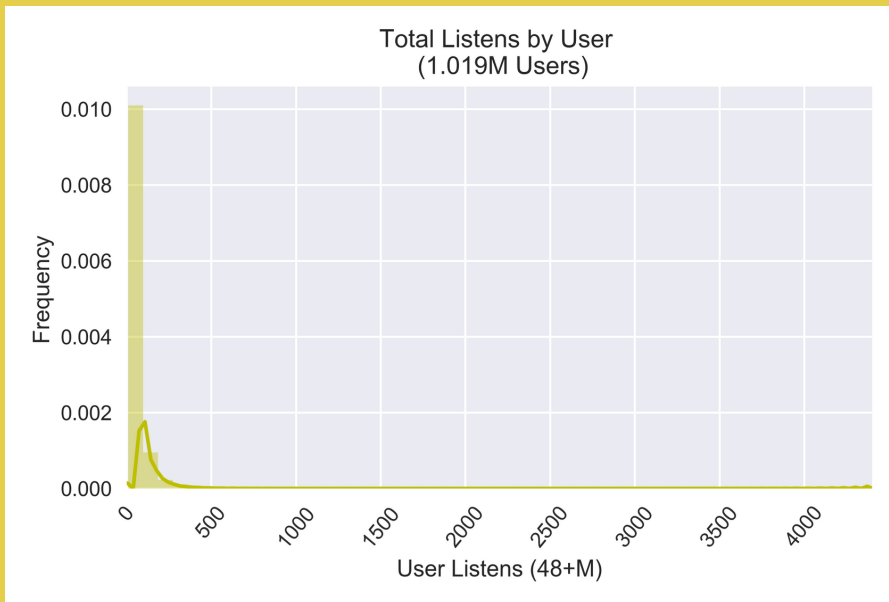
Total Song Listens: 48,373,586

# SKEWED USER ENGAGEMENT

Average listens per user is 5.7, standard deviation is 5.8.

10.8% of users have greater than 100 total listens.

1.4% of users had greater than 500 total listens.



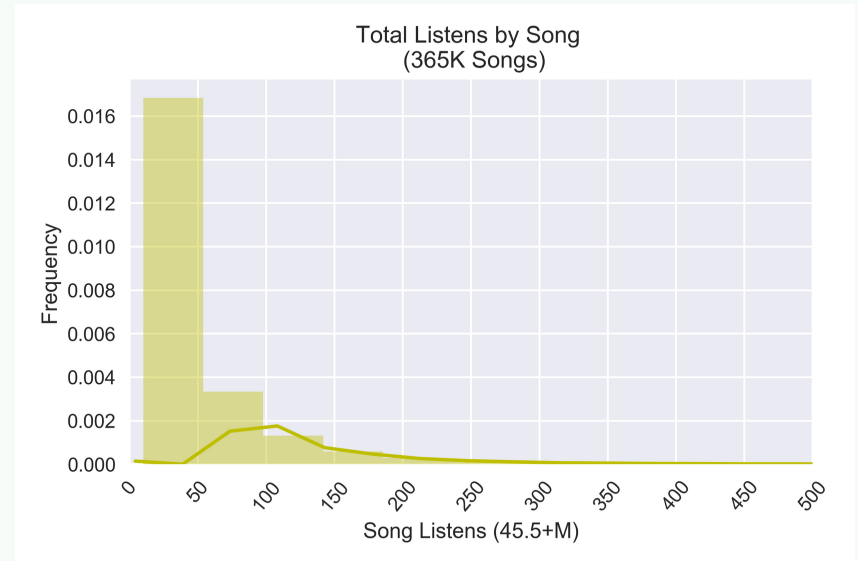
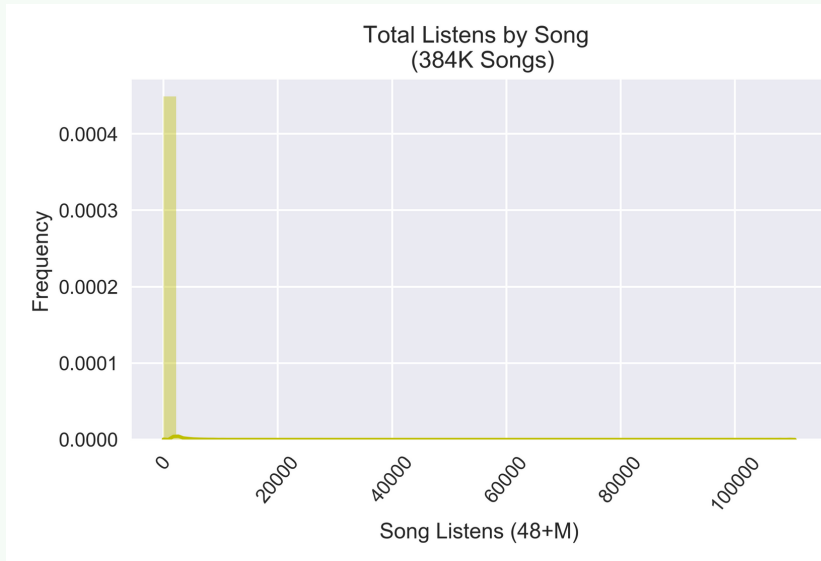
Min = 1, Max = 4,400

# SKEWED SONG ENGAGEMENT

Average listens per song is 125.8, standard deviation is 799.02

16.9% of songs have greater than a 100 total listens.

42% of the total song catalog has been listened to.



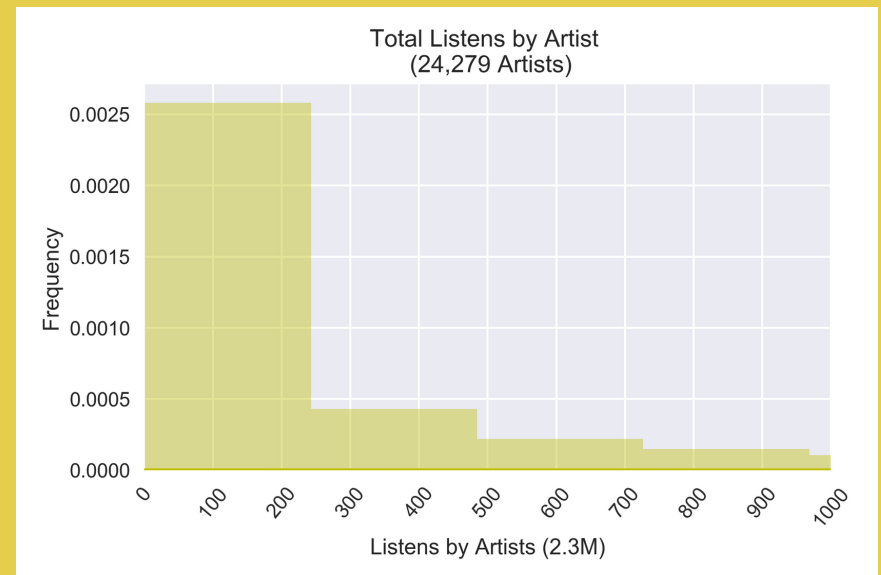
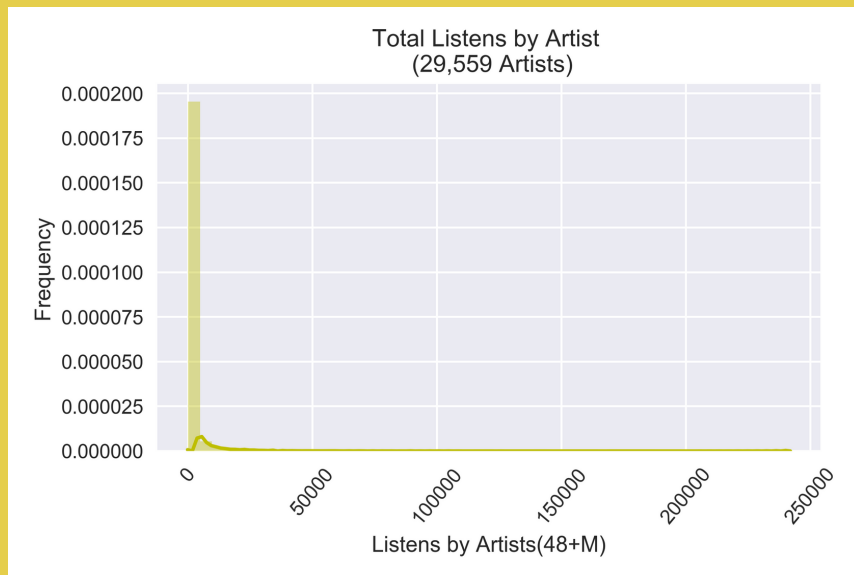
Min = 1, Max = 110,479

# SKEWED ARTIST ENGAGEMENT

Average listens per artist is 1367, standard deviation is 6,498

17.9% of artists have more than a 1,000 total listens.

Artists with listens below 1,000 represent 48% of total listens.



Min = 1, Max = 241,823

# LONG TAILS WITH & VERY SPARSE MATRIX

**The listens by song, user and artist suggest:**

- A small proportion of users skews the distribution and creates a long tail. However, they do not account for a large fraction of listens.
- A minority of songs and artists account for a large fraction of listens.

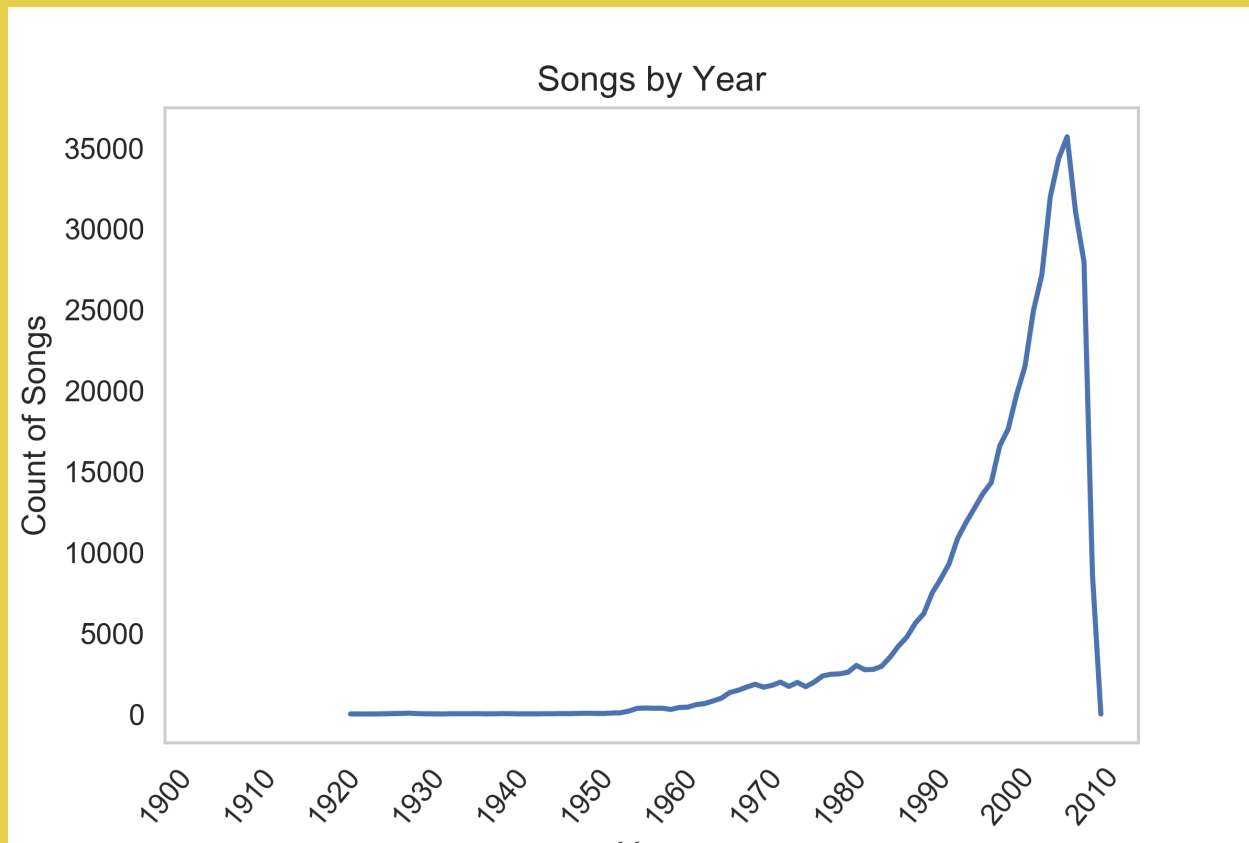
**A very sparse dataset:**

- > Forty-two percent of the available catalog has been listened to.
- > The average user has used the service 6 times and the most engaged user 4,400 times.
- > As a result, the user and song matrix is extremely sparse.

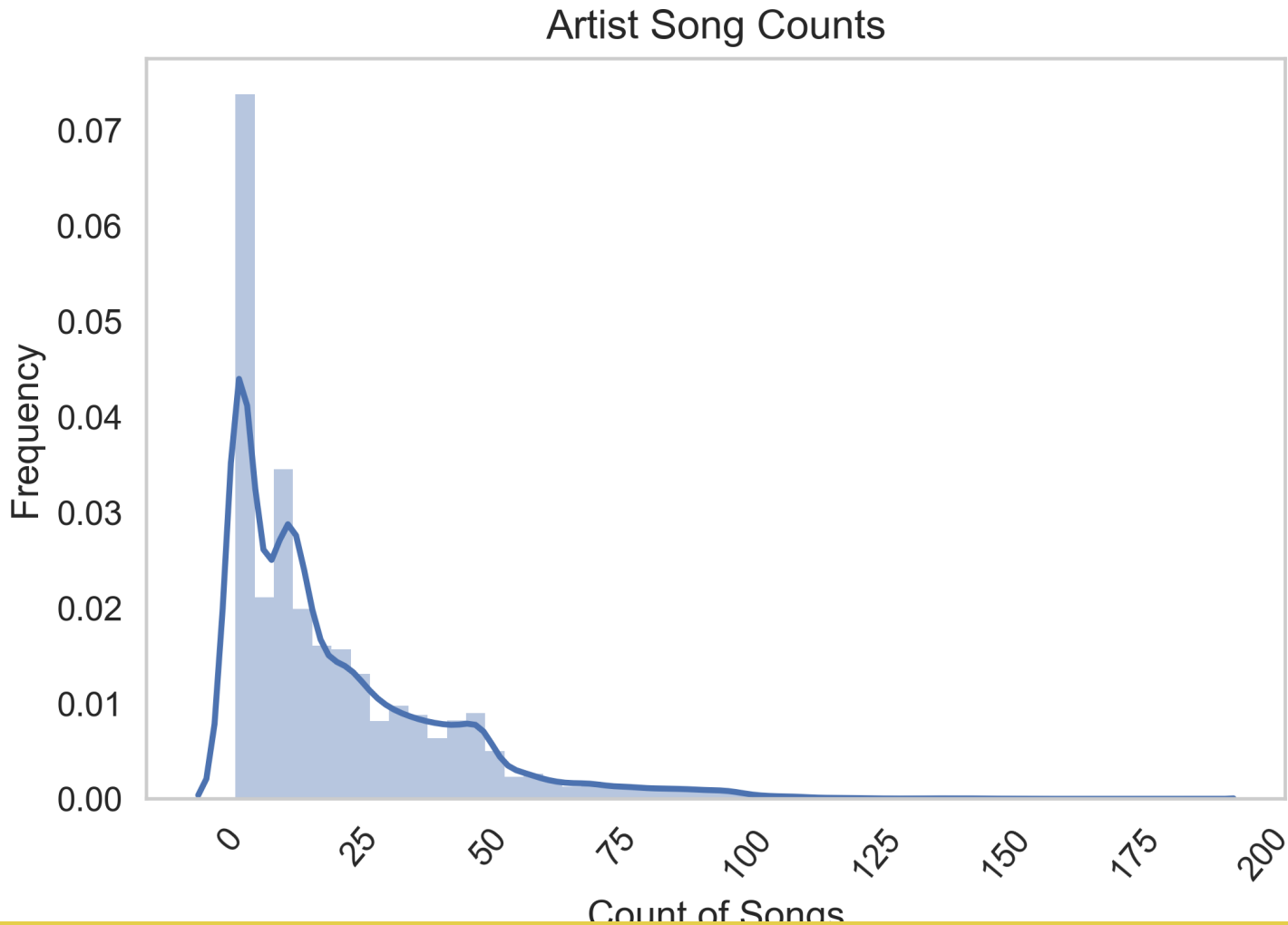


# AVAILABLE SONG YEAR DATA SUGGESTS SONGS BETWEEN 1990 - 2012 ARE OVERLY REPRESENTED IN THE DATASET

49.5% of song years are missing

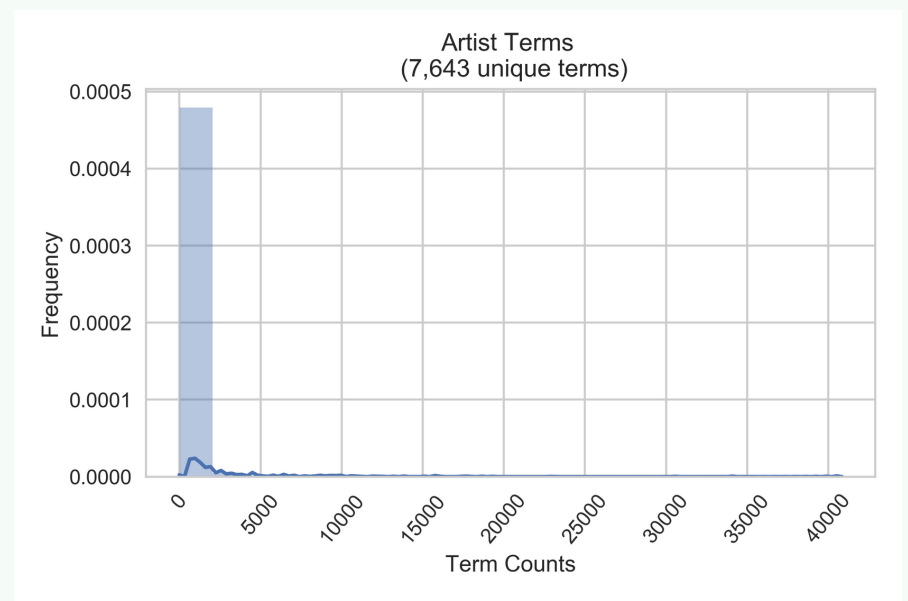
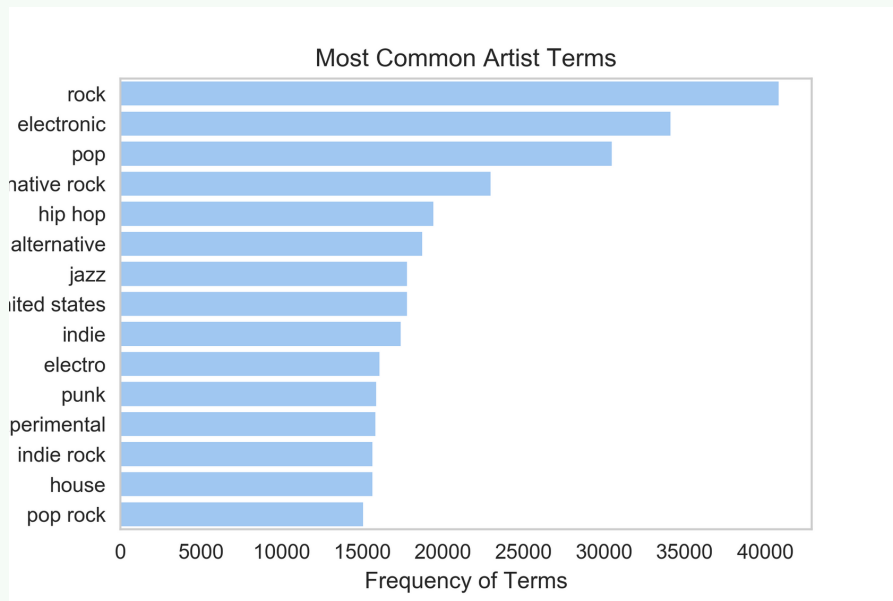


# SOME SKEWNESS IN SONG PER ARTIST DISTRIBUTION

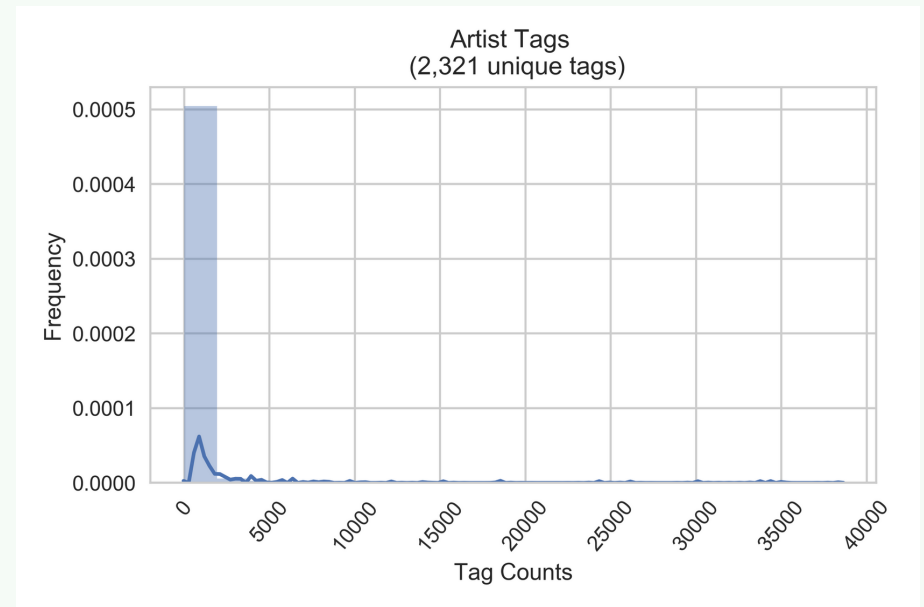
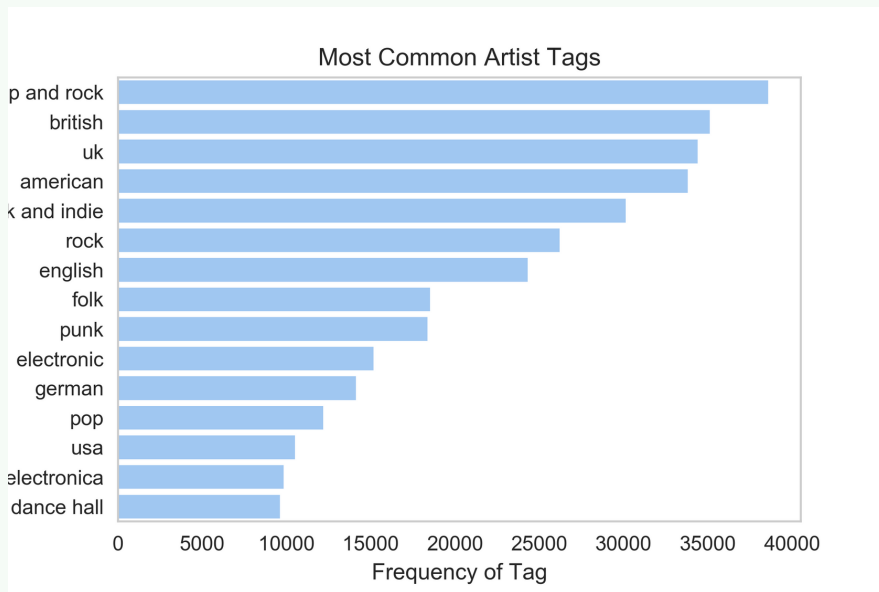




# ARTIST TERMS

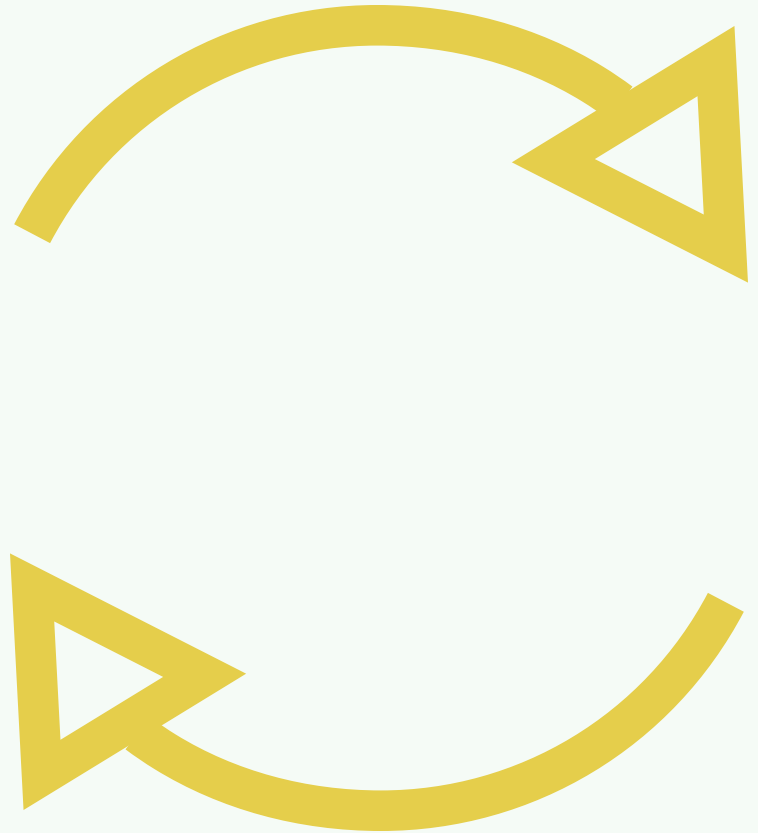


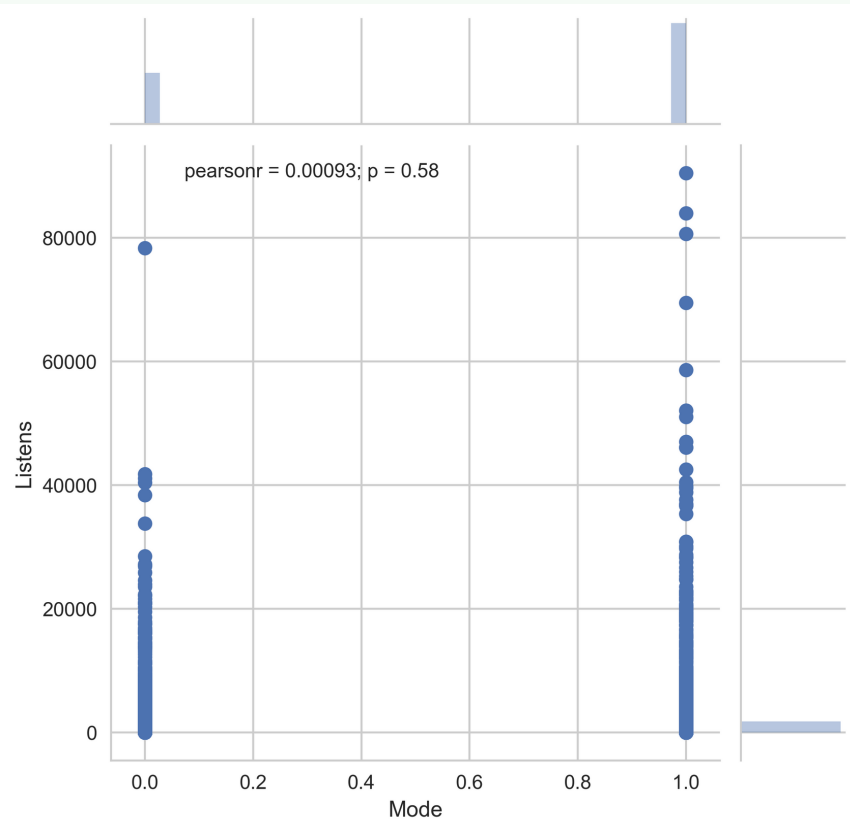
# ARTIST TAGS



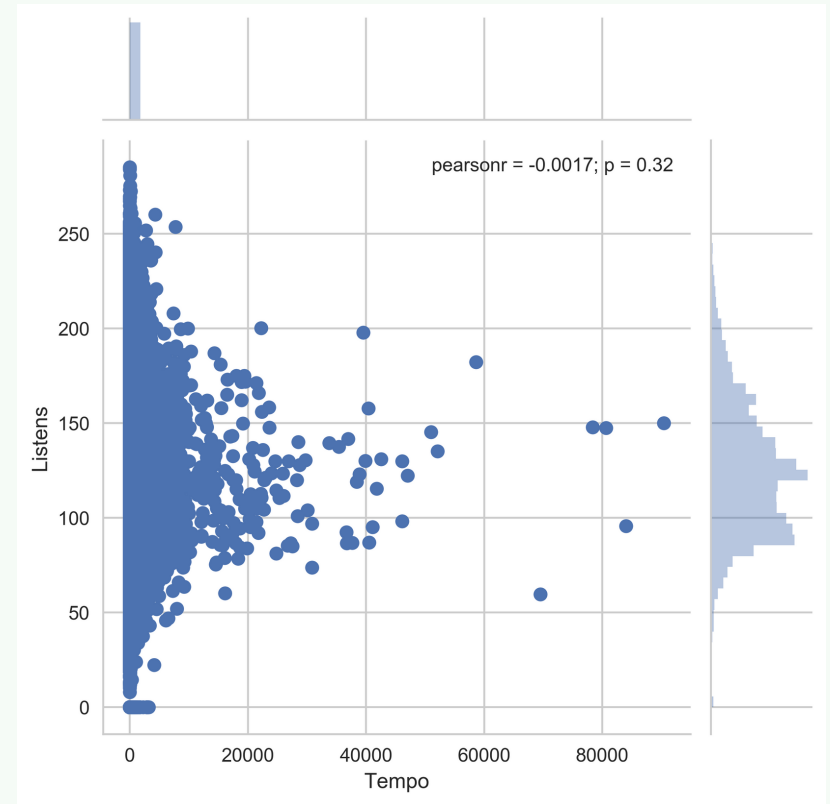
---

# RELATIONSHIPS BETWEEN BASIC SONG FEATURES AND SONG LISTENS

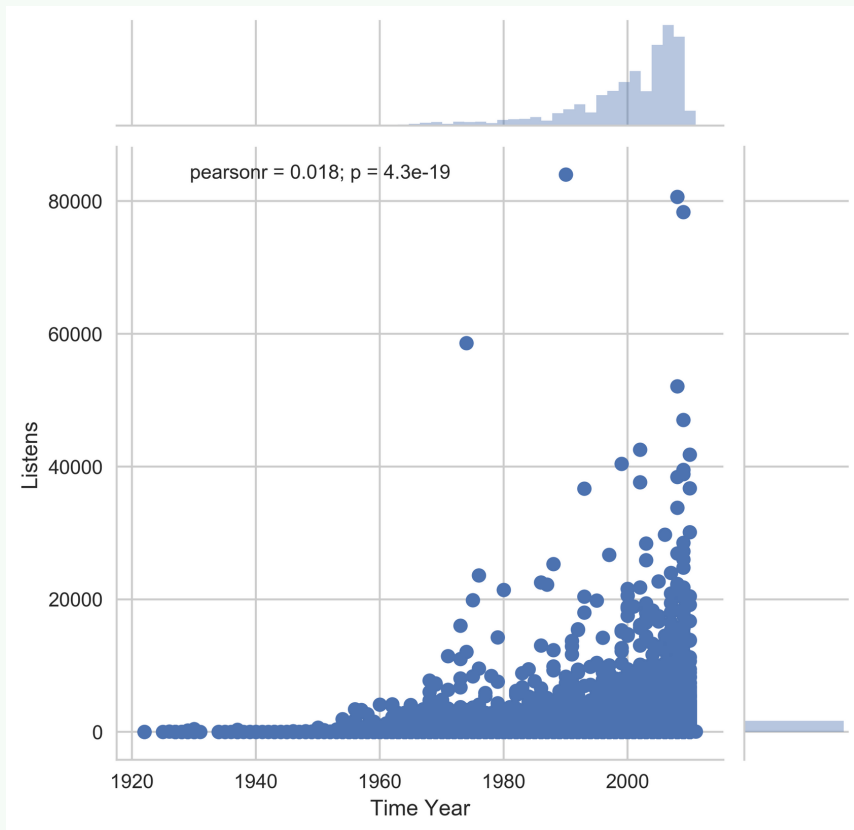




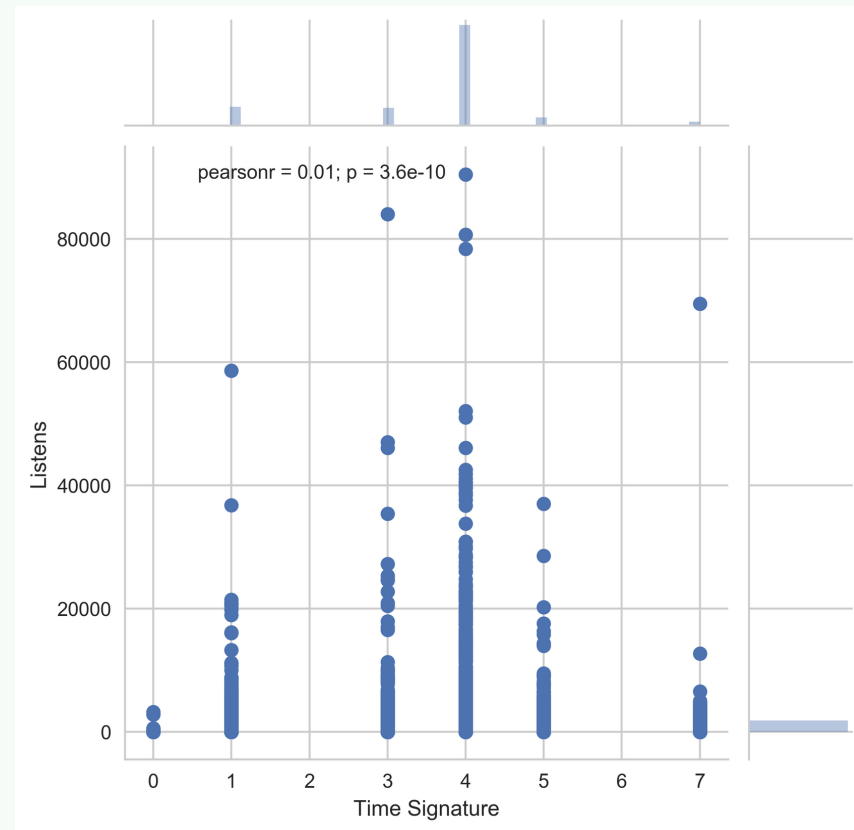
**Song Mode**



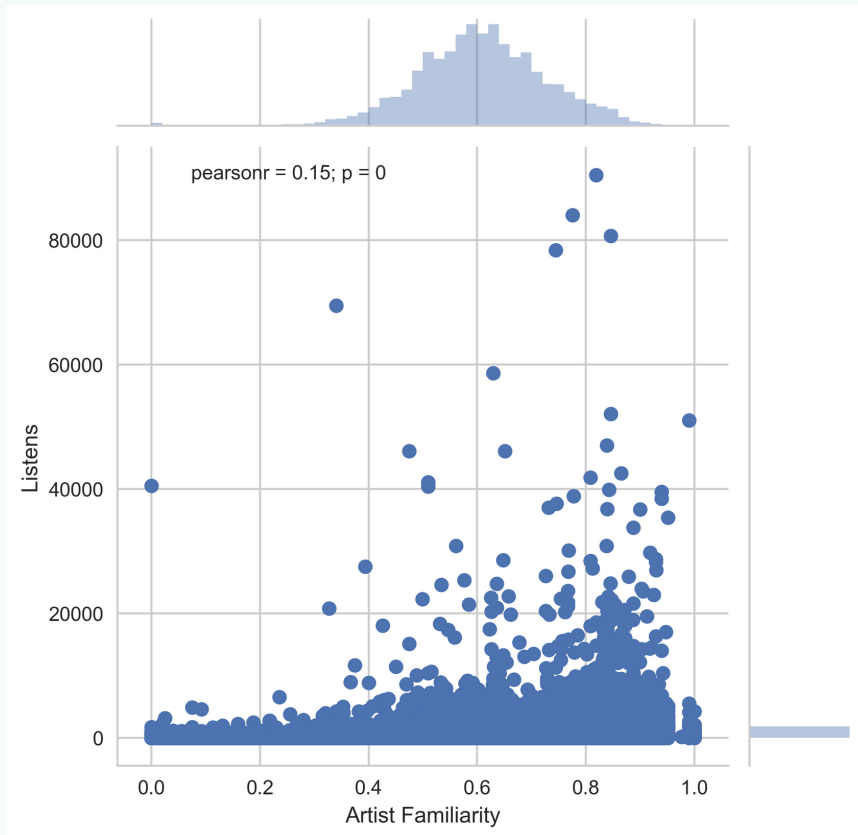
**Song Tempo**



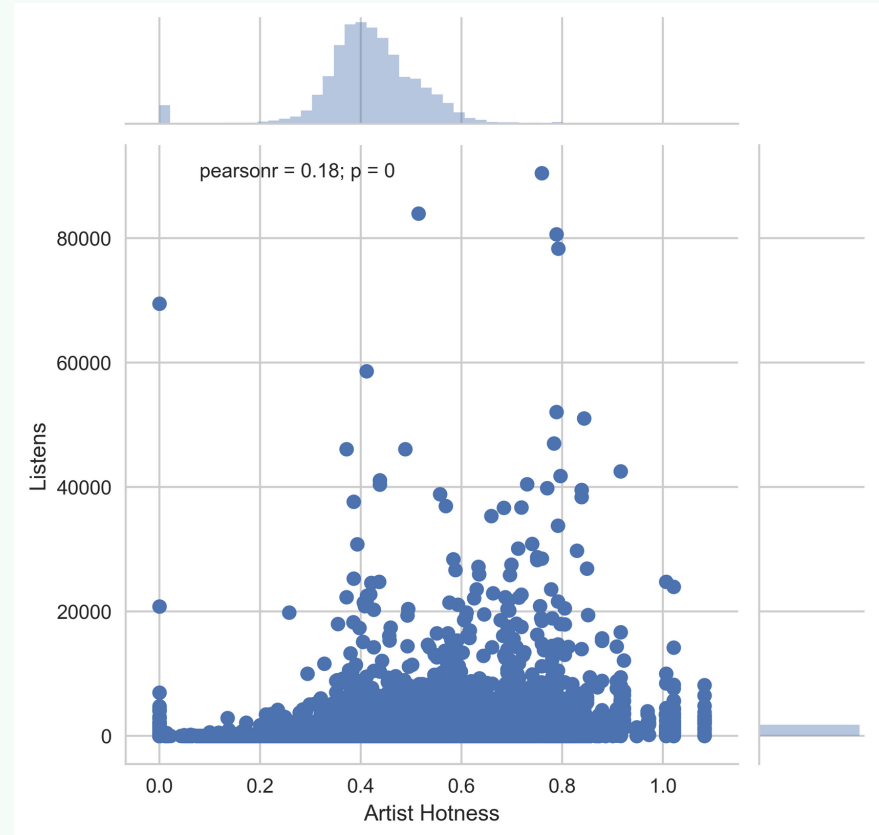
**Year**



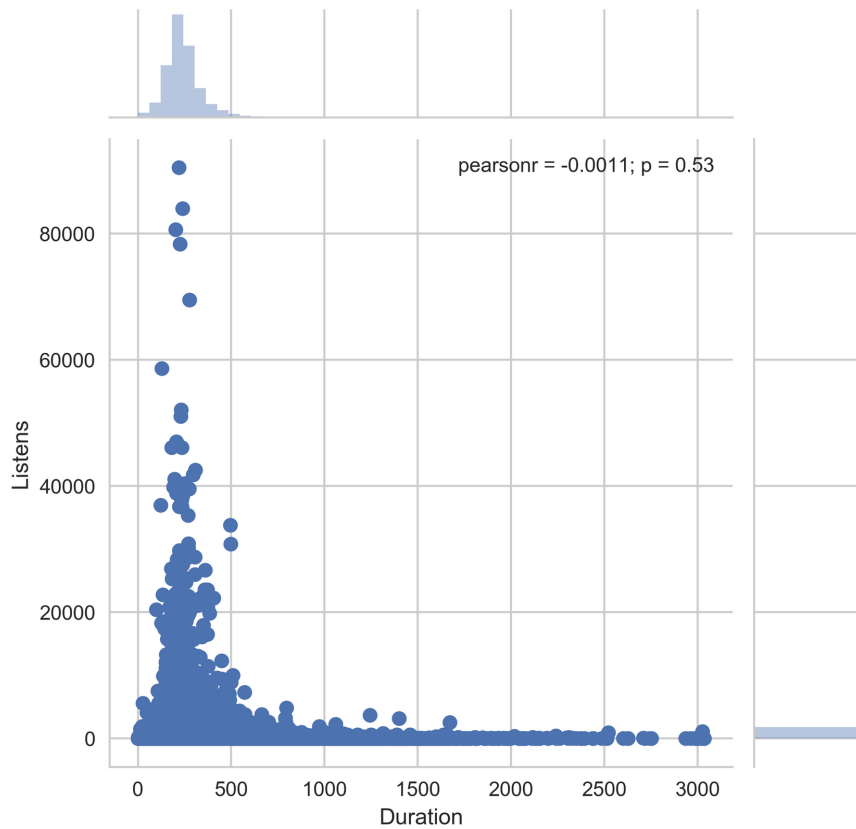
**Time Signature**



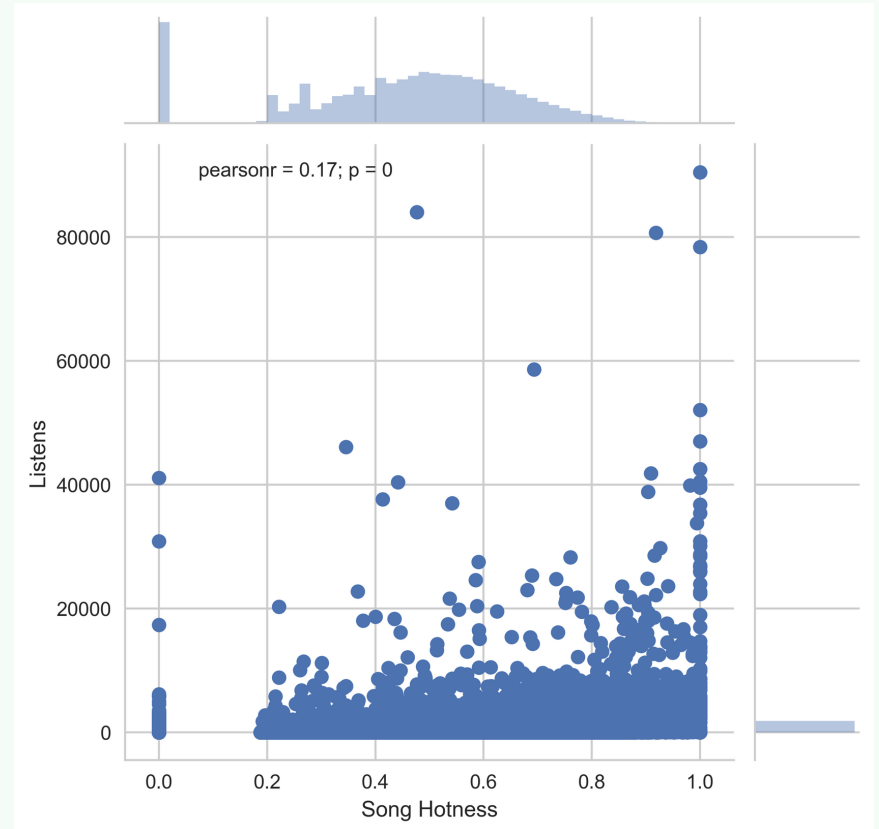
**Artist Familiarity**



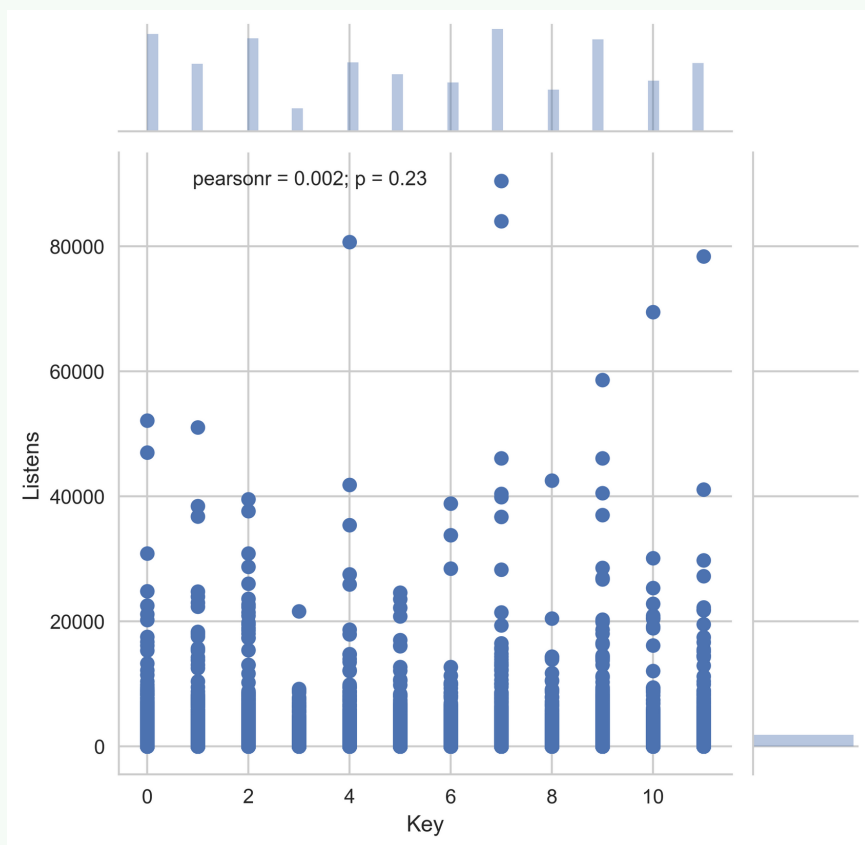
**Artist Hotness**



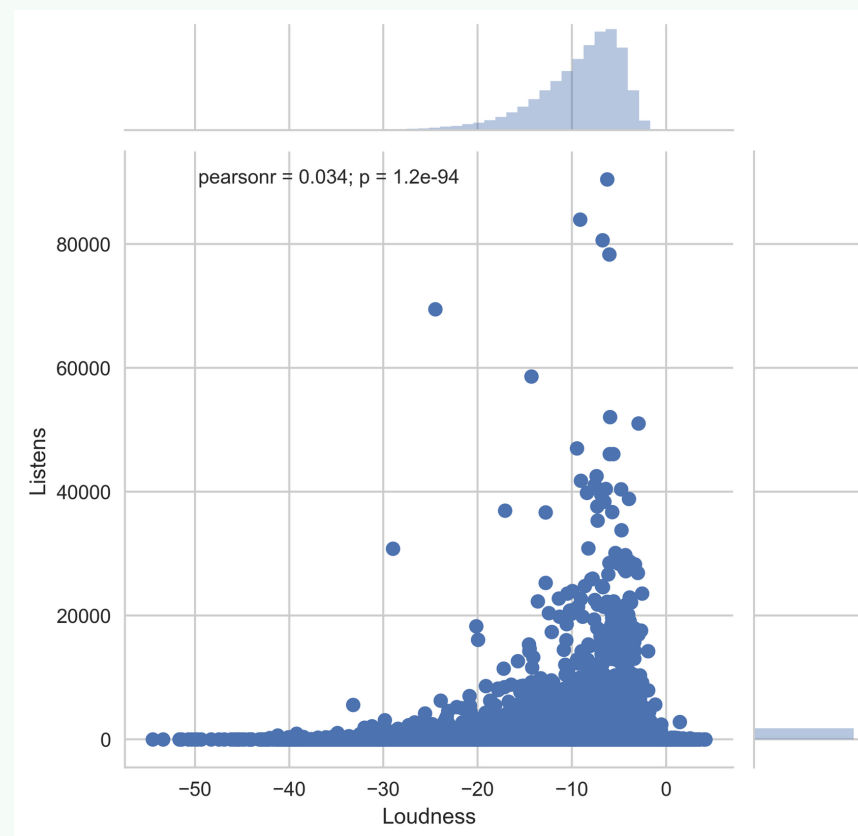
Duration



Song Hotness



Key



Loudness