Analysis of Spotify Music Data

SML310 Final Project

Maximilian Sporer

December 1, 2020

I pledge my honour that this paper represents my own work in accordance with University regulations.

- *Maximillian Sporer*

Introduction

As of 2019, the global music streaming market size was estimated at 20.95 billion USD and has been projected to grow to 24.40 billion USD in 2020[1]. Spotify is the most popular of these streaming platforms, having 36% market share and 286 million monthly active users as per their Q1 2020 report[2]. Part of Spotify's appeal to consumers is the personalization that their platform offers. An example of this personalization is the "Discover Weekly" playlist, which is offered to every user. The description reads "Your weekly mixtape of fresh music. Enjoy new music and deep cuts picked for you. Updates every Monday." These playlists are individualized for each user based on their listening activities on the platform. This project will attempt to gain more insight into how these playlists are generated for each user, specifically by analyzing the metrics by which Spotify defines each song on their platform.

Dataset

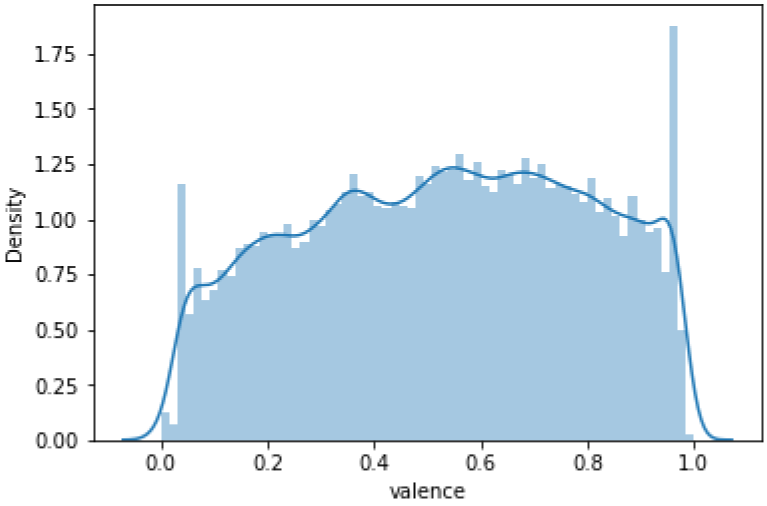
In order to better understand how Spotify is recommending these songs, the following dataset will be analyzed:

Spotify Dataset 1921 - 2020, 160k+ Tracks
https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

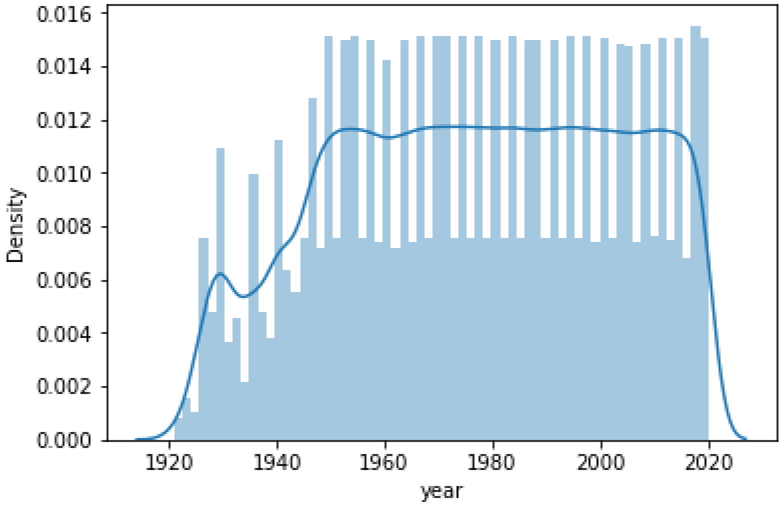
As of its most recent update, this dataset contains 170,653 observations and 18 features. These features[3] (with their density plots) are:

- valence - a value from 0.0 to 1.0 describing how "positive" the track is. Tracks with a higher valence are described as more happy or bright and those with a lower valence are described as more sad or gloomy.

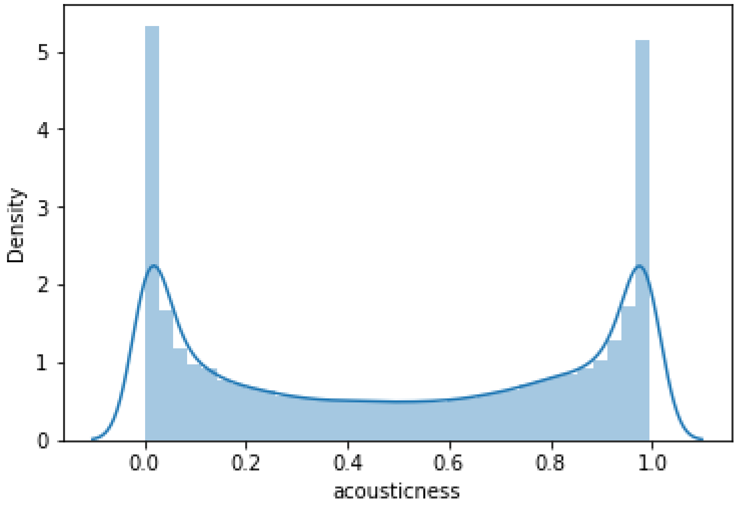
[1] Grand View Research (2020)
[2] Iqbal (2020)
[3] https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/



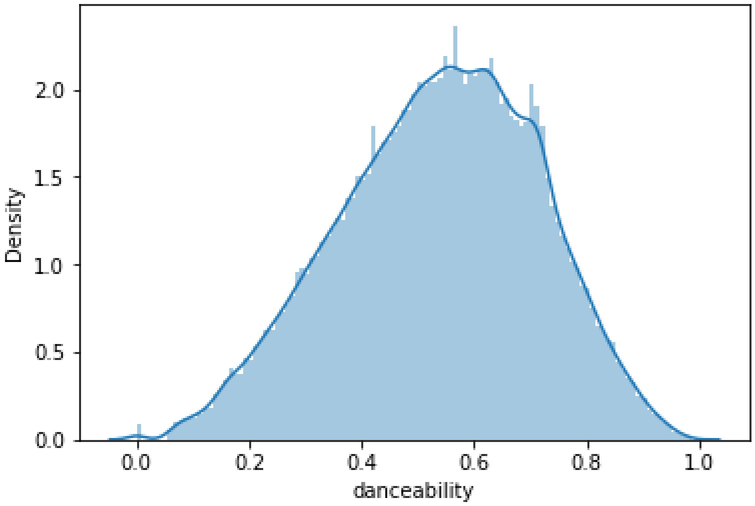
- year - the year in which the track was released.



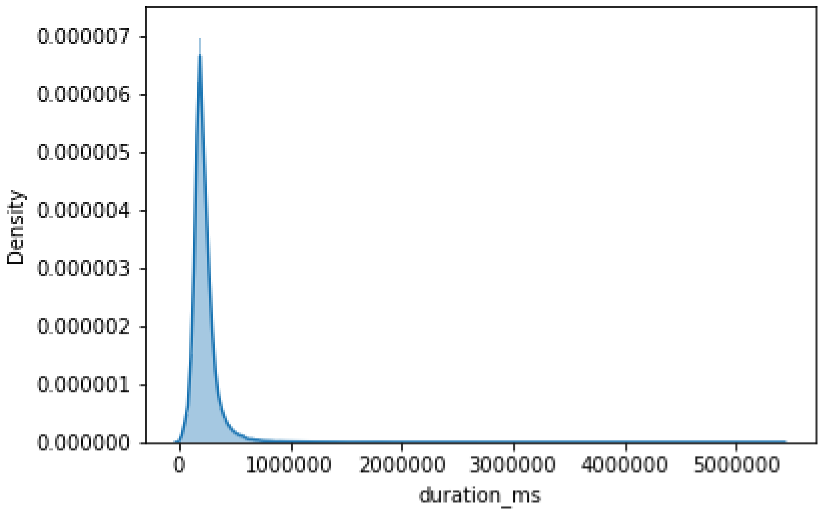
- acousticness - a probability value from 0.0 to 1.0 indicating whether the track is acoustic or not. A value of 1.0 would indicate a high level of confidence that the track is acoustic.



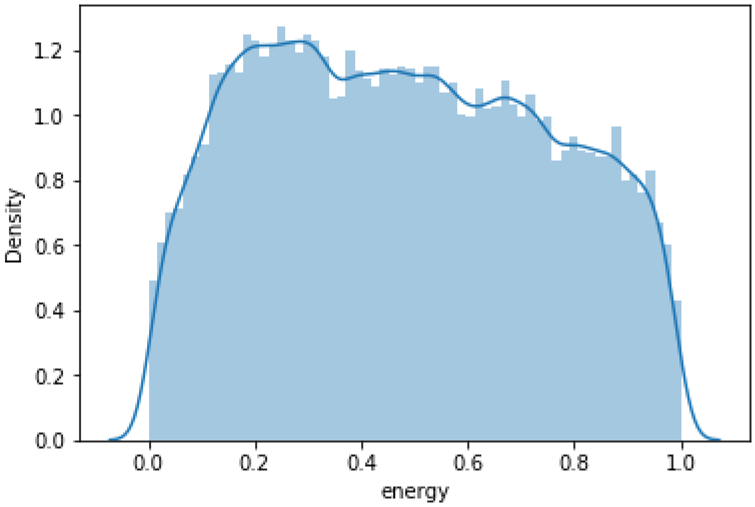
- danceability - a value from 0.0 to 1.0 describing how suitable the track is for dancing. The value is derived from a combination of musical elements such as tempo, rhythm stability, beat strength, and overall regularity. A value of 1.0 would indicate that the track is very suitable for dancing.



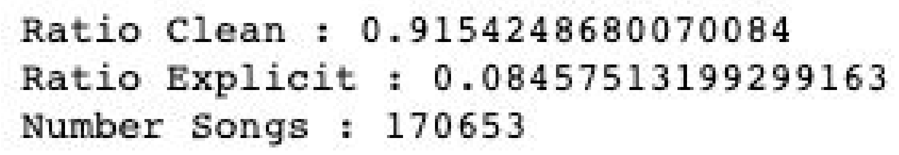
- duration_ms - the length of the track measured in milliseconds



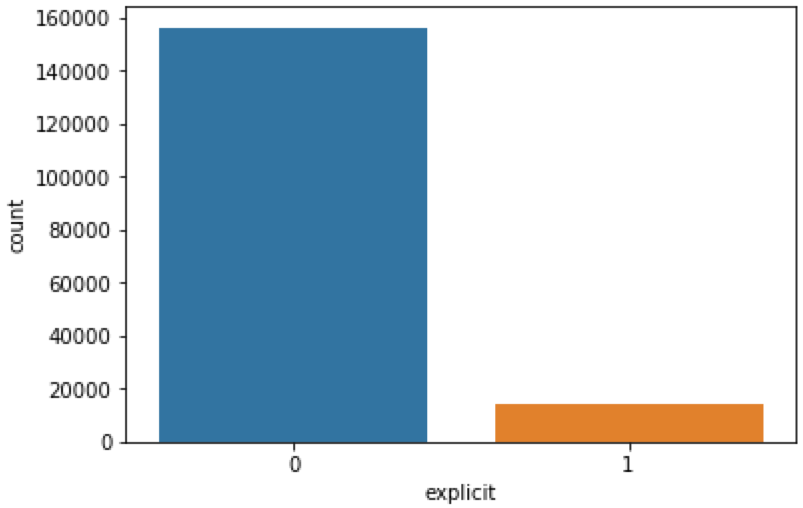
- energy - a value from 0.0 to 1.0 representing the measure of the tracks intensity and activity. Higher energy values correspond to tracks that feel faster, louder, and more noisy (like death metal, for example).



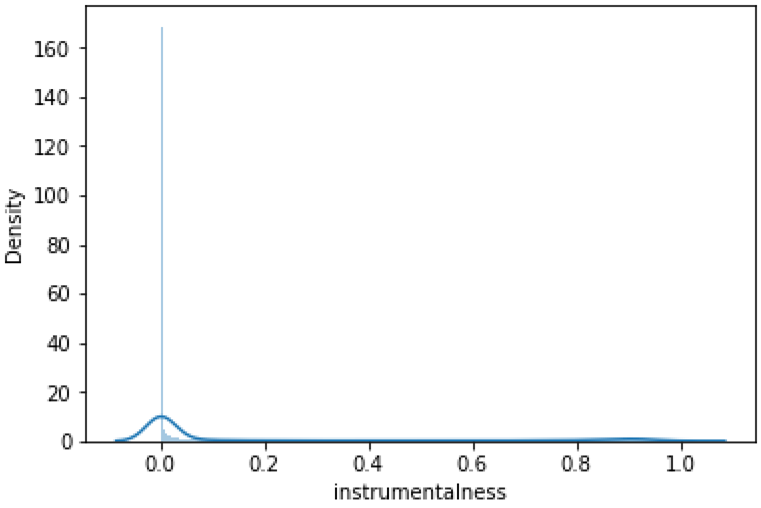
- explicit - a boolean indicating whether the track contains expletives. A value of 1 indicates that the track is explicit.



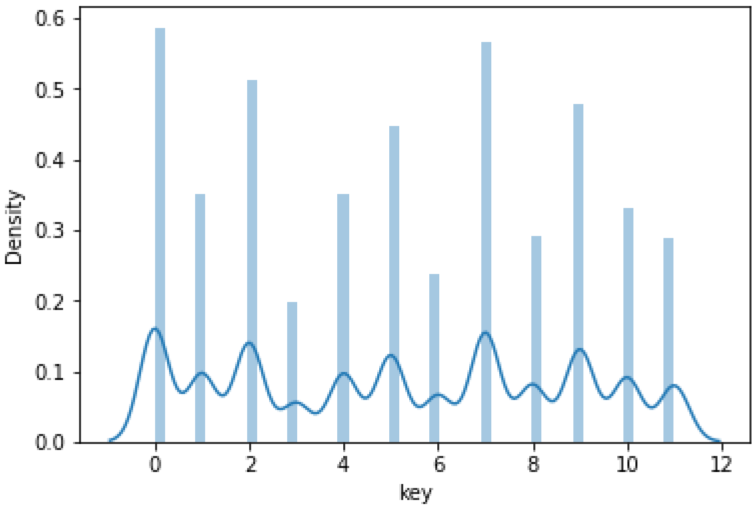
```
Ratio Clean : 0.9154248680070084
Ratio Explicit : 0.08457513199299163
Number Songs : 170653
```



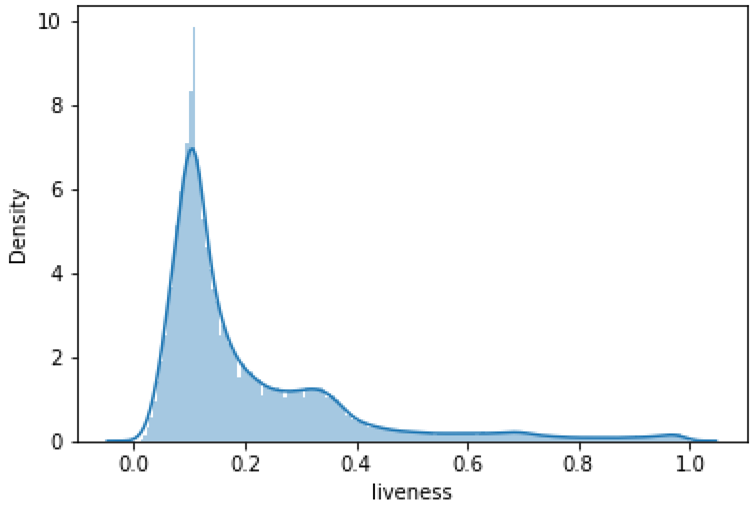
- instrumentalness - a probability value from 0.0 to 1.0 indicating whether the track likely contains vocals or not. A value of 1.0 would indicate a high level of confidence that the track contains no vocals.



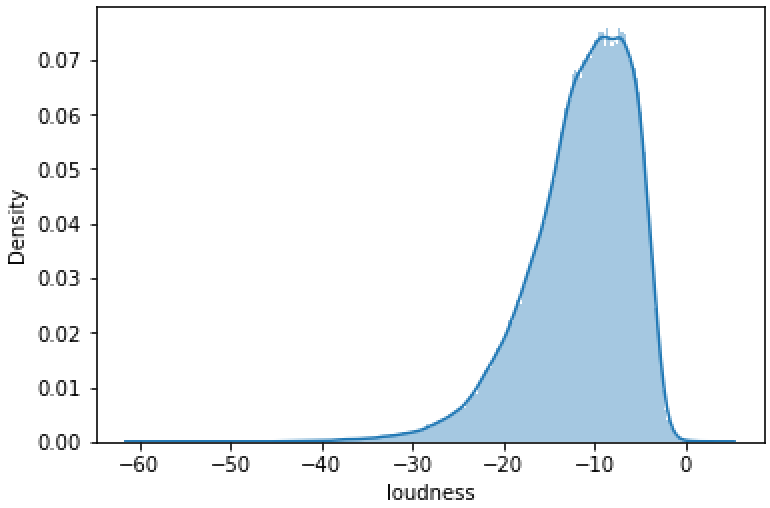
- key - the estimated overall key of the track. Integer values from 0 to 11 map to pitches in standard Pitch Class notation (0 = C , 1 = C # / D ♭ , 2 = D, etc.)



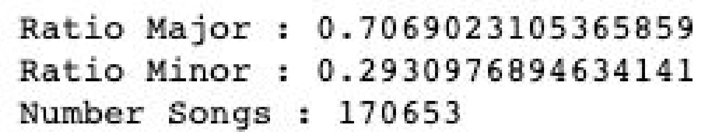
- liveness - a probability value from 0.0 to 1.0 indicating whether an audience was likely present during the tracks recording. A value of 1.0 would indicate a high level of confidence that an audience was present.



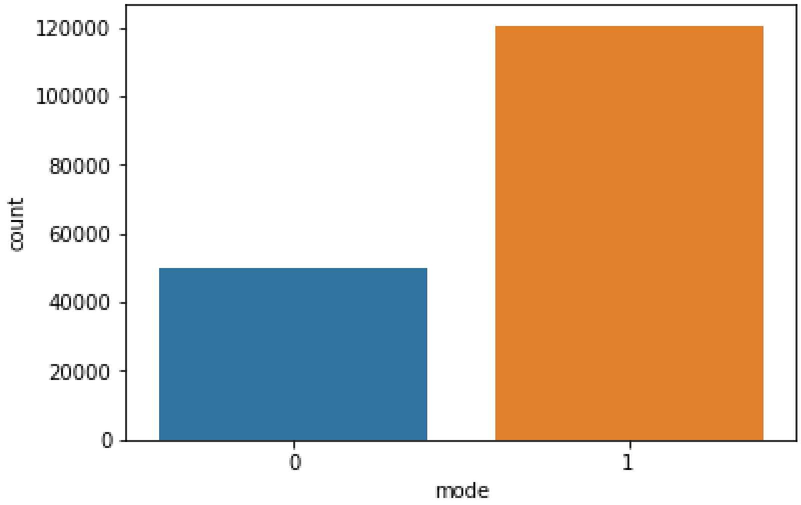
- loudness - a measure of the volume of the track measured in decibels. Values range from -60.0 to 3.86.



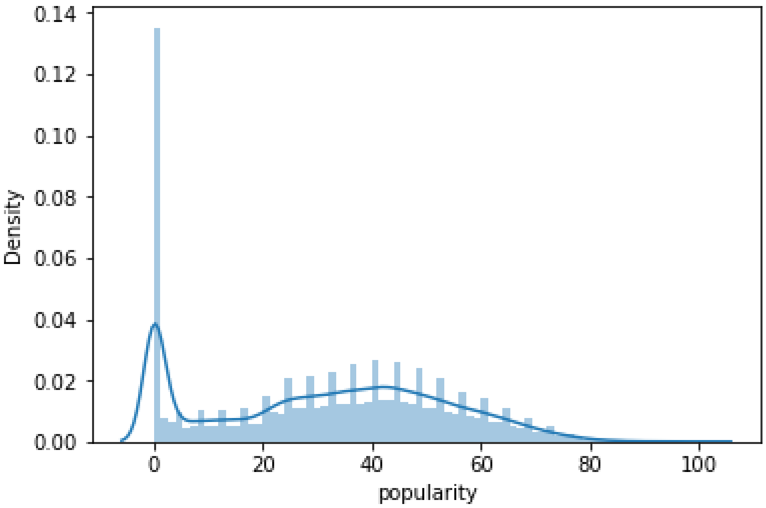
- mode - a boolean indicating the modality of the track, i.e. whether the track is major or minor. A value of 1 indicates that the track is major.



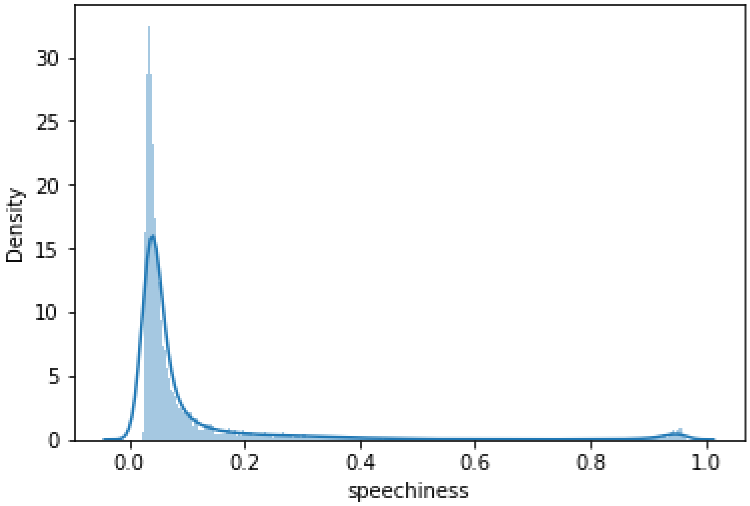
```
Ratio Major : 0.7069023105365859
Ratio Minor : 0.2930976894634141
Number Songs : 170653
```



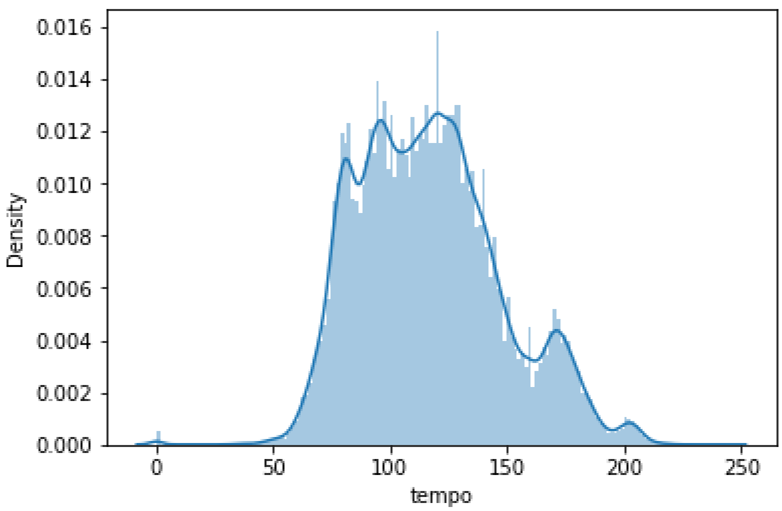
- popularity - a value between 0 to 100 indicating the popularity of the track, with 100 being the most popular. Popularity is calculated using the number of times the track has been played and how recently these plays occurred.



- speechiness - a value from 0.0 to 1.0 indicating the presence of spoken words. This differs from instrumentalness in that speechiness captures tracks along the lines of talk shows, audio books, poetry, etc. as opposed to vocals in a song. A value of 1.0 indicates a high level of speechiness.



- tempo - the estimated tempo of the track in beats per minute



- release_date - the date the track was released
- name - the title of the track
- artists - the artists credited for the track
- id - the id of the track in the Spotify database

Research Questions

This paper will explore the topic of interest guided by the following research question:

> Which musical features are good predictors of a track's popularity?

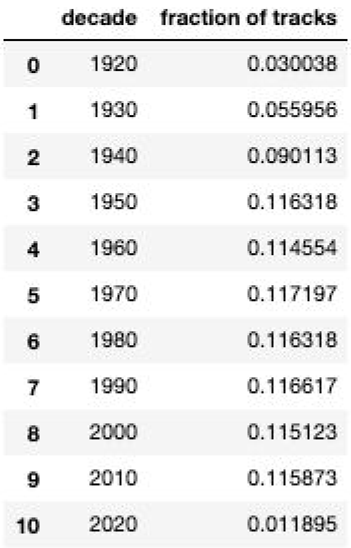
All data analysis, visualization, and modeling were performed using Python in a Jupyter Notebook. The notebook (ipynb file) has been submitted as a separate file and any methods / calculations that are not explained in this paper can be found in the notebook.

In the remaining sections, a track's popularity (a value from 0 to 100) will be estimated based on the musical features of the tracks using Ordinary Least Squares Regression. Before diving into the modeling, I will explain some of the exploratory data analysis I performed.

Exploratory Data Analysis - Distributions

The distribution plots for each feature are included in the section titled “Dataset”. Some of these distributions are noteworthy because they may have implications on later analysis and/or modeling.

The distributions of the year feature (fig. 2 pp. 2) shows that each year does not contribute an equal number of tracks to the dataset.

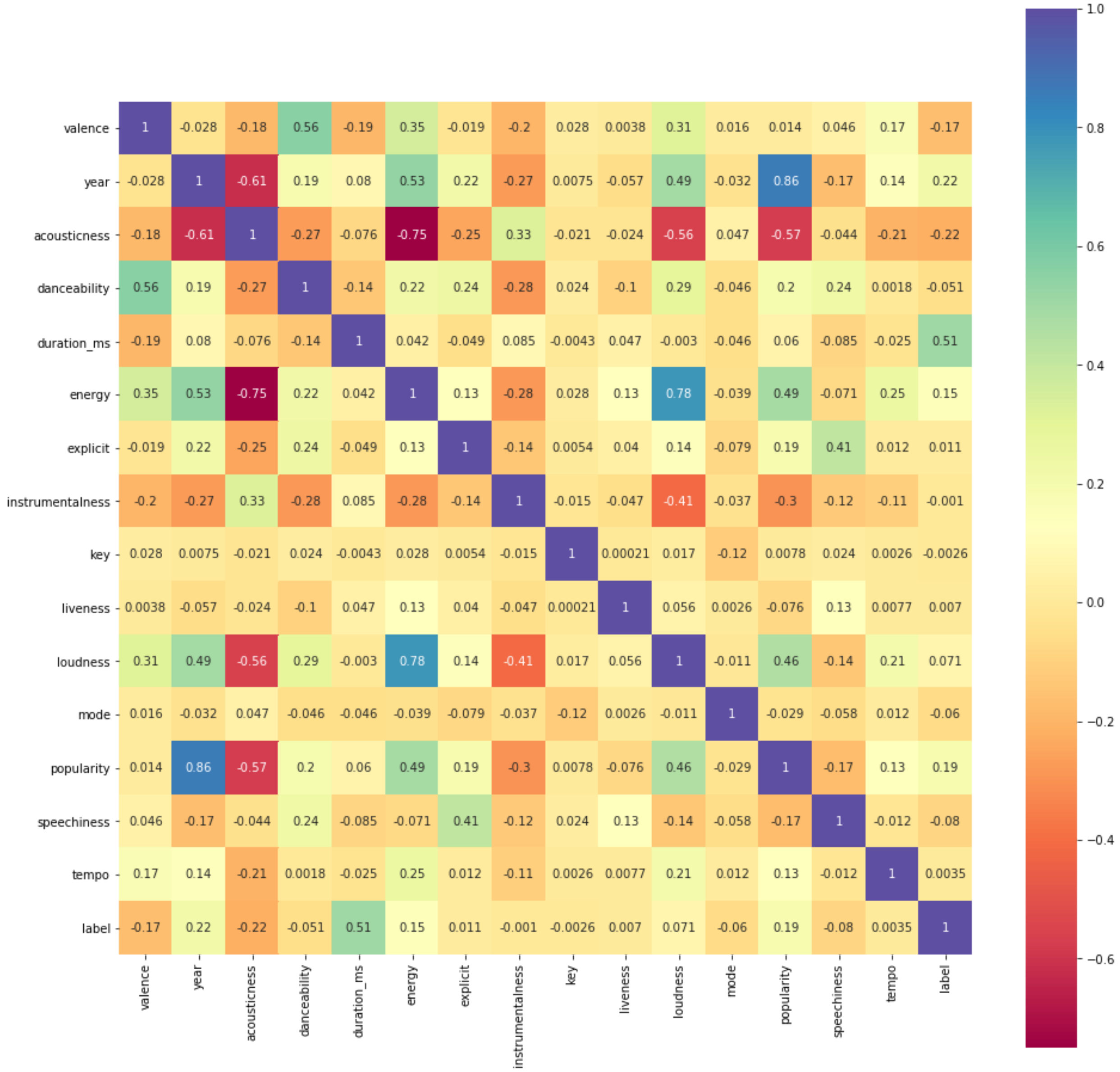


	decade	fraction of tracks
0	1920	0.030038
1	1930	0.055956
2	1940	0.090113
3	1950	0.116318
4	1960	0.114554
5	1970	0.117197
6	1980	0.116318
7	1990	0.116617
8	2000	0.115123
9	2010	0.115873
10	2020	0.011895

This may be due to the fact that more recent tracks are already digitized and uploaded to Spotify. It is probably more difficult to digitize and upload tracks from the 1920's.

Also of note is the distributions of the instrumentalness and speechiness features (fig. 2 pp. 5 and fig. 2 pp. 8, respectively), which follow a similar pattern. Both distributions show centers close to 0.1 (0.167 and 0.098, respectively). This indicates that the dataset is comprised mostly of musical tracks but also contains a considerable amount of spoken word tracks. For example, two of the tracks having a speechiness value greater than 0.95 are "If I Ran the Zoo" by Dr. Seuss and "Exorcist - Remastered Version" by Richard Pryor.

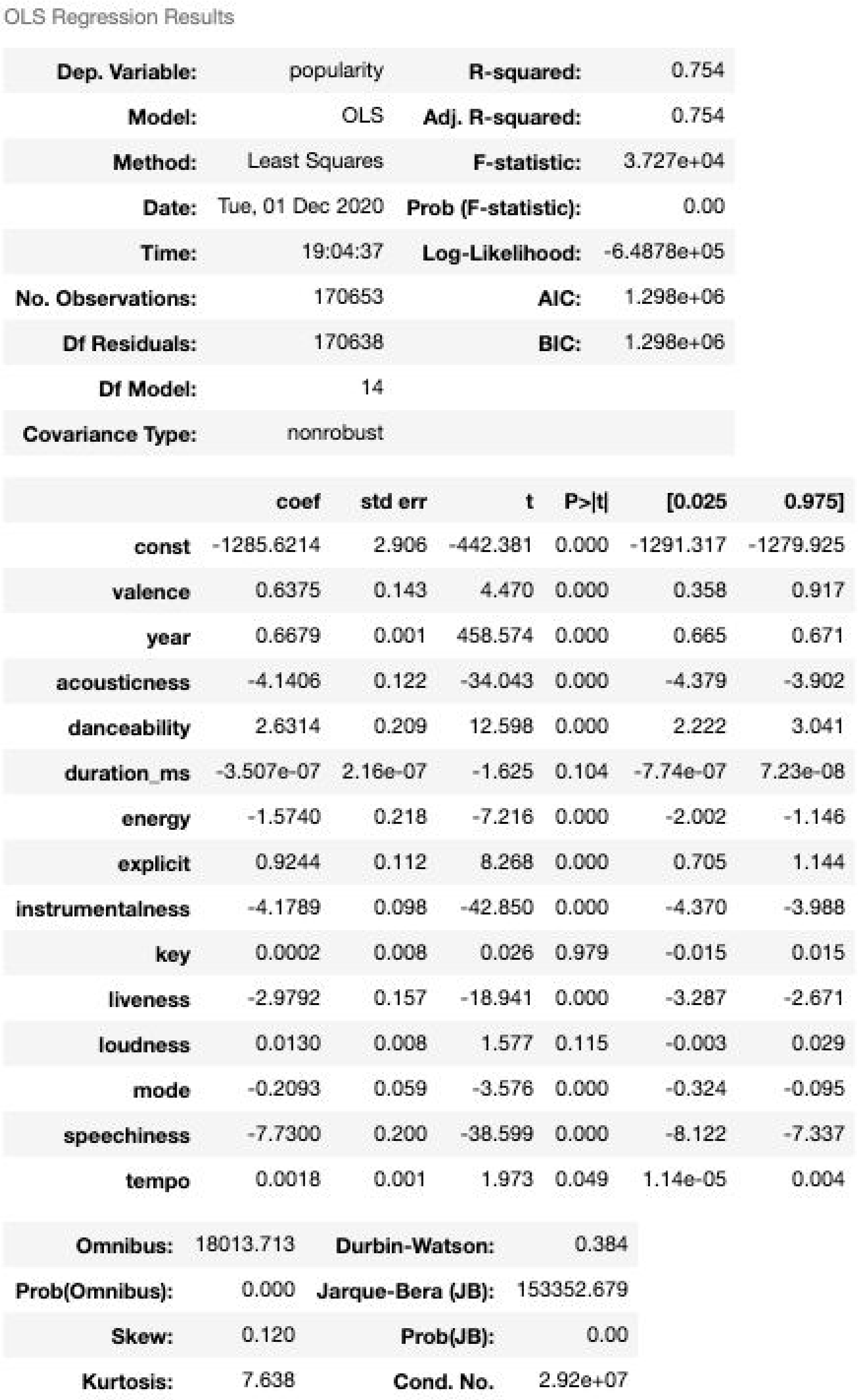
Exploratory Data Analysis - Correlation Matrix



The plot above shows a strong correlation between year and popularity, which makes sense. One would expect recently released music to make up the bulk of plays on Spotify platform. However, this may not be very interesting to our analysis because you don't need a machine learning model to tell you this. Popularity also has high positive correlations with energy and loudness and high negative correlations with acousticness and instrumentalness.

OLS Regression - Results and Analysis

An Ordinary Least Squares linear regression was performed to model the popularity of a track. Below are the results using all available data:



OLS Regression Results

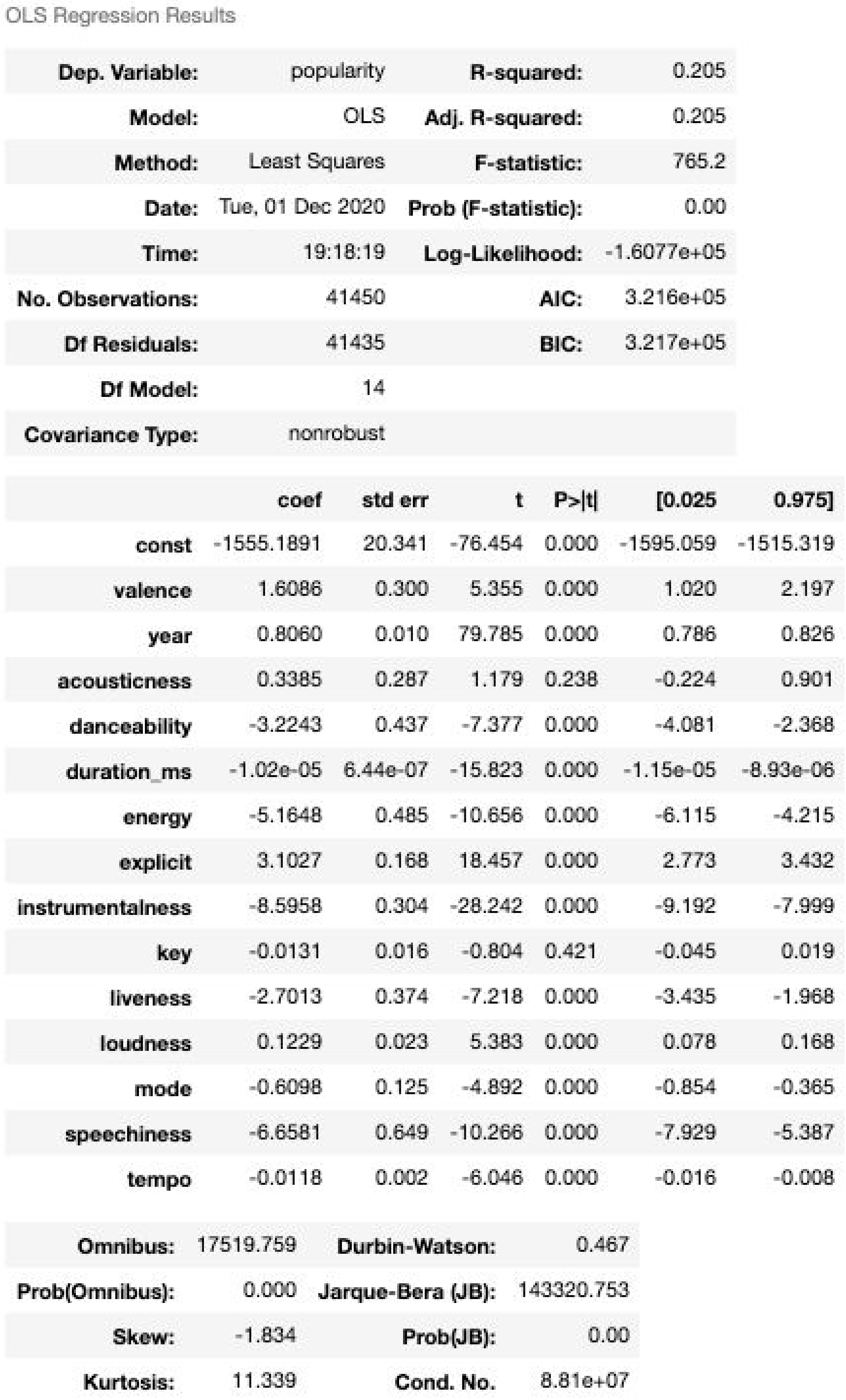
Dep. Variable:	popularity	**R-squared:**	0.754
Model:	OLS	**Adj. R-squared:**	0.754
Method:	Least Squares	**F-statistic:**	3.727e+04
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	19:04:37	**Log-Likelihood:**	-6.4878e+05
No. Observations:	170653	**AIC:**	1.298e+06
Df Residuals:	170638	**BIC:**	1.298e+06
Df Model:	14		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-1285.6214	2.906	-442.381	0.000	-1291.317	-1279.925
valence	0.6375	0.143	4.470	0.000	0.358	0.917
year	0.6679	0.001	458.574	0.000	0.665	0.671
acousticness	-4.1406	0.122	-34.043	0.000	-4.379	-3.902
danceability	2.6314	0.209	12.598	0.000	2.222	3.041
duration_ms	-3.507e-07	2.16e-07	-1.625	0.104	-7.74e-07	7.23e-08
energy	-1.5740	0.218	-7.216	0.000	-2.002	-1.146
explicit	0.9244	0.112	8.268	0.000	0.705	1.144
instrumentalness	-4.1789	0.098	-42.850	0.000	-4.370	-3.988
key	0.0002	0.008	0.026	0.979	-0.015	0.015
liveness	-2.9792	0.157	-18.941	0.000	-3.287	-2.671
loudness	0.0130	0.008	1.577	0.115	-0.003	0.029
mode	-0.2093	0.059	-3.576	0.000	-0.324	-0.095
speechiness	-7.7300	0.200	-38.599	0.000	-8.122	-7.337
tempo	0.0018	0.001	1.973	0.049	1.14e-05	0.004

Omnibus:	18013.713	**Durbin-Watson:**	0.384
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	153352.679
Skew:	0.120	**Prob(JB):**	0.00
Kurtosis:	7.638	**Cond. No.**	2.92e+07

At an alpha level of 0.001, the statistical significant predictors of a track's popularity with a positive correlation are valence, year, danceability, and explicit. Those with a negative correlation are acousticness, energy, instrumentalness, liveness, mode, and speechiness. From these results, we can surmise that Spotify users currently prefer happy, recently released music that contain expletives and that you can dance to!

However, because I discovered a high correlation between year and popularity during the exploratory data analysis, I thought it may be illuminative to run an OLS regression only on tracks released on or after the year 2000. Below are the results:



OLS Regression Results

Dep. Variable:	popularity	**R-squared:**	0.205
Model:	OLS	**Adj. R-squared:**	0.205
Method:	Least Squares	**F-statistic:**	765.2
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	19:18:19	**Log-Likelihood:**	-1.6077e+05
No. Observations:	41450	**AIC:**	3.216e+05
Df Residuals:	41435	**BIC:**	3.217e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-1555.1891	20.341	-76.454	0.000	-1595.059	-1515.319
valence	1.6086	0.300	5.355	0.000	1.020	2.197
year	0.8060	0.010	79.785	0.000	0.786	0.826
acousticness	0.3385	0.287	1.179	0.238	-0.224	0.901
danceability	-3.2243	0.437	-7.377	0.000	-4.081	-2.368
duration_ms	-1.02e-05	6.44e-07	-15.823	0.000	-1.15e-05	-8.93e-06
energy	-5.1648	0.485	-10.656	0.000	-6.115	-4.215
explicit	3.1027	0.168	18.457	0.000	2.773	3.432
instrumentalness	-8.5958	0.304	-28.242	0.000	-9.192	-7.999
key	-0.0131	0.016	-0.804	0.421	-0.045	0.019
liveness	-2.7013	0.374	-7.218	0.000	-3.435	-1.968
loudness	0.1229	0.023	5.383	0.000	0.078	0.168
mode	-0.6098	0.125	-4.892	0.000	-0.854	-0.365
speechiness	-6.6581	0.649	-10.266	0.000	-7.929	-5.387
tempo	-0.0118	0.002	-6.046	0.000	-0.016	-0.008

Omnibus:	17519.759	**Durbin-Watson:**	0.467
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	143320.753
Skew:	-1.834	**Prob(JB):**	0.00
Kurtosis:	11.339	**Cond. No.**	8.81e+07

At an alpha level of 0.001, the statistical significant predictors of a track's popularity with a positive correlation are valence, year, explicit, and loudness. Those with a negative correlation are danceability, energy, instrumentalness, liveness, mode, speechiness, and tempo.

Comparing OLS results run on the entire dataset versus the dataset pared down by tracks released on or after 2000, we can first notice a drop in R-squared from 0.754 to 0.205. This may indicate that as we move to a more recent dataset, listening trends among Spotify users may be explained by other factors beyond raw musical features. What this dataset fails to address is the social aspect to music consumption that platforms like Spotify introduce. Spotify doubles as a form of social media in that users can follow their friends and see what songs they are listening to. By only considering raw musical features, the above modeling does not capture how interactions between users affects the popularity of tracks.

Conclusion

The above analysis is ultimately oversimplified. Peoples' interaction with music is certainly multifaceted and the popularity of a track cannot be fully predicted based on the track's raw musical features and metadata alone. I believe this dataset fully captures the intrinsic properties of each track that lend themselves to being enjoyed. But what it fails to address are the extrinsic properties that convince a person to listen to a certain song, such as what they're friends are listening to, who they are with, what they are doing, and even what the weather is like outside.

Works Cited

Grand View Research. (2020, October). Music Streaming Market Size & Share Report, 2020-2027. Retrieved December 01, 2020, from https://www.grandviewresearch.com/industry-analysis/music-streaming-market

Iqbal, M. (2020, October 30). Spotify Usage and Revenue Statistics (2020). Retrieved December 01, 2020, from https://www.businessofapps.com/data/spotify-statistics/

Spotify for Developers, Get Audio Features for a Track : https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/