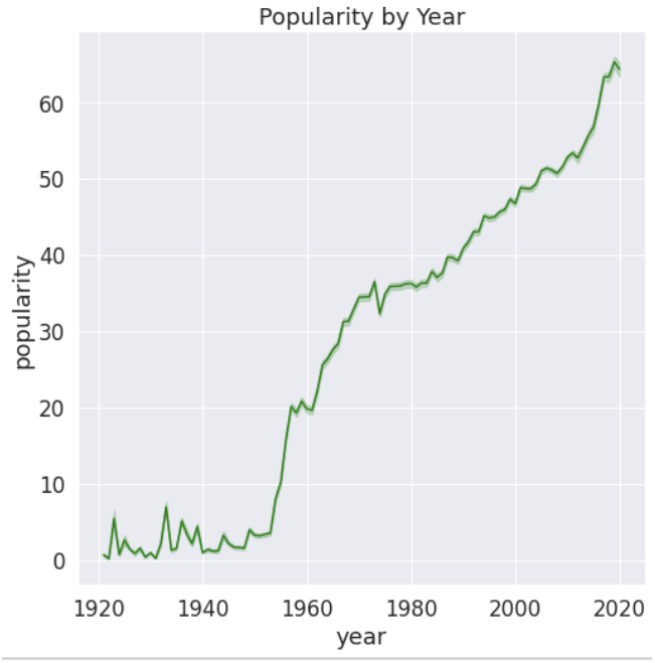
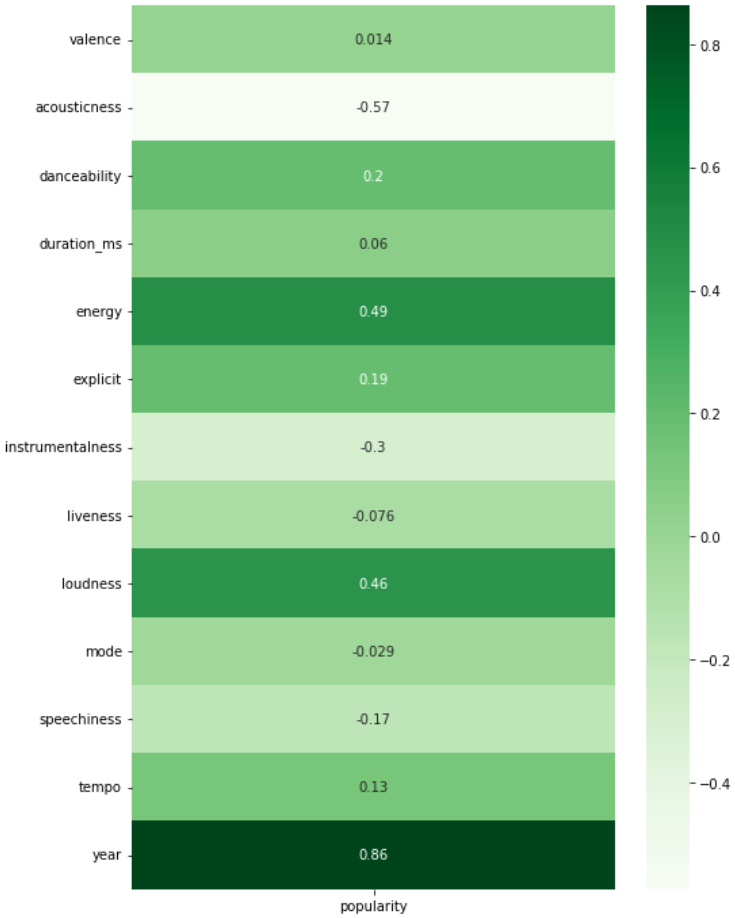
Spotify is a multimedia platform that gives you access to millions of different songs and other content from creators all over the world all from an app on your phone. In this article, we will analyze a dataset, provided by Spotify API, that includes the name of a song, the artist, its release date,the popularity of the song and features of the track.

The focus of this article is showing how we predicted popularity based on different features through a linear model and built a model that creates a playlist based on a listener's song of choice and its features. The features of our dataset were: valence, acousticness, danceability, energy, explicit, instrumentalness, liveness, loudness, mode, speechiness, tempo,year, duration in ms,name, artists, ID, key, release_date, and popularity. The data was cleaned by dropping "ID" and "key" columns, and non-numerical values were encoded into integers for better analysis. Duplicates were also dropped from the original dataset which included 170,653 rows and 19 columns originally and 168,694 rows and 17 columns were left after cleaning. The linear model was split into three popularity data sets concatenated together; above 60, below and equal to 60 and above 0, and below and equal to 0. That way we balanced the data by making these three subsets and sampling them to the lowest value counts out of the three. The reasoning behind undersampling was that after taking a deeper look into "popularity distribution we noticed 16% of the values were 0, and only 9% of the values were above 9%. After splitting the "release_date" column with the potential of changing it into a ‘datetime’ object we found that 30% of the data did not have a month or day of the release. We also split the "artists" column into two columns: one with the first artist mentioned and the second one containing the rest of the artists. Our idea was to improve our linear model by using encoded artists either all of them together or separated. We found that 79.5 % of songs only had one artist so we decided to use the original ‘artists’ column throughout our analysis.

Since popularity will be the main focus of our analysis, we wanted to shed some light on how popularity is measured. Popularity is measured by the total number of plays the track has had and how recent those plays are. Due to this, popularity is heavily

influenced by the year of song release because newer songs have a higher count of more recent plays, as seen in the figure below.



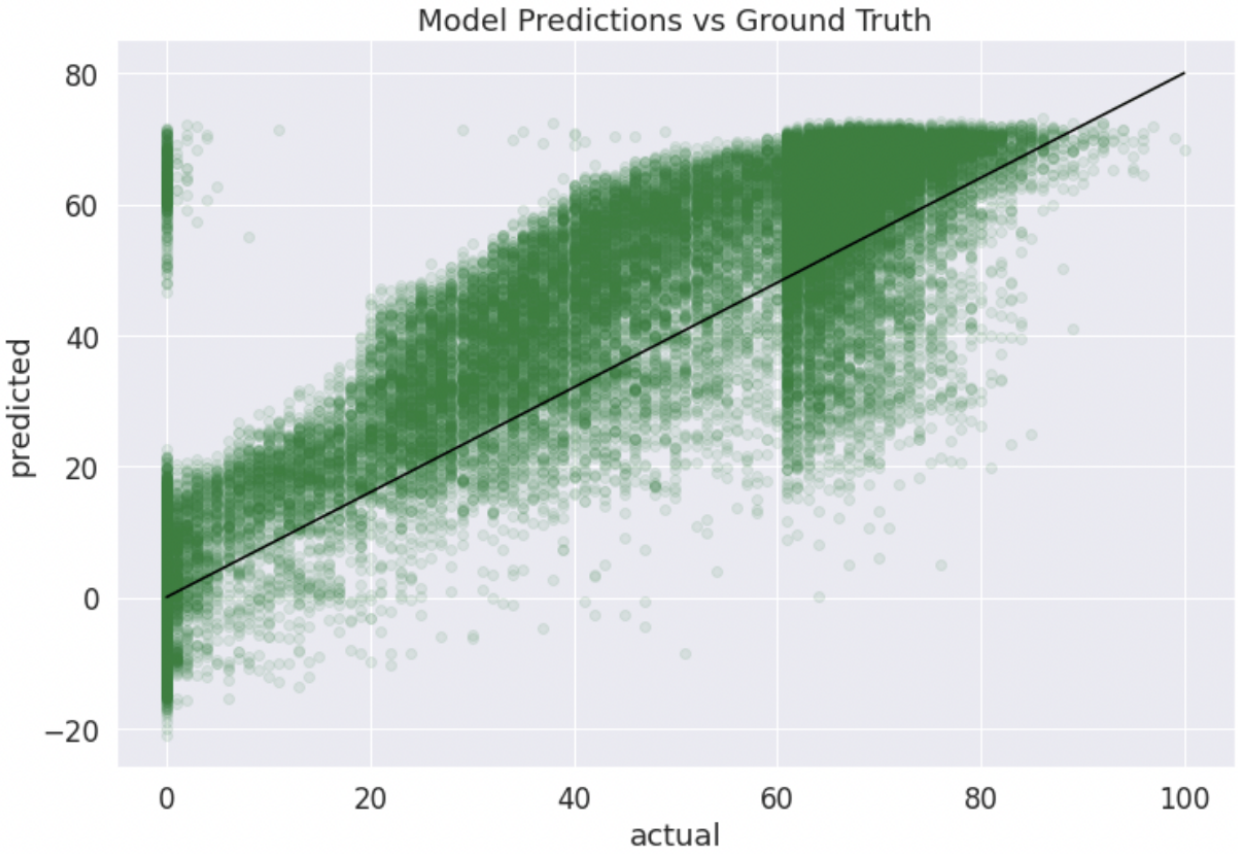


Focusing on popularity, looking at a correlation model to the left, we see high correlations with acousticness, energy, loudness, and the highest correlation with year.

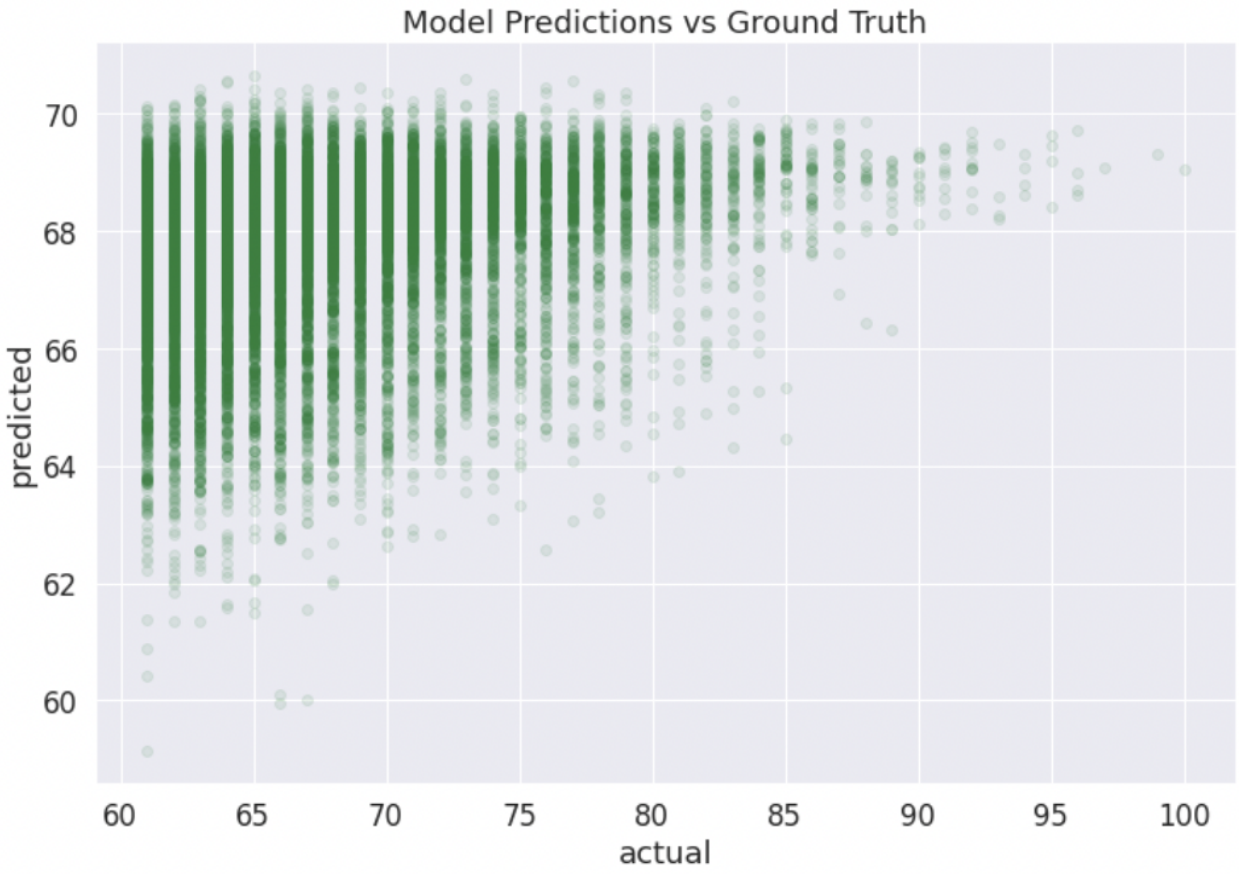
Furthermore, we see this high correlation due to the way the popularity is scaled having a heavy bias with year.

Linear Regression Model

We further looked into what popularity was and how we could predict popularity based on their features. To do this, we build a linear model with the features year, danceability, energy, instrumentalness, liveness, loudness, acousticness, speechiness, and valence. To conduct this model, due to the total popularity values ranging above 10 to 60 with an average of 31 , the model was split and trained into three different popularity data sets. The data sets were split into popularity of above 60, below and equal to 60 and above 0, and below and equal to 0. These data sets were then concat into one balanced data set with equal shape size and used for the linear model to predict popularity. With this trained dataset, the linear model predicted a 82% pearson strong positive correlation, R, as seen below, with a mean absolute error of 8.7.



However, the model did not predict any values above 80, as seen in the graph, while showing actual values up to 100. So to further understand what was happening to the values above 60, we conducted a linear model based on only above 60 popularity. As seen in the figure below, this model did not predict values above 75, further telling that due to there being so few values above 75 the model does not think it makes sense to predict higher than 75.



Song Recommendation Model

When building a song recommendation model we decided to focus on 'valence','year','acousticness' , 'danceability', 'energy', 'explicit','liveness','loudness','mode','speechiness' as our X variable, and ‘name’ being our y variable. Features used in X were, in our opinion, what makes a song unique. We decided to use those because we believed those would best describe someone’s taste in music. We used sklearn’s KNearest Neighbors to find 10 songs with most similar features as our recommendations to the listener. Here is an example of our model predictions.

```
[12] X=df[['valence','year','acousticness','danceability','energy','explicit','liveness','loudness','mode','speechiness']]
     y = df['name']
     neigh = NearestNeighbors(n_neighbors=10)
     neigh.fit(X,y)
     ind=neigh.kneighbors(X.iloc[[57331]])[1]
     for i in ind:
       print(df.iloc[i][['name', 'artists']])
```

	name	artists
57830	When You Love Me	['Memphis Minnie']
39126	New Orleans Stop Time	['Memphis Minnie']
1591	Pálida Noche - Instrumental (Remasterizado)	['Francisco Canaro']
39299	No Seas Malita - Instrumental (Remasterizado)	['Francisco Canaro']
20628	No Seas Amlita	['Francisco Canaro']
75774	Quand Il Joue De L'accordéon	['Fréhel']
39184	Farolito de papel - Instrumental (Remasterizado)	['Francisco Canaro']
57871	En la Trampa - Remasterizado	['Francisco Canaro', 'Charlo']
20610	Flores Secas - Remasterizado	['Francisco Canaro', 'Charlo']
75666	La Guardia Vieja - Instrumental (Remasterizado)	['Francisco Canaro']

After looking at our data, cleaning, analysing and making models we found that newer songs tend to be more popular. A few features that help a song become more popular are: more energy and loudness, and less acousticness.
We would suggest adding more newer songs to the dataset to balance it out with the older songs which are less popular. Since Spotify was founded in 2006 it is reasonable that songs after 2006 have more play counts that the ones made in the early 1920's.

Cover credits: [Spotify Premium Wallpaper 68954 1920x1277px (hdwallsource.com)]


Spotify®