Spotify music popularity prediction

Team 5 - Meghana Gantla, Adarsh Kotla, Hang Wang

Index:

Chapter 1: Introduction of Project and Previous Research:

The charts are an obvious way to gauge music's popularity. Since 1958, Billboard has published weekly charts of the "Hot 100" songs. Chart rankings are based on a mix of physical and digital sales, radio play, and internet streaming.

In recent years the music industry has shifted to a place where the popularity of a song is the most important factor and is influenced to a very high degree by having references from the most trending aspects of the pop culture throughout the social media platforms. And after the rise of digital albums, there is an almost 80 percent drop in the physical record sales of the albums from the artists. And the control on the popularity of the songs is mostly controlled and determined by online streaming services like Spotify, apple music and etc.

We in this project wanted to consider the Spotify app's song data from 2018. The whole of the music industry wants a hit formula and there has been significant research conducted in this area there are various researches with many findings even as intuitive as having the pronoun 'you' in the song would increase its popularity. We wanted to particularly look into what parts of the song structure contribute to the popularity of the song. We planned the project as first trying to find the variables that play a considerable role in determining the popularity of a song and eliminating the others. The dataset also contained unnecessary variables that wouldn't make a difference if removed which will be elaborated on in the following chapters. As with any project, we followed the project cycle of understanding, collecting, preparing, modeling, and visualizing.

Chapter 2: SMART Question.

"In 2018, Did the audio features of a song control its popularity?"

Our SMART question asks, whether audio features of a song have significance in determining the popularity of a song and if so, we want to find out which ones. In the data set, we have several audio features, song names, and names of the artists as well. We are aware that in current times the artist influences the popularity of a given song competently. Our research is focused on how much the other

features can influence the popularity, therefore, we have excluded the artist name and track name features from our research.

Chapter 3: Description of Dataset and cleaning:

This data set has around 115K Observations and contains 17 columns representing the features of a song. We have 11 numerical continuous variables and three categorical variables of interest, we dropped 3 irrelevant variables (track_id, track_name, artist_name) and removed around 1k duplicates in the data set. In the data, we have,

- Acousticness: This is a measure of how acoustic the song is according to Spotify. This feature has a range of 0 - 1 for this data set.
- Danceability: This is a measure of how suitable the music is for dancing. This feature ranges from 0 - 1
- Duration: This is a metric for the length of the song and it ranges from 3203 ms to 5610020 ms.
- Energy: This is a measure of how upbeat/energetic the song is and it has a range of 0 -1 for this data set
- Instrumentalness: This represents the usage of instruments in the making of said song and has a range of 0 - 1
- Liveness: This metric represents how lively the song is and ranges from 0 - 1
- Loudness: As the name suggests, this metric represents the sound of the song and has a range of about -60 - 2
- Speechiness: This is a measure of the number of words used in the song and has a range of 0 - 1
- Tempo: Tempo is another feature of the song and has a range of about 0 - 250 for this data
- Valence: Its a measure from 0 - 1 representing the positiveness of a song
- Popularity: This is our variable of interest, representing how popular a song is on Spotify. It has a range of 0 - 100.

Below we have complete descriptive statistics of all the numerical columns,

	Minimum	**Maximum**	**Mean**	**Standard deviation**	**25%**	**50%**	**75%**
Acousticness	0.0	0.996	0.335	0.343	0.029	0.194	0.620
Danceability	0.0	0.996	0.582	0.189	0.461	0.606	0.728
Duration	3203.0	5610020.0	212546.16	124320.83	164049.0	201773.0	240268.5
Energy	0.0	1.0	0.572	0.258	0.401	0.605	0.776
Instrumentalness	0.0	1.0	0.230	0.363	0.0	0.0	0.491
Liveness	0.0	0.999	0.194	0.167	0.097	0.123	0.234
Loudness	-60.0	1.806	-9.945	6.503	-11.845	-7.992	-5.701

Speechiness	0.0	0.966	0.112	0.124	0.038	0.056	0.129
Tempo	0.0	249.983	119.60	30.151	96.131	4.0	139.78
Valence	0.0	1.0	0.438	0.259	0.222	0.419	0.637
Popularity	0	100.00	24.235	17.931	10.00	22.00	35.00

We checked the data for any null values in the columns and have found none. Below are the data types for each column in the data and the number of null values in each of them.

```
RangeIndex: 116372 entries, 0 to 116371
Data columns (total 17 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   artist_name       116372 non-null  object
 1   track_id          116372 non-null  object
 2   track_name        116372 non-null  object
 3   acousticness      116372 non-null  float64
 4   danceability      116372 non-null  float64
 5   duration_ms       116372 non-null  int64
 6   energy            116372 non-null  float64
 7   instrumentalness  116372 non-null  float64
 8   key               116372 non-null  int64
 9   liveness          116372 non-null  float64
 10  loudness          116372 non-null  float64
 11  mode              116372 non-null  int64
 12  speechiness       116372 non-null  float64
 13  tempo             116372 non-null  float64
 14  time_signature    116372 non-null  int64
 15  valence           116372 non-null  float64
 16  popularity        116372 non-null  int64
dtypes: float64(9), int64(5), object(3)
memory usage: 15.1+ MB
```

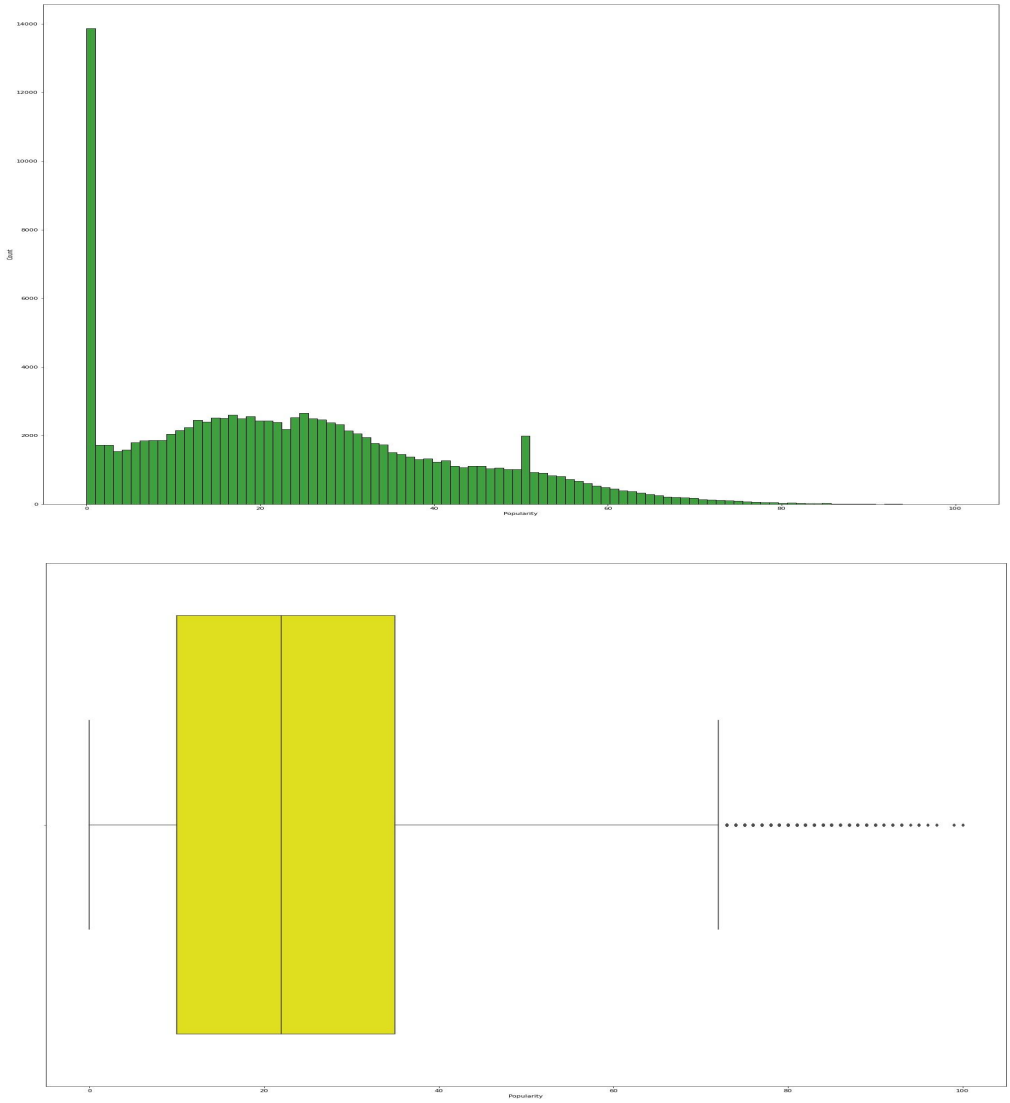
```
artist_name         0
track_id            0
track_name          0
acousticness        0
danceability        0
duration_ms         0
energy              0
instrumentalness    0
key                 0
liveness            0
loudness            0
mode                0
speechiness         0
tempo               0
time_signature      0
valence             0
popularity          0
dtype: int64
```

Chapter 4: EDA and Observations:

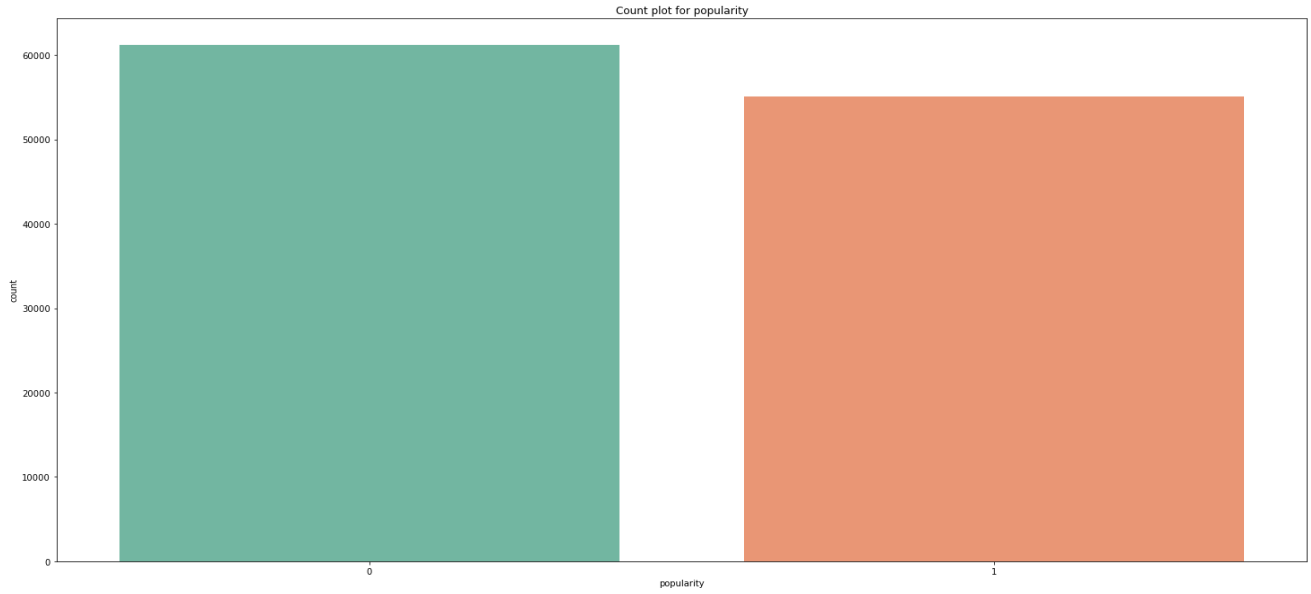
4.1 Popularity distribution:

Popularity is our variable of interest, looking at the distribution we can understand that it has a range of 0 - 100 and a mean of 24.235 for this data set. We can also observe a slight skew in the histogram (4.1.1fig) which indicates the presence of outliers in the column and this observation is further strengthened by looking at the boxplot. Going into our research, we have decided to convert the popularity feature, which is numerical into a categorical feature using its mean as the reference point. By doing this we split the popularity into “above average” and “below average” categories. After converting, below is the distribution. In the graph (4.1.2fig), 0 suggests below-average popularity and 1 suggests above-average popularity.

(4.1.1fig)



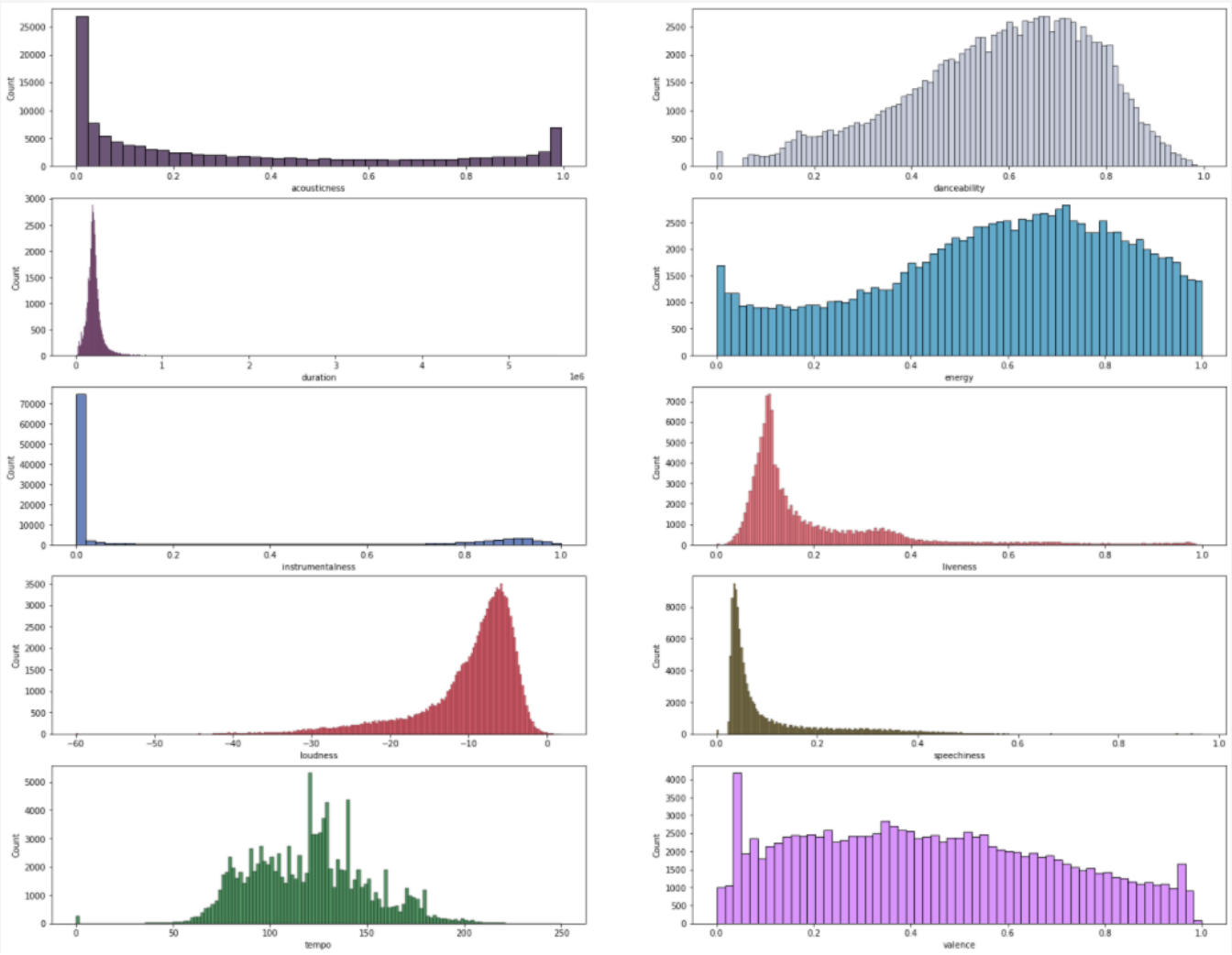
(4.1.2fig)



4.2 Distribution of all variables:

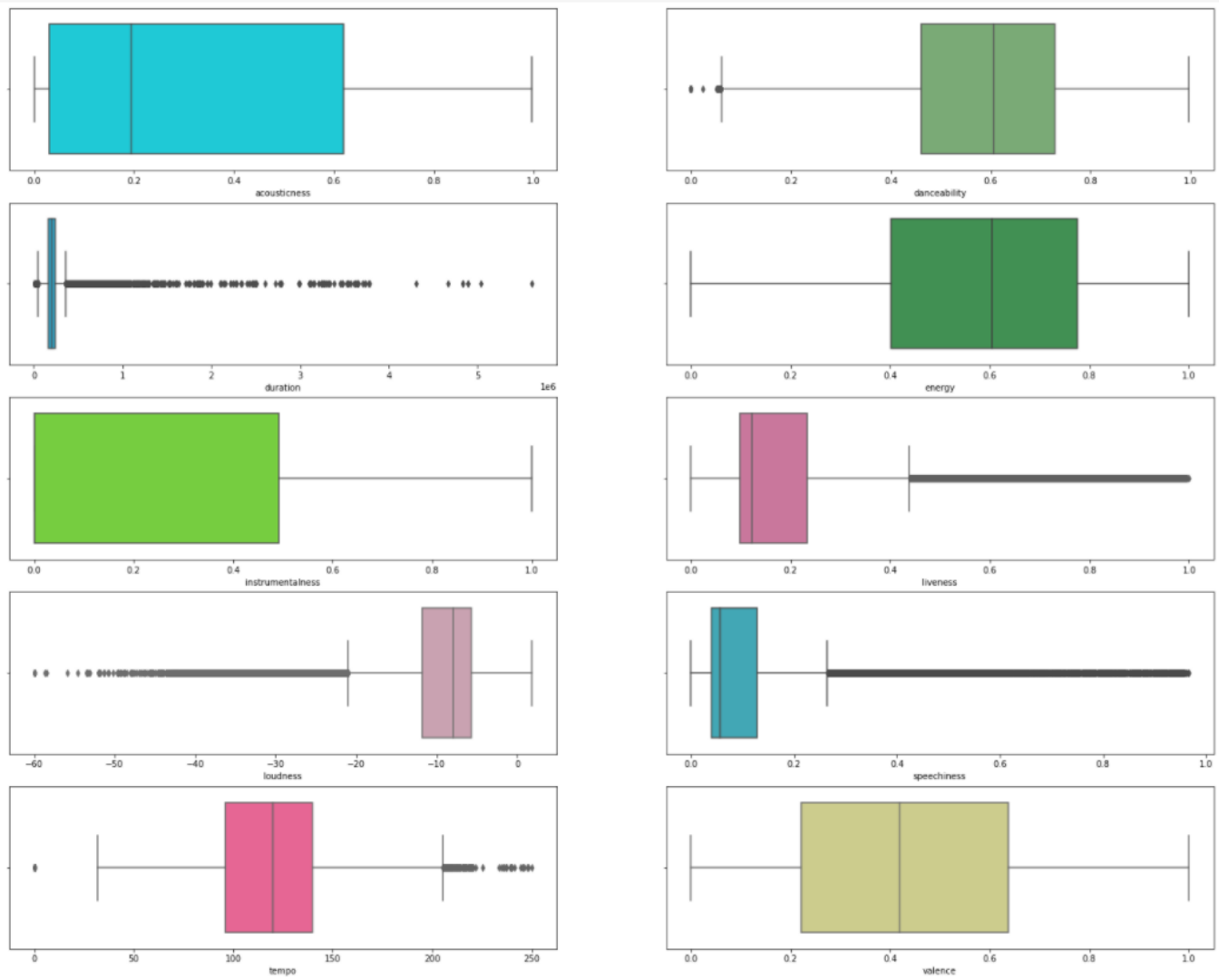
Going further, we want to understand all the features better before we can start our analysis. Below are the histograms for all the numerical variables in the data set displaying their distribution. It is clear that some of the variables have outliers from the way the graphs are skewed to one side. To observe the distribution better we have looked into boxplots and this has further strengthened our assumptions.

(4.2.1 fig)



From the plots below we can understand the range and observe the presence of outliers. There seem to be high numbers of outliers in the durations column, some in loudness, tempo, and danceability columns. We will be dealing with these outliers further into the preprocessing.

(4.2.2 fig)



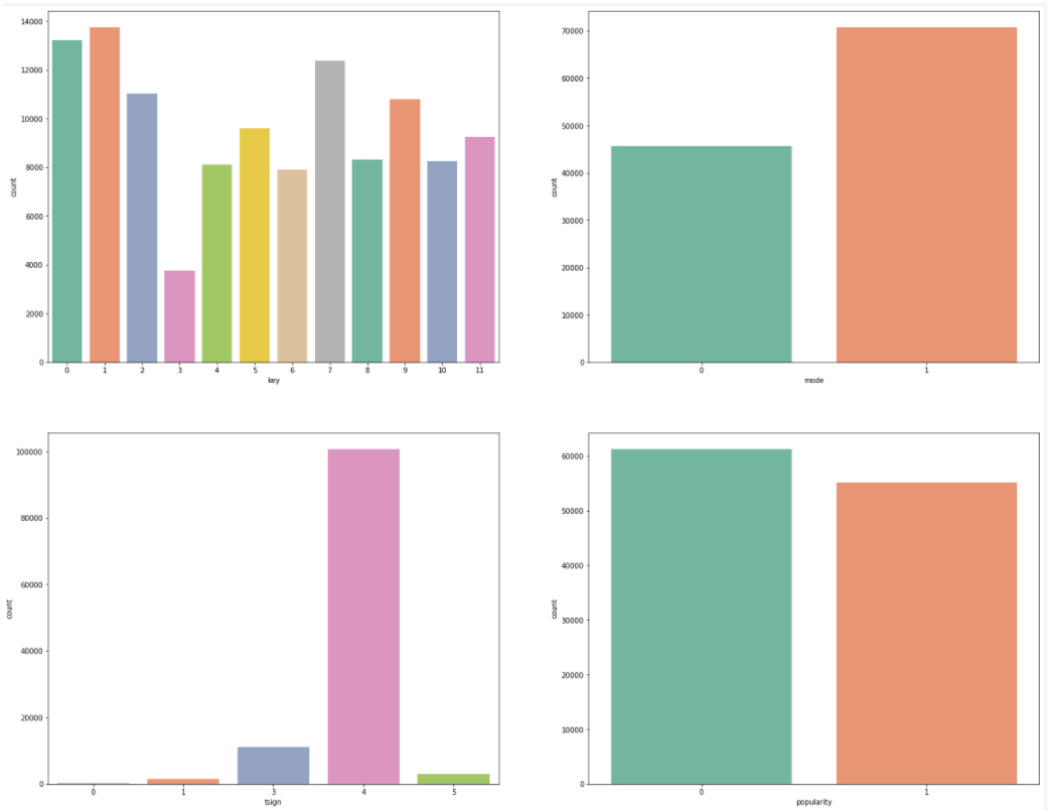
4.3: Categorical features:

We have a total of 4 categorical variables in our data set that are relevant to our research. We have,

- Key: We have 12 keys starting from 0 - 11 and they represent the key in which the song is in
- Mode: We have two modes 0 and 1 and they represent the mode of the song
- Time signature: We have 5 of them in this data and it is a notational constraint to represent the number of beats in each measure bar

Below we have the count plots for each of these variables to show how the data is distributed amongst each category. Key = 1, tsign = 4, mode = 1 have the highest number of observations in their categories. We also have popularity which is somewhat close to equal distribution as the median and mode of the variable are very close to each other. However, more songs fall under the below-average (0) popular category than the songs that fall under the above-average category (1).

(4.3.1 fig)



4.4: Correlation of variables:

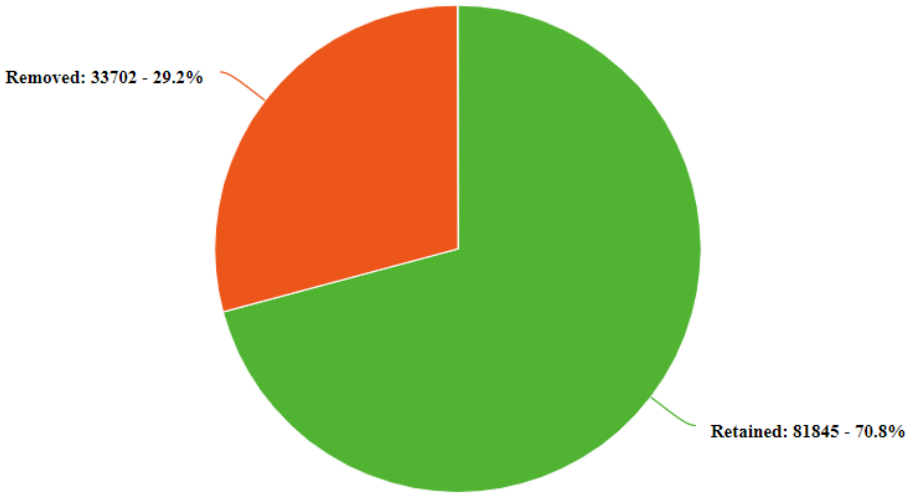
Interpreting the correlations, we do not see any of the variables having a strong direct correlation with the target variable. Danceability, instrumentalness, and loudness are very weakly correlated. Let's see if this changes after we are done preprocessing this data.

(4.4.1 fig)

4.5: Removing outliers:

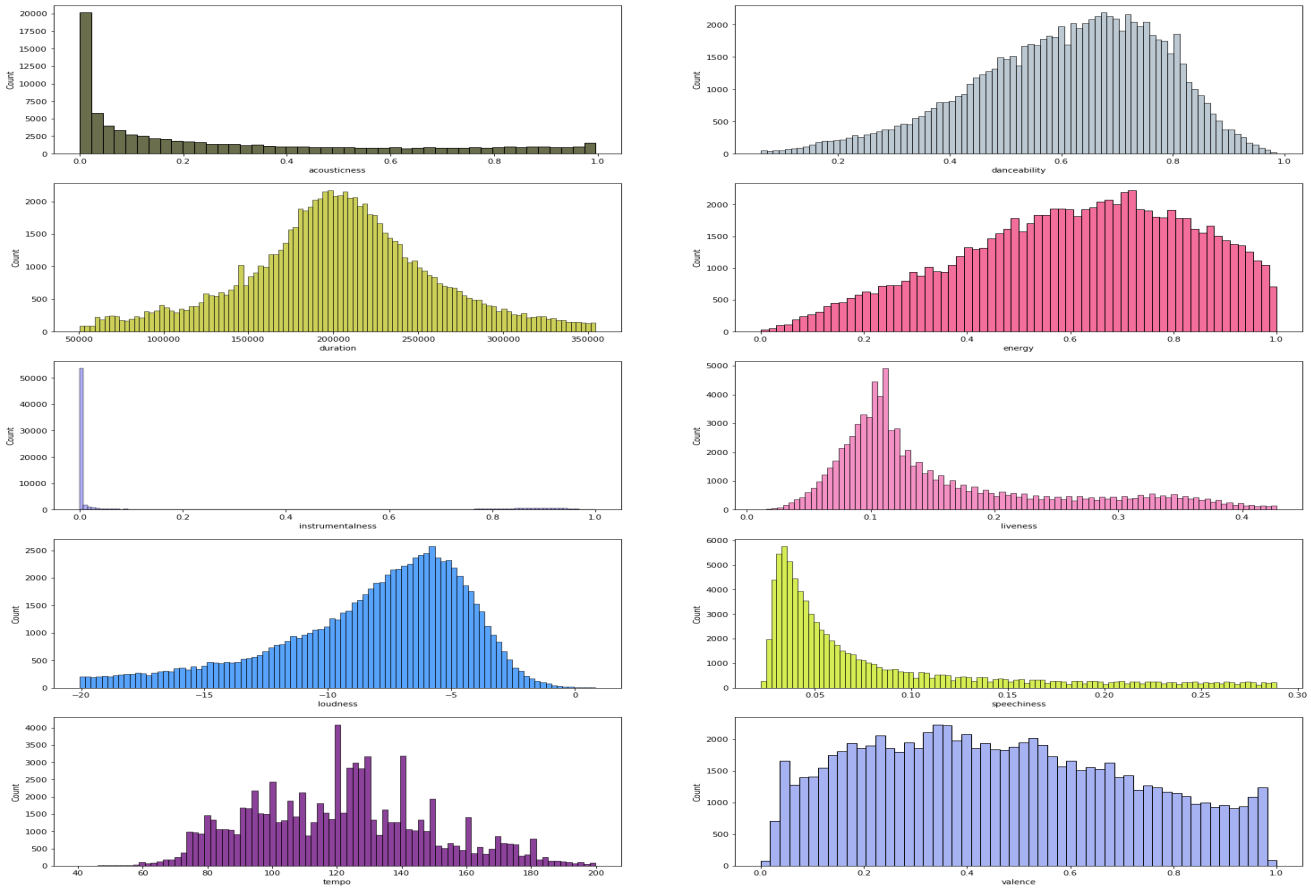
Here we have removed the outliers from all the features except the instrumentalness as it has unique distribution and most of the data is distributed on either ends of the range. We are applying 0.75 and 0.25 quantiles to remove the outliers and are left with about 81k observations that we will use for our predictive modeling.

(4.5.1 fig)



4.6 Numerical features after removing outliers:

(4.6.1 fig)



Above we have the histograms for all the numerical columns after finishing the preprocessing segment. The distribution is now closer to normal in most features and the skew has decreased in the remaining representing the changes after the outliers were removed.

4.7: Correlation of variables after removing outliers:

After the preprocessing, the correlation coefficients have changed a bit. Previous correlations of greater than 0.1 between danceability - popularity and instrumentalness - popularity have gone down to 0.07 and 0.08 respectively. This could yield some interesting findings.

(4.7.1 fig)

Chapter 5: Models and predictions:

The object of this project is to find which factors influence a song's popularity. From the experience, several reasons including the popularity of the singer, tunes, type of lyrics, and length of the song, can lead to a song being popular. To explore the most influential variables of the song's popularity, this project uses data from Spotify. We used a regression model, logistic model, decision tree, and Knn to describe and simulate the relationship between popularity and other variables including acoustics danceability duration_ms energy instrumentalness loudness Liveness mode valence tempo, and speechiness. Those variables describe a song in 13 aspects.

The first method we use is the linear regression model. A linear regression model is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables. And linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in

predicting an outcome (dependent) variable? (2) Which variables, in particular, are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? In our project, we see popularity as the response variable and see the other 13 variables as explanatory.

(1) This formula this the base model of the linear regression model, and the result of this regression model is in table 1. Following the most important criterion (John, R. C. S.,1983, p424)

(2) Then we got which means the model is not following linear regression, then we delete variables, key, and tempo of which the P-value is bigger than 0.05, to adjust the model.

(5.1.1 Table)

	coef	**std err**	**t**	**P>|t|**	**0.025**	**0.975**
Intercept	34.7007	0.498	69.634	0.000	33.724	35.677
key	-0.0162	0.014	-1.139	0.255	-0.044	0.012
mode	-0.3397	0.106	-3.211	0.001	-0.547	-0.132
danceability	5.9222	0.345	17.161	0.000	5.246	6.599
duration_ms	-1.545e-06	4.12e-07	-3.747	0.000	-2.35e-06	-7.37e-07
acousticness	1.5730	0.217	7.251	0.000	1.148	1.998
energy	-4.2910	0.372	11.520	0.000	-5.021	-3.561
liveness	-2.1845	0.321	-6.802	0.000	-2.814	-1.555
loudness	0.6537	0.014	45.711	0.000	0.626	0.682
speechiness	-6.1727	0.435	14.190	0.000	-7.025	-5.320
tempo	-0.0017	0.002	-0.998	0.318	-0.005	0.002
valence	-4.8436	0.231	20.976	0.000	-5.296	-4.391
instrumentalness	-6.1746	0.169	36.476	0.000	-6.506	-5.843

Then of the modify model is same as before, which is , and all the P-value are less than 0.05. Both results show that linear regression is not a good model to predict the popularity of the song.

(5.1.2 Table)

	coef	**std err**	**t**	**P>|t|**	**0.025**	**0.975**
Intercept	34.3889	0.439	78.301	0.000	33.528	35.250
key	-0.3198	0.104	-3.073	0.002	-0.524	-0.116

mode	5.9455	0.345	17.257	0.000	5.270	6.621
danceability	-1.548e-06	4.12e-07	-3.753	0.000	-2.36e-06	-7.39e-07
duration_ms	1.5871	0.216	7.339	0.000	1.163	2.011
acousticness	-4.3124	0.372	11.588	0.000	-5.042	-3.583
energy	-2.1661	0.321	-6.757	0.000	-2.795	-1.538
liveness	0.6528	0.014	45.772	0.000	0.625	0.681
loudness	-6.1878	0.435	14.240	0.000	-7.039	-5.336
speechiness	-4.8645	0.230	21.105	0.000	-5.316	-4.413
tempo	34.3889	0.439	78.301	0.000	33.528	35.250
valence	-0.3198	0.104	-3.073	0.002	-0.524	-0.116
instrumentalness	-6.1752	0.169	36.498	0.000	-6.507	-5.844

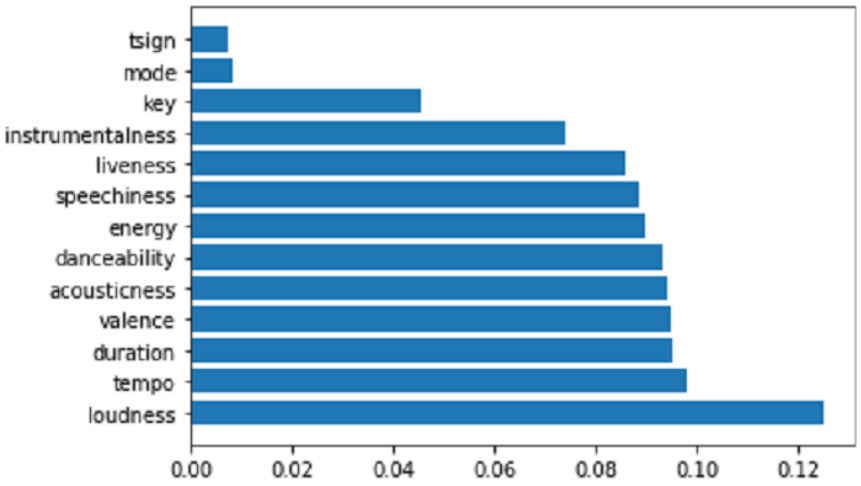
After that, the split the whole data into two parts and the parameter(train-size) = 0.08, which means the 80% of whole data is train data size and the 20%of whole data is test data size. This means the result is also not good, getting the same conclusion.

Then we tried the general linear model, and logistic model, to simulate the data. Data are unbalanced on Y if y = 1 occurs relatively few times or if y = 0 occurs relatively few times. This limits the number of predictors for which effects can be estimated precisely. (Agresti, A. 2018, P 138), Before that, we have already changed the popularity column from scalar to catalog variable. We settled the popularity as 0 if the number of popularity is smaller than 24 and settled the popularity as 1 if the number of popularity is bigger than 24. After that, we got a balance data of y which can get a meaningful result.

(3) & (4) Our logistic model is built based on the formula (3) and (4). To simulate the model, we split data as same as before. Then we got an accuracy of just 5%, which means the logistic model is not a fit for the data, we can not use the logistic model to predict a song's popularity.

Decision tree learning or induction of decision trees is one of the predictive modeling approaches used in statistics, data mining, and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves) Then we use a decision tree to simulate the data which used the same way to split data. After this supervised learning, the accuracy of test data by using a decision tree is just 51.29%, Then we considered the feature importance of the model. 'loudness' is the most influential variable compared with the other variables on the popularity. Even though all the variables are not very influential.

(5.1.1 fig)



Then we try the k-nearest neighbors (KNN) model. KNN is a data classification method for estimating the likelihood that a data point will become a member of one group or another based on what group the data points nearest to it belong to. We also used the same data way to split data. On the first try, we set k as 10. After the simulation, we got an accuracy is 61.039%. The number is not very high, so we try to set k as 5 and get the accuracy as 65.936%.

Reference:

https://www.theverge.com/tldr/2018/2/5/16974194/spotify-recommendation-algorithm-playlist-hack-nelson
John, R. C. S. (1983). Applied linear regression models.
Agresti, A. (2018). An introduction to categorical data analysis. John Wiley & Sons.