Alan Wu
Principles of Data Science Capstone
Pascal Wallisch

Description of dataset:

This dataset consists of data on 52,000 songs that were randomly picked from a variety of genres sorted in alphabetical order (a as in "acoustic" to h as in "hiphop"). For the purposes of this analysis, I can assume that the data for one song is independent of data from other songs. This data is stored in the file "*spotify52kData.csv*", as follows:

Row 1: Column headers

Row 2-52001: Specific individual songs

Column 1: *songNumber* – the track ID of the song, from 0 to 51999.

Column 2: *artist(s)* – the artist(s) who are credited with creating the song.

Column 3: *album_name* – the name of the album

Column 4: *track_name* – the title of the specific track corresponding to the track ID

Column 5: *popularity* – this is an important metric provided by spotify, an integer from 0 to 100, where a higher number corresponds to a higher number of plays on spotify.

Column 6: *duration* – this is the duration of the song in ms. A ms is a millisecond. There are a thousand milliseconds in a second and 60 seconds in a minute.

Column 7: *explicit* – this is a binary (Boolean) categorical variable. If it is true, the lyrics of the track contain explicit language, e.g. foul language, swear words or otherwise content that some consider to be indecent.

Column 8: *danceability* – this is an audio feature provided by the Spotify API. Method of achieving this metric is unknown. It tries to quantify how easy it is to dance to the song (presumably capturing tempo and beat), and varies from 0 to 1.

Column 9: *energy* - this is an audio feature provided by the Spotify API. It tries to quantify how "hard" a song goes. Intense songs have more energy, softer/melodic songs lower energy, it varies from 0 to 1

Column 10: *key* – what is the key of the song, from A to G# (mapped to categories 0 to 11).

Column 11: *loudness* – average loudness of a track in dB (decibels)

Column 12: *mode* – this is a binary categorical variable. 1 = song is in major, 0 =song is in minor

Column 13: *speechiness* – quantifies how much of the song is spoken, varying from 0 (fully instrumental songs) to 1 (songs that consist entirely of spoken words).

Column 14: *acousticness* – varies from 0 (song contains exclusively synthesized sounds) to 1 (song features exclusively acoustic instruments like acoustic guitars, pianos or orchestral instruments)

Column 15: *instrumentalness* – basically the inverse of speechiness, varying from 1 (for songs without any vocals) to 0.

Column 16: *liveness* - this is an audio feature provided by the Spotify API. It tries to quantify how likely the recording was live in front of an audience (values close to 1) vs. how likely it was recorded in a studio without a live audience (values close to 0).
Column 17: *valence* - this is an audio feature provided by the Spotify API. It tries to quantify how uplifting a song is. Songs with a positive mood =close to 1 and songs with a negative mood =close to 0
Column 18: *tempo* – speed of the song in beats per minute (BPM)
Column 19: *time_signature* – how many beats there are in a measure (usually 4 or 3)
Column 20: *track_genre* – genre assigned by spotify, e.g. “blues” or “classica

10 questions to consider:

1) Consider the 10 song features duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo. Is any of these features reasonably distributed normally?
2) Is there a relationship (positive/negative) between song length and popularity of a song?
3) Are explicitly rated songs more popular than songs that are not explicit?
4) Are songs in major key more popular than songs in minor key?
5) Substantiate or refute: energy is believed to largely reflect the “loudness” of a song.
6) Which of the 10 individual (single) song features from question 1 predicts popularity best? How good is this “best” model?
7) Building a model that uses *all* of the song features from question 1, how well can you predict popularity now? How much (if at all) is this model improved compared to the best model in question 6). How do you account for this?
8) When considering the 10 song features above, how many meaningful principal components can you extract? What proportion of the variance do these principal components account for?
9) Can you predict whether a song is in major or minor key from valence? If so, how good is this prediction? If not, is there a better predictor?
10) Which is a better predictor of whether a song is classical music – duration or the principal components you extracted in question 8?

Introduction/initial data processing

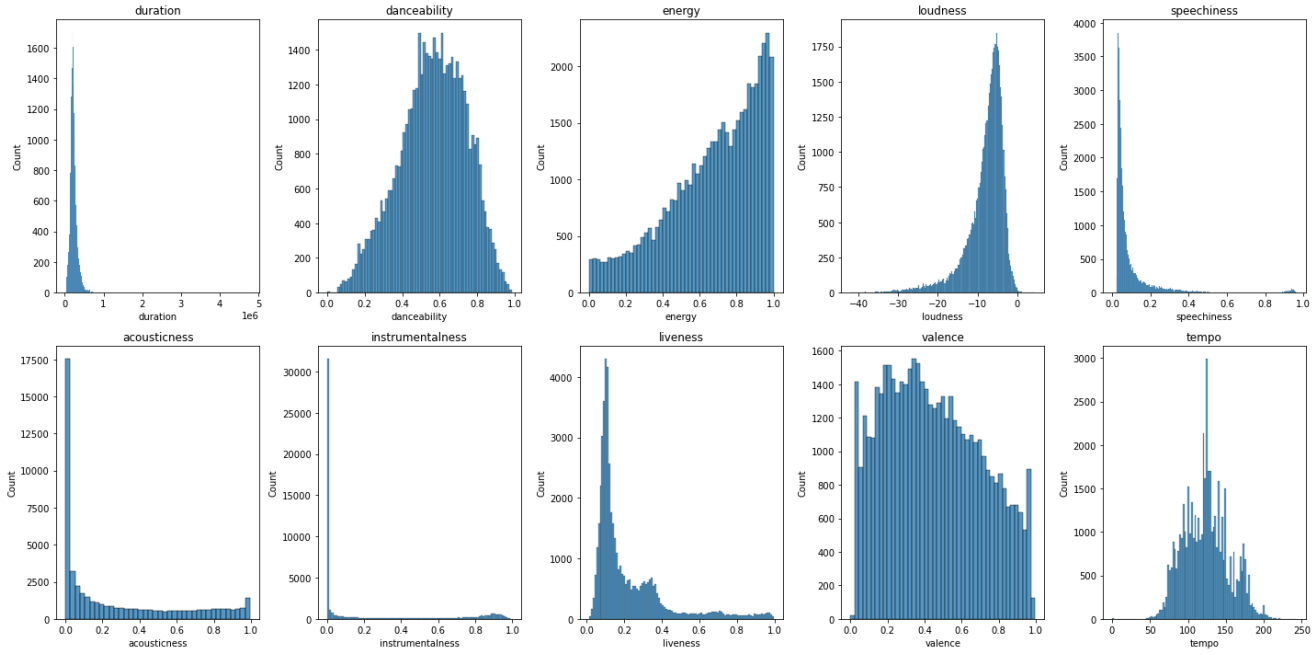
No missing data was found in the dataset, so replacing NaN values was not necessary. Duplicates of songs were found in the form of identical songs duplicated in differently named albums. An example is shown on the right:

28	Jason Mraz	Christmas Time	Winter Wonderland
29	Jason Mraz	Perfect Chr…	Winter Wonderland
30	Jason Mraz	Merry Christmas	Winter Wonderland
31	Jason Mraz	Christmas M…	Winter Wonderland

I dropped the duplicates by finding songs that had identical values in all but "songNumber" and "album_name" and removing them from the dataset. RNG was seeded using my N-number ***11262055***.

```
#data cleaning
missing_data = data.isnull().sum() #no missing data found in dataset
#drop duplicate songs
columns_remove = ['album_name', 'songNumber']
duplicates = data[data.duplicated(data.columns.difference(columns_remove), keep = False )]
data = data.drop_duplicates(data.columns.difference(columns_remove), keep='first')
```

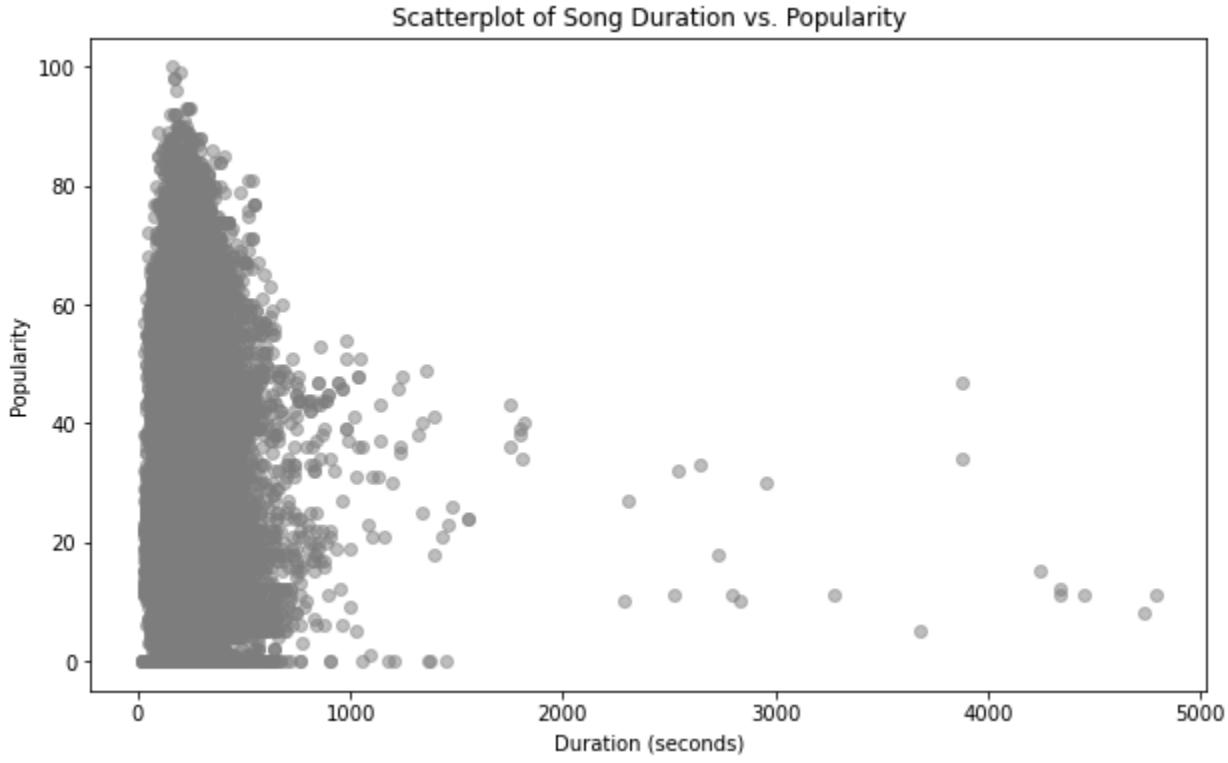
Q1: I visualized the distributions of the 10 song features (duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo) with a histogram for each one by using pyplot. X axis corresponds to feature value, Y axis corresponds to frequency.



An immediate look at the figure shows no feature follows a normal distribution. I also conducted a Shapiro-Wilk test and found that none of the features were normally distributed ($p < 0.05$).

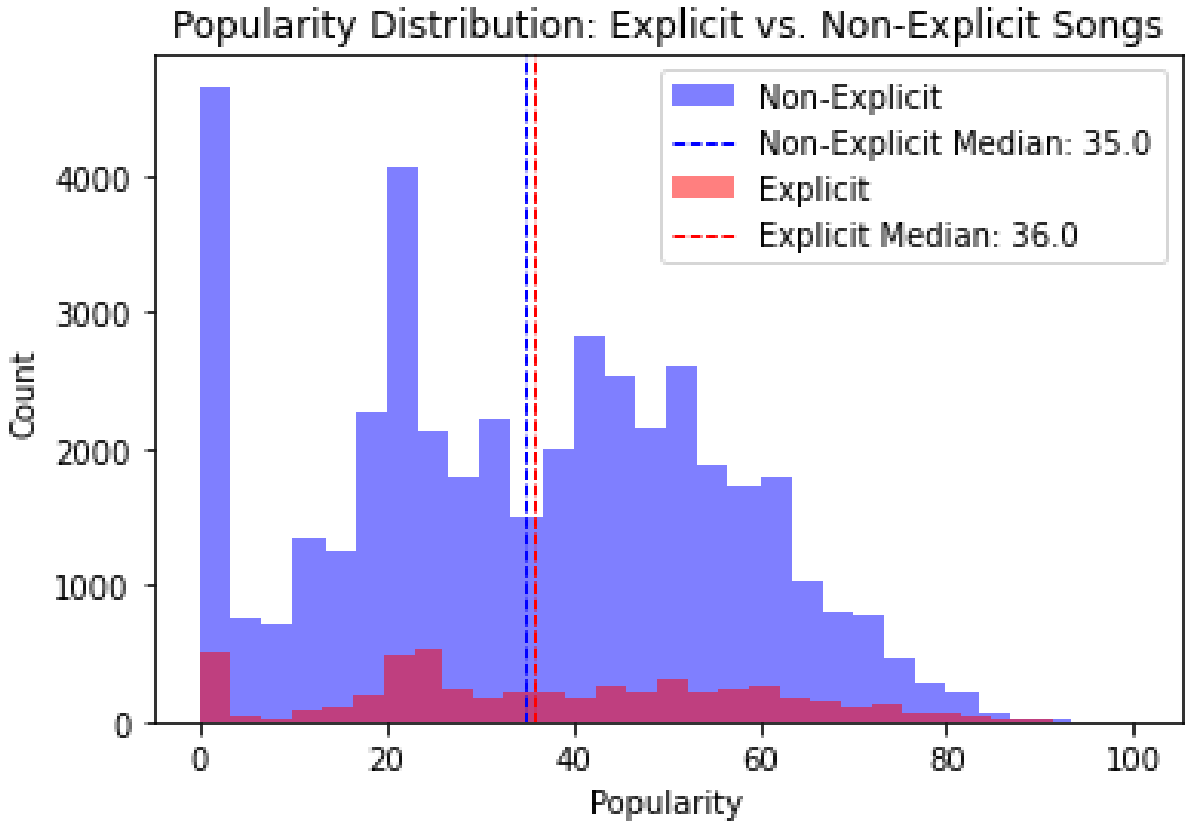
Feature	Statistic	p-value	Normally Distributed
duration	0.644040	0.000000	No
danceability	0.989707	0.000000	No
energy	0.930382	0.000000	No
loudness	0.837497	0.000000	No
speechiness	0.498481	0.000000	No
acousticness	0.795932	0.000000	No
instrumentalness	0.605978	0.000000	No
liveness	0.742436	0.000000	No
valence	0.964832	0.000000	No
tempo	0.990196	0.000000	No

Q2: I wanted to determine if there is a relationship between song length and song popularity. I calculated the Pearson correlation coefficient to determine the strength of the linear relationship. The r value was -0.081533, indicating a very weak negative correlation between song length and popularity. Both the coefficient and the scatterplot below demonstrate there is little to no correlation between these variables. X axis is duration converted from ms to seconds for ease of interpretation.



Q3: I wanted to find if explicitly rated songs are more popular than non-explicit songs. We can't assume that song popularity is normally distributed, so a parametric T-test is not appropriate. Instead, I can use a Mann-Whitney U Test because it is a non-parametric test optimal for two independent sample groups and it is robust to extreme values. I split my dataset into the two

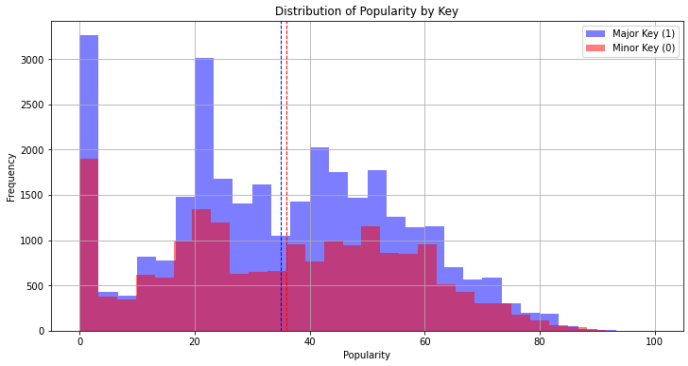
categories (explicit/non-explicit songs) that I defined. I performed a one-sided U Test to see if there was a significant difference in popularity between the two groups. I came back with a p-value of 4.983748064755883e-19, far less than 0.05, indicating significance. As seen from the population distribution histogram below, the median of popularity for explicit songs is greater than non-explicit.



It can be concluded that explicit songs are more popular than non-explicit rated songs.

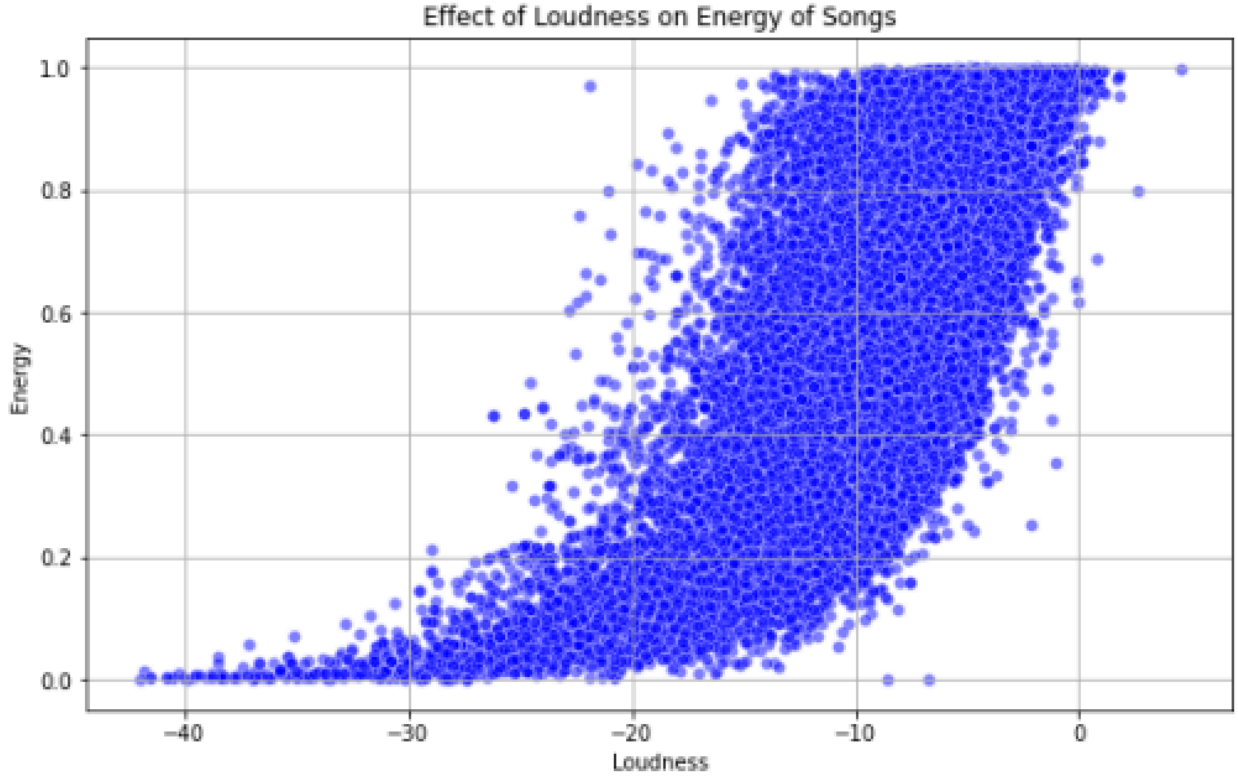
Q4: I wanted to find if major key songs are more popular than minor key songs. Again, we cannot assume normal distribution of song popularity so we can perform a Mann-Whitney U Test. I split the dataset into two categories (songs with major or minor key) and performed a one-sided U test (because we want to find if major key is MORE popular than minor key) to see if there was a significant difference in popularity between the two groups. The resulting p-value from the calculation was:

```
Q4
Mann-Whitney U Test: U=281050716.5, p-value=0.9978254725
```



Indicating that p-value (which was > 0.05) was not significant. As a result, we cannot conclude that major key songs are more popular than minor key ones.

Q5: To determine if energy reflects the loudness of a song, I constructed a scatterplot with loudness on the X axis and energy on the Y axis to visualize if there were any relationships.



I conducted a Pearson correlation coefficient calculation, which returned a r-value of 0.7757248496000904, indicating a strong positive linear correlation between loudness and energy. However, I noticed a slight curve and wondered if it could also be represented with a Spearman correlation coefficient. The rho-value returned from the Spearman coefficient calculation was 0.7335199900517353, indicating a strong positive correlation. As a result, we can refute any assumptions that energy does not reflect song loudness.

Q6: The goal is to find which feature from the 10 features mentioned in Q1 best predict song popularity. To do this, I needed to run a simple linear regression model on these features and determine the coefficient of determination(R^2), essentially what percentage of variance for popularity is determined by that specific feature. I separated the 10 features data from the rest of the dataset and standardized it by z-scoring. I split the data into training and test sets using my seed for RNG created through my N-number "11262055", which was important to allow my training/testing datasets to be representative of the entire dataset.

```
Feature: duration
RMSE: 20.99793531200686
(COD) R^2: 0.0061831704927546305

Feature: danceability
RMSE: 21.04186997341388
(COD) R^2: 0.0020200293060080865

Feature: energy
RMSE: 20.994823442306444
(COD) R^2: 0.006477713672508045

Feature: loudness
RMSE: 21.012236540405876
(COD) R^2: 0.004828976257334583

Feature: speechiness
RMSE: 21.01012444140977
(COD) R^2: 0.0050290305521009104

Feature: acousticness
RMSE: 21.04847982397533
(COD) R^2: 0.0013929429727163045
```

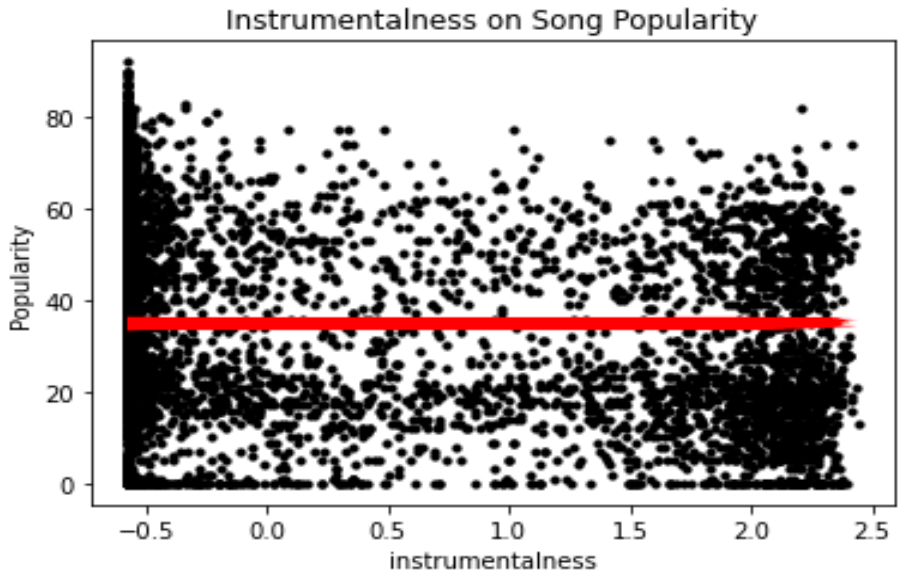
```
Feature: instrumentalness
RMSE: 20.666924499183644
(COD) R^2: 0.03726920076434148

Feature: liveness
RMSE: 21.03633692407543
(COD) R^2: 0.002544806469223393

Feature: valence
RMSE: 21.06190620735007
(COD) R^2: 0.00011855569727403648

Feature: tempo
RMSE: 21.0691555608065
(COD) R^2: -0.000569866459021684
```

As you can see from the metrics generated from my regression, the instrumentalness feature appears to be the best predictor of popularity, with both the lowest RMSE and R^2 values. This means that instrumentalness explains about 3.7% of variance that is observed in song popularity. As you can see below with my scatterplot showing the relationship between instrumentalism and popularity, despite instrumentalness being the best predictor for RMSE, you can still visually see a lot of error. Combined with the fact that this predictor only accounts for 3.7% of the variance, it is not an ideal model.

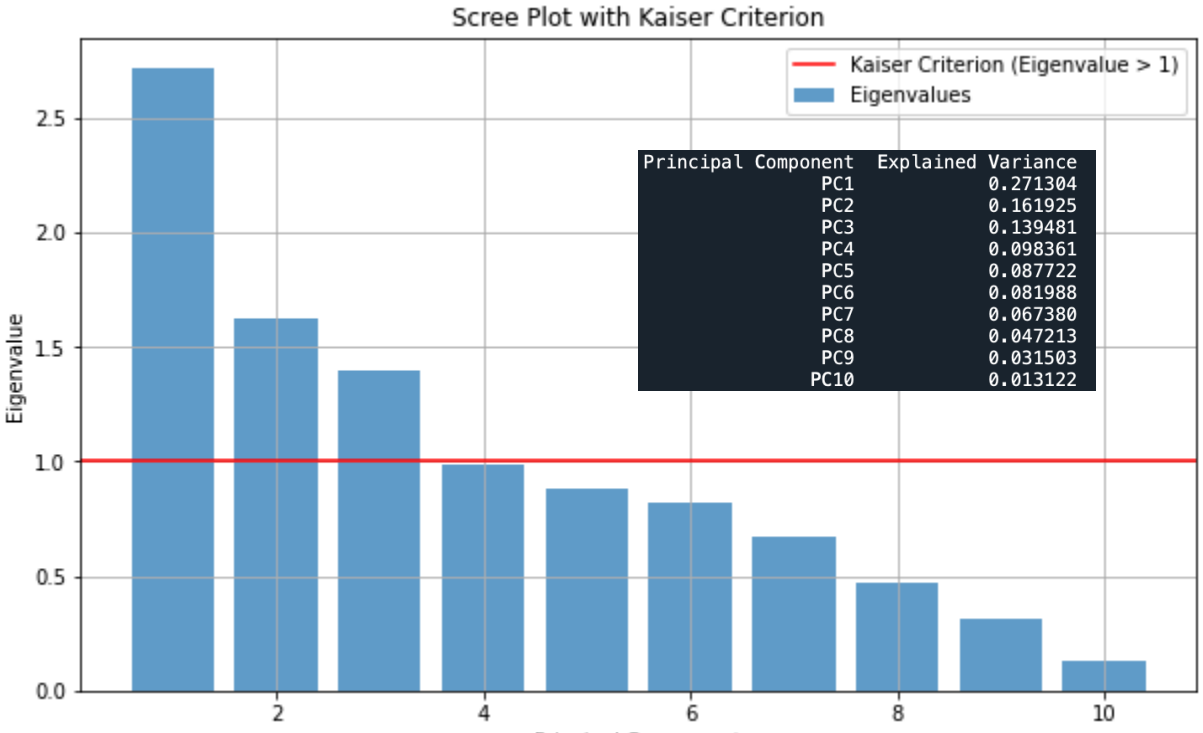


Q7: In order to build a model that incorporates all 10 features mentioned in Q1 to predict song popularity, it is necessary to construct a multiple regression model. Again, the song feature data is standardized and split into training and testing data according to the seed. Using all 10 features to predict song popularity, there is little difference:

```
MULTIPLE REGRESSION model for predicting popularity with all 10 features
RMSE of model using all 10 features: 20.26300387967229
RMSE Improvement: 0.4039206195113536
R^2 of  model using all 10 features: 0.07453326059282628
R^2 Improvement 0.0372640598284848
```

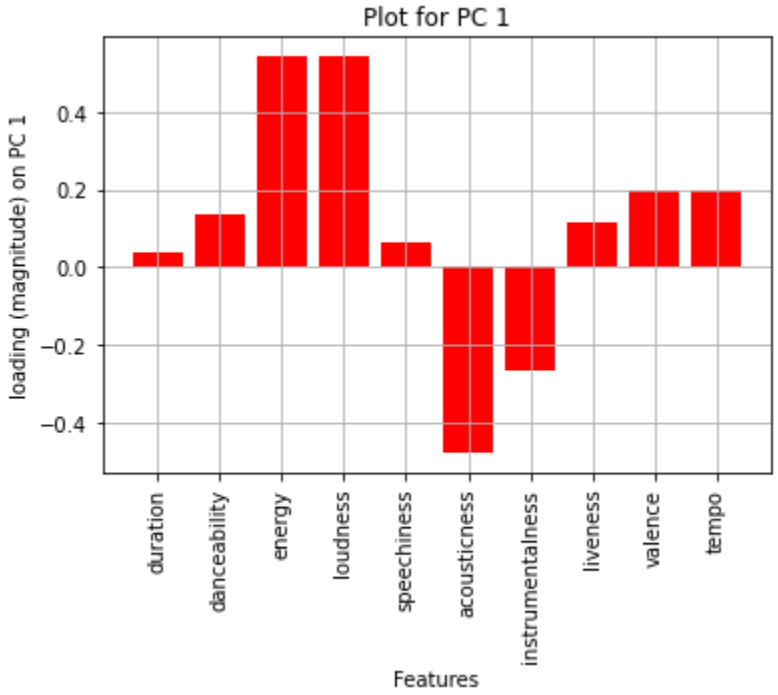
According to my multiple regression model, the R^2 value is now 0.074, meaning that the 10 features account for about 7.4% of the variance in song popularity, an improvement of about 3.7% from the previous sole predictor, instrumentalness. Furthermore, RMSE has improved by just 0.4, meaning error has not reduced by much by adding other features. The model has a low percentage of overall variance and has low predictive power. The reason why this is is because of the multidimensionality of data, especially the real-world settings such as music taste, where there are many niche factors that may affect an individual's opinion of music. Each feature only captures a certain amount of variance and we must also account for collinearity which complicates the interpretation. It's very possible that many of the features are collinear (aka highly correlated with each other) and that makes the model less robust.

Q8: To extract meaningful principal components from our 10 features, a principal component analysis (PCA) must be conducted. After first standardizing the data across the features, PCA was performed and I used the explained variance function to find and display the explained variance ratio of each of the 10 components, all of the ratios adding up to 1.0 (100%). Using this info, I also created a scree plot with the eigenvalues of each of these principal components.



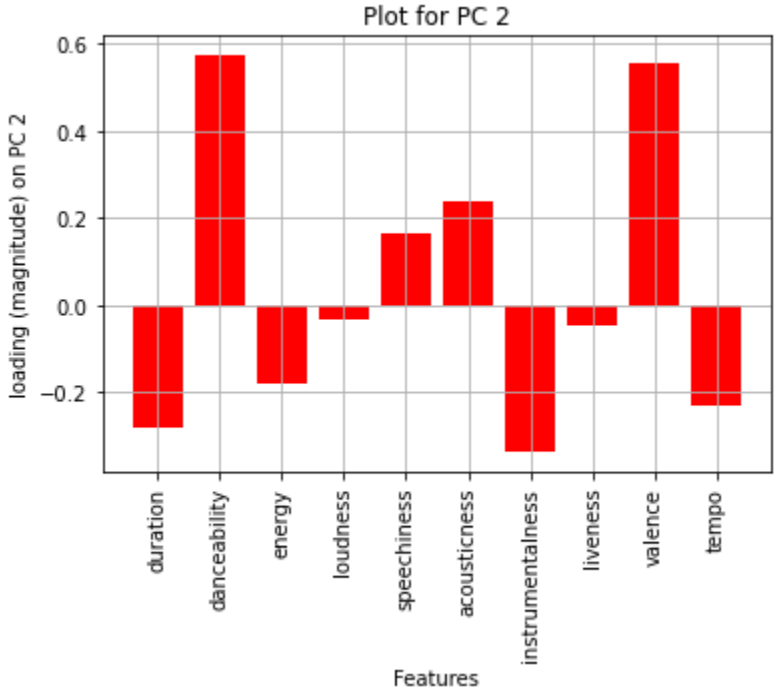
Principal Component	Explained Variance
PC1	0.271304
PC2	0.161925
PC3	0.139481
PC4	0.098361
PC5	0.087722
PC6	0.081988
PC7	0.067380
PC8	0.047213
PC9	0.031503
PC10	0.013122

I chose the Kaiser criterion (which keeps only principal components with eigenvalues greater than 1). The reasoning for my cutoff of eigenvalue 1.0 is because that means that the component explains more variance than simply a single original variable (itself). As a result, I got 3 meaningful principal components above this cutoff, accounting for about 57.2% of variance.



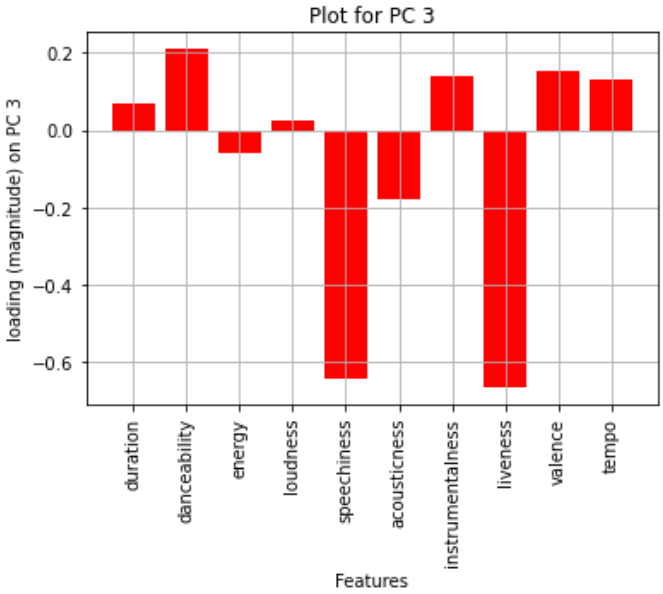
PC 1 primary features: weighted for energy, loudness

PC 1 = “overall vibe of the song, does it go hard?”



PC 2 primary features: heavily weighted for danceability, valence

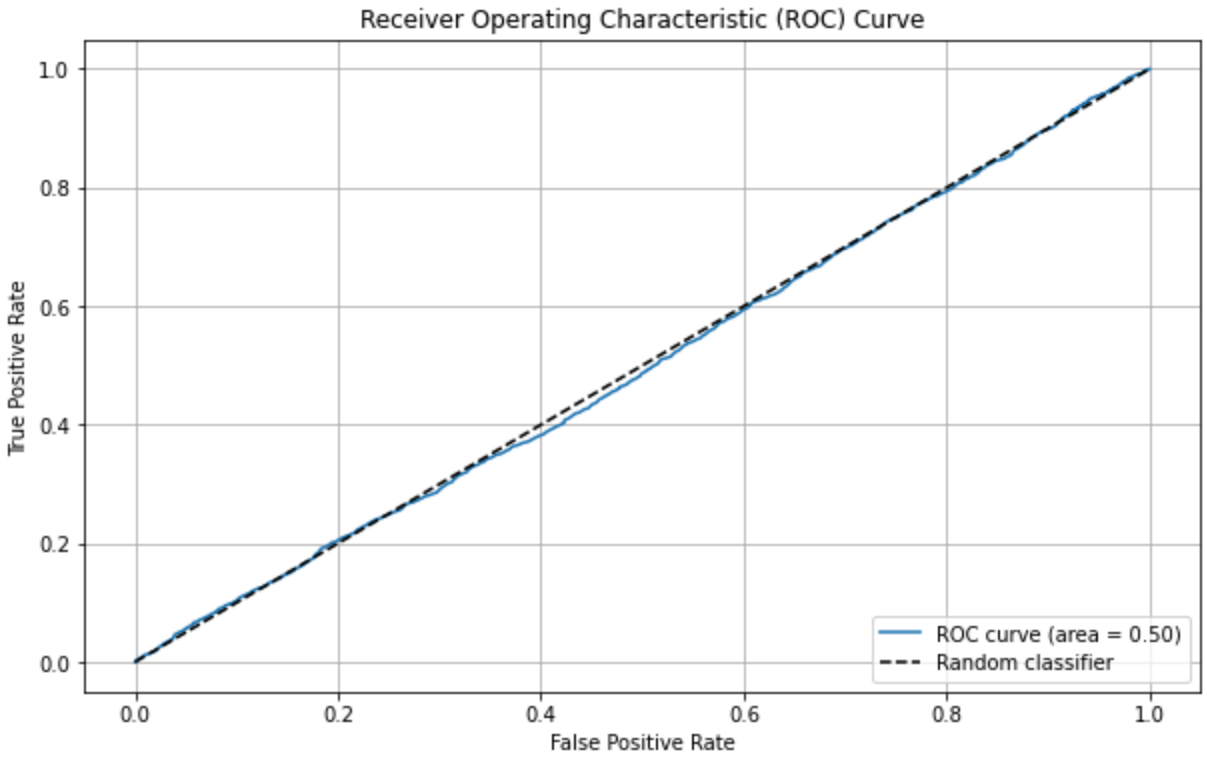
PC 2 = “dance song? is the song good for the club?”



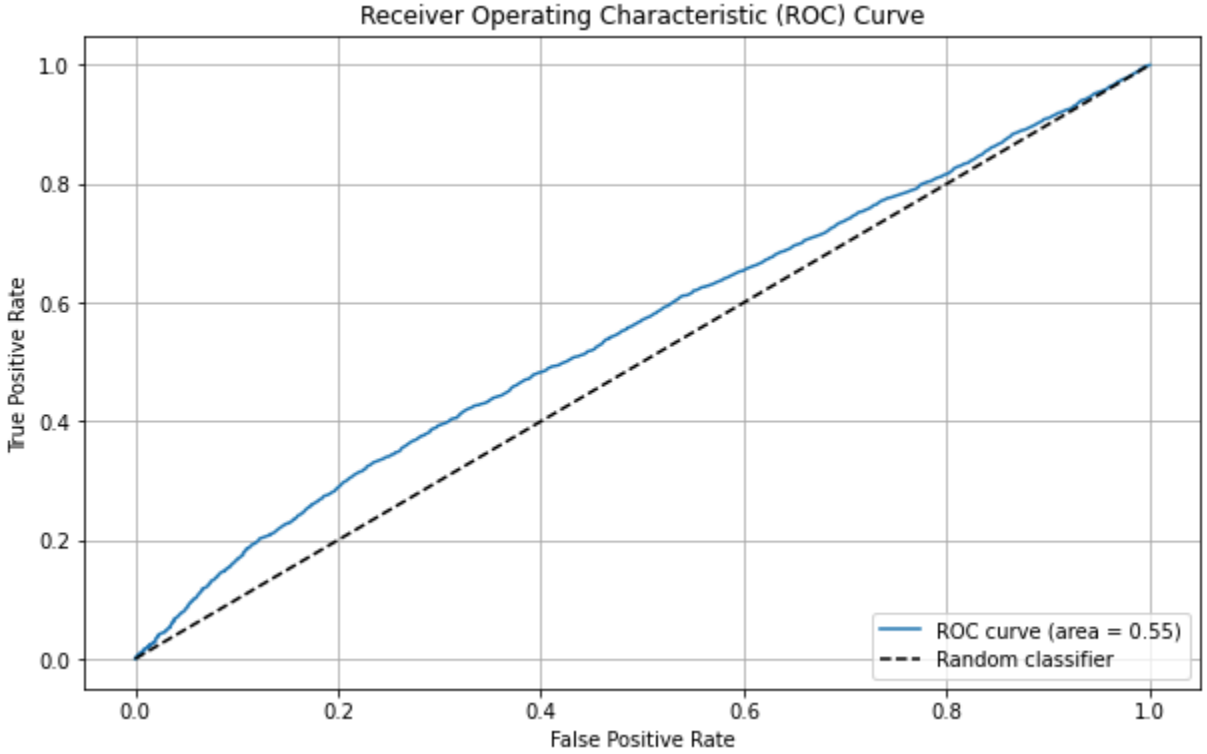
PC 3 primary features: heavily weighted against speechiness, liveness

PC 3 = “Is the song melodic? well produced/studio version?”

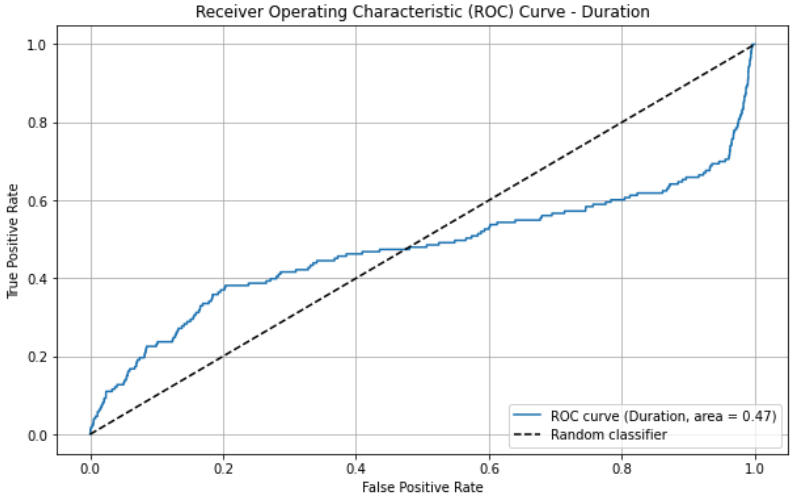
Q9: The next question is whether a song’s key (major or minor) can be predicted based on the valence (how uplifting the song is). After standardizing data and splitting it into test and training set using my random seed for a logistic regression, I imported and used the Synthetic Minority Oversampling technique (SMOTE), because I noticed the class imbalance (the number of minor key songs is less than major key songs) that could lead to model inaccuracy due to training data having less data from minor key songs. Displaying and printing out my data received from the logistic regression and ROC model, I found my AUROC score to be 0.496255. Below is a visualization of the ROC curve for valence’s ability to predict mode(song key):



As seen by the visualization and ROC score being very close to 0.5, is essentially guessing by chance which means valence is not an effective predictor of song key. However, a feature with a slightly higher AUC that can be used to predict song keys better is energy, with an AUC of 0.55279, slightly greater than 0.5. This is a better predictor based on my model as seen below:

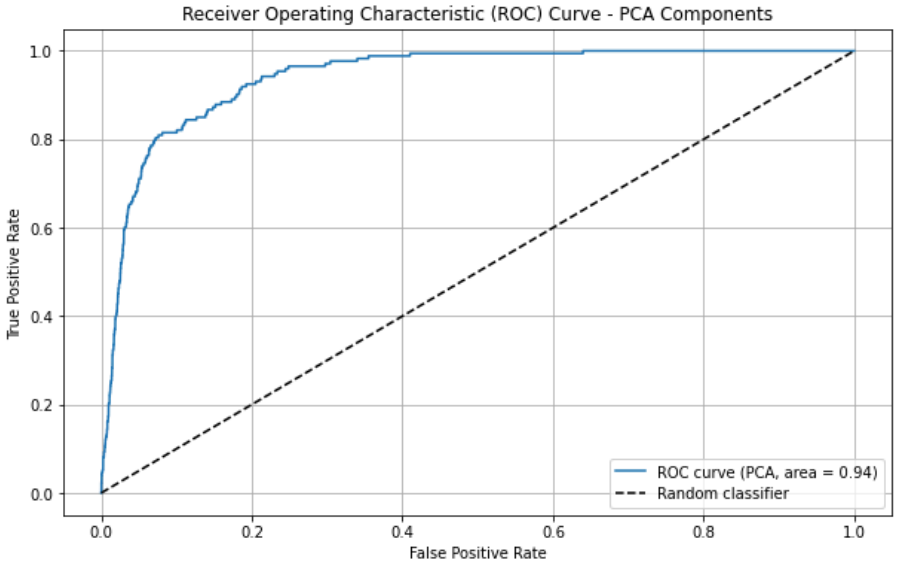


<u>Q10)</u> This last question pertains to finding which is a better predictor of whether a song is classical – duration or the principal components extracted from question 8. I extracted the data, creating a new column specifically for the "classical" songs found within the "track_genre" column. I needed to convert the genre label to a binary numerical label indicating true or false to a classical song. Again, using train, test, split with my unique seed for the logistic regression model, I plotted the ROC curve for duration as a predictor of classical music:



With an AUROC score of 0.47, duration is shown to not be a good predictor of whether or not a song is classical.

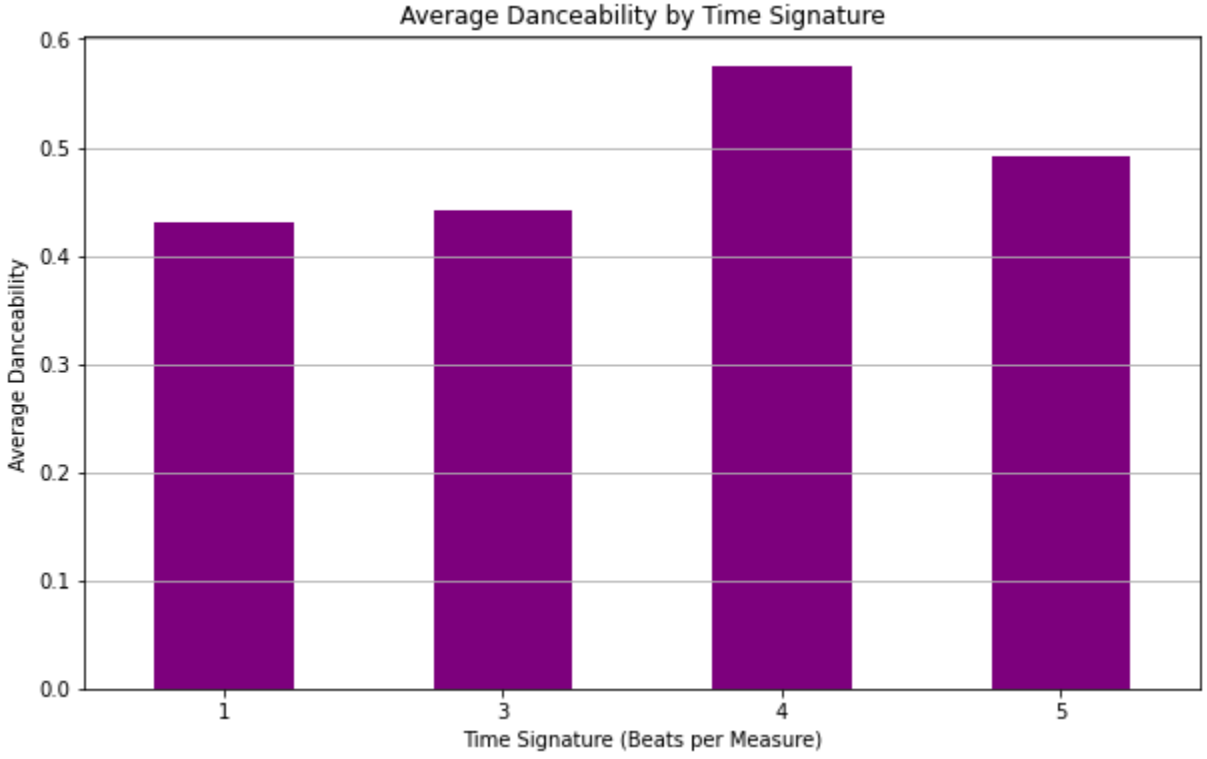
However, by using the same principal components that extracted from question 8, the ROC curve is shown below:



With an AUROC score of 0.94, using the PCs extracted from Q8 are far better predictors for determining whether a song is classical.

EXTRA CREDIT:

I wanted to see if the beats per measure (time_signature) had any correlation with the danceability of a song. Beats per measure varied from 1 to 5, with the vast majority of songs either being 3 or 4 beats per measure. X-axis: time sig, Y-axis: Avg danceability of the time sig groups



I noticed that the typical 4 beats per measure had the highest danceability values. This makes sense because most modern pop/dance music is in the 4/4 time signature. My Pearson correlation coefficient r was about 0.16, showing that there is a weak positive linear relationship between time signature and danceability. It isn't a strong predictor of danceability.