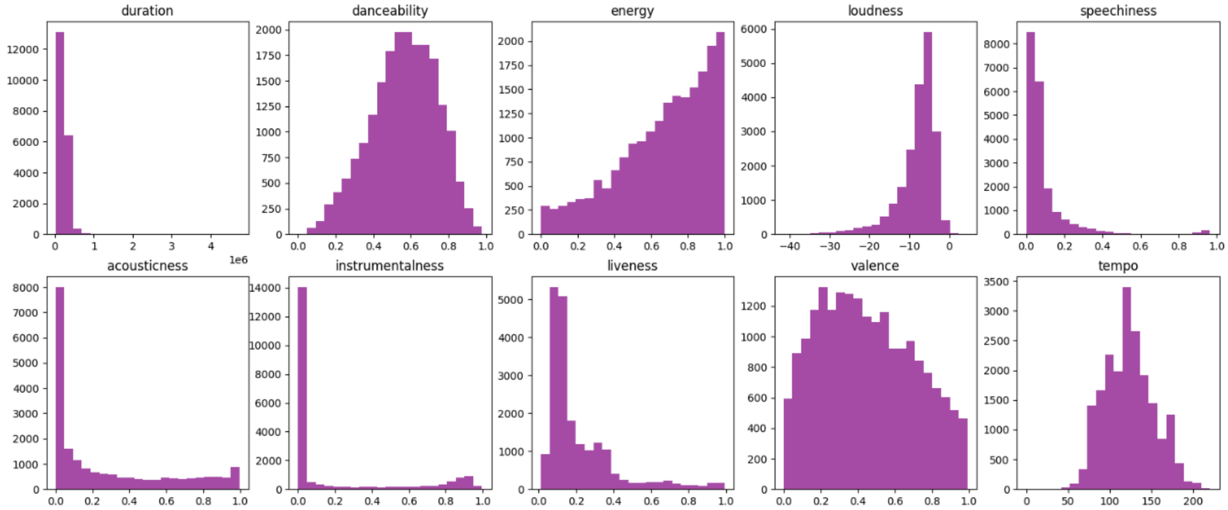
PODS Capstone
Ritti Bhogal
Professor Pascal Wallisch
May 14, 2024

Pre-Processing

The preprocessing I performed included removing rows with qualitative data that was ultimately not relevant to what I was analyzing. This included the rows album_name, artists, and track_name. Furthermore, I decided to only work with a sample of the data, or 20,000 songs, rather than all 52,000 songs since it would be more practical when it comes to simulating real world data science problems (how often do you get 20,000 participants, let alone 52,000 in a study), and a smaller dataset is easier to manage.

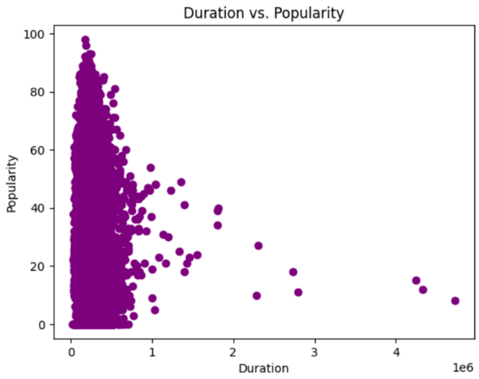
Questions

1. **Consider the 10 song features duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo. Are any of these features reasonably distributed normally? If so, which one? [Suggestion: Include a 2x5 figure with histograms for each feature)**



Not a single one of the features is distributed normally, with the closest to a normal distribution perhaps being danceability or tempo. Nearly all the histograms are skewed. Duration, speechiness, acousticness, instrumentalness, liveness, and valence are all skewed right. This is perhaps due to there being several 0 or near 0 values for each of the features. On the other hand, danceability, energy, and loudness are skewed left. This is interesting because this implies that there are songs in the sample that have high energy, are danceable, and are pretty loud, which is somewhat reflective of most pop songs. Tempo is the only distribution that lies somewhat in the middle, with songs between 50 and 200 beats.

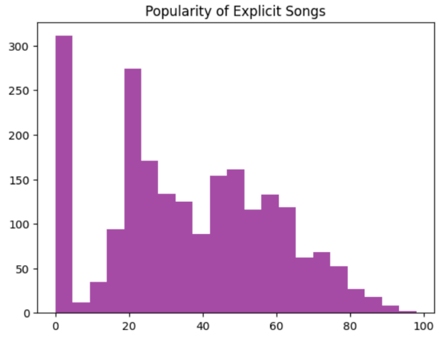
2. **Is there a relationship between song length and popularity of a song? If so, if the relationship positive or negative? [Suggestion: Include a scatterplot]**

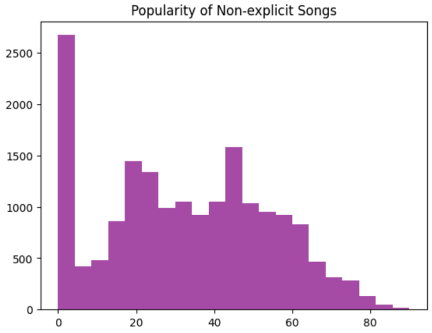


Based on the scatterplot above, it looks like duration and popularity aren't very correlated. In fact, the Pearson correlation coefficient for the relationship is -0.054651, which is nearly no correlation. However, what this correlation coefficient doesn't capture is the clustering of data points between the 0 and 1 minutes duration. It seems that anything beyond 1 minute doesn't reach a popularity score above 60. This could be indicative of how short form content like Tik Toks with relatively short audios are becoming increasingly popular.

3. **Are explicitly rated songs more popular than songs that are not explicit? [Suggestion: Do a suitable significance test, be it parametric, non-parametric or permutation]**

For this question, I decided to perform the Mann-Whitney U test. This is because the popularity distribution for both explicit and non-explicit songs is not normally distributed, and the samples of explicit and non-explicit songs are independent from each other.

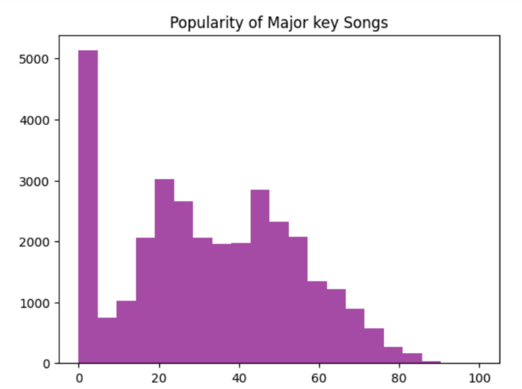


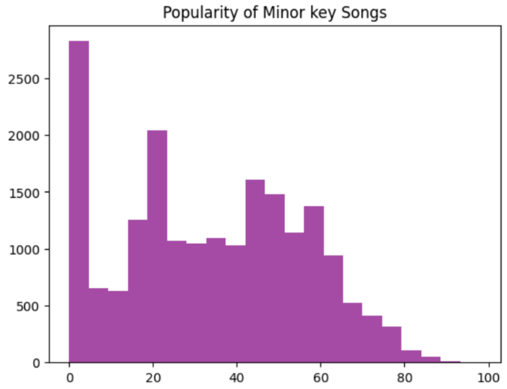


The null hypothesis I tested was that explicit songs are as popular as non-explicit songs. After performing the test, I got a p-value of 3.626992895736226e-08, indicating that the distributions are highly dissimilar, meaning that there is a statistically significant difference between the popularity of explicit and non-explicit songs, so I need to reject the null hypothesis. To see which one is more popular, I computed the mean and median of each popularity distribution to determine the one with the higher popularity rating. The explicit song popularity distribution has a higher mean (around 36) than non-explicit songs (around 33), and the median popularity of explicit songs (34) is slightly higher than the median of non-explicit songs (33). It appears that based on the central tendency statistics (more importantly the median, as this is what the Mann-Whitney test uses), explicit songs are more popular than non-explicit songs.

4. **Are songs in major key more popular than songs in minor key? [Suggestion: Do a suitable significance test, be it parametric, non-parametric or permutation]**

I performed a Mann-Whitney U test for similar reasons as the previous question: the popularity distribution for both minor and major key songs is not normally distributed, and the samples of minor and major key songs are independent from each other.

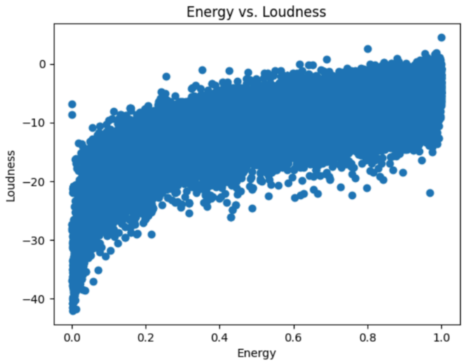

Popularity of Major key Songs


Popularity of Minor key Songs

The null hypothesis is that neither is more popular than the other. After performing the Mann-Whitney U test, I got a p value of around 0.00115, indicating that the effect was statistically significant, and I can reject the null hypothesis that minor and major key songs have similar popularities. To determine the key type with higher popularity, I compared means and medians, which showed me that the minor key song popularity distribution has a higher mean and median (around 33.9 and 34) compared to the major key song popularity distribution (around 32.9 and 33). So, it appears that minor key songs are more popular than major key songs.

5. **Energy is believed to largely reflect the "loudness" of a song. Can you substantiate (or refute) that this is the case? [Suggestion: Include a scatterplot]**

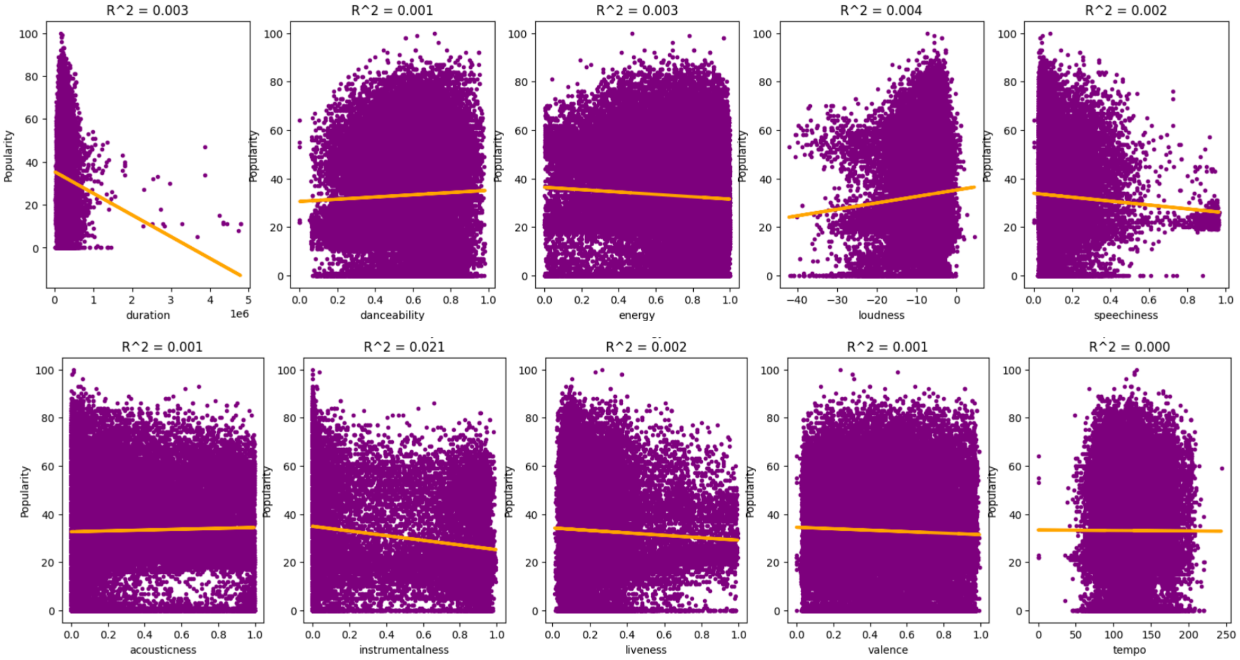
The distribution of loudness to energy is shown below.


Energy vs. Loudness

Just by observing the scatterplot, there is clearly a positive relationship between loudness and energy, however it is non-linear. To represent the relationship in numerical form, I calculated the Spearman correlation coefficient for the distribution and got a correlation coefficient of approximately 0.73394, indicating that energy and loudness are indeed reflective of each other.

6. **Which of the 10 individual (single) song features from question 1 predicts popularity best? How good is this "best" model?**

To capture the predictability of each feature in terms of popularity, I performed a linear regression using each feature and plotted the coefficient of determination, or R^2, of each model to measure the goodness of fit. The regressions are shown below.

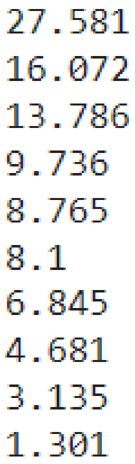


Based on the output above, it is clear that while not a single feature is a good predictor of popularity due to the extremely low R^2 all across, the feature that explains the variance of popularity the best is instrumentalness, with an R^2 of 0.021.

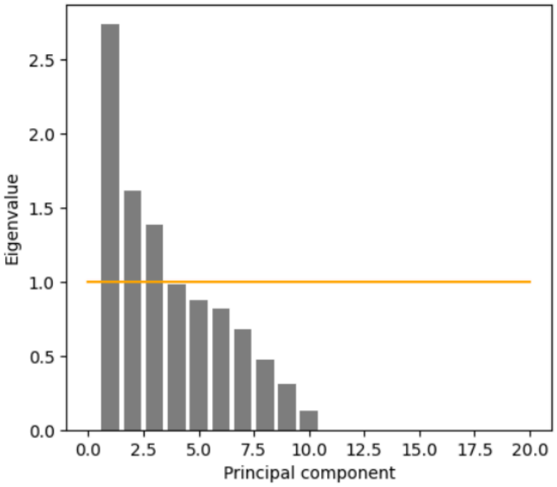
7. **Building a model that uses *all* of the song features from question 1, how well can you predict popularity now? How much (if at all) is this model improved compared to the best model in question 6). How do you account for this?**

When performing a linear regression using all of the variables, my model gets an R^2 value of around 0.04687. This is nearly double the value of the best model in the previous question. As we know, adding more predictors causes R^2 to increase because we are better able to capture different relationships between variables and popularity for a given song. However, it's also important to be conscious of overfitting. Fortunately, with such a low R^2 value in the first place, this isn't really a concern.

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?**



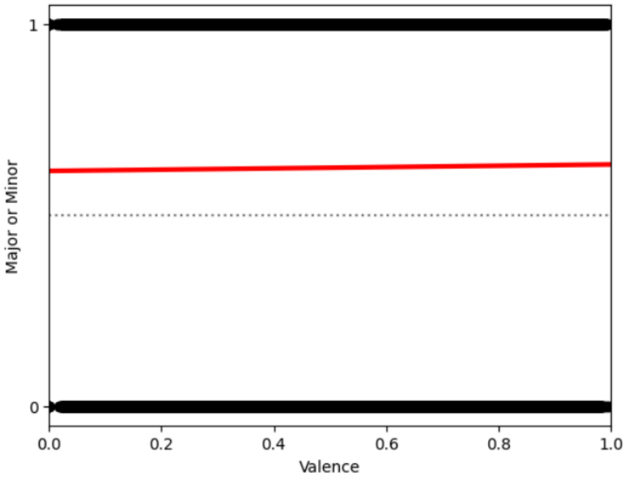
```
27.581
16.072
13.786
9.736
8.765
8.1
6.845
4.681
3.135
1.301
```



I extracted 3 principal components in my analysis, as per the kaiser criterion. The 3 principal components with eigenvalues greater than 1 explain the most variance in data. I also examined the respective loadings for each principal component and found that the feature that made up the most of each of three components was loudness, danceability, and liveness. In total, the 3 principal components account for approximately 58% of all the variance.

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? [Suggestion: It might be nice to show the logistic regression once you are done building the model]**

I built a logistic regression model by using valence to classify songs as major or minor keys. The results are shown below. I set major to class 1 and minor to class 0.



While the graph doesn’t look perfect, it seems that the classifier is biased to select a class as having a major key. The AUC score I got was around 51%, which further rectifies the inherent bias of the song (if it always picks major, it is guaranteed a 50% true positive rate).