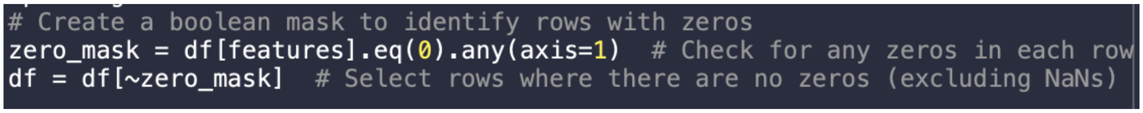
Aaron Munford
Pascal Wallisch
Principles of Data Science
06/25/2024

Analysis of Spotify Data

This analysis utilized both NumPy and pandas for data manipulation. I identified missing values by finding rows with zeros in specific features and excluded those rows to ensure data quality. When necessary, I used Principal Component Analysis (PCA) to identify uncorrelated features that capture most of the data's variance. This helped to reduce the number of features for analysis and potentially improve model performance. I also standardized numerical features using techniques like StandardScaler to ensure all features are on a similar scale. Lastly, I seeded "randomState" as my N number.



```
# Create a boolean mask to identify rows with zeros
zero_mask = df[features].eq(0).any(axis=1)  # Check for any zeros in each row
df = df[~zero_mask]  # Select rows where there are no zeros (excluding NaNs)
```

Figure 1: *Cleaning zeros row-wise from the NumPy array "df"*

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? If so, which one?
Using matplotlib I created created a 2x5 grid of subplots for each of the 10 features as seen in figure 2:

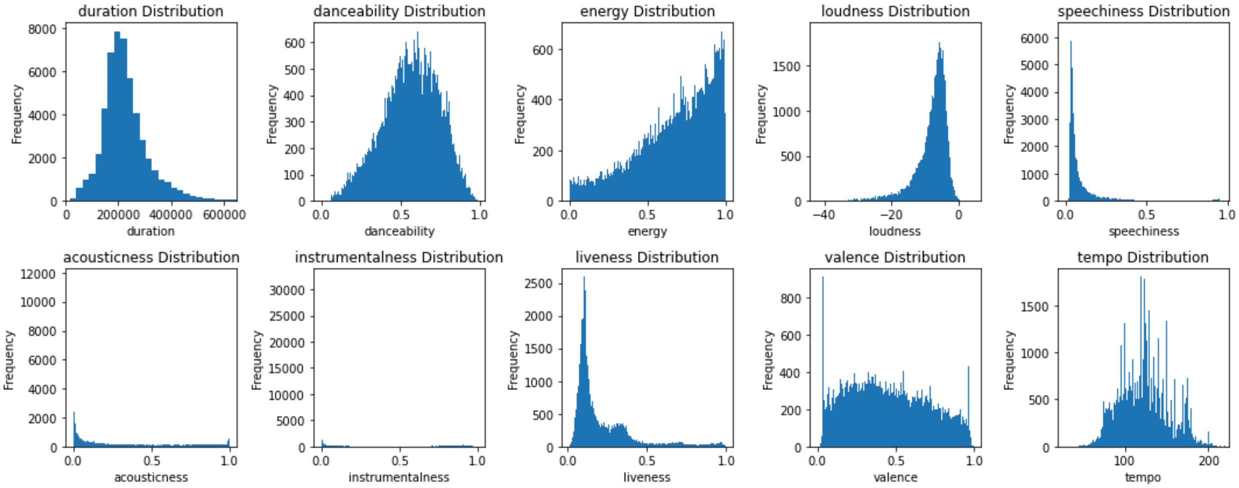


Figure 2: *Distributions of 10 Song Features*

Danceability appears to be the closest to a normal distribution, with a bell-shaped curve centered around a specific value.

2) Is there a relationship between song length and popularity of a song? If so, if the relationship positive or negative?

I created a new dataframe ‘data’, so that I could isolate only the 'duration' and 'popularity' columns. I then used NumPy to calculate the Pearson correlation coefficient of **-0.085**. Additionally, using the stats package, I calculated a Spearman's rank correlation coefficient of **-0.086** as I wanted to see if there was possibly a nonlinear monotonic relationship. I then graphed a scatter plot of the data as shown in Figure 2.

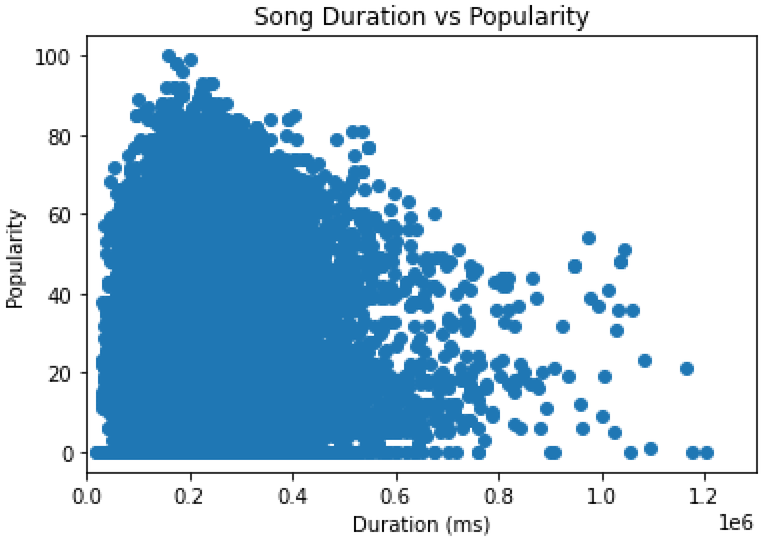


Figure 3: *Scatter plot illustrating the correlation between Song Duration and Song Popularity*

It can be noted that the vast majority of songs (popular or unpopular) are between 0 and 800,000ms (13.33 minutes). There is an ever-so-slight negative relationship between Song Duration and Popularity, though it is so small that it is likely negligible. It can be seen that songs longer than 1,000,000ms are less 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]
I first divided the data into two groups: Songs with explicit ratings and songs with a non-explicit rating. I also extracted the 'popularity' values for each group to compare their popularity distributions. I then performed a Mann-Whitney U test to compare the medians of the two groups' popularity distributions. I chose this test because the data is not normally distributed. Also, because we are comparing medians, our test is more robust to extreme values that might skew means. The explicit songs had a median of **34.0** whereas the nonexplicit songs had a popularity rating median of **30.0**. Running the Mann-Whitney U test gave me a p-value of **3.0679199339114678e-19** which is much smaller than p=0.05. This means that there is a statistically significant difference in median popularity, with explicit songs being more popular than nonexplicit songs.

4) Are songs in major key more popular than songs in minor key?
To determine this I first divided the data into two groups: Songs with a major key (mode = 1) and songs with a minor key (mode = 0). This separation allows me to compare the popularity distribution within

each key category. I chose to perform a Mann-Whitney U test since we can't use a parametric test because the data is not normally distributed. While the median popularity for songs in a minor key is slightly higher (30.0) compared to major key songs (29.0), the test resulted in a p-value of **0.259.** This value is greater than our significance level of 0.05, indicating that there is not enough evidence to conclude a statistically significant difference in median popularity between major and minor key songs based on this data.

5) Energy is believed to largely reflect the "loudness" of a song. Can you substantiate (or refute) that this is the case?
To find out if this is the case, I created a scatter plot to visualize the distribution of energy vs. loudness as shown in Figure 4.

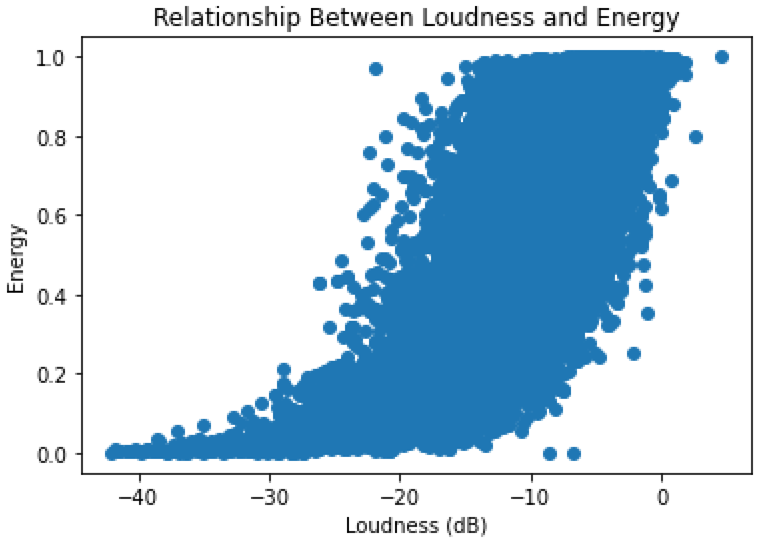


Figure 4: *Scatter plot illustrating the correlation between Average Loudness (decibels) and Energy*

The scatter plot appears to show an exponential growth pattern, suggesting a stronger increase in energy with higher average loudness values. I also calculated the Pearson correlation coefficient and Spearman's rank correlation and obtained values of **0.813** and **0.771** respectively. These values both indicate a moderate to strong positive correlation between energy and loudness. Based on the positive correlation coefficients and the visualization of the data, the analysis provides evidence to substantiate the claim that energy largely reflects 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?
I used NumPy to calculate the correlation between each feature and popularity and found that out of all the features instrumentalness had the highest correlation of **-0.14.** Therefore I would claim that out of all the features, instrumentalness predicts popularity best. However, it's important to note that this is a weak negative correlation, suggesting a very slight association between less instrumental music and higher popularity. I then divided the data into training and testing sets using train_test_split. I then created a

separate linear regression model for each feature using the training data. The performance of each model was then evaluated on the testing set using two metrics: R-squared and Root Mean Squared Error (RMSE). Ultimately, the model that utilized instrumentalness as the predictor variable achieved the highest R-squared value of all the models at **0.0198** and the lowest RMSE value of **21.385.** It should be noted that the R-squared value is quite low, suggesting that instrumentalness alone explains only a small portion of the variance in popularity.

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?
I input all 10 features from question 1 into a linear regression model that was created and fit the training data. I obtained an R-squared value of **0.062** and an RMSE of **20.70**. This is slightly higher than the best single-feature model's R-squared reported in question 6 of **0.028** which indicates a slightly better ability to explain the variation in song popularity. Also, the RMSE is slightly lower than the RMSE of the best single-feature model of **21.14.** This implies a slightly smaller prediction error. An increase in model performance is typically expected as we input more features in the model, however, in this case, the improvement is minimal. This is likely because the original ten features have weak correlations with song popularity, suggesting that other factors likely play a more significant role.

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?
Since the data is not normally distributed nor do the different variables have equal variance, I z-scored the data before doing PCA. I then created a scree plot to visualize the sorted eigenvalues:

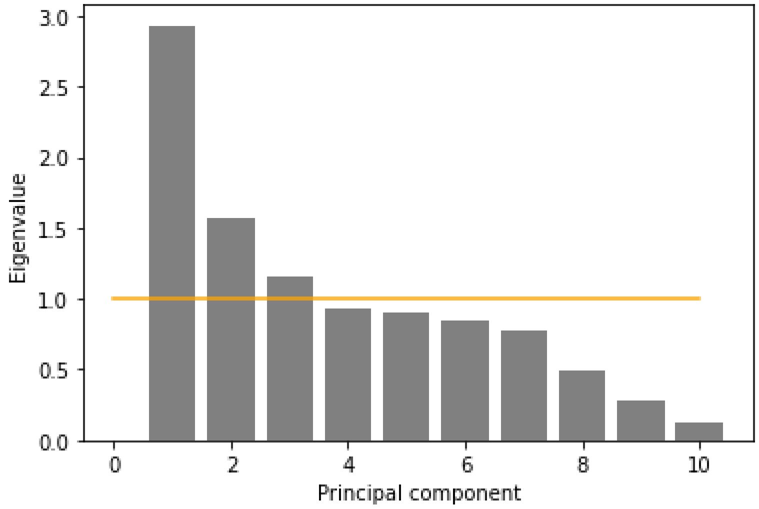


Figure 5: *Eigenvalues of principal components that encompass 10 song features*

To pick the number of factors I interpret meaningfully, I used the Kaiser criterion, meaning that I will only be looking at the principal components with an eigenvalue greater than 1. This means that in this case, I will be looking at 3 components.

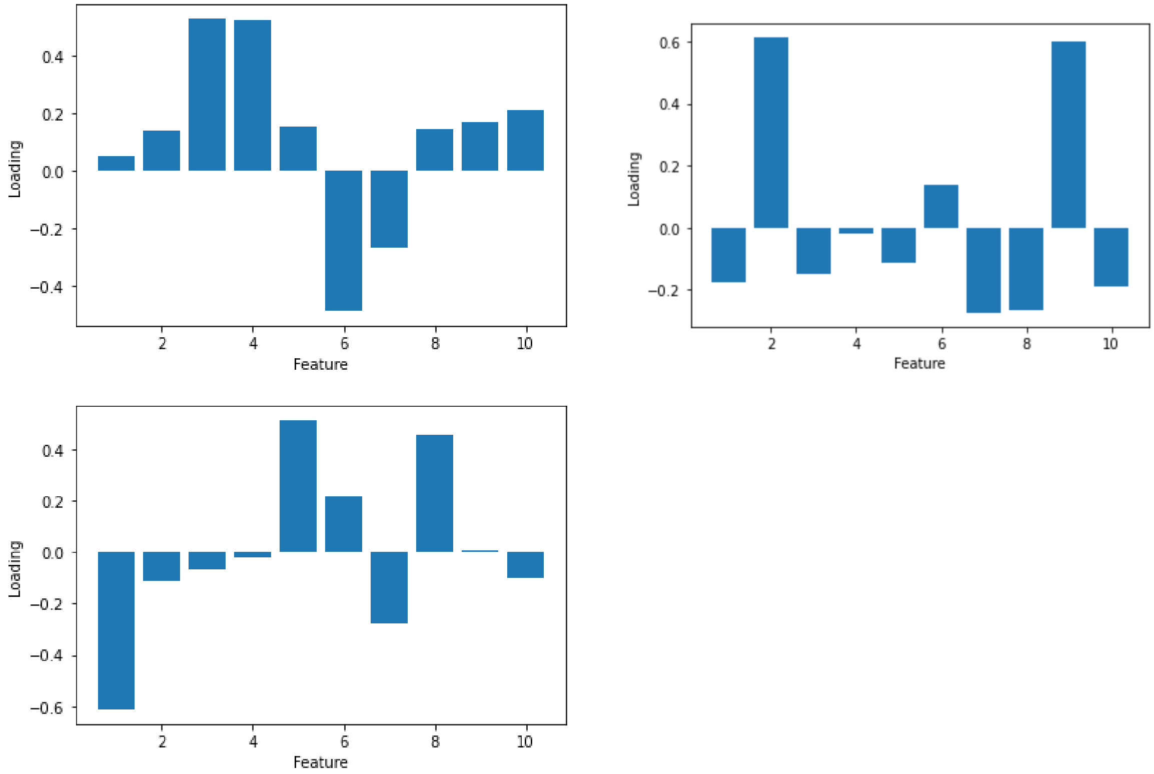


Figure 6: *Component breakdown for all 10 song features. Component 1 (top left), Component 2 (top right), Component 3 (bottom left).*

In this case, the first component accounts for 29.33% of the variance, the second component accounts for 15.63%, and the third component accounts for 11.5%. These 3 components together explain 56.51% of 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]

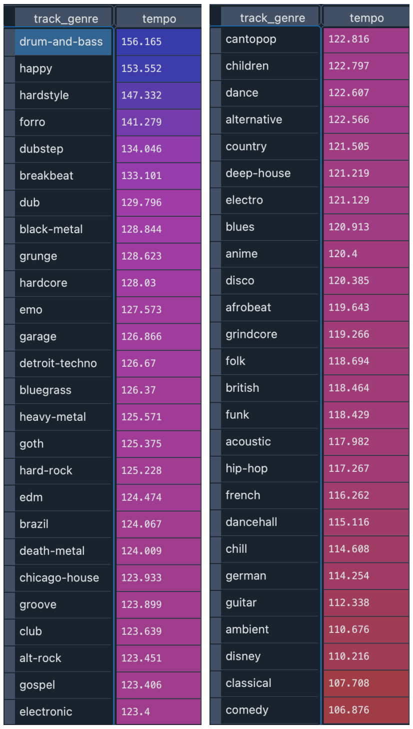
I built a logistic regression model to predict song key (major/minor) based on the valence feature. I evaluated the model using several metrics:

- Accuracy (**62.4%):** This metric reflects the overall proportion of correctly classified songs in terms of their key (major or minor). While not exceptionally high, it suggests the model can make some accurate predictions.
- Precision (**62.4%)**: This value indicates that when the model predicts a song to be in a major key, it's often correct.
- Recall: Proportion of true positives identified by the model **(1.0).** This indicates that the model can find all instances of major and minor keys.

- ROC AUC Score (**0.6069)** suggests the model performs slightly better than random chance at predicting song key based on valence.

10) Which is a better predictor of whether a song is classical music – duration or the principal components you extracted in question 8?
I first converted the qualitative genre label 'classical' (or others) into a binary numerical label (1 for classical, 0 for not classical). I then split the data into training and testing sets for both X_duration (containing duration) and X_pca (containing the first 3 principal components).Because the dataset has an imbalanced distribution, with fewer classical songs compared to other genres, I used class weighting in the logistic regression models to help compensate for the imbalance by giving higher weights to the underrepresented class (classical). I assessed the models using accuracy, precision, recall, and ROC AUC score. Ultimately, my results showed a higher ROC AUC score for a duration of **0.597** compared to PCA of **0.523**. Because of this, we can conclude that duration is a slightly better predictor of whether a song is classical music or not than the principal components I extracted in question 8.

Extra credit: Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions [Suggestion: Do something with the number of beats per measure, something with the key, or something with the song or album titles]
I calculated the average tempo for each genre in the dataset and sorted them by fastest first. The results show a clear trend: genres like hardstyle and dubstep have much higher average tempos compared to chill or ambient music. However what was most interesting was that the two genres with the highest BPM were drum-and-bass and happy, while German had the 6th lowest BPM of the genres.



track_genre	tempo
drum-and-bass	156.165
happy	153.552
hardstyle	147.332
forro	141.279
dubstep	134.046
breakbeat	133.101
dub	129.796
black-metal	128.844
grunge	128.623
hardcore	128.03
emo	127.573
garage	126.866
detroit-techno	126.67
bluegrass	126.37
heavy-metal	125.571
goth	125.375
hard-rock	125.228
edm	124.474
brazil	124.067
death-metal	124.009
chicago-house	123.933
groove	123.899
club	123.639
alt-rock	123.451
gospel	123.406
electronic	123.4
cantopop	122.816
children	122.797
dance	122.607
alternative	122.566
country	121.505
deep-house	121.219
electro	121.129
blues	120.913
anime	120.4
disco	120.385
afrobeat	119.643
grindcore	119.266
folk	118.694
british	118.464
funk	118.429
acoustic	117.982
hip-hop	117.267
french	116.262
dancehall	115.116
chill	114.608
german	114.254
guitar	112.338
ambient	110.676
disney	110.216
classical	107.708
comedy	106.876

***Figure** 7: Genres of music ranked in descending order of tempo (BPM)*