Austin Wu

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? [Suggestion: Include a 2x5 figure with histograms for each feature)

A group of blue and white graphs

Description automatically generated

*Preprocessing: load the data, apply transformation to handle skewed data (log transformation for high skewness, square root transformation for moderate skewness, no transformation for dancebility, energy, and valence), plot*

*Looking at the 10 song features 4 features: duration, danceability, valence and tempo seems to show relatively symmetrical and have somewhat of a reasonable normal distribution. Energy and acousticness are slightly skewed right. Loudness, speechiness, instrumentalness, and liveness are moderately skewed right suggesting that the dataset contains songs that are more “chill” vibe and with less instruments involved.*

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

*A blue and white diagram

Description automatically generated with medium confidenceA graph with blue dots

Description automatically generated*

*Preprocessing: inspect and clean data, transform duration, correlation analysis*

*Looking at the graph on the left, the data is concentrated around shorter durations between 0 to 400,000ms which equates to 6.67 minutes, which suggests that most songs in the dataset have a short duration as most conventional pop songs does. The popularity scores mostly spread across 0 to 80 scores, so while song duration are mostly less than 6.67 mins, the popularity scores for most songs varies. Also, there is also no sign of linear relationship. There are some outliers spread across 2-5,000,000ms, but they generally have low popularity.*

*The right graph also shows a scatterplot of duration and popularity but in logarithmic scale to better visualize the relationship between the two variables. The graph indicates there is not a strong linear relationship, as the points are dispersed across the popularity range for most durations. Furthermore, the Pearson correlation coefficient is approx.*

*-0.055, which again shows a very weak negative correlation showing that song duration has an insignificant effect/influence on the popularity of the song.*

1. 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]

*Preprocessing: load data, find IQR, filter out outliers, conduct t-test, calculate mean and median, plot box plot*

*A diagram of a song

Description automatically generated with medium confidence*

*T-statistic: 9.843878032402834*

*P-value: 7.627809438725804e-23*

*Explicit Songs - Mean: 35.81311416830445, Median: 34.0*

*Non-Explicit Songs - Mean: 32.78772009223939, Median: 33.0*

*The box plot shows the distribution of popularity for explicit and non-explicit songs. Visually, it seems that explicit songs have a slightly higher median popularity compared to non-explicit songs.*

*T-statistic: The t-statistic is 9.84, which indicates a difference between the means of the two groups. P-value: The p-value is extremely small (7.63e-23), which means the difference in popularity between explicit and non-explicit songs is statistically significant. There is a statistically significant difference in the popularity of explicit and non-explicit songs, with explicit songs being more popular on average.*

1. 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]

*Preprocessing: load data, find IQT, filter out outliers, conduct t-test, calculate mean and median, plot box plot*

*A chart of song and song key

Description automatically generated with medium confidence*

*T-statistic: -4.841848673146735*

*P-value: 1.2900559655636913e-06*

*Major Key Songs - Mean: 32.754268424465096, Median: 32.0*

*Minor Key Songs - Mean: 33.70651231577337, Median: 34.0*

*Box Plot: The box plot shows the distribution of popularity for songs in major and minor keys. Visually, the median popularity for songs in a minor key appears to be slightly higher than for songs in a major key.*

*T-statistic: The t-statistic is -4.84, indicating a difference between the means of the two groups.P-value: The p-value is extremely small (1.29e-06), which means the difference in popularity between songs in major and minor keys is statistically significant. There is a statistically significant difference in the popularity of songs in major and minor keys. So no, major key songs are not more popular than songs in minor keys.*

1. 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]

A diagram of a graph

Description automatically generated

*Pearson Correlation Coefficient:*

*Value: 0.775*

*P-Value:0.0000*

*OLS Regression Results:*

*R^2: 0.600*

*Loudness Coefficient: 0.7749*

*P-Value: 0.0000*

*CI analysis:*

|  |  |  |
| --- | --- | --- |
|  | *0* | *1* |
| *Intercept (const)* | *-0.005433* | *0.005433* |
| *Loudness* | *0.769448* | *0.780314* |

*Preprocessing: plot scatterplot between energy and loudness, calculate Pearson correlation, regression analysis, and confidence interval analysis.*

*The Pearson correlation coefficient Is relatively high 0.775 indicates there is a strong positive relationship between energy and loudness. The R^2 value is 0.6, which means that 60% of the variability in energy can be explained by loudness alone. 0.7749 means that for each unit increase in loudness, energy is expected to increase by about 0.7749 units on average. P value 0.000 means that loudness is statistically significant predictor of energy. Furthermore with confidence interval check the Loudness interval is between 0.769448 and 0.780314 which indicates a high degree of precision in the estimates and excludes 0. There we can substantiate the case that energy largely reflect the “loudness” of a song.*

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

*A screenshot of a computer

Description automatically generatedPreprocessing: load data, feature loop ( loop through each feature to build a simple linear regression model, then store R^2), identify best feature (find the highest R^2 value), model summary*

*The best model out of the 10 individual song features from question 1 is instrumentalness, which has a R^2 value of 0.0210. The simple OLS linear regression only explains 2% of variance which is very low in the general sense of R^2 analysis, but comparing to other features (Duration: 0.002987, Danceability: 0.001381, Energy: 0.003128, Loudness: 0.003625, Speechiness: 0.002355, Acousticness: 0.000688, Liveness: 0.001922, Valence: 0.001279, Tempo: 0.000007) Instumentalness is relatively a better indicator. This also indicates that a combination of features might provide a better prediction.*

1. 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?

|  |  |
| --- | --- |
| *R²: 0.4266144146232478* | *Improvement in R²: 0.40550974967957054* |
| *RMSE: 16.42163200390007* | *Improvement in RMSE: 5.034975638934402* |

*Preprocessing: load data, apply log and square root transformation to the features, remove outliers with IQR, train test split, compare the two model’s R^2 and RMSE.*

*First, I tried to use OLS multiple linear regression but yielded also a fairly low R^2 around 0.055 with high RMSE so I turned my way to other regression methods, running Ridge, Lasso, XGBoost, but with random forest the R^2 and RMSE improved the most under multiple regression. With random forest the Percentage Improvement in R^2: approximately*

*1921.89% Percentage Improvement in RMSE: approximately 23.47%, which are substantial amounts. The advantage of using random forest can better explain non-linear relationship which OLS cannot. Also the use of multiple decision trees and averaging results reduces the risk of overfitting.*

1. 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?

A graph of a number of components

Description automatically generated

*Number of meaningful principal components: 4*

*Proportion of variance explained by these components: 0.9670313422796835*

*Preprocessing: load and clean data, apply transformations, remove outlier IQR, perform PCA, determine the number of PCA. Plot histogram for visualization. Results: We have extracted 4 meaningful principal components and can explain at least 96.7% of variance.*

1. 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]

A graph of a curve

Description automatically generated

*Preprocessing: Data Processing (outlier removal), Feature & target variable preparation(feature matric, target variable), model training and evaluation (train-test split, Random Forest classification, model evaluation, ROC curve.)*

*Accuracy: 0.7460118861432593*

*ROC AUC Score: 0.8057045124784626*

*Confusion Matrix:*

|  |  |
| --- | --- |
| *1591* | *1827* |
| *609* | *5564* |

*The ROC AUC score of 0.8057 indicates good model performance in distinguishing between major and minor keys. In question 9 I also implimented Random Forest model, which have shown better results. The model has an accuracy of 74.6% and an ROC AUC score of 0.8057. It is quite effective in predicting whether a song is in a major or minor key. The confusion matrix shows that the model performs well in predicting both classes, with a higher true positive rate for major key songs, correctly predicted 5564 songs as major key, and correctly predicted 1591 songs as minor key.*

*The high ROC AUC score suggests that the model has a good balance between sensitivity (true positive rate) and specificity (true negative rate). This performance improvement over previous models is likely due to the Random Forest's ability to handle non-linear relationships and interactions between features, making it a robust choice for classification tasks.*

1. Which is a better predictor of whether a song is classical music – duration or the principal components you extracted in question 8? [Suggestion: You might have to convert the qualitative genre label to a binary numerical label (classical or not)]

A graph of a curve

Description automatically generated

Duration Model - Accuracy: 0.9950589896137945, ROC AUC: 0.6894217962823557

PCA Model - Accuracy: 0.9950589896137945, ROC AUC: 0.8196810138729185

Both models have a very high accuracy of about 99.51%, which indicates that the vast majority of predictions are correct.

The ROC AUC score of the PCA model (0.8197) is higher than that of the Duration model (0.6894), indicating that the PCA model is better at distinguishing between classical and non-classical songs.

Preprocessing: Converting genre to binary label ((lambda x: 1 if 'classical' in x.lower() else 0)), data transformation, remove outliters, extract principal components of 4, feature matrix, train test split, logistic regression with duration & PCA, evaluate models.

Using the principal components extracted from PCA provides a better prediction of whether a song is classical music compared to using only the song's duration. The PCA model has a significantly higher ROC AUC score, indicating better overall performance in distinguishing between the two classes. This improvement can be attributed to the PCA model's ability to capture more complex patterns and interactions between the features, which a single feature like duration cannot achieve.

Extra credit: Title Analysis

* Identify the most common words in album titles.
* Analyze the distribution of album title lengths.
* Perform sentiment analysis on album titles and correlate sentiment with popularity.

Preprocessing: load data, preprocessing album titles, perform sentiment analysis (VADER)

A chart of a number of words

Description automatically generated

The bar plot shows the most frequent words found in album titles. Words like "2022," "vol," "Christmas," "hits," and "original" appear frequently, indicating that many albums are recent, part of a series, themed around holidays, or collections of popular songs.

A graph of a number of words

Description automatically generated

The histogram shows the distribution of album title lengths (number of words). Most album titles are relatively short, typically between 1 to 5 words, with very few titles longer than 10 words. This suggests a preference for concise album titles.

A diagram of a group of squares

Description automatically generatedThe box plot shows the distribution of popularity scores for album titles categorized by sentiment (Negative, Neutral, Positive). Despite some variations, there is no clear indication that positive sentiments correlate strongly with higher popularity. The average popularity for each sentiment category is:

Negative: 37.05

Neutral: 32.93

Positive: 31.88

Negative sentiment titles have a slightly higher average popularity, but the differences are not substantial.