Table of Contents

Table of Contents..1
Group Number and Names.. 1
Description of Dataset..1
Variables:... 2
Summary Statistics:.. 4
Types of Figures to Plot:... 5
Models.. 6
Evaluation Metrics..7
Workload Sharing...7
Report of Our Observations... 10

Group Number and Names

Group # 4

Members: Sarayu Ramachandra, Eleanor Madderra**, Sachi Satish Kumar, Snarr Pung, Senait Pirani

Description of Dataset

https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs
This dataset contains information about 28,356 songs on Spotify that are in playlists that are reported to be 5 different genres (pop, r&b, latin, rap, and edm).
We want to be able to predict the genre of a playlist a song is in based on data about the song including danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, valence, and tempo. These variables are described below.

Exploratory Data Analytics

Variables:

- **Danceability**
  - Double
  - Quantitative variable
  - Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
  - Some genres might be considered more ‘danceable’ than others. We may be able to predict the genre of the playlist a song is in depending on how ‘danceable’ a track is considered to be.
- **Energy**
  - Double
  - Quantitative variable
  - Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
  - Some genres might be considered more energetic than others. We may be able to predict the genre of the playlist a song is in depending on how energetic a track is considered to be.
- **Key**
  - Double
  - Quantitative variable
  - The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
  - Some genres might use one key more than other genres. We may be able to predict the genre of the playlist a song is in depending on what key a song is in.
- **Loudness**
  - Double
  - Quantitative variable
  - The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative

    loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
    - Some genres might be louder than others. We may be able to predict the genre of the playlist a song is in depending on how loud a track is.
- **Speechiness**
    - Double
    - Quantitative variable
    - Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
    - Some genres might contain more spoken words than others. We may be able to predict the genre of the playlist a song is in depending on the ‘speechiness’ of a track.
- **Acousticness**
    - Double
    - Quantitative variable
    - A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
    - Some genres might be more likely to be acoustic than others. We may be able to predict the genre of the playlist a song is in depending on the acoustic confidence of a track.
- **Instrumentalness**
    - Double
    - Quantitative variable
    - Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
    - Some genres might contain more instrumentals than others. We may be able to predict the genre of the playlist a song is in depending on the ‘instrumentalness’ of a track.
- **Valence**
    - Double

    - Quantitative variable
    - A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
    - Some genres might be considered more positive than others. We may be able to predict the genre of the playlist a song is in depending on the valence of a track.
- **Tempo**
    - Double
    - Quantitative variable
    - The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
    - Some genres could tend to have a faster tempo than others. We may be able to predict the genre of the playlist a song is in depending on the tempo of a track.
- **Playlist_genre**
    - Character
    - Categorical variable
    - The genre of the playlist of the song
    - This is how we will be able to determine genre.

Summary Statistics:

- Mean
    - Determine the typical value for specific features
        - Helping to identify genre-specific features
- Median
    - Provide a central value that isn't as heavily affected by outliers
        - Helping to identify genre-specific features
- Standard deviation
    - Used to determine how spread our or concentrated a specific variable is within genres
- Min and max value
    - Provide range
    - Show the boundaries of our variables and how they differ from one another between genres
- Q1 and Q3
    - We can determine where most of a variable falls

    - Help us understand the distribution of data
    - Can determine if certain variables have concentrated values within a certain range
    - We can determine the typical range of each variable in each genre which can be helpful for predictions

Types of Figures to Plot:

There may be multiple versions of each figure (one for each genre/variable) depending on the type of figure. These figures will help us visualize the effect a variable may have on the genre of the playlist.

- Scatter plot (Senait)
    - Including line of fit
    - Can be used to determine if there is a correlation between *x* variable and genre
    - We will compare the scatterplots to each other
- Histogram (Sachi)
    - Can be used to see how the data is distributed and determine if there is a skew
    - There will be 9 figures, one for each variable. They will be compared to each genre
    - Mean or median of variables vs genre
- Bar chart (Sarayu)
    - We can compare the average tempo, valence, etc. for each of the five genres
    - Mean or median of variables vs genre
- Pie chart (Eleanor)
    - Can be used to see what genres are most likely when *x* variable is (for example) above 0.75, etc
    - Can help us determine if there is a correlation between certain variables and genre
- Radar chart (Snarr)
    - Can feature multiple quantitative variables at the same time
    - Each corner will be a variable and each color will represent the genre

Models

- Logistic Regression (Senait Pirani)
    - Will let us quickly be able to tell which variables are strong predictors of genre
    - Calculates the probability of each genre by modeling the relationship between the variables and the genre classes
        - Allows us to see how each variable contributes to the prediction
- Neural Network (Sachi)
    - Mechanism: Composed of layers of interconnected neurons, each layer transforms input data in complex ways to detect intricate patterns.
    - High Dimensional Data: Neural networks excel with high-dimensional data by capturing complex relationships across multiple features.
    - Non-linear Relationships: Neural networks are powerful for datasets with non-linear patterns, as they can learn and model intricate dependencies between features (like tempo, energy, etc.) and targets (genres).
    - Handling Noise: They perform well with noisy data by adjusting weights during training to focus on meaningful patterns and ignore irrelevant noise.
    - Overfitting Prevention: With techniques like dropout and regularization, neural networks reduce overfitting, which is particularly useful when training on large datasets of varying songs.
    - Suitability for Categorical Targets: Neural networks are well-suited for multi-class classification, making them effective for predicting categorical targets such as genres
    - This makes neural networks a strong choice for accurately predicting song genres based on complex musical attributes.
- Support Vector Machine (SVM) (Eleanor)
    - Can separate classes
    - SVM can distinguish between genres based on song features/variables
    - SVM can use different kernel functions to handle non-linear relationships
        - Linear, polynomial, radial basis function
        - Beneficial for complex datasets
    - Finds an optimal line that maximizes distances between each class
- K Nearest Numbers Classifier (Snarr)
    - Relatively accurate for its simplicity.
    - Alongside feature selection, can be adjusted with different k values and choosing either uniform or distance-based weights for each neighbor.
    - Does not perform well with many dimensions; therefore, should try to reduce the number of features used to lower the dimensions.

    - Prone to overfitting with low values of k, and to underfitting with very large k values.
- Random Forest Classifier (Sarayu)
    - Combines multiple decision trees to improve prediction accuracy
    - works well with high dimensional data (datasets with many features)
    - The features in our dataset likely have non linear relationships, and random forest classifiers are good at finding non linear patterns
    - random forest classifiers are good at handling outliers and noise, which can be prevalent in a dataset with so many datapoints, and about songs as they can vary widely
    - random forest classifiers have a lower risk of overfitting than individual decision trees since they aggregate predictions of multiple decision trees
    - random forest classifiers are known to work well with categorical targets (genres in our case)

Evaluation Metrics

We will be evaluating our models by assessing these metrics: accuracy, precision, recall, F1-scores, and confusion matrix.
By evaluating these metrics, we will be able to tell how accurately the different variables are at predicting the genre of the playlist.

Workload Sharing

Name	**Workload**
Eleanor (20%)	- Upload documents to eLC - Dataset research - Project proposal document - Description of dataset - Summary statistics - Figures to plot - Models - Evaluation metrics - Review and edit document - Jupyter Notebook - Pie Chart - Write short blurb interpreting figure

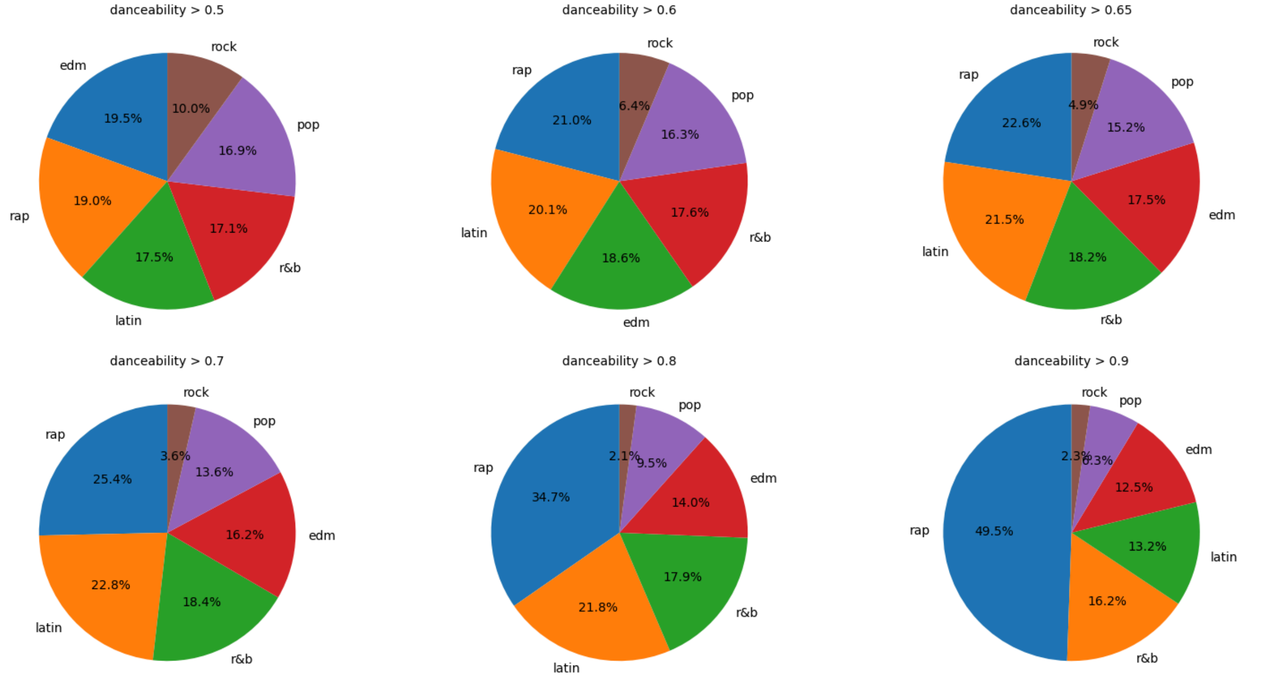
Name	Workload
	- Support Vector Machine (SVM) - Write short blurb interpreting model - Before submission: - Run all lines in order and make sure everything runs and prints correctly - Make sure file has a relevant name - Project report document - Help where needed when necessary - Review and edit document - Confirm all requirements are met and document is titled correctly before submission
Sarayu (20%)	- Dataset research - Project proposal document - Models - Review and edit document - Jupyter Notebook - Figure 2 - Write short blurb interpreting figure - Model 2 - Write short blurb interpreting model - Compile all models into one notebook - The notebook should already contain all of the figures - Ensure the notebook is easily readable - Markdown formatting and grammar - Project report document - Expand on model interpretations - Including our interpretation of all models together - Help where needed when necessary - Review and edit document
Sachi (20%)	- Dataset research - Project proposal document - Models - Review and edit document - Jupyter Notebook - Figure 3 - Write short blurb interpreting figure - Model 3 - Write short blurb interpreting model - Compile all figures into one notebook - Project report document

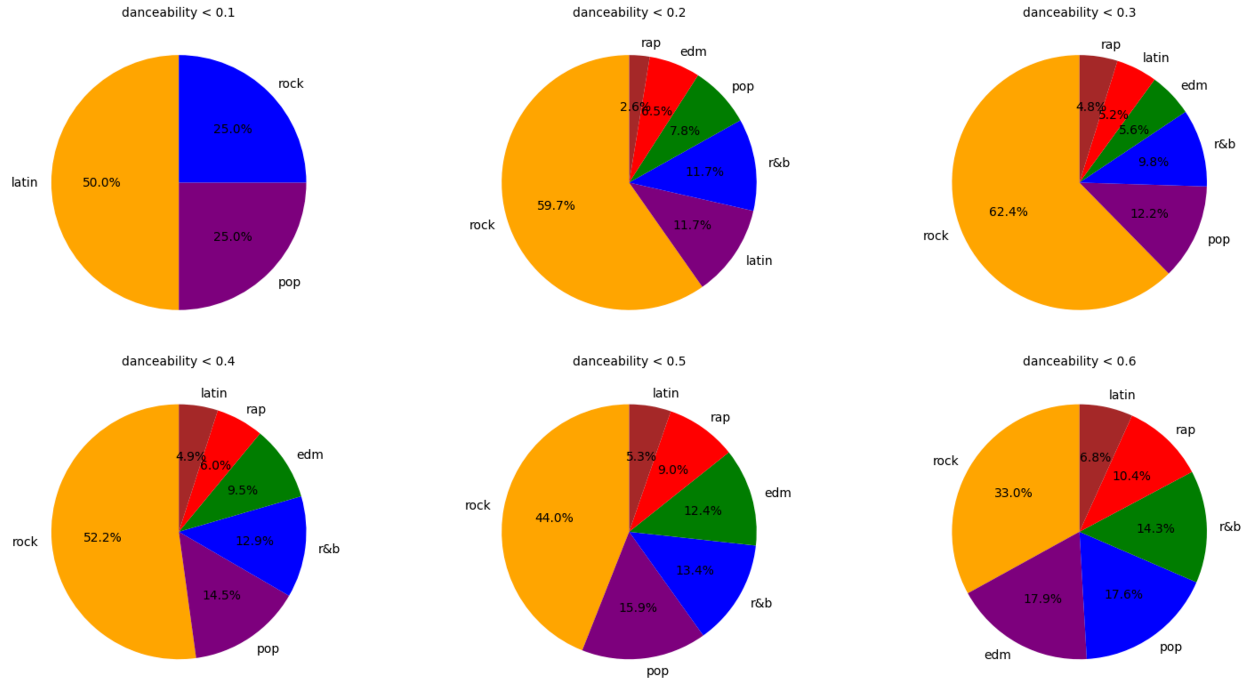
Name	Workload
	- Expand on figure interpretations - Including our interpretation of all figures together - Help where needed when necessary - Review and edit document
Snarr (20%)	- Dataset research - Project proposal document - Models - Review and edit document - Jupyter Notebook - Figure 4 - Write short blurb interpreting figure - Model 4 - Write short blurb interpreting model - Add description on how to run the notebook to the top - Project report document - Final conclusions - Can any or all of our variables predict the genre of a playlist a song is in? - What we learned section - Review and edit document
Senait (20%)	- Dataset research - Project proposal document - Report of our observations - Review and edit document - Jupyter Notebook - Scatter plot - Write short blurb interpreting figure - Model 5 (Logistic Regression) - Write short blurb interpreting model - Project report document - Provide details of the entire process of data analysis - Include previously made figures and results for support - Review and edit document

Report of Our Observations

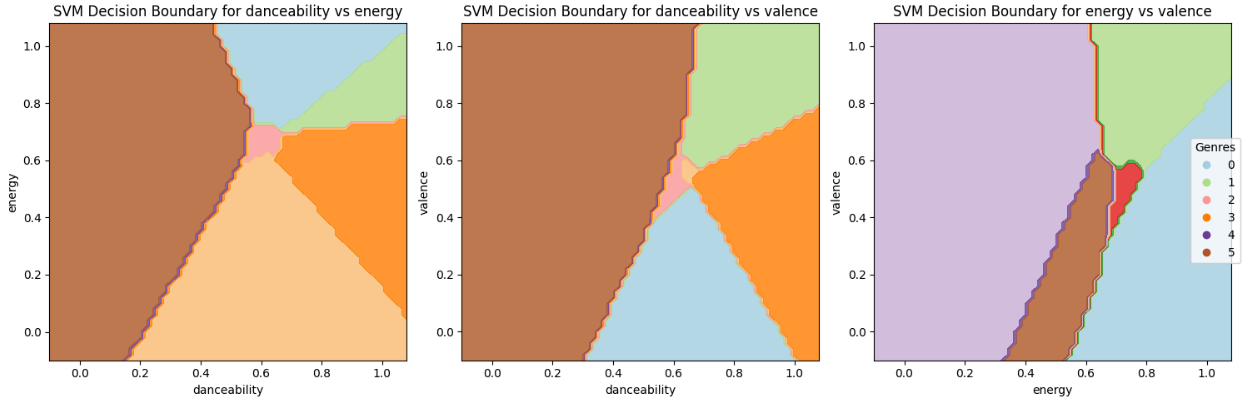
Our chosen dataset is a feasible choice to accomplish our goals of predictive modeling. Looking at the pie charts below, there are clear changes in genre as danceability changes, which suggests that danceability is a relevant variable that can be used to differentiate between genres. This variability proves that our chosen dataset contains potentially statistically significant relationships between our genre and its corresponding independent variables. Additionally, the variability between danceability at different thresholds suggests the potential for feature selection and other analysis, which implies that models trained on this dataset could be able to highlight more subtle patterns in genre prediction

Proportion of songs when danceability value is above or below the specified threshold:





The SVM decision boundary plots shown below provide a clear visual of how genres can be separated based on combinations of features like danceability, energy, and valence. The distinct regions in the plots show that there are meaningful relationships between these features and the genres, making this dataset promising for building predictive models. The complex shapes of the boundaries suggest that genre classification isn't strictly linear, which demonstrates the possibility of correctly forecasting genres, and implies that there is sufficient complexity in this dataset to enable more in-depth pattern exploration through modeling.


SVM Decision Boundary for danceability vs energy
energy
danceability
SVM Decision Boundary for danceability vs valence
valence
danceability
SVM Decision Boundary for energy vs valence
valence
energy
Genres