**Project stage I report**

**Dataset Selection & Project Setup**

**Total: 100 points**

Group **1** (Saranya Pettela, Varun Reddy Pakeru, Keerthi devireddy)

Project title/topic: **Spotify Song Popularity Prediction**

Dataset source (url or resource): [SpotifyDataset](https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db#SpotifyFeatures.csv)

**Task 1 Problem Framing** (10 points)

Your understanding of the project (what you want to study or predict).

Write 4-6 sentences describing: the task (classification/regression/….); who/what benefits from this study; why this dataset fits (size, feature).

Our project focuses on predicting the popularity of songs on Spotify using audio and track features. Since popularity is represented as a numerical score between 0 and 100, this task can be framed as a regression problem. However, for interpretability, we can also treat it as a classification task (e.g., low, medium, high popularity). This study benefits music streaming services, producers, and artists by helping them understand what features contribute to popular tracks. The dataset is large (over 200k tracks) and includes diverse features like danceability, energy, loudness, valence, and tempo, making it suitable for building predictive models.

**Task 2 Dataset exploration** (20 points)

Show the first 10 rows of your dataset; create a dataset dictionary:

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Type: Int / float /… | examples | Missing % |
| Feature 1 |  |  |  |
| Feature 2 |  |  |  |
| … |  |  |  |
| Target variables |  |  |  |

A screenshot of a computer

AI-generated content may be incorrect.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated | A screenshot of a graph  Description automatically generated |

**Dataset Dictionary**

| **Feature** | **Type** | **Example** | **Missing %** |
| --- | --- | --- | --- |
| genre | object | Movie | 0% |
| artist\_name | object | Henri Salvador | 0% |
| track\_name | object | C'est beau de faire un Show | 0% |
| track\_id | object | 0BRjO6ga9RKCKjfDqeFgWV | 0% |
| popularity | int64 | 0 | 0% |
| acousticness | Float 64 | 0.611 | 0% |
| Danceability | float64 | 0.389 | 0% |
| duration\_ms | int64 | 99373 | 0% |
| Energy | float64 | 0.910 | 0% |
| Instrumentalness | float64 | 0.0 | 0% |
| key | object | C# | 0% |
| liveness | float64 | 0.346 | 0% |
| loudness | float64 | -1.828 | 0% |
| mode | object | Major | 0% |
| speechiness | float64 | 0.0525 | 0% |
| tempo | float64 | 166.969 | 0% |
| time\_signature | object | 4/4 | 0% |
| valence | float64 | 0.814 | 0% |
|  |  |  |  |

A screenshot of a computer

Description automatically generated

**Task 3 Feature exploration** (40 points)

Deal with the missing values/outliers; and then create a clean dataset for the next stage analysis.

* Each member works on **at least one important feature** (≥3 features per team total).
* In your summary, clearly states each member’s contributions.

Handling Missing Values:

* The dataset is very clean — no missing values in any feature (0%).
* Therefore, no imputation is required.

Outlier Detection & Handling:

1. Track Duration (duration\_ms)
   * Typical songs range from ~2 to 7 minutes (120,000–420,000 ms).
   * Found 150 tracks shorter than 30 seconds (likely interludes or errors).
   * Found 626 tracks longer than 15 minutes (could be live sets or incorrect entries).
   * Decision: remove extreme outliers outside 30 seconds–15 minutes.
2. Loudness (loudness)
   * Loudness is usually between -60 dB and 0 dB.
   * Found 91 tracks outside this range.
   * Decision: clip values to the valid range [-60, 0] to reduce skew.
3. Popularity (popularity)
   * Scale: 0 to 100.
   * Distribution: mean ~41, median 43, std ~18.
   * Skewed slightly toward mid-range; no invalid values found.
   * Decision: keep as is.
4. Other features (danceability, energy, valence, etc.)
   * All are scaled between 0 and 1.
   * No missing or invalid values.

**Member Contributions**:

* *Saranya Pettela*: Inspected duration\_ms, flagged and handled extreme short/long tracks.
* *Varunreddy Pakeru*: Checked loudness, identified unrealistic values, proposed clipping.
* *Keerthi Devireddy*: Validated popularity distribution, confirmed no missing values, ensured categorical features (key, mode, time\_signature) are consistent.

**Task 4 Save cleaned data** (10 points)

* Save your final cleaned dataset as **newdata.csv** (UTF-8, includes header).
* Place it in your GitHub repo (e.g., /data/newdata.csv). If data is too large/private, include a script to reproduce it and a README with retrieval steps.

**Task 5 Delivery** (20 points)

* **Notebook**: Use Markdown cells to explain every step (why you did it, not just what). Upload to GitHub.
* Export and upload **HTML or PDF** in canvas, together with project stage I summary.