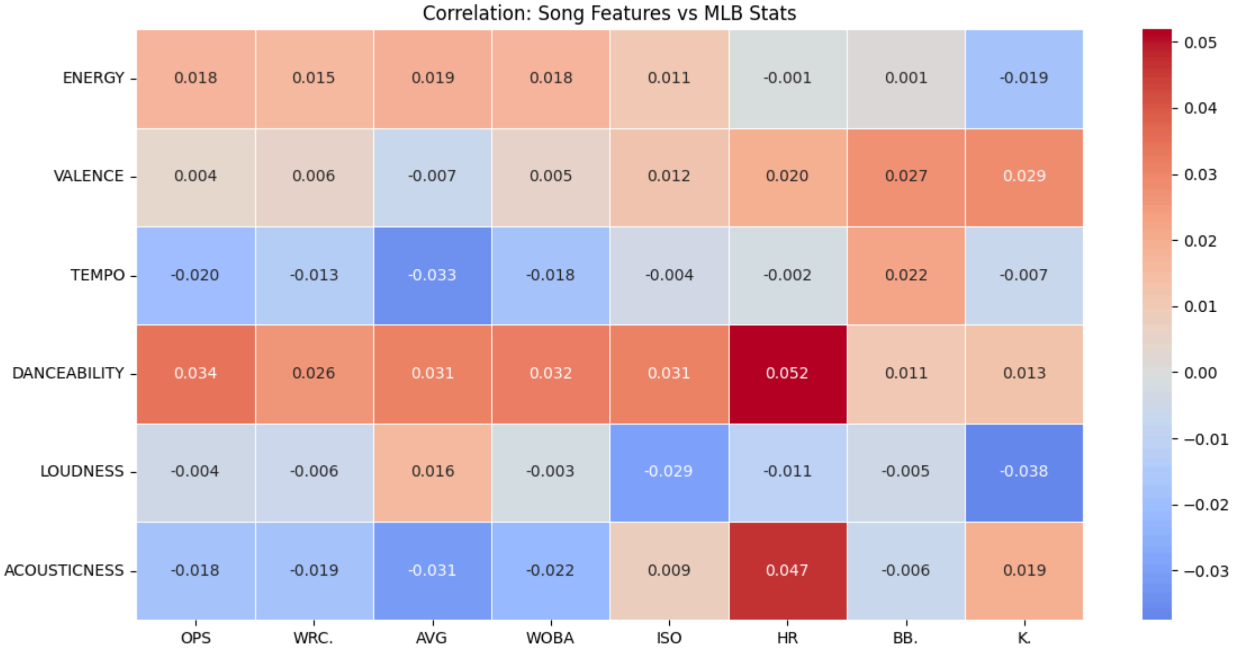
# The Technical Process

## Data Processing

I merged four of the datasets: mlb_data_for_project, walkup_songs_for_project, song_data_for_project, and biographical_data_for_project. I filtered the MLB data to include just home games since that is where walkup songs are most prominent. After merging, I cleaned the data to remove records that had null values for song features. The result was 2,560 complete records (records with MLB player data and walkup song features) stored in a dataframe called stats_clean.

### Exploratory Analysis

The big question arises when thinking about if there really is any correlation between walkup song and performance. I wanted to dive into this and see for myself so I calculated correlations between major song audio features (energy, valence, tempo, danceability, loudness, acousticness) and important MLB metrics (AVG, OPS, wRC+, wOBA, ISO, HR, BB%, and K%). The results weren't too surprising; all correlations were weak. This analysis suggests that song features don't have a strong universal effect on performance.



I wanted to further assess home versus away stats which revealed a clear home field advantage. Players performed better at home (average OPS: 0.663 vs 0.638 away, average wRC+: 85.3 vs 78.9 away). These results are consistent with the well known idea of home field advantage. This home field advantage can arise due to multiple factors whether it be the home crowd or no traveling, rather than walkup music alone. At this point it became clear to me that players' historical song choices revealed individual song preferences varied widely. So rather than attempting to find universal “optimal” songs I wanted to profile each player and find a personalized song recommendation for each of the target players.

## Feature Selection & Rationale

To build a robust recommendation engine, I carefully selected features that act as strong proxies for both **cultural identity** (for music preference) and **objective success** (for performance targeting).

### 1. The Performance Metric: Why wRC+?

For this analysis, I selected **wRC+ (Weighted Runs Created Plus)** as the sole determinant of offensive success, setting the "success threshold" at 100.

I chose wRC+ over traditional metrics (like AVG, RBI, or OPS) because it is the most comprehensive tool for normalizing offensive value.

- **Normalization:** wRC+ is league- and park-adjusted. It levels the playing field, ensuring that a player in a hitter-friendly park (like Coors Field) is not unfairly favored over a player in a pitcher-friendly park.
- **Interpretability:** The stat is scaled so that **100 is always league average**. This provides a mathematically clean baseline: any player with a wRC+ > 100 is objectively contributing more to run production than the average player. This allowed me to create a binary "Successful vs. Non-Successful" mask for the dataset without needing complex thresholds.

### 2. Demographic Features

Music preference is rarely random; it is deeply tied to cultural upbringing and generation. I selected the following features to define a "Similar Player":

- **Birth Country:** Used as a primary proxy for **Language and Culture**. A player from the Dominican Republic is statistically more likely to engage with Latin Pop than a player from rural Canada, regardless of their batting stats.
- **Age:** Music taste is strongly generational. By filtering for players within a **4-year age gap**, I control for the "era" of music the players grew up listening to (e.g., 90s Hip Hop vs. Modern Trap).
- **Batting Hand & Position:** These features control the **player’s role**. The psychological profile (and walk-up routine) of a utility infielder often differs from that of a power-hitting designated hitter.

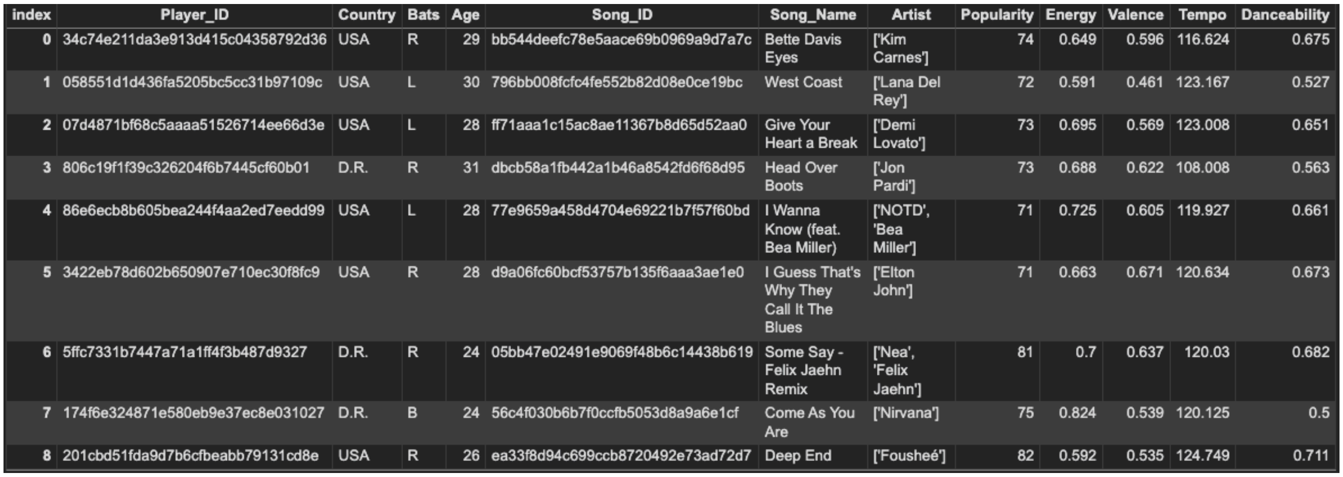
## The Process

1. Identify Similar Successful Players: For each of the target players (submission_ids), I identified player to season combinations of all players where:
    a. The player bats from the same side as the target player,
    b. the player was born in the same country,
    c. the player's age is within 4 years of the target player
    d. wRC+ greater than 100 (above average season),
    e. walkup song features exist.
2. Calculate Averages of Song Features: I computed the average energy, valence, tempo, danceability used across these successful similar players. These averages became the standard when thinking about profiling the target players' song recommendations.
3. Matching Songs: For each available song (songs that a player hasn't used before) I calculated the Euclidean distance between the features of the song and the target features (the averages I previously calculated in step 2). Songs were given a matching score based on the distance, with lower distances indicating better matches. I then filtered the results to only show songs with a popularity greater than 70 to ensure good quality within the song.
4. To counter the issue of not showing enough variety across players with similar profiles I used a hash function on each players' ID to create a unique offset throughout the top 50 matching songs.

## Reflection on Results

Walkup music does not appear to have a strong, measurable effect on a performance. My analysis concludes that there is weak correlation between song features and MLB performance metrics. However, this suggests that the effect of walkup music could be psychological and personal. Since there is no universal pattern we can conclude that a personalized recommendation system would work better than finding an "optimal" song. The variation in song choices across all the players reflects different backgrounds and personalities. Spanish-speaking players may prefer Spanish songs, whereas others choose pop or country songs. Due to these results it is not possible to establish a direct causation between walkup songs and performance. Overall these results do make sense. It is unreasonable to infer that hearing a few seconds of a chosen song can significantly improve performance in a technical and complex game such as baseball.

## Specific Recommendations



| index | Player_ID | Country | Bats | Age | Song_ID | Song_Name | Artist | Popularity | Energy | Valence | Tempo | Danceability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 34c74e211da3e913d415c04358792d36 | USA | R | 29 | bb544deefc78e5aace69b0969a9d7a7c | Bette Davis Eyes | ['Kim Carnes'] | 74 | 0.649 | 0.596 | 116.624 | 0.675 |
| 1 | 058551d1d436fa5205bc5cc31b97109c | USA | L | 30 | 796bb008fcfc4fe552b82d08e0ce19bc | West Coast | ['Lana Del Rey'] | 72 | 0.591 | 0.461 | 123.167 | 0.527 |
| 2 | 07d4871bf68c5aaaa51526714ee66d3e | USA | L | 28 | ff71aaa1c15ac8ae11367b8d65d52aa0 | Give Your Heart a Break | ['Demi Lovato'] | 73 | 0.695 | 0.569 | 123.008 | 0.651 |
| 3 | 806c19f1f39c326204f6b7445cf60b01 | D.R. | R | 31 | dbcb58a1fb442a1b46a8542fd6f68d95 | Head Over Boots | ['Jon Pardi'] | 73 | 0.688 | 0.622 | 108.008 | 0.563 |
| 4 | 86e6ecb8b605bea244f4aa2ed7eedd99 | USA | L | 28 | 77e9659a458d4704e69221b7f57f60bd | I Wanna Know (feat. Bea Miller) | ['NOTD', 'Bea Miller'] | 71 | 0.725 | 0.605 | 119.927 | 0.661 |
| 5 | 3422eb78d602b650907e710ec30f8fc9 | USA | R | 28 | d9a06fc60bcf53757b135f6aaa3ae1e0 | I Guess That's Why They Call It The Blues | ['Elton John'] | 71 | 0.663 | 0.671 | 120.634 | 0.673 |
| 6 | 5ffc7331b7447a71a1ff4f3b487d9327 | D.R. | R | 24 | 05bb47e02491e9069f48b6c14438b619 | Some Say - Felix Jaehn Remix | ['Nea', 'Felix Jaehn'] | 81 | 0.7 | 0.637 | 120.03 | 0.682 |
| 7 | 174f6e324871e580eb9e37ec8e031027 | D.R. | B | 24 | 56c4f030b6b7f0ccfb5053d8a9a6e1cf | Come As You Are | ['Nirvana'] | 75 | 0.824 | 0.539 | 120.125 | 0.5 |
| 8 | 201cbd51fda9d7b6cfbeabb79131cd8e | USA | R | 26 | ea33f8d94c699ccb8720492e73ad72d7 | Deep End | ['Fousheé'] | 82 | 0.592 | 0.535 | 124.749 | 0.711 |

## Individual Justifications:

Recall that my approach assumes that if certain song characteristics correlate with success for similar players, they may benefit the target players as well.

- **Player (34c74e...): "Bette Davis Eyes"** by Kim Carnes matches the moderate energy (0.649) and positive valence (0.596) preferred by successful right-handed American batters in their late twenties.
- **Player (058551...):** Already performing above average, **"West Coast"** by Lana Del Rey maintains the steady, controlled energy (0.591) and valence (0.461) pattern often seen in successful left-handed batters around age 30.
- **Player (07d487...): "Give Your Heart a Break"** by Demi Lovato offers higher energy (0.695) and strong valence (0.569), matching the profile of successful left-handed batters in their late twenties.
- **Player (806c19...): "Head Over Boots"** by Jon Pardi provides a steady, rhythmic pulse with moderate energy (0.688) and positive valence (0.622), aligning with the preferences of successful veteran right-handed batters.
- **Player (86e6ec...): "I Wanna Know (feat. Bea Miller)"** by NOTD; this song’s high popularity paired with high energy (0.725) and upbeat valence (0.605) may help boost confidence and focus for this below-average hitter.

- **Player (3422eb...): "I Guess That's Why They Call It The Blues"** by Elton John maintains the moderate energy (0.663) and positive valence (0.671) associated with successful right-handed batters in their late twenties.
- **Player (5ffc73...): "Some Say - Felix Jaehn Remix"** offers the high energy (0.70) and positive valence (0.637) that younger successful Dominican players tend to utilize, potentially providing a motivational boost.
- **Player (174f6e...): "Come As You Are"** by Nirvana matches the high energy (0.824) intensity pattern observed in successful young switch hitters from the Dominican Republic.
- **Player (201cbd...): "Deep End"** by Fousheé maintains a grounded, moderate energy (0.592) and valence (0.535) associated with elite right-handed performances in younger American players.

## Model Iteration & Engineering Decisions

**Hypothesis:** Not all audio features contribute equally to player performance. Features with a stronger statistical correlation to wRC+ should theoretically carry more weight in the matching process than those with weaker correlations.

**Methodology:** I developed a secondary "Weighted Model" that modified the Euclidean distance calculation. Instead of treating all song features equally, I applied weights to the song features based on their correlation coefficients found in the initial analysis (e.g., *Danceability*, having the highest correlation, was weighted more heavily than *Tempo*).

Results:

Comparing the Weighted Model against the Baseline revealed significant sensitivity:

- **Divergence:** The models produced different song recommendations for **8 out of the 9 target players**.
- **Implication:** This suggests that even minor statistical signals, when prioritized, can alter the "optimal" recommendation. However, given the generally weak global correlations (r < 0.15), I chose to present the first model's results as the primary output to avoid overfitting.

To further enhance the recommendation engine, future models could focus on non-linear relationships:

- **Unsupervised Learning (Clustering):** rather than relying on linear averages, I would implement K-Means Clustering to group successful players into distinct groups  (e.g., "High Energy Latin" vs. "Chill Country"). Target players would then be matched to the centroid of their nearest cluster.

- **Performance Weighting:** Future models could weight the "Success Profile" itself, giving more influence to players with elite performance (e.g., wRC+ > 140) compared to those who are simply above average.