Stage 1 Proposal

TeamID: Team-082
Team Members: Luca Listi (lucal2), Ben Hesse (bhesse2), Tina Ding (landing2),
Yihan Gao (yihang5) (not contributing)

Project Title: Primary Key Predictions

Project Summary:
Our project's goal is to develop a web-based tool that provides historical analysis and predictive modeling for professional sports games. Using datasets from the NBA, our tool will allow users to analyze past game performances, compare player statistics, and forecast future outcomes. By integrating multiple data sources, including historical game records, real-time updates, and social media sentiment, the application will offer fans a comprehensive tool for analysis and predictions.

The platform will be valuable for sports analysts, fans, and sports bettors who want data-driven insights for upcoming professional sports games. Users will be able to view interactive charts, compare teams and players, and use AI-driven predictive models to estimate game results.

Description of Application:
The Professional Sport Game Output Prediction and Analysis Tool will be a web-based platform designed for sports enthusiasts, analysts, bettors, and fantasy league players. The application will collect and process historical and real-time sports data to provide insights into past performances and predict future outcomes.

Our tool will allow users to:

- **View interactive visualizations**
  - Showcasing past team & player performance
- **Compare head-to-head matchups**
  - Allowing users to analyze key statistics for the given matchup
- **Use machine learning models**
  - Forecast outcomes of upcoming games with the help of machine learning
- **Search & Filter Data**
  - Users can look through past sports data and set filters to look at specific stats or metrics

Creative Components (Technically Challenging Features)
The creative components include:

1. **Real-time data integration & live updates:** This can integrate live APIs to get the most recent data during ongoing matches. It can display live updates on player performance and adjust predictions
2. **Predictive Model Dashboard:** The application may utilize this feature to visualize factors that most influenced the outcome. This enables users to comprehend the underlying mechanics of the AI-driven model.
3. **Sentiment Analysis:** The application may incorporate sentiment analysis by fetching data from social media platforms or news feeds to assess public opinion on teams or players. Then, it can overlay this with the historical data visualization.

Usefulness

- **Functions of the Website:**
    1. View historical game performance chart
    2. Predictive analytics and outcome forecasting
    3. Search and compare players or teams
- **Are there any similar websites/applications out there?**
    - There are a handful of different applications that perform similar functions as our project. *NBA Stats*, for instance, offers interactive dashboards and explores player and team performance data, featuring advanced filtering and visualization options. In addition, *TeamRankings* combines historical performance data with predictive analytics to predict game outcomes across various sports.
- **If so, what are they, and how is yours different?**
    - For example, one application that uses league data to predict NBA player performance is *showstone.io*. The main way we are going to differentiate our project from other applications is by using social media sentiment to provide a more comprehensive overview of each player. To do this we can use a library like PRAW (Python Reddit API Wrapper) to extract comments mentioning players or teams from relevant subreddits. Then we can use sentiment analysis tools to score the sentiment on those posts.

Realness

- Describe data sources
    - **From where**
        - Our *primary* data sources are from public sport datasets, we are scraping data from Basketball Reference.
        - Our *secondary* data sources include APIs, such as The Sports DB) for real-time updates and PRAW for social media sentiment analysis.
    - **What format?**
        - The format of data sources are CSV
    - **Data size for NBA Player Dataset**
        - **Size**

        - 635 (number of rows/players)
    - **Cardinality**
        - Players: 534 unique players
        - Teams: 32 unique teams
        - Positions: 5 unique positions
        - Games Player (G): 56 unique values
        - Points Per Game (PTS): 219 Unique values
    - **Degree**
        - 30 (number of attributes of rows/players)
- **Data size for January NBA Games Dataset**
    - **Size**
        - 277 (number of rows, each representing a different game)
    - **Cardinality**
        - Date: 31 unique values (31 different days)
        - Start (ET): 18 unique values (implies 18 different game start times)
        - Visitor/Neutral: 30 unique values (implies 30 different teams played as visitors)
        - PTS: 50 unique values (Team 1s final score)
        - Home/Neutral: 30 unique values (implies 30 different home teams)
        - PTS.1: 57 unique values (Team 2s, opposing teams, score)
        - Attendance: 142 unique values (suggests 142 different attendance numbers)
        - LOG: 42 unique values (Length of Game)
        - Arena: 31 unique values (implies games were played in 31 different arenas)
    - **Degree**
        - 12 (columns used)
- **Information the Data Source Capture:**

  In terms of the **historical sports datasets,** the data source captures following information:
    - *Game-related statistics*: date and time of the game, home and away team identifier, location, and the final scores and outcome (overtime status, and win/loss).
    - *Team-related statistics*: total points scored, numbers of rebounds/assists, shooting percentages, numbers of fouls committed
    - *Player-related statistics:* number of points/assist/blocks, minutes played. The shooting status such as three-point percentage. It may also have some other metrics like usage rates.

  In terms of the **real-time data,** the data source captures following information:

- *Real-time game updates:* it provides ongoing game scores and any statistical updates. It also gives current player performance.
- *Current team/player data:* it showcases any adjustments to current rosters

In terms of the **social media data** (via PRAW for Reddit sentiment analysis):

- *User-generated content:* it shows any comments and posts associated with players, team, and matches.
- *Sentiment metrics:* current trends in public opinion that can be overlaid on historical and real-time stats.

Functionality that the Website Offers

- **Users will be able to:**
  - Search for teams & players and view detailed stats
  - Compare two teams or players across multiple metrics
  - View interactive historical game trends & charts
  - Generate predictive insights for upcoming games
- **CRUD Features**
  - **Create:** Users can save favorite teams/players
  - **Read:** Users can view and analyze data through charts & tables
  - **Update:** Users can adjust filters and customize reports
  - **Delete:** Users can remove saved teams/players from their watchlist
- **Low-Fidelity UI Mockup**

Project Work Distribution

- **Ben Hesse**
  - Backend API integration, setting up database storage
- **Luca Listi**
  - Frontend Development (UI Design, data visualizations)
- **Tina Ding**
  - Data collection, cleaning, and statistical analysis
- **Yihan Gao**
  - Nothing, team member has yet to contribute to a single Group activity or work on the group project. Also has not responded to any of our emails or texts.