Iteration #04
Dataset and Project Overview

Tea Adams, Shriniketh Mukundan, Jay Shah

1. Dataset Description

Dataset Link:
(2010-11 NBA Schedule — Basketball-Reference.com, Injury History — Kaggle).

Description and Source:
Our dataset includes historical NBA data from 2010–2020, consisting of games played, player injury records over the same period, player stats and draft stats for the players. The data captures game-level information (date, teams, scores, attendance) and player-level statistics (minutes played, points, assists, blocks, steals, etc.), alongside injury occurrences. We plan to add player rosters and demographic information (e.g., age, position) to create a complete dataset for analysis.

Relevance and Suitability:
This dataset is well-suited for our project because it provides extensive, structured, and time-bound data on NBA player performance and injuries. It enables us to explore relationships between playing time, workload, and injury likelihood. Since the goal is to predict and prevent injuries, the combination of performance statistics and historical injury data provides the essential variables for training a predictive model that can generate actionable insights for teams and medical staff.

2. Tools and Methodologies

Our team will use a combination of Excel, SQL, Python, and Tableau to manage, analyze, and visualize the data.

Data Management & Cleaning: SQL (PostgreSQL) will be used for relational database design, data cleaning, and preprocessing, while Excel will support quick data exploration and basic validation tasks.

Modeling & Analysis: Python with pandas, NumPy, and scikit-learn will be used to perform exploratory data analysis and build a logistic regression model, chosen for its interpretability and effectiveness in binary classification (injured vs. not injured).

Visualization & Reporting: Tableau will be used to create interactive dashboards visualizing injury trends by player, position, and team workload, enabling easy insights for both technical and non-technical audiences.

Collaboration: GitHub will serve as our version control system, and shared cloud notebooks (e.g., via Databricks or local Jupyter) will facilitate collaborative development and reproducibility.

Tools, Frameworks, Libraries:

Project Area	**Tools/Frameworks**	**Reasoning**
Database	PostgreSQL, SQLAlchemy, Python pandas, Prefect, Azure DataBricks	Will be helpful for data mining, transformation, and modeling
Predictive Modeling	Python - NumPy, matplotlib, pandas, scikit learn, etc.	Helpful for building models, evaluating, and interpreting
Visualizations	Tableau, Power BI	Provide friendly visuals for injury trends
Collab Tools	GitHub, Google Drive	Easy ways to collaborate with one another

3. Preliminary Timeline

Team Progress Tracker (Google Sheets).

Week	Dates	Phase	Owner	Key Tasks	Deliverables	Status
1	September 29	Project Setup & Scoping	Team	Define the project objectives and scope, identify and document data sources, outline KPIs and success metrics, and set up data storage and version control.	Project charter and defined objectives document	Complete
2	October 6	Data Acquisition	Jay	Collect raw data from all identified sources, verify data access and permissions, assess data quality, and document data fields and formats.	Compiled raw dataset with data source documentation	Complete
3	October 13	Data Cleaning I	Jay	Handle missing values, standardize data types and formats, and identify and remove duplicates or outliers.	Initial cleaned dataset (version 1)	Complete
4	October 20	Data Cleaning II & Integration	Jay	Merge data from multiple sources, validate consistency across key fields, and finalize the data dictionary.	Fully cleaned and integrated master dataset with data dictionary	Complete
5	October 27	Data Compilation & Exploration	Team	Conduct exploratory data analysis (EDA), generate summary statistics and visualizations, and identify correlations, patterns, and anomalies.	Exploratory data analysis summary and visual insights	Complete
6	November 3	Pre-Modeling Preparation	Tea	Engineer relevant features, split the dataset into training and test sets, and normalize or encode variables as needed for modeling.	Model-ready dataset with engineered features	In Progress
7	November 10	Model Development	Tea	Develop baseline models using selected algorithms, evaluate model performance metrics, and refine parameters for improved accuracy.	Baseline model results and performance summary	In Progress
8	November 17	Model Validation & Refinement	Tea	Perform cross-validation and error analysis, compare alternative models, and document final model performance and rationale.	Validated final model with documentation of evaluation metrics	Not Started
9	November 24	Reporting & Visualization Build	Shriniketh	Create dashboards and visualizations, summarize key insights, and draft the written report or presentation materials.	Draft report and visualization dashboard	Not Started
10	December 1	Final Review & Presentation	Team	Review and quality-check all outputs, finalize and deliver the presentation to stakeholders, and archive project files and documentation.	Final report, presentation deck, and archived project files	Not Started

Figure 1: Progress Tracker

4. Team Member Contributions

Tea Adams – Data Analyst / Predictive Modeling Lead
Tea's main responsibilities include Python-based data analysis, predictive modeling, and interpretation of results. She brings prior experience as a Data Analyst, which is crucial for building and tuning the logistic regression model, evaluating model accuracy, and translating results into actionable insights.

Jay Shah – Data Engineer / Database Designer
Jay focuses on SQL database design, data cleaning, and schema creation. His experience with data aggregation and SQL enables efficient integration of player, game, and injury datasets. Jay ensures that all datasets are consistent, complete, and structured for analysis.

Shriniketh Mukundan – Data Visualization / Reporting Lead
Shriniketh leads visualization and reporting. Leveraging his background in business intelligence and Power BI, he will develop dashboards that illustrate player injury risk, trends by position, and team-level insights.

5. Progress and Next Steps

Collaboration & Progress So Far:
The team has successfully cleaned and combined the datasets, and several new columns have been created to support predictive modeling. Our logistic regression model is currently underway and is expected to be completed by the end of the week. Team members are collaborating closely through GitHub and shared notebooks, ensuring that all data preparation, modeling, and visualization work is synchronized and reproducible.

Here is the link to our repository: NBA Injury Modeling GitHub Repository.