STA160 Project Plan

Team Number: 5
Team Name (optional): Team Sharks

1. Team Members

- **Leonel Garibay-Estrada** (Data Science, 4th Year, skill highlights)
- **Jiarui Hou:** (Statistics, 4th year, Modeling and ML)
- **Tyler Venner*** (Statistics, 4th, ML and probability)
- **Macy Chen** (Statistics, 4th, Data analysis and ML)
- **Alberto Ramirez** (Statistics, 4th year, ML and regression)

** denotes team leader.*

2. Motivation & Problem Statement

What question or problem are you addressing? Why is it important (academic, practical, or social relevance)? What is the intended audience/use?

We are trying to discover the patterns of systematic bias between players' actual contract and expected market value. After accounting for a player's performance and true playing role, do systematic market biases or inefficiencies remain—and how have they evolved over time? Does a player's age, nationality, draft status, or other factors not related to on-court performance explain the variance in salaries?

Intention is to help better understand what aspects of performance and contextual factors lead to certain contracts and can be used to identify certain players who are under/overpaid. This is important for NBA teams to try to identify if their contracts reflect the value the player brings. It is also important for NBA players and fans to understand individual player value differences.

3. Data Sources

- **Datasets: NBA Wrapper API, https://www.basketball-reference.com/, https://www.spotrac.com/nba/ .** We will have performance stats (X) and contract value stats (Y) and contextual variables (Z).
- **Access:** player stats (X) API Wrapper `nba_api` which gives player stats. And web scraping for contract evaluations and non-performance stats.
- **Preprocessing:** Since we use multiple sources, we need to combine datasets via key. That key will be the players names, and will smartly use a secondary key for name conflicts. Also, some names may contain middle names or alternative names, and we will review on a case by case basis each difference. We will then identify performance

metrics, feature engineering other metrics, and identify any contextual features.
- **Ethics/Privacy:** All data (performance and contextual stats) used are publicly available information.

4. Methods & Tools

- **Methods:** Our workflow is broken down into four parts:
    - Role Discovery: We will use Unsupervised Learning (k-means or other clustering models) on performance matrix X to partition players into k latent roles, C. We will use PCA for visualization.
    - Role interpretation: We will cross-reference our latent roles C with traditional positions (point guard, center) to create true role labels.
    - Inefficiency Modeling: We then build k separate multiple regression models (Linear, Ridge, Lasso) to predict $log(Y)$ from X, one for each role. The residuals will be our mispricing (or inefficiency) metric.
    - Bias Analysis: We will standardize the residuals and use a final linear model to test if the residuals is explained by contextual features Z.
- **Tooling:** Python (Scikit-Learn, Pandas, Numpy, Matplotlib, Seaborn), Streamlit, PowerBI
- **Computing:** Works on any computer with CPU (no cloud).
- **Reproducibility:** Git/Github, Requirements.txt for dependencies

5. Expected Outcomes

What deliverables will you produce (figures, models, dashboards, docs)? How will you evaluate success (metrics, baselines, ablations, interpretability)?

The project will produce a combination of analytical models, figures, and interactive dashboards to communicate findings about player market efficiency and bias: we will produce clustering and regression models to identify latent player roles and predict expected salaries based on performance, as well as a bias analysis model to test how contextual factors like nationality, draft status, or age influence residual salary differences. The results will be communicated through spider plots, PCA scatter plots, box and violin plots for salary distributions, coefficient plots for bias effects, and an interactive dashboard that highlights over- and under-valued players.

We will have:

1. A set of k player roles, identified via unsupervised learning, visualized with spider plots.
2. A ranked list of the most over and under valued players in the league, relative to their player role.
3. A final interpretable model that quantifies impact of bias variables (like team name) on player contract valuations.
4. An interactive dashboard which allows users to explore roles, find specific players, and identify market inefficiencies.

6. Risks & Mitigations

Identify key risks (data availability/quality, scope creep, feasibility) and mitigation plans (backup datasets, simplified scope, staged milestones).

Multicollinearity could exist due to possible correlations between certain performance metrics and contextual variables. We can mitigate this by building and analyzing correlation matrices before any modeling is done.

Data availability and quality. Our quality of our data depends on if we can correctly scrap high quality data. Also, merging by player name may be challenging. We will look into fuzzy string matching for this problem, and use effective fallback strategies if the datasets do not find matching paired names. Some players could have missing values for certain variables. During data cleaning, we will identify any missing data and reference backup datasets as needed.

7. Timeline & Milestones

Customize weeks/dates to match the course deadlines.

Week / Date	Goal / Deliverable	Responsible
Week 3	**Defining topic/Project Plan**	**Everybody gets together and discusses project plans.**
Week 4	**`nba_api` pipeline built. Salary Data Scraped.**	**Each week, each team member is responsible for a related module with defined input/output. So everybody.**
Week 5	**Data cleaning and Feature engineering (X, Y, Z). Clustering complete. Intermediate project presentation.**	**Each week, each team member is responsible for a related module with defined input/output. So everybody.**
Week 6	**K salary models built.**	**Each week, each team member is responsible for a related module with defined input/output. So**

		everybody.
Week 7	**Systemic Bias Modeling Complete**	**Each week, each team member is responsible for a related module with defined input/output. So everybody.**
Week 8	**Visualizations and Dashboards completed**	**Each week, each team member is responsible for a related module with defined input/output. So everybody.**
Week 9	**Analysis and Presentation completed. Turn in the final report.**	**Each team member will write a part of the final report.**
Week 10	**Practice Presentation. And presentation due.**	**Team members will come together to practice the presentation.**

8. Division of Work

Focus on tasks first, then the owner(s). Use one or many members per task as needed.

We will follow standard software development best practices and split up the tasks into modules which have defined input and output with only one particular purpose. Each team member is responsible for their respective modules. In addition, if a team member needs help, other members will help the other team member. So, the division of work is dynamic and changing week to week depending on that week's needs and goals and modules. In the end, we will have a list of modules each person built and their contributions.

9. Constraints & Policies (Course-Specific)

- **Collaboration:** Schedule weekly meetings to work on the project (in-person or via zoom) and identify that week's modules and how the modules talk to each other. Communicate with one another on discord for any updates and issues. Make sure we complete each week's goal. Help any team member who needs help.
- **AI Usage:** Potential troubleshooting on programming errors.
- **Attribution:** Cite all sources.