

Source Validation

Checkpoint 3

Overview.....	1
Group Work Option.....	1
Assignment Objectives.....	1
What to Include in Your Checkpoint.....	2
Submission Details.....	3
Grading & Rubric (10 points total).....	3
Additional Notes.....	3

Overview

In this checkpoint, you will critically evaluate the quality and reliability of the dataset(s) selected for your project. A high-quality dataset is essential for credible analysis and trustworthy insights. In your memo, assess completeness, accuracy, consistency, and potential biases in the data. Explain why evaluating data quality is important: it helps prevent misleading conclusions and identifies limitations before you begin modeling. Write as if you are reporting to a project stakeholder, emphasizing that understanding data limitations upfront is critical for trustworthy analysis.

Group Work Option

You may work in groups of up to 3 students, but each student must maintain their own GitHub repository. Collaboration on topic and data collection is encouraged, but each student must submit an individual memo and ensure their repository contains all deliverables.

Assignment Objectives

By completing this checkpoint, you will:

- 1. Assess and summarize data sources** - clearly describe origin and contents of dataset(s).
- 2. Evaluate data completeness and consistency** - check for missing values, coverage gaps, and inconsistent formatting or units.
- 3. Identify biases or limitations** - recognize any data quality issues, anomalies, or sample biases that could affect your analysis.

- 4. Develop an initial cleaning plan** - outline how you will address data issues. **5. Write a professional memo** - communicate your findings in a concise, formal report.

DATA 6560 - Sports Analytics - Checkpoint 3 Fall 2025 What to Include in Your Checkpoint

Project Title & Data Overview	Provide your working project title and a brief context. Summarize the dataset(s) you are evaluating and why they are relevant.	What is your project about, and what data are you using?
Data Source Summary	Describe each data source and how it was collected or obtained.	Who collected the data? Is it official league data, crowd-sourced, or from a third party? When and how were the data recorded or compiled?
Data Structure & Content	Outline the structure and key variables of the dataset(s).	How many records (games, player entries, plays) are there? What are the main variables, units, and formats? Are multiple tables or files involved, and how are they related?
Data Completeness & Consistency	Assess whether the data cover the full scope needed and use consistent conventions.	Are there missing values or gaps (by team, season, date, etc.)? Are all fields consistently formatted (units, categories)? Do you see any duplicate or mismatched records?
Quality Issues & Potential Biases	Identify any data issues or biases.	Are there out-of-range values, errors, or irregular entries? Could the data be biased (over- or under-represent certain teams, games, or time periods)? How might these issues affect your analysis?
Initial Cleaning / Preparation Plan	Outline steps you will take to clean and prepare the data.	How will you handle missing or inconsistent data? What transformations or encoding steps might be needed (parsing dates, converting units)? What tools or methods (scripts, software) will you use?

Next Steps	Summarize your immediate tasks following this checkpoint.	What actions will you take next (gather additional data, begin cleaning, perform exploratory checks)? What are your priorities for the coming week?
------------	---	--

DATA 6560 - Sports Analytics - Checkpoint 3 Fall 2025 Submission Details

Document Type	Google Doc (Professional Memo Format)
Submission	Share your Google Doc link on Google Classroom and upload a PDF copy to the /reports folder in your GitHub repo.
File Naming	Use the format LastName_FirstName_CP3.pdf (Doe_Jane_CP3.pdf). Include your name(s), project title, and checkpoint heading at the top of the memo. EACH STUDENT MUST UPLOAD TO INDIVIDUAL REPO
Format Tips	Include your name(s), project title, and checkpoint heading at the top. Use headings or bullets to organize content clearly.

Grading & Rubric (10 points total)

Your Checkpoint 3 will be evaluated on the following criteria:

Data Source Clarity & Relevance	Data sources and their relevance to the project are clearly described.	2
Completeness & Consistency	Assessment of missing data, consistency, and accuracy is thorough and clear.	2
Bias & Limitations	Potential biases, errors, and data limitations are identified and discussed.	2
Cleaning Plan Feasibility	Proposed data cleaning steps are practical and well-justified.	2
Professional Structure & Writing	Memo is well-organized, concise, and professionally written (grammar, style).	2

Additional Notes

- Keep your writing professional and focused. Aim for clarity and completeness over length.
- Write as if you are addressing a sports executive or coach. Be concise and avoid jargon. •
- Clearly state any assumptions or uncertainties about the data.
- This memo contributes to your semester project portfolio; maintain consistency with previous checkpoints and your repository's organization.

1. Data Source Overview

DATA 6560 - Sports Analytics - Checkpoint 3 Fall 2025

The dataset, titled *NBA Players Performance and Salaries*, was sourced from Kaggle and covers the **2017–2018 NBA season**. It merges player salary data with performance statistics, enabling analysis of how player efficiency aligns with compensation.

Key contents include:

- **Player Info:** Name, team, position, age
- **Salary:** Total salary for the season
- **Performance Metrics:** Games played, minutes, points, rebounds, assists, steals, blocks, turnovers, shooting percentages, and advanced stats like PER (Player Efficiency Rating)

2. Data Completeness and Consistency

Findings:

- **Missing Values:**
 - A few players have missing values for advanced stats like PER or shooting percentages, often due to limited minutes played.
- **Formatting Issues:**
 - Salary values are stored as strings with dollar signs and commas (e.g., "\$5,000,000"), requiring conversion to numeric format.
 - Team names and positions are mostly consistent but may need standardization for grouping.
- **Coverage Gaps:**
 - Players with very low minutes (e.g., <100 minutes) may skew efficiency metrics due to small sample sizes.

3. Biases and Limitations

- **Sample Bias:** Players with minimal playing time may appear highly efficient or inefficient due to small sample sizes.
- **Salary Lag:** Some contracts reflect past performance or market dynamics, not current-season output.
- **Role Differences:** Comparing bench players to starters without adjusting for usage or minutes

can misrepresent value.

- **No Contextual Factors:** The dataset lacks injury data, defensive assignments, or team context, which can influence performance.

4. Initial Data Cleaning Plan

To prepare the data for analysis, I will:

- **Convert Salary to Numeric:** Strip symbols and commas, convert to float.
 - **Filter by Minutes Played:** Exclude players with fewer than 250 minutes to reduce noise.
 - **Handle Missing Values:** Drop or impute missing performance metrics based on position averages.
- DATA 6560 - Sports Analytics - Checkpoint 3 Fall 2025
- **Create Salary Tiers:** Use quantiles (e.g., top 25%, middle 50%, bottom 25%) to define salary groups.
 - **Normalize Efficiency Metrics:** Standardize PER and other stats for fair comparison across roles and minutes.

5. Conclusion

This dataset provides a strong foundation for analyzing the relationship between salary and player efficiency. By addressing formatting issues, filtering outliers, and accounting for sample bias, the analysis will offer meaningful insights into how NBA teams might identify undervalued talent and optimize roster spending.