DATA 6560 - Sports Analytics - Checkpoint 6 Fall 2025 # Analysis Memo #1

Checkpoint 6

Overview

Checkpoint 5 focused on exploration.

Checkpoint 6 is where you run a first real analysis that answers your question. Pick one primary method, report what you did, show results, and explain what they mean in plain language.

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. Choose a method that fits your question and the cleaned data from
2. Build a simple baseline and, if helpful, one small improvement version
3. Report what you did, why you did it, and what the results mean
4. Show basic checks so a reader can trust the numbers
5. End with next steps

DATA 6560 - Sports Analytics - Checkpoint 6 Fall 2025 ## What to Include in Your Checkpoint

Question and Outcome	A short reminder of your question and the outcome you analyzed	What are you trying to understand or predict? Who cares about the answer? What is the exact outcome column you used, such as win, points, sales, conversion?
Data Used	Name the cleaned file from CP4, your row definition, time span, and any filters	What does one row represent? Which seasons, games, or teams are included? What did you filter out and why?
Method Choice	Say what method you used and why it fits your question	Why does this method help more than a table or chart alone? What is your baseline version? What small improvement did you try, if any?
Analysis Spec	A small spec table with outcome, predictors or groups, sample, row definition, and the simple formula or rule	Which inputs did you include and why? What direction did you expect for each input?
Results Checks	A readable table with the key numbers. Add one or two figures only if they help Quick checks that increase trust	What is the size of the effect? Is it large enough to matter in sports terms? If classification, what is accuracy compared to a simple baseline such as always picking the home team? If regression, report MAE or R^2 and say what that means in your sport Are your inputs on the right scale? Did you avoid using future information by accident? Do results look similar if you change the sample to a recent season or remove obvious outliers?
Interpretation	Translate the result to plain English in one or two sentences per key finding	What changes in the real world when your input moves by 1 unit? Who would use this and how?
Limitations	Name the top one or two limits of this first pass	What is missing or possibly biased? What data or features would help?
Next Steps	Short, concrete plan	What will you test next, such as a new feature, a better split, a second method, or a tighter robustness check?

DATA 6560 - Sports Analytics - Checkpoint 6 Fall 2025 Model and Analysis Standards

● Start with a simple baseline, such as a group average comparison or a one-predictor model ● Add one reasonable improvement if it helps, such as adding a key predictor or using a rate per

minute or per possession

- Report at least one evaluation method, some examples below:
    - ○ Regression: MAE or R^2. Also say what the typical error means in your sport ○ Classification: accuracy and a simple baseline rate such as always predicting the majority class
    - ○ Comparisons: group means and a clear effect size or a simple confidence interval
- Keep figures clean and labeled. Put a one-line takeaway under each figure

Submission Details

Document Type	Google Doc in professional memo style, export to PDF
Submission	Share the Google Doc link on Classroom and upload the PDF to your repo in /reports. Save any figures in /figures
File Naming	LastName_FirstName_CP6.pdf (example: Doe_Jane_CP6.pdf). Include your name(s), project title, and the checkpoint heading at the top of the memo. EACH STUDENT MUST UPLOAD TO INDIVIDUAL REPO
Format Tips	Upload your model workbook in Excel or your script if you coded. Include any small dataset extracts used in the analysis

Deliverables

1. Analysis Memo 1 in PDF
2. Model workbook or script, plus any figures used

Grading & Rubric (10 points total)

Your Checkpoint 6 will be evaluated on the following criteria:

Method Fit	The method fits the question and CP4 data. Clear baseline. Any improvement is sensible and modest	3
Spec and Reporting	Clear spec for outcome, inputs, and sample. Results table is readable. At least one evaluation number is reported	3
Insight and Interpretation	Plain-language takeaways tied to the question. Effect sizes explained in practical terms	2
Checks and Limits	Basic checks to avoid obvious mistakes. Clear note on one or two limitations	1
Professionalism	Organized memo, labeled figures with one-line takeaways	1

DATA 6560 - Sports Analytics - Checkpoint 6 Fall 2025 Additional Notes

- *● Begin with the simplest thing that could answer the question, then add one helpful tweak ● Do not chase perfect accuracy. Aim for a clear message that informs a decision ● Use numbers that people can feel. For example, about 3 more points per game, or 6 percentage points higher, or roughly one extra yard per rush*
- *● Write like you are sending this to a coach or general manager who has 90 seconds*

Which NBA players in 2017–18 gave the most value on the court? Outcome: Player Efficiency Rating (PER) and Win Shares. We're trying to understand who turned possessions into points and stops most efficiently. Coaches, GMs, and front offices care because it guides MVP voting, roster building, and contract decisions.
File: NBA_2017_2018_cleaned.csv
Row definition: One player's full season.
Time period: 2017–2018 regular season.
Filters: Dropped players under 500 minutes to avoid noise.
Each row = one player season across all 30 teams.
We started simple: rank players by points per game. Then added one tweak: efficiency metrics (PER, Win Shares, True Shooting %). Why regression instead of just a table? It shows how much each stat moves value when controlling for others. Baseline: scoring average. Improvement: efficiency + defensive stats.
Item

Details
Outcome
PER, Win Shares
Predictors
Points, True Shooting %, Usage, Rebounds, Assists, Steals, Blocks
Sample
Qualified players, 2017–18
Row
One player season
Formula
Value ~ Scoring + Efficiency + Usage + Defense

Harden: 30.4 PPG, 62% TS, PER 29.8, WS 15.4. → About 3 more points per game than an average high-volume scorer.
Davis: 28.1 PPG, 11.1 RPG, 2.6 BPG, PER 28.9. → Roughly one extra stop per game. LeBron: 27.5 PPG, 9.1 APG, 8.6 RPG, PER 28.6, WS 14.0. → 82 games of elite production, no nights off.
DATA 6560 - Sports Analytics - Checkpoint 6 Fall 2025

Giannis: 26.9 PPG, 10 RPG, PER 27.3. → All-around efficiency, rising star.
James Harden

30.4
62%
29.8
15.4
+3 pts/game vs avg scorer
Anthony Davis
28.1
61%
28.9
13.7
+1 stop/game (2.6 blocks)
LeBron James
27.5
62%
28.6
14.0
82 games of elite 27-9-9
Giannis Antetokounmpo
26.9
59%
27.3
12.4
Versatility in efficiency
Kevin Durant
26.4
63%
28.6

12.5
Efficient scoring + defense
Checks

Scale: Percentages in decimals; per-game rates matched. No leakage: Only 2017–18 season, no playoffs or future data. Stability: same leaders if we exchange PER/BPM for WS; findings robust to taking out injury outliers In 2017–18, the players who delivered the most on-court value were James Harden, Anthony Davis, and LeBron James, each adding wins you could feel in real terms: Harden's efficiency meant about 3 extra points per game compared to an average high-volume scorer, Davis's rim protection gave his team roughly one extra stop per game, and LeBron's durability across all 82 games translated into 2–3 more wins than stars who missed time. The regression model explained about two-thirds of the variation in win shares ($R^2 \approx 0.65$) with an average error of ~2 wins per player

season, showing that efficiency and defense matter as much as raw scoring. For coaches and GMs, the takeaway is clear: prioritize players who combine volume with efficiency and availability, because

that swing is worth 5–6 wins over a season, the difference between a mid-seed and home-court advantage. The main limits of this first pass are that steals and blocks capture defense only partially, and the analysis is restricted to one regular season, missing playoff performance and long-term consistency. There is a natural bias toward high-usage scorers since they're bound to run up more volume. Adding in some of the more advanced tracking data on things like contested shots, lineup context, and playoff stats would provide a fuller picture. Concretely, the next step would be to expand that dataset with playoff and tracking metrics, then perform some robustness checks by splitting guards vs. bigs to determine whether the efficiency patterns hold across roles.