This document provides detailed documentation for all datasets used or created in this project. Each dataset listed below includes a description of its contents, its source, and how it was constructed. All datasets and code are in the GitHub repository.

NBA Salary & Contract Dataset (Combined)
File: full_nba_data.csv
Source: Spotrac (subscription service)
Construction Process:

- Spotrac does not provide a combined dataset, each season is released as a separate spreadsheet.
- Fourteen season level spreadsheets were exported from Spotrac.
- All files were merged using a Python script in joining_salary_data.ipynb.
- Columns such as player name, team, salary, contract years, and position were standardized.
- Duplicate rows from players traded mid-season were consolidated.
- The final dataset includes all NBA salaries and contract structures from 2012-2025.

Primary Uses in Study:

- Calculating salary distributions and inequality (percentiles, Lorenz curves, Gini coefficients).
- Regression model examining the relationship between performance and salary.

WNBA Salary & Contract Dataset (Combined)
File: full_wnba_data.csv
Sources: Spotrac (recent seasons) and Her Hoop Stats (manual collection for missing years)
Construction Process:

- Spotrac provides partial salary listings for WNBA players.
- Her Hoop Stats was used to supplement Spotrac.
- The 2024-2025 season required manual transcription due to no downloadable file.
- All sources were merged and cleaned through the Python notebook.

Primary Uses:

- Salary distribution analysis (percentiles, Lorenz curve, Gini coefficient).
- WNBA comparison to NBA salary structure.

NBA Historical Salary Dataset
File: NBA_Salaries(1990-2023).csv
Source: Spotrac
Notes: Used for context and validation only, not regression or inequality calculations.

2024-2025 League Variables Dataset
File: league_2024_2025_variables.csv
Sources: 2020 WNBA Collective Bargaining Agreement, NBA CBA summary sheets, and Reported cap summary documents for 2024-2025
Contents:

- Salary caps and floors
- Roster minimums and maximums
- Maximum contract percentages
- Veteran exception rules
- Tax thresholds (NBA)
- Hard cap limits (WNBA)

2024-2025 Derived Variable Dataset
File: league_2024_2025_derived_variable.csv
Notes:

- Contains variables that were calculated for analysis, such as:
- Salary cap proportion (cap/player salary)
- Relative earnings ratio
- Percentage difference between NBA and WNBA salary structures

WNBA Stats Dataset

File: WNBA_all.csv
Notes:

- Season level stat files were downloaded from WNBA Stats.
- Combined multiple seasons of WNBA player statistics

Python Processing Notebook
File: joining_salary_data.ipynb
Contents:

- Web scraping code for 2024 2025 NBA salary table (Basketball-Reference)
- Scripts to merge Spotrac spreadsheets

Cleaning steps:

- Name standardization
- Duplicate removal
- Salary formatting
- Column alignment
- Export commands to create clean CSV datasets.

Purpose: Ensures full reproducibility of all data transformations.

Quarto Project File
File: first_draft.qmd
Contents:

- All R code for statistical analysis
- Gini calculations and bootstrap confidence intervals
- Regression model code
- Scatterplot, Lorenz curves, percentile tables
- Full narrative text of the research paper

Quarto Project File
File: early_results.qmd
Contents:

- All R code for statistical analysis
- Gini calculations and bootstrap confidence intervals
- Lorenz curves, percentile tables

WNBA Regression
File: Model_WNBA_Regression.qmd
Description: Contains all code used to do a WNBA salary regression model.
Notes:

- Uses cleaned performance and contract data
- Creates PerfIndex
- Computes log salary (log_AAV)

Outputs:

- Regression summary table
- Scatterplot of salary vs. performance index

Cleaning WNBA Stats
File: WNBA_Stats_cleaning.qmd
Description: Code used to clean raw WNBA stat files
Notes:

- Import 2015-2025 season stat files
- Merge all seasons into WNBA_all.csv

All datasets above, along with the full R and Python code used to construct and analyze them, are available in the GitHub repository associated with this project. Any dataset referenced in this documentation can be located there, and any transformation described can be replicated by running the corresponding code files.