[DSCI 510] Final Project : 2024 WNBA Salary-Performance Analysis

Name	: Fatia Marwa Nastitie
USC ID	: 3143982880
Github	: https://www.github.com/fatiabob
Data Source	: player stats and contracts

1. Introduction

The WNBA 2023-2024 season has been taking its toll towards the sports industry this year drawing attention not just from fans but experts debating about team performance, statistical player abilities and their total values in terms of salary cap. The main objective of this project is to provide a thorough, data-driven analysis of WNBA player statistics. The goal is to offer valuable insights into the individual performances of players, their contributions relative to their salaries and the team's salary cap, and the broader trends observed across the league. By uncovering hidden patterns, correlations, and nuanced details in the salary and performance data, this project would like to answer the following questions:

- How is the relationship between player statistics metrics and their contracts?
- Which WNBA players are providing the most (or least) value relative to their current salaries?

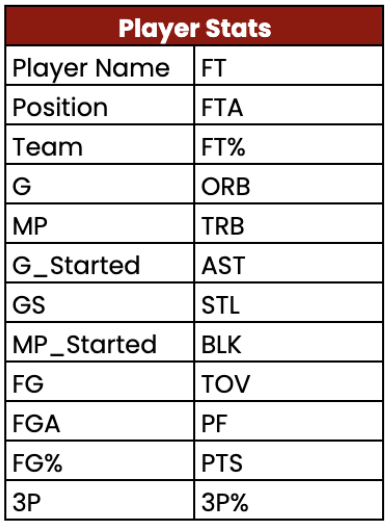
2. Data Collection

The data collected for this project comes from two primary sources: basketball-reference.com and spotrac.com. These two sources are widely regarded as the most comprehensive and up-to-date source for player statistics and player contract data in the WNBA. To gather the necessary data, I utilized web scraping techniques (Requests + BeautifulSoup4), as taught in weeks 9 and 10 of the course, to extract the relevant information from these websites.

In total, the dataset comprises 183 players and 34 initial columns, covering a wide range of statistical categories and contract-related information. The original plan was to use datasets from the WNBA and ESPN websites. However, during development, I discovered that player valuation data wasn't accessible through the requests/BeautifulSoup library. This led to using a dual approach: requests/BeautifulSoup for player statistics and Selenium for contract data, ensuring successful data collection from both sources

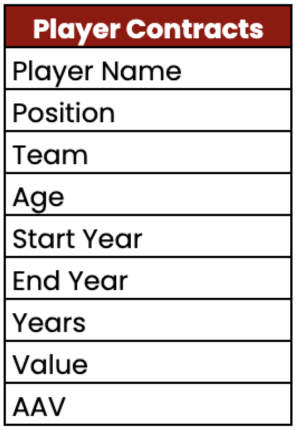
3. Data Cleaning

The initial data collected for this project was in the form of HTML code, as it was retrieved through web scraping technique. To prepare the data for analysis, the project parsed the HTML content to identify and extract the relevant information, such as player statistics and contract details. The extracted data was then converted into a CSV (Comma-Separated Values) format. The scope metrics of this project were player-level data of statistics performance and contracts as below:



Player Stats	
Player Name	FT
Position	FTA
Team	FT%
G	ORB
MP	TRB
G_Started	AST
GS	STL
MP_Started	BLK
FG	TOV
FGA	PF
FG%	PTS
3P	3P%

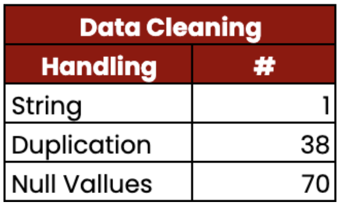
Image 1. Player Statistics Metrics



Player Contracts
Player Name
Position
Team
Age
Start Year
End Year
Years
Value
AAV

Image 2. Player Valuation Metrics

In the first step, we found that our data having some issues while we process it so we had to perform several cleaning data approach which were shown below:

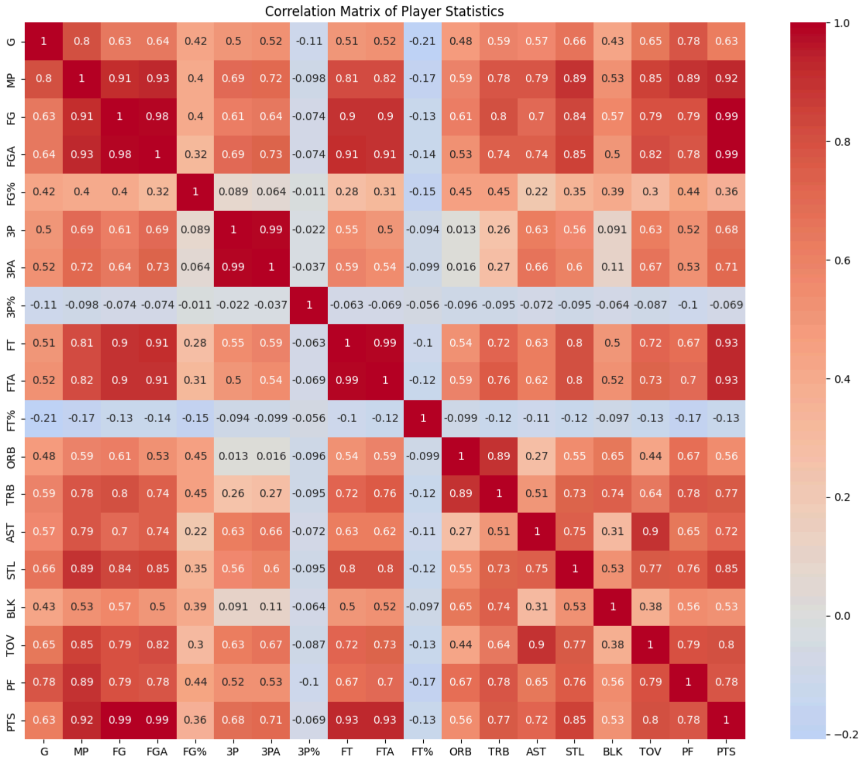


Data Cleaning	
Handling	#
String	1
Duplication	38
Null Vallues	70

String handling was performed on player_name columns so that both of the data sets had the same format of player name when we joined both of the dataset. In terms of duplication, we found that there are several players that have changed teams mid-season so duplication on stats was found and we handled it by only using the total number of stats of the player the whole season and using their latest team. Furthermore, ~30% of total players' salaries were unknown due to their free agent or missing half of the season, so we won't use their data for this analysis. This process resulted in our data to shape 160 rows and 33 columns that will be used on the data analysis part.

4. Data Analysis and Visualization

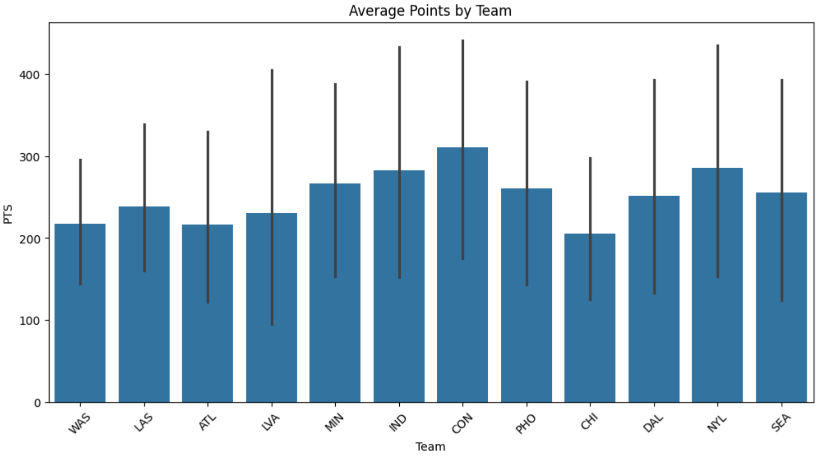
In this step, we would like to start with player stats correlation analysis using heat-map visualizations to know which are the metrics indicating a strong correlation between variables. We select on using numerical stats columns :



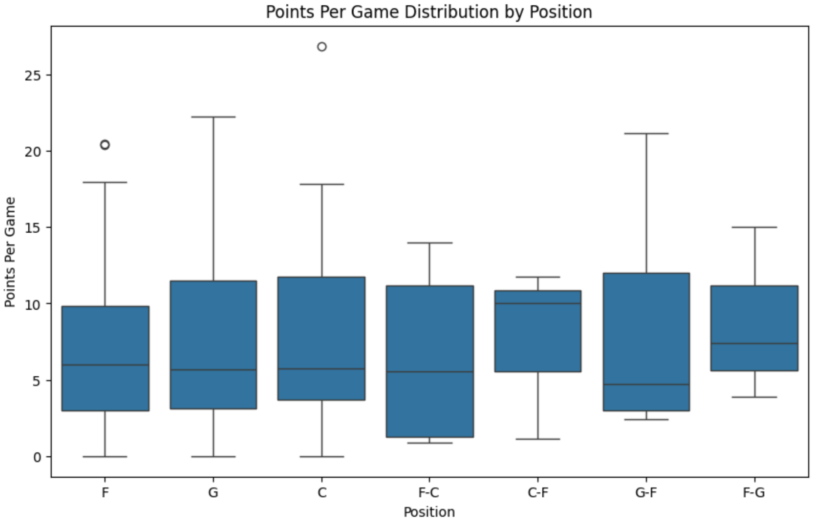
Based on the correlation matrix we noticed there are several insight we found:

- The data shows a positive correlation between Points (PTS) and metrics like Steals (STL), Turnovers (TOV), and Personal Fouls (PF). This suggests that high-scoring players tend to be more aggressive, leading to increased steals, turnovers, and fouls.
- Rebounds (REB) and blocks (BLK) are strongly correlated, as players who excel at rebounding are often effective shot-blockers as well.

Next, we want to analyze the average points scored per game by each team in the WNBA league:



Based on the bar plot above, it can be seen that The Indiana Fever (IND), Connecticut Sun (CON), and New York Liberty (NYL) have the highest average points per game, scoring around 300 points on average.

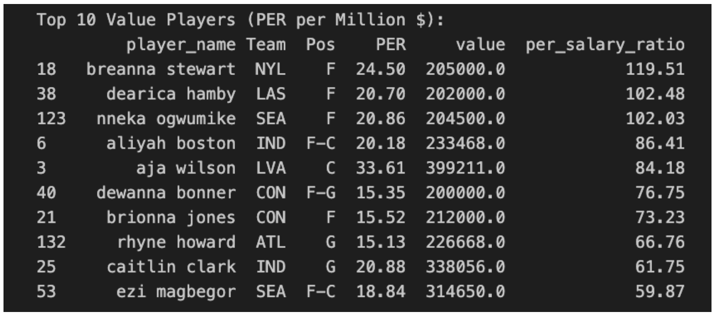


Based on the boxplot above, it can be concluded that Guards (G) have the highest median points per game, with a wide interquartile range, indicating a diverse scoring output among guards. The plot also reveals some outliers, represented by the circles, which indicate players with exceptionally high or low points per game compared to their position peers.

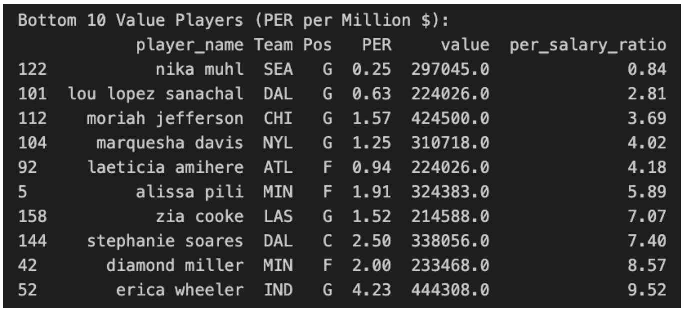
Next, we wanted to analyze the player efficiency rating by calculating the new metrics called "PER"' using this formula :

$$PER = (PTS + TRB + AST + STL + BLK - TOV - (FGA - FG) - (FTA - FT)) / G$$

This formula provides a comprehensive way to evaluate a player's overall statistical contributions, taking into account both positive and negative factors that contribute to a player's efficiency on the court.

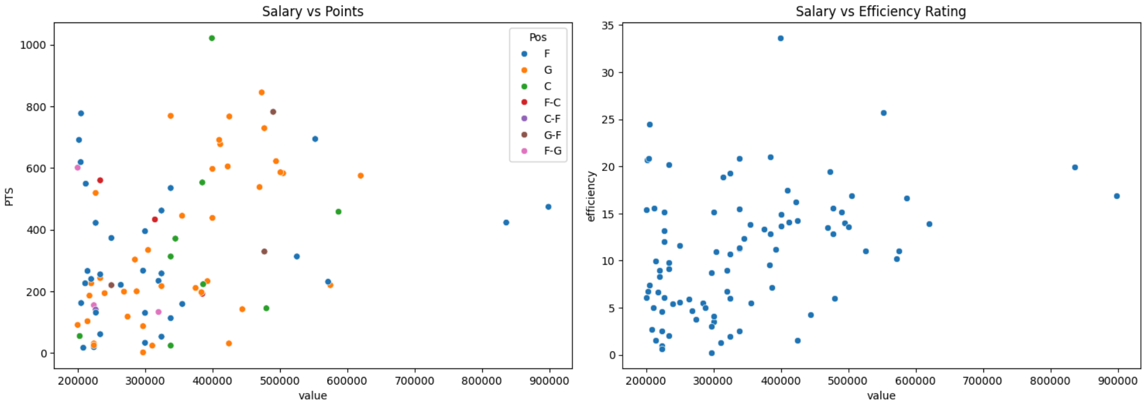


```
Top 10 Value Players (PER per Million $):
            player_name Team  Pos    PER     value  per_salary_ratio
18   breanna stewart  NYL    F  24.50  205000.0            119.51
38     dearica hamby  LAS    F  20.70  202000.0            102.48
123   nneka ogwumike  SEA    F  20.86  204500.0            102.03
6      aliyah boston  IND  F-C  20.18  233468.0             86.41
3         aja wilson  LVA    C  33.61  399211.0             84.18
40    dewanna bonner  CON  F-G  15.35  200000.0             76.75
21     brionna jones  CON    F  15.52  212000.0             73.23
132     rhyne howard  ATL    G  15.13  226668.0             66.76
25      caitlin clark  IND    G  20.88  338056.0             61.75
53      ezi magbegor  SEA  F-C  18.84  314650.0             59.87
```



```
Bottom 10 Value Players (PER per Million $):
            player_name Team Pos   PER     value  per_salary_ratio
122           nika muhl  SEA   G  0.25  297045.0              0.84
101  lou lopez sanachal  DAL   G  0.63  224026.0              2.81
112    moriah jefferson  CHI   G  1.57  424500.0              3.69
104     marquesha davis  NYL   G  1.25  310718.0              4.02
92     laeticia amihere  ATL   F  0.94  224026.0              4.18
5           alissa pili  MIN   F  1.91  324383.0              5.89
158           zia cooke  LAS   G  1.52  214588.0              7.07
144    stephanie soares  DAL   C  2.50  338056.0              7.40
42       diamond miller  MIN   F  2.00  233468.0              8.57
52        erica wheeler  IND   G  4.23  444308.0              9.52
```

The per salary ratio captures this efficiency more accurately than just looking at the raw salary amounts. It can be seen that Aja Wilson (LVA) PER of 33.61 and a per salary ratio of 84.18 provides the most value per million dollars of salary. She is greatly outperforming her contract and providing immense value this season. Aliyah Boston (IND), PER of 20.18 and a per salary ratio of 86.41, is another player who stands out for her high PER relative to her salary. On the other hand, players who are considered the "least valuable" on the list because their PER is extremely low compared to their salaries are Nika Muhl (SEA), Lou Lopez Sanacha (DAL), and Moriah Jefferson (CHI) providing less than a dollar and a half worth of value per million dollars earned.



The scatter plot showed a potential relationship between salary and efficiency rating showing a clear positive correlation, with higher-paid players generally having higher efficiency ratings. However, there is also a fair amount of variation, suggesting that factors beyond just salary contribute to a player's on-court efficiency. In terms of salary and points, a scatter plot shows players across various positions (Forwards, Guards, Centers) are distributed throughout the plot, indicating that high-scoring players are not limited to a specific position.

5. Conclusion

The analysis revealed several key insights into WNBA player salaries and performance metrics. There is a notable positive correlation between salary and player efficiency ratings, indicating that higher-paid players tend to perform better on the court. However, variations in the data highlight that salary alone does not fully account for efficiency, suggesting room for improvement in aligning player contracts with performance. Players like Aja Wilson and Aliyah Boston stood out for their high efficiency relative to their salaries, while others, such as Nika Muhl and Lou Lopez, provided significantly less value in comparison.

6. Future Work

With additional time, I would enhance the analysis by incorporating linear regression models to quantify the relationship between player salaries and their performance metrics turning descriptive insights into predictive and actionable outcomes. For next time, incorporating official WNBA or ESPN datasets could improve the analysis's reliability and accuracy.

7. Reference

- https://github.com/erilu/web-scraping-NBA-statistics
- https://github.com/Chisuso/NBA-PLAYERS-AND-TEAMS-?tab=readme-ov-file#use-case-3-relationship-between-games-started-and-assists
- https://towardsdatascience.com/how-to-use-selenium-to-web-scrape-with-example-80f9b23a843a
- https://www.scrapingbee.com/blog/selenium-python/
- https://github.com/logan-lauton/nba_webscrape/
- https://pypi.org/project/beautifulsoup4/
- https://www.spotrac.com/wnba/contracts
- https://www.basketball-reference.com/wnba/years/2024_advanced.html
- https://seaborn.pydata.org/generated/seaborn.heatmap.html
- https://www.basketball-reference.com/about/per.html