Data, Intelligence and Analytics Report

NBA Player Performance 2023 - 2024

Author:
Ainul Marzuki Febri

Abstract

An analysis of NBA player performance from November 2023 to January 2024, using monthly data of 5522 entries, revealed some important findings. Correlation results show that metrics such as 2-point shots (2P), 3-point shots (3P), free throws (FT), and field goals (FG) have a strong positive correlation with points per game (PTS), which is expected. Interestingly, turnovers (TOV) and personal fouls (PF) also show a moderately strong positive correlation with PTS, indicating that players with high points tend to play aggressively, thus are more likely to commit turnovers or fouls. Positional analysis revealed that players who play near the ring, such as power forwards (PF), small forwards (SF), and centers (C), do more blocking, rebounding, and stealing, while shooting guards (SG) and point guards (PG) focus more on scoring points from long range or providing assists. A comparison of team performance shows that Miami Heat (MIA) consistently increased the number of points per game, while other teams such as TOT showed a significant performance spike at the end of the analysis period. ANOVA tests showed no significant differences in PTS, assists (AST), total rebounds (TRB), and PF metrics on a monthly basis, except for TOV. The regression model identified that turnovers (TOV), playing time (MP), as well as contributions from the PF-SF and SG-PG positions, had a significant influence on the number of points scored, with the Indiana Pacers (IND) as the team with the highest points.

Background

This NBA player performance analysis was conducted to better understand the factors that influence player and team performance during the regular season period from November 2023 to January 2024. In a professional sports competition like the NBA, understanding performance metrics and game dynamics is critical for coaches, team managers and data analysts to develop effective strategies, identify strengths and weaknesses, and improve overall performance.

The data used in this analysis includes various player statistics, including points per game (PTS), assists (AST), blocks (BLK), steals (STL), rebounds (REB), 2-point shots (2P), 3-point shots (3P), free throws (FT), field goals (FG), turnovers (TOV), and personal fouls (PF). These metrics were chosen for their relevance in evaluating a player's efficiency and effectiveness on the court. For example, metrics such as 2P, 3P, FT, and FG give an idea of the ability to score points, while AST, BLK, STL, and REB provide insight into a player's contribution in the aspects of defense and ball distribution.

In addition, this analysis also considers the playing positions and specific contributions of various roles in the team, such as power forwards (PF), small forwards (SF), centers (C), shooting guards (SG), and point guards (PG). Understanding how each position contributes to the team's overall performance can help in determining a more optimal strategy and player rotation. For example, PF, SF, and C players tend to block and rebound more often, while SG and PG focus more on scoring points and providing assists.

This monthly performance analysis also aims to identify trends and changes in team performance throughout the season. By observing performance from November 2023 to January 2024, this analysis can provide insights into how teams and players adapt to changing competition conditions, as well as identify factors that contribute to improved or decreased performance.

Overall, this analysis aims to provide deep insights into the performance dynamics of NBA players and teams, and support better decision-making in game strategy and team management. This analysis is also expected to help teams optimize game strategies, improve the effectiveness of defense and attack, and maximize the potential of individual players to achieve the best performance.

Purpose of Analysis

1. Identifying Key Factors
2. Analyzing Correlation Between Variables
3. Evaluate Performance Based on Position
4. Comparing Team Performance
5. Assessing Monthly Performance Changes
6. Perform regression analysis

Data Processing

Columns	Description
Rk	Rank
Player	Player's name
Pos	Position
Age	Player's age
Tm	Team
G	Games played
GS	Games started
MP	Minutes played per game
FG	Field goals per game
FGA	Field goal attempts per game
FG%	Field goal percentage
3P	3-point field goals per game
3PA	3-point field goal attempts per game
3P%	3-point field goal percentage
2P	2-point field goals per game
2PA	2-point field goal attempts per game
2P%	2-point field goal percentage
eFG%	Effective field goal percentage
FT	Free throws per game
FTA	Free throw attempts per game
FT%	Free throw percentage
ORB	Offensive rebounds per game
DRB	Defensive rebounds per game
TRB	Total rebounds per game
AST	Assists per game
STL	Steals per game
BLK	Blocks per game
TOV	Turnovers per game
PF	Personal fouls per game
PTS	Points per game

- **Importing and combine monthly dataset**

The basketball match data for 3 months, each consisting of 11 matches, will be imported one by one. Prior to the import process, each match data will be expanded with the addition of a new column that includes the match date. This step allows for a more in-depth analysis based on time, making it possible to identify trends and patterns that may have developed over that time period. Once the data import is complete and the date column is added, further analysis can be performed to understand the dynamics and development of team and player performance over the specified time period.

- **Dealing with Missing Values**

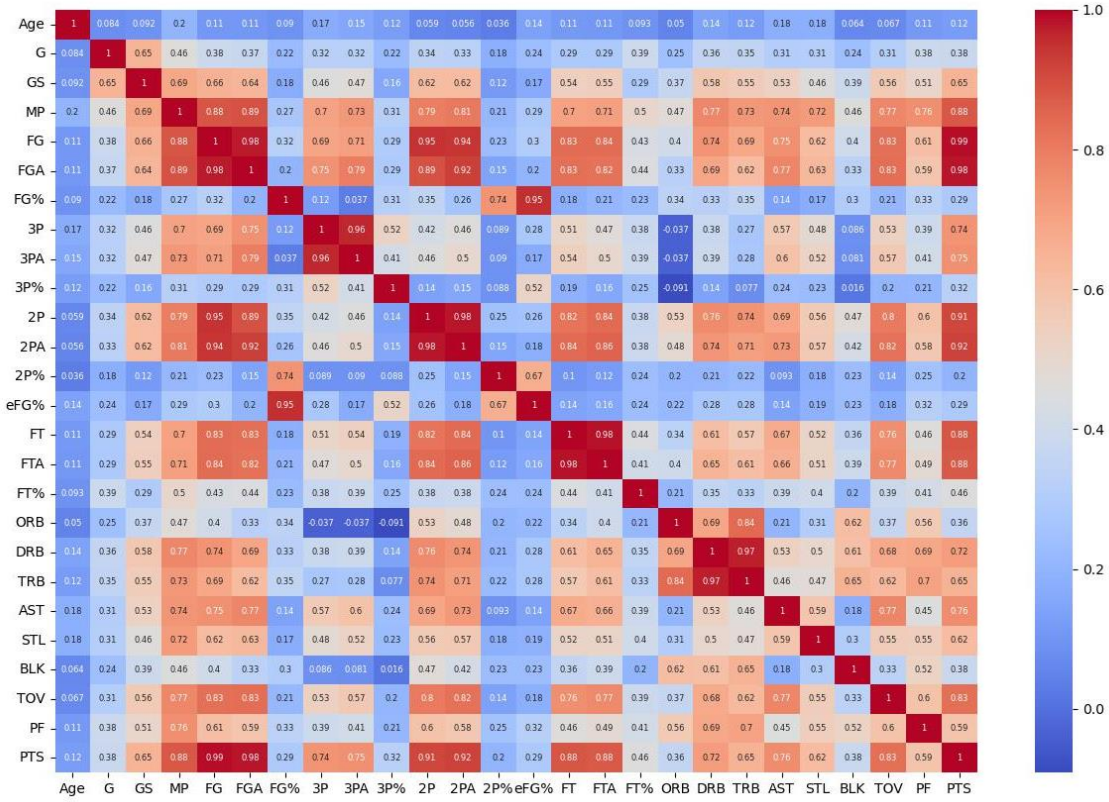
After identification, it was found that there were missing values in several key columns, including FG%, 2P%, 3P%, FT%, and eFG%. The missing values are caused by dividing the printed scores by the scoring attempts. However, overall, the scoring attempts have a value of 0, so dividing

any number by 0 results in a NaN (Not a Number) value. To handle this issue, the step taken was to replace the NaN value with 0. This ensured consistency in data analysis and prevented distortion of the results caused by missing values in the key variable. By replacing the right values, further analysis can be continued to gain accurate and reliable insights from the available data.

Exploratory Data Analysis (EDA)

- **Correlation**

Correlation analysis is a statistical method used to measure and evaluate the relationship between two or more variables. The goal is to determine if there is a significant relationship between the variables and how strong the relationship is. Correlations can range from -1 to 1, where a positive value indicates a direct relationship, a negative value indicates an inverse relationship, and a value of zero indicates no relationship. This analysis is useful for identifying patterns and directing decision-making based on the relationships found in the data.



- **Age vs. Performance Metrics**: There are weak to moderate positive correlations between age and various performance metrics such as games played (G), games started (GS), minutes played (MP), field goals made (FG), free throws made (FT), and points scored (PTS). This suggests that older players tend to have more experience and contribute more to their teams in terms of playing time and scoring.

- **Playing Time Metrics:** Games played (G), games started (GS), and minutes played (MP) show strong positive correlations with each other. The analysis also showed a positive correlation between G, GS, and MP with points per game (PTS). This means that players who play more often and get more time on the court also tend to score more points.

- **Shooting Metrics**: Field goals made (FG), field goal attempts (FGA), and points scored (PTS) exhibit strong positive correlations with each other, as expected. Similarly, there are strong positive correlations between free throws made (FT), free throw attempts (FTA), and points scored (PTS).

- **Efficiency Metrics**: Effective field goal percentage (eFG%) shows moderate to strong positive correlations with field goal percentage (FG%) and points scored (PTS). This suggests that players with higher shooting efficiency tend to score more points.

- **Rebounding and Defense**: There are moderate positive correlations between offensive rebounds (ORB), defensive rebounds (DRB), total rebounds (TRB), steals (STL), and blocks (BLK). This indicates that players who excel in one aspect of rebounding or defense often perform well in other related aspects.

- **Turnovers and Fouls**: Turnovers (TOV) and personal fouls (PF) show moderate positive correlations with each other and with various other metrics such as field goal attempts (FGA), points scored (PTS), and assists (AST). This suggests that players who are more involved in offensive plays and scoring also tend to commit more turnovers and fouls.

- **Top Player**

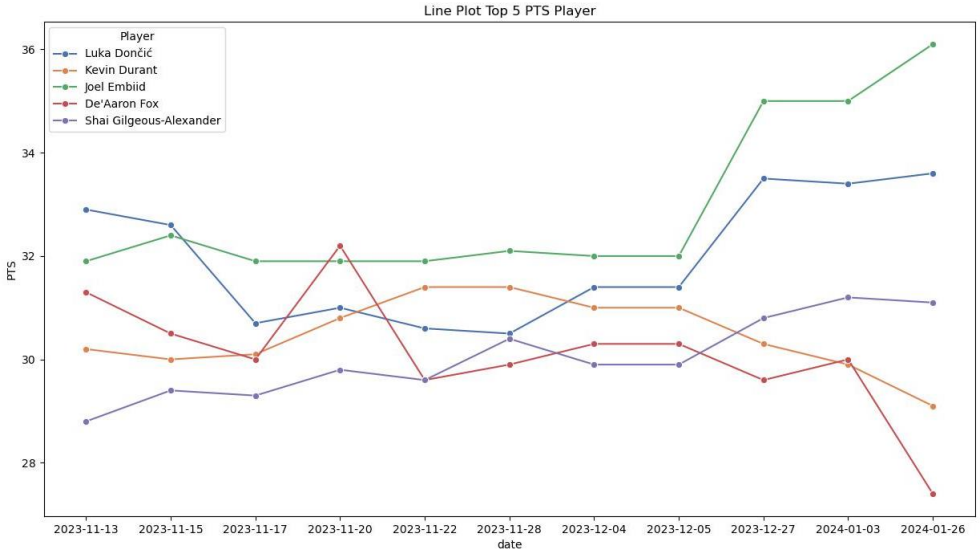
In analyzing the performance of players who performed above average on various metrics such as points per game (PTS), assists (AST), 2-point field goals (2P), 3-point field goals (3P), and effective field goal percentage (eFG%), it was found that these players not only excelled in scoring points, but also showed significant contributions in ball distribution and shooting efficiency. Players with high performance on PTS tend to be the main scorers that the team can rely on in various game situations. Those who excel in AST show the ability to not only score points themselves, but also create opportunities for their teammates through effective assists. Meanwhile, players with high 2P and 3P percentages show expertise in shooting from both inside and outside the three-point line, increasing the team's offensive threat from various distances. Lastly, players with a high eFG% show great efficiency in scoring points, maximizing scoring opportunities from every shot they take.

- **Based on average point per game**

Player	PTS	Pos	Tm	G	GS	TOV	PF	TRB
Joel Embiid	32.927273	C	PHI	16.545455	16.545455	3.727273	2.772727	11.445455
Luka Dončić	31.963636	PG	DAL	18.727273	18.727273	4.072727	1.800000	8.263636
Kevin Durant	30.472727	PF	PHO	17.818182	17.818182	3.736364	1.763636	6.745455
De'Aaron Fox	30.100000	PG	SAC	13.181818	13.181818	2.272727	3.009091	4.354545
Shai Gilgeous-Alexander	30.018182	PG	OKC	18.363636	18.363636	2.236364	2.309091	6.190909

  - From the analysis, it was found that the 5 players with the highest average points had above-average playing time, along with a high number of turnovers (TOV), personal fouls (PF), and total rebounds (TRB). This finding is interesting as it suggests a link between players' productivity in scoring points and their overall level of engagement in the game. While these players scored at a high rate, they were also involved in other aspects of the game, such as committing turnovers, earning personal fouls, and being active in grabbing rebounds. This highlights the importance of a balance between offensive productivity and involvement in defense and overall team play.
  - Although they are categorized as players with the highest average game points, the stability of points performance is not always consistent. For example, De'Aaron Fox had a notable drop in performance in recent matches, showing significant fluctuations in performance. On the other hand, Joel Embiid showed the opposite by experiencing a significant increase in points performance from the last 4 matches. This finding underscores the importance of looking at not only the average performance of players, but also the fluctuations or trends in their performance over time. This allows

teams to be more responsive in game strategy and player rotation management, as well as.



-
- **Based on Average assist per game**

Player	AST	Pos	Tm	G	GS	TOV	PF	TRB	STL	BLK
Tyrese Haliburton	12.145455	PG	IND	16.545455	16.545455	2.436364	1.136364	3.990909	1.027273	0.663636
Trae Young	10.781818	PG	ATL	17.909091	17.909091	4.090909	1.590909	2.872727	1.518182	0.054545
Devin Booker	9.018182	SG	PHO	11.454545	11.454545	3.590909	3.336364	5.936364	0.581818	0.272727
Nikola Jokić	8.954545	C	DEN	20.181818	20.181818	3.109091	2.418182	13.027273	1.090909	0.809091
Fred VanVleet	8.581818	PG	HOU	17.363636	17.363636	1.645455	1.945455	3.854545	0.763636	0.436364

- The data shows the significant contributions of five NBA players in different aspects of the game. Tyrese Haliburton of the Indiana Pacers stood out with the highest assists and relatively low turnovers, showing playmaking efficiency. Trae Young of the Atlanta Hawks also recorded high assists but with more turnovers, showing greater aggressiveness in the game. Devin Booker of the Phoenix Suns not only contributed in assists but also showed better rebounding ability compared to other guards. Nikola Jokić of the Denver Nuggets, a center, showed his versatility with outstanding rebounding and assist statistics. Fred VanVleet of the Houston Rockets showed efficiency in assists with a low number of turnovers. In general, assists are generally filled by players who are positioned away from the ring such as point guard (PG) and shooting guard (SG). This data emphasizes the importance of the guard's role in playmaking and ball distribution.

- **Based on Average effective field goal percentage**

Top average eFG%

Player	FG	3P	FGA	eFG%
Chris Livingston	0.881818	0.4	0.945455	1.127273
Drew Peterson	1.000000	1.0	1.000000	1.500000
Terquavion Smith	0.700000	0.7	1.000000	1.000000
Terry Taylor	0.363636	0.0	0.427273	0.952727
Udoka Azubuike	0.818182	0.0	0.890909	0.959455

Analysis of the data shows that Chris Livingston, Drew Peterson, Terquavion Smith, Terry Taylor, and Udoka Azubuike have high shooting effectiveness based on field effectiveness (eFG%). Nonetheless, it is important to note that the number of shot attempts of each player varies. This suggests that while their eFG% is high, the relatively small number of attempts might affect the conclusion about their prominence as players with the highest eFG% in general.

Top Average PTS

Player	FG	3P	FGA	eFG%
De'Aaron Fox	10.781818	3.145455	22.163636	0.556909
Joel Embiid	10.972727	1.145455	21.390909	0.539182
Kevin Durant	10.463636	2.100000	20.172727	0.570909
Luka Dončić	10.918182	4.036364	21.981818	0.590000
Shai Gilgeous-Alexander	11.090909	1.218182	20.745455	0.563455

From the analysis, it can be concluded that players like Chris Livingston, Drew Peterson, Terquavion Smith, Terry Taylor, and Udoka Azubuike show high shooting effectiveness based on field effectiveness (eFG%). Nonetheless, a comparison with the top players in average points shows a significant difference in point contribution. De'Aaron Fox, Joel Embiid, Kevin Durant, Luka Dončić, and Shai Gilgeous-Alexander have significantly higher point averages, despite their lower eFG%. Therefore, although shot efficiency is important, the number of points generated from those shots is also a very important factor in assessing a player's overall performance. This shows that the evaluation of a player's performance does not rely solely on shot efficiency, but also considers the contribution of the total points generated by the player.

- **Based on 2P and 3P percentage**

Top 2P% Player with Minimal 10 Attempt, 2PA

Player	2P%	Pos	Tm
Nikola Jokić	0.647636	C	DEN
Giannis Antetokounmpo	0.641091	PF	MIL
LeBron James	0.639909	PF	LAL
Domantas Sabonis	0.617182	C	SAC
Alperen Şengün	0.612818	C	HOU
Tobias Harris	0.607600	PF	PHI
Jayson Tatum	0.600727	PF	BOS
Marvin Bagley III	0.592000	C	DET
Luka Dončić	0.580727	PG	DAL
Shai Gilgeous-Alexander	0.576818	PG	OKC

- From the data, it can be seen that players with the highest two-point shooting percentage (2P%), such as Nikola Jokić, Giannis Antetokounmpo, and LeBron James, have an average 2P% above 60%, showing consistency and accuracy in completing two-point shots from various distances. This illustrates their effectiveness in scoring points through two-point shots. More interestingly, all the players with the highest 2P% are either forwards (PF) or centers (C), including Domantas Sabonis, Alperen Şengün, and Tobias Harris. This suggests that players in the forward or center position tend to have an advantage in finishing two-point shots, possibly because they often operate in areas closer to the basket and have greater physical strength to finish shots in traffic.

Top 3P% Player with Minimal 8 Attempt, 3PA

Player	3P%	Pos	Tm
Stephen Curry	0.433818	PG	GSW
Tyrese Haliburton	0.427833	PG	IND
Tyler Herro	0.413000	SG	MIA
Lauri Markkanen	0.404400	PF	UTA
Luka Dončić	0.400364	PG	DAL
Trey Murphy III	0.400000	SF	NOP
Anfernee Simons	0.396000	SG	POR
Tyrese Maxey	0.396000	PG	PHI
Paul George	0.394100	PF	LAC
Klay Thompson	0.383000	SF	GSW

- From the data, it can be seen that some NBA basketball players have a fairly high three-point shooting percentage (3P%). Stephen Curry stands out as the player with the highest 3P%, followed by Tyrese Haliburton, Tyler Herro, and Lauri Markkanen. This shows their accuracy and consistency in completing three-point shots from various positions on the court. Interestingly, some players with high 3P%, such as Luka Dončić and Paul George, play in different positions such as Point Guard (PG) and Power Forward (PF), showing their versatility and ability to score points from different positions on the court.

- **Position**

In basketball, player positions are typically divided into five main roles: Center (C), Shooting Guard (SG), Power Forward (PF), Point Guard (PG), and Small Forward (SF). Here is an explanation of each position, including combination positions:

1. **Center (C)**:
    - Position: Usually the tallest player on the team.
    - Role: Responsible for rebounds, blocking shots, and playing near the basket on both offense and defense.
2. **Shooting Guard (SG)**:
    - Position: Typically a skilled long-range shooter.
    - Role: Scoring points through outside shooting, dribbling, and assisting on defense.
3. **Power Forward (PF)**:
    - Position: Generally strong and plays near the basket, but can also shoot mid-range shots.
    - Role: Rebounding, physical play inside, and scoring points from close to the basket.

4. **Point Guard (PG)**:

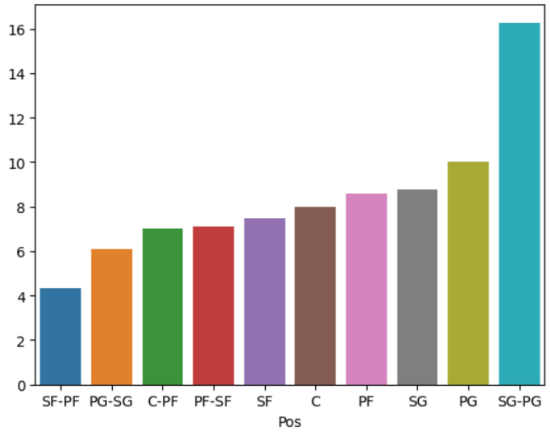
    - Position: Usually the team's playmaker.
    - Role: Dribbling the ball, organizing the offense, providing assists, and often scoring points.

5. **Small Forward (SF)**:

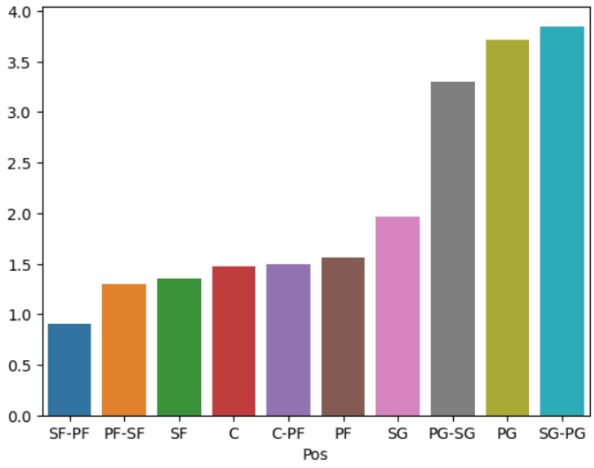
    - Position: A versatile player, effective in both offense and defense.
    - Role: Scoring points, defending, and rebounding.

Analysis

- **POS vs PTS**
    - In summary, the data shows that pure **Point Guards and hybrid Shooting Guard-Point Guards** have the highest average points per game, indicating their crucial roles in offensive plays. On the other hand, versatile players who switch between **forward positions** tend to have lower scoring averages, likely due to their broader responsibilities on the court.

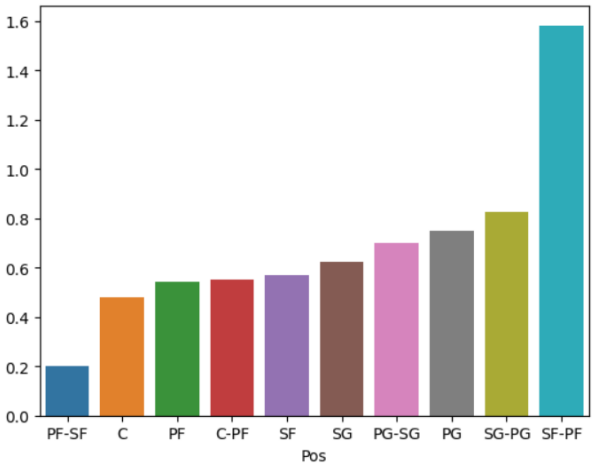


- **POS vs AST**

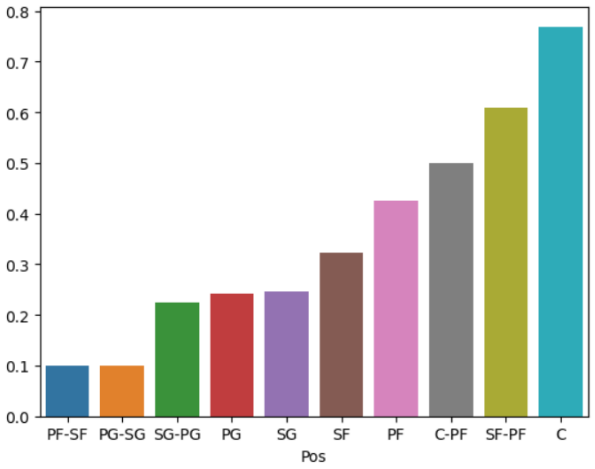


    - **Highest Assists**: Point guards (PG) and hybrid players like SG-PG and PG-SG have the highest average assists, highlighting their central role in ball distribution and offensive facilitation.
    - **Moderate Assists**: Positions like shooting guard (SG) and power forward (PF) contribute moderately to assists, balancing scoring and playmaking responsibilities.
    - **Lowest Assists**: Positions like center (C) and small forward-power forward (SF-PF) have the lowest average assists, focusing more on scoring, defense, and other responsibilities rather than facilitating the offense.

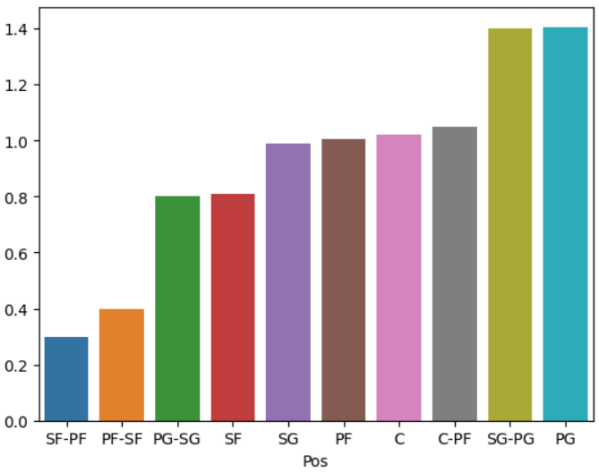
- **POS vs STL**
  - **Highest Steals**: Players in hybrid forward positions (SF-PF) and guard positions (SG-PG, PG) tend to have the highest average steals. These players are often involved in active perimeter defense and intercepting passes.
  - **Moderate Steals**: Small forwards (SF) and shooting guards (SG) have a moderate number of steals, balancing their defensive roles between the perimeter and interior.
  - **Lowest Steals**: Centers (C) and power forwards-small forwards (PF-SF) have the lowest average steals, as they focus more on interior defense and rebounding than on playing the passing lanes.



- **POS vs BLK**
  - **Highest Blocks**: Centers (C) have the highest average blocks, indicating their primary role in defending the frontcourt and blocking opponent shots.
  - **Lowest Blocks**: Point guards (PG) and combinations like PG-SG have relatively low average blocks, as their primary focus is on organizing offense and playing in the backcourt.

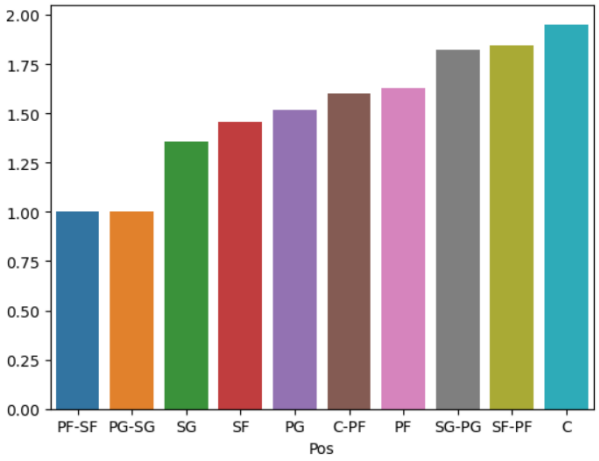


- **POS vs TOV**
  - **Highest Turnovers**: Point guards (PG) and combinations like SG-PG have the highest average turnovers, reflecting their primary role in ball-handling and initiating offensive plays.
  - **Lowest Turnovers**: Positions like PF-SF and SF-PF have the lowest average turnovers, likely due to

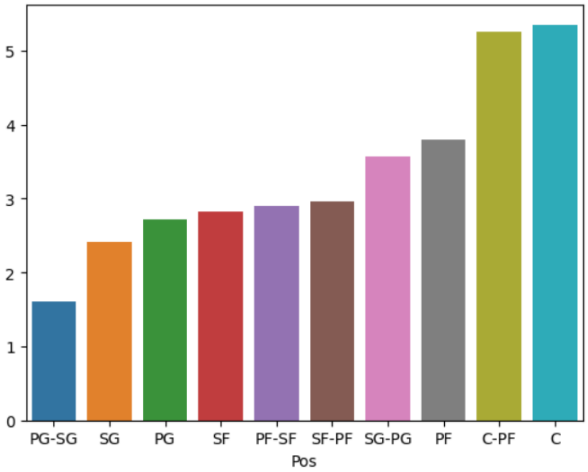


their focus on scoring and less involvement in ball-handling duties compared to guards.

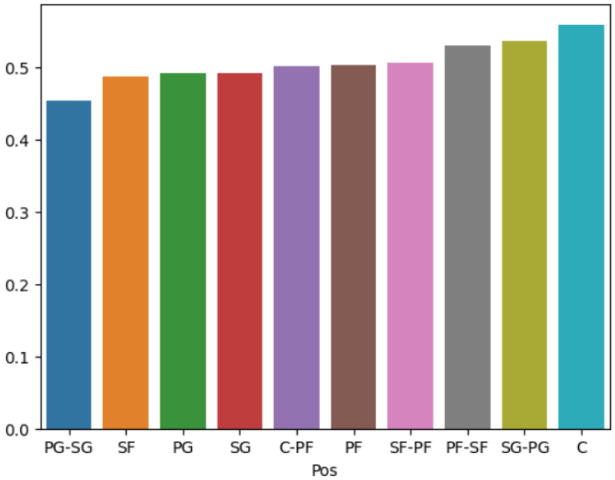
- **POS vs PF**
  - From this data, it can be seen that players in frontcourt positions **(C, PF, SF)** tend to have higher personal foul rates compared to players in backcourt positions **(PG, SG)**, due to their involvement in physical play near the basket. Players with positional flexibility **(e.g. C-PF, SF-PF)** have variations in personal foul rates, depending on their playing style and role in the team.



- **POS vs TRB**
  - From the data, we can see that a player's position has an effect on their ability to collect rebounds in a game. Positions closer to the basket **(such as C and PF)** tend to have higher average TRBs, while positions further away from the basket **(such as PG and SG)** tend to have lower average TRBs. This shows the importance of a player's role in the team and how their position affects their contribution in certain aspects of the game, such as collecting rebounds.

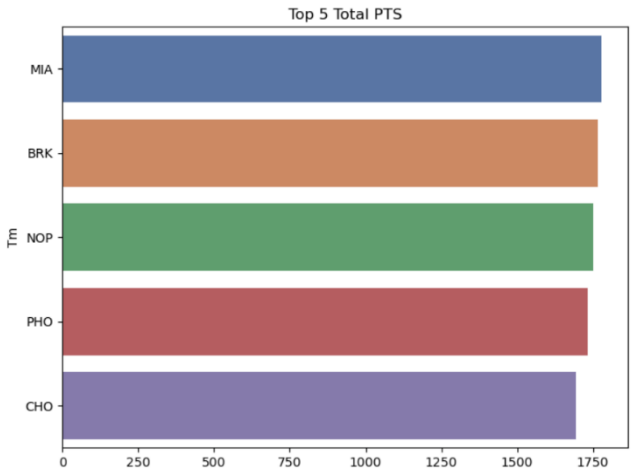


- **POS vs eFG%**

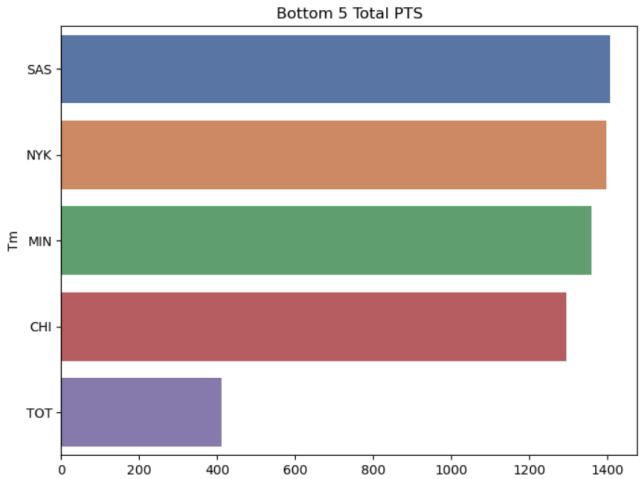


  - Players at the **Center (C)** position have **the highest average eFG%**, indicating that they are most efficient at scoring points, likely because they often shoot from close range.
  - Players in combined positions such as **PF-SF and SG-PG** also showed **high efficiency**. **Point Guard (PG) and PG-SG** combined positions have a lower eFG%, suggesting that players in these positions may take more shots from longer distances or in more challenging situations.

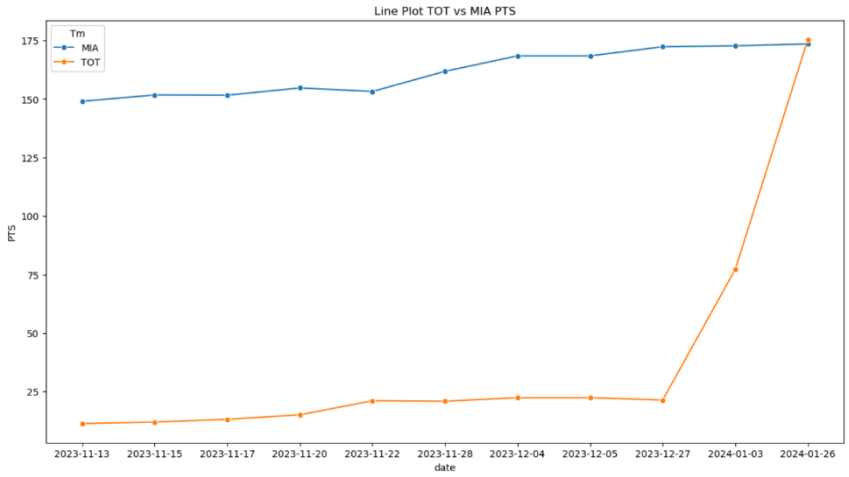
- **Team Comparison**
  - **Top team PTS**



  - **Bottom team PTS**



    - Miami Heat (MIA) was the team with the highest total score during the match, reaching 1777 points. This puts MIA at the top of the scoring table, with a not-so-significant difference compared to the other high-performing teams. However, there is a stark contrast when comparing MIA to the TOT team, who only managed to accumulate 412 points over the same period. TOT's point total is about four times less than that of MIA, showing a significant difference in offensive productivity between the two teams.
  - **Comparison between TOT (Lowest Total PTS) vs MIA (Highest Total PTS)**



- Although the TOT team overall had a much lower total score compared to the Miami Heat (MIA), there was a significant improvement in performance in the last three games of the TOT team. This improvement is so prominent that the number of points per game of the TOT team in the last three matches is able to rival the MIA team, which has the highest total PTS throughout the season. This phenomenon suggests that the TOT team has found a more effective strategy or rhythm of play in the later stages of the analysis period, allowing them to compete more closely with the top teams.

MIA vs TOT

Tm	MIA	TOT
Age	26.893617	29.600000
G	11.382979	16.087500
GS	5.877660	5.475000
MP	20.976596	15.728750
FG	3.402128	1.896250
FGA	7.517553	4.117500
FG%	0.433319	0.433563
3P	1.110106	0.661250
3PA	2.959043	1.825000
3P%	0.354090	0.292725
2P	2.297340	1.230000
2PA	4.560638	2.287500
2P%	0.502915	0.533563
eFG%	0.511670	0.518025
FT	1.543617	0.708750
FTA	1.838830	0.962500
FT%	0.698080	0.642850
ORB	0.853191	0.790000
DRB	2.989894	1.882500
TRB	3.836702	2.667500
AST	2.247872	1.111250
STL	0.756383	0.637500
BLK	0.288830	0.391250
TOV	1.125000	0.521250
PF	1.580319	1.627500
PTS	9.453723	5.156250

MIA

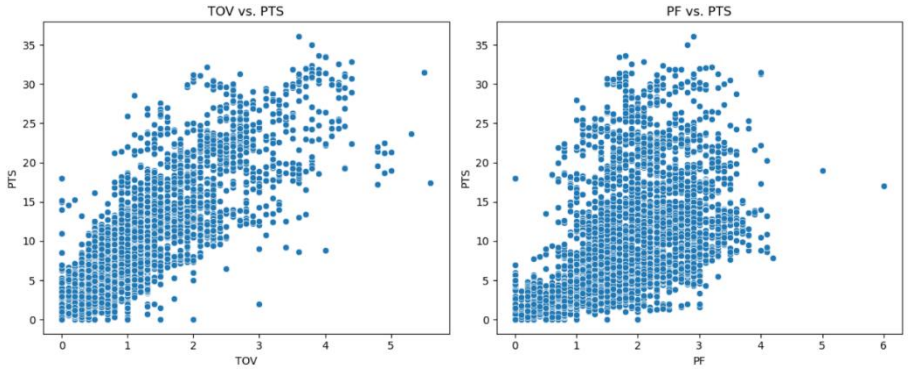
- **Age and Experience**: MIA players have an average age of 26.89 years, younger compared to the TOT team. This may indicate more energy and long-term development potential.
- **Playing Time and Efficiency**: MIA players play more minutes per game (20.98 minutes) and produce more points per game (9.45 PTS) compared to TOT players. Additionally, MIA had a slightly better shooting percentage (eFG% 51.17% vs. TOT 51.80%).
- **Shooting Ability**: On average, MIA players took more shot attempts, both 2P (4.56) and 3P (2.96), and scored more points from those attempts.
- **Rebounds and Assists**: MIA also excelled in rebounding (3.84 TRB per game) and assists (2.25 AST per game) statistics, showing a greater contribution in team play and ball possession.
- **Turnovers and Fouls**: Despite being more productive, MIA players tend to commit more turnovers (1.13) and personal fouls (1.58) than TOT, which could be an area for improvement.

TOT

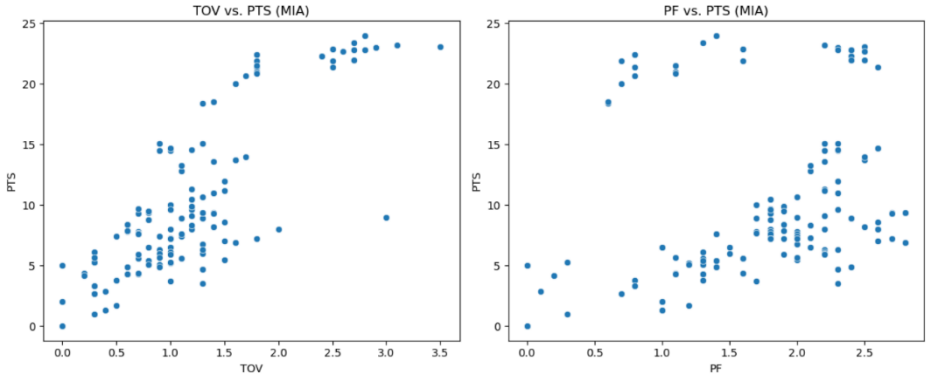
- **Age and Experience**: The TOT team has players with an average age of 29.60 years, older than MIA. This may indicate more experience but could also mean potential physical decline.
- **Playing Time and Efficiency**: TOT players play fewer minutes per game (15.73 minutes) and score fewer points per game (5.16 PTS). Despite less playing time, TOT has a shooting efficiency that is competitive with MIA.
- **Shooting Ability**: TOT had a slightly higher 2P shooting percentage (53.36%) but they made fewer shot attempts overall.
- **Rebounds and Assists**: TOT had lower rebounding and assist numbers (2.67 TRB and 1.11 AST per game), which indicates a smaller contribution in these aspects.
- **Turnovers and Fouls**: TOT players committed fewer turnovers (0.52) and fouls (1.63) compared to MIA, indicating a neater and less aggressive game.

- **Interesting Correlation Between Turnovers and Fouls with Point per game**
  - From the analysis of the data obtained, we can draw some interesting conclusions regarding the relationship between turnovers (TOV), personal fouls (PF), and points per game (PTS) in basketball games, specifically for certain players and teams.
  - Miami (MIA) players seem to have high turnover (TOV) and personal foul (PF) rates. However, despite this, the team is still one of the teams with the highest point totals. This shows that high TOV and PF do not always have a negative impact on the total points produced by the team.

Scatter plot TOV and PF with PTS (Dataset)



MIA Team



Interesting Correlation:

Most likely, the positive correlation between TOV and PF with PTS indicates that teams that generate more turnovers and personal fouls also tend to score more points. This may be because the team plays an aggressive style of play, creating more attacking opportunities and exploiting the opponent's mistakes to score points.

However, TOV (Turnover) and PF (Personal Foul) can be detrimental factors for their own team. Although the positive correlation between TOV, PF, and PTS indicates a statistical relationship among the variables, it is important to remember that this relationship does not necessarily signify an advantage for the team.

- **Loss of Possession**: Turnovers can lead to loss of possession, giving the opponent the opportunity to score points through quick counter-attacks or attacking opportunities resulting from the loss of the ball.
- **Gives Opponent Advantage**: Excessive PF can give the opposing team an advantage by providing easy free throw opportunities. This can increase the opponent's point total and hurt the team that committed the foul. Personal fouls can also cause key players in the team to get into foul

trouble and be forced out of the game, reducing the strength and depth of the team.

Statistical Analysis

Anova test, or analysis of variance, is a statistical technique used to compare means between three or more groups. In the context of analyzing basketball player performance on a monthly basis, the Anova test can be used to assess whether there are significant differences in player performance, such as points per game (PTS), assists (AST), rebounds (TRB), turnovers (TOV), and personal fouls (PF), between specific months. By applying the Anova test to players' performance data from month to month, we can determine if there are significant differences in players' performance over time, which can provide valuable insights into the factors that affect their performance during a particular season or period.

```
Performance Metrics: PTS
F value: 0.090345554537282424
P-value: 0.9136167862972281
There is no significant difference in PTS between months.

Performance Metrics: AST
F value: 0.0004731220415519137
P-value: 0.9995269889903534
There is no significant difference in AST between months.

Performance Metrics: TRB
F value: 0.6943142803230007
P-value: 0.49946040804588765
There is no significant difference in TRB between months.

Performance Metrics: TOV
F value: 3.1022358637154617
P-value: 0.04502698062745895
There is a significant difference in TOV between months.

Performance Metrics: PF
F value: 0.6494016808451478
P-value: 0.5223981303559532
There is no significant difference in PF between months.
```

- Anova test
  - **Summary Conclusion:**
    - From the ANOVA test results of the tested player performance metrics (PTS, AST, TRB, TOV, PF), it was found that only turnovers per game (TOV) showed significant differences between the months analyzed. This indicates that player performance in terms of turnovers varies significantly each month, while the other performance metrics, namely points per game (PTS), assists (AST), rebounds (TRB), and personal fouls (PF), tend to be consistent throughout the months. In other words, players show significant fluctuations in the number of turnovers from month to month, which may reflect changes in game strategy, physical condition, or other factors that affect ball possession. In contrast, the consistency in the PTS, AST, TRB, and PF metrics suggests that these aspects of performance are more stable, and not significantly affected by time factors or monthly changes.

Categorical Variables and Dummy Variables

Before conducting regression analysis, it is important to convert all categorical variables that will be used as predictors into dummy variables. Dummy variables are binary variables that indicate the presence or absence of a condition or category. This process is necessary because some regression algorithms require the predictor variables to be in binary form to produce accurate results. By converting the predictor variables into dummy variables, we can describe certain characteristics or categories in our data more effectively, which in turn allows the regression analysis to provide a deeper understanding of the relationship between the observed variables.

- Position
  Position has 10 unique values ['C', 'SG', 'PF', 'PG', 'SF', 'SF-PF', 'C-PF', 'SG-PG', 'PF-SF', 'PG-SG'].
- Team
  Team has 31 unique values ['TOR', 'MIA', 'UTA', 'MEM', 'MIN', 'PHO', 'CLE', 'MIL', 'ORL', 'NYK', 'WAS', 'POR', 'DET', 'CHO', 'PHI', 'BOS', 'SAS', 'SAC', 'TOT', 'LAC', 'OKC', 'ATL', 'CHI', 'DEN', 'BRK', 'HOU', 'IND', 'LAL', 'DAL', 'GSW', 'NOP'].

If we have a categorical variable with k categories, the dummy coding approach will result in k binary variables, where each variable represents one category. However, if we include all k binary variables into our model, we will fall into the "dummy variable trap".

Dummy variable trap occurs when we have collinearity between dummy variables. In other words, one dummy variable can be predicted from another. This can cause problems in some analysis methods, especially in regression analysis, where it can interfere with the interpretation of the coefficients.

So, to avoid the dummy variable trap, we need to drop one of the dummy variables. Drop first is one way to avoid this problem by removing the first one dummy variable from each dummy set. This makes each dummy variable a reference variable relative to the category that does not belong to the deleted dummy variable. Thus, we can avoid the multicollinearity problem in the regression model.

First Regression Analysis

In the context of data analysis, we want to understand what factors affect the number of points scored by basketball players ('PTS'). Therefore, we collect data on player statistics, including variables such as age, team, performance metrics like assist, block, steal, rebound, shooting attempt etc. These variables are considered as predictors or independent variables (X) that might affect the number of points scored ('PTS'), which is the dependent variable or what we want to predict (Y).

Using linear regression or other analytical methods, we can explore the relationship between predictor variables (X) with the dependent variable (Y) 'PTS'. This allows us to identify how strong or weak the influence of each predictor variable is on the number of points scored by basketball players, and thus, provides valuable insights in understanding the factors that contribute to player performance.

- **Split the X and y variable**

  X = All variable expect ‘PTS’
  Y= Variable ‘PTS’

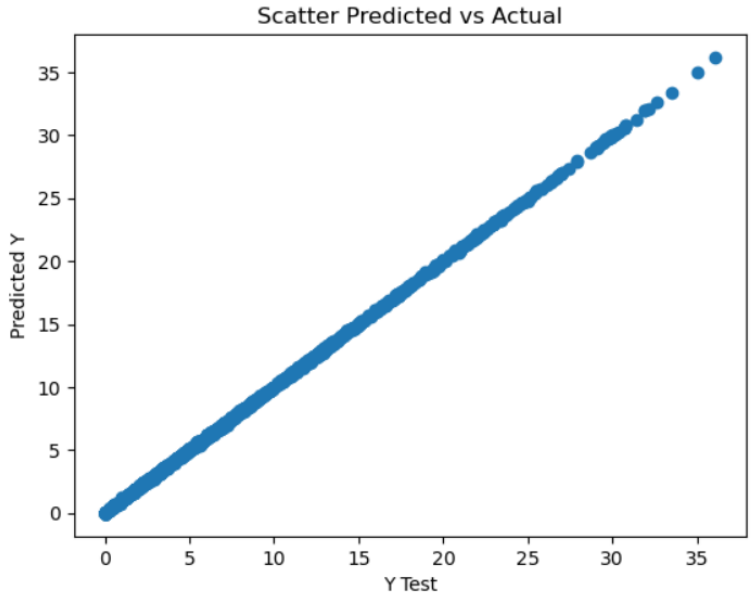
- **Split the train and test data**

  We divided the data into training data (70%) and test data (30%). This way, we can use the training data to train the model, so that it can accurately predict the test data that was not used in the training process. This division is important to test the performance of the model on data that has not been seen before, so that we can evaluate how well our model can handle new data. By training the model on training data and then testing on test data, we can ensure that our model not only learns from specific data, but can also accurately generalize the patterns found to new data.

- **Linear Regression**

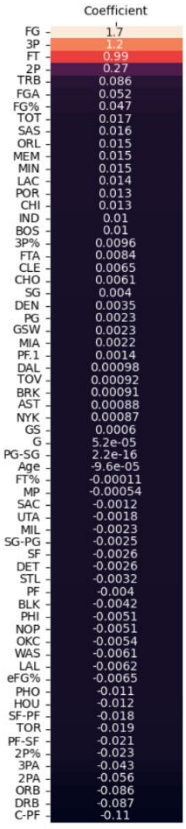
  We use linear regression to predict the response variable based on one or more predictor variables. In the context of dividing the data into training and test data, linear regression is used to train the model on 70% of the training data, where the model will learn the relationship patterns between the predictor variables and the response variable. After training the model, we tested its performance on 30% of the test data that was not used during the training process. The aim is to ensure that our model can produce accurate and reliable predictions when applied to new data.

- **Prediction Result**



The scatter plot shows us an almost perfect line between the values predicted by the model and the actual values of the test data. This indicates that our linear regression model has a good ability to predict the response variable based on the predictor variables used. The high similarity between the predicted and actual values indicates that our model is able to capture the patterns present in the data and produce accurate estimates. This provides additional confidence that our linear regression model is an effective and reliable model to use in predicting future response variables.

- **Coeff Conclusion :**



  - Positive coefficients indicate that an increase in the corresponding independent variable leads to an increase in the dependent variable, while negative coefficients indicate the opposite.
  - The magnitude of the coefficient indicates the strength of the impact. Larger magnitudes suggest a stronger influence on the dependent variable.
  - Coefficients closer to zero suggest a weaker impact on the dependent variable.
  - Variables with coefficients close to zero or very small magnitudes may have little to no impact on the dependent variable in the context of the model.

- **For example:**

1. **FG (Field Goals Made)**: The coefficient of approximately 1.731 suggests that for every one-unit increase in the number of Field Goals Made (FG), there is an associated increase of approximately 1.731 units in Points Per Game, holding all other factors constant. This indicates that efficiency in scoring field goals can significantly enhance a player's performance in scoring points.

2. **C-PF (Center - Power Forward)**: The coefficient for "C-PF" is approximately -0.111. This indicates that for every one-unit increase in the "C-PF"

variable (which likely represents a combination or hybrid role between Center and Power Forward positions), there is an associated decrease of approximately 0.111 units in the dependent variable (presumably Points Per Game), holding all other factors constant.

The provided metrics are:

1. **MAE (Mean Absolute Error)**: 0.0485

    - MAE measures the average of the absolute differences between predictions and actual values.
    - A lower MAE value indicates that the model has lower prediction errors on average.
    - In this context, an MAE of 0.0485 indicates that the average absolute difference between predictions and actual PTS values is around 0.0485.
2. **MSE (Mean Squared Error)**: 0.00451

    - MSE measures the average of the squared differences between predictions and actual values.
    - A lower MSE value indicates that the model has lower prediction errors on average, with an emphasis on larger errors.
    - In this context, an MSE of 0.00451 indicates that the average squared difference between predictions and actual PTS values is around 0.00451.
3. **RMSE (Root Mean Squared Error)**: 0.0672

    - RMSE is the square root of MSE, providing a more intuitively interpretable error in the same units as the target variable.
    - A lower RMSE value indicates that the model has lower prediction errors on average, with an emphasis on larger errors.
    - In this context, an RMSE of 0.0672 indicates that the average difference between predictions and actual PTS values is around 0.0672, in the same units as PTS.

- **Conclusion**

    This model is almost perfect. I think this is where data leakage occurs. Data leakage occurs when information from outside the training set is used to build the model, which can lead to overly optimistic and unrealistic results when applied to new data. In the context of predicting NBA player performance (PTS or Points Per Game), we must ensure that the predictor variables do not contain information that directly reflects future target values that should not be known at the time of prediction.

    Let's review some variables:

    - **2P (2-point field goals per game) and 3P (3-point field goals per game):** 2P and 3P: These variables are a direct part of the total points scored by players. Since PTS is calculated as a result of different types of shots, including 2P and 3P, using these variables as predictors directly will lead to data leakage. This is because the model will 'see' part of the information that it is trying to predict.
    - **Other Variables:** FG (Field Goals), FT (Free Throws), FGA (Field Goal Attempts), FTA (Free Throw Attempts), etc.: These variables can also contribute directly to PTS and may cause data leakage if used as predictors.

Second Regression Analysis

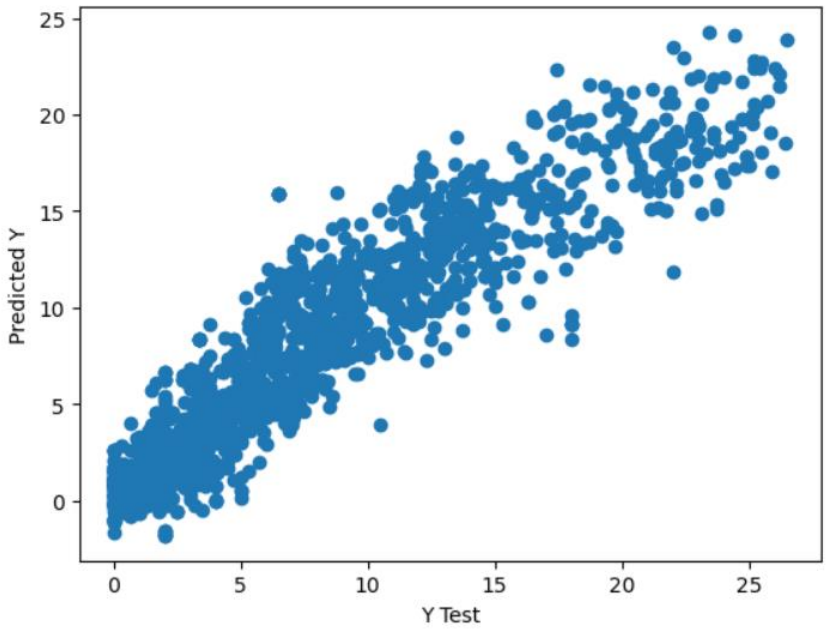
- **Removing outliers in PTS**

  Use IQR to remove outliers from the 'PTS' variable to improve the accuracy of the model:
  - Calculate Q1 (first quartile) and Q3 (third quartile) of 'PTS'.
  - Calculate the IQR (interquartile range) as the difference between Q3 and Q1.
  - Define the lower limit as Q1 - (1.5 * IQR) and the upper limit as Q3 + (1.5 * IQR).
  - Delete all 'PTS' values that fall outside this range.
- **Split the X and y Variable**

  X = All variable expect 'PTS', 'FG', '2P', '3P', 'FT', 'FGA', 'FTA', '2PA', '3PA', 'FG%', '2P%', '3P%', 'FT%', 'eFG%'

  Y= Variable 'PTS'
- **Split train and test data, and create the model with Linear Regression**

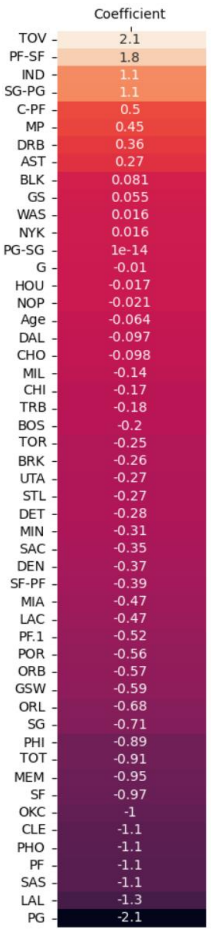
  The method is the same as the first regression.
- **Prediction Result**



Although the scatter plot does not show a perfect line between the value predicted by the model and the actual value of the test data, there is still a positive correlation between the two. Although there is variation in the model predictions compared to the actual values, the presence of a positive correlation indicates that the linear regression model is still able to capture the pattern of relationships that exist between the predictor and response variables. Although not perfect, the ability of the model to predict in the correct direction is an indication that the model is still useful in providing useful estimates. Nonetheless, it is important to still consider the variation and uncertainty that may occur in the model's predictions.

- **Coeff Conclusion**

  Positive



  - **TOV(Turnovers per game): 2.083** A positive coefficient indicates that the higher the number of turnovers a player commits, the higher the points he scores per game. This could mean that players with high turnovers tend to be more aggressive in scoring points, but also have a higher risk of losing the ball.
  - **MP (Minutes Played): 0.453** The positive coefficient indicates that the more playing time a player has, the more points he scores per game. This makes sense, as the longer a player is on the field, the more opportunities he has to score points.
  - **PF-SF and SC-PG: 1.06 and 0.5** The positive coefficient (1.08) indicates that players playing at the PF-SF position have a considerable impact on scoring points for their team per game. A positive coefficient (0.5) indicates that players playing in the SG-PG position also have a significant contribution in scoring points for their team per game. Players in this position are often responsible for the creation of attacks and the completion of attacks, which allows them to have more opportunities to score points.
  - **DRB (Defensive rebounds per game): 0.36** The positive coefficient (0.36) for Defensive Rebounds (DRB) per game indicates that the number of defensive rebounds grabbed by a player has a positive influence on the number of points scored by their team per game. This suggests that players who are active in grabbing defensive rebounds have a tendency to give their team more opportunities to initiate counterattacks or stop the opponent's offense, which can ultimately result in more points for their team.
  - **AST (Assists per game): 0.26** This shows that the number of assists a player makes has a positive impact on their team's attacking productivity. This suggests that players who are active in providing assists tend to boost their team's attack efficiency, helping their teammates to score points better.
  - **BLK (Blocks per game): 0.08** This shows that the number of blocks made by a player also has a positive impact on the team's success in scoring points. This is because a block can disrupt or even stop an opponent's attack, giving the team the opportunity to win the ball back and start a counterattack or even stop an opponent's attack outright.
  - **IND (Indiana Pacers): 1.08** this shows that the Indiana Pacers team tends to score more points compared to other teams. This could be due to several factors, such as effective attack strategies, individual players' ability to score points, or the team's superiority in overcoming the opponent's defense.

  Negative

  - **PG, PF, SF and others** Negative coefficients for player positions such as PG (Point Guard), PF (Power Forward), SF (Small Forward), and others with negative coefficients indicate that the presence or involvement of players in these positions tends to have a lower impact on the number of points scored by the team in each match. This interpretation could mean that players playing in these positions may have roles that focus more on other aspects of the game besides scoring points, such as organizing the offense, defending, or contributing in other ways such as rebounding or assists.
  - **Another baskeball team such as TOT, LAL, SAS and others** A negative coefficient indicates that these teams have a lower influence on the number of

points scored by the team in each match. This could be due to various factors such as poor offensive performance, strong defense, inefficient placement of players, or different game strategies.

The provided metrics are:

1. **MAE (Mean Absolute Error)**: 1.823

    - MAE measures the average of the absolute differences between predictions and actual values.
    - A lower MAE value indicates that the model has lower prediction errors on average.
    - In this context, an MAE of 1.823 indicates that the average absolute difference between predictions and actual PTS values is around 1.823.
2. **MSE (Mean Squared Error)**: 6.019

    - MSE measures the average of the squared differences between predictions and actual values.
    - A lower MSE value indicates that the model has lower prediction errors on average, with an emphasis on larger errors.
    - In this context, an MSE of 6.019 indicates that the average squared difference between predictions and actual PTS values is around 6.019.
3. **RMSE (Root Mean Squared Error)**: 2.453

    - RMSE is the square root of MSE, providing a more intuitively interpretable error in the same units as the target variable.
    - A lower RMSE value indicates that the model has lower prediction errors on average, with an emphasis on larger errors.
    - In this context, an RMSE of 2.453 indicates that the average difference between predictions and actual PTS values is around 2.453, in the same units as PTS.

Model Predict Demonstration with random player data

Column	Value	Column	Value
Age	23	C-PF	False
G	29	PF.1	True
GS	0	PF-SF	False
MP	10.4	PG	False
ORB	1.2	PG-SG	False
DRB	1.1	SF	False
TRB	2.4	SF-PF	False
AST	0.5	SG	False
STL	0.2	SG-PG	False
BLK	0.7	BOS	False
TOV	0.7	BRK	False
PF	1.6	CHI	False

Based on the data provided, the model was run and produced a predicted value of 3.10414 42, while the actual observed value was 3.4. The data used to run the model includes various attribute s, including age (23 years old), number of games (29), number of games started (0), average playing t ime (10.4 minutes), offensive rebounds (1.2), defensive rebounds (1.1), total rebounds (2.4), assists (0 .5), steals (0.2), blocks (0.7), turnovers (0.7), and personal fouls (1.6). In addition, there are several pl ayer position attributes such as C-PF, PF.1, PF-SF, PG, PG-SG, SF, SF-PF, SG, SG-PG, as well as t eam attributes such as BOS, BRK, and CHI, all of which are False.

The prediction results show that the model has a fairly good accuracy, although there is a sl ight difference between the predicted value and the actual value. The difference between the predicte d value and the actual value is about 0.2958558.

Conclusion

Based on the analysis conducted, there are several findings as follows:

- From the correlation analysis of each variable, it is found that variables directly related to points per game such as **2P, 3P, FT, FG** and others have a strong positive correlation which makes sense. **The interesting** findings are that **TOV and PF** also have a fairly strong positive correlation with PTS or points per game, because generally TOV and PF can harm the team and lose the opportunity to score points, but on the other hand this also indicates that players who have high points tend to **play aggressively** so as not to rule out the possibility of turnovers or fouls. For example, **Joel Embiid**, who has an average point per game of 32.9, has TOV and PF above the average of the entire dataset.
- If we look at players who have a high percentage of shots that reach **100% without a miss**, they tend to have **very small** shot attempts, for example if a player makes 2 shot attempts and 2 goals, his shooting percentage is 100%, which is not comparable to a person who makes 20 shots and 10 goals only has **50%**. So this **percentage cannot represent** the shooting ability of a basketball player.
- Position-based analysis, players who are near the ring such as **PF, SF and C**, tend to block, rebound or steal to defend or receive passes from **SG, PG. Because SG and PG** tend to be far from the ring, these positions are more focused on scoring points from afar or driving to the position of **PF or SF or C** to score points up close as assists.
- Conducting a performance comparison analysis between the team with the highest number of scores, **Miami Heat (MIA)**, and another team, **team TOT**, shows an interesting trend in their performance. Team MIA **consistently saw an increase** in the number of points scored per game in each match. Their steady and increasing performance makes them the team with the highest number of scores overall. On the other hand, **team TOT** showed a **significant spike** in performance in their last three matches. Although it did not initially register as high a number of points per game as MIA, the drastic improvement in their last three matches made **the TOT team able to rival**, and even approach, the performance of the MIA team in terms of the number of points scored. This trend suggests that the TOT team may have found a more **effective strategy** or form of play in the latter stages of the season, allowing them to compete with top teams like the MIA.
- Based on **Anova test**, there is no significant difference between **PTS, AST, TRB, PF** have no significant difference every month except **TOV**.
- **The first regression** produces an almost perfect model in predicting points per game, because there are several variables that strongly describe the prediction of points per game such as **2P, 3P, FT, FG** and others that cause the **data leakage**. Judging from the coefficients between variables in the model **dominated** by FG, FT, 2P and 3P which clearly describe the points per game. Therefore, we should discard them.
- **The second regression makes more sense**. The coefficient analysis shows that certain factors have a significant influence on the number of points scored by the team in each match. Players with high **turnovers (TOV)** and **more playing time (MP)** tend to score more points. **The PF-SF and SG-PG** positions contribute **significantly** to scoring points, while defensive rebounds (DRB), assists (AST), and blocks (BLK) also have a **positive impact**. **The Indiana Pacers (IND) team** scored more points than any other team. **In contrast**, players at positions such as PG, PF, and SF, as well as teams such as TOT, LAL, and SAS, tend to have a lower influence on the number of points scored, possibly due to a focus on other aspects or a less effective offensive strategy. **The MAE, MSE and RMSE** metrics are also relatively small.

Data : https://www.kaggle.com/datasets/bryanchungweather/nba-player-stats-dataset-for-the-2023-2024

Jupyter Notebook: https://github.com/AinulMr/NBA_Player_Performance

Google Slide :
https://drive.google.com/drive/folders/134cAUFBhs2KDWX3gTny05UZG4uXz6juw?usp=sharing