Data, Intelligence and Analytics Report

NBA Player Performance 2023-2024

Ainul Marzuki Febri
Data Enthusiast

Our Data
Column

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

Preparing Data

Mars

01

Importing dataset each month

Mercury

02

Adding date column before combine the dataset

Jupiter

03

Save the dataset ready for use

Dealing with missing value

FG%

Field goal percentage (112)

FT%

Free throw percentage (827)

2P%

2 point percentage (238)

3P%

3 point percentage (633)

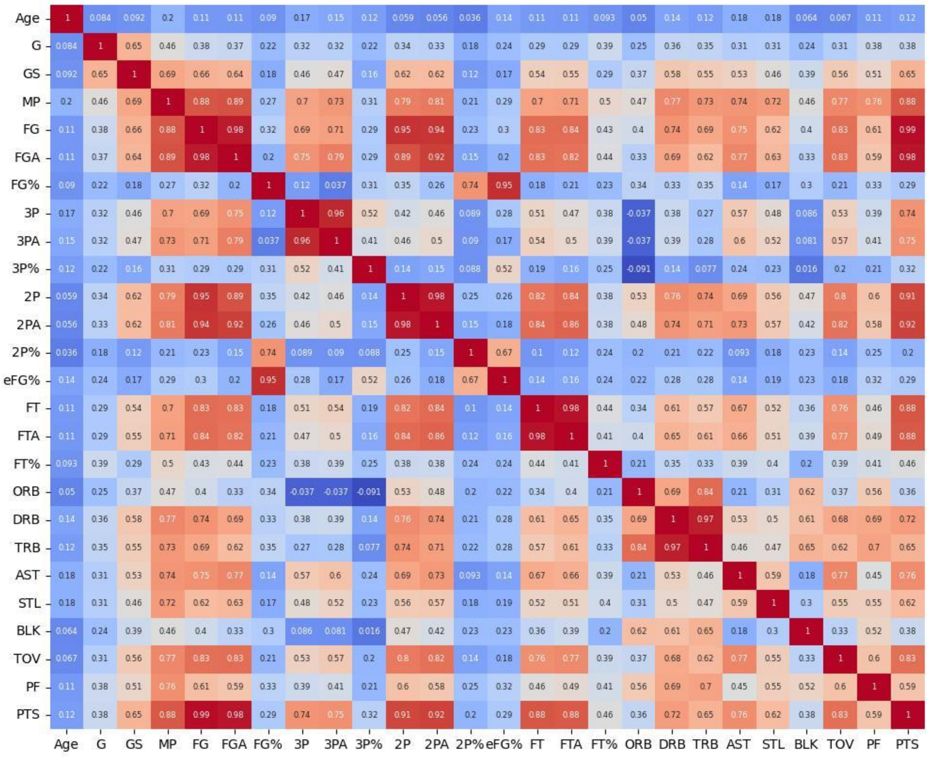
eFG%

Effective gield goal percentage (112)

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

Exploratory Data Analysis (EDA)

Correlation



Age vs. Performance Metrics

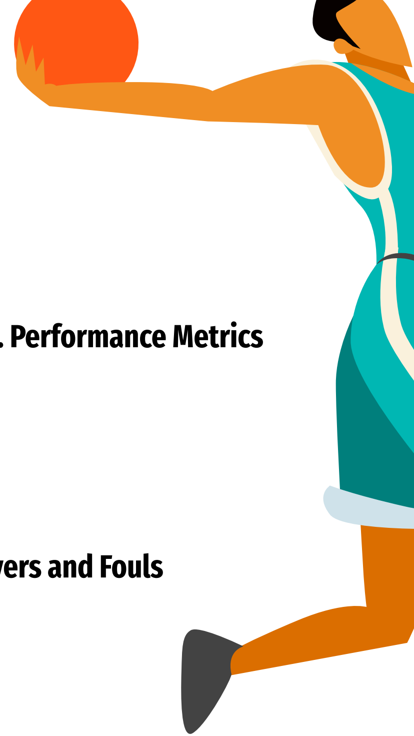
Playing Time Metrics

Turnovers and Fouls

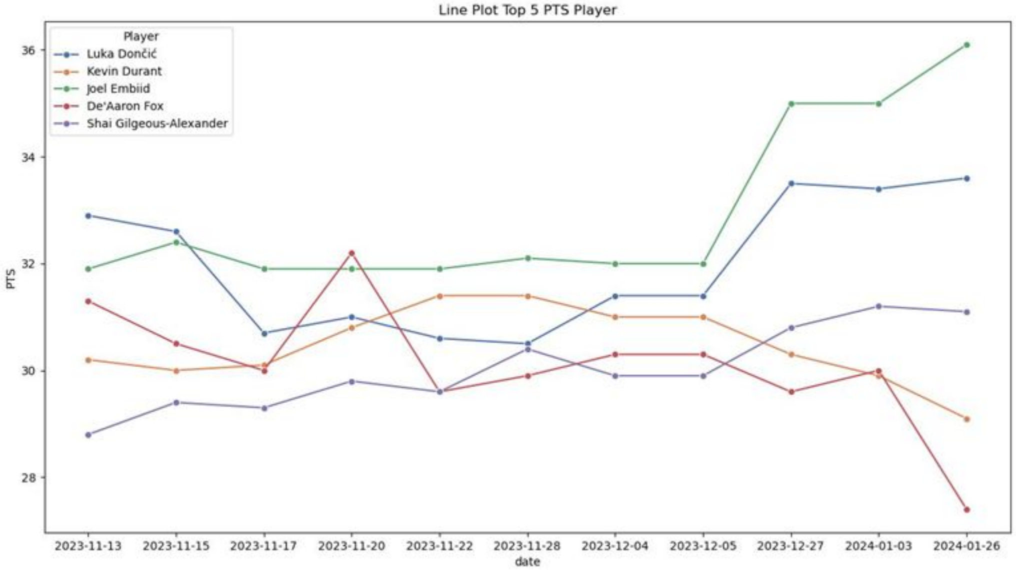
Shooting Metrics

Efficiency Metrics

Rebounding and Defense



Top Player Based on PTS



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).

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

Top Player based on AST

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

AST | **PG** | **SG** | **C**

The data shows the significant contributions of five NBA players in different aspects of the game. Tyrese Haliburton (Indiana Pacers) stands out with the highest assists and low turnovers, showing playmaking efficiency. Trae Young (Atlanta Hawks) has high assists but with more turnovers, reflecting his aggressiveness. Devin Booker (Phoenix Suns) contributed in assists and rebounds. Nikola Jokić (Denver Nuggets), a center, showed his versatility with outstanding rebounds and assists. Fred VanVleet (Houston Rockets) was efficient in assists with low turnovers. In general, assists generally come from players positioned away from the ring such as point guards (PG) and shooting guards (SG), emphasizing the important role of guards in ball distribution.

Average effective field goal percentage

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

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.

TOP

2P

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

Minimal 2PA = 10 Attempt

3P

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

Minimal 3PA = 8 Attempt

The data shows that three-point shooting percentage (3P%) is dominated by players positioned away from the hoop, such as point guards (PG) and shooting guards (SG). Players like Stephen Curry and Tyrese Haliburton stand out in this category, reflecting their ability to score from long distances. In contrast, the two-point shooting percentage (2P%) is dominated by players who are close to the hoop, such as centers (C) and shooting guards (SG).

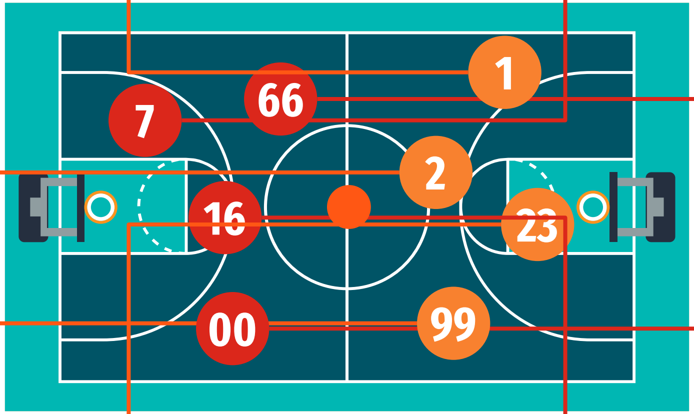
Position Performance

Position Most PST
SG and PG

Position Most AST
SG and PG

Position Most STL
SF and PF

Position Most BLK
C, SF and PF



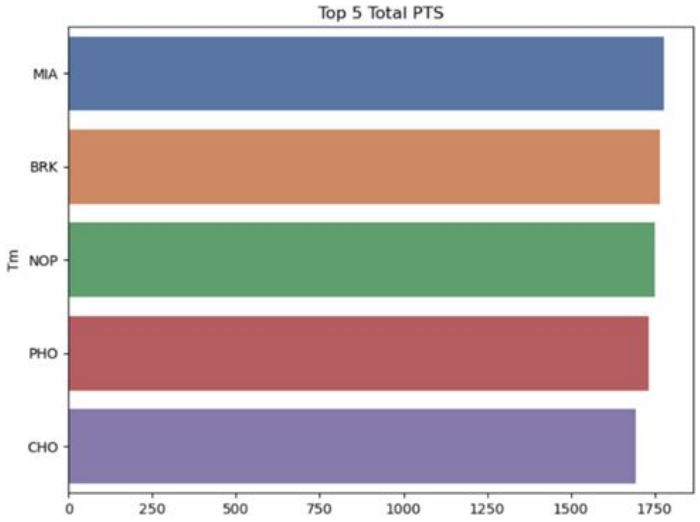
Position Most TOV
SG and PG

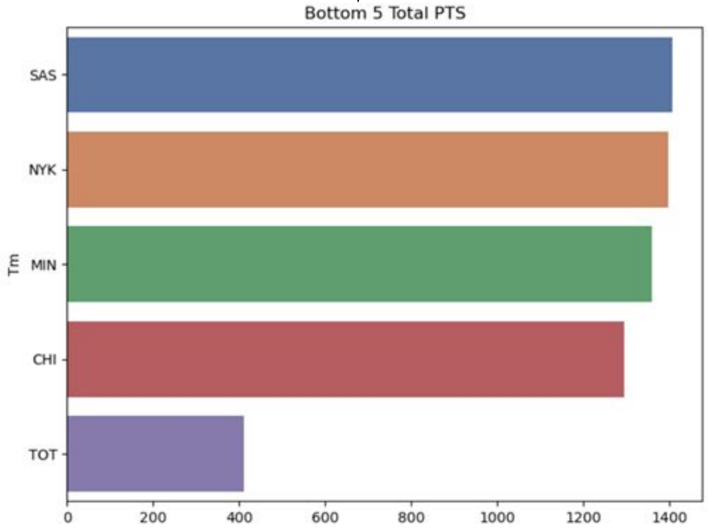
Position Most PF
C, SF and PF

Position Most TRB
C and PF

Position Most eFG%
C, SG and PG

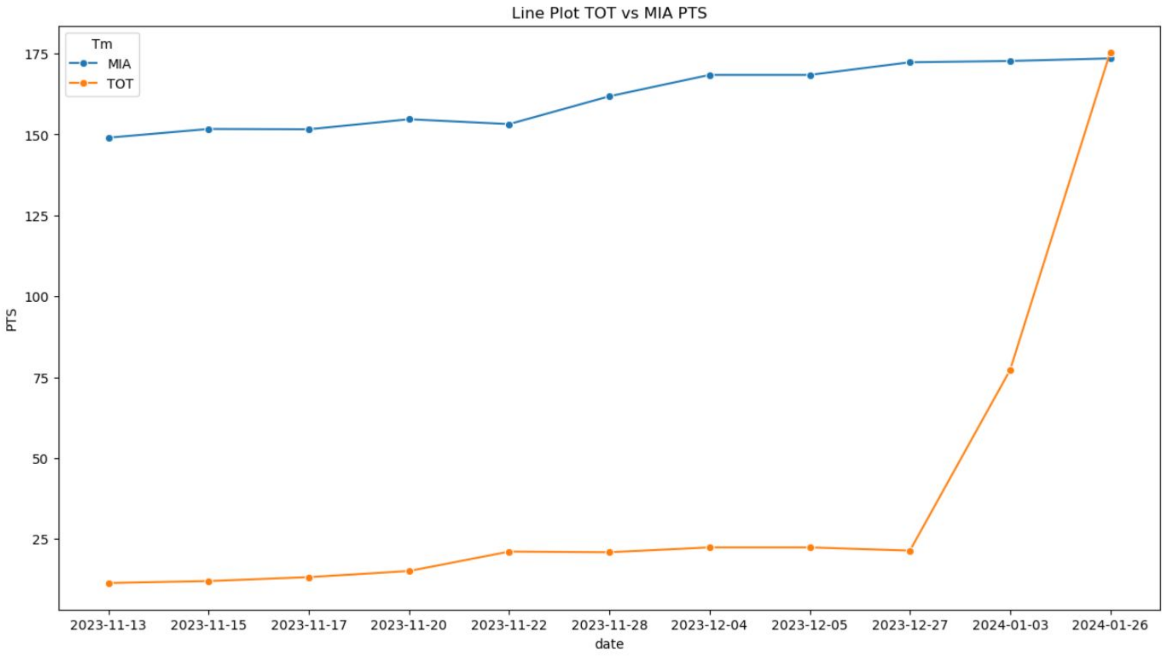
Team Comparison



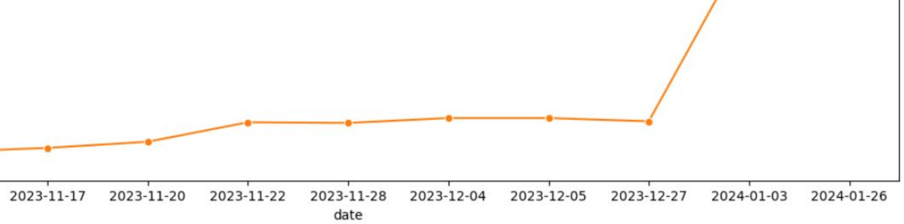


Miami Heat (MIA) was the team with the highest total score during the match, reaching 1777 points. This puts MIA at the top of the scoring table, with a not-so-significant difference compared to the other high-performing teams. However, there is a stark contrast when comparing MIA to the TOT team, who 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.

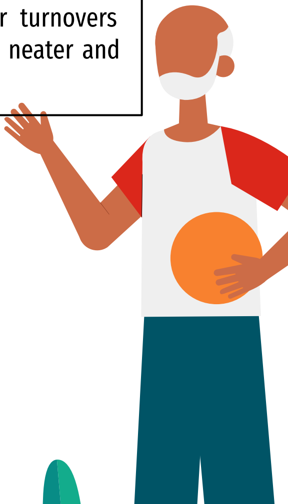


MIA

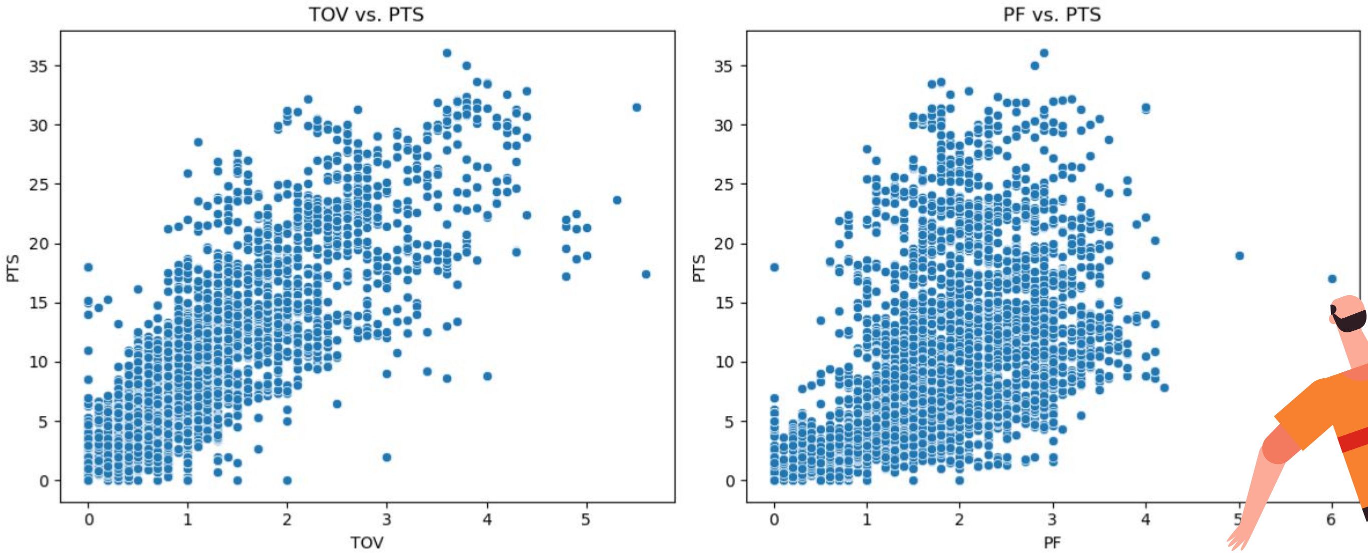
- Age and Experience: MIA players have an average age of 26.89 years, younger compared to the TOT team. This may indicate more energy and long-term development potential.
- Playing Time and Efficiency: MIA players play more minutes per game (20.98 minutes) and produce more points per game (9.45 PTS) 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.

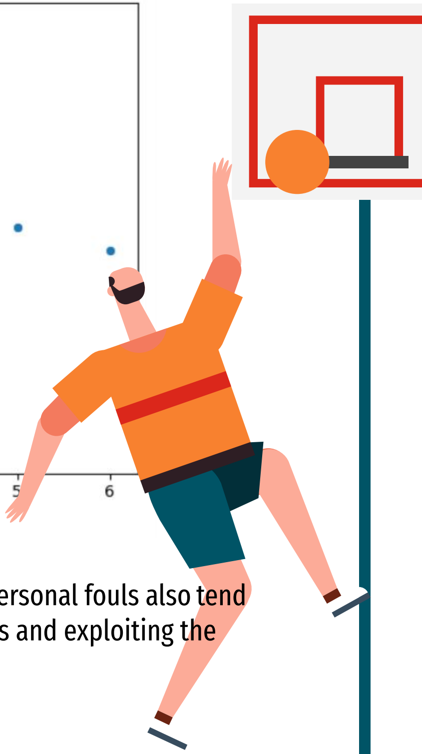
TOT

- 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
TOV and PF with PTS

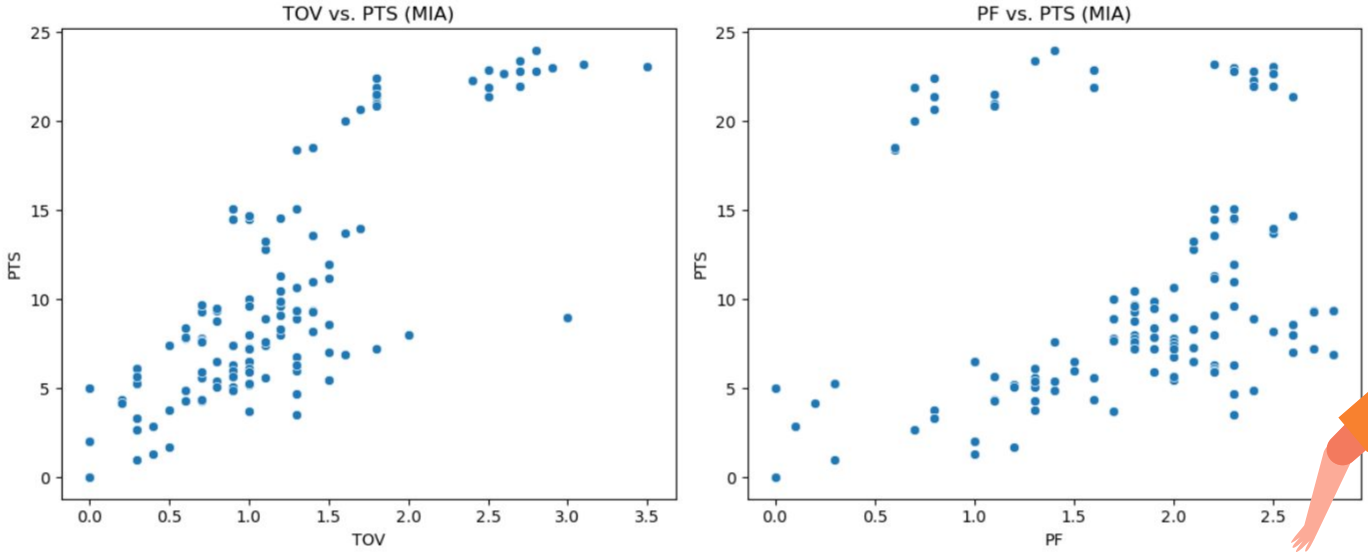




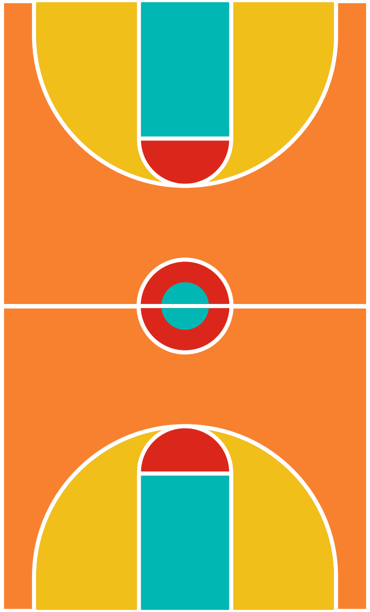
Most likely, the positive correlation between TOV and PF with PTS indicates that teams that generate more turnovers and personal fouls also tend to score more points. This may be because the team plays an aggressive style of play, creating more attacking opportunities and exploiting the opponent's mistakes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

MIA Team

TOV and PF with PTS



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



Statistical Analysis (Anova test)

Anova Test

```
Performance Metrics: PTS
F value: 0.09034554537282424
P-value: 0.9136167862972281
There is no significant difference in PTS between months.

Performance Metrics: AST
F value: 0.0004731220415519137
P-value: 0.999526989903534
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.
```

Summary Conclusion:

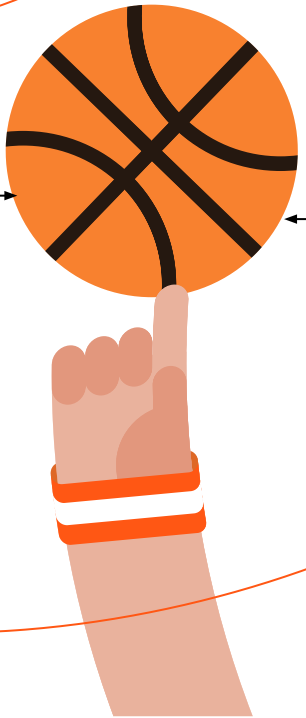
- Of the performance metrics tested (PTS, AST, TRB, TOV, PF), only turnovers per game (TOV) showed significant differences between months. For the other performance metrics (PTS, AST, TRB, PF), there were no significant differences between the months.
- This means that player performance in terms of turnovers varies significantly between months, while performance in terms of points, assists, rebounds, and personal fouls tends to be consistent throughout the months.

Categorical Variables

Categorical Variables

Before conducting regression analysis, it is important to convert all categorical variables that will be used as predictors into dummy variables

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".



Team

Team has 31 unique values

Position

Position has 10 unique values

So, to avoid the dummy variable trap, we need to drop one of the dummy variables.

Regression
Analysis

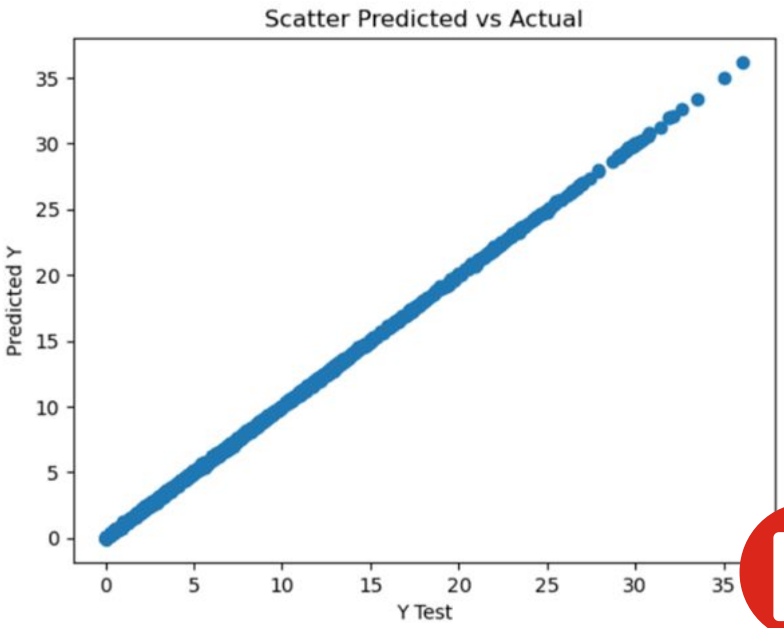
First Regression Analysis

Split Data and Train Test Data

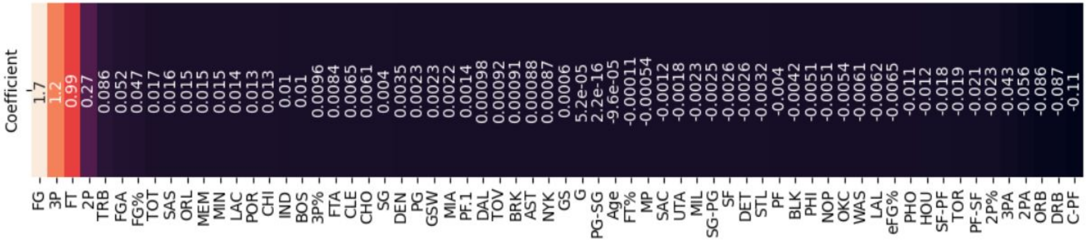
X = All variable expect 'PTS'
Y= Variable 'PTS'

We divided the data into training data (70%) and test data (30%)

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.



Coeff First Regression Analysis



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.

It can be seen that the positive coefficient is driven by FG, 3P, 2P and FT. It makes sense because those are the parameters that directly describe points per game (PTS).

Evaluation First Model

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.
   - 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.

BUT 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:

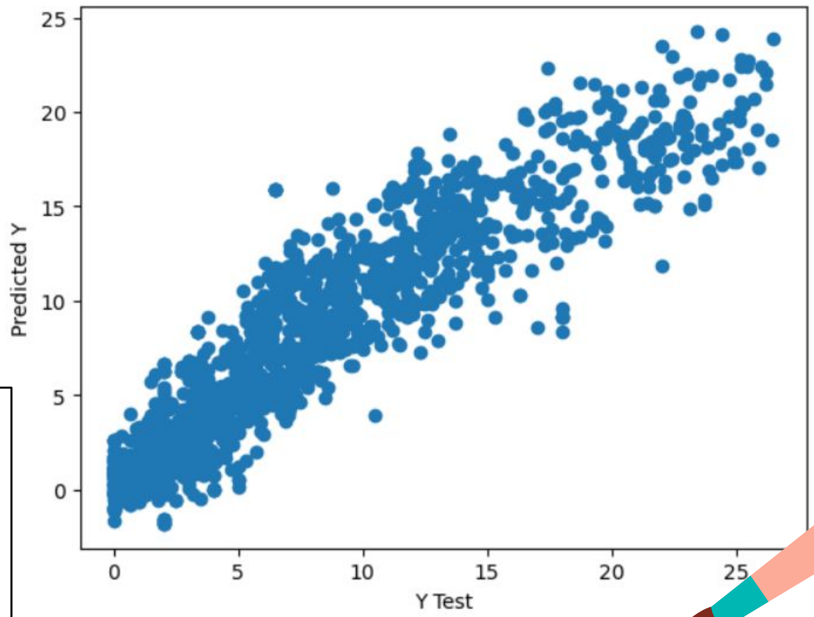
- **2P (2-point field goals per game) and 3P (3-point field goals per game):** 2P and 3P: These variables are a direct part of the total points scored by players. Since PTS is calculated as a result of different types of shots, including 2P and 3P, using these variables as predictors directly will lead 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.
- So there is no valuable insight from this model

Second Regression Analysis

Removing Outlier in PTS and Split Data and Train Test Data

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

We divided the data into training data (70%) and test data (30%)



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.

Coeff Second Regression Analysis

Positive:

- Turnovers (TOV): More turnovers correlate with more points, indicating scoring aggressiveness.
- Minutes Played (MP): More playing time correlates with more points.
- PF-SF and SG-PG positions: Players in these positions contribute significantly in scoring points.
- Defensive Rebounds (DRB), Assists (AST), Blocks (BLK): Defensive activities and assists increase the team's points.
- Indiana Pacers (IND): This team scores more points than any other team.

Negatives:

- PG, PF, SF positions: Players in these positions focus on other aspects besides scoring points.
- Other Basketball Teams (TOT, LAL, SAS, etc.): These teams score fewer points due to various factors such as game strategy and efficiency.

Evaluation Second Model

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.
    - 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 Demonstration

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

Predict value vs Actual Value

```
predicted_value = lm.predict(new_player.values.reshape(1, 51))
actual_value = df.iloc[random_ind]['PTS']
print ('Predicted value', predicted_value)
print ('Actual value', actual_value)
```

```
Predicted value [3.1041442]
Actual value 3.4
```

Based on the data provided, the model was run and produced a predicted value of **3.1041442**, while the actual observed value was **3.4**. The data used to run the model includes various attributes, including age (23 years old), number of games (29), number of games started (0), average playing time (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 player position attributes such as C-PF, PF.1, PF-SF, PG, PG-SG, SF, SF-PF, SG, SG-PG, as well as team 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 slight difference between the predicted value and the actual value. The difference between the predicted value and the actual value is about **0.2958558.**

Conclusion

Based on the analysis:

1. **Correlation Analysis:** Variables like 2P, 3P, FT, FG show strong positive correlations with points per game (PTS). Surprisingly, TOV and PF also correlate positively with PTS, suggesting aggressive play styles can lead to higher scoring despite risks of turnovers and fouls.
2. **Shooting Efficiency:** High shooting percentages (e.g., 100%) often correlate with low shot attempts, which can skew perceptions of shooting ability compared to players with higher volume shooting.
3. **Positional Analysis:** PF, SF, and C focus on inside scoring, while SG and PG contribute more from perimeter shooting and playmaking roles, highlighting different scoring dynamics based on position.
4. **Team Performance:** Miami Heat (MIA) consistently scored the most points per game, while team TOT showed a late-season surge, indicating potential strategic adjustments or improved performance towards the end of the season.
5. **Regression Analysis:** The first regression model was flawed due to data leakage from variables like FG, FT, 2P, and 3P. The second regression model highlighted factors like TOV, MP, PF-SF, SG-PG positions, DRB, AST, BLK influencing higher points per game. Indiana Pacers (IND) stood out for scoring prowess, while PG, PF, SF positions and teams like TOT, LAL, SAS showed lesser impact on scoring.
6. **Statistical Metrics:** MAE, MSE, and RMSE metrics indicated relatively small errors in predictions, emphasizing the model's accuracy in forecasting points per game.

Thanks! :)

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

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