

Analyzing factors towards Minutes Played per player in Professional Basketball: A Predictive Modelling Approach

Ali Ammar, Arudhra Venkatachalam, Surabhi Borase, Sai Sathyanarayanan

Mujazi Kekepuram

Group: 21

MSc Data Science group project

April 2024



UNIVERSITY OF
BIRMINGHAM

University of Birmingham Dubai

Acknowledgements

We would like to acknowledge and thank our supervisor, **Syed Fawadh Hussain , Ph.D.(CS)** for his advice and support throughout the duration of our group project. His feedback was integral to the group investigation of our research topic both in terms of dashboard and the formulation of the final report.

We would also like to thank the Data.World website and author “**Eduardo Tocco**” for the availability of the data used in this project freely and publicly available for our analysis.

Abstract

In recent years, the integration of advanced analytics into sports has revolutionized the approach to player evaluation and game strategy, particularly in basketball. This study explores the influence of turnovers (TOV), personal fouls (PF), and player positions (POS) on the minutes played (MP) by players in NBA games. Utilizing data spanning from the 1997-98 to the 2021-22 NBA seasons, we applied multiple machine learning models to analyze the relationship between these variables and their impact on player court time. Our analysis employed models including XGBoost, Random Forest, and Support Vector Regressors, which revealed that the combined factors of TOV, PF, and POS could predict up to 80% of the variability in MP, indicating a strong predictive power. The study confirmed that turnovers and personal fouls significantly decrease the minutes a player plays, with turnovers having a slightly more pronounced effect than fouls. The position of a player also plays a crucial role, although it has less predictive power when considered alone. The findings underscore the importance of nuanced metrics like turnovers and fouls in shaping player utilization strategies. They also highlight the evolving role of player positions in modern basketball, reflecting a shift towards more dynamic and positionless gameplay. This study contributes to the ongoing discourse in sports analytics by providing actionable insights that can help coaches and sports strategists optimize player performance and team efficiency in competitive basketball environments.

Keywords: Analytics, NBA, Strategy, Player Position, Turnover, Personal Foul, and Minute Played

Table of Contents

1	Introduction.....	5
2	Background Research	6
3	Question Development	8
4	Retrieving the Data	8
4.1	Data Description	9
4.2	Data Preprocessing & Feature Engineering.....	9
4.3	Addressing the Research Question	9
5	Rationale for the Exploratory Data Analysis	10
5.1	Univariate Analysis:.....	10
5.2	Bivariate Analysis	11
5.3	Position-Specific Trends:	13
5.4	Multivariate Analysis (Correlation Matrix)	14
5.5	Position-Specific Insights:	15
5.6	Hypothesis Formulation.....	15
5.7	Hypothesis Testing:.....	16
5.8	Why ANOVA and not others like T-Test, Chi-Square test etc?	16
5.9	One Way ANOVA:	16
5.10	Two Way ANOVA:	17
5.11	Regression Analysis:	18
6	Rationale for Data Modelling	18
7	Results.....	20
7.1	Statistical Significance of Variables.....	21
7.2	Impact of Individual and Combined Variables	21
7.3	Combination Effects:	21
9.3	Combination Effects	21
10.	Discussion of Interpretation of Results: How Has the Question Been Answered	22
10.1	Turnovers (TOV) and Personal Fouls (PF)	22
10.2	Role of Player Positions (POS)	22
10.3	Comprehensive Data Analysis	22
8	Conclusion	22
9	Group Work.....	23
10	Contribution by the each member.....	23
	References.....	24
11	Appendix.....	25

1 Introduction

All sports have undergone a revolution over the past decade due to disruptive technology, primarily the emergence of artificial intelligence and the utilization of sports analytics. These advancements, through both qualitative (visualization) and quantitative data, have transformed the way teams approach player evaluation and strategic decision-making. With an abundance of statistics available on players, teams, games, and seasons, there exists a prime opportunity to extract valuable insights that can enhance team performance and optimize player utilization. This has enabled data scientists to use data mining to create statistically significant models to predict effective techniques for players in the competitive landscape, thereby elevating their game. Basketball is one such sport where data-driven decisions are increasingly used to train players.

Drawing inspiration from the rich history of sports analytics, particularly in disciplines like baseball with the advent of sabermetrics, basketball has seen a surge in innovative methodologies for player evaluation. Early pioneers such as Oliver and Hollinger (2007) have advocated for evaluating players on a per-minute basis, paving the way for advanced metrics like the Player Efficiency Rating (PER).

The exploration of various metrics in basketball analytics and their impact on player performance has become increasingly significant. One such unexplored area that our research aims to delve into is understanding how turnovers and personal fouls per game influence the minutes played by players, with considerations based on their respective positions on the basketball court. This study is pivotal, as understanding the nuanced effects of these variables can lead to more informed coaching strategies and player development plans.

Moreover, analyzing the interplay between turnovers, personal fouls, and minutes played, conditioned on the player's position, can offer insights into positional demands and how they dictate player behavior on the court. This aspect is critical in the era of positionless basketball, where the traditional roles of positions are evolving. By examining these relationships, our research contributes to the broader conversation on optimal player utilization and strategic game planning in modern basketball.

The remainder of this paper is organized as follows: Section 2 provides a review of related literature in basketball analytics, highlighting key studies that have shaped our understanding of performance metrics and their implications. In Section 3, we detail the methodology and data used for our analysis, emphasizing the statistical techniques employed to dissect the relationships

between turnovers, personal fouls, and minutes played. Section 4 presents the findings of our study, including exploratory data analysis and statistical modeling techniques. This section aims to unpack the complexities of our research question and offer evidence-based insights. Finally, in Section 5, we discuss the implications of our findings and outline potential avenues for future research in this domain, suggesting how this work can be extended or applied in practical settings to improve team performance and player development strategies.

2 Background Research

Most of the literature has relied on traditional methods to analyze sports analytics to reduce turnovers, enhance performance, team selection, and improve the timing of players in basketball games. Turnovers, more common than one might expect given the high skill level of modern players, significantly impact offensive efficiency, overshadowing even shot percentages and rebound rates. As such, minimizing turnovers should be a primary focus for any offense aiming for victory. To this end, statistical analysis techniques such as Spearman's Correlation Test and Wilcoxon's Nonparametric Test have been utilized in SPSS to underline their significance.

A study by Daniel & Kylie (2014) found that the depth of player position change significantly positively affects game results, with the relationship conditioned by the number of personal fouls, the team's overall strength, and home game advantage. A larger rotation was shown to significantly improve the chances of winning, as demonstrated through logistic regression analysis on a broad dataset.

Sarlis and Tjortjis (2020) employed Machine Learning (ML) and Data Mining (DM) techniques on sports data from 2017 to 2020 to predict player performance using both background and advanced basketball metrics, such as player position, minutes played, and turnovers, in National Basketball Association (NBA) and Euroleague games. Their findings underscored that a balanced team rotation and role distribution are crucial for team success, highlighting a notable shift in strategies compared to the previous decade.

Zhang et al. (2018) conducted a clustering analysis to identify NBA players with similar attributes, disregarding traditional positions. By incorporating anthropometric properties and experience into their analysis, they identified five distinct clusters but found no correlation between team performance and player configurations based on these clusters.

Oskan et al. (2022) aimed to bridge the gap in NBA game outcome prediction methods by adopting a complex system approach, similar to previous works by Lutz (2012), Oh et al. (2015), and Kuehn (2017). By identifying player stereotypes independently of traditional positions and using clustering techniques like k-means for optimal predictors, their model achieved a prediction success rate of approximately 71% across an entire season. This success rate not only surpassed human experts but also rivaled top-performing models in similar research settings.

Most recently, Bian et al. (2024) applied a unified approach featuring a combination of supervised and unsupervised machine learning models. This approach incorporates dimension reduction techniques from unsupervised learning, such as principal components analysis (PCA), to streamline NBA player performance data, followed by clustering techniques for player categorization. Supervised learning, specifically neural networks, is then employed for season outcome prediction and optimal team roster formation, offering a novel perspective in the evaluation of team and player performance.

Building on this foundation, the integration of advanced analytics into basketball decision-making processes represents a pivotal shift towards a more data-driven approach. This evolution reflects not only in strategic game planning but also in player development and team management. For instance, the analysis of turnovers and personal fouls, as explored in the aforementioned studies, underscores the importance of nuanced player metrics in enhancing team performance. Such insights offer coaches and analysts the ability to tailor strategies that mitigate weaknesses and leverage strengths more effectively. Furthermore, the conditioning of these metrics on players' positions introduces a layer of complexity that reflects the dynamic nature of modern basketball. As teams strive for optimization in every aspect of the game, the role of comprehensive analytics, encompassing both traditional and innovative metrics, becomes increasingly indispensable. This burgeoning field promises to uncover new dimensions of player evaluation and game theory, potentially redefining basketball strategies for the future. Through the meticulous examination of player performance data, our research aims to contribute to this ongoing discourse, providing actionable insights that can influence both the theoretical and practical realms of basketball analytics.

3 Question Development

After considerable discussion and evaluation of various potential topics, our team unanimously decided to focus our research on analyzing the dynamics of basketball gameplay. Specifically, we are interested in exploring the relationship between turnovers, personal fouls, and the amount of time players spend on the court, taking into account their specific positions.

Our research will delve into the following question: **"How turnovers and personal fouls per game affect the minutes played do, conditioned on the player's position?"** and **"How do other performance metrics such as points scored (PTS), rebounds (TRB), assists (AST), and efficiency ratings influence the Minutes Played (MP) across different player positions in the NBA?"**. These question stems from our curiosity about the strategic aspects of basketball and the role of player behavior in game management. By examining how the frequency of turnovers and personal fouls might influence a coach's decision on player court time across different positions—whether guards, forwards, or centers—we aim to uncover patterns that could suggest deeper tactical insights.

To tackle this inquiry, we will gather detailed game-by-game data that includes each player's position, the number of minutes they played, and their recorded turnovers and fouls per game. Our approach will involve data cleaning and transformation to ensure accuracy and relevance, followed by rigorous statistical analysis. We plan to use visualization techniques to represent our findings effectively and build predictive models that explore the relationships between these variables.

The results of this study are intended to be comprehensive enough to inform basketball fans and coaches alike, providing a deeper understanding of how player discipline and game dynamics might affect playing time decisions. This exploration is not only about uncovering hidden statistical relationships but also about enhancing our understanding of basketball strategy through the lens of data analytics.

4 Retrieving the Data

The data for this research comes from two primary sources, both offering extensive datasets on NBA player statistics spanning from the 1997-98 season through the 2021-22 season. The primary dataset, created by Eduardo Tocco <https://data.world/etocco/nba-player-stats>, is publicly

available for use since 2021 and was sourced from www.sports-reference.com using a custom scraper. This dataset was specifically compiled for the analysis of basketball player history and is notable for its comprehensive coverage of player statistics across multiple seasons, providing a rich foundation for this study.

4.1 Data Description

The datasets include a wide array of basketball metrics, such as turnovers (TOV), personal fouls (PF), and minutes played (MP), alongside more advanced basketball metrics like player positions, which are crucial for this study's focus on understanding how TOV and PF per game affect the minutes played, conditioned on the player's position. However, a limitation of the current dataset is that it only contains data up until the year 2022, with no updates post-2022 as of the last creation in 2021. Despite this, the dataset's depth and breadth offer a solid basis for conducting detailed analyses of player performance trends over the years.

4.2 Data Preprocessing & Feature Engineering

From the research questions, we are analyzing primarily three variables namely Turnovers TOV, PF and POS towards MP. Hence to proceed with the analysis, we selected only the mentioned variables and ignored the rest for now. And all null values have been replaced with 0 instead of removing them due to limited data availability.

4.3 Addressing the Research Question

For research question one, it was a straightforward question and did not require much of data engineering works. We removed null values and teams values 'TOT' and 'SEA' as these were not real teams and had some positional values which were in combination and in real life players do not really play combination positions <https://jr.nba.com/basketball-positions/> . We selected only those features which were suitable for our research question one which is TOV, MP, POS and PF and we did not take other variables into account.

For question two, it was an extension of question one where we dealt with other parameters with which we would be able to predict MP other than the variables used in research question one. From the definition of Player Efficiency Ratings 'PER'

[https://en.wikipedia.org/wiki/Efficiency_\(basketball\)](https://en.wikipedia.org/wiki/Efficiency_(basketball)) we had to create a new variable named

"PER" in our dataset with the available formula

[https://en.wikipedia.org/wiki/Efficiency_\(basketball\)](https://en.wikipedia.org/wiki/Efficiency_(basketball)) .

5 Rationale for the Exploratory Data Analysis

We started with Univariate analysis of TOV, PF and MP categorized on Positions, to understand how values are distributed for each position and to find patterns like which position has higher number of PF, TOV and MP mapping to the definition of each position.

5.1 Univariate Analysis:

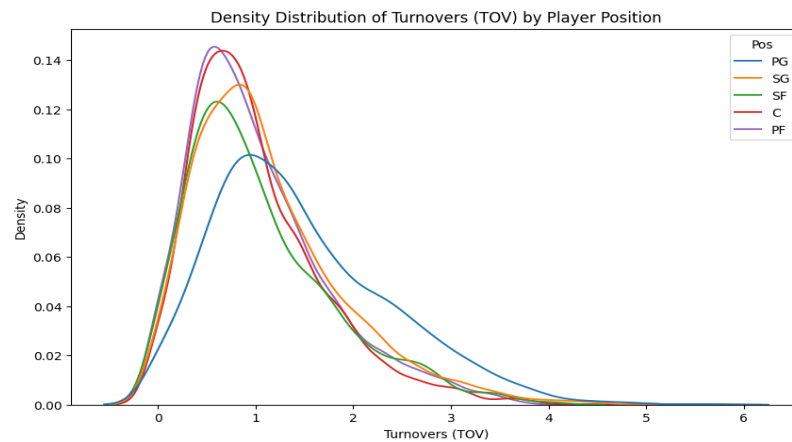


Figure 1. Density distribution of TOV by player position.

The above graph shows that the TOV data normally distributed across the given positions. Positions PF and C show some overlap indicating that the players in these positions pose similar characteristics in terms of TOV.

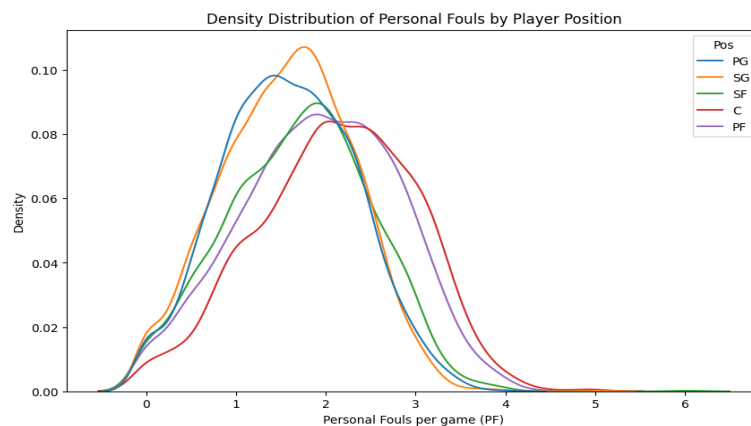


Figure 2. Density distribution of POV by player position.

The above graph shows the distribution of PF for different positions. Players in position C, SG and PG tend to have more Fouls according to the roles mentioned. Players in position C tend to play more aggressively in defensive nature, hence there are higher chances of fouls and as expected, the graph is peaking at mean value 2 for C, but surprisingly, players who are in position SG and PG are having mean value of Foul close to 1.5 even though they are the one playing predominantly 3Ps and long shoots.

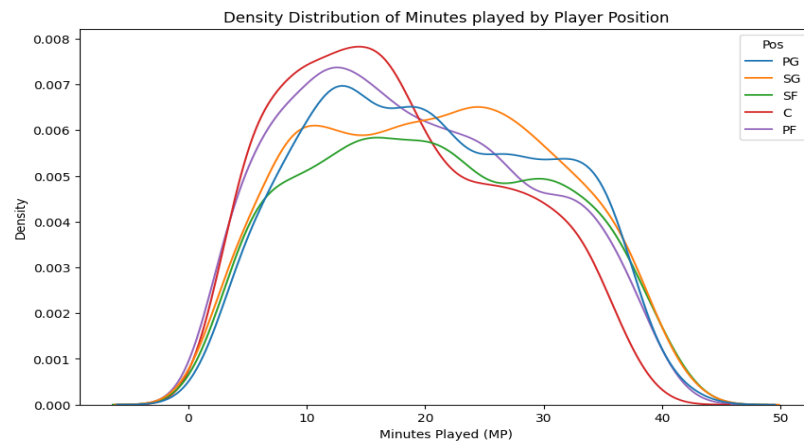


Figure 3. Density distribution of MP by player position

The above graph shows the distribution of Minutes played for each different position. The players in positions C tend to have less MP as their mean values is close to 15. The players in position SG are having the highest mean value 25 , indicating that these are the most prominent players even though their average fouls rate is 1.5 from the previous graph .

This analysis was done to understand how the mean values differ for independent variables TOV and PF and dependent variable MP on different positions and as visualized the mean values differ for different positions.

5.2 Bivariate Analysis

After studying the distribution of values of each variable independently, we moved forward with bivariate analysis to understand the relation between two variables and how they are correlated to each other. We used a scatter plot for analysis as our interest was in the relationship of two variables (non-time series).

MP vs TOV: The figure 4 show the relationship between MP and TOV. For each position, we are witnessing a positive relationship. As MP increases, the TOV also increases,

with few outliers. Also the trend is not linear as the rate of increase of TOV changes as MP increases.

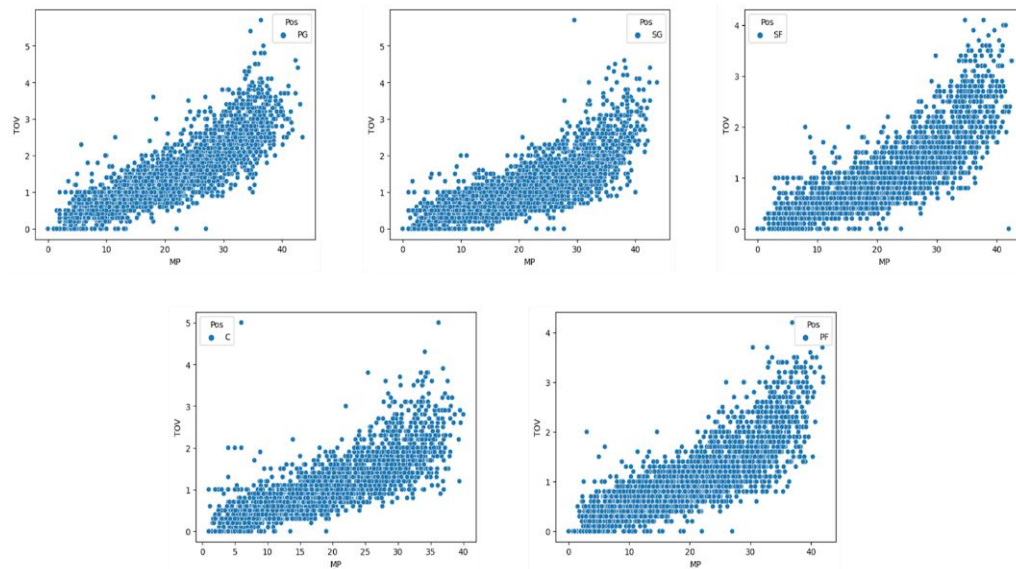


Figure 4. Relationship between MP and TOV

Figure 5 shows the relationship between MP and PF. As observed, we are witnessing a positive relation between MP and PF as well. We also observe that for all POS, the PF value is 0 even for higher MP, suggesting the existence of more efficient players.

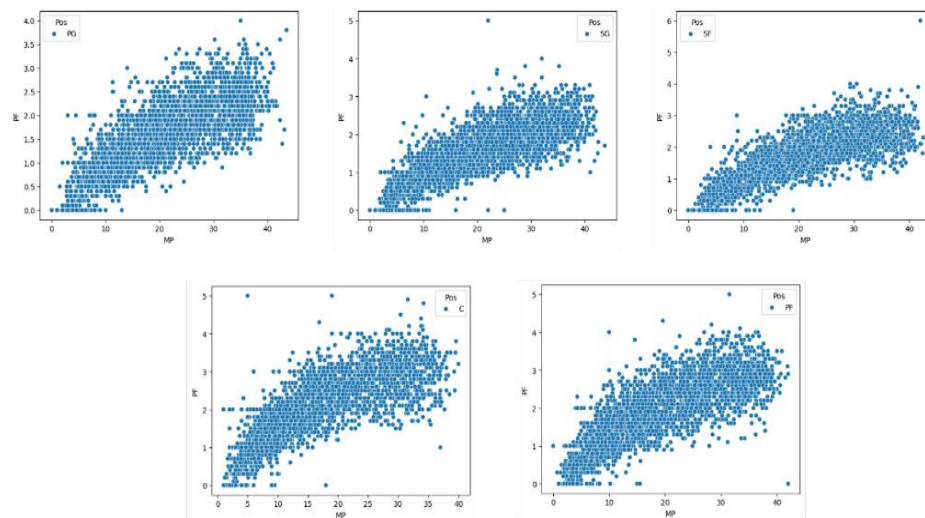


Figure 5. Relationship between MP and PF.

The below plot 6 shows the relation between two independent variables TOV and PF. This will be one of our interaction terms in the upcoming analysis to see the combined effect of TOV and PF on MP.

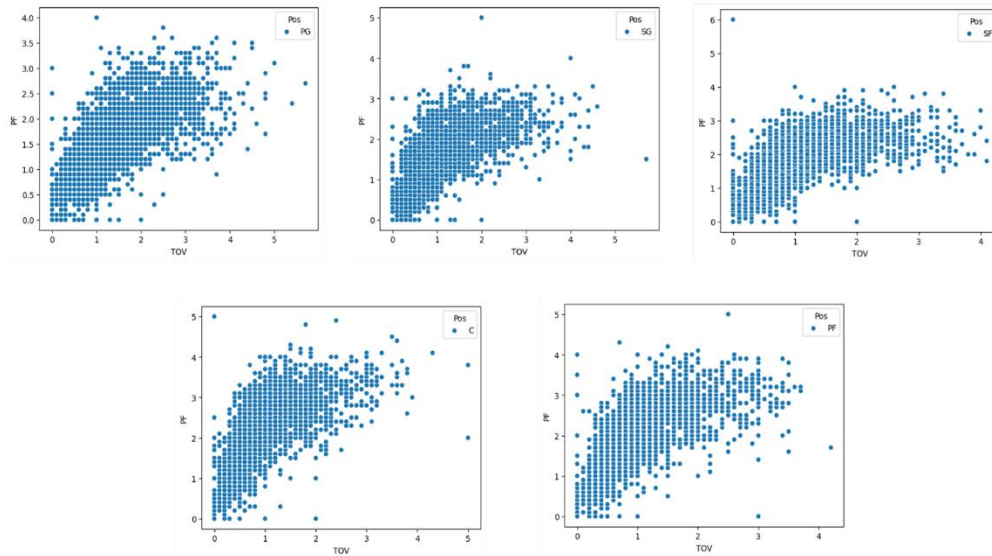


Figure 6. Relationship between TOV and PF.

Relationship between TOV and PF: Across all positions, there does not appear to be a strong, clear linear relationship between turnovers and personal fouls. While there's a broad trend that more turnovers can be associated with more personal fouls (and vice versa), the scatter plots reveal a lot of variability and no distinct pattern that holds consistently across all positions.

5.3 Position-Specific Trends:

PGs: The Point Guards show a somewhat more pronounced upward trend, possibly indicating that as PGs are often ball handlers, their turnovers and personal fouls may slightly increase together.

SGs and SFs: The Shooting Guards and Small Forwards have a large spread in both TOV and PF with no clear direction, which may suggest that other factors influence their turnovers and fouls more than their position does.

Cs: Centres show a wide spread primarily in the PF direction, which could indicate that their role in the paint and under the basket could result in varying levels of PF while TOV remains less varied.

PFs : Similar to Centers, Power Forwards show a wider spread in personal fouls, likely due to their defensive roles and physical play under the basket.

Variability: There's a high degree of variability in the data, especially for the Small Forward, Center, and Power Forward positions, indicating that there are likely many factors at play in determining the number of turnovers and fouls a player commits, and these factors may vary from player to player even within the same position.

To summarize, the bivariate analysis was conducted to check the relation between two variables and conclude whether any linear relationship is visible or not. Linear trend was visible between TOV vs MP and PF vs MP, but not a clear linearity in TOV vs PF.

5.4 Multivariate Analysis (Correlation Matrix)

From the previous section, we tried to understand data distribution within a specific variable and relationship between two variables conditioned on different positions. Now we will be plotting a correlation matrix for each POS to see how all three variables are correlated and how the correlation differs for each POS.

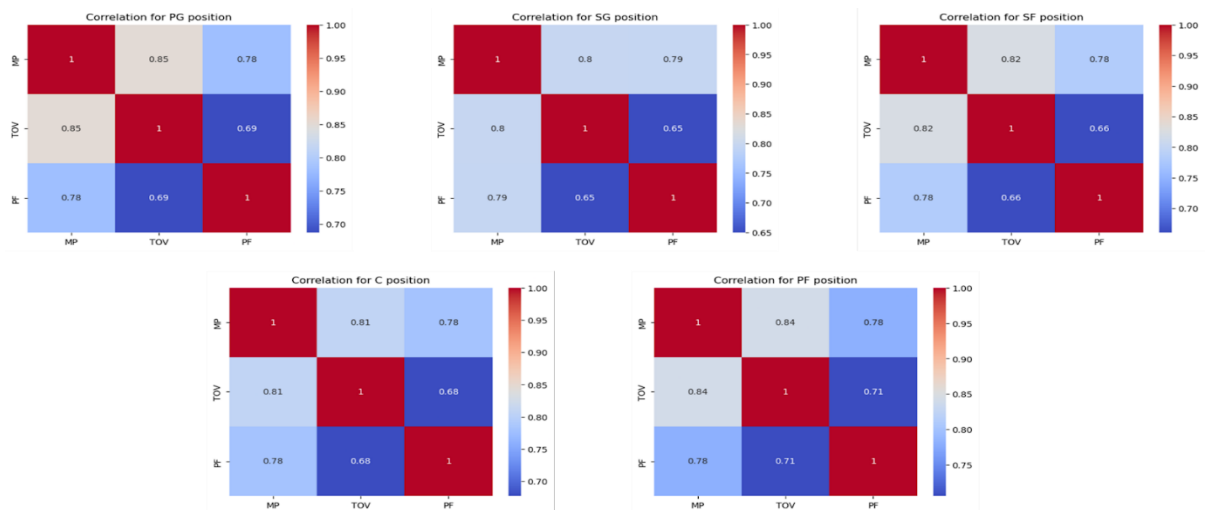


Figure 7. Correlation matrices between TOV, PF and MP on all Positions.

From the above matrices, the relation between TOV and MP is the highest ranging from 0.8 to 0.85. This suggests that players who play more minutes are typically likely to have more number of TOVs due to increased ball handling opportunities.

The relation PF and MP is also high, dwindling from 0.78 to 0.79. This can indicate that players who play more in the court are likely to make more number of PFs due to defensive strategies.

The relation between TOV and PF is also high but not as high as TOV vs MP and PF vs MP. The relationship score ranges between 0.65 to 0.71. While the relationship is lower, we can say that more number of TOVs means more number of PFs, but this is not always the case and this can be verified in the bivariate analysis between TOV and PF, where the relationship is not linear.

5.5 Position-Specific Insights:

- **PG Position:** Point Guards show the strongest correlations among the three variables, which aligns with the nature of their role, involving extensive ball-handling and playmaking, leading to more opportunities for both turnovers and fouls.
- **SG and SF Positions:** Shooting Guards and Small Forwards have slightly lower but still significant correlations. This might reflect their roles as scorers and wing defenders, where they are less likely to handle the ball as much as Point Guards but still engage in substantial offensive and defensive actions.
- **C Position:** Centers have the lowest correlation between TOV and PF, which could be due to their position mainly under the basket, leading to less ball-handling and varied defensive responsibilities compared to guards.
- **PF Position:** Power Forwards' correlation patterns are similar to Centers, indicating that while they are involved in physical play that can lead to fouls, their turnovers are not as closely related to their minutes played as with the guard positions.

From the above analysis, we can conclude that the relationship between the three variables TOV, PF and MP is different across different positions and the impact factor is wide enough.

5.6 Hypothesis Formulation

6.6.1 H0 (Null Hypothesis): The player's position does not moderate the relationship between minutes played and turnovers/personal fouls.

6.6.2 H1 (Alternate Hypothesis): The player's position significantly moderates the relationship between minutes played and turnovers/personal fouls, with some positions showing stronger relationships than others.

5.7 Hypothesis Testing:

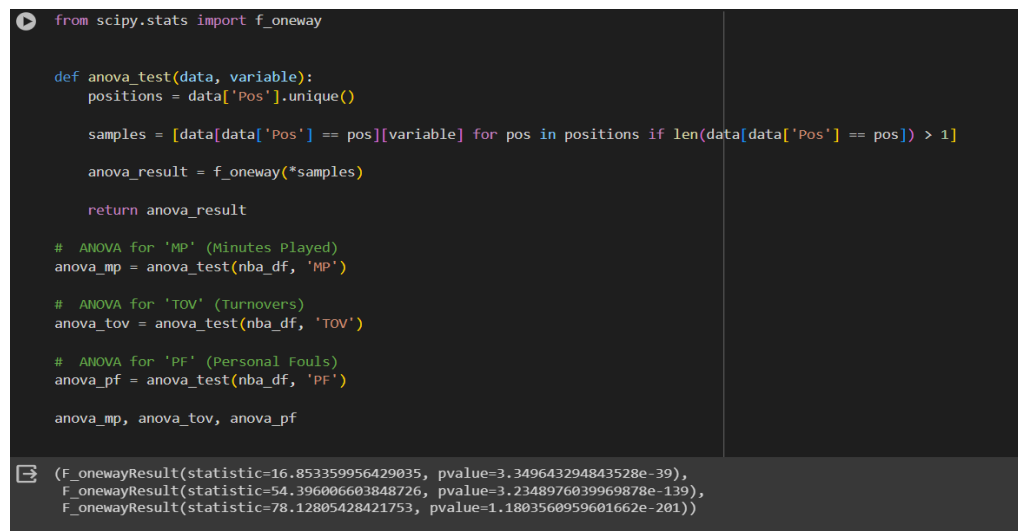
After conducting Univariate, Bivariate and Correlation analysis, we formed the hypothesis as stated in the previous section. In order to prove it, we performed ANOVA(Analysis Of Variables) statistical test both One Way and Two Way to check whether our alternate hypothesis is correct or not.

5.8 Why ANOVA and not others like T-Test, Chi-Square test etc?

Well, we have our categorical variable POS which has more than two categories (C, PF, SF, SG, PG and combinational positions). Since we have quantitative variables (TOV, PF and MP) and a categorical variable with more than 2 categories, we preferred the ANOVA test. T-Test is generally preferred when there is a categorical variable, Chi-Square is performed when we have only two categories in the categorical variable.

We performed a One Way ANOVA test to check the mean distribution of each variable independently by taking sample data from each position. We performed a Two Way ANOVA test to check the mean distribution of each variable along with Interaction terms towards the dependent variable MP. Below are the results observed?

5.9 One Way ANOVA:



```

from scipy.stats import f_oneway

def anova_test(data, variable):
    positions = data['Pos'].unique()

    samples = [data[data['Pos'] == pos][variable] for pos in positions if len(data[data['Pos'] == pos]) > 1]

    anova_result = f_oneway(*samples)

    return anova_result

# ANOVA for 'MP' (Minutes Played)
anova_mp = anova_test(nba_df, 'MP')

# ANOVA for 'TOV' (Turnovers)
anova_tov = anova_test(nba_df, 'TOV')

# ANOVA for 'PF' (Personal Fouls)
anova_pf = anova_test(nba_df, 'PF')

anova_mp, anova_tov, anova_pf

```

```

(F_onewayResult(statistic=16.853359956429035, pvalue=3.349643294843528e-39),
 F_onewayResult(statistic=54.396006603848726, pvalue=3.2348976039969878e-139),
 F_onewayResult(statistic=78.12805428421753, pvalue=1.1803560959601662e-201))

```

Figure 8. One Way ANOVA test with f-statistic value and p-value

The figure 8 shows the code snippet of One Way ANOVA and their results. We have used the **f_oneway()** function from the scipy.stats library to perform the test. We defined a function

which first creates a sample data for each variable on different positions and then the `f_oneway()` function is called to perform the test and the result is returned to the specific variable.

From the results, it is evident that p-value is significantly less than threshold value **0.05** . This indicates that mean values significantly differ across all positions and are not just random.

5.10 Two Way ANOVA:

In order to prove the hypothesis, One Way ANOVA is not enough as it doesn't deal with interaction terms and has a dependent variable. So we went ahead with performing Two Way ANOVA which includes Interaction terms as well.

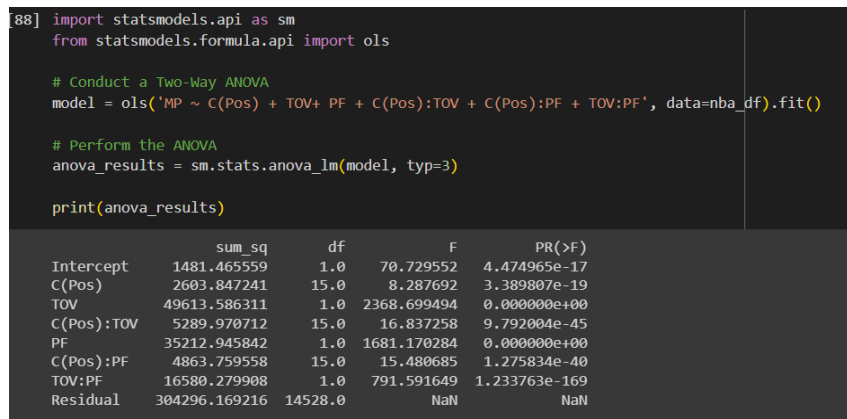


Figure 9. Two Way ANOVA test with f-statistic value and p-value

Two ANOVA Test was implemented using the OLS (Ordinary Least Squares) and `annova_lm()` method from StatsModels library. We used the formula approach in OLS to specify a model that includes both main and interaction effects.

The formula used is '**MP ~ C(Pos) + TOV +PF + C(Pos):TOV + C(POS):PF + TOV:PF**', which states that MP is the dependent variable and after the ~ are independent variables with C(Pos) which depicts that Pos will be treated as Categorical variable , C(Pos) : TOV will be combinational effect of TOV of multiple categories on dependent variable, C(Pos) : PF will be combinational effect of PF of multiple categories on dependent variable and finally TOV:PF will be combinational effect of TOV and PF on dependent variable. After this, the model is fit to the data with the formula. By using the `annova_lm()` method with input parameters model and `typ = 3`, we got the results as mentioned. The `typ` parameter specifies the type of sum of squares to be used and typically can be either 1,2 or 3. Type 1 is generally used when the interaction terms are not included or for One Way Anova test. Type 2 is used when we have interaction terms and when

subgroups have the same number of observations (balanced system) and type 3 is used when we have interaction terms and number of observations differ in subgroups. Since we have done grouping on **Positions** and the number of observations differ for each category, we went ahead with type 3.

From these results, we can conclude that both individual factors (player position, turnovers, and personal fouls) and their interactions are important in predicting the minutes played in NBA games. The significant interaction terms suggest that the role of turnovers and fouls in determining playing time is complex and influenced by the position a player occupies.

5.11 Regression Analysis:

$$\text{MP} = 2.663 - 2.846*\text{POS_C} - 0.580*\text{POS_PF} - 0.429*\text{POS_PG} + 1.821*\text{POS_SF} + 2.034*\text{POS_SG} + 6.949*\text{TOV} + 5.140*\text{PF}$$

After conducting the ANOVA test, we wanted to further dig down the experiment. From the ANOVA test, we concluded that all the variables are statistically significant towards predicting Minutes Played (MP). But, we didn't know to what extent each variable's contribution is. Hence, to quantify the relationship, we performed regression analysis using Ridge regression. The reason behind using ridge regression is the feasibility of adding a penalty term which helps in normalizing the constants and avoids overfitting. We quantified the relationship and came to an equation form ($y = mx+c$). The constant value 2.663 mentions the value of MP when all other parameter values are 0. Turnovers (TOV) is having a positive constant which means that for every single value rise in Turnovers, there will be an increase in the value of MP by 6.949, similarly for every single value rise in Personal Fouls (PF), there will be an increase in the value of MP by 5.140. The same holds with other values as well.

6 Rationale for Data Modelling

After going through the Hypothesis testing and accepting our alternate hypothesis, we moved to the modelling part. We will predict the Minutes Played with Independent variables as TOV, PF and POS. After going through a literature review, we found that most of them did comparative analysis of regression models. Most prominent ones were XGBoost Tree regressor, XGBoost Linear Regression, Random Forest Regressor, SVR and Linear Regression. Hence we will also be implementing them in our research question.

The modelling part was done in two ways. The first way was to fit the data into all the models without hyperparameters, so that model will get trained with hyperparameter values.

In the second way, we used GridSearchCV and RandomSearchCV to get the best hyperparameter values by feeding some predefined set of values. We used GridSearchCV for XGBoost Tree Regressor and XGBoost Linear Ensemble model and RandomSearchCV for Random Forest Regressor and Support Vector Regressor. The reason for using RandomSearchCV over GridSearchCV for RandomForest and Support Vector Regressor is due to computation complexity and the cost of computation is very high.

The metrics used for checking model performance are MAE (Mean Absolute Error) , MSE (Mean Squared Error) and R2 Score.

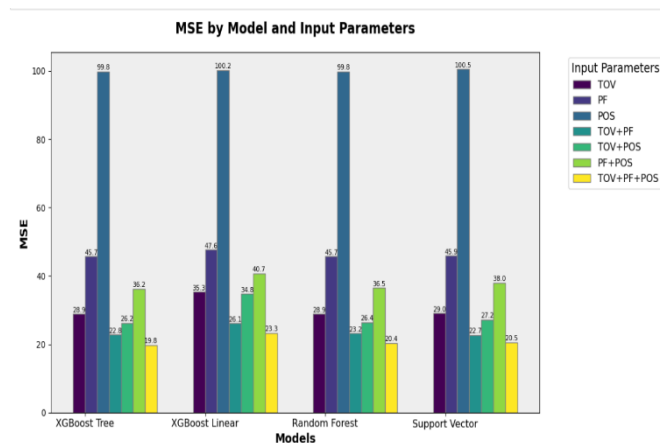


Figure 12. MSE by model.

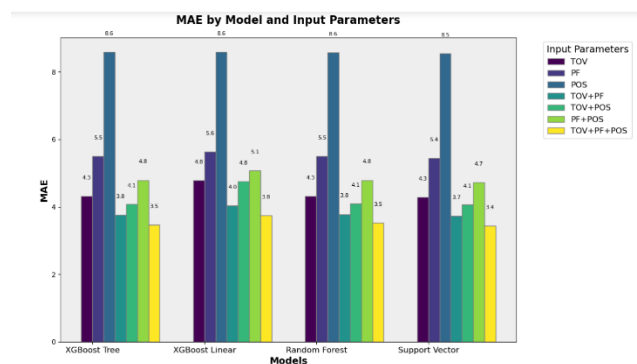


Figure 13. Mean absolute error and input parameters

INPUT PARAMETERS / MODELS	XGBOOST TREE	XGBOOST LINEAR	RANDOM FOREST	SUPPORT VECTOR REGRESSOR
TOV	0.70	0.65	0.70	0.70
PF	0.53	0.51	0.53	0.53
POS	0.014	0.012	0.014	0.013
TOV+PF	0.77	0.74	0.76	0.77
TOV+POS	0.73	0.65	0.72	0.72
PF+POS	0.63	0.59	0.63	0.62
TOV+PF+POS	0.80	0.77	0.79	0.80

Table 2. R2-Score for each model and its related input parameters.

To prove that all input parameters are required to predict the Minutes Played (MP) value, we did isolated modelling where the input parameters were either given as a single parameter or as a combinational parameter. From the above charts and table, it is evident that MSE and MAE is least for input parameter **TOV+PF+POS** and highest for input parameter **POS**, which means that the variation in Minutes Played (MP) value predicted value and true value will be less for input parameter **TOV+PF+POS** and the variation will be high when only **POS** is used as input parameter. The R2 Score for input parameter **TOV+PF+POS** is the highest with values close to 0.80 for all models which indicates that 80% of the variance in target variable (MP) is predictable from **TOV+PF+POS** as a whole and explains a larger portion of the variability in the outcome variable as compared to the R2-Score of input variable **POS** alone which has a score of only 0.014 which means only 1.4 % of variance in the target variable (MP) is predictable from **POS**.

7 Results

The analysis conducted in this study applied various machine learning models to assess the impact of turnovers (TOV), personal fouls (PF), and player positions (POS) on the minutes played

(MP) from RQ-1 and impact of Assists (AST), Point Scored (PTS), Total Rebounds (TRB) and Player Efficiency Rating (PER) by players in NBA basketball.

7.1 Statistical Significance of Variables

The combination of TOV, PF, and POS yielded the highest R2 score across models in RQ-1 indicating robust predictability and in similar fashion, the combination of AST, PTS, TRB and PER yielded the highest R2 Score across all 4 models in RQ-2. These scores suggest that TOV, PF and POS when combined explain approximately 80% of the variability in minutes played and AST, PTS, TRB and PER when combined explain approximately 92% of variability in minutes played.

7.2 Impact of Individual and Combined Variables

Turnovers (TOV) alone had R2 scores ranging from 0.65 to 0.70, showing a substantial but incomplete predictive power in RQ-1. Points Scored (PTS), Total Rebounds (TRB) and Player Efficiency Rating (PER) had similar R2 scores ranging from 0.80 to 0.87 in RQ-2.

Player Position (POS) alone: Demonstrated the least predictive power with an R2 score around 0.014, indicating a very minor standalone impact on minutes played.

7.3 Combination Effects:

Combined variable's R2 scores ranged from 0.59 to 0.80, with PF+POS showing the least (0.59) and TOV+PF+POS showing the highest (0.80) for RQ-1. Testing on multiple combinations of input variables for RQ-2 were not done due to multiple variables taken into account, but the combination of all variables showed a significant R2-Score of 0.93.

- **Turnovers (TOV) alone:** R2 scores ranged from 0.65 to 0.70, showing a substantial but incomplete predictive power.
- **Personal Fouls (PF) alone:** R2 scores were slightly lower, ranging from 0.51 to 0.53.
- **Player Position (POS) alone:** Demonstrated the least predictive power with an R2 score around 0.014, indicating a very minor standalone impact on minutes played.

9.3 Combination Effects:

- **TOV + PF:** Combined R2 scores ranged from 0.74 to 0.77, suggesting that these two variables together provide a stronger basis for predicting minutes played.
- **TOV + POS:** Showed improved prediction with R2 scores from 0.65 to 0.73.

- **PF + POS:** Similarly improved, with R2 scores from 0.59 to 0.63.

10. Discussion of Interpretation of Results: How Has the Question Been Answered?

The research question aimed to explore how turnovers and personal fouls per game affect the minutes played, conditioned on the player's position. Our findings from the models provide a substantive answer:

10.1 Turnovers (TOV) and Personal Fouls (PF):

Both metrics significantly impact playing time, with turnovers showing slightly stronger predictive power. High occurrences of turnovers and fouls lead to reduced playing time as they negatively affect team performance. This aligns with traditional basketball strategies that emphasize possession retention and disciplined play.

10.2 Role of Player Positions (POS):

The position factor alone showed minimal direct impact ($R^2 = 0.014$), suggesting that while position influences playing time, it does not strongly dictate it without the interaction of other factors like TOV and PF.

10.3 Comprehensive Data Analysis: By employing advanced statistical and machine learning techniques, our research has confirmed the importance of these metrics and provided a predictive framework for player utilization. The highest R2 score of 0.80 for the combination of TOV, PF, and POS indicates that these three factors together provide a comprehensive model for predicting minutes played.

8 Conclusion

With the help of statistical analysis, we were able to prove that all the variables are statistically significant. We also performed isolated modeling to check whether either one of the input variables is enough to predict or all whether all parameters are required. We concluded that in all the evaluation metric values (MAE, MSE, R2-Score) used, the combination of all variables gave the least errors and best scores evident from the bar charts and tabular results. Thus with the already available ML models and by fine tuning them, we were able to predict the Minutes Played (MP) by each player for a single season based on two different sets of input parameters with first set being "Turnovers (TOV), Personal Fouls per game (PF) and Positions (POS)" and second set "Assists (AST), Points Scored (PTS), Total Rebounds (TRB), Player Efficiency Ratings (PER) and

Positions (POS)" , with second set of parameters predicting Minutes Played (MP) much closer to the true value due to different variables involved.

9 Group Work

At the start of the project, the group decided to approach the project in a way that each team member would interact with all the aspects of the project, while taking on more responsibilities in each preferred area. Broadly, each team member was responsible for the analysis. A more detailed report of individual contributions is listed as following:

10 Contribution by the each member

Here is our division of task for this group project:

- **Arudhra Venkatachalam**

Data retrieving and data transformation, Data refinement and further exploration. Data visualization, Data Modelling, Report writing for EDA.

- **Surabhi Borase**

Data modelling, Creation of PPT, Rationale for Data Modelling/- Experimentation, Results leading to answering the question, and Discussion of interpretation of results, Conclusion and Abstract.

- **Ali Ammar**

Exploration of problem statement, Poster Creation of Project, Writing the report sections: Abstract, Introduction, Background Research. Question Development, Results and Discussion of the results, Summary, and References. Report formatting and compilation of all the work is also done by Ali Ammar

- **Mujazi Kekepuram and Sai Sathyanarayanan:** Data cleaning and additional research work

References

1. Nistala, A., & Guttag, J. (2019, March). Using deep learning to understand patterns of player movement in the NBA. In *Proceedings of the MIT Sloan Sports Analytics Conference* (pp. 1-14).
2. Vinué Visús, G., & Epifanio, I. (2017). Archetypoid analysis for sports analytics.
3. Wang, K. C., & Zemel, R. (2016, March). Classifying NBA offensive plays using neural networks. In *Proceedings of MIT Sloan sports analytics conference* (Vol. 4).
4. Sill, J. (2010, March). Improved NBA adjusted+/-using regularization and out-of-sample testing. In *Proceedings of the 2010 MIT Sloan sports analytics conference*.
5. Piette, J., Pham, L., & Anand, S. (2011, March). Evaluating basketball player performance via statistical network modeling. In *The 5th MIT Sloan sports analytics conference* (pp. 4-5).
6. Sarlis, V., & Tjortjis, C. (2020). Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, 93, 101562.
7. Oskan, C., & Onay, C. (2022). Predicting the winning team in basketball: A novel approach. *Heliyon*, 8(12).
8. Ke, Y., Bian, R., & Chandra, R. (2024). A unified machine learning framework for basketball team roster construction: NBA and WNBA. *Applied Soft Computing*, 153, 111298.
9. Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *Journal of quantitative analysis in sports*, 3(3).
10. Tyshanti Montgomery, Jason Lin, Hongyou Chen, and Hao Cui (2021). NBA Data Visualization,
https://www.stat.cmu.edu/capstoneresearch/fall2022/315files_f22/team17.html

11 Appendix

- Notebook link for Research Question-1

Code Link :

https://colab.research.google.com/drive/1EBj_ELuTWZaBOI8abzcaPtk61Xi3XHiA?usp=sharing

Modeling :

https://colab.research.google.com/drive/1NOtK7nofEy0JFtd3WXnAax60Rrr_Xce?usp=sharing

- Notebook link for Research Question 2 :

Code Link :

<https://colab.research.google.com/drive/1GjZPNjHDSbosLj-qUUFgs88nJwuKSQib?usp=sharing>

Modeling :

https://colab.research.google.com/drive/1EBj_ELuTWZaBOI8abzcaPtk61Xi3XHiA?usp=sharing