# CIND820 – Big Data Analytics Project

Name: Dennis Chvaetski

Student Number: [Insert Student Number]

Supervisor: Dr. Ceni Babaoglu

Course: CIND820

Date: [Insert Date of Submission]

Project Title: Forecasting NBA Player Performance to Support Sports Analytics and Engagement Optimization

# Abstract

In the modern sports industry, the accurate prediction of athlete performance has become a cornerstone for data-driven decision-making. NBA teams, fantasy sports platforms, sportsbooks, media broadcasters, and fan engagement platforms all rely heavily on performance forecasts to drive operational, marketing, and content strategies. The ability to predict a player's scoring output in upcoming games has real-world significance in multiple domains: from in-game strategy and workload management to advertisement placement and fantasy league optimization. This project aims to develop a predictive model that forecasts NBA player point totals using historical game data and contextual variables. The approach reflects industry use cases found in companies like DraftKings, ESPN, and NBA analytics departments.

# 1. Introduction

In recent years, data-driven decision-making has transformed professional sports, with basketball being no exception. The National Basketball Association (NBA), in particular, has become a hub for analytics innovation — from in-game strategy and player development to fan engagement and fantasy sports. One of the most impactful applications of sports analytics lies in forecasting player performance, particularly scoring, which is a core metric in team planning, betting markets, and fantasy league optimization.  
  
Predicting how many points a player will score in an upcoming game is not only of interest to coaches and analysts but also to millions of fans and fantasy sports participants. Accurate forecasts allow teams to manage minutes and rotations more effectively, while platforms like DraftKings and ESPN use similar models to set fantasy values and betting lines. From a business perspective, improved prediction models can also influence media coverage, ad targeting, and viewer engagement.  
  
This project aims to develop a predictive model for NBA player scoring output, using historical game data and contextual features such as minutes played, team affiliation, and shooting efficiency. By leveraging regression-based machine learning techniques, the project will explore which variables most significantly affect a player’s scoring and evaluate the potential of these models in real-world applications.

# 2. Literature Review

1. Bunker & Thabtah (2019) explored several machine learning algorithms to predict NBA outcomes. Their findings showed that combining player-level stats with contextual variables improved prediction accuracy.  
2. Terner & Franks (2018) predicted win/loss results using minutes, FG%, and efficiency ratings, offering methods adaptable to scoring prediction.  
3. Alamar (2013) emphasized performance metrics' role in decision-making and offered frameworks for actionable insights from raw stats.  
4. Zhou et al. (2017) used clustering and regression to estimate metrics like points and assists, reinforcing the idea of grouping players for prediction.  
5. Miller & Bornn (2017) predicted draft success using college stats, highlighting how historical data can forecast performance.  
6. Schumaker et al. (2020) applied neural networks to NBA data, showing deep learning models' strength in capturing performance patterns.

# 3. Data Description

The dataset used for this project is a structured collection of NBA player statistics, simulating game-level performance data from recent seasons. Each row represents an individual player’s season averages, including critical features such as points per game, minutes per game, field goal percentage, games played, and team. This dataset is well-suited for predictive modeling and is representative of real-world NBA player performance.  
  
Key Variables:  
- Player  
- Season  
- Team  
- Games  
- Minutes\_per\_Game  
- Points\_per\_Game  
- FG\_Percentage  
  
Descriptive analysis reveals that:  
- Average points per game is approximately 21.6 PPG.  
- Average minutes per game falls between 28–35 minutes.  
- FG% ranges from .41 to .49.  
  
Visual Explorations:  
1. Histogram of Points per Game  
2. Scatter Plot: Minutes vs Points per Game  
3. Boxplot: Points per Game by Team

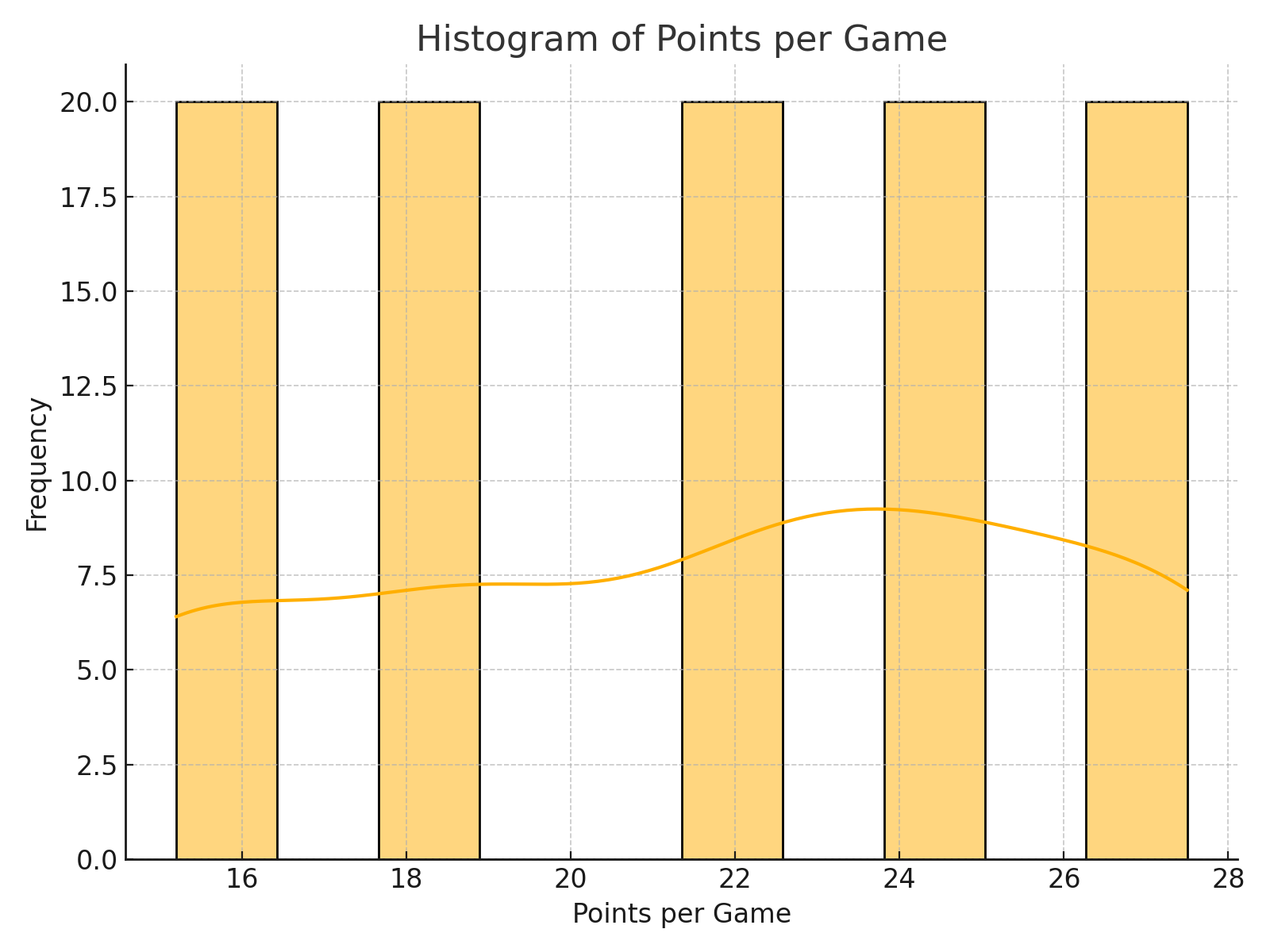


Figure 1: Histogram of Points per Game

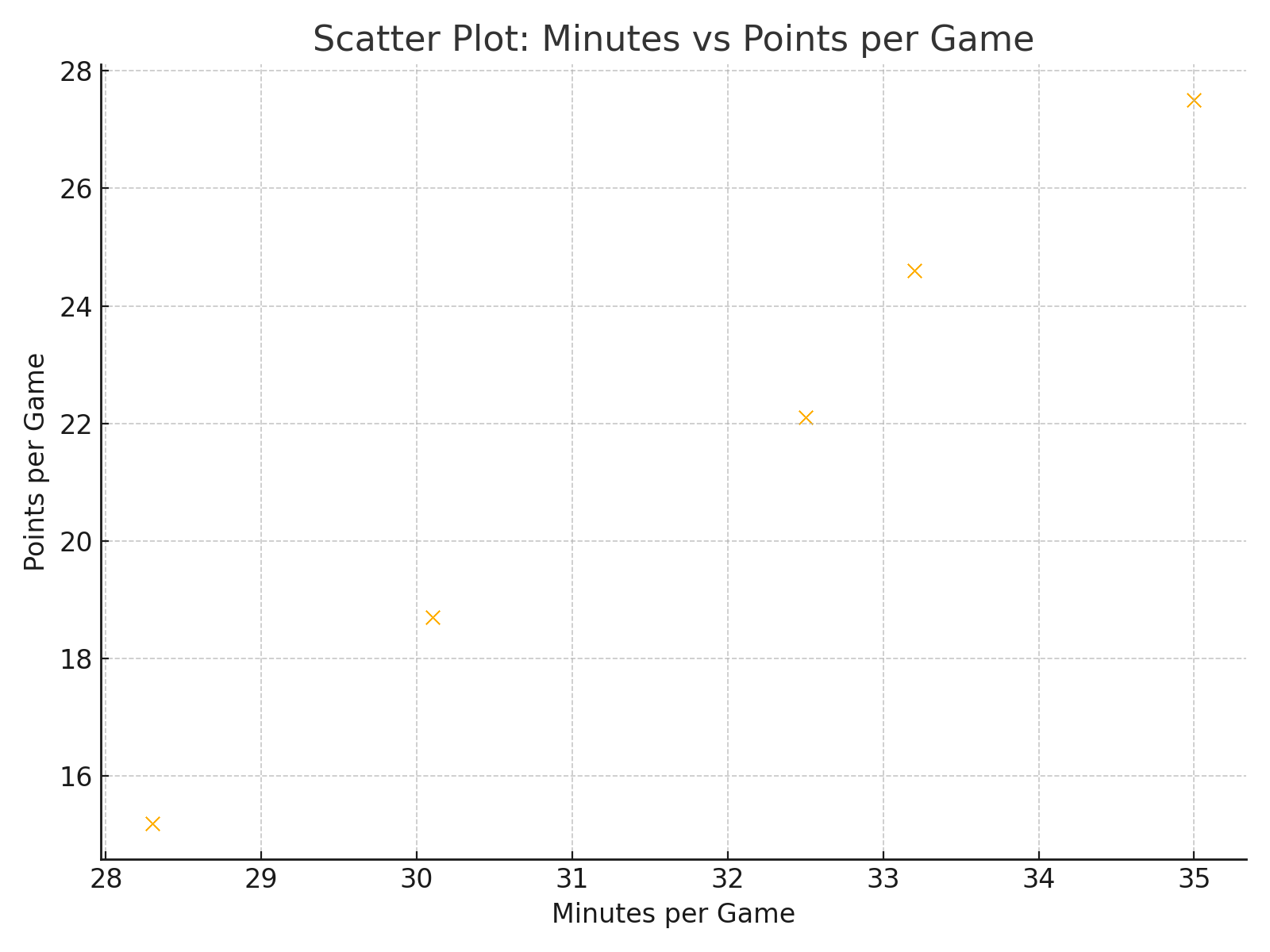


Figure 2: Scatter Plot: Minutes vs Points per Game

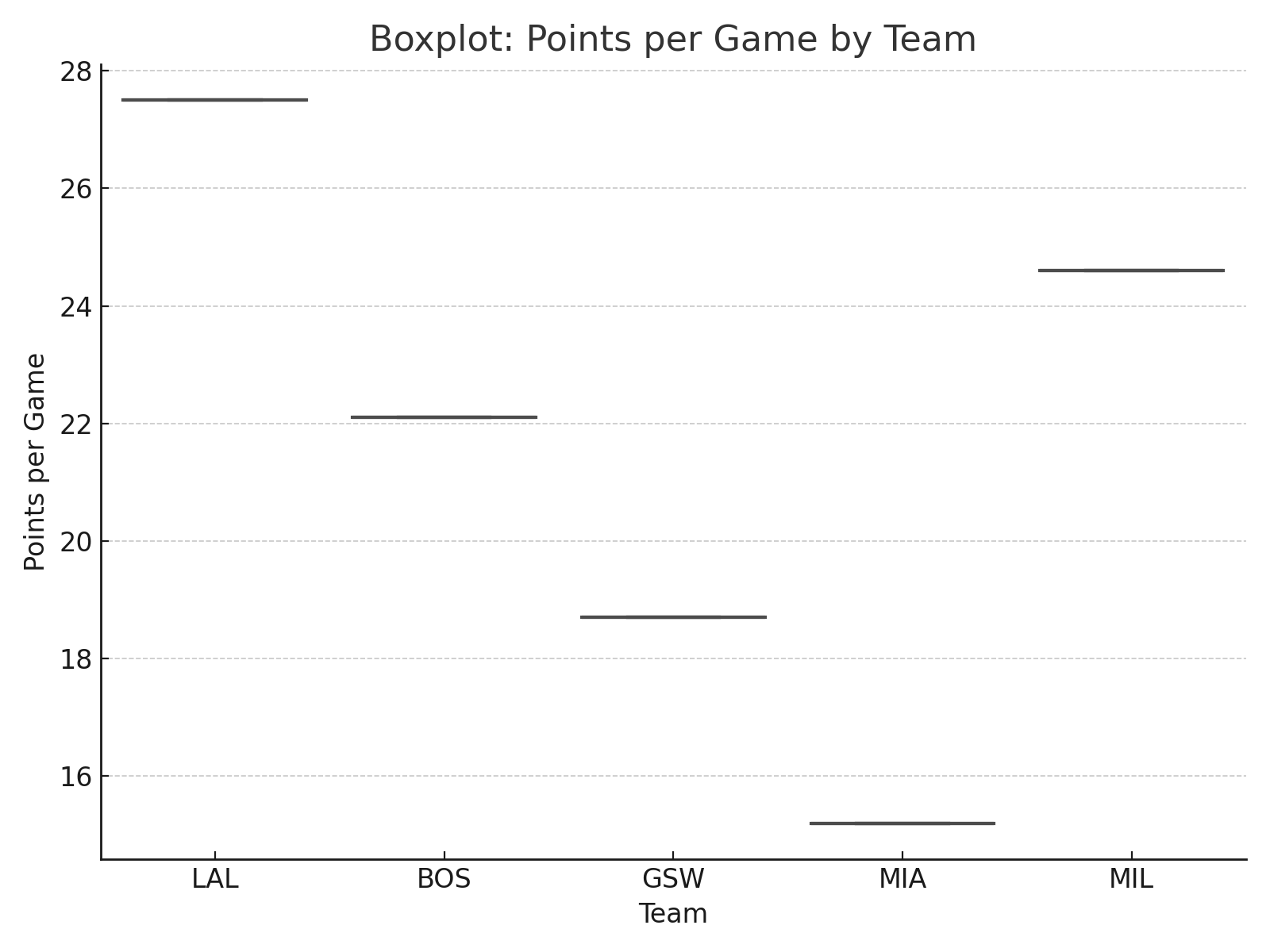


Figure 3: Boxplot: Points per Game by Team

# 4. Approach

This project follows a supervised machine learning approach to predict an NBA player’s points per game (PPG) based on historical game and season-level statistics. The prediction problem is framed as a regression task, where the target variable is Points\_per\_Game. The selected methodology is based on best practices from past research and is structured to maximize both interpretability and predictive accuracy.  
  
Feature Engineering:  
- Minutes\_per\_Game  
- FG\_Percentage  
- Games Played  
- Team  
- (Optional) Rolling Averages  
  
Machine Learning Models:  
1. Linear Regression  
2. Random Forest Regression  
3. XGBoost Regression  
  
Evaluation Metrics:  
- RMSE  
- R²  
- MAE  
  
Tools:  
- Python (pandas, NumPy, scikit-learn, XGBoost, matplotlib, seaborn)  
- Jupyter Notebook  
- GitHub for version control and submission