**EXPLORATORY DATA ANALYSIS ON PLAYER DATASET**

Sirichandana Bomma

Department of Information Science

University of North Texas Denton, Texas, USA

Email: SirichandanaBomma@my.unt.edu

I’ve chosen to take player data for this research project.

**Introduction:**

Basketball is a quick-paced, thrilling sport that has millions of admirers all around the world. The National Basketball Association (NBA) is one of the most well-known professional basketball associations. Some of the world's most gifted athletes compete in the NBA, and their statistics reflect their outstanding on-court performance. In this research article, we examine 492 NBA players' 2014–2015 season statistics. I investigate several facets of these statistics, such as the data distribution, the relationships between various variables, and the performance of the players in various game scenarios.

I’m planning to analyse the data and I’ve chosen to do analysis based on the following question. What factors contribute to player performance in a sports league?

**Dataset Description:**

The dataset contains statistics for 492 NBA players during the 2014-2015 season. The variables in the dataset are Name, Age, Team, Games played, Rebounds, Assists, Steals, Blocks, Turnovers, and Points. The data was collected from multiple sources and then combined to form this dataset.

**Methodology:**

1. **Exploratory Data Analysis (EDA):**
2. **Data Collection:**

I haven’t done any further collection of data because the dataset is already available in canvas (given by professor).

1. **Data Cleaning and Data Pre-processing:**

First, I’ve checked for duplicates, but there are no duplicates found in the dataset. Then, I’ve checked for values, and I did not find any null values. The data is already cleaned.

1. **Data Analysis:**

I’m following the below steps of analysis.

1. **Distribution of Data:**

I begin by analyzing the distribution of the data. The distribution of each variable is an essential aspect of data analysis because it provides insights into the data's central tendency and variability. I’ll use histograms and box plots to visualize the distribution of the data.

1. **Correlation Analysis:**

Next, I analyze the correlation between different variables in the dataset. Correlation analysis is a useful tool for identifying the relationships between different variables and their strength. I’ll use scatter plots and correlation coefficients to analyze the correlation between the variables in the dataset.

1. **Performance Analysis:**

Finally, I analyze the players' performance in different game scenarios. I’ll compare the players' statistics in games won and lost, as well as home and away games. This analysis can provide insights into the players' performance under different conditions.

**Results:**

The following is the output for Distribution of Data.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Fig 1: Distribution of Data

We can see from the correlation matrix that the variables "Games," "Rebounds," "Assists," and "Steals" have a substantial positive association with the variable "Points." In contrast to the variable "Assists," which has a somewhat positive correlation with the variable "Turnovers," the variable "Rebounds" has a large positive correlation with the variables "Blocks" and "Turnovers".

Following is the result for Correlation Analysis.

Chart, treemap chart

Description automatically generated

Fig 2: Correlation Analysis

The heatmap shown above reveals a strong link between the columns "Rebounds" and "Points" (correlation coefficient of 0.87). This shows that a player's point and rebound totals are reliable measures of their effectiveness.

Additionally, we can see that the columns "Games" and "Rebounds" and "Assists" have moderately positive correlations with each other (correlation values of 0.44 and 0.48, respectively). This implies that a player's performance may also be influenced by the number of games they play.

Chart, scatter chart

Description automatically generated

Fig 1: Performance Analysis

**Conclusion:**

By examining the dataset's distribution of data, correlations between various variables, and player performance in various gaming scenarios, I have done an exploratory data analysis (EDA). The goal of the analysis is to pinpoint the elements that influence player performance in a certain sports league. To visualize and evaluate the data, the author will make use of histograms, box plots, scatter plots, and correlation coefficients. The results of this study will offer insightful information about NBA players' performances, which will be helpful for basketball fans, team managers, and analysts.

**References:**

National Basketball Association (NBA) - About: https://www.nba.com/about

Basketball-Reference: NBA Stats and History: https://www.basketball-reference.com/

Dataset source: Provided by professor via Canvas Learning Management System.

Wickham, H., & Grolemund, G. (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, Inc.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 56-61.