**Universitat Politécnica de Catalunya**

**Multivariate Analysis**

**Final Project**

**Professor:**

Acar Denizli Nihan

**Group members:**

William Alejandro Cueva

Nashly Erielis González

**Group 11 and 12**

**Interfaz de usuario gráfica, Texto

Descripción generada automáticamenteTexto

Descripción generada automáticamenteDecember 23, 2024**

Index

[1. Introduction 3](#_Toc185614139)

[2. Exploratory Data Analysis 3](#_Toc185614140)

[3. Hotelling T 2 Test 4](#_Toc185614141)

[4. MANOVA 5](#_Toc185614142)

[5. Principal Component Analysis 5](#_Toc185614143)

[7. Multidimensional Scaling 5](#_Toc185614144)

[9. Multiple Correspondence Analysis 5](#_Toc185614145)

[10. Cluster Analysis 6](#_Toc185614146)

[11. Discriminant Analysis 6](#_Toc185614147)

[12. Conclusions: 6](#_Toc185614148)

1. **Introduction**

The present work aims to perform a multivariate analysis of the performance of outstanding NBA players. This analysis explores general patterns, identifies outliers, and segments players into groups based on their performance and physical characteristics. Additionally, the association between different metrics is evaluated and performances are compared between teams and universities. The variables of special interest are the performance metrics in the game of basketball. For the present work, these variables correspond to the number of games played throughout the season (gp), average number of points scored per game (pts), average number of rebounds per game (reb), average number of assists per game (ast), percentage of available offensive rebounds that the player captured while on the court (oreb\_pct), percentage of available defensive rebounds that the player captured while on the court (dreb\_pct), a measure of the player's shooting efficiency that takes into account free throws, 2- and 3-point shots (ts\_pct), and percentage of field goals from teammates assisted by the player while on the court (ast\_pct). On the other hand, physical variables such as age, height, and weight are available. Finally, categorical variables such as university, school, year of contract, round in which the player plays, and jersey number are also available.

1. **Exploratory Data Analysis**
   1. Data Description

We used a data set that contains general information about the top NBA players (college, team), and their physical characteristics and performance (height, weight, points, rebounds, assists, and TS%).

* 1. Distributions

The distribution of the variables we consider most important is presented below, while the remaining variables have been included in the appendix.

* 1. Missing Values and Imputation

The metric variables did not present missing values, so it was not necessary to carry out the imputation process. The variables year of call-up to the team, jersey number and game round presented missing values. However, they were not included in the study because they do not present any relationship with the other variables.

* 1. Correlations:
  2. Multivariate Outliers

The aim of this study is to discover whether there are players within the data set with the most outstanding performance metrics. To do this, multivariate outliers will be identified. The Mahalanobis distance metric was used to identify those players with multivariate atypical characteristics. According to the R report, 236 players are identified with characteristics different from the others. These make up 5.54% of the dataset.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figure

Next, the maximum values of the players' performance metrics are compared to highlight some of the features of this group of players with atypical values. Class “0” refers to players not considered as atypical, while class “1” refers to players with atypical characteristics. These players have a maximum value lower than the maximum games played of “normal” players. To belong to the group of special players, they must stand out in at least the highest points obtained, the highest number of rebounds, the highest number of assists, and the highest percentage of offensive and defensive rebounds.

|  |  |  |
| --- | --- | --- |
| **Cluster** | **0** | **1** |
| Gp | 82 | 80 |
| Pts | 28.1 | 28.7 |
| Reb | 12.4 | 14.4 |
| Ast | 9.7 | 11.1 |
| oreb\_pct | 0.183 | 0.5 |
| dreb\_pct | 0.303 | 0.5 |

Table

1. **Hotelling T 2 Test**

Through this objective, we intend to determine whether universities with a greater number of outstanding players present a difference between performance metrics. To do this, the T2 Hotelling test was applied. Kentucky and Duke were identified as the universities with the most standout players.

|  |  |  |
| --- | --- | --- |
| **None** | **Kentucky** | **Duke** |
| 87 | 21 | 19 |

Table

To apply Hotelling's T2 test, we proceeded to verify whether the performance metrics data fit a normal distribution, where the following hypotheses are contrasted:

* ***H0:*** The data of the i-th variable of the j-th university fits a Normal distribution.
* ***H1:*** The data of the i-th variable of the j-th university does NOT fit a Normal distribution.

The following table shows the p-values for the Kolmogorov test by university. It can be observed that the p-values for the game performance metrics are greater than 0.05, that is, the variables fit the Normal distribution.

|  |  |  |
| --- | --- | --- |
|  | **Kentucky** | **Duke** |
| **Variable** | **Pvalue** | **Pvalue** |
| gp | 0.3992 | 0.478 |
| pts | 0.6101 | 0.4093 |
| reb | 0.8768 | 0.4347 |
| ast | 0.05158 | 0.3965 |
| oreb\_pct | 0.4932 | 0.1451 |
| dreb\_pct | 0.8059 | 0.4920 |
| ts\_pct | 0.3673 | 0.6856 |
| ast\_pct | 0.1524 | 0.702 |

Table

With this result, the hypothesis test of the vector of means of performance metrics for both groups is applied. The Hotelling d test is not significant, that is, no differences are found in the vector of means of the performance metrics of the players at the universities of Kentucky and Duke.

1. **MANOVA**
2. **Principal Component Analysis**
3. ~~Factor Analysis~~
4. **Multidimensional Scaling**
5. ~~Correspondence Analysis~~
6. **Multiple Correspondence Analysis**
7. **Cluster Analysis**
8. **Discriminant Analysis**
9. **Conclusions:**

From the analysis of NBA data it is concluded that:

* The variables associated with the age of the players have a positive impact on the team, while the variables of percentage of offensive rebounds and passes to the player are related in a negative way. For both height and weight, these variables are positively related to percentage of offensive and defensive rebounds, rebounds, while they are negatively related to number of assists and passes to the player.
* The characteristics that generate players with metrics with atypical values are those players who have a maximum of points obtained, number of rebounds, number of assists, percentage of offensive or defensive rebounds.
* There are no significant differences in performance metrics between Kentucky and Duke, which have a higher number of standout players.
* There are no significant differences in performance metrics between DAL, CHA, MIN, NYK teams, which feature a higher number of featured players.
* Performance metrics, as well as physical variables such as age, height and weight, are explained by 75.8% by the dimensions of Size and Rebounding, Production in Play, Efficiency and Offensive Use and Experience.
* Two segments are generated in which players can be grouped as: consistent players and support and reserve players.
* An association is found between high levels of points obtained with low levels of offensive rebounds, low levels of defensive rebounds and regular levels of assists, which reflects that the points obtained by the players are due to collective work and not to the individual talent of the players.
* The transformed variables for points scored, rebounds and assists influence the player's impact on the team by 62.9%.

1. Bibliography:

**Appendix**

1. Exploratory Data Analysis
   1. Distributions
   2. Missing Values and Imputation
   3. Multivariate Outliers
2. Hotelling T 2 Test
3. MANOVA
4. Principal Component Analysis
5. Factor Analysis
6. Multidimensional Scaling
7. ~~Correspondence Analysis~~
8. Multiple Correspondence Analysis
9. Cluster Analysis
10. Discriminant Analysis
11. Conclusions:
12. Bibliography: