**All\_Seasons**

**Exploratory Analysis**

William Muckler, wmuckler@bellarmine.edu

Sean Hand, shand@bellarmine.edu

1. **INTRODUCTION**

Short description of the data set including a reference to where it can be found and why you chose it.

Our data set was found in Kaggle. We chose this specific set because of the large amount of data and number of variables. We believed this would give us lots of information to work with and compare to each other. We are also both interested in the basketball statistics we were able to pull from this data set. Data includes several variables of NBA players from 1996-2022. This data includes categorical and numerical data.

Link: https://www.kaggle.com/code/justinas/nba-height-and-weight-analysis

1. **DATA SET DESCRIPTION**

Narrative summary of the data set: e.g. this data set contains 398 samples with 7 columns with various data types. A complete listing is shown in **Table 1**. For data types you want to indicate two things (nominal, ordinal, interval, or ratio) and the Pandas data type. For example, age might be ratio/int32. For missing data, indicate what percentage of data from that column are missing. Ensure you check to for NaN, NA, or any other indicators that actually mean missing data.

Our data set contains 11,145 samples with 22 columns of variables. The complete list of columns and data type is found in our table 1. This was a large set with many variables. The only column we found to have missing data was college. This made sense and we attributed this to players who took an international route or played in the NBA directly after high school. Although there were many complete entries, we ran into a problem while working through the data. Although our data was complete, except the college column, we realized that these were sampled from each players career. This meant that we would not be able to find statistics on any one player but we would have to rely on single year or more general statistics.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| Variable Name | Data Type | Missing Data (%) |
| Index | Int64 | 0% |
| player\_name | Object | 0% |
| Team\_abbreviation | Object | 0% |
| Age | Float64 | 0% |
| Player\_height | Float64 | 0% |
| Player\_weight | Float64 | 0% |
| College | Object | 15.11% |
| Country | Object | 0% |
| Draft\_year | Object | 0% |
| Draft\_round | Object | 0% |
| Draft\_number | Object | 0% |
| Gp | Int64 | 0% |
| Pts | Float64 | 0% |
| Reb | Float64 | 0% |
| Ast | Float64 | 0% |
| Net\_rating | Float64 | 0% |
| Oreb\_pct | Float64 | 0% |
| Dreb\_pct | Float64 | 0% |
| Usg\_pct | Float64 | 0% |
| Ts\_pct | Float64 | 0% |
| Ast\_pct | Float64 | 0% |
| season | object | 0% |

1. **Data Set Summary Statistics**

There were 14 variables that were continuous and 8 that were objects. We are able to retrieve the summary statistics for continuous variables. When discussing the object variables, we listed the top 10 for each variable. If we were to list every possibility, some variable tables would be hundreds or thousands of entries. Name is an example of this where we don’t need the full list, but the top 10 may be helpful in understanding our data.

**Table 2: Summary Statistics for all\_seasons.csv (name of dataset)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable Name | Count | Mean | Standard Deviation | Min | 25th | 50th | 75th | Max |
| Unnamed | 11145 | 5572 | 3217.45 | 0 | 2786 | 5572 | 8358 | 11144 |
| Age | 11145 | 27.17 | 4.34 | 18 | 24 | 27 | 30 | 44 |
| Player\_height | 11145 | 200.81 | 9.19 | 160 | 195.58 | 200.66 | 208.28 | 231.14 |
| Player\_weight | 11145 | 100.64 | 12.58 | 60.32 | 90.71 | 99.79 | 109.32 | 163.29 |
| Gp | 11145 | 52 | 25.07 | 1 | 32 | 58 | 74 | 85 |
| Pts | 11145 | 8.13 | 5.94 | 0 | 3.5 | 6.6 | 11.5 | 36.1 |
| Reb | 11145 | 3.56 | 2.49 | 0 | 1.8 | 3 | 4.7 | 16.3 |
| Ast | 11145 | 1.8 | 1.8 | 0 | .6 | 1.2 | 2.4 | 11.7 |
| Net\_rating | 11145 | -2.15 | 12.15 | -200 | -6.3 | -1.3 | 3.2 | 300 |
| Oreb\_pct | 11145 | .0556 | .0439 | 0 | .022 | .043 | .086 | 1 |
| Dreb\_pct | 11145 | .1418 | .0631 | 0 | .096 | .132 | .182 | 1 |
| Usg\_pct | 11145 | .1856 | .053 | 0 | .15 | .182 | .218 | 1 |
| Ts\_pct | 11145 | .508 | .0989 | 0 | .478 | .521 | .557 | 1.5 |
| Ast\_pct | 11145 | .131 | .095 | 0 | .065 | .102 | .178 | 1 |

There should be a table for **EACH** categorical variable.

Table 3: Proportions for team\_abbreviation (n=yyy)

|  |  |  |
| --- | --- | --- |
| Category | Frequency | Proportion (%) |
| TOR | 390 | 3.5% |
| CLE | 390 | 3.5% |
| LAC | 389 | 3.49% |
| MIA | 387 | 3.47% |
| DAL | 384 | 3.45% |
| ATL | 383 | 3.44% |
| PHI | 380 | 3.41% |
| WAS | 379 | 3.4% |
| HOU | 378 | 3.39% |
| SAS | 377 | 3.38% |

Table 4: Proportions for college

|  |  |  |
| --- | --- | --- |
| Category | Frequency | Proportion (%) |
| unknown | 1684 | 15.12% |
| Kentucky | 360 | 3.23% |
| Duke | 331 | 2.97% |
| North Carolina | 318 | 2.85% |
| UCLA | 280 | 2.51% |
| Arizona | 257 | 2.31% |
| Kansas | 251 | 2.25% |
| Connecticut | 220 | 1.97% |
| Georgia Tech | 180 | 1.62% |
| Florida | 173 | 1.55% |

Table 5: Proportions for country

|  |  |  |
| --- | --- | --- |
| Category | Frequency | Proportion (%) |
| USA | 9410 | 84.43% |
| France | 153 | 1.37% |
| Canada | 140 | 1.26% |
| Spain | 79 | .71% |
| Brazil | 78 | .7% |
| Australia | 74 | .64% |
| Slovenia | 67 | .6% |
| Turkey | 63 | .57% |
| Croatia | 62 | .56% |
| Argentina | 60 | .54% |

Table 6: Proportions for season

|  |  |  |
| --- | --- | --- |
| Category | Frequency | Proportion |
| 2017-18 | 540 | 4.85% |
| 2018-19 | 530 | 4.76% |
| 2019-20 | 514 | 4.61% |
| 2014-15 | 492 | 4.41% |
| 2016-17 | 486 | 4.36% |
| 2013-14 | 482 | 4.32% |
| 2011-12 | 478 | 4.29% |
| 2015-16 | 476 | 4.27% |
| 2012-13 | 469 | 4.21% |
| 2004-05 | 464 | 4.16% |

Table 7: Proportions for name

|  |  |  |
| --- | --- | --- |
| Category | Frequency | Proportion |
| Vince Carter | 22 | .02% |
| Dirk Nowitzki | 21 | .02% |
| Kevin Garnett | 20 | .02% |
| Kobe Bryant | 20 | .02% |
| Tyson Chandler | 19 | .02% |
| Jason Terry | 19 | .02% |
| Paul Pierce | 19 | .02% |
| Jamal Crawford | 19 | .02% |
| Tim Duncan | 19 | .02% |
| Jermaine O’Neal | 18 | .02% |

Table 8: Proportions for draft\_year

|  |  |  |
| --- | --- | --- |
| Category | Frequency | Proportion |
| Undrafted | 1942 | 17.42% |
| 1998 | 454 | 4.07% |
| 2003 | 430 | 3.86% |
| 2005 | 420 | 3.77% |
| 1996 | 406 | 3.64% |
| 2001 | 403 | 3.62% |
| 2008 | 381 | 3.42% |
| 1999 | 366 | 3.28% |
| 2000 | 364 | 3.27% |
| 2004 | 362 | 3.25% |

Table 9: Proportions for draft\_round

|  |  |  |
| --- | --- | --- |
| Category | Frequency | Proportion |
| 1 | 6513 | 58.44% |
| 2 | 2629 | 23.59% |
| Undrafted | 1959 | 17.58% |
| 3 | 20 | .18% |
| 4 | 12 | .11% |
| 6 | 5 | .04% |
| 7 | 5 | .04% |
| 8 | 2 | .02% |

Table 10: Proportions for draft\_number

|  |  |  |
| --- | --- | --- |
| Category | Frequency | Proportion |
| Undrafted | 1959 | 17.58% |
| 1 | 320 | 2.87% |
| 5 | 320 | 2.87% |
| 4 | 311 | 2.79% |
| 3 | 299 | 2.68% |
| 2 | 299 | 2.68% |
| 9 | 282 | 2.53% |
| 10 | 278 | 2.49% |
| 7 | 272 | 2.44% |
| 8 | 258 | 2.31% |

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn’t make sense)

A screenshot of a graph

Description automatically generated

1. **DATA SET GRAPHICAL EXPLORATION**

Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

The most interesting distribution we found was between height and rebounds. While there definitely was a trend that increased the number of rebounds as height increased, it wasn’t as cut and dry as we thought it would be initially. The was a lot of outliers that showed those with shorter height having more rebounds than someone who was taller than then. We chalked this up to being a difference in athleticism and players who are taller tend to be less athletic than their shorter counterparts meaning they perhaps had increased jumping abilities compared.

A graph of blue bars

AI-generated content may be incorrect.

1. **SUMMARY OF FINDINGS**

Finish up with a paragraph or two of summarizing your findings about this data set.

Unfortunately, the main issue with our data set as mentioned during our presentation is that it isn’t complete. For us to be able to do exactly like we wanted to we would need data for the entirety of a player’s career. However, the data we are given only samples a random number of seasons for a certain player. This ends up giving us a random collection of data. There also isn’t a specific number of seasons for players. For example, we may have only gotten two or three seasons for one player and then five for another. Therefore, comparing a player’s points vs age for example isn’t something we can do.

What we are able to do with this data set is use the samples we are given and make inferences. One thing that made complete sense after looking at it was the fact that as a player gets taller their weight increases. We were also able to find a relationship between rebounds and height. Another thing that we came to realize was the players in the NBA and the college they attended. Because European colleges aren’t on the list and there is a large number of players in the NBA from Europe, the unknown category was by far the largest (over 40% of our data).