Data wrangling - project report

A report about scraping, cleaning, visualising and analysing basketball statistics

Project group members

Wouter van Zeijl, 2533591
Group 36

Research question

Which factors are related to an NBA player being nominated MVP?

Sub Questions

Does the position you play influence your ability to be MVP?
Do your MVP dreams fade away when you reach the age of 30?
Can you still reach MVP if you don't play that much?
Does a good player in a bad team get extra credits?

Data sources

All the basketball data has been scraped from the website www.basketball-reference.com. We created 3 datasets from the following sources:

- https://www.basketball-reference.com/awards/awards_{}.html
- https://www.basketball-reference.com/leagues/NBA_{}_per_game.html
- https://www.basketball-reference.com/leagues/NBA_{}_standings.html

Where each container is formatted to contain the year 1985 to 2023. This gives us a total 37 * 3 = 111 scraped web pages with data. This data has last been scraped on 2 february 2023 15:00 gmt + 2 and is stored in 3 different files. Moreover, these 3 files contain 37 scraped html files each and have been transformed into 3 different datasets. Finally, these 3 datasets have been merged to result in 1 final dataset.

We also scraped a table from wikipedia that contained a lot of abbreviations for each team name. We did not store this data, however we used it to map the above datasets in order to successfully merge them together.
https://en.wikipedia.org/wiki/Wikipedia:WikiProject_National_Basketball_Association/National_Basketball_Association_team_abbreviations

Data wrangling methods

Data acquisition

Method / Packages	Explanation
Requests	We used the requests library to get the html of a webpage and store it locally after encoding the url request to utf-8.
VPN	Our loop to scrape the pages contains 37 get requests, therefore we make too many requests to the website and we get blocked after 30 iterations for a certain amount of time. Therefore we have to rerun the last 7 iterations from a different IP using a VPN. Another option would be to use the time package to slow down the requests.
BeautifulSoup	BeautifulSoup allows for parsing html and extracting the table that we need based on a html ID.
Webdriver from the package Selenium	The table of the second webpage is loaded in with javascript on the client side. This prevents us from scraping the web page with a get request. We are able to work around this issue with the following steps: First download the chromedriver.exe from the website https://chromedriver.chromium.org/downloads. With the correct chrome version (in our case version 109), we are able to run the webdriver. The selenium webdriver offers automated test software which opens a standalone server that the webdriver uses to launch Google Chrome. From this standalone server we are able to scrape the web pages that we need.
Time	We need the time package because when we open the website on our standalone server we have to add commands to scroll down so the table gets loaded in properly on the client side and then wait for 2 seconds to make sure the table gets loaded in fully before we scrape it.

Data cleaning

Method / Packages	Explanation
Merge	We merge the dataframe players with the dataframe mvp on the player and year column. We use the argument outer to get the union of the keys of both dataframes. This gives us a lot of NaN values because the mvp dataframe contains far fewer rows than the dataframe players. (This is because the mvp dataframe only contains information about basketball players that have gained at least 1 mvp point). The NaN values are

	filled with 0.
Gather abbreviations by hand	As previously mentioned we scraped data from wikipedia to get a list of abbreviations and put them in a dictionary so we can map them in order to successfully merge the dataframes. However, this wikipedia was far from complete so we decided to manually add the missing 10 keys and 2 values to the dictionary.
To numeric	A lot of columns in our final dataframe had the type object, however we wanted them as strings, therefore we converted them with pandas.to_numeric

Data Analysis and visualisation

Method / Packages	**Explanation**
Scatter plots and heatmaps	We made scatter plots to see if we can come up with hints and answers to our research question. Finally we used seaborn to create a boxplot and a heatmap.

Besides what is described in the above 3 tables we used the obvious data wrangling tools like pandas, numpy, dropping duplicates, replacing headers and special characters in string objects.

Conclusion

We are happy to conclude that we are able to find factors that influence the MVP status of NBA players and therefore are able to answer our research question. We were surprised to find that the amount of free throw attempts is the most influential factor to the MVP score. We have the hypothesis that free throw attempts result in a lot of individual screen time of a player. Therefore, the player becomes more recognizable and is more prone to receive MVP points. However, the other high correlated factors such as points scored per game and 2-point field goals per game were less surprising. But we expected that the 3-point field goals per game would have a higher correlation with the MVP score than the 2-point field goals per game. Simply because 3 is the higher number. This is not the case, the 2-point factor has about double the correlation than the 3-point factor has with the MVP score.

Our conducted research has a possible limitation. We are dealing with data from 1985 till 2022, but we did not investigate if people awarded NBA players MVP points for different reasons than they do in 2022. There could be a subcultural difference between the 80’s and the last era.

Further research can be conducted to test our hypothesis. Image recognition can be used to count the amount of time a player is visibly on the screen. Then this newly created time factor can be tested for correlation with the MVP score. Moreover, if you are into gambling you can use this data to create a machine learning model and try to forecast which player is most likely

to end up as MVP. You could then use this information to bet with your friends or gambling companies.

Appendix

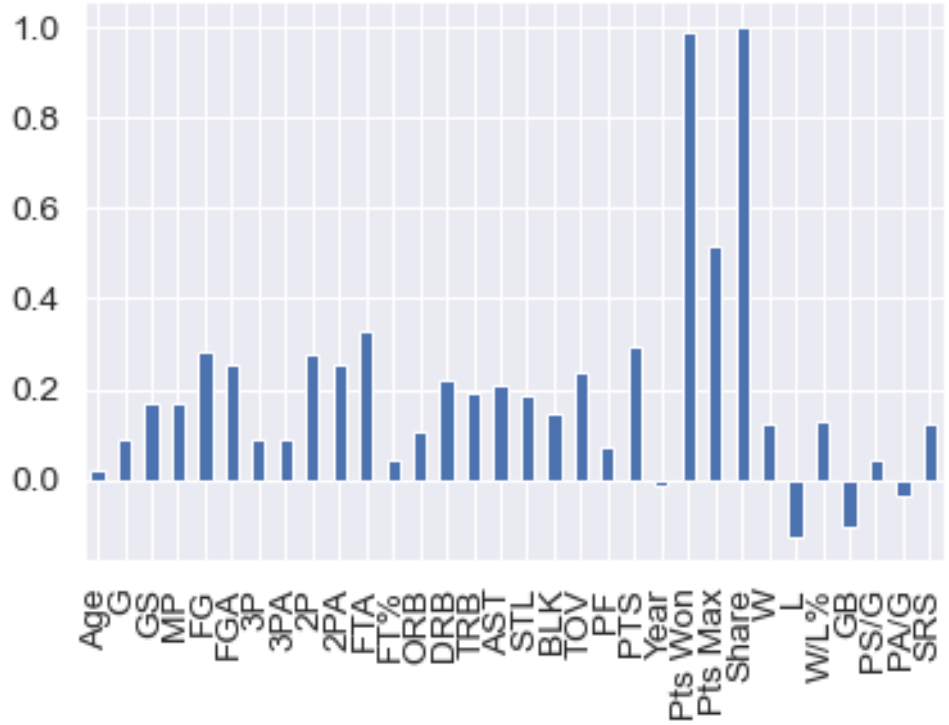

Fig 1. The correlation of every column against the Share column. The share column is the percentage of MVP points gained divided by the amount of MVP points obtainable.