Text

Description automatically generated

**Table of contents**

Introduction3

Data and data sources 4

Data cleaning and Exploratory data analysis6

Python cleaning process6

Data visualization9

Database4

Entities. ER Model5

SQL Queries 5

Conclusion4

**Introduction**

**Performances of the Milwaukee bucks and its superstar Giannis Antetokounmpo**

The NBA (National Basketball Association) is a professional basketball league in North America and the biggest basketball league in the world. The league is composed of 30 teams (29 in the United States and 1 in Canada).

Unlike many team sports where each team has developed an identity of play, the style of play of NBA teams varies from year to year and mainly depends on the star player. It only takes one trade to make a champion team at the bottom of the standings the next year.

Passionate about basketball I am following the NBA and most particularly the Bucks, the team of Milwaukee city. Giannis Antetokounmpo is a Greek player who was drafted by the bucks in 2013, at the age of 18. He is now 28 and one of the best players in the NBA with two MVP (Most Valuable Trophies) in 2019 and 2020.

Having learnt the basic tools of data analysis I now want to do my own approach on analyzing data about the evolution of the Bucks

The goal of this project was to analyze the weight of Giannis Antetokounmpo performances on the results of the Milwaukee Bucks team by comparing the player individuals’ statistics and the team statistics. I also want to show the evolution of the Milwaukee Bucks across the years team since Giannis started being in the team in 2013.

The plan of the project was to choose a data source that was able to provide all the game results and statistics of the Milwaukee bucks since 2013.

After export of the data into CSV files, imported the files into a new script of Python, to produce data cleaning (by managing with outliers, missing values, and features that were not relevant for the study) and visualizations, to perform a primary data analysis.

**DATA AND DATA SOURCES**

The data sources are from (<https://www.kaggle.com/datasets/nathanlauga/nba-games?resource=download>)

It is composed of 5 CSV files: games.csv, games\_details.csv, players.csv, ranking.csv, teams.csv.

I have only used these three files: games.csv, games\_details.csv.

I needed to merge these two tables to compare games statistics with players statistics

|  |  |  |
| --- | --- | --- |
| **file** | Games.csv | Games\_details.csv |
| **Content** | Every NBA game from 2003 to 2022 with stats of the two teams | Every NBA game from 2003 to 2022 with stats for each player of the two teams |
| **shape** | 25 796 rows,  21 columns | 645 953 rows,  29 columns |

**Description of some not-explicit columns**:

*Games.csv*

'PTS\_home/away': Total points scored by the home/away team

'AST\_ home/away': Total assists (pass before a shot made) by the home/away team

'REB\_ home/away': Total rebound by the home/away team

'FG\_PCT\_ home/away': percentage (accuracy) of shot made by the home/away team

'FT\_PCT\_ home/away': percentage (accuracy) of free throws made by the home/away team

'FG3\_PCT\_ home/away': percentage (accuracy) of 3 points shots made by the home/away team

Games\_details.csv

‘START\_POSITION': role of the player when he started the game

'MIN': total minutes played by a player during a game

'FGA': (Field goal attempts) number of shots attempts by a player

'FGM': (Field goal made) number of shots made by a player

'FG\_PCT':(Field goal percentage) = FGM / FGA

'OREB': offensive rebounds by a player

'DREB': defensive rebounds by a player

'REB': total rebounds by a player

'AST': total assists by a player

'STL': total steal by a player

'BLK': total block by a player

'TO': total turnover by a player

'PTS': total points by a player

'PLUS\_MINUS'**:** used to measure a player's impact on the game, represented by the difference between their team's total scoring versus their opponent's when the player is in the game.

Data cleaning and Exploratory data analysis

Python cleaning process (cleaning of the two files separately)

Games.csv

* Import of the libraries Pandas and NumPy
* Import files with pandas to read csv
* Listed the name of the columns to verifier
* Delete unnecessary columns (columns with one unique value and columns that I don’t want to use)
* Look for missing values in the data frame: there only missing values for the 2003 season. We want to focus on the year where Giannis Antetokounmpo was playing (2013-2022)
* Look for outliers with boxplots visualization and drop the outliers
* Create a new data frame with the data from 2013 to 2022 only
* Export the data frame to csv (“cleangames.csv”) to perform visualization

A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

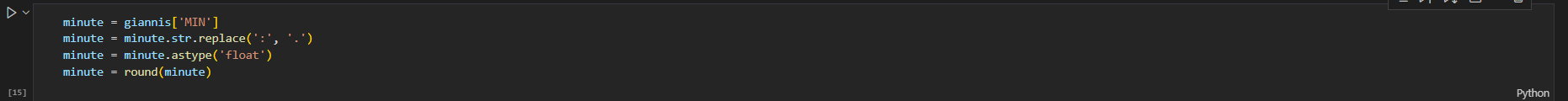
Gamesdetails.csv

* Import of the libraries Pandas and NumPy
* Import files with pandas to read csv
* Listed the name of the columns to verifier
* Look for the columns type
* Delete unnecessary columns
* Look for and manage with the missing values
* Drop the rows with too many missing values
* Convert the ‘MIN’ column from object type to float type
* Had to replace Nan values by ‘F’ the “START\_POSITION” column because Giannis Antetokounmpo role is Forward
* Create a new data frame (called giannis) which focus on the game where the Milwaukee Bucks and Giannis Antetokounmpo were playing
* Export the giannis data frame to csv (“cleangames.csv”) to perform visualization

Text

Description automatically generated

A screenshot of a computer

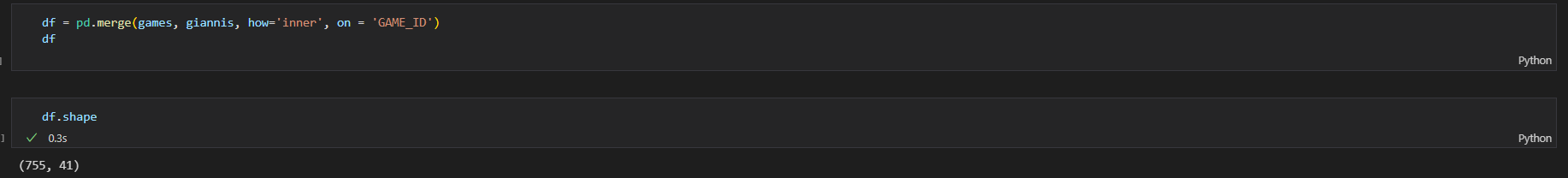
Description automatically generated with medium confidence

Graphical user interface

Description automatically generated with medium confidence



Merge the two data frame and perform visualization



**Data visualization**

**Visualization about the NBA League: Chart

Description automatically generated**

The previous lockout in 1998–99 had shortened the season to 50 games. During the lockout, teams could not trade, sign, or contact players.

**Chart, pie chart

Description automatically generated** **Chart, box and whisker chart

Description automatically generated**

We can see that when a team plays Home, its win rate is higher. And the main stats (points, rebound and assists) in the boxplot follow the tendency by being a little higher for home team.

**Visualization about Giannis performances:**

**Chart, bar chart, histogram

Description automatically generatedChart, funnel chart

Description automatically generated**

Giannis was getting better and improving his game year by year until he become the best player in the league in 2019 and 2020.

**Giannis weight on the bucks’ team:**

**Chart, scatter chart

Description automatically generated**

In the two graphs below, we compare Giannis evolution as a player with the win rate of the Milwaukee bucks.

These two graphs almost have the same behavior. So, it shows that Giannis has a huge impact on his team. He is a ‘game changer’

**Chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Chart, bar chart

Description automatically generated** **Chart, bar chart

Description automatically generated**

**Database**

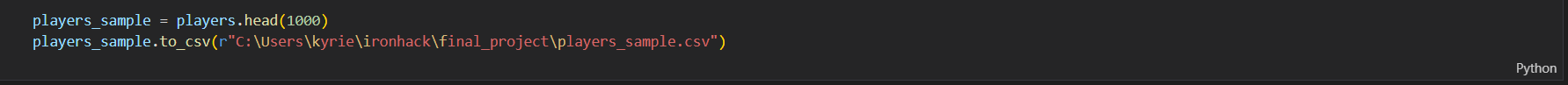
**Entities. ER models**

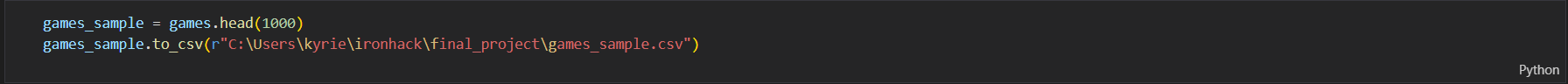
**Table

Description automatically generated**

**SQL Queries**

I had some problems while importing my data in MySQL because the file was really big, so I used sample of my two data frame and export them as csv:





Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Conclusion**

The data I collected for this project was quite comprehensive. Indeed, the NBA datasets are fed very seriously and regularly by the NBA and by the fan’s associations. For my project I used one dataset about the history of all the NBA games since 2003 and my second dataset was about the players statistics for all these games.

I cleaned the data from both files step by step and one dataset after the other. The first step was to select the columns I wanted to deal with and to drop the not relevant ones. Then I had to manage with some missing values and to drop some outliers. In the end I merge my two data frames to do some charts who compare Giannis Antetokounmpo with the Milwaukee Bucks team.

The visualization was about to show some characteristics of an NBA season and then I tried to show the correlation with Giannis Antetokounmpo statistics in game with the performances of his team. Some charts are clear by showing graphs with a linear behavior.

Finally, we can conclude that Giannis Antetokounmpo had a huge impact on the Milwaukee Bucks dominations this last few years as well as the Milwaukee team had an impact on the Giannis Antetokounmpo