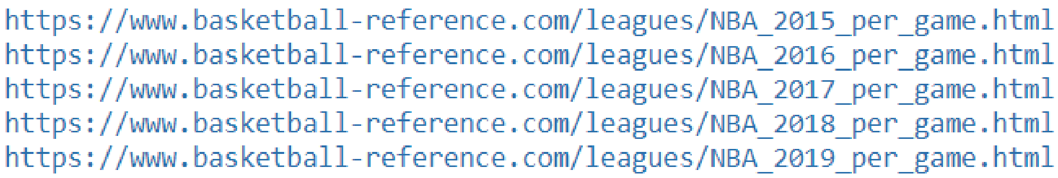
Project Name:

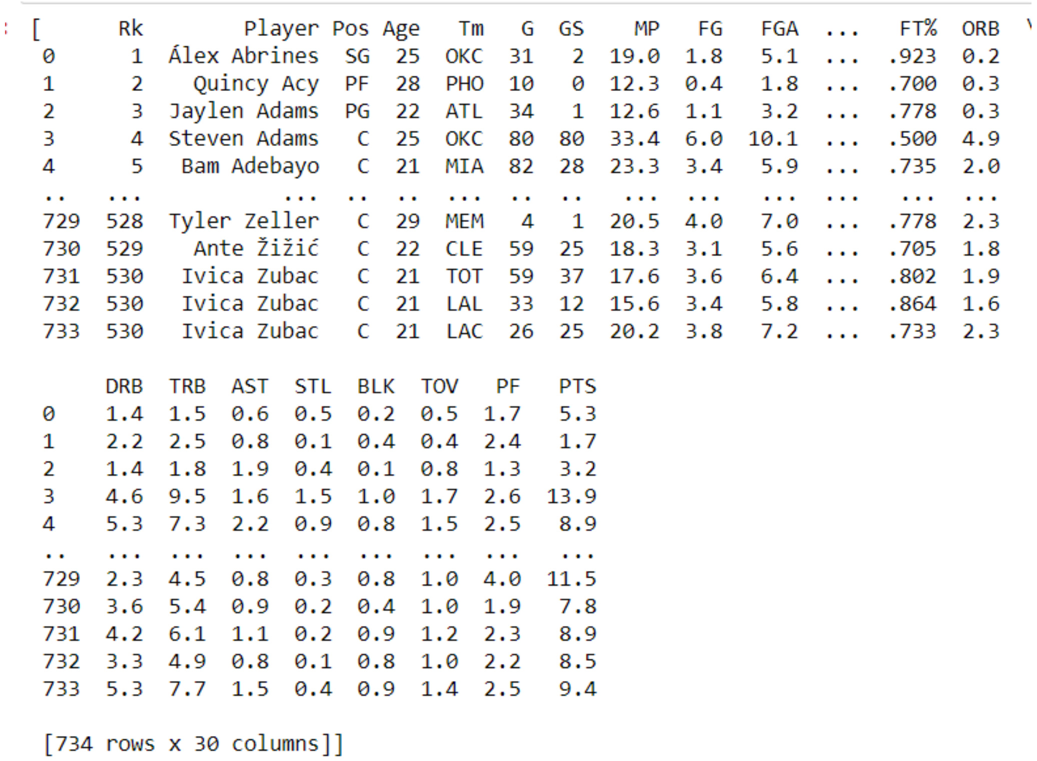
Web-Scraping Data-Cleaning EDA Visualization

Table of Contents

Demo



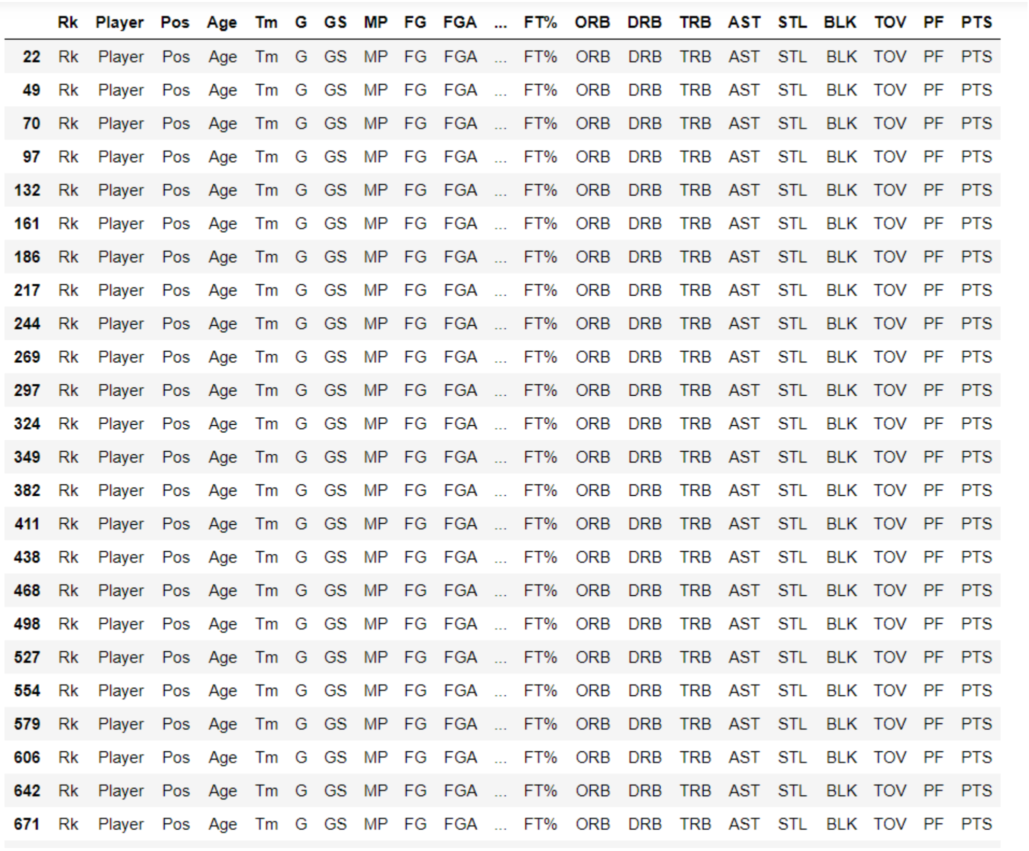
```
https://www.basketball-reference.com/leagues/NBA_2015_per_game.html
https://www.basketball-reference.com/leagues/NBA_2016_per_game.html
https://www.basketball-reference.com/leagues/NBA_2017_per_game.html
https://www.basketball-reference.com/leagues/NBA_2018_per_game.html
https://www.basketball-reference.com/leagues/NBA_2019_per_game.html
```



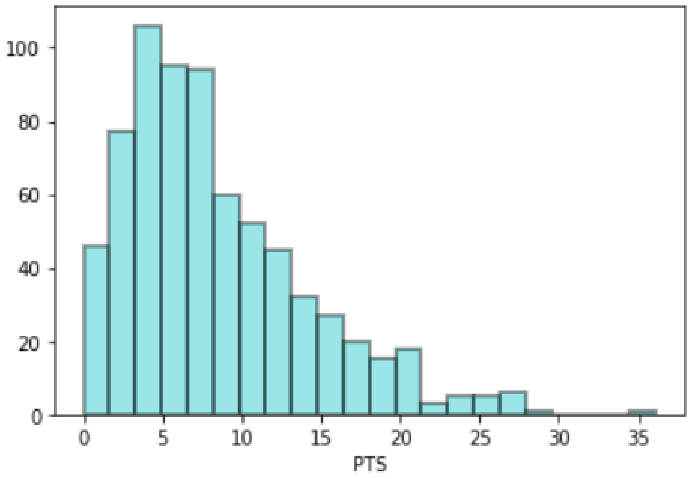
```
[      Rk          Player Pos Age   Tm   G  GS    MP   FG   FGA  ...   FT%  ORB  \
 0      1    Álex Abrines  SG  25  OKC  31   2  19.0  1.8   5.1  ...  .923  0.2
 1      2      Quincy Acy  PF  28  PHO  10   0  12.3  0.4   1.8  ...  .700  0.3
 2      3    Jaylen Adams  PG  22  ATL  34   1  12.6  1.1   3.2  ...  .778  0.3
 3      4    Steven Adams   C  25  OKC  80  80  33.4  6.0  10.1  ...  .500  4.9
 4      5     Bam Adebayo   C  21  MIA  82  28  23.3  3.4   5.9  ...  .735  2.0
 ..   ...             ...  ..  ..  ...  ..  ..   ...  ...   ...  ...   ...  ...
 729  528    Tyler Zeller   C  29  MEM   4   1  20.5  4.0   7.0  ...  .778  2.3
 730  529      Ante Žižić   C  22  CLE  59  25  18.3  3.1   5.6  ...  .705  1.8
 731  530     Ivica Zubac   C  21  TOT  59  37  17.6  3.6   6.4  ...  .802  1.9
 732  530     Ivica Zubac   C  21  LAL  33  12  15.6  3.4   5.8  ...  .864  1.6
 733  530     Ivica Zubac   C  21  LAC  26  25  20.2  3.8   7.2  ...  .733  2.3

      DRB  TRB  AST  STL  BLK  TOV   PF   PTS
 0    1.4  1.5  0.6  0.5  0.2  0.5  1.7   5.3
 1    2.2  2.5  0.8  0.1  0.4  0.4  2.4   1.7
 2    1.4  1.8  1.9  0.4  0.1  0.8  1.3   3.2
 3    4.6  9.5  1.6  1.5  1.0  1.7  2.6  13.9
 4    5.3  7.3  2.2  0.9  0.8  1.5  2.5   8.9
 ..   ...  ...  ...  ...  ...  ...  ...   ...
 729  2.3  4.5  0.8  0.3  0.8  1.0  4.0  11.5
 730  3.6  5.4  0.9  0.2  0.4  1.0  1.9   7.8
 731  4.2  6.1  1.1  0.2  0.9  1.2  2.3   8.9
 732  3.3  4.9  0.8  0.1  0.8  1.0  2.2   8.5
 733  5.3  7.7  1.5  0.4  0.9  1.4  2.5   9.4

 [734 rows x 30 columns]]
```



	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
22	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
49	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
70	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
97	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
132	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
161	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
186	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
217	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
244	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
269	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
297	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
324	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
349	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
382	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
411	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
438	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
468	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
498	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
527	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
554	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
579	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
606	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
642	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
671	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS



Overview

Web Scraping is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format.
it automates the gathering and dissemination of information. In the wrong hands, it can lead to theft of intellectual property or an unfair competitive edge. Therefore, before you scrape you need be careful and scrape only legal sites. Web scraping extracts underlying HTML code and, with it, data stored in a database. The scraper can then replicate entire website content elsewhere. Whether a website can be scraped or not, can check or know if a website allows scraping either by python or any tool or language, all you need do is to check the websites robots. txt file by going to websiteName. tld/robots.

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mis-labeled.

EDA is essentially a type of storytelling for statisticians.

It allows us to uncover patterns and insights, often with visual methods, within data.

EDA is often the first step of the data modelling process.

This repository contains the code for right from web scraping till Visualization using python's various libraries.

It used Pandas and Seaborn libraries.

These libraries help to perform individually one particular functionality.

Pandas objects rely heavily on Numpy objects.

Seaborn is data visualization library based on matplotlib.

The purpose of creating this repository is to gain insights into process from data collection to data visualization.

These python libraries raised knowledge in discovering these libraries with practical use of it.

It leads to growth in my ML repository.

This above few screenshots will help you to understand flow of output.

Motivation

Web-scraping provides one of the great tools to automate most of the things a human does while browsing. Web-scraping is used in an enterprise in a variety of ways − Data for Research, Products prices & popularity comparison, SEO Monitoring, Sales and Marketing. The reason behind building this is, to maximize I as analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract. It is a way of visualizing, summarizing and interpreting the information that is hidden in rows and column format. EDA is understanding the data sets by summarizing their main characteristics often plotting them visually. This step **is** very important for me especially when I arrive at modelling the data in order to apply Machine learning. Data visualization is the discipline of trying to understand data by placing it in a visual context so that patterns, trends and correlations that might not otherwise be detected can be exposed. Visualization through visual imagery has been an effective way to communicate both abstract and concrete ideas since the dawn of humanity. A good visualization tells a story, removing the noise from data and highlighting the useful information. Effective data visualization is a delicate balancing act between form and function. Even statistically, it is said that child from 0-5 years of age can remember 92% of things that have seen in form of image as cartoons rather than only read as text such as dialogues of cartoon characters. For example, I do not remember all dialogues of Tom-Jerry Cartoon but I definitely remember how they look and that is because I saw their visual picture. Till now I have practiced one by one each of the skills as so now I thought to make a project by combining them which includes web scraping, data cleaning, exploratory data analysis and visualization. As an employee of company, one should be self-directed in cases and hence making this project will lead me to the aim. Hence, I continue to gain knowledge while practicing the same and spread intellectual wings in tech-heaven.

Technical Aspect

Pandas module mainly works with the tabular data. It contains Data Frame and Series. Pandas is 18 to 20 times slower than Numpy. Pandas is seriously a game changer when it comes to cleaning, transforming, manipulating and analyzing data.
Seaborn provides a high-level interface for drawing attractive and informative statistical graphics. It provides a variety of visualization patterns and visualize random distributions.

Installation

Using intel core i5 9th generation with NVIDIA GFORECE GTX1650.
Windows 10 Environment Used.
Already Installed Anaconda Navigator for Python 3.x
The Code is written in Python 3.8.
If you don't have Python installed then please install Anaconda Navigator from its official site.
If you are using a lower version of Python you can upgrade using the pip package, ensuring you have the latest version of pip, *python* -m pip install --*upgrade pip and press Enter*.

Run/How to Use/Steps

Keep your internet connection on while running or accessing files and throughout too.
Follow this when you want to perform from scratch.
Open Anaconda Prompt, Perform the following steps:

```
cd <PATH>
pip install pandas
pip install seaborn
```

You can also create requirement.txt file as, pip freeze > requirements.txt
run files.

Follow this when you want to just perform on local machine.
Download ZIP File.
Right-Click on ZIP file in download section and select Extract file option, which will unzip file.
Move unzip folder to desired folder/location be it D drive or desktop etc.
Open Anaconda Prompt, write cd <PATH> and press Enter.
eg: cd C:\Users\Monica\Desktop\Projects\Python Projects 1\allin1\Project_1_WebScraping+DataCleaning+EDA+Visualization

In Anconda Prompt, pip install -r requirements.txt to install all packages.
Open in Jupyter Notebook, <filename>.ipynb
That is,

Open in Jupyter Notebook, WebScraping_DataCleaning_EDA_Visualization.ipynb
Please be careful with spellings or numbers while typing filename and easier is just copy filename and then run it to avoid any silly errors.
Note: cd <PATH>
[Go to Folder where file is. Select the path from top and right-click and select copy option and paste it next to cd one space <path> and press enter, then you can access all files of that folder]   [cd means change directory]

Directory Tree/Structure of Project

Folder: allin1 > Project_1_WebScraping+DataCleaning+EDA+Visualization
WebScraping_DataCleaning_EDA_Visualization.ipynb

To Do/Future Scope

Can add ML models.

Technologies Used/System Requirements/Tech Stack

pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

seaborn

Credits

Data Professor Channel