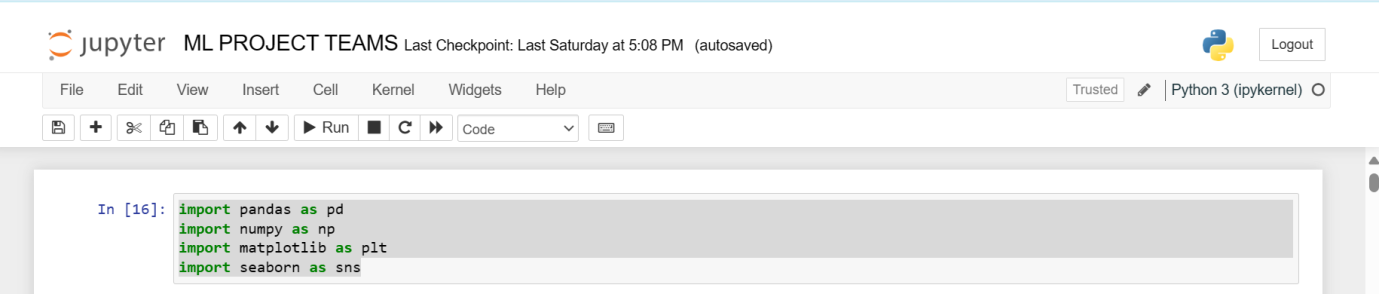
**Summary of the project**

The “Team .csv “dataset comprises various attributes pertaining to different teams. These attributes may include (team, age, country, year, event, height, weight, medals, Prev medals) and contains **2144 rows and 11 columns**. Leveraging machine learning algorithms on this dataset can unveil valuable insights and facilitate informed decision-making processes.

To begin the analysis, exploratory data analysis **(EDA)** techniques can be employed to understand the structure of the dataset. This involves tasks such as data cleaning, handling missing values, and visualizing distributions of features.

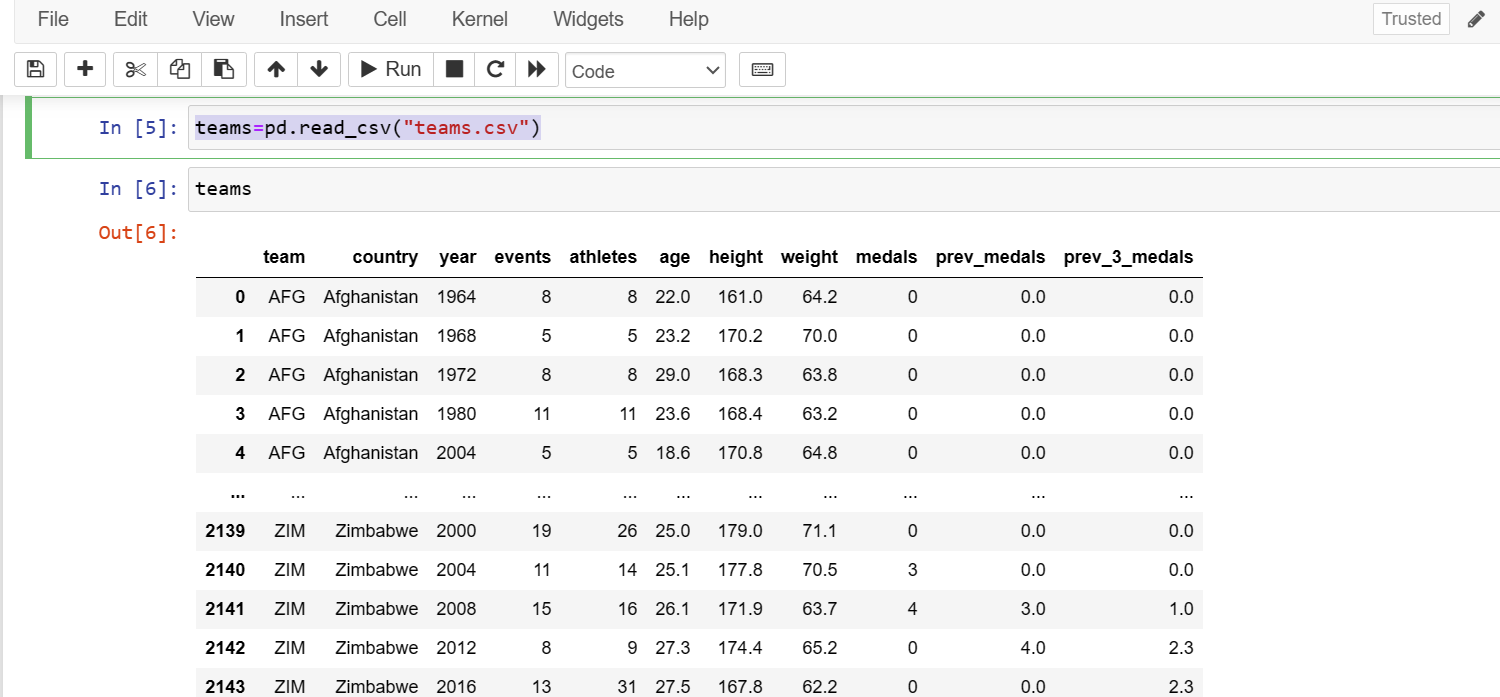
Here’s a step-by-step detail following below: -

* **Importing libraries**
* **pandas:** It’s particularly useful for data manipulation, analysis, and cleaning tasks.
* **NumPy:** Offers support for large, multi-dimensional arrays and matrices.
* **Matplotlib**: It's highly customizable and suitable for creating static, interactive, and publication-quality plots.
* **Seaborn**: It simplifies the process of creating complex visualizations such as heatmaps, pair plots, and categorical plots.



**Step-2🡪**

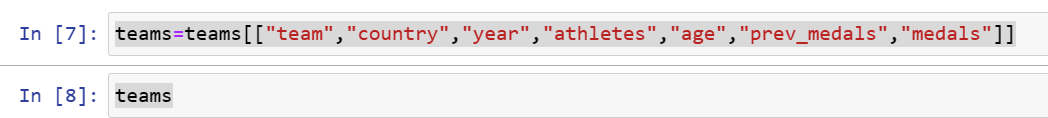
Loading the data from the “team.csv” file into a Pandas Data Frame named ‘teams’



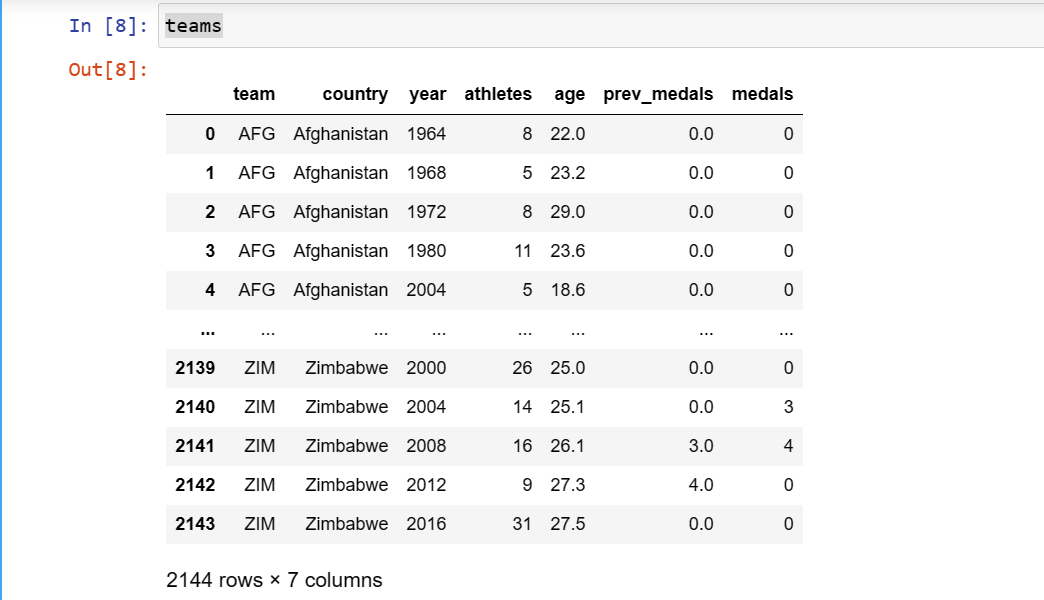
2144 rows × 11 columns

**Step 3🡪**

selecting specific columns from the DataFrame teams and reassigned it to the same variable, keeping only columns: "team", "country", "year", "athletes", "age", "Prev medals", and "medals".

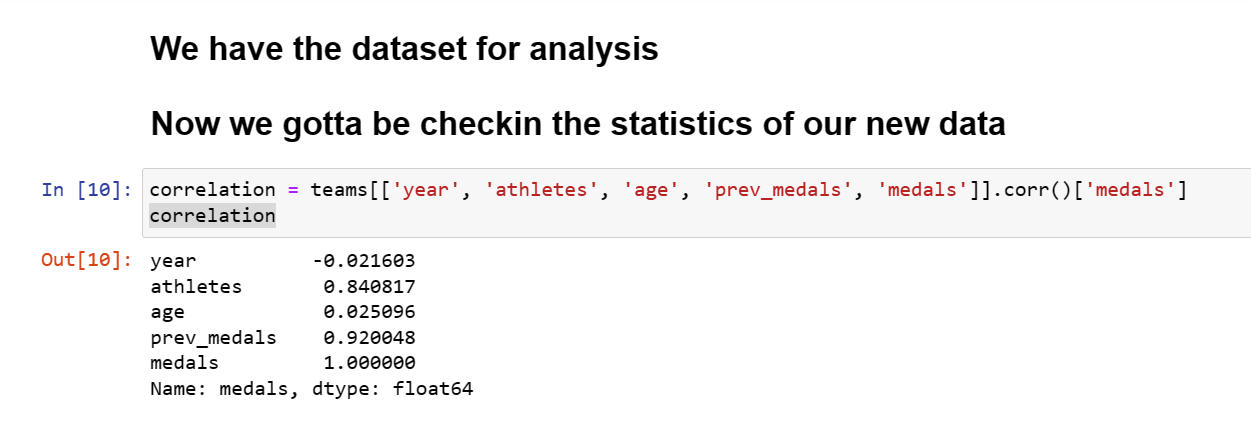


Followed by Printing out the variable teams after the column selection operation would display the DataFrame with only the specified columns: "team", "country", "year", "athletes", "age", "Prev medals", and "medals".

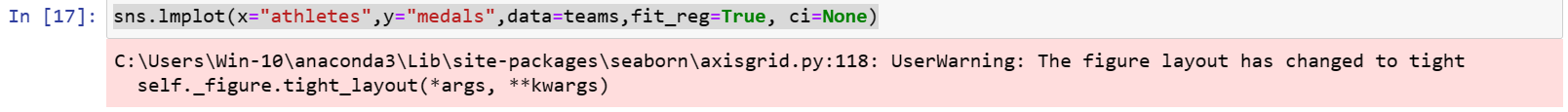


**STEP 4 🡪** **Performing Explanatory Data Analysis: -**

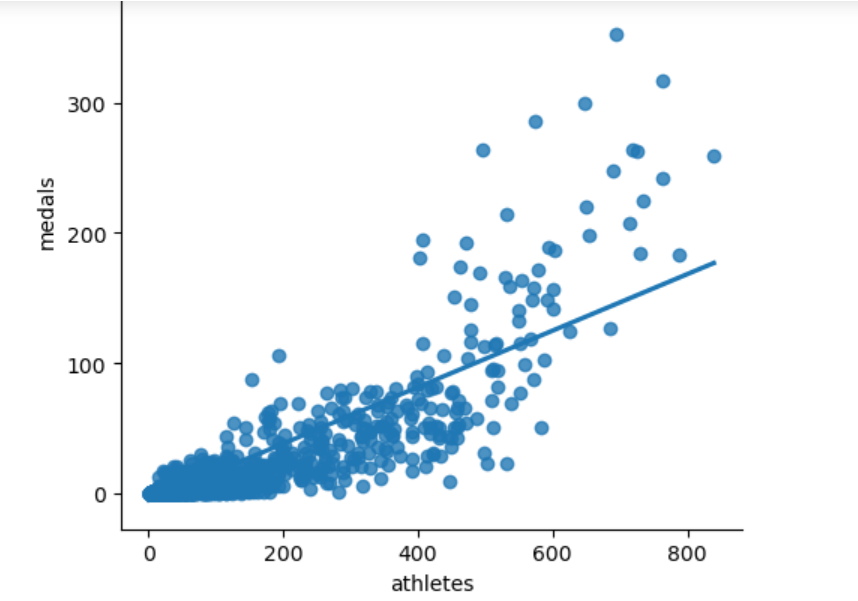
Explanatory Data Analysis (EDA) is an approach to analysing data sets to summarize their main characteristics, often employing graphical techniques. The goal of EDA is to gain insights into the data and uncover patterns, anomalies, relationships, and other interesting features. It is typically one of the initial steps in the data analysis process, preceding more formal statistical modeling or hypothesis testing



# Visually analysing strong correlation with medals

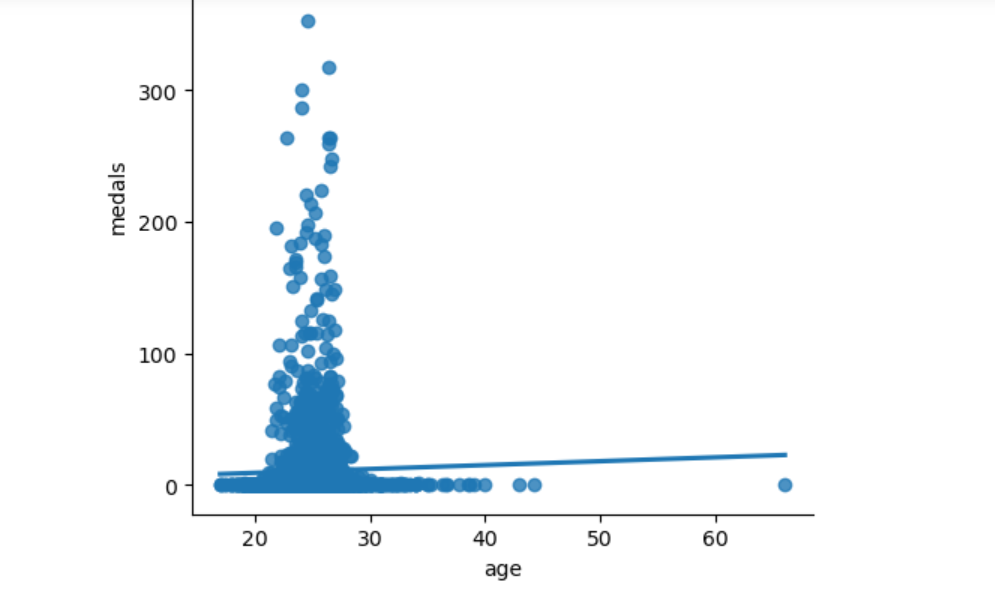


* OUTPUT



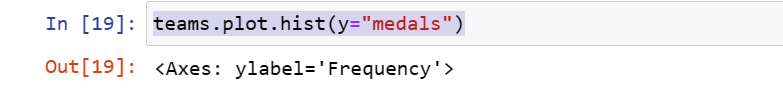
**Step 5🡪**

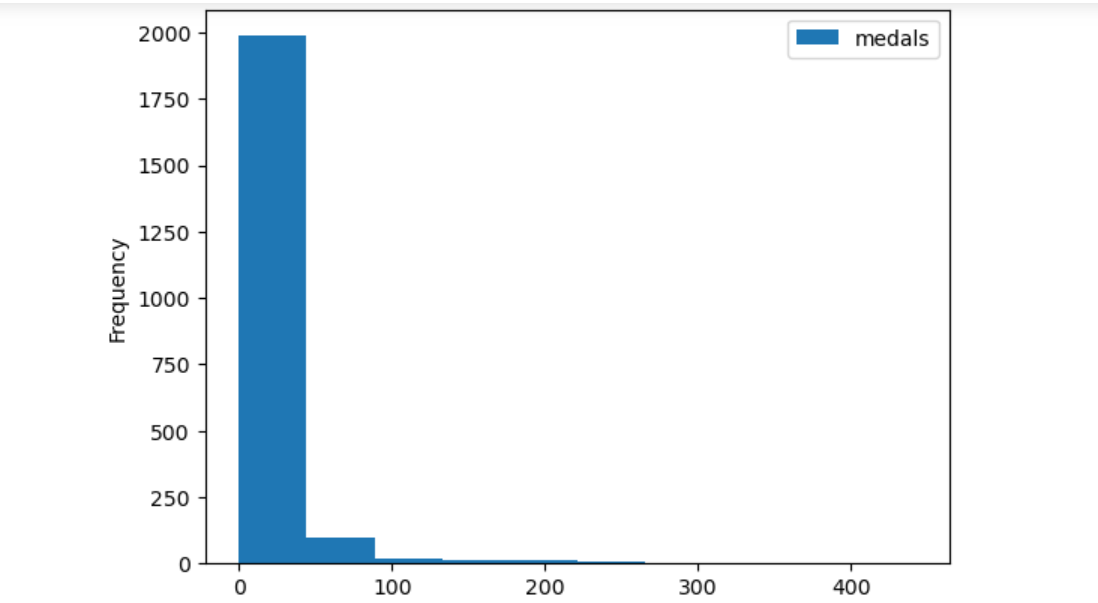
The code creates a scatter plot with a linear regression line to visualize the relationship between athletes' age and the number of medals won, using data from the DataFrame teams.



**Step 6 🡪**

The code shows the frequency or count of different medal counts.





Step 7 🡪

**Data cleaning process**

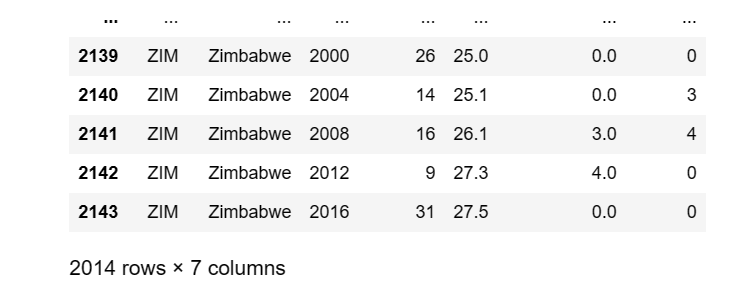
Data cleaning involves identifying and correcting errors, inconsistencies, and missing values in a dataset to improve its quality and reliability for analysis. As in the dataset alots of missing value present so first we have to check how many are present (missing values) in our data set.

**Source Code:-**  teams[teams.isnull().any(axis=1)]

This code filters rows in the 'teams' data frame containing any null values across columns, highlighting incomplete or missing data entries.



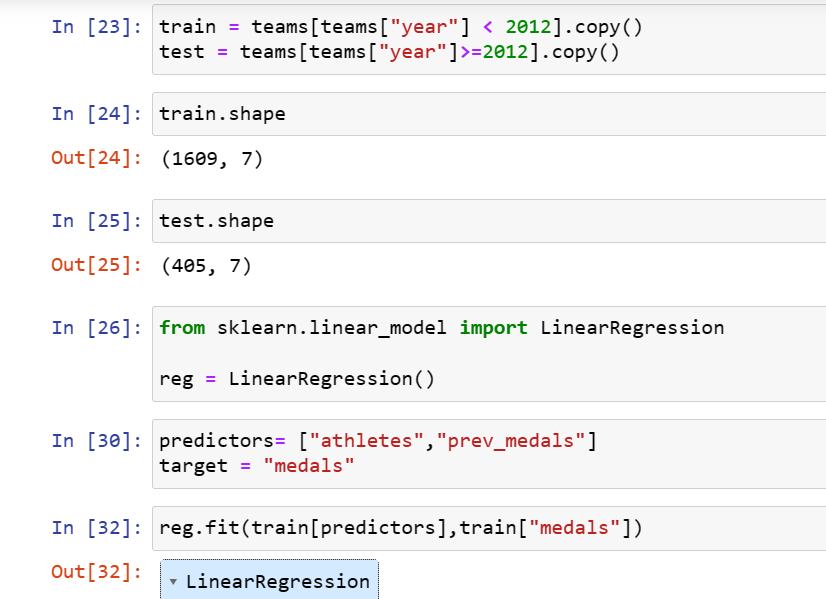


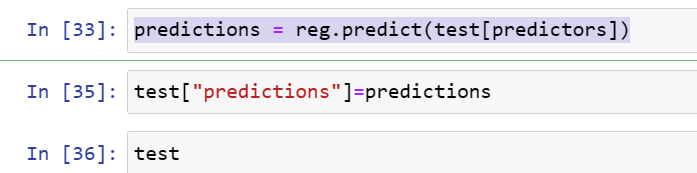


**Step 8 🡪**

**Linear Rgression:-** Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.

Involves with the training and testing of the models

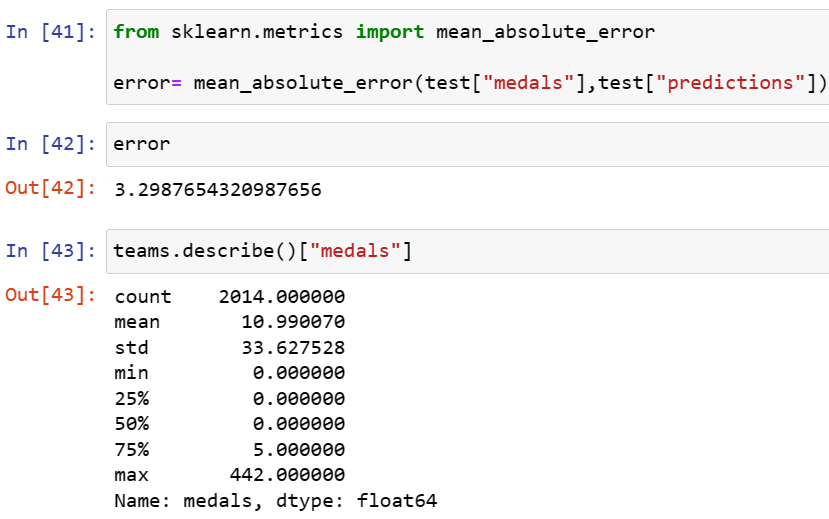


The code generates predictions for the target variable using a trained regression model on the predictor variables from the test dataset.



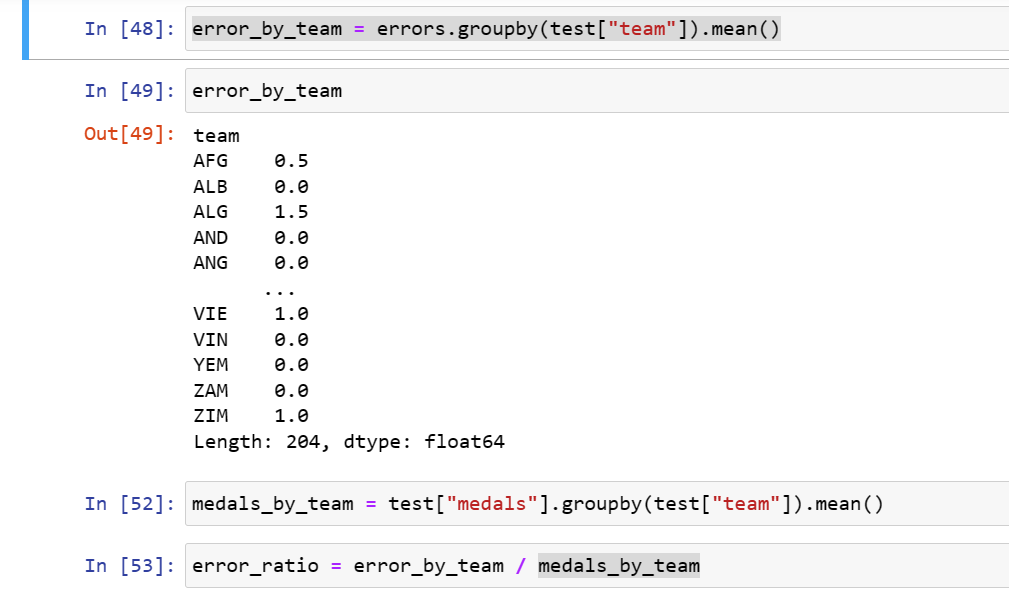
**Step 9🡪**

The code snippet modifies the values in the "predictions" column of he DataFrame "test" by setting any values less than 0 to 0

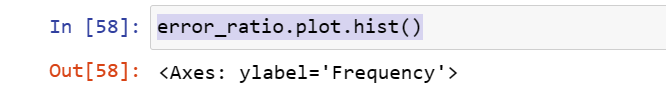


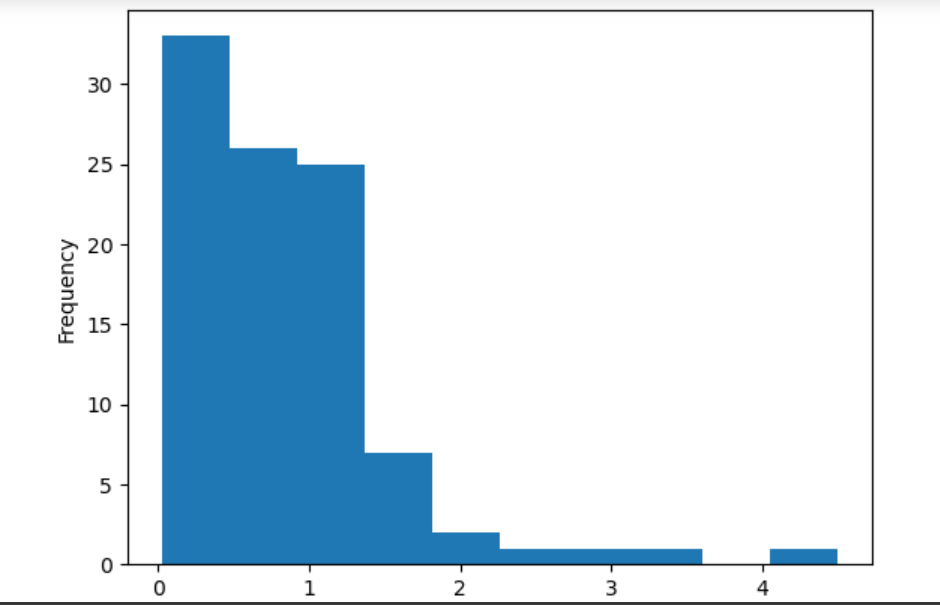
**Step 10🡪**

The code calculates the mean of errors grouped by the "team" column in the DataFrame "test" and assigns the result to the variable "error\_by\_team".



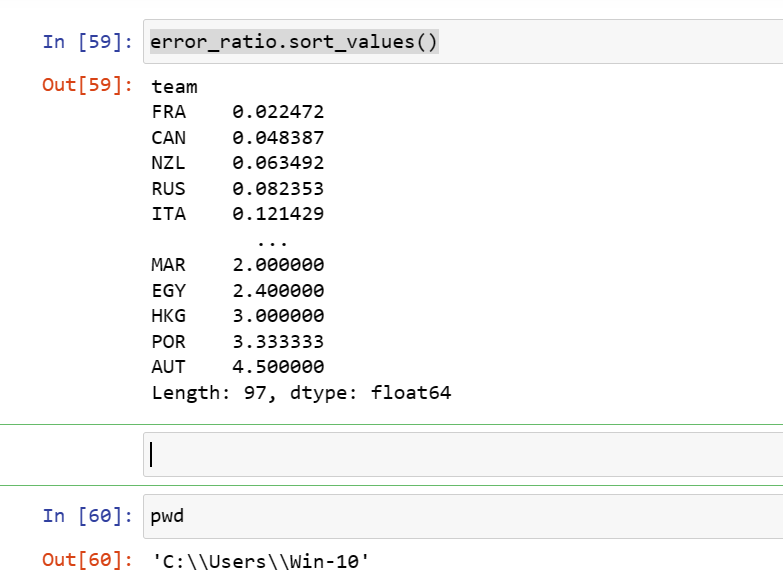
Now creating a histogram to visualize the distribution of values.





**Step 11🡪**

The code sorts the values in the “error\_ratio” variable in ascending order.



**Conclusion**

This project utilized Python libraries including Pandas, NumPy, Matplotlib, and Seaborn for data analysis, visualization, model building, and performance evaluation. Despite challenges, such as data quality issues and model optimization, the project successfully extracted insights and trends from the dataset. Through regression analysis and group-wise aggregation, meaningful patterns were uncovered. The project showcased the effectiveness of these libraries in extracting actionable insights from data, facilitating informed decision-making processes. By leveraging these tools and techniques, valuable contributions were made towards understanding the underlying dynamics of the dataset and building predictive models to support data-driven decision-making in relevant domains.