1. Title: Football player state analysis

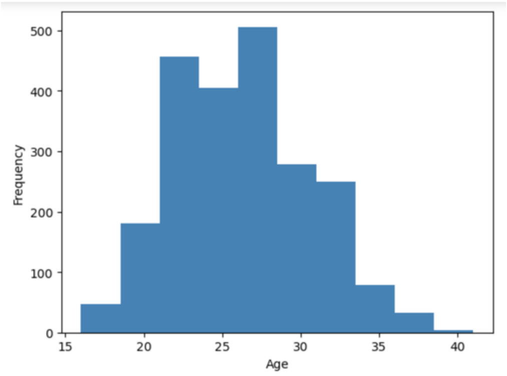
I get the data from Kaggle (I will put the link in the end). This dataset describes the data of 2022-2023 football player state. When I download this dataset, The data is messy. I preprocess the dataset first.

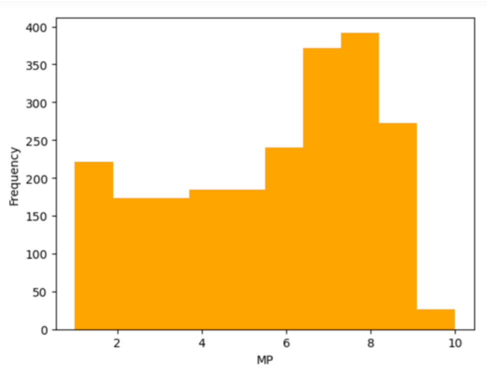
After that, I also notice that some data are not reasonable. such as some player age over 100 or lower than 16. The number of goals by someone are higher than shoot. I replace these error value by average value.

2. the main figure produced in Python

This part I use three kinds of image. They are histogram, word cloud map, kernel density map and network map.

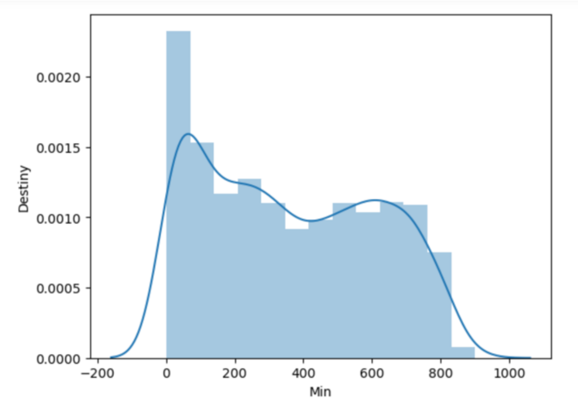
Histogram:

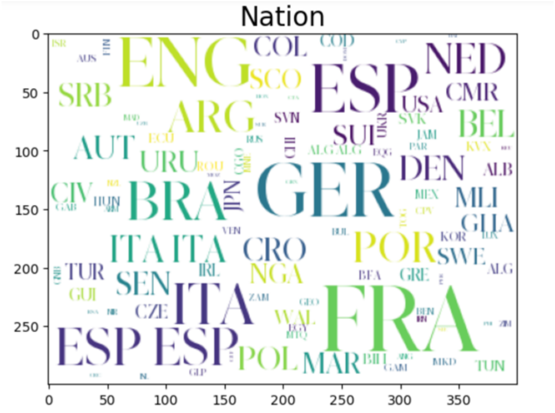




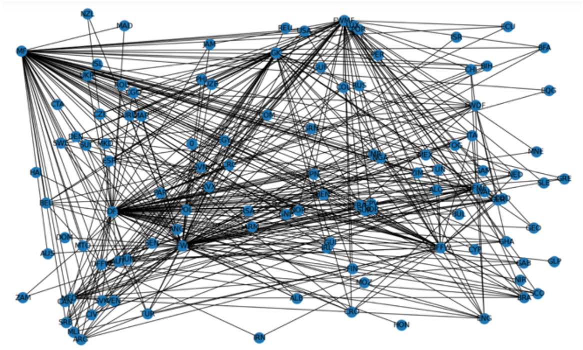
Kernel map

Word cloud map:





Network map:



3. the legend explaining the visualization components in the figure

In histogram, I draw the players age and matches played frequency. The blue histogram x axis means the players' age, y axis means the frequency of the players in this age. The yellow histogram x axis means the number of matches played; y axis means the frequency in this match number. The light blue image is kernel density map. X axis means how much time about playing time. Y axis means the density (proportional distribution). The blue curve shows us the trend. It is clearer to get the overall game time distribution of all members. The word cloud map shows us the nations where these players come from. I choose different color to easily distinguish. The words bigger mean more players come from this country. Finally, I draw a network map. It describes the players' cities in different countries. If the black line more density, it means this county have more football player.

4. findings text introducing highlights of the produced figure in bulletin points.

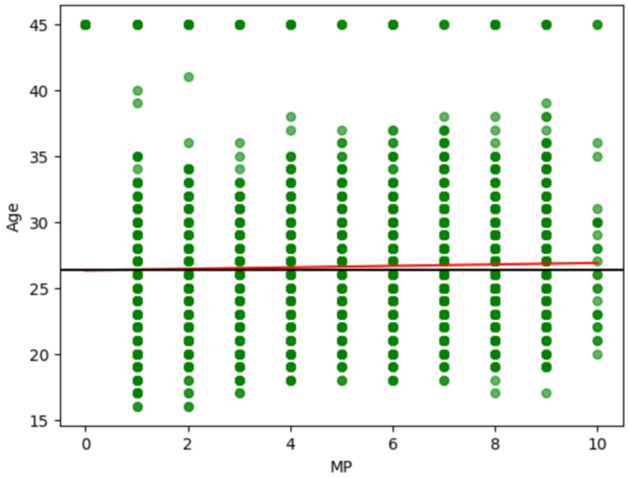
In the histogram and kernel density map, we can easily get the distribution of each kind of data. Which variable has more density. The distribution is Gauss or flat. The word cloud map and network map make us easily get the distribution by text instead number. All of these images make us understand dataset more comprehensively.

5. data and method text describing the data and method used in this process

After the first time I visualize the data, I notice that some data are obvious error, such as some player age over 100 or lower than 16. The number of goals by someone are higher than shoot. I replace these error value by average value. After preprocess, there are no weird data in the map. In these images, I just choose some kinds of data, they are Age, MP (Matches played), Min: Minutes played, Nation and Pos: Position.

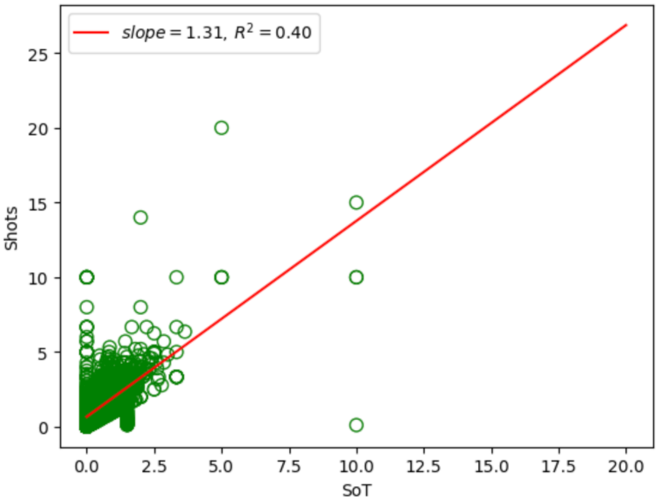
6. A significance statement on why the presented figure is an important topic.

In order to finish this part, I do some more analysis to presented next figure is an important topic. At the first part, I draw two histograms about age and match played. Then I want to analysis the relationship between them.



This image shows us the relationship between match played and Age. I use the scatter plot and trend line. As usual we think id players older, they will play less match. However, we notice that the red line in the image coefficient is very small. Then we get the Age does not relate to the match played.

In this dataset, we also have other data. Here is the relationship between shots and shot of target (SoT):



The X axis means the number of goal and Y axis means the number of total shots. From the image, we know that these two data are positive correlation. It means if players want to get more score they need shot more.

My topic is football players state analysis. I not only analysis the distribution of some single data, but also analysis some two kinds of data correlation. They both important to learn and analysis dataset.

Dataset link in Kaggle: https://www.kaggle.com/datasets/vivovinco/20222023-football-player-stats

GitHub link: https://github.com/chriswong-6/2415.git