Emilio Padrón Molina

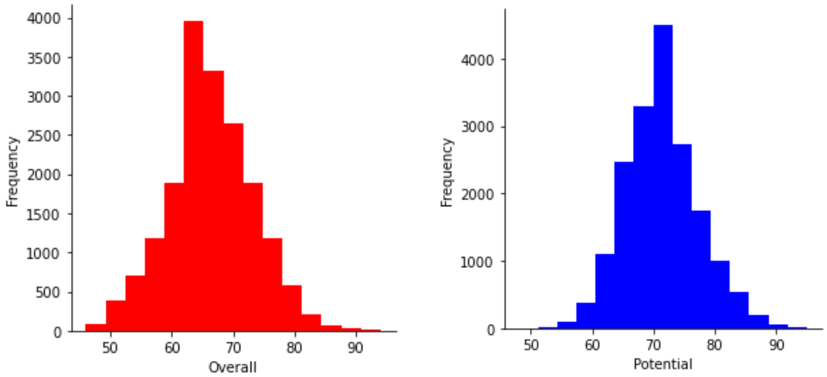
Exploratory Data Analysis for ML Report

For this report, I decided to work with the FIFA 19 dataset. This dataset, obtained from Kaggle (https://www.kaggle.com/karangadiya/fifa19), contains information from all of the active professional soccer players in the world. Aside from their names, age, nationality and club, there are other specific attributes such as their overall score, potential, stamina, and specific abilities. This set has 18,207 rows (corresponding to specific players), and 89 columns (representing their attributes). Some of the columns are not very useful for an analysis (such as photo or flag); therefore, the first step in cleaning the data should rely on dropping columns that are not needed for the analysis.

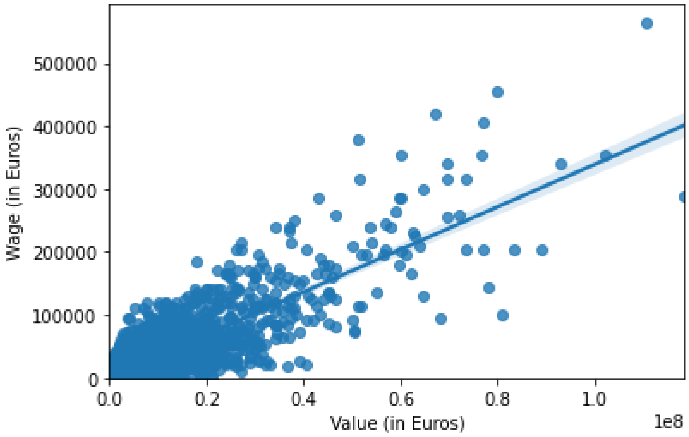
After importing the data and dropping the useless columns, it is necessary to deal with missing values. For this specific set, there are many features that contain null values. In the case of the 'Club' column, one can assume that if it a value is empty, it means that the player doesn't have a club; therefore, one can assign a 'No-Club' string for these cases. Moreover, the 'Preferred Foot', 'Position', 'Height', 'Weight' columns don’t have many missing values, so one can just delete the observations that don't contain this info. On the other hand, the 'Release Clause' column has a lot missing values; therefore, it is easier to just drop the entire column as a whole. Additionally, for the columns that represent specific abilities, the missing values could just be replaced by the mean of the column; nonetheless, this wasn’t necessary given that these observations coincided with the observations that were deleted previously.

Furthermore, it is relevant to check the data types of each column. In this specific case, it can be seen that the 'Value', 'Wage', 'Height' and 'Weight' columns are considered as objects rather than numbers; therefore, it is necessary to change them to a numerical type. Similarly, for the attributes regarding abilities, given that their values correspond to integers, it is required to switch from a float64 type to an int64 type.

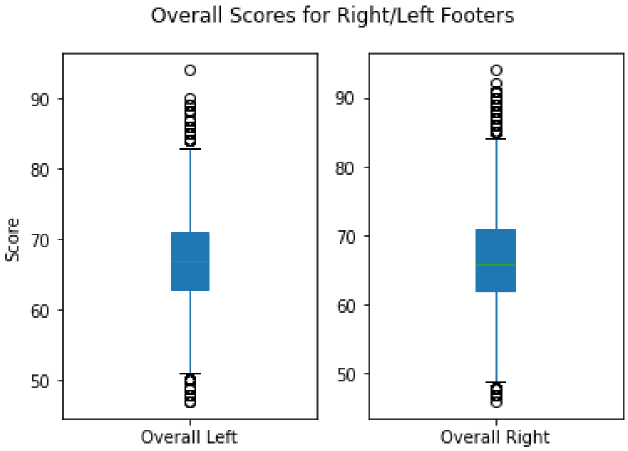
At this point of the exploration, it is relevant to look at some visuals to obtain some insights of the set. First, it is thought appropriate to look at the histograms of the overall scores and potential of the players. Considering 15 bins for each case, the result is the following:



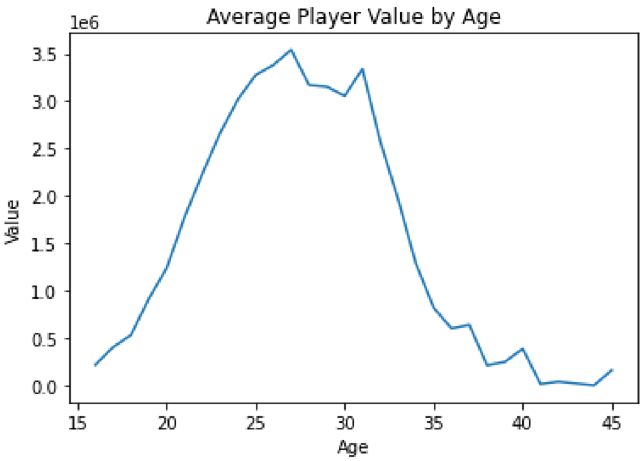
These graphs show that the overall and potential scores seem to follow a normal distribution. However, these distributions are not centered at 50 as one would expect; while the overall quality appears to be centered around 65, the center of potential feature is located around 72. Also, it is considered relevant to analyze the player's wage with respect to their value. This relationship can be visualized with the next figure:



Even though the wage of a player appears to be positively correlated with his value, given the dispersion of points, one cannot clearly determine that this relation is linear, or even if it actually follows a pattern. Furthermore, one may ask oneself whether right footers are better than left footers, or vice versa. This can be answered with the help of the following boxplots:



In this figure, it is evident that both, the inter quartile range and the distribution in general, are similar between right and left footers. This indicates that one cannot say that there is a significant difference in talent based on the player's preferred foot, at least visually. Additionally, as a last visual, it is believed that looking at the average wages of players based on their age might show an interesting insight. For this, a line plot can be built:



This graphic shows that while salary tends to increase for promising youngsters less than 25 years old, players between 25-32 years of age will have achieved what will probably be their highest salary and afterwards, once they pass the 32 years threshold, their wage will continue to decrease until the end of their careers.

After finishing with the visual analysis, it becomes pertinent to conduct some type of hypothesis testing. For this, three different tests were designed:

1. Soccer fans in Mexico believe that their Federation has a special talent in the formation of goalkeepers; in other words, they say that Mexican goal keepers are greater than average. Is this true?

   This can be put as:

   $$H_0 : \bar{x}_{Mexico} = \mu$$
   $$H_A : \bar{x}_{Mexico} > \mu$$

   where μ is the average score of all goalkeepers and x_Mexico is the mean score of Mexican goalkeepers.

2. It is evident that strikers receive more prices than any other position. However, is it true that the also get payed better than the rest of the positions?

   This can be put as:

   $$H_0 : \bar{x}_{ST} = \mu$$
   $$H_A : \bar{x}_{ST} > \mu$$

   were μ is the average salary of soccer players and x_ST is the mean salary of strikers.

3. When it comes to penalties, are left footers better than right footers?

   This can be put as:

   $$H_0 : \bar{x}_{Left} = \mu$$
   $$H_A : \bar{x}_{Left} \neq \mu$$

   were μ is the average penalty score of all soccer players and x_Left is the mean penalty score of left footers.

One can solve any of these tests by calculating their respective μ, and using the "stats.ttest_1samp" object from Scipy. For instance, in the case of the first test in the previous paragraph, μ=64.603 and a Mexican goalkeepers' dataset was passed through the already mentioned object. Using a significance level of 5%, this resulted in the acceptance of the null hypothesis; i.e., with 95% certainty, one cannot say that Mexican goalkeepers are significantly better than the rest.

The next step that needs to be taken with this dataset is to simplify the age and nationality data by dividing their respective values into bins. For example, in the case of age, these bins can be: 15-18, 19-22, 23-26, and so on. Furthermore, the numerical features such as the specific abilities, height, weight, wage, and value should be scaled before generating a machine learning model. Similarly, it is necessary to encode the different categorical values depending on the nature of the feature; for example: preferred foot with binary encoding and position with one-hot-encoding.

Finally, it is worth noting that this dataset seems to be quite complete to the degree that rather than adding more data, the challenge relies on deciding which features to drop. Even though this may not appear to be a significant problem, the selection of specific features greatly depends on the type of analysis or model that needs to be constructed and, therefore, a filtering of this sort can't happen beforehand.