What kind of problem,

Project BMI

Initial Data cleaning:

1. Standardizing data
2. Checking data shape
3. Checking for null values
4. Dropping columns that are not needed
5. Removing outliers from data

Analyzing data:

Df1 = Started by choosing columns I wanted to work with to analyze the data to answer some of the questions below:

Best player OVA - Messi

Best total stats – L. Suarez

Best goalkeeper – M.Neuer

Fastest player – Adama Traore

Most goals – D.Kulusevski

I was curious to compare the difference between Messi who is best OVA and Suarez who has best total stats. Messi is best OVA but in comparison to Suarez, Suarez is a better defender and is more powerful and may know the game a little better than Messi.

Best goalkeeper is M.Neuer who is 6’4 and weighs 203lbs. He is not the tallest or heaviest but comparing his total stats with player T.Courtois who is 6’6 and 212lbs, Neuer is significantly better at skill level and mentality than he is with strength. He is also a more aggressive player.

BMI dataset:

I then changed the dataset to only show columns that would have been more significant to BMI. I wanted to see if the BMI of the athlete influenced their game. I removed the null values and standardized the data in the data set.

I changed the height and weight column to integers so I could calculate BMI. I’ve cleaned the data by removing outliers and standardized the data by using log or square transformation if applicable.

Checked with a heatmap if there was any correlation between BMI and the other columns. There is only a strong correlation to weight which is to be expected otherwise there is a very weak positive correlation to OVA, power and age. I would consider that BMI does not have a strong correlation to athletes overall performance or doesn’t affect specific performance stats.

Training model:

Gender was removed from dataset because all results in column are male. Which leaves me with only 3 categorical columns.

Scaled the numerical features. Then proceeded to drop dummy columns from categorical variables.

Encoded the categorical variables and then concatenated the numerical and categorical.

Next I trained the model for linear regression to see if there a relationship between two variables.