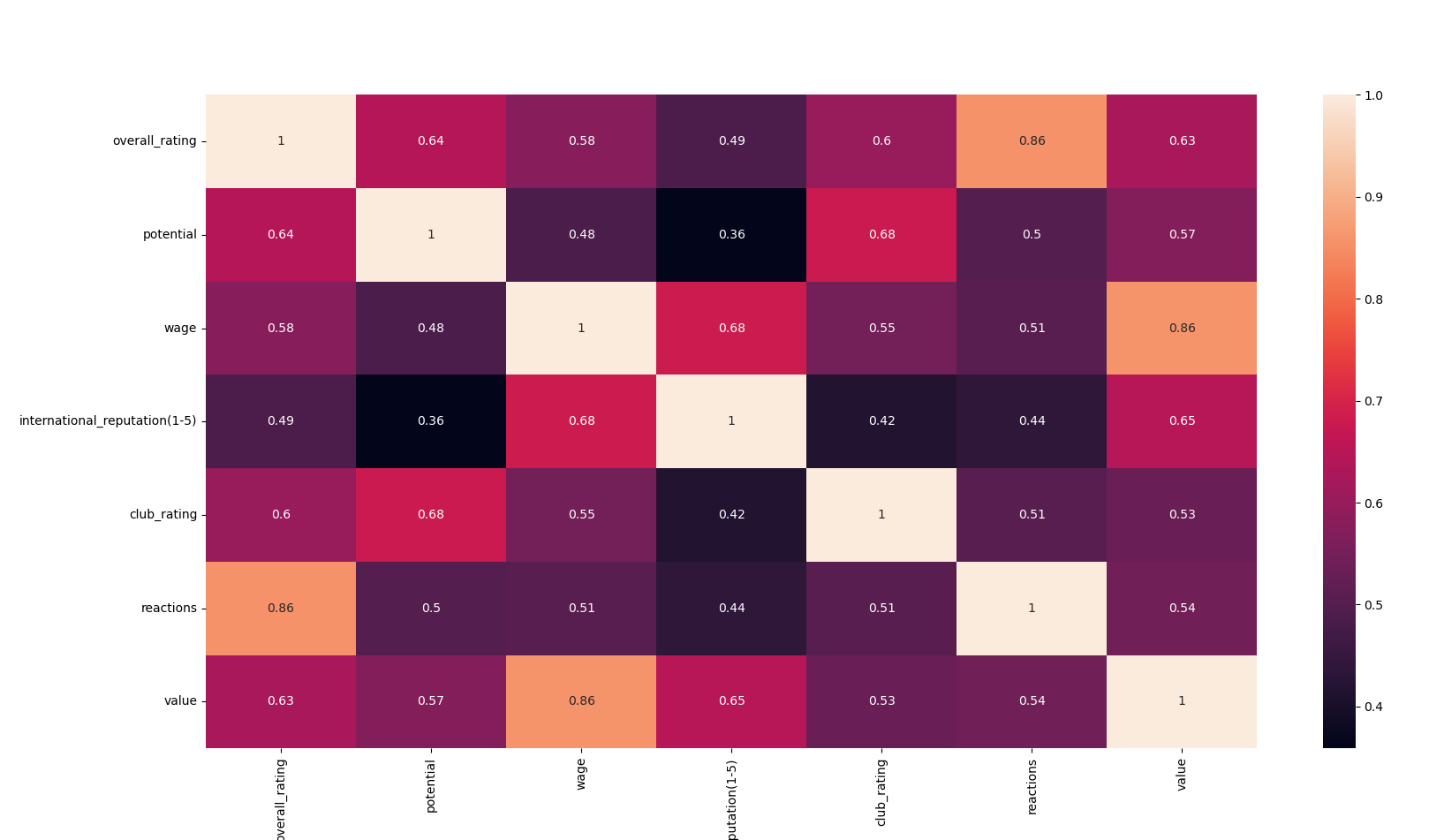
**Analysis:**

* The data consists of
  + - multiple features for a player that affect his value in the game, such as his club’s information (jersey number, position, etc.).
  + his national team’s information (rating, position, etc.).
  + and normal required information about the player such as (age, positions, nationality, etc.).
* The first three columns in the file (id, name, full name) were dropped early on since they have no effect on the value.
* The fourth column (birthdate) was dropped at the start; due to the player’s age existing in the file, which means it wouldn’t have affected the player’s value.
* Analysis is applied by dropping the features which have more than 10% missing values.
* Rows with missing values are then dropped which are approximately 300 rows (which is considered almost 2% of the whole data).

**Preprocessing:**

* Label encoding was applied to multiple features such as: nationality, preferred foot, body\_type, club\_team, club\_position through a built-in function.
* Work\_rate: We applied ordinal encoding to the column by manually mapping the values such that: High – 3, Medium – 2, Low – 1. The values are then summed since the players have two value per row.
* Positions: We applied One Hot Encoding by dividing each distinct values to columns.
* **Feature Selection:**

Feature Selection was applied depending on Correlation between features and value, we chose the top features depending on which of them has Correlation greater than 0.5.

The top features were (Overall rating, wage, potential, international reputation [1-5], club rating, reactions)

* ***Feature Scaling*** was applied to the columns of the top features to have their values in range of [0,1].

**Regression:**

* Polynomial regression and multi variable linear regression were applied on the data.
* The size of the test data was 20% while the training data was 80% as it was the most appropriate percentage to get best results.
* We changed the value of Random\_state to (120) as it gave the least MSE with the two mentioned regression techniques.
* With Polynomial Regression we chose its degree = 3, it gave the best results, and any higher or lowers degrees gave us higher MSE and lower accuracy.
* The difference between each model is that Polynomial regression gave the best results since it had the least MSE and the highest R2 accuracy score (0.94) but it had higher training time. While with Linear Regression had a higher MSE and had an R2 accuracy of (0.79) with lower training time.

Text

Description automatically generated

**Conclusion:**

* We thought it would be better to start by preprocessing all the features in the data, but later we realized it wasn’t the best idea since many of those features wouldn’t be used later the code, so we decided to start by analyzing the data to drop some features and work on preprocessing the remaining ones later.
* During the analysis phase, we chose to drop columns with 10% missing values at the start, then drop rows with any missing values which ended up being over (3000) rows which is 20% precent of the data, so we lowered the percentage of the missing value in the columns to 10%.