FIFA19 Data Wrangling Report Summary

In order to prepare my data for analysis I performed the following data cleaning steps to ensure my data are as analysis friendly as possible. After exploring the header, I removed the following attributes -- "ID', 'Photo', 'Flag', 'Club Logo' , 'Loaned From', 'Real Face'-- due to their repetitive nature. Meaning, they do not provide additional information. I cleaned the very important 'Value and Wage' columns by converting them from categorical (100k) to numerical (100,000). I did the same by removing the "lbs" from the weight column. Workrate required splitting into two unique columns as it included dual information in one column. The column 'Contract Valid Until' was inconsistent as some included months while others did not. I converted these dates to only years to keep them uniform because specific monthly contract dates would not influence the player's value. There were also some minor issues such as players' height which I converted to numerical values by converting to inches only.

There were some missing values that also needed to be examined. The 'Loaned From' column had many missing values, therefore I removed this column completely as this attribute is not important to begin with. Goalkeepers had missing values for field player columns. I replaced these with zeros as goalkeepers spend most of their time in their box. Some of the players had missing "Club" values which are replaced with "Freelance" to make it more complete and intuitive. In addition, 48 missing values from most categories. Depending on their data type I filled them with either mean, mode, or median. Because there is only 48 of them, even if these fillers are not precise, they won't affect the model significantly.

Regarding outliers, they will stay in the dataset as they provide valuable insights in player market value. It is important to look at those outliers and explore why some of those players have such high market values. Despite this, I still explored the number of outliers using the interquartile range. I found that there were 2031 outliers for 'Wage' and 2487 outliers for 'Value'.