For this project, we will be looking at FIFA 22 players’ dataset which has all the soccer related attributes along with values (Euro) & wages (Euro) for each FIFA registered player. There were 89 attributes in total and 19260 rows in this dataset.

As part of exploratory data analysis phase, I tried to look at the distribution of age, growth and playing positions of the players.

Below are a couple of graphs to visualize my findings -

***Players growth by Age:***

Chart, bar chart

Description automatically generated

As players grow older, their growth rate declines.

***Players’ growth by position:***

Chart, bar chart

Description automatically generated

GK has the highest maximum growth, but CAM has highest average growth.

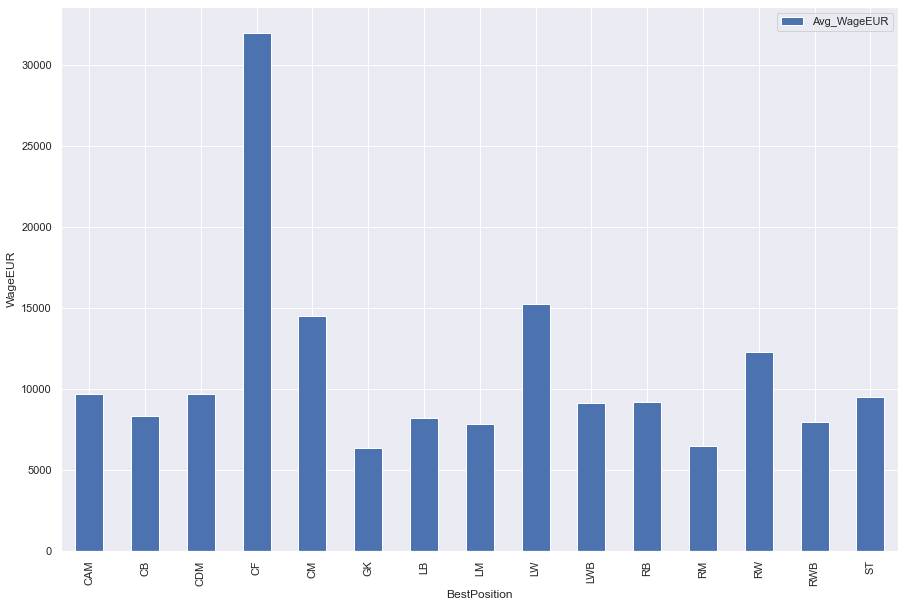
***Players’ values by position:***

Chart, bar chart

Description automatically generated

CF (Center Forward) & LW (Left Winger) tend to have higher values (Euro in millions) in the transfer market.

***Players’ wages by position:***



CF (Center Forward) & LW (Left Winger) tend to have higher wages (in Euro) as well.

***Problem statements:***

* Does position have an impact on the wage and the valuation of a player?
* Does growth have an impact on the wage and the valuation of a player?

***Findings:***

Based on my findings, position has little to no impact on wage or valuation of a player (correlations < .02). Growth might have little impact on wage or valuation of a player (correlations <.2 but >.1). But it is strongly & negatively correlated with Age (-.86). So, in a way age also very little impact on either value of wage of a player.

***Feature selection:***

I performed feature selection using f\_classif method which suggested dropping attributes with very little variances. Thus, I brought the total number of attributes down to 26. Furthermore, I used DBSCAN (used knee method to find eps) to identify outliers. There were 19 outliers, but I did not drop them for clustering purposes as they were not that drastic.

***Unsupervised K-means Clustering:***

Using the elbow method, I was able to find that the optimum number of clusters for this dataset should be 4.

Chart, scatter chart

Description automatically generated

Cluster 0 had only goal keepers. Cluster 1 had players with good playmaking abilities which included midfielders and wing backs. Cluster 2 had players with good defending abilities. Cluster 3 had players with good goal scoring abilities.

***Modeling & prediction:***

I was looking for a model that can predict the best position for a player based on all the selected features. I created a categorical attribute which basically represents the best positions of each player with an integer number. This attribute was passed as predicted variable for decision tree classifier. This model had only 62% accuracy.

Then I used all the original attributes for random forest classification model, which had 72.461% accuracy. After being hyper tuned using GridSearchCV, the model gave me 72.731% accuracy.