Published in Analytics Vidhya

HARSHITA GARG Follow

Feb 21, 2021 · 8 min read ★ · Listen

FIFA19 dataset analysis

How do different factors affect the wages of players in the FIFA19 game dataset

FIFA is the Fédération Internationale de Football Association and FIFA 19 is part of the FIFA series of football video games. It is one of the best-selling video games of all times and is extremely popular among children and adults alike.

wages? Secondly, does the height, weight, or position of players affect their wages? Which clubs spend the maximum amount of money on their players? And finally, players of which countries get paid the most amount of money?

Data Preparation

We start with reading the .csv file and putting it into a data frame. There are nearly 18,000 rows and 89 columns in the dataset. Carefully looking at the data frame reveals that nearly 48 rows have missing data for around 50% of the variables. These rows were removed from the frame.

```
#Read data into a frame
import pandas as pd
import numpy as np
data = pd.read_csv("C:/Users/~/data.csv")
print(data.head())
#remove rows with many null values. 48 such rows are deleted
clean_data = data.dropna(how='all', subset=['Positioning', 'Vision',
'Marking', 'StandingTackle'])
clean_data.shape
```

The wages column has values stored in the format ‘€565K’. This column is modified to get rid of the ‘€’ sign and ‘K’ and converted to an integer value. Similar preprocessing is done on the Value column. Also, the height is present in the data frame in inches as ‘5'7’. This is converted to centimeters. And weight, given like ‘159lbs’, is converted to integer value after getting rid of ‘lbs’.

```
#convert height into cms
Height = ['0\'0' if isinstance(val, float) else val for val in
clean_data['Height']]
Height = [(float(val[0])*12 + float(val[2:]))*2.5 if len(val) > 2 else
0 for val in Height]
```


```
#convert Weight to number
wt = clean_data['Weight'].str.replace(r'\D+', '')
clean_data['Weight'] = wt.astype(int)

#convert the value column to number
def convert_scale(value):
    if value.endswith("M"):
        return float(value[:-1]) * 10**6
    elif value.endswith("K"):
        return float(value[:-1]) * 10**3
    else:
        return float(value)

clean_data["Value"] = clean_data["Value"].str.replace('€',
'').apply(convert_scale)
print(clean_data["Value"])
```

1. Highest earning players

Now let's begin the analysis of the dataset by first looking at the details of the highest paid players.

```
#Details of the Heighest paid players
clean_data.nlargest(10, 'Wage')[['Name','Club', 'Wage', 'Overall',
'Potential', 'Nationality']]
```


	Name	Club	Wage	Overall	Potential	Nationality
0	L. Messi	FC Barcelona	565000	94	94	Argentina
7	L. Suárez	FC Barcelona	455000	91	91	Uruguay
6	L. Modrić	Real Madrid	420000	91	91	Croatia
1	Cristiano Ronaldo	Juventus	405000	94	94	Portugal
8	Sergio Ramos	Real Madrid	380000	91	91	Spain
4	K. De Bruyne	Manchester City	355000	91	92	Belgium
11	T. Kroos	Real Madrid	355000	90	90	Germany
36	G. Bale	Real Madrid	355000	88	88	Wales
5	E. Hazard	Chelsea	340000	91	91	Belgium
32	Coutinho	FC Barcelona	340000	88	89	Brazil

10 highest earning players

Here are the details of top 10 players who are paid the maximum wages in the game. Messi is the top earning player and he also has the maximum overall and potential rankings.

One thing worth noticing is that 3 out of the 10 highest paid players belong to FC Barcelona and 4 out of top 10 earning players belong to the club Real Madrid.

Similar analysis of the details of players with top 10 Overall ranking is given below.

```
#Details of the Heighest paid players
clean_data.nlargest(10, 'Overall')[['Name','Club', 'Wage', 'Overall',
'Potential', 'Nationality']]
```

	Name	Club	Wage	Overall	Potential	Nationality
0	L. Messi	FC Barcelona	565000	94	94	Argentina
1	Cristiano Ronaldo	Juventus	405000	94	94	Portugal
2	Neymar Jr	Paris Saint-Germain	290000	92	93	Brazil
3	De Gea	Manchester United	260000	91	93	Spain
4	K. De Bruyne	Manchester City	355000	91	92	Belgium
5	E. Hazard	Chelsea	340000	91	91	Belgium
6	L. Modrić	Real Madrid	420000	91	91	Croatia
7	L. Suárez	FC Barcelona	455000	91	91	Uruguay
8	Sergio Ramos	Real Madrid	380000	91	91	Spain
9	J. Oblak	Atlético Madrid	94000	90	93	Slovenia

Top 10 players by Overall Ranking

We can see in this table that Neymar and De Gea are amongst the top 10 players in terms of Overall and Potential ranking, but they are not amongst the highest paid players. J. Oblak's Overall ranking is 90, which is very close to some of the other players and Potential ranking as 93, which is better than a lot of highest paid players, he still gets paid a lot less than some of the other players.

2. Club wise analysis

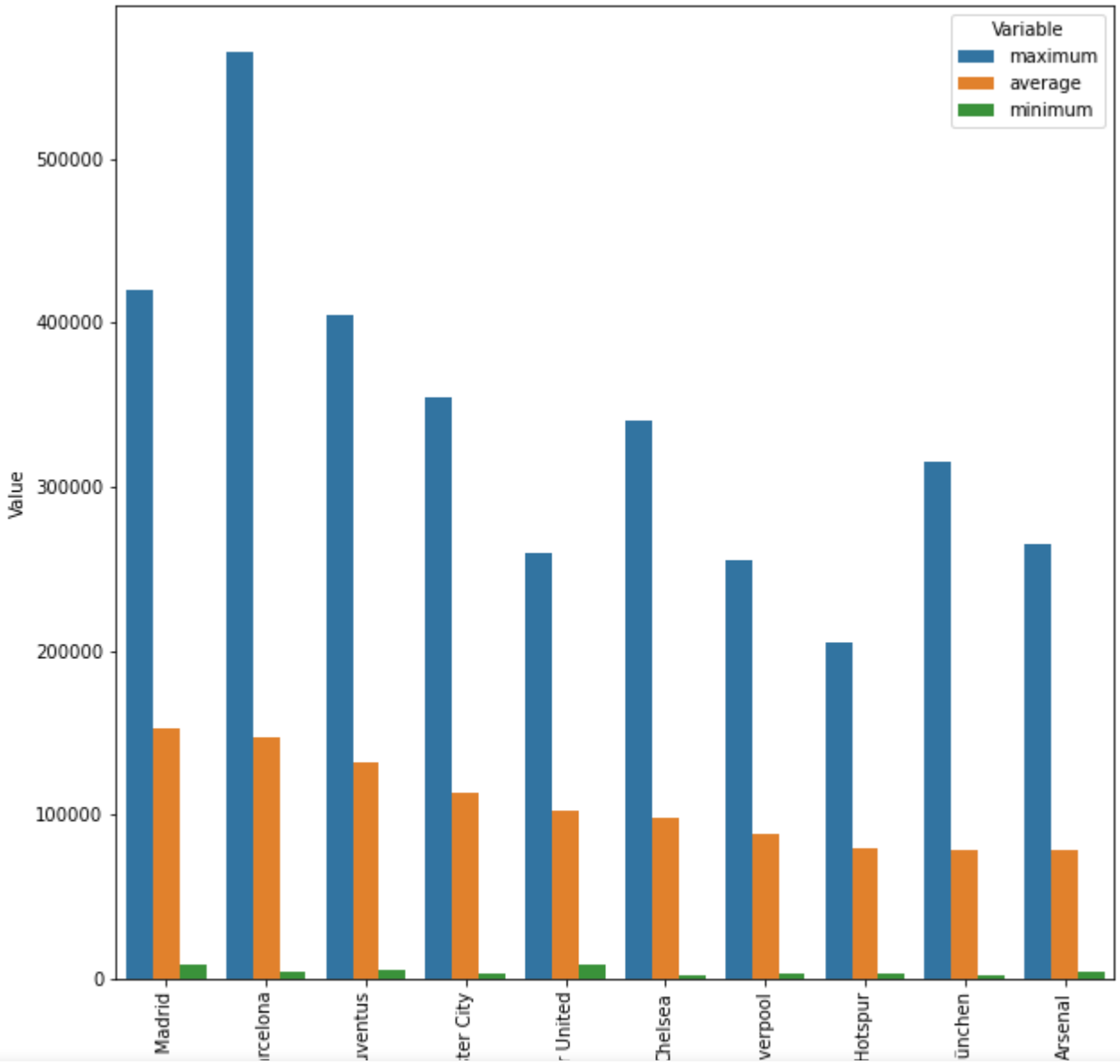
There are total 651 clubs in the dataset. Let's see which clubs spend the most amount of money on the wages of their players. I tried to create a bar plot between the maximum amount of money spent by the clubs on their players, minimum and average amount of money spent on the wages of players by each club. Top 10 clubs by the decreasing mean wages are:

```
clubs = clean_data.groupby(["Club"], as_index= False)
['Wage'].agg(['max','mean','min'])
clubs.columns =
```

```
#Get data for top 10 clubs
clubs_top10 = clubs.nlargest(10, 'average')

#Draw the bar plots
import seaborn as sns
import matplotlib.pyplot as plt

fig, ax1 = plt.subplots(figsize=(10, 10))
plt.xticks(rotation = 90)
tidy = clubs_top10.melt(id_vars='Club').rename(columns=str.title)
sns.barplot(x='Club', y='Value', hue='Variable', data=tidy, ax=ax1)
```



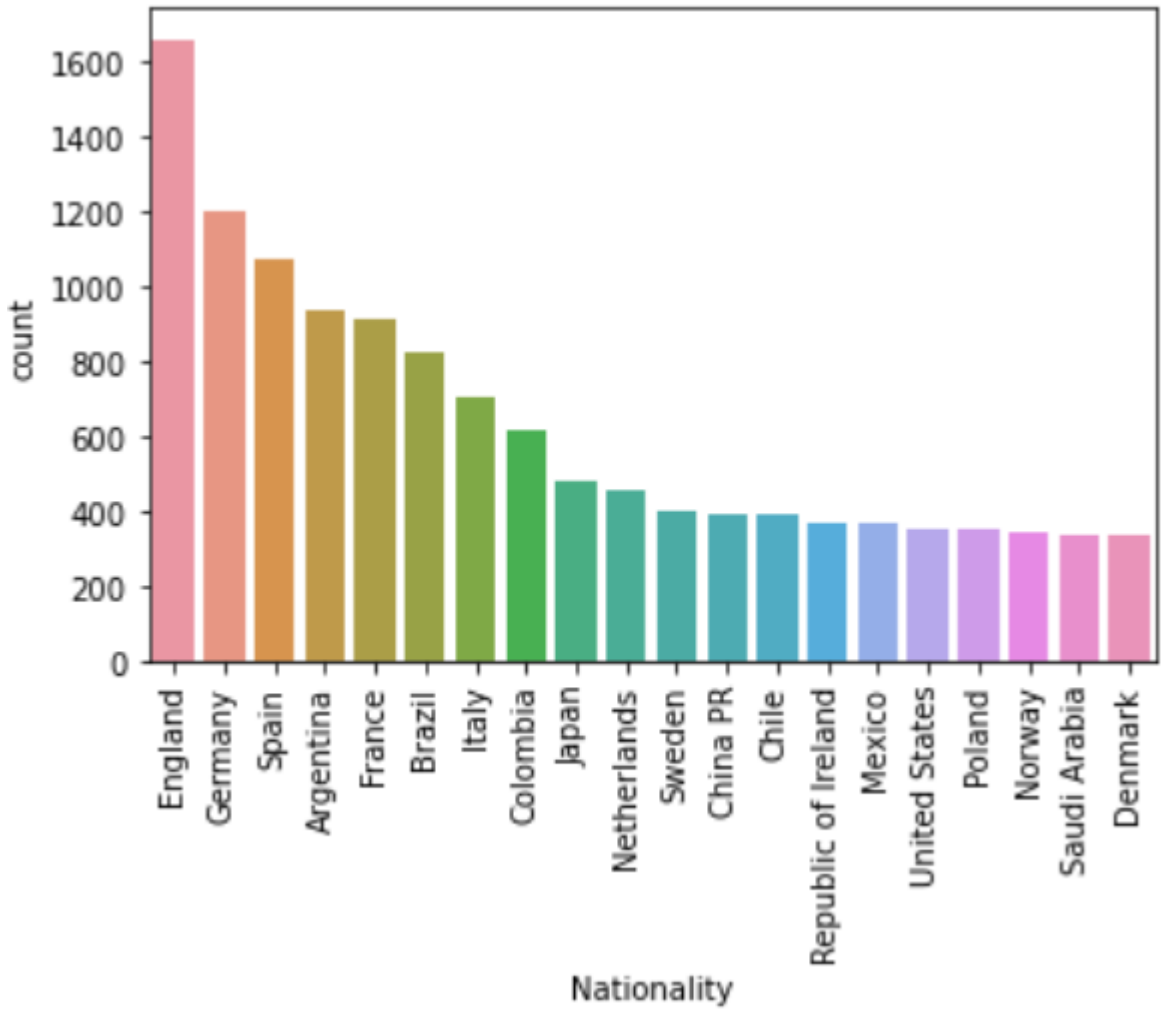
Notice the difference between the wages of highest paid and the lowest paid players. The mean salaries are also way below the maximum wages, showing that only a handful of players get paid the most amount of money.

Clubs spending most money on their players are Real Madrid, FC Barcelona, Juventus, Manchester City and Manchester United. Their mean values vary a lot, showing that the club of a player is an important deciding factor in their wages.

3. Country wise analysis

There are players from 164 different countries in the dataset. Top 20 countries, that contribute most number of players in the dataset are given in the plot below.

```
sns.countplot(x = 'Nationality', data=clean_data,
order=clean_data['Nationality'].value_counts().iloc[:20].index)
plt.xticks(rotation = 90)
```

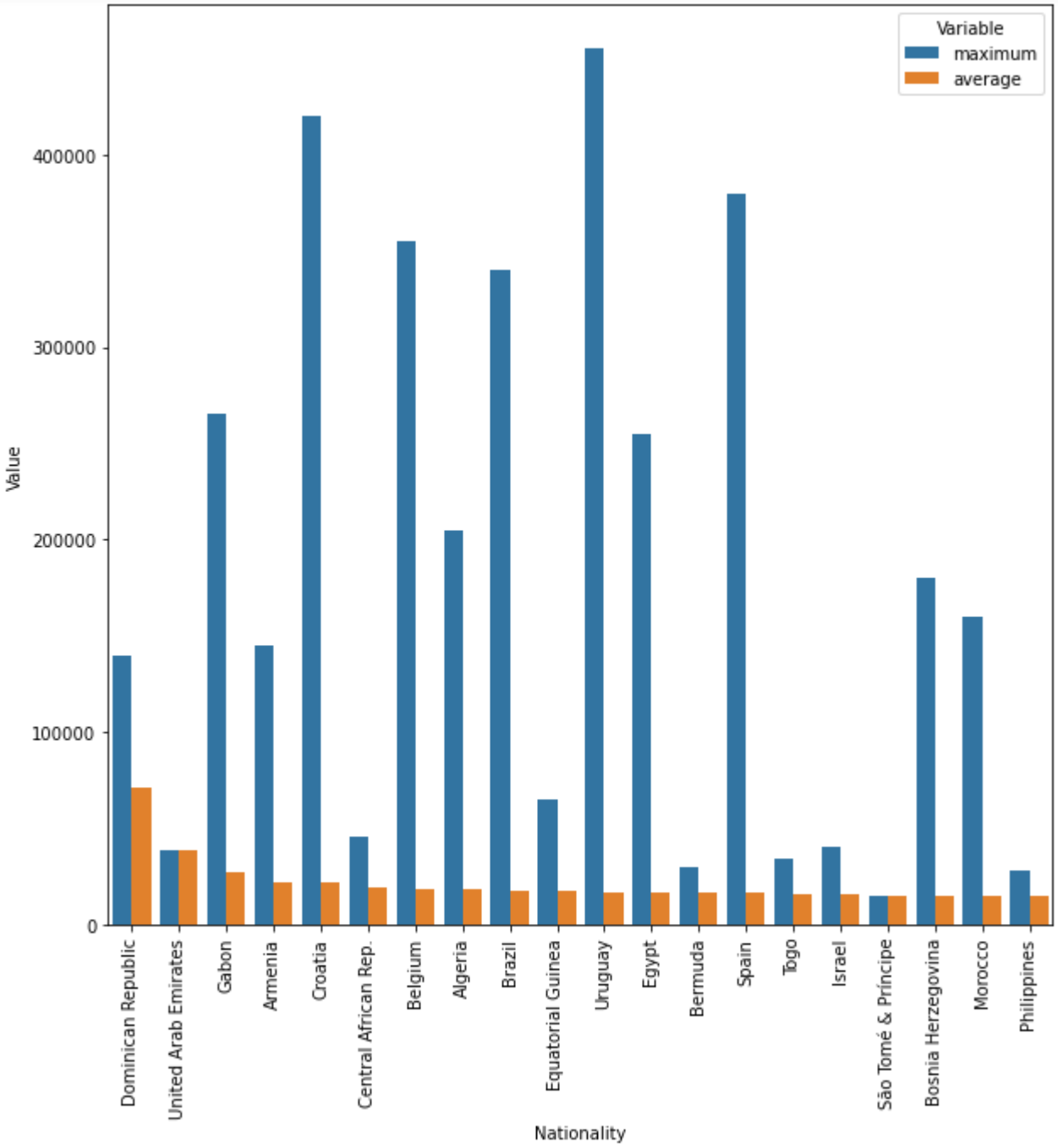



Top 20 countries where most players come from

A bar chart is then drawn to see the maximum and mean wages earned by the players of different countries, arranged in the descending order of mean wages.

```
countries = clean_data.groupby(["Nationality"], as_index= False)
['Wage'].agg(['max','mean'])
countries.columns = (countries.columns.str.replace('max','maximum')
                    .str.replace('mean','average'))
countries = countries.reset_index()
countries_top20 = countries.nlargest(20, 'average')

fig, ax1 = plt.subplots(figsize=(10, 10))
plt.xticks(rotation = 90)
```



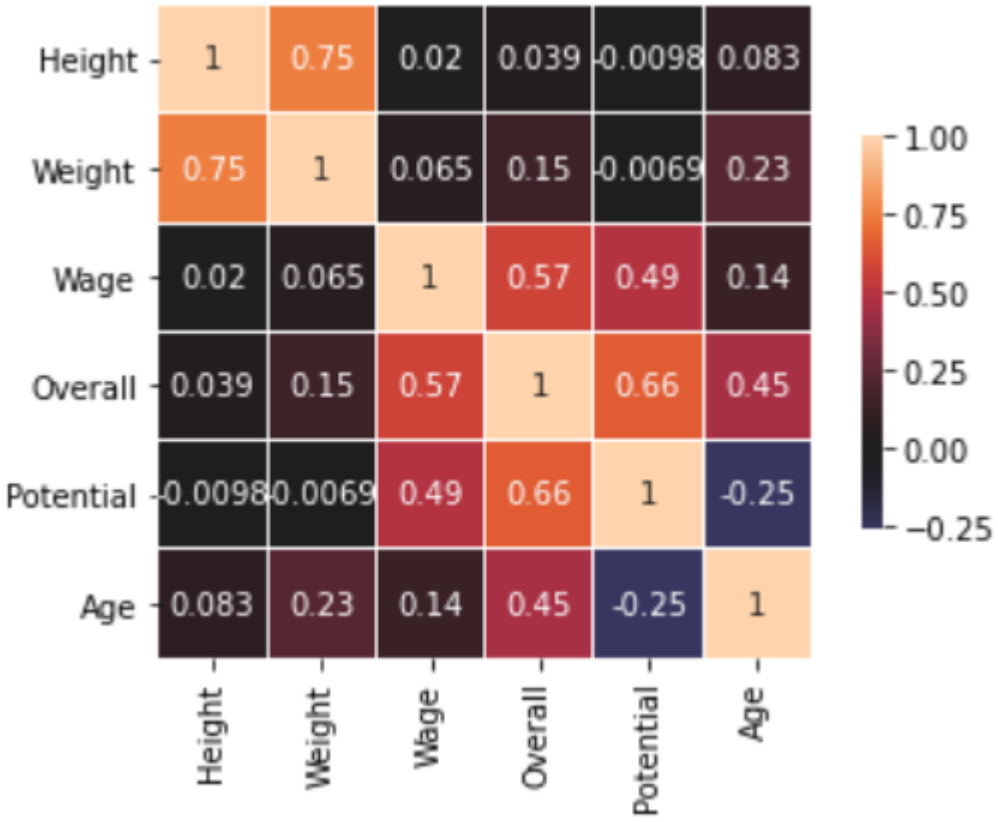
Country wise maximum and mean wages of players

This plot once again, highlights the difference between the highest salary and mean salaries of the players. This plot also shows that value of mean wages could be affected by the outliers.

Players from different countries have different mean wages, indicating that the

In order to study the effect of various factors on the wages of a player, let's draw a correlation plot between the variables- Height, Weight, Age, Wage, Overall ranking and Potential ranking.

```
selected_attr = ['Height', 'Weight', 'Wage', 'Overall', 'Potential',
'Age']
corr = clean_data[selected_attr].corr()
sns.heatmap(corr,  center=0, annot=True,square=True, linewidths=.05,
cbar_kws={"shrink": .6})
```



Correlation plot of different variables

We can draw following conclusions from this correlation plot:-

i. As expected, there is a strong positive correlation between height and weight, meaning

iii. Height, weight and age does not seem to affect a player's wages much.

iv. There is a positive correlation between potential and overall ranking.

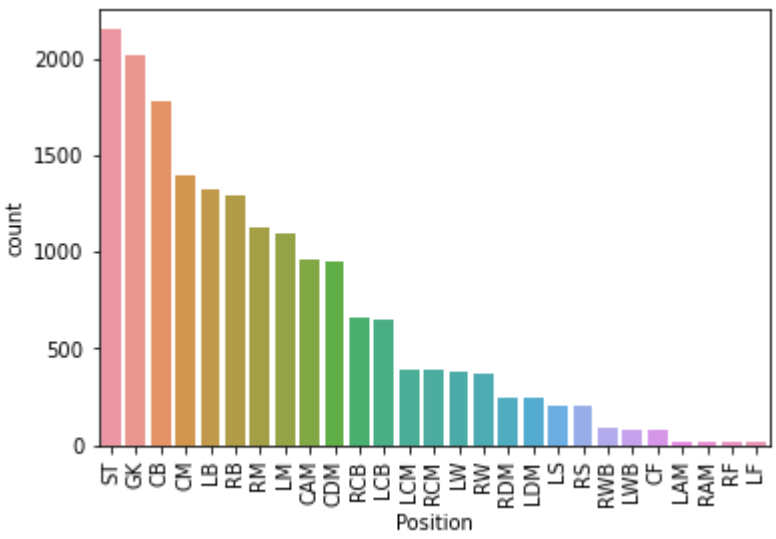
v. There is a small negative correlation between potential ranking and age of a player, meaning as the age increases, potential ranking decreases.

5. Effect of player position on his wages

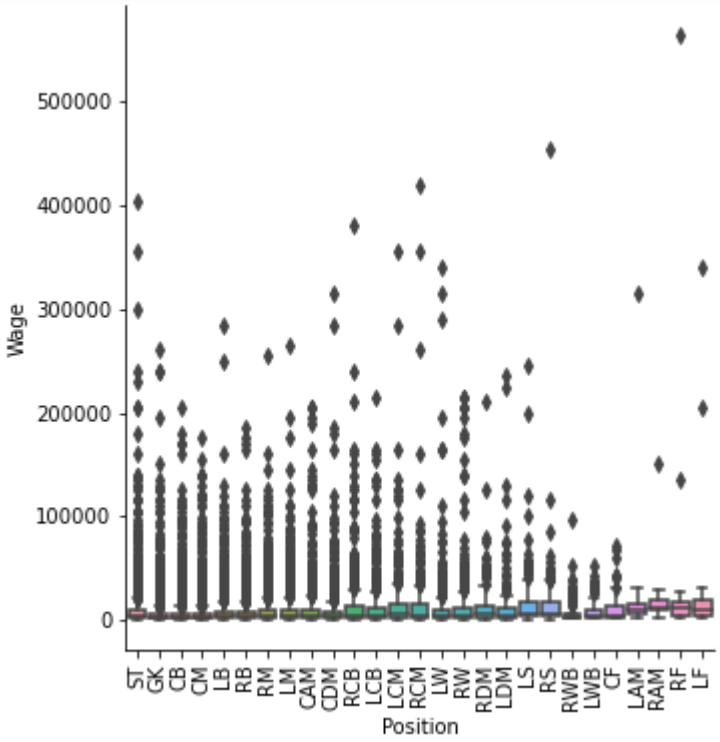
For analyzing how player position affects his wages in the game, we have drawn 2 plots. Bar graph for the number of players playing at each position and box plots for wages of players at every position.

```
#does position affect overall rating and Wage
import matplotlib.pylab as plt
sns.countplot(clean_data['Position'], order =
clean_data['Position'].value_counts().index)
plt.xticks(rotation=90)

sns.catplot(x = 'Position', y = 'Wage', data = clean_data, kind =
'box', order = clean_data['Position'].value_counts().index )
plt.xticks(rotation=90)
```



Number of players playing at each position



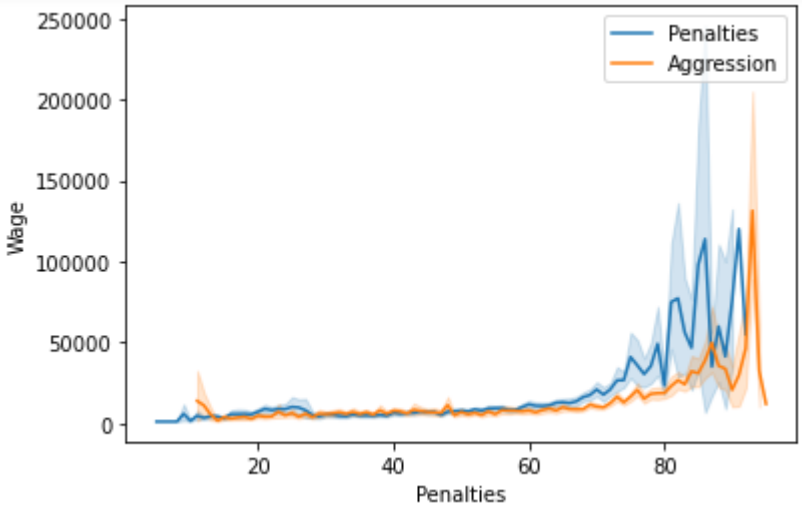
Wages of players playing at different positions

The boxplots for players playing at different positions have quite a few outliers for the value of wages. But we can notice a trend by observing both the adjoining plots that as the number of players playing at each position is decreasing, their mean wage (showed by the line inside the box plot) in increasing. This means that players playing at less common positions like LF and RF get paid more than the ones playing at the more common ones like ST and GK.

6. Effect of aggression and penalties on wages of players

Next, line plots are drawn between the aggression, penalties and wages of players.

```
sns.lineplot(x= "Penalties", y = "Wage", data = clean_data)
sns.lineplot(x= "Aggression", y = "Wage", data = clean_data)
plt.legend(labels = ['Penalties', 'Aggression'])
```



Penalties and Aggression by Wages

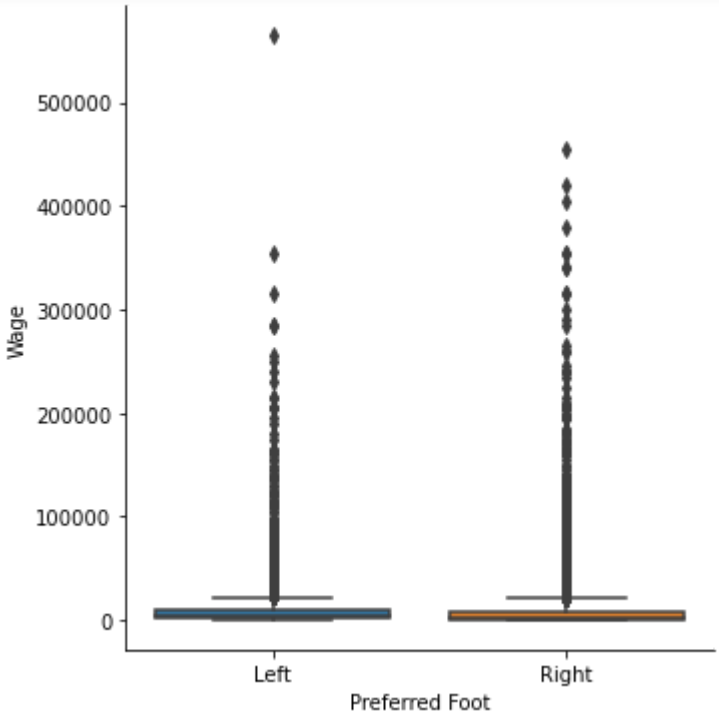
In this plot we can not see any strong relationship between penalties and wages. Wages seem to be going up with penalties, but the trend is not uniform and keeps changing.

Similarly, for aggression, the plot peaks and dips without showing a clear strong relationship. So we can conclude that there is very weak or no relation between either aggression and wages or penalties and wages.

7. Effect of preferred foot on wages of a player

Lastly, I drew boxplots to see the effect of preferred foot on the wages of a player.

```
sns.catplot(x = "Preferred Foot", y = "Wage", kind = "box", data =
clean_data)
```



The long whiskers on the top of the boxplots show the presence of outliers. The median value of wages inside the box seems to be approximately the same for both the values of preferred foot. So we can say that preferred foot does not affect the wages of a player much.

Conclusion

After the detailed analysis done above we can conclude that the following factors affect the wages of players in the FIFA19 game- Their club, nationality, overall ranking and position they play on. Age, height, weight, penalties, aggression and preferred foot have little or no effect on the wages of the players.

Detailed Python code for the plots plotted here could be found here. If you liked this article, please don't forget to clap and follow. Some of my other popular articles could be