FIFA Player Performance Analyzing

Project Participants

Name	Enrollment No.
Pareshkumar Sharma	185223693104

Synopsis

FIFA is an international association football competition contested by the senior men's national teams of the members of the Federation International de Football Association (FIFA), the sport's global governing body.

For this project I have selected FIFA 19 complete player dataset that contains all the player’s performance data. After detailed analysis this project, the extracted information will help the football team managers to improve their team performance. This performance is also takes important role to select any team or club or player for awarding cup or prizes.

This project contains this analysis modules:-

- Top 10 players of world
- What would be the best team
- Top 10 teams
- 10 poorest player of world
- Most paid Players
- Left and right footed players count, dribbling, ball control abilities

Description

This database contains 88 columns and 18207 rows that contains all player's performance like nationality, position wise performance, his club, strength etc.

This dataset is use full for football manager and any award giving organization that want to know the top players, top football clubs, and top nations.

FIFA is an international association football competition contested by the senior men's national teams of the members of the Federation International de Football Association (FIFA), the sport's global governing body. The championship has been awarded every four years since the inaugural tournament in 1930, except in 1942 and 1946 when it was not held because of the Second World War. The current champion is France, which won its second title at the 2018 tournament in Russia. All football clubs of world are participating in this FIFA tournament.

For this project I have selected FIFA 19 complete player dataset that contains all the player's performance data. After detailed analysis this project, the extracted information will help the football team managers to improve their team performance. This performance is also takes important role to select any team or club or player for awarding cup or prizes.

Plagiarism report

Understanding Data

Column Description and Basic Questions

Analysis on **Nationality** of Players, how many players are coming from each nation and how many nations are from best performance players are coming

Analysis on **Potential** of player, who is the player who has max performance. Top 10 players, Top 10 clubs, Top 10 Nations

Analysis Worlds **football club**, total players in each club, total wage of all players of a club, total clubs from each nation

Analysis on value and **wage** of player, who has maximum wage, wage by nationality, wage by club is came just previous analysis

Analysis **Preferred foot** of player, how many players are playing from left foot, and how many are righty, nation wise player's foot, and maximum left or right footed players.

International **reputation** of player, total players on each category

Week foot of player, player's histogram

Skill moves of the player

How many players have **real face?**

Position of player

Jersey Number of player

Joined date of player on to the club

Loaned From other club

Contact of players.

Height of players

Weight of players

Performance on different **position** of ground like LS (Left Striker), ST (Striker), RS (Right Striker), LW (Left Wing), LF (Left Forward), CF (Center Forward), RF (Right Forward), RW (Right Wing), LAM (Left Attacking Midfield), CAM (Center Attacking Midfield), RAM (Right Attacking Midfield), LM (Left Midfield), LCM (Left Center Midfield), CM (Center Midfield), RCM (Right Center Midfield), RM (Right Midfield), LWB (Left Wing Back), LDM (Left Defensive Midfield), CDM (Center Defensive Midfield), RDM(Right Defensive Midfield), RWB(Right Wing Back), LB(Left Back), LCB(Left Center Back), CB(Center Back), RCB(Right Center Back), RB(Right Back)

Performance on different **task** like Crossing, Finishing, Heading Accuracy, Short Passing, Volleys, Dribbling, Curve, Free Kick Accuracy, Long Passing, Ball Control, Acceleration, Sprint Speed, Agility, Reactions, Balance, Shot Power, Jumping, Stamina, Strength, Longshot, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Marking, Standing Tackle, Sliding Tackle, GK Diving, GK Handling, GK Kicking, GK Positioning, GK Reflexes.

Big Questions: -

- Which nation is paying maximum wage and getting the best performance from players.
- Which are the players who can be best in world and have all abilities as any best footballer should have.
- Which club has best amount of player and which are the club that should be awarded by their players performance.
- Which jersey numbers are famous, what is number of player in each jersey number? Any club has two players with one number of jersey?
- Which player are good in specific action of football? Like they are in range of almost up of 85 or 90 in performance of particular position or action.
- How many players are loaned?
- Total how much wage is invested in all players?

Data Processing and Cleaning

Data is processed using data frames in all languages, but all languages have different way and methods to process the data, in this project I am using both Python and R technologies because both have their own advantages and disadvantages.

For processing I'm using module pandasql in Python. I am storing my processed data using two format one is txt (in some analysis) and second is excel file (in some analysis) using python's xlwt module.

For analysis on accurate data I have removed the 2085 players only when I performed operation on these columns 'RS', 'LW', 'LF', 'CF', 'RF', 'RW', 'LAM', 'CAM', 'RAM', 'LM', 'LCM', 'CM', 'RCM', 'RM', 'LWB', 'LDM', 'CDM', 'RDM', 'RWB', 'LB', 'LCB', 'CB', 'RCB', I removed it because these have null values. After doing it 16122 players are in dataset for analysis.

Apart from this processing i can't find any cleaning process for data till pre visualization, if needed in post visualization I will point it at top of any code for that query.

Minimum Hardware Requirement

Ram	2 GB
Processor	Dual Core 1.60 GHz or Higher
Resolution	1024 X 768 or higher

Software Used

- Python 3.7 (32 bit)
- MS Excel 2016

Module Used in Python

pandas:

It is the module used to read and manipulate data in excel or csv file using dataframe objects. It provides function and classes for manipulate the data in dataframes such as read_excel(), concat(), max(), count(), sum(), groupby() etc.

pandasql:

Pandasql is needed if we want to get data or filter data in a form of sql language. It is providing sqldf() in which we can fetch data and filter it using select statements.

pylab:

In Python pylab is provided by matplotlib module that contains number of methods for draw and save graphs or visualize the data in graphical representation. We can draw simple dot plote, line plot, bar graph, histograms and pie charts using this module.

seaborn:

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. It provide almost everything for data visualization with very good colors and all eye catching representation. We can draw countplot, swormplot, factorplot, heatmap graphs using this module.

Algorithm

Linear Regression

It is the most well-known and popular algorithm in machine learning and statistics. This model will assume a linear relationship between the input and the output variable. It is represented in the form of linear equation which has a set of inputs and a predictive output. Then it will estimate the values of coefficient used in the representation.

Classification Algorithm

In this data is classified in different categories and further analysis done on each category for particular result. For example we can classify players of Fifa 19 in good players or weak players etc.

Collaborative Algorithm

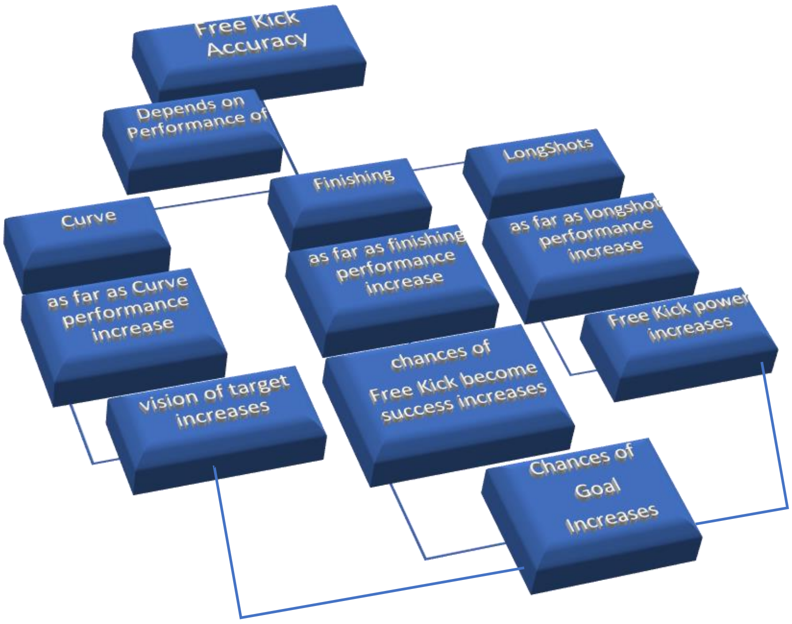
This is used for predicting some output or information based on different inputs are came, for example in University maximum student learning masters in Mathematics from 8 years it's may happen that next year also be max students from mathematics.

Decision Tree:

Decision trees are used for both classification and regression problems, Decision tress often mimic the human level thinking so it's so simple to understand the data and make some good interpretations.

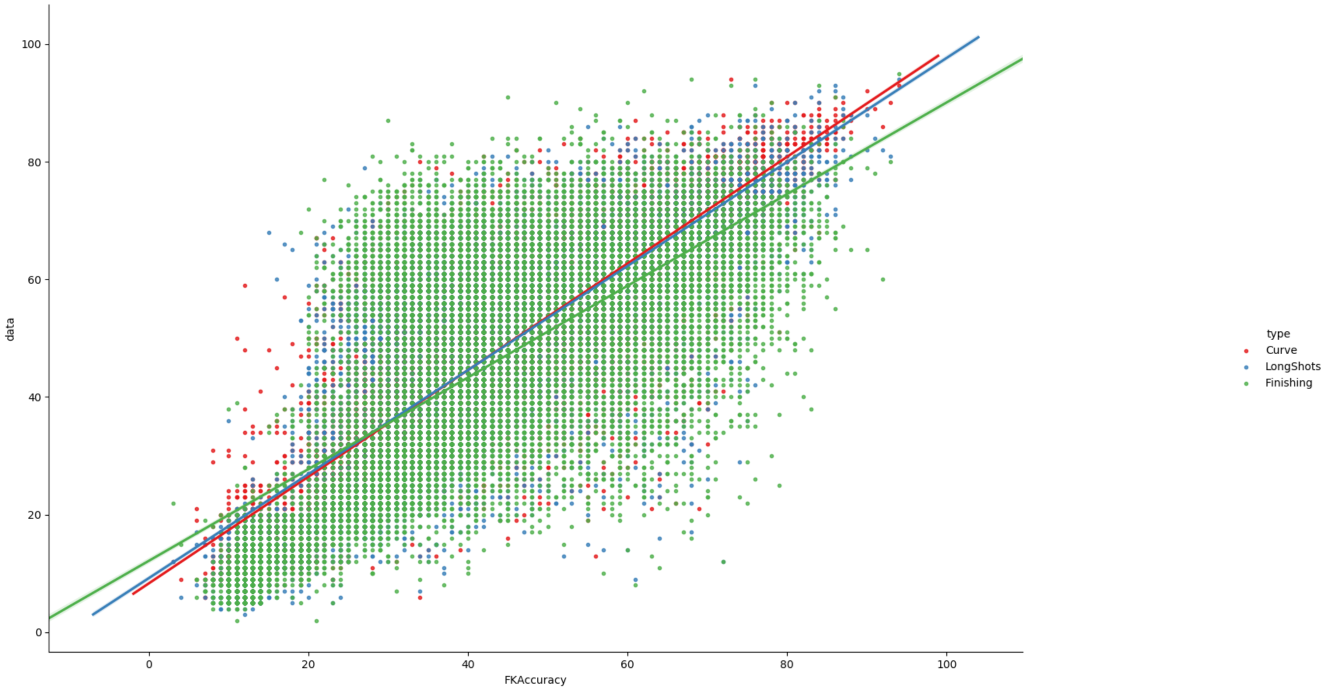
Decision Tree – Making decision about Players Free Kick Accuracy

As this graph represents we can take a decision of Free Kick Accuracy of Player on the performance of player's Curve, Long Shots and Finishing.


Free Kick Accuracy
Depends on Performance of
Curve
Finishing
LongShots
as far as Curve performance increase
as far as finishing performance increase
as far as longshot performance increase
vision of target increases
chances of Free Kick become success increases
Free Kick power increases
Chances of Goal Increases

Python Code

```
Import seaborn as sns
Imprt matplotlib.pyplot as plt
reg = p.read_excel(r"Football Data New
Excel Format.xlsx", "Reggression" )
sns.lmplot("FKAccuracy","data",hue="type",data=reg)
plt.show()
```



Conclusion:

The **Red** dots representing the **vision clarity**, **ORANGE** dots representing the **Power** and **GREEN** dots representing **chances of success** of players while free kick is concerned. The green dots representing players who are equal in these 3 things, in graph players with green dots and having more than 60 FK Accuracy are eligible for taking free kick.

As the lines of Curve, Long Shots and Finishing represents FK Accuracy highly depend on these 3 things.

Pre Visualization

The dataset named Fifa 19 Complete player dataset is downloaded from Kaggle.com.

In this dataset there are many numeric columns, I need to check for Null values, and different values available on that columns like

- Age
- Preferred Foot
- Total Players
- Clubs
- Etc.

For that I performed these following visualizations you can see it from next page.

Player Visualization

- **Total players in world : 18207**

```
#players is data frame taken from pandas.read_excel () function
>>> players["ID"].count()
```

- **Player with left preferred foot: 4211**

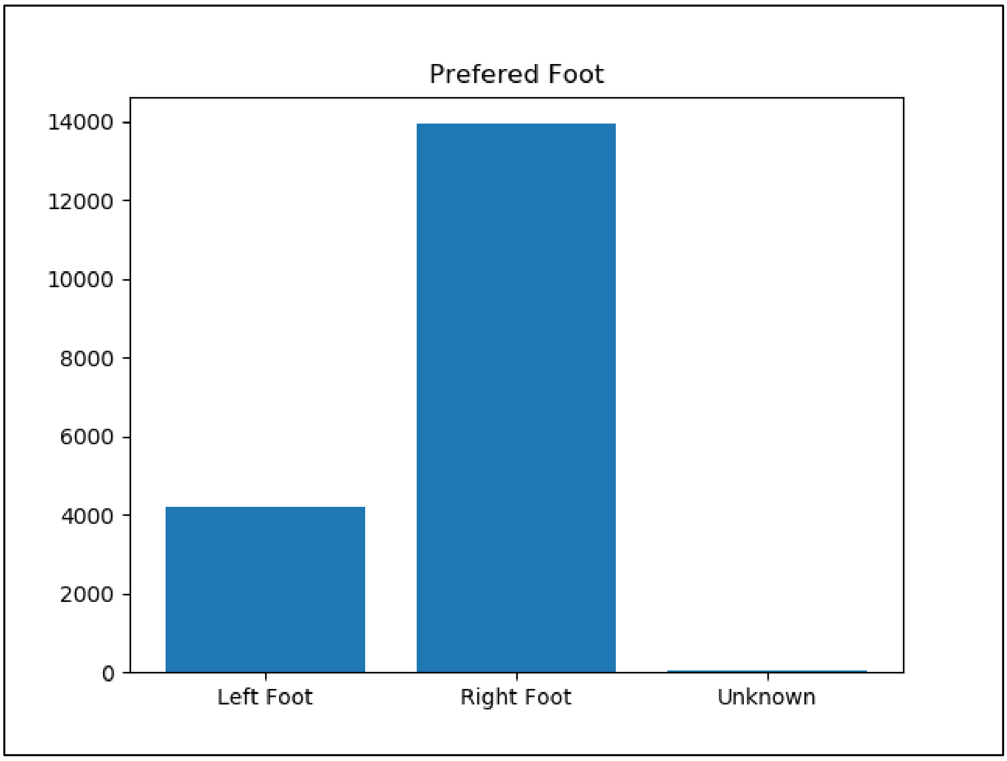
```
>>> players["Preferred Foot"][players["Preferred Foot"]=="Left"].count()
```

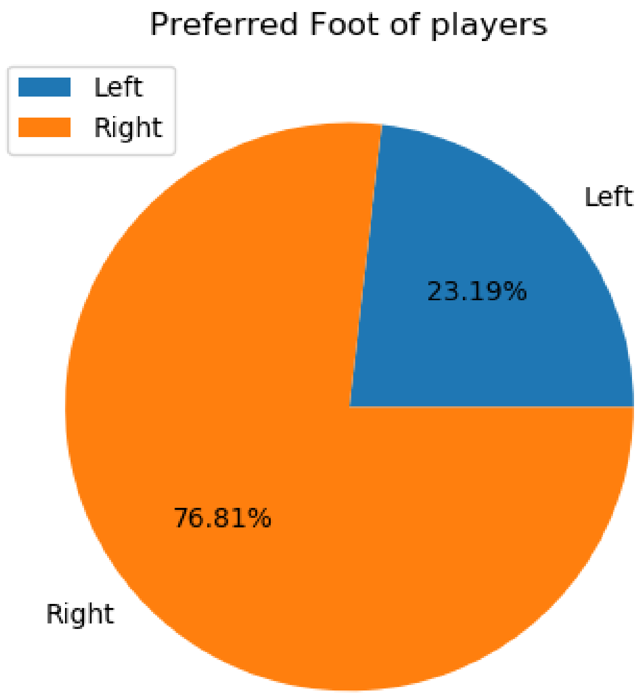
- **Player with right preferred foot: 13948**

```
>>> players["Preferred Foot"][players["Preferred Foot"]=="Right"].count()
```

- **Player with unknown preferred foot: 48**
-

```
>>> players["Preferred Foot"][players["Preferred Foot"].isna()].count()
>>> x
['Left Foot', 'Right Foot', 'Unknown']
>>> y
[4211, 13948, 48]
>>> py.bar(x,y)
<BarContainer object of 3 artists>
>>> py.title("Prefered Foot")
Text(0.5, 1.0, 'Prefered Foot')
>>> py.show()
>>> g.pie([4211,13948],labels=["Left","Right"],autopct="%.2f%%")
([<matplotlib.patches.Wedge object at 0x145731D0>, <matplotlib.patches.Wedge object at
0x145735F0>], [Text(0.8207745153900478, 0.7323449971743047, 'Left'), Text(-
0.8207745496735903, -0.732344958751076, 'Right')], [Text(0.4476951902127533,
0.39946090754962066, '23.19%'), Text(-0.4476952089128674, -0.3994608865914959,
'76.81%')])
>>> g.title("Preferred Foot of players")
Text(0.5, 1.0, 'Preferred Foot of players')
>>> g.legend()
<matplotlib.legend.Legend object at 0x14567B70>
>>> g.show()
```





Conclusion

- ✓ As per the analysis more than 75% of players of the world are having Right as preferred foot

Club Player Visualization

- **Total players which are not part of any club are: 241**

```
>>> len(players["Club"][players["Club"].isna()])
```

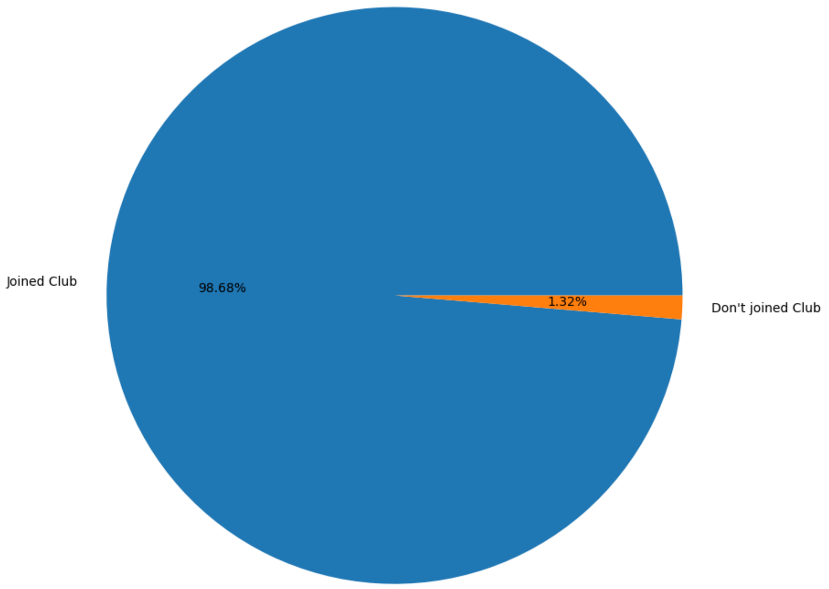
- **Total Players have joined club: 17966**

```
>>> players["Club"].count()
```

- **Total Clubs in world : 651**

```
>>> len(players["Club"].unique())
#it gives 652 but one subtracted which is “nan”
>>> x
['Joined Club', "Don't joined Club"]
>>> y
[17966, 241]
>>> py.pie(y,autopct="%.2f%%",labels=x)
>>> py.title("Player participation in club")
>>> py.show()
```

- **Graph**



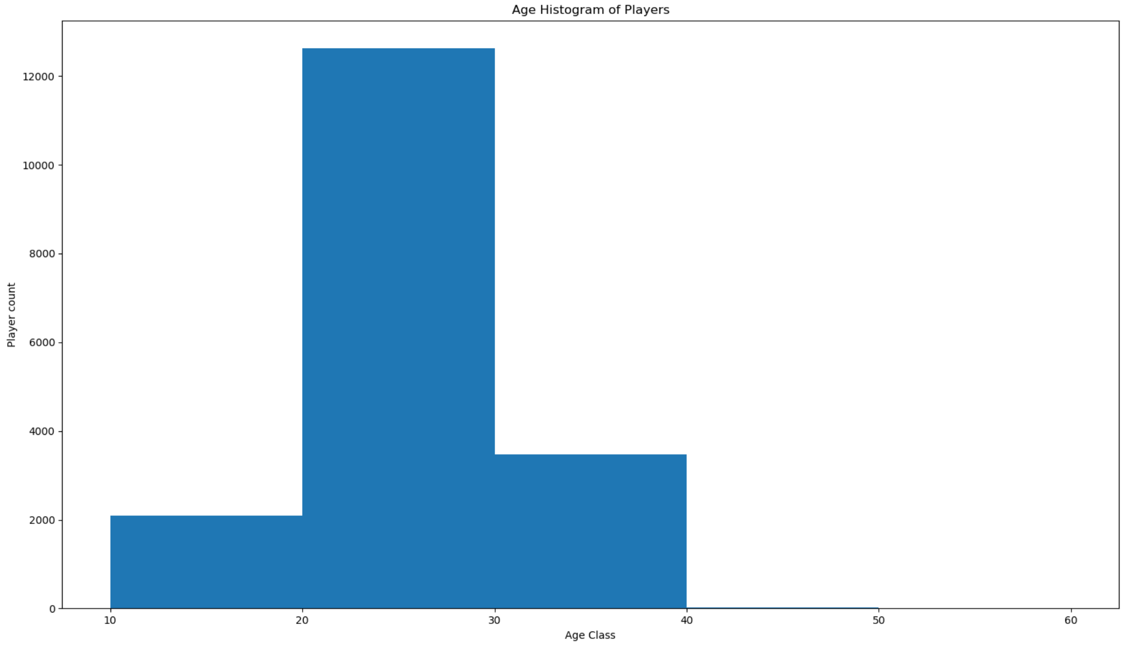
- **Conclusion**

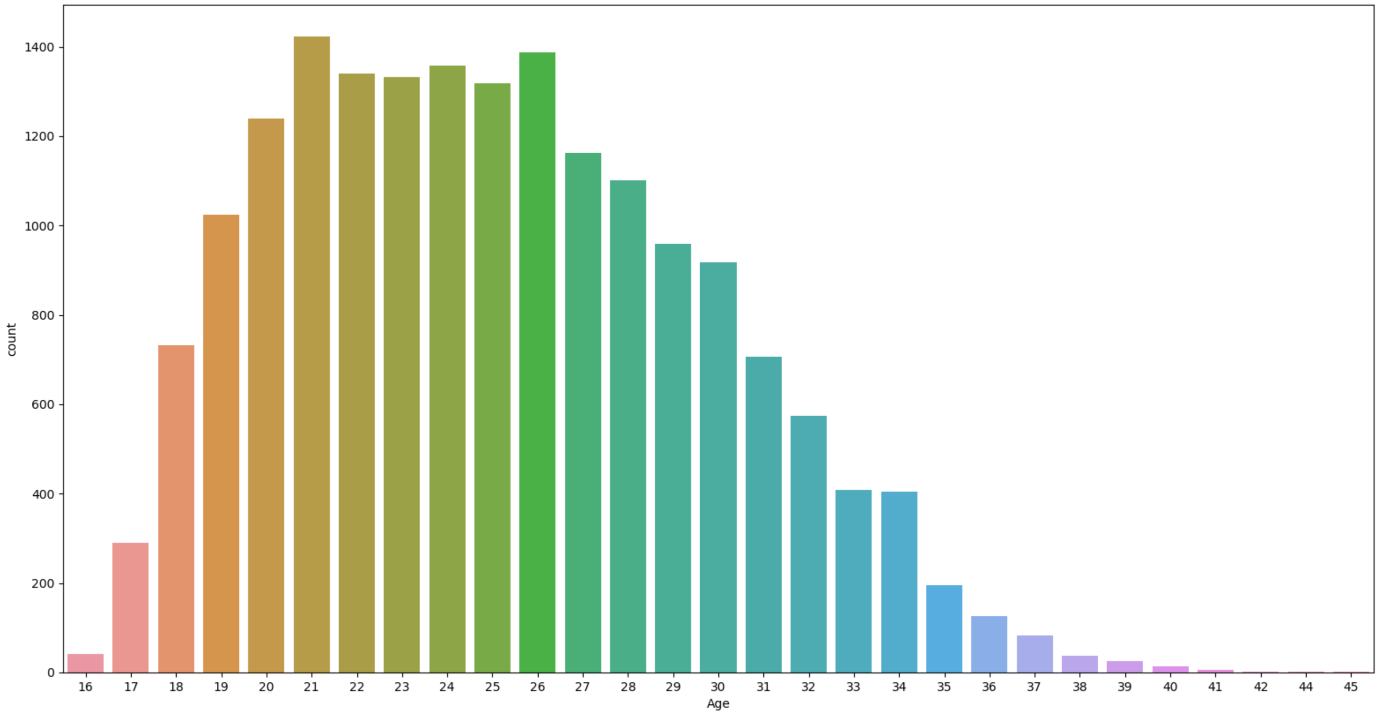
Very less amount of player are who didn’t joined any club. It 98% players joined clubs.

Age Visualization

- **Histogram Graph of Player age**

```
>>> ages = list(players["Age"])

>>> len(ages)

18207

>>> g.hist(ages,[10,20,30,40,50,60])

(array([ 2087., 12621.,  3477.,    22.,     0.]), array([10, 20, 30, 40, 50, 60]), <a list of 5 Patch objects>)

>>> g.title("Age Histogram of Players")

Text(0.5, 1.0, 'Age Histogram of Players')

>>> g.xlabel("Age Class")Text(0.5, 0, 'Age Class')

>>> g.ylabel("Player count")

Text(0, 0.5, 'Player count')

>>> g.show()

#the following is age data

>>> age_count = {}

>>> for x in range(16,46):

    count = 0

    for xx in ages:

        if(x==xx):

            count+=1

    age_count.update({x:count})

>>> age_count

{16: 42, 17: 289, 18: 732, 19: 1024, 20: 1240, 21: 1423, 22: 1340, 23: 1332, 24: 1358, 25: 1319, 26: 1387, 27: 1162, 28: 1101, 29: 959, 30: 917, 31: 707, 32: 574, 33: 408, 34: 404, 35: 196, 36: 127, 37: 82, 38: 37, 39: 25, 40: 13, 41: 5, 42: 1, 43: 0, 44: 2, 45: 1}

>>> max(age_count.values())

1423

>>> min(age_count.values())

0
```





- **Conclusion**

most of players are with age class, between 20 – 30, max age is 45, min age is 16, max amount of players are with 21 age which are 1423

Post- Visualization

- Nationality

Analysis on Nationality of Players, how many players are coming from each nation and how many nations are from best performance players are coming

Q.1 Total nations have participated in Fifa 19.

<u>Python</u>

```
import pandas as p
import pandasql as sql
import pylab as g

# read excel file
nation_data = p.read_excel(r"..\Football Data New Excel Format.xlsx","Nation")

# total records
# output 18207
print(nation_data.count())

# no null value data found
# output 0
print(nation_data["Nationality"][nation_data["Nationality"].isna()].count())

#unique nation
un_na = set(nation_data["Nationality"])
count = len(un_na)
#output 164
```

<u>Conclusion</u>

This year 18207 player are taking participation in Fifa 19 from 164 different nations

Q.2 How many players from each nation?

Python

```
import pandas as p
import pandasql as sql
import pylab as g

# read excel file
nation_data = p.read_excel(r"..\Football Data New Excel Format.xlsx","Nation")

# total records
# output 18207
print(nation_data.count())

# no null value data found
# output 0
nation = nation_data["Nationality"]
print(nation_data["Nationality"][nation.isna()].count())

#unique nation
#output 164
un_na = set(nation)
count = len(un_na)

pl_dict = {}
for u_n in un_na:
    count = 0
    for n in nation:
        if n == u_n:
            count+=1
    pl_dict.update({u_n:count})

# x is nation name
x = list(pl_dict.keys())
# y is player count
y = list(pl_dict.values())
g.bar(list(range(1,165)),y)
##count = 1
##while count < 165 :
##    g.bar(count,y[count-1])
##    count += 1
##g.legend(x)
g.title("Total player from each Nation")
g.savefig(".\\Total_Payers Each Nation.png")
#g.show()
g.close()
g.barh(list(range(1,165)),y)
g.title("Total player from each Nation")
```

```
g.savefig(".\\Total_Payers H Bar.png")
#g.show()
g.close()

# sorting dictionary in python

def srt(d,base="k|v",reversed1=False):
    assert type(d)==dict,"Dictionary argument is required"
    assert base.upper()=="K|V" or base.upper()=="K" or base.upper()=="V", "enter valid \"base\" argument"

    sorted_dict = {}

    if base.upper()=="K|V" or base.upper()=="K":

        for x in sorted(d.keys(),reverse=reversed1):
            sorted_dict.update({x:d[x]})
            print(x)
    else:

        empty_dict = {}

        for x in d:
            empty_dict.update({d[x]:x})

        for x in sorted(empty_dict,reverse=reversed1):
            sorted_dict.update({empty_dict[x]:x})

    return sorted_dict

sorted_nation = srt(pl_dict,"v",False)
file = open("nation_player_ascending.txt","w")
for x in sorted_nation:
    file.write(str(x)+","+str(sorted_nation[x])+"\n")
file.close()
sorted_nation = srt(pl_dict,"v",True)
file = open("nation_player_descending.txt","w")
for x in sorted_nation:
    file.write(str(x)+","+str(sorted_nation[x])+"\n")
file.close()
```