

# FIFA Data Analysis



By: Sanskar Adhikari, Ankit Kafle, Anish Bhurtyal,  
Aayush Shrestha

# Introduction

- FIFA is a non-profit organization that acts as an international governing body of association football, futsal and beach soccer.
- It was founded in 1904 and its headquarterd in Zürich, Switzerland.



## Intro cont

- FIFA is in charge of organizing and promoting association football's biggest international competitions, including the World Cup, and the Women's World Cup.
- For this, it has licensed FIFA soccer game under EA(Electronic Arts) sports.



# Purpose

- The player ratings are one of the components of the FIFA.
- This helps managers pick who to select on any particular game by determining who is the best player on each installment.
- In this project, we aim to explore different factors of a player and analyze if different factors has influence with the overall performance in the pitch.



# Data set info

Data set source:

<https://raw.githubusercontent.com/4m4n5/fifa18-all-player-statistics/master/2019/data.csv>

```
In [3]: raw_df.shape
```

```
Out[3]: (18207, 88)
```

# Overview

```
for col in raw_df.columns:  
    print(col,end='    ')
```

Unnamed: 0	Name	Age	Photo	Nationality	Flag	Overall	Potential	Club	Club Logo	Value	Wage	Spec							
ial	Preferred Foot	International Reputation	Weak Foot	Skill Moves	Work Rate	Body Type	Real Face	Positio											
n	Jersey Number	Joined	Loaned From	Contract Valid Until	Height	Weight	LS	ST	RS	LW	LF	CF							
RF	RW	LAM	CAM	RAM	LM	LCM	CM	RCM	RM	LWB	LDM	CDM	RDM	RWB	LB	LCB	CB	RCB	R
B	Crossing	Finishing	HeadingAccuracy	ShortPassing	Volleys	Dribbling	Curve	FKAccuracy	LongPassing										
BallControl	Acceleration	SprintSpeed	Agility	Reactions	Balance	ShotPower	Jumping	Stamina	Strength										
LongShots	Aggression	Interceptions	Positioning	Vision	Penalties	Composure	Marking	StandingTackle	Sli										
dingTackle	GK Diving	GK Handling	GK Kicking	GK Positioning	GK Reflexes	Release Clause													

# Data Cleaning

```
df = raw_df.drop_duplicates()
```

```
df.drop(columns=['Unnamed: 0', 'Photo', 'Flag', 'Club Logo', 'Real Face', 'Position',  
                'Jersey Number', 'Joined', 'Loaned From', 'Contract Valid Until',  
                'LS', 'ST', 'RS', 'LW', 'LF', 'CF', 'RF', 'RW', 'LAM', 'CAM', 'RAM',  
                'LM', 'LCM', 'CM', 'RCM', 'RM', 'LWB', 'LDM', 'CDM', 'RDM', 'RWB', 'LB',  
                'LCB', 'CB', 'RCB', 'RB', 'Loaned From', 'Release Clause'], inplace = True)
```

```
df.columns
```

```
]: Index(['Name', 'Age', 'Nationality', 'Overall', 'Potential', 'Club', 'Value',  
        'Wage', 'Special', 'Preferred Foot', 'International Reputation',  
        'Weak Foot', 'Skill Moves', 'Work Rate', 'Body Type', 'Height',  
        'Weight', 'Crossing', 'Finishing', 'HeadingAccuracy', 'ShortPassing',  
        'Volleys', 'Dribbling', 'Curve', 'FKAccuracy', 'LongPassing',  
        'BallControl', 'Acceleration', 'SprintSpeed', 'Agility', 'Reactions',  
        'Balance', 'ShotPower', 'Jumping', 'Stamina', 'Strength', 'LongShots',  
        'Aggression', 'Interceptions', 'Positioning', 'Vision', 'Penalties',  
        'Composure', 'Marking', 'StandingTackle', 'SlidingTackle', 'GKDividing',  
        'GKHandling', 'GK Kicking', 'GK Positioning', 'GK Reflexes'],  
        dtype='object')
```

# Analyzing Numerical Data

```
df['Height'].sample(5)
```

```
ID
242385      6'1
213697      6'0
224520      5'8
227813      5'9
211322      5'11
Name: Height, dtype: object
```

```
df['Value'].sample(5)
```

```
ID
174665      €250K
202769      €575K
236506      €1.4M
207862      €15.5M
169388      €575K
Name: Value, dtype: object
```

```
df['Weight'].sample(5)
```

```
ID
234701      157lbs
209892      154lbs
192101      176lbs
215501      137lbs
229718      154lbs
Name: Weight, dtype: object
```



```
df['Height'].sample(5)
```

```
ID
200765    180.340000
211037    177.800000
208386    187.960000
233782    180.779725
244942    190.500000
Name: Height, dtype: float64
```

```
df['Weight'].sample(5)
```

```
ID
243427    78.017824
245253    64.863656
233466    84.821704
236579    78.925008
219814    76.203456
Name: Weight, dtype: float64
```

```
df['Value'].sample(5)
```

```
ID
213242    6500000
181971     775000
241096    1800000
242603     240000
169432     725000
Name: Value, dtype: int32
```

# Preferred Foot Analysis

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Int64Index: 18207 entries, 158023 to 246269
```

```
Data columns (total 84 columns):
```

#	Column	Non-Null Count	Dtype
0	Name	18207 non-null	object
1	Age	18207 non-null	int64
2	Nationality	18207 non-null	object
3	Overall	18207 non-null	int64
4	Potential	18207 non-null	int64
5	Club	17966 non-null	object
6	Value	18207 non-null	object
7	Wage	18207 non-null	object
8	Special	18207 non-null	int64
9	Preferred Foot	18159 non-null	object

```
df['Preferred Foot'].isna().sum()
```

```
48
```

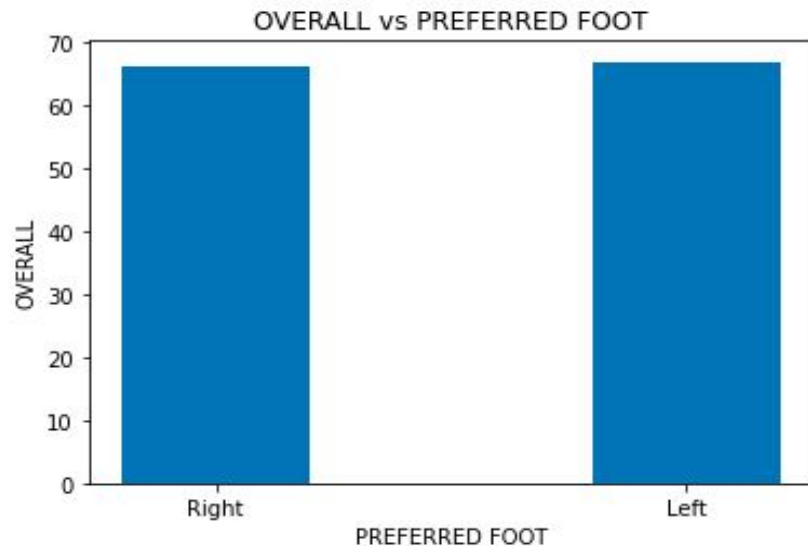
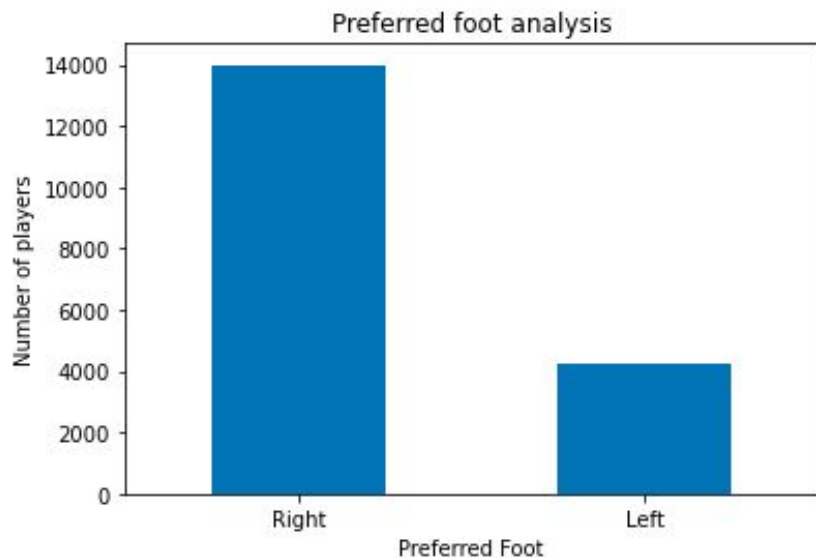
# Cleaning the missing “Preferred Foot” data

```
df['Preferred Foot'] = df['Preferred Foot'].fillna(df['Preferred Foot'].mode()[0])
```

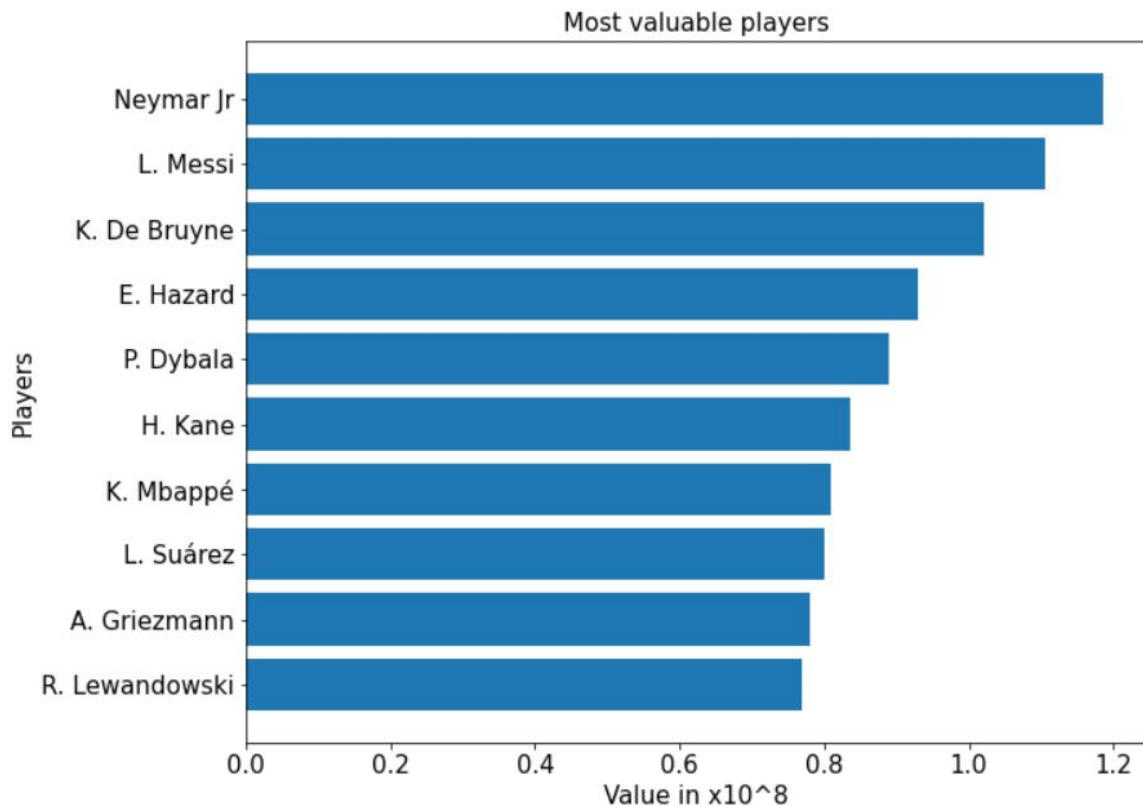
```
df['Preferred Foot'].isna().sum()
```

0

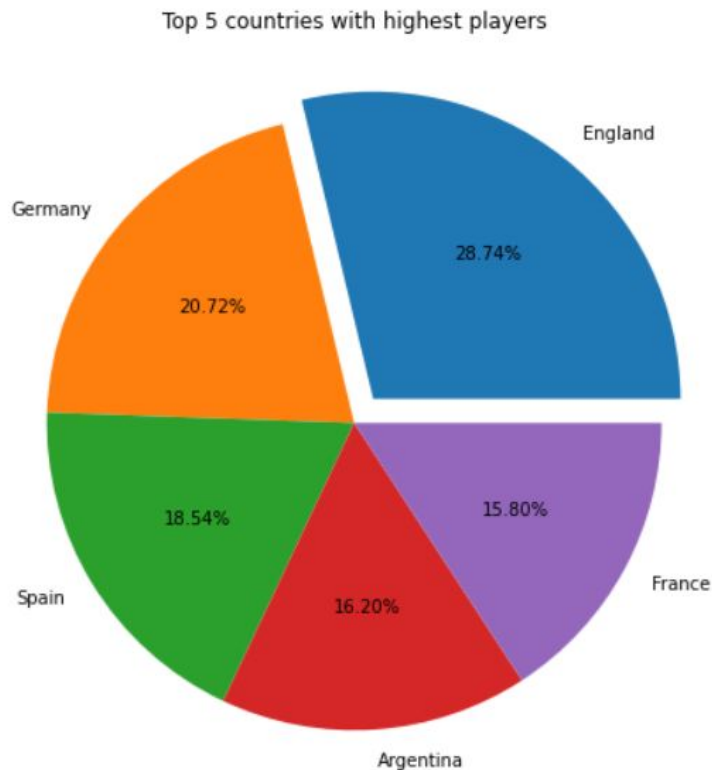
# Preferred foot



# Visualizing Amounts

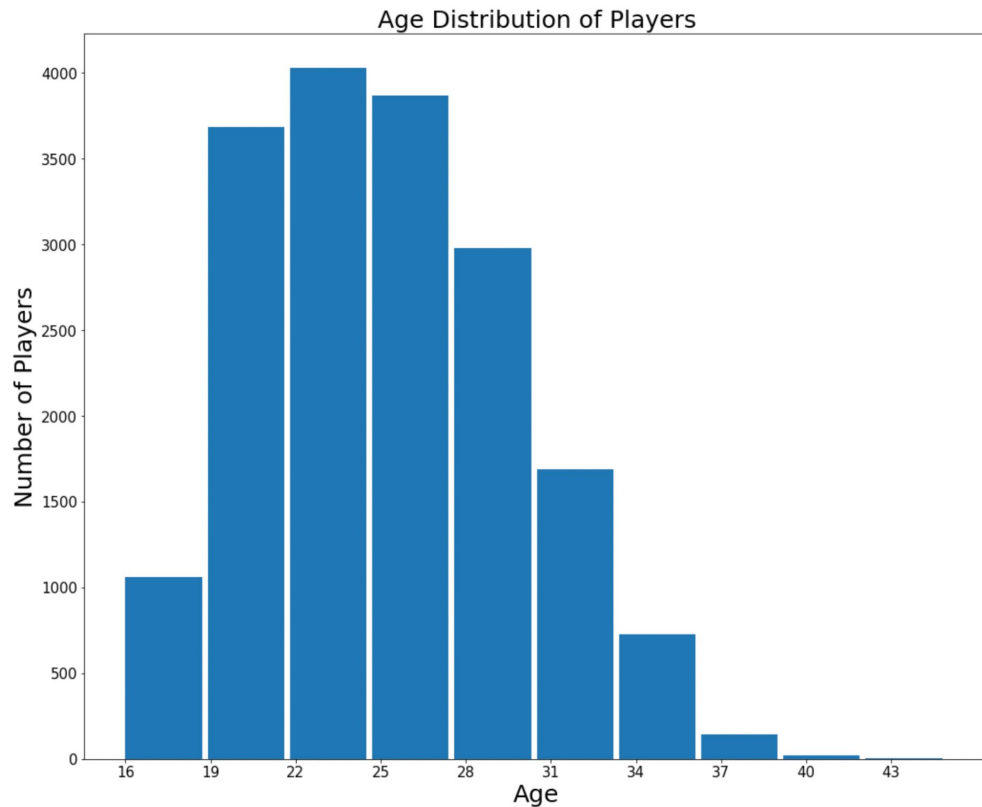


# Visualizing Proportions



	index	total
0	England	1662
1	Germany	1198
2	Spain	1072
3	Argentina	937
4	France	914

# Analysing Age Distribution



# Oldest and Youngest players

```
oldest5 = df.nlargest(5, "Age")  
oldest5[['Name', 'Age', 'Nationality']]
```

ID	Name	Age	Nationality
140029	O. Pérez	45	Mexico
51963	T. Warner	44	Trinidad & Tobago
53748	K. Pilkington	44	England
140183	S. Narazaki	42	Japan
156092	J. Villar	41	Paraguay

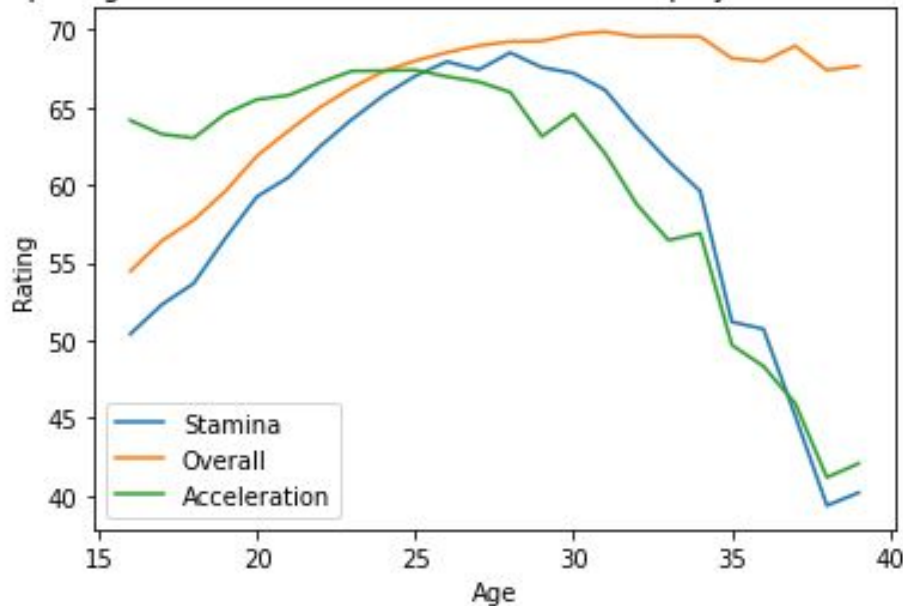
```
youngest5 = df.nsmallest(5, "Age")  
youngest5[['Name', 'Age', 'Nationality']]
```

ID	Name	Age	Nationality
241266	W. Geubbels	16	France
244403	A. Taoui	16	France
245616	Pelayo Morilla	16	Spain
246465	Guerrero	16	Spain
246594	H. Massengo	16	France



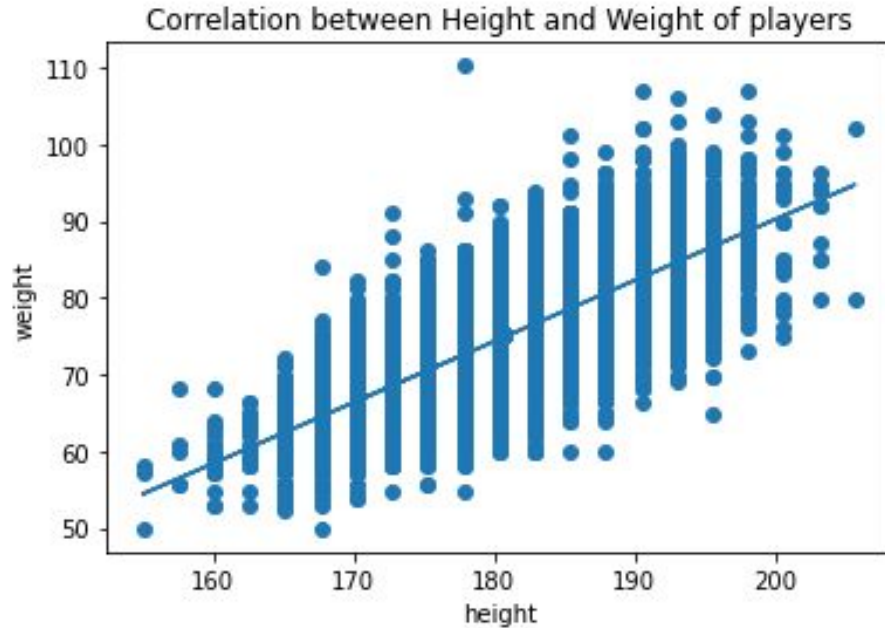
# Age vs Stamina, Overall and Acceleration

Comparing Stamina, Overall and Acceleration of players according to age



# Visualizing x-y relations:

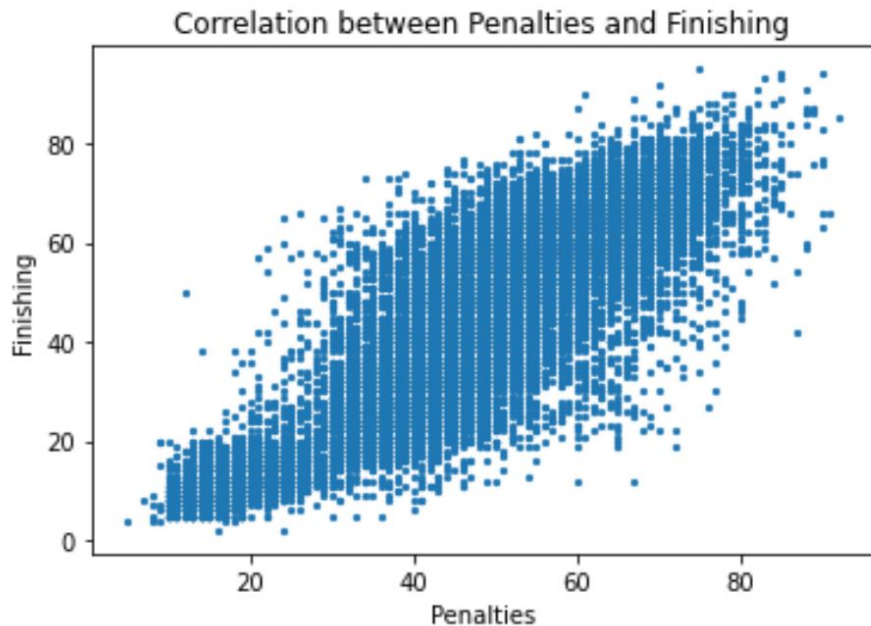
## Height vs Weight



```
print(df['Height'].corr(df["Weight"]))
```

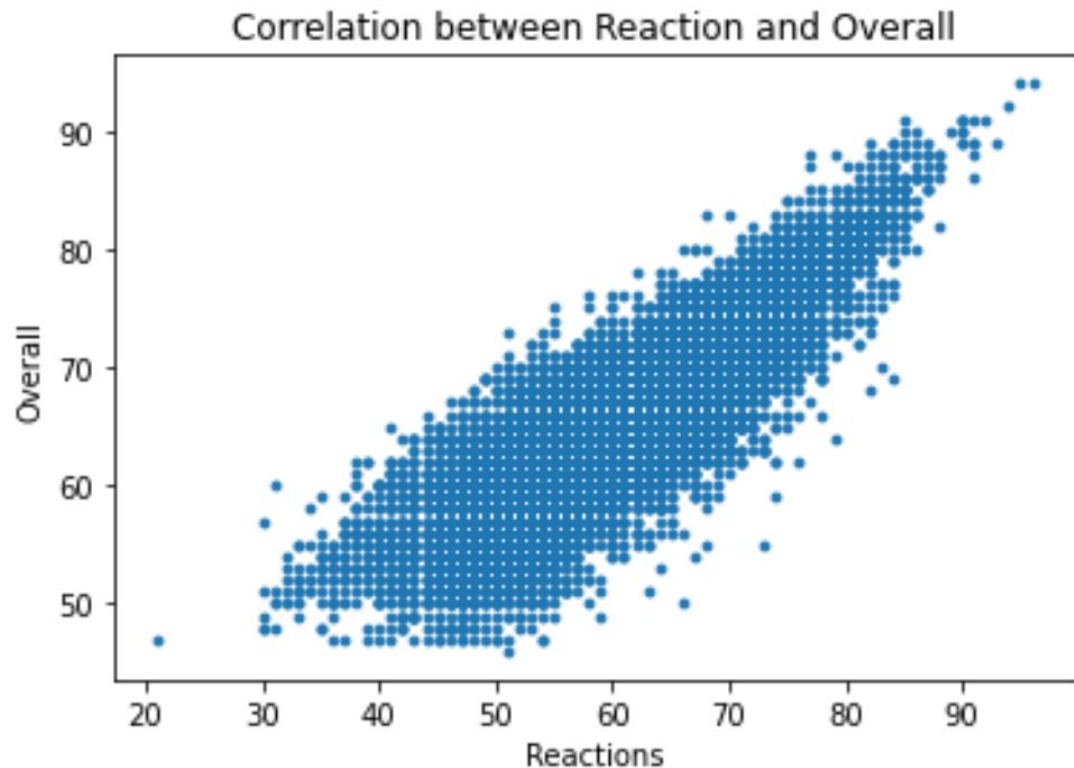
0.7546776706360377

# Penalties vs Finishing



```
| print(df['Penalties'].corr(df['Finishing']));  
0.8378270334761936
```

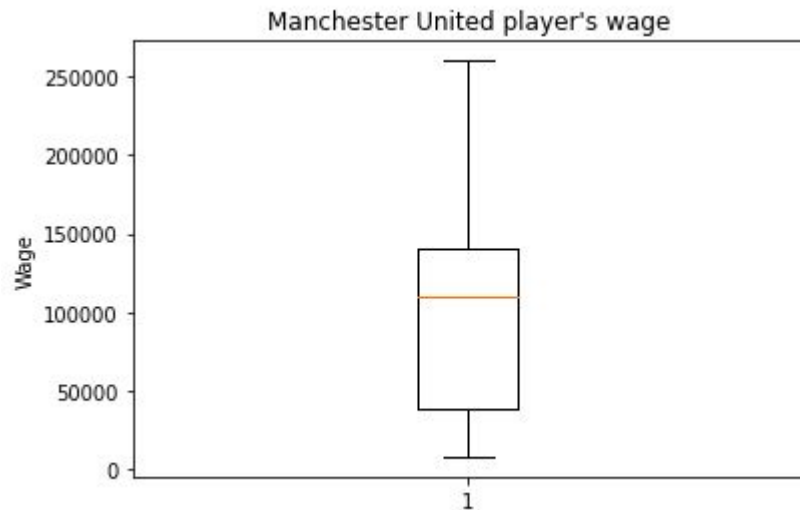
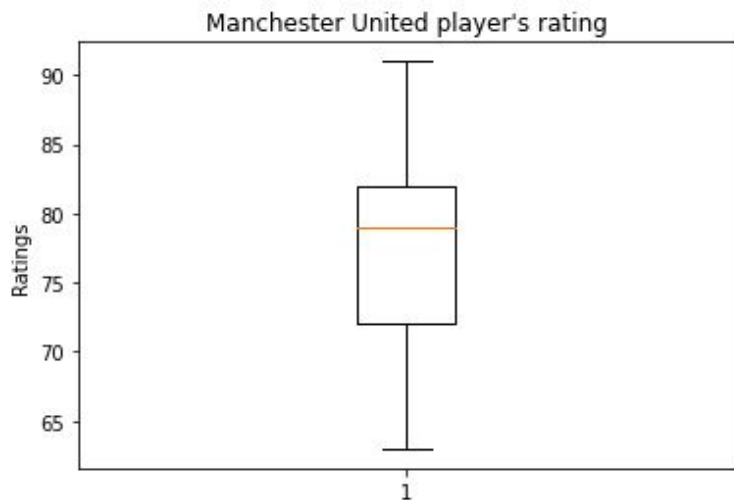
# Reaction vs Overall



```
print(df['Reactions'].corr(df['Overall']));
```

0.8500449067063615

# Specific club analysis



# Conclusion

- A. Data Preparation
- B. Data Cleaning
- C. Data Analytics
- D. Data Visualization

## Data Preparation

