

Data Science Project

FIFA-19



21-05-2020

Project by:

Prasfur Tiwari

Ravi Sista

INDEX

Introduction	3
Data Analysis	7
Data Collection/Gathering	7
Data Cleaning/Filtering	7
Data Exploration	10
Data Visualization	22
Model Fitting	42
Regression	43
Clustering	46
Visualizing Target Cluster	49
Observations	50

INTRODUCTION

FIFA 19 is a football simulation video game developed by EA Vancouver as part of Electronic Arts' FIFA series. It is the 26th instalment in the FIFA series, and was released on 28 September 2018 for PlayStation 3, PlayStation 4, Xbox 360, Xbox One, Nintendo Switch, and Microsoft Windows.

In addition to Classic Kick-Off, which is simply a normal match without added visuals or modified rules, there are five new match types that players can choose from UEFA Champions League, House Rules, Best of Series, Home & Away, and Cup Finals. Each type has a twist on the normal match experience in Kick-Off, keeping things interesting and fresh every time you play.

- **UEFA Champions League** — The most prestigious club competition in the world is just as integrated into Kick-Off as it is to the rest of FIFA 19. From group stage matches to the Final, you can set up a custom Champions League match with specific visuals, rules, and more that provide a realistic, immersive tournament experience.
- **House Rules** — If you've ever wanted to play a match in FIFA with different rules—or no rules at all—the House Rules match type is just for you. You can set up a match with a selection of custom rules, including No Rules, Survival, Long Range, First to . . . , and Headers & Volleys.
 - **Survival Mode**— Each time a user scores a goal, a random player from the scoring club is removed (excluding the goalkeeper) to create a challenge for the player with a score advantage.
 - **No Rules** — Anything goes in this match type, in which there are no offside calls, fouls, or bookings.
 - **Long Range**— Any goal scored inside the box will count as one goal, but goals scored from outside the box count as two goals.
 - **First to. . .**— This match type lets you set a custom win condition, whether it's first to score (golden goal), first to three goals, etc. The match will still play to the clock and go through full time, as well as extra time and penalties, if you choose.
 - **Headers & Volleys**— You can only score in this match type with a header or a volley. Free kicks and penalties also count, but any other goal scored using your feet outside of a volley will be disallowed.

- **Best of Series** — Play classic matches in a three- or five-match series to determine an overall winner.
- **Home & Away** — This is a two-legged match type in which you play one home and one away match to determine the overall winner. The winner is determined by the aggregate score, which is the team that has scored the most combined goals from those two matches. If the teams are level after two matches, the team with the most away goals will be determined the winner. If the teams are still level, the match will go to extra time, then to a penalty shootout.
- **Cup Finals**— Play your match as one of a handful of real-life cup finals, including the Champions League Final, Europa League Final, FA Cup Final, and others. Official kits, badges, match balls, and authentic broadcast overlays (for a select few tournaments) provide an authentic cup final experience.



Now you can track your Kick-Off mode record and stats, including detailed analytics, from all matches played within the mode. Use the stats to analyze and refine your game plan, tweak your pre-game tactics, and prepare for every match. This advanced system tracks all available Kick Off mode gameplay information about you and your opponent. With a host of detailed, immersive information, FIFA 19 allows you to approach every Kick-Off match just like you would an actual game of football, utilizing tactics and strategies gleaned from your stats.

You can view your objective stats or compare them to any opponent that you've faced in Kick-Off mode.

***Note that Stat Tracking works slightly differently on each console. See the Account Linking section for more details.**

Tracked Stats

- Wins
- Losses
- Draws
- Win %
- Goals Scored
- Goals Allowed
- Goal Differential
- Goal Types:
 - Inside box
 - Outside of the box
 - Penalties
 - Free Kicks
- Goal Heatmap on net (where on net it was scored)
- Shot on Target % (comparing total shots vs shots on target)
- Total Shots on Target
- Total Shots
- Average Possession %
- Average possession % in areas
- Pass % completion

Tracked Milestones (for head-to-head comparisons)

- Previous five match results
- Two-legged match wins
- Best of three wins
- Best of five wins
- Cup Final wins
- Fastest goal scored
- Biggest win (biggest goal differential in a match)



New features which are added to the game are:

- **Active Touch System**

The new Active Touch System fundamentally changes the way you receive and strike the ball, providing closer control, improved fluidity, more creativity and increased player personality.

- **Dynamic Tactics**

A re-imagined system gives players the tools to set multiple tactical approaches, offering in-depth customization pre-match, as well as more options for dynamic in-match adjustments simply from the D-Pad.

- **Timed Finishing**

Whether it's a hit from outside the box, a precision header, or a deft touch, timed finishing adds a new layer of control to your chances in front of goal.

- **50/50 Battles**

With 50/50 Battles, user reactions and player attributes determine the outcome for winning loose balls across the pitch. With increased teammate intelligence and spatial awareness, every challenge counts in the fight for possession.

- **Real Player Motion Technology**

The game-changing animation system, which brought player personality and increased fidelity in movement to EA SPORTS FIFA, returns with increased coverage across the pitch. Enhanced animations for tactical shielding, impact balancing, and physical jostles bring realistic player movement, responsiveness and personality to new heights.

DATA ANALYSIS

Data collection/Gathering:

The dataset of Players information, provided by the EA sports was taken from **Kaggle**. A 'search' query was made in the Jupyter notebook with Python kernel, so as to search the players information and perform analysis on it.

The dataset was imported in the Jupyter notebook using the '**pandas**' module.

Data cleaning:

For the analysis of the data, data cleaning is one of the steps in the procedure. For analysing the data, we only require that information which will be helpful to us in producing some insights.

So, we've to filter out or remove the unnecessary information from the dataset. (e.g) Columns with missing values or columns with no meaningful information are considered to be unnecessary information.

For checking if there are any missing values in our dataset, we use the '**heat-map**' for visualization.

	ID	Name	Age	Photo	Nationality	Flag	Overall	Potential	Club
0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	94	FC Barcelona https://cdn.sofifa.org/
1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Juventus https://cdn.sofifa.org/
2	190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	https://cdn.sofifa.org/flags/54.png	92	93	Paris Saint-Germain https://cdn.sofifa.org/
3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	93	Manchester United https://cdn.sofifa.org/
4	192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	https://cdn.sofifa.org/flags/7.png	91	92	Manchester City https://cdn.sofifa.org/

5 rows × 88 columns

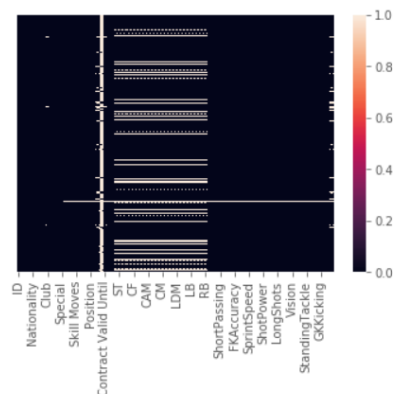
```
# Checking the number of missing values
df.isnull().sum()
```

```
ID          0
Name         0
Age          0
Photo        0
Nationality  0
...
GKHandling   48
GKKicking    48
GKPositioning 48
GKReflexes   48
Release Clause 1564
Length: 88, dtype: int64
```

```
#Checking extent of null values
```

```
sns.heatmap(df.isnull(),yticklabels=False)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x16933bf3708>
```



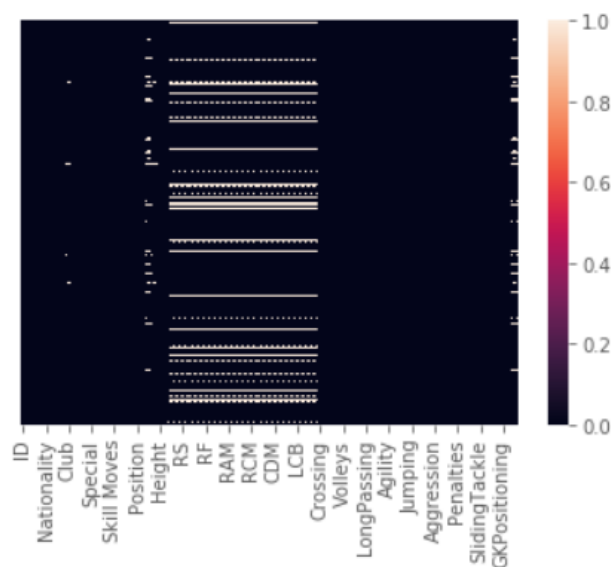
With the help of the heatmap, we can say that there are some missing values present in our dataset. Since, the column 'Loaned From' has more than 75% of values missing, therefore it is for the best that we should remove this column.

To fill up the missing values, we will enter the mean of the values in the place of the missing values.

```
#Checking extent of null values
```

```
sns.heatmap(df.isnull(),yticklabels=False)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x23edc54d748>
```

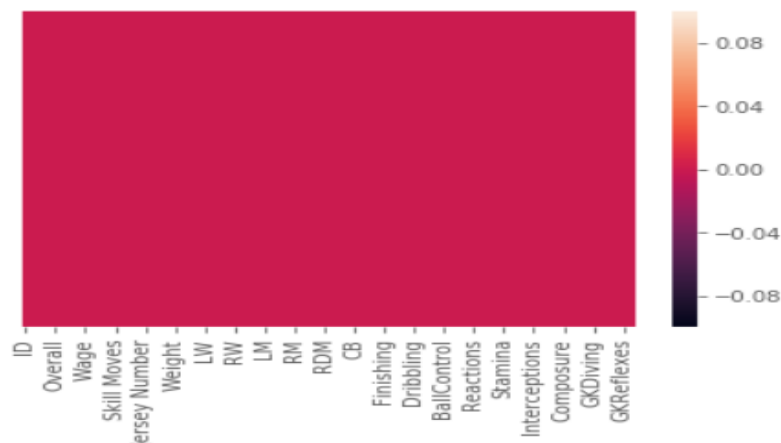


Still some columns have missing values; even after we replaced them with mean values. This may be because those columns have strings. So, we will replace those missing values with 'unassigned'.

```
#there are still cells in which the mean value could not be assigned.
df.fillna("Unassigned",inplace=True)
```

```
#Checking extent of null values
sns.heatmap(df.isnull(),yticklabels=False)
```

<matplotlib.axes._subplots.AxesSubplot at 0x23edb802128>



Since, there are no missing values according to the heatmap, we now have complete data, which implies that our insights generated will be more accurate.

Also, the data set consists of some more unnecessary information, so will remove it.

```
# Final data clean-up procedure
df.drop(['Photo', 'Flag', 'Club Logo', 'Real Face', 'Special'],axis=1,inplace=True)
df.head()
```

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	...	Composure	Marking	StandingTackle	SlidingTackle	GKDiv
0	158023	L. Messi	31	Argentina	94	94	FC Barcelona	€110.5M	€565K	Left	...	96.0	33.0	28.0	26.0	
1	20801	Cristiano Ronaldo	33	Portugal	94	94	Juventus	€77M	€405K	Right	...	95.0	28.0	31.0	23.0	
2	190871	Neymar Jr	26	Brazil	92	93	Paris Saint-Germain	€118.5M	€290K	Right	...	94.0	27.0	24.0	33.0	
3	193080	De Gea	27	Spain	91	93	Manchester United	€72M	€260K	Right	...	68.0	15.0	21.0	13.0	€
4	192985	K. De Bruyne	27	Belgium	91	92	Manchester City	€102M	€355K	Right	...	88.0	68.0	58.0	51.0	1

5 rows × 82 columns

Data exploration & analysis:

Before doing the analysis of the data, we've to explore the dataset first.

```
# exploring the data set
df.head()
```

	Unnamed: 0	ID	Name	Age	Photo	Nationality	Flag	Overall	Potential	Club	...	CoI
0	0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	94	FC Barcelona	...	
1	1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Juventus	...	
2	2	190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	https://cdn.sofifa.org/flags/54.png	92	93	Paris Saint-Germain	...	
3	3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	93	Manchester United	...	
4	4	192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	https://cdn.sofifa.org/flags/7.png	91	92	Manchester City	...	

5 rows × 89 columns

In the above picture we can see the first 5 tuples containing information of players present in FIFA 19.

The dimensions of the dataset are 18207 Rows & 89 Columns.

```
df.shape
```

```
(18207, 89)
```

➔ The columns present in the dataset are as follows:

```
# Listing out the columns
df.columns
```

```
Index(['Unnamed: 0', 'ID', 'Name', 'Age', 'Photo', 'Nationality', 'Flag',
      'Overall', 'Potential', 'Club', 'Club Logo', 'Value', 'Wage', 'Special',
      'Preferred Foot', 'International Reputation', 'Weak Foot',
      'Skill Moves', 'Work Rate', 'Body Type', 'Real Face', 'Position',
      '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', 'RB', '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',
      'GKKicking', 'GKPositioning', 'GKReflexes', 'Release Clause'],
      dtype='object')
```

Viewing data types of columns:

To see the data type of the columns we use dtypes method.







```
df.dtypes
ID                int64
Name              object
Age              int64
Photo            object
Nationality       object
Flag             object
Overall          int64
Potential        int64
Club             object
Club Logo        object
Value            object
Wage             object
Special          int64
Preferred Foot    object
International Reputation float64
Weak Foot        float64
Skill Moves      float64
Work Rate        object
Body Type        object
Real Face        object
Position         object
Jersey Number    float64
Joined           object
Contract Valid Until object
Height           object
Weight           object
LS              object
ST              object
RS              object
LW              object
```

➔ Now let us perform some analysis on the dataset.

- **Best players according to particular field positions:**

	Position	Name	Age	Club	Nationality
17	CAM	A. Griezmann	27	Atlético Madrid	France
12	CB	D. Godín	32	Atlético Madrid	Uruguay
20	CDM	Sergio Busquets	29	FC Barcelona	Spain
271	CF	Luis Alberto	25	Lazio	Spain
67	CM	Thiago	27	FC Bayern München	Spain
3	GK	De Gea	27	Manchester United	Spain
28	LAM	J. Rodríguez	26	FC Bayern München	Colombia
35	LB	Marcelo	30	Real Madrid	Brazil
24	LCB	G. Chiellini	33	Juventus	Italy
11	LCM	T. Kroos	28	Real Madrid	Germany
14	LDM	N. Kanté	27	Chelsea	France
5	LF	E. Hazard	27	Chelsea	Belgium
33	LM	P. Aubameyang	29	Arsenal	Gabon
21	LS	E. Cavani	31	Paris Saint-Germain	Uruguay
2	LW	Neymar Jr	26	Paris Saint-Germain	Brazil
474	LWB	N. Schulz	25	TSG 1899 Hoffenheim	Germany
129	RAM	J. Cuadrado	30	Juventus	Colombia
69	RB	Azpilicueta	28	Chelsea	Spain
8	RCB	Sergio Ramos	32	Real Madrid	Spain
4	RCM	K. De Bruyne	27	Manchester City	Belgium
45	RDM	P. Pogba	25	Manchester United	France
0	RF	L. Messi	31	FC Barcelona	Argentina
25	RM	K. Mbappé	19	Paris Saint-Germain	France
7	RS	L. Suárez	31	FC Barcelona	Uruguay
56	RW	Bernardo Silva	23	Manchester City	Portugal
450	RWB	M. Ginter	24	Borussia Mönchengladbach	Germany
1	ST	Cristiano Ronaldo	33	Juventus	Portugal

- Player's best features:

<p>Best Crossing</p>  <p>91 CAM</p> <p>DE BRUYNE</p> <table> <tr> <td>77 PAC</td> <td>87 DRI</td> </tr> <tr> <td>86 SHO</td> <td>60 DEF</td> </tr> <tr> <td>92 PAS</td> <td>78 PHY</td> </tr> </table>	77 PAC	87 DRI	86 SHO	60 DEF	92 PAS	78 PHY	<p>Best Finishing</p>  <p>94 CF</p> <p>MESSI</p> <table> <tr> <td>88 PAC</td> <td>96 DRI</td> </tr> <tr> <td>91 SHO</td> <td>32 DEF</td> </tr> <tr> <td>88 PAS</td> <td>61 PHY</td> </tr> </table>	88 PAC	96 DRI	91 SHO	32 DEF	88 PAS	61 PHY
77 PAC	87 DRI												
86 SHO	60 DEF												
92 PAS	78 PHY												
88 PAC	96 DRI												
91 SHO	32 DEF												
88 PAS	61 PHY												
<p>Best Heading Accuracy</p>  <p>86 CB</p> <p>NALDO</p> <table> <tr> <td>63 PAC</td> <td>62 DRI</td> </tr> <tr> <td>69 SHO</td> <td>87 DEF</td> </tr> <tr> <td>64 PAS</td> <td>74 PHY</td> </tr> </table> <p>fifadataba.com</p>	63 PAC	62 DRI	69 SHO	87 DEF	64 PAS	74 PHY	<p>Best Short Passing</p>  <p>91 CM</p> <p>MODRIĆ</p> <table> <tr> <td>76 PAC</td> <td>91 DRI</td> </tr> <tr> <td>76 SHO</td> <td>70 DEF</td> </tr> <tr> <td>90 PAS</td> <td>67 PHY</td> </tr> </table>	76 PAC	91 DRI	76 SHO	70 DEF	90 PAS	67 PHY
63 PAC	62 DRI												
69 SHO	87 DEF												
64 PAS	74 PHY												
76 PAC	91 DRI												
76 SHO	70 DEF												
90 PAS	67 PHY												
<p>Best Volleys</p>  <p>89 ST</p> <p>CAVANI</p> <table> <tr> <td>76 PAC</td> <td>80 DRI</td> </tr> <tr> <td>87 SHO</td> <td>52 DEF</td> </tr> <tr> <td>72 PAS</td> <td>83 PHY</td> </tr> </table> <p>fifadataba.com</p>	76 PAC	80 DRI	87 SHO	52 DEF	72 PAS	83 PHY	<p>Best Dribbling</p>  <p>94 CF</p> <p>MESSI</p> <table> <tr> <td>88 PAC</td> <td>96 DRI</td> </tr> <tr> <td>91 SHO</td> <td>32 DEF</td> </tr> <tr> <td>88 PAS</td> <td>61 PHY</td> </tr> </table>	88 PAC	96 DRI	91 SHO	32 DEF	88 PAS	61 PHY
76 PAC	80 DRI												
87 SHO	52 DEF												
72 PAS	83 PHY												
88 PAC	96 DRI												
91 SHO	32 DEF												
88 PAS	61 PHY												

Best Curve



Best Freekick Accuracy



Best Long Passing



Best Ball Control



Best Acceleration



Best Sprint Speed



Best Agility



Best Reactions



Best Balance



Best Shot Power



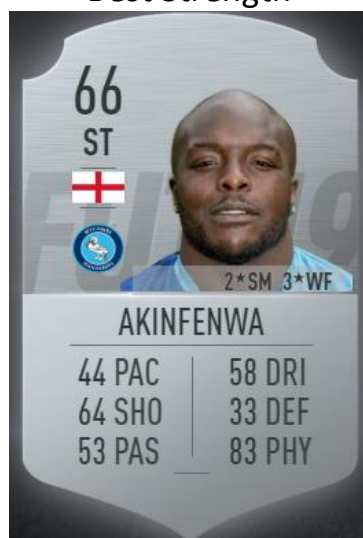
Best Jumping



Best Stamina



Best Strength



Best Long Shots



Best Aggression



Best Interceptions



Best Positioning



Best Vision



Best Penalties



Best Composure



Best Marking



Best Standing Tackle



Best GK Diving



Best GK Handling



Best GK Kicking



Best GK Positioning



Best GK Reflexes



- Best players from each position with their age, nationality, club based on their potential scores:

	Position	Name	Age	Club	Nationality
31	CAM	C. Eriksen	26	Tottenham Hotspur	Denmark
42	CB	S. Umtiti	24	FC Barcelona	France
27	CDM	Casemiro	26	Real Madrid	Brazil
350	CF	A. Milik	24	Napoli	Poland
78	CM	S. Milinković-Savić	23	Lazio	Serbia
3	GK	De Gea	27	Manchester United	Spain
28	LAM	J. Rodríguez	26	FC Bayern München	Colombia
35	LB	Marcelo	30	Real Madrid	Brazil
77	LCB	M. Škriniar	23	Inter	Slovakia
11	LCM	T. Kroos	28	Real Madrid	Germany
14	LDM	N. Kanté	27	Chelsea	France
15	LF	P. Dybala	24	Juventus	Argentina

- Features according to field positions:

Position CAM: Balance, Agility, Acceleration
 Position CB: Jumping, Aggression, HeadingAccuracy
 Position CDM: Aggression, Jumping, Balance
 Position CF: Agility, Balance, Acceleration
 Position CM: Balance, Agility, Acceleration
 Position GK: GKReflexes, GKDiving, GKPositioning
 Position LAM: Agility, Balance, Acceleration
 Position LB: Acceleration, Balance, Agility
 Position LCB: Jumping, Aggression, HeadingAccuracy
 Position LCM: Balance, Agility, BallControl
 Position LDM: Aggression, BallControl, LongPassing
 Position LF: Balance, Agility, Acceleration
 Position LM: Acceleration, Agility, Balance
 Position LS: Acceleration, Agility, Finishing
 Position LW: Acceleration, Agility, Balance
 Position LWB: Acceleration, Agility, Balance
 Position RAM: Agility, Balance, Acceleration
 Position RB: Acceleration, Balance, Jumping
 Position RCB: Jumping, Aggression, HeadingAccuracy
 Position RCM: Agility, Balance, BallControl
 Position RDM: Aggression, Jumping, BallControl
 Position RF: Agility, Acceleration, Balance
 Position RM: Acceleration, Agility, Balance
 Position RS: Acceleration, Agility, Jumping
 Position RW: Acceleration, Agility, Balance
 Position RWB: Acceleration, Agility, Balance
 Position ST: Acceleration, Jumping, Finishing
 Position Unassigned: Acceleration, Balance, Jumping

- **Top 5 Expensive clubs:**

```
# Top five the most expensive clubs
df.groupby(['Club'])['Value'].sum().sort_values(ascending = False).head()
```

```
Club
Real Madrid      874425000.0
FC Barcelona     852600000.0
Manchester City   786555000.0
Juventus          704475000.0
FC Bayern München 679025000.0
```

- **Bottom 5 expensive clubs:**

```
# Top five the less expensive clubs
df.groupby(['Club'])['Value'].sum().sort_values().head()
```

```
Club
Unassigned      0.0
Bray Wanderers  1930000.0
Limerick FC     2040000.0
Derry City      2795000.0
Bohemian FC     3195000.0
```

- **Top 5 clubs with best Overall:**

```
# Top five teams with the best players
df.groupby(['Club'])['Overall'].max().sort_values(ascending = False).head()
```

```
Club
Juventus      94
FC Barcelona  94
Paris Saint-Germain 92
Manchester United 91
Manchester City 91
```

- **Top 5 Elder Players:**

	Name	Club	Nationality	Overall	Age
4741	O. Pérez	Pachuca	Mexico	71	45
18183	K. Pilkington	Cambridge United	England	48	44
17726	T. Warner	Accrington Stanley	Trinidad & Tobago	53	44
10545	S. Narazaki	Nagoya Grampus	Japan	65	42
7225	C. Muñoz	CD Universidad de Concepción	Argentina	68	41

- **Top 5 Young Players:**

	Name	Club	Nationality	Overall	Age
18206	G. Nugent	Tranmere Rovers	England	46	16
17743	J. Olstad	Sarpsborg 08 FF	Norway	52	16
13293	H. Massengo	AS Monaco	France	62	16
16081	J. Italiano	Perth Glory	Australia	58	16
18166	N. Ayéva	Örebro SK	Sweden	48	16

- **Top 5 Free-kick takers:**

	Name	Club	Nationality	Overall	Age	FKAccuracy
0	L. Messi	FC Barcelona	Argentina	94	31	94.0
293	S. Giovinco	Toronto FC	Italy	82	31	93.0
72	M. Pjanić	Juventus	Bosnia Herzegovina	86	28	92.0
1113	E. Bardhi	Levante UD	FYR Macedonia	77	22	91.0
449	H. Çalhanoğlu	Milan	Turkey	80	24	90.0

- **Top 5 Penalty takers:**

	Name	Club	Nationality	Overall	Age	BallControl
0	L. Messi	FC Barcelona	Argentina	94	31	96.0
2	Neymar Jr	Paris Saint-Germain	Brazil	92	26	95.0
30	Isco	Real Madrid	Spain	88	26	95.0
13	David Silva	Manchester City	Spain	90	32	94.0
5	E. Hazard	Chelsea	Belgium	91	27	94.0

- **Players with best Ball Control:**

	Name	Club	Nationality	Overall	Age	Penalties
206	M. Balotelli	OGC Nice	Italy	83	27	92.0
118	Fabinho	Liverpool	Brazil	84	24	91.0
16	H. Kane	Tottenham Hotspur	England	89	24	90.0
823	R. Jiménez	Wolverhampton Wanderers	Mexico	78	27	90.0
945	L. Baines	Everton	England	77	33	90.0

- **Players with best Sprint speed:**

	Name	Club	Nationality	Overall	Age	SprintSpeed
55	L. Sané	Manchester City	Germany	86	22	96.0
25	K. Mbappé	Paris Saint-Germain	France	88	19	96.0
1968	Adama	Wolverhampton Wanderers	Spain	75	22	96.0
36	G. Bale	Real Madrid	Wales	88	28	95.0
10928	Maicon	Livorno	Brazil	65	25	95.0

DATA VISUALIZATION

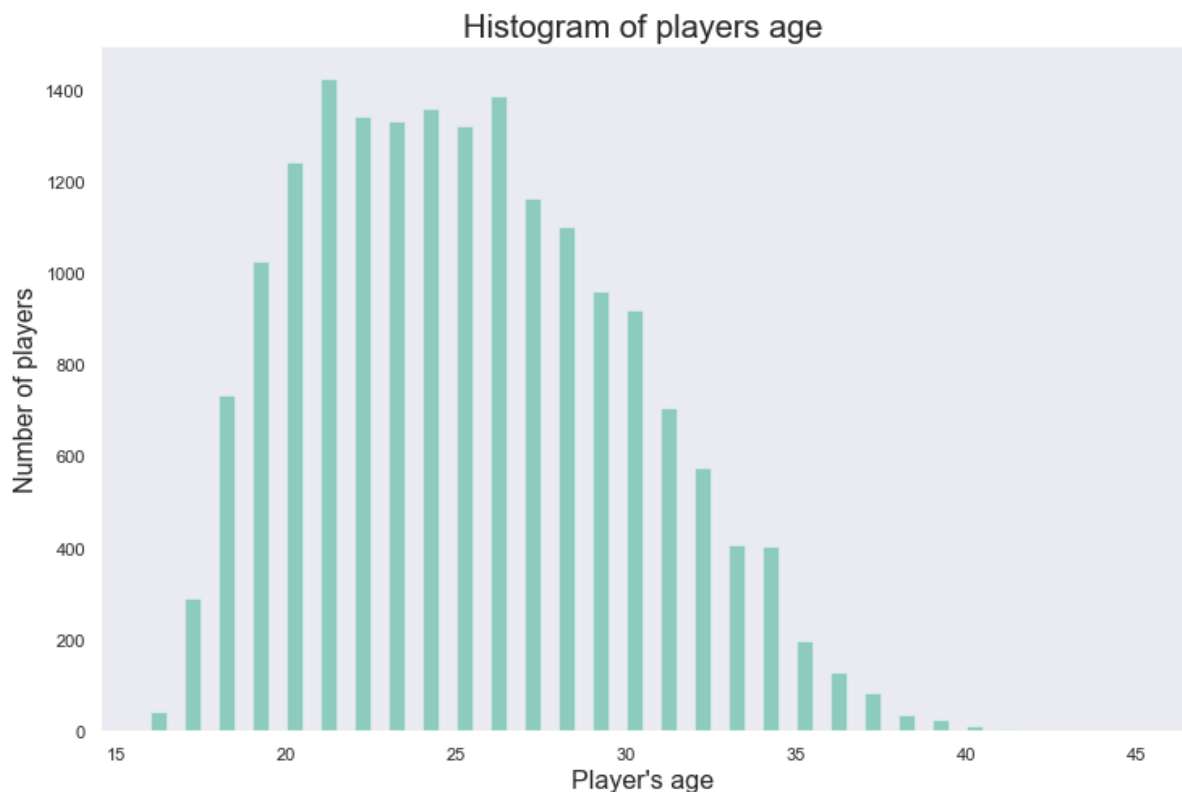
By having a look at a Dataset, we cannot understand the attributes. If we represent these attributes in graphical way, then we could get some meaningful insights out from it. This graphical representation of the data is called as **Data Visualization**.

It involves producing images that communicate relationships among the represented data to the viewers of the images. This communication is achieved through the use of a systematic mapping between graphic marks & data values in the creation of the visualization.

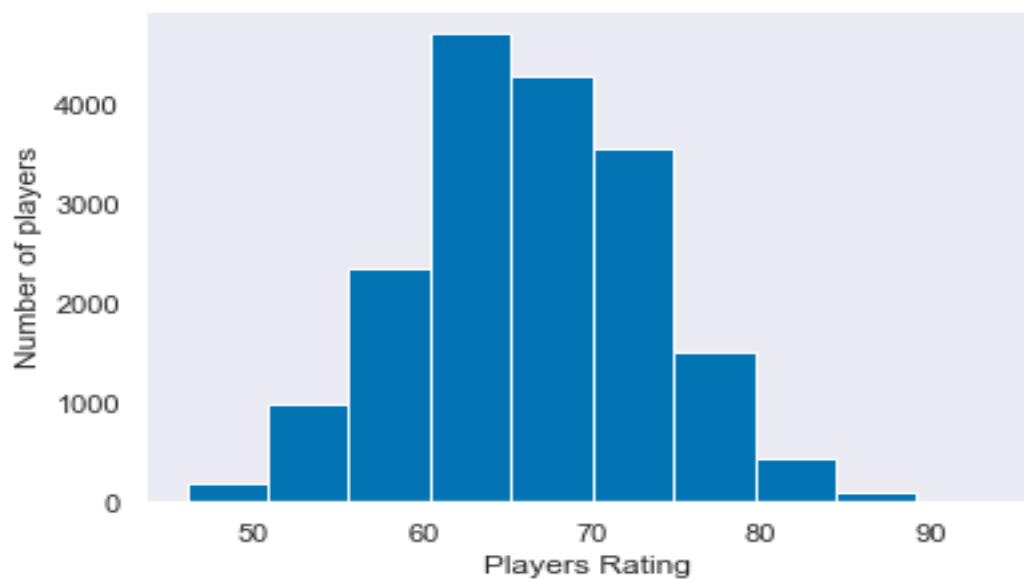
Data Visualization can be done through visualization tools such as [Tableau](#), [Power BI](#), etc., and also with the help of the Statistical tools such as [Python & R](#).

In this project we've done the following visualizations:

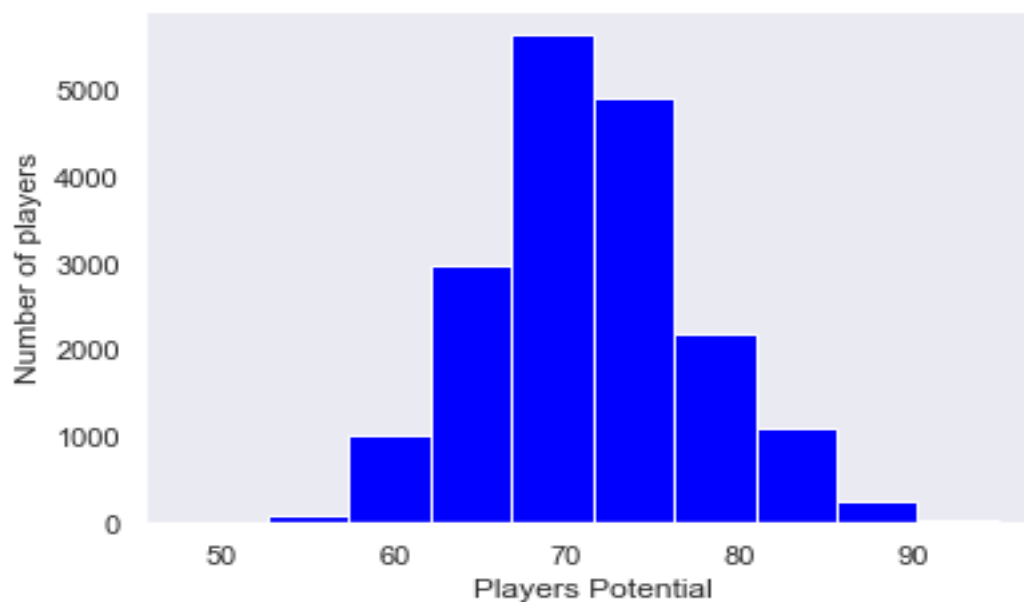
- **Age distribution of players:**



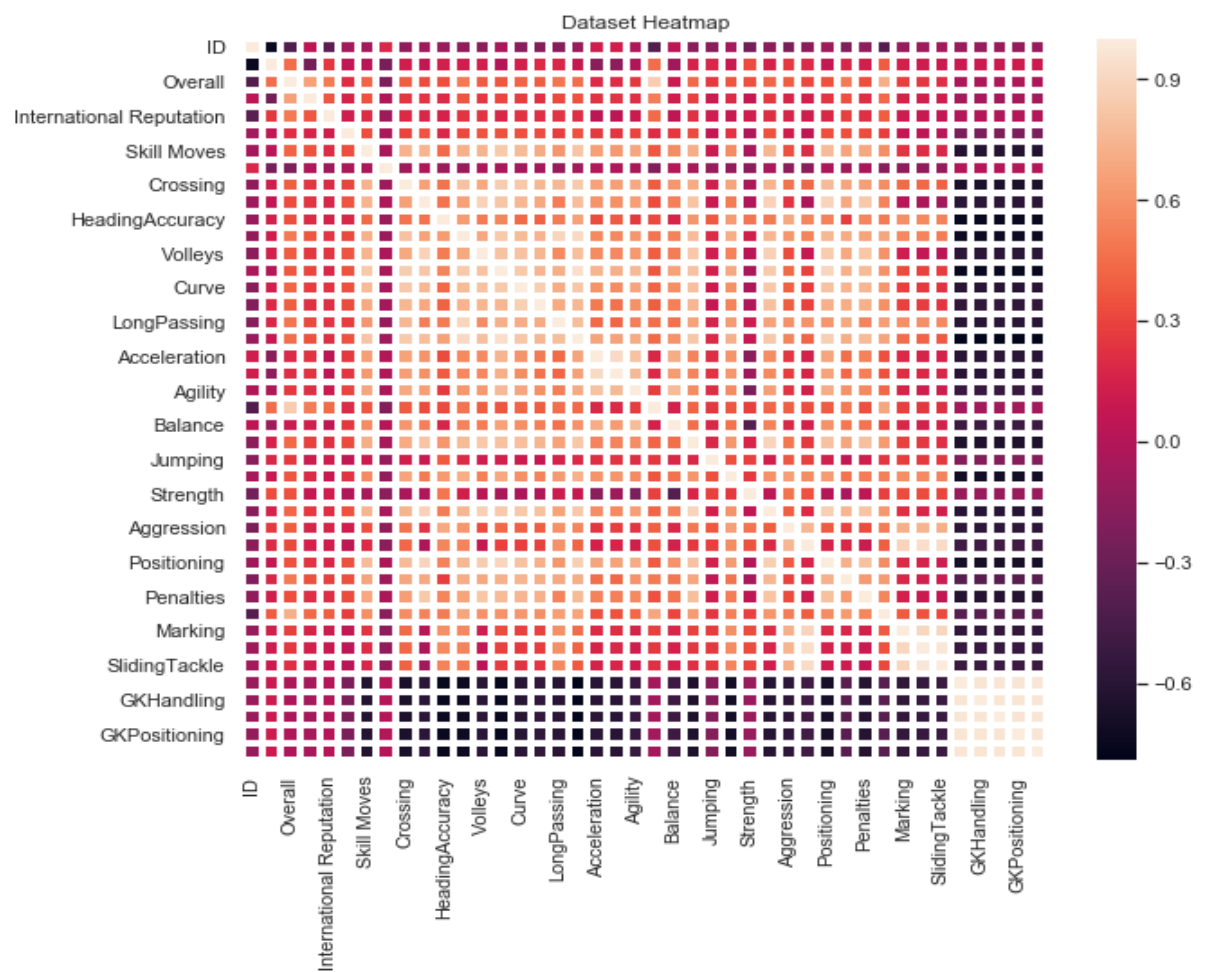
- Overall ratings of players:



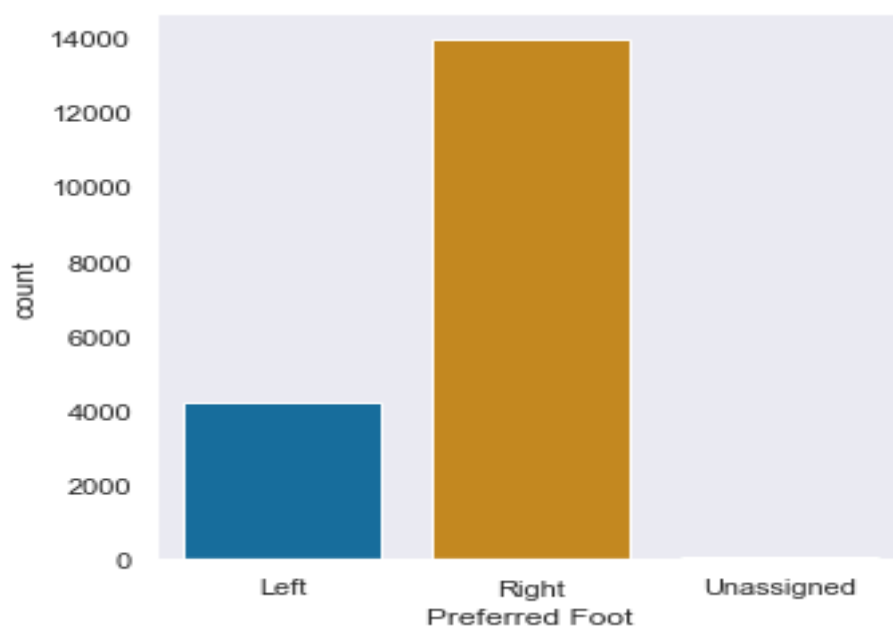
- Distribution of players Potential rating:



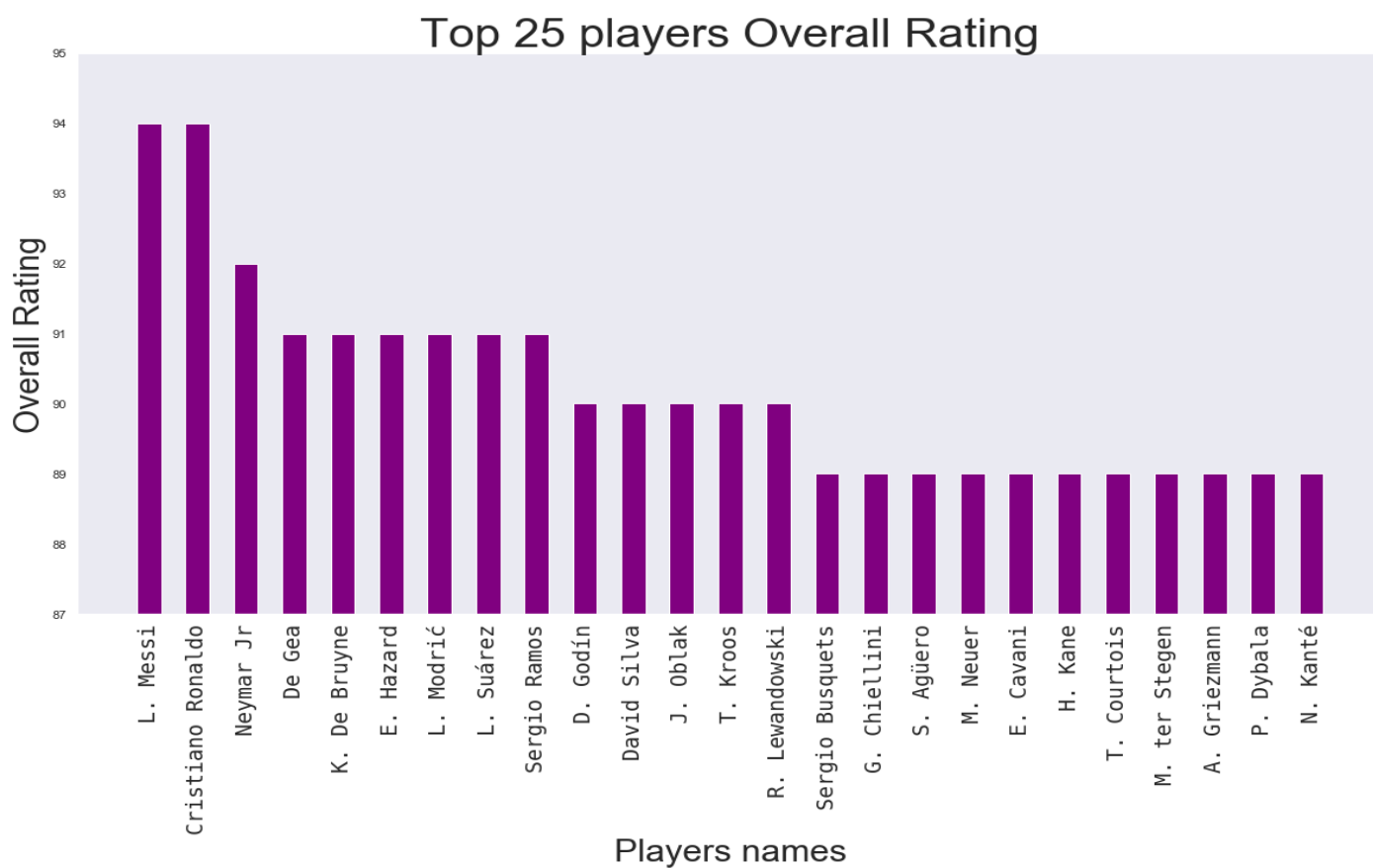
- Heat-map of the dataset:



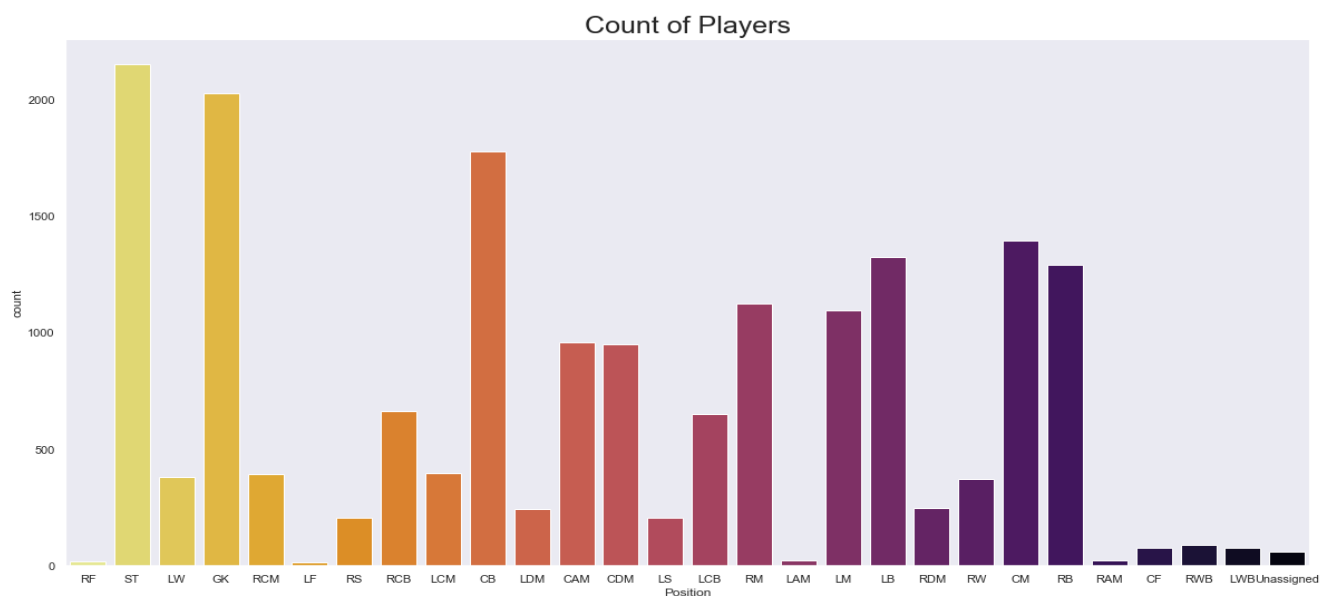
- Players foot preference:



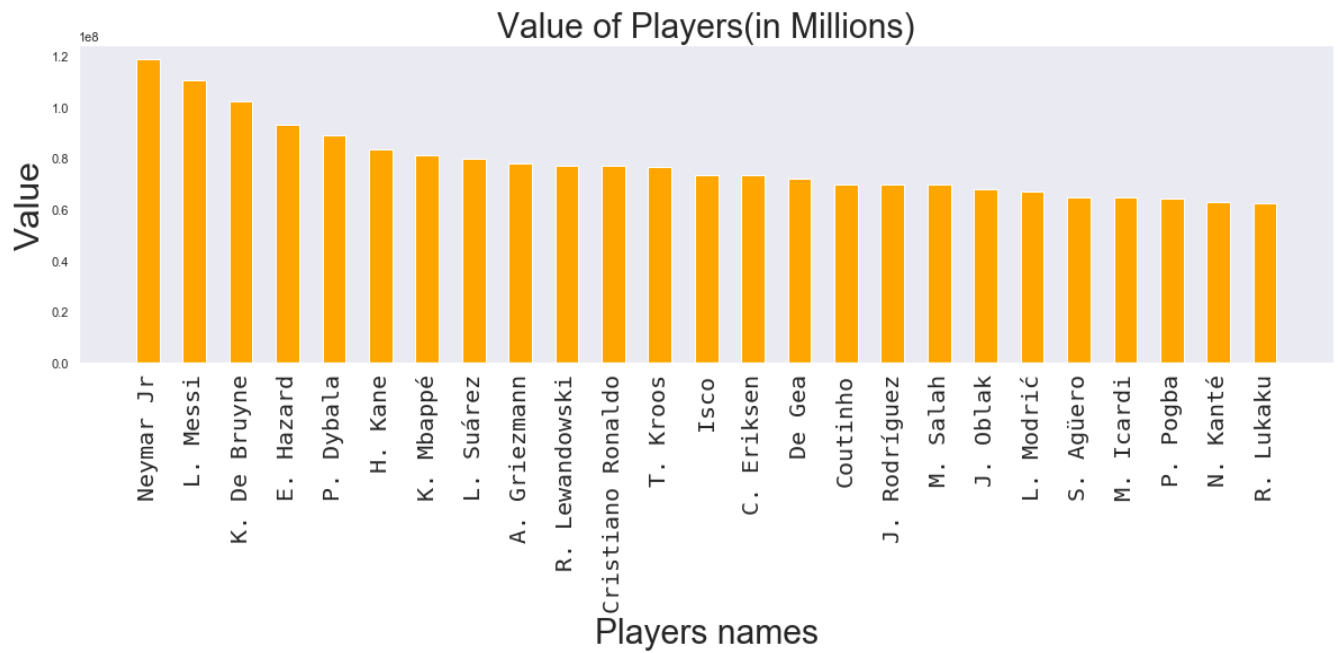
- **Top 25 players according to their overall:**



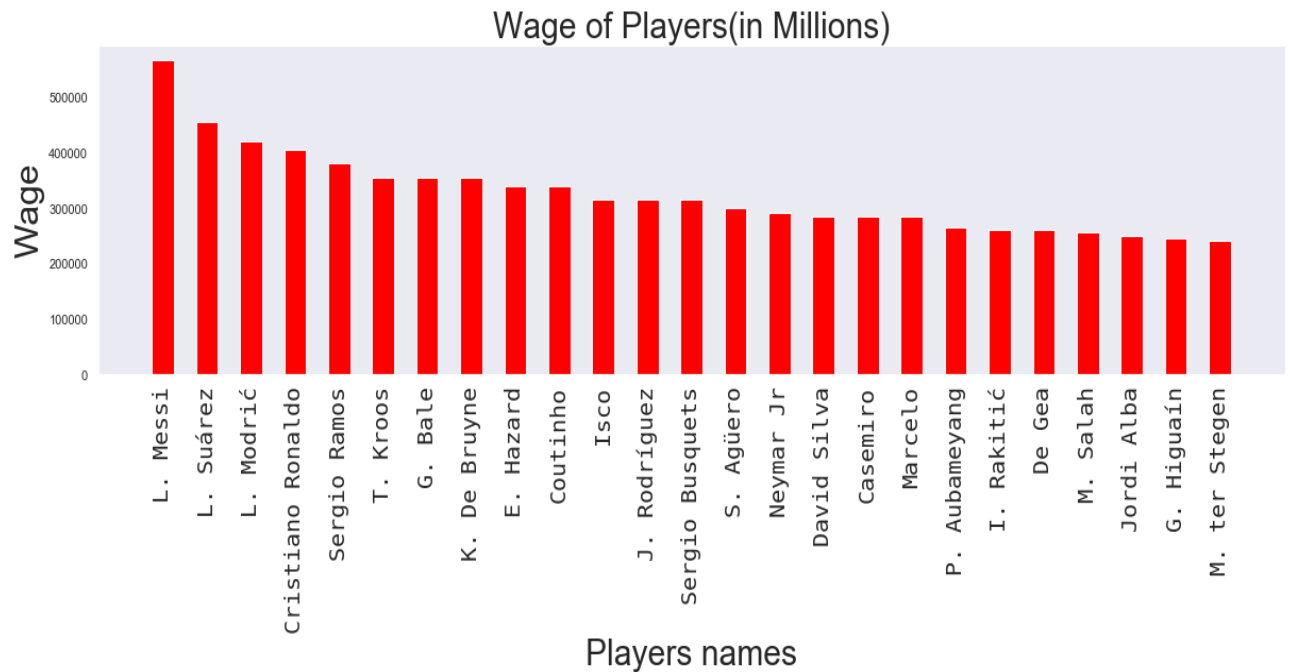
- **Distribution of players according to their field positions:**



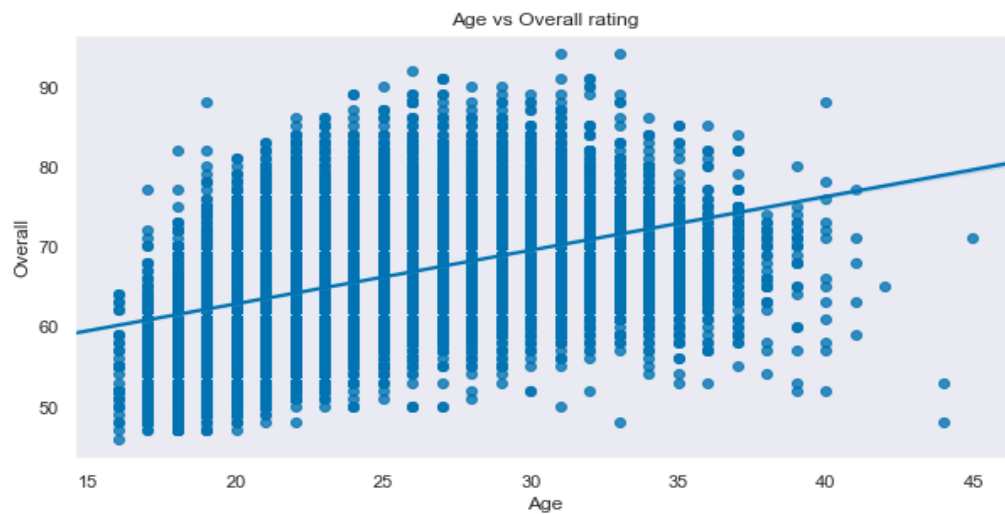
- **Distribution of Value (in millions) of players:**



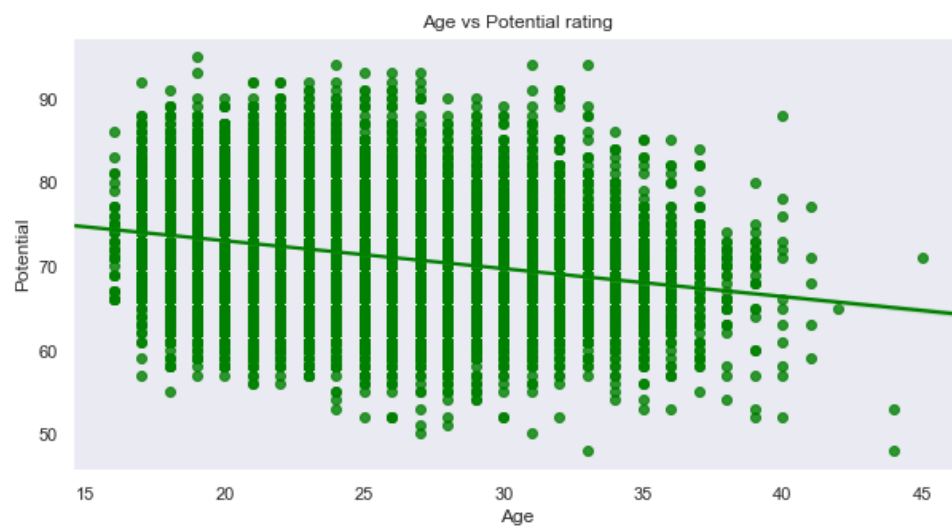
- **Wage (in millions) distribution of players:**



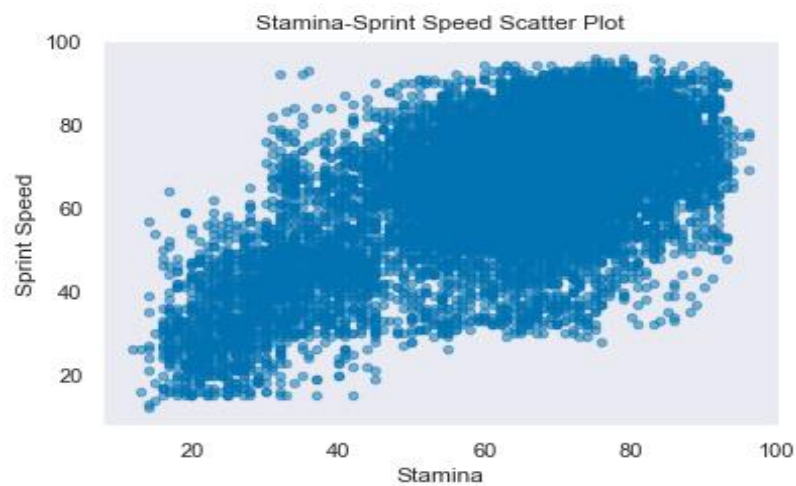
- **Relation between Age & Overall rating:**



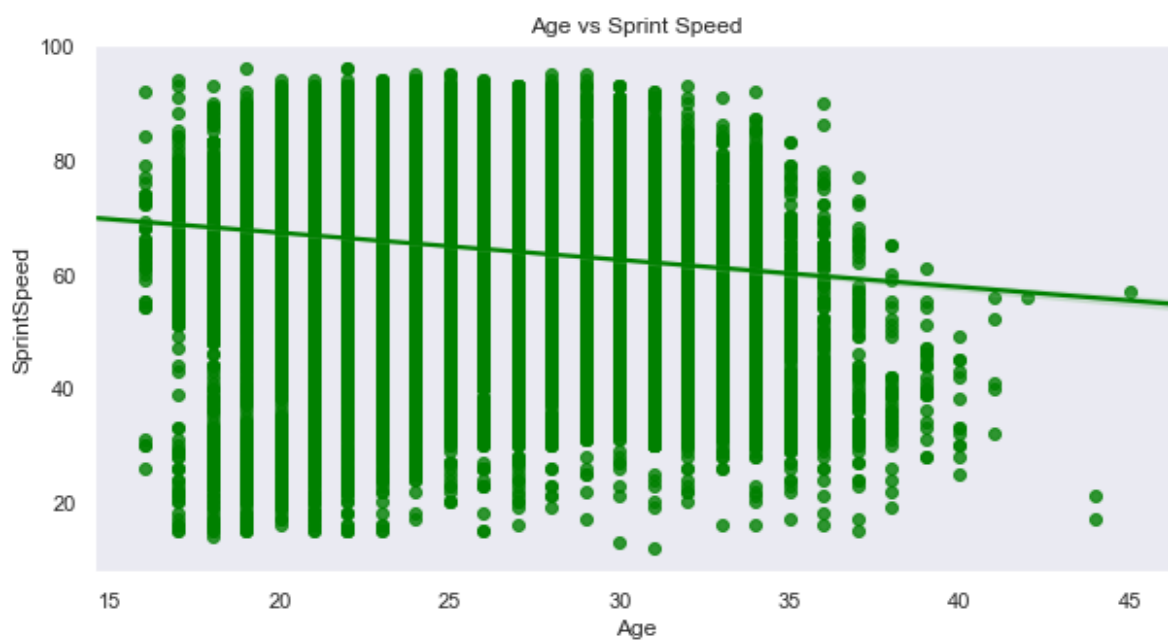
- **Relation between Age & Potential rating:**



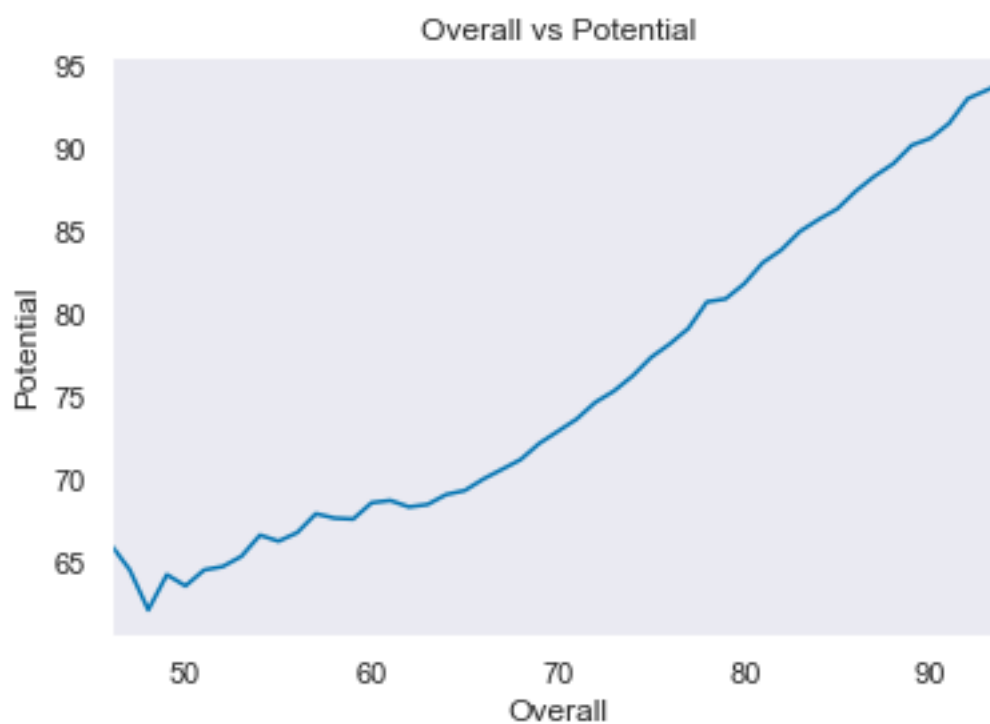
- **Relation between Stamina & Sprint speed:**



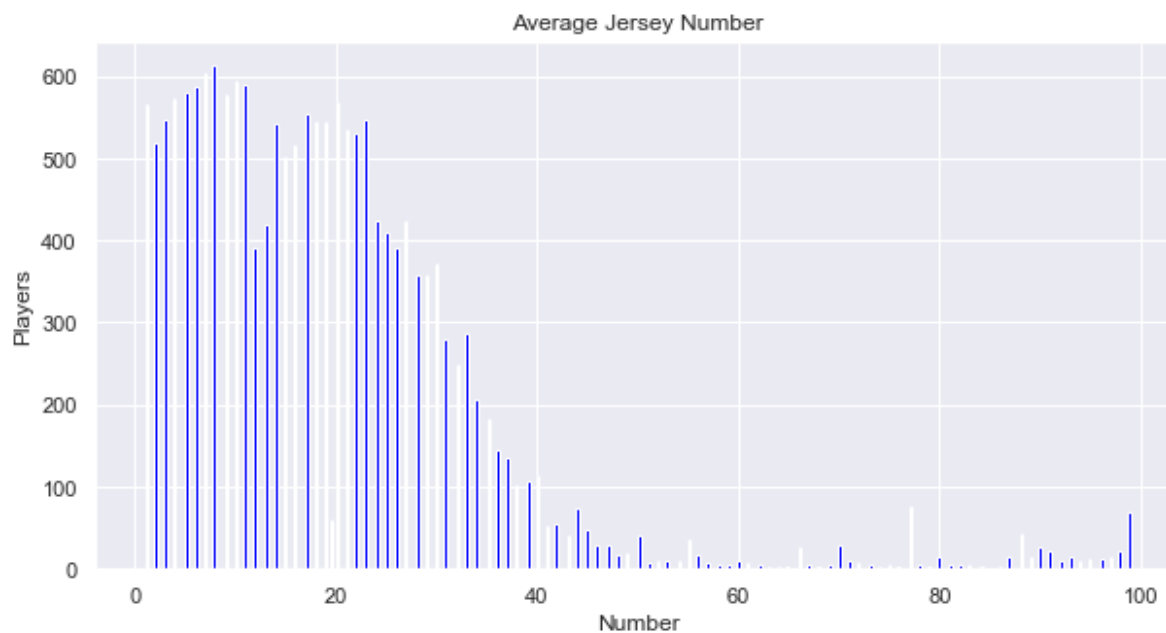
- **Relation between Age & Sprint speed:**



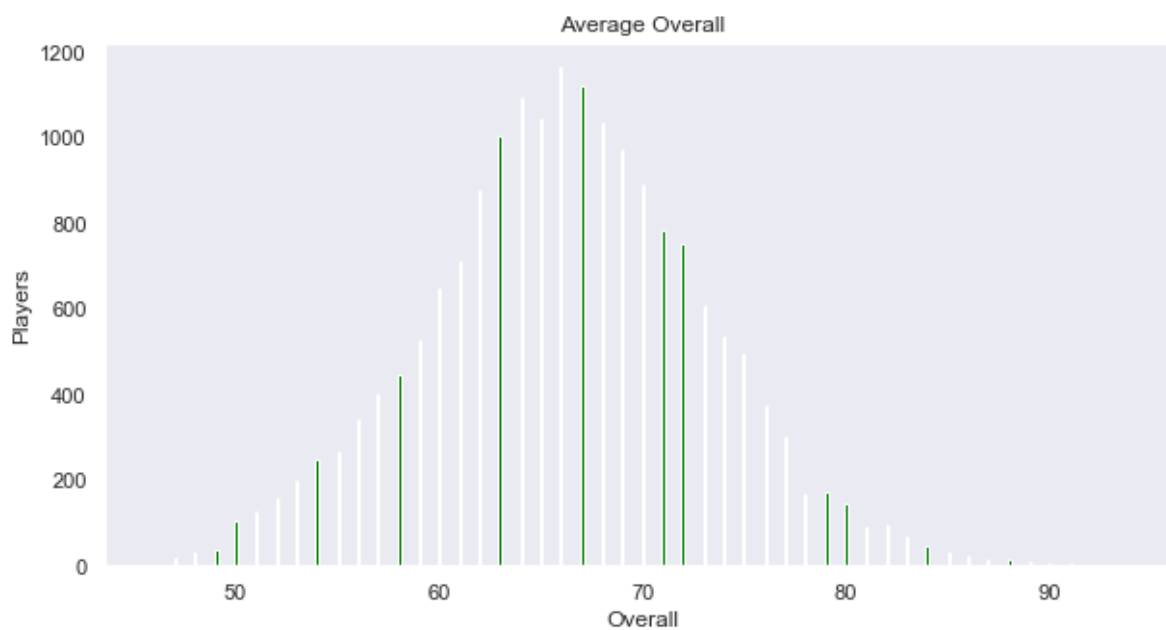
- **Relation between Overall rating & the Potential rating:**



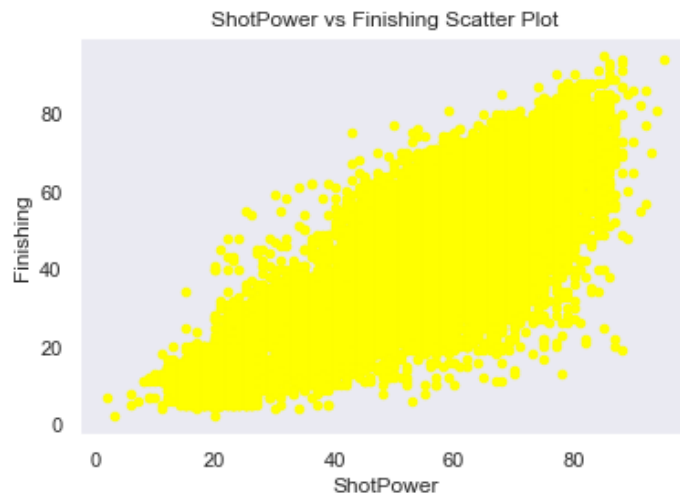
- **Distribution of players jersey number:**



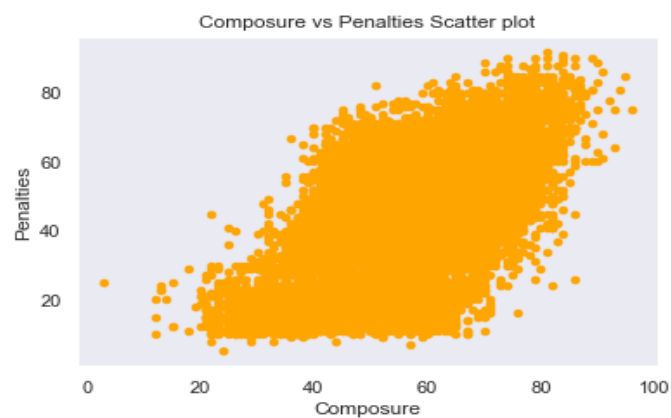
- **Distribution of Average overall of players:**



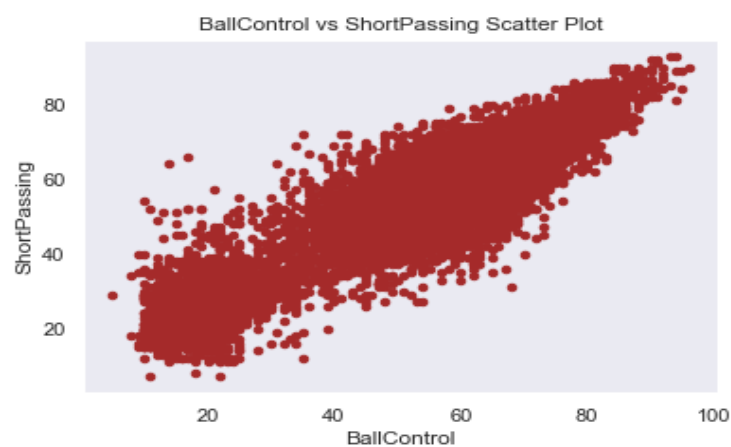
- **Relation between Shot power & Finishing:**



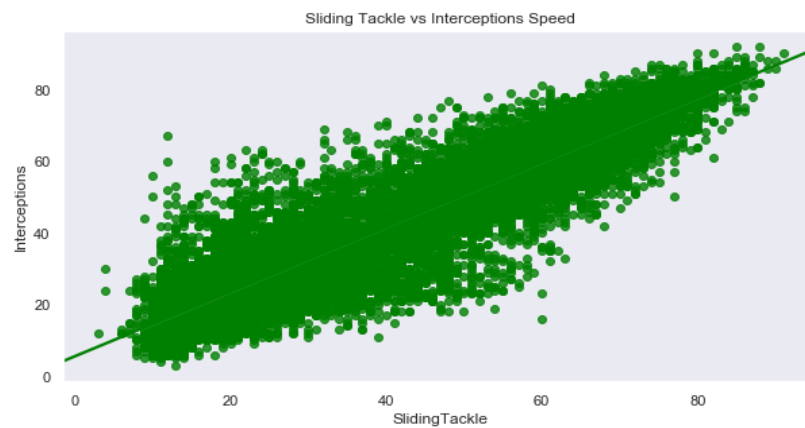
- **Relation between Composure & Penalties:**



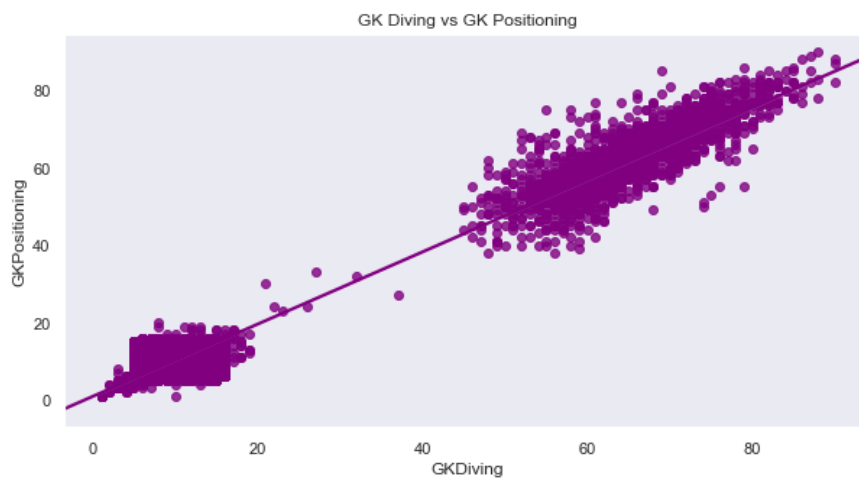
- **Relation between Ball control & Short passing:**



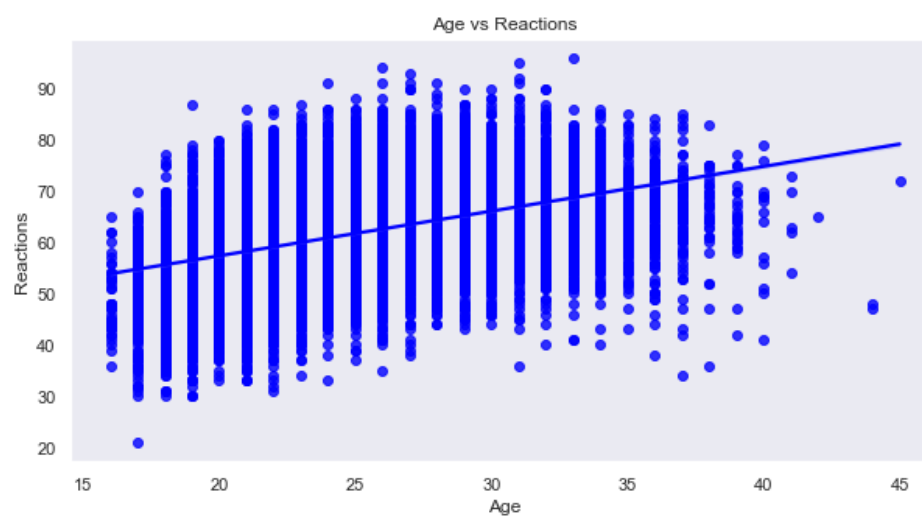
- **Relation between Sliding tackle & Interceptions:**



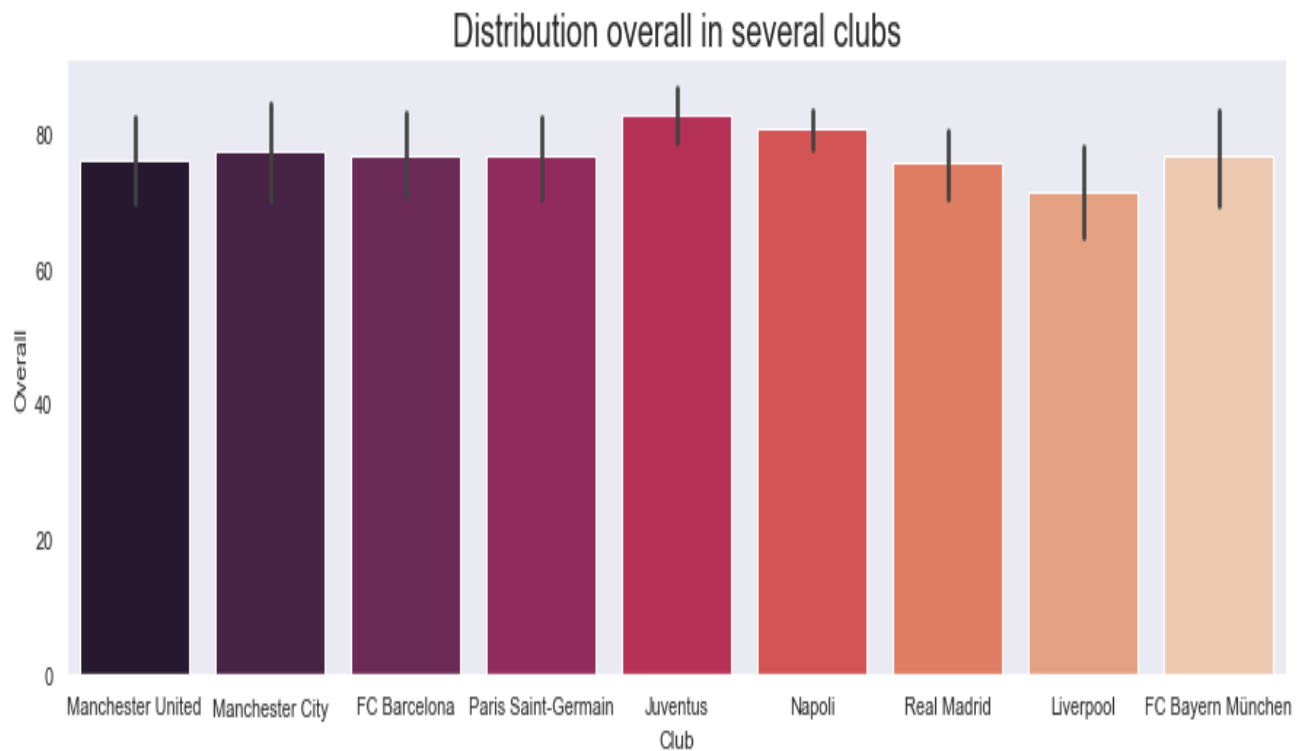
- **Relation between GK Diving & GK Positioning:**



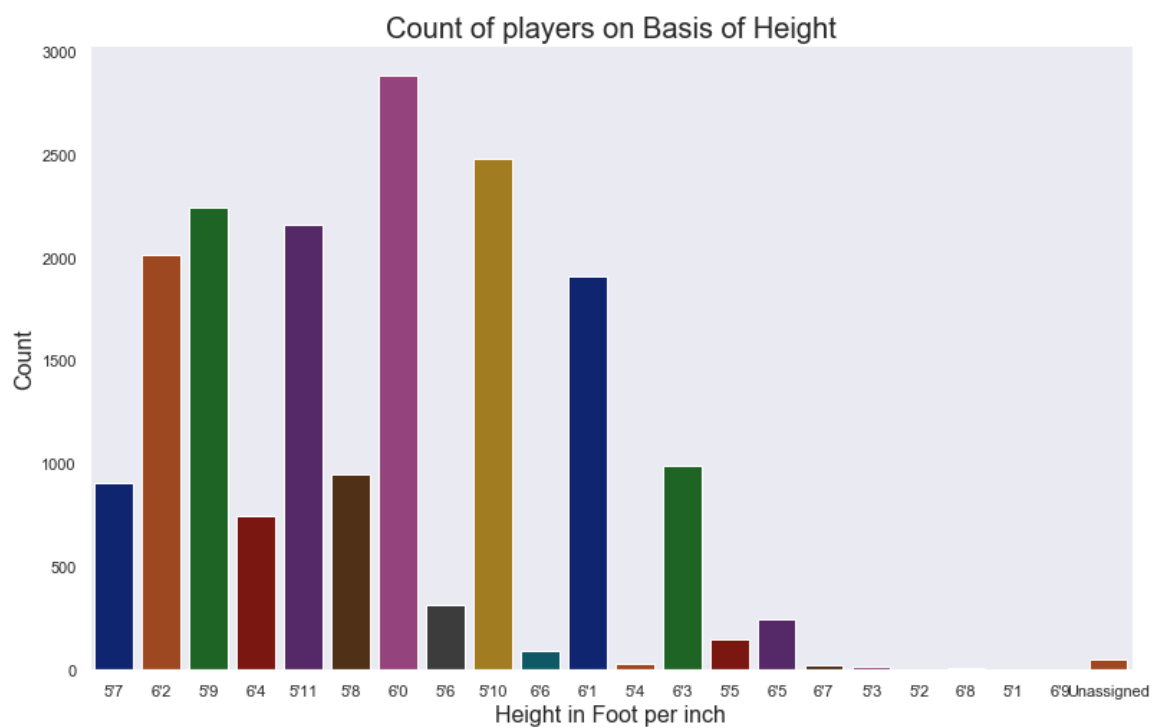
- **Relation between Age & Reactions:**



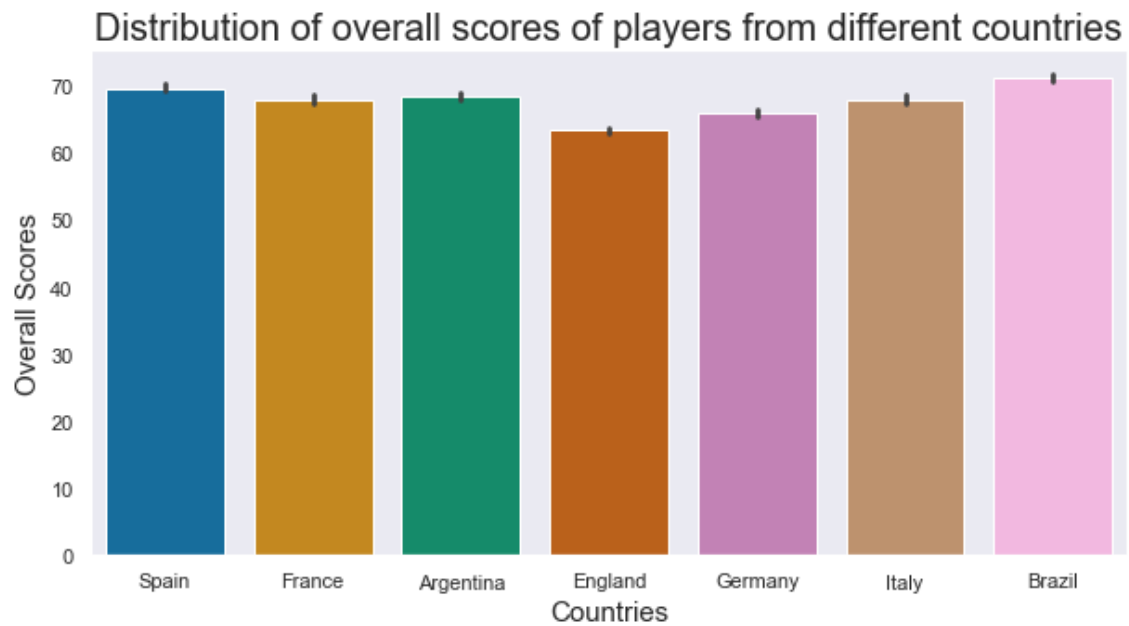
- Overall Distribution of clubs:



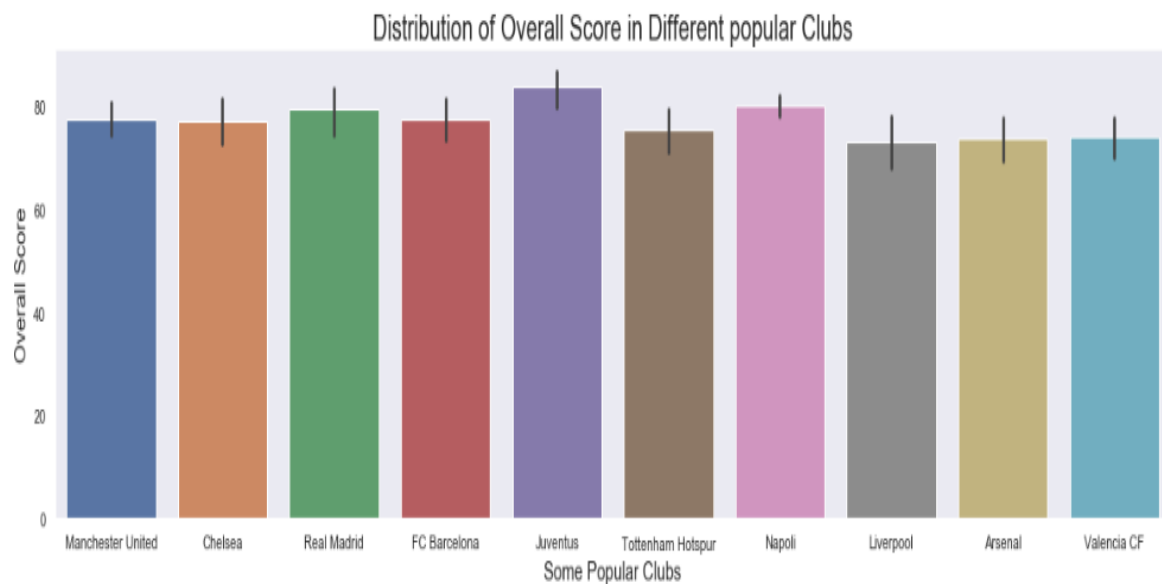
- Based on height:



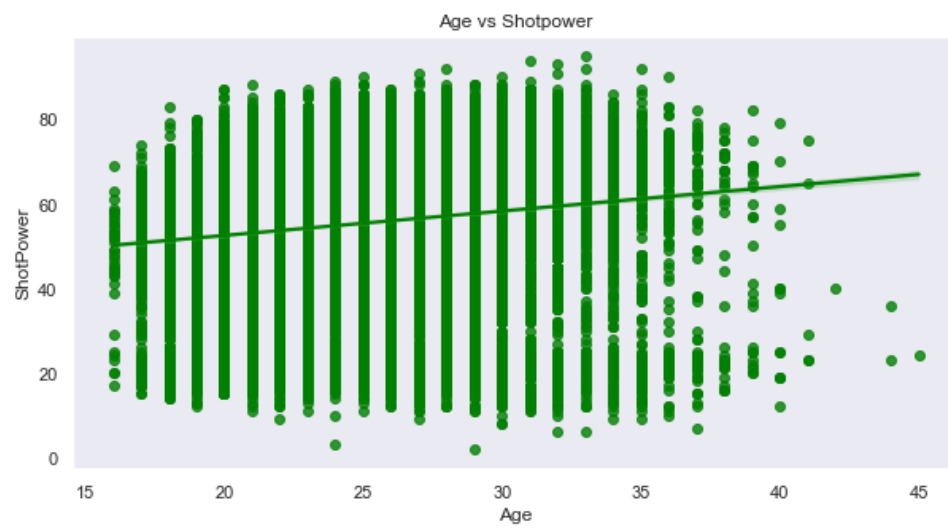
- Every Nations' Player and their overall scores:



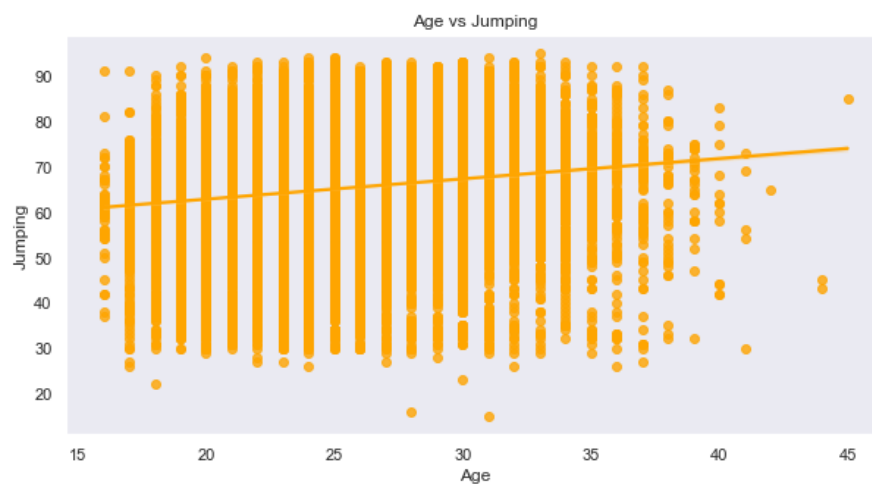
- Overall scores per club:



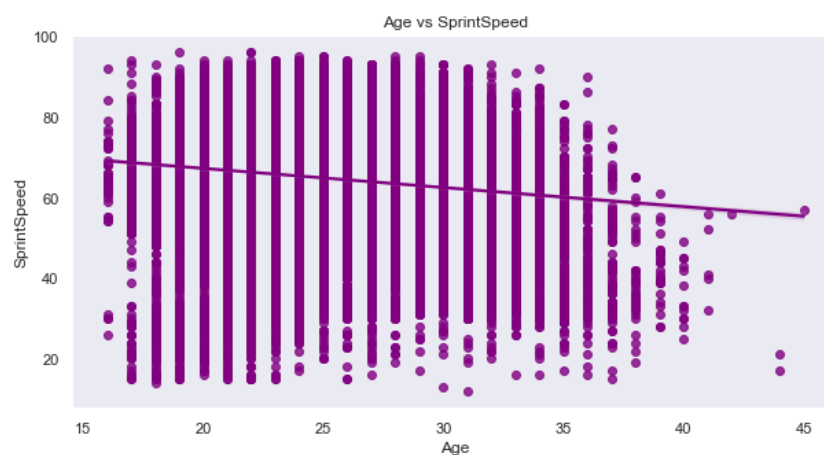
- **Relation between Age & Shot-power:**



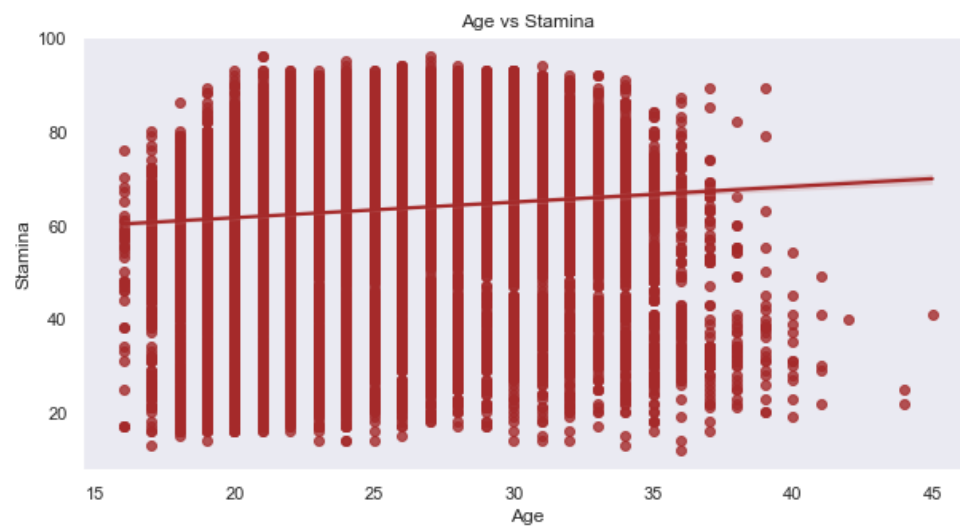
- **Relation between Age & Jumping:**



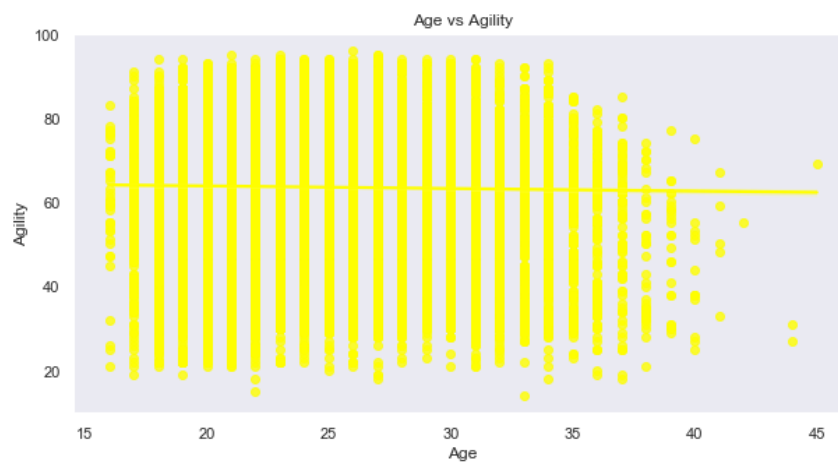
- **Relation between Age & Sprint-Speed:**



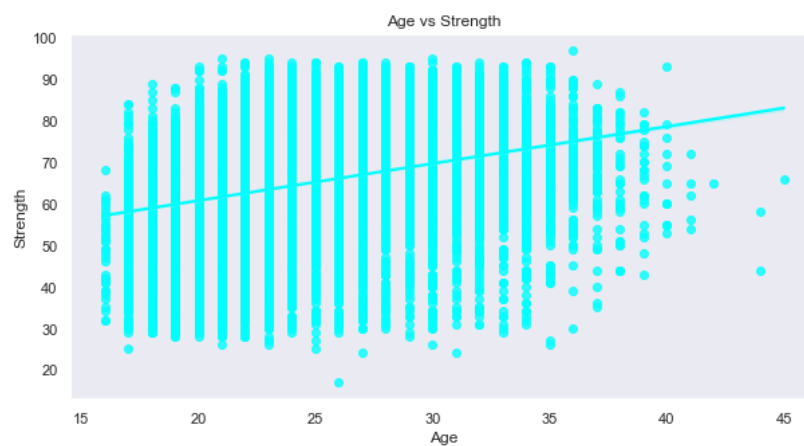
- **Relation between Age & Stamina:**



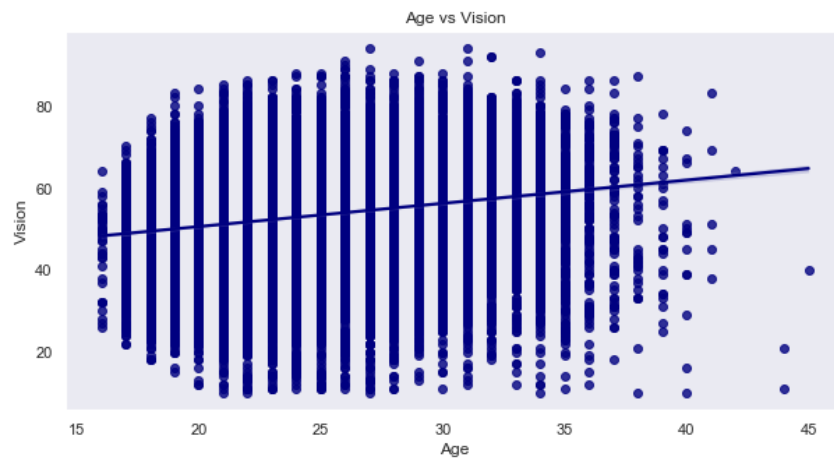
- **Relation between Age & Agility:**



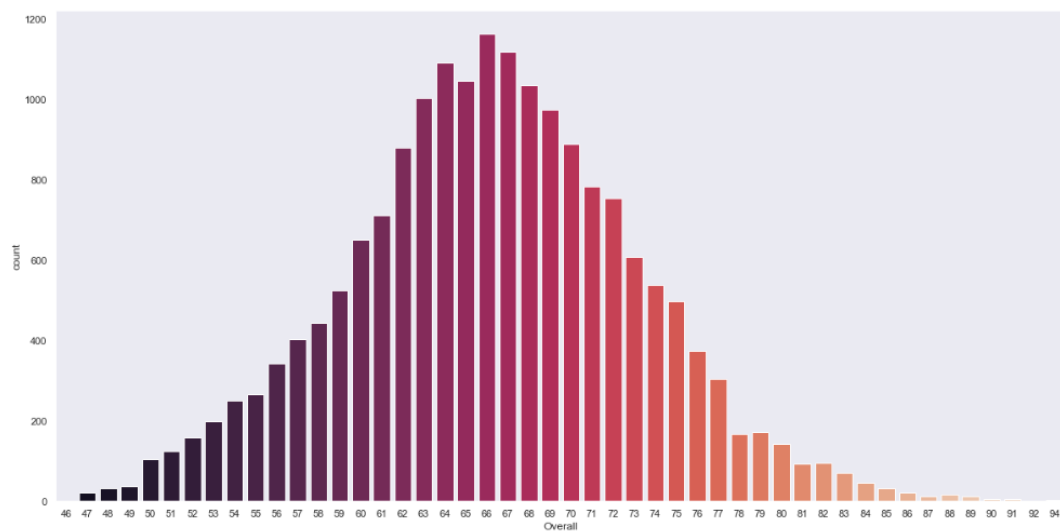
- **Relation between Age & Strength:**



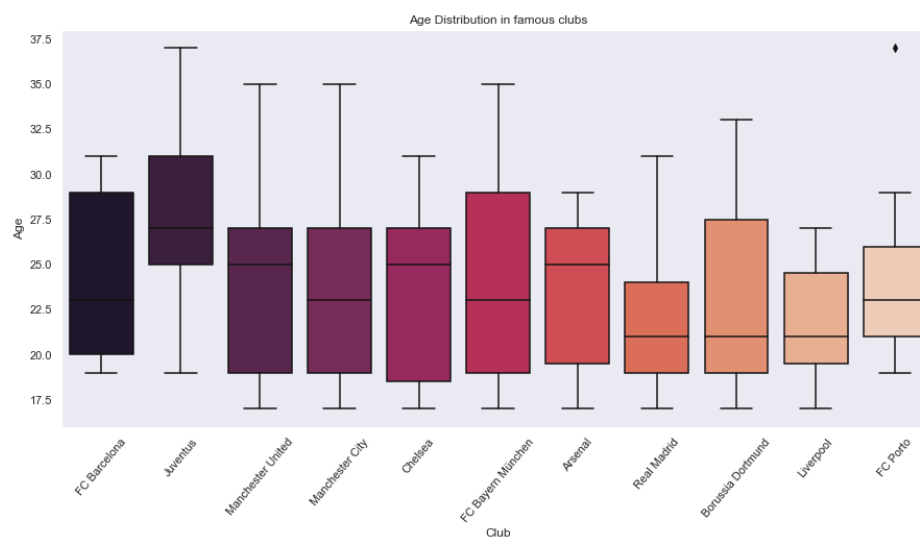
- **Relation between Age & Vision:**



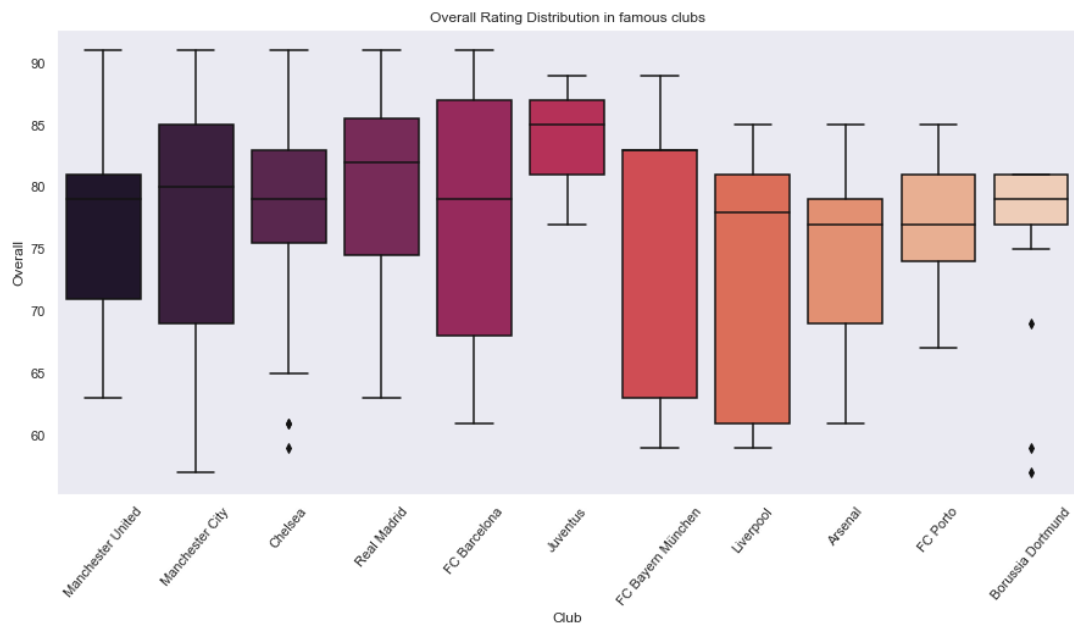
- **Distribution of players according to their Overall:**



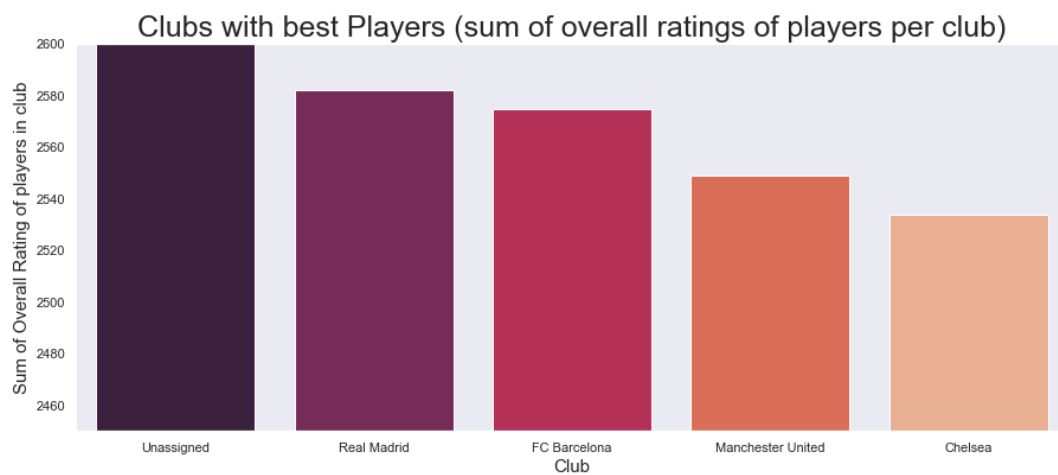
- **Age distribution among famous clubs:**



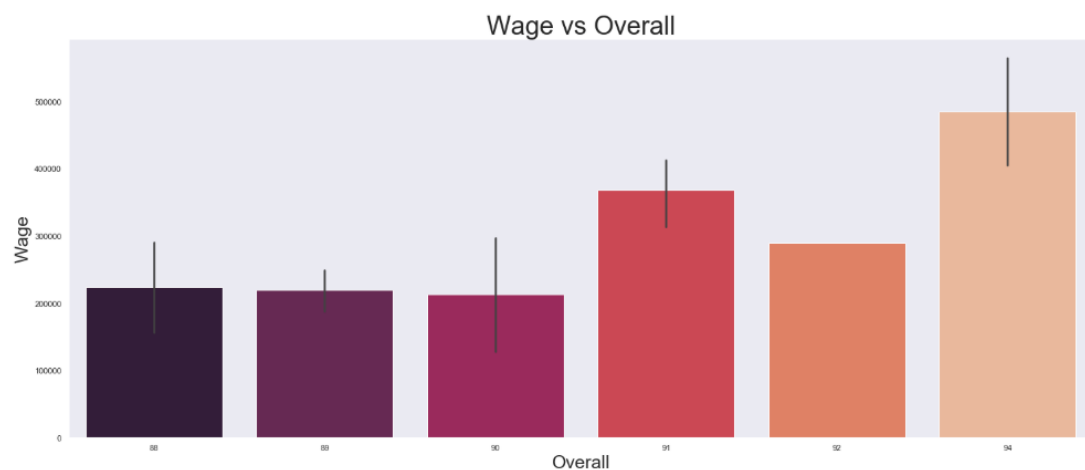
- Overall Rating of the clubs:



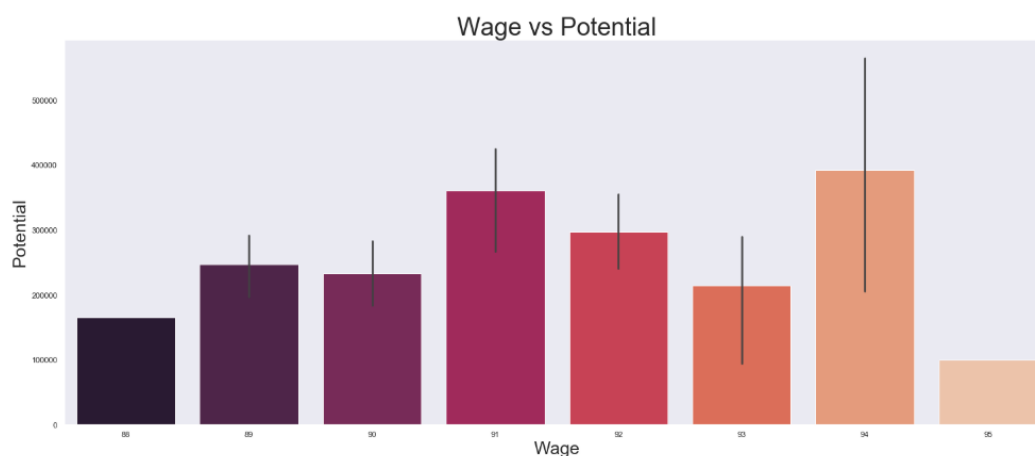
- Best club:



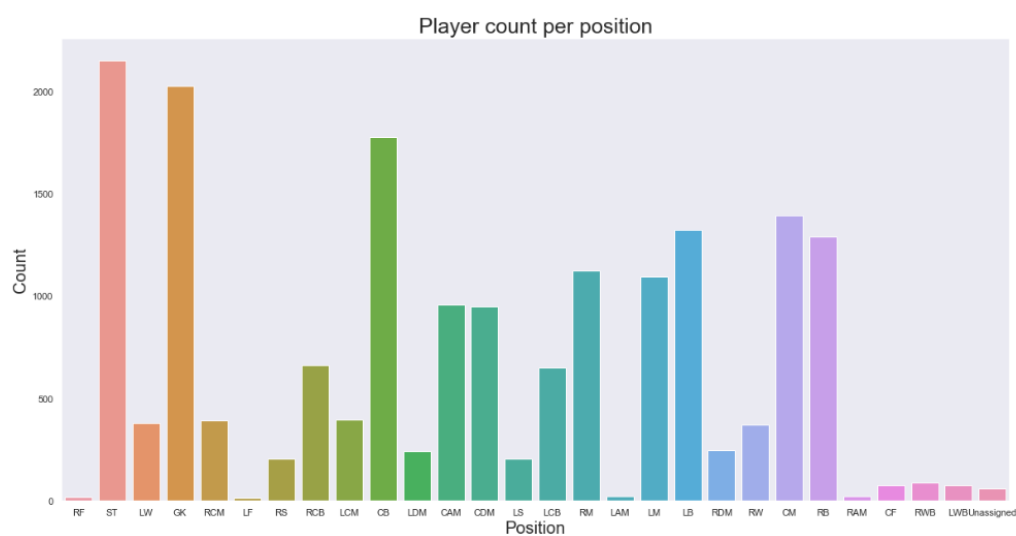
- Comparing Wage and Overall:



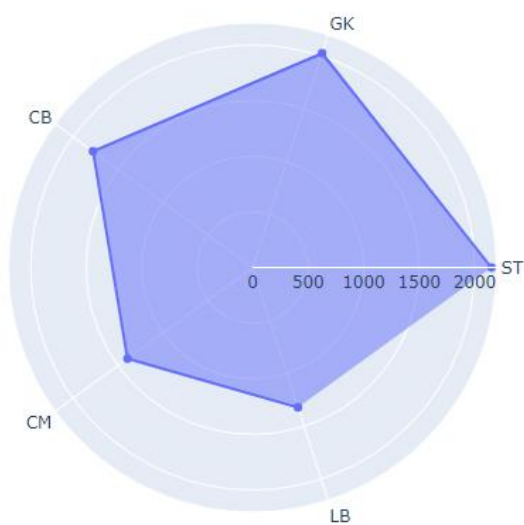
- **Comparing Wage and Potential:**



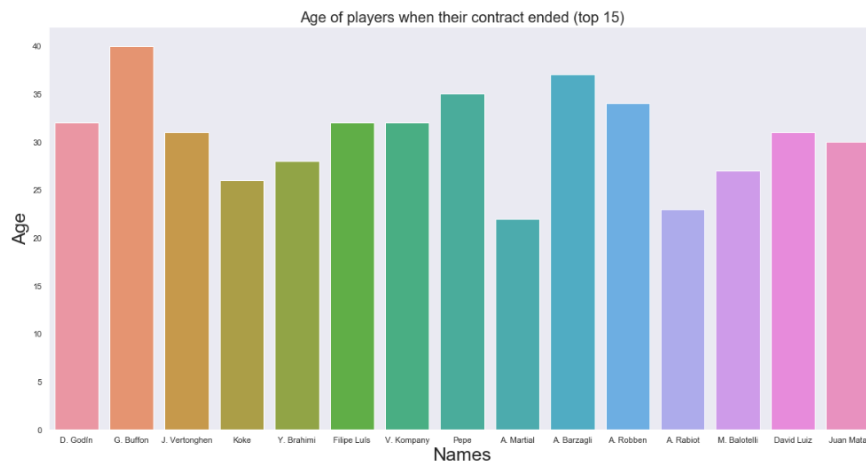
- **Player count per position:**



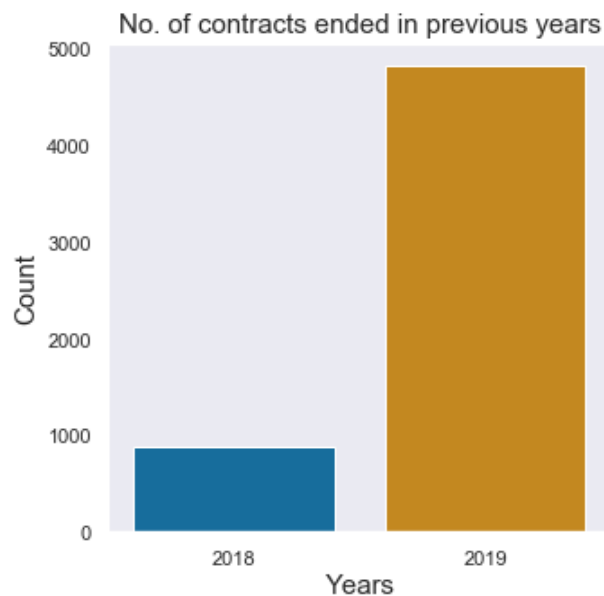
- **Radar plot for top 5 positions:**



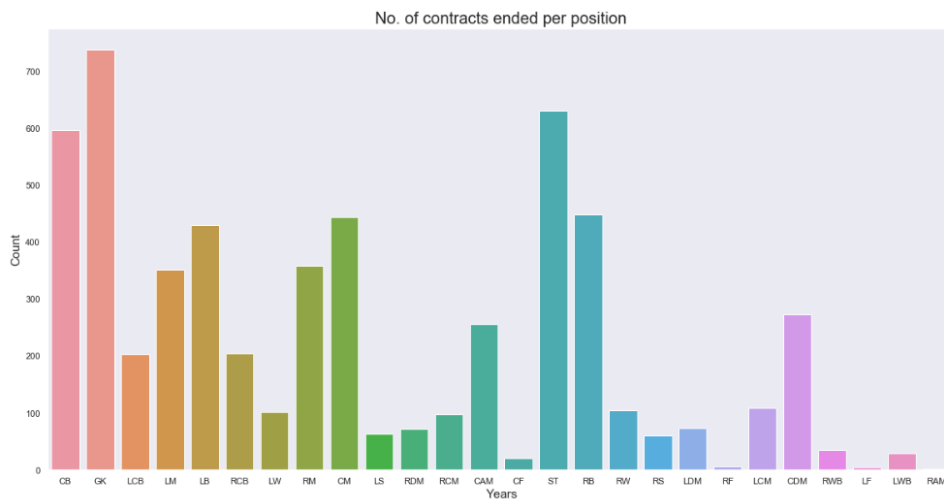
- **Players' age when their contract ended (2019):**



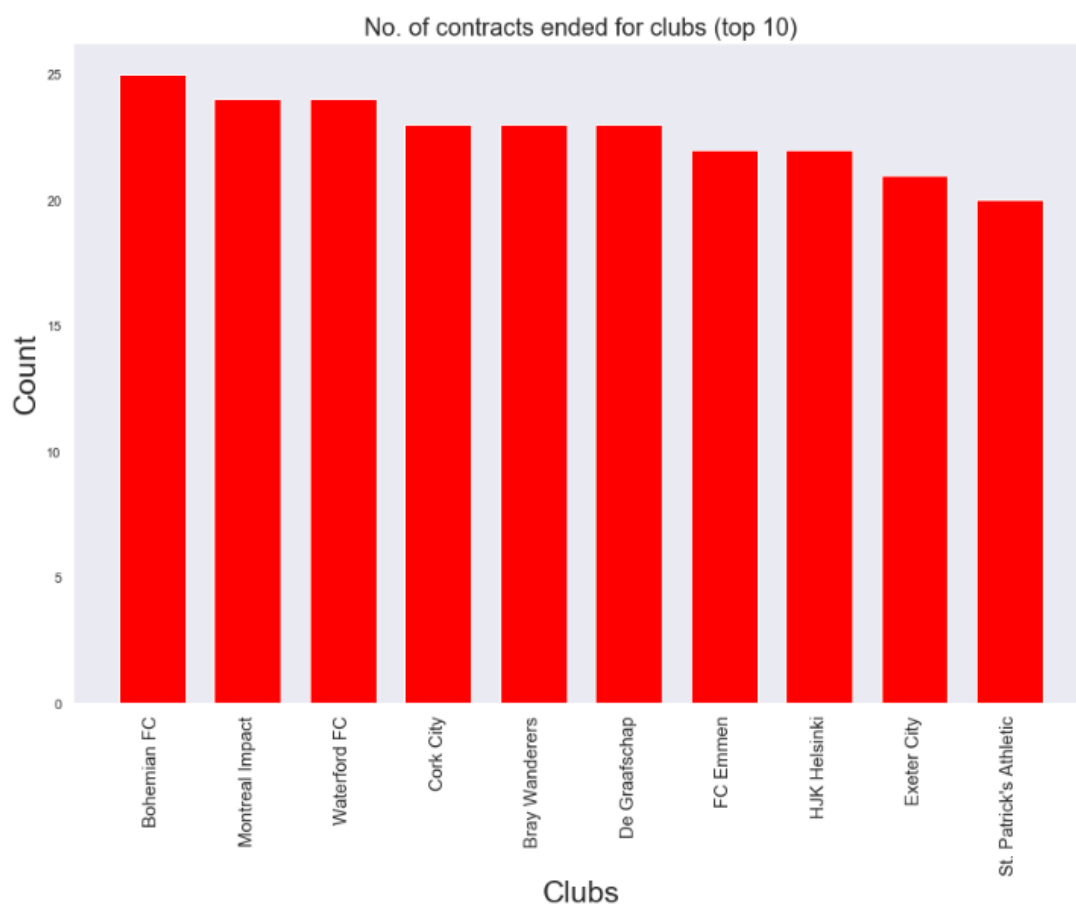
- **Contracts ended per year:**



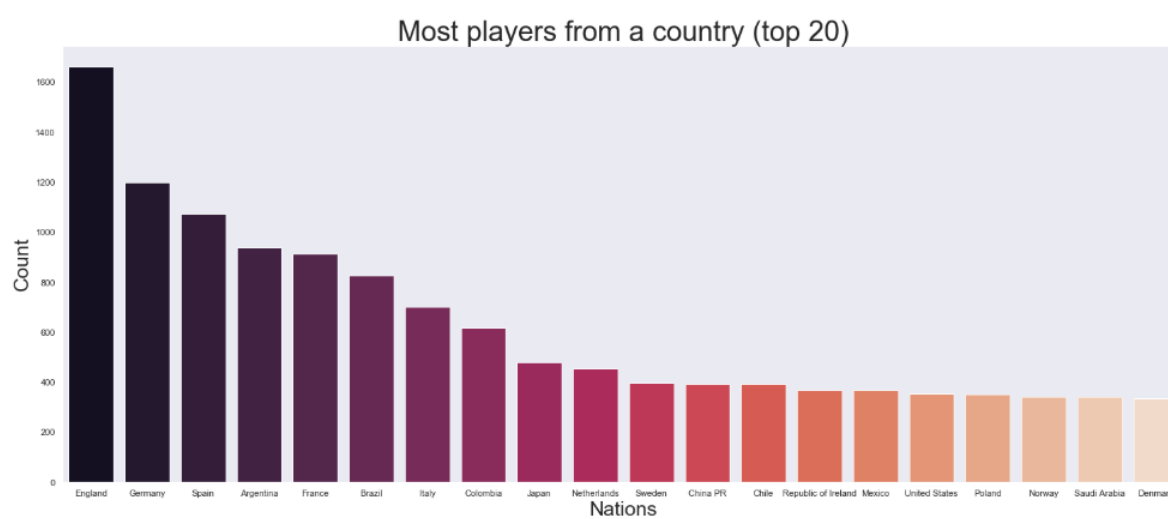
- **Contracts ended per position:**



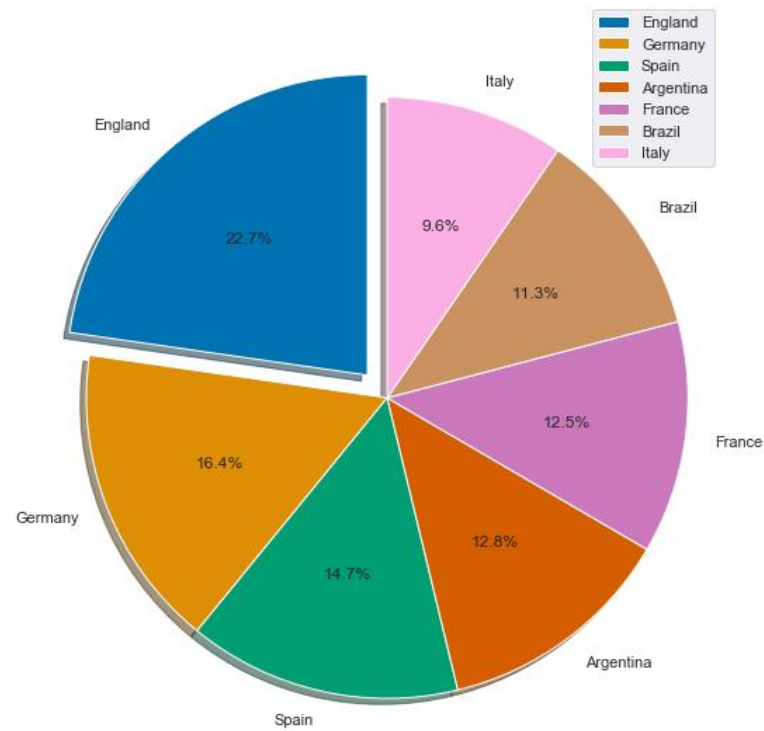
- Contracts ended per club:



- Most players from a country:

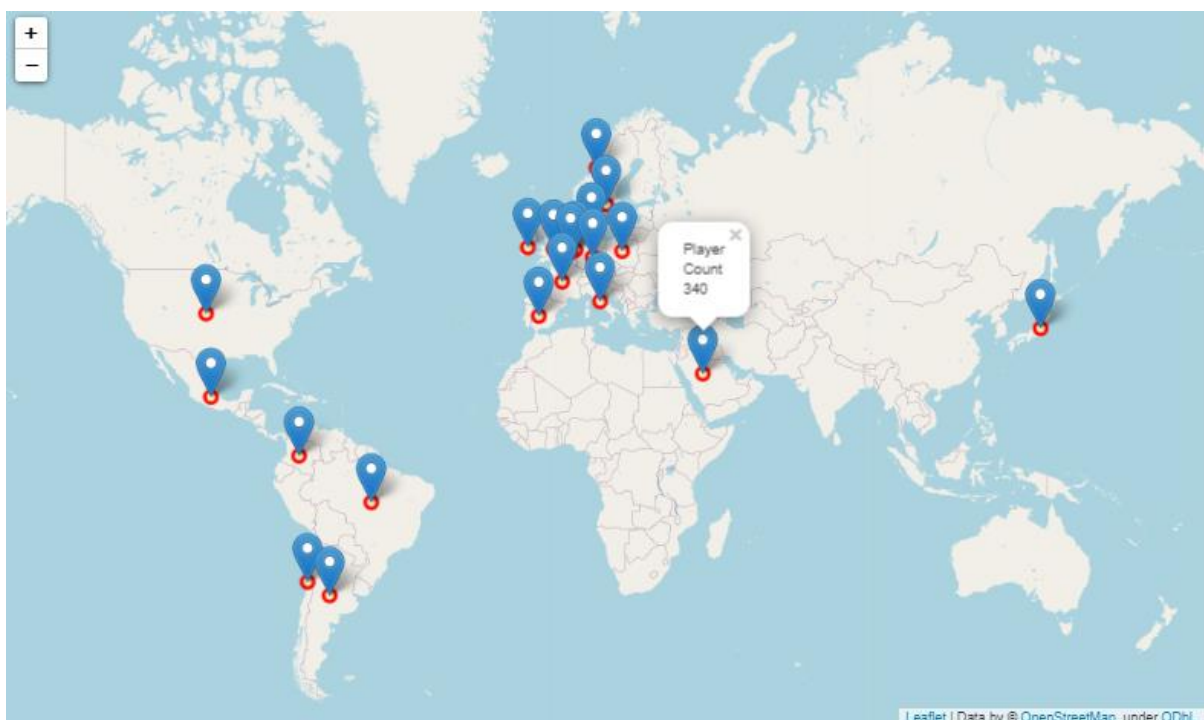


- Pie chart depiction of countries with most players:



This signifies that England and Germany have most footballers.

- World map for number of players per nation:



DATA MODELING

Here, we apply some machine learning algorithms so as to fit a model and get some insights from our data and make certain predictions.

In this project, we used 2 machine learning algorithms:

1. Regression: In this, we used certain factors, like potential, skills, etc. and fit a *linear regression* model so as to predict the '**overall**' of a player. The main module used was '**sklearn**'.

The reason to use a linear regression algorithm was based on the regression plots obtained during data visualization. We saw nearly all relations to be linear; hence this algorithm was the best fit.

We then found the accuracy of our model to know how well it predicts an unknown set of values.

2. Clustering: Here, we used the '**k-means**' clustering algorithm so as to segregate our data set into 4 parts, based on the 'potential' and 'overall' of the players.

The 4 groups are:

- Best
- Good
- Average
- Below Average

We then attached the obtained labels with the data set so that whenever any player is inspected, one can get a clue about the performance and quality of his play.

Lastly, we visualized our clusters in the form of a scatter plot, with the usage of color combinations which helps to identify the clusters easily.

Regression

We started by extracting the useful features that qualified to become the independent variables, i.e. the predictors. We selected the 'overall' column as our dependent variable.

```
predictors=df[['Overall','Potential','Value','Wage','Skill Moves',
'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','GKPositioning','GKReflexes']]
```

Train and test data split: This was done so as to separate the data-set for training and testing purposes. We use the:

1. **Training set:** To fit the regression model (75% of data set here)
2. **Testing set:** To check accuracy of our model (25% of data set here)

These splits were created using the random function.

Fitting the model: After the split, we fit the model using the 'linear regression' algorithm. This was easily done using the pre-defined functions supported by scikit-learn module in python.

We fit the model using our training data set.

- The coefficients obtained for each predictor were:

```
Coefficients: [ 2.10627489e-01  8.75778492e-08  6.24602506e-06  9.76594170e-01
 3.51105247e-02  1.64723949e-02  7.31253206e-02  4.99534810e-02
-3.19766787e-03 -1.73395578e-02  1.00924451e-03  9.01523418e-03
-6.34858080e-03  1.01765810e-01  1.01395955e-02  1.35312230e-02
 1.54747760e-02  2.27641840e-01 -1.61914151e-02  1.60970659e-02
 8.97213032e-03  1.91617039e-02  4.72023743e-02 -5.94666950e-03
 1.17998109e-02  1.16937932e-02 -3.47153836e-02 -2.52632988e-02
 6.85185081e-03  9.44627274e-02  2.97385569e-02  7.27159395e-03
-2.21729650e-02  5.74494470e-02  5.80808423e-02  2.74748760e-02
 6.88187245e-02  6.15259115e-02]
```

- The intercept of the model was:

```
Intercept: 4.169329184804496
```

Prediction: After fitting the model using the training data set, we used our test data set to check accuracy of our model.

The **predict** function supported by scikit-learn was used to do this.

```
Y=regr.predict(test[['Potential', 'Value', 'Wage', 'Skill Moves', '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']])
```

Comparison: We compared the predicted 'overall' values with the original values of our test data set. This can give the audience an obscure, yet helpful idea about the accuracy of our model. If the predicted and actual values are close enough, one can easily get a glimpse of how accurate our model predicted.

The top 5 values were generated:

Predicted Values	Actual Values
98	91
94	90
91	90
89	89
93	89

Model Evaluation: Using several classes and functions of scikit-learn module, we found the accuracy and error in our project. We also found several other deciders such as:

- Mean Absolute Error
- Residual Sum of Squares
- Variance Score
- Mean Squared Error
- R^2 Score

We found the **R^2 Score** value of our model, which gave us the accuracy probability of our project. This helped us to find the accuracy percentage.

Model Summary: The summary of the errors and accuracy in our project was:

Feature	Value
Residual Sum Of Squares	4.863
Variance Score	0.900
Mean Absolute Error	1.736
Mean Squared Error	2.205
R2-Score Percentage	88.857

Since the R^2 Score's percentage came out to be approximately 89%, we conclude that our model predicts with a good perfection. Hence, it can utterly be used to predict the 'overall' of a particular player based on his skills and potential. Using this, one can recognize the quality of the player.

This concludes the regression model fitting, with a successful and high prediction rate.

In the next phase, we apply 'Clustering' algorithm to classify our players based on their play. We can also use the result of our regression model to select the label and category of our player, because ultimately the clusters will be based on the overall and potential of the players.

Clustering

As already mentioned above, generated 4 clusters based on the 'potential' and 'overall' of the players.

The algorithm used was 'k-means' clustering, where 'k' is the number of clusters one wishes to generate.

Extracting required data: We started by selecting only those columns of our data set that we wish to keep as a factor for our clusters.

We created another data frame which consisted of only 'potential' and 'overall' column.

Data Normalization/Standardization: So as to fit our model with suitable values, we normalized the values in a specific and standard range.

This was done using the '*preprocessing*' class of scikit-learn.

After processing the data, we got values as:

```
array([[ 3.69809177],
       [ 3.69809177],
       [ 3.53512784],
       ...,
       [-0.70193445],
       [-0.86489839],
       [-0.86489839]])
```

Fitting the model: We applied the 'k-means' clustering here and generated the labels for our data set.

The generated labels were in a range: **0-3**.

The generated list of labels for each row was concatenated with extracted data set. This can easily determine the labels of their respected row.

Clusters: Based on the 4 clusters generated, we calculated the mean of potential and overall of each of them. This helped us to assign classes to our clusters.

- **Cluster 1:** The first 5 results for first cluster were:

	Name	Overall	Potential	Labels
737	Sidney Pessinho	78	78	0
738	Everticinho	78	78	0
739	Claudio Coíntra	78	78	0
740	Ronaldo Esler	78	78	0
749	M. Díaz	78	78	0

Mean Potential of cluster: 74.2297538351766
 Mean Overall of cluster: 68.18444523724581
 Members in cluster 1: 5606

- **Cluster 2:** The first 5 results for second cluster were:

	Name	Overall	Potential	Labels
3968	Allison Sireo	71	71	1
3969	Z. Stieber	71	71	1
3971	O. Skúlason	71	71	1
3973	M. Pektemek	71	71	1
3974	R. Schüller	71	71	1

Mean Potential of cluster: 68.60502936304773
 Mean Overall of cluster: 64.44285499171812
 Members in cluster 1: 6641

- **Cluster 3:** The first 5 results for third cluster were:

	Name	Overall	Potential	Labels
0	L. Messi	94	94	2
1	Cristiano Ronaldo	94	94	2
2	Neymar Jr	92	93	2
3	De Gea	91	93	2
4	K. De Bruyne	91	92	2

Mean Potential of cluster: 81.20923722883136
 Mean Overall of cluster: 73.55213435969209
 Members in cluster 1: 2858

- **Cluster 4:** The first 5 results for fourth cluster were:

	Name	Overall	Potential	Labels
9928	J. Akinde	65	65	3
9929	D. McGregor	65	65	3
9932	Teixeira José	65	65	3
9933	A. Fernández	65	65	3
9938	A. Considine	65	65	3

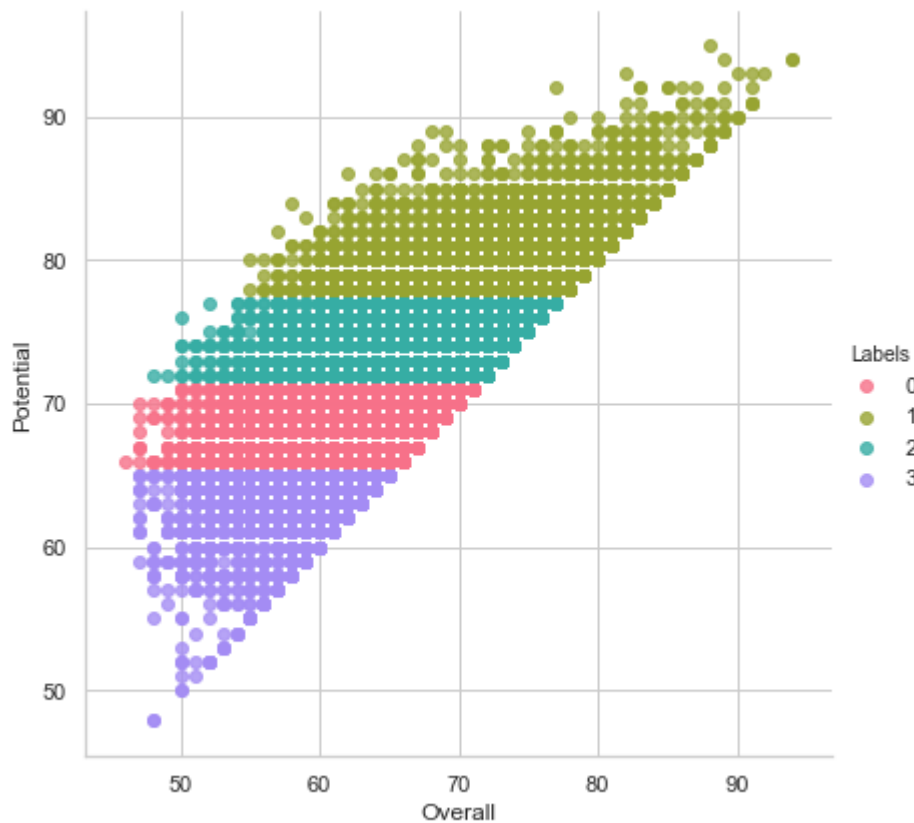
Mean Potential of cluster: 62.687943262411345
 Mean Overall of cluster: 59.82882011605416
 Members in cluster 1: 3102

Hence, the mean values of the 4 clusters clearly signified the correct segregation of our data set. This helped us to assign the classes to the clusters without any perplexity.

Summary: The report of our model was:

	Tags	Potential Mean	Overall Mean	Players	Cluster
Index					
1	Best	81.969	74.302	2311	3
2	Good	74.565	68.380	6153	1
3	Average	68.605	64.443	6641	2
4	Below Average	62.688	59.829	3102	4

Visualizing Clusters: This was the final job of our project. We visualized the 4 clusters in the form of a scatter plot, with different colour codes.



OBSERVATIONS

We made several observations from our project. Some of the useful insights are:

- Most of the players belonged to England and Germany.
- Most of the players had the field position 'ST'.
- Most of the contracts ended in the year 2019
- Most contracts ended for the position 'GK'.
- Wages of players depended on their overall performance than their potential.
- Real Madrid and FC Barcelona had best players among all.
- O. Perez of Pachuca club is the eldest player (45 years).
- G. Nugent of Tranmere Rovers club is the youngest player (16 years).
- L. Messi of FC Barcelona club is the best free-kick taker.
- M. Balotelli of OGC Nice club is the best Penalty taker.
- L. Messi has the best ball control.
- L Sane of Manchester City is the quickest player.

Project URL:

<https://nbviewer.jupyter.org/github/Prasfur/Prasfurs-Projects/blob/master/FIFA19%20-%20Notebook.ipynb>