DWM EXP3: Data Preprocessing

Aim: Given a case study of given data set. You are expected to perform data preprocessing/cleaning using Python/R. Quote your observations after the preprocessing(null values, Useless columns removed, etc).

Name: Meet Dave UID:2018140015 Batch:A TE-IT

In [1]:
```python
import pandas as pd
```

In [2]:
```python
df=pd.read_csv('data.csv')
df.head()
```

Out[2]:

	Unnamed: 0	ID	Name	Age	Photo	Nationality	Flag	Overall	Potenti
0	0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	9
1	1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	9
2	2	190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	https://cdn.sofifa.org/flags/54.png	92	9
3	3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	9
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	9

5 rows × 89 columns

In [3]:
```python
#Dropping the Unnamed:0 Column
df=df.drop(['Unnamed: 0'],axis=1)
df.head()
```

Out[3]:

	ID	Name	Age	Photo	Nationality	Flag	Overall	Potential	Cl
0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	94	F Barcelo
1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Juvent
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 Sai Germa
3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	93	Manches Unit
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	Manches C

5 rows × 88 columns

In [4]:
```python
df.shape
```

Out[4]:
```
(18207, 88)
```

In [5]:
```python
df.isna().sum()
```

Out[5]:
```
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
```

In [6]:
```python
#changing options to display all columns and rows
pd.options.display.max_columns= None
pd.options.display.max_rows=None
```

In [7]:
```python
df.isna().sum()
```

Out[7]:
```
ID                              0
Name                            0
Age                             0
Photo                           0
Nationality                     0
Flag                            0
Overall                         0
Potential                       0
Club                          241
Club Logo                       0
Value                           0
Wage                            0
Special                         0
Preferred Foot                 48
International Reputation       48
Weak Foot                      48
Skill Moves                    48
Work Rate                      48
Body Type                      48
Real Face                      48
Position                       60
Jersey Number                  60
Joined                       1553
Loaned From                 16943
Contract Valid Until          289
Height                         48
Weight                         48
LS                           2085
ST                           2085
RS                           2085
LW                           2085
LF                           2085
CF                           2085
RF                           2085
RW                           2085
LAM                          2085
CAM                          2085
RAM                          2085
LM                           2085
LCM                          2085
CM                           2085
RCM                          2085
RM                           2085
LWB                          2085
LDM                          2085
CDM                          2085
RDM                          2085
RWB                          2085
LB                           2085
LCB                          2085
CB                           2085
RCB                          2085
RB                           2085
Crossing                       48
Finishing                      48
HeadingAccuracy                48
ShortPassing                   48
Volleys                        48
Dribbling                      48
Curve                          48
FKAccuracy                     48
LongPassing                    48
BallControl                    48
Acceleration                   48
SprintSpeed                    48
Agility                        48
Reactions                      48
Balance                        48
ShotPower                      48
Jumping                        48
Stamina                        48
Strength                       48
LongShots                      48
Aggression                     48
Interceptions                  48
Positioning                    48
Vision                         48
Penalties                      48
Composure                      48
Marking                        48
StandingTackle                 48
SlidingTackle                  48
GKDiving                       48
GKHandling                     48
GKKicking                      48
GKPositioning                  48
GKReflexes                     48
Release Clause               1564
dtype: int64
```

In [8]:
```python
#Out of 18207 rows, nearly 17000 rows are empty ofLoaned from column
df=df.drop(['Loaned From'],axis=1)
df.head()
```

Out[8]:

	ID	Name	Age	Photo	Nationality	Flag	Overall	Potential	Cl
0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	94	F Barcelo
1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Juvent
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 Sai Germa
3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	93	Manches Unit
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	Manches C

Our prediction is not dependent on player's id,nationality,photo,name,jersey number,club,club logo,flag and real face

In [9]:
```python
df=df.drop(['ID','Name','Photo','Nationality','Flag','Body Type',
            'Jersey Number','Club','Club Logo','Flag','Real Face'],axis=1)
df.head()
```

Out[9]:

	Age	Overall	Potential	Value	Wage	Special	Preferred Foot	International Reputation	Weak Foot	Skill Moves	Work Rate	Position	Joined	Contract Valid Until	Height
0	31	94	94	€110.5M	€565K	2202	Left	5.0	4.0	4.0	Medium/ Medium	RF	Jul 1, 2004	2021	5'7
1	33	94	94	€77M	€405K	2228	Right	5.0	4.0	5.0	High/ Low	ST	Jul 10, 2018	2022	6'2
2	26	92	93	€118.5M	€290K	2143	Right	5.0	5.0	5.0	High/ Medium	LW	Aug 3, 2017	2022	5'9
3	27	91	93	€72M	€260K	1471	Right	4.0	3.0	1.0	Medium/ Medium	GK	Jul 1, 2011	2020	6'4
4	27	91	92	€102M	€355K	2281	Right	4.0	5.0	4.0	High/ High	RCM	Aug 30, 2015	2023	5'11

In [10]:
```python
df=df.dropna()
df.shape
```

Out[10]:
```
(14743, 77)
```

In [11]:
```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 14743 entries, 0 to 18206
Data columns (total 77 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       14743 non-null  int64
 1   Overall                   14743 non-null  int64
 2   Potential                 14743 non-null  int64
 3   Value                     14743 non-null  object
 4   Wage                      14743 non-null  object
 5   Special                   14743 non-null  int64
 6   Preferred Foot            14743 non-null  object
 7   International Reputation  14743 non-null  float64
 8   Weak Foot                 14743 non-null  float64
 9   Skill Moves               14743 non-null  float64
 10  Work Rate                 14743 non-null  object
 11  Position                  14743 non-null  object
 12  Joined                    14743 non-null  object
 13  Contract Valid Until      14743 non-null  object
 14  Height                    14743 non-null  object
 15  Weight                    14743 non-null  object
 16  LS                        14743 non-null  object
 17  ST                        14743 non-null  object
 18  RS                        14743 non-null  object
 19  LW                        14743 non-null  object
 20  LF                        14743 non-null  object
 21  CF                        14743 non-null  object
 22  RF                        14743 non-null  object
 23  RW                        14743 non-null  object
 24  LAM                       14743 non-null  object
 25  CAM                       14743 non-null  object
 26  RAM                       14743 non-null  object
 27  LM                        14743 non-null  object
 28  LCM                       14743 non-null  object
 29  CM                        14743 non-null  object
 30  RCM                       14743 non-null  object
 31  RM                        14743 non-null  object
 32  LWB                       14743 non-null  object
 33  LDM                       14743 non-null  object
 34  CDM                       14743 non-null  object
 35  RDM                       14743 non-null  object
 36  RWB                       14743 non-null  object
 37  LB                        14743 non-null  object
 38  LCB                       14743 non-null  object
 39  CB                        14743 non-null  object
 40  RCB                       14743 non-null  object
 41  RB                        14743 non-null  object
 42  Crossing                  14743 non-null  float64
 43  Finishing                 14743 non-null  float64
 44  HeadingAccuracy           14743 non-null  float64
 45  ShortPassing              14743 non-null  float64
 46  Volleys                   14743 non-null  float64
 47  Dribbling                 14743 non-null  float64
 48  Curve                     14743 non-null  float64
 49  FKAccuracy                14743 non-null  float64
 50  LongPassing               14743 non-null  float64
 51  BallControl               14743 non-null  float64
 52  Acceleration              14743 non-null  float64
 53  SprintSpeed               14743 non-null  float64
 54  Agility                   14743 non-null  float64
 55  Reactions                 14743 non-null  float64
 56  Balance                   14743 non-null  float64
 57  ShotPower                 14743 non-null  float64
 58  Jumping                   14743 non-null  float64
 59  Stamina                   14743 non-null  float64
 60  Strength                  14743 non-null  float64
 61  LongShots                 14743 non-null  float64
 62  Aggression                14743 non-null  float64
 63  Interceptions             14743 non-null  float64
 64  Positioning               14743 non-null  float64
 65  Vision                    14743 non-null  float64
 66  Penalties                 14743 non-null  float64
 67  Composure                 14743 non-null  float64
 68  Marking                   14743 non-null  float64
 69  StandingTackle            14743 non-null  float64
 70  SlidingTackle             14743 non-null  float64
 71  GKDiving                  14743 non-null  float64
 72  GKHandling                14743 non-null  float64
 73  GKKicking                 14743 non-null  float64
 74  GKPositioning             14743 non-null  float64
 75  GKReflexes                14743 non-null  float64
 76  Release Clause            14743 non-null  object
dtypes: float64(37), int64(4), object(36)
memory usage: 8.8+ MB
```

In [12]:
```python
position = ['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']
def player_number_change(x):
     if type(x) == str:
        return float(x.split("+")[0])
     else:
        return 0.0
for i in position:
     df[i]=df[i].apply(player_number_change)
```

In [13]:
```python
def player_height_change(height):
    h = height.split("'")
    return float(h[0])*12 + float(h[1])
df["Height"]=df["Height"].apply(player_height_change)
df["Weight"] = df["Weight"].apply(lambda x : float(x[:-3]) )
df['Preferred Foot']=df['Preferred Foot'].apply(lambda x: float(1) if x=='Right' else float(0))
df.head()
```

Out[13]:

	Age	Overall	Potential	Value	Wage	Special	Preferred Foot	International Reputation	Weak Foot	Skill Moves	Work Rate	Position	Joined	Contract Valid Until	Height
0	31	94	94	€110.5M	€565K	2202	0.0	5.0	4.0	4.0	Medium/ Medium	RF	Jul 1, 2004	2021	67.0
1	33	94	94	€77M	€405K	2228	1.0	5.0	4.0	5.0	High/ Low	ST	Jul 10, 2018	2022	74.0
2	26	92	93	€118.5M	€290K	2143	1.0	5.0	5.0	5.0	High/ Medium	LW	Aug 3, 2017	2022	69.0
4	27	91	92	€102M	€355K	2281	1.0	4.0	5.0	4.0	High/ High	RCM	Aug 30, 2015	2023	71.0
5	27	91	91	€93M	€340K	2142	1.0	4.0	4.0	4.0	High/ Medium	LF	Jul 1, 2012	2020	68.0

In [22]:
```python
def player_money_change(e):
    if e[-1] == 'K':
        return float(e[1:-1])/1000.0
    else:
        return float(e[1:-1])

for money in ['Value','Wage','Release Clause']:
     df[money]=df[money].apply(player_money_change)
df.head()
```

Out[22]:

	Age	Overall	Potential	Value	Wage	Special	Preferred Foot	International Reputation	Weak Foot	Skill Moves	Work Rate	Position	Joined	Contract Valid Until	Height	W
0	31	94	94	110.5	0.565	2202	0.0	5.0	4.0	4.0	Medium/ Medium	RF	Jul 1, 2004	2021	67.0	
1	33	94	94	77.0	0.405	2228	1.0	5.0	4.0	5.0	High/ Low	ST	Jul 10, 2018	2022	74.0	
2	26	92	93	118.5	0.290	2143	1.0	5.0	5.0	5.0	High/ Medium	LW	Aug 3, 2017	2022	69.0	
4	27	91	92	102.0	0.355	2281	1.0	4.0	5.0	4.0	High/ High	RCM	Aug 30, 2015	2023	71.0	
5	27	91	91	93.0	0.340	2142	1.0	4.0	4.0	4.0	High/ Medium	LF	Jul 1, 2012	2020	68.0	

In [23]:
```python
df['Work Rate'].value_counts()
```

Out[23]:
```
Medium/ Medium    7119
High/ Medium      2886
Medium/ High      1572
High/ High         931
Medium/ Low        769
High/ Low          621
Low/ Medium        413
Low/ High          404
Low/ Low            28
Name: Work Rate, dtype: int64
```

In [24]:
```python
values = {"Medium/ Medium":1,"High/ Medium":2,"Medium/ High":3,"High/ High":4,
          "Medium/ Low":5,"High/ Low":6,"Low/ Medium":7,"Low/ High":8,"Low/ Low":9}
df["Work Rate"]=df["Work Rate"].replace(values)
df.head()
```

Out[24]:

	Age	Overall	Potential	Value	Wage	Special	Preferred Foot	International Reputation	Weak Foot	Skill Moves	Work Rate	Position	Joined	Contract Valid Until	Height	Weig
0	31	94	94	110.5	0.565	2202	0.0	5.0	4.0	4.0	1	RF	Jul 1, 2004	2021	67.0	159
1	33	94	94	77.0	0.405	2228	1.0	5.0	4.0	5.0	6	ST	Jul 10, 2018	2022	74.0	183
2	26	92	93	118.5	0.290	2143	1.0	5.0	5.0	5.0	2	LW	Aug 3, 2017	2022	69.0	150
4	27	91	92	102.0	0.355	2281	1.0	4.0	5.0	4.0	4	RCM	Aug 30, 2015	2023	71.0	154
5	27	91	91	93.0	0.340	2142	1.0	4.0	4.0	4.0	2	LF	Jul 1, 2012	2020	68.0	163

In [25]:
```python
df.describe()
```

Out[25]:

	Age	Overall	Potential	Value	Wage	Special	Preferred Foot	International Reputation	Weak Foot	
count	14743.000000	14743.000000	14743.000000	14743.000000	14743.000000	14743.000000	14743.000000	14743.000000	14743.000000	1
mean	25.114088	66.381808	71.334871	2.551617	0.009991	1666.474259	0.753985	1.117073	3.001221	
std	4.594359	6.889961	6.099177	5.833752	0.022834	198.177615	0.430702	0.400780	0.635514	
min	16.000000	46.000000	48.000000	0.010000	0.001000	1000.000000	0.000000	1.000000	1.000000	
25%	21.000000	62.000000	67.000000	0.350000	0.001000	1524.000000	1.000000	1.000000	3.000000	
50%	25.000000	66.000000	71.000000	0.725000	0.003000	1669.000000	1.000000	1.000000	3.000000	
75%	28.000000	71.000000	75.000000	2.200000	0.009000	1806.000000	1.000000	1.000000	3.000000	
max	39.000000	94.000000	95.000000	118.500000	0.565000	2346.000000	1.000000	5.000000	5.000000	

Conclusion: Data Cleaning and Preprocessing was done on the given dataset