ML Lab

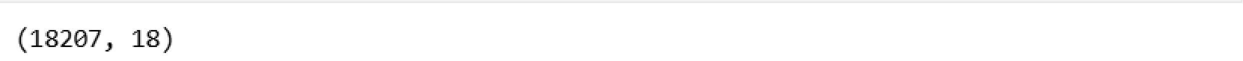
Assignment-2

Q1. Load a CSV dataset and Perform EDA on it

import pandas as pd

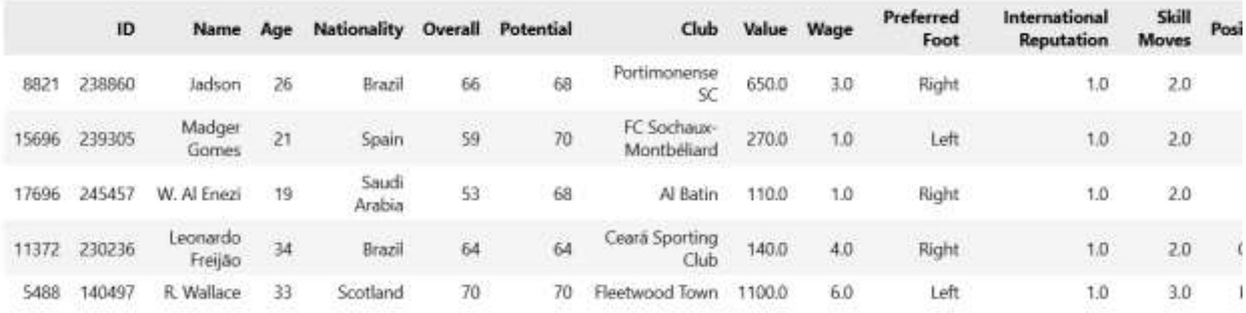
df = pd.read_csv('fifa_eda.csv')

df.shape



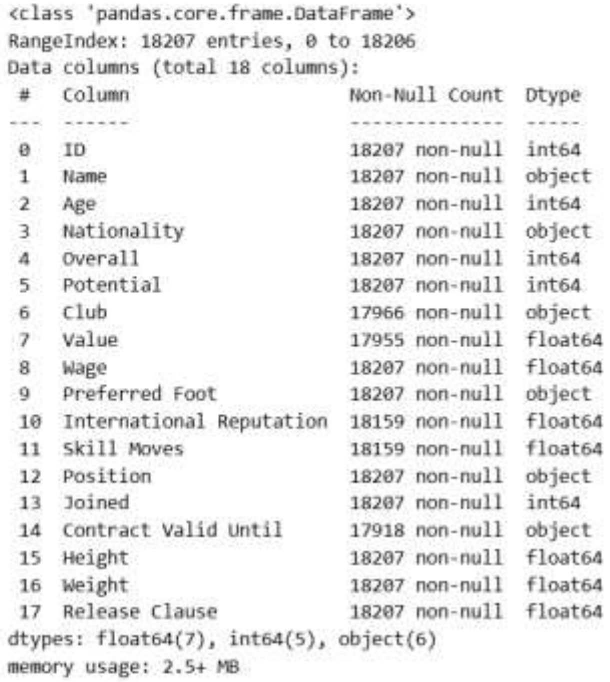
```
(18207, 18)
```

df.sample(5)



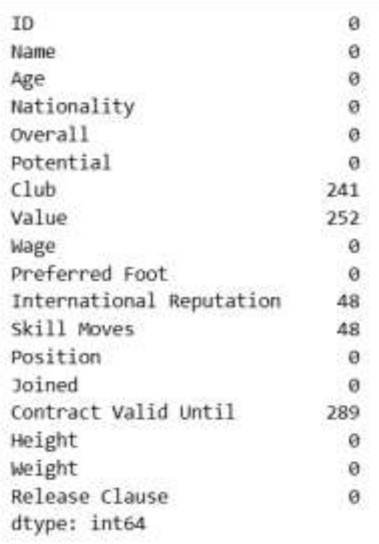
	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	International Reputation	Skill Moves	Posi
8821	238860	Jadson	26	Brazil	66	68	Portimonense SC	650.0	3.0	Right	1.0	2.0	
15696	239305	Madger Gomes	21	Spain	59	70	FC Sochaux-Montbéliard	270.0	1.0	Left	1.0	2.0	
17696	245457	W. Al Enezi	19	Saudi Arabia	53	68	Al Batin	110.0	1.0	Right	1.0	2.0	
11372	230236	Leonardo Freijão	34	Brazil	64	64	Ceará Sporting Club	140.0	4.0	Right	1.0	2.0	
5488	140497	R. Wallace	33	Scotland	70	70	Fleetwood Town	1100.0	6.0	Left	1.0	3.0	

df.info()



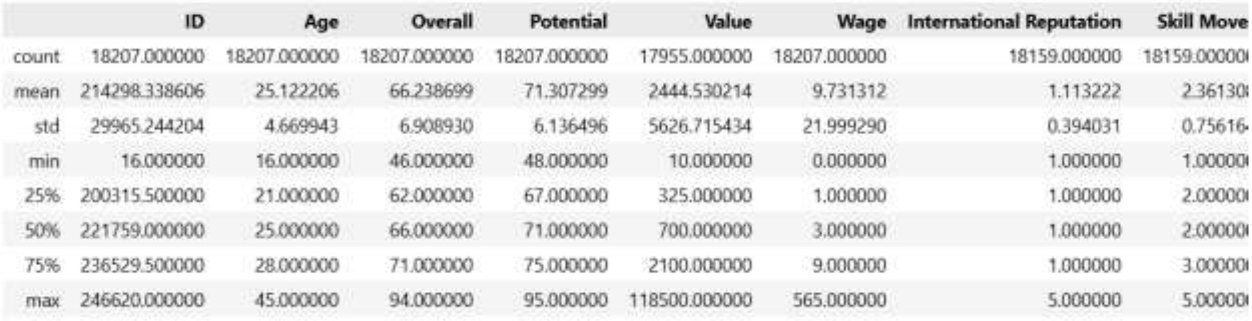
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18207 entries, 0 to 18206
Data columns (total 18 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   ID                        18207 non-null  int64  
 1   Name                      18207 non-null  object 
 2   Age                       18207 non-null  int64  
 3   Nationality               18207 non-null  object 
 4   Overall                   18207 non-null  int64  
 5   Potential                 18207 non-null  int64  
 6   Club                      17966 non-null  object 
 7   Value                     17955 non-null  float64
 8   Wage                      18207 non-null  float64
 9   Preferred Foot            18207 non-null  object 
 10  International Reputation  18159 non-null  float64
 11  Skill Moves               18159 non-null  float64
 12  Position                  18207 non-null  object 
 13  Joined                    18207 non-null  int64  
 14  Contract Valid Until      17918 non-null  object 
 15  Height                    18207 non-null  float64
 16  Weight                    18207 non-null  float64
 17  Release Clause            18207 non-null  float64
dtypes: float64(7), int64(5), object(6)
memory usage: 2.5+ MB
```

df.isnull().sum()



```
ID                          0
Name                        0
Age                         0
Nationality                 0
Overall                     0
Potential                   0
Club                      241
Value                     252
Wage                        0
Preferred Foot              0
International Reputation   48
Skill Moves                48
Position                    0
Joined                      0
Contract Valid Until      289
Height                      0
Weight                      0
Release Clause              0
dtype: int64
```

df.describe()

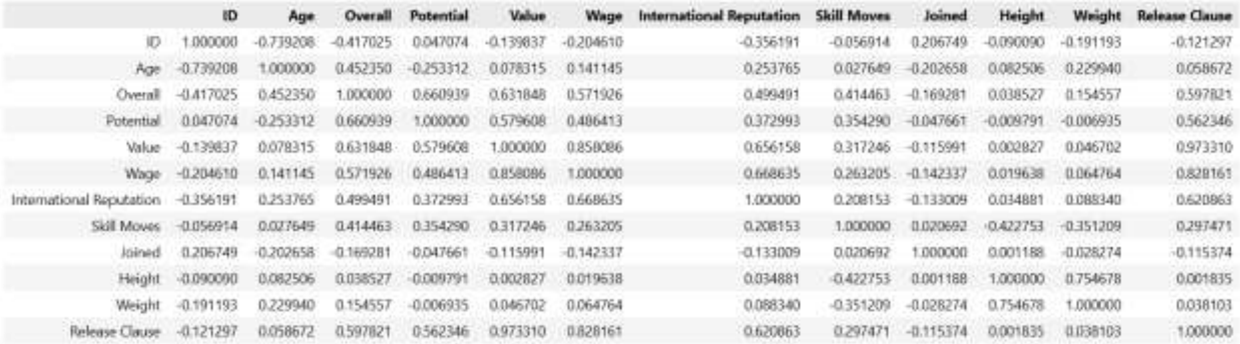


	ID	Age	Overall	Potential	Value	Wage	International Reputation	Skill Move
count	18207.000000	18207.000000	18207.000000	18207.000000	17955.000000	18207.000000	18159.000000	18159.00000
mean	214298.338606	25.122206	66.238699	71.307299	2444.530214	9.731312	1.113222	2.36130
std	29965.244204	4.669943	6.908930	6.136496	5626.715434	21.999290	0.394031	0.75616
min	16.000000	16.000000	46.000000	48.000000	10.000000	0.000000	1.000000	1.00000
25%	200315.500000	21.000000	62.000000	67.000000	325.000000	1.000000	1.000000	2.00000
50%	221759.000000	25.000000	66.000000	71.000000	700.000000	3.000000	1.000000	2.00000
75%	236529.500000	28.000000	71.000000	75.000000	2100.000000	9.000000	1.000000	3.00000
max	246620.000000	45.000000	94.000000	95.000000	118500.000000	565.000000	5.000000	5.00000

df.duplicated().sum()

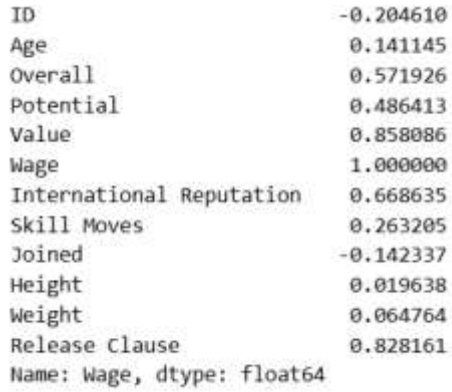
```
np.int64(0)
```

df.drop(columns=["Name", "Nationality", "Club", "Preferred Foot", "Position","Contract Valid Until"], inplace=True) df.corr()



	ID	Age	Overall	Potential	Value	Wage	International Reputation	Skill Moves	Joined	Height	Weight	Release Clause
ID	1.000000	-0.739208	-0.417025	0.047074	-0.139837	-0.204610	-0.356191	-0.056914	0.206749	-0.090090	-0.191193	-0.121297
Age	-0.739208	1.000000	0.452350	-0.253312	0.078315	0.141145	0.253765	0.027649	-0.202658	0.082506	0.229940	0.058672
Overall	-0.417025	0.452350	1.000000	0.660939	0.631848	0.571926	0.499491	0.414463	-0.169281	0.038527	0.154557	0.597821
Potential	0.047074	-0.253312	0.660939	1.000000	0.579608	0.486413	0.372993	0.354290	-0.047661	-0.009791	-0.006935	0.562346
Value	-0.139837	0.078315	0.631848	0.579608	1.000000	0.858086	0.656158	0.317246	-0.115991	0.002827	0.046702	0.973310
Wage	-0.204610	0.141145	0.571926	0.486413	0.858086	1.000000	0.668635	0.263205	-0.142337	0.019638	0.064764	0.828161
International Reputation	-0.356191	0.253765	0.499491	0.372993	0.656158	0.668635	1.000000	0.208153	-0.133009	0.034881	0.088340	0.620863
Skill Moves	-0.056914	0.027649	0.414463	0.354290	0.317246	0.263205	0.208153	1.000000	0.020692	-0.422753	-0.351209	0.297471
Joined	0.206749	-0.202658	-0.169281	-0.047661	-0.115991	-0.142337	-0.133009	0.020692	1.000000	0.001188	-0.028274	-0.115374
Height	-0.090090	0.082506	0.038527	-0.009791	0.002827	0.019638	0.034881	-0.422753	0.001188	1.000000	0.754678	0.001835
Weight	-0.191193	0.229940	0.154557	-0.006935	0.046702	0.064764	0.088340	-0.351209	-0.028274	0.754678	1.000000	0.038103
Release Clause	-0.121297	0.058672	0.597821	0.562346	0.973310	0.828161	0.620863	0.297471	-0.115374	0.001835	0.038103	1.000000

df.corr()["Wage"]



```
ID                         -0.204610
Age                         0.141145
Overall                     0.571926
Potential                   0.486413
Value                       0.858086
Wage                        1.000000
International Reputation    0.668635
Skill Moves                 0.263205
Joined                     -0.142337
Height                      0.019638
Weight                      0.064764
Release Clause              0.828161
Name: Wage, dtype: float64
```

!pip install seaborn

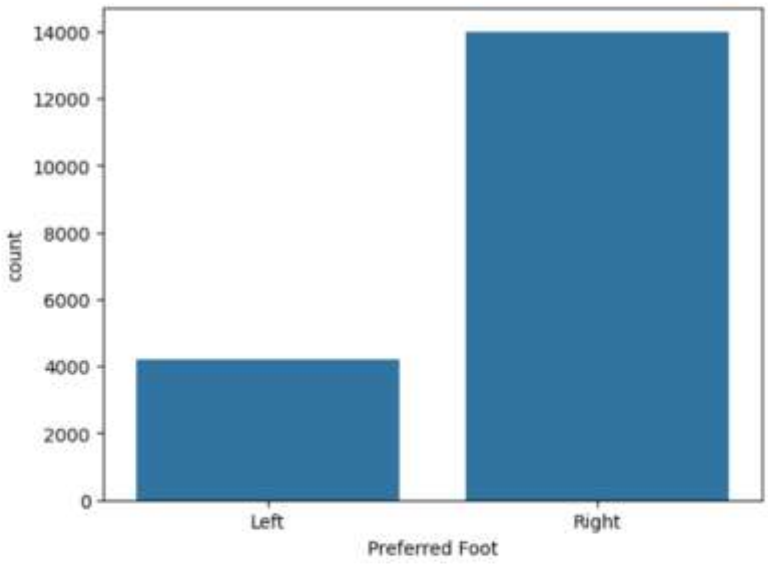
import seaborn as sns

import matplotlib.pyplot as plt

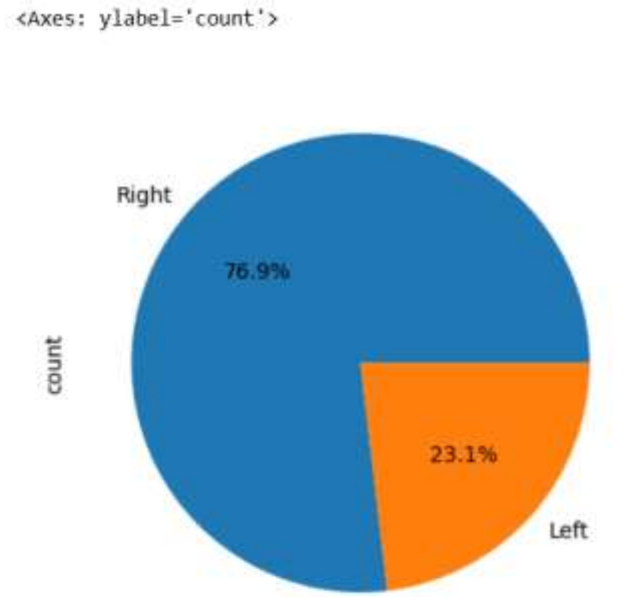
df2=pd.read_csv('fifa_eda.csv')

sns.countplot(x=df2["Preferred Foot"])

```
<Axes: xlabel='Preferred Foot', ylabel='count'>
```

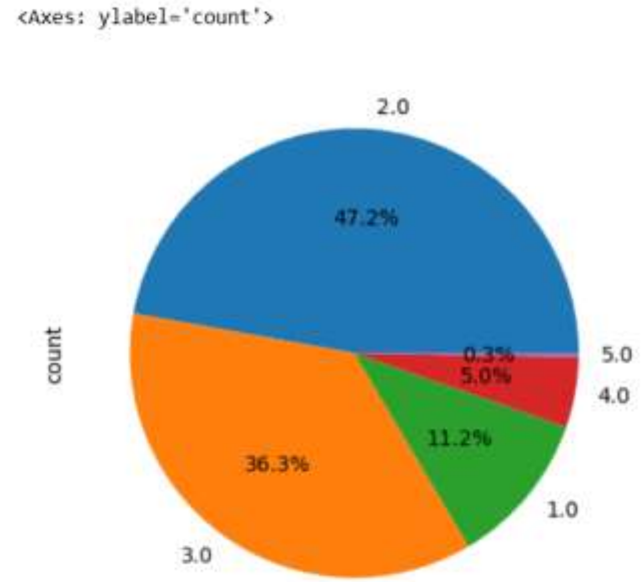


df2['Preferred Foot'].value_counts().plot(kind='pie', autopct='%1.1f%%')



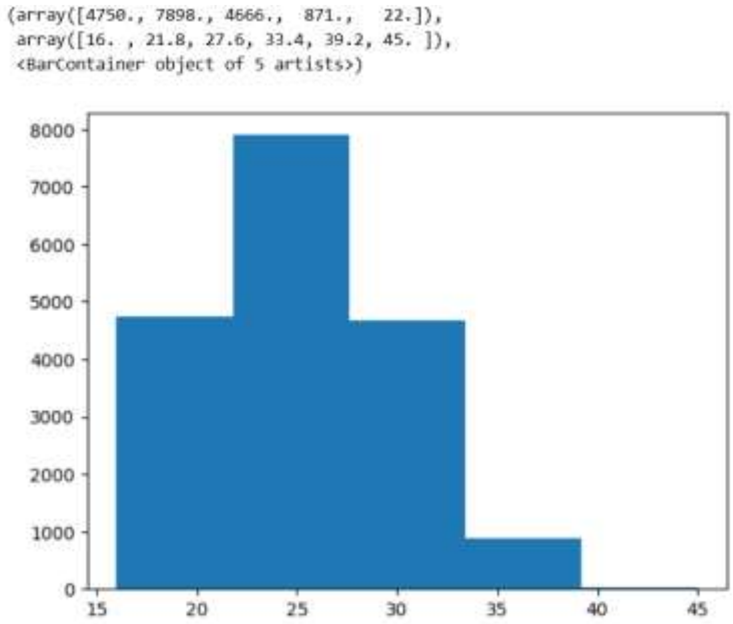
```
<Axes: ylabel='count'>
```

df2['Skill Moves'].value_counts().plot(kind='pie', autopct='%1.1f%%')



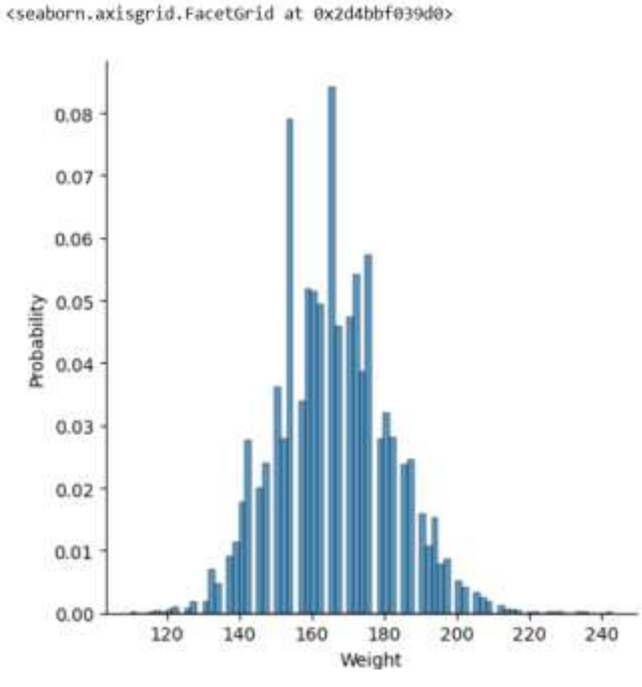
```
<Axes: ylabel='count'>
```

plt.hist(df2['Height'],bins=5)



```
(array([4750., 7898., 4666.,  871.,   22.]),
 array([16. , 21.8, 27.6, 33.4, 39.2, 45. ]),
 <BarContainer object of 5 artists>)
```

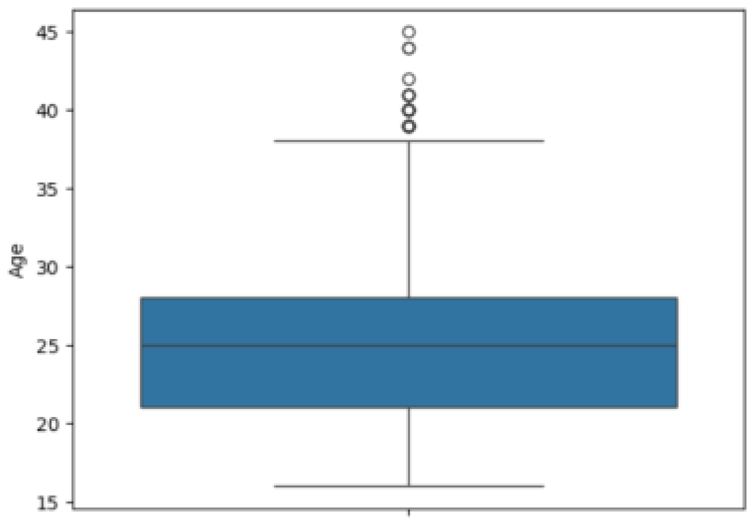
```
sns.displot(data=df2, x="Weight", stat="probability")
```



```
<seaborn.axisgrid.FacetGrid at 0x2d4bbf039d0>
```

```
sns.boxplot(df2['Age'])
```

```
<Axes: ylabel='Age'>
```

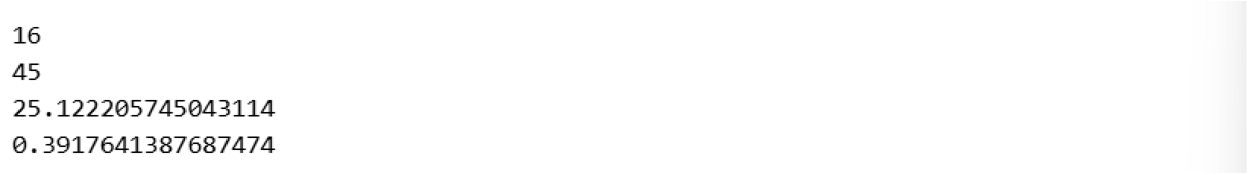


print(df['Age'].min())

print(df['Age'].max())

print(df['Age'].mean())

print(df['Age'].skew())



```
16
45
25.122205745043114
0.3917641387687474
```