EXPLORATORY DATA ANALYSIS

The main purpose of EDA is to detect any errors, outliers as well as to understand different patterns in the data. It allows Analysts to understand the data better before making any assumptions. The outcomes of EDA helps businesses to know their customers, expand their business and take decisions accordingly.

Importing libraries

In [33]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

Loading Dataset

This Dataset of various football leagues and the clubs as well as player market values, squad size gives you an overall information about the club finnacials as well as how wealthy the league is.

In [34]:

```
data=pd.read_csv('club.csv')
```

In [35]:

```
data
```

Out[35]:

	Unnamed: 0	Club Name	Competition Name	Squad Size	Average Age Of Players	Market Value Of Club In Millions(£)	Average Market Value Of Players In Millions(£)	Market Value Of Top 18 Players In Millions(£)
0	0	Manchester City	Premier League	24	27.1	970.02	40.42	920.70
1	1	Paris Saint-Germain	Ligue 1	36	26.1	891.18	24.76	801.00
2	2	Manchester United	Premier League	29	27.9	820.13	28.28	742.50
3	3	Chelsea FC	Premier League	27	26.9	802.35	29.72	737.10
4	4	Liverpool FC	Premier League	27	27.2	779.85	28.88	715.95
...	...	...	...	...	...	...	...	...
95	95	Levante UD	LaLiga	27	28.1	89.19	3.30	82.35
96	96	FC Metz	Ligue 1	29	25.3	89.19	3.08	79.74
97	97	Clube Atlético Mineiro	Série A	29	27.5	88.61	3.06	76.46
98	98	Lokomotiv Moscow	Premier Liga	29	25.0	87.32	3.01	77.85
99	99	Genoa CFC	Serie A	34	27.6	86.94	2.56	72.27

100 rows × 8 columns

To check the first 5 record of dataset

In [36]:

```
data.head(5)
```

Out[36]:

	Unnamed: 0	Club Name	Competition Name	Squad Size	Average Age Of Players	Market Value Of Club In Millions(£)	Average Market Value Of Players In Millions(£)	Market Value Of Top 18 Players In Millions(£)
0	0	Manchester City	Premier League	24	27.1	970.02	40.42	920.70
1	1	Paris Saint-Germain	Ligue 1	36	26.1	891.18	24.76	801.00
2	2	Manchester United	Premier League	29	27.9	820.13	28.28	742.50
3	3	Chelsea FC	Premier League	27	26.9	802.35	29.72	737.10
4	4	Liverpool FC	Premier League	27	27.2	779.85	28.88	715.95

Changing column names

In [59]:

```
o.","Club_Name","Competition_Name","Squad_size","Age","Market_Value_of_club","Market_Value_c
```

In [60]:

```
data.columns=Col
```

In [61]:

```
data.head()
```

Out[61]:

	S.No.	Club_Name	Competition_Name	Squad_size	Age	Market_Value_of_club	Market_Value
0	0	Manchester City	Premier League	24	27.1	970.02	
1	1	Paris Saint-Germain	Ligue 1	36	26.1	891.18	
2	2	Manchester United	Premier League	29	27.9	820.13	
3	3	Chelsea FC	Premier League	27	26.9	802.35	
4	4	Liverpool FC	Premier League	27	27.2	779.85	

I got my column names.

In [62]:

```
data.drop(['S.No.'], axis = 1)
```

Out[62]:

	Club_Name	Competition_Name	Squad_size	Age	Market_Value_of_club	Market_Value_of_pl
0	Manchester City	Premier League	24	27.1	970.02	
1	Paris Saint-Germain	Ligue 1	36	26.1	891.18	
2	Manchester United	Premier League	29	27.9	820.13	
3	Chelsea FC	Premier League	27	26.9	802.35	
4	Liverpool FC	Premier League	27	27.2	779.85	
...	...	...	...	...	...	
95	Levante UD	LaLiga	27	28.1	89.19	
96	FC Metz	Ligue 1	29	25.3	89.19	
97	Clube Atlético Mineiro	Série A	29	27.5	88.61	
98	Lokomotiv Moscow	Premier Liga	29	25.0	87.32	
99	Genoa CFC	Serie A	34	27.6	86.94	

100 rows × 7 columns

Checking for missing values

In [63]:

```
data.isnull().sum()
```

Out[63]:

```
S.No.                     0
Club_Name                 0
Competition_Name          0
Squad_size                0
Age                       0
Market_Value_of_club      0
Market_Value_of_players   0
Market_Value_of_TOP18     0
dtype: int64
```

There is no missing value

checking the data types of each attribute.

In [64]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 8 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   S.No.                    100 non-null    int64  
 1   Club_Name                100 non-null    object 
 2   Competition_Name         100 non-null    object 
 3   Squad_size               100 non-null    int64  
 4   Age                      100 non-null    float64
 5   Market_Value_of_club     100 non-null    float64
 6   Market_Value_of_players  100 non-null    float64
 7   Market_Value_of_TOP18    100 non-null    float64
dtypes: float64(4), int64(2), object(2)
memory usage: 6.4+ KB
```

In [65]:

```
#Checking for wrong entries like symbols -,?,#,*,etc.
for col in data.columns:
    print('{} : {}'.format(col,data[col].unique()))
```

```
S.No. : [ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 2
2 23
 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71
 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95
 96 97 98 99]
Club_Name : ['Manchester City' 'Paris Saint-Germain' 'Manchester United' 'Ch
elsea FC'
 'Liverpool FC' 'Bayern Munich' 'Real Madrid' 'Atlético de Madrid'
 'Tottenham Hotspur' 'FC Barcelona' 'Borussia Dortmund' 'Juventus FC'
 'Arsenal FC' 'Leicester City' 'Inter Milan' 'SSC Napoli' 'RB Leipzig'
 'AC Milan' 'AS Roma' 'Everton FC' 'Sevilla FC' 'Atalanta BC'
 'Aston Villa' 'Bayer 04 Leverkusen' 'Wolverhampton Wanderers'
 'Real Sociedad' 'AS Monaco' 'West Ham United' 'Ajax Amsterdam'
 'Olympique Lyon' 'Villarreal CF' 'SS Lazio' 'VfL Wolfsburg' 'SL Benfica'
 'Borussia Mönchengladbach' 'LOSC Lille' 'FC Porto' 'Leeds United'
 'Valencia CF' 'Brighton & Hove Albion' 'Southampton FC' 'ACF Fiorentina'
 'Olympique Marseille' 'Crystal Palace' 'Newcastle United' 'OGC Nice'
 'Real Betis Balompié' 'US Sassuolo' 'Stade Rennais FC' 'Athletic Bilbao'
 'TSG 1899 Hoffenheim' 'Eintracht Frankfurt' 'Brentford FC' 'Sporting CP'
 'Norwich City' 'Torino FC' 'Shakhtar Donetsk' 'VfB Stuttgart'
 'Zenit St. Petersburg' 'Club Brugge KV' 'Getafe CF' 'Red Bull Salzburg'
 'PSV Eindhoven' 'Clube de Regatas do Flamengo' 'Fulham FC'
 'Cagliari Calcio' 'Bologna FC 1909' 'Watford FC' 'Burnley FC'
 'Sociedade Esportiva Palmeiras' 'Celta de Vigo' 'KRC Genk' 'Dynamo Kyiv'
 'SC Freiburg' 'Fenerbahce SK' 'Besiktas JK' 'Spartak Moscow'
 'Sheffield United' 'Hertha BSC' 'UC Sampdoria' 'SC Braga'
 'AFC Bournemouth' 'Olympiacos Piraeus' 'FK Krasnodar'
 'Feyenoord Rotterdam' 'Club Atlético River Plate' 'GNK Dinamo Zagreb'
 'Galatasaray A.S.' 'Hellas Verona' '1.FSV Mainz 05' 'Rangers FC'
 'Udinese Calcio' 'RCD Espanyol Barcelona' 'AS Saint-Étienne'
 'FC Girondins Bordeaux' 'Levante UD' 'FC Metz' 'Clube Atlético Mineiro'
 'Lokomotiv Moscow' 'Genoa CFC']
Competition_Name : ['Premier League' 'Ligue 1' 'Bundesliga' 'LaLiga' 'Serie
A' 'Eredivisie'
 'Liga Bwin' 'Premier Liga' 'Jupiler Pro League' 'Série A' 'Championship'
 'Süper Lig' 'Super League 1' 'Liga Profesional' '1.HNL' 'Premiership']
Squad_size : [24 36 29 27 26 22 30 34 25 28 33 32 31]
Age : [27.1 26.1 27.9 26.9 27.2 27.3 28.2 25.6 25.9 24.8 27.5 25.4 29.  27.7
 24.1 26.8 25.2 28.  28.3 24.6 26.4 26.2 24.4 28.9 25.7 28.1 27.6 25.3
 27.  26.6 25.  24.  28.4 28.7 27.4 23.7 26.5 25.5 25.8 23.5 24.3 23.
 30.1 27.8 23.3 23.6]
Market_Value_of_club : [970.02 891.18 820.13 802.35 779.85 756.45 680.4  67
1.31 627.3  592.2
 543.51 542.61 507.15 493.29 473.31 467.01 447.8  429.03 386.33 385.88
 374.31 373.82 371.79 348.71 345.69 345.6  331.47 318.83 304.65 301.82
 300.96 280.17 257.42 252.9  244.35 242.28 239.27 232.83 232.02 231.57
 229.68 228.42 225.27 224.51 220.86 202.73 198.63 195.21 182.75 182.07
 182.05 180.41 179.96 178.38 167.81 167.33 165.6  156.78 156.33 150.08
 141.3  138.96 138.69 133.74 133.34 132.46 130.95 130.77 128.88 127.17
 120.87 119.7  118.94 114.17 114.08 112.86 112.05 110.43 108.72 107.24
 104.   102.24 101.93 100.58 100.04  98.82  98.73  98.42  95.58  94.5
  92.7   92.25  90.09  89.55  89.19  88.61  87.32  86.94]
Market_Value_of_players : [40.42 24.76 28.28 29.72 28.88 29.09 25.2  30.51 2
6.14 19.74 15.99 21.7
```

```
 19.51 17.62 16.9  17.96 15.44 14.79 12.88 14.84 14.4  14.38 14.87 12.45
 13.29 11.43 13.28 12.19 10.41 11.58  8.49  8.04  8.72  8.73  9.32  8.86
  9.7   8.59  8.58  8.83  8.79  8.34  8.63  7.51  7.09  7.23  6.53  7.
  5.52  5.64  5.81  7.43  5.99  5.58  5.71  4.75  7.11  5.    5.65  4.79
  5.14  4.18  4.44  4.73  4.37  4.85  5.45  4.16  5.78  4.83  3.99  4.57
  3.57  3.56  4.7   4.    3.81  4.03  3.83  3.2   4.25  3.59  3.45  3.29
  3.18  3.79  3.82  3.26  3.69  3.34  2.99  3.3   3.08  3.06  3.01  2.56]
Market_Value_of_TOP18 : [920.7  801.   742.5  737.1  715.95 726.21 610.2  65
1.6  558.   535.5
 503.1  506.7  453.6  447.3  455.85 444.6  408.15 389.7  349.02 360.9
 341.1  342.9  354.42 320.4  332.37 324.45 291.6  308.25 282.6  286.29
 257.67 232.56 215.1  231.66 231.75 217.8  228.15 216.   213.12 214.2
 214.02 210.6  209.7  204.3  188.82 181.8  178.83 175.95 171.   157.32
 162.68 168.75 170.1  154.35 150.3  149.85 140.4  154.17 136.71 135.45
 123.84 128.7  117.   118.8  122.22 121.32 115.2  124.65 113.36 123.3
 112.23 106.65 112.05 102.78 103.23 108.9  102.51  99.63  96.66  97.2
  99.81  87.93  98.19  96.21  93.17  89.46  86.13  88.74  89.82  88.83
  86.31  83.97  81.18  78.75  82.35  79.74  76.46  77.85  72.27]
```

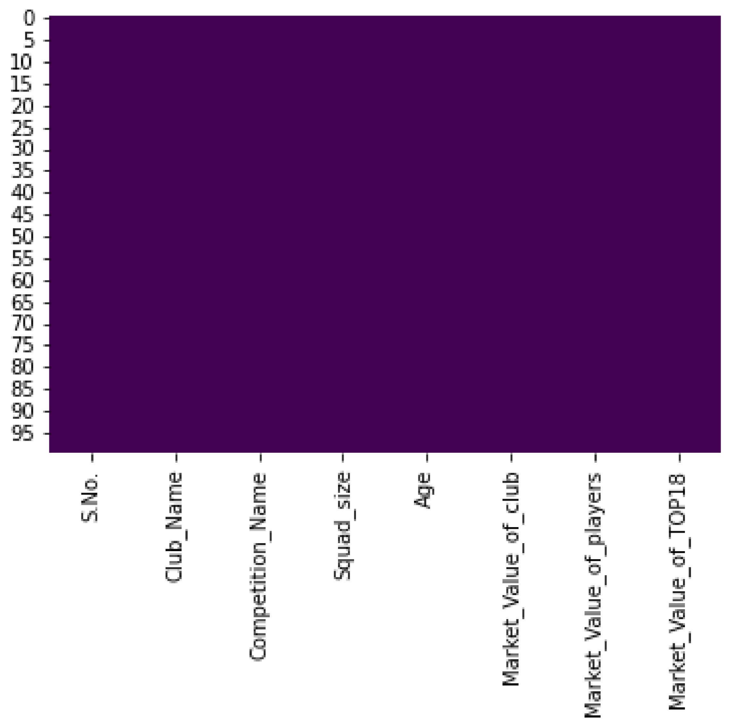
Visualizing missing value

In [66]:

```
sns.heatmap(data.isnull(),cbar=False,cmap='viridis')
```

Out[66]:

```
<AxesSubplot:>
```



As there is no missing value hence we got blank graph.

Asking Analytical Questions and Visualizations

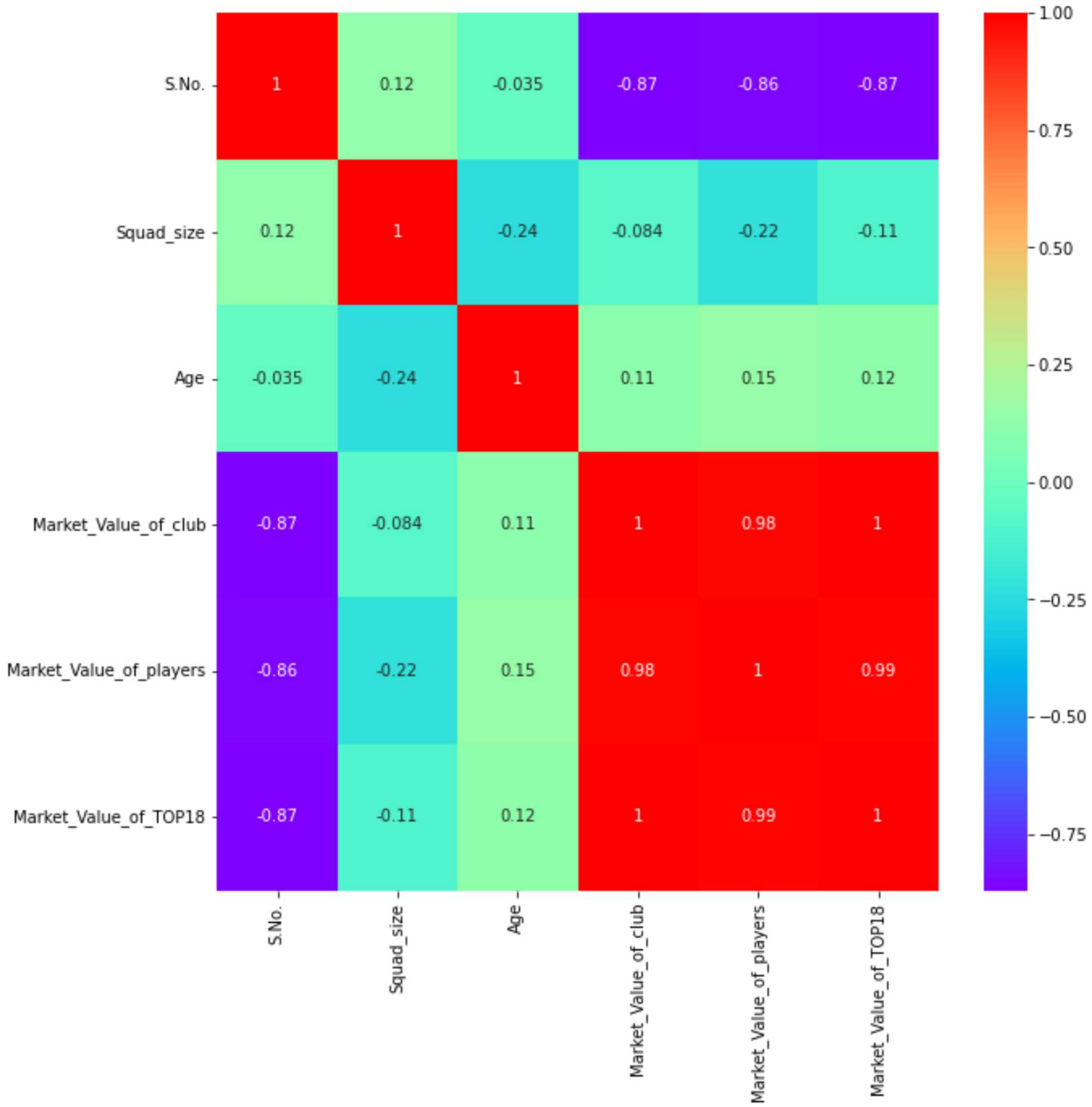
How market value of player affect the squad size?

In [67]:

```
plt.figure(figsize=(10,10))
sns.heatmap(data.corr(),cbar=True,annot=True,cmap='rainbow')
```

Out[67]:

```
<AxesSubplot:>
```



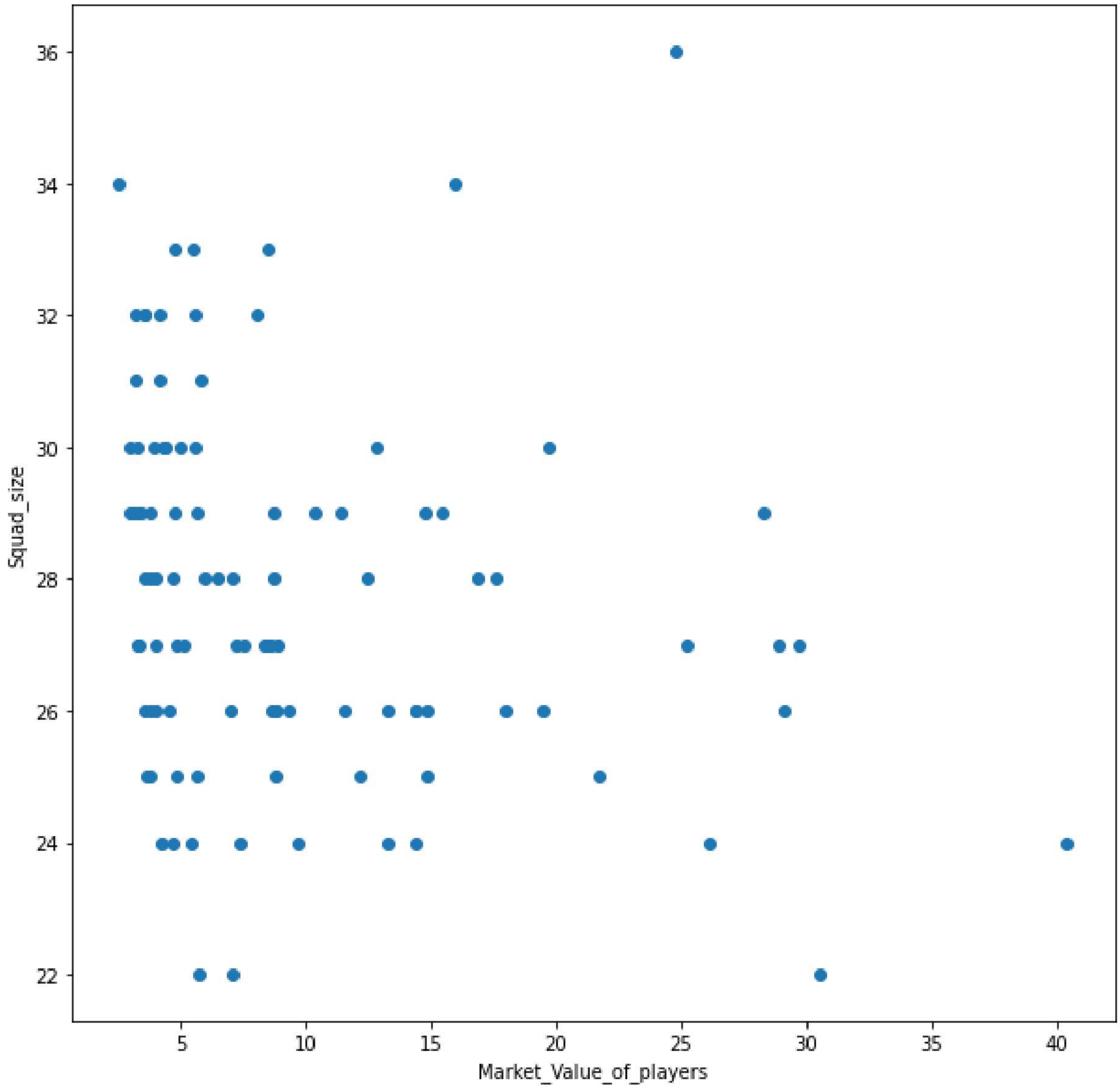
How does the Market Value of players affect the Squad Size?

In [73]:

```
plt.figure(figsize=(10,10))
plt.scatter(x='Market_Value_of_players',y='Squad_size',data=data)
plt.xlabel('Market_Value_of_players')
plt.ylabel('Squad_size')
```

Out[73]:

```
Text(0, 0.5, 'Squad_size')
```



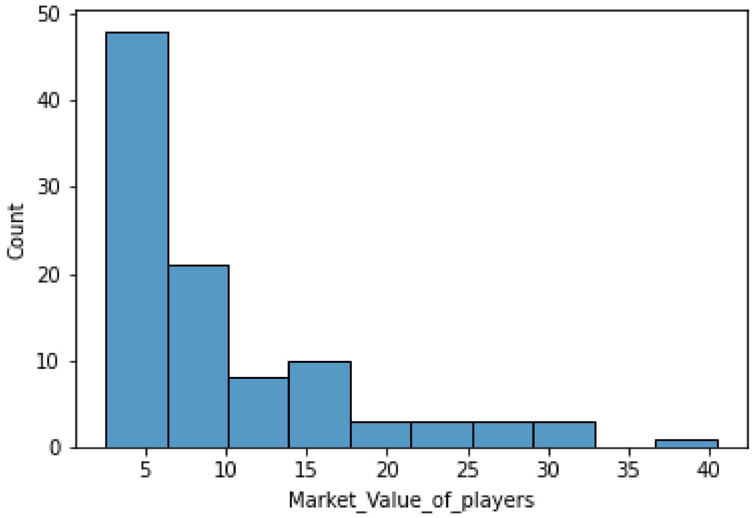
It is clear that lower the Market Value of player bigger the Squad size.

In [74]:

```
sns.histplot(data.Market_Value_of_players,bins=10)
```

Out[74]:

```
<AxesSubplot:xlabel='Market_Value_of_players', ylabel='Count'>
```



The average count between 5-10 is 50 and it is positively skewed.

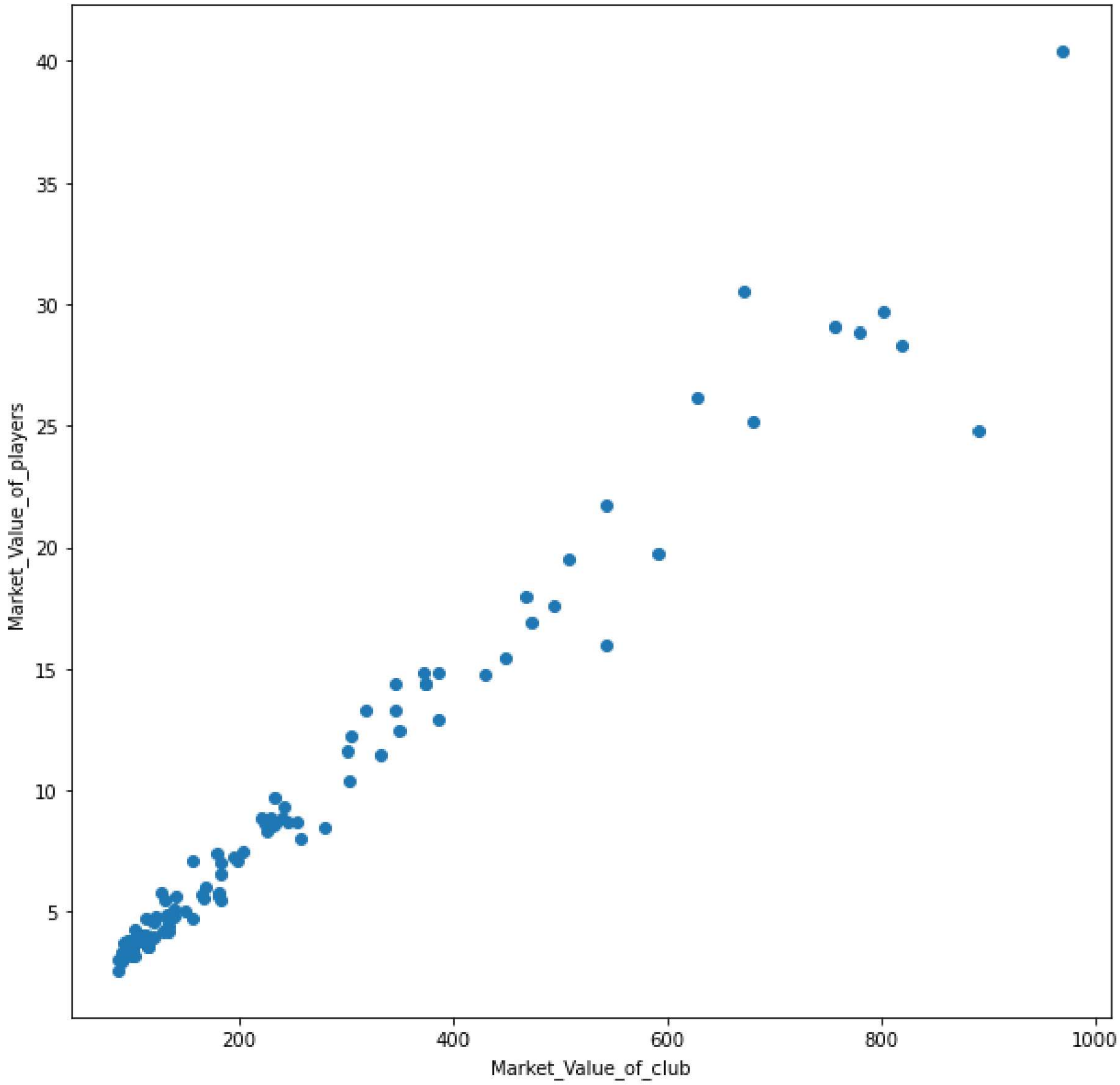
What is the relation between Market_Value_of_club and Market_Value_of_players?

In [75]:

```
plt.figure(figsize=(10,10))
plt.scatter(x='Market_Value_of_club',y='Market_Value_of_players',data=data)
plt.xlabel('Market_Value_of_club')
plt.ylabel('Market_Value_of_players')
```

Out[75]:

```
Text(0, 0.5, 'Market_Value_of_players')
```



More the market value of player, more the market value of club.

Scatter plot showing Age vs market value of player.

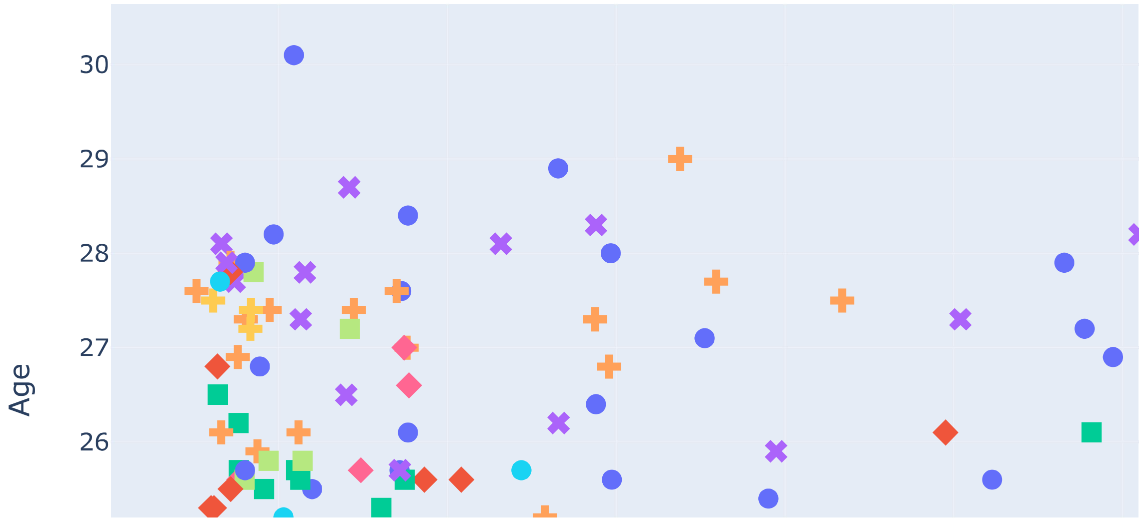
In [87]:

```
!pip install plotly
import plotly.express as px
```

```
Requirement already satisfied: plotly in c:\users\gourj\anaconda3\lib\site-p
ackages (5.5.0)
Requirement already satisfied: six in c:\users\gourj\anaconda3\lib\site-pack
ages (from plotly) (1.16.0)
Requirement already satisfied: tenacity>=6.2.0 in c:\users\gourj\anaconda3\l
ib\site-packages (from plotly) (8.0.1)
```

In [88]:

```
fig = px.scatter(data, y="Age", x="Market_Value_of_players", color="Competition_Name", symb
fig.update_traces(marker_size=10)
fig.show()
```



3D Squad Size vs Market Value vs Average Age

The above graph has ploted squad in the x axis, market value of club in y and average player age in z axis. This data can be interpreted such that :-

1. more the club value, lesser the squad size and lesser the average age of players : Club may have high valued youngsters and have a team that can player for years to come
2. more the club value, lesser the squad size and more the average age of players : Club may high valued veterans or mix of old and fresh talent and have a team that has highly skilled individual with equal youth and experience
3. less the club value, more the squad size : Club has a lot of players, but none of them are of great value
4. more the club value, more the squad size : Club may have a lot of players which may include some high valued talents or not

In [89]:

```
fig = px.scatter_3d(data, x='Squad_size', y='Market_Value_of_club', z='Age', color='Club_Na
fig.show()
```