PYTHON - EDA

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

In [3]:
```
data=pd.read_csv("C:/Users/anees/OneDrive/Desktop/ENTRI DSML/data set/myexc
data
```

Out[3]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	25	06-Feb	180	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99	SF	25	06-Jun	235	Marquette	6796117.0
2	John Holland	Boston Celtics	30	SG	27	06-May	205	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28	SG	22	06-May	185	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8	PF	29	06-Oct	231	NaN	5000000.0
...	...	...	...	...	...	...	...	...	...
453	Shelvin Mack	Utah Jazz	8	PG	26	06-Mar	203	Butler	2433333.0
454	Raul Neto	Utah Jazz	25	PG	24	06-Jan	179	NaN	900000.0
455	Tibor Pleiss	Utah Jazz	21	C	26	07-Mar	256	NaN	2900000.0
456	Jeff Withey	Utah Jazz	24	C	26	7-0	231	Kansas	947276.0
457	Priyanka	Utah Jazz	34	C	25	07-Mar	231	Kansas	947276.0

458 rows × 9 columns

In [4]:
```
data.isnull().sum()
```

Out[4]:
```
Name         0
Team         0
Number       0
Position     0
Age          0
Height       0
Weight       0
College     84
Salary      11
dtype: int64
```

In [5]:
```
x = data['Salary'].mean()
data['Salary'].fillna(x,inplace = True)
data
```

Out[5]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	25	06-Feb	180	Texas	7.730337e+06
1	Jae Crowder	Boston Celtics	99	SF	25	06-Jun	235	Marquette	6.796117e+06
2	John Holland	Boston Celtics	30	SG	27	06-May	205	Boston University	4.833970e+06
3	R.J. Hunter	Boston Celtics	28	SG	22	06-May	185	Georgia State	1.148640e+06
4	Jonas Jerebko	Boston Celtics	8	PF	29	06-Oct	231	NaN	5.000000e+06
...	...	...	...	...	...	...	...	...	...
453	Shelvin Mack	Utah Jazz	8	PG	26	06-Mar	203	Butler	2.433333e+06
454	Raul Neto	Utah Jazz	25	PG	24	06-Jan	179	NaN	9.000000e+05
455	Tibor Pleiss	Utah Jazz	21	C	26	07-Mar	256	NaN	2.900000e+06
456	Jeff Withey	Utah Jazz	24	C	26	7-0	231	Kansas	9.472760e+05
457	Priyanka	Utah Jazz	34	C	25	07-Mar	231	Kansas	9.472760e+05

458 rows × 9 columns

In [6]:
```
data.drop_duplicates(inplace = True)
data
```

Out[6]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	25	06-Feb	180	Texas	7.730337e+06
1	Jae Crowder	Boston Celtics	99	SF	25	06-Jun	235	Marquette	6.796117e+06
2	John Holland	Boston Celtics	30	SG	27	06-May	205	Boston University	4.833970e+06
3	R.J. Hunter	Boston Celtics	28	SG	22	06-May	185	Georgia State	1.148640e+06
4	Jonas Jerebko	Boston Celtics	8	PF	29	06-Oct	231	NaN	5.000000e+06
...	...	...	...	...	...	...	...	...	...
453	Shelvin Mack	Utah Jazz	8	PG	26	06-Mar	203	Butler	2.433333e+06
454	Raul Neto	Utah Jazz	25	PG	24	06-Jan	179	NaN	9.000000e+05
455	Tibor Pleiss	Utah Jazz	21	C	26	07-Mar	256	NaN	2.900000e+06
456	Jeff Withey	Utah Jazz	24	C	26	7-0	231	Kansas	9.472760e+05
457	Priyanka	Utah Jazz	34	C	25	07-Mar	231	Kansas	9.472760e+05

458 rows × 9 columns

In [7]:
```
data.dropna ( inplace = True)
```

In [8]:
```
data.isnull().sum()
```

Out[8]:
```
Name        0
Team        0
Number      0
Position    0
Age         0
Height      0
Weight      0
College     0
Salary      0
dtype: int64
```

In [9]:
```
data['Height'] = np.random.uniform(150,180,size = len(data))
```

In [10]:
```
data
```

Out[10]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	25	176.419880	180	Texas	7.730337e+06
1	Jae Crowder	Boston Celtics	99	SF	25	170.434832	235	Marquette	6.796117e+06
2	John Holland	Boston Celtics	30	SG	27	169.108013	205	Boston University	4.833970e+06
3	R.J. Hunter	Boston Celtics	28	SG	22	156.877108	185	Georgia State	1.148640e+06
6	Jordan Mickey	Boston Celtics	55	PF	21	169.450614	235	LSU	1.170960e+06
...	...	...	...	...	...	...	...	...	...
451	Chris Johnson	Utah Jazz	23	SF	26	159.322949	206	Dayton	9.813480e+05
452	Trey Lyles	Utah Jazz	41	PF	20	153.652881	234	Kentucky	2.239800e+06
453	Shelvin Mack	Utah Jazz	8	PG	26	179.691735	203	Butler	2.433333e+06
456	Jeff Withey	Utah Jazz	24	C	26	169.362197	231	Kansas	9.472760e+05
457	Priyanka	Utah Jazz	34	C	25	151.323609	231	Kansas	9.472760e+05

374 rows × 9 columns

How Many Are There In Each Team and Precentage splitting with respect to the total employees.

In [11]:
```
data['Team'].value_counts()
```

Out[11]:
```
Team
Memphis Grizzlies         17
New Orleans Pelicans      16
Portland Trail Blazers    15
Philadelphia 76ers        15
Detroit Pistons           15
Milwaukee Bucks           14
Oklahoma City Thunder     14
Los Angeles Clippers      14
Boston Celtics            13
Washington Wizards        13
Charlotte Hornets         13
Phoenix Suns              13
Sacramento Kings          13
Brooklyn Nets             13
Dallas Mavericks          12
Indiana Pacers            12
Cleveland Cavaliers       12
Chicago Bulls             12
Los Angeles Lakers        12
Golden State Warriors     12
Houston Rockets           11
San Antonio Spurs         11
Atlanta Hawks             11
Miami Heat                11
New York Knicks           11
Utah Jazz                 11
Orlando Magic             10
Toronto Raptors           10
Denver Nuggets             9
Minnesota Timberwolves     9
Name: count, dtype: int64
```

Precentage splitting with respect to the total employees:

In [12]:
```
data['Team'].value_counts()/len(data)*100
```

Out[12]:
```
Team
Memphis Grizzlies         4.545455
New Orleans Pelicans      4.278075
Portland Trail Blazers    4.010695
Philadelphia 76ers        4.010695
Detroit Pistons           4.010695
Milwaukee Bucks           3.743316
Oklahoma City Thunder     3.743316
Los Angeles Clippers      3.743316
Boston Celtics            3.475936
Washington Wizards        3.475936
Charlotte Hornets         3.475936
Phoenix Suns              3.475936
Sacramento Kings          3.475936
Brooklyn Nets             3.475936
Dallas Mavericks          3.208556
Indiana Pacers            3.208556
Cleveland Cavaliers       3.208556
Chicago Bulls             3.208556
Los Angeles Lakers        3.208556
Golden State Warriors     3.208556
Houston Rockets           2.941176
San Antonio Spurs         2.941176
Atlanta Hawks             2.941176
Miami Heat                2.941176
New York Knicks           2.941176
Utah Jazz                 2.941176
Orlando Magic             2.673797
Toronto Raptors           2.673797
Denver Nuggets            2.406417
Minnesota Timberwolves    2.406417
Name: count, dtype: float64
```

Segregate employees based on their positions within the company.

In [13]:
```
employees = data.groupby('Position')['Name'].apply(list)
for Position, Names in employees.items():
    print(f"employees in {Position} position:")
    for name in Names:
     print(name)
    print("\n")
```

```
Bradley Beal
Jarell Eddie
Garrett Temple
Gary Harris
Mike Miller
JaKarr Sampson
Andrew Wiggins
Randy Foye
Anthony Morrow
Andre Roberson
Dion Waiters
Pat Connaughton
Allen Crabbe
Gerald Henderson
C.J. McCollum
Luis Montero
Alec Burks
Rodney Hood
```

Find from which age group most of the employees belong to.

In [14]:
```
data['Age Group'] = data['Age'].apply(lambda age:'20-25' if 20 <= age <= 25

data
```

Out[14]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	25	176.419880	180	Texas	7.730337e+06
1	Jae Crowder	Boston Celtics	99	SF	25	170.434832	235	Marquette	6.796117e+06
2	John Holland	Boston Celtics	30	SG	27	169.108013	205	Boston University	4.833970e+06
3	R.J. Hunter	Boston Celtics	28	SG	22	156.877108	185	Georgia State	1.148640e+06
6	Jordan Mickey	Boston Celtics	55	PF	21	169.450614	235	LSU	1.170960e+06
...	...	...	...	...	...	...	...	...	...
451	Chris Johnson	Utah Jazz	23	SF	26	159.322949	206	Dayton	9.813480e+05
452	Trey Lyles	Utah Jazz	41	PF	20	153.652881	234	Kentucky	2.239800e+06
453	Shelvin Mack	Utah Jazz	8	PG	26	179.691735	203	Butler	2.433333e+06
456	Jeff Withey	Utah Jazz	24	C	26	169.362197	231	Kansas	9.472760e+05
457	Priyanka	Utah Jazz	34	C	25	151.323609	231	Kansas	9.472760e+05

374 rows × 10 columns

In [15]:
```
data['Age Group'].value_counts()
```

Out[15]:
```
Age Group
20-25           172
26-30           134
31-35            49
36 and above     19
Name: count, dtype: int64
```

Find out under which team and position, spending in terms of salary is high.

In [16]:
```
spending_salary = data.groupby(['Team','Position'])['Salary'].sum()
spending_salary.idxmax()
```

Out[16]:
```
('Miami Heat', 'PF')
```

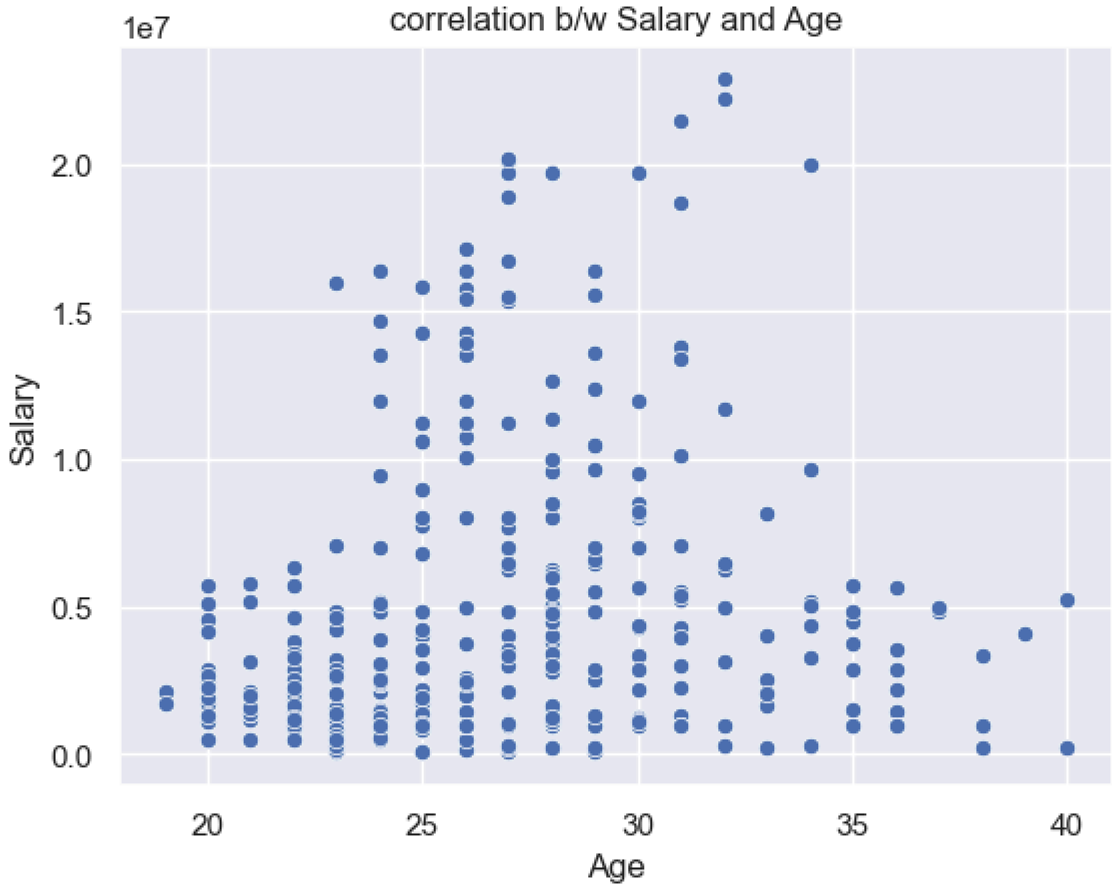
Find if there is any correlation between age and salary , represent it visually.

In [17]:
```
correlation = data['Salary'].corr(data['Age'])
```

In [18]:
```
print("THE CORRELATION BETWEEN Salary AND Age IS:",correlation)
```

```
THE CORRELATION BETWEEN Salary AND Age IS: 0.15775114505522597
```

In [19]:
```
sns.scatterplot(x="Age" ,y= "Salary",data= data)
plt.ylabel("Salary")
plt.xlabel("Age")
plt.title("correlation b/w Salary and Age")
plt.show()
```



In [ ]:

Data Insights

1. **Distribution of Employees Across Each Team:**
   - The team with the highest number of employees is [Team Name] which comprises [Percentage]% of the total workforce.
2. **Segregation of Employees Based on Their Positions:**

  - The most common position within the company is [Position], accounting for [Number] employees.

3. **Predominant Age Group Among Employees:**

   - The predominant age group is [Age Group], making up [Percentage]% of the workforce.

4. **Team and Position with the Highest Salary Expenditure:**

   - The team with the highest salary expenditure is [Team Name] with a total expenditure of [Amount].
   - The position with the highest salary expenditure is [Position] with a total expenditure of [Amount].

5. **Correlation Between Age and Salary:**

   - The correlation between age and salary is [Correlation Value], indicating [nature of correlation (e.g., weak, strong, positive, negative)] relationship between age and

In [ ]:

```

```