Module End Project

Providing a dataset of employees working in ABC company. It consists of 458 rows and 9 columns. The company needs the detailed report and explanation of their employees in each team, also need to identify the following:

1.How many are there in each Team and the percentage splitting with respect to the total employees.

2.Segregate the employees w.r.t different positions.

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

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

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

Before doing the above questions,perform pre-processing of the dataset. Also, the column height is having incorrect data, changing the data of that particular column with any random numbers between 150 and 180.

Detailed report of employees of ABC company.

Performing EDA- Exploratory Data Analysis on employees dataset.

1.Import libraries

In [63]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

2.Load the dataset

In [64]:
```
df=pd.read_csv('myexcel - myexcel.csv.csv')
print(df)
```

```
                Name            Team  Number Position  Age  Height  Weight  \
0     Avery Bradley  Boston Celtics       0       PG   25  06-Feb     180
1       Jae Crowder  Boston Celtics      99       SF   25  06-Jun     235
2      John Holland  Boston Celtics      30       SG   27  06-May     205
3       R.J. Hunter  Boston Celtics      28       SG   22  06-May     185
4     Jonas Jerebko  Boston Celtics       8       PF   29  06-Oct     231
..              ...             ...     ...      ...  ...     ...     ...
453   Shelvin Mack       Utah Jazz       8       PG   26  06-Mar     203
454      Raul Neto       Utah Jazz      25       PG   24  06-Jan     179
455   Tibor Pleiss       Utah Jazz      21        C   26  07-Mar     256
456    Jeff Withey       Utah Jazz      24        C   26     7-0     231
457       Priyanka       Utah Jazz      34        C   25  07-Mar     231

                College     Salary
0                 Texas  7730337.0
1             Marquette  6796117.0
2     Boston University        NaN
3         Georgia State  1148640.0
4                   NaN  5000000.0
..                  ...        ...
453             Butler  2433333.0
454                NaN   900000.0
455                NaN  2900000.0
456             Kansas   947276.0
457             Kansas   947276.0

[458 rows x 9 columns]
```

3.Understand the dataset

In [7]:
```
df.columns
```

Out[7]:
```
Index(['Name', 'Team', 'Number', 'Position', 'Age', 'Height', 'Weight',
       'College', 'Salary'],
      dtype='object')
```

In [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 458 entries, 0 to 457
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Name      458 non-null    object 
 1   Team      458 non-null    object 
 2   Number    458 non-null    int64  
 3   Position  458 non-null    object 
 4   Age       458 non-null    int64  
 5   Height    458 non-null    object 
 6   Weight    458 non-null    int64  
 7   College   374 non-null    object 
 8   Salary    447 non-null    float64
dtypes: float64(1), int64(3), object(5)
memory usage: 32.3+ KB
```

In [9]:
```
df.head()
```

Out[9]:

	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

In [10]:
```
df.tail()
```

Out[10]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
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

In [11]:
```
df.describe()
```

Out[11]:

	Number	Age	Weight	Salary
count	458.000000	458.000000	458.000000	4.470000e+02
mean	17.713974	26.934498	221.543668	4.833970e+06
std	15.966837	4.400128	26.343200	5.226620e+06
min	0.000000	19.000000	161.000000	3.088800e+04
25%	5.000000	24.000000	200.000000	1.025210e+06
50%	13.000000	26.000000	220.000000	2.836186e+06
75%	25.000000	30.000000	240.000000	6.500000e+06
max	99.000000	40.000000	307.000000	2.500000e+07

In [13]:
```
df.describe(include='all')
```

Out[13]:

	Name	Team	Number	Position	Age	Height	Weight	Colleg
count	458	458	458.000000	458	458.000000	458	458.000000	37
unique	458	30	NaN	5	NaN	18	NaN	11
top	Avery Bradley	New Orleans Pelicans	NaN	SG	NaN	06-Sep	NaN	Kentuck
freq	1	19	NaN	102	NaN	59	NaN	2
mean	NaN	NaN	17.713974	NaN	26.934498	NaN	221.543668	NaN
std	NaN	NaN	15.966837	NaN	4.400128	NaN	26.343200	NaN
min	NaN	NaN	0.000000	NaN	19.000000	NaN	161.000000	NaN
25%	NaN	NaN	5.000000	NaN	24.000000	NaN	200.000000	NaN
50%	NaN	NaN	13.000000	NaN	26.000000	NaN	220.000000	NaN
75%	NaN	NaN	25.000000	NaN	30.000000	NaN	240.000000	NaN
max	NaN	NaN	99.000000	NaN	40.000000	NaN	307.000000	NaN

4.Handling missing values

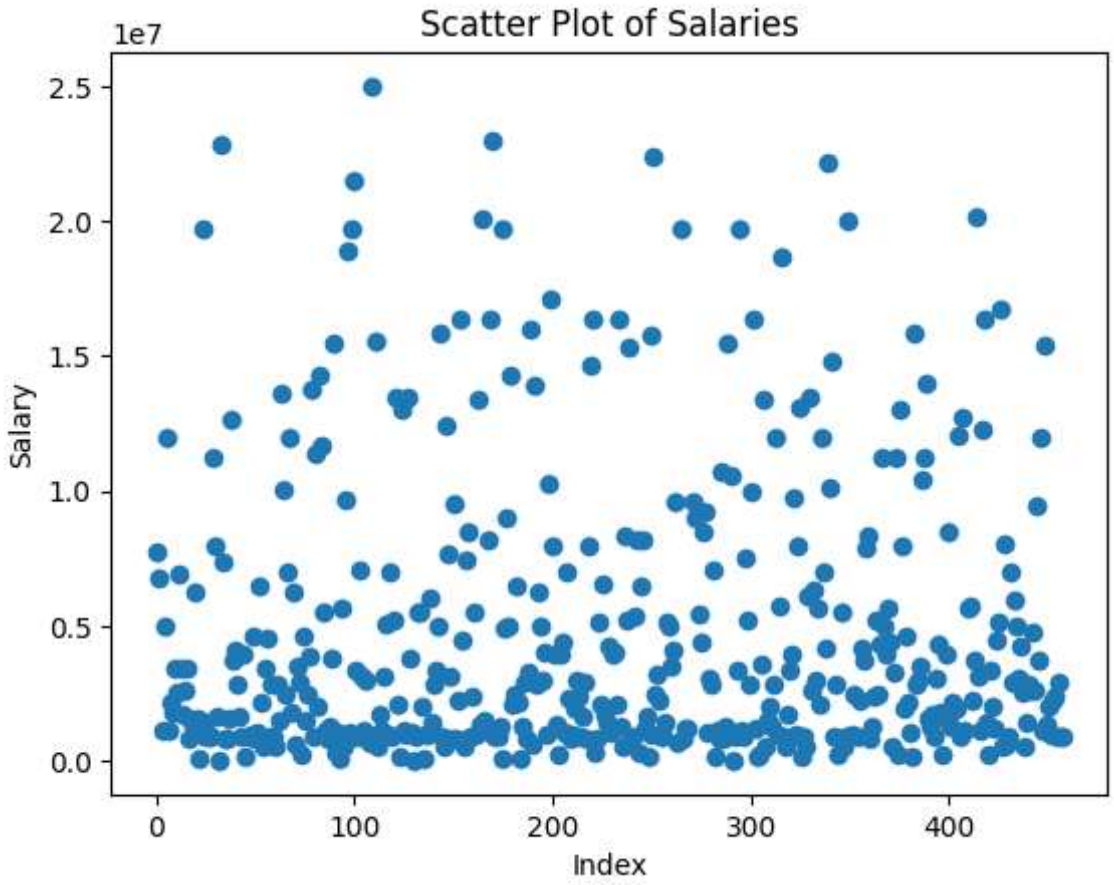
In [65]:
```
df.isnull().sum()
```

Out[65]:
```
Name          0
Team          0
Number        0
Position      0
Age           0
Height        0
Weight        0
College      84
Salary       11
dtype: int64
```

Data story : Columns College has 84 and Salary has 11 null values

Salary distribution of different players

In [81]:
```
plt.scatter(range(len(df['Salary'])),df['Salary'])
plt.xlabel('Index')
plt.ylabel('Salary')
plt.title('Scatter Plot of Salaries')
plt.show()
```



In [18]:
```
# filling the null values in Salary column with 0.
df['Salary'].fillna(0,inplace=True)
```

In [20]:
```
# filling the null values in 'College' column with 'Unknown'
df['College'].fillna('Unknown',inplace=True)
```

5. Handling duplicate values.

In [21]:
```
duplicate=df[df.duplicated()]
print(duplicate)
```

```
Empty DataFrame
Columns: [Name, Team, Number, Position, Age, Height, Weight, College, Salary]
Index: []
```

Data Story: No duplicate rows in this dataset.

6.Filling column 'Height' with random values between 150-180.

In [66]:
```
df['Height']=np.random.uniform(150,180,len(df))
df['Height']
```

Out[66]:
```
0      178.490587
1      179.247674
2      167.286347
3      166.998973
4      162.660089
          ...
453    154.156904
454    152.899924
455    162.200439
456    152.957354
457    170.784500
Name: Height, Length: 458, dtype: float64
```

Table of Contents

1.How many are there in each Team and the percentage splitting with respect to the total employees.

In [44]:
```
df['Team'].nunique()
```

Out[44]:
```
30
```

In [4]:
```
# to find how many are there in each Team
Team_counts=df['Team'].value_counts()
print(Team_counts)
```

```
Team
New Orleans Pelicans      19
Memphis Grizzlies         18
Utah Jazz                 16
New York Knicks           16
Milwaukee Bucks           16
Brooklyn Nets             15
Portland Trail Blazers    15
Oklahoma City Thunder     15
Denver Nuggets            15
Washington Wizards        15
Miami Heat                15
Charlotte Hornets         15
Atlanta Hawks             15
San Antonio Spurs         15
Houston Rockets           15
Boston Celtics            15
Indiana Pacers            15
Detroit Pistons           15
Cleveland Cavaliers       15
Chicago Bulls             15
Sacramento Kings          15
Phoenix Suns              15
Los Angeles Lakers        15
Los Angeles Clippers      15
Golden State Warriors     15
Toronto Raptors           15
Philadelphia 76ers        15
Dallas Mavericks          15
Orlando Magic             14
Minnesota Timberwolves    14
Name: count, dtype: int64
```

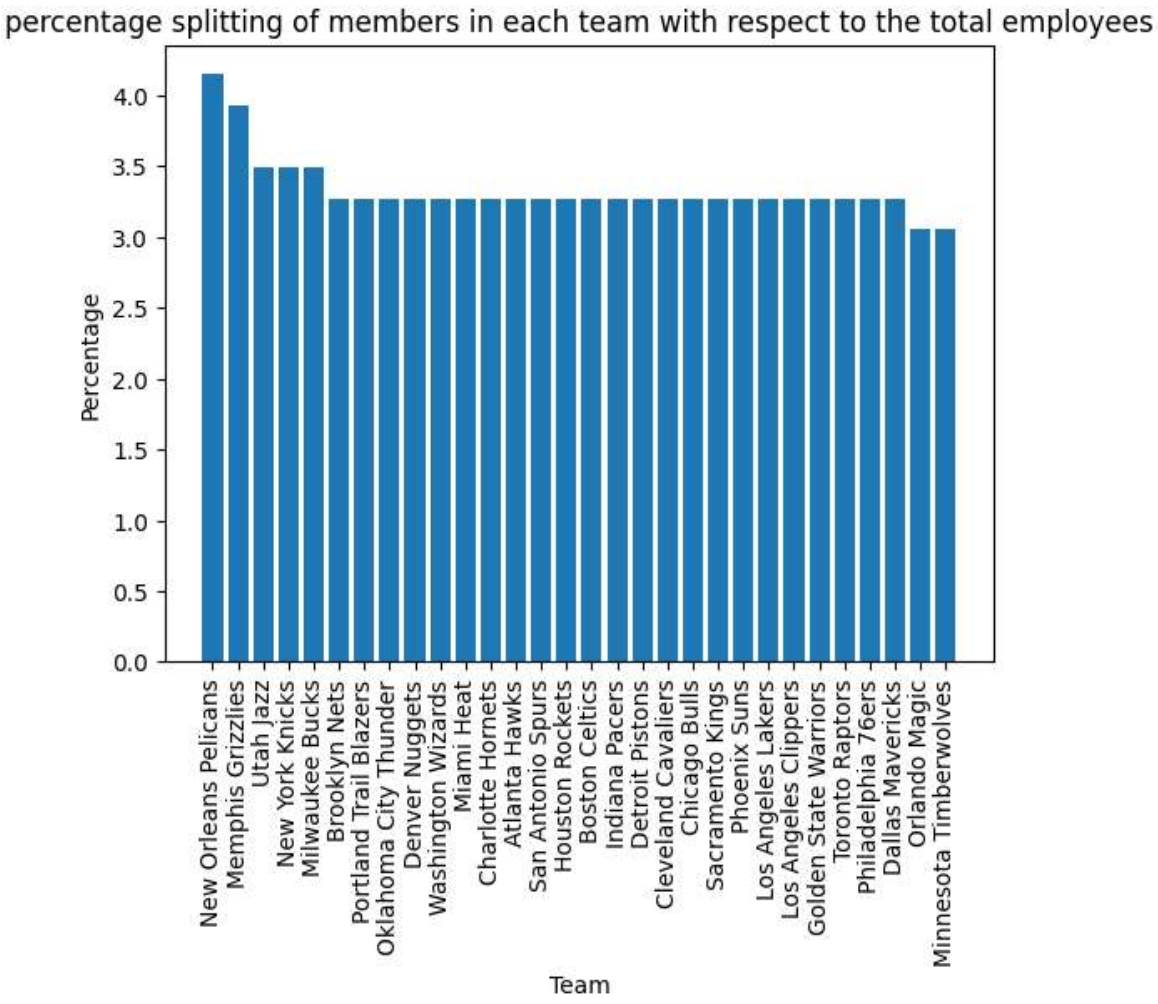
In [25]:
```
# total employees
df['Name'].nunique()
```

Out[25]: 458

In [5]:
```
#percentage splitting with respect to the total employees
percentage=Team_counts*(100/458)
print(percentage)
```

```
Team
New Orleans Pelicans      4.148472
Memphis Grizzlies         3.930131
Utah Jazz                 3.493450
New York Knicks           3.493450
Milwaukee Bucks           3.493450
Brooklyn Nets             3.275109
Portland Trail Blazers    3.275109
Oklahoma City Thunder     3.275109
Denver Nuggets            3.275109
Washington Wizards        3.275109
Miami Heat                3.275109
Charlotte Hornets         3.275109
Atlanta Hawks             3.275109
San Antonio Spurs         3.275109
Houston Rockets           3.275109
Boston Celtics            3.275109
Indiana Pacers            3.275109
Detroit Pistons           3.275109
Cleveland Cavaliers       3.275109
Chicago Bulls             3.275109
Sacramento Kings          3.275109
Phoenix Suns              3.275109
Los Angeles Lakers        3.275109
Los Angeles Clippers      3.275109
Golden State Warriors     3.275109
Toronto Raptors           3.275109
Philadelphia 76ers        3.275109
Dallas Mavericks          3.275109
Orlando Magic             3.056769
Minnesota Timberwolves    3.056769
Name: count, dtype: float64
```

In [6]:
```
# Visualising the  percentage splitting of members in each team with respect to
x=percentage.index
y=percentage.values
plt.bar(x,y)
plt.xticks(rotation='vertical')
plt.xlabel('Team')
plt.ylabel('Percentage')
plt.title('percentage splitting of members in each team with respect to the tota
plt.show()
```



Data Story: Team 'New Orleans Pelicans' have highest number of players.Most of the team have a percentage share of players between 3 and 3.5

2.Segregate the employees w.r.t different positions.

In [60]:
```
df['Position'].unique()
```

Out[60]:
```
array(['PG', 'SF', 'SG', 'PF', 'C'], dtype=object)
```

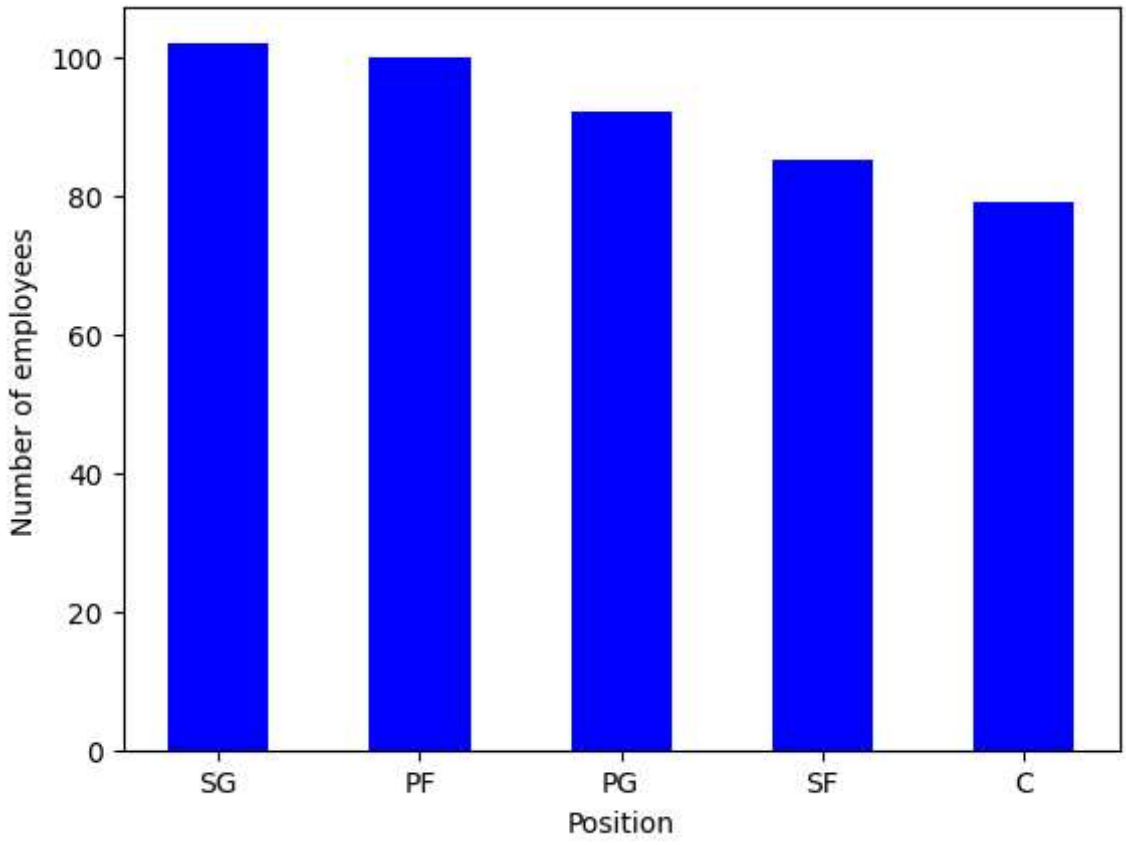
In [27]:
```
position=df['Position'].value_counts()
print(position)
```

```
Position
SG    102
PF    100
PG     92
SF     85
C      79
Name: count, dtype: int64
```

In [47]:
```
grouped= df.groupby('Position')
grouped_to_string=df.groupby('Position')['Name'].apply(list).reset_index()
print(grouped_to_string)
```

	Position	Name
0	C	[Kelly Olynyk, Jared Sullinger, Tyler Zeller, ...
1	PF	[Jonas Jerebko, Amir Johnson, Jordan Mickey, C...
2	PG	[Avery Bradley, Terry Rozier, Marcus Smart, Is...
3	SF	[Jae Crowder, Thanasis Antetokounmpo, Carmelo ...
4	SG	[John Holland, R.J. Hunter, Evan Turner, James...

In [41]:
```
plt.bar(position.index,position.values,color='b',width=0.5)
plt.xlabel('Position')
plt.ylabel('Number of employees')
plt.show()
```



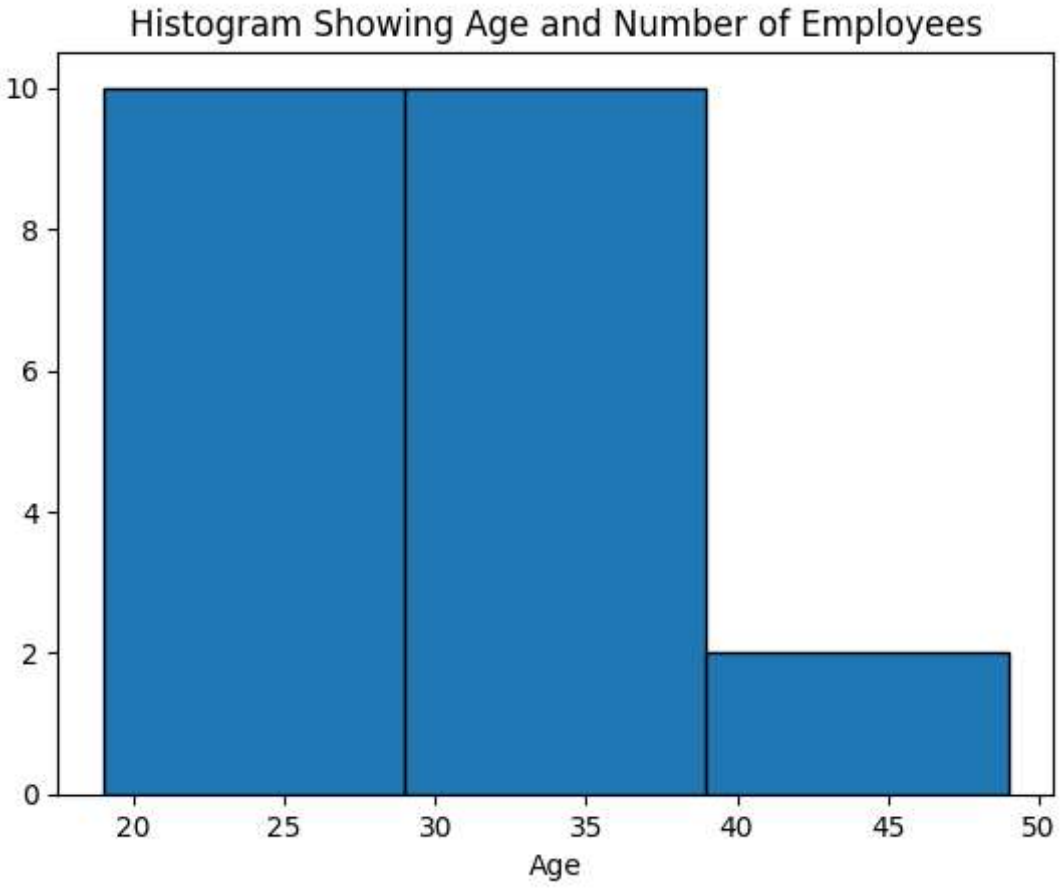
Data Story: Position 'SG' has most number of players

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

In [30]:
```
data=df['Age'].value_counts()
data
```

Out[30]:
```
Age
24    47
25    46
27    41
23    41
26    36
28    31
30    31
29    28
22    26
31    22
20    19
21    19
33    14
32    13
34    10
36    10
35     9
37     4
38     4
40     3
39     2
19     2
Name: count, dtype: int64
```

In [83]:
```
plt.hist(data.index,bins=[19,29,39,49],edgecolor='black')
plt.xlabel('Age')
plt.title('Histogram Showing Age and Frequency of Employees')
plt.show()
```



Data Story : Most number of players are in between 19-29 and 29-39 age group.A Few are in the age between 39 and 49.

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

In [61]:
```
group_data=df.groupby(['Team','Position'])['Salary'].sum()
group_data
```

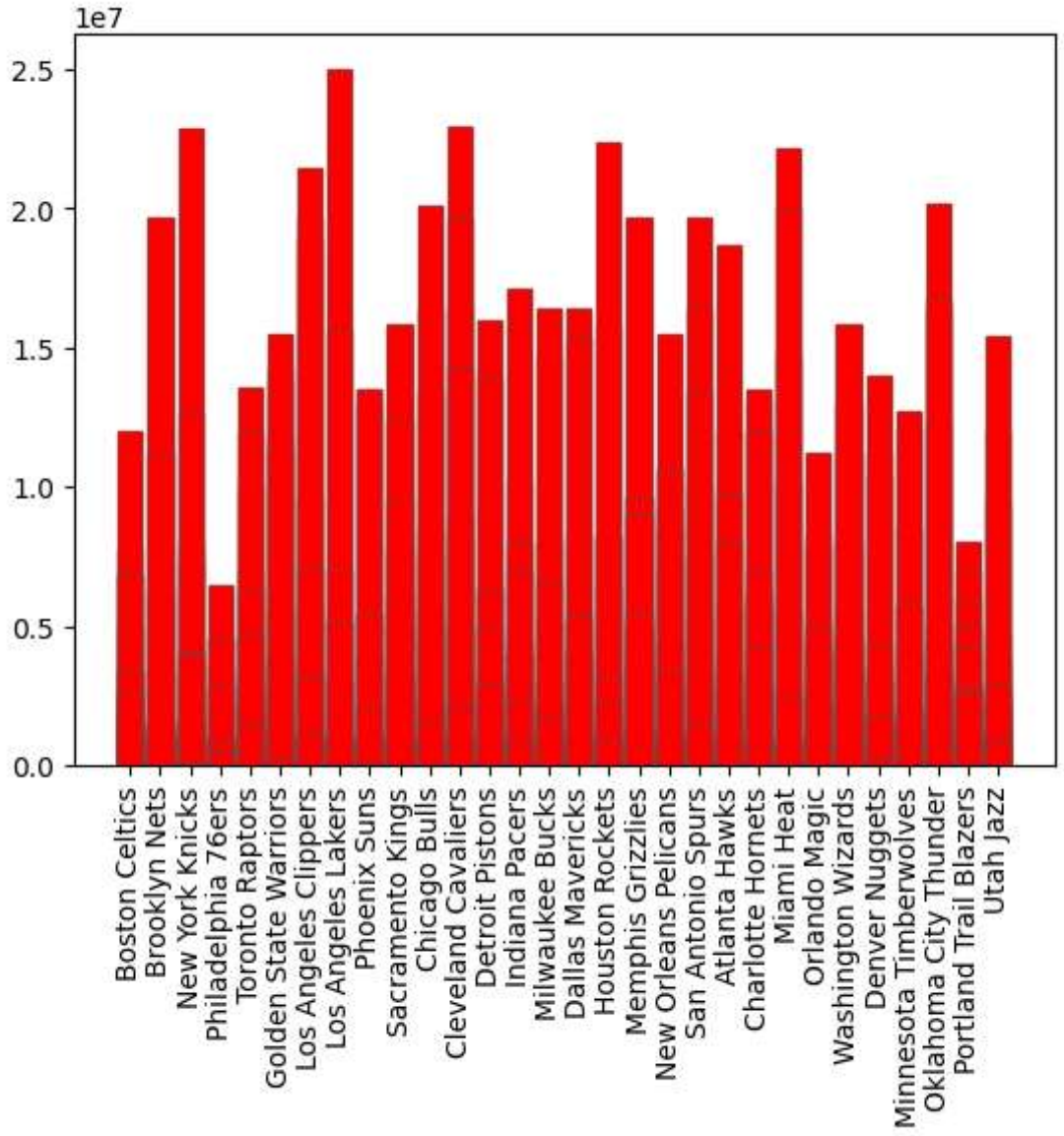
Out[61]:
```
Team                Position
Atlanta Hawks       C           22756250.0
                    PF          23952268.0
                    PG           9763400.0
                    SF           6000000.0
                    SG          10431032.0
                                    ...
Washington Wizards  C           24490429.0
                    PF          11300000.0
                    PG          18022415.0
                    SF          11158800.0
                    SG          11356992.0
Name: Salary, Length: 149, dtype: float64
```

In [34]:
```
#Sorting results in descending order of total salary
sorted_data=group_data.sort_values(ascending=False)
print(sorted_data.head())
```

```
Team                Position
Los Angeles Lakers  SF          31866445.0
Miami Heat          PF          31538671.0
Houston Rockets     SG          28122883.0
Phoenix Suns        PG          28002998.0
Denver Nuggets      SF          27982771.0
Name: Salary, dtype: float64
```

Data Story: Team Los Angeles Lakers with position SF, spends the highest for salary.

In [55]:
```
# Plot showing Team and spending on salary
plt.bar(df['Team'],df['Salary'],color='r',edgecolor='brown')
plt.xticks(rotation='vertical')
plt.show()
```



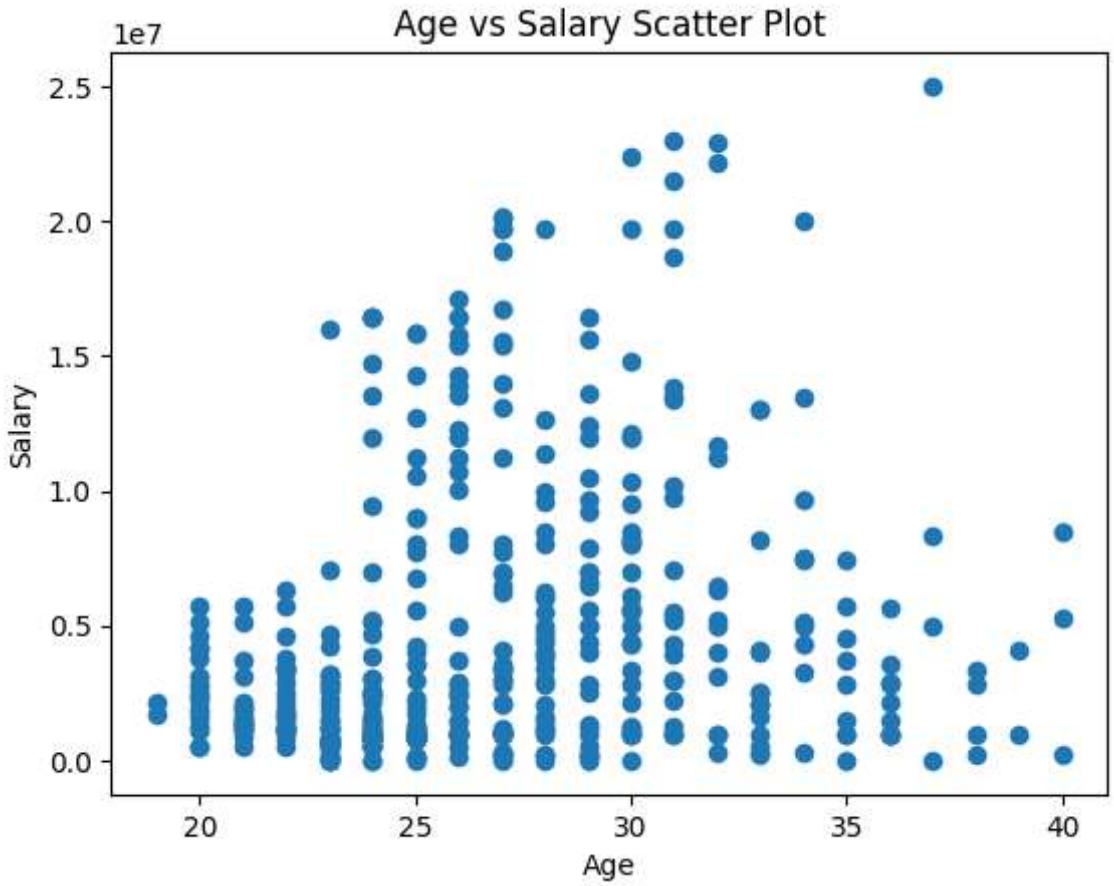
Data Story: Los Angeles Lakers spends more on salary

5. Find if there is any correlation between age and salary. Represent it visually.

In [40]:
```
correlation=df['Age'].corr(df['Salary'])
print(correlation)
```

```
0.2050096028480935
```

In [42]:
```
plt.scatter(df['Age'],df['Salary'])
plt.title('Age vs Salary Scatter Plot')
plt.xlabel('Age')
plt.ylabel('Salary')
plt.show()
```



Data Story: There is a weak positive correlation between Age and Salary. As age increases, salary tends to increase slightly, but the relationship is not very strong

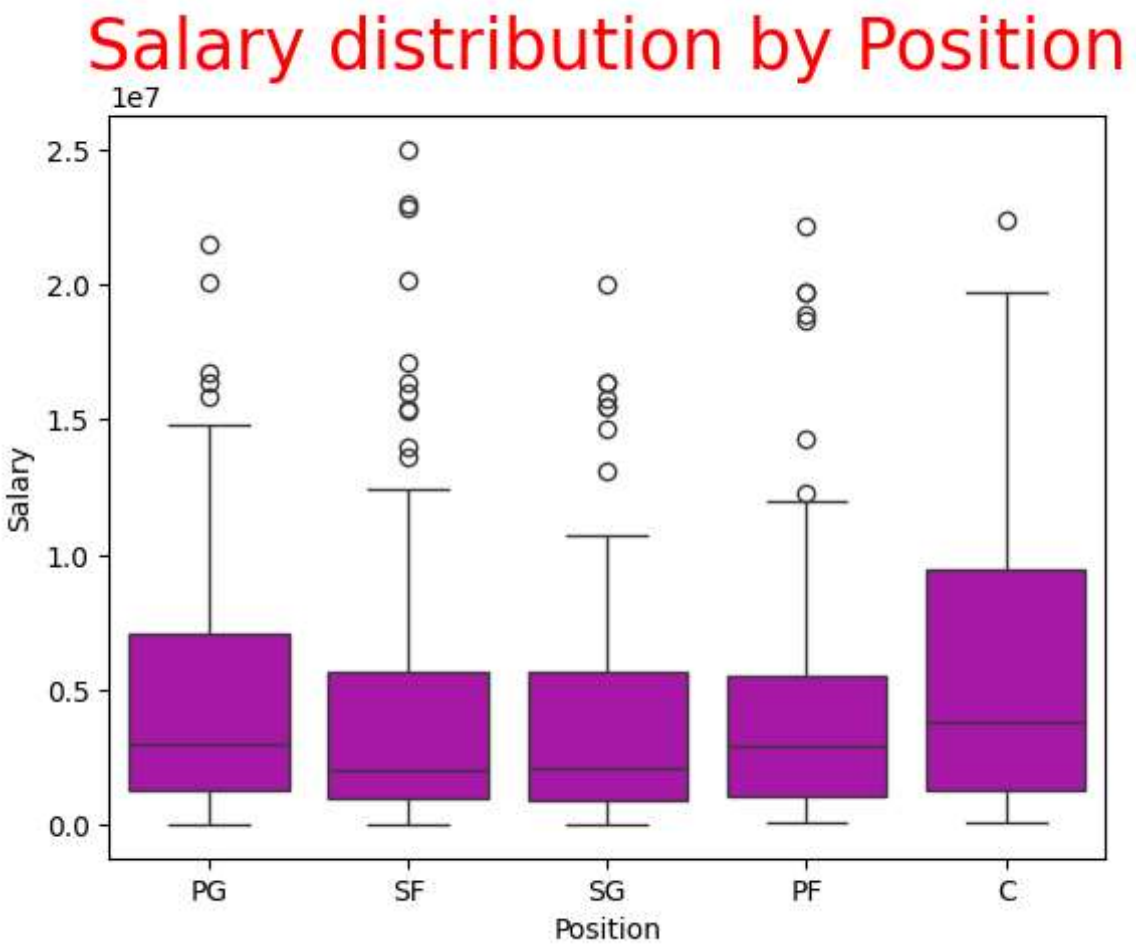
Exploring some possible correlations.

Relationship between a Employees' position and their salary

In [74]:
```
position_salary=df.groupby('Position')['Salary'].sum()
position_salary
```

Out[74]:
```
Position
C     466377332.0
PF    442560850.0
PG    446848971.0
SF    408020976.0
SG    396976258.0
Name: Salary, dtype: float64
```

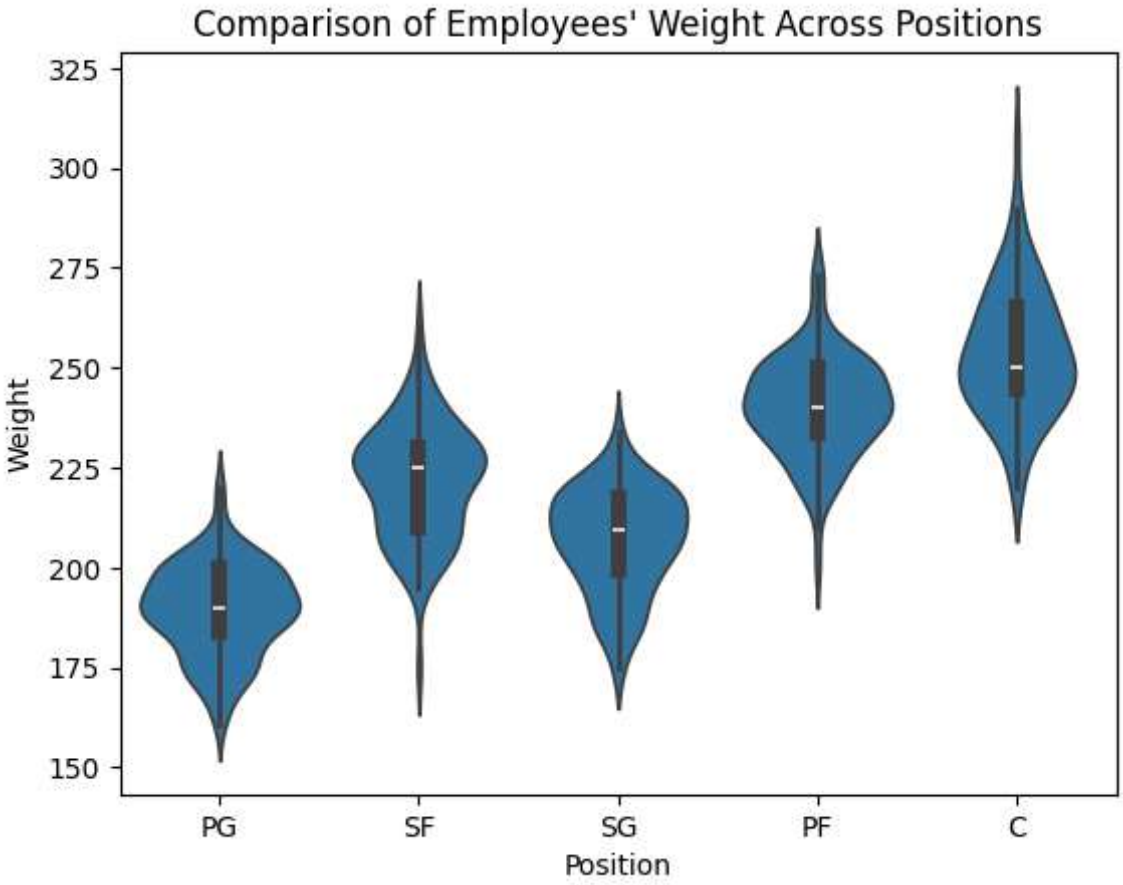
In [71]:
```
#visual representation using box plot
sns.boxplot(x='Position',y='Salary',data=df,color='m')
plt.title('Salary distribution by Position',color='r',size=25)
plt.xlabel('Position')
plt.ylabel('Salary')
plt.show()
```



Data Story : Employees in position 'C' tend to get higher salaries than other positions.

Relationship between a player's weight and their position.

In [84]:
```
sns.violinplot(x='Position',y='Weight',data=df)
plt.title("Comparison of Employees' Weight Across Positions")
plt.show()
```



Data Story: The violin plot shows that there is a relation between weight and position they occupy. Players with more weight tend to occupy position 'C' while players with less weight tend to occupy position 'PG'

Unveiling Insights:Exploratory Data Analysis of the dataset

The dataset of employees consists of 458 rows and 9 columns. Columns include Name, Team, Number, Position, Age, Height, Weight, College, Salary of the players are present

1. Team 'New Orleans Pelicans' has the highest number of players. Most of the teams have a percentage share of players between 3 and 3.5

2. segregated players w.r.t different positions and found out that most numbers of players have 'SG' position followed by 'PF'.Position 'C' has the least number of players.

3. Most of the players are in the age group of 19-29 and 29-39.

4.Team Los Angeles Lakers with position SF, spends the highest for salary

5. while finding the correlation between age and salary we found that there is a weak positive correlation exists between Age and Salary.As age increases, salary tends to increase slightly, but the relationship is not very strong.

6.The Box plot showing the relationship between an employee's position and their salary shows that employees in position 'C' tend to earn more salary than players in any other position.

7. The violin plot comparing weight and positions shows that there is a relation between weight and the position employees' occupy.