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

In [2]:
```
df=pd.read_csv('C:/Users/ahsan/Downloads/myexcel - myexcel.csv.csv')
```

In [3]:
```
pd.DataFrame(df)
```

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]:
```
df['Height'] = np.random.randint(150, 181, size=len(df))
```

In [5]:
```
df['Height']
```

Out[5]:
```
0      179
1      168
2      166
3      155
4      162
      ... 
453    174
454    163
455    167
456    175
457    165
Name: Height, Length: 458, dtype: int32
```

In [ ]:

1. Determine the distribution of employees across each team and calculate the percentage split relative to the total number of employees

In [6]:

```
df
```

Out[6]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	25	179	180	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99	SF	25	168	235	Marquette	6796117.0
2	John Holland	Boston Celtics	30	SG	27	166	205	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28	SG	22	155	185	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8	PF	29	162	231	NaN	5000000.0
...	...	...	...	...	...	...	...	...	...
453	Shelvin Mack	Utah Jazz	8	PG	26	174	203	Butler	2433333.0
454	Raul Neto	Utah Jazz	25	PG	24	163	179	NaN	900000.0
455	Tibor Pleiss	Utah Jazz	21	C	26	167	256	NaN	2900000.0
456	Jeff Withey	Utah Jazz	24	C	26	175	231	Kansas	947276.0
457	Priyanka	Utah Jazz	34	C	25	165	231	Kansas	947276.0

458 rows × 9 columns

In [7]:
```
x=df['Team']
distribution=x.value_counts()
distribution
```

```
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 [8]:
```
x=df['Name']
count=x.value_counts()
total_emply=sum(count)
total_emply
```

Out[8]: 458

In [9]:
```
percentage=(distribution/total_emply)*100
percentage
```

Out[9]:
```
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
```

2. Segregate employees based on their positions within the company. (2 marks)

In [10]:
```
x=df['Position']
position_counts=x.value_counts()
position_counts_df = pd.DataFrame(position_counts).reset_index()
position_counts_df.columns = ['Position', 'Employees']
print(position_counts_df)
```

```
  Position  Employees
0       SG        102
1       PF        100
2       PG         92
3       SF         85
4        C         79
```

3. Identify the predominant age group among employees. (2 marks)

In [11]:
```
x=df['Age']
age_counts=x.value_counts()
predominant_age = age_counts.idxmax()
predominant_age_count = age_counts.max()
print('The predominant age group is',predominant_age,'with',predominant_age
```

```
The predominant age group is 24 with 47 employees
```

4. Discover which team and position have the highest salary expenditure. (2 marks)

In [12]:
```
salary_expenditure = df.groupby(['Team', 'Position'])['Salary'].sum().reset
max_expenditure = salary_expenditure.loc[salary_expenditure['Salary'].idxma
print(f"Team and Position with the highest salary expenditure:\n{max_expend
```

```
Team and Position with the highest salary expenditure:
Team        Los Angeles Lakers
Position                    SF
Salary              31866445.0
Name: 67, dtype: object
```

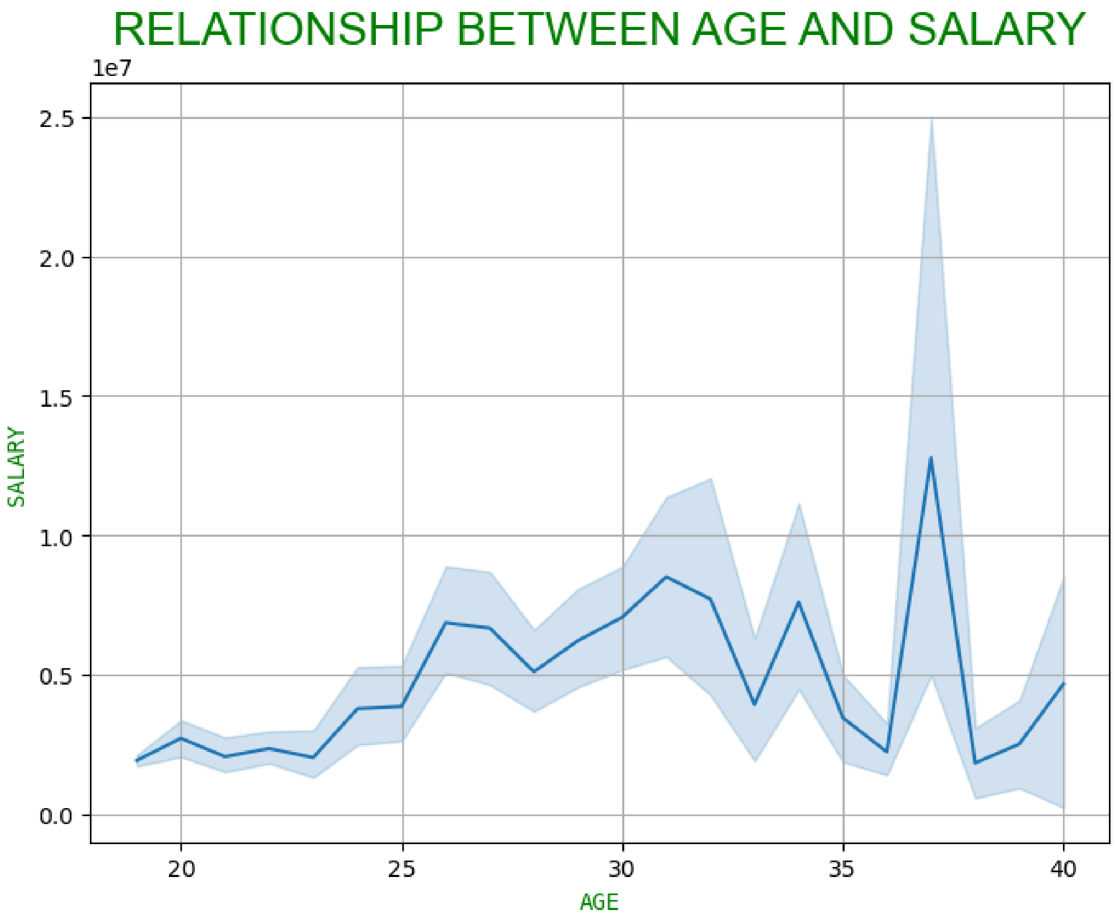
5. Investigate if there's any correlation between age and salary, and represent it visually. (2 marks)

In [13]:
```
correlation = df['Age'].corr(df['Salary'])
print("Correlation is",correlation)
```

```
Correlation is 0.21400941226570974
```

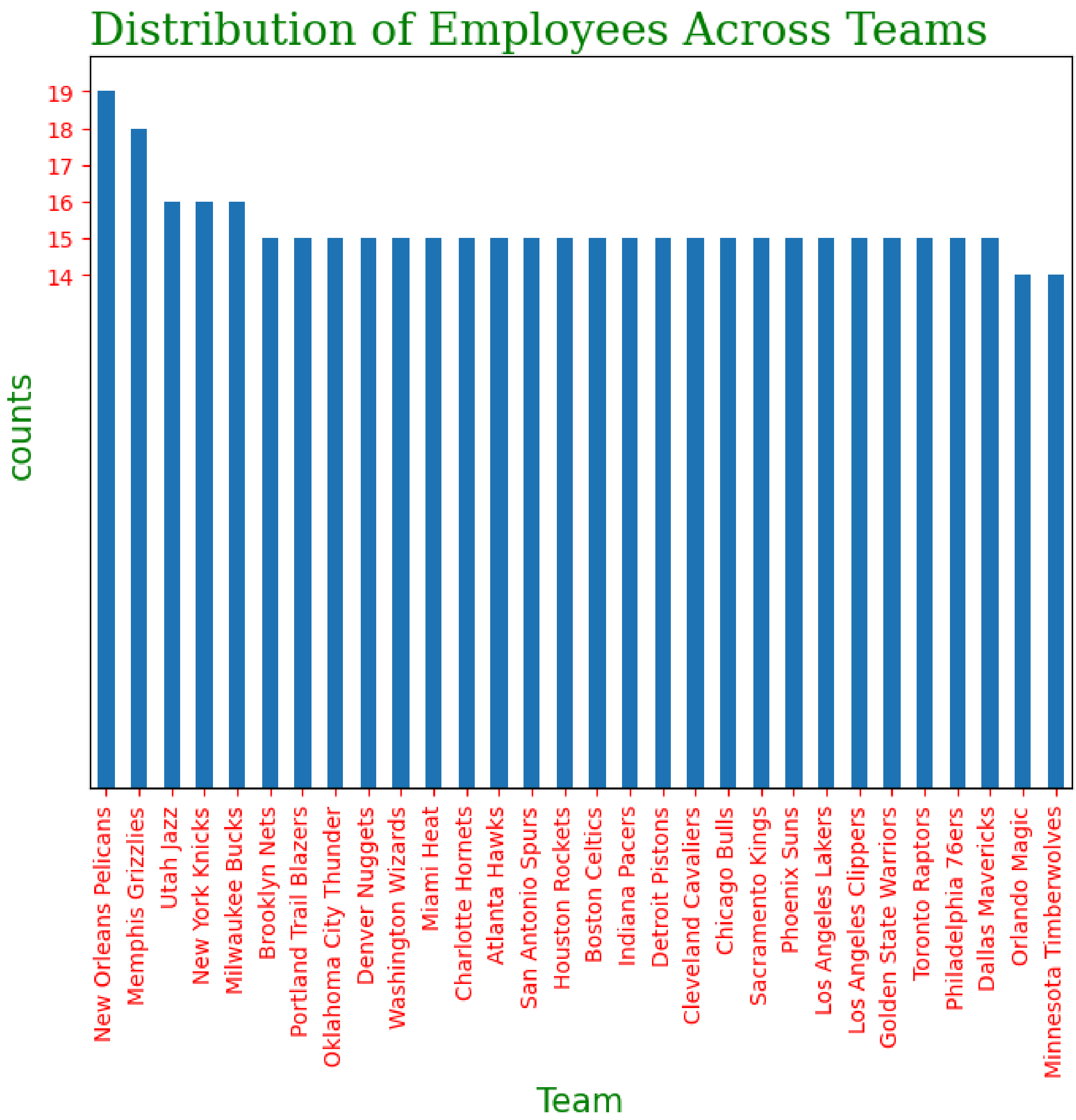
In [14]:
```
#visual representation

import matplotlib.pyplot as plt
import seaborn as sns
x=df['Age']
y=df['Salary']
```

In [25]:
```
plt.figure(figsize=(8,6))
sns.lineplot(x=x, y=y, data=df)
font1={'family':'Arial','color':'green','size':20}
font2={'family':'monospace','color':'green','size':10}
plt.title('RELATIONSHIP BETWEEN AGE AND SALARY',fontdict=font1)
plt.xlabel('AGE',fontdict=font2)
plt.ylabel('SALARY',fontdict=font2)
plt.grid(True)
plt.show()
```



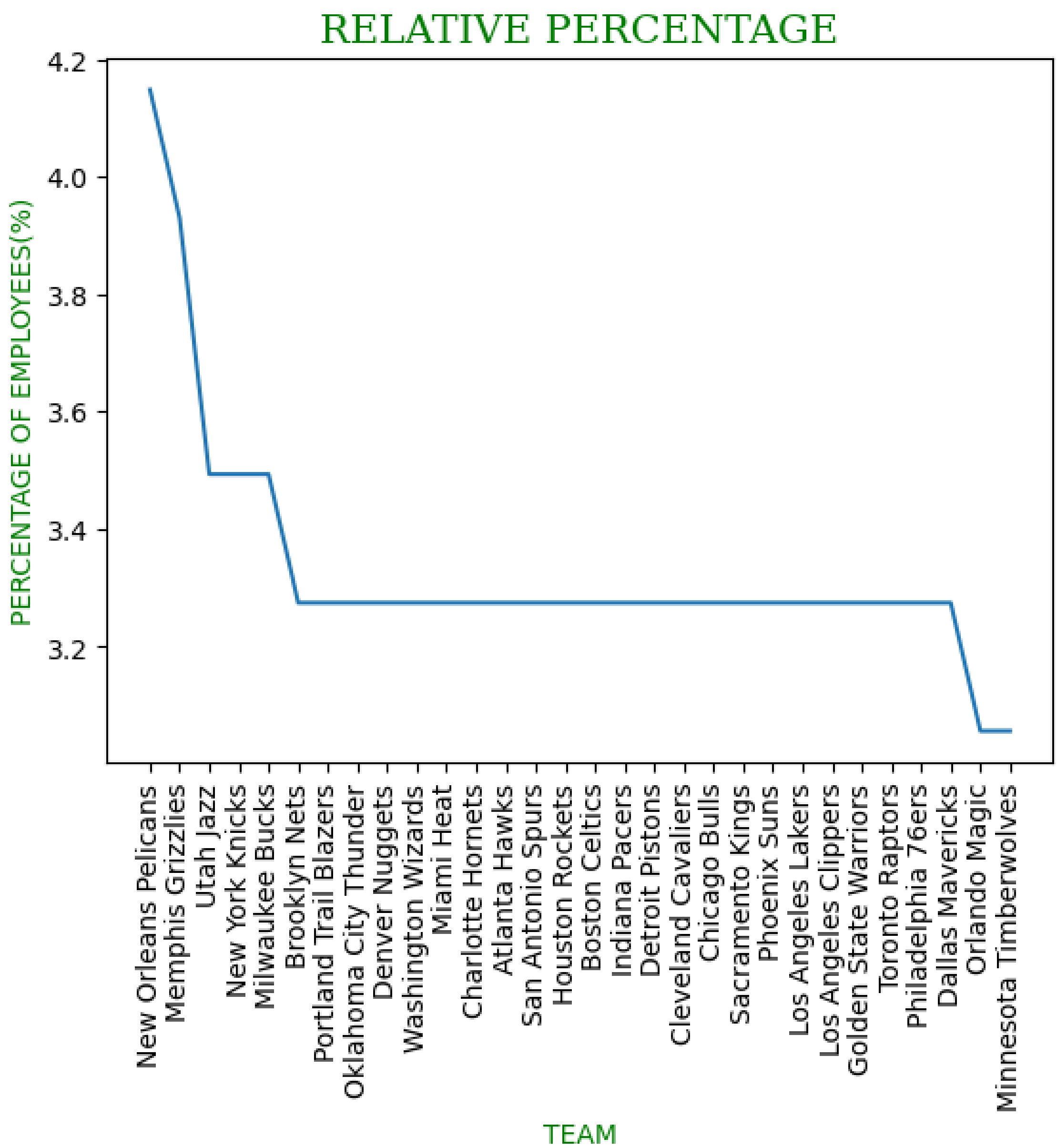
In [16]:
```
#visualization for the first question
```

In [17]:
```
plt.figure(figsize=(8,6))
team_counts=df['Team'].value_counts()
team_counts.plot(kind='bar')
font1={'family':'serif','color':'green','size':20}
font2={'family':'sans-serif','color':'green','size':15}
plt.xlabel('Team',fontdict=font2)
plt.ylabel('counts',fontdict=font2)
plt.tick_params(direction='out', colors='red')
plt.yticks([14,15,16,17,18,19])
plt.title('Distribution of Employees Across Teams',fontdict=font1,loc ='lef
plt.show()
```



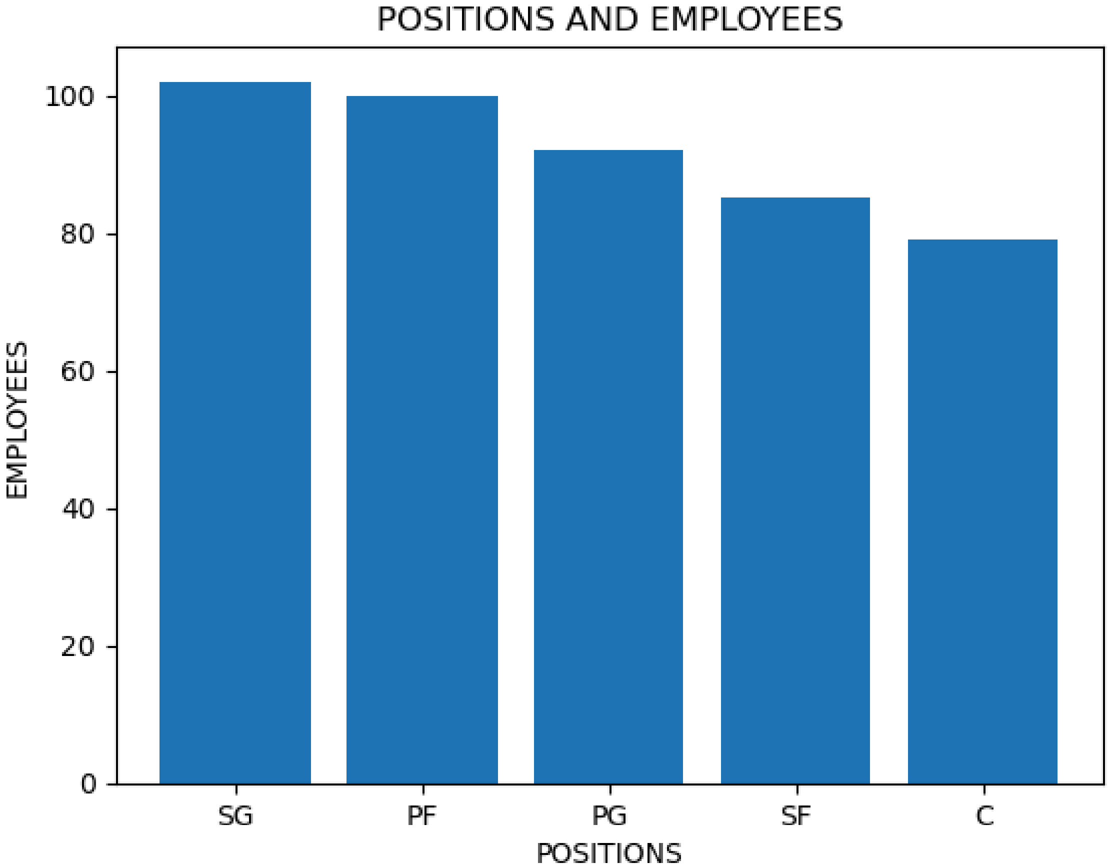
In [18]:
```
c=df['Team'].value_counts()
k=c.keys()
```

In [19]:
```
import seaborn as sns
import matplotlib.pyplot as plt


sns.lineplot(x=k,y=percentage)
font1={'family':'serif','color':'green','size':15}
font2={'family':'sans-serif','color':'green','size':10}
plt.title('RELATIVE PERCENTAGE',fontdict=font1)
plt.xlabel('TEAM',fontdict=font2)
plt.xticks(rotation=90)
plt.ylabel('PERCENTAGE OF EMPLOYEES(%)',fontdict=font2)
plt.show()
```



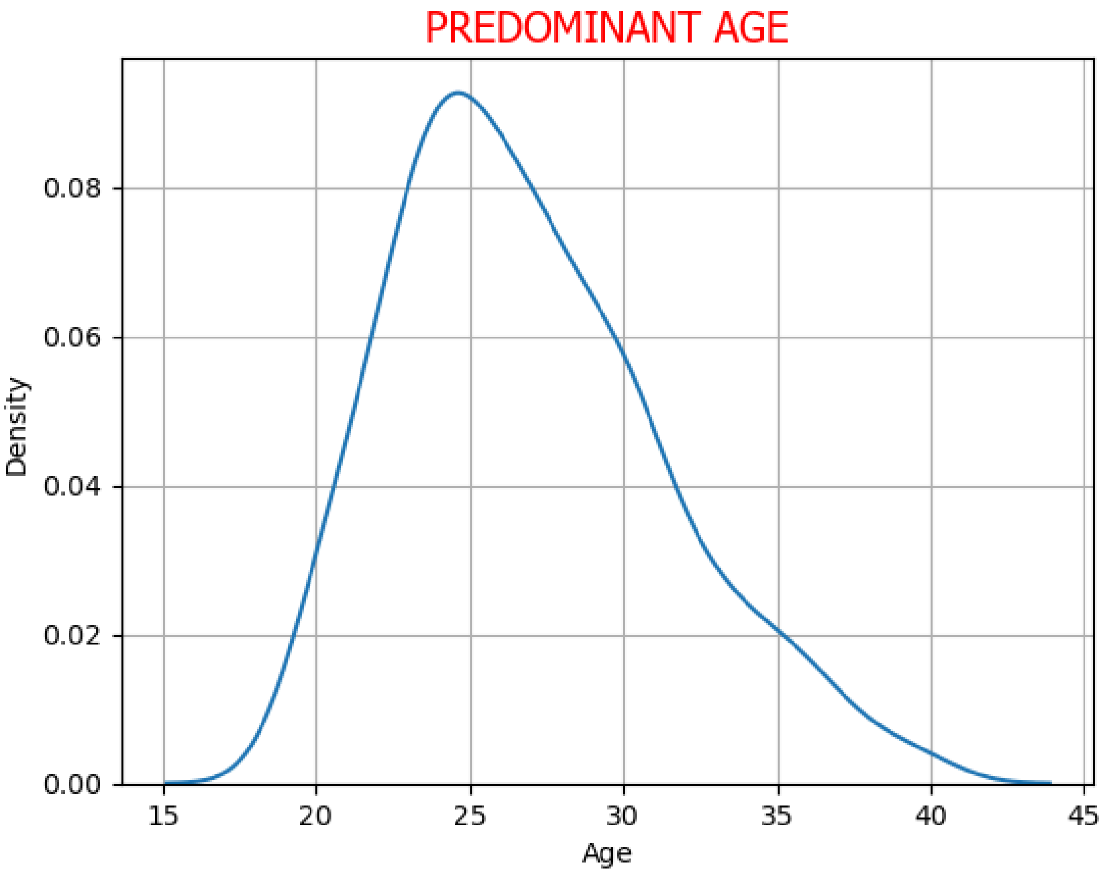
In [20]:
```
#visualization for the second question
```

In [21]:
```
plt.bar(position_counts_df['Position'],position_counts_df['Employees'])
plt.title('POSITIONS AND EMPLOYEES')
plt.xlabel('POSITIONS')
plt.ylabel('EMPLOYEES')
plt.show()
```

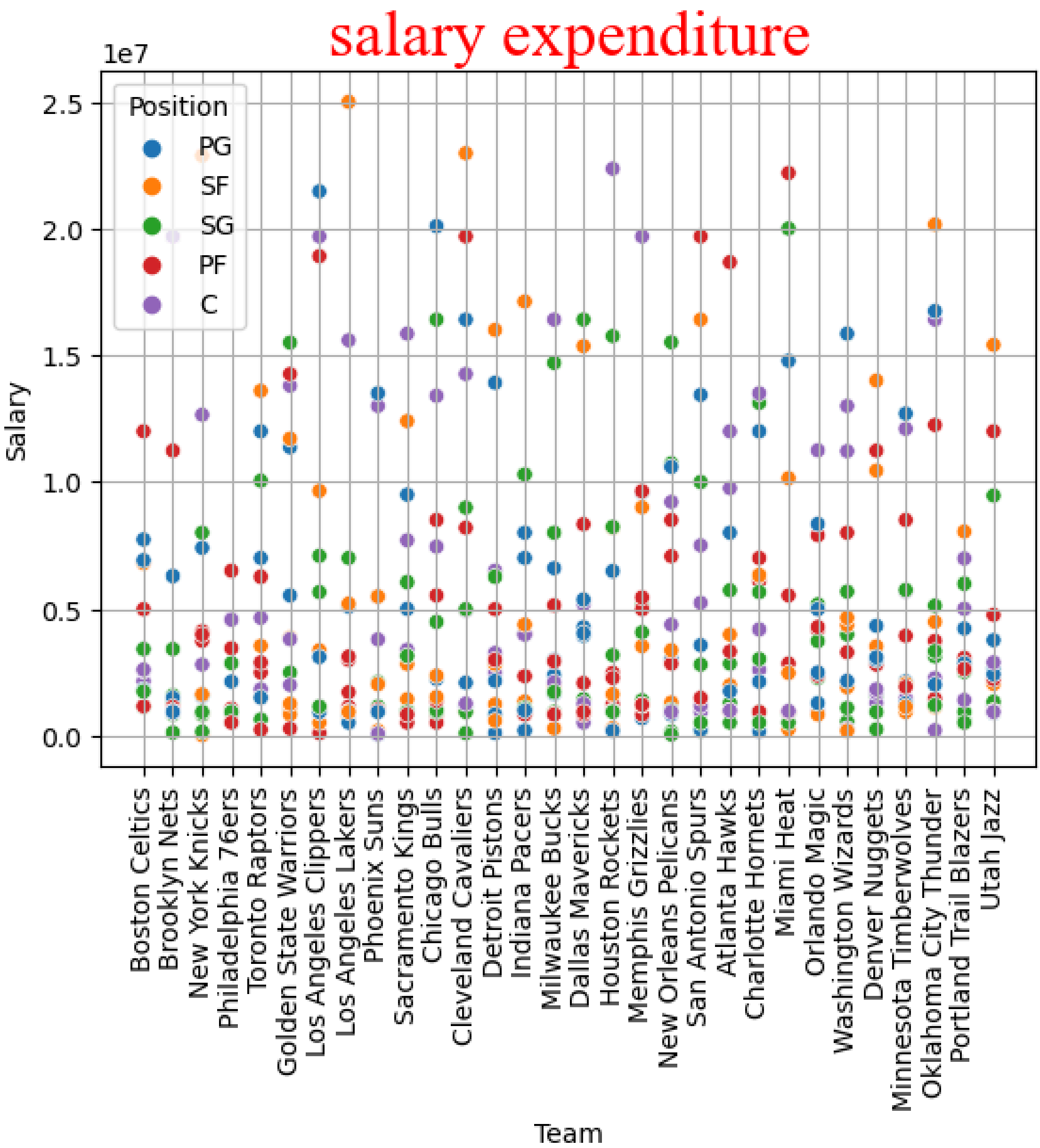


In [22]:
```
#visualization for the question three
```

In [23]:
```
sns.kdeplot(df['Age'])
font={'family':'Tahoma','color':'red','size':15}
plt.title('PREDOMINANT AGE',fontdict=font)
plt.grid(True)
plt.show()
```



In [24]:
```
sns.scatterplot(x=df['Team'],y=df['Salary'],data=df,hue=df['Position'])
font={'family':'Times New Roman','color':'red','size':25}
plt.title('salary expenditure',fontdict=font)
plt.xticks(rotation=90)
plt.grid(True)
plt.show()
```



In [36]:
```
legendary_player=df['Salary']==max(df['Salary'])
df[legendary_player]
```

Out[36]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
109	Kobe Bryant	Los Angeles Lakers	24	SF	37	154	212	NaN	25000000.0

From the graph we can say that Orleans pelicans has just over 4% of employees,while Orlando Magic and Minnesota Timberwolves have just under 1% of employees.And the other teams appear relatively even in height, suggesting a fairly uniform distribution of employees across different teams.

The actual counts verify the comparable representation of SG and PF, showing that their numbers are quite close. It is accurate to indicate that PG is greater than SF and C but somewhat lower than SG and PF.The least are C and SF.

The predominant age group among employees is 24 years old, comprising 47 individuals, which suggests a strong interest among adults in basketball.

The plot shows a right-skewed distribution. It rises steeply to the peak at age 24 and then gradually declines. Ages below 24 are less common, and there's a rapid increase in density up to 24.Ages above 24 show a more gradual decline, suggesting a steady decrease in frequency as age increases.

The graph(relationship between age and salary) shows that, in general, salary tends to increase with age. As individuals gain more work experience, their earnings typically rise.And there's a significant peak in salary between ages 35 and 40. This suggests that professionals in this age range tend to earn the most.However, the sharp drop after this peak indicates that there might be other factors at play, such as retirement or career shifts.

Los Angeles Lakers(SF) has highest salary expenditure.The team allocates significant resources to small forwards, possibly emphasizing star players or key contributors in that position.

Kobe Bryant, a small forward for the Los Angeles Lakers, earned a salary of approximately $25,000,000. At around 37 years old, this high salary reflects his extensive experience and exceptional skills as a player.