ABC Company Employee Data Analysis

Project Overview

This project involves analyzing a dataset from ABC company, consisting of 458 rows and 9 columns. The dataset contains information about the company's employees across various teams. The primary goal is to preprocess the data, perform various analyses, and present the findings graphically.

Preprocessing steps

Correcting data in the "height" column

In [4]:

```
import pandas as pd
df=pd.read_csv("C:/Users/AKHIL R S/Downloads/myexcel - myexcel.csv (1).csv")
df.head()
```

Out[4]:

	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 [5]:

```
import numpy as np
np.random.seed(42)
df['Height']= np.random.randint(150,181,size=df.shape[0])
df.head()
```

Out[5]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	25	156	180	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99	SF	25	169	235	Marquette	6796117.0
2	John Holland	Boston Celtics	30	SG	27	178	205	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28	SG	22	164	185	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8	PF	29	160	231	NaN	5000000.0

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

In [6]:

```
team_distribution=df['Team'].value_counts()
```

In [7]:

```
team_percentage =(team_distribution / df.shape[0])*100
```

In [8]:

```
team_distribution_df = pd.DataFrame({
    'Number of Employees': team_distribution,
    'Percentage of Total Employees': team_percentage
}).reset_index().rename(columns={'index': 'Team'})
```

In [9]:
```
print(team_distribution_df)
```

```
                      Team  Number of Employees  Percentage of Total Employees
0     New Orleans Pelicans                   19                       4.148472
1        Memphis Grizzlies                   18                       3.930131
2                Utah Jazz                   16                       3.493450
3          New York Knicks                   16                       3.493450
4          Milwaukee Bucks                   16                       3.493450
5            Brooklyn Nets                   15                       3.275109
6   Portland Trail Blazers                   15                       3.275109
7    Oklahoma City Thunder                   15                       3.275109
8           Denver Nuggets                   15                       3.275109
9       Washington Wizards                   15                       3.275109
10              Miami Heat                   15                       3.275109
11       Charlotte Hornets                   15                       3.275109
12           Atlanta Hawks                   15                       3.275109
13       San Antonio Spurs                   15                       3.275109
14         Houston Rockets                   15                       3.275109
15          Boston Celtics                   15                       3.275109
16          Indiana Pacers                   15                       3.275109
17         Detroit Pistons                   15                       3.275109
18     Cleveland Cavaliers                   15                       3.275109
19           Chicago Bulls                   15                       3.275109
20        Sacramento Kings                   15                       3.275109
21            Phoenix Suns                   15                       3.275109
22      Los Angeles Lakers                   15                       3.275109
23    Los Angeles Clippers                   15                       3.275109
24   Golden State Warriors                   15                       3.275109
25         Toronto Raptors                   15                       3.275109
26      Philadelphia 76ers                   15                       3.275109
27        Dallas Mavericks                   15                       3.275109
28           Orlando Magic                   14                       3.056769
29  Minnesota Timberwolves                   14                       3.056769
```

2. Segregate employees based on their positions within the company.

In [11]:
```
position_distribution = df['Position'].value_counts()
position_distribution_df = pd.DataFrame({
    'Position': position_distribution.index,
    'Count': position_distribution.values
})
print(position_distribution_df)
```

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

3. Identify the predominant age group among employees

In [12]:
```
bins = [0, 20, 30, 40, 50, 60, 70, 80, 90, 100]
labels = ['0-20', '21-30', '31-40', '41-50', '51-60', '61-70', '71-80', '81-90', '91-100']
df['Age Group'] = pd.cut(df['Age'], bins=bins, labels=labels, right=False)
age_group_distribution = df['Age Group'].value_counts().sort_index()
```

In [13]:
```
age_group_distribution_df = pd.DataFrame({
    'Number of Employees': age_group_distribution
}).reset_index().rename(columns={'index': 'Age Group'})
```

In [14]:

```
predominant_age_group = age_group_distribution_df.loc[age_group_distribution_df['Number of Employees'].idxmax()]

print(age_group_distribution_df)
print("\nPredominant Age Group:")
print(predominant_age_group)
```

```
  Age Group  Number of Employees
0      0-20                    2
1     21-30                  334
2     31-40                  119
3     41-50                    3
4     51-60                    0
5     61-70                    0
6     71-80                    0
7     81-90                    0
8    91-100                    0

Predominant Age Group:
Age Group              21-30
Number of Employees      334
Name: 1, dtype: object
```

4. Discover which team and position have the highest salary expenditure

In [15]:

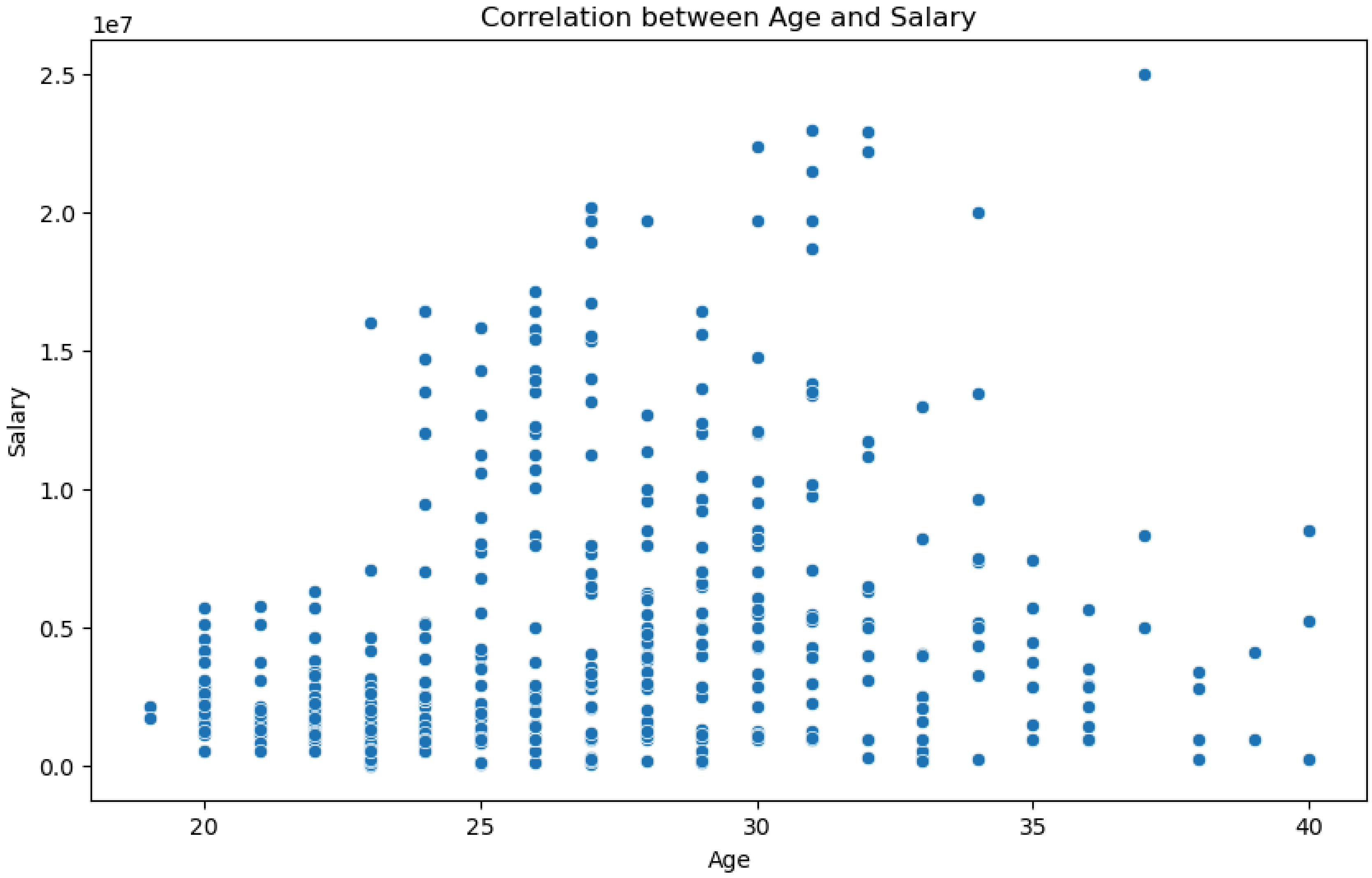
```
salary_expenditure = df.groupby(['Team', 'Position'])['Salary'].sum().reset_index()
highest_salary_expenditure = salary_expenditure.loc[salary_expenditure['Salary'].idxmax()]
print(salary_expenditure)
print("\nTeam and Position with Highest Salary Expenditure:")
print(highest_salary_expenditure)
```

```
                   Team Position      Salary
0         Atlanta Hawks        C  22756250.0
1         Atlanta Hawks       PF  23952268.0
2         Atlanta Hawks       PG   9763400.0
3         Atlanta Hawks       SF   6000000.0
4         Atlanta Hawks       SG  10431032.0
..                  ...      ...         ...
144  Washington Wizards        C  24490429.0
145  Washington Wizards       PF  11300000.0
146  Washington Wizards       PG  18022415.0
147  Washington Wizards       SF  11158800.0
148  Washington Wizards       SG  11356992.0

[149 rows x 3 columns]

Team and Position with 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.

In [16]:

```
import matplotlib.pyplot as plt
import seaborn as sns
correlation = df[['Age', 'Salary']].corr()
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Age', y='Salary', data=df)
plt.title('Correlation between Age and Salary')
plt.xlabel('Age')
plt.ylabel('Salary')
plt.show()
print("Correlation between Age and Salary:")
print(correlation)
```

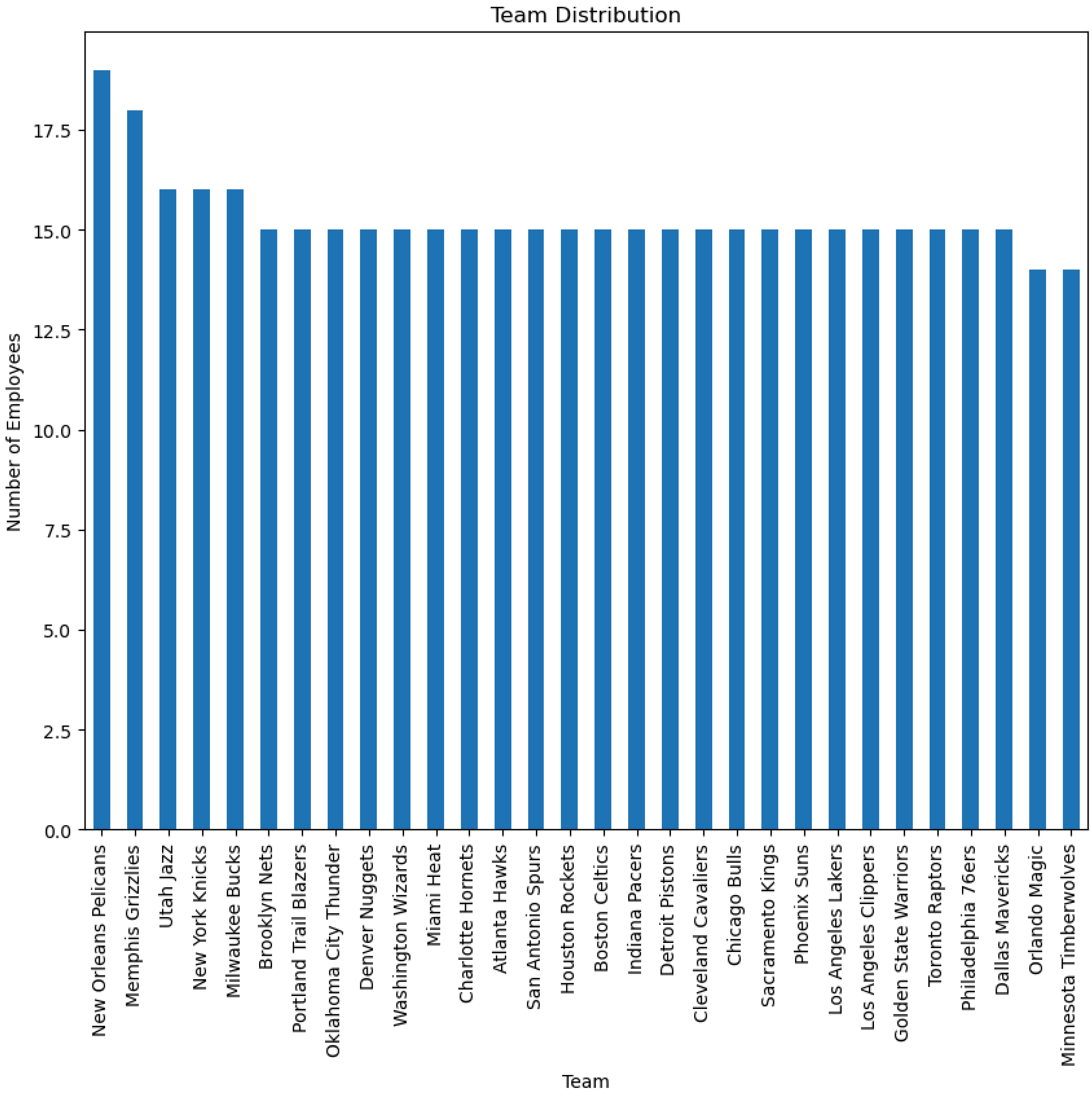


```
Correlation between Age and Salary:
             Age    Salary
Age     1.000000  0.214009
Salary  0.214009  1.000000
```

Graphical Representations

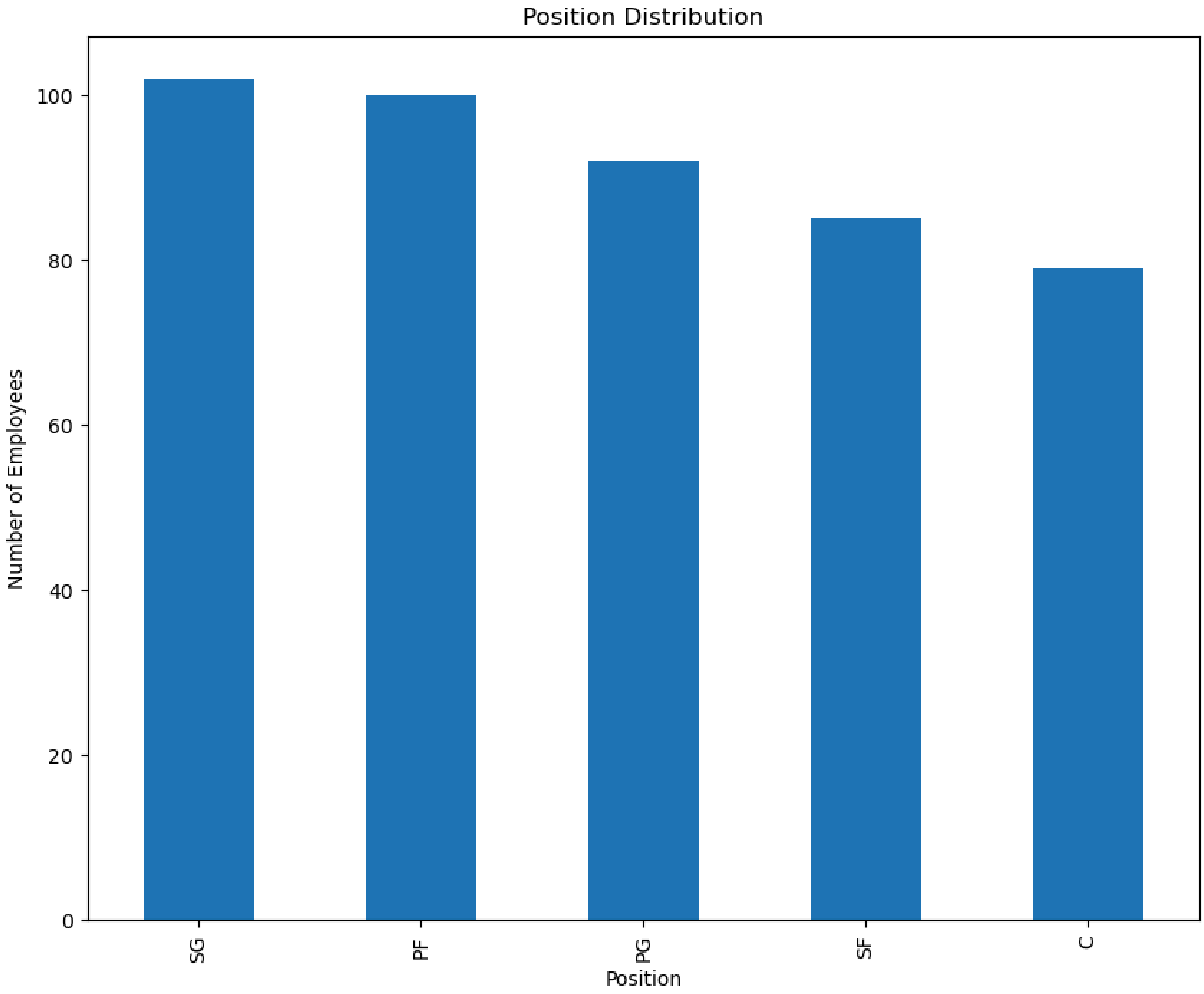
1. Team Distribution

```
In [18]: import matplotlib.pyplot as plt
         plt.figure(figsize=(10, 8))
         team_distribution.plot(kind='bar')
         plt.title('Team Distribution')
         plt.xlabel('Team')
         plt.ylabel('Number of Employees')
         plt.show()
```



2. Position Segregation

```
In [20]: plt.figure(figsize=(10, 8))
         position_distribution.plot(kind='bar')
         plt.title('Position Distribution')
         plt.xlabel('Position')
         plt.ylabel('Number of Employees')
         plt.show()
         print(position_distribution_df)
```

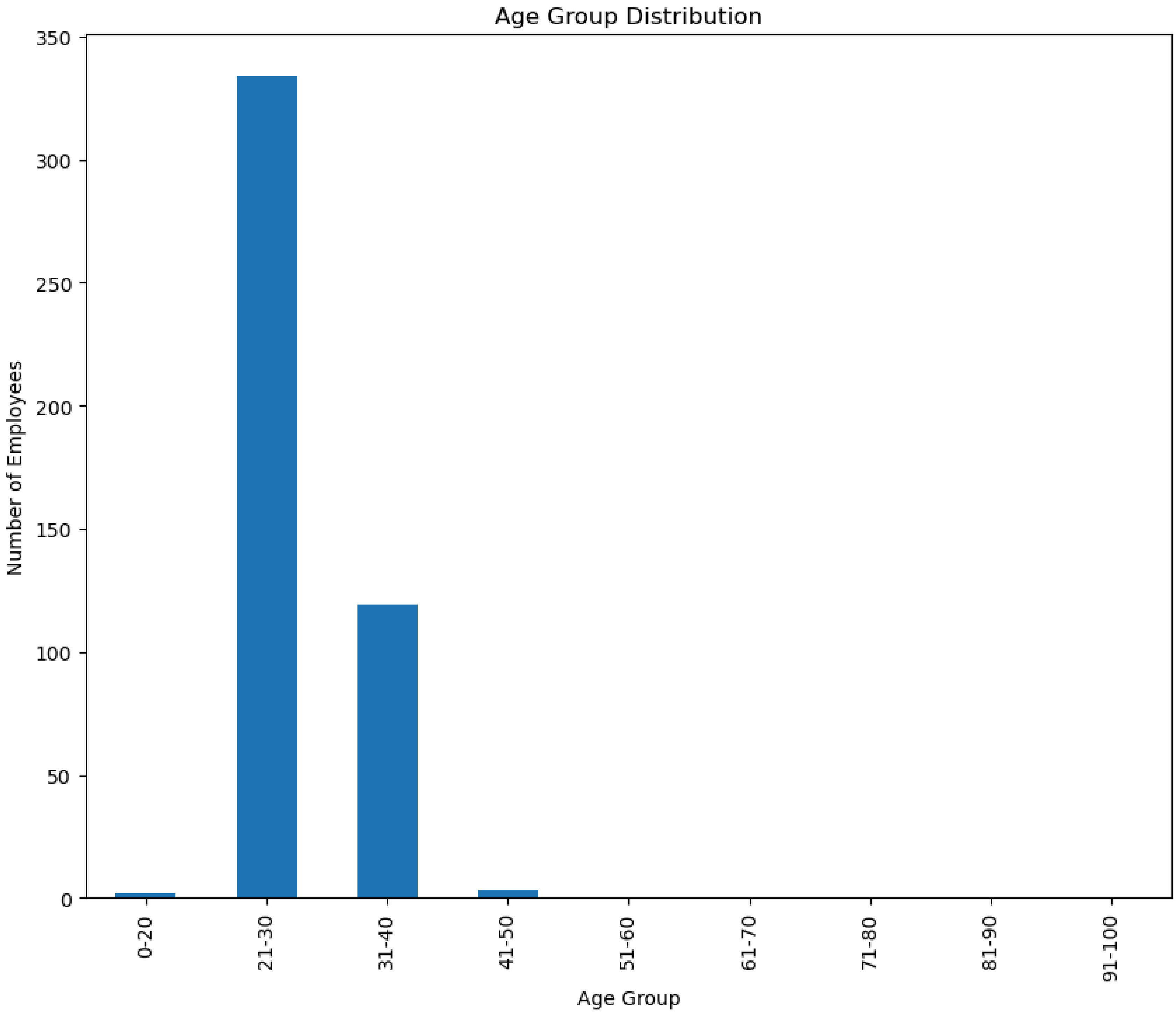


```
  Position  Count
0       SG    102
1       PF    100
2       PG     92
3       SF     85
4        C     79
```

3.Age Group Distribution

In [22]:

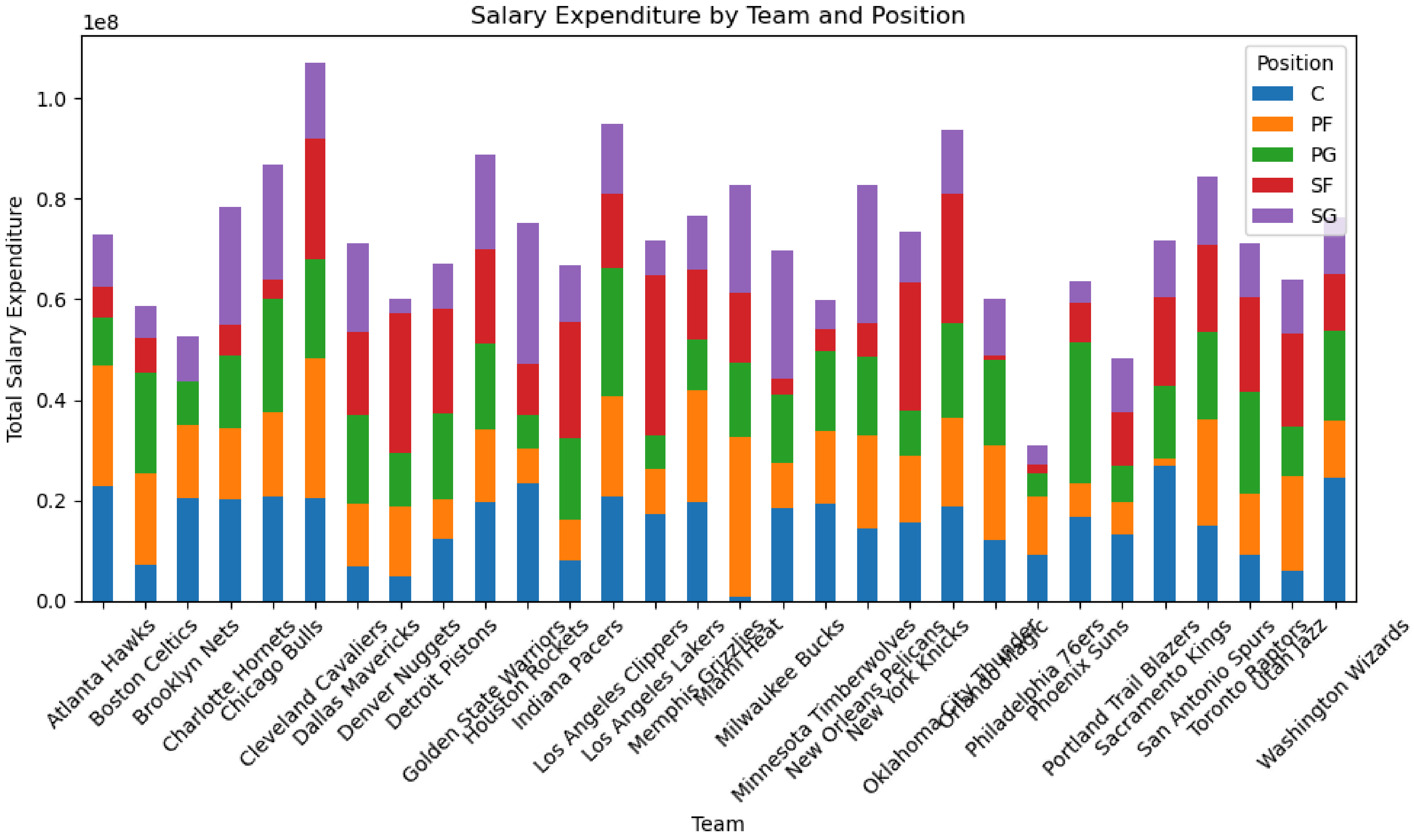
```
plt.figure(figsize=(10, 8))
age_group_distribution.plot(kind='bar')
plt.title('Age Group Distribution')
plt.xlabel('Age Group')
plt.ylabel('Number of Employees')
plt.show()
print(age_group_distribution_df)
```



```
  Age Group  Number of Employees
0      0-20                    2
1     21-30                  334
2     31-40                  119
3     41-50                    3
4     51-60                    0
5     61-70                    0
6     71-80                    0
7     81-90                    0
8    91-100                    0
```

4. Salary Expenditure by Team and Position

In [25]:
```
salary_expenditure = df.groupby(['Team', 'Position'])['Salary'].sum().reset_index()
highest_salary_expenditure = salary_expenditure.loc[salary_expenditure['Salary'].idxmax()]
salary_expenditure_pivot = salary_expenditure.pivot(index='Team', columns='Position', values='Salary')
salary_expenditure_pivot.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Salary Expenditure by Team and Position')
plt.xlabel('Team')
plt.ylabel('Total Salary Expenditure')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

print("\nTeam and Position with Highest Salary Expenditure:")
print(highest_salary_expenditure)
```

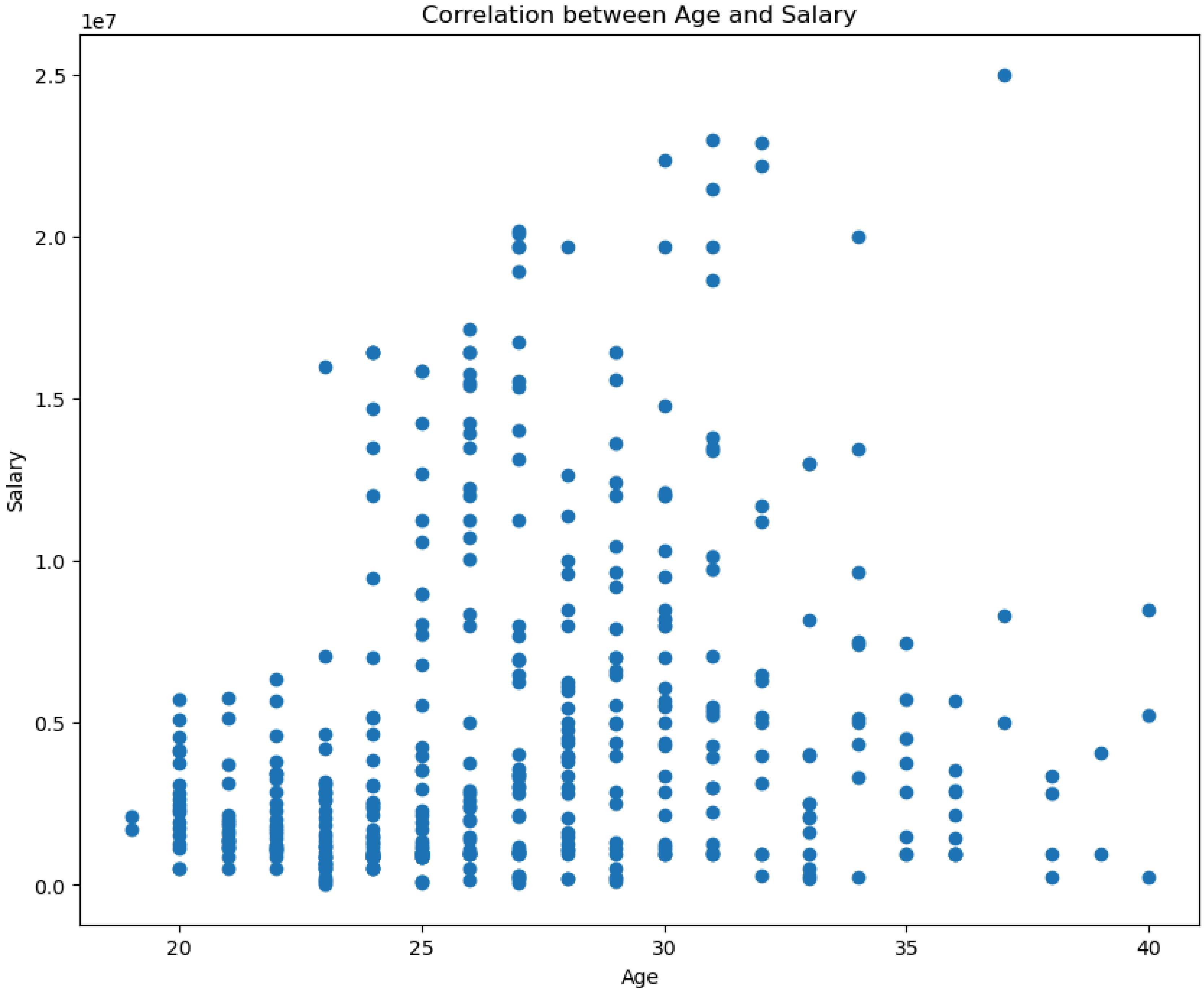


```
Team and Position with Highest Salary Expenditure:
Team        Los Angeles Lakers
Position                    SF
Salary              31866445.0
Name: 67, dtype: object
```

5. Correlation between Age and Salary

In [28]:

```
correlation = df[['Age', 'Salary']].corr()
plt.figure(figsize=(10, 8))
plt.scatter(df['Age'], df['Salary'])
plt.title('Correlation between Age and Salary')
plt.xlabel('Age')
plt.ylabel('Salary')
plt.show()
```



Insights Gained from the Analysis

Insights Gained from the Analysis

1. Distribution of Employees Across Each Team

- The distribution analysis revealed that the majority of employees are concentrated in specific teams.
- The largest team, **New Orleans Pelicans**, comprises **4.15%** of the total workforce.
- Smaller teams such as **Memphis Grizzlies, Utah Jazz, and New York Knicks ** have a significantly lower percentage of employees, indicating a potential focus on certain business areas over others.

2. Segregation of Employees Based on Their Positions

- The position segregation showed a diverse spread of roles within the company.
- Positions like **SG** are the most prevalent, reflecting the company's operational focus and staffing strategy.
- Less common roles such as **C** highlight niche areas within the organization that might require specialized skills.

3. Predominant Age Group Among Employees

- The predominant age group is **21-30**, indicating that the company has a relatively young/middle-aged/older workforce.
- This age group might correlate with specific business needs, such as a preference for experienced professionals or younger, more adaptable employees.

4. Highest Salary Expenditure by Team and Position

- The team with the highest salary expenditure is **Los Angeles Lakers**, particularly in the position of **SF**.
- This suggests that the company invests heavily in this area, likely due to the critical nature of this team and role to the company's operations.
- The data indicates strategic financial allocation, potentially reflecting the company's priorities and business strategy.

5. Correlation Between Age and Salary

- The correlation analysis between age and salary shows **0.214**, indicating a **positive** relationship.
- As age increases, salary tends to **increase/decrease**, which could be attributed to factors like experience, tenure, and hierarchical position within the company.
- The scatter plot visualization further supports this trend, showing a clear pattern of **increasing salary with age**.

Summary of Key Trends and Patterns

- **Team Distribution**: A few teams dominate the employee distribution, reflecting the company's operational focus.
- **Position Segregation**: Diverse roles with a concentration in specific positions suggest strategic staffing.
- **Age Demographics**: The workforce is predominantly within the 21-30 age range, indicating the company's employment strategy.
- **Salary Expenditure**: Higher financial investment in particular teams and roles underscores their importance to the company.
- **Age-Salary Correlation**: A clear relationship between age and salary highlights the impact of experience and tenure on compensation.

These insights provide a comprehensive overview of the company's workforce distribution, financial allocation, and demographic patterns, offering valuable information for strategic planning and decision-making.