Module 4: Python Project

Import Libraries

In [95]:
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [4]:
```python
data = pd.read_csv('abc_company.csv')
```

In [5]:
```python
data.head
```

Out[5]:
```
<bound method NDFrame.head of               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]>
```

In [6]:
```python
data.tail
```

Out[6]:
```
<bound method NDFrame.tail of               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]>
```

In [66]:
```python
data.info
```

Out[66]:
```
<bound method DataFrame.info of               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  height
0                Texas  7730337.0     171
1            Marquette  6796117.0     176
2    Boston University        NaN     174
3        Georgia State  1148640.0     151
4                  NaN  5000000.0     154
..                 ...        ...     ...
453             Butler  2433333.0     179
454                NaN   900000.0     161
455                NaN  2900000.0     177
456             Kansas   947276.0     162
457             Kansas   947276.0     150

[458 rows x 10 columns]>
```

In [76]:
```python
data.isnull()
```

Out[76]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary	height
0	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	True	False	False
...	...	...	...	...	...	...	...	...	...	...
453	False	False	False	False	False	False	False	False	False	False
454	False	False	False	False	False	False	False	True	False	False
455	False	False	False	False	False	False	False	True	False	False
456	False	False	False	False	False	False	False	False	False	False
457	False	False	False	False	False	False	False	False	False	False

458 rows × 10 columns

In [80]:
```python
# --- Null Value Handling ---
# 1. Check for Null Values
print("Null Values:\n", data.isnull().sum())  # Print the number of nulls in each column
```

```
Null Values:
 Name         0
Team         0
Number       0
Position     0
Age          0
Height       0
Weight       0
College     84
Salary      11
height       0
dtype: int64
```

In [97]:
```python
data['Salary'].fillna(data['Salary'] .median (), inplace=True)  # Fill with median (or some other appropriate value)
```

In [131…
```python
data['College'].fillna(0, inplace=True)  # Fill with 0 (or some other appropriate value)
```

In [99]:
```python
# Preprocessing: Correcting the 'height' column
data['height'] = np.random.randint(150, 181, size=len(data))  # Generates random integers between 150 and 180 (inclusive)
```

Analysis Tasks:

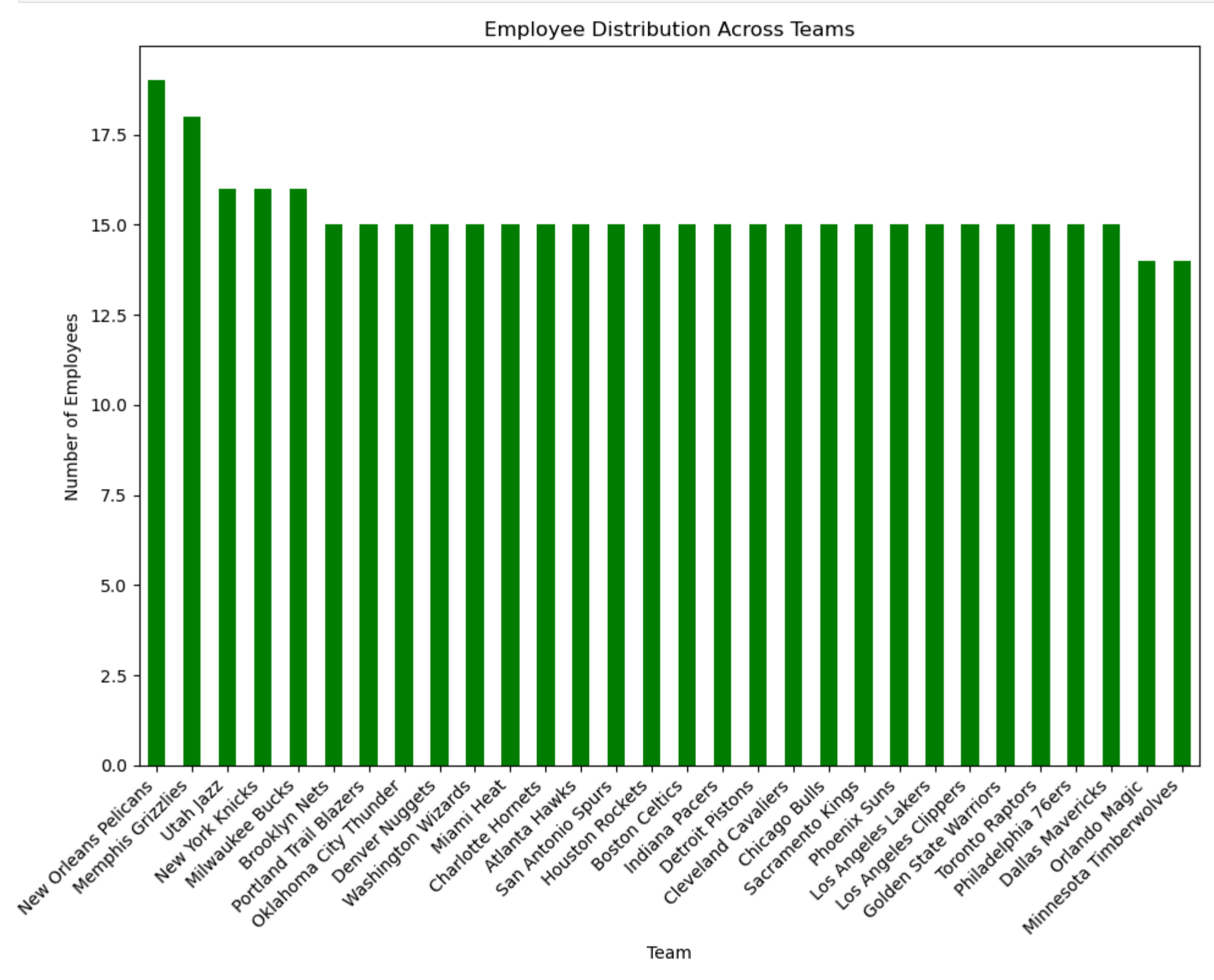
In [133…
```python
# 1. Employee Distribution Across Teams
team_distribution = data['Team'].value_counts()
team_percentage = (team_distribution / len(data)) * 100
```

In [135…
```python
print("\nEmployee Distribution Across Teams:")
print(team_distribution)
print("\nPercentage Split:")
print(team_percentage)
```

```
Employee Distribution Across Teams:
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

Percentage Split:
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 [137…
```python
plt.figure(figsize=(10, 8))
team_distribution.plot(kind='bar', color='Green')
plt.title('Employee Distribution Across Teams')
plt.xlabel('Team')
plt.ylabel('Number of Employees')
plt.xticks(rotation=45, ha='right') # Rotate x-axis labels for readability
plt.tight_layout() # Adjust layout to prevent labels from overlapping
plt.show()
```

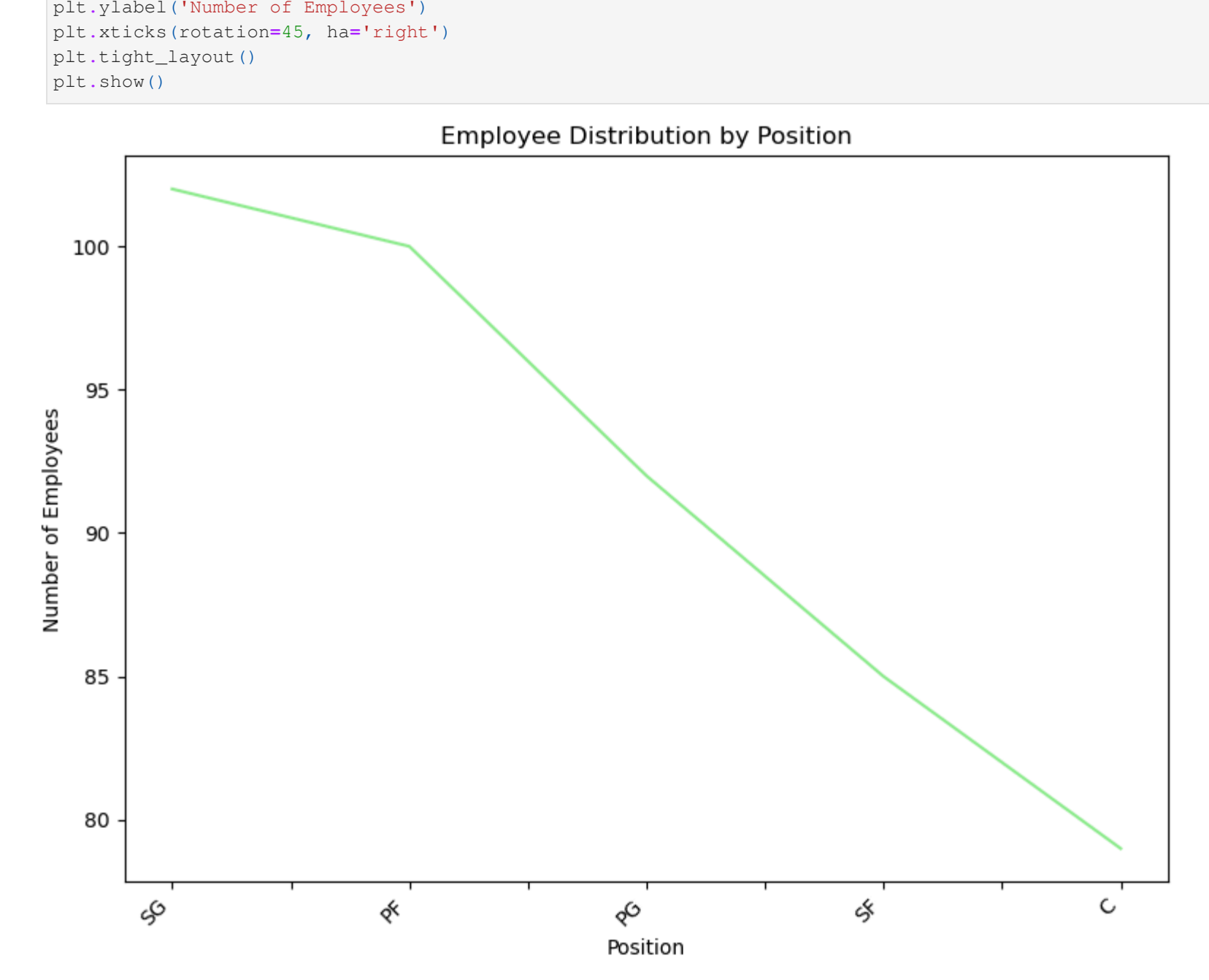


In [138…
```python
# 2. Employee Segregation by Position
position_distribution = data['Position'].value_counts()
```

In [141…
```python
print("\nEmployee Segregation by Position:")
print(position_distribution)
```

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

In [143…
```python
plt.figure(figsize=(8, 6))
position_distribution.plot(kind='line', color='lightgreen')
plt.title('Employee Distribution by Position')
plt.xlabel('Position')
plt.ylabel('Number of Employees')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



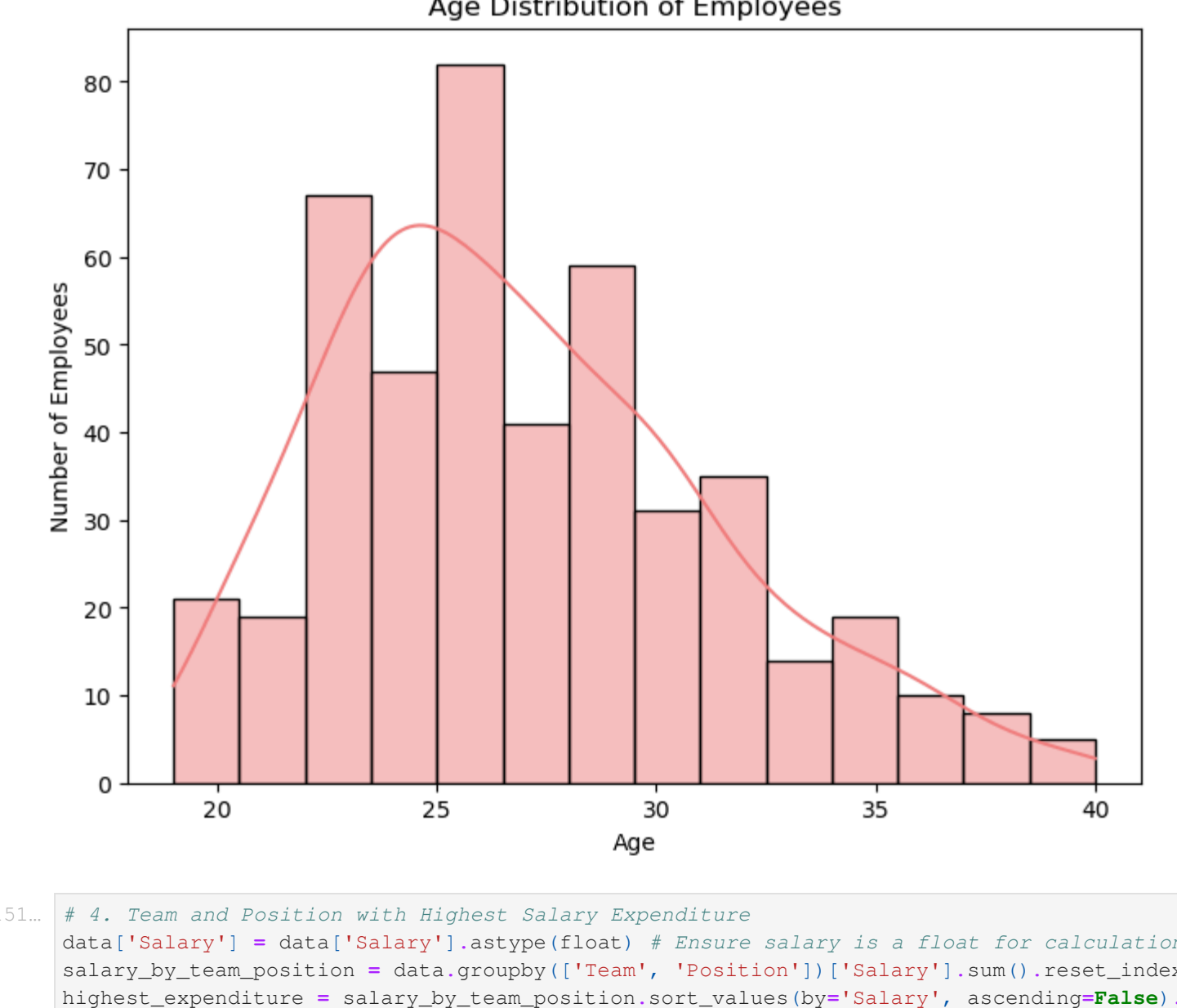
In [145…
```python
# 3. Predominant Age Group
age_group = data['Age'].describe()   # Use describe() for summary stats including median
median_age = data['Age'].median()
```

In [147…
```python
print("\nAge Statistics:")
print(age_group)
print(f"\nMedian Age: {median_age}")  # Median is a better measure of central tendency than mean if data is skewed.
```

```
Age Statistics:
count    458.000000
mean      26.934498
std        4.400128
min       19.000000
25%       24.000000
50%       26.000000
75%       30.000000
max       40.000000
Name: Age, dtype: float64

Median Age: 26.0
```

In [149…
```python
plt.figure(figsize=(8, 6))
sns.histplot(data['Age'], kde=True, color='lightcoral') # Histogram with KDE for better visualization of distribution
plt.title('Age Distribution of Employees')
plt.xlabel('Age')
plt.ylabel('Number of Employees')
plt.show()
```



In [151…
```python
# 4. Team and Position with Highest Salary Expenditure
data['Salary'] = data['Salary'].astype(float) # Ensure salary is a float for calculations
salary_by_team_position = data.groupby(['Team', 'Position'])['Salary'].sum().reset_index()
highest_expenditure = salary_by_team_position.sort_values(by='Salary', ascending=False).iloc[0]
```

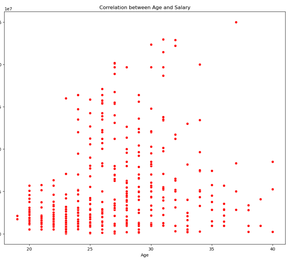
In [153…
```python
print("\nTeam and Position with Highest Salary Expenditure:")
print(highest_expenditure)
```

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

In [155…
```python
# 5. Correlation between Age and Salary
correlation = data['Age'].corr(data['Salary'])
print(f"\nCorrelation between Age and Salary: {correlation}")
```

```
Correlation between Age and Salary: 0.20912419115196068
```

In [161…
```python
plt.figure(figsize=(12, 10))
sns.scatterplot(x='Age', y='Salary', data=data, color='red')
plt.title('Correlation between Age and Salary')
plt.xlabel('Age')
plt.ylabel('Salary')
plt.show()
```


Correlation between Age and Salary
1e7
2.5
2.0
1.5
1.0
0.5
0.0
Salary
20
25
30
35
40
Age