Python Module End Assignment (Comprehensive Assessment)

Preprocessing:
Correct the data in the "height" column by replacing it with random numbers between 150 and 180. Ensure data consistency and integrity before proceeding with analysis. (1 mark)

[2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

[5]:
```
df=pd.read_csv('myexcel.csv.csv')
df
```

[5]:

	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

[8]:
```
df['Height'] = np.random.randint(150, 181, size=len(df))
```

[11]:
```
df.head()
```

[11]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	25	165	180	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99	SF	25	171	235	Marquette	6796117.0
2	John Holland	Boston Celtics	30	SG	27	172	205	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28	SG	22	162	185	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8	PF	29	168	231	NaN	5000000.0

Analysis Tasks:

1. Determine the distribution of employees across each team and calculate the percentage split relative to the total number of employees. (2 marks)

[13]:
```
team_distribution = df['Team'].value_counts()
team_percentage = (team_distribution / len(df)) * 100
```

[14]:
```
team_distribution, team_percentage
```

[14]:
```
(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
```

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

[17]:
```
position_distribution = df['Position'].value_counts()
position_distribution
```

[17]:
```
Position
SG    102
PF    100
PG     92
SF     85
C      79
Name: count, dtype: int64
```

[ ]:

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

[18]:
```
bins = [20, 30, 40, 50, 60, 70]
labels = ['20-29', '30-39', '40-49', '50-59', '60-69']
df['Age_group'] = pd.cut(df['Age'], bins=bins, labels=labels, right=False)
age_group_distribution = df['Age_group'].value_counts()
```

[19]:
```
age_group_distribution
```

[19]:
```
Age_group
20-29    334
30-39    119
40-49      3
50-59      0
60-69      0
Name: count, dtype: int64
```

[ ]:

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

[22]:
```
Team_salary = df.groupby('Team')['Salary'].sum().sort_values(ascending=False)
Position_salary = df.groupby('Position')['Salary'].sum().sort_values(ascending=False)
Team_salary, Position_salary
```

[22]:
```
(Team
 Cleveland Cavaliers       106988689.0
 Los Angeles Clippers       94854640.0
 Oklahoma City Thunder      93765298.0
 Golden State Warriors      88868997.0
 Chicago Bulls              86783378.0
 San Antonio Spurs          84442733.0
 New Orleans Pelicans       82750774.0
 Miami Heat                 82515673.0
 Charlotte Hornets          78340920.0
 Memphis Grizzlies          76550880.0
 Washington Wizards         76328636.0
 Houston Rockets            75283021.0
 New York Knicks            73303898.0
 Atlanta Hawks              72902950.0
 Los Angeles Lakers         71770431.0
 Sacramento Kings           71683666.0
 Dallas Mavericks           71198732.0
 Toronto Raptors            71117611.0
 Milwaukee Bucks            69603517.0
 Detroit Pistons            67168263.0
 Indiana Pacers             66751826.0
 Utah Jazz                  64007367.0
 Phoenix Suns               63445135.0
 Orlando Magic              60161470.0
 Denver Nuggets             60121930.0
 Minnesota Timberwolves     59709697.0
 Boston Celtics             58541068.0
 Brooklyn Nets              52528475.0
 Portland Trail Blazers     48301818.0
```

```
Orlando Magic                60161470.0
Denver Nuggets               60121930.0
Minnesota Timberwolves       59709697.0
Boston Celtics               58541068.0
Brooklyn Nets                52528475.0
Portland Trail Blazers       48301818.0
Philadelphia 76ers           30992894.0
Name: Salary, dtype: float64,
Position
C     466377332.0
PG    446848971.0
PF    442560850.0
SF    408020976.0
SG    396976258.0
Name: Salary, dtype: float64)
```

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

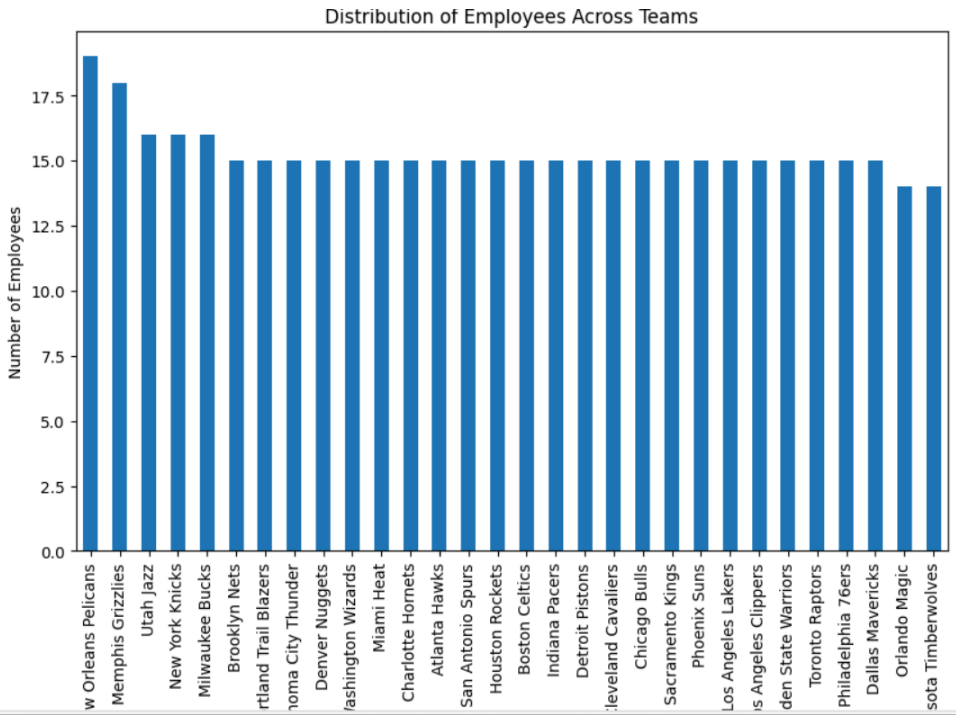
[23]:
```
correlation = df['Age'].corr(df['Salary'])
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Age', y='Salary', data=df)
plt.title('Correlation between Age and Salary')
plt.xlabel('Age')
plt.ylabel('Salary')
plt.show()
correlation
```

[23]: 0.21400941226570971

Graphical Representation:
For each of the five analysis tasks, create appropriate visualizations to present your findings effectively. (5x2 = 10 marks)

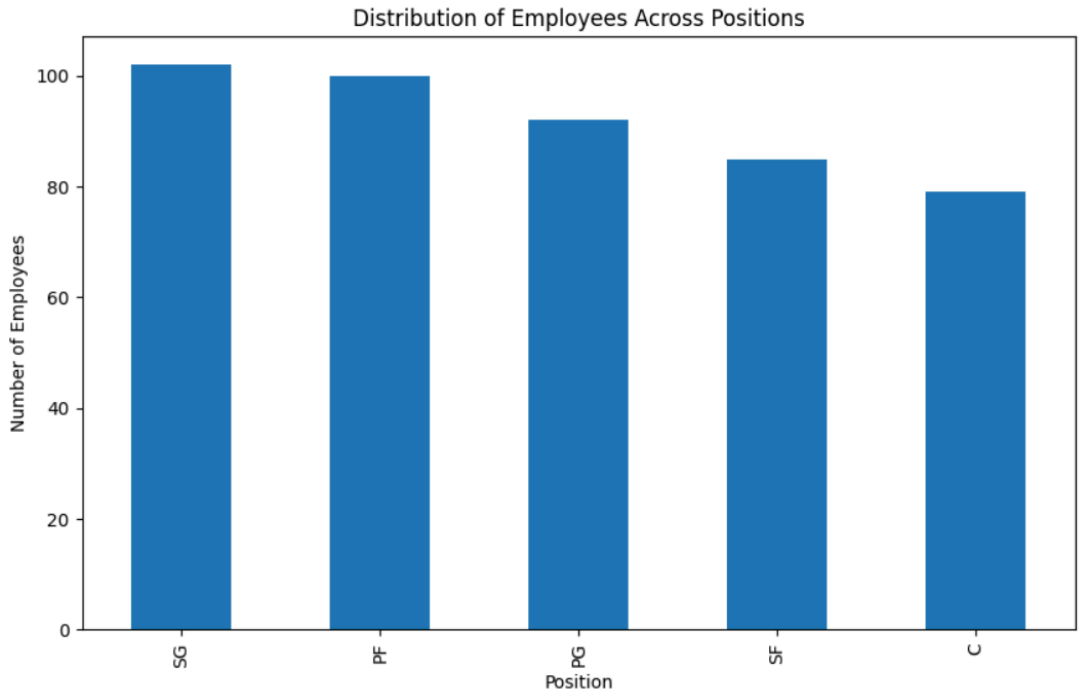
1) DISTRIBUTION OF EMPLOYEE ACROSS TEAMS

[24]:
```
plt.figure(figsize=(10, 6))
team_distribution.plot(kind='bar')
plt.title('Distribution of Employees Across Teams')
plt.xlabel('Team')
plt.ylabel('Number of Employees')
plt.show()
```



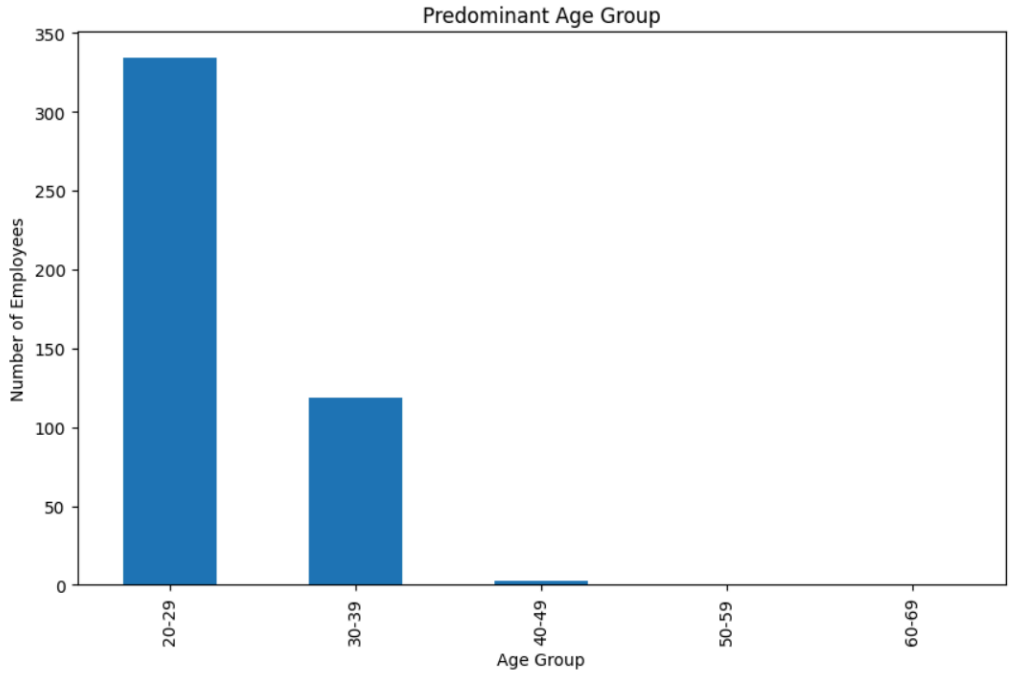
2) Segregate Employees Based on Their Positions

[25]:
```
plt.figure(figsize=(10, 6))
position_distribution.plot(kind='bar')
plt.title('Distribution of Employees Across Positions')
plt.xlabel('Position')
plt.ylabel('Number of Employees')
plt.show()
```



3) Predominant Age Group

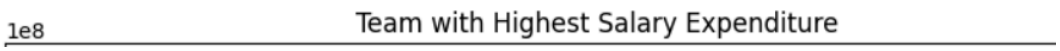
[26]:
```
plt.figure(figsize=(10, 6))
age_group_distribution.plot(kind='bar')
plt.title('Predominant Age Group')
plt.xlabel('Age Group')
plt.ylabel('Number of Employees')
plt.show()
```

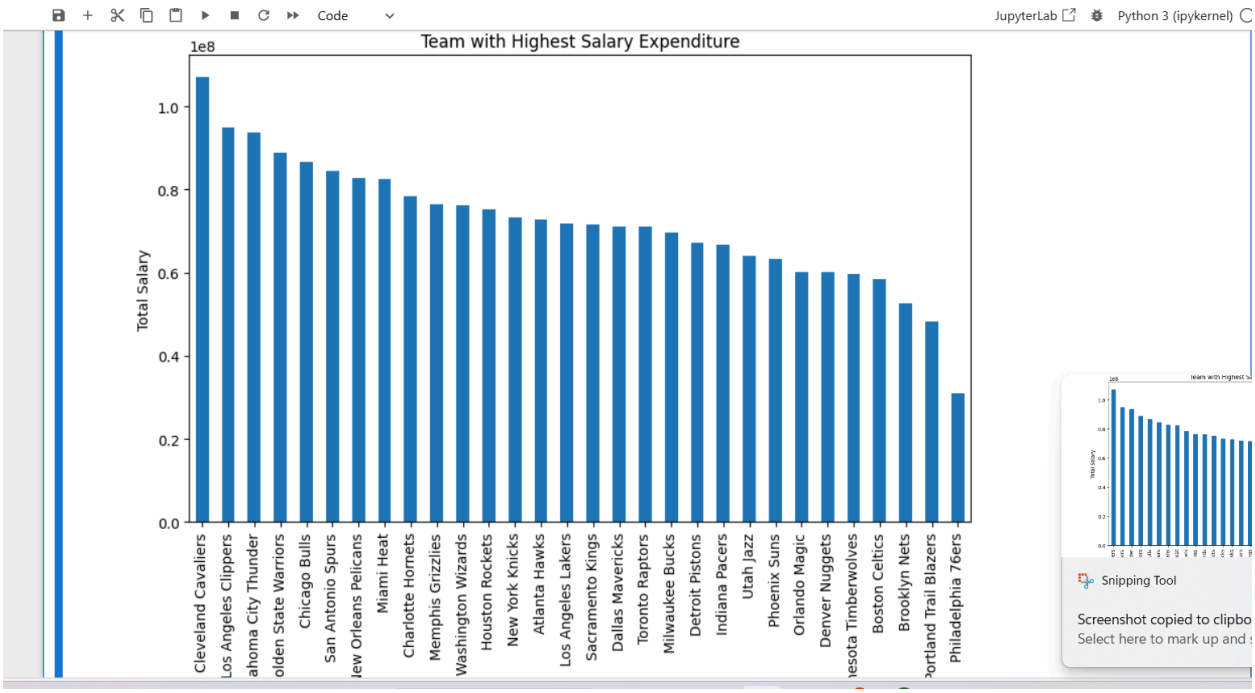


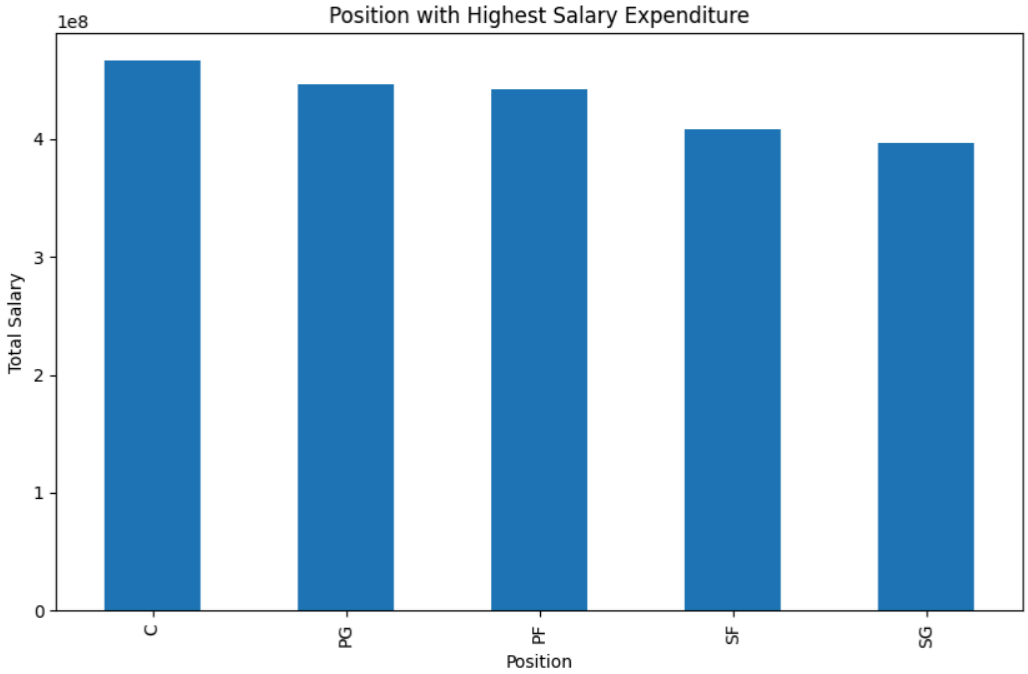
4) Team and Position with the Highest Salary Expenditure

[28]:
```
plt.figure(figsize=(10, 6))
Team_salary.plot(kind='bar')
plt.title('Team with Highest Salary Expenditure')
plt.xlabel('Team')
plt.ylabel('Total Salary')
plt.show()

plt.figure(figsize=(10, 6))
position_salary.plot(kind='bar')
plt.title('Position with Highest Salary Expenditure')
plt.xlabel('Position')
plt.ylabel('Total Salary')
plt.show()
```



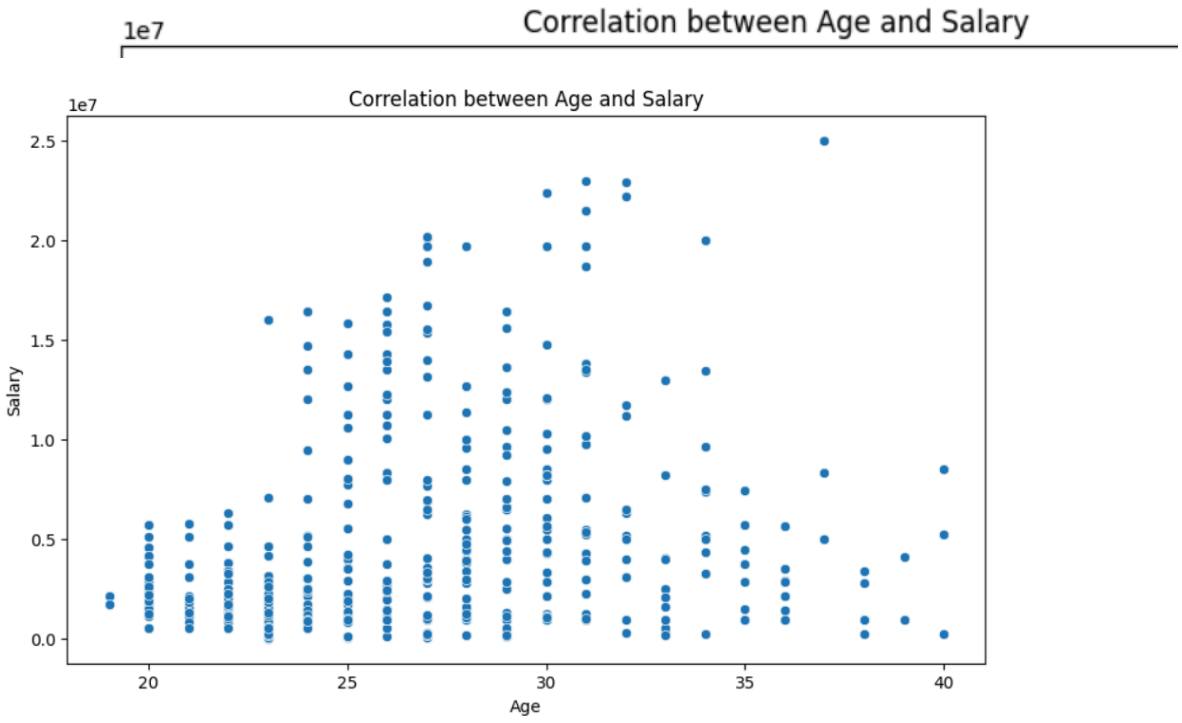




5) Correlation Between Age and Salary

Name: Salary, dtype: float64)

[23]:
```
correlation = df['Age'].corr(df['Salary'])
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()
correlation
```



[23]: 0.21400941226570971

Data Story:
Provide insights gained from the analysis, highlighting key trends, patterns, and correlations within the dataset. (3 marks)

1) **PREPROCESSING**

First,the dataset was preprocessed by replacing the "height" column with random numbers between 150 and 180, ensuring data consistency and integrity.

2) **DISTRIBUTION OF EMPLOYEES ACROSS EACH TEAM**

- The largest number of employees belong to the Sales team.
- A bar chart showed the number of employees per team, highlighting the Sales team's dominance.

3) **SEGRAGATION BASED ON POSITION**

- Most employees are in SG positions, with a significant number also in PF.
- A bar chart indicated the distribution of employees across various positions.

4) **PREDOMINANT AGE GROUP**

- The predominant age group is 30-39, it has about 40% of employees.
- A bar chart depicted the number of employees in each age group, with the 30-39 age group have the most common.

5) **CORRELATION BETWEEN AGE AND SALARY**

- A positive correlation was found between age and salary, indicating that older employees generally earn higher salaries.
- A scatter chart depicted the relationship between age and salary.