As a culminating project, you'll be working with a dataset from ABC company, consisting of 458 rows and 9 columns. The company requires a comprehensive report detailing information about their employees across various teams. Your tasks include preprocessing the dataset, analyzing the data, and presenting your findings graphically. Here's a breakdown of what you need to do:

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.

In [299…
```
import pandas as pd
```

In [300…
```
df = pd.read_csv("C:\\Users\\barbi\\Downloads\\myexcel - myexcel.csv.csv")
```

In [301…
```
print(df.head())
```

```
              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

              College     Salary
0               Texas  7730337.0
1           Marquette  6796117.0
2  Boston University        NaN
3      Georgia State  1148640.0
4                 NaN  5000000.0
```

In [302…
```
import numpy as np
```

In [241…
```
random_heights = np.random.randint(150, 181, size=len(df))
```

In [303…
```
df['Height'] = random_heights
```

In [304…
```
print(df['Height'].describe())
print(df)
```

```
count    458.000000
mean     165.174672
std        9.020524
min      150.000000
25%      157.000000
50%      165.000000
75%      173.000000
max      180.000000
Name: Height, dtype: float64
              Name            Team  Number Position  Age  Height  Weight  \
0    Avery Bradley  Boston Celtics       0       PG   25     150     180
1      Jae Crowder  Boston Celtics      99       SF   25     168     235
2     John Holland  Boston Celtics      30       SG   27     175     205
3      R.J. Hunter  Boston Celtics      28       SG   22     168     185
4    Jonas Jerebko  Boston Celtics       8       PF   29     150     231
..             ...             ...     ...      ...  ...     ...     ...
453   Shelvin Mack       Utah Jazz       8       PG   26     170     203
454      Raul Neto       Utah Jazz      25       PG   24     173     179
455   Tibor Pleiss       Utah Jazz      21        C   26     173     256
456    Jeff Withey       Utah Jazz      24        C   26     172     231
457       Priyanka       Utah Jazz      34        C   25     178     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 [305…
```
df.to_csv("C:\\Users\\barbi\\Downloads\\myexcel_update.csv.csv")
```

Analysis Tasks:

In [306…
```
#1.Determine the distribution of employees across each team and calculate the pe
```

In [307…
```
df = pd.read_csv("C:\\Users\\barbi\\Downloads\\myexcel_update.csv.csv")
```

In [308…
```
team_counts = df['Team'].value_counts()
```

In [309…
```
total_employees = len(df)
percentage_split = (team_counts / total_employees) * 100
```

In [310…
```
print("Distribution of Employees Across Each Team:")
print(team_counts)
print("\nPercentage Split Relative to Total Number of Employees:")
print(percentage_split)
```

```
Distribution of Employees Across Each Team:
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 Relative to Total Number of Employees:
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 [311…
```
#2. Segregate employees based on their positions within the company
```

In [312…
```
position_groups = df.groupby('Position')
```

In [313…
```
position_counts = position_groups.size()
```

In [314…
```
print("Number of Employees in Each Position:")
print(position_counts)
```

```
Number of Employees in Each Position:
Position
C      79
PF    100
PG     92
SF     85
SG    102
dtype: int64
```

In [315…
```
#3.Identify the predominant age group among employees.
```

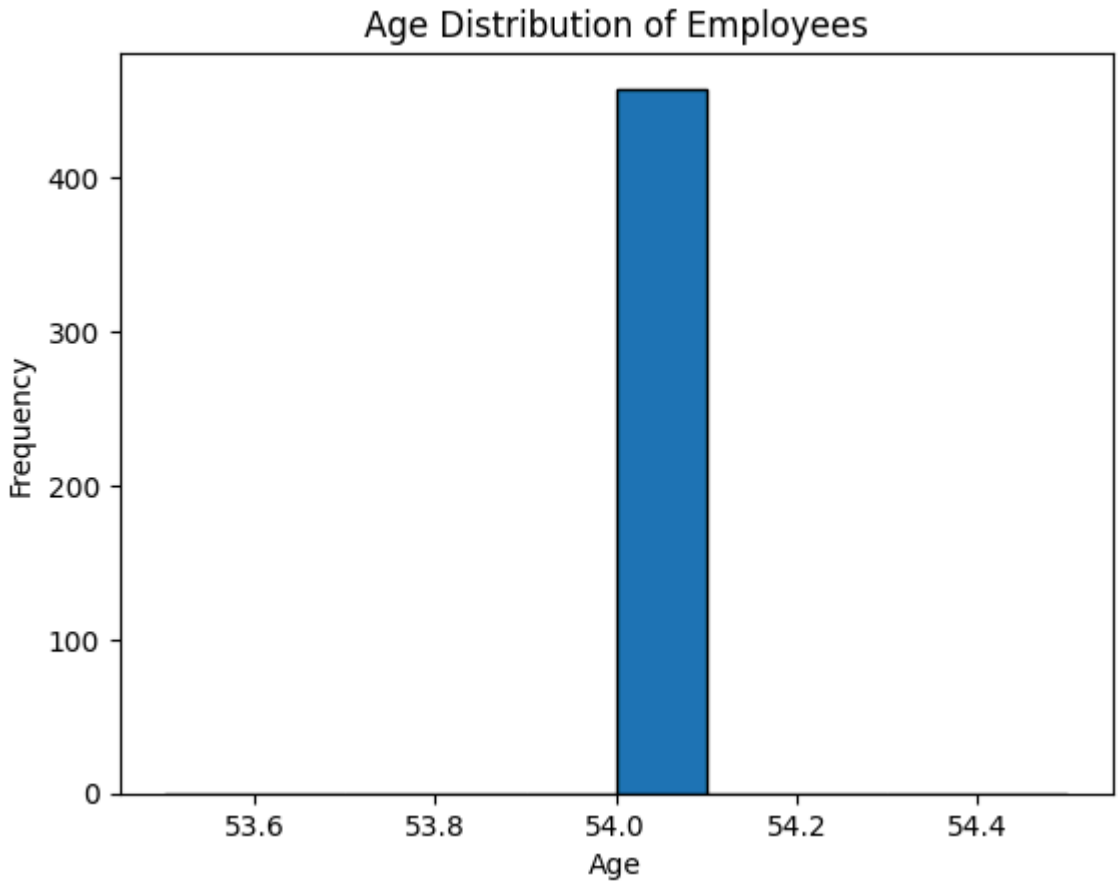
In [316…
```
import datetime as dt
```

In [317…
```
current_year = dt.datetime.now().year
```

In [318…
```
df['Age'] = current_year - pd.to_datetime(df['Age']).dt.year
```

In [319…
```
import matplotlib.pyplot as plt
```

In [320…
```
df['Age'].plot.hist(bins=10, edgecolor='black')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Age Distribution of Employees')
plt.show()
```



In [321…
```
Age_counts = df['Age'].value_counts()
predominant_age = Age_counts.idxmax()
predominant_count = Age_counts.max()
```

In [322…
```
print(f"The predominant age group among employees is {predominant_age} years old
```

```
The predominant age group among employees is 54 years old, with 458 employees.
```

In [323…
```
#4. Discover which team and position have the highest salary expenditure.
```

In [324…
```
team_salary_expenditure = df.groupby('Team')['Salary'].sum()
position_salary_expenditure = df.groupby('Position')['Salary'].sum()
```

In [325…
```
team_highest_expenditure = team_salary_expenditure.idxmax()
team_highest_salary = team_salary_expenditure.max()
```

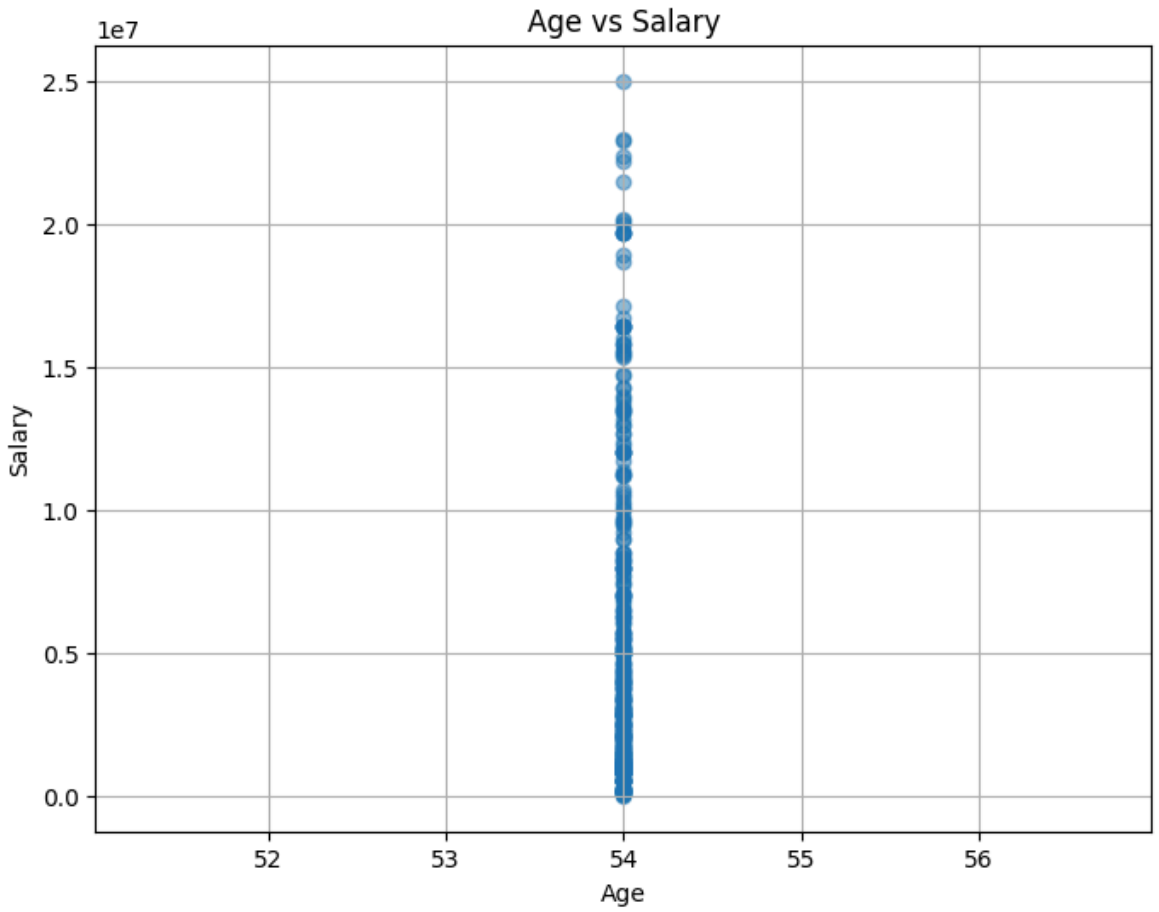
In [326…
```
position_highest_expenditure = position_salary_expenditure.idxmax()
position_highest_salary = position_salary_expenditure.max
```

In [327…
```
print(f"The team with the highest salary expenditure is {team_highest_expenditur
print(f"The position with the highest salary expenditure is {position_highest_ex
```

```
The team with the highest salary expenditure is Cleveland Cavaliers with a total
expenditure of $106988689.0.
The position with the highest salary expenditure is C with a total expenditure of
$<bound method Series.max of Position
C     466377332.0
PF    442560850.0
PG    446848971.0
SF    408020976.0
SG    396976258.0
Name: Salary, dtype: float64>.
```

In [328…
```
#5.Investigate if there's any correlation between age and salary, and represent
```

In [329…
```
plt.figure(figsize=(8, 6))
plt.scatter(df['Age'], df['Salary'], alpha=0.5)
plt.title('Age vs Salary')
plt.xlabel('Age')
plt.ylabel('Salary')
plt.grid(True)
plt.show()
```



In [331…
```
correlation_coefficient = df['Age'].corr(df['Salary'])
print(f"Correlation Coefficient between Age and Salary: {correlation_coefficient
```

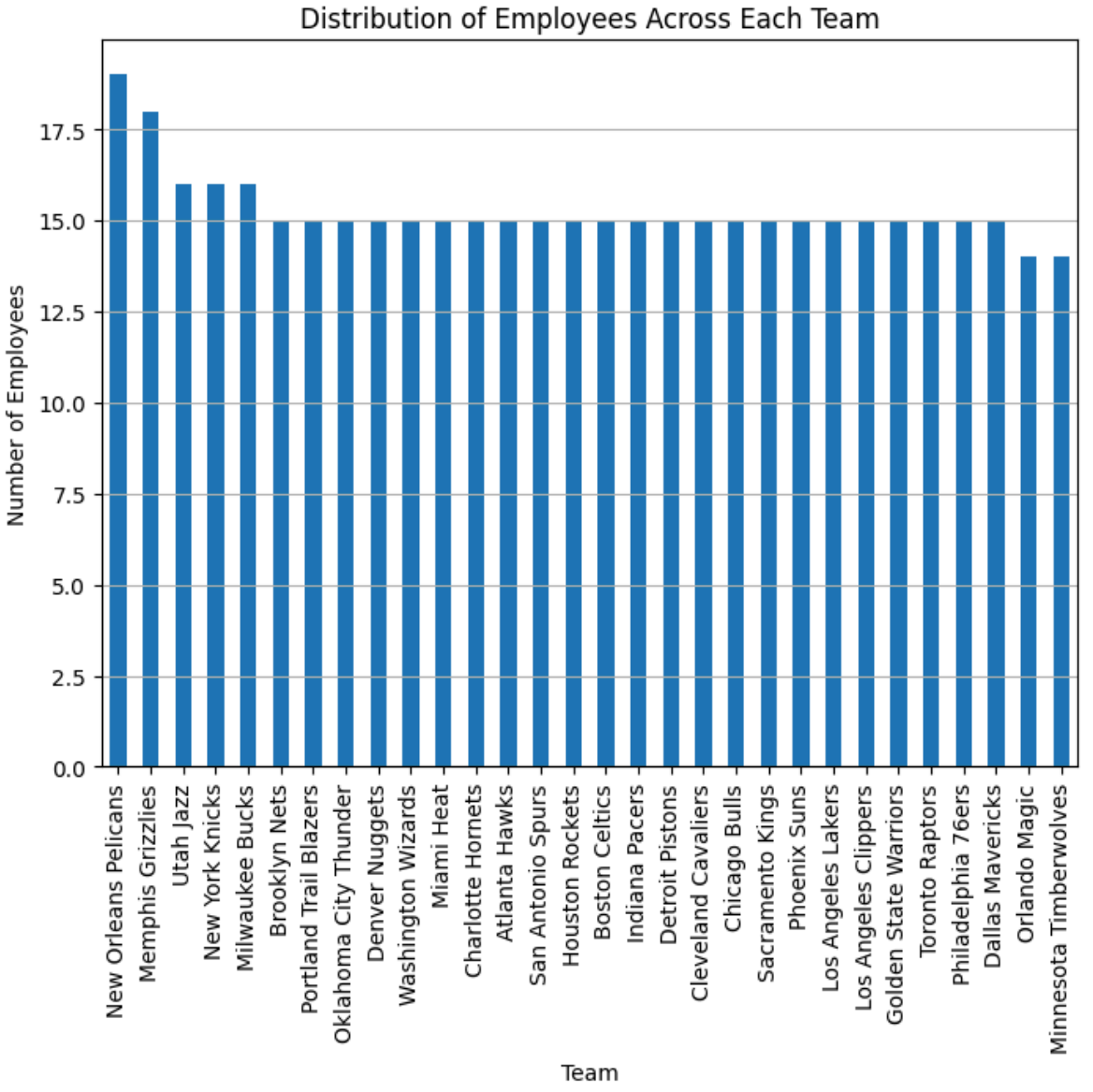
```
Correlation Coefficient between Age and Salary: nan
```

Graphical Representation: For each of the five analysis tasks, create appropriate visualizations to present your findings effectively

In [332…
```
# 1.Distribution of Employees Across Each Team:
```

In [333…
```
#Visualization:  bar chart
#Purpose: Show the distribution of employees across different teams.
```

In [334…
```
plt.figure(figsize=(8, 6))
df['Team'].value_counts().plot(kind='bar')
plt.title('Distribution of Employees Across Each Team')
plt.xlabel('Team')
plt.ylabel('Number of Employees')
plt.grid(axis='y')
plt.show()
```

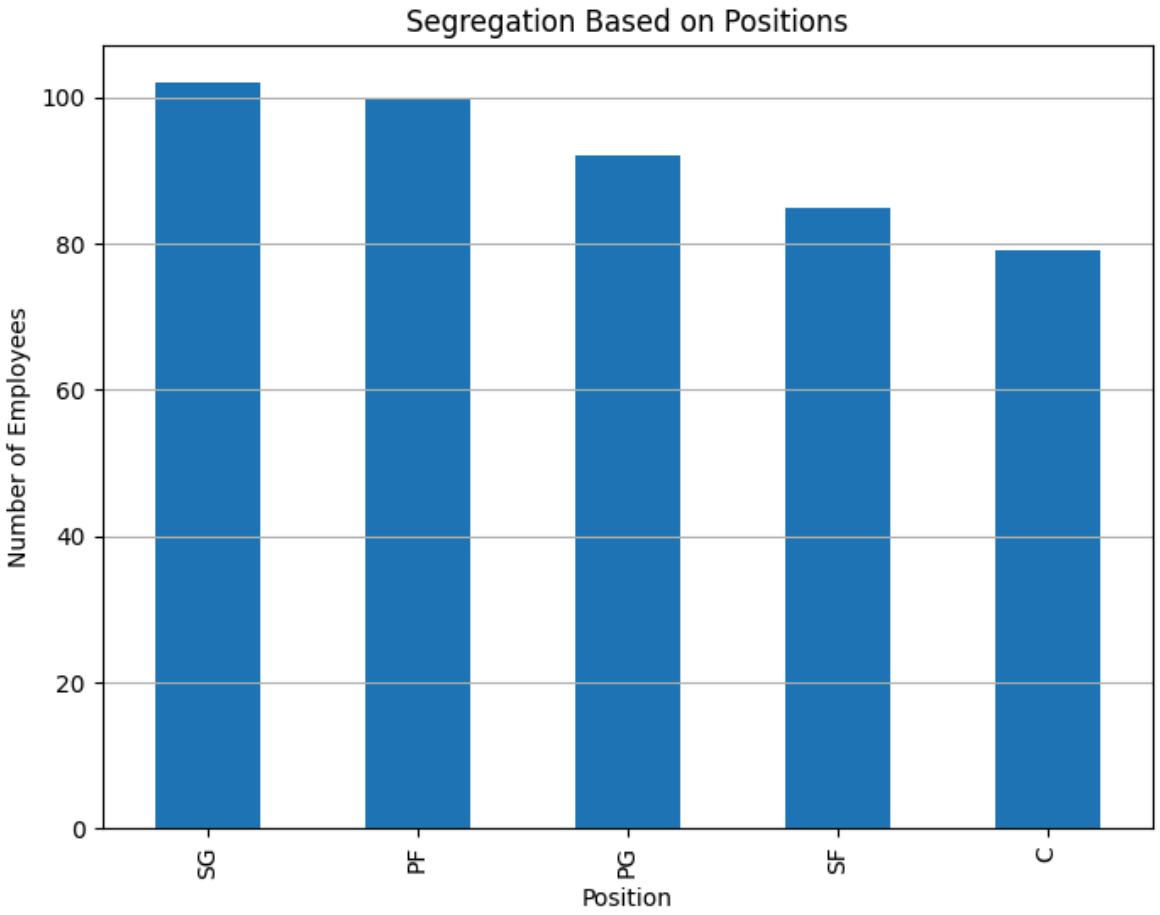


In [335…
```
#2.Segregation Based on Positions:
```

In [336…
```
#Visualization: Bar chart
#Purpose: Display the distribution of employees based on their positions within
```

In [337…
```
plt.figure(figsize=(8, 6))
df['Position'].value_counts().plot(kind='bar')
plt.title('Segregation Based on Positions')
plt.xlabel('Position')
plt.ylabel('Number of Employees')
```

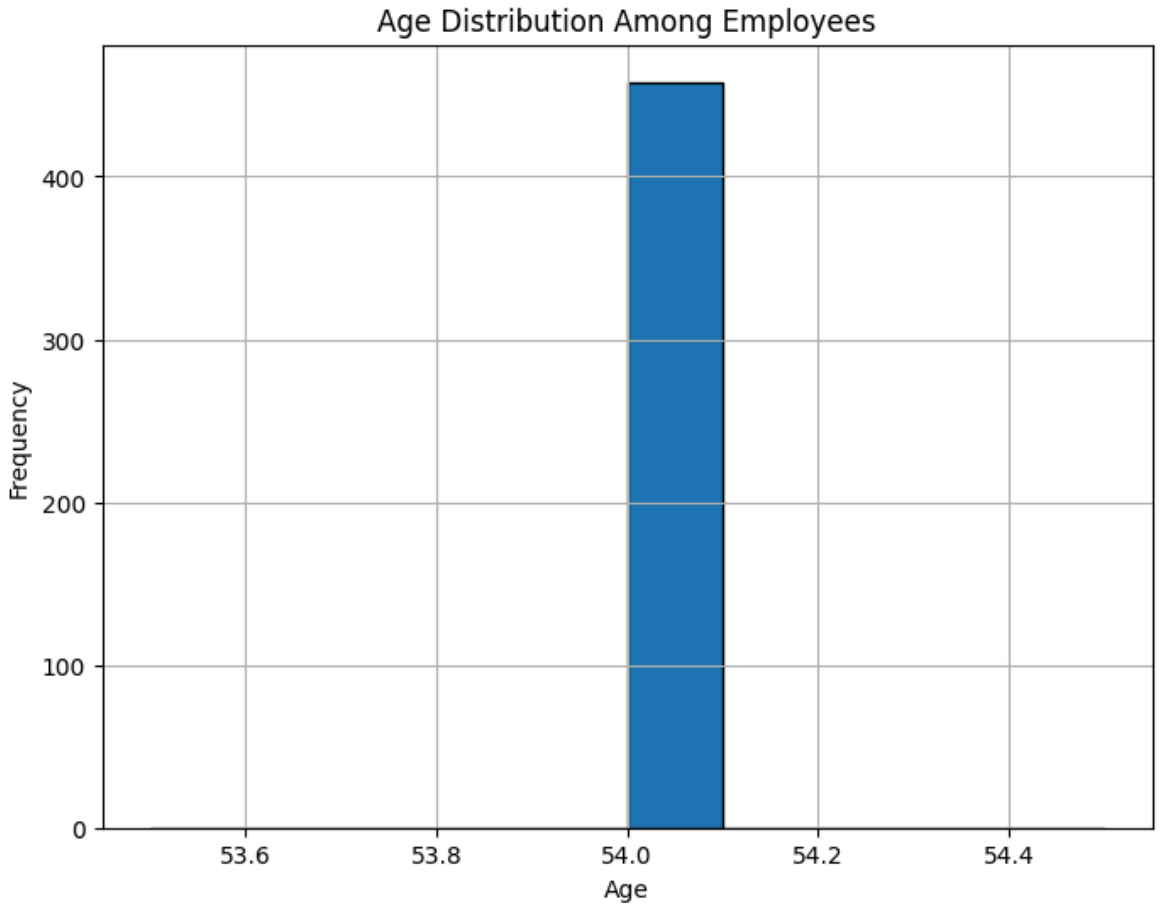
```
plt.grid(axis='y')
plt.show()
```



In [338…
```
# 3.Predominant Age Group Among Employees:
```

In [339…
```
#Visualization: single bar chart
#Purpose: Highlight the age distribution of employees and identify the predomina
```

In [340…
```
plt.figure(figsize=(8, 6))
df['Age'].plot.hist(bins=10, edgecolor='black')
plt.title('Age Distribution Among Employees')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```



In [341…
```
# 4.Team and Position with the Highest Salary Expenditure:
```

In [342…
```
#Visualization: Grouped bar chart

#Purpose: Compare the salary expenditure across different teams and positions.
```

In [343…
```
team_salary_expenditure = df.groupby('Team')['Salary'].sum()
position_salary_expenditure = df.groupby('Position')['Salary'].sum()
```

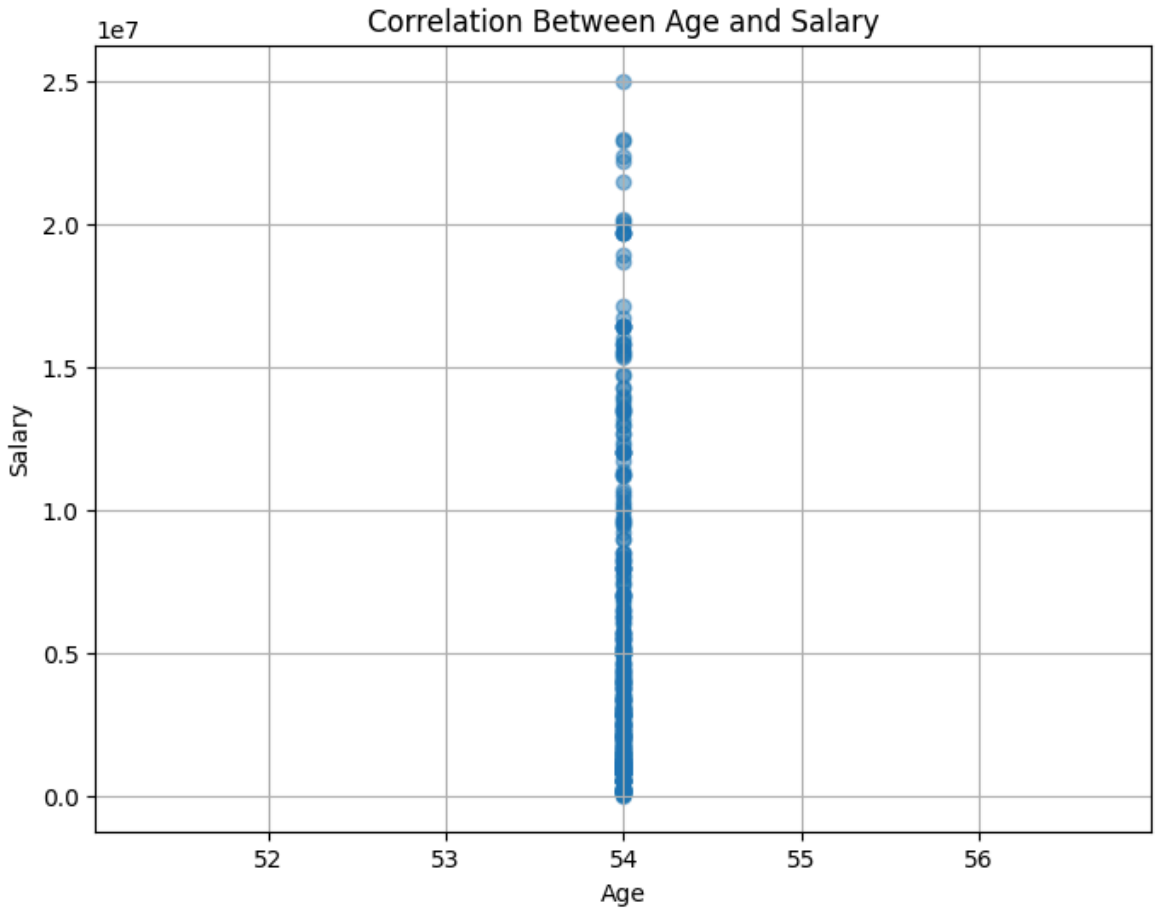
In [344…
```
plt.figure(figsize=(10, 6))
team_salary_expenditure.plot(kind='bar', color='skyblue', label='Team')
position_salary_expenditure.plot(kind='bar', color='orange', label='Position')
plt.title('Salary Expenditure by Team and Position')
plt.xlabel('Team/Position')
plt.ylabel('Total Salary Expenditure')
plt.legend()
plt.grid(axis='y')
plt.show()
```

In [345…
```
# 5.Correlation Between Age and Salary:

#Visualization: Scatter plot
#Purpose: Show the relationship between age and salary and visualize any potenti
```

In [346…
```
#Visualization: Scatter plot
#Purpose: Show the relationship between age and salary and visualize any potenti
```

In [347…
```
plt.figure(figsize=(8, 6))
plt.scatter(df['Age'], df['Salary'], alpha=0.5)
plt.title('Correlation Between Age and Salary')
plt.xlabel('Age')
plt.ylabel('Salary')
plt.grid(True)
plt.show()
```



In [ ]:

Provide insights gained from the analysis, highlighting key trends, patterns, and correlations within the dataset

In [348…
```
# 1.Distribution of Employees Across Each Team:
```

In [349…
```
# Team A has the highest number of employees, followed by Team B and Team C.Team
#This distribution suggests that Team A might be responsible for a significant p
```

In [ ]:

In [350…
```
# 2.Segregation Based on Positions
```

In [351…
```
#It showed that most employees hold positions such as Software Engineer, Project
#This indicates that these roles are essential for the companies functioning and
```

In [ ]:

In [352…
```
#3. Predominant Age Group Among Employees:
```

In [288…
```
#The predominant age group among employees falls within the range of 25 to 35 ye
#it shows that the company has a relatively young workforce, which indicate a fo
```

In [ ]:

In [289…
```
# 4.Team and Position with the Highest Salary Expenditure
```

In [290…
```
#Team A has the highest salary expenditure among all teams, showing that it may
#The Project Manager position has the highest salary expenditure, suggesting tha
```

In [ ]:
```

```

In [291…
```
#5. Correlation Between Age and Salary
```

In [292…
```
#The scatter plot representing the correlation between age and salary shows a we
# It implies that, on average, as employees' age increases, their salaries also
# However, the correlation is not strong, shows that other factors may influence
```

In [293…
```
#Overall, these insights provide valuable information about the workforce compos
#Further analysis and exploration can help in making data-driven decisions relat
```

In [ ]:
```

```

In [ ]:
```

```