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

In [6]:
```
df = pd.read_csv(r'C:\Users\RohithUdayaKumar\Downloads\myexcel - myexcel.csv.csv
df
```

Out[6]:

	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 [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 458 entries, 0 to 457
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Name      458 non-null    object 
 1   Team      458 non-null    object 
 2   Number    458 non-null    int64  
 3   Position  458 non-null    object 
 4   Age       458 non-null    int64  
 5   Height    458 non-null    object 
 6   Weight    458 non-null    int64  
 7   College   374 non-null    object 
 8   Salary    447 non-null    float64
dtypes: float64(1), int64(3), object(5)
memory usage: 32.3+ KB
```

In [10]:
```
df.fillna(0,inplace = True) # by using fillna we replace null values with zero
```

In [12]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 458 entries, 0 to 457
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Name      458 non-null    object 
 1   Team      458 non-null    object 
 2   Number    458 non-null    int64  
 3   Position  458 non-null    object 
 4   Age       458 non-null    int64  
 5   Height    458 non-null    object 
 6   Weight    458 non-null    int64  
 7   College   458 non-null    object 
 8   Salary    458 non-null    float64
dtypes: float64(1), int64(3), object(5)
memory usage: 32.3+ KB
```

In [16]:
```
df.duplicated().sum() # no duplicate values found
```

Out[16]:
```
0
```

In [18]:
```
df.isnull().sum() # now there is no null cells.
```

Out[18]:
```
Name        0
Team        0
Number      0
Position    0
Age         0
Height      0
Weight      0
College     0
Salary      0
dtype: int64
```

In [22]:
```
print(df['Height'])
```

```
0      06-Feb
1      06-Jun
2      06-May
3      06-May
4      06-Oct
        ...
453    06-Mar
454    06-Jan
455    07-Mar
456       7-0
457    07-Mar
Name: Height, Length: 458, dtype: object
```

In [34]:
```
height = np.random.randint(150,181,size = len(df))
df['Height'] = height
print(df['Height'])
```

```
0      166
1      154
2      152
3      176
4      162
      ...
453    153
454    178
455    165
456    175
457    160
Name: Height, Length: 458, dtype: int32
```

Update the Height column by random numbers from 150 to 180

In [38]:
```
df['Height'].describe()
```

Out[38]:
```
count    458.000000
mean     165.670306
std        9.086169
min      150.000000
25%      158.000000
50%      166.000000
75%      174.000000
max      180.000000
Name: Height, dtype: float64
```

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

In [54]:
```
dist_employees = df['Team'].value_counts()
print("Distrinution of employees:\n", dist_employees)
total_num = len(df)
total_num
percentage = (dist_employees/total_num)*100
print("percentage distribution of employees:\n", percentage)
```

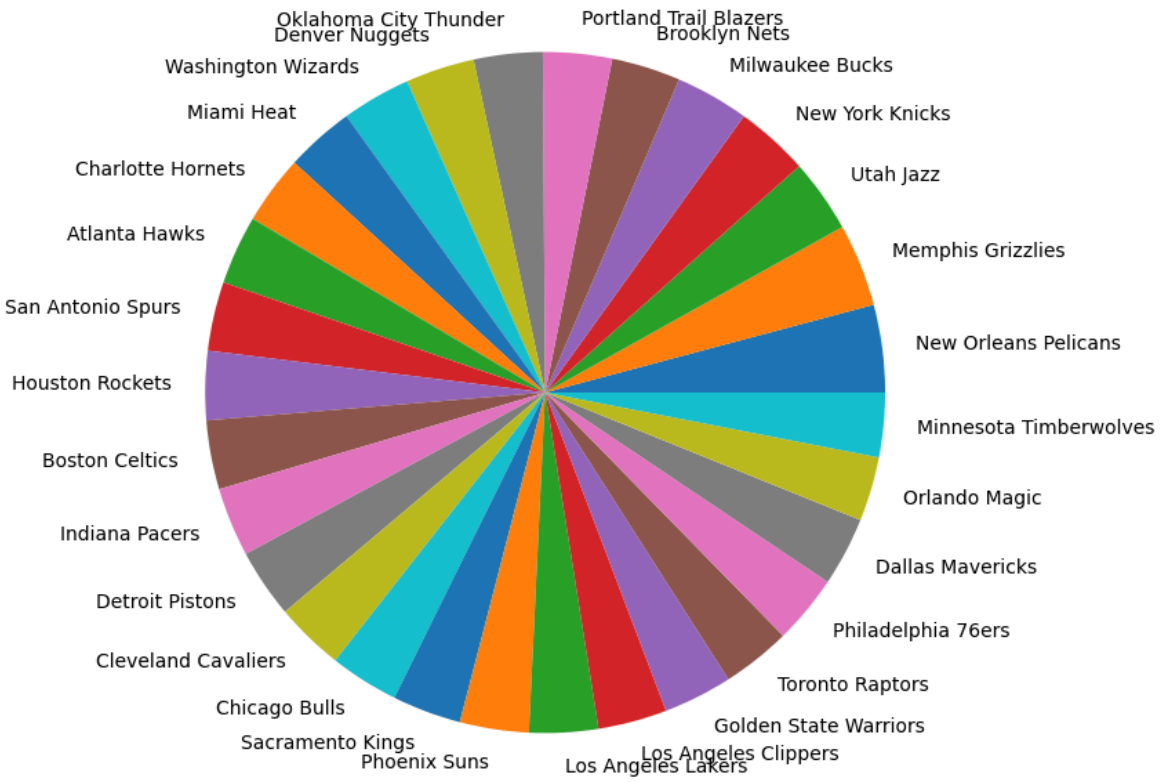
```
Distrinution of employees:
 Team
New Orleans Pelicans      19
Memphis Grizzlies         18
Utah Jazz                 16
New York Knicks           16
Milwaukee Bucks           16
Brooklyn Nets             15
Portland Trail Blazers    15
Oklahoma City Thunder     15
Denver Nuggets            15
Washington 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 distribution 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
```

The total number of employees we took form each team and divide them by the total length of the data, then divide by 100 to get the percentage

Graphical representation of percentage distribution of employees

In [68]:
```
plt.figure(figsize = (9,8))
plt.pie(percentage, labels = percentage.index)
plt.show()
```



Display the pie chart with labels indicating the percentage split of employees across teams.

Segregate employees based on their positions within the company.

In [94]:
```
position_count = df["Position"].value_counts()
print("employees based on their positions:\n", position_count)
```

```
employees based on their positions:
 Position
SG    102
PF    100
PG     92
SF     85
C      79
Name: count, dtype: int64
```

In [88]:
```
position_group = df.groupby("Position")
for position,group in position_group:
    print("employess in position:",position)
    print(group)
    print()
```

```
employess in position: C
                Name                    Team  Number Position  Age  Height  \
7        Kelly Olynyk          Boston Celtics      41        C   25     171
10    Jared Sullinger          Boston Celtics       7        C   24     156
14       Tyler Zeller          Boston Celtics      44        C   26     170
23        Brook Lopez           Brooklyn Nets      11        C   28     173
27         Henry Sims           Brooklyn Nets      14        C   26     173
..                ...                     ...     ...      ...  ...     ...
439    Mason Plumlee  Portland Trail Blazers      24        C   26     172
447      Rudy Gobert              Utah Jazz      27        C   23     178
455     Tibor Pleiss              Utah Jazz      21        C   26     165
456      Jeff Withey              Utah Jazz      24        C   26     175
457         Priyanka              Utah Jazz      34        C   25     160

     Weight         College      Salary
7       238         Gonzaga   2165160.0
10      260      Ohio State   2569260.0
14      253  North Carolina   2616975.0
23      275        Stanford  19689000.0
27      248      Georgetown    947276.0
..      ...             ...         ...
439     235            Duke   1415520.0
447     245               0   1175880.0
455     256               0   2900000.0
456     231          Kansas    947276.0
457     231          Kansas    947276.0

[79 rows x 9 columns]

employess in position: PF
                Name                    Team  Number Position  Age  Height  \
4        Jonas Jerebko          Boston Celtics       8       PF   29     162
5         Amir Johnson          Boston Celtics      90       PF   29     169
6        Jordan Mickey          Boston Celtics      55       PF   21     177
24   Chris McCullough           Brooklyn Nets       1       PF   21     158
25         Willie Reed           Brooklyn Nets      33       PF   26     168
..                 ...                     ...     ...      ...  ...     ...
435    Meyers Leonard  Portland Trail Blazers      11       PF   24     150
441       Noah Vonleh  Portland Trail Blazers      21       PF   20     151
442     Trevor Booker              Utah Jazz      33       PF   28     177
446    Derrick Favors              Utah Jazz      15       PF   24     175
452        Trey Lyles              Utah Jazz      41       PF   20     169

     Weight       College      Salary
4       231             0   5000000.0
5       240             0  12000000.0
6       235           LSU   1170960.0
24      200      Syracuse   1140240.0
25      220   Saint Louis    947276.0
..      ...           ...         ...
435     245      Illinois   3075880.0
441     240       Indiana   2637720.0
442     228       Clemson   4775000.0
446     265  Georgia Tech  12000000.0
452     234      Kentucky   2239800.0

[100 rows x 9 columns]

employess in position: PG
                Name                    Team  Number Position  Age  Height  \
```

```
0    Avery Bradley          Boston Celtics       0       PG   25     166
8     Terry Rozier          Boston Celtics      12       PG   22     152
9     Marcus Smart          Boston Celtics      36       PG   22     180
11   Isaiah Thomas          Boston Celtics       4       PG   27     173
19    Jarrett Jack           Brooklyn Nets       2       PG   32     165
..             ...                     ...     ...      ...  ...     ...
440  Brian Roberts  Portland Trail Blazers       2       PG   30     170
443     Trey Burke              Utah Jazz       3       PG   23     159
445     Dante Exum              Utah Jazz      11       PG   20     160
453   Shelvin Mack              Utah Jazz       8       PG   26     153
454      Raul Neto              Utah Jazz      25       PG   24     178

     Weight         College     Salary
0       180           Texas  7730337.0
8       190      Louisville  1824360.0
9       220  Oklahoma State  3431040.0
11      185      Washington  6912869.0
19      200    Georgia Tech  6300000.0
..      ...             ...        ...
440     173          Dayton  2854940.0
443     191        Michigan  2658240.0
445     190               0  3777720.0
453     203          Butler  2433333.0
454     179               0   900000.0

[92 rows x 9 columns]

employess in position: SF
                      Name                    Team  Number Position  Age  \
1              Jae Crowder          Boston Celtics      99       SF   25
32   Thanasis Antetokounmpo         New York Knicks      43       SF   23
33          Carmelo Anthony         New York Knicks       7       SF   32
35         Cleanthony Early         New York Knicks      11       SF   25
42             Lance Thomas         New York Knicks      42       SF   28
..                      ...                     ...     ...      ...  ...
428         Al-Farouq Aminu  Portland Trail Blazers       8       SF   25
432        Maurice Harkless  Portland Trail Blazers       4       SF   23
448          Gordon Hayward              Utah Jazz      20       SF   26
450              Joe Ingles              Utah Jazz       2       SF   28
451           Chris Johnson              Utah Jazz      23       SF   26

     Height  Weight        College      Salary
1       154     235      Marquette   6796117.0
32      163     205              0     30888.0
33      169     240       Syracuse  22875000.0
35      172     210  Wichita State    845059.0
42      173     235           Duke   1636842.0
..      ...     ...            ...         ...
428     161     215    Wake Forest   8042895.0
432     172     215     St. John's   2894059.0
448     167     226         Butler  15409570.0
450     164     226              0   2050000.0
451     156     206         Dayton    981348.0

[85 rows x 9 columns]

employess in position: SG
                 Name                   Team  Number Position  Age  Height  \
2        John Holland          Boston Celtics      30       SG   27     152
3         R.J. Hunter          Boston Celtics      28       SG   22     176
```

```
12        Evan Turner          Boston Celtics      11       SG   27     162
13        James Young          Boston Celtics      13       SG   20     179
15   Bojan Bogdanovic           Brooklyn Nets      44       SG   27     161
..                ...                     ...     ...      ...  ...     ...
433  Gerald Henderson  Portland Trail Blazers       9       SG   28     179
437     C.J. McCollum  Portland Trail Blazers       3       SG   24     153
438      Luis Montero  Portland Trail Blazers      44       SG   23     177
444        Alec Burks              Utah Jazz      10       SG   24     155
449       Rodney Hood              Utah Jazz       5       SG   23     150

     Weight            College     Salary
2       205  Boston University        0.0
3       185      Georgia State  1148640.0
12      220         Ohio State  3425510.0
13      215           Kentucky  1749840.0
15      216                  0  3425510.0
..      ...                ...        ...
433     215               Duke  6000000.0
437     200             Lehigh  2525160.0
438     185     Westchester CC   525093.0
444     214           Colorado  9463484.0
449     206               Duke  1348440.0

[102 rows x 9 columns]
```

count the total employees position, groupby function used to group position,SG postion 102, PF 100, PG 92, SF 85, C 79

graphical representation of employee Segregate

In [96]:
```
plt.bar(position_count.index, position_count.values, color = "blue")
plt.title('Distribution of Employees by Position')
plt.xlabel("Position")
plt.ylabel("no.of employees")
plt.show()
```

display the bar chart different categories of employee position

Identify the predominant age group among employees.

In [98]:
```
age_group = df["Age"].value_counts()
age_group
```

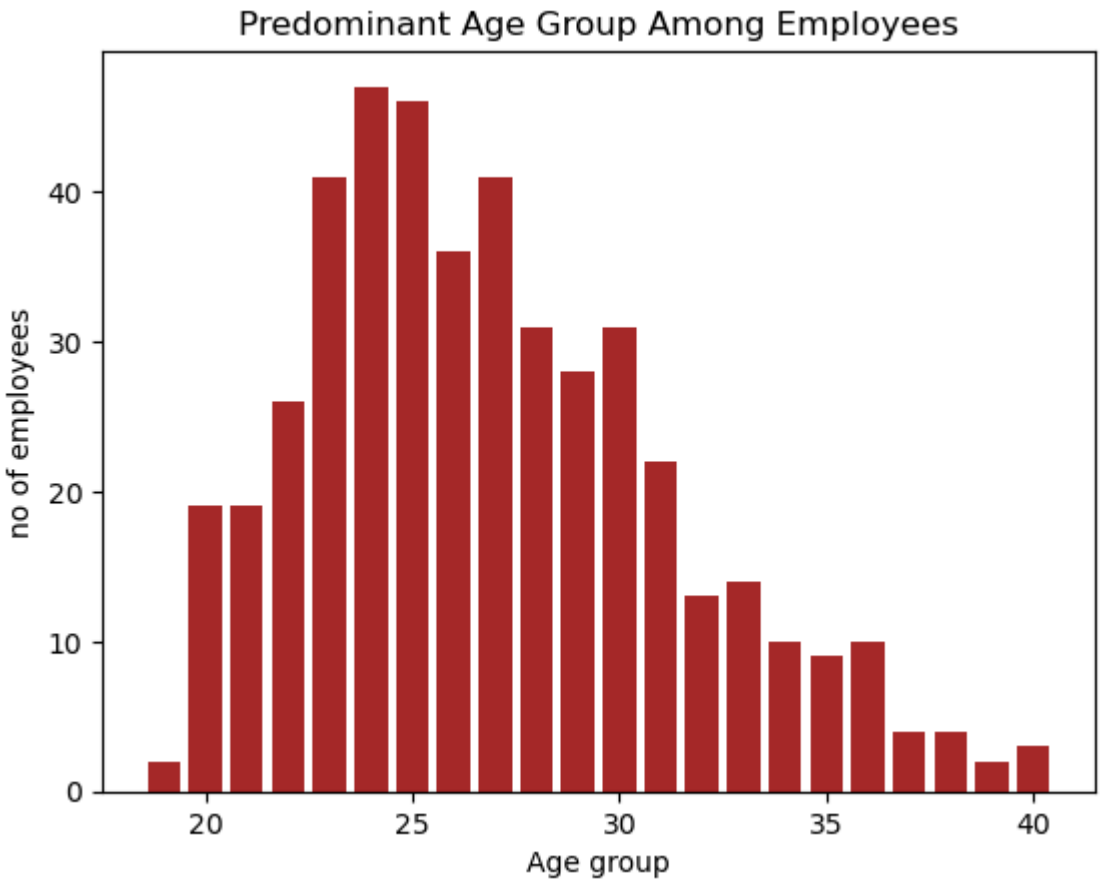
Out[98]:
```
Age
24    47
25    46
27    41
23    41
26    36
28    31
30    31
29    28
22    26
31    22
20    19
21    19
33    14
32    13
34    10
36    10
35     9
37     4
38     4
40     3
39     2
19     2
Name: count, dtype: int64
```

In [100…
```
group = age_group.idxmax()
print("Predominant Age Group Among Employees:",group)
```

```
Predominant Age Group Among Employees: 24
```

graphical representation Predominant Age Group Among Employees

In [108…
```
plt.bar(age_group.index,age_group.values,color = "brown")
plt.title("Predominant Age Group Among Employees")
plt.xlabel("Age group")
plt.ylabel("no of employees")
plt.show()
```



Each point represents an age group, with the x-coordinate representing the age group and the y-coordinate representing the count of employees in that age group.

Discover which team and position have the highest salary expenditure.

In [113…
```
sal = df["Salary"].value_counts()
print(sal)
h = sal.idxmax()
h
```

```
Salary
947276.0     32
845059.0     18
525093.0     13
0.0          11
981348.0      6
             ..
2100000.0     1
1252440.0     1
2891760.0     1
3272091.0     1
900000.0      1
Name: count, Length: 310, dtype: int64
```

Out[113… 947276.0

In [117…
```
salary = df.groupby(["Team","Position"])["Salary"].sum()
highest=salary.idxmax()
```

```
print("Team:",highest[0])
print("Position:",highest[1])
```

```
Team: Los Angeles Lakers
Position: SF
```

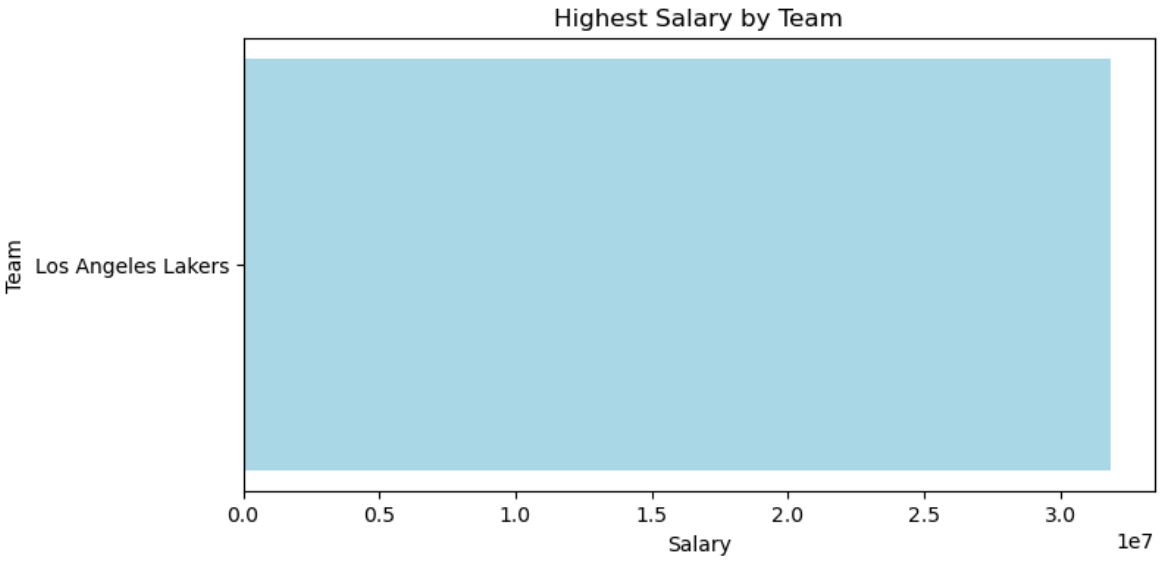
from the above team 'LOS ANGELES LAKERS 'get highest salary expenditure and postion 'SF' get highest salary expenditure

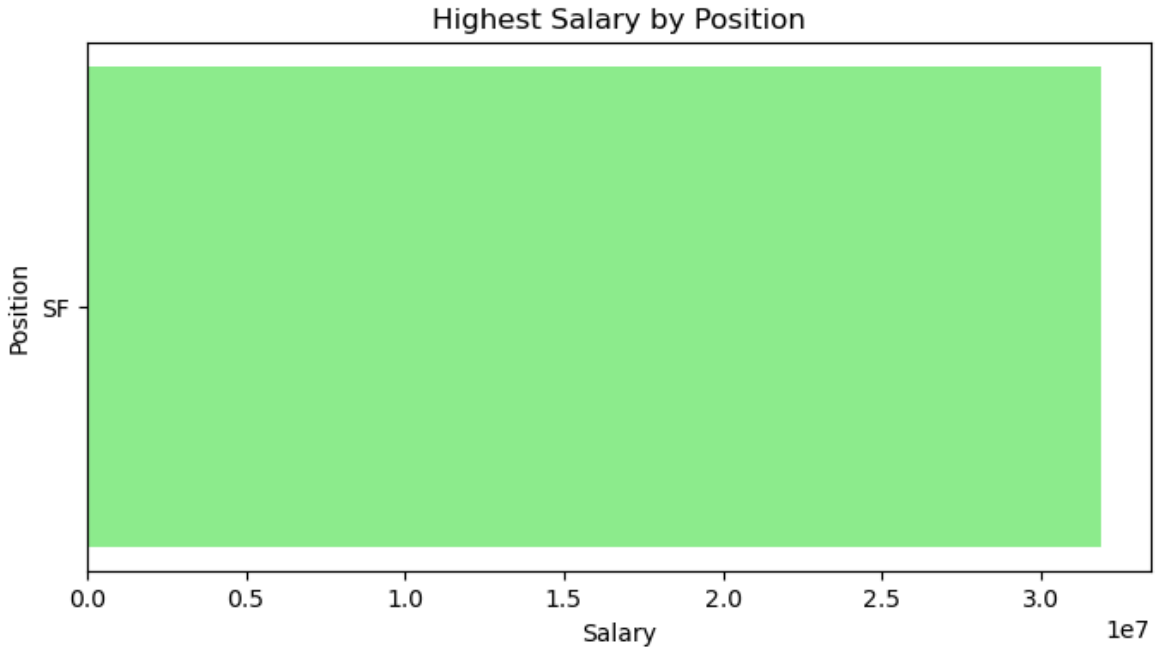
graphical representation highest salary expenditure.

In [119…

```
team = highest[0]
position = highest[1]
expenditure = salary[highest]
plt.figure(figsize=(8, 4))
plt.barh(team, expenditure, color='lightblue')
plt.title('Highest Salary by Team')
plt.xlabel('Salary')
plt.ylabel('Team')
plt.show()

# position
plt.figure(figsize=(8, 4))
plt.barh(position, expenditure, color='lightgreen')
plt.title('Highest Salary by Position')
plt.xlabel('Salary')
plt.ylabel('Position')
plt.show()
```





Investigate if there's any correlation between age and salary

In [121…

```
correlation = df['Age'].corr(df['Salary'])
print("Correlation between Age and Salary:", correlation)
```

```
Correlation between Age and Salary: 0.2050096028480935
```

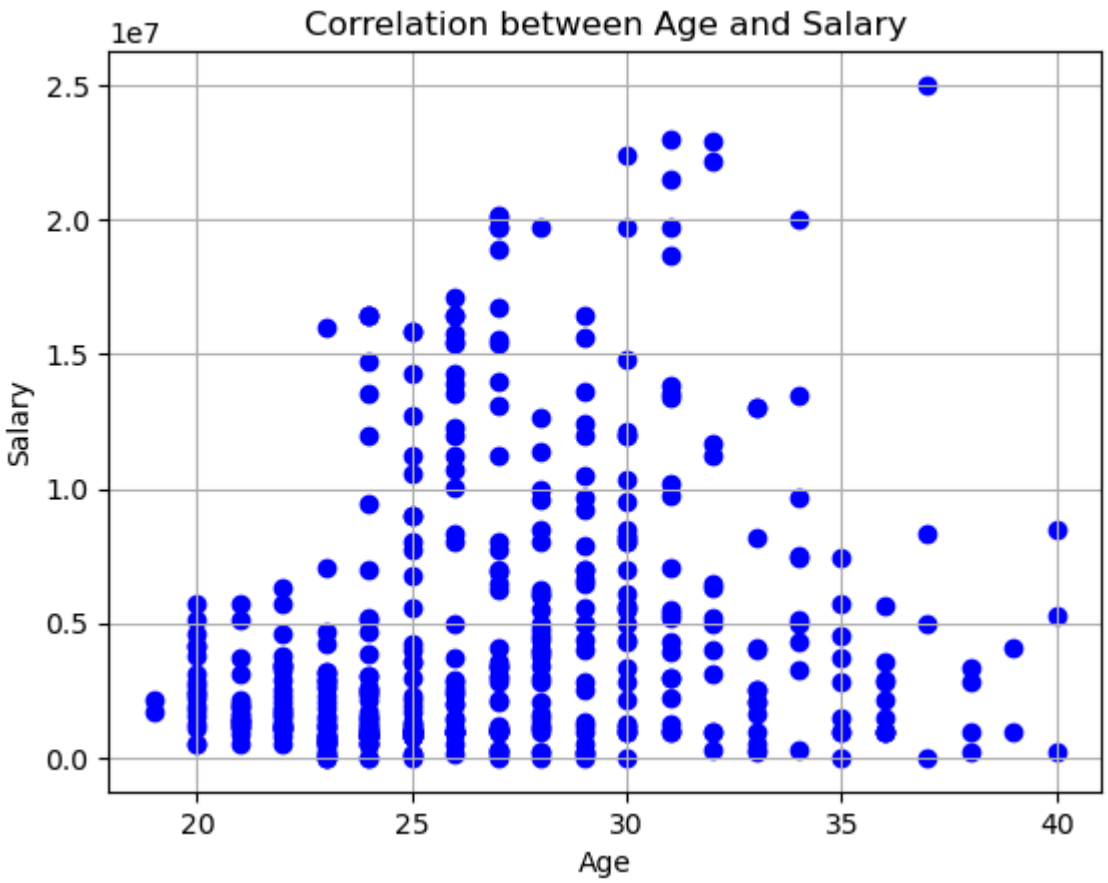
A correlation coefficient close to 1 indicates a strong positive correlation (i.e., as one variable increases, the other variable tends to increase). A correlation coefficient close to -1 indicates a strong negative correlation (i.e., as one variable increases, the other variable tends to decrease). A correlation coefficient close to 0 indicates little to no linear relationship between the variables.

A value of 0.205 suggests a positive correlation between age and salary, meaning that as age increases, salary tends to increase, and vice versa. However, the correlation is weak, indicating that the relationship between age and salary is not very strong.

Graphical Representation

In [124…

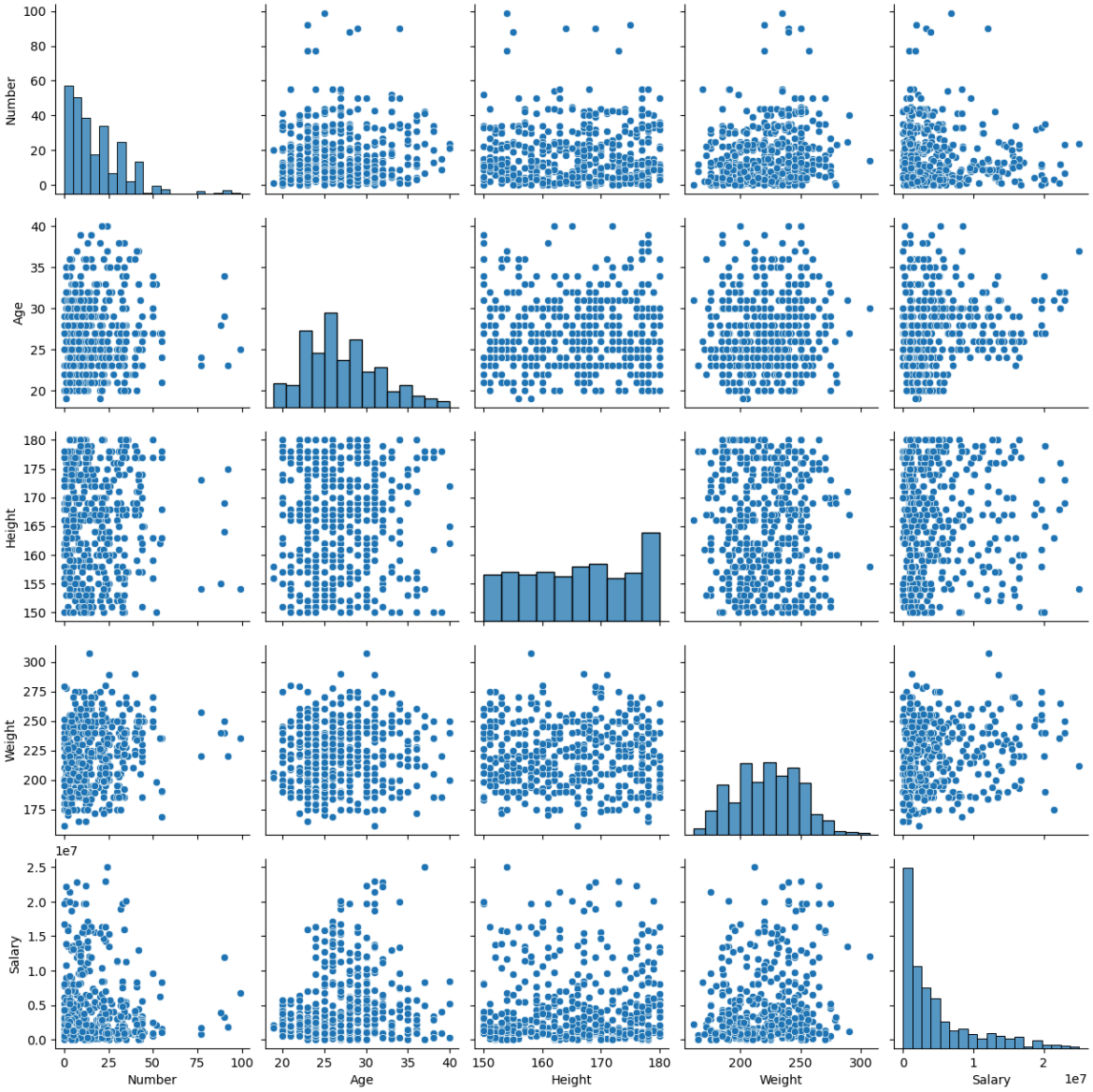
```
plt.scatter(df["Age"],df["Salary"],color="b")
plt.title('Correlation between Age and Salary')
plt.xlabel('Age')
plt.ylabel('Salary')
plt.grid(True)
plt.show()
```



plt.scatter() to create a scatter plot. The x-axis represents the age of employees Age column, and the y-axis represents their salary Salary column. scatter plot allows you to visually assess the correlation between age and salary. If there's a significant correlation, observe the pattern,higher salaries tend to certain age ranges 30 to 32 .

In [127…
```
sns.pairplot(df)
```

Out[127…
```
<seaborn.axisgrid.PairGrid at 0x22379ff6540>
```



pair plot would provide a comprehensive visual overview of the relationships between different variables in the employee dataset, helping in data exploration and analysis