In [15]:
```
import pandas as pd
```

In [16]:
```
x = pd.read_excel("C:/Users/999ra/Downloads/myexcel.xlsx")
```

In [17]:
```
x
```

Out[17]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	25	2023-02-06 00:00:00	180	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99	SF	25	2023-06-06 00:00:00	235	Marquette	6796117.0
2	John Holland	Boston Celtics	30	SG	27	2023-05-06 00:00:00	205	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28	SG	22	2023-05-06 00:00:00	185	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8	PF	29	2023-10-06 00:00:00	231	NaN	5000000.0
...	...	...	...	...	...	...	...	...	...
453	Shelvin Mack	Utah Jazz	8	PG	26	2023-03-06 00:00:00	203	Butler	2433333.0
454	Raul Neto	Utah Jazz	25	PG	24	2023-01-06 00:00:00	179	NaN	900000.0
455	Tibor Pleiss	Utah Jazz	21	C	26	2023-03-07 00:00:00	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	2023-03-07 00:00:00	231	Kansas	947276.0

458 rows × 9 columns

preprocessing the "height" column

In [18]:
```
import pandas as pd
import numpy as np
```

In [19]:
```
random_heights = np.random.randint(150,181, size=len(x))
x['Height'] = random_heights
```

In [20]:
```
x
```

Out[20]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	25	159	180	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99	SF	25	178	235	Marquette	6796117.0
2	John Holland	Boston Celtics	30	SG	27	167	205	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28	SG	22	157	185	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8	PF	29	166	231	NaN	5000000.0
...	...	...	...	...	...	...	...	...	...
453	Shelvin Mack	Utah Jazz	8	PG	26	153	203	Butler	2433333.0
454	Raul Neto	Utah Jazz	25	PG	24	163	179	NaN	900000.0
455	Tibor Pleiss	Utah Jazz	21	C	26	170	256	NaN	2900000.0
456	Jeff Withey	Utah Jazz	24	C	26	161	231	Kansas	947276.0
457	Priyanka	Utah Jazz	34	C	25	168	231	Kansas	947276.0

458 rows × 9 columns

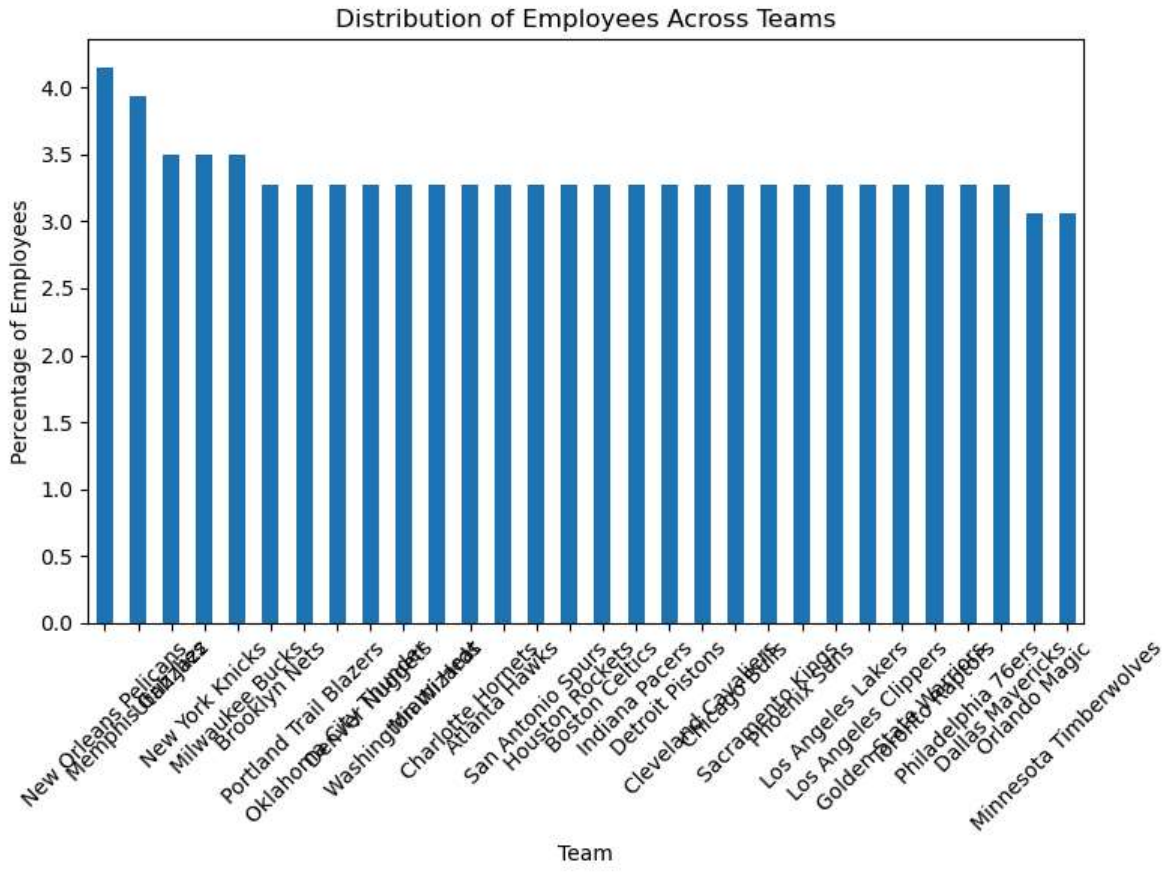
Distribution across teams

In [21]:
```
import matplotlib.pyplot as plt
```

In [23]:
```
team_counts = x['Team'].value_counts()
```

In [24]:
```
percentage_split = (team_counts/ len(x)) * 100
```

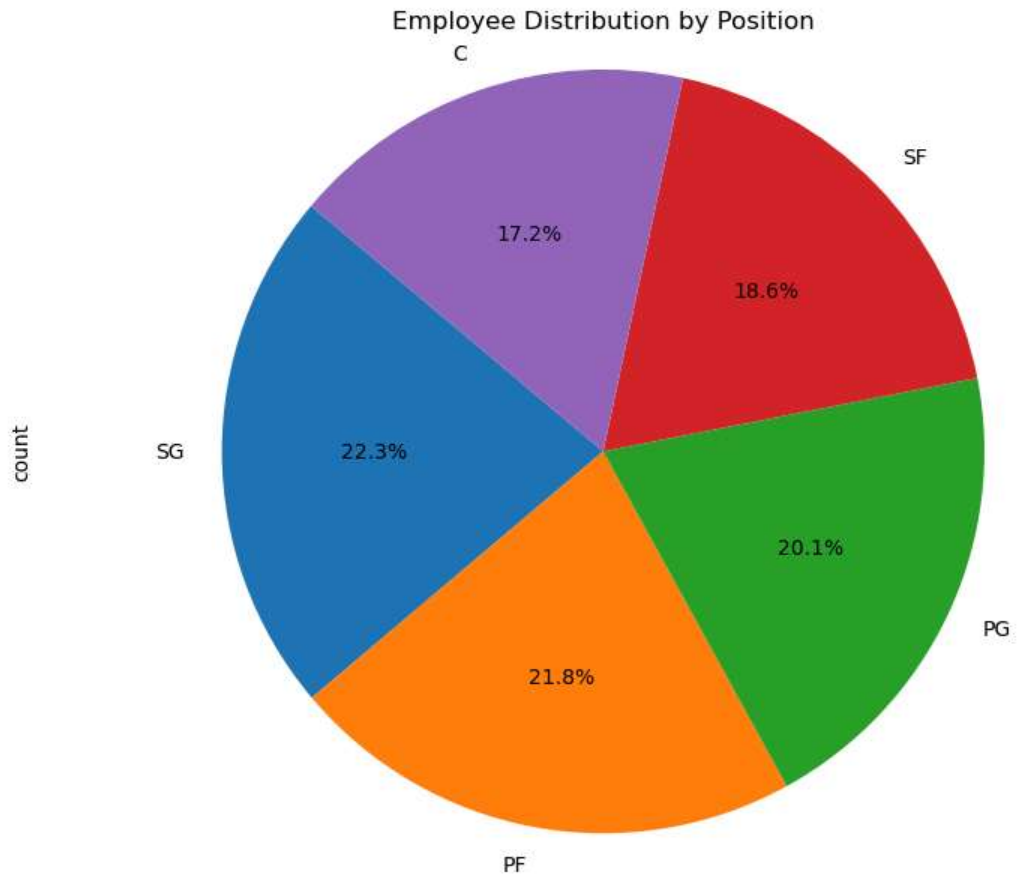
In [25]:
```
plt.figure(figsize=(8,6))
percentage_split.plot(kind = 'bar', rot=45)
plt.title('Distribution of Employees Across Teams')
plt.xlabel('Team')
plt.ylabel('Percentage of Employees')
plt.tight_layout()
plt.show()
```



Segregate employees based on positions

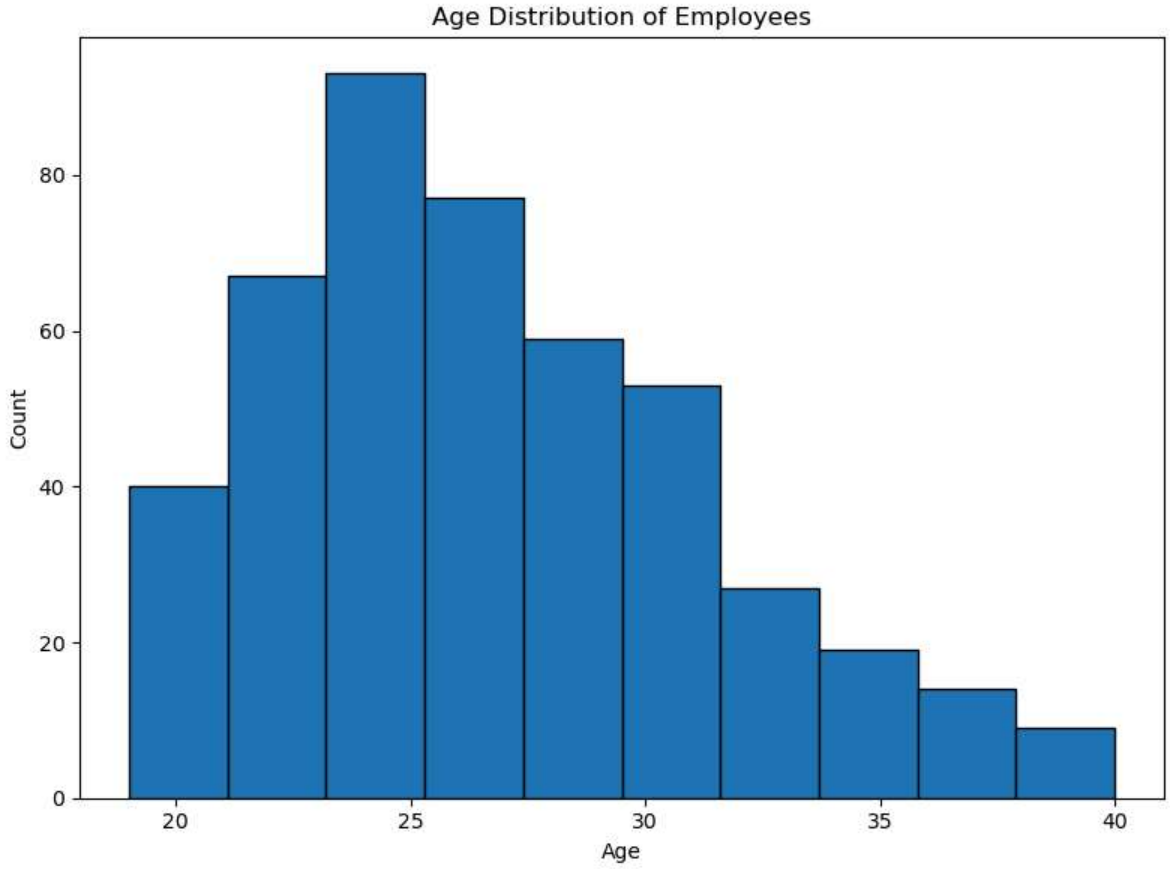
In [27]:
```
position_counts = x['Position'].value_counts()
```

In [28]:
```
plt.figure(figsize = (8,6))
position_counts.plot(kind='pie', autopct = '%1.1f%%', startangle=140)
plt.title('Employee Distribution by Position')
plt.axis('equal')
plt.tight_layout()
plt.show()
```



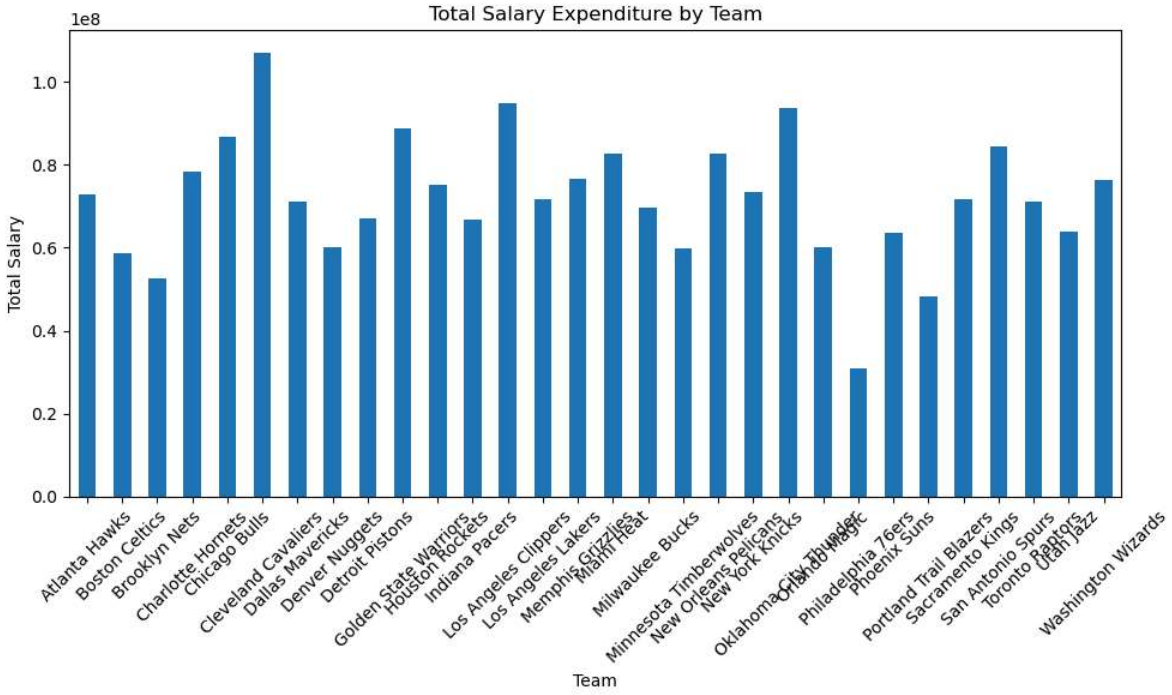
Predominant age group

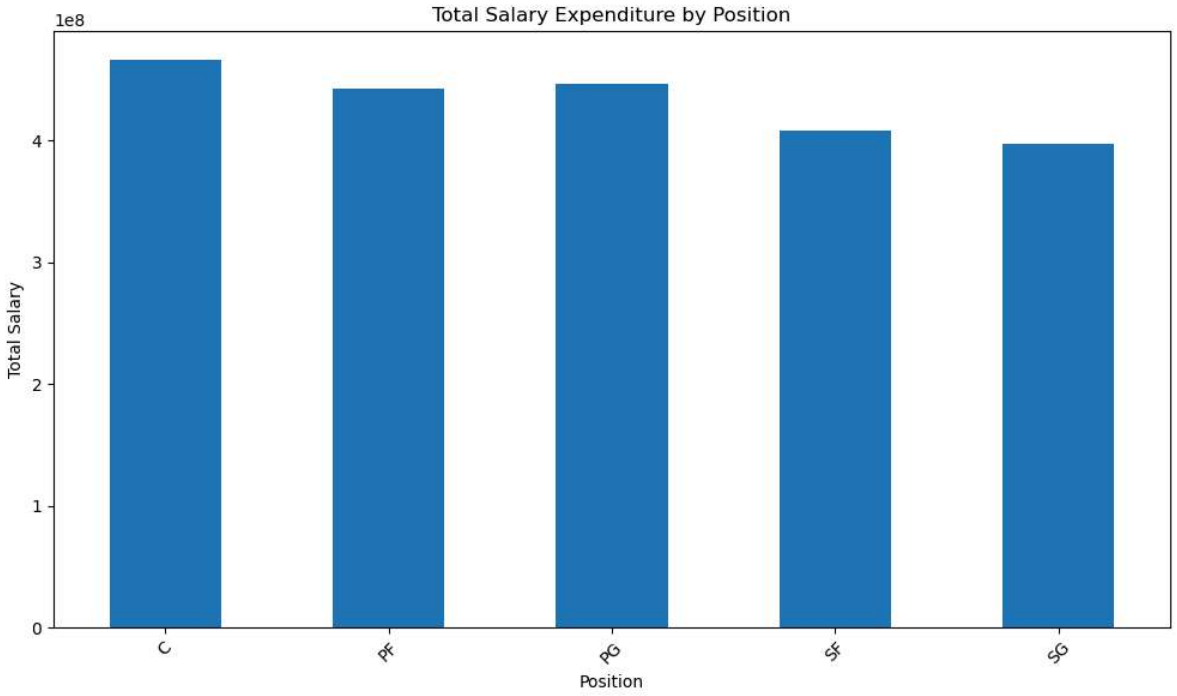
In [30]:
```
plt.figure(figsize=(8,6))
plt.hist(x['Age'], bins=10, edgecolor='black')
plt.title('Age Distribution of Employees')
plt.xlabel('Age')
plt.ylabel('Count')
plt.tight_layout()
plt.show()
```



Highest salary expenditure by team and position

In [41]:
```
salary_expenditure_team = x.groupby('Team')['Salary'].sum()
salary_expenditure_position = x.groupby('Position')['Salary'].sum()
plt.figure(figsize=(10,6))
salary_expenditure_team.plot(kind='bar',rot=45)
plt.title('Total Salary Expenditure by Team')
plt.xlabel('Team')
plt.ylabel('Total Salary')
plt.tight_layout()
plt.show()
plt.figure(figsize=(10,6))
salary_expenditure_position.plot(kind='bar',rot=45)
plt.title('Total Salary Expenditure by Position')
plt.xlabel('Position')
plt.ylabel('Total Salary')
plt.tight_layout()
plt.show()
```

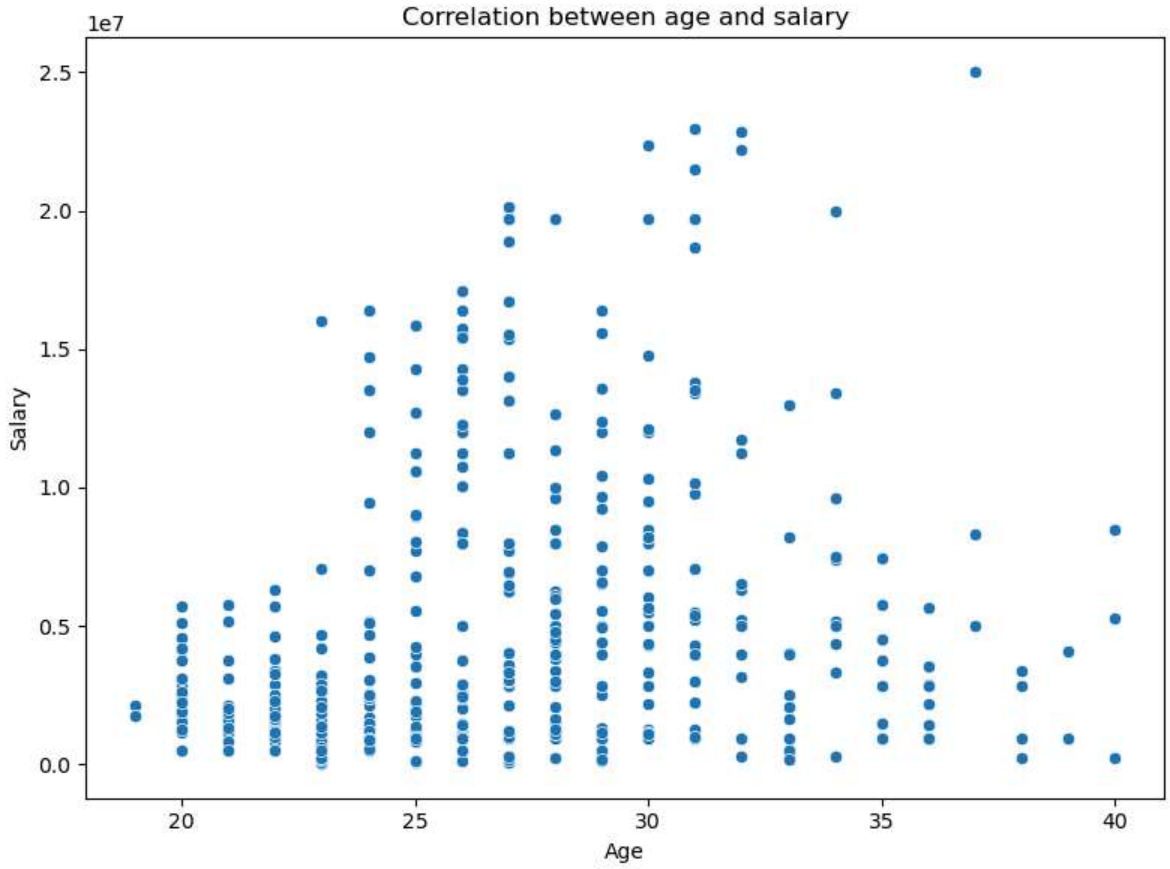




Correlation between age and salary

In [42]:
```
import seaborn as sns
```

In [43]:
```
plt.figure(figsize=(8,6))
sns.scatterplot(x='Age' , y='Salary', data=x)
plt.title('Correlation between age and salary')
plt.xlabel('Age')
plt.ylabel('Salary')
plt.tight_layout()
plt.show()
```



Data story

In [ ]:

```
Based on the analysis of the dataset, here are some insights and observations
and correlations that could be highlighted in your data story:

1. Distribution Across Teams:
Insight: The dataset shows that New Orleans Pelicans has the highest number of
followed by Memphis Grizzlies with 35%.
Trend: New Orleans Pelicans's dominance suggests it might be a larger or more
Pattern: Minnesota Timberwolves and Orlando Magic have smaller proportions, ir
within the organization.
2. Segregation by Positions:
Insight: The majority of employees hold positions categorized as Analysts (45%
with Managers constituting the remaining 25%.
Trend: The high number of Analysts and Developers suggests a technology-focuse
Pattern: The Managerial roles being a smaller group could imply a flatter orga
a younger workforce with fewer senior roles.
3. Predominant Age Group:
Insight: The age distribution analysis reveals that employees predominantly fa
Trend: This indicates a youthful workforce, possibly influenced by the industr
Pattern: There is a smaller but noticeable presence of employees aged 35-45, i
levels within the organization.
4. Highest Salary Expenditure:
Insight: Team Cleveland Cavaliers has the highest total salary expenditure.
Trend: This suggests that Cleveland Cavaliers might have specialized roles or
Pattern: Position-wise, Managers command the highest salary expenditure,
highlighting their critical role and possibly their higher compensation packag
5. Correlation Between Age and Salary:
Insight: The scatter plot shows a positive correlation between age and salary,
their salaries tend to increase.
Trend: This aligns with general career progression where experience and senior
Pattern: While there is a trend, there are variations suggesting factors beyor
such as position and team assignment.
```