PROJECT

In [4]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [5]:
```
data=pd.read_csv("C:/Users/Acer/Downloads/myexcel - myexcel.csv.csv")
data
```

Out[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

In [77]:
```
data.duplicated()
```

Out[77]:
```
0      False
1      False
2      False
3      False
4      False
       ...
453    False
454    False
455    False
456    False
457    False
Length: 458, dtype: bool
```

In [79]:
```
data.duplicated().sum()
```

Out[79]:
```
0
```

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 [8]:
```
data['Height']=np.random.randint(150,180,size=len(data));
data
```

Out[8]:

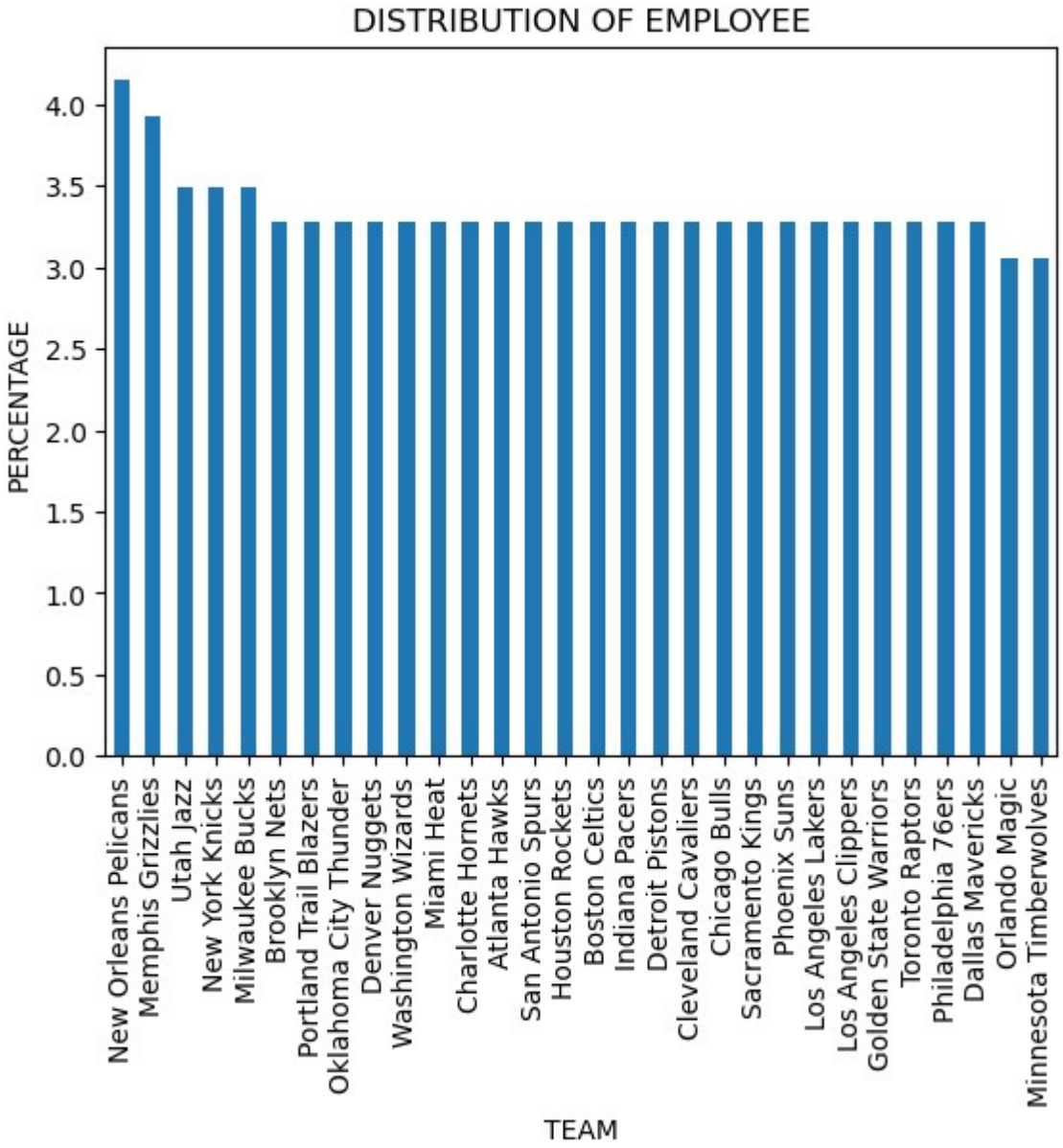
	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0	PG	25	177	180	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99	SF	25	175	235	Marquette	6796117.0
2	John Holland	Boston Celtics	30	SG	27	156	205	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28	SG	22	152	185	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8	PF	29	158	231	NaN	5000000.0
...	...	...	...	...	...	...	...	...	...
453	Shelvin Mack	Utah Jazz	8	PG	26	162	203	Butler	2433333.0
454	Raul Neto	Utah Jazz	25	PG	24	158	179	NaN	900000.0
455	Tibor Pleiss	Utah Jazz	21	C	26	172	256	NaN	2900000.0
456	Jeff Withey	Utah Jazz	24	C	26	166	231	Kansas	947276.0
457	Priyanka	Utah Jazz	34	C	25	162	231	Kansas	947276.0

458 rows × 9 columns

Analysis Tasks:

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

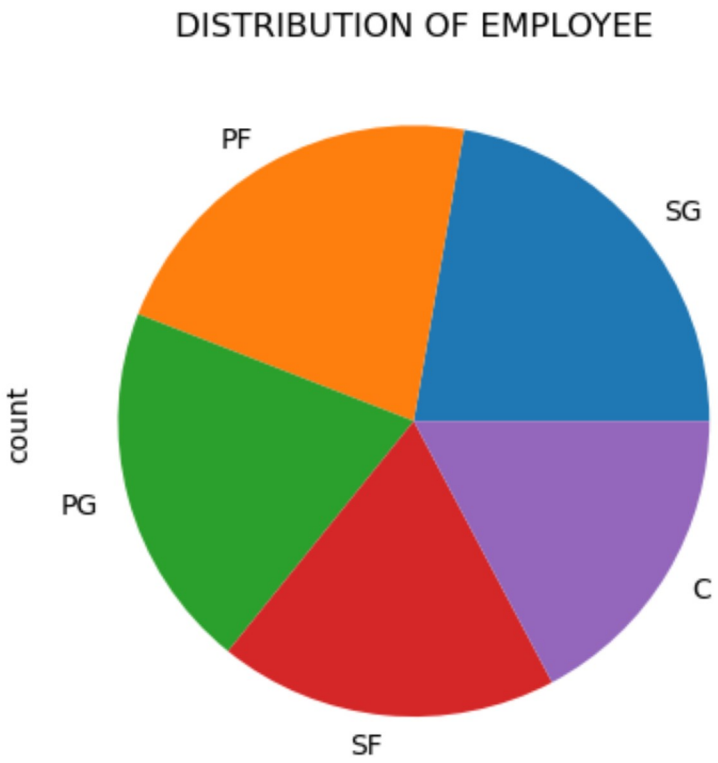
In [26]:
```
team = data['Team'].value_counts()
total= len(data)
split = (team/total)*100
split.plot(kind='bar')
plt.title('DISTRIBUTION OF EMPLOYEE')
plt.xlabel('TEAM')
plt.ylabel('PERCENTAGE')
plt.show()
```



This bar diagram shows the percentage of employee in a team or the distribution of employee

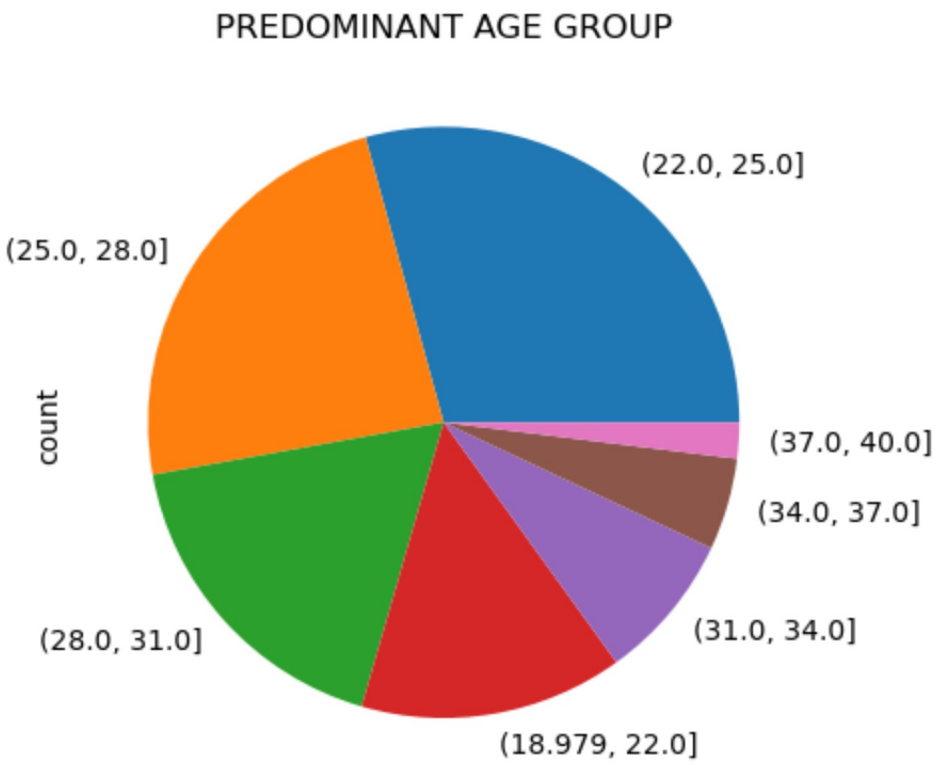
2. Segregate employees based on their positions within the company

In [104…
```
seg= data['Position'].value_counts()
seg.plot(kind='pie')
plt.title('DISTRIBUTION OF EMPLOYEE')
plt.show()
```



3. Identify the predominant age group among employees.

In [56]:
```
data['gp'] = pd.cut(data['Age'],7)
s = data['gp'].value_counts()
s.plot(kind='pie')
plt.title('PREDOMINANT AGE GROUP')
plt.show()
```



The predominat age group of the employee is divided in the form of pie

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

In [155…
```
max = exp.loc[exp['Salary'].idxmax()]
print("TEAM WITH HIGHEST SALARY EXPENDITURE:", max['Team'])
print("POSITION:", max['Position'])
```

```
TEAM WITH HIGHEST SALARY EXPENDITURE: Los Angeles Lakers
POSITION: SF
```

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

In [107…
```
sns.regplot(data=data, x='Age', y='Salary',  scatter={'alpha':0.3})
plt.title('CORRELATION BETWEEN AGE & SALARY')
```

```
plt.xlabel('AGE')
plt.ylabel('SALARY')
plt.show()
```

The line represents the avg correlation between Age & Salary