** Day - 14 Project **

In [546]:
```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

In [548]:
```
df=pd.read_csv('HR dataset.csv')
df
```

Out[548]:

	Unnamed: 0	Employee_ID	Full_Name	Department	Job_Title	Hire_Date
0	0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10
1	1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02
2	2	EMP0000003	Alyssa Martinez	HR	HR Manager	2023-03-20
3	3	EMP0000004	Nicholas Valdez	IT	Software Engineer	2023-10-12
4	4	EMP0000005	Joel Hendricks	Operations	Logistics Coordinator	2024-12-09
...	...	...	...	...	...	...
1999995	1999995	EMP1999996	Cody Russell	Operations	Logistics Coordinator	2010-08-31
1999996	1999996	EMP1999997	Tracey Smith	IT	Software Engineer	2021-05-07
1999997	1999997	EMP1999998	Tracy Lee	Sales	Business Development Manager	2024-05-29
1999998	1999998	EMP1999999	Michael Roberson	IT	Software Engineer	2023-02-14
1999999	1999999	EMP2000000	Angela Lambert	HR	Talent Acquisition Specialist	2020-11-11

2000000 rows × 12 columns

In [549]:
```
# getting basic information about the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000000 entries, 0 to 1999999
Data columns (total 12 columns):
 #   Column             Dtype 
---  ------             ----- 
 0   Unnamed: 0         int64 
 1   Employee_ID        object
 2   Full_Name          object
 3   Department         object
 4   Job_Title          object
 5   Hire_Date          object
 6   Location           object
 7   Performance_Rating  int64 
 8   Experience_Years    int64 
 9   Status             object
 10  Work_Mode          object
 11  Salary_INR         int64 
dtypes: int64(4), object(8)
memory usage: 183.1+ MB
```

In [550]:
```
# removing unwanted column from the dataframe
df.drop('Unnamed: 0',axis =1, inplace=True)
```

In [551]:
```
# change the data-type of Date column
df['Hire_Date']=pd.to_datetime(df['Hire_Date'])
```

In [552]:
```
df.head()
```

Out[552]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Pe
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	
2	EMP0000003	Alyssa Martinez	HR	HR Manager	2023-03-20	Port Christinaport, Saudi Arabia	
3	EMP0000004	Nicholas Valdez	IT	Software Engineer	2023-10-12	Port Shelbychester, Antigua and Barbuda	
4	EMP0000005	Joel Hendricks	Operations	Logistics Coordinator	2024-12-09	Lake Kimberly, Palestinian Territory	

In [553]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000000 entries, 0 to 1999999
Data columns (total 11 columns):
 #   Column             Dtype         
---  ------             -----         
 0   Employee_ID        object        
 1   Full_Name          object        
 2   Department         object        
 3   Job_Title          object        
 4   Hire_Date          datetime64[ns]
 5   Location           object        
 6   Performance_Rating  int64         
 7   Experience_Years    int64         
 8   Status             object        
 9   Work_Mode          object        
 10  Salary_INR         int64         
dtypes: datetime64[ns](1), int64(3), object(7)
memory usage: 167.8+ MB
```

In [554]:
```
df['Performance_Rating'].unique()
```

Out[554]:
```
array([5, 2, 1, 4, 3], dtype=int64)
```

In [555]:
```
df['Performance_Rating'].value_counts()
```

Out[555]:
```
Performance_Rating
4    400529
2    400174
3    399814
1    399756
5    399727
Name: count, dtype: int64
```

In [556]:
```
df['Performance_Rating'].mean()
```

Out[556]:
```
3.0001485
```

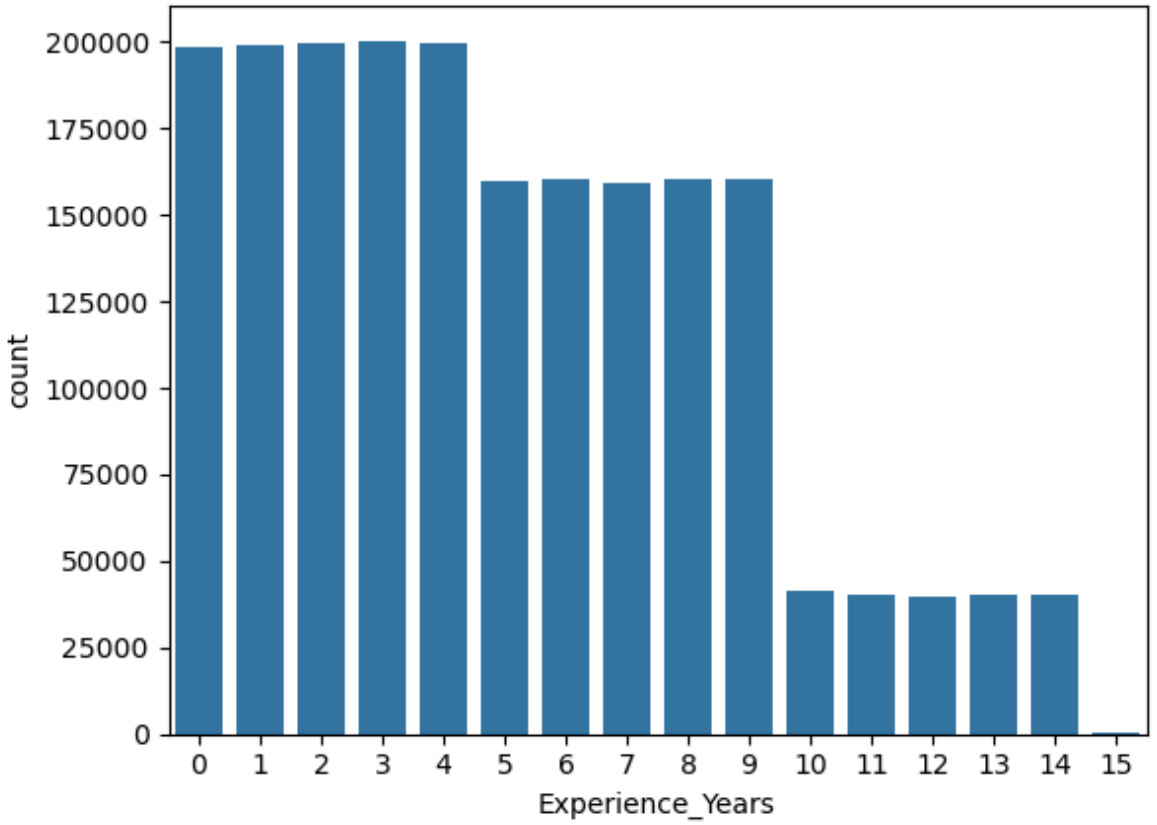
In [557]:
```
df['Experience_Years'].nunique()
```

Out[557]:
```
16
```

In [558]:
```
df['Experience_Years'].unique()
```

Out[558]:
```
array([14,  7,  2,  1,  0,  4,  9,  5,  6,  8,  3, 10, 11, 12, 13, 15],
      dtype=int64)
```

In [559]:
```
sns.countplot(x='Experience_Years', data=df)
plt.show()
```



In [560]:
```
df['Experience_Years'].value_counts()
```

Out[560]:
```
Experience_Years
3    200522
2    199924
4    199866
1    199162
0    198775
6    160410
9    160223
8    160212
5    160112
7    159005
10    41209
13    40149
11    40146
14    40005
12    39709
15      571
Name: count, dtype: int64
```

In [561]:
```
# Consider the columns having data-type 'object' only
df.select_dtypes( include ='object')
```

Out[561]:

	Employee_ID	Full_Name	Department	Job_Title	Location	Sta
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	Isaacland, Denmark	Resig
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	Anthonyside, Costa Rica	Ac
2	EMP0000003	Alyssa Martinez	HR	HR Manager	Port Christinaport, Saudi Arabia	Ac
3	EMP0000004	Nicholas Valdez	IT	Software Engineer	Port Shelbychester, Antigua and Barbuda	Ac
4	EMP0000005	Joel Hendricks	Operations	Logistics Coordinator	Lake Kimberly, Palestinian Territory	Ac
...	...	...	...	...	...	
1999995	EMP1999996	Cody Russell	Operations	Logistics Coordinator	Casefurt, Serbia	Ac
1999996	EMP1999997	Tracey Smith	IT	Software Engineer	Dannyport, Kuwait	Ac
1999997	EMP1999998	Tracy Lee	Sales	Business Development Manager	Craighaven, Nigeria	Ac
1999998	EMP1999999	Michael Roberson	IT	Software Engineer	Jonathanmouth, Djibouti	Ret
1999999	EMP2000000	Angela Lambert	HR	Talent Acquisition Specialist	Morganchester, Canada	Ac

2000000 rows × 7 columns

In [562]:
```
df.select_dtypes( include ='number')
```

Out[562]:

	Performance_Rating	Experience_Years	Salary_INR
0	5	14	1585363
1	2	7	847686
2	1	2	1430084
3	1	1	990689
4	5	0	535082
...	...	...	...
1999995	3	14	657648
1999996	3	4	1030109
1999997	5	1	1313085
1999998	4	2	1479727
1999999	1	4	993718

2000000 rows × 3 columns

Q.1) What is the distribution of Employee Status(Active, Resigned, Retired, Terminated) ?

In [564]:
```
df.head(2)
```

Out[564]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

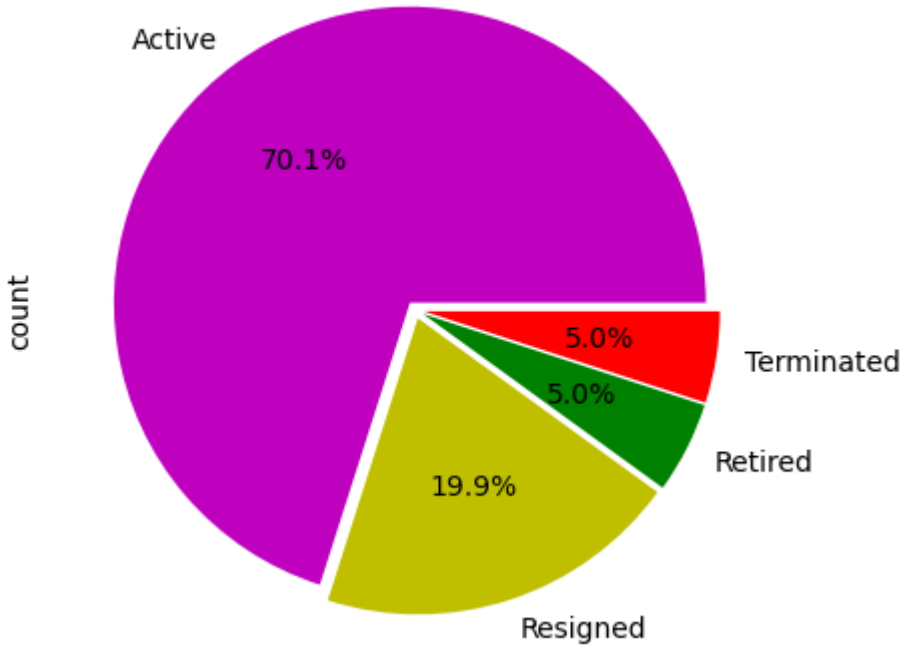
In [565]:
```
status=df['Status'].value_counts()
status
```

Out[565]:
```
Status
Active       1401558
Resigned      398660
Retired        99912
Terminated     99870
Name: count, dtype: int64
```

In [566]:
```
type(status)
```

Out[566]:
```
pandas.core.series.Series
```

In [567]:
```
status.plot(kind ='pie', colors='mygr', autopct='%1.1f%%',explode=(0.03,0.03,0.03,0.03))
plt.show()
```



Q.2) What is the distribution of work modes(On-site,Remote)?

In [569]:
```
df.head(2)
```

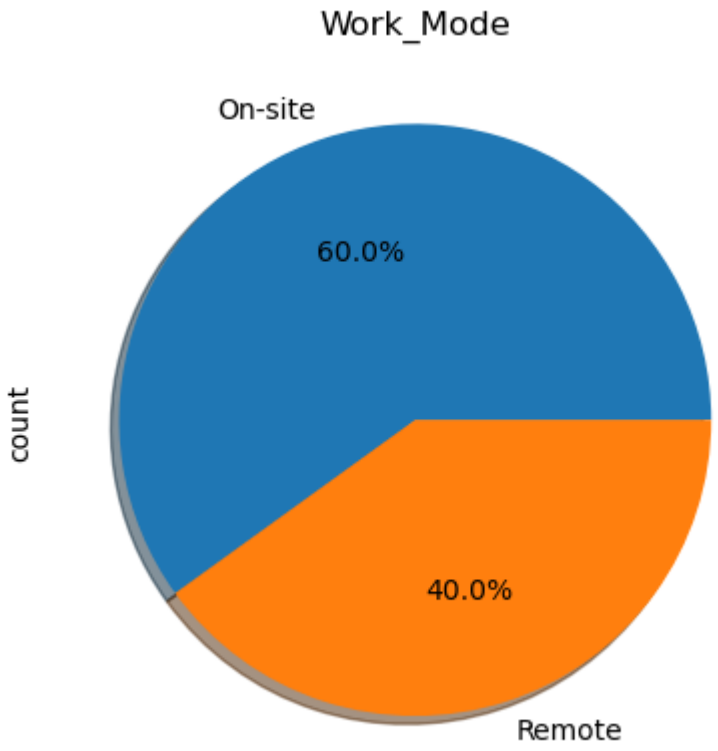
Out[569]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [570]:
```
work=df['Work_Mode'].value_counts()
work
```

Out[570]:
```
Work_Mode
On-site    1199109
Remote      800891
Name: count, dtype: int64
```

In [571]:
```
work.plot(kind='pie', color='cr',autopct='%1.1f%%', shadow=True)
plt.title("Work_Mode")
plt.show()
```



Q.3) How many employees are there in each department?

In [573]:
```
df.head(2)
```

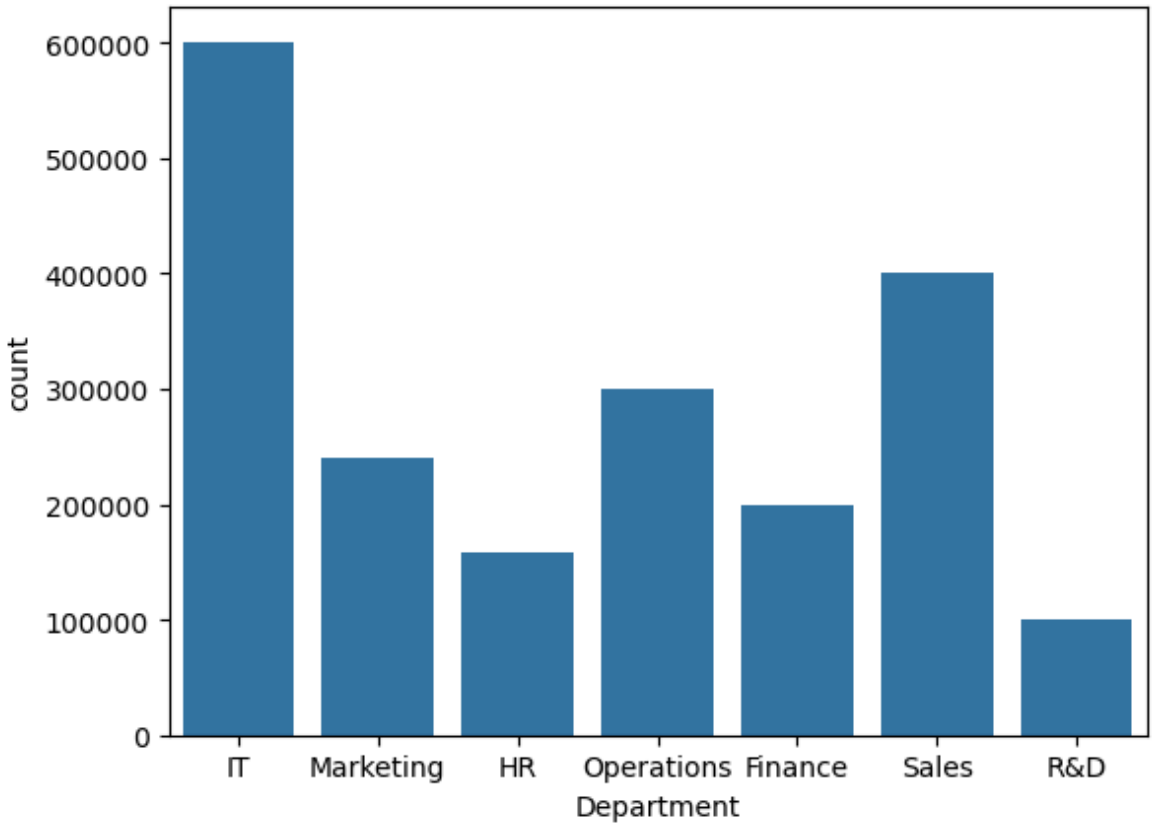
Out[573]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [574]:
```
df['Department'].value_counts()
```

Out[574]:
```
Department
IT            601042
Sales         400031
Operations    300095
Marketing     240081
Finance       199873
HR            159119
R&D            99759
Name: count, dtype: int64
```

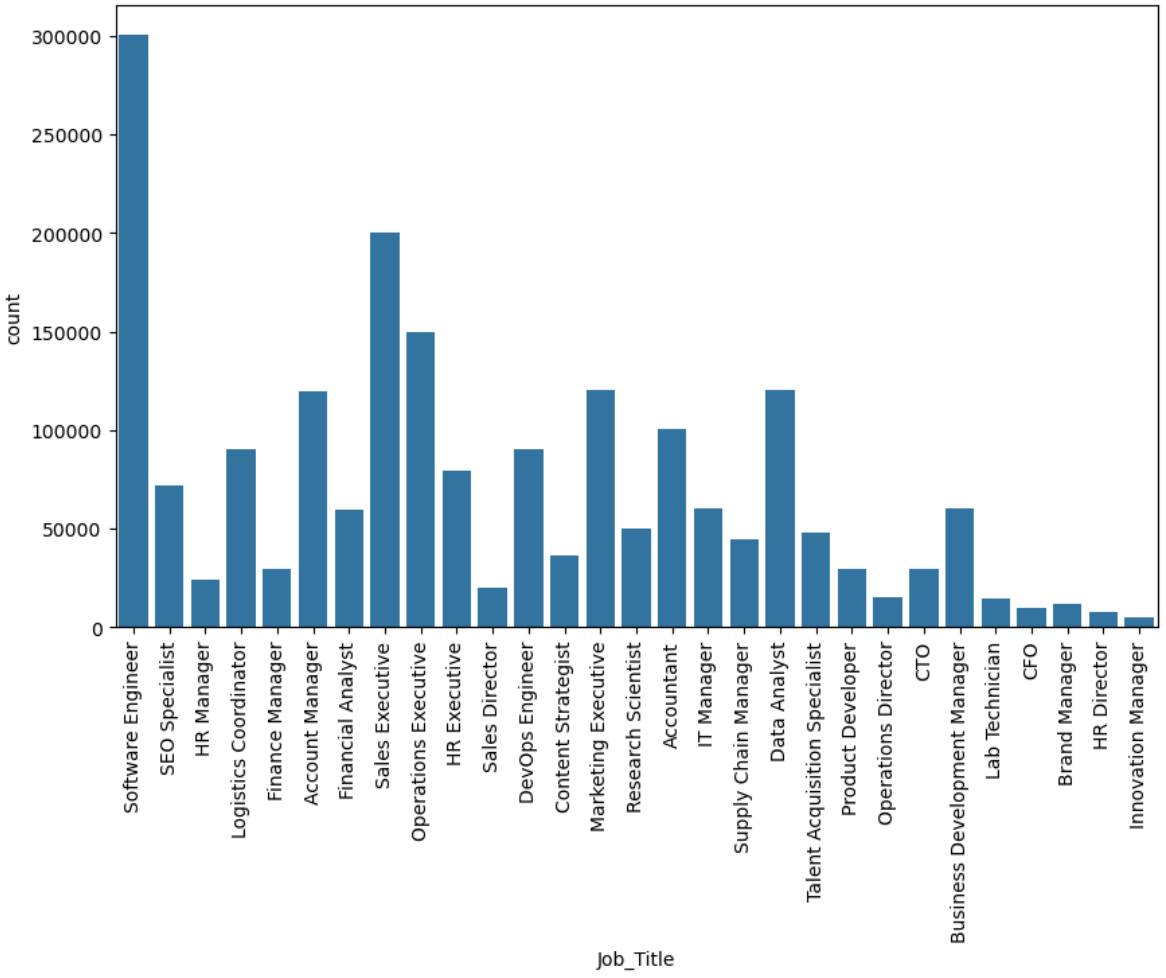
In [575]:
```
sns.countplot(x='Department',data=df)
plt.show()
```



In [576]:
```
df['Job_Title'].value_counts()
```

Out[576]:
```
Job_Title
Software Engineer              300358
Sales Executive                199982
Operations Executive           150058
Data Analyst                   120375
Marketing Executive            120154
Account Manager                119929
Accountant                     100307
DevOps Engineer                 90197
Logistics Coordinator           90188
HR Executive                    79348
SEO Specialist                  71692
Business Development Manager    60233
IT Manager                      60224
Financial Analyst               59815
Research Scientist              50017
Talent Acquisition Specialist   47994
Supply Chain Manager            44935
Content Strategist              36154
CTO                             29888
Product Developer               29872
Finance Manager                 29799
HR Manager                      23841
Sales Director                  19887
Operations Director             14914
Lab Technician                  14829
Brand Manager                   12081
CFO                              9952
HR Director                      7936
Innovation Manager               5041
Name: count, dtype: int64
```

In [577]:
```
plt.figure(figsize =(10,6))
sns.countplot(x='Job_Title',data=df)
plt.xticks(rotation='vertical')
plt.show()
```



Q.4) What is the average salary by Department?

In [579]:
```
df.head(2)
```

Out[579]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [580]:
```
dept=df.groupby('Department')['Salary_INR'].mean()/1000
dept
```

Out[580]:
```
Department
Finance       940.411743
HR            743.853561
IT           1129.858151
Marketing     769.936152
Operations    754.626253
R&D           800.377157
Sales         792.957860
Name: Salary_INR, dtype: float64
```

In [581]:
```
type(dept)
```

Out[581]: pandas.core.series.Series

In [582]:
```
plt.figure(figsize=(10,6))
dept.plot(x=dept.index, y=dept.values, kind ='bar', color='cyan',legend= True, width =0.3 )
plt.grid()
plt.title("Average Salary in Departments")
plt.ylabel("Salary")
plt.show()
```

Q.5) Which job title has highest average salary ?

In [584]:
```
df.head(2)
```

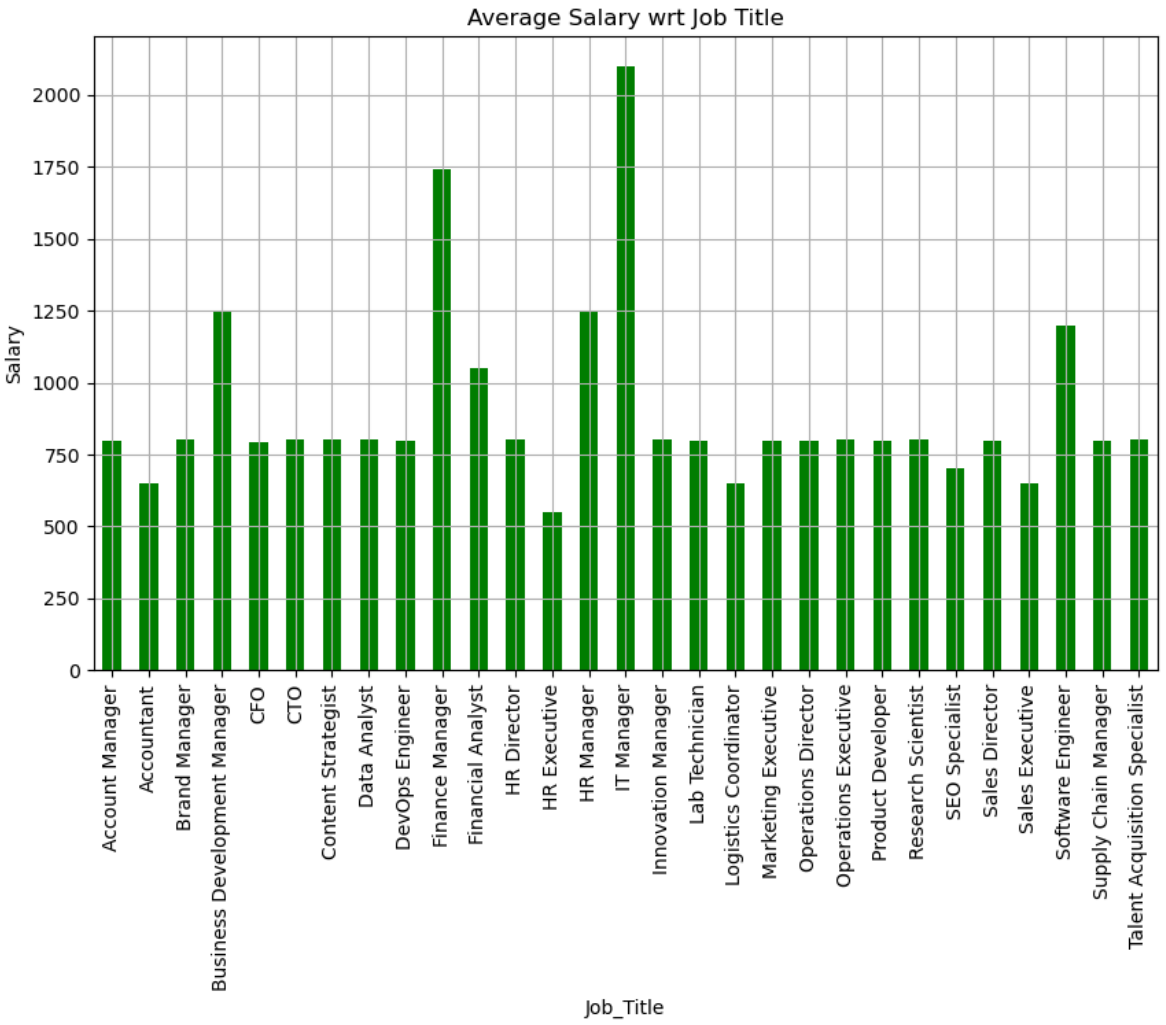
Out[584]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [585]:
```
salary=df.groupby('Job_Title')['Salary_INR'].mean()/1000
salary
```

Out[585]:
```
Job_Title
Account Manager                799.373734
Accountant                     650.076482
Brand Manager                  803.127787
Business Development Manager    1252.016231
CFO                            795.015873
CTO                            801.402754
Content Strategist             800.760030
Data Analyst                   800.996380
DevOps Engineer                799.949184
Finance Manager                1743.241525
Financial Analyst              1051.522903
HR Director                    800.694437
HR Executive                   550.548859
HR Manager                     1252.401915
IT Manager                     2098.155777
Innovation Manager             801.870103
Lab Technician                 800.181468
Logistics Coordinator          649.631726
Marketing Executive            798.780404
Operations Director            798.298093
Operations Executive           800.350915
Product Developer              798.652261
Research Scientist             801.314879
SEO Specialist                 700.456337
Sales Director                 799.069374
Sales Executive                650.237755
Software Engineer              1199.260843
Supply Chain Manager           798.168555
Talent Acquisition Specialist  801.422237
Name: Salary_INR, dtype: float64
```

In [586]:
```
plt.figure(figsize=(10,6))
salary.plot(x=salary.index, y=salary.values, kind ='bar', color='g')
plt.grid()
plt.title("Average Salary wrt Job Title")
plt.ylabel("Salary")
plt.show()
plt.show()
```



Q.6) What is the average salary in different Departments bases on Job Title ?

In [588]:
```
df.head(2)
```

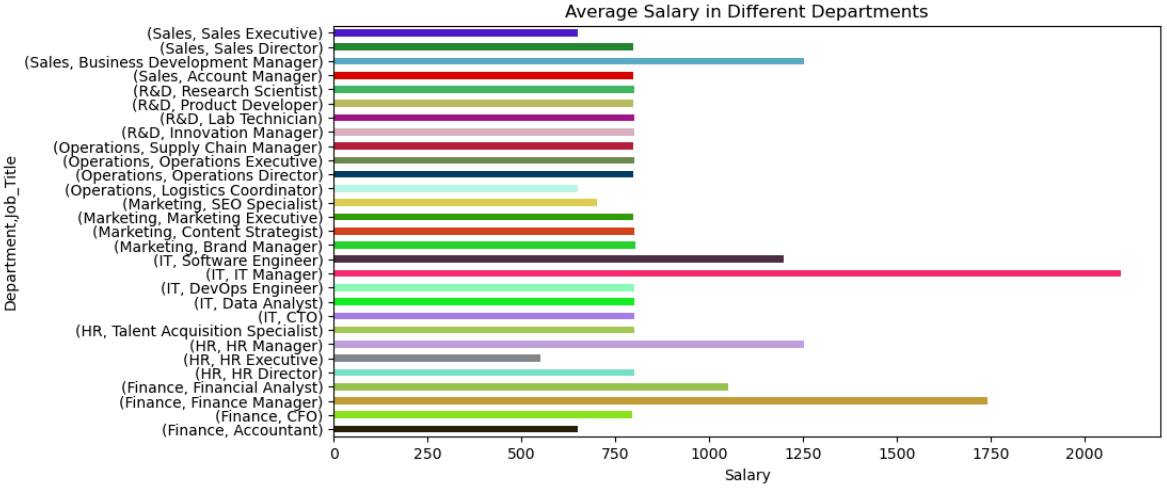
Out[588]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [589]:
```
dept_job=df.groupby(['Department', 'Job_Title'])['Salary_INR'].mean()/1000
dept_job
```

Out[589]:
```
Department  Job_Title
Finance     Accountant                    650.076482
            CFO                           795.015873
            Finance Manager              1743.241525
            Financial Analyst            1051.522903
HR          HR Director                   800.694437
            HR Executive                  550.548859
            HR Manager                   1252.401915
            Talent Acquisition Specialist  801.422237
IT          CTO                           801.402754
            Data Analyst                  800.996380
            DevOps Engineer               799.949184
            IT Manager                   2098.155777
            Software Engineer            1199.260843
Marketing   Brand Manager                 803.127787
            Content Strategist            800.760030
            Marketing Executive           798.780404
            SEO Specialist                700.456337
Operations  Logistics Coordinator         649.631726
            Operations Director           798.298093
            Operations Executive          800.350915
            Supply Chain Manager          798.168555
R&D         Innovation Manager            801.870103
            Lab Technician                800.181468
            Product Developer             798.652261
            Research Scientist            801.314879
Sales       Account Manager               799.373734
            Business Development Manager  1252.016231
            Sales Director                799.069374
            Sales Executive               650.237755
Name: Salary_INR, dtype: float64
```

In [590]:
```
import random
num_bars=len(dept_job)
random_colors=[f'#{random.randint(0,0xFFFFFF):06x}' for _ in range(num_bars)]
dept_job.plot(kind='barh',figsize=(10,5), color=random_colors)
plt.title('Average Salary in Different Departments')
plt.xlabel("Salary")
plt.savefig('new_chart.png')
plt.show()
```



Q.7)How many employees Resigned & Terminated in each department ?

In [592]:
```
df.head(2)
```

Out[592]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [594]:
```
df.Status.unique()
```

Out[594]:
```
array(['Resigned', 'Active', 'Terminated', 'Retired'], dtype=object)
```

In [597]:
```
R=df[df['Status']== 'Resigned']
R
```

Out[597]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Loca
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaac Den
8	EMP0000009	Cathy Thompson	Finance	Financial Analyst	2018-05-29	S Cathe E
11	EMP0000012	Kevin Lowe	Sales	Account Manager	2024-07-02	East
16	EMP0000017	Robert Martin	Operations	Logistics Coordinator	2025-05-13	Laurah Afghan
19	EMP0000020	Donald Hoffman	Marketing	Content Strategist	2022-04-01	South Ja New Zea
...	...	...	...	...	...	
1999976	EMP1999977	Angela Curtis	Operations	Operations Executive	2021-08-07	Jeremiahb Rw
1999983	EMP1999984	Joshua Ponce	Sales	Account Manager	2020-05-08	North Tr Vene
1999985	EMP1999986	Aaron Montgomery	Marketing	Marketing Executive	2017-06-03	Maddenm E
1999986	EMP1999987	Mason Parker	Operations	Operations Executive	2018-02-27	Jose Came
1999989	EMP1999990	Adrian Lopez	Sales	Sales Executive	2017-07-25	N Elizabeth Mor

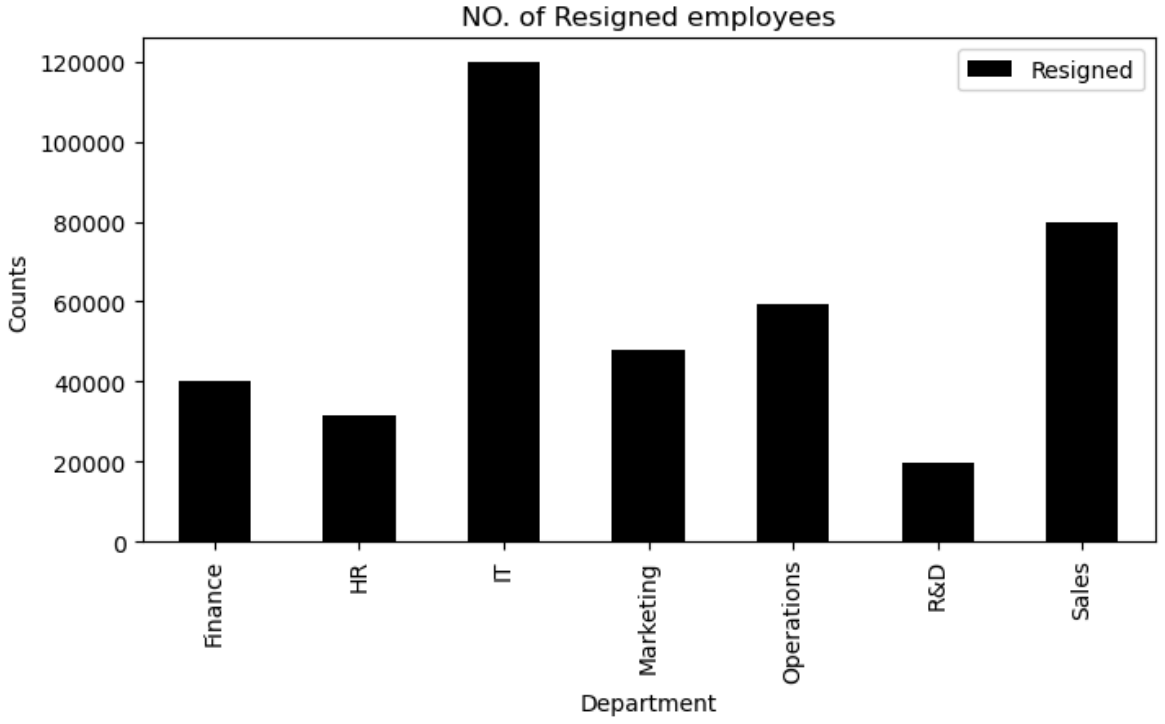
398660 rows × 11 columns

In [598]:
```
R_emp=R.groupby('Department')['Status'].count()
R_emp
```

Out[598]:
```
Department
Finance       40238
HR            31736
IT           119852
Marketing     47793
Operations    59397
R&D           19919
Sales         79725
Name: Status, dtype: int64
```

In [599]:
```
# R_emp.groupby('Department')['Work_Mode'].count()
```

In [600]:
```
plt.figure(figsize=(8,4))
R_emp.plot(x=R_emp.index, y=R_emp.values,kind='bar',color='black', legend =True,label='Resig
plt.title("NO. of Resigned employees")
plt.ylabel("Counts")
plt.show()
```



In [601]:
```
df.head(2)
```

Out[601]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [602]:
```
df_Terminated=df[df['Status']=='Terminated']
df_Terminated
```

Out[602]:

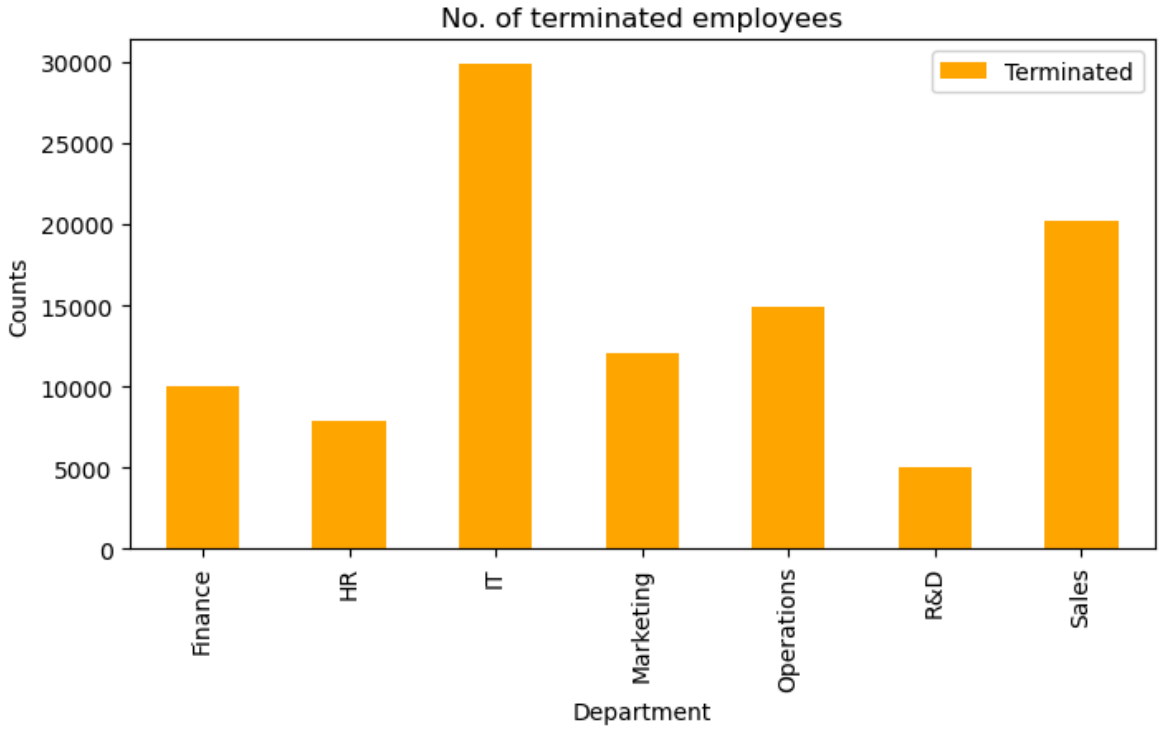
	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Locatic
20	EMP0000021	Mr. Billy Rodgers DDS	Marketing	Marketing Executive	2017-10-12	We Bryantc Saint Mart
33	EMP0000034	Steve Carlson	IT	Software Engineer	2020-04-25	Grahamfu Jamai
56	EMP0000057	Claire Martinez	IT	DevOps Engineer	2020-01-17	Garciatc Libyan Ara Jamahiri
100	EMP0000101	Johnny Shepard	Finance	Accountant	2023-02-02	Nor Briannatow Cul
121	EMP0000122	Vanessa Brown	IT	Data Analyst	2017-08-14	South Teres Liechtenste
...	...	...	...	...	...	
1999912	EMP1999913	Stefanie Valentine	Marketing	Content Strategist	2016-05-04	New Aarontc Andor
1999936	EMP1999937	Lisa Gordon	Finance	Financial Analyst	2025-02-25	Baxtermout Qat
1999947	EMP1999948	John Johnson	Sales	Sales Executive	2019-11-13	Maryboroug Nep
1999981	EMP1999982	Mindy Campbell	Sales	Account Manager	2018-07-16	Sharoncheste Belgiu
1999993	EMP1999994	Ashley Fuller	IT	DevOps Engineer	2018-06-09	Dylanhave Bermu

99870 rows × 11 columns

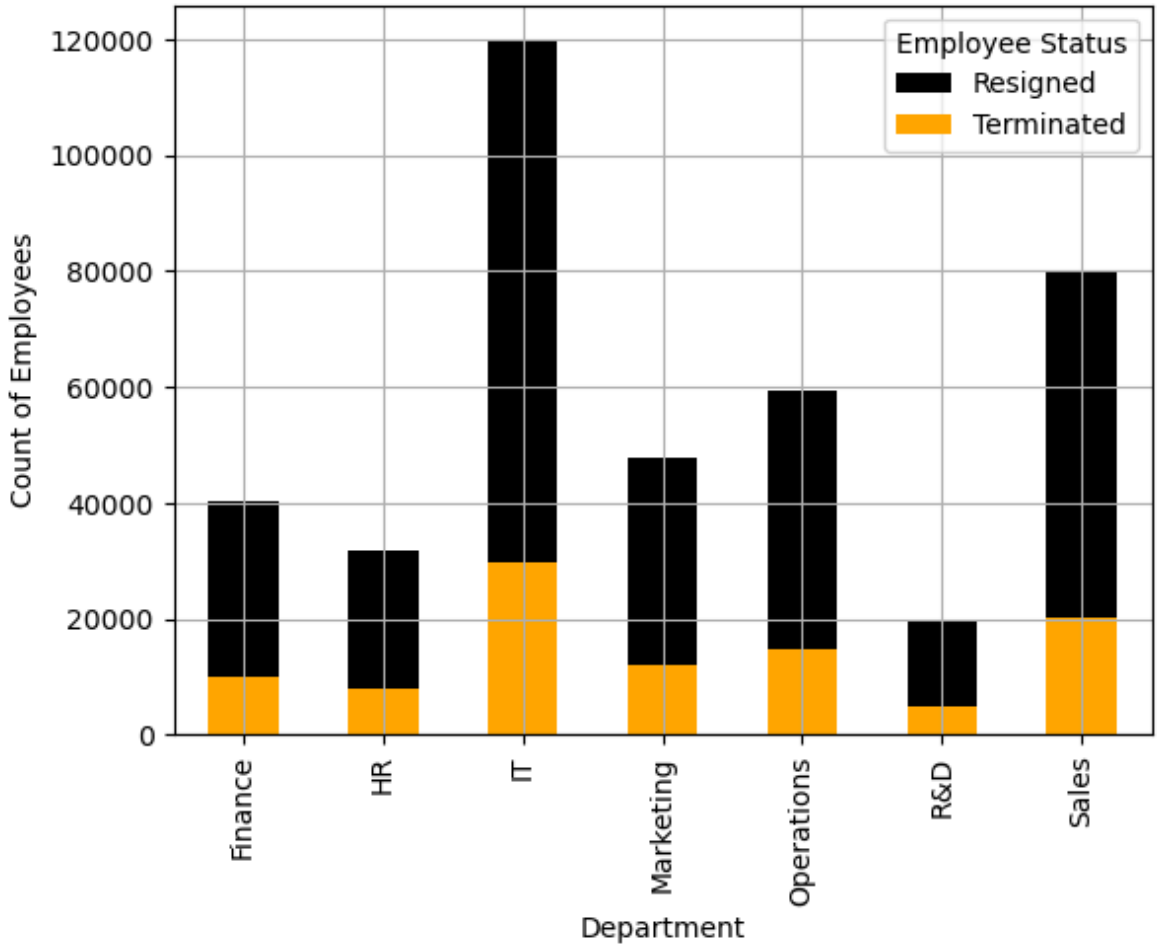
In [604]:
```
T_emp=df_Terminated.groupby('Department')['Status'].count()
T_emp
```

Out[604]:
```
Department
Finance       9988
HR            7861
IT           29881
Marketing    12044
Operations   14884
R&D           4998
Sales        20214
Name: Status, dtype: int64
```

In [607]:
```
plt.figure(figsize=(8,4))
T_emp.plot(x=T_emp.index, y=T_emp.values, kind ='bar', color='orange', legend=True, label ='T
plt.title('No. of terminated employees')
plt.ylabel("Counts")
plt.show()
```



In [612]:
```
R_emp.plot(x=R_emp.index, y=R_emp.values,kind='bar',color='black', legend =True,label='Resig
T_emp.plot(x=T_emp.index, y=T_emp.values, kind ='bar', color='orange', legend=True, label ='T
plt.legend(title='Employee Status')
plt.ylabel("Count of Employees")
plt.grid()
plt.show()
```



Q.8)How does Salary very with years of experience ?

In [615]:
```
df.head(2)
```

Out[615]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [616]:
```
df['Experience_Years'].nunique()
```

Out[616]: 16

In [617]:
```
df.groupby('Experience_Years')['Salary_INR'].mean()
```

Out[617]:
```
Experience_Years
0    896737.454775
1    895903.759824
2    896755.652313
3    896861.245240
4    897944.573965
5    896484.084828
6    896012.632467
7    895722.673960
8    897148.361090
9    898482.940577
10   895662.027882
11   901452.750112
12   896432.933416
13   898790.197041
14   895610.790251
15   895647.401051
Name: Salary_INR, dtype: float64
```

Q.9) What is the average performance rating by department?

In [619]:
```
df.head(2)
```

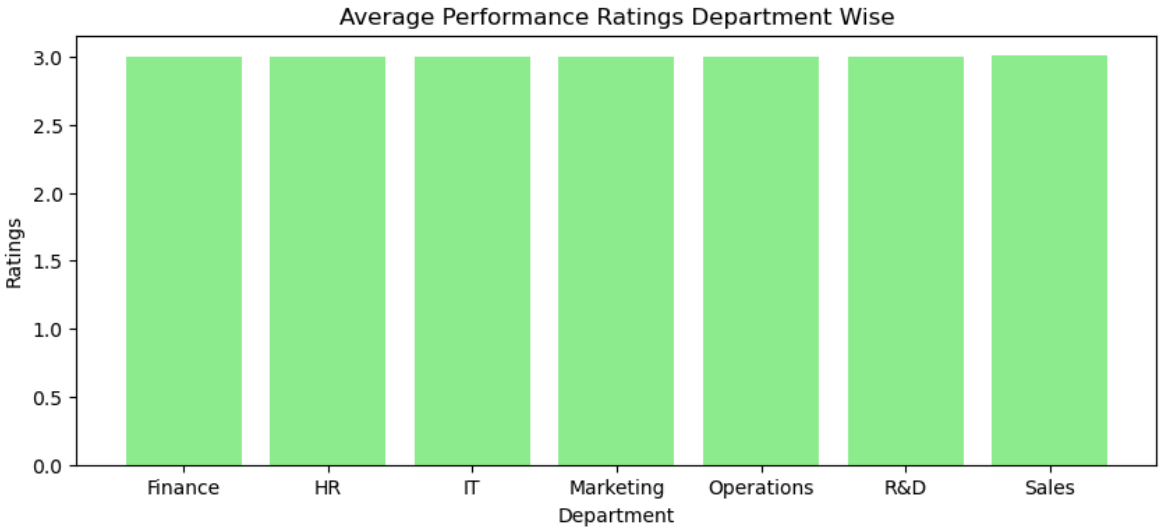
Out[619]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [620]:
```
PR=df.groupby('Department')['Performance_Rating'].mean()
PR
```

Out[620]:
```
Department
Finance       2.996818
HR            2.995670
IT            2.998216
Marketing     3.004736
Operations    2.996081
R&D           3.001885
Sales         3.006362
Name: Performance_Rating, dtype: float64
```

In [621]:
```
plt.figure(figsize=(10,4))
plt.bar(PR.index, PR.values, color='lightgreen')
plt.title("Average Performance Ratings Department Wise")
plt.ylabel("Ratings")
plt.xlabel("Department")
plt.show()
```



Q.10) Which Country have the highest concentration of employees?

In [623]:
```
df.head(2)
```

Out[623]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [624]:
```
df['Country']=df['Location'].apply (lambda x:str(x.split(',')[1]))
```

In [625]:
```
df.head()
```

Out[625]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Pe
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	
2	EMP0000003	Alyssa Martinez	HR	HR Manager	2023-03-20	Port Christinaport, Saudi Arabia	
3	EMP0000004	Nicholas Valdez	IT	Software Engineer	2023-10-12	Port Shelbychester, Antigua and Barbuda	
4	EMP0000005	Joel Hendricks	Operations	Logistics Coordinator	2024-12-09	Lake Kimberly, Palestinian Territory	

In [626]:
```
df.Country.nunique()
```

Out[626]: 243

In [627]:
```
df.Country.value_counts()
```

Out[627]:
```
Country
Congo                  16286
Korea                  16285
Sri Lanka               8409
Switzerland             8391
British Virgin Islands  8373
                       ...
Guinea-Bissau           7983
Kazakhstan              7973
Montenegro              7972
Bhutan                  7971
Palestinian Territory   7895
Name: count, Length: 243, dtype: int64
```

Q.11) Is there a correlation b/w performance rating and salary?

In [629]:
```
df.head(2)
```

Out[629]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [630]:
```
df['Performance_Rating'].corr(df['Salary_INR'])
```

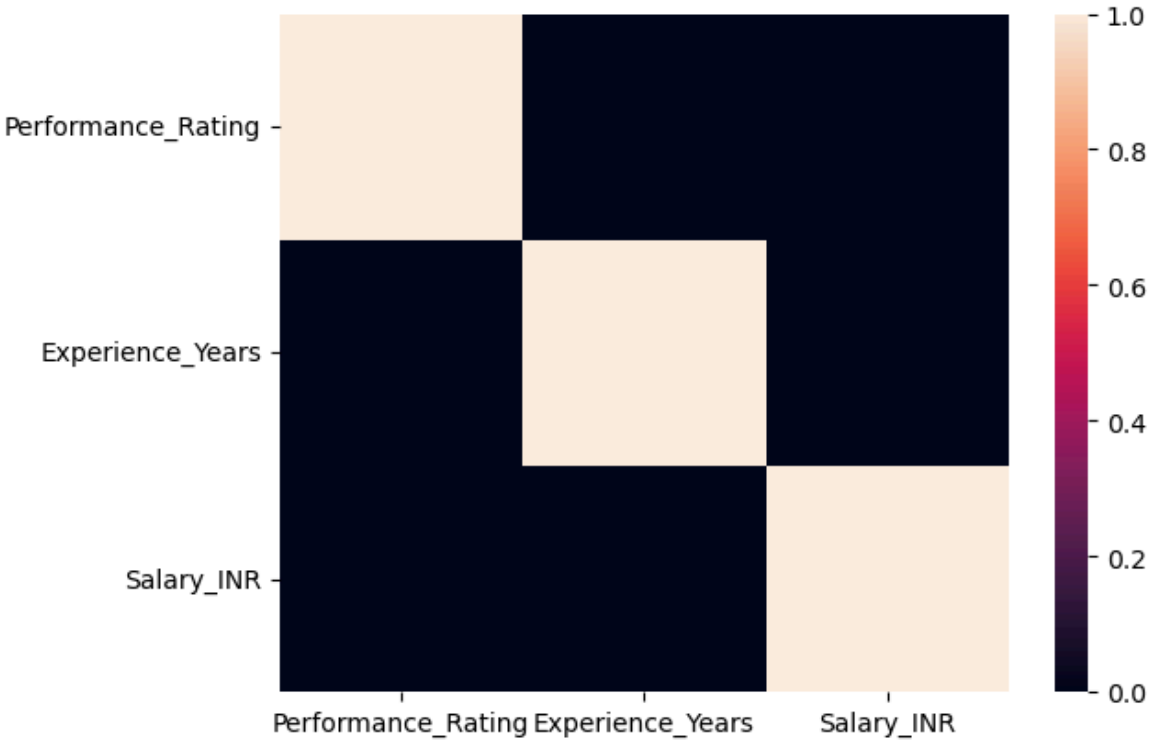
Out[630]: -0.00020919799940916222

In [631]:
```
# Alternated Command to show Correlation
df[['Performance_Rating', 'Salary_INR']].corr()
```

Out[631]:

	Performance_Rating	Salary_INR
Performance_Rating	1.000000	-0.000209
Salary_INR	-0.000209	1.000000

In [632]:
```
# Showing Corrrelation with heatmap
sns.heatmap(df.corr(numeric_only= True))
plt.show()
```



Q.12) How has the number of hires changed over time(per year)?

In [634]:
```
df.head(2)
```

Out[634]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Location	Perfor
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	Isaacland, Denmark	
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	Anthonyside, Costa Rica	

In [636]:
```
df.Hire_Date.dtype
```

Out[636]: dtype('<M8[ns]')

In [638]:
```
df.insert(5, 'Year',df['Hire_Date'].dt.year)
```

In [639]:
```
df.head()
```

Out[639]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Year	Locatic
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	2011	Isaaclan Denma
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	2018	Anthonysic Costa Ri
2	EMP0000003	Alyssa Martinez	HR	HR Manager	2023-03-20	2023	Pc Christinapo Saudi Arab
3	EMP0000004	Nicholas Valdez	IT	Software Engineer	2023-10-12	2023	Pc Shelbycheste Antigua ar Barbu
4	EMP0000005	Joel Hendricks	Operations	Logistics Coordinator	2024-12-09	2024	Lal Kimber Palestini Territo

In [643]:
```
df.Year.unique()
```

Out[643]:
```
array([2011, 2018, 2023, 2024, 2021, 2016, 2020, 2015, 2025, 2022, 2017,
       2019, 2014, 2013, 2012, 2010])
```

In [646]:
```
df.Year.nunique()
```

Out[646]: 16

In [649]:
```
hire=df.groupby('Year')['Employee_ID'].count()
hire
```

Out[649]:
```
Year
2010    15520
2011    40089
2012    39765
2013    39988
2014    40202
2015    85984
2016   160249
2017   160363
2018   159658
2019   160202
2020   175460
2021   199366
2022   201373
2023   198982
2024   200001
2025   122798
Name: Employee_ID, dtype: int64
```

In [652]:
```
plt.figure(figsize=(10,4))
hire.plot(x=hire.index,y=hire.values, kind='bar', color ='violet')
plt.grid(True, color='r')
plt.title("No. of Employee Hired in any Year")
plt.ylabel("Count")
plt.show()
```

Q.13) Compare Salaries of Remote vs. On-site employees- Is there a significant difference?

In [656]:
```
df.head(2)
```

Out[656]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Year	Location
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	2011	Isaacland, Denmark
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	2018	Anthonyside, Costa Rica

In [657]:
```
df.groupby('Work_Mode')['Salary_INR'].mean()
```

Out[657]:
```
Work_Mode
On-site    896835.945792
Remote     896965.326373
Name: Salary_INR, dtype: float64
```

Q.14) Find the top 10 employees with the highest salary in each dapertment.

In [659]:
```
df.head(2)
```

Out[659]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Year	Location
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	2011	Isaacland, Denmark
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	2018	Anthonyside, Costa Rica

In [660]:
```
top_10 = df.groupby('Department').apply(
    lambda x: x.nlargest(10, 'Salary_INR'),
    include_groups=False
)
```

In [661]:
```
top_10.head(10)
```

Out[661]:

		Employee_ID	Full_Name	Job_Title	Hire_Date	Year	Lo
Department							
Finance	**888712**	EMP0888713	Christopher Sloan	Finance Manager	2011-07-19	2011	East A Pc
	695808	EMP0695809	Todd Rodgers	Finance Manager	2019-12-27	2019	North
	459273	EMP0459274	Angela Payne	Finance Manager	2021-08-12	2021	Rave Isle c
	750893	EMP0750894	Nina Lara	Finance Manager	2021-10-19	2021	Christ Am
	780290	EMP0780291	Brittany Thompson	Finance Manager	2021-07-23	2021	Meliss Mart
	1316795	EMP1316796	Larry Wilson	Finance Manager	2015-01-30	2015	Lopezr Philj
	737507	EMP0737508	Alexis Schroeder	Finance Manager	2024-10-28	2024	Teresar Cam
	781352	EMP0781353	Sarah Jones	Finance Manager	2018-04-02	2018	South
	803785	EMP0803786	Jose Anderson	Finance Manager	2020-11-17	2020	Bryar R Fede
	905337	EMP0905338	Jennifer Dominguez	Finance Manager	2018-03-22	2018	Port Jer

In [662]:
```
top_10.tail(10)
```

Out[662]:

Department		Employee_ID	Full_Name	Job_Title	Hire_Date	Year	
Sales	**1729875**	EMP1729876	Hector Love	Business Development Manager	2020-01-02	2020	
	3493	EMP0003494	Tracy Hill	Business Development Manager	2017-07-23	2017	Ste
	161163	EMP0161164	Mark Mccann	Business Development Manager	2025-06-09	2025	Jer
	50430	EMP0050431	Christine Wood	Business Development Manager	2021-12-05	2021	
	339734	EMP0339735	Sarah Watson	Business Development Manager	2021-04-05	2021	
	86194	EMP0086195	Gabrielle Phelps	Business Development Manager	2015-11-23	2015	M Ce
	1116580	EMP1116581	Kimberly Mullen	Business Development Manager	2025-01-09	2025	
	1760918	EMP1760919	Christopher Farmer	Business Development Manager	2013-01-12	2013	
	1878661	EMP1878662	Margaret Gardner	Business Development Manager	2025-04-23	2025	G
	1333220	EMP1333221	Benjamin Jones	Business Development Manager	2017-11-14	2017	E

Q.15) Identify departments with the highest attrition rate (Resigned %)

In [664]:
```
df.head(2)
```

Out[664]:

	Employee_ID	Full_Name	Department	Job_Title	Hire_Date	Year	Location
0	EMP0000001	Joshua Nguyen	IT	Software Engineer	2011-08-10	2011	Isaacland, Denmark
1	EMP0000002	Julie Williams	Marketing	SEO Specialist	2018-03-02	2018	Anthonyside, Costa Rica

In [665]:
```
dept_counts=df.groupby('Department')['Status'].agg(total_emp='count',resigned=lambda x:(x==
dept_counts
```

Out[665]:

Department	total_emp	resigned
Finance	199873	40238
HR	159119	31736
IT	601042	119852
Marketing	240081	47793
Operations	300095	59397
R&D	99759	19919
Sales	400031	79725

In [666]:
```
type(dept_counts)
```

Out[666]: pandas.core.frame.DataFrame

In [669]:
```
# calculate attrition rate
dept_counts['attrition_rate_%']=(dept_counts['resigned']/dept_counts['total_emp'])*100
```

In [670]:
```
dept_counts
```

Out[670]:

Department	total_emp	resigned	attrition_rate_%
Finance	199873	40238	20.131784
HR	159119	31736	19.944821
IT	601042	119852	19.940703
Marketing	240081	47793	19.907031
Operations	300095	59397	19.792732
R&D	99759	19919	19.967121
Sales	400031	79725	19.929705

In [671]:
```
# Sort by attrition rate (highest first)
dept_counts.sort_values('attrition_rate_%',ascending=False)
```

Out[671]:

Department	total_emp	resigned	attrition_rate_%
Finance	199873	40238	20.131784
R&D	99759	19919	19.967121
HR	159119	31736	19.944821
IT	601042	119852	19.940703
Sales	400031	79725	19.929705
Marketing	240081	47793	19.907031
Operations	300095	59397	19.792732

** END - PROJECT ***

In [ ]: