---------------------------------Exploratory Data Analysis-------------------------------------

-------For this EDA data is taken from Kaggle, This data set is about HR---------

In [1]:

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [72]:

```python
hr = pd.read_csv("C:/Users/DELL/Desktop/Atomcamp Python/data_science.csv")
```

In [22]:

```python
hr.head(10)
```

Out[22]:

	Unnamed: 0	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio
0	0	2020	MI	FT	Data Scientist	70000	EUR	79833	DE	0
1	1	2020	SE	FT	Machine Learning Scientist	260000	USD	260000	JP	0
2	2	2020	SE	FT	Big Data Engineer	85000	GBP	109024	GB	50
3	3	2020	MI	FT	Product Data Analyst	20000	USD	20000	HN	0
4	4	2020	SE	FT	Machine Learning Engineer	150000	USD	150000	US	50
5	5	2020	EN	FT	Data Analyst	72000	USD	72000	US	100
6	6	2020	SE	FT	Lead Data Scientist	190000	USD	190000	US	100
7	7	2020	MI	FT	Data Scientist	11000000	HUF	35735	HU	50
8	8	2020	MI	FT	Business Data Analyst	135000	USD	135000	US	100
9	9	2020	SE	FT	Lead Data Engineer	125000	USD	125000	NZ	50

In [23]:

```python
hr.shape
```

Out[23]:

```
(607, 12)
```

In [34]:

```python
hr.drop(['Unnamed: 0', 'salary'], axis= 1, inplace= True)
```

In [35]:

```
hr
```

Out[35]:

work_year	experience_level	employment_type	job_title	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_
2020	MI	FT	Data Scientist	EUR	79833	DE	0	DE	
2020	SE	FT	Machine Learning Scientist	USD	260000	JP	0	JP	
2020	SE	FT	Big Data Engineer	GBP	109024	GB	50	GB	
2020	MI	FT	Product Data Analyst	USD	20000	HN	0	HN	
2020	SE	FT	Machine Learning Engineer	USD	150000	US	50	US	
...	...	...	...	...	...	...	...	...	
2022	SE	FT	Data Engineer	USD	154000	US	100	US	
2022	SE	FT	Data Engineer	USD	126000	US	100	US	
2022	SE	FT	Data Analyst	USD	129000	US	0	US	
2022	SE	FT	Data Analyst	USD	150000	US	100	US	
2022	MI	FT	AI Scientist	USD	200000	IN	100	US	

ws × 10 columns

Average Salary by each job title

In [41]:

```
age_salaries = hr.groupby('job_title')['salary_in_usd'].mean().reset_index().sort_values(by='salary_in_usd', ascending=Fa
```

In [42]:

```
average_salaries
```

Out[42]:

	job_title	salary_in_usd
14	Data Analytics Lead	405000.000000
45	Principal Data Engineer	328333.333333
28	Financial Data Analyst	275000.000000
46	Principal Data Scientist	215242.428571
25	Director of Data Science	195074.000000
16	Data Architect	177873.909091
3	Applied Data Scientist	175655.000000
2	Analytics Engineer	175000.000000
23	Data Specialist	165000.000000
29	Head of Data	160162.600000
41	Machine Learning Scientist	158412.500000
21	Data Science Manager	158328.500000
24	Director of Data Engineering	156738.000000
30	Head of Data Science	146718.750000
4	Applied Machine Learning Scientist	142068.750000
33	Lead Data Engineer	139724.500000
15	Data Analytics Manager	127134.285714
9	Cloud Data Engineer	124647.000000
18	Data Engineering Manager	123227.200000
44	Principal Data Analyst	122500.000000
36	ML Engineer	117504.000000
40	Machine Learning Manager	117104.000000
34	Lead Data Scientist	115190.000000
17	Data Engineer	112725.000000
48	Research Scientist	109019.500000
22	Data Scientist	108187.832168
11	Computer Vision Software Engineer	105248.666667
49	Staff Data Scientist	105000.000000
38	Machine Learning Engineer	104880.146341
39	Machine Learning Infrastructure Engineer	101145.000000
6	Big Data Architect	99703.000000
12	Data Analyst	92893.061856
32	Lead Data Analyst	92203.000000
42	Marketing Data Analyst	88654.000000
35	Lead Machine Learning Engineer	87932.000000
37	Machine Learning Developer	85860.666667
31	Head of Machine Learning	79039.000000
8	Business Data Analyst	76691.200000
20	Data Science Engineer	75803.333333
5	BI Data Analyst	74755.166667
19	Data Science Consultant	69420.714286
1	AI Scientist	66135.571429
13	Data Analytics Engineer	64799.250000
27	Finance Data Analyst	61896.000000
26	ETL Developer	54957.000000
7	Big Data Engineer	51974.000000
10	Computer Vision Engineer	44419.333333
43	NLP Engineer	37236.000000
47	Product Data Analyst	13036.000000
0	3D Computer Vision Researcher	5409.000000

Top 10 job titles by salary

In [43]:

```
top_10_salaries = average_salaries.sort_values(by='salary_in_usd', ascending=False).head(10)
```

In [44]:

```
top_10_salaries
```

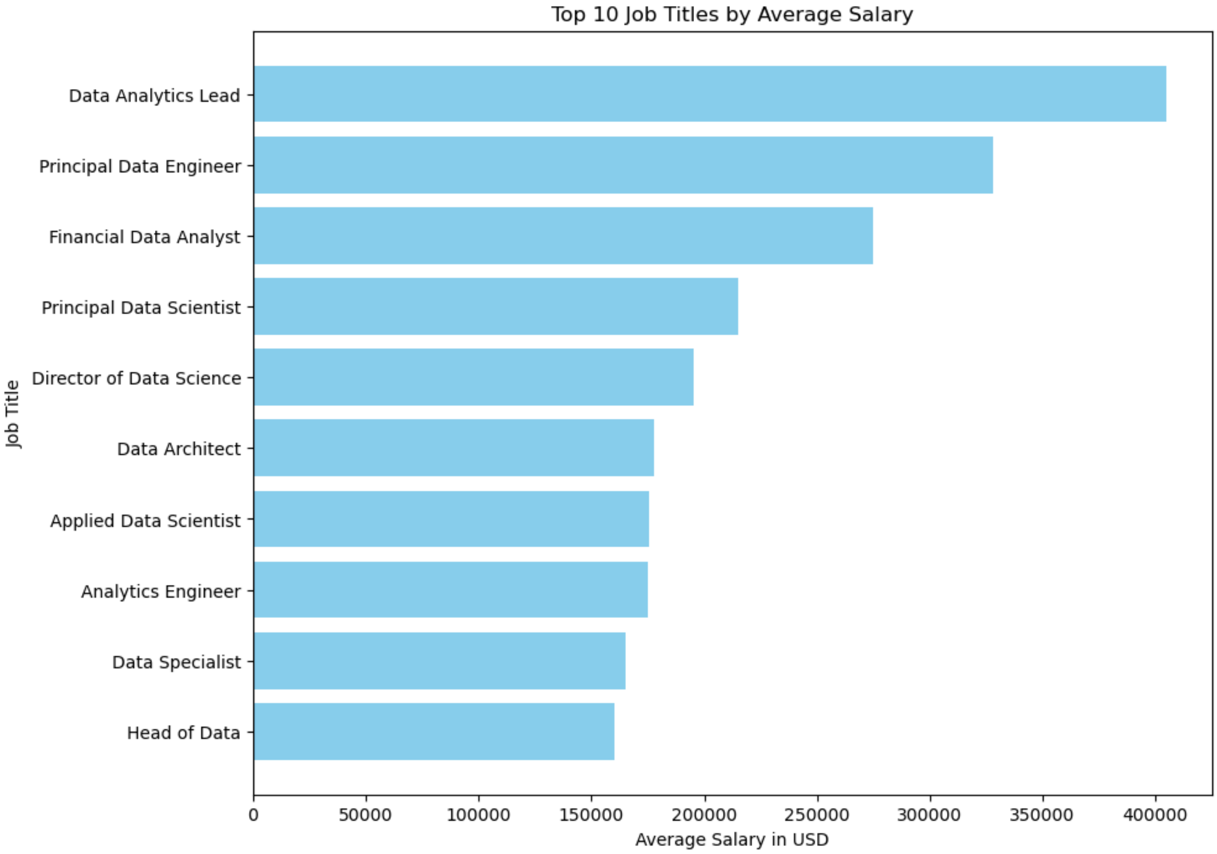
Out[44]:

	job_title	salary_in_usd
14	Data Analytics Lead	405000.000000
45	Principal Data Engineer	328333.333333
28	Financial Data Analyst	275000.000000
46	Principal Data Scientist	215242.428571
25	Director of Data Science	195074.000000
16	Data Architect	177873.909091
3	Applied Data Scientist	175655.000000
2	Analytics Engineer	175000.000000
23	Data Specialist	165000.000000
29	Head of Data	160162.600000

Bar Chart for top 10 job title by salary

In [45]:

```
plt.figure(figsize=(10, 8))
plt.barh(top_10_salaries['job_title'], top_10_salaries['salary_in_usd'], color='skyblue')
plt.xlabel('Average Salary in USD')
plt.ylabel('Job Title')
plt.title('Top 10 Job Titles by Average Salary')
plt.gca().invert_yaxis()
plt.show()
```



Ratio of remote employees based on company size

In [57]:

```
remt_ratio = hr.groupby('company_size')['remote_ratio'].mean().reset_index().round(2)
```

In [58]:

```
remt_ratio
```

Out[58]:

	company_size	remote_ratio
0	L	68.69
1	M	72.55
2	S	69.88

In [62]:

```python
work_years = hr.groupby('work_year')['salary_in_usd'].mean().reset_index().round(2)
```

In [63]:

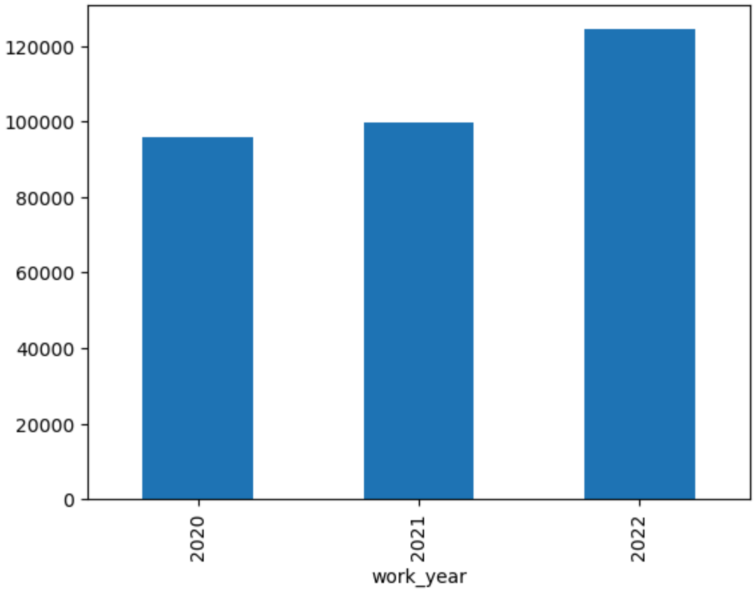
```python
work_years
```

Out[63]:

	work_year	salary_in_usd
0	2020	95813.00
1	2021	99853.79
2	2022	124522.01

In [69]:

```python
work_years = hr.groupby('work_year')['salary_in_usd'].mean().round(2)
work_years.plot(kind='bar')
plt.show()
```



In [76]:

```python
df3 = hr.company_size.value_counts()
df3
```

Out[76]:

```
company_size
M    326
L    198
S     83
Name: count, dtype: int64
```

In [77]:

```python
df3.index.to_list()
```

Out[77]:

```
['M', 'L', 'S']
```

In [82]:

```python
values2 = df3.to_list()
values2
```

Out[82]:

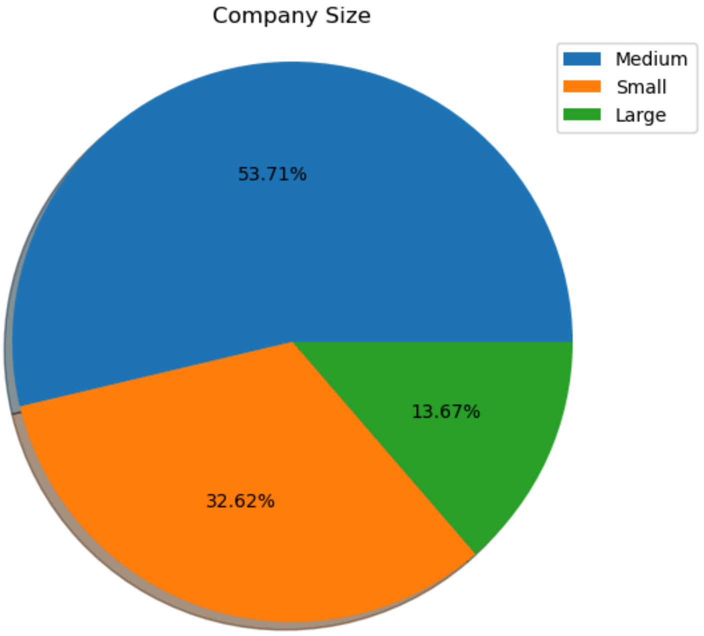
```
[326, 198, 83]
```

In [88]:

```python
labels_for_company = ['Medium','Small','Large']
```

In [92]:

```
plt.figure(figsize=(8,6))
plt.pie(x=values2, labels=None, autopct='%1.2f%%', shadow=True)
plt.legend(labels_for_company, loc = 'upper right')
plt.axis('equal')
plt.title('Company Size')
plt.show()
```



In [99]:

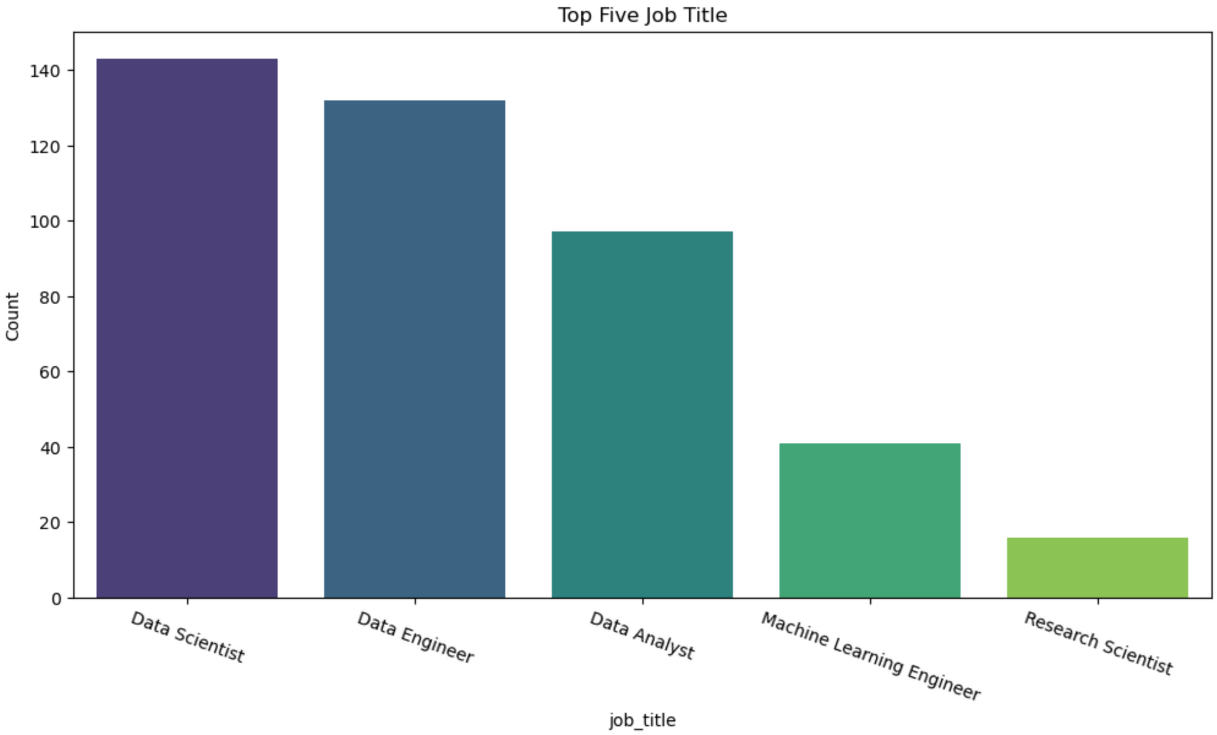
```
df4= hr.job_title.value_counts().head(5)
df4
```

Out[99]:

```
job_title
Data Scientist              143
Data Engineer               132
Data Analyst                 97
Machine Learning Engineer    41
Research Scientist           16
Name: count, dtype: int64
```

In [104]:

```
plt.figure(figsize=(12,6))
sns.barplot(x=df4.index, y=df4.values,palette='viridis')
plt.title('Top Five Job Title')
plt.ylabel('Count')
plt.xticks(rotation = -20)
plt.show()
```



In [142]:

```
df5= hr.experience_level.value_counts()
df5
```

Out[142]:

```
experience_level
Senior level       280
Middle Level       213
Entry Level         88
Executive Level     26
Name: count, dtype: int64
```

In [144]:

```
exp_map = {
    'SE': 'Senior level',
    'MI': 'Middle Level',
    'EN': 'Entry Level',
    'EX': 'Executive Level'
}
```

In [150]:

```
df5.index.to_list()
```

Out[150]:

```
['Senior level', 'Middle Level', 'Entry Level', 'Executive Level']
```

In [145]:

```

hr['experience_level'].replace(exp_map, inplace=True)
```

In [151]:

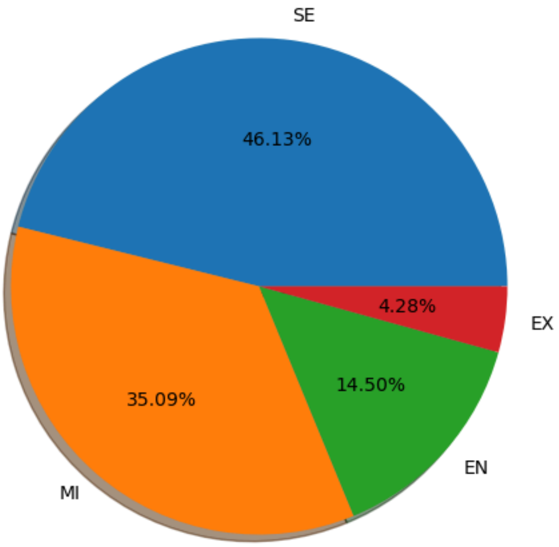
```
values = df5.values
values
```

Out[151]:

```
array([280, 213,  88,  26], dtype=int64)
```

In [155]:

```
plt.figure(figsize=(6,12))
plt.pie(x = values, labels= labels, autopct = '%1.2f%%',shadow=True)
plt.show()
```



In [ ]:

```

```