Salary Analysis of Data Jobs

Welcome to the **Salary Analysis of Data Jobs** task ! This task provides valuable insights into the compensation and job roles of employees across various industries and regions. Whether you're an HR analyst, data scientist, or someone interested in understanding salary trends, this task offers a wealth of information to explore and analyze.

Content:

The dataset contains the following fields:

work_year: The year of employment.
experience_level: The experience level of the employee (e.g., entry-level, mid-level, senior).
employment_type: The type of employment (e.g., full-time, part-time, contract).
job_title: The job title or position of the employee within the company.
salary: The salary amount in the local currency.
salary_currency: The currency in which the salary is denoted.
salary_in_usd: The equivalent salary amount in USD (United States Dollars).
employee_residence: The location of the employee's residence.
remote_ratio: The percentage of remote work allowed for the position.
company_location: The location of the company.
company_size: The size of the company (e.g., small, medium, large).

Objectives

This can be utilized for various purposes, including but not limited to:

Analyzing salary trends across different job titles and experience levels.

Investigating the impact of remote work on compensation.

Comparing salary levels between full-time and part-time employment.

Understanding the correlation between company size and employee salaries.

Predictive analysis for forecasting salaries based on experience and job roles.

Work Flow

1. Gathering the data
2. Cleaning and Transformation of data
3. Conducting Exploratory data analysis
4. Data Visualization
5. Conducting Forecasting for Salary
6. Creating a Document for showcasing the task

Skip towards the next page for Python code and output

Here is the code:

Importing the libraries and reading the data

```
#Importing Libraries and Reading the data
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df=pd.read_csv('/content/ds_sals.csv')
df
```

	Emp_id	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	c
0	0	2020	MI	FT	Data Scientist	70000	EUR	79833	DE	0	DE	
1	1	2020	SE	FT	Machine Learning Scientist	260000	USD	260000	JP	0	JP	
2	2	2020	SE	FT	Big Data Engineer	85000	GBP	109024	GB	50	GB	
3	3	2020	MI	FT	Product Data Analyst	20000	USD	20000	HN	0	HN	
4	4	2020	SE	FT	Machine Learning Engineer	150000	USD	150000	US	50	US	
...	...	...	...	...	...	...	...	...	...	...	...	
602	602	2022	SE	FT	Data Engineer	154000	USD	154000	US	100	US	
603	603	2022	SE	FT	Data Engineer	126000	USD	126000	US	100	US	
604	604	2022	SE	FT	Data Analyst	129000	USD	129000	US	0	US	

Knowing about the dataset

```
[2] print(df.info())
    print(df.shape)
    print(df.size)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 607 entries, 0 to 606
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Emp_id              607 non-null    int64
 1   work_year           607 non-null    int64
 2   experience_level    607 non-null    object
 3   employment_type     607 non-null    object
 4   job_title           607 non-null    object
 5   salary              607 non-null    int64
 6   salary_currency     607 non-null    object
 7   salary_in_usd       607 non-null    int64
 8   employee_residence  607 non-null    object
 9   remote_ratio        607 non-null    int64
 10  company_location    607 non-null    object
 11  company_size        607 non-null    object
dtypes: int64(5), object(7)
memory usage: 57.0+ KB
None
(607, 12)
7284
```

```
df.describe()
```

	Emp_id	work_year	salary	salary_in_usd	remote_ratio
count	607.000000	607.000000	6.070000e+02	607.000000	607.00000
mean	303.000000	2021.405272	3.240001e+05	112297.869852	70.92257
std	175.370085	0.692133	1.544357e+06	70957.259411	40.70913
min	0.000000	2020.000000	4.000000e+03	2859.000000	0.00000
25%	151.500000	2021.000000	7.000000e+04	62726.000000	50.00000
50%	303.000000	2022.000000	1.150000e+05	101570.000000	100.00000
75%	454.500000	2022.000000	1.650000e+05	150000.000000	100.00000
max	606.000000	2022.000000	3.040000e+07	600000.000000	100.00000

EDA

```
[4] print(df['experience_level'].unique())
    print('------------------------------\n------------------------------')
    print(df['employment_type'].unique())
    print('------------------------------\n------------------------------')
    print(df['job_title'].unique())
    print('------------------------------\n------------------------------')
    print(df['salary_currency'].unique())
    print('------------------------------\n------------------------------')
    print(df['employee_residence'].unique())
```

```
['MI' 'SE' 'EN' 'EX']
------------------------------
------------------------------
['FT' 'CT' 'PT' 'FL']
------------------------------
------------------------------
['Data Scientist' 'Machine Learning Scientist' 'Big Data Engineer'
 'Product Data Analyst' 'Machine Learning Engineer' 'Data Analyst'
 'Lead Data Scientist' 'Business Data Analyst' 'Lead Data Engineer'
 'Lead Data Analyst' 'Data Engineer' 'Data Science Consultant'
 'BI Data Analyst' 'Director of Data Science' 'Research Scientist'
 'Machine Learning Manager' 'Data Engineering Manager'
 'Machine Learning Infrastructure Engineer' 'ML Engineer' 'AI Scientist'
 'Computer Vision Engineer' 'Principal Data Scientist'
 'Data Science Manager' 'Head of Data' '3D Computer Vision Researcher'
 'Data Analytics Engineer' 'Applied Data Scientist'
 'Marketing Data Analyst' 'Cloud Data Engineer' 'Financial Data Analyst'
 'Computer Vision Software Engineer' 'Director of Data Engineering'
 'Data Science Engineer' 'Principal Data Engineer'
 'Machine Learning Developer' 'Applied Machine Learning Scientist'
 'Data Analytics Manager' 'Head of Data Science' 'Data Specialist'
 'Data Architect' 'Finance Data Analyst' 'Principal Data Analyst'
 'Big Data Architect' 'Staff Data Scientist' 'Analytics Engineer'
 'ETL Developer' 'Head of Machine Learning' 'NLP Engineer'
 'Lead Machine Learning Engineer' 'Data Analytics Lead']
------------------------------
------------------------------
['EUR' 'USD' 'GBP' 'HUF' 'INR' 'JPY' 'CNY' 'MXN' 'CAD' 'DKK' 'PLN' 'SGD'
 'CLP' 'BRL' 'TRY' 'AUD' 'CHF']
------------------------------
------------------------------
['DE' 'JP' 'GB' 'HN' 'US' 'HU' 'NZ' 'FR' 'IN' 'PK' 'PL' 'PT' 'CN' 'GR'
 'AE' 'NL' 'MX' 'CA' 'AT' 'NG' 'PH' 'ES' 'DK' 'RU' 'IT' 'HR' 'BG' 'SG'
 'BR' 'IQ' 'VN' 'BE' 'UA' 'MT' 'CL' 'RO' 'IR' 'CO' 'MD' 'KE' 'SI' 'HK'
 'TR' 'RS' 'PR' 'LU' 'JE' 'CZ' 'AR' 'DZ' 'TN' 'MY' 'EE' 'AU' 'BO' 'IE'
 'CH']
```

```
[5] #Job titles ML enginer and Machine Learning engineer have no difference so we should replace Machine Engineers with ML engineer
    df['job_title'].replace('Machine Learning Engineer','ML Engineer',inplace=True)
```

```
[6] print(df['company_location'].unique())
    print('------------------------------\n------------------------------')
    print(df['company_size'].unique())
```

```
['DE' 'JP' 'GB' 'HN' 'US' 'HU' 'NZ' 'FR' 'IN' 'PK' 'CN' 'GR' 'AE' 'NL'
 'MX' 'CA' 'AT' 'NG' 'ES' 'PT' 'DK' 'IT' 'HR' 'LU' 'PL' 'SG' 'RO' 'IQ'
 'BR' 'BE' 'UA' 'IL' 'RU' 'MT' 'CL' 'IR' 'CO' 'MD' 'KE' 'SI' 'CH' 'VN'
 'AS' 'TR' 'CZ' 'DZ' 'EE' 'MY' 'AU' 'IE']
------------------------------
------------------------------
['L' 'S' 'M']
```

```
[7] df['job_title'].value_counts()
    #There are more data Scientists >> Data Engineers >> Data Analysts >> ML Engineers
```

job_title	count
Data Scientist	143
Data Engineer	132
Data Analyst	97
ML Engineer	47
Research Scientist	16
Data Science Manager	12
Data Architect	11
Big Data Engineer	8
Machine Learning Scientist	8
Data Analytics Manager	7
Principal Data Scientist	7

AI Scientist	7
Data Science Consultant	7
Director of Data Science	7
Computer Vision Engineer	6
BI Data Analyst	6
Lead Data Engineer	6
Data Engineering Manager	5
Business Data Analyst	5
Head of Data	5
Applied Data Scientist	5
Applied Machine Learning Scientist	4
Head of Data Science	4
Analytics Engineer	4
Data Analytics Engineer	4
Machine Learning Developer	3
Machine Learning Infrastructure Engineer	3
Lead Data Scientist	3
Computer Vision Software Engineer	3
Lead Data Analyst	3
Data Science Engineer	3

Finding the average salary of employees from each job

```
#Now finding the avg salary of each role from the given Dataset
df.groupby('job_title')['salary_in_usd'].mean().sort_values(ascending=False)
```

job_title	salary_in_usd
Data Analytics Lead	405000.000000
Principal Data Engineer	328333.333333
Financial Data Analyst	275000.000000
Principal Data Scientist	215242.428571
Director of Data Science	195074.000000
Data Architect	177873.909091
Applied Data Scientist	175655.000000
Analytics Engineer	175000.000000
Data Specialist	165000.000000
Head of Data	160162.600000
Machine Learning Scientist	158412.500000
Data Science Manager	158328.500000
Director of Data Engineering	156738.000000
Head of Data Science	146718.750000
Applied Machine Learning Scientist	142068.750000
Lead Data Engineer	139724.500000
Data Analytics Manager	127134.285714
Cloud Data Engineer	124647.000000
Data Engineering Manager	123227.200000
Principal Data Analyst	122500.000000
Machine Learning Manager	117104.000000
Lead Data Scientist	115190.000000
Data Engineer	112725.000000
Research Scientist	109019.500000

Data Scientist	108187.832168
ML Engineer	106491.702128
Computer Vision Software Engineer	105248.666667
Staff Data Scientist	105000.000000
Machine Learning Infrastructure Engineer	101145.000000
Big Data Architect	99703.000000
Data Analyst	92893.061856
Lead Data Analyst	92203.000000
Marketing Data Analyst	88654.000000
Lead Machine Learning Engineer	87932.000000
Machine Learning Developer	85860.666667
Head of Machine Learning	79039.000000
Business Data Analyst	76691.200000
Data Science Engineer	75803.333333
BI Data Analyst	74755.166667
Data Science Consultant	69420.714286
AI Scientist	66135.571429
Data Analytics Engineer	64799.250000
Finance Data Analyst	61896.000000
ETL Developer	54957.000000

Now finding the average salaries of Data Analysts by experience levels

```
#Data Analyst Salaries for different job experiences
df[df['job_title']=='Data Analyst'].groupby('experience_level')['salary_in_usd'].mean()
```

experience_level	salary_in_usd
EN	53960.666667
EX	120000.000000
MI	71699.206897
SE	111922.629630

dtype: float64

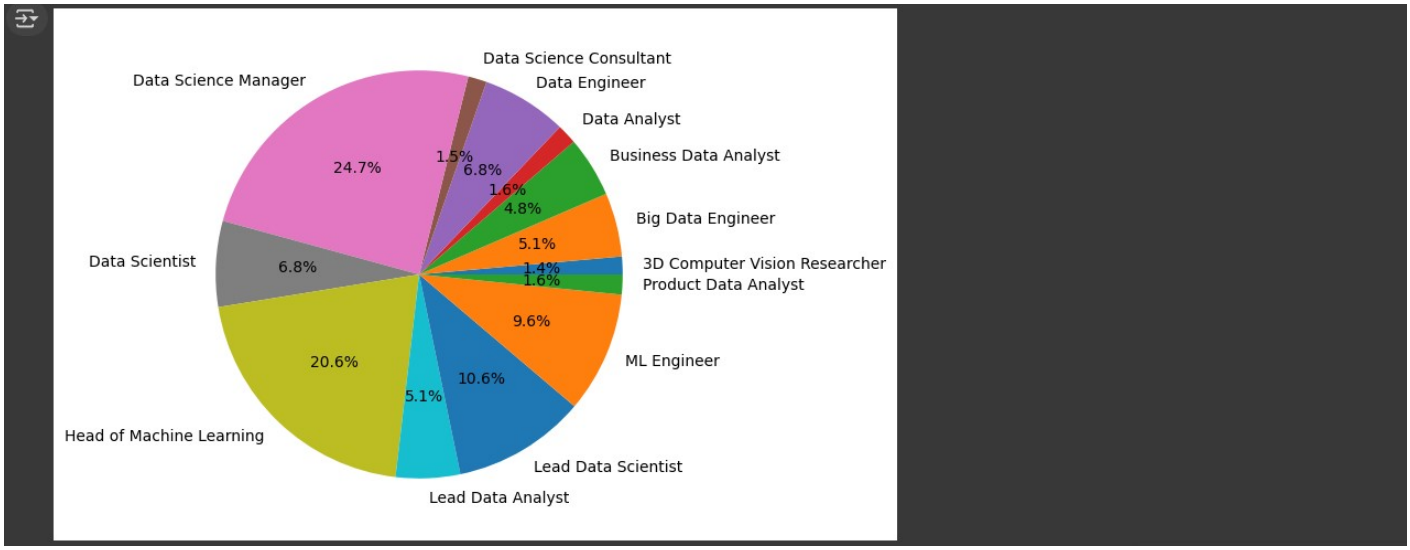
```
#now finding the the average salaries for ML Engineers different company sizes in India
df[(df['job_title']=='ML Engineer') & (df['company_location']=='IN')].groupby('company_size')['salary_in_usd'].mean()
#seems like there are no medium sized companyies related to ML engineers in India In 2020,2021,2022
```

company_size	salary_in_usd
L	45303.5
S	20000.0

dtype: float64

Now finding how much each job role is contributing the India of all job experiences

```
#Total avg Revenue of Indians in all jobs of all all experineces
india=df[df['company_location']=='IN'].groupby('job_title')['salary_in_usd'].mean().reset_index()
india.columns=['Job','Salary(USD)']
plt.figure(figsize=(12,6))
plt.pie(india['Salary(USD)'],labels=india['Job'],autopct='%1.1f%%')
plt.show()
```



Now we are retrieving all the Indians data who are working in other countries

```
[12] #Now we are retriving all the indians who are doing their job in other countries
     df[(df['employee_residence']=='IN')&(df['company_location']!='IN')]
```

	Emp_id	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	c
73	73	2021	EX	FT	BI Data Analyst	150000	USD	150000	IN	100	US	
179	179	2021	MI	FT	Data Scientist	420000	INR	5679	IN	100	US	
198	198	2021	SE	FT	Data Science Manager	4000000	INR	54094	IN	50	US	
213	213	2021	EN	FT	Big Data Engineer	435000	INR	5882	IN	0	CH	
244	244	2021	EN	FT	AI Scientist	1335000	INR	18053	IN	100	AS	
606	606	2022	MI	FT	AI Scientist	200000	USD	200000	IN	100	US	

The average AI Scientist salary over the years

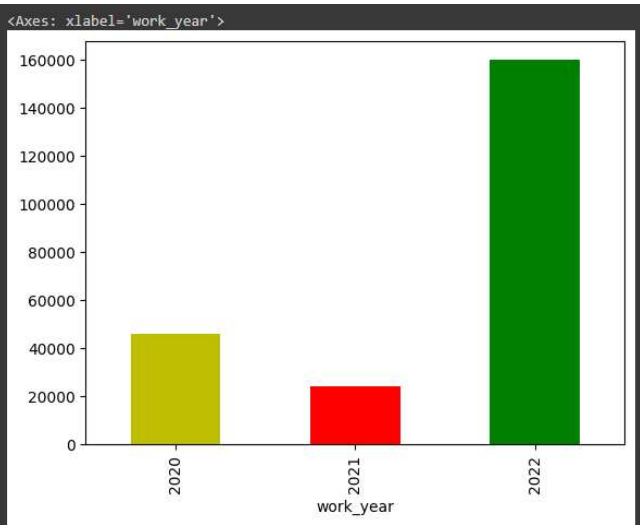
```
#average salaries of AI engineers over the years
df[df['job_title']=='AI Scientist'].groupby('work_year')['salary_in_usd'].mean()
```

work_year	salary_in_usd
2020	45896.00
2021	24263.25
2022	160000.00

dtype: float64

```
[14] df[df['job_title']=='AI Scientist'].groupby('work_year')['salary_in_usd'].mean().plot(
         kind='bar',
         x='work_year',
         y='salary_in_usd',
         color=['y','r','g'],
         )
```

```
<Axes: xlabel='work_year'>
```



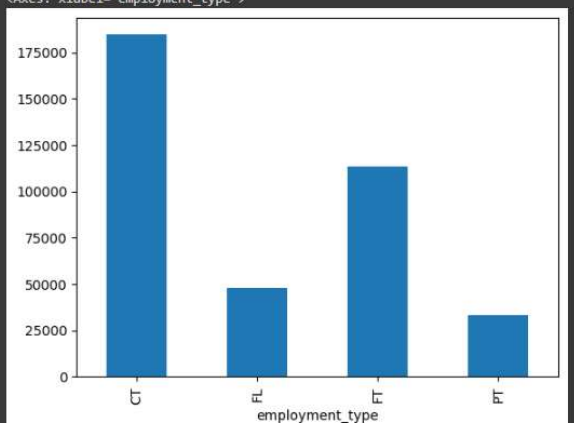
Now finding the average salary of different employee types (Part time, Full time)

```
[15] #Now We try to retrive the average salary of all kinds of job experiences and all jobs by Job type
     df.groupby('employment_type')['salary_in_usd'].mean()
```

employment_type	salary_in_usd
CT	184575.000000
FL	48000.000000
FT	113468.073129
PT	33070.500000

dtype: float64

```
[16] df.groupby('employment_type')['salary_in_usd'].mean().plot(kind='bar')
<Axes: xlabel='employment_type'>
```



Now finding the average salary of part time and full time data analysts

```
[17] #What is the average salary of part time data analysts
     df[(df['job_title']=='Data Analyst') & (df['employment_type']=='PT')]['salary_in_usd'].mean()
10354.0

     #What is the average salary of full  time time data analysts
     df[(df['job_title']=='Data Analyst') & (df['employment_type']=='FT')]['salary_in_usd'].mean()
93752.84375
```

Now finding the average salary of data scientists across the globe

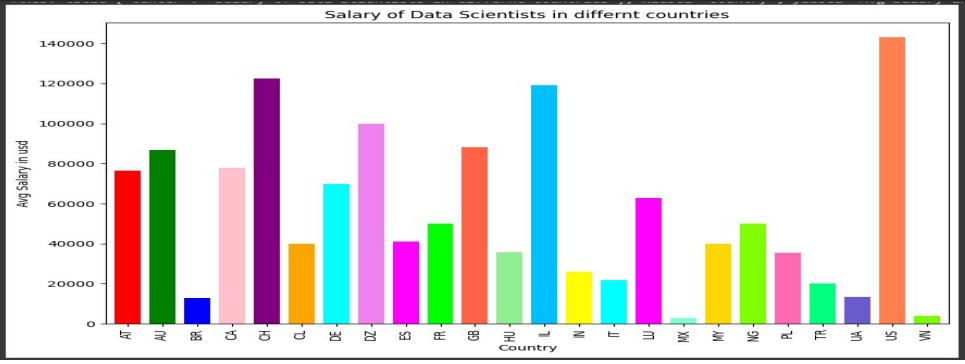
```
#comparison of average salary of Data analysts in differnt countries
df[df['job_title']=='Data Scientist'].groupby('company_location')['salary_in_usd'].mean().sort_values(ascending=False)
```

company_location	salary_in_usd
US	143115.678571
CH	122346.000000
IL	119059.000000
DZ	100000.000000
GB	88177.363636
AU	86703.000000
CA	77787.000000
AT	76352.000000
DE	69640.142857
LU	62726.000000
FR	50085.571429
NG	50000.000000
ES	41136.666667
CL	40038.000000
MY	40000.000000
HU	35735.000000
PL	35590.000000
IN	26108.250000
IT	21669.000000
TR	20171.000000
UA	13400.000000
BR	12901.000000
VN	4000.000000
MX	2859.000000

dtype: float64

Plotting the same using bar graph

```
[20] df[df['job_title']=='Data Scientist'].groupby('company_location')['salary_in_usd'].mean().plot(
        kind='bar',
        xlabel='Country',
        ylabel='Avg Salary in usd',
        title='Salary of Data Scientists in differnt countries',
        width=0.75,
        figsize=(10,7),
        color = [
        'red', 'green', 'blue', 'pink', 'purple', 'orange',
        'cyan', 'violet', 'magenta', 'lime', 'tomato', 'lightgreen',
        'deepskyblue', 'yellow', 'aqua', 'fuchsia', 'aquamarine', 'gold',
        'chartreuse', 'hotpink', 'springgreen', 'slateblue', 'coral', 'lawngreen']
    )
```



Remote ratio analysis

```
[21] #Remote ratio analysis
     df['remote_ratio'].value_counts()
     # 0- Complete job on sight
     #50 - Hybrid (some job in Company and some in Home)
     #100 percent remote  (work from home)
```

remote_ratio	count
100	381
0	127
50	99

dtype: int64

Info about all employees who does work from home from their own countries for overseas companies

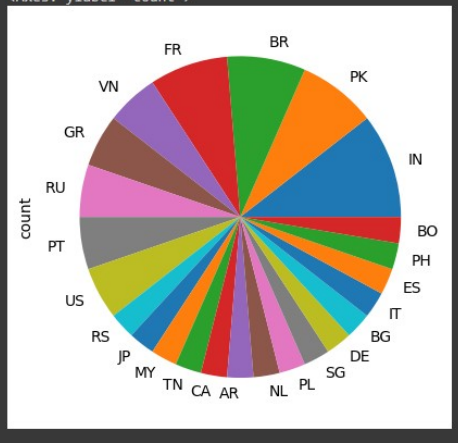
```
df[(df['remote_ratio']==100) & (df['employee_residence']!=df['company_location'])]
#This is the info of the people who are living in thier own countries and doing work from home for oversea companies
```

	Emp_id	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_locatio
17	17	2020	SE	FT	Big Data Engineer	100000	EUR	114047	PL	100	G
19	19	2020	MI	FT	Lead Data Engineer	56000	USD	56000	PT	100	U
32	32	2020	SE	FT	Data Scientist	60000	EUR	68428	GR	100	U
40	40	2020	MI	FT	Data Scientist	45760	USD	45760	PH	100	U
53	53	2020	EN	FT	Data Engineer	48000	EUR	54742	PK	100	D
54	54	2020	SE	FL	Computer Vision Engineer	60000	USD	60000	RU	100	U
61	61	2020	MI	FT	Data Engineer	130800	USD	130800	ES	100	U
73	73	2021	EX	FT	BI Data Analyst	150000	USD	150000	IN	100	U
84	84	2021	EX	FT	Director of Data Science	130000	EUR	153667	IT	100	P

Pie plot showing how many employees from each country doing WFH from their own country to overseas company

```
[23] df[(df['remote_ratio']==100) & (df['employee_residence']!=df['company_location'])]['employee_residence'].value_counts().plot(kind='pie')
```

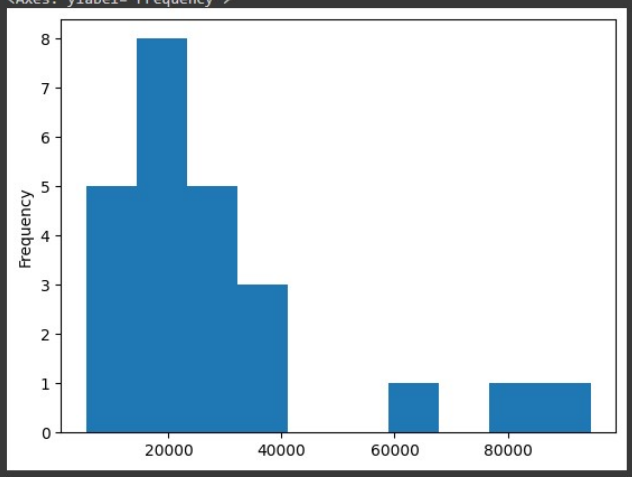
<Axes: ylabel='count'>



Average salary analysis of all jobs and all experiences by histogram

```
[ ] df[df['company_location']=='IN']['salary_in_usd'].plot(kind='hist',bins=10)
```

```
<Axes: ylabel='Frequency'>
```



Adding the conditional column for remote ratio

```
[25] df['job_kind']=np.where(df['remote_ratio']==100,'WFH',np.where(df['remote_ratio']==50,'Hybrid','OnSight'))
     df
```

erience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size	job_kind
MI	FT	Data Scientist	70000	EUR	79833	DE	0	DE	L	OnSight
SE	FT	Machine Learning Scientist	260000	USD	260000	JP	0	JP	S	OnSight
SE	FT	Big Data Engineer	85000	GBP	109024	GB	50	GB	M	Hybrid
MI	FT	Product Data Analyst	20000	USD	20000	HN	0	HN	S	OnSight
SE	FT	ML Engineer	150000	USD	150000	US	50	US	L	Hybrid
...	...	...	...	...	...	...	...	...	...	...
SE	FT	Data Engineer	154000	USD	154000	US	100	US	M	WFH
SE	FT	Data Engineer	126000	USD	126000	US	100	US	M	WFH
SE	FT	Data Analyst	129000	USD	129000	US	0	US	M	OnSight
SE	FT	Data Analyst	150000	USD	150000	US	100	US	M	WFH
MI	FT	AI	200000	USD	200000	IN	100	US	L	WFH

Now we will find how many employees from each job working in different types (WFH, Hybrid, OnSight)

```
[26]
```

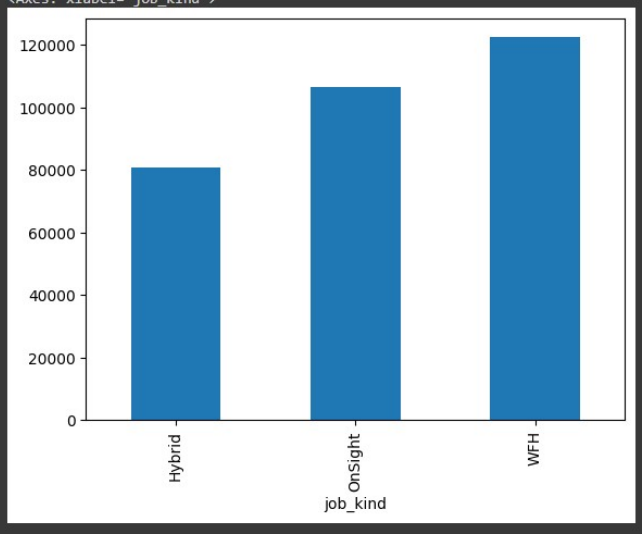
job_title	job_kind	count
Data Engineer	WFH	93
Data Scientist	WFH	79
Data Analyst	WFH	70
Data Scientist	OnSight	39
Data Engineer	OnSight	27
ML Engineer	WFH	26
Data Scientist	Hybrid	25

Data Engineer	OnSight	27
ML Engineer	WFH	26
Data Scientist	Hybrid	25
Data Analyst	OnSight	21
ML Engineer	Hybrid	13
Data Engineer	Hybrid	12
Data Architect	WFH	11
Data Science Manager	WFH	9
Research Scientist	Hybrid	9
ML Engineer	OnSight	8
Data Analytics Manager	WFH	6
Data Analyst	Hybrid	6
Principal Data Scientist	WFH	6
AI Scientist	WFH	5
Business Data Analyst	WFH	4
Data Science Consultant	WFH	4
Machine Learning Scientist	WFH	4
Research Scientist	WFH	4
Head of Data	WFH	4
Data Analytics Engineer	WFH	3

Now we will find the average salary for each job kind

```
#Now we will see the average salary of all job kinds of all jobs
df.groupby('job_kind')['salary_in_usd'].mean().plot(kind='bar')
```

```
<Axes: xlabel='job_kind'>
```



Now we will forecast the Average salary of Data analysts across the globe depending on experience and year
Preparing data first

```
#Now forecasting the salary of data analysts for different experiences accross the globe
#first we should know the data
forecast_data=df[df['job_title']=='Data Analyst'].groupby(['work_year','experience_level'])['salary_in_usd'].mean().reset_index()
forecast_data
```

	work_year	experience_level	salary_in_usd
0	2020	EN	44768.000000
1	2020	MI	46586.333333
2	2021	EN	58242.666667
3	2021	MI	83788.428571
4	2021	SE	103904.250000
5	2022	EN	59500.000000
6	2022	EX	120000.000000
7	2022	MI	71210.473684
8	2022	SE	112564.100000

Importing the model and fitting the data to model

```
[29] #importing the model and perform multiple linear regression
     from sklearn.linear_model import LinearRegression
     lr=LinearRegression()
     fd=pd.get_dummies(forecast_data,columns=['experience_level'])
     x=fd.drop(columns=['salary_in_usd'])
     y=fd['salary_in_usd']
     lr.fit(x,y)
```

```
LinearRegression
LinearRegression()
```

Preparing the data for future prediction

```
[30] future_years = [2022,2023, 2024, 2025,2026,2027]
     experience_levels = ['EN', 'MI', 'SE','EX']

     #Generate the data for prediction
     future_data = pd.DataFrame({
         'work_year': np.repeat(future_years, len(experience_levels)),
         'experience_level': experience_levels * len(future_years)
     })

     #Encode the Experience column for future data
     future_data_encoded = pd.get_dummies(future_data, columns=['experience_level'])

     #Make sure all columns match
     future_data_encoded = future_data_encoded.reindex(columns=x.columns, fill_value=0)

     #Predict future salaries
     future_data['Predicted Salary'] = lr.predict(future_data_encoded)

     future_data
```

[30]

	work_year	experience_level	Predicted Salary
0	2022	EN	63878.236745
1	2022	MI	76903.093052
2	2022	SE	113088.182261
3	2022	EX	120000.000000
4	2023	EN	73586.251267
5	2023	MI	86611.107574
6	2023	SE	122796.196784
7	2023	EX	129708.014522
8	2024	EN	83294.265789
9	2024	MI	96319.122097
10	2024	SE	132504.211306
11	2024	EX	139416.029045
12	2025	EN	93002.280312
13	2025	MI	106027.136619
14	2025	SE	142212.225828
15	2025	EX	149124.043567
16	2026	EN	102710.294834
17	2026	MI	115735.151142
18	2026	SE	151920.240351
19	2026	EX	158832.058090
20	2027	EN	112418.309357
21	2027	MI	125443.165664
22	2027	SE	161628.254873
23	2027	EX	168540.072612

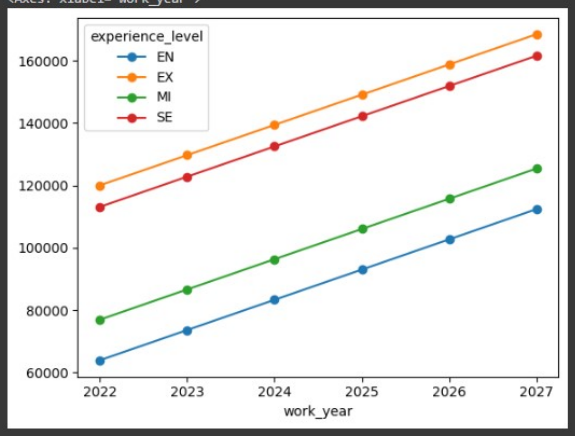
So it is the predicted salary of data analysts of different experiences in coming years

Note: this forecasting may not be true because it is the data collected from small amount of employees

Now plotting the forecast data into a line graph

```
[31] forecastchart=future_data.pivot(index='work_year',columns='experience_level',values='Predicted Salary')
     forecastchart.plot(kind='line',marker='o')
```

```
<Axes: xlabel='work_year'>
```



We can see there is very sharp rise in salary of Data Analysts in upcoming years, but this is almost true but only according the give data. This may varies with another sample of employees if we take thousands and lakhs of sample size.

So, Thats'it friends, meet you in next document, Thank You