0hjsubxdq

February 14, 2025

[38]:
```
import pandas as pd
```

[16]:
```
df = pd.read_csv(r"C:\Users\Welcome\Documents\dsbda2\employee_salaries.csv")
df
```

[16]:

	work_year	experience_level	employment_type	job_title	\
0	2023	SE	FT	Principal Data Scientist	
1	2023	MI	CT	ML Engineer	
2	2023	MI	CT	ML Engineer	
3	2023	SE	FT	Data Scientist	
4	2023	SE	FT	Data Scientist	
…	…	…	…	…	
3750	2020	SE	FT	Data Scientist	
3751	2021	MI	FT	Principal Data Scientist	
3752	2020	EN	FT	Data Scientist	
3753	2020	EN	CT	Business Data Analyst	
3754	2021	SE	FT	Data Science Manager	

	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	\
0	80000	EUR	85847	ES	100	
1	30000	USD	30000	US	100	
2	25500	USD	25500	US	100	
3	175000	USD	175000	CA	100	
4	120000	USD	120000	CA	100	
…	…	…	…	…	…	
3750	412000	USD	412000	US	100	
3751	151000	USD	151000	US	100	
3752	105000	USD	105000	US	100	
3753	100000	USD	100000	US	100	
3754	7000000	INR	94665	IN	50	

	company_location	company_size
0	ES	L
1	US	S
2	US	S
3	CA	M
4	CA	M

```
…                …           …
3750             US          L
3751             US          L
3752             US          S
3753             US          L
3754             IN          L

[3755 rows x 11 columns]
```

[ ]:
```
avg_salary_job = df.groupby('job_title')['salary_in_usd'].mean()
print(avg_salary_job)
```

```
job_title
3D Computer Vision Researcher     21352.250000
AI Developer                     136666.090909
AI Programmer                     55000.000000
AI Scientist                     110120.875000
Analytics Engineer               152368.631068
                                        …
Research Engineer                163108.378378
Research Scientist               161214.195122
Software Data Engineer            62510.000000
Staff Data Analyst                15000.000000
Staff Data Scientist             105000.000000
Name: salary_in_usd, Length: 93, dtype: float64
```

[ ]:
```
avg_salary_year = df.groupby('work_year')['salary_in_usd'].mean()
print(avg_salary_year)
```

```
work_year
2020     92302.631579
2021     94087.208696
2022    133338.620793
2023    149045.541176
Name: salary_in_usd, dtype: float64
```

[21]:
```
from sklearn import preprocessing
import pandas as pd
```

[40]:
```
# one hot encoding
from sklearn import preprocessing
enc = preprocessing.OneHotEncoder()
enc_df = pd.DataFrame(enc.fit_transform(df[['employment_type']]).toarray())
enc_df
```

[40]:

	0	1	2	3
0	0.0	0.0	1.0	0.0

```
1     1.0  0.0  0.0  0.0
2     1.0  0.0  0.0  0.0
3     0.0  0.0  1.0  0.0
4     0.0  0.0  1.0  0.0
…     …    …    …    …
3750  0.0  0.0  1.0  0.0
3751  0.0  0.0  1.0  0.0
3752  0.0  0.0  1.0  0.0
3753  1.0  0.0  0.0  0.0
3754  0.0  0.0  1.0  0.0

[3755 rows x 4 columns]
```

[41]:
```
df_encode =df.join(enc_df)
df_encode
```

[41]:
```
      work_year experience_level employment_type                 job_title  \
0          2023               SE              FT  Principal Data Scientist
1          2023               MI              CT               ML Engineer
2          2023               MI              CT               ML Engineer
3          2023               SE              FT            Data Scientist
4          2023               SE              FT            Data Scientist
…           …                  …               …                         …
3750       2020               SE              FT            Data Scientist
3751       2021               MI              FT  Principal Data Scientist
3752       2020               EN              FT            Data Scientist
3753       2020               EN              CT     Business Data Analyst
3754       2021               SE              FT      Data Science Manager

       salary salary_currency  salary_in_usd employee_residence  remote_ratio  \
0       80000             EUR          85847                 ES           100
1       30000             USD          30000                 US           100
2       25500             USD          25500                 US           100
3      175000             USD         175000                 CA           100
4      120000             USD         120000                 CA           100
…       …                 …               …                  …             …
3750   412000             USD         412000                 US           100
3751   151000             USD         151000                 US           100
3752   105000             USD         105000                 US           100
3753   100000             USD         100000                 US           100
3754  7000000             INR          94665                 IN            50

     company_location company_size    0    1    2    3
0                  ES            L  0.0  0.0  1.0  0.0
1                  US            S  1.0  0.0  0.0  0.0
2                  US            S  1.0  0.0  0.0  0.0
3                  CA            M  0.0  0.0  1.0  0.0
```

```
4                   CA             M  0.0  0.0  1.0  0.0
…                 …            …   …  …   …   …
3750                US             L  0.0  0.0  1.0  0.0
3751                US             L  0.0  0.0  1.0  0.0
3752                US             S  0.0  0.0  1.0  0.0
3753                US             L  1.0  0.0  0.0  0.0
3754                IN             L  0.0  0.0  1.0  0.0

[3755 rows x 15 columns]
```