datascience-job

March 18, 2024

[88]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

[89]:
```
data = pd.read_csv('C:\\Users\\kishu\\Downloads\\Datascience\\ds_salaries.csv')
data=data.iloc[:,1:]
```

[90]:
```
data.head()
```

[90]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2020	MI	FT	Data Scientist	70000	EUR	79833	DE	0	DE	L
1	2020	SE	FT	Machine Learning Scientist	260000	USD	260000	JP	0	JP	S
2	2020	SE	FT	Big Data Engineer	85000	GBP	109024	GB	50	GB	M
3	2020	MI	FT	Product Data Analyst	20000	USD	20000	HN	0	HN	S
4	2020	SE	FT	Machine Learning Engineer	150000	USD	150000	US	50	US	L

[91]:
```
data.shape
```

[91]:
```
(607, 11)
```

[92]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 607 entries, 0 to 606
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   work_year           607 non-null    int64 
 1   experience_level    607 non-null    object
 2   employment_type     607 non-null    object
 3   job_title           607 non-null    object
 4   salary              607 non-null    int64 
 5   salary_currency     607 non-null    object
 6   salary_in_usd       607 non-null    int64 
 7   employee_residence  607 non-null    object
 8   remote_ratio        607 non-null    int64 
 9   company_location    607 non-null    object
 10  company_size        607 non-null    object
dtypes: int64(4), object(7)
memory usage: 52.3+ KB
```

[93]:
```
data.isnull().sum()
```

[93]:
```
work_year             0
experience_level      0
employment_type       0
job_title             0
salary                0
salary_currency       0
salary_in_usd         0
employee_residence    0
remote_ratio          0
company_location      0
company_size          0
dtype: int64
```

[94]:
```
data.describe()
```

[94]:

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

[95]:
```
data.columns
```

[95]:
```
Index(['work_year', 'experience_level', 'employment_type', 'job_title',
       'salary', 'salary_currency', 'salary_in_usd', 'employee_residence',
       'remote_ratio', 'company_location', 'company_size'],
      dtype='object')
```

[96]:
```
for i in data.columns:
    if data[i].dtype=='object':
        print(data[i],data[i].value_counts())
```

```
0      MI
1      SE
2      SE
3      MI
4      SE
       ..
602    SE
603    SE
604    SE
605    SE
606    MI
Name: experience_level, Length: 607, dtype: object experience_level
SE    280
MI    213
EN     88
EX     26
Name: count, dtype: int64
0      FT
1      FT
2      FT
3      FT
4      FT
       ..
602    FT
603    FT
604    FT
605    FT
606    FT
Name: employment_type, Length: 607, dtype: object employment_type
FT    588
PT     10
CT      5
FL      4
Name: count, dtype: int64
0                   Data Scientist
1      Machine Learning Scientist
```

```
2               Big Data Engineer
3            Product Data Analyst
4       Machine Learning Engineer
                   …
602                  Data Engineer
603                  Data Engineer
604                   Data Analyst
605                   Data Analyst
606                   AI Scientist
Name: job_title, Length: 607, dtype: object job_title
Data Scientist                                143
Data Engineer                                 132
Data Analyst                                   97
Machine Learning Engineer                      41
Research Scientist                             16
Data Science Manager                           12
Data Architect                                 11
Big Data Engineer                               8
Machine Learning Scientist                      8
Principal Data Scientist                        7
AI Scientist                                    7
Data Science Consultant                         7
Director of Data Science                        7
Data Analytics Manager                          7
ML Engineer                                     6
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
Principal Data Engineer                         3
Principal Data Analyst                          2
ETL Developer                                   2
Product Data Analyst                            2
Director of Data Engineering                    2
Financial Data Analyst                          2
```

```
Cloud Data Engineer                          2
Lead Machine Learning Engineer               1
NLP Engineer                                 1
Head of Machine Learning                     1
3D Computer Vision Researcher                1
Data Specialist                              1
Staff Data Scientist                         1
Big Data Architect                           1
Finance Data Analyst                         1
Marketing Data Analyst                       1
Machine Learning Manager                     1
Data Analytics Lead                          1
Name: count, dtype: int64
0      EUR
1      USD
2      GBP
3      USD
4      USD
      …
602    USD
603    USD
604    USD
605    USD
606    USD
Name: salary_currency, Length: 607, dtype: object salary_currency
USD    398
EUR     95
GBP     44
INR     27
CAD     18
JPY      3
PLN      3
TRY      3
CNY      2
MXN      2
HUF      2
DKK      2
SGD      2
BRL      2
AUD      2
CLP      1
CHF      1
Name: count, dtype: int64
0      DE
1      JP
2      GB
3      HN
4      US
```

```
      ..
602    US
603    US
604    US
605    US
606    IN
Name: employee_residence, Length: 607, dtype: object employee_residence
US    332
GB     44
IN     30
CA     29
DE     25
FR     18
ES     15
GR     13
JP      7
PT      6
BR      6
PK      6
NL      5
PL      4
IT      4
RU      4
AE      3
AT      3
VN      3
TR      3
AU      3
RO      2
BE      2
SG      2
SI      2
DK      2
HU      2
NG      2
MX      2
BO      1
MY      1
TN      1
IE      1
DZ      1
AR      1
CZ      1
JE      1
LU      1
PR      1
RS      1
EE      1
```

```
CL      1
HK      1
KE      1
MD      1
CO      1
IR      1
CN      1
MT      1
UA      1
IQ      1
HN      1
BG      1
HR      1
PH      1
NZ      1
CH      1
Name: count, dtype: int64
0      DE
1      JP
2      GB
3      HN
4      US
       ..
602    US
603    US
604    US
605    US
606    US
Name: company_location, Length: 607, dtype: object company_location
US    355
GB     47
CA     30
DE     28
IN     24
FR     15
ES     14
GR     11
JP      6
NL      4
AT      4
PT      4
PL      4
LU      3
PK      3
BR      3
AE      3
MX      3
AU      3
```

```
TR      3
DK      3
IT      2
CZ      2
SI      2
RU      2
CH      2
NG      2
CN      2
BE      2
VN      1
EE      1
AS      1
DZ      1
MY      1
MD      1
KE      1
SG      1
CO      1
IR      1
CL      1
MT      1
IL      1
UA      1
IQ      1
RO      1
HR      1
NZ      1
HU      1
HN      1
IE      1
Name: count, dtype: int64
0      L
1      S
2      M
3      S
4      L
      ..
602    M
603    M
604    M
605    M
606    L
Name: company_size, Length: 607, dtype: object company_size
M    326
L    198
S     83
Name: count, dtype: int64
```

[97]:
```
for column in data.columns:
    if(data[column].dtype=='object'):
        unique_values = data[column].unique()
        print(f"Unique values in '{column}'is {len(unique_values)} and values␣
 ↪are {unique_values}")
```

```
Unique values in 'experience_level'is 4 and values are ['MI' 'SE' 'EN' 'EX']
Unique values in 'employment_type'is 4 and values are ['FT' 'CT' 'PT' 'FL']
Unique values in 'job_title'is 50 and values are ['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']
Unique values in 'salary_currency'is 17 and values are ['EUR' 'USD' 'GBP' 'HUF'
'INR' 'JPY' 'CNY' 'MXN' 'CAD' 'DKK' 'PLN' 'SGD'
 'CLP' 'BRL' 'TRY' 'AUD' 'CHF']
Unique values in 'employee_residence'is 57 and values are ['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']
Unique values in 'company_location'is 50 and values are ['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']
Unique values in 'company_size'is 3 and values are ['L' 'S' 'M']
```

[98]:
```
data['experience_level'] = data['experience_level'].map({'MI':'Intermediate␣
 ↪Level','EX':'Expert Level','SE':'Senior Level','EN':'Entry Level'})
```

```
data['employment_type'] = data['employment_type'].map({'PT':'Part Time','FT':
 ↪'Full Time','FL':'Freelance','CT':'Contratual'})
data['company_size'] = data['company_size'].map({'M':'Medium','S':'Small','L':
 ↪'Large'})
data['remote_ratio'] = data['remote_ratio'].map({0: 'No Remote',50: 'Partially␣
 ↪Remote',100: 'Fully Remote'})
```

[99]:
```
data.head()
```

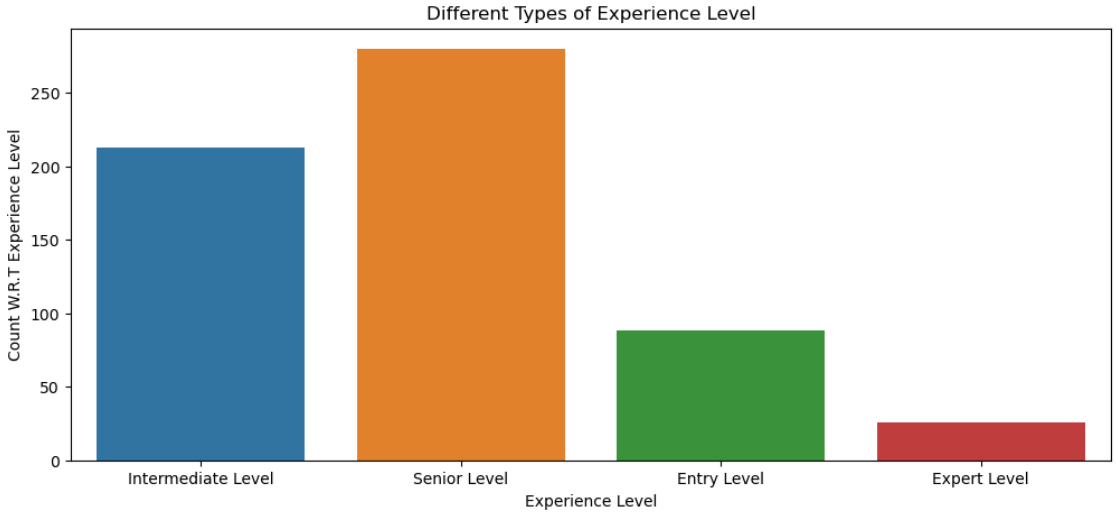
[99]:
```
   work_year    experience_level employment_type                 job_title  \
0       2020  Intermediate Level       Full Time              Data Scientist
1       2020        Senior Level       Full Time  Machine Learning Scientist
2       2020        Senior Level       Full Time           Big Data Engineer
3       2020  Intermediate Level       Full Time        Product Data Analyst
4       2020        Senior Level       Full Time   Machine Learning Engineer

   salary salary_currency  salary_in_usd employee_residence      remote_ratio  \
0   70000             EUR          79833                 DE         No Remote
1  260000             USD         260000                 JP         No Remote
2   85000             GBP         109024                 GB  Partially Remote
3   20000             USD          20000                 HN         No Remote
4  150000             USD         150000                 US  Partially Remote

  company_location company_size
0               DE        Large
1               JP        Small
2               GB       Medium
3               HN        Small
4               US        Large
```

[100]:
```
data.
 ↪drop(columns=['salary_currency','salary','employee_residence'],axis=1,inplace=True)
data
```

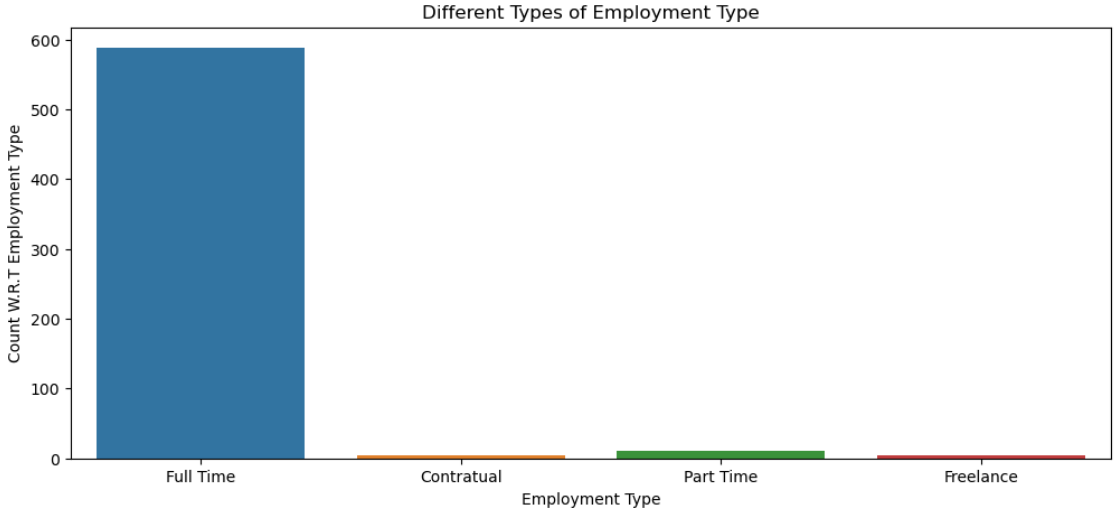
[100]:
```
     work_year    experience_level employment_type  \
0         2020  Intermediate Level       Full Time
1         2020        Senior Level       Full Time
2         2020        Senior Level       Full Time
3         2020  Intermediate Level       Full Time
4         2020        Senior Level       Full Time
..         …                   …             …
602       2022        Senior Level       Full Time
603       2022        Senior Level       Full Time
604       2022        Senior Level       Full Time
605       2022        Senior Level       Full Time
606       2022  Intermediate Level       Full Time
```

```
                    job_title  salary_in_usd      remote_ratio  \
0               Data Scientist          79833         No Remote
1    Machine Learning Scientist         260000         No Remote
2             Big Data Engineer         109024  Partially Remote
3          Product Data Analyst          20000         No Remote
4     Machine Learning Engineer         150000  Partially Remote
..                          …            …               …
602              Data Engineer         154000      Fully Remote
603              Data Engineer         126000      Fully Remote
604               Data Analyst         129000         No Remote
605               Data Analyst         150000      Fully Remote
606               AI Scientist         200000      Fully Remote

    company_location company_size
0                 DE        Large
1                 JP        Small
2                 GB       Medium
3                 HN        Small
4                 US        Large
..               …          …
602               US       Medium
603               US       Medium
604               US       Medium
605               US       Medium
606               US        Large

[607 rows x 8 columns]
```

[101]:
```
plt.figure(figsize=(12,5))
sns.countplot(x='experience_level', data=data)
plt.title("Different Types of Experience Level")
plt.xlabel('Experience Level')
plt.ylabel('Count W.R.T Experience Level')
plt.show()
```

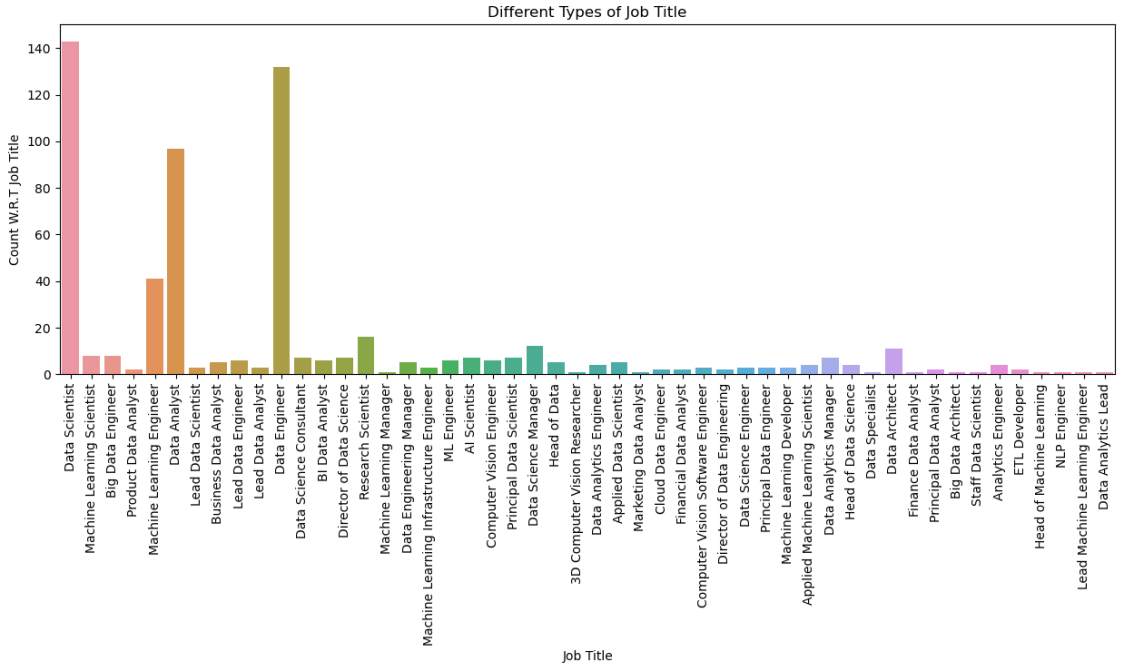


[102]:
```
plt.figure(figsize=(12,5))
sns.countplot(x='employment_type', data=data)
plt.title("Different Types of Employment Type")
plt.xlabel('Employment Type')
plt.ylabel('Count W.R.T Employment Type')
plt.show()
```

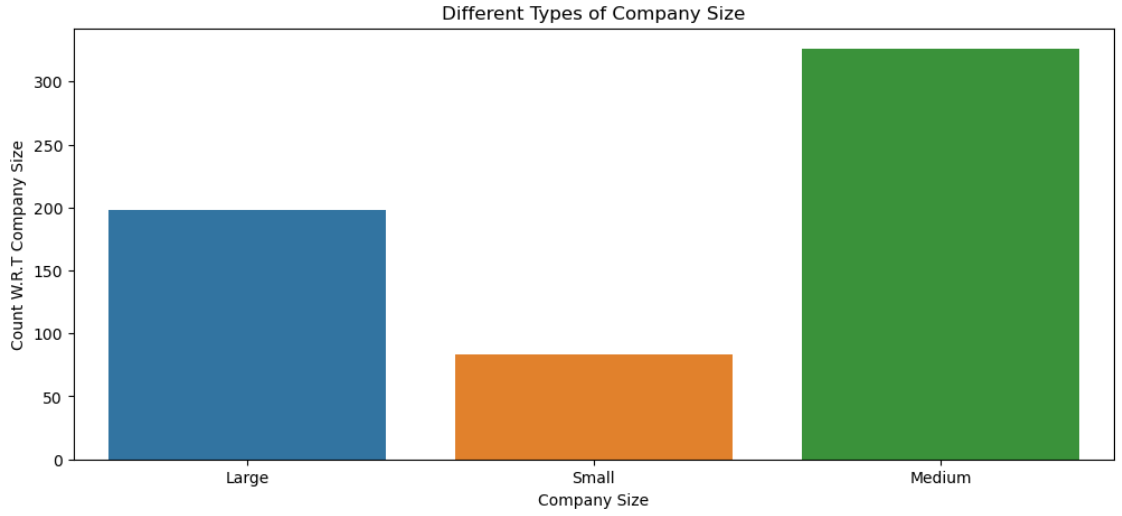


[103]:
```
plt.figure(figsize=(15,5))
sns.countplot(x='job_title', data=data)
plt.title("Different Types of Job Title")
plt.xlabel('Job Title')
plt.ylabel('Count W.R.T Job Title')
```

```
plt.xticks(rotation=90)
plt.show()
```



[104]:
```
plt.figure(figsize=(12,5))
sns.countplot(x='company_size', data=data)
plt.title("Different Types of Company Size")
plt.xlabel('Company Size')
plt.ylabel('Count W.R.T Company Size')
plt.show()
```



[105]:
```
X = data.drop(['salary_in_usd'], axis=1)
y = data['salary_in_usd']
```

[106]:
```
one_hot_encoded = pd.get_dummies(data=data)
```

[107]:
```
one_hot_encoded
```

[107]:

	work_year	salary_in_usd	experience_level_Entry Level	\
0	2020	79833	False	
1	2020	260000	False	
2	2020	109024	False	
3	2020	20000	False	
4	2020	150000	False	
..	…	…	…	
602	2022	154000	False	
603	2022	126000	False	
604	2022	129000	False	
605	2022	150000	False	
606	2022	200000	False	

	experience_level_Expert Level	experience_level_Intermediate Level	\
0	False	True	
1	False	False	
2	False	False	
3	False	True	
4	False	False	
..	…	…	
602	False	False	
603	False	False	
604	False	False	
605	False	False	
606	False	True	

	experience_level_Senior Level	employment_type_Contratual	\
0	False	False	
1	True	False	
2	True	False	
3	False	False	
4	True	False	
..	…	…	
602	True	False	
603	True	False	
604	True	False	
605	True	False	
606	False	False	

```
     employment_type_Freelance  employment_type_Full Time  \
0                        False                       True
1                        False                       True
2                        False                       True
3                        False                       True
4                        False                       True
..                         …                          …
602                      False                       True
603                      False                       True
604                      False                       True
605                      False                       True
606                      False                       True

     employment_type_Part Time  …  company_location_RU  company_location_SG  \
0                        False  …                False                False
1                        False  …                False                False
2                        False  …                False                False
3                        False  …                False                False
4                        False  …                False                False
..                         …  …                   …                    …
602                      False  …                False                False
603                      False  …                False                False
604                      False  …                False                False
605                      False  …                False                False
606                      False  …                False                False

     company_location_SI  company_location_TR  company_location_UA  \
0                  False                False                False
1                  False                False                False
2                  False                False                False
3                  False                False                False
4                  False                False                False
..                   …                    …                    …
602                False                False                False
603                False                False                False
604                False                False                False
605                False                False                False
606                False                False                False

     company_location_US  company_location_VN  company_size_Large  \
0                  False                False               True
1                  False                False              False
2                  False                False              False
3                  False                False              False
4                   True                False               True
..                   …                    …                  …
602                 True                False              False
```

```
603                 True                 False               False
604                 True                 False               False
605                 True                 False               False
606                 True                 False                True

     company_size_Medium  company_size_Small
0                  False               False
1                  False                True
2                   True               False
3                  False                True
4                  False               False
..                   …                   …
602                 True               False
603                 True               False
604                 True               False
605                 True               False
606                False               False

[607 rows x 116 columns]
```

[108]:
```
# Split the data into training and testing set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(one_hot_encoded, y,␣
 ↪test_size=0.2, random_state=0)
```

[109]:
```
print(X_train.shape)
print(y_train.shape)
print(X_test.shape)
print(y_test.shape)
```

```
(485, 116)
(485,)
(122, 116)
(122,)
```

[110]:
```
X
```

[110]:
```
     work_year    experience_level employment_type  \
0         2020  Intermediate Level       Full Time
1         2020        Senior Level       Full Time
2         2020        Senior Level       Full Time
3         2020  Intermediate Level       Full Time
4         2020        Senior Level       Full Time
..         …                   …             …
602       2022        Senior Level       Full Time
603       2022        Senior Level       Full Time
604       2022        Senior Level       Full Time
```

```
605       2022        Senior Level       Full Time
606       2022  Intermediate Level       Full Time

                     job_title      remote_ratio company_location  \
0               Data Scientist         No Remote               DE
1    Machine Learning Scientist         No Remote               JP
2            Big Data Engineer  Partially Remote               GB
3         Product Data Analyst         No Remote               HN
4     Machine Learning Engineer  Partially Remote               US
..                          …               …              …
602             Data Engineer      Fully Remote               US
603             Data Engineer      Fully Remote               US
604              Data Analyst         No Remote               US
605              Data Analyst      Fully Remote               US
606              AI Scientist      Fully Remote               US

    company_size
0          Large
1          Small
2         Medium
3          Small
4          Large
..           …
602       Medium
603       Medium
604       Medium
605       Medium
606        Large

[607 rows x 7 columns]
```

[111]:
```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
```

[112]:
```
lreg = LinearRegression()
lreg.fit(X_train, y_train)
```

[112]:
```
LinearRegression()
```

[113]:
```
y_pred=lreg.predict(X_test)
```

[114]:
```
print(y_test,y_pred)
```

```
575    140000
52      45896
530     85000
345    156600
```

```
55     148261
        …
479    120000
293     90000
532    214000
278     20171
165    165000
Name: salary_in_usd, Length: 122, dtype: int64 [140000.  45896.  85000. 156600.
148261.  21669. 113000.  52351. 200000.
 183600. 116000. 200000.  90734. 211500.  60000.  87932. 116150. 136994.
 120000. 130000. 105000. 200000. 110000. 120000.  69741.  87425.  12901.
  60000.  24823.  65438. 108800.  78791.  37300.  33808.  62000.  60000.
 132000.  90320. 104702. 181940.  72212.  96113.  85000. 136600. 150000.
 324000.  45807. 192400. 260000.  32974. 109280. 124333. 109280.   9466.
 405000. 152000.  64849. 105000. 135000. 174000.  50000.  59102. 158200.
 114047. 105000.   6072. 115000.  28476. 125000.  81000. 164996. 123000.
  67000.  36643. 250000. 416000.  54957.  46759. 154600.  45618. 130000.
  98158. 145000.  90320. 170000.  31875. 210000.  39916. 147800.  71786.
   6072.  70000. 165000. 270000. 137141.  49268. 150000. 161342.  19609.
 423000. 145000.  56738.  21637.  25000.  18442.   8000. 120000.  15966.
 190000.   9272.  37825. 170000.  53192.  51321. 170000. 209100. 210000.
 120000.  90000. 214000.  20171. 165000.]
```

[115]:
```
r2 = r2_score(y_test, y_pred)
```

[117]:
```
print(r2*100)
```

```
100.0
```