Data Science Job salaries

Data science job salries encompassing the experience level,employment time,work year,salary,salary in dollars,remote ration,company location,employee's residence etc.

IMPORTS

In [ ]:
```
import pandas as pd
import opendatasets as od
```

In [ ]:
```
datasets_url= 'https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries'
od.download(datasets_url)
```

```
Skipping, found downloaded files in ".\data-science-job-salaries" (use force=True to
force download)
```

In [ ]:
```
data_dir=".\data-science-job-salaries"
```

In [ ]:
```
import os
os.listdir(data_dir)
```

Out[ ]:
```
['ds_salaries.csv']
```

In [ ]:
```
ds_job_salary_df= pd.read_csv("data-science-job-salaries/ds_salaries.csv")
ds_job_salary_df
os.getcwd()
```

Out[ ]:
```
'c:\\Users\\OLUWASORE\\Documents\\Data Analytics'
```

In [ ]:
```
ds_job_salary_df= pd.read_csv("data-science-job-salaries/ds_salaries.csv")
ds_job_salary_df
```

Out[ ]:

	Unnamed: 0	work_year	experience_level	employment_type	job_title	salary	salary_currency	sal
0	0	2020	MI	FT	Data Scientist	70000	EUR	
1	1	2020	SE	FT	Machine Learning Scientist	260000	USD	
2	2	2020	SE	FT	Big Data Engineer	85000	GBP	
3	3	2020	MI	FT	Product Data Analyst	20000	USD	
4	4	2020	SE	FT	Machine Learning Engineer	150000	USD	
...	...	...	...	...	...	...	...	
602	602	2022	SE	FT	Data Engineer	154000	USD	
603	603	2022	SE	FT	Data Engineer	126000	USD	
604	604	2022	SE	FT	Data Analyst	129000	USD	
605	605	2022	SE	FT	Data Analyst	150000	USD	
606	606	2022	MI	FT	AI Scientist	200000	USD	

607 rows × 12 columns

DATA CLEANING

In [ ]:
```
ds_job_salary_df.shape
```

Out[ ]:
```
(607, 12)
```

In [ ]:
```
ds_job_salary_df.columns
```

Out[ ]:
```
Index(['Unnamed: 0', 'work_year', 'experience_level', 'employment_type',
       'job_title', 'salary', 'salary_currency', 'salary_in_usd',
       'employee_residence', 'remote_ratio', 'company_location',
       'company_size'],
      dtype='object')
```

In [ ]:
```
ds_job_sal_df=ds_job_salary_df.copy()
```

In [ ]:
```
ds_job_sal_df
```

Out[ ]:

	Unnamed: 0	work_year	experience_level	employment_type	job_title	salary	salary_currency	sal
0	0	2020	MI	FT	Data Scientist	70000	EUR	
1	1	2020	SE	FT	Machine Learning Scientist	260000	USD	
2	2	2020	SE	FT	Big Data Engineer	85000	GBP	
3	3	2020	MI	FT	Product Data Analyst	20000	USD	
4	4	2020	SE	FT	Machine Learning Engineer	150000	USD	
...	...	...	...	...	...	...	...	
602	602	2022	SE	FT	Data Engineer	154000	USD	
603	603	2022	SE	FT	Data Engineer	126000	USD	
604	604	2022	SE	FT	Data Analyst	129000	USD	
605	605	2022	SE	FT	Data Analyst	150000	USD	
606	606	2022	MI	FT	AI Scientist	200000	USD	

607 rows × 12 columns

In [ ]:
```
ds_job_sal_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 607 entries, 0 to 606
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Unnamed: 0          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
```

In [ ]:
```
ds_job_sal_df.drop('Unnamed: 0',axis=1,inplace=True)
```

In [ ]:
```
ds_job_sal_df.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
```

In [ ]:
```
ds_job_sal_df.isnull().sum()
```

Out[ ]:
```
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
```

In [ ]:
```
ds_job_sal_df.salary
```

Out[ ]:
```
0       70000
1      260000
2       85000
3       20000
4      150000
        ...
602    154000
603    126000
604    129000
605    150000
606    200000
Name: salary, Length: 607, dtype: int64
```

In [ ]:
```
ds_job_sal_df.remote_ratio
```

Out[ ]:
```
0        0
1        0
2       50
3        0
4       50
       ...
602    100
603    100
604      0
605    100
606    100
Name: remote_ratio, Length: 607, dtype: int64
```

In [ ]:
```
ds_job_sal_df.company_location.nunique()
```

Out[ ]:
```
50
```

In [ ]:
```
ds_job_sal_df.company_location.value_counts()
```

Out[ ]:
```
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: company_location, dtype: int64
```

In [ ]:
```
ds_job_sal_df.salary_currency.unique()
```

Out[ ]:
```
array(['EUR', 'USD', 'GBP', 'HUF', 'INR', 'JPY', 'CNY', 'MXN', 'CAD',
       'DKK', 'PLN', 'SGD', 'CLP', 'BRL', 'TRY', 'AUD', 'CHF'],
      dtype=object)
```

In [ ]:
```
ds_job_sal_df.salary_currency.nunique()
```

Out[ ]:
```
17
```

In [ ]:
```
ds_job_sal_df.drop('salary',axis=1,inplace=True)
```

In [ ]:
```
ds_job_sal_df.work_year.value_counts()
```

Out[ ]:
```
2022    318
2021    217
2020     72
Name: work_year, dtype: int64
```

In [ ]:
```
ds_job_sal_df.company_size.value_counts()
```

Out[ ]:
```
M    326
L    198
S     83
Name: company_size, dtype: int64
```

DATA MANIPULATION WITH VISUALIZATION

In [ ]:
```
ds_job_sal_df.describe()
```

Out[ ]:

	work_year	salary_in_usd	remote_ratio
count	607.000000	607.000000	607.00000
mean	2021.405272	112297.869852	70.92257
std	0.692133	70957.259411	40.70913
min	2020.000000	2859.000000	0.00000
25%	2021.000000	62726.000000	50.00000
50%	2022.000000	101570.000000	100.00000
75%	2022.000000	150000.000000	100.00000
max	2022.000000	600000.000000	100.00000

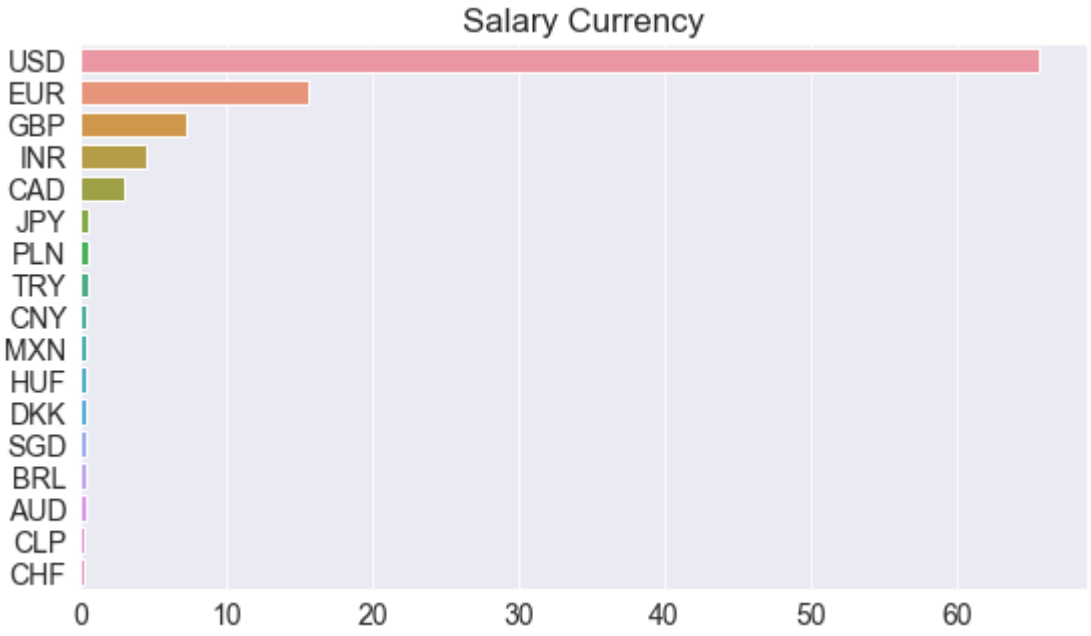
In [ ]:
```
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
```

In [ ]:
```
sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (9, 5)
matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

In [ ]:
```
salary_currency=ds_job_sal_df.salary_currency.value_counts()
salary_currency_percentage=salary_currency*100/ds_job_sal_df.salary_currency.count()
salary_currency_percentage
```

Out[ ]:
```
USD    65.568369
EUR    15.650741
GBP     7.248764
INR     4.448105
CAD     2.965404
JPY     0.494234
PLN     0.494234
TRY     0.494234
CNY     0.329489
MXN     0.329489
HUF     0.329489
DKK     0.329489
SGD     0.329489
BRL     0.329489
AUD     0.329489
CLP     0.164745
CHF     0.164745
Name: salary_currency, dtype: float64
```

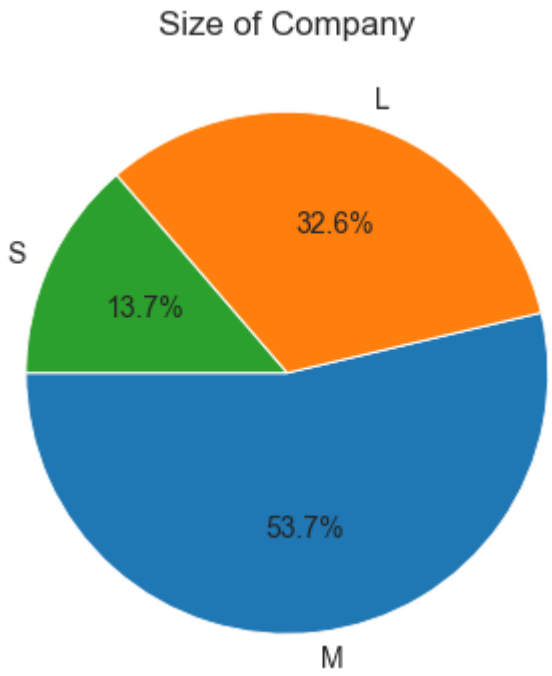
In [ ]:
```
sns.barplot(x=salary_currency_percentage,y=salary_currency_percentage.index)
plt.title('Salary Currency')
plt.ylabel(None);
plt.xlabel(None);
```



66% receives their salaries normally in USD,which shows USD is a preffered choice for global Exchange in data science which makes it encouraging to work internationally without having currency issues and salary marginalisation.

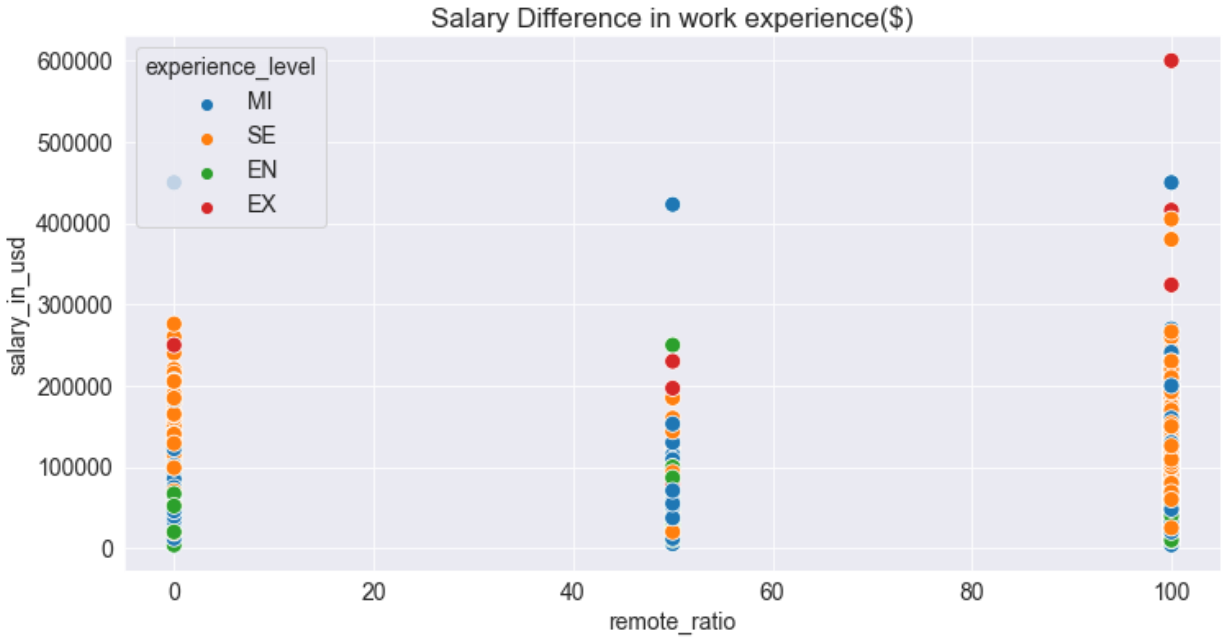
In [ ]:
```
Company_size=ds_job_sal_df.company_size.value_counts()
#visualization

plt.figure(figsize=(12,6))
plt.title('Size of Company')
plt.pie(Company_size,labels=Company_size.index,autopct='%1.1f%%',startangle=180,);
```



More than half of data science companies are Medium.It is noticed the Medium companies dominates the Data science world followed by a large margin whiles small companies(start ups) accross the world being little which shows enough capital is needed to run the Companies effectively.

In [ ]:
```
plt.figure(figsize=(12, 6))
plt.title('Salary Difference in work experience($)')
sns.scatterplot(x=ds_job_sal_df.remote_ratio, y=ds_job_sal_df.salary_in_usd, hue=ds_jo
```

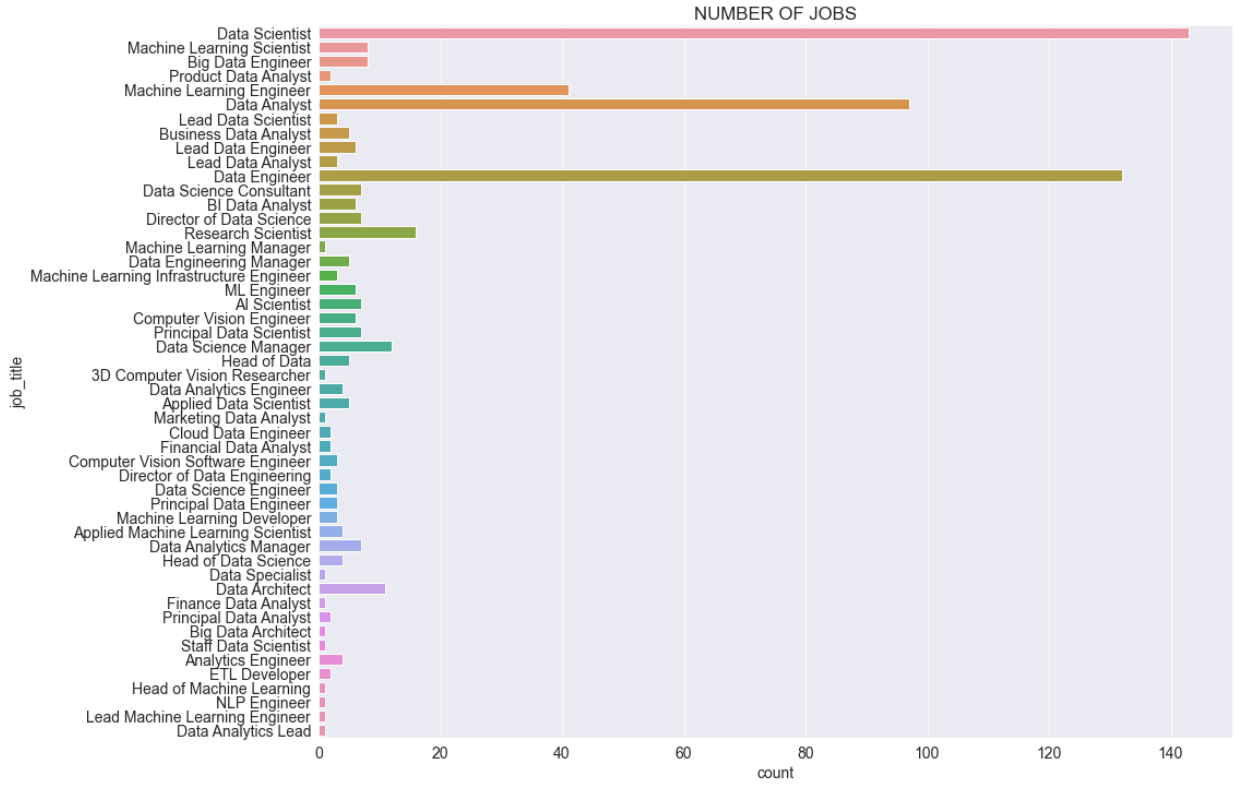


Only EX earns $600,000 while MI, EX, SE earn within 500,000-$300,000 while EN earnings are below 300,000 as the seen in the plot above. The higher your experience level,the higher your salary earnings. geting to EX makes your salary earnings not less than $200,000.

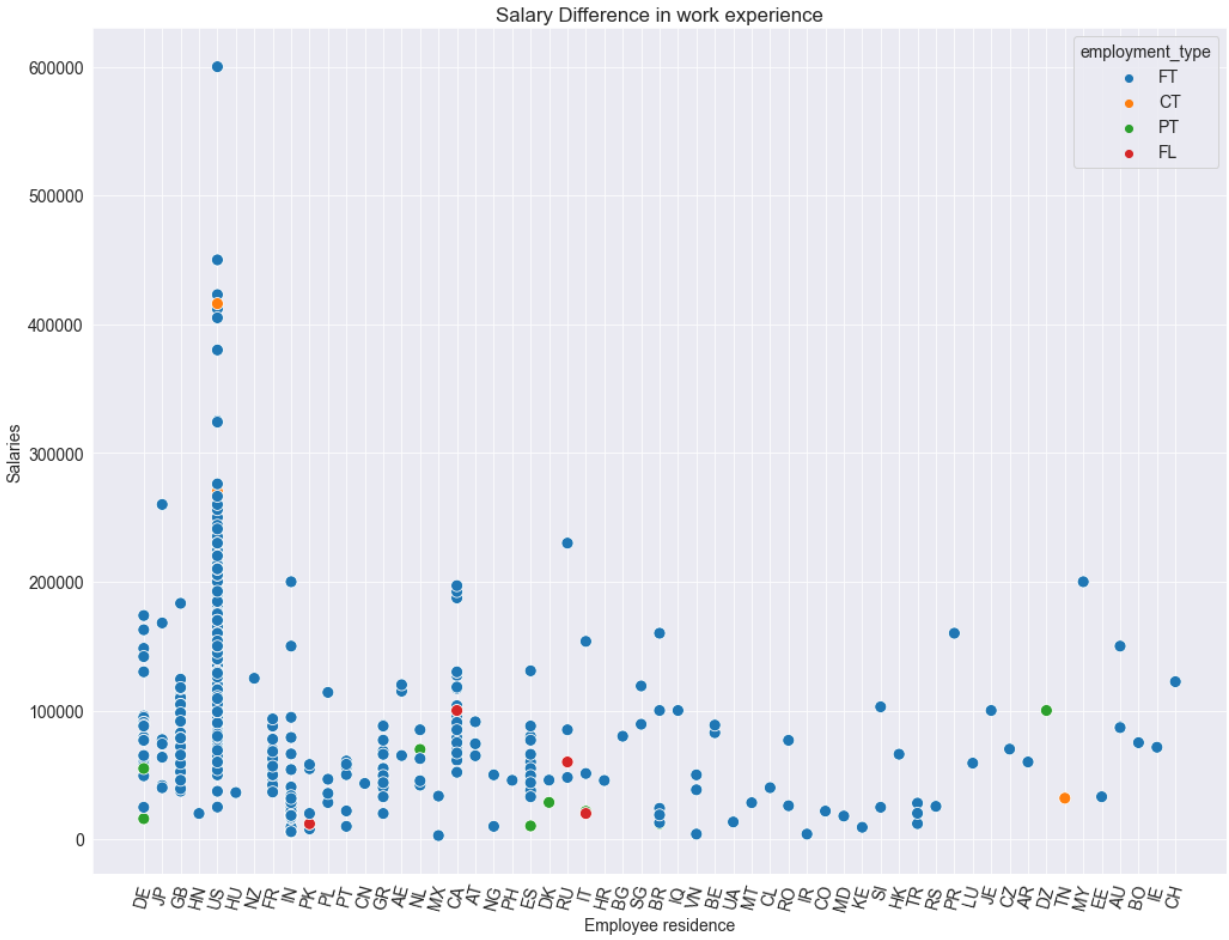
In [ ]:
```
plt.figure(figsize=(12,8))
plt.title('Company size\'s salary')
plt.xlabel('Salary in USD')
plt.ylabel('Count')
sns.histplot(x=ds_job_sal_df.salary_in_usd,hue=ds_job_sal_df.company_size);
```

Large companies pay higher salaries compared to Meduims. Only Large companies pays 600,000 which shows they can afford the most skilled and qualified in the Data science field. Small & Medium companies are still able within the the range of 400,000 below.

In [ ]:
```
plt.figure(figsize=(15,12))
plt.title('NUMBER OF JOBS')
plt.xlabel(None)
sns.countplot(y=ds_job_sal_df.job_title);
```



In [ ]:
```
plt.figure(figsize=(17,13))
plt.title('Salary Difference in work experience')
plt.xlabel('Employee residence')
plt.ylabel('Salaries')
plt.xticks(rotation=75)
sns.scatterplot(x=ds_job_sal_df.employee_residence, y=ds_job_sal_df.salary_in_usd, hue
```

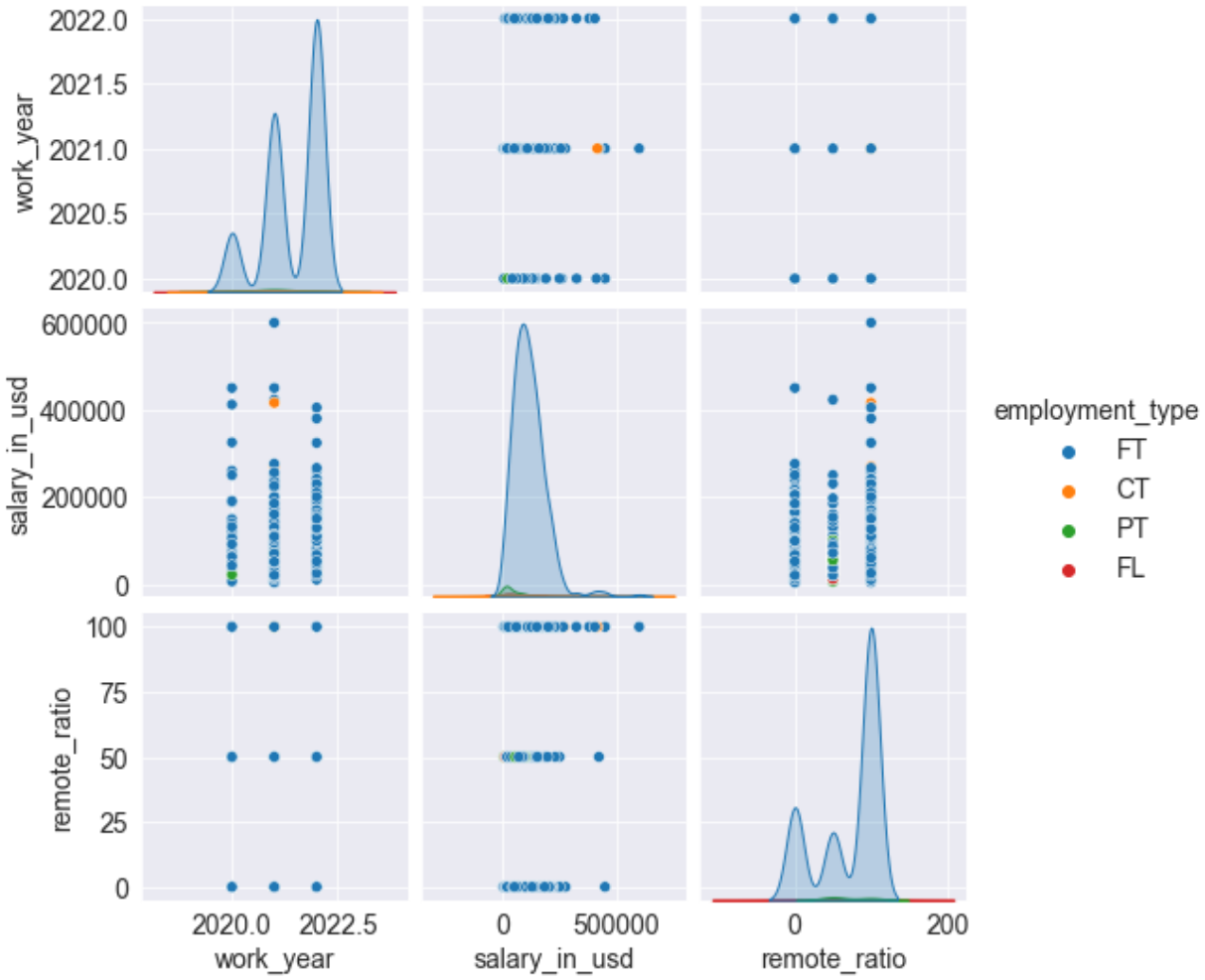


The US as the highest FT employed type with a higher salary earnings unlike the rest of them which shows the marginlisation in salry earnings and effects of other currencies to the Dollar.

In [ ]:
```
plt.figure(figsize=(16,14))
sns.pairplot(ds_job_sal_df,hue='employment_type')
```

Out[ ]:
```
<seaborn.axisgrid.PairGrid at 0x32e1076a30>

<Figure size 1152x1008 with 0 Axes>
```



In [ ]:
```
ds_job_sal_df['job_branch']=ds_job_sal_df['job_title']
ds_job_sal_df
```

Out[ ]:

	work_year	experience_level	employment_type	job_title	salary_currency	salary_in_usd	employe
0	2020	MI	FT	Data Scientist	EUR	79833	
1	2020	SE	FT	Machine Learning Scientist	USD	260000	
2	2020	SE	FT	Big Data Engineer	GBP	109024	
3	2020	MI	FT	Product Data Analyst	USD	20000	
4	2020	SE	FT	Machine Learning Engineer	USD	150000	
...	...	...	...	...	...	...	
602	2022	SE	FT	Data Engineer	USD	154000	
603	2022	SE	FT	Data Engineer	USD	126000	
604	2022	SE	FT	Data Analyst	USD	129000	
605	2022	SE	FT	Data Analyst	USD	150000	
606	2022	MI	FT	AI Scientist	USD	200000	

607 rows × 12 columns

In [ ]:
```
ds_job_sal_df
```

Out[ ]:

	work_year	experience_level	employment_type	job_title	salary_currency	salary_in_usd	employe
0	2020	MI	FT	Data Scientist	EUR	79833	
1	2020	SE	FT	Machine Learning Scientist	USD	260000	
2	2020	SE	FT	Big Data Engineer	GBP	109024	
3	2020	MI	FT	Product Data Analyst	USD	20000	
4	2020	SE	FT	Machine Learning Engineer	USD	150000	
...	...	...	...	...	...	...	
602	2022	SE	FT	Data Engineer	USD	154000	
603	2022	SE	FT	Data Engineer	USD	126000	
604	2022	SE	FT	Data Analyst	USD	129000	
605	2022	SE	FT	Data Analyst	USD	150000	
606	2022	MI	FT	AI Scientist	USD	200000	

607 rows × 11 columns

EXPLORATIVE ANALYSIS

Q1:what is the salaries of Job branch with in respect to their Employment type?

In [ ]:
```
ds_job_sal_df['job_branch'].replace(['Data Scientist',
                                     'Research Scientist',
                                     'Data Science Manager',
                                     'Machine Learning Scientist',
                                     'Principal Data Scientist',
                                     'AI Scientist',
                                     'Data Science Consultant',
                                     'Director of Data Science',
                                     'Applied Data Scientist',
                                     'Applied Machine Learning Scientist',
                                     'Head of Data Science',
                                     'Lead Data Scientist',
                                     'Data Specialist',
```

```
                                    'Staff Data Scientist',
                                    'Machine Learning Manager'],'Data Science related
```

In [ ]:
```
ds_job_sal_df.job_branch.value_counts()
```

Out[ ]:
```
Data Science related jobs                    226
Data Engineer                                132
Data Analyst                                  97
Machine Learning Engineer                     41
Data Architect                                11
Big Data Engineer                              8
Data Analytics Manager                         7
BI Data Analyst                                6
Computer Vision Engineer                       6
ML Engineer                                    6
Lead Data Engineer                             6
Data Engineering Manager                       5
Head of Data                                   5
Business Data Analyst                          5
Analytics Engineer                             4
Data Analytics Engineer                        4
Machine Learning Infrastructure Engineer       3
Machine Learning Developer                     3
Lead Data Analyst                              3
Computer Vision Software Engineer              3
Data Science Engineer                          3
Principal Data Engineer                        3
Principal Data Analyst                         2
ETL Developer                                  2
Cloud Data Engineer                            2
Director of Data Engineering                   2
Financial Data Analyst                         2
Product Data Analyst                           2
Finance Data Analyst                           1
Marketing Data Analyst                         1
Big Data Architect                             1
3D Computer Vision Researcher                  1
Head of Machine Learning                       1
NLP Engineer                                   1
Lead Machine Learning Engineer                 1
Data Analytics Lead                            1
Name: job_branch, dtype: int64
```

In [ ]:
```
ds_job_sal_df['job_branch'].replace(['Data Analyst',
                                    'Data Analytics Manager',
                                    'BI Data Analyst',
                                    'Head of Data',
                                    'Business Data Analyst',
                                    'Lead Data Analyst',
                                    'Financial Data Analyst',
                                    'Product Data Analyst',
                                    'Principal Data Analyst',
                                    'Marketing Data Analyst',
                                    'Finance Data Analyst',
                                    '3D Computer Vision Researcher','Data Analytics Le
```

In [ ]:
```
ds_job_sal_df.job_branch.value_counts()
```

Out[ ]:
```
Data Science related jobs                  226
Data Analysis related jobs                 132
Data Engineer                              132
Machine Learning Engineer                   41
Data Architect                              11
Big Data Engineer                            8
Lead Data Engineer                           6
ML Engineer                                  6
Computer Vision Engineer                     6
Data Engineering Manager                     5
Analytics Engineer                           4
Data Analytics Engineer                      4
Principal Data Engineer                      3
Machine Learning Developer                   3
Computer Vision Software Engineer            3
Data Science Engineer                        3
Machine Learning Infrastructure Engineer     3
Director of Data Engineering                 2
Cloud Data Engineer                          2
ETL Developer                                2
Big Data Architect                           1
Head of Machine Learning                     1
NLP Engineer                                 1
Lead Machine Learning Engineer               1
Data Analytics Lead                          1
Name: job_branch, dtype: int64
```

In [ ]:
```
ds_job_sal_df['job_branch'].replace(['Lead Machine Learning Engineer',
                                     'NLP Engineer',
                                     'Head of Machine Learning',
                                     'Big Data Architect',
                                     'Director of Data Engineering',
                                     'Cloud Data Engineer',
                                     'ETL Developer',
                                     'Principal Data Engineer',
                                     'Data Science Engineer',
                                     'Computer Vision Software Engineer',
                                     'Machine Learning Developer',
                                     'Machine Learning Infrastructure Engineer',
                                     'Data Analytics Engineer',
                                     'Analytics Engineer',
                                     'Data Engineering Manager',
                                     'Lead Data Engineer',
                                     'ML Engineer',
                                     'Computer Vision Engineer',
                                     'Big Data Engineer',
                                     'Data Architect',
                                     'Machine Learning Engineer',
                                     'Data Engineer'],'Data Engineering related jobs',i
```

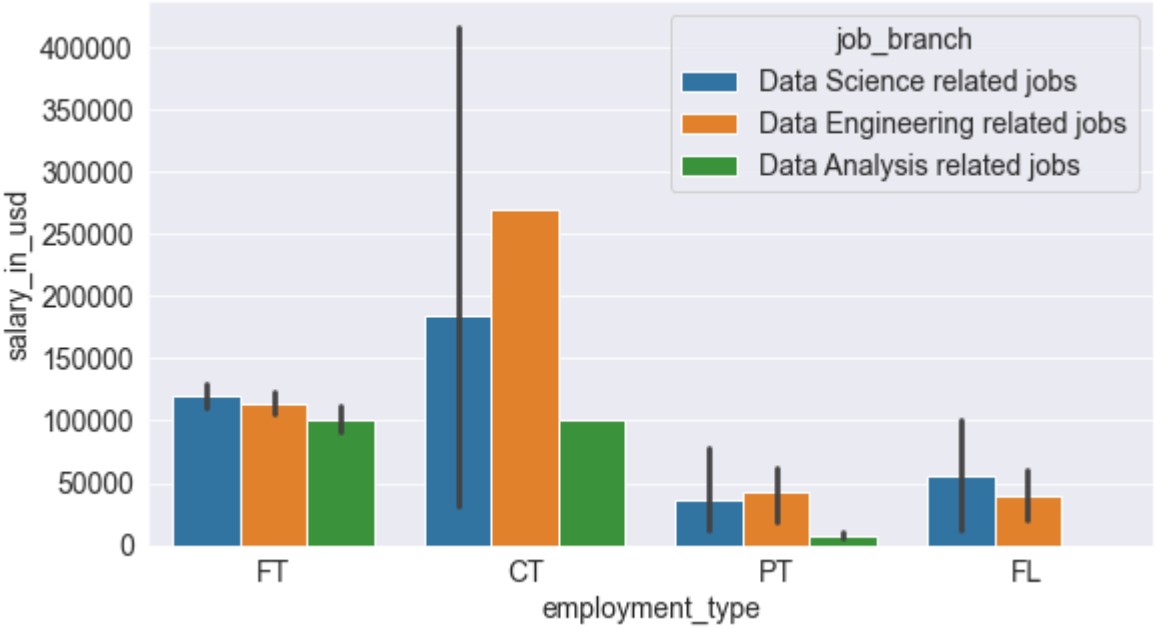
In [ ]:
```
ds_job_sal_df.job_branch.value_counts()
```

Out[ ]:
```
Data Engineering related jobs    248
Data Science related jobs        226
Data Analysis related jobs       133
Name: job_branch, dtype: int64
```

In [ ]:
```
ds_job_sal_df.groupby('employment_type').mean()['salary_in_usd'].value_counts()
```

Out[ ]:
```
184575.000000    1
48000.000000     1
113468.073129    1
33070.500000     1
Name: salary_in_usd, dtype: int64
```

In [ ]:
```
sns.barplot(x=ds_job_sal_df.employment_type,y=ds_job_sal_df.salary_in_usd,hue=ds_job_s
```



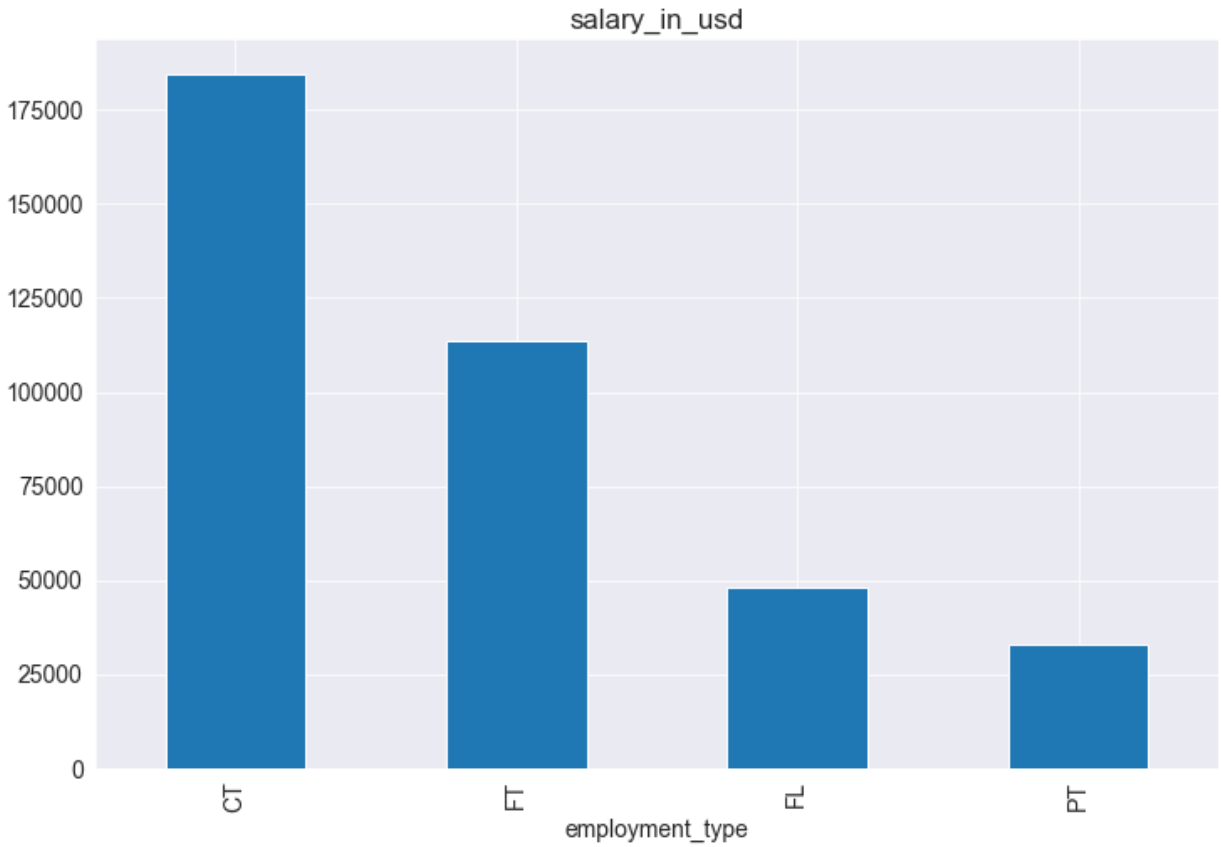
As noticed Data science has a higher earning for FT & FL employees, while Data engineering has a higher earing for CT and PT which shows data engineers prefer being employed using contracts and part time working schedule due to the nature of their job,while Data Science are employed full time or freelancing.

Q2: Which Employment type earns more?,and what is the average salaries earned?

In [ ]:
```
avg_sal_csize = ds_job_sal_df.groupby('employment_type').mean()['salary_in_usd'].sort_
avg_sal_csize
```

Out[ ]:
```
employment_type
CT    184575.000000
FT    113468.073129
FL     48000.000000
PT     33070.500000
Name: salary_in_usd, dtype: float64
```

In [ ]:
```
avg_sal_csize.plot(kind='bar',subplots=True,figsize=(12,8));
```



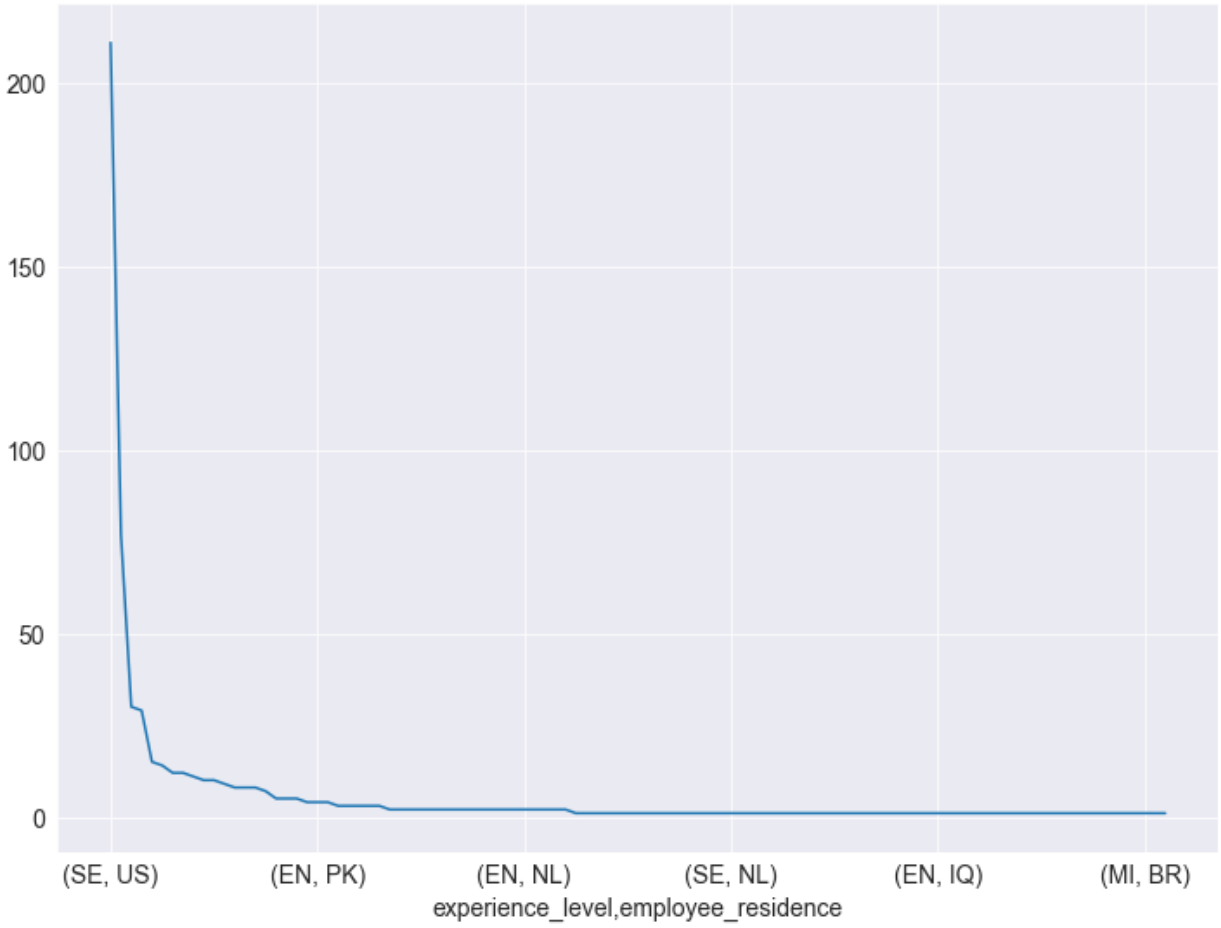
CT earns more since contract jobs are extremely tedious with a time frame to finish compared to others

Q3: where do the most experienced data scientist reside?

In [ ]:
```
most_experienced=ds_job_sal_df[['experience_level','employee_residence']].value_counts
most_experienced
```

Out[ ]:
```
experience_level  employee_residence
SE                US                   211
MI                US                    77
                  GB                    30
EN                US                    29
EX                US                    15
                                       ...
MI                CL                     1
                  CH                     1
                  BR                     1
                  BO                     1
SE                VN                     1
Length: 103, dtype: int64
```

In [ ]:
```
most_experienced.plot(x='experience_level',figsize=(12,9),grid=True);
```



the US dominates the data science as of today with a lot of Skilled and well qualoified employees

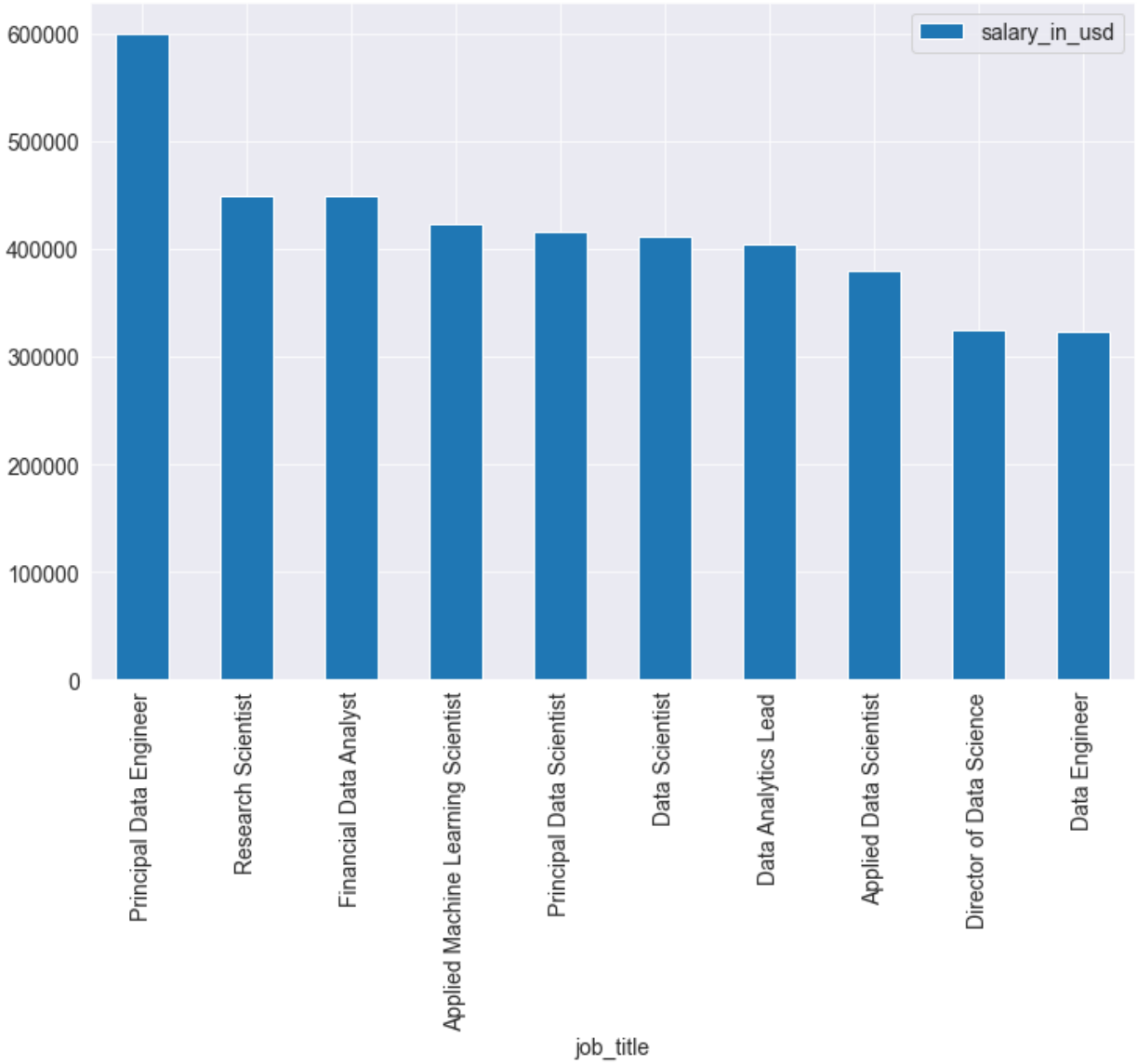
Q4: - what are highest paying jobs in Data science?

In [ ]:
```
highest_paid=ds_job_sal_df.nlargest(n=10,columns=['salary_in_usd'])[['job_title','sala
highest_paid
```

Out[ ]:

	job_title	salary_in_usd
252	Principal Data Engineer	600000
33	Research Scientist	450000
97	Financial Data Analyst	450000
157	Applied Machine Learning Scientist	423000
225	Principal Data Scientist	416000
63	Data Scientist	412000
523	Data Analytics Lead	405000
519	Applied Data Scientist	380000
25	Director of Data Science	325000
482	Data Engineer	324000

In [ ]:
```
highest_paid.plot(x='job_title',kind='bar',figsize=(12,8));
```



The Principal Data Engineer earns the highest salary in this survey.

In [ ]:
```
ds_job_sal_df
```

Out[ ]:

	work_year	experience_level	employment_type	job_title	salary_currency	salary_in_usd	employe
0	2020	MI	FT	Data Scientist	EUR	79833	
1	2020	SE	FT	Machine Learning Scientist	USD	260000	
2	2020	SE	FT	Big Data Engineer	GBP	109024	
3	2020	MI	FT	Product Data Analyst	USD	20000	
4	2020	SE	FT	Machine Learning Engineer	USD	150000	
...	...	...	...	...	...	...	
602	2022	SE	FT	Data Engineer	USD	154000	
603	2022	SE	FT	Data Engineer	USD	126000	
604	2022	SE	FT	Data Analyst	USD	129000	
605	2022	SE	FT	Data Analyst	USD	150000	
606	2022	MI	FT	AI Scientist	USD	200000	

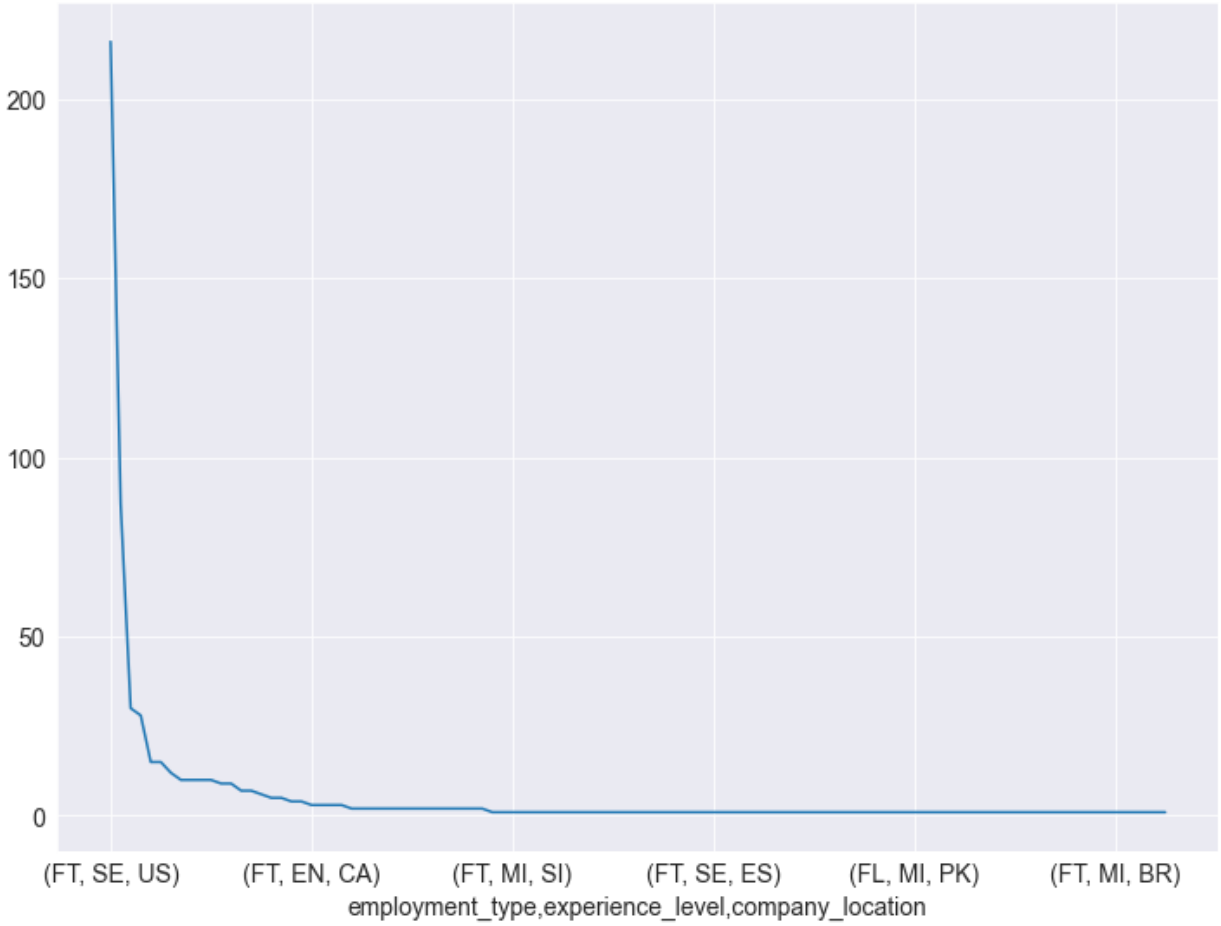
607 rows × 11 columns

Q5: Where to look for jobs based on higher Experience level with an expected maximum salary?

In [ ]:
```
max_sal_job=ds_job_sal_df[['employment_type','experience_level','company_location']].v
max_sal_job
```

Out[ ]:
```
employment_type  experience_level  company_location
FT               SE                US                 216
                 MI                US                  87
                                   GB                  30
                 EN                US                  28
                 EX                US                  15
                                                     ...
                 MI                BE                   1
                                   AU                   1
                                   AE                   1
                 EX                PL                   1
PT               MI                NL                   1
Length: 106, dtype: int64
```

In [ ]:
```
max_sal_job.plot(x='experience_level',figsize=(12,9),grid=True);
```



It is no suprise the US Dominates this too as one most populous country with the Largest data network.

INFERENCES AND CONCLUSION

Based on the survey it is noticed the US dominates the Data science industry.

CT employees earns while pt earns the least.

Data scientist /engineers/analyst are demanded more in the indusrtry.

The higher your experience level ,the higher your salary