In [228…

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotnine as p9
import numpy as np
import os
import sqlalchemy
```

In [229…

```
from sqlalchemy import create_engine
```

In [230…

```
os.getcwd()
```

Out[230…

```
'/home/smartc78/DSE5002/Project 1'
```

In [231…

```
df = pd.read_csv('r project data-1-1.csv')
```

In [232…

```
print(f"Shape: {df.shape} (rows, columns)")
```

```
Shape: (607, 12) (rows, columns)
```

In [233…

```
column_types = df.dtypes
print(column_types)
```

```
Unnamed: 0             int64
work_year              int64
experience_level      object
employment_type       object
job_title             object
salary                 int64
salary_currency       object
salary_in_usd          int64
employee_residence    object
remote_ratio           int64
company_location      object
company_size          object
dtype: object
```

In [234…

```
df['remote_ratio'] = df['remote_ratio'].astype(str)
print("New data type of 'remote_ratio':", df['remote_ratio'].dtype)
```

```
New data type of 'remote_ratio': object
```

In [235…

```
df['work_year'] = df['work_year'].astype(str)
print("New data type of 'work_year':", df['work_year'].dtype)
```

```
New data type of 'work_year': object
```

In [236…

```
column_types = df.dtypes
print(column_types)
```

```
Unnamed: 0             int64
work_year             object
experience_level      object
employment_type       object
job_title             object
salary                 int64
salary_currency       object
salary_in_usd          int64
employee_residence    object
remote_ratio          object
company_location      object
company_size          object
dtype: object
```

In [237…

```python
print("df:")
df.info()
```

```
df:
<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    object
 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    object
 10  company_location    607 non-null    object
 11  company_size        607 non-null    object
dtypes: int64(3), object(9)
memory usage: 57.0+ KB
```

In [238…

```python
print("\nMissing values per column:")
print(df.isnull().sum())
```

```
Missing values per column:
Unnamed: 0            0
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 [239…

```
print("\nDescriptive statistics for numerical columns:")
print(df.describe())
```

```
Descriptive statistics for numerical columns:
       Unnamed: 0        salary  salary_in_usd
count  607.000000  6.070000e+02     607.000000
mean   303.000000  3.240001e+05  112297.869852
std    175.370085  1.544357e+06   70957.259411
min      0.000000  4.000000e+03    2859.000000
25%    151.500000  7.000000e+04   62726.000000
50%    303.000000  1.150000e+05  101570.000000
75%    454.500000  1.650000e+05  150000.000000
max    606.000000  3.040000e+07  600000.000000
```

In [240…

```
print("\nValue counts for categorical columns:")
for column in df.select_dtypes(include='object').columns:
    print(f"\n--- {column} ---")
    print(df[column].value_counts())
```

```
Value counts for categorical columns:

--- work_year ---
work_year
2022    318
2021    217
2020     72
Name: count, dtype: int64

--- experience_level ---
experience_level
SE    280
MI    213
EN     88
EX     26
Name: count, dtype: int64

--- employment_type ---
employment_type
FT    588
PT     10
CT      5
FL      4
Name: count, dtype: int64

--- job_title ---
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
Machine Learning Scientist                    8
Big Data Engineer                             8
Director of Data Science                      7
AI Scientist                                  7
Principal Data Scientist                      7
Data Science Consultant                       7
Data Analytics Manager                        7
BI Data Analyst                               6
Computer Vision Engineer                      6
ML Engineer                                   6
Lead Data Engineer                            6
Applied Data Scientist                        5
Business Data Analyst                         5
Data Engineering Manager                      5
Head of Data                                  5
Data Analytics Engineer                       4
Head of Data Science                          4
Applied Machine Learning Scientist            4
Analytics Engineer                            4
Machine Learning Developer                    3
Data Science Engineer                         3
Lead Data Analyst                             3
```

```
Machine Learning Infrastructure Engineer      3
Lead Data Scientist                           3
Principal Data Engineer                       3
Computer Vision Software Engineer             3
Product Data Analyst                          2
ETL Developer                                 2
Cloud Data Engineer                           2
Financial Data Analyst                        2
Director of Data Engineering                  2
Principal Data Analyst                        2
Machine Learning Manager                      1
Marketing Data Analyst                        1
3D Computer Vision Researcher                 1
Finance Data Analyst                          1
Data Specialist                               1
Staff Data Scientist                          1
Big Data Architect                            1
Head of Machine Learning                      1
NLP Engineer                                  1
Lead Machine Learning Engineer                1
Data Analytics Lead                           1
Name: count, dtype: int64

--- salary_currency ---
salary_currency
USD    398
EUR     95
GBP     44
INR     27
CAD     18
JPY      3
PLN      3
TRY      3
HUF      2
MXN      2
CNY      2
SGD      2
DKK      2
AUD      2
BRL      2
CLP      1
CHF      1
Name: count, dtype: int64

--- employee_residence ---
employee_residence
US    332
GB     44
IN     30
CA     29
DE     25
FR     18
ES     15
GR     13
JP      7
PT      6
```

```
PK      6
BR      6
NL      5
IT      4
RU      4
PL      4
AE      3
TR      3
AU      3
VN      3
AT      3
DK      2
NG      2
HU      2
MX      2
SI      2
RO      2
BE      2
SG      2
PH      1
CN      1
HN      1
NZ      1
UA      1
IQ      1
CL      1
MT      1
IR      1
CO      1
HR      1
BG      1
KE      1
MD      1
RS      1
HK      1
LU      1
JE      1
CZ      1
PR      1
AR      1
DZ      1
MY      1
TN      1
EE      1
BO      1
IE      1
CH      1
Name: count, dtype: int64

--- remote_ratio ---
remote_ratio
100    381
0      127
50      99
Name: count, dtype: int64
```

```
--- company_location ---
company_location
US    355
GB     47
CA     30
DE     28
IN     24
FR     15
ES     14
GR     11
JP      6
NL      4
PT      4
PL      4
AT      4
MX      3
DK      3
AE      3
PK      3
LU      3
TR      3
BR      3
AU      3
RU      2
CN      2
CH      2
BE      2
NG      2
SI      2
IT      2
CZ      2
NZ      1
HU      1
HN      1
SG      1
HR      1
MT      1
IL      1
UA      1
RO      1
IQ      1
MD      1
CL      1
IR      1
VN      1
KE      1
CO      1
AS      1
DZ      1
EE      1
MY      1
IE      1
Name: count, dtype: int64

--- company_size ---
company_size
```

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

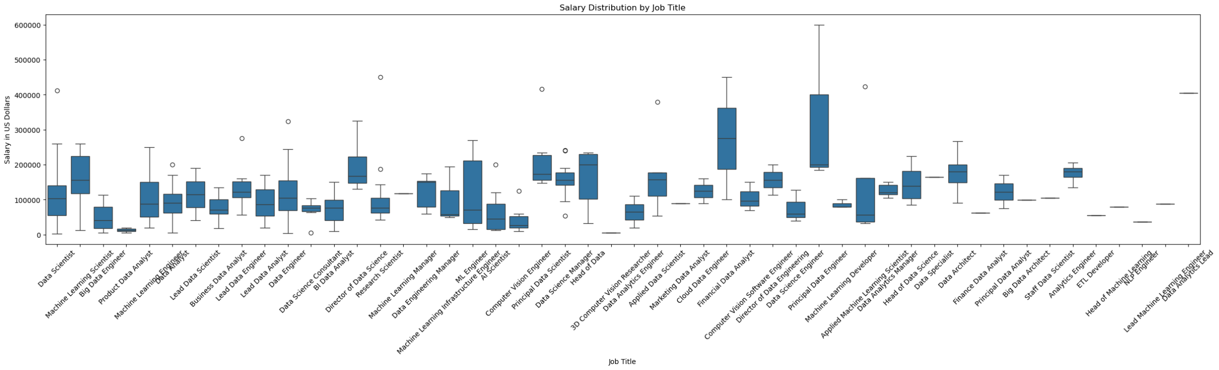
In [241…

```
print(df)
```

```
     Unnamed: 0  work_year experience_level employment_type  \
0             0       2020               MI              FT
1             1       2020               SE              FT
2             2       2020               SE              FT
3             3       2020               MI              FT
4             4       2020               SE              FT
..          ...        ...              ...             ...
602         602       2022               SE              FT
603         603       2022               SE              FT
604         604       2022               SE              FT
605         605       2022               SE              FT
606         606       2022               MI              FT

                      job_title  salary salary_currency  salary_in_usd  \
0                Data Scientist   70000             EUR          79833
1     Machine Learning Scientist  260000             USD         260000
2             Big Data Engineer   85000             GBP         109024
3          Product Data Analyst   20000             USD          20000
4      Machine Learning Engineer  150000             USD         150000
..                          ...     ...             ...            ...
602               Data Engineer  154000             USD         154000
603               Data Engineer  126000             USD         126000
604                Data Analyst  129000             USD         129000
605                Data Analyst  150000             USD         150000
606                AI Scientist  200000             USD         200000

    employee_residence remote_ratio company_location company_size
0                   DE            0               DE            L
1                   JP            0               JP            S
2                   GB           50               GB            M
3                   HN            0               HN            S
4                   US           50               US            L
..                 ...          ...              ...          ...
602                 US          100               US            M
603                 US          100               US            M
604                 US            0               US            M
605                 US          100               US            M
606                 IN          100               US            L

[607 rows x 12 columns]
```

In [242…

```
plt.figure(figsize=(30, 6))
sns.boxplot(x='job_title', y='salary_in_usd', data=df)

plt.title('Salary Distribution by Job Title')
plt.xlabel('Job Title')
plt.ylabel('Salary in US Dollars')
plt.xticks(rotation=45)
plt.show()
```



In [243…

```
jtdf=df.groupby("job_title")
print(jtdf.head())
```

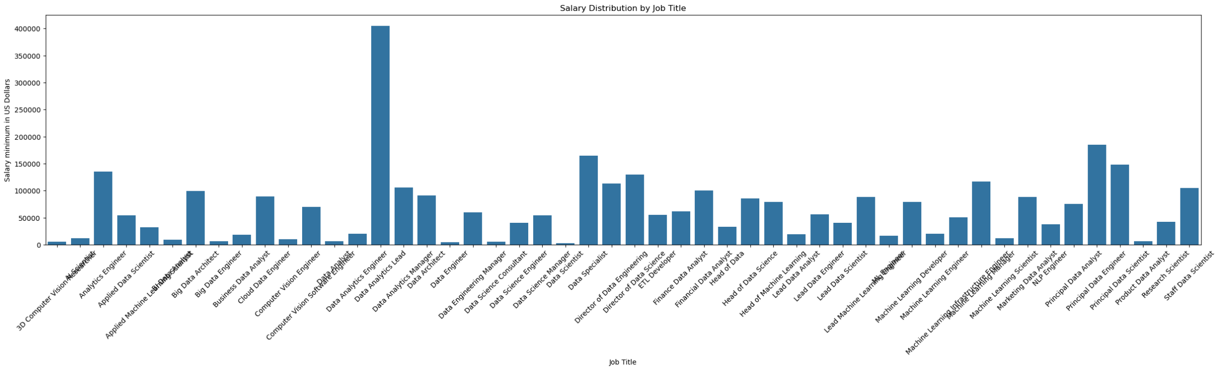
```
     Unnamed: 0 work_year experience_level employment_type  \
0             0      2020               MI              FT   
1             1      2020               SE              FT   
2             2      2020               SE              FT   
3             3      2020               MI              FT   
4             4      2020               SE              FT   
..          ...       ...              ...             ...   
519         519      2022               SE              FT   
523         523      2022               SE              FT   
525         525      2022               SE              FT   
560         560      2022               SE              FT   
561         561      2022               SE              FT   

                      job_title  salary salary_currency  salary_in_usd  \
0                Data Scientist   70000             EUR          79833   
1    Machine Learning Scientist  260000             USD         260000   
2             Big Data Engineer   85000             GBP         109024   
3          Product Data Analyst   20000             USD          20000   
4     Machine Learning Engineer  150000             USD         150000   
..                          ...     ...             ...            ...   
519      Applied Data Scientist  380000             USD         380000   
523         Data Analytics Lead  405000             USD         405000   
525      Applied Data Scientist  177000             USD         177000   
560          Analytics Engineer  205300             USD         205300   
561          Analytics Engineer  184700             USD         184700   

    employee_residence remote_ratio company_location company_size  
0                   DE            0               DE            L  
1                   JP            0               JP            S  
2                   GB           50               GB            M  
3                   HN            0               HN            S  
4                   US           50               US            L  
..                 ...          ...              ...          ...  
519                 US          100               US            L  
523                 US          100               US            L  
525                 US          100               US            L  
560                 US            0               US            M  
561                 US            0               US            M  

[170 rows x 12 columns]
```

```
In [244…  agg_jtdf = df.groupby('job_title')['salary_in_usd'].agg(salary_min='min', sa
          print(agg_jtdf)
```

```
                                   job_title  salary_min  salary_max
0              3D Computer Vision Researcher        5409        5409
1                               AI Scientist       12000      200000
2                         Analytics Engineer      135000      205300
3                     Applied Data Scientist       54238      380000
4         Applied Machine Learning Scientist       31875      423000
5                            BI Data Analyst        9272      150000
6                         Big Data Architect       99703       99703
7                          Big Data Engineer        5882      114047
8                      Business Data Analyst       18442      135000
9                        Cloud Data Engineer       89294      160000
10                  Computer Vision Engineer       10000      125000
11         Computer Vision Software Engineer       70000      150000
12                              Data Analyst        6072      200000
13                   Data Analytics Engineer       20000      110000
14                       Data Analytics Lead      405000      405000
15                    Data Analytics Manager      105400      150260
16                            Data Architect       90700      266400
17                             Data Engineer        4000      324000
18                  Data Engineering Manager       59303      174000
19                   Data Science Consultant        5707      103000
20                     Data Science Engineer       40189      127221
21                      Data Science Manager       54094      241000
22                            Data Scientist        2859      412000
23                           Data Specialist      165000      165000
24              Director of Data Engineering      113476      200000
25                  Director of Data Science      130026      325000
26                             ETL Developer       54957       54957
27                      Finance Data Analyst       61896       61896
28                    Financial Data Analyst      100000      450000
29                              Head of Data       32974      235000
30                      Head of Data Science       85000      224000
31                  Head of Machine Learning       79039       79039
32                         Lead Data Analyst       19609      170000
33                        Lead Data Engineer       56000      276000
34                       Lead Data Scientist       40570      190000
35            Lead Machine Learning Engineer       87932       87932
36                               ML Engineer       15966      270000
37                Machine Learning Developer       78791      100000
38                 Machine Learning Engineer       20000      250000
39  Machine Learning Infrastructure Engineer       50180      195000
40                  Machine Learning Manager      117104      117104
41                Machine Learning Scientist       12000      260000
42                    Marketing Data Analyst       88654       88654
43                              NLP Engineer       37236       37236
44                    Principal Data Analyst       75000      170000
45                   Principal Data Engineer      185000      600000
46                  Principal Data Scientist      148261      416000
47                      Product Data Analyst        6072       20000
48                        Research Scientist       42000      450000
49                      Staff Data Scientist      105000      105000
```

In [245…

```
plt.figure(figsize=(30, 6))
sns.barplot(x='job_title', y='salary_min', data=agg_jtdf)

plt.title('Salary Distribution by Job Title')
plt.xlabel('Job Title')
plt.ylabel('Salary minimum in US Dollars')
plt.xticks(rotation=45)

plt.show()
```



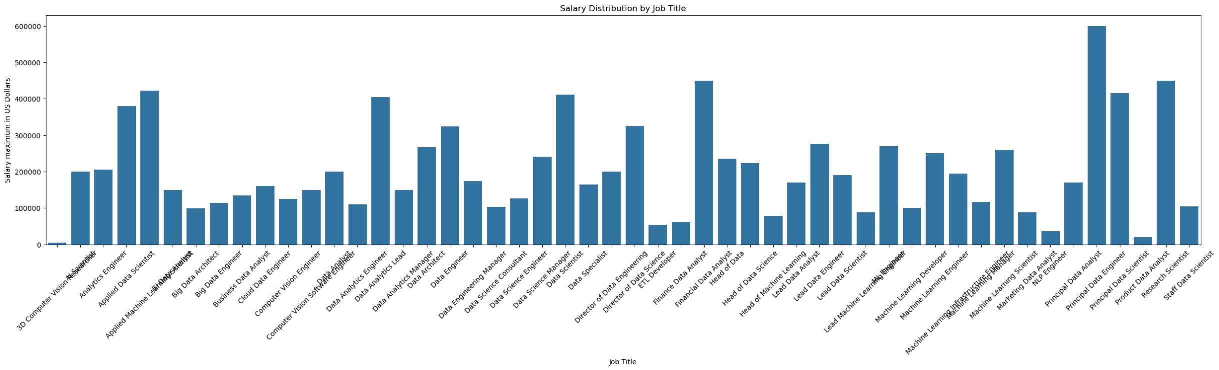
In [246…

```
sortminjtdf= agg_jtdf.sort_values(by='salary_min', ascending=True)
print(sortminjtdf)
```

```
                                   job_title  salary_min  salary_max
22                            Data Scientist        2859      412000
17                             Data Engineer        4000      324000
0              3D Computer Vision Researcher        5409        5409
19                   Data Science Consultant        5707      103000
7                          Big Data Engineer        5882      114047
12                              Data Analyst        6072      200000
47                      Product Data Analyst        6072       20000
5                            BI Data Analyst        9272      150000
10                  Computer Vision Engineer       10000      125000
1                               AI Scientist       12000      200000
41                Machine Learning Scientist       12000      260000
36                               ML Engineer       15966      270000
8                      Business Data Analyst       18442      135000
32                         Lead Data Analyst       19609      170000
13                   Data Analytics Engineer       20000      110000
38                 Machine Learning Engineer       20000      250000
4         Applied Machine Learning Scientist       31875      423000
29                              Head of Data       32974      235000
43                              NLP Engineer       37236       37236
20                     Data Science Engineer       40189      127221
34                       Lead Data Scientist       40570      190000
48                        Research Scientist       42000      450000
39  Machine Learning Infrastructure Engineer       50180      195000
21                      Data Science Manager       54094      241000
3                     Applied Data Scientist       54238      380000
26                             ETL Developer       54957       54957
33                        Lead Data Engineer       56000      276000
18                  Data Engineering Manager       59303      174000
27                      Finance Data Analyst       61896       61896
11         Computer Vision Software Engineer       70000      150000
44                    Principal Data Analyst       75000      170000
37                Machine Learning Developer       78791      100000
31                  Head of Machine Learning       79039       79039
30                      Head of Data Science       85000      224000
35            Lead Machine Learning Engineer       87932       87932
42                    Marketing Data Analyst       88654       88654
9                        Cloud Data Engineer       89294      160000
16                            Data Architect       90700      266400
6                         Big Data Architect       99703       99703
28                    Financial Data Analyst      100000      450000
49                      Staff Data Scientist      105000      105000
15                    Data Analytics Manager      105400      150260
24              Director of Data Engineering      113476      200000
40                  Machine Learning Manager      117104      117104
25                  Director of Data Science      130026      325000
2                         Analytics Engineer      135000      205300
46                  Principal Data Scientist      148261      416000
23                           Data Specialist      165000      165000
45                   Principal Data Engineer      185000      600000
14                       Data Analytics Lead      405000      405000
```

In [247…

```
plt.figure(figsize=(30, 6))
sns.barplot(x='job_title', y='salary_max', data=agg_jtdf)

plt.title('Salary Distribution by Job Title')
plt.xlabel('Job Title')
plt.ylabel('Salary maximum in US Dollars')
plt.xticks(rotation=45)

plt.show()
```



In [248…

```
sortmaxjtdf= agg_jtdf.sort_values(by='salary_max', ascending=True)
print(sortmaxjtdf)
```

```
                                   job_title  salary_min  salary_max
0              3D Computer Vision Researcher        5409        5409
47                      Product Data Analyst        6072       20000
43                              NLP Engineer       37236       37236
26                             ETL Developer       54957       54957
27                      Finance Data Analyst       61896       61896
31                  Head of Machine Learning       79039       79039
35            Lead Machine Learning Engineer       87932       87932
42                    Marketing Data Analyst       88654       88654
6                         Big Data Architect       99703       99703
37                Machine Learning Developer       78791      100000
19                   Data Science Consultant        5707      103000
49                      Staff Data Scientist      105000      105000
13                   Data Analytics Engineer       20000      110000
7                          Big Data Engineer        5882      114047
40                  Machine Learning Manager      117104      117104
10                  Computer Vision Engineer       10000      125000
20                     Data Science Engineer       40189      127221
8                      Business Data Analyst       18442      135000
5                            BI Data Analyst        9272      150000
11         Computer Vision Software Engineer       70000      150000
15                    Data Analytics Manager      105400      150260
9                        Cloud Data Engineer       89294      160000
23                           Data Specialist      165000      165000
32                         Lead Data Analyst       19609      170000
44                    Principal Data Analyst       75000      170000
18                  Data Engineering Manager       59303      174000
34                       Lead Data Scientist       40570      190000
39  Machine Learning Infrastructure Engineer       50180      195000
1                               AI Scientist       12000      200000
12                              Data Analyst        6072      200000
24              Director of Data Engineering      113476      200000
2                         Analytics Engineer      135000      205300
30                      Head of Data Science       85000      224000
29                              Head of Data       32974      235000
21                      Data Science Manager       54094      241000
38                 Machine Learning Engineer       20000      250000
41                Machine Learning Scientist       12000      260000
16                            Data Architect       90700      266400
36                               ML Engineer       15966      270000
33                        Lead Data Engineer       56000      276000
17                             Data Engineer        4000      324000
25                  Director of Data Science      130026      325000
3                     Applied Data Scientist       54238      380000
14                       Data Analytics Lead      405000      405000
22                            Data Scientist        2859      412000
46                  Principal Data Scientist      148261      416000
4         Applied Machine Learning Scientist       31875      423000
28                    Financial Data Analyst      100000      450000
48                        Research Scientist       42000      450000
45                   Principal Data Engineer      185000      600000
```

In [249…

```
exdf=df.groupby("experience_level")
print(exdf.head())
```

```
    Unnamed: 0 work_year experience_level employment_type  \
0            0      2020               MI              FT   
1            1      2020               SE              FT   
2            2      2020               SE              FT   
3            3      2020               MI              FT   
4            4      2020               SE              FT   
5            5      2020               EN              FT   
6            6      2020               SE              FT   
7            7      2020               MI              FT   
8            8      2020               MI              FT   
9            9      2020               SE              FT   
10          10      2020               EN              FT   
11          11      2020               MI              FT   
12          12      2020               EN              FT   
16          16      2020               EN              FT   
18          18      2020               EN              FT   
25          25      2020               EX              FT   
41          41      2020               EX              FT   
73          73      2021               EX              FT   
74          74      2021               EX              FT   
84          84      2021               EX              FT   

                     job_title    salary salary_currency  salary_in_usd  \
0               Data Scientist     70000             EUR          79833   
1   Machine Learning Scientist    260000             USD         260000   
2            Big Data Engineer     85000             GBP         109024   
3         Product Data Analyst     20000             USD          20000   
4    Machine Learning Engineer    150000             USD         150000   
5                 Data Analyst     72000             USD          72000   
6          Lead Data Scientist    190000             USD         190000   
7               Data Scientist  11000000             HUF          35735   
8        Business Data Analyst    135000             USD         135000   
9           Lead Data Engineer    125000             USD         125000   
10              Data Scientist     45000             EUR          51321   
11              Data Scientist   3000000             INR          40481   
12              Data Scientist     35000             EUR          39916   
16               Data Engineer   4450000             JPY          41689   
18     Data Science Consultant    423000             INR           5707   
25    Director of Data Science    325000             USD         325000   
41    Data Engineering Manager     70000             EUR          79833   
73             BI Data Analyst    150000             USD         150000   
74                Head of Data    235000             USD         235000   
84    Director of Data Science    130000             EUR         153667   

   employee_residence remote_ratio company_location company_size  
0                  DE            0               DE            L  
1                  JP            0               JP            S  
2                  GB           50               GB            M  
3                  HN            0               HN            S  
4                  US           50               US            L  
5                  US          100               US            L  
6                  US          100               US            S  
7                  HU           50               HU            L  
8                  US          100               US            L  
9                  NZ           50               NZ            S  
10                 FR            0               FR            S  
```

```
11                  IN           0             IN            L
12                  FR           0             FR            M
16                  JP         100             JP            S
18                  IN          50             IN            M
25                  US         100             US            L
41                  ES          50             ES            L
73                  IN         100             US            L
74                  US         100             US            L
84                  IT         100             PL            L
```

In [250…

```
agg_exdf = df.groupby(['experience_level','employment_type','job_title','sal
print(agg_exdf)
```

```
    experience_level employment_type                           job_title  \
0                 EN              CT  Applied Machine Learning Scientist   
1                 EN              CT               Business Data Analyst   
2                 EN              FT                        AI Scientist   
3                 EN              FT                        AI Scientist   
4                 EN              FT              Applied Data Scientist   
..               ...             ...                                 ...   
523               SE              FT                  Research Scientist   
524               SE              FT                  Research Scientist   
525               SE              FT                  Research Scientist   
526               SE              FT                  Research Scientist   
527               SE              FT                  Research Scientist   

     salary_in_usd  salary_min  salary_max  
0            31875       31875       31875  
1           100000      100000      100000  
2            18053       18053       18053  
3            45896       45896       45896  
4           110037      110037      110037  
..             ...         ...         ...  
523          50000       50000       50000  
524          60757       60757       60757  
525          93427       93427       93427  
526          96113       96113       96113  
527         144000      144000      144000  

[528 rows x 6 columns]
```

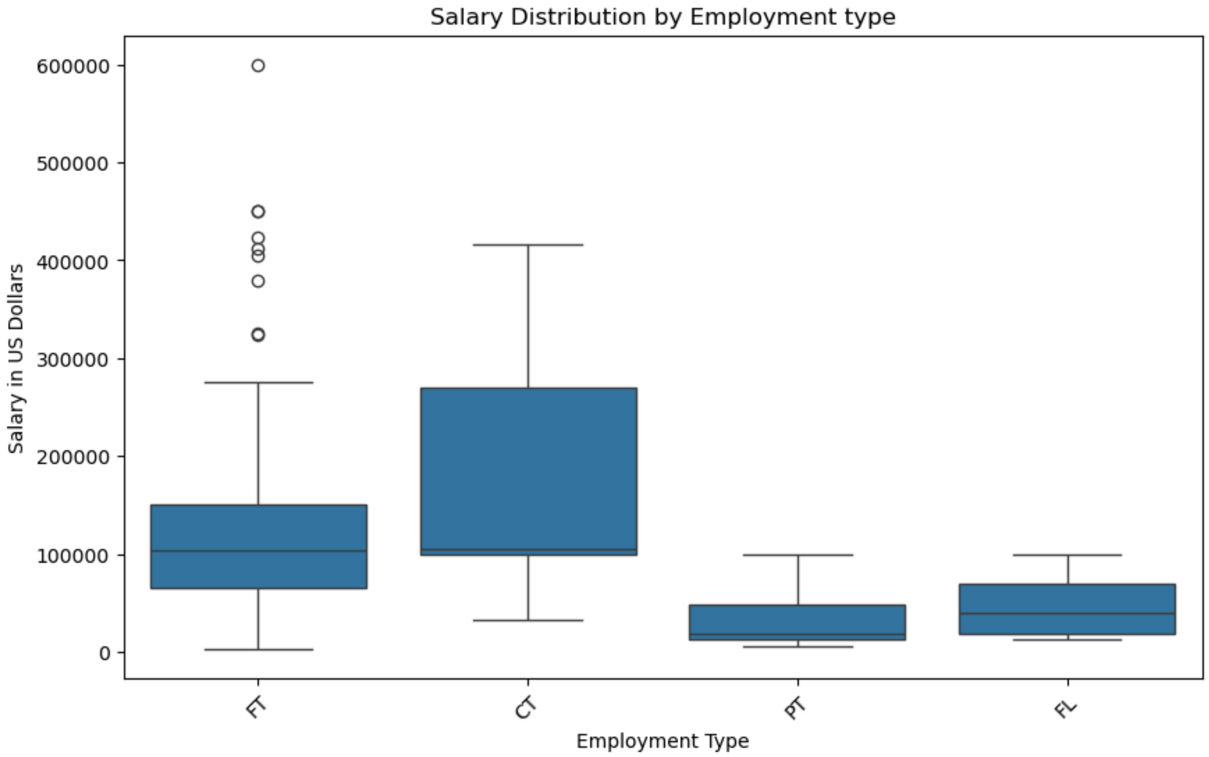
In [251…

```
et_cat = df['employment_type'].unique()
print(et_cat)
```

```
['FT' 'CT' 'PT' 'FL']
```

In [252…

```
plt.figure(figsize=(10, 6))
sns.boxplot(x='employment_type', y='salary_in_usd', data=df)

plt.title('Salary Distribution by Employment type')
plt.xlabel('Employment Type')
plt.ylabel('Salary in US Dollars')
plt.xticks(rotation=45)

plt.show()
```



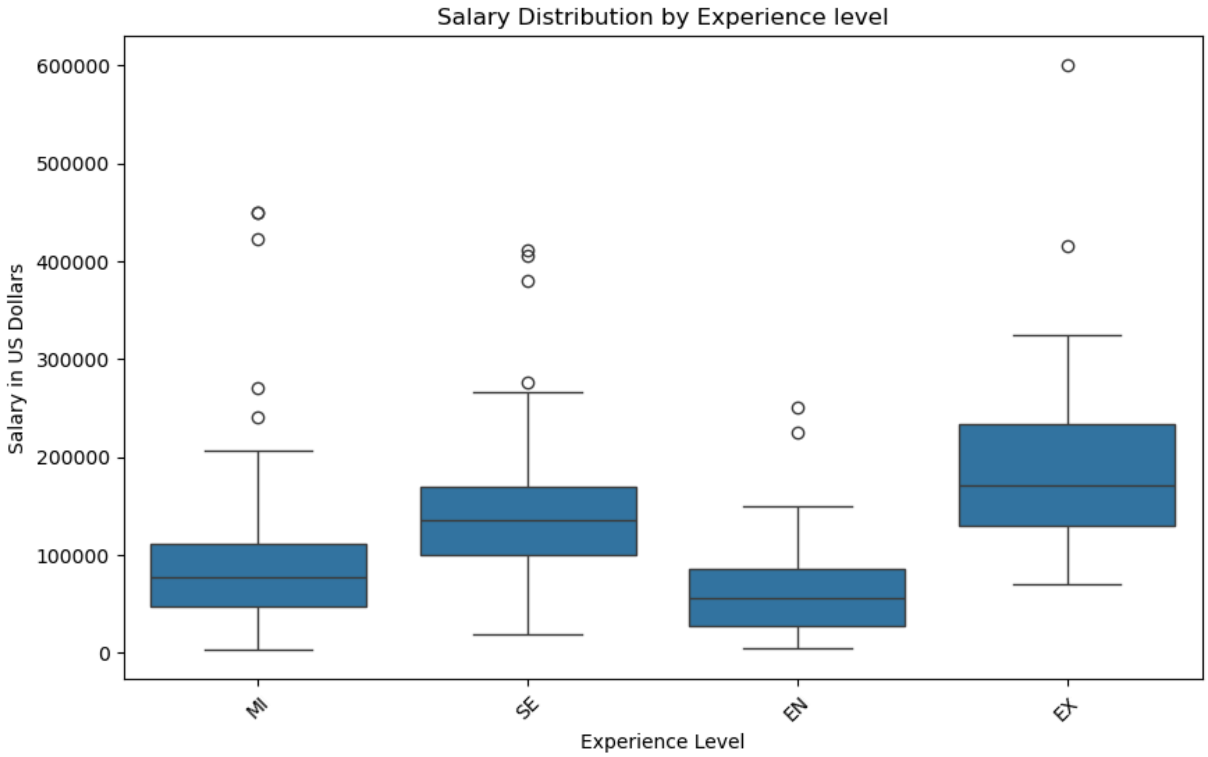
In [253…

```
ex_cat = df['experience_level'].unique()
print(ex_cat)
```

```
['MI' 'SE' 'EN' 'EX']
```

In [254…

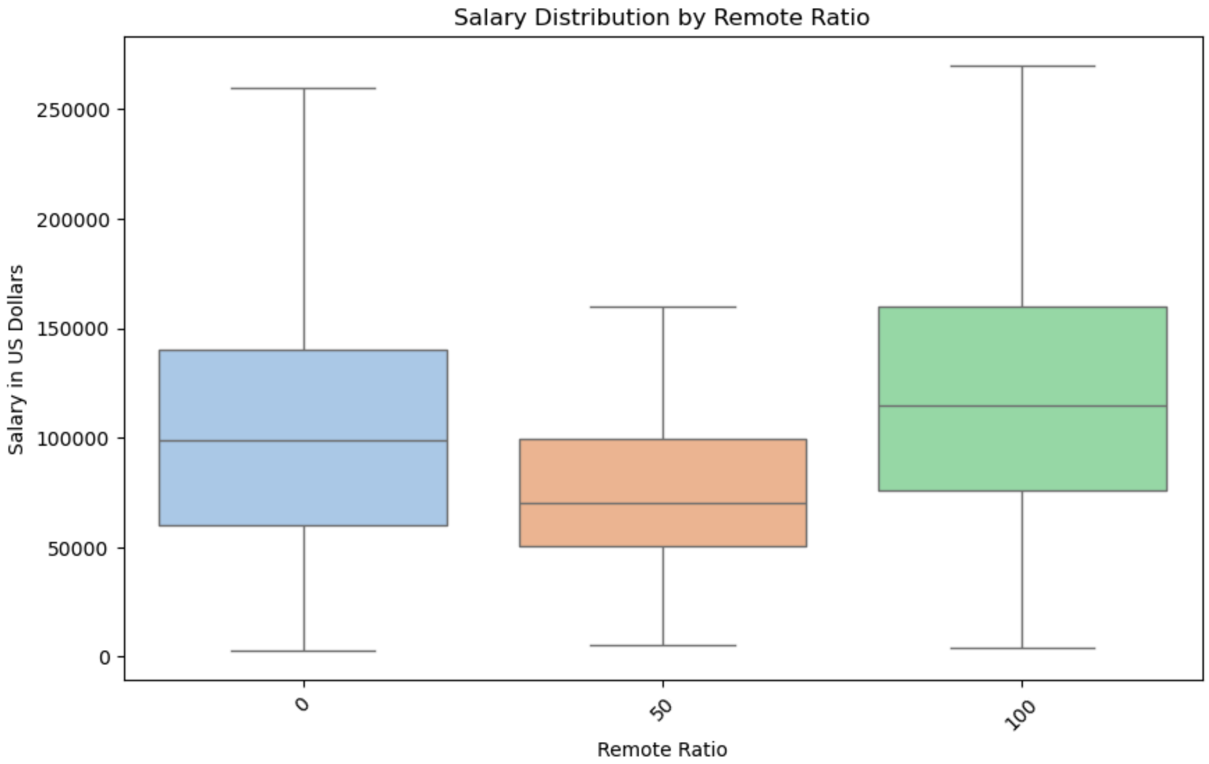
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='experience_level', y='salary_in_usd', data=df)

plt.title('Salary Distribution by Experience level')
plt.xlabel('Experience Level')
plt.ylabel('Salary in US Dollars')
plt.xticks(rotation=45)

plt.show()
```



```
In [255…  rr_cat = df['remote_ratio'].unique()
          print(rr_cat)
```

```
['0' '50' '100']
```

```
In [322…  plt.figure(figsize=(10, 6))
          sns.boxplot(x='remote_ratio', y='salary_in_usd', data=df,palette='pastel',sh

          plt.title('Salary Distribution by Remote Ratio')
          plt.xlabel('Remote Ratio')
          plt.ylabel('Salary in US Dollars')
          plt.xticks(rotation=45)

          plt.show()
```



In [257…

```
jt_cat = df['job_title'].unique()
print(jt_cat)
```

```
['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']
```

In [258…

```
    num_unique_categories = df['job_title'].nunique()
    print(f"Number of unique categories: {num_unique_categories}")
```

```
Number of unique categories: 50
```

In [259…

```
df1= df.copy()
```

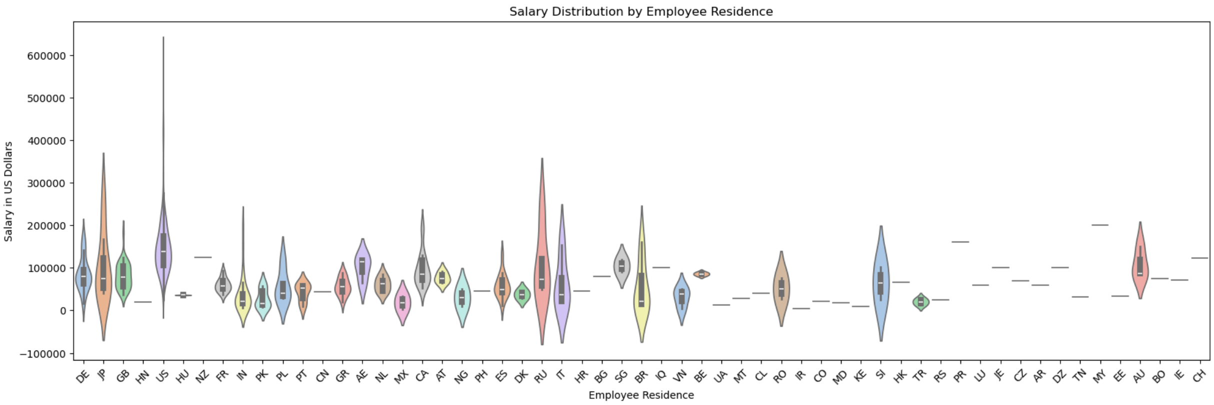
In [260…

```
print(df1)
```

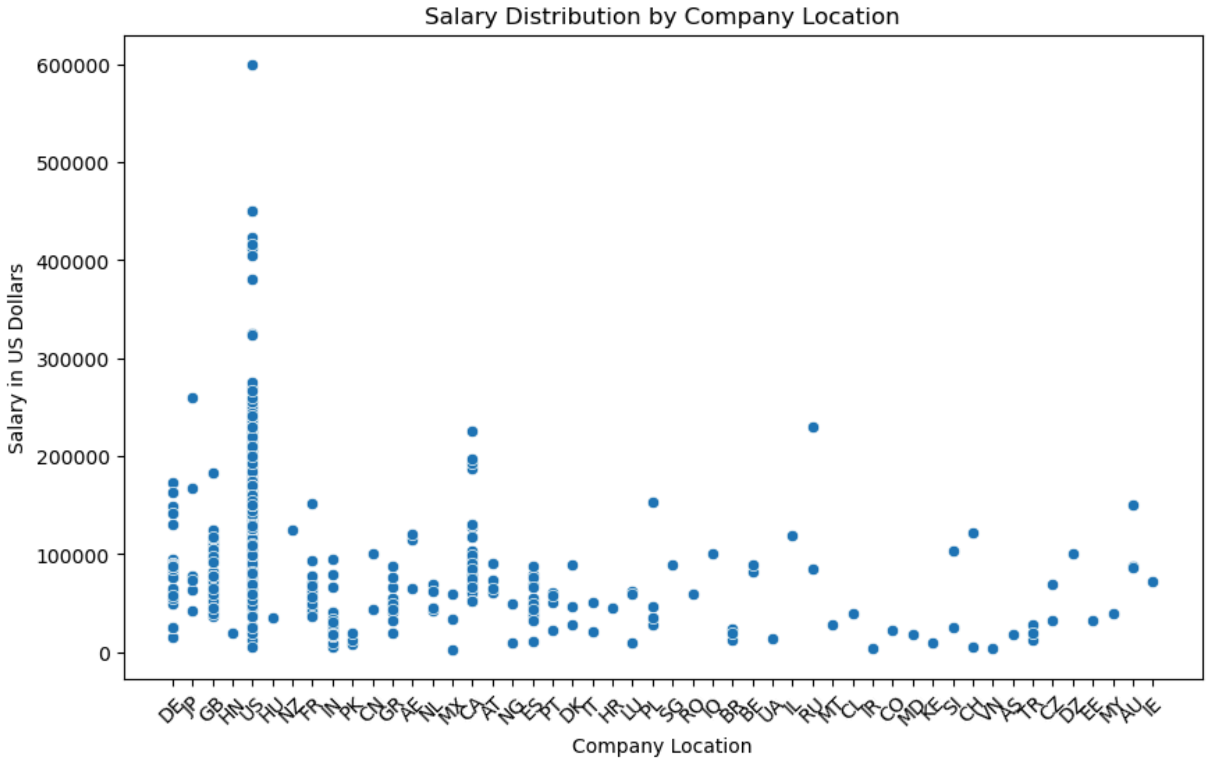
```
     Unnamed: 0 work_year experience_level employment_type  \
0             0      2020               MI              FT
1             1      2020               SE              FT
2             2      2020               SE              FT
3             3      2020               MI              FT
4             4      2020               SE              FT
..          ...       ...              ...             ...
602         602      2022               SE              FT
603         603      2022               SE              FT
604         604      2022               SE              FT
605         605      2022               SE              FT
606         606      2022               MI              FT

                      job_title  salary salary_currency  salary_in_usd  \
0                Data Scientist   70000             EUR          79833
1    Machine Learning Scientist  260000             USD         260000
2             Big Data Engineer   85000             GBP         109024
3          Product Data Analyst   20000             USD          20000
4     Machine Learning Engineer  150000             USD         150000
..                          ...     ...             ...            ...
602               Data Engineer  154000             USD         154000
603               Data Engineer  126000             USD         126000
604                Data Analyst  129000             USD         129000
605                Data Analyst  150000             USD         150000
606                AI Scientist  200000             USD         200000

    employee_residence remote_ratio company_location company_size
0                   DE            0               DE            L
1                   JP            0               JP            S
2                   GB           50               GB            M
3                   HN            0               HN            S
4                   US           50               US            L
..                 ...          ...              ...          ...
602                 US          100               US            M
603                 US          100               US            M
604                 US            0               US            M
605                 US          100               US            M
606                 IN          100               US            L

[607 rows x 12 columns]
```

In [261…

```
plt.figure(figsize=(20, 6))
sns.violinplot(x='employee_residence', y='salary_in_usd', data=df, palette='

plt.title('Salary Distribution by Employee Residence')
plt.xlabel('Employee Residence')
plt.ylabel('Salary in US Dollars')
plt.xticks(rotation=45)

plt.show()
```



```
In [262…  plt.figure(figsize=(10, 6))
          sns.scatterplot(x='company_location', y='salary_in_usd', data=df)

          plt.title('Salary Distribution by Company Location')
          plt.xlabel('Company Location')
          plt.ylabel('Salary in US Dollars')
          plt.xticks(rotation=45)

          plt.show()
```



```
In [263…  remove_employment_types = ['FL', 'CT', 'PT']
          df1ft= df1[~df1['employment_type'].isin(remove_employment_types)]
          print(df1ft)
```

```
     Unnamed: 0 work_year experience_level employment_type  \
0             0      2020               MI              FT
1             1      2020               SE              FT
2             2      2020               SE              FT
3             3      2020               MI              FT
4             4      2020               SE              FT
..          ...       ...              ...             ...
602         602      2022               SE              FT
603         603      2022               SE              FT
604         604      2022               SE              FT
605         605      2022               SE              FT
606         606      2022               MI              FT

                      job_title  salary salary_currency  salary_in_usd  \
0                Data Scientist   70000             EUR          79833
1    Machine Learning Scientist  260000             USD         260000
2             Big Data Engineer   85000             GBP         109024
3          Product Data Analyst   20000             USD          20000
4     Machine Learning Engineer  150000             USD         150000
..                          ...     ...             ...            ...
602               Data Engineer  154000             USD         154000
603               Data Engineer  126000             USD         126000
604                Data Analyst  129000             USD         129000
605                Data Analyst  150000             USD         150000
606                AI Scientist  200000             USD         200000

    employee_residence remote_ratio company_location company_size
0                   DE            0               DE            L
1                   JP            0               JP            S
2                   GB           50               GB            M
3                   HN            0               HN            S
4                   US           50               US            L
..                 ...          ...              ...          ...
602                 US          100               US            M
603                 US          100               US            M
604                 US            0               US            M
605                 US          100               US            M
606                 IN          100               US            L

[588 rows x 12 columns]
```

In [264…

```
selected_columns = ['Unnamed: 0', 'work_year', 'experience_level', 'job_titl
df2= df1ft[selected_columns].copy()
print(df2)
```

```
     Unnamed: 0 work_year experience_level                   job_title  \
0             0      2020               MI              Data Scientist
1             1      2020               SE  Machine Learning Scientist
2             2      2020               SE           Big Data Engineer
3             3      2020               MI        Product Data Analyst
4             4      2020               SE   Machine Learning Engineer
..          ...       ...              ...                         ...
602         602      2022               SE               Data Engineer
603         603      2022               SE               Data Engineer
604         604      2022               SE                Data Analyst
605         605      2022               SE                Data Analyst
606         606      2022               MI                AI Scientist

     salary_in_usd company_size
0            79833            L
1           260000            S
2           109024            M
3            20000            S
4           150000            L
..             ...          ...
602         154000            M
603         126000            M
604         129000            M
605         150000            M
606         200000            L

[588 rows x 6 columns]
```

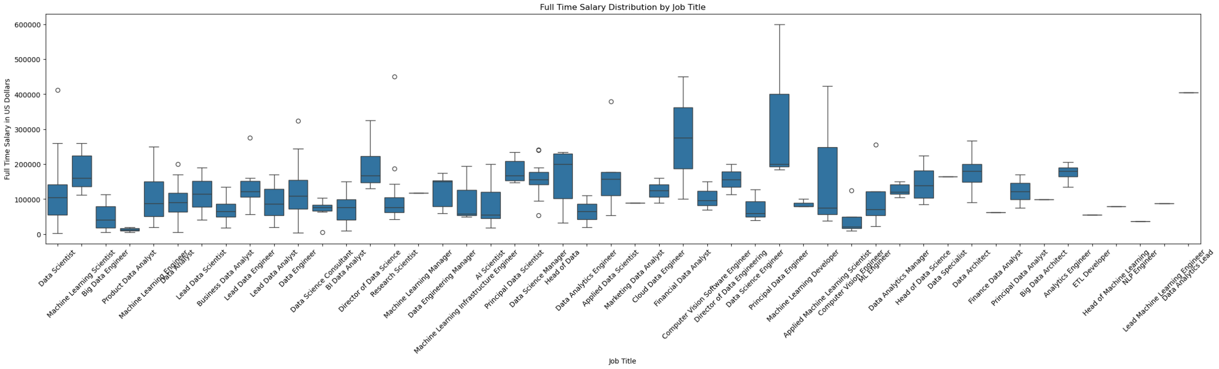
In [265…

```
remove_over_under_qualified = ['EN', 'EX']
df3= df2[~df2['experience_level'].isin(remove_over_under_qualified)]
print(df3)
```

```
     Unnamed: 0  work_year experience_level                  job_title  \
0             0       2020               MI             Data Scientist   
1             1       2020               SE  Machine Learning Scientist   
2             2       2020               SE          Big Data Engineer   
3             3       2020               MI       Product Data Analyst   
4             4       2020               SE  Machine Learning Engineer   
..          ...        ...              ...                        ...   
602         602       2022               SE              Data Engineer   
603         603       2022               SE              Data Engineer   
604         604       2022               SE               Data Analyst   
605         605       2022               SE               Data Analyst   
606         606       2022               MI               AI Scientist   

     salary_in_usd company_size  
0            79833            L  
1           260000            S  
2           109024            M  
3            20000            S  
4           150000            L  
..             ...          ...  
602         154000            M  
603         126000            M  
604         129000            M  
605         150000            M  
606         200000            L  

[484 rows x 6 columns]
```

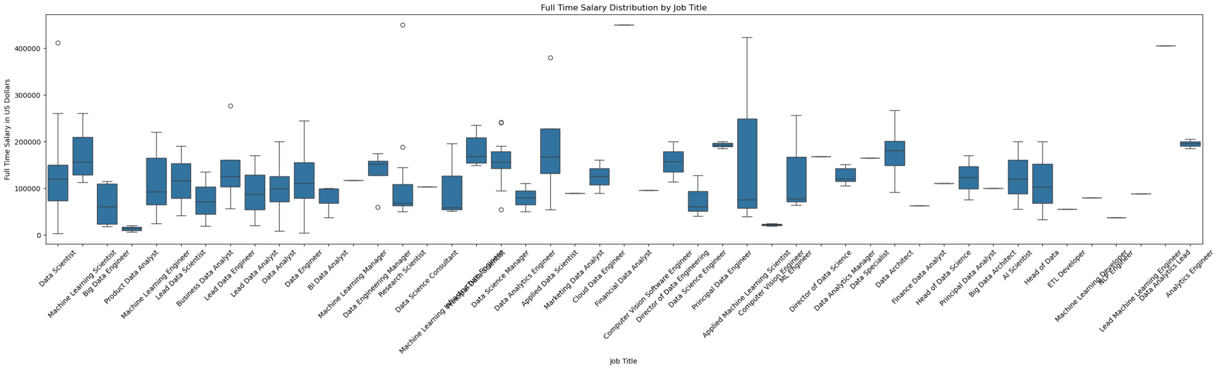
In [266…

```python
plt.figure(figsize=(30, 6))
sns.boxplot(x='job_title', y='salary_in_usd', data=df2)

plt.title('Full Time Salary Distribution by Job Title')
plt.xlabel('Job Title')
plt.ylabel('Full Time Salary in US Dollars')
plt.xticks(rotation=45)

plt.show()
```



In [267…

```python
plt.figure(figsize=(30, 6))
sns.boxplot(x='job_title', y='salary_in_usd', data=df3)

plt.title('Full Time Salary Distribution by Job Title')
plt.xlabel('Job Title')
plt.ylabel('Full Time Salary in US Dollars')
plt.xticks(rotation=45)
plt.show()
```



In [268…

```
print("\nDescriptive statistics for numerical columns:")
print(df.describe())
```

```
Descriptive statistics for numerical columns:
       Unnamed: 0        salary  salary_in_usd
count  607.000000  6.070000e+02     607.000000
mean   303.000000  3.240001e+05  112297.869852
std    175.370085  1.544357e+06   70957.259411
min      0.000000  4.000000e+03    2859.000000
25%    151.500000  7.000000e+04   62726.000000
50%    303.000000  1.150000e+05  101570.000000
75%    454.500000  1.650000e+05  150000.000000
max    606.000000  3.040000e+07  600000.000000
```

In [269…

```
print("\nDescriptive statistics for numerical columns:")
print(df3.describe())
```

```
Descriptive statistics for numerical columns:
       Unnamed: 0  salary_in_usd
count  484.000000     484.000000
mean   322.082645  117477.055785
std    173.946047   64959.617369
min      0.000000    2859.000000
25%    178.500000   71933.000000
50%    334.500000  111350.000000
75%    469.250000  154150.000000
max    606.000000  450000.000000
```

```
"""df vs df3
```

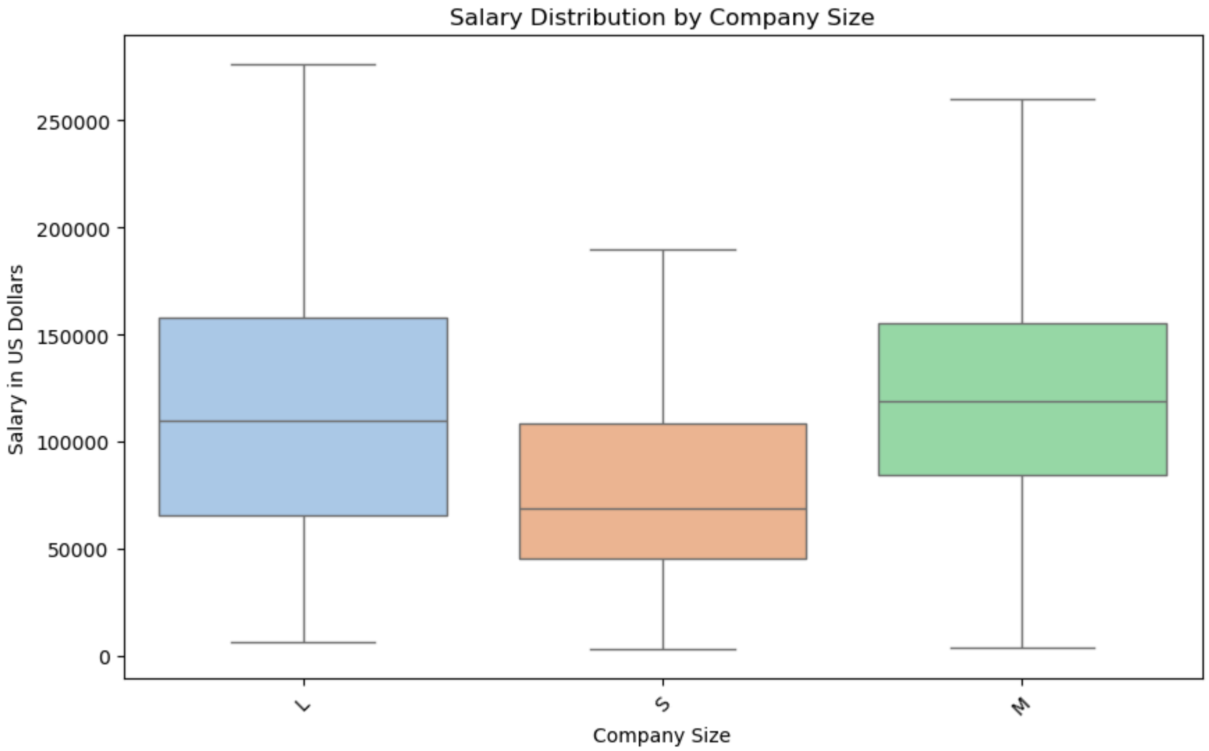
In [270…

```
df3sort= df3.sort_values(by='salary_in_usd', ascending=True)
print(df3sort)
```

```
     Unnamed: 0 work_year experience_level  \
176         176      2021               MI   
185         185      2021               MI   
179         179      2021               MI   
21           21      2020               MI   
15           15      2020               MI   
..          ...       ...              ...   
523         523      2022               SE   
63           63      2020               SE   
157         157      2021               MI   
97           97      2021               MI   
33           33      2020               MI   

                              job_title  salary_in_usd company_size  
176                      Data Scientist           2859            S  
185                       Data Engineer           4000            M  
179                      Data Scientist           5679            S  
21                 Product Data Analyst           6072            L  
15                         Data Analyst           8000            L  
..                                  ...            ...          ...  
523                 Data Analytics Lead         405000            L  
63                       Data Scientist         412000            L  
157  Applied Machine Learning Scientist         423000            L  
97               Financial Data Analyst         450000            L  
33                   Research Scientist         450000            M  

[484 rows x 6 columns]
```

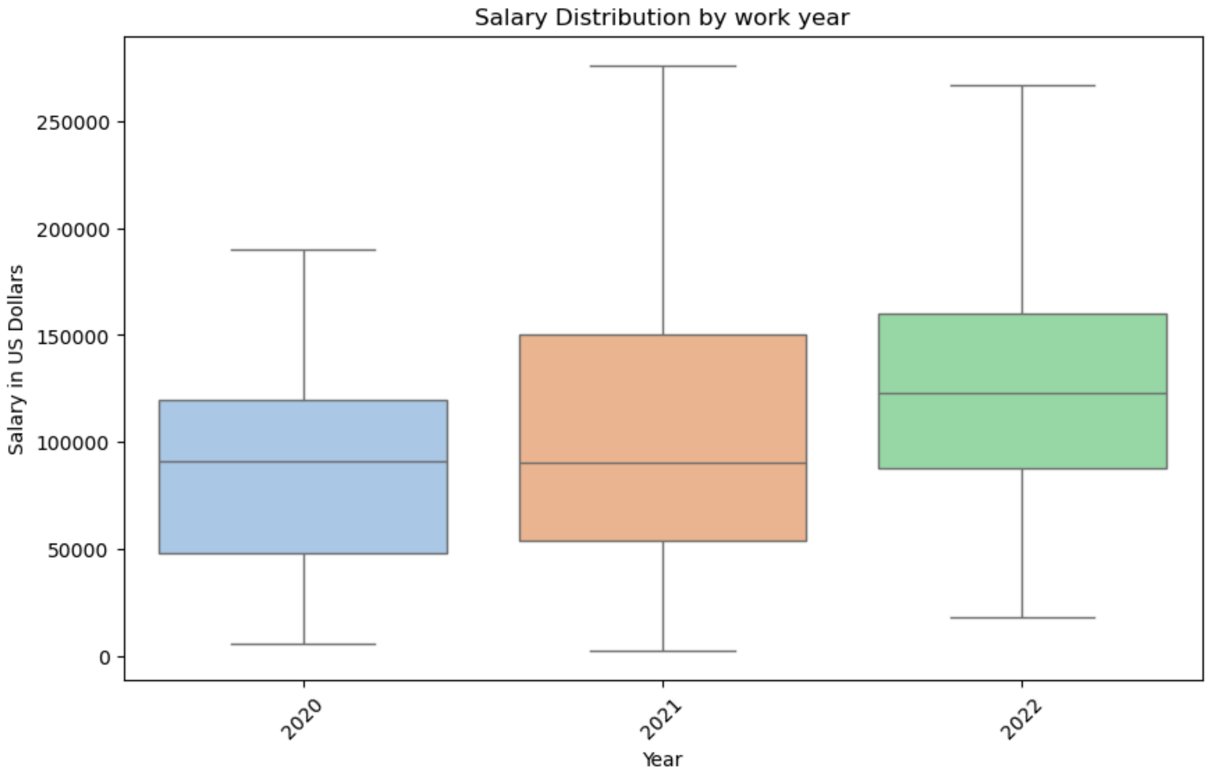
In [271…

```python
plt.figure(figsize=(10, 6))
sns.boxplot(x='company_size', y='salary_in_usd',hue='company_size', data=df3

plt.title('Salary Distribution by Company Size')
plt.xlabel('Company Size')
plt.ylabel('Salary in US Dollars')
plt.xticks(rotation=45)

plt.show()
```



In [272…

```
plt.figure(figsize=(10, 6))
sns.boxplot(x='work_year', y='salary_in_usd',hue= 'work_year', data=df3, sho

plt.title('Salary Distribution by work year')
plt.xlabel('Year')
plt.ylabel('Salary in US Dollars')
plt.xticks(rotation=45)

plt.show()
```



In [273…

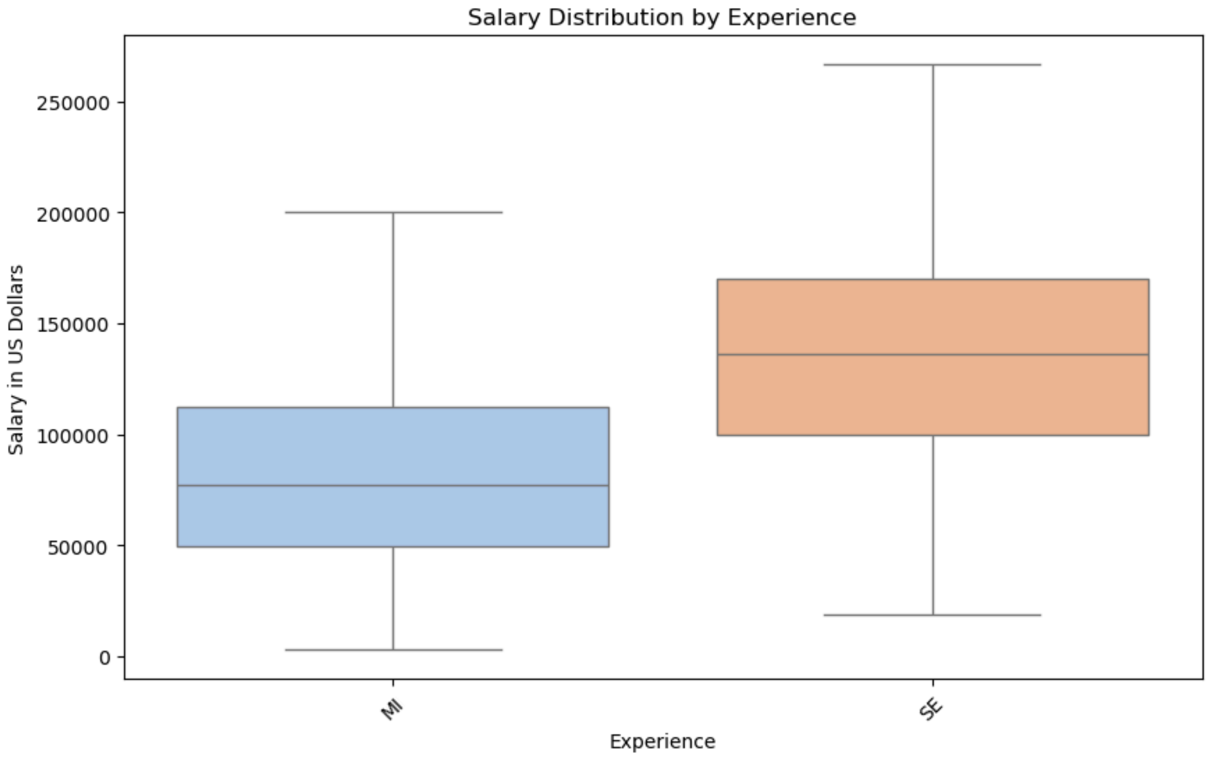
```
print("\nValue counts for categorical columns:")
for column in df3.select_dtypes(include='object').columns:
    print(f"\n--- {column} ---")
    print(df3[column].value_counts())
```

```
Value counts for categorical columns:

--- work_year ---
work_year
2022    282
2021    153
2020     49
Name: count, dtype: int64

--- experience_level ---
experience_level
SE    278
MI    206
Name: count, dtype: int64

--- job_title ---
job_title
Data Scientist                              120
Data Engineer                               113
Data Analyst                                 83
Machine Learning Engineer                    32
Research Scientist                           12
Data Science Manager                         12
Data Architect                               11
Data Analytics Manager                        7
Principal Data Scientist                      6
Machine Learning Scientist                    6
Big Data Engineer                             5
Lead Data Engineer                            5
Data Engineering Manager                      4
Applied Data Scientist                        4
Lead Data Scientist                           3
AI Scientist                                  3
Head of Data                                  3
Lead Data Analyst                             3
BI Data Analyst                               3
Business Data Analyst                         3
ML Engineer                                   3
Data Analytics Engineer                       3
Machine Learning Infrastructure Engineer      3
Applied Machine Learning Scientist            3
Data Science Engineer                         3
Machine Learning Developer                    2
Product Data Analyst                          2
Principal Data Engineer                       2
Computer Vision Engineer                      2
Cloud Data Engineer                           2
Director of Data Engineering                  2
Principal Data Analyst                        2
ETL Developer                                 2
Analytics Engineer                            2
Machine Learning Manager                      1
Computer Vision Software Engineer             1
Financial Data Analyst                        1
Data Science Consultant                       1
Marketing Data Analyst                        1
```

```
Big Data Architect                              1
Head of Data Science                            1
Finance Data Analyst                            1
Data Specialist                                 1
Director of Data Science                        1
NLP Engineer                                    1
Lead Machine Learning Engineer                  1
Data Analytics Lead                             1
Name: count, dtype: int64

--- company_size ---
company_size
M    280
L    154
S     50
Name: count, dtype: int64
```

In [274…

```
plt.figure(figsize=(10, 6))
sns.boxplot(x='experience_level', y='salary_in_usd', hue='experience_level',

plt.title('Salary Distribution by Experience')
plt.xlabel('Experience')
plt.ylabel('Salary in US Dollars')
plt.xticks(rotation=45)

plt.show()
```



In [275…

```
df3_mi = df3[df3['experience_level'] == 'MI'].copy()
df3_se = df3[df3['experience_level'] == 'SE'].copy()
print("DataFrame df3_mi:")
print(df3_mi)
print("\nDataFrame df3_se:")
print(df3_se)
```

```
DataFrame df3_mi:
     Unnamed: 0 work_year experience_level              job_title  \
0             0      2020               MI         Data Scientist   
3             3      2020               MI   Product Data Analyst   
7             7      2020               MI         Data Scientist   
8             8      2020               MI  Business Data Analyst   
11           11      2020               MI         Data Scientist   
..          ...       ...              ...                    ...   
567         567      2022               MI           Data Analyst   
586         586      2022               MI           Data Analyst   
598         598      2022               MI         Data Scientist   
599         599      2022               MI         Data Scientist   
606         606      2022               MI           AI Scientist   

     salary_in_usd company_size  
0            79833            L  
3            20000            S  
7            35735            L  
8           135000            L  
11           40481            L  
..             ...          ...  
567          65438            M  
586          45807            M  
598         160000            M  
599         130000            M  
606         200000            L  

[206 rows x 6 columns]

DataFrame df3_se:
     Unnamed: 0 work_year experience_level                   job_title  \
1             1      2020               SE  Machine Learning Scientist   
2             2      2020               SE           Big Data Engineer   
4             4      2020               SE   Machine Learning Engineer   
6             6      2020               SE         Lead Data Scientist   
9             9      2020               SE          Lead Data Engineer   
..          ...       ...              ...                         ...   
597         597      2022               SE                Data Analyst   
602         602      2022               SE               Data Engineer   
603         603      2022               SE               Data Engineer   
604         604      2022               SE                Data Analyst   
605         605      2022               SE                Data Analyst   

     salary_in_usd company_size  
1           260000            S  
2           109024            M  
4           150000            L  
6           190000            S  
9           125000            S  
..             ...          ...  
597         170000            M  
602         154000            M  
603         126000            M  
604         129000            M  
605         150000            M  
```

```
[278 rows x 6 columns]
```

In [276…

```
print("\nValue counts for categorical columns:")
for column in df3_mi.select_dtypes(include='object').columns:
    print(f"\n--- {column} ---")
    print(df3_mi[column].value_counts())
```

```
Value counts for categorical columns:

--- work_year ---
work_year
2022    89
2021    85
2020    32
Name: count, dtype: int64

--- experience_level ---
experience_level
MI    206
Name: count, dtype: int64

--- job_title ---
job_title
Data Scientist                              59
Data Engineer                               50
Data Analyst                                29
Machine Learning Engineer                   12
Research Scientist                           7
Machine Learning Scientist                   3
BI Data Analyst                              3
Business Data Analyst                        3
Applied Machine Learning Scientist           3
Big Data Engineer                            3
Data Architect                               3
Product Data Analyst                         2
AI Scientist                                 2
Data Science Manager                         2
Applied Data Scientist                       2
ETL Developer                                2
Lead Data Analyst                            2
Machine Learning Infrastructure Engineer     2
ML Engineer                                  2
Financial Data Analyst                       1
Data Analytics Engineer                      1
Cloud Data Engineer                          1
Lead Data Scientist                          1
Lead Data Engineer                           1
Data Engineering Manager                     1
Data Science Consultant                      1
Head of Data Science                         1
Computer Vision Software Engineer            1
Data Science Engineer                        1
Principal Data Scientist                     1
Machine Learning Developer                   1
NLP Engineer                                 1
Principal Data Analyst                       1
Head of Data                                 1
Name: count, dtype: int64

--- company_size ---
company_size
M    95
L    82
```

```
S    29
Name: count, dtype: int64
```

In [277…

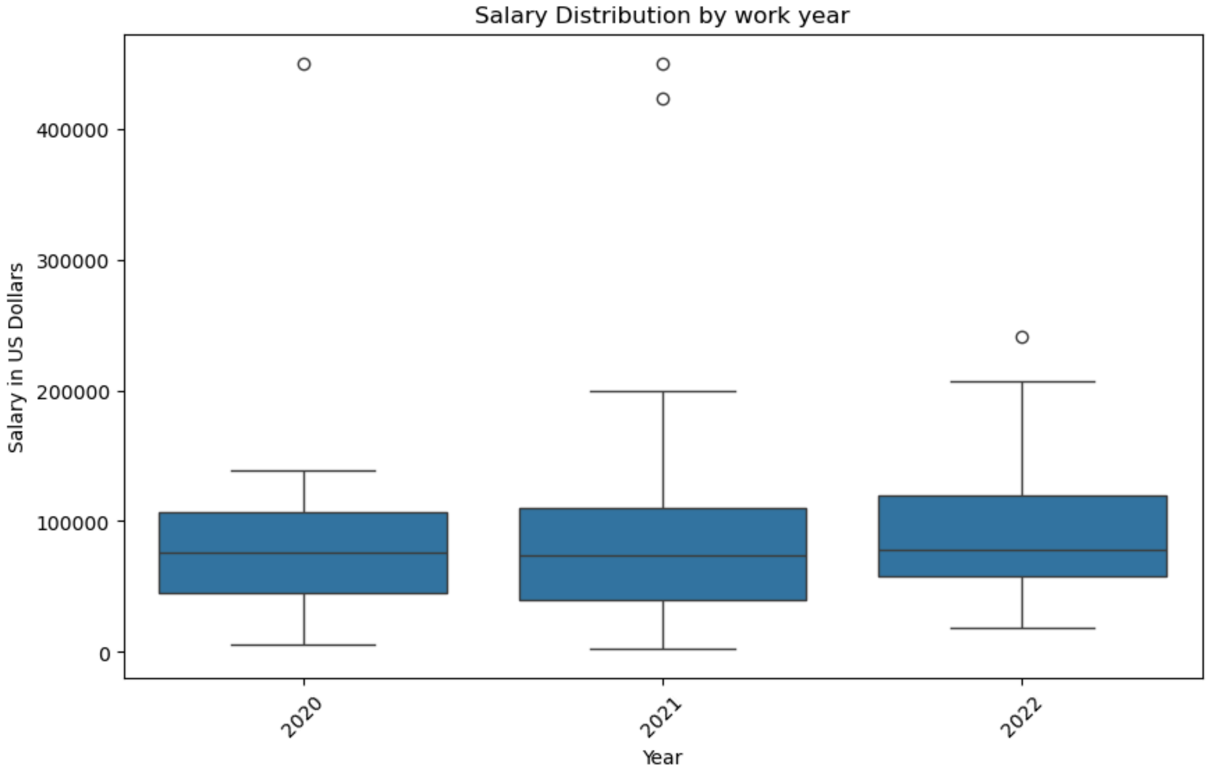
```
print("\nValue counts for categorical columns:")
for column in df3_se.select_dtypes(include='object').columns:
    print(f"\n--- {column} ---")
    print(df3_se[column].value_counts())
```

```
Value counts for categorical columns:

--- work_year ---
work_year
2022    193
2021     68
2020     17
Name: count, dtype: int64

--- experience_level ---
experience_level
SE    278
Name: count, dtype: int64

--- job_title ---
job_title
Data Engineer                               63
Data Scientist                              61
Data Analyst                                54
Machine Learning Engineer                   20
Data Science Manager                        10
Data Architect                               8
Data Analytics Manager                       7
Principal Data Scientist                     5
Research Scientist                           5
Lead Data Engineer                           4
Data Engineering Manager                     3
Machine Learning Scientist                   3
Big Data Engineer                            2
Lead Data Scientist                          2
Data Science Engineer                        2
Computer Vision Engineer                     2
Applied Data Scientist                       2
Head of Data                                 2
Principal Data Engineer                      2
Director of Data Engineering                 2
Data Analytics Engineer                      2
Analytics Engineer                           2
Director of Data Science                     1
Marketing Data Analyst                       1
Lead Data Analyst                            1
Machine Learning Manager                     1
Cloud Data Engineer                          1
Data Specialist                              1
Finance Data Analyst                         1
Big Data Architect                           1
Principal Data Analyst                       1
ML Engineer                                  1
Machine Learning Infrastructure Engineer     1
Lead Machine Learning Engineer               1
AI Scientist                                 1
Machine Learning Developer                   1
Data Analytics Lead                          1
Name: count, dtype: int64

--- company_size ---
```

```
company_size
M    185
L     72
S     21
Name: count, dtype: int64
```

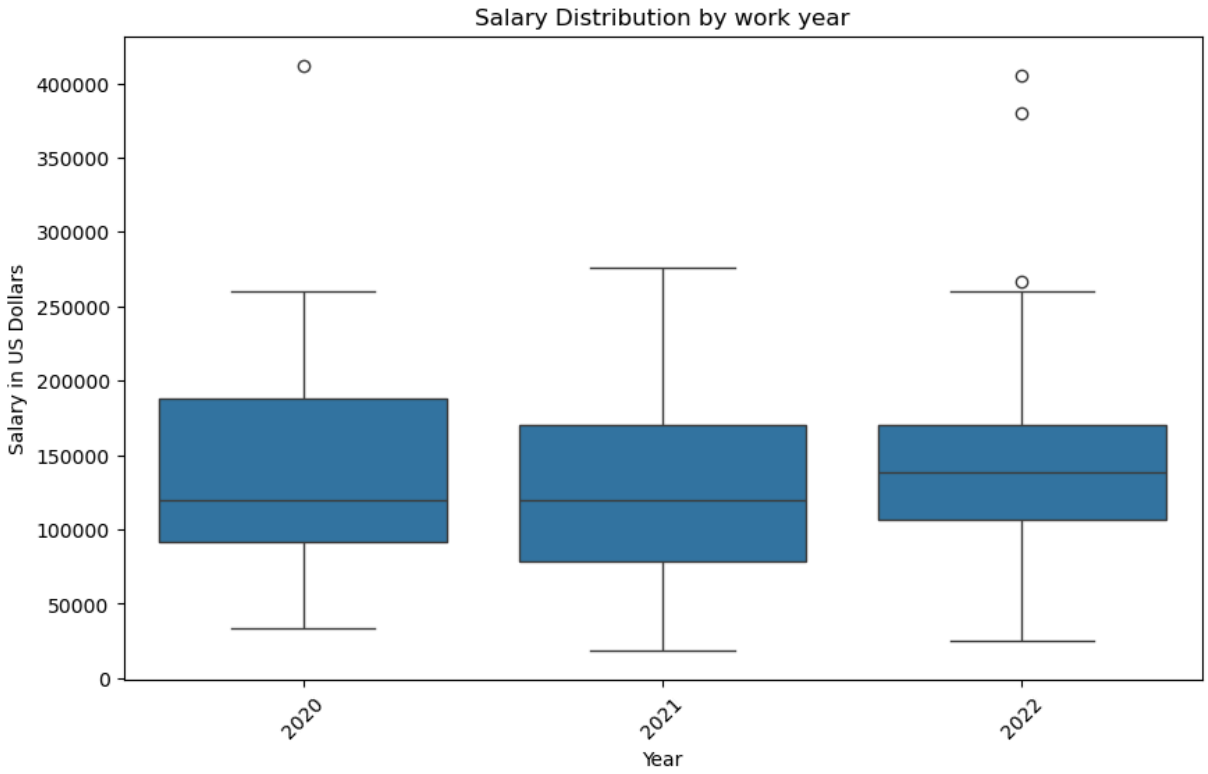
In [278…

```
plt.figure(figsize=(10, 6))
sns.boxplot(x='work_year', y='salary_in_usd', data=df3_mi)

plt.title('Salary Distribution by work year')
plt.xlabel('Year')
plt.ylabel('Salary in US Dollars')
plt.xticks(rotation=45)

plt.show()
```



In [279…

```
plt.figure(figsize=(10, 6))
sns.boxplot(x='work_year', y='salary_in_usd', data=df3_se)

plt.title('Salary Distribution by work year')
plt.xlabel('Year')
plt.ylabel('Salary in US Dollars')
plt.xticks(rotation=45)

plt.show()
```



In [280…

```
print("\nDescriptive statistics for numerical columns:")
print(df3_mi.describe())
```

```
Descriptive statistics for numerical columns:
       Unnamed: 0  salary_in_usd
count  206.000000     206.000000
mean   270.194175   88403.169903
std    171.177048   63002.949437
min      0.000000    2859.000000
25%    119.250000   49461.000000
50%    246.000000   77161.000000
75%    429.750000  112225.000000
max    606.000000  450000.000000
```

In [281…

```
print("\nDescriptive statistics for numerical columns:")
print(df3_se.describe())
```

```
Descriptive statistics for numerical columns:
       Unnamed: 0  salary_in_usd
count  278.000000     278.000000
mean   360.532374  139021.014388
std    166.095414   57670.092013
min      1.000000   18907.000000
25%    240.750000  100000.000000
50%    363.500000  136300.000000
75%    518.250000  170000.000000
max    605.000000  412000.000000
```

In [316…

```
selected_columns5 = ['Unnamed: 0', 'work_year', 'experience_level', 'job_tit
df5= df1ft[selected_columns5].copy()
print(df5)
```

```
     Unnamed: 0 work_year experience_level                   job_title  \
0             0      2020               MI              Data Scientist
1             1      2020               SE  Machine Learning Scientist
2             2      2020               SE           Big Data Engineer
3             3      2020               MI        Product Data Analyst
4             4      2020               SE   Machine Learning Engineer
..          ...       ...              ...                         ...
602         602      2022               SE               Data Engineer
603         603      2022               SE               Data Engineer
604         604      2022               SE                Data Analyst
605         605      2022               SE                Data Analyst
606         606      2022               MI                AI Scientist

     salary_in_usd employee_residence
0            79833                 DE
1           260000                 JP
2           109024                 GB
3            20000                 HN
4           150000                 US
..             ...                ...
602         154000                 US
603         126000                 US
604         129000                 US
605         150000                 US
606         200000                 IN

[588 rows x 6 columns]
```

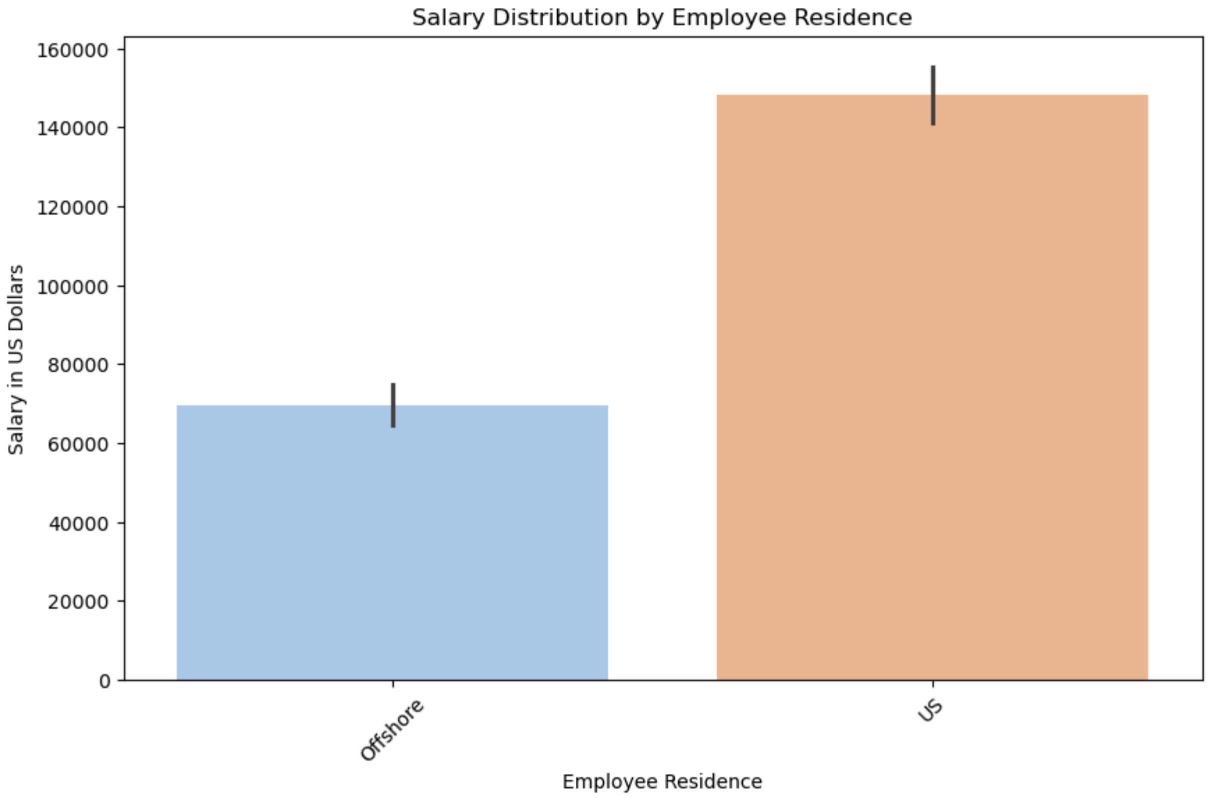
In [317…

```
df5.loc[df5['employee_residence'] != 'US', 'employee_residence'] = 'Offshore
print(df5)
```

```
     Unnamed: 0 work_year experience_level                  job_title  \
0             0      2020               MI             Data Scientist   
1             1      2020               SE  Machine Learning Scientist   
2             2      2020               SE          Big Data Engineer   
3             3      2020               MI       Product Data Analyst   
4             4      2020               SE   Machine Learning Engineer   
..          ...       ...              ...                        ...   
602         602      2022               SE              Data Engineer   
603         603      2022               SE              Data Engineer   
604         604      2022               SE               Data Analyst   
605         605      2022               SE               Data Analyst   
606         606      2022               MI               AI Scientist   

     salary_in_usd employee_residence  
0            79833           Offshore  
1           260000           Offshore  
2           109024           Offshore  
3            20000           Offshore  
4           150000                 US  
..             ...                ...  
602         154000                 US  
603         126000                 US  
604         129000                 US  
605         150000                 US  
606         200000           Offshore  

[588 rows x 6 columns]
```

In [319…

```
plt.figure(figsize=(10, 6))
sns.barplot(x='employee_residence', y='salary_in_usd', data=df5, palette='pa

plt.title('Salary Distribution by Employee Residence')
plt.xlabel('Employee Residence')
plt.ylabel('Salary in US Dollars')
plt.xticks(rotation=45)

plt.show()
```


Salary Distribution by Employee Residence
Salary in US Dollars
160000
140000
120000
100000
80000
60000
40000
20000
0
Offshore
US
Employee Residence

In [ ]: