Data Science Salary Prediction

May 2, 2024

1 Predicting Data Science Salaries Worldwide

Robinson Kao

[1]:
```
import random
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import OneHotEncoder
```

1.1 Import dataset CSV file as a dataframe.

[2]:
```
df = pd.read_csv(r"Data Science Jobs Salaries.csv")
```

[3]:
```
df
```

[3]:
```
    work_year experience_level employment_type                  job_title  \
0       2021e               EN              FT   Data Science Consultant
1        2020               SE              FT             Data Scientist
2       2021e               EX              FT     Head of Data Science
3       2021e               EX              FT               Head of Data
4       2021e               EN              FT Machine Learning Engineer
..        …                 …               …                          …
240      2020               SE              FT             Data Scientist
241     2021e               MI              FT   Principal Data Scientist
242      2020               EN              FT             Data Scientist
243      2020               EN              CT      Business Data Analyst
244     2021e               SE              FT       Data Science Manager

      salary salary_currency  salary_in_usd employee_residence  remote_ratio  \
0      54000             EUR          64369                 DE            50
1      60000             EUR          68428                 GR           100
2      85000             USD          85000                 RU             0
3     230000             USD         230000                 RU            50
4     125000             USD         125000                 US           100
```

```
..       …             …            …                …           …
240    412000             USD         412000                 US           100
241    151000             USD         151000                 US           100
242    105000             USD         105000                 US           100
243    100000             USD         100000                 US           100
244  7000000             INR          94917                 IN            50

    company_location company_size
0                  DE            L
1                  US            L
2                  RU            M
3                  RU            L
4                  US            S
..               …          …
240                US            L
241                US            L
242                US            S
243                US            L
244                IN            L

[245 rows x 11 columns]
```

2 Data Preprocessing

2.1 Drop unnecessary columns

[4]:

```
df.drop(columns=['salary_currency', 'salary'], inplace=True)
```

[5]:

```
df = df.dropna()
```

2.2 One-hot encode categorical variables

[6]:

```
df = pd.get_dummies(df)
```

[7]:

```
for col in df.columns:
    df.rename(columns={col : '_'.join(col.split())}, inplace=True)
df.columns
```

[7]:

```
Index(['salary_in_usd', 'remote_ratio', 'work_year_2020', 'work_year_2021e',
       'experience_level_EN', 'experience_level_EX', 'experience_level_MI',
       'experience_level_SE', 'employment_type_CT', 'employment_type_FL',
       …
       'company_location_RU', 'company_location_SG', 'company_location_SI',
       'company_location_TR', 'company_location_UA', 'company_location_US',
       'company_location_VN', 'company_size_L', 'company_size_M',
       'company_size_S'],
      dtype='object', length=144)
```

2.3 Create n-1 columns for categorical variables to prevent the dummy variable trap. Drop 1 of each.

[8]:
```
df.drop(columns=['work_year_2020', 'experience_level_EN', 'employment_type_CT',␣
 ↪'job_title_3D_Computer_Vision_Researcher',
'employee_residence_AE', 'company_location_AE', 'company_size_L'], inplace=True)
```

2.4 Process DataFrame into more consistent, clear categorical variables.

[9]:
```
df
```

[9]:
```
     salary_in_usd  remote_ratio  work_year_2021e  experience_level_EX  \
0            64369            50                1                    0
1            68428           100                0                    0
2            85000             0                1                    1
3           230000            50                1                    1
4           125000           100                1                    0
..             …           …              …                  …
240         412000           100                0                    0
241         151000           100                1                    0
242         105000           100                0                    0
243         100000           100                0                    0
244          94917            50                1                    0

     experience_level_MI  experience_level_SE  employment_type_FL  \
0                      0                    0                   0
1                      0                    1                   0
2                      0                    0                   0
3                      0                    0                   0
4                      0                    0                   0
..                   …                  …                 …
240                    0                    1                   0
241                    1                    0                   0
242                    0                    0                   0
243                    0                    0                   0
244                    0                    1                   0

     employment_type_FT  employment_type_PT  job_title_AI_Scientist  …  \
0                     1                   0                       0  …
1                     1                   0                       0  …
2                     1                   0                       0  …
3                     1                   0                       0  …
4                     1                   0                       0  …
..                  …                 …                    …  …
```

```
240                   1                   0                      0  …
241                   1                   0                      0  …
242                   1                   0                      0  …
243                   0                   0                      0  …
244                   1                   0                      0  …

     company_location_PT  company_location_RU  company_location_SG  \
0                      0                    0                    0
1                      0                    0                    0
2                      0                    1                    0
3                      0                    1                    0
4                      0                    0                    0
..                   …                  …                  …
240                    0                    0                    0
241                    0                    0                    0
242                    0                    0                    0
243                    0                    0                    0
244                    0                    0                    0

     company_location_SI  company_location_TR  company_location_UA  \
0                      0                    0                    0
1                      0                    0                    0
2                      0                    0                    0
3                      0                    0                    0
4                      0                    0                    0
..                   …                  …                  …
240                    0                    0                    0
241                    0                    0                    0
242                    0                    0                    0
243                    0                    0                    0
244                    0                    0                    0

     company_location_US  company_location_VN  company_size_M  company_size_S
0                      0                    0               0               0
1                      1                    0               0               0
2                      0                    0               1               0
3                      0                    0               0               0
4                      1                    0               0               1
..                   …                  …             …             …
240                    1                    0               0               0
241                    1                    0               0               0
242                    1                    0               0               1
243                    1                    0               0               0
244                    0                    0               0               0

[245 rows x 137 columns]
```

[10]:

```python
# Unique values of each column
print('COLUMN NULL COUNT')
column_types = {}
for col in df.columns:
    # print(i)
    # print(df[i].unique())
    column_types[col] = type(df[col][0])
    null_count = sum(df[col].isna())
    print(col + ':', null_count)

    # print('-----')

print('\nCOLUMN DATA TYPE')
for col, type in column_types.items():
    print(col + ':', type)
```

```
COLUMN NULL COUNT
salary_in_usd: 0
remote_ratio: 0
work_year_2021e: 0
experience_level_EX: 0
experience_level_MI: 0
experience_level_SE: 0
employment_type_FL: 0
employment_type_FT: 0
employment_type_PT: 0
job_title_AI_Scientist: 0
job_title_Applied_Data_Scientist: 0
job_title_Applied_Machine_Learning_Scientist: 0
job_title_BI_Data_Analyst: 0
job_title_Big_Data_Architect: 0
job_title_Big_Data_Engineer: 0
job_title_Business_Data_Analyst: 0
job_title_Cloud_Data_Engineer: 0
job_title_Computer_Vision_Engineer: 0
job_title_Computer_Vision_Software_Engineer: 0
job_title_Data_Analyst: 0
job_title_Data_Analytics_Engineer: 0
job_title_Data_Analytics_Manager: 0
job_title_Data_Architect: 0
job_title_Data_Engineer: 0
job_title_Data_Engineering_Manager: 0
job_title_Data_Science_Consultant: 0
job_title_Data_Science_Engineer: 0
job_title_Data_Science_Manager: 0
job_title_Data_Scientist: 0
job_title_Data_Specialist: 0
```

```
job_title_Director_of_Data_Engineering: 0
job_title_Director_of_Data_Science: 0
job_title_Finance_Data_Analyst: 0
job_title_Financial_Data_Analyst: 0
job_title_Head_of_Data: 0
job_title_Head_of_Data_Science: 0
job_title_Lead_Data_Analyst: 0
job_title_Lead_Data_Engineer: 0
job_title_Lead_Data_Scientist: 0
job_title_ML_Engineer: 0
job_title_Machine_Learning_Engineer: 0
job_title_Machine_Learning_Infrastructure_Engineer: 0
job_title_Machine_Learning_Scientist: 0
job_title_Manager_Data_Science: 0
job_title_Marketing_Data_Analyst: 0
job_title_Principal_Data_Analyst: 0
job_title_Principal_Data_Engineer: 0
job_title_Principal_Data_Scientist: 0
job_title_Product_Data_Analyst: 0
job_title_Research_Scientist: 0
job_title_Staff_Data_Scientist: 0
employee_residence_AT: 0
employee_residence_BE: 0
employee_residence_BG: 0
employee_residence_BR: 0
employee_residence_CA: 0
employee_residence_CL: 0
employee_residence_CN: 0
employee_residence_CO: 0
employee_residence_DE: 0
employee_residence_DK: 0
employee_residence_ES: 0
employee_residence_FR: 0
employee_residence_GB: 0
employee_residence_GR: 0
employee_residence_HK: 0
employee_residence_HR: 0
employee_residence_HU: 0
employee_residence_IN: 0
employee_residence_IR: 0
employee_residence_IT: 0
employee_residence_JE: 0
employee_residence_JP: 0
employee_residence_KE: 0
employee_residence_LU: 0
employee_residence_MD: 0
employee_residence_MT: 0
employee_residence_MX: 0
```

```
employee_residence_NG: 0
employee_residence_NL: 0
employee_residence_NZ: 0
employee_residence_PH: 0
employee_residence_PK: 0
employee_residence_PL: 0
employee_residence_PR: 0
employee_residence_PT: 0
employee_residence_RO: 0
employee_residence_RS: 0
employee_residence_RU: 0
employee_residence_SG: 0
employee_residence_SI: 0
employee_residence_TR: 0
employee_residence_UA: 0
employee_residence_US: 0
employee_residence_VN: 0
company_location_AS: 0
company_location_AT: 0
company_location_BE: 0
company_location_BR: 0
company_location_CA: 0
company_location_CH: 0
company_location_CL: 0
company_location_CN: 0
company_location_CO: 0
company_location_DE: 0
company_location_DK: 0
company_location_ES: 0
company_location_FR: 0
company_location_GB: 0
company_location_GR: 0
company_location_HR: 0
company_location_HU: 0
company_location_IL: 0
company_location_IN: 0
company_location_IR: 0
company_location_IT: 0
company_location_JP: 0
company_location_KE: 0
company_location_LU: 0
company_location_MD: 0
company_location_MT: 0
company_location_MX: 0
company_location_NG: 0
company_location_NL: 0
company_location_NZ: 0
company_location_PK: 0
```

```
company_location_PL: 0
company_location_PT: 0
company_location_RU: 0
company_location_SG: 0
company_location_SI: 0
company_location_TR: 0
company_location_UA: 0
company_location_US: 0
company_location_VN: 0
company_size_M: 0
company_size_S: 0

COLUMN DATA TYPE
salary_in_usd: <class 'numpy.int64'>
remote_ratio: <class 'numpy.int64'>
work_year_2021e: <class 'numpy.uint8'>
experience_level_EX: <class 'numpy.uint8'>
experience_level_MI: <class 'numpy.uint8'>
experience_level_SE: <class 'numpy.uint8'>
employment_type_FL: <class 'numpy.uint8'>
employment_type_FT: <class 'numpy.uint8'>
employment_type_PT: <class 'numpy.uint8'>
job_title_AI_Scientist: <class 'numpy.uint8'>
job_title_Applied_Data_Scientist: <class 'numpy.uint8'>
job_title_Applied_Machine_Learning_Scientist: <class 'numpy.uint8'>
job_title_BI_Data_Analyst: <class 'numpy.uint8'>
job_title_Big_Data_Architect: <class 'numpy.uint8'>
job_title_Big_Data_Engineer: <class 'numpy.uint8'>
job_title_Business_Data_Analyst: <class 'numpy.uint8'>
job_title_Cloud_Data_Engineer: <class 'numpy.uint8'>
job_title_Computer_Vision_Engineer: <class 'numpy.uint8'>
job_title_Computer_Vision_Software_Engineer: <class 'numpy.uint8'>
job_title_Data_Analyst: <class 'numpy.uint8'>
job_title_Data_Analytics_Engineer: <class 'numpy.uint8'>
job_title_Data_Analytics_Manager: <class 'numpy.uint8'>
job_title_Data_Architect: <class 'numpy.uint8'>
job_title_Data_Engineer: <class 'numpy.uint8'>
job_title_Data_Engineering_Manager: <class 'numpy.uint8'>
job_title_Data_Science_Consultant: <class 'numpy.uint8'>
job_title_Data_Science_Engineer: <class 'numpy.uint8'>
job_title_Data_Science_Manager: <class 'numpy.uint8'>
job_title_Data_Scientist: <class 'numpy.uint8'>
job_title_Data_Specialist: <class 'numpy.uint8'>
job_title_Director_of_Data_Engineering: <class 'numpy.uint8'>
job_title_Director_of_Data_Science: <class 'numpy.uint8'>
job_title_Finance_Data_Analyst: <class 'numpy.uint8'>
job_title_Financial_Data_Analyst: <class 'numpy.uint8'>
job_title_Head_of_Data: <class 'numpy.uint8'>
```

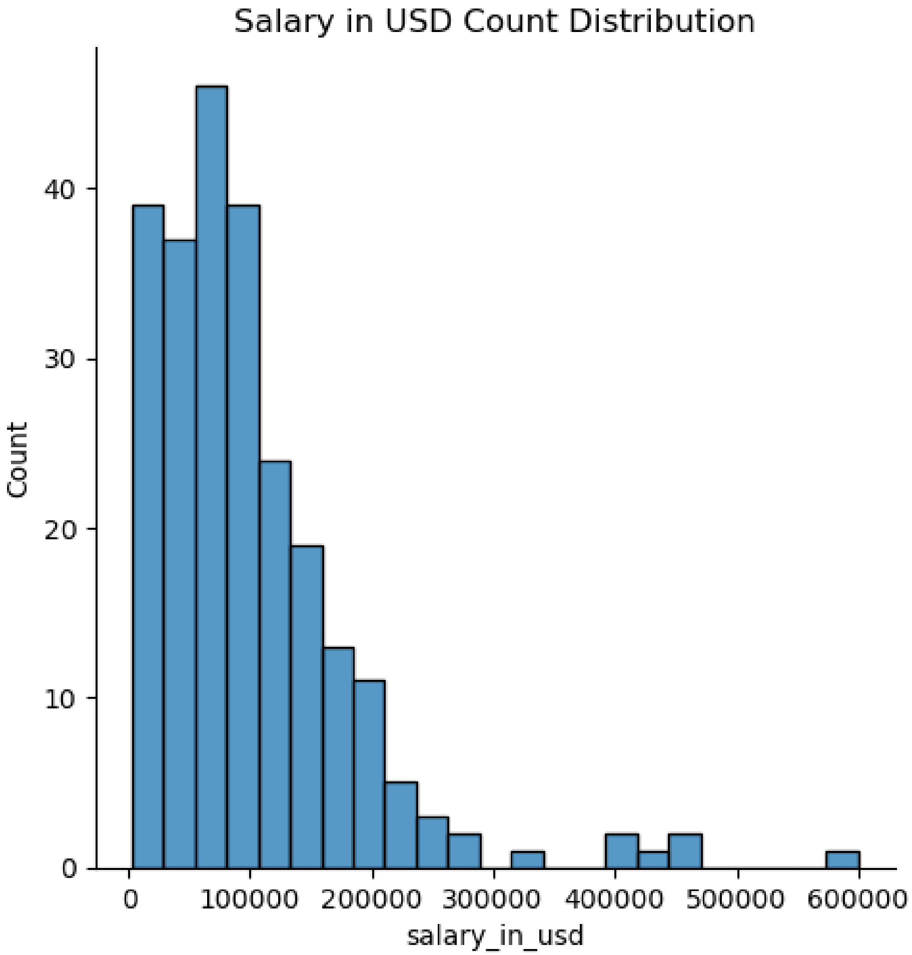
```
job_title_Head_of_Data_Science: <class 'numpy.uint8'>
job_title_Lead_Data_Analyst: <class 'numpy.uint8'>
job_title_Lead_Data_Engineer: <class 'numpy.uint8'>
job_title_Lead_Data_Scientist: <class 'numpy.uint8'>
job_title_ML_Engineer: <class 'numpy.uint8'>
job_title_Machine_Learning_Engineer: <class 'numpy.uint8'>
job_title_Machine_Learning_Infrastructure_Engineer: <class 'numpy.uint8'>
job_title_Machine_Learning_Scientist: <class 'numpy.uint8'>
job_title_Manager_Data_Science: <class 'numpy.uint8'>
job_title_Marketing_Data_Analyst: <class 'numpy.uint8'>
job_title_Principal_Data_Analyst: <class 'numpy.uint8'>
job_title_Principal_Data_Engineer: <class 'numpy.uint8'>
job_title_Principal_Data_Scientist: <class 'numpy.uint8'>
job_title_Product_Data_Analyst: <class 'numpy.uint8'>
job_title_Research_Scientist: <class 'numpy.uint8'>
job_title_Staff_Data_Scientist: <class 'numpy.uint8'>
employee_residence_AT: <class 'numpy.uint8'>
employee_residence_BE: <class 'numpy.uint8'>
employee_residence_BG: <class 'numpy.uint8'>
employee_residence_BR: <class 'numpy.uint8'>
employee_residence_CA: <class 'numpy.uint8'>
employee_residence_CL: <class 'numpy.uint8'>
employee_residence_CN: <class 'numpy.uint8'>
employee_residence_CO: <class 'numpy.uint8'>
employee_residence_DE: <class 'numpy.uint8'>
employee_residence_DK: <class 'numpy.uint8'>
employee_residence_ES: <class 'numpy.uint8'>
employee_residence_FR: <class 'numpy.uint8'>
employee_residence_GB: <class 'numpy.uint8'>
employee_residence_GR: <class 'numpy.uint8'>
employee_residence_HK: <class 'numpy.uint8'>
employee_residence_HR: <class 'numpy.uint8'>
employee_residence_HU: <class 'numpy.uint8'>
employee_residence_IN: <class 'numpy.uint8'>
employee_residence_IR: <class 'numpy.uint8'>
employee_residence_IT: <class 'numpy.uint8'>
employee_residence_JE: <class 'numpy.uint8'>
employee_residence_JP: <class 'numpy.uint8'>
employee_residence_KE: <class 'numpy.uint8'>
employee_residence_LU: <class 'numpy.uint8'>
employee_residence_MD: <class 'numpy.uint8'>
employee_residence_MT: <class 'numpy.uint8'>
employee_residence_MX: <class 'numpy.uint8'>
employee_residence_NG: <class 'numpy.uint8'>
employee_residence_NL: <class 'numpy.uint8'>
employee_residence_NZ: <class 'numpy.uint8'>
employee_residence_PH: <class 'numpy.uint8'>
employee_residence_PK: <class 'numpy.uint8'>
```

```
employee_residence_PL: <class 'numpy.uint8'>
employee_residence_PR: <class 'numpy.uint8'>
employee_residence_PT: <class 'numpy.uint8'>
employee_residence_RO: <class 'numpy.uint8'>
employee_residence_RS: <class 'numpy.uint8'>
employee_residence_RU: <class 'numpy.uint8'>
employee_residence_SG: <class 'numpy.uint8'>
employee_residence_SI: <class 'numpy.uint8'>
employee_residence_TR: <class 'numpy.uint8'>
employee_residence_UA: <class 'numpy.uint8'>
employee_residence_US: <class 'numpy.uint8'>
employee_residence_VN: <class 'numpy.uint8'>
company_location_AS: <class 'numpy.uint8'>
company_location_AT: <class 'numpy.uint8'>
company_location_BE: <class 'numpy.uint8'>
company_location_BR: <class 'numpy.uint8'>
company_location_CA: <class 'numpy.uint8'>
company_location_CH: <class 'numpy.uint8'>
company_location_CL: <class 'numpy.uint8'>
company_location_CN: <class 'numpy.uint8'>
company_location_CO: <class 'numpy.uint8'>
company_location_DE: <class 'numpy.uint8'>
company_location_DK: <class 'numpy.uint8'>
company_location_ES: <class 'numpy.uint8'>
company_location_FR: <class 'numpy.uint8'>
company_location_GB: <class 'numpy.uint8'>
company_location_GR: <class 'numpy.uint8'>
company_location_HR: <class 'numpy.uint8'>
company_location_HU: <class 'numpy.uint8'>
company_location_IL: <class 'numpy.uint8'>
company_location_IN: <class 'numpy.uint8'>
company_location_IR: <class 'numpy.uint8'>
company_location_IT: <class 'numpy.uint8'>
company_location_JP: <class 'numpy.uint8'>
company_location_KE: <class 'numpy.uint8'>
company_location_LU: <class 'numpy.uint8'>
company_location_MD: <class 'numpy.uint8'>
company_location_MT: <class 'numpy.uint8'>
company_location_MX: <class 'numpy.uint8'>
company_location_NG: <class 'numpy.uint8'>
company_location_NL: <class 'numpy.uint8'>
company_location_NZ: <class 'numpy.uint8'>
company_location_PK: <class 'numpy.uint8'>
company_location_PL: <class 'numpy.uint8'>
company_location_PT: <class 'numpy.uint8'>
company_location_RU: <class 'numpy.uint8'>
company_location_SG: <class 'numpy.uint8'>
company_location_SI: <class 'numpy.uint8'>
```

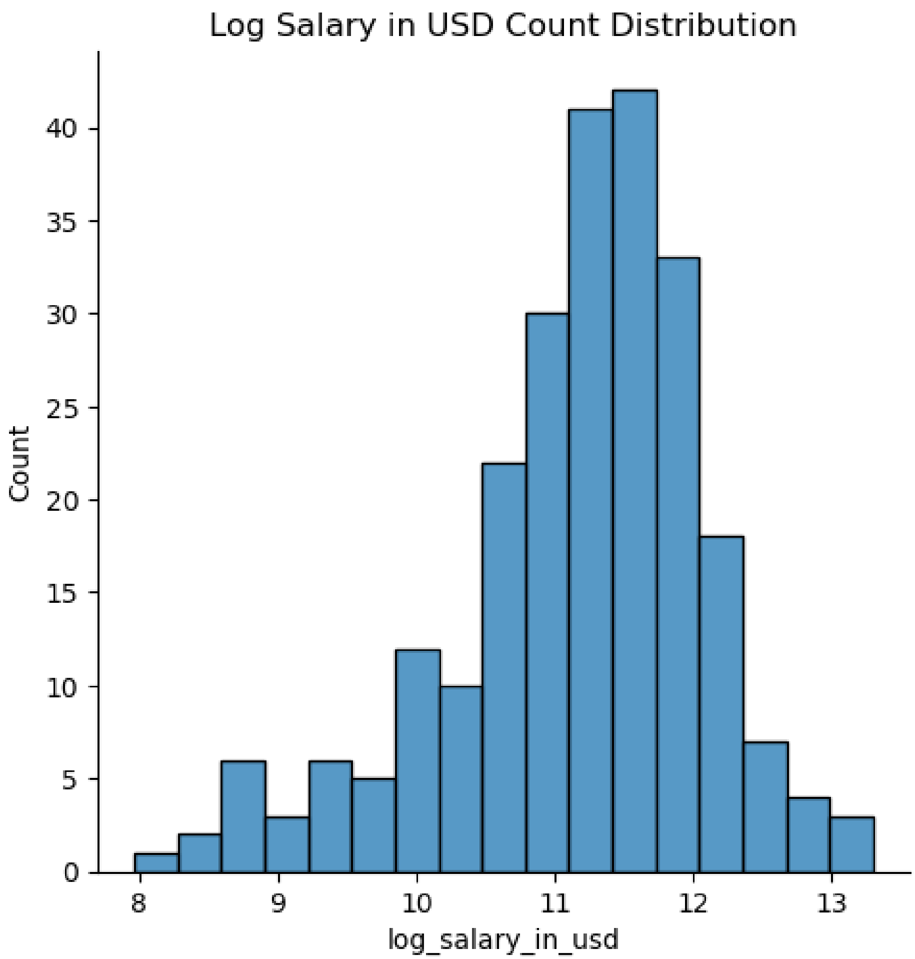
```
company_location_TR: <class 'numpy.uint8'>
company_location_UA: <class 'numpy.uint8'>
company_location_US: <class 'numpy.uint8'>
company_location_VN: <class 'numpy.uint8'>
company_size_M: <class 'numpy.uint8'>
company_size_S: <class 'numpy.uint8'>
```

2.5 Split dataset into train and test sets.

[10]:
```
sns.displot(df, x='salary_in_usd')
plt.title('Salary in USD Count Distribution')
plt.show();
```

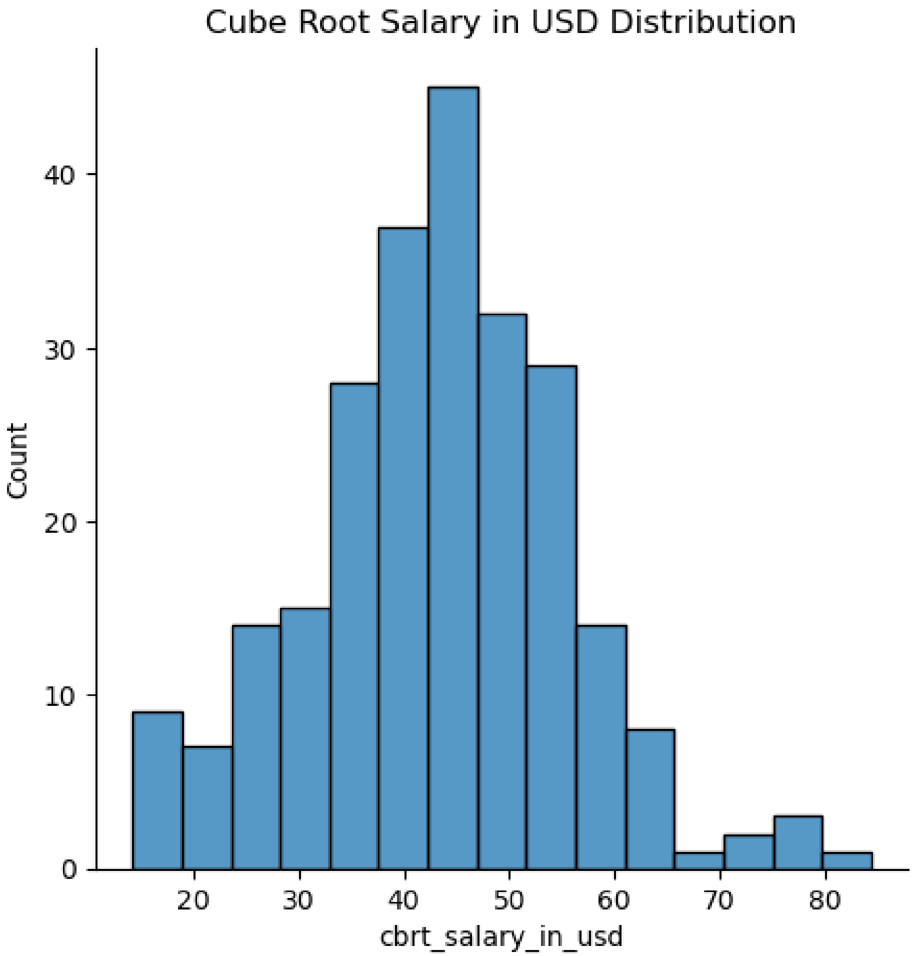


[11]:
```
df['log_salary_in_usd'] = np.log(df['salary_in_usd'].values)
sns.displot(df, x='log_salary_in_usd')
plt.title('Log Salary in USD Count Distribution');
```



[12]:

```
df['cbrt_salary_in_usd'] = np.array(df['salary_in_usd'].values)**(1/3)  # Cube␣
  ↪Root
sns.displot(data=df, x='cbrt_salary_in_usd')
plt.title('Cube Root Salary in USD Distribution')
plt.show();
```



[14]:
```
# train test split
df_train, df_test = train_test_split(df, test_size=0.3, random_state=88)
df_train_y_actual = df_train['cbrt_salary_in_usd']
df_test_y_actual = df_test['cbrt_salary_in_usd']
df_train.shape, df_test.shape
```

[14]:
```
((171, 138), (74, 138))
```

[15]:
```
# target_column = 'log_salary_in_usd'
target_column = 'cbrt_salary_in_usd'
feature_columns = [col for col in df_train.columns if col != target_column]
all_features = ' + '.join(feature_columns)
all_features
linreg = smf.ols(formula = target_column + ' ~ ' + all_features,
                       data = df_train).fit()
print(linreg.summary())
linreg.summary()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     cbrt_salary_in_usd   R-squared:                       0.981
Model:                            OLS   Adj. R-squared:                  0.954
Method:                 Least Squares   F-statistic:                     36.26
Date:                Mon, 08 May 2023   Prob (F-statistic):           1.09e-37
Time:                        01:04:21   Log-Likelihood:                -341.37
No. Observations:                 171   AIC:                             884.7
Df Residuals:                      70   BIC:                             1202.
Df Model:                         100
Covariance Type:            nonrobust
================================================================================
======================================
                                                         coef    std err
t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
--------------------------------------
Intercept                                             21.1044      4.810
4.388      0.000      11.512      30.697
salary_in_usd                                       9.947e-05    6.4e-06
15.532      0.000    8.67e-05       0.000
remote_ratio                                          -0.0072      0.010
-0.746      0.458      -0.026       0.012
work_year_2021e                                       -0.0626      0.707
-0.089      0.930      -1.473       1.348
experience_level_EX                                    2.7895      2.897
0.963      0.339      -2.989       8.568
experience_level_MI                                    0.8204      0.860
0.954      0.343      -0.894       2.535
experience_level_SE                                    2.0696      1.071
1.932      0.057      -0.067       4.206
employment_type_FL                                    -7.3516      6.302
-1.167      0.247     -19.920       5.217
employment_type_FT                                     6.9583      4.491
1.549      0.126      -1.999      15.916
employment_type_PT                                     7.5817      5.592
1.356      0.179      -3.571      18.734
job_title_AI_Scientist                                 1.6816      2.945
0.571      0.570      -4.193       7.556
job_title_Applied_Data_Scientist                       7.3678      5.312
1.387      0.170      -3.227      17.963
job_title_Applied_Machine_Learning_Scientist          -2.9716      3.282
-0.906      0.368      -9.517       3.574
job_title_BI_Data_Analyst                              1.5084      3.069
0.491      0.625      -4.613       7.630
job_title_Big_Data_Architect                          -0.0997      3.425
-0.029      0.977      -6.931       6.732
job_title_Big_Data_Engineer                           -1.1441      2.279
```

```
-0.502      0.617      -5.690       3.402
job_title_Business_Data_Analyst                          -2.017e-11   2.29e-11
-0.882      0.381   -6.58e-11    2.54e-11
job_title_Cloud_Data_Engineer                            -1.663e-12   4.04e-12
-0.412      0.682   -9.72e-12    6.39e-12
job_title_Computer_Vision_Engineer                           0.1228      3.424
0.036      0.971      -6.706       6.951
job_title_Computer_Vision_Software_Engineer                  1.0060      2.969
0.339      0.736      -4.916       6.928
job_title_Data_Analyst                                      -0.4791      1.190
-0.403      0.688      -2.852       1.894
job_title_Data_Analytics_Engineer                            1.3717      2.883
0.476      0.636      -4.378       7.121
job_title_Data_Analytics_Manager                             2.0740      2.921
0.710      0.480      -3.751       7.900
job_title_Data_Architect                                     2.9565      2.843
1.040      0.302      -2.715       8.628
job_title_Data_Engineer                                      1.8334      0.868
2.111      0.038       0.102       3.565
job_title_Data_Engineering_Manager                           1.7737      2.353
0.754      0.453      -2.919       6.467
job_title_Data_Science_Consultant                           -0.2006      1.586
-0.126      0.900      -3.363       2.962
job_title_Data_Science_Engineer                              1.1235      2.146
0.524      0.602      -3.156       5.403
job_title_Data_Science_Manager                               5.5431      1.558
3.559      0.001       2.437       8.650
job_title_Data_Scientist                                     0.4854      0.844
0.575      0.567      -1.199       2.169
job_title_Data_Specialist                                    1.5853      2.816
0.563      0.575      -4.032       7.203
job_title_Director_of_Data_Engineering                       2.3658      2.102
1.125      0.264      -1.827       6.558
job_title_Director_of_Data_Science                           1.5090      2.599
0.581      0.563      -3.674       6.692
job_title_Finance_Data_Analyst                              -2.6147      2.993
-0.874      0.385      -8.585       3.355
job_title_Financial_Data_Analyst                            -3.7327      3.383
-1.103      0.274     -10.479       3.014
job_title_Head_of_Data                                      -4.0535      3.443
-1.177      0.243     -10.920       2.813
job_title_Head_of_Data_Science                            1.571e-12   2.61e-12
0.602      0.549   -3.63e-12    6.77e-12
job_title_Lead_Data_Analyst                                 -0.5254      2.201
-0.239      0.812      -4.916       3.865
job_title_Lead_Data_Engineer                                 0.9303      2.127
0.437      0.663      -3.312       5.173
job_title_Lead_Data_Scientist                                2.4796      2.972
```

```
0.834      0.407      -3.449       8.408
job_title_ML_Engineer                                  1.9840      2.958
0.671      0.505      -3.916       7.885
job_title_Machine_Learning_Engineer                    1.5101      1.023
1.477      0.144      -0.529       3.550
job_title_Machine_Learning_Infrastructure_Engineer     3.2782      2.880
1.138      0.259      -2.465       9.021
job_title_Machine_Learning_Scientist                   3.5321      3.386
1.043      0.301      -3.222      10.286
job_title_Manager_Data_Science                         1.2411      2.830
0.439      0.662      -4.404       6.886
job_title_Marketing_Data_Analyst                       2.8309      2.806
1.009      0.317      -2.766       8.428
job_title_Principal_Data_Analyst                       3.1727      2.885
1.100      0.275      -2.580       8.926
job_title_Principal_Data_Engineer                    -12.9105      4.506
-2.865      0.006     -21.898      -3.923
job_title_Principal_Data_Scientist                     3.4608      1.562
2.216      0.030       0.346       6.576
job_title_Product_Data_Analyst                        -8.5013      3.114
-2.730      0.008     -14.712      -2.290
job_title_Research_Scientist                          -0.3904      1.632
-0.239      0.812      -3.644       2.864
job_title_Staff_Data_Scientist                      5.487e-12   6.22e-12
0.882      0.381   -6.91e-12    1.79e-11
employee_residence_AT                                  5.4057      5.513
0.981      0.330      -5.590      16.401
employee_residence_BE                                  3.8739      1.555
2.491      0.015       0.773       6.975
employee_residence_BG                                  4.3732      3.261
1.341      0.184      -2.130      10.877
employee_residence_BR                                 -9.7715      5.431
-1.799      0.076     -20.602       1.059
employee_residence_CA                                  7.6433      4.696
1.628      0.108      -1.722      17.008
employee_residence_CL                              -1.066e-12   9.75e-13
-1.093      0.278   -3.01e-12    8.79e-13
employee_residence_CN                                 -1.6487      2.402
-0.686      0.495      -6.440       3.143
employee_residence_CO                                 -0.9162      1.516
-0.604      0.548      -3.940       2.108
employee_residence_DE                                  5.2784      3.439
1.535      0.129      -1.580      12.137
employee_residence_DK                                 -0.6969      2.652
-0.263      0.794      -5.986       4.593
employee_residence_ES                                  4.1136      2.727
1.508      0.136      -1.326       9.553
employee_residence_FR                                  4.0950      2.514
```

```
1.629      0.108      -0.918       9.108
employee_residence_GB                                        -4.1012      2.972
-1.380      0.172     -10.029       1.826
employee_residence_GR                                         1.5076      3.174
0.475      0.636      -4.822       7.837
employee_residence_HK                                      3.739e-12   3.78e-12
0.990      0.326   -3.79e-12    1.13e-11
employee_residence_HR                                         0.5372      1.522
0.353      0.725      -2.498       3.572
employee_residence_HU                                        -3.0715      4.195
-0.732      0.467     -11.439       5.296
employee_residence_IN                                        -9.3334      2.367
-3.943      0.000     -14.054      -4.613
employee_residence_IR                                      1.732e-12   1.67e-12
1.035      0.304    -1.6e-12    5.07e-12
employee_residence_IT                                         6.1201      3.425
1.787      0.078      -0.710      12.950
employee_residence_JE                                         5.2538      2.466
2.131      0.037       0.336      10.172
employee_residence_JP                                         2.5501      1.000
2.550      0.013       0.555       4.545
employee_residence_KE                                        -3.9494      2.099
-1.882      0.064      -8.136       0.237
employee_residence_LU                                     -1.202e-12   1.31e-12
-0.921      0.360   -3.81e-12     1.4e-12
employee_residence_MD                                        -1.2283      1.815
-0.677      0.501      -4.848       2.392
employee_residence_MT                                        -1.2776      1.433
-0.892      0.376      -4.135       1.580
employee_residence_MX                                        -4.2061      1.150
-3.659      0.000      -6.499      -1.913
employee_residence_NG                                        -0.5572      1.041
-0.535      0.594      -2.634       1.520
employee_residence_NL                                         7.1599      4.622
1.549      0.126      -2.058      16.378
employee_residence_NZ                                     -1.386e-12   1.48e-12
-0.936      0.352   -4.34e-12    1.57e-12
employee_residence_PH                                         0.6806      3.022
0.225      0.822      -5.347       6.708
employee_residence_PK                                        -9.0381      5.540
-1.632      0.107     -20.087       2.010
employee_residence_PL                                        -0.7609      2.618
-0.291      0.772      -5.983       4.461
employee_residence_PR                                         5.8451      3.493
1.673      0.099      -1.122      12.813
employee_residence_PT                                         2.4410      3.807
0.641      0.524      -5.152      10.034
employee_residence_RO                                        -6.2619      3.747
```

```
-1.671      0.099     -13.736       1.212
employee_residence_RS                                 -1.1814      4.679
-0.253      0.801     -10.513       8.150
employee_residence_RU                                 16.0680      5.896
2.725      0.008       4.309      27.827
employee_residence_SG                                  5.1570      1.522
3.389      0.001       2.122       8.192
employee_residence_SI                                 -1.6038      1.472
-1.090      0.280      -4.539       1.331
employee_residence_TR                                 -1.7678      1.120
-1.578      0.119      -4.003       0.467
employee_residence_UA                                 -2.6728      1.488
-1.796      0.077      -5.640       0.295
employee_residence_US                                  4.8100      1.480
3.251      0.002       1.859       7.761
employee_residence_VN                                 -7.7642      4.263
-1.821      0.073     -16.266       0.737
company_location_AS                                    5.6315      4.870
1.156      0.251      -4.081      15.344
company_location_AT                                    0.7221      4.493
0.161      0.873      -8.239       9.684
company_location_BE                                    3.8739      1.555
2.491      0.015       0.773       6.975
company_location_BR                                    4.7085      6.168
0.763      0.448      -7.593      17.010
company_location_CA                                    0.7494      4.269
0.176      0.861      -7.765       9.264
company_location_CH                                   -0.0416      4.015
-0.010      0.992      -8.049       7.966
company_location_CL                                 5.331e-18   2.13e-17
0.251      0.803   -3.71e-17    4.77e-17
company_location_CN                                    3.6050      1.456
2.476      0.016       0.701       6.509
company_location_CO                                   -0.9162      1.516
-0.604      0.548      -3.940       2.108
company_location_DE                                    0.3665      3.602
0.102      0.919      -6.818       7.551
company_location_DK                                    2.1340      1.939
1.100      0.275      -1.734       6.002
company_location_ES                                   -1.3657      3.140
-0.435      0.665      -7.629       4.897
company_location_FR                                   -0.8660      2.490
-0.348      0.729      -5.833       4.101
company_location_GB                                   10.4463      2.778
3.760      0.000       4.905      15.988
company_location_GR                                    1.1235      2.146
0.524      0.602      -3.156       5.403
company_location_HR                                    0.5372      1.522
```

```
0.353      0.725     -2.498       3.572
company_location_HU                                         0           0
nan        nan          0           0
company_location_IL                                    5.1570      1.522
3.389      0.001      2.122       8.192
company_location_IN                                    7.3119      2.343
3.120      0.003      2.638      11.985
company_location_IR                                         0           0
nan        nan          0           0
company_location_IT                                         0           0
nan        nan          0           0
company_location_JP                                    2.5501      1.000
2.550      0.013      0.555       4.545
company_location_KE                                   -3.9494      2.099
-1.882      0.064      -8.136       0.237
company_location_LU                                         0           0
nan        nan          0           0
company_location_MD                                   -1.2283      1.815
-0.677      0.501      -4.848       2.392
company_location_MT                                   -1.2776      1.433
-0.892      0.376      -4.135       1.580
company_location_MX                                   -4.2061      1.150
-3.659      0.000      -6.499      -1.913
company_location_NG                                   -0.5572      1.041
-0.535      0.594      -2.634       1.520
company_location_NL                                   -4.6550      5.471
-0.851      0.398     -15.567       6.257
company_location_NZ                                         0           0
nan        nan          0           0
company_location_PK                                    0.1987      6.375
0.031      0.975     -12.515      12.913
company_location_PL                                    0.5999      2.935
0.204      0.839     -5.255       6.455
company_location_PT                                    1.5683      5.002
0.314      0.755     -8.408      11.544
company_location_RU                                   -4.0535      3.443
-1.177      0.243     -10.920       2.813
company_location_SG                                         0           0
nan        nan          0           0
company_location_SI                                   -1.6038      1.472
-1.090      0.280      -4.539       1.331
company_location_TR                                   -1.7678      1.120
-1.578      0.119      -4.003       0.467
company_location_UA                                   -2.6728      1.488
-1.796      0.077      -5.640       0.295
company_location_US                                    2.6906      1.394
1.931      0.058     -0.089       5.470
company_location_VN                                   -3.7090      4.790
```

```
-0.774      0.441     -13.263       5.845
company_size_M                                               -1.5362      0.860
-1.785      0.079      -3.252       0.180
company_size_S                                               -0.8028      0.962
-0.835      0.407      -2.721       1.116
==============================================================================
Omnibus:                       24.077   Durbin-Watson:                   2.067
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               67.549
Skew:                          -0.519   Prob(JB):                     2.15e-15
Kurtosis:                       5.899   Cond. No.                     5.87e+20
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The smallest eigenvalue is 9.49e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

```
[15]: <class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:     cbrt_salary_in_usd   R-squared:                       0.981
Model:                            OLS   Adj. R-squared:                  0.954
Method:                 Least Squares   F-statistic:                     36.26
Date:                Mon, 08 May 2023   Prob (F-statistic):           1.09e-37
Time:                        01:04:21   Log-Likelihood:                -341.37
No. Observations:                 171   AIC:                             884.7
Df Residuals:                      70   BIC:                             1202.
Df Model:                         100
Covariance Type:            nonrobust
================================================================================
======================================
                                                          coef    std err
t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
--------------------------------------
Intercept                                              21.1044      4.810
4.388      0.000      11.512      30.697
salary_in_usd                                        9.947e-05    6.4e-06
15.532       0.000    8.67e-05       0.000
remote_ratio                                           -0.0072      0.010
-0.746       0.458      -0.026       0.012
work_year_2021e                                        -0.0626      0.707
-0.089       0.930      -1.473       1.348
experience_level_EX                                     2.7895      2.897
0.963      0.339      -2.989       8.568
```

```
experience_level_MI                                    0.8204      0.860
0.954      0.343      -0.894       2.535
experience_level_SE                                    2.0696      1.071
1.932      0.057      -0.067       4.206
employment_type_FL                                    -7.3516      6.302
-1.167      0.247     -19.920       5.217
employment_type_FT                                     6.9583      4.491
1.549      0.126      -1.999      15.916
employment_type_PT                                     7.5817      5.592
1.356      0.179      -3.571      18.734
job_title_AI_Scientist                                 1.6816      2.945
0.571      0.570      -4.193       7.556
job_title_Applied_Data_Scientist                       7.3678      5.312
1.387      0.170      -3.227      17.963
job_title_Applied_Machine_Learning_Scientist          -2.9716      3.282
-0.906      0.368      -9.517       3.574
job_title_BI_Data_Analyst                              1.5084      3.069
0.491      0.625      -4.613       7.630
job_title_Big_Data_Architect                          -0.0997      3.425
-0.029      0.977      -6.931       6.732
job_title_Big_Data_Engineer                           -1.1441      2.279
-0.502      0.617      -5.690       3.402
job_title_Business_Data_Analyst                    -2.017e-11   2.29e-11
-0.882      0.381   -6.58e-11    2.54e-11
job_title_Cloud_Data_Engineer                      -1.663e-12   4.04e-12
-0.412      0.682   -9.72e-12    6.39e-12
job_title_Computer_Vision_Engineer                     0.1228      3.424
0.036      0.971      -6.706       6.951
job_title_Computer_Vision_Software_Engineer            1.0060      2.969
0.339      0.736      -4.916       6.928
job_title_Data_Analyst                                -0.4791      1.190
-0.403      0.688      -2.852       1.894
job_title_Data_Analytics_Engineer                      1.3717      2.883
0.476      0.636      -4.378       7.121
job_title_Data_Analytics_Manager                       2.0740      2.921
0.710      0.480      -3.751       7.900
job_title_Data_Architect                               2.9565      2.843
1.040      0.302      -2.715       8.628
job_title_Data_Engineer                                1.8334      0.868
2.111      0.038       0.102       3.565
job_title_Data_Engineering_Manager                     1.7737      2.353
0.754      0.453      -2.919       6.467
job_title_Data_Science_Consultant                     -0.2006      1.586
-0.126      0.900      -3.363       2.962
job_title_Data_Science_Engineer                        1.1235      2.146
0.524      0.602      -3.156       5.403
job_title_Data_Science_Manager                         5.5431      1.558
```

```
3.559      0.001       2.437       8.650
job_title_Data_Scientist                               0.4854      0.844
0.575      0.567      -1.199       2.169
job_title_Data_Specialist                              1.5853      2.816
0.563      0.575      -4.032       7.203
job_title_Director_of_Data_Engineering                 2.3658      2.102
1.125      0.264      -1.827       6.558
job_title_Director_of_Data_Science                     1.5090      2.599
0.581      0.563      -3.674       6.692
job_title_Finance_Data_Analyst                        -2.6147      2.993
-0.874      0.385      -8.585       3.355
job_title_Financial_Data_Analyst                      -3.7327      3.383
-1.103      0.274     -10.479       3.014
job_title_Head_of_Data                                -4.0535      3.443
-1.177      0.243     -10.920       2.813
job_title_Head_of_Data_Science                      1.571e-12   2.61e-12
0.602      0.549   -3.63e-12    6.77e-12
job_title_Lead_Data_Analyst                           -0.5254      2.201
-0.239      0.812      -4.916       3.865
job_title_Lead_Data_Engineer                           0.9303      2.127
0.437      0.663      -3.312       5.173
job_title_Lead_Data_Scientist                          2.4796      2.972
0.834      0.407      -3.449       8.408
job_title_ML_Engineer                                  1.9840      2.958
0.671      0.505      -3.916       7.885
job_title_Machine_Learning_Engineer                    1.5101      1.023
1.477      0.144      -0.529       3.550
job_title_Machine_Learning_Infrastructure_Engineer     3.2782      2.880
1.138      0.259      -2.465       9.021
job_title_Machine_Learning_Scientist                   3.5321      3.386
1.043      0.301      -3.222      10.286
job_title_Manager_Data_Science                         1.2411      2.830
0.439      0.662      -4.404       6.886
job_title_Marketing_Data_Analyst                       2.8309      2.806
1.009      0.317      -2.766       8.428
job_title_Principal_Data_Analyst                       3.1727      2.885
1.100      0.275      -2.580       8.926
job_title_Principal_Data_Engineer                    -12.9105      4.506
-2.865      0.006     -21.898      -3.923
job_title_Principal_Data_Scientist                     3.4608      1.562
2.216      0.030       0.346       6.576
job_title_Product_Data_Analyst                        -8.5013      3.114
-2.730      0.008     -14.712      -2.290
job_title_Research_Scientist                          -0.3904      1.632
-0.239      0.812      -3.644       2.864
job_title_Staff_Data_Scientist                      5.487e-12   6.22e-12
0.882      0.381   -6.91e-12    1.79e-11
```

```
employee_residence_AT                                  5.4057      5.513
0.981      0.330      -5.590      16.401
employee_residence_BE                                  3.8739      1.555
2.491      0.015       0.773       6.975
employee_residence_BG                                  4.3732      3.261
1.341      0.184      -2.130      10.877
employee_residence_BR                                 -9.7715      5.431
-1.799      0.076     -20.602       1.059
employee_residence_CA                                  7.6433      4.696
1.628      0.108      -1.722      17.008
employee_residence_CL                              -1.066e-12   9.75e-13
-1.093      0.278   -3.01e-12    8.79e-13
employee_residence_CN                                 -1.6487      2.402
-0.686      0.495      -6.440       3.143
employee_residence_CO                                 -0.9162      1.516
-0.604      0.548      -3.940       2.108
employee_residence_DE                                  5.2784      3.439
1.535      0.129      -1.580      12.137
employee_residence_DK                                 -0.6969      2.652
-0.263      0.794      -5.986       4.593
employee_residence_ES                                  4.1136      2.727
1.508      0.136      -1.326       9.553
employee_residence_FR                                  4.0950      2.514
1.629      0.108      -0.918       9.108
employee_residence_GB                                 -4.1012      2.972
-1.380      0.172     -10.029       1.826
employee_residence_GR                                  1.5076      3.174
0.475      0.636      -4.822       7.837
employee_residence_HK                               3.739e-12   3.78e-12
0.990      0.326   -3.79e-12    1.13e-11
employee_residence_HR                                  0.5372      1.522
0.353      0.725      -2.498       3.572
employee_residence_HU                                 -3.0715      4.195
-0.732      0.467     -11.439       5.296
employee_residence_IN                                 -9.3334      2.367
-3.943      0.000     -14.054      -4.613
employee_residence_IR                               1.732e-12   1.67e-12
1.035      0.304    -1.6e-12    5.07e-12
employee_residence_IT                                  6.1201      3.425
1.787      0.078      -0.710      12.950
employee_residence_JE                                  5.2538      2.466
2.131      0.037       0.336      10.172
employee_residence_JP                                  2.5501      1.000
2.550      0.013       0.555       4.545
employee_residence_KE                                 -3.9494      2.099
-1.882      0.064      -8.136       0.237
employee_residence_LU                              -1.202e-12   1.31e-12
```

```
-0.921      0.360   -3.81e-12     1.4e-12
employee_residence_MD                                 -1.2283      1.815
-0.677      0.501      -4.848       2.392
employee_residence_MT                                 -1.2776      1.433
-0.892      0.376      -4.135       1.580
employee_residence_MX                                 -4.2061      1.150
-3.659      0.000      -6.499      -1.913
employee_residence_NG                                 -0.5572      1.041
-0.535      0.594      -2.634       1.520
employee_residence_NL                                  7.1599      4.622
1.549      0.126      -2.058      16.378
employee_residence_NZ                              -1.386e-12   1.48e-12
-0.936      0.352   -4.34e-12    1.57e-12
employee_residence_PH                                  0.6806      3.022
0.225      0.822      -5.347       6.708
employee_residence_PK                                 -9.0381      5.540
-1.632      0.107     -20.087       2.010
employee_residence_PL                                 -0.7609      2.618
-0.291      0.772      -5.983       4.461
employee_residence_PR                                  5.8451      3.493
1.673      0.099      -1.122      12.813
employee_residence_PT                                  2.4410      3.807
0.641      0.524      -5.152      10.034
employee_residence_RO                                 -6.2619      3.747
-1.671      0.099     -13.736       1.212
employee_residence_RS                                 -1.1814      4.679
-0.253      0.801     -10.513       8.150
employee_residence_RU                                 16.0680      5.896
2.725      0.008       4.309      27.827
employee_residence_SG                                  5.1570      1.522
3.389      0.001       2.122       8.192
employee_residence_SI                                 -1.6038      1.472
-1.090      0.280      -4.539       1.331
employee_residence_TR                                 -1.7678      1.120
-1.578      0.119      -4.003       0.467
employee_residence_UA                                 -2.6728      1.488
-1.796      0.077      -5.640       0.295
employee_residence_US                                  4.8100      1.480
3.251      0.002       1.859       7.761
employee_residence_VN                                 -7.7642      4.263
-1.821      0.073     -16.266       0.737
company_location_AS                                    5.6315      4.870
1.156      0.251      -4.081      15.344
company_location_AT                                    0.7221      4.493
0.161      0.873      -8.239       9.684
company_location_BE                                    3.8739      1.555
2.491      0.015       0.773       6.975
```

```
company_location_BR                                    4.7085      6.168
0.763      0.448      -7.593      17.010
company_location_CA                                    0.7494      4.269
0.176      0.861      -7.765       9.264
company_location_CH                                   -0.0416      4.015
-0.010      0.992      -8.049       7.966
company_location_CL                                 5.331e-18   2.13e-17
0.251      0.803   -3.71e-17    4.77e-17
company_location_CN                                    3.6050      1.456
2.476      0.016       0.701       6.509
company_location_CO                                   -0.9162      1.516
-0.604      0.548      -3.940       2.108
company_location_DE                                    0.3665      3.602
0.102      0.919      -6.818       7.551
company_location_DK                                    2.1340      1.939
1.100      0.275      -1.734       6.002
company_location_ES                                   -1.3657      3.140
-0.435      0.665      -7.629       4.897
company_location_FR                                   -0.8660      2.490
-0.348      0.729      -5.833       4.101
company_location_GB                                   10.4463      2.778
3.760      0.000       4.905      15.988
company_location_GR                                    1.1235      2.146
0.524      0.602      -3.156       5.403
company_location_HR                                    0.5372      1.522
0.353      0.725      -2.498       3.572
company_location_HU                                         0          0
nan        nan          0           0
company_location_IL                                    5.1570      1.522
3.389      0.001       2.122       8.192
company_location_IN                                    7.3119      2.343
3.120      0.003       2.638      11.985
company_location_IR                                         0          0
nan        nan          0           0
company_location_IT                                         0          0
nan        nan          0           0
company_location_JP                                    2.5501      1.000
2.550      0.013       0.555       4.545
company_location_KE                                   -3.9494      2.099
-1.882      0.064      -8.136       0.237
company_location_LU                                         0          0
nan        nan          0           0
company_location_MD                                   -1.2283      1.815
-0.677      0.501      -4.848       2.392
company_location_MT                                   -1.2776      1.433
-0.892      0.376      -4.135       1.580
company_location_MX                                   -4.2061      1.150
```

```
-3.659      0.000      -6.499      -1.913
company_location_NG                                   -0.5572      1.041
-0.535      0.594      -2.634       1.520
company_location_NL                                   -4.6550      5.471
-0.851      0.398     -15.567       6.257
company_location_NZ                                         0          0
nan        nan          0          0
company_location_PK                                    0.1987      6.375
0.031      0.975     -12.515      12.913
company_location_PL                                    0.5999      2.935
0.204      0.839      -5.255       6.455
company_location_PT                                    1.5683      5.002
0.314      0.755      -8.408      11.544
company_location_RU                                   -4.0535      3.443
-1.177      0.243     -10.920       2.813
company_location_SG                                         0          0
nan        nan          0          0
company_location_SI                                   -1.6038      1.472
-1.090      0.280      -4.539       1.331
company_location_TR                                   -1.7678      1.120
-1.578      0.119      -4.003       0.467
company_location_UA                                   -2.6728      1.488
-1.796      0.077      -5.640       0.295
company_location_US                                    2.6906      1.394
1.931      0.058      -0.089       5.470
company_location_VN                                   -3.7090      4.790
-0.774      0.441     -13.263       5.845
company_size_M                                        -1.5362      0.860
-1.785      0.079      -3.252       0.180
company_size_S                                        -0.8028      0.962
-0.835      0.407      -2.721       1.116
==============================================================================
Omnibus:                       24.077   Durbin-Watson:                   2.067
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               67.549
Skew:                          -0.519   Prob(JB):                     2.15e-15
Kurtosis:                       5.899   Cond. No.                     5.87e+20
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The smallest eigenvalue is 9.49e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
"""
```

2.6 Define OSR2

[16]:

```python
# compute out-of-sample R-squared using the test set
def OSR2(model, df_train, df_test, dependent_var):
    y_test = df_test[dependent_var]
    y_pred = model.predict(df_test)
    SSE = np.sum((y_test - y_pred)**2)
    SST = np.sum((y_test - np.mean(df_train[dependent_var]))**2)
    return 1 - SSE/SST
```

[17]:

```python
OSR2(linreg, df_train, df_test, target_column)
```

[17]:

```
0.7394346635275322
```

2.7 Feature Selection: Use VIF to keep good features.

[18]:

```python
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor

def VIF(df, columns):
    values = sm.add_constant(df[columns]).values
    num_columns = len(columns) + 1
    vif = [variance_inflation_factor(values, i) for i in range(num_columns)]
    return pd.Series(vif[1:], index=columns)
```

2.8 Feature Selection: Identify and eliminate high P-Value features

[19]:

```python
def get_formula(features, target):
    features = [f for f in features if f != target]
    sum_features = ' + '.join(features)
    formula = ' ~ '.join([target, sum_features])
    return formula
```

[20]:

```python
def filter_feature(features, model, df_train, target):
    p_values = {}
    features = [f for f in model.params.index if f != 'Intercept']

    #
    for feat in features:
        p_values[feat] = model.pvalues.loc[feat]

    worst_feat = max(features, key=lambda f: p_values[f])

    print('WORST:', worst_feat, '-->', p_values[worst_feat])
    new_features = [f for f in features if f != worst_feat]

    new_model = smf.ols(formula = get_formula(new_features, target),
                            data = df_train).fit()


    return new_features, new_model, p_values[worst_feat]
```

[21]:
```
features = df_train.columns
model = linreg
p_value = float('inf')
models = {} # (formula, model)

while p_value > 0.05:
    features, model, p_value = filter_feature(features, model, df_train,␣
 ↪target_column)
    formula = get_formula(features, target_column)
```

```
WORST: company_location_CH --> 0.9917592810503727
WORST: job_title_Big_Data_Architect --> 0.9765611842923134
WORST: job_title_Data_Science_Consultant --> 0.978899374584294
WORST: company_location_PK --> 0.9751836220509351
WORST: company_location_CL --> 0.957280923019161
WORST: company_location_DE --> 0.9496630619778473
WORST: job_title_Computer_Vision_Engineer --> 0.937177139133903
WORST: work_year_2021e --> 0.9307019294832726
WORST: company_location_CA --> 0.9163775632788269
WORST: company_location_PL --> 0.9154613441099624
WORST: company_location_AT --> 0.9138669495937461
WORST: job_title_Research_Scientist --> 0.8906949325926343
WORST: job_title_Lead_Data_Analyst --> 0.9021539749130271
WORST: job_title_Data_Analyst --> 0.8980333768007385
WORST: company_location_PT --> 0.8149601176331617
WORST: company_location_HR --> 0.8150797190579797
WORST: job_title_Business_Data_Analyst --> 0.9202698235782161
WORST: employee_residence_HR --> 0.815079718972983
WORST: employee_residence_PH --> 0.9664697386595282
WORST: employee_residence_GR --> 0.8082642134004365
WORST: company_location_FR --> 0.6357524503392765
WORST: job_title_Cloud_Data_Engineer --> 0.9677198019038615
WORST: company_location_ES --> 0.7197205854084017
WORST: employee_residence_HK --> 0.9557818595661532
WORST: employee_residence_CL --> 0.9968502734836702
WORST: employee_residence_PL --> 0.6909557178946283
WORST: employee_residence_LU --> 0.9084730230282493
WORST: employee_residence_NZ --> 0.859307410598456
WORST: job_title_Head_of_Data_Science --> 0.7062949780221364
WORST: job_title_Big_Data_Engineer --> 0.7005089358665094
WORST: employee_residence_IR --> 0.6787798946154769
WORST: job_title_Staff_Data_Scientist --> 0.8792821168890983
```

```
WORST: company_location_HU --> 0.9272219478075725
WORST: company_location_IR --> 0.7744399997154151
WORST: employee_residence_RS --> 0.6787171506466998
WORST: company_location_IT --> 0.9016017008083217
WORST: job_title_Lead_Data_Engineer --> 0.6564565220889897
WORST: job_title_Computer_Vision_Software_Engineer --> 0.6686086550600603
WORST: employee_residence_DK --> 0.6523054428761657
WORST: job_title_Manager_Data_Science --> 0.6151040095322029
WORST: job_title_Data_Specialist --> 0.5535398310173966
WORST: employee_residence_NG --> 0.5415779583160978
WORST: company_location_LU --> 0.8670642552949288
WORST: company_location_NG --> 0.5415779580918841
WORST: employee_residence_CO --> 0.5750071339008482
WORST: company_location_CO --> 0.5750071338989861
WORST: job_title_Data_Analytics_Engineer --> 0.5316336596350422
WORST: job_title_ML_Engineer --> 0.519202249363028
WORST: job_title_Data_Analytics_Manager --> 0.5195121681344219
WORST: job_title_BI_Data_Analyst --> 0.5232920903632007
WORST: employee_residence_HU --> 0.5374865851441789
WORST: employee_residence_GB --> 0.5133020818990319
WORST: company_location_MT --> 0.5055915920116858
WORST: employee_residence_MT --> 0.5055915920465585
WORST: job_title_AI_Scientist --> 0.5075837760551528
WORST: remote_ratio --> 0.5354983173300677
WORST: employee_residence_CN --> 0.5378249025811064
WORST: job_title_Lead_Data_Scientist --> 0.4954750579896604
WORST: job_title_Data_Engineering_Manager --> 0.4632261450010775
WORST: job_title_Director_of_Data_Science --> 0.5523598612457511
WORST: job_title_Director_of_Data_Engineering --> 0.5021124501705805
WORST: job_title_Principal_Data_Analyst --> 0.4680701470325329
WORST: job_title_Machine_Learning_Infrastructure_Engineer --> 0.4856626869788869
WORST: employee_residence_SI --> 0.46050380187407713
WORST: company_location_SI --> 0.4605038018713522
WORST: job_title_Data_Scientist --> 0.4189258110639156
WORST: job_title_Marketing_Data_Analyst --> 0.40253054007965905
WORST: company_location_BR --> 0.41746030310376336
WORST: employee_residence_MD --> 0.3166879555856055
WORST: company_location_MD --> 0.31668795558148877
WORST: job_title_Data_Architect --> 0.3066123768719811
WORST: employee_residence_RO --> 0.317503351071737
WORST: job_title_Machine_Learning_Engineer --> 0.21901309459396162
WORST: employee_residence_BG --> 0.23769372183539975
WORST: job_title_Machine_Learning_Scientist --> 0.20449851004080027
WORST: job_title_Principal_Data_Scientist --> 0.23958043514479097
WORST: job_title_Data_Science_Engineer --> 0.19366650816381334
WORST: company_location_GR --> 0.1936665080268255
WORST: company_location_CN --> 0.18029210215867833
WORST: employee_residence_PT --> 0.20593951540626168
```

```
WORST: company_size_M --> 0.17824332034677423
WORST: job_title_Finance_Data_Analyst --> 0.17971681822916835
WORST: company_size_S --> 0.19610909727278517
WORST: company_location_RU --> 0.15365049646812592
WORST: job_title_Head_of_Data --> 0.15365049647602486
WORST: employment_type_FL --> 0.36007373222483796
WORST: company_location_NL --> 0.14438500198718263
WORST: job_title_Data_Engineer --> 0.21137667285849507
WORST: job_title_Applied_Data_Scientist --> 0.12074026356174956
WORST: employee_residence_PR --> 0.08765391158777444
WORST: employee_residence_VN --> 0.08750858729921741
WORST: employee_residence_AT --> 0.0711152222596231
WORST: company_location_DK --> 0.07944439646037656
WORST: employee_residence_IT --> 0.09345170753000334
WORST: employee_residence_ES --> 0.14387124916723143
WORST: employee_residence_FR --> 0.1827550794812623
WORST: job_title_Applied_Machine_Learning_Scientist --> 0.08398503678915212
WORST: job_title_Financial_Data_Analyst --> 0.06962268170459035
WORST: experience_level_MI --> 0.05951294621155277
WORST: company_location_AS --> 0.06033434807966886
WORST: employee_residence_RU --> 0.04207439881923185
```

[22]:

```python
best_features = features + ['employee_residence_RU']
print(best_features)
best_linreg = smf.ols(formula=get_formula(best_features, target_column),
                 data=df_train).fit()
print(best_linreg.summary())
```

```
['salary_in_usd', 'experience_level_EX', 'experience_level_SE',
'employment_type_FT', 'employment_type_PT', 'job_title_Data_Science_Manager',
'job_title_Principal_Data_Engineer', 'job_title_Product_Data_Analyst',
'employee_residence_BE', 'employee_residence_BR', 'employee_residence_CA',
'employee_residence_DE', 'employee_residence_IN', 'employee_residence_JE',
'employee_residence_JP', 'employee_residence_KE', 'employee_residence_MX',
'employee_residence_NL', 'employee_residence_PK', 'employee_residence_SG',
'employee_residence_TR', 'employee_residence_UA', 'employee_residence_US',
'company_location_BE', 'company_location_GB', 'company_location_IL',
'company_location_IN', 'company_location_JP', 'company_location_KE',
'company_location_MX', 'company_location_NZ', 'company_location_SG',
'company_location_TR', 'company_location_UA', 'company_location_US',
'company_location_VN', 'employee_residence_RU']
                            OLS Regression Results
==============================================================================
Dep. Variable:     cbrt_salary_in_usd   R-squared:                       0.966
Model:                            OLS   Adj. R-squared:                  0.959
Method:                 Least Squares   F-statistic:                     143.4
Date:                Mon, 08 May 2023   Prob (F-statistic):           3.01e-90
Time:                        01:04:35   Log-Likelihood:                -391.77
```

```
No. Observations:                 171   AIC:                             841.5
Df Residuals:                     142   BIC:                             932.7
Df Model:                          28
Covariance Type:            nonrobust
================================================================================
=====================
                                        coef    std err          t      P>|t|
[0.025      0.975]
--------------------------------------------------------------------------------
---------------------
Intercept                            23.1892      1.824     12.712      0.000
19.583      26.795
salary_in_usd                         0.0001   3.52e-06     28.838      0.000
9.46e-05       0.000
experience_level_EX                   4.2508      1.177      3.612      0.000
1.925       6.577
experience_level_SE                   2.8723      0.496      5.792      0.000
1.892       3.853
employment_type_FT                    6.7035      1.745      3.842      0.000
3.254      10.153
employment_type_PT                    5.6287      2.302      2.445      0.016
1.078      10.180
job_title_Data_Science_Manager        3.3110      1.286      2.575      0.011
0.769       5.853
job_title_Principal_Data_Engineer   -16.0164      3.072     -5.214      0.000
-22.089      -9.944
job_title_Product_Data_Analyst       -8.2223      2.746     -2.994      0.003
-13.651      -2.793
employee_residence_BE                 2.8731      1.336      2.150      0.033
0.232       5.515
employee_residence_BR                -8.3231      1.649     -5.046      0.000
-11.584      -5.063
employee_residence_CA                 4.7788      1.408      3.394      0.001
1.996       7.562
employee_residence_DE                 3.7198      0.851      4.373      0.000
2.038       5.401
employee_residence_IN                -9.6288      1.439     -6.692      0.000
-12.473      -6.784
employee_residence_JE                 6.3717      2.675      2.382      0.019
1.083      11.660
employee_residence_JP                 1.8942      0.711      2.664      0.009
0.489       3.299
employee_residence_KE                -4.9128      1.333     -3.685      0.000
-7.548      -2.277
employee_residence_MX                -4.9726      0.958     -5.189      0.000
-6.867      -3.078
employee_residence_NL                 3.9997      1.669      2.397      0.018
0.701       7.298
```

```
employee_residence_PK                 -9.9922      2.048     -4.878      0.000
-14.042      -5.943
employee_residence_SG                  3.6133      1.341      2.694      0.008
0.962       6.264
employee_residence_TR                 -2.3577      0.957     -2.465      0.015
-4.249      -0.467
employee_residence_UA                 -3.7505      1.333     -2.813      0.006
-6.386      -1.115
employee_residence_US                  3.1694      0.883      3.589      0.000
1.424       4.915
company_location_BE                    2.8731      1.336      2.150      0.033
0.232       5.515
company_location_GB                    3.3960      0.900      3.775      0.000
1.618       5.174
company_location_IL                    3.6133      1.341      2.694      0.008
0.962       6.264
company_location_IN                    5.5856      1.600      3.490      0.001
2.422       8.749
company_location_JP                    1.8942      0.711      2.664      0.009
0.489       3.299
company_location_KE                   -4.9128      1.333     -3.685      0.000
-7.548      -2.277
company_location_MX                   -4.9726      0.958     -5.189      0.000
-6.867      -3.078
company_location_NZ                         0          0        nan        nan
0           0
company_location_SG                         0          0        nan        nan
0           0
company_location_TR                   -2.3577      0.957     -2.465      0.015
-4.249      -0.467
company_location_UA                   -3.7505      1.333     -2.813      0.006
-6.386      -1.115
company_location_US                    2.1380      0.786      2.721      0.007
0.585       3.691
company_location_VN                  -14.4247      2.667     -5.408      0.000
-19.697      -9.152
employee_residence_RU                  4.3178      2.105      2.051      0.042
0.157       8.479
==============================================================================
Omnibus:                       25.929   Durbin-Watson:                   1.946
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               43.882
Skew:                          -0.772   Prob(JB):                     2.96e-10
Kurtosis:                       4.943   Cond. No.                     6.31e+21
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

```
[2] The smallest eigenvalue is 8.2e-32. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

[23]:

```
OSR2(best_linreg, df_train, df_test, target_column)
```

[23]:

```
0.8035179130896895
```

2.9 Remove 0 coefficient and NaN p-value features

[24]:

```
best_features.remove('company_location_NZ')
best_features.remove('company_location_SG')
best_linreg2 = smf.ols(formula=get_formula(best_features, target_column),
                 data=df_train).fit()
print(best_linreg2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     cbrt_salary_in_usd   R-squared:                       0.966
Model:                            OLS   Adj. R-squared:                  0.959
Method:                 Least Squares   F-statistic:                     143.4
Date:                Mon, 08 May 2023   Prob (F-statistic):           3.01e-90
Time:                        01:04:36   Log-Likelihood:                -391.77
No. Observations:                 171   AIC:                             841.5
Df Residuals:                     142   BIC:                             932.7
Df Model:                          28
Covariance Type:            nonrobust
================================================================================
=====================
                                        coef    std err          t      P>|t|
[0.025      0.975]
--------------------------------------------------------------------------------
---------------------
Intercept                            23.1892      1.824     12.712      0.000
19.583      26.795
salary_in_usd                         0.0001   3.52e-06     28.838      0.000
9.46e-05       0.000
experience_level_EX                   4.2508      1.177      3.612      0.000
1.925       6.577
experience_level_SE                   2.8723      0.496      5.792      0.000
1.892       3.853
employment_type_FT                    6.7035      1.745      3.842      0.000
3.254      10.153
employment_type_PT                    5.6287      2.302      2.445      0.016
1.078      10.180
job_title_Data_Science_Manager        3.3110      1.286      2.575      0.011
0.769       5.853
job_title_Principal_Data_Engineer   -16.0164      3.072     -5.214      0.000
```

```
-22.089      -9.944
job_title_Product_Data_Analyst       -8.2223      2.746     -2.994      0.003
-13.651      -2.793
employee_residence_BE                 2.8731      1.336      2.150      0.033
0.232       5.515
employee_residence_BR                -8.3231      1.649     -5.046      0.000
-11.584      -5.063
employee_residence_CA                 4.7788      1.408      3.394      0.001
1.996       7.562
employee_residence_DE                 3.7198      0.851      4.373      0.000
2.038       5.401
employee_residence_IN                -9.6288      1.439     -6.692      0.000
-12.473      -6.784
employee_residence_JE                 6.3717      2.675      2.382      0.019
1.083      11.660
employee_residence_JP                 1.8942      0.711      2.664      0.009
0.489       3.299
employee_residence_KE                -4.9128      1.333     -3.685      0.000
-7.548      -2.277
employee_residence_MX                -4.9726      0.958     -5.189      0.000
-6.867      -3.078
employee_residence_NL                 3.9997      1.669      2.397      0.018
0.701       7.298
employee_residence_PK                -9.9922      2.048     -4.878      0.000
-14.042      -5.943
employee_residence_SG                 3.6133      1.341      2.694      0.008
0.962       6.264
employee_residence_TR                -2.3577      0.957     -2.465      0.015
-4.249      -0.467
employee_residence_UA                -3.7505      1.333     -2.813      0.006
-6.386      -1.115
employee_residence_US                 3.1694      0.883      3.589      0.000
1.424       4.915
company_location_BE                   2.8731      1.336      2.150      0.033
0.232       5.515
company_location_GB                   3.3960      0.900      3.775      0.000
1.618       5.174
company_location_IL                   3.6133      1.341      2.694      0.008
0.962       6.264
company_location_IN                   5.5856      1.600      3.490      0.001
2.422       8.749
company_location_JP                   1.8942      0.711      2.664      0.009
0.489       3.299
company_location_KE                  -4.9128      1.333     -3.685      0.000
-7.548      -2.277
company_location_MX                  -4.9726      0.958     -5.189      0.000
-6.867      -3.078
company_location_TR                  -2.3577      0.957     -2.465      0.015
```

```
-4.249      -0.467
company_location_UA                      -3.7505      1.333     -2.813      0.006
-6.386      -1.115
company_location_US                       2.1380      0.786      2.721      0.007
0.585       3.691
company_location_VN                     -14.4247      2.667     -5.408      0.000
-19.697      -9.152
employee_residence_RU                     4.3178      2.105      2.051      0.042
0.157       8.479
==============================================================================
Omnibus:                       25.929   Durbin-Watson:                   1.946
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               43.882
Skew:                          -0.772   Prob(JB):                     2.96e-10
Kurtosis:                       4.943   Cond. No.                     4.06e+21
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The smallest eigenvalue is 1.98e-31. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

[25]:
```python
OSR2(best_linreg2, df_train, df_test, target_column)
```

[25]:
```
0.8035179130896896
```

2.10 Predict test set and compare to results.

[26]:
```python
y_pred = best_linreg2.predict(df_test)
y_pred
```

[26]:
```
201    52.284403
14     45.656034
12     42.383180
226    42.813622
140    48.904506
         …
154    43.422710
18     41.290900
74     28.740143
138    47.325491
121    44.656114
Length: 74, dtype: float64
```

[27]:
```python
df_test
```

```
[27]:
     salary_in_usd  remote_ratio  work_year_2021e  experience_level_EX  \
201         140000           100                1                    0
14          103000           100                0                    0
12           75966           100                1                    0
226          75000             0                1                    0
140         135000           100                0                    0
..              …           …              …                  …
154          81000            50                1                    0
18           60000           100                1                    0
74           28475           100                1                    0
138          96357            50                1                    0
121          93150             0                1                    0

     experience_level_MI  experience_level_SE  employment_type_FL  \
201                    0                    1                   0
14                     1                    0                   0
12                     1                    0                   0
226                    1                    0                   0
140                    1                    0                   0
..                   …                  …                 …
154                    0                    0                   0
18                     0                    0                   0
74                     0                    0                   0
138                    0                    1                   0
121                    1                    0                   0

     employment_type_FT  employment_type_PT  job_title_AI_Scientist  …  \
201                   1                   0                       0  …
14                    1                   0                       0  …
12                    1                   0                       0  …
226                   1                   0                       0  …
140                   1                   0                       0  …
..                  …                 …                     …  …
154                   1                   0                       0  …
18                    1                   0                       0  …
74                    1                   0                       0  …
138                   1                   0                       0  …
121                   1                   0                       0  …

     company_location_RU  company_location_SG  company_location_SI  \
201                    0                    0                    0
14                     0                    0                    0
12                     0                    0                    0
226                    0                    0                    0
140                    0                    0                    0
..                   …                  …                  …
154                    0                    0                    0
```

```
18                     0                    0                    0
74                     0                    0                    0
138                    0                    0                    0
121                    0                    0                    0

     company_location_TR  company_location_UA  company_location_US  \
201                    0                    0                    1
14                     0                    0                    1
12                     0                    0                    0
226                    0                    0                    1
140                    0                    0                    1
..                   …                  …                  …
154                    0                    0                    1
18                     0                    0                    1
74                     0                    0                    0
138                    0                    0                    0
121                    0                    0                    1

     company_location_VN  company_size_M  company_size_S  cbrt_salary_in_usd
201                    0               0               0           51.924941
14                     0               0               0           46.875481
12                     0               0               0           42.351918
226                    0               0               0           42.171633
140                    0               0               0           51.299278
..                   …            …            …                 …
154                    0               0               1           43.267487
18                     0               0               1           39.148676
74                     0               1               0           30.536640
138                    0               0               0           45.845258
121                    0               1               0           45.330894

[74 rows x 138 columns]
```

2.11 Predict my own entries (randomized).

2.11.1 Read CSV again with new dataframe, dropping unnecessary features.

[28]:

```
df_copy = pd.read_csv(r"Data Science Jobs Salaries.csv").
 ↪drop(columns=['salary_currency', 'salary_in_usd', 'salary'])
df_copy
```

[28]:

```
   work_year experience_level employment_type                  job_title  \
0      2021e               EN              FT    Data Science Consultant
1       2020               SE              FT             Data Scientist
2      2021e               EX              FT       Head of Data Science
3      2021e               EX              FT               Head of Data
4      2021e               EN              FT  Machine Learning Engineer
```

```
..        …              …             …                        …
240      2020               SE              FT             Data Scientist
241     2021e               MI              FT   Principal Data Scientist
242      2020               EN              FT             Data Scientist
243      2020               EN              CT      Business Data Analyst
244     2021e               SE              FT       Data Science Manager

    employee_residence  remote_ratio company_location company_size
0                   DE            50               DE            L
1                   GR           100               US            L
2                   RU             0               RU            M
3                   RU            50               RU            L
4                   US           100               US            S
..                 …           …              …          …
240                 US           100               US            L
241                 US           100               US            L
242                 US           100               US            S
243                 US           100               US            L
244                 IN            50               IN            L

[245 rows x 8 columns]
```

2.11.2 Get unique values of each of our features to randomize.

[29]:
```
possible_WY = df_copy['work_year'].unique()
possible_ET = df_copy['employment_type'].unique()
possible_ER = df_copy['employee_residence'].unique()
possible_RR = df_copy['remote_ratio'].unique()
possible_CL = df_copy['company_location'].unique()
possible_CS = df_copy['company_size'].unique()
```

2.11.3 Make an empty dataframe to add our imaginary data scientists to.

[30]:
```
random_df_copy = df_copy.copy()
random_df_copy = random_df_copy[0:0]
random_df_copy
```

[30]:
```
Empty DataFrame
Columns: [work_year, experience_level, employment_type, job_title,
employee_residence, remote_ratio, company_location, company_size]
Index: []
```

2.11.4 Create 200 imaginary data scientists by randomizing possible values from our raw dataset.

[31]:

```python
for i in range(200):

    EN_JT_index = random.randint(0, len(df_copy.index)-1)

    random_WY = random.choice(possible_WY)
    random_EL = df_copy.iloc[EN_JT_index]['experience_level']
    random_ET = random.choice(possible_ET)
    random_JT = df_copy.iloc[EN_JT_index]['job_title']
    random_ER = random.choice(possible_ER)
    random_RR = random.choice(possible_RR)
    random_CL = random.choice(possible_CL)
    random_CS = random.choice(possible_CS)

    random_df_copy.loc[len(random_df_copy.index)] = [random_WY, random_EL,␣
 ↪random_ET, random_JT,
                                                      random_ER, random_RR,␣
 ↪random_CL, random_CS]

random_df_copy
```

[31]:

```
    work_year experience_level employment_type                   job_title  \
0       2021e               EN              FL              Data Scientist
1       2021e               SE              FT  Machine Learning Scientist
2        2020               EN              CT     Data Science Consultant
3        2020               MI              PT               Data Engineer
4        2020               MI              PT                Data Analyst
..        …                …               …                           …
195     2021e               SE              FL        Data Science Manager
196      2020               SE              CT    Data Engineering Manager
197      2020               SE              PT              Data Scientist
198      2020               EN              CT              Data Scientist
199     2021e               EN              FL              Data Scientist

    employee_residence  remote_ratio company_location company_size
0                   UA            50               SI            L
1                   US           100               ES            L
2                   TR             0               CN            L
3                   JP             0               IR            M
4                   SG             0               ES            M
..                 …            …                …           …
195                 GR            50               IL            L
196                 MT            50               RU            L
197                 US            50               CH            M
198                 DK           100               NZ            S
```

```
199                  JP            50               PK           S

[200 rows x 8 columns]
```

2.11.5 Create dummy variables for our features to use in making predictions.

[32]:

```
random_df_copy = pd.get_dummies(random_df_copy)
random_df_copy
```

[32]:

```
     remote_ratio  work_year_2020  work_year_2021e  experience_level_EN  \
0              50               0                1                    1
1             100               0                1                    0
2               0               1                0                    1
3               0               1                0                    0
4               0               1                0                    0
..            …             …              …                  …
195            50               0                1                    0
196            50               1                0                    0
197            50               1                0                    0
198           100               1                0                    1
199            50               0                1                    1

     experience_level_EX  experience_level_MI  experience_level_SE  \
0                      0                    0                    0
1                      0                    0                    1
2                      0                    0                    0
3                      0                    1                    0
4                      0                    1                    0
..                   …                  …                  …
195                    0                    0                    1
196                    0                    0                    1
197                    0                    0                    1
198                    0                    0                    0
199                    0                    0                    0

     employment_type_CT  employment_type_FL  employment_type_FT  …  \
0                     0                   1                   0  …
1                     0                   0                   1  …
2                     1                   0                   0  …
3                     0                   0                   0  …
4                     0                   0                   0  …
..                  …                 …                 …  …
195                   0                   1                   0  …
196                   1                   0                   0  …
197                   0                   0                   0  …
198                   1                   0                   0  …
199                   0                   1                   0  …
```

```
     company_location_RU  company_location_SG  company_location_SI  \
0                      0                    0                    1
1                      0                    0                    0
2                      0                    0                    0
3                      0                    0                    0
4                      0                    0                    0
..                   …                    …                    …
195                    0                    0                    0
196                    1                    0                    0
197                    0                    0                    0
198                    0                    0                    0
199                    0                    0                    0

     company_location_TR  company_location_UA  company_location_US  \
0                      0                    0                    0
1                      0                    0                    0
2                      0                    0                    0
3                      0                    0                    0
4                      0                    0                    0
..                   …                    …                    …
195                    0                    0                    0
196                    0                    0                    0
197                    0                    0                    0
198                    0                    0                    0
199                    0                    0                    0

     company_location_VN  company_size_L  company_size_M  company_size_S
0                      0               1               0               0
1                      0               1               0               0
2                      0               1               0               0
3                      0               0               1               0
4                      0               0               1               0
..                   …              …               …               …
195                    0               1               0               0
196                    0               1               0               0
197                    0               0               1               0
198                    0               0               0               1
199                    0               0               0               1

[200 rows x 138 columns]
```

2.11.6 Rename columns to include _ in existing ones with spaces (to match the features in the model).

[33]:
```
for col in random_df_copy.columns:
    random_df_copy.rename(columns={col : '_'.join(col.split())}, inplace=True)
```

2.11.7 Add 0 columns for features which weren't randomly chosen from possible values. If this happens, the dummy variables for those possible categories are not in our testing set and cannot work with the model as intended.

[34]:
```
for c in df.columns:
    if c not in random_df_copy.columns:
        random_df_copy[c] = 0
```

2.11.8 Remove unneeded columns to match model perfectly.

[35]:
```
to_drop = ['work_year_2020', 'experience_level_EN', 'employment_type_CT',␣
 ↪'job_title_3D_Computer_Vision_Researcher',
'employee_residence_AE', 'company_location_AE', 'company_size_L',␣
 ↪'log_salary_in_usd', 'sqrt_salary_in_usd', 'cbrt_salary_in_usd']
for c in to_drop:
    if c in random_df_copy.columns:
        random_df_copy.drop(columns=[c], inplace=True)
```

2.11.9 Print the processed, randomized dataframe of 200 data scientists.

[36]:
```
random_df_copy
```

[36]:
```
     remote_ratio  work_year_2021e  experience_level_EX  experience_level_MI  \
0              50                1                    0                    0
1             100                1                    0                    0
2               0                0                    0                    0
3               0                0                    0                    1
4               0                0                    0                    1
..            …                …                    …                    …
195            50                1                    0                    0
196            50                0                    0                    0
197            50                0                    0                    0
198           100                0                    0                    0
199            50                1                    0                    0

     experience_level_SE  employment_type_FL  employment_type_FT  \
0                      0                   1                   0
1                      1                   0                   1
2                      0                   0                   0
3                      0                   0                   0
```

```
4                      0                   0                   0
..                   …                 …                 …
195                    1                   1                   0
196                    1                   0                   0
197                    1                   0                   0
198                    0                   0                   0
199                    0                   1                   0

     employment_type_PT  job_title_AI_Scientist  \
0                     0                       0
1                     0                       0
2                     0                       0
3                     1                       0
4                     1                       0
..                  …                     …
195                   0                       0
196                   0                       0
197                   1                       0
198                   0                       0
199                   0                       0

     job_title_Applied_Data_Scientist  …  company_location_US  \
0                                   0  …                    0
1                                   0  …                    0
2                                   0  …                    0
3                                   0  …                    0
4                                   0  …                    0
..                                …  …                  …
195                                 0  …                    0
196                                 0  …                    0
197                                 0  …                    0
198                                 0  …                    0
199                                 0  …                    0

     company_location_VN  company_size_M  company_size_S  salary_in_usd  \
0                      0               0               0              0
1                      0               0               0              0
2                      0               0               0              0
3                      0               1               0              0
4                      0               1               0              0
..                   …             …             …            …
195                    0               0               0              0
196                    0               0               0              0
197                    0               1               0              0
198                    0               0               1              0
199                    0               0               1              0
```

```
     job_title_Big_Data_Architect  \
0                               0
1                               0
2                               0
3                               0
4                               0
..                            …
195                             0
196                             0
197                             0
198                             0
199                             0

     job_title_Machine_Learning_Infrastructure_Engineer  \
0                                                     0
1                                                     0
2                                                     0
3                                                     0
4                                                     0
..                                                  …
195                                                   0
196                                                   0
197                                                   0
198                                                   0
199                                                   0

     job_title_Marketing_Data_Analyst  job_title_Principal_Data_Engineer  \
0                                   0                                  0
1                                   0                                  0
2                                   0                                  0
3                                   0                                  0
4                                   0                                  0
..                                …                                  …
195                                 0                                  0
196                                 0                                  0
197                                 0                                  0
198                                 0                                  0
199                                 0                                  0

     job_title_Staff_Data_Scientist
0                                 0
1                                 0
2                                 0
3                                 0
4                                 0
..                              …
195                               0
```

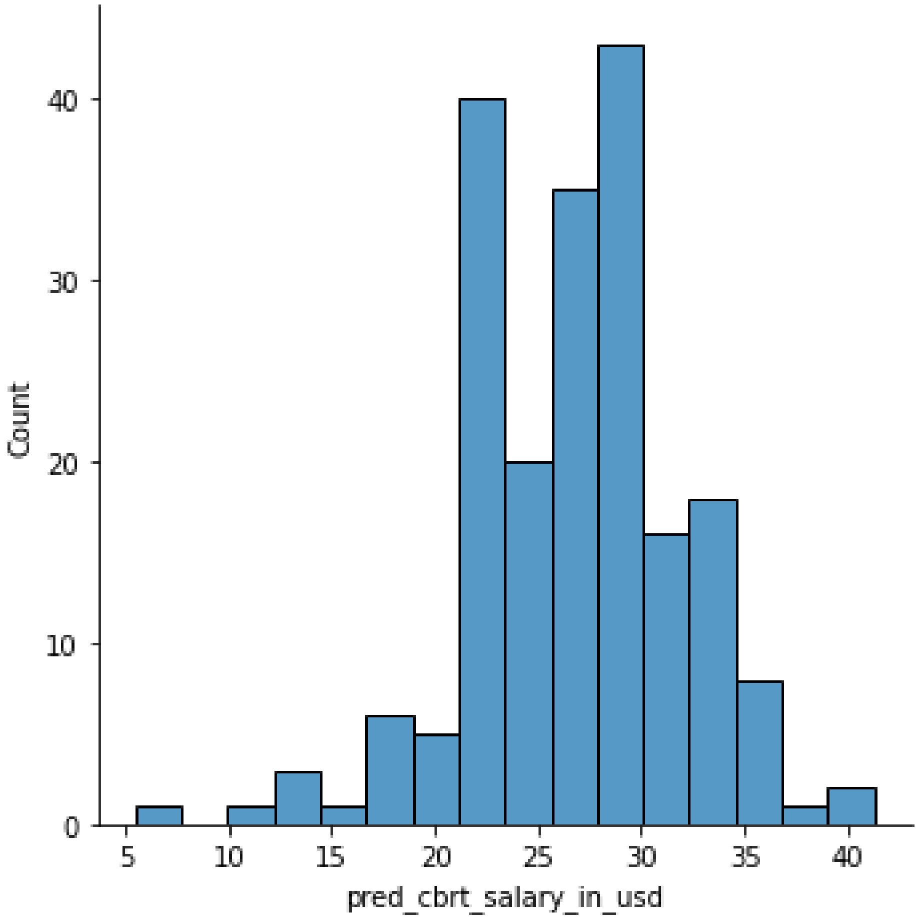
```
196                                  0
197                                  0
198                                  0
199                                  0

[200 rows x 137 columns]
```

2.11.10 Make predictions for our 200 random data scientist's salaries.

[37]:
```
y_pred_rand = best_linreg2.predict(random_df_copy)
# y_pred_rand = np.rint(y_pred_rand).astype(int)
y_pred_rand
```

[37]:
```
0      19.438706
1      35.934362
2      20.831496
3      30.711979
4      32.431146
         …
195    32.985860
196    26.061482
197    34.859494
198    23.189158
199    25.083323
Length: 200, dtype: float64
```

[38]:
```
random_df_copy['pred_cbrt_salary_in_usd'] = y_pred_rand
sns.displot(data=random_df_copy, x = 'pred_cbrt_salary_in_usd')
plt.show();
```



[39]:

```
negative_predictions = random_df_copy[random_df_copy['pred_cbrt_salary_in_usd']␣
 ↪< 0]
display(negative_predictions)
```

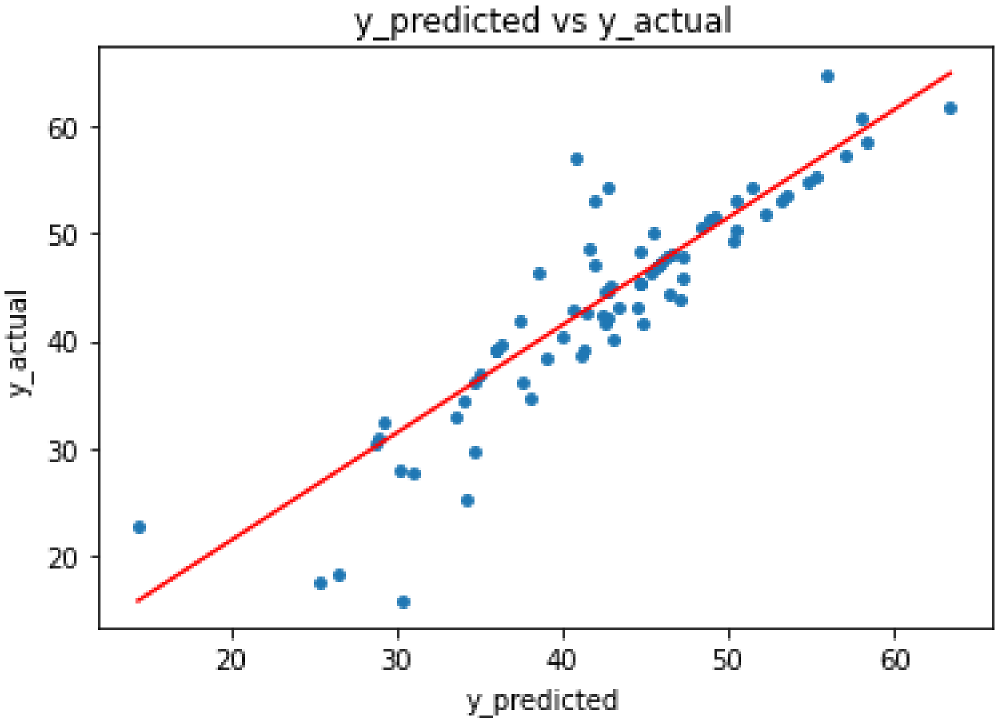
```
Empty DataFrame
```

```
Columns: [remote_ratio, work_year_2021e, experience_level_EX,␣
↪experience_level_MI, experience_level_SE, employment_type_FL,␣
↪employment_type_FT, employment_type_PT, job_title_AI_Scientist,␣
↪job_title_Applied_Data_Scientist,␣
↪job_title_Applied_Machine_Learning_Scientist, job_title_BI_Data_Analyst,␣
↪job_title_Big_Data_Engineer, job_title_Business_Data_Analyst,␣
↪job_title_Cloud_Data_Engineer, job_title_Computer_Vision_Engineer,␣
↪job_title_Computer_Vision_Software_Engineer, job_title_Data_Analyst,␣
↪job_title_Data_Analytics_Engineer, job_title_Data_Analytics_Manager,␣
↪job_title_Data_Architect, job_title_Data_Engineer,␣
↪job_title_Data_Engineering_Manager, job_title_Data_Science_Consultant,␣
↪job_title_Data_Science_Engineer, job_title_Data_Science_Manager,␣
↪job_title_Data_Scientist, job_title_Data_Specialist,␣
↪job_title_Director_of_Data_Engineering, job_title_Director_of_Data_Science,␣
↪job_title_Finance_Data_Analyst, job_title_Financial_Data_Analyst,␣
↪job_title_Head_of_Data, job_title_Head_of_Data_Science,␣
↪job_title_Lead_Data_Analyst, job_title_Lead_Data_Engineer,␣
↪job_title_Lead_Data_Scientist, job_title_ML_Engineer,␣
↪job_title_Machine_Learning_Engineer, job_title_Machine_Learning_Scientist,␣
↪job_title_Manager_Data_Science, job_title_Principal_Data_Analyst,␣
↪job_title_Principal_Data_Scientist, job_title_Product_Data_Analyst,␣
↪job_title_Research_Scientist, employee_residence_AT, employee_residence_BE,␣
↪employee_residence_BG, employee_residence_BR, employee_residence_CA,␣
↪employee_residence_CL, employee_residence_CN, employee_residence_CO,␣
↪employee_residence_DE, employee_residence_DK, employee_residence_ES,␣
↪employee_residence_FR, employee_residence_GB, employee_residence_GR,␣
↪employee_residence_HK, employee_residence_HR, employee_residence_HU,␣
↪employee_residence_IN, employee_residence_IR, employee_residence_IT,␣
↪employee_residence_JE, employee_residence_JP, employee_residence_KE,␣
↪employee_residence_LU, employee_residence_MD, employee_residence_MT,␣
↪employee_residence_MX, employee_residence_NG, employee_residence_NL,␣
↪employee_residence_NZ, employee_residence_PH, employee_residence_PK,␣
↪employee_residence_PL, employee_residence_PR, employee_residence_PT,␣
↪employee_residence_RO, employee_residence_RS, employee_residence_RU,␣
↪employee_residence_SG, employee_residence_SI, employee_residence_TR,␣
↪employee_residence_UA, employee_residence_US, employee_residence_VN,␣
↪company_location_AS, company_location_AT, company_location_BE,␣
↪company_location_BR, company_location_CA, company_location_CH,␣
↪company_location_CL, company_location_CN, company_location_CO,␣
↪company_location_DE, company_location_DK, …]
Index: []

[0 rows x 138 columns]
```

[41]:

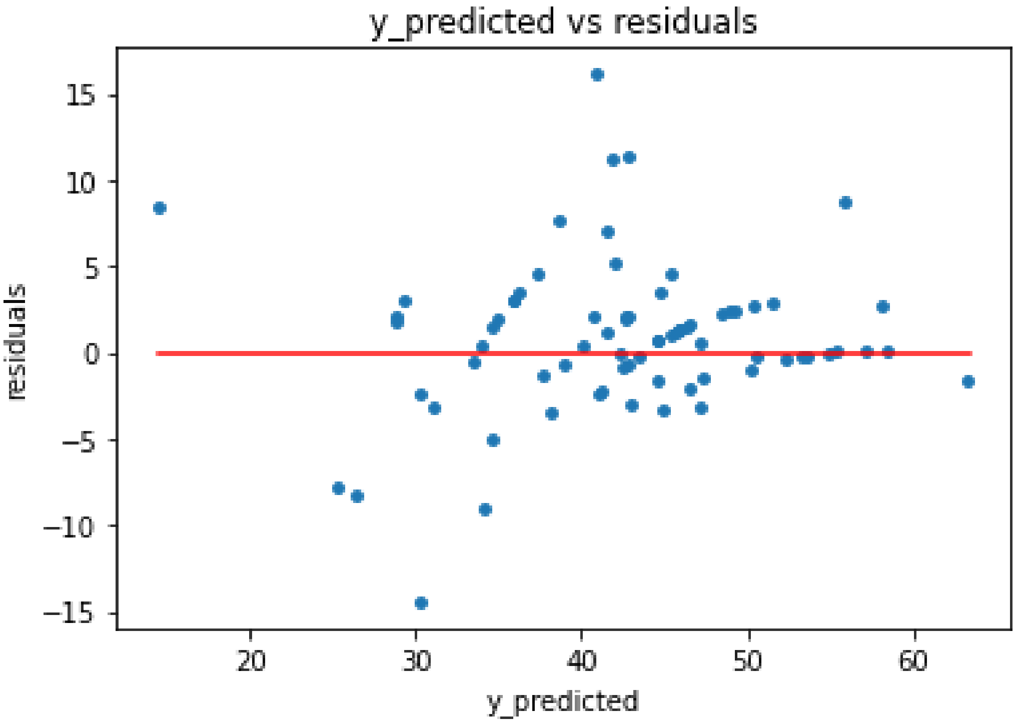
```
x_min = np.min(y_pred)
x_max = np.max(y_pred)
y_min = np.min(df_test_y_actual)
```

```
y_max = np.max(df_test_y_actual)

x = np.arange(x_min, x_max + 1)
y = np.arange(y_min, y_max + 1)


plt.scatter(x=y_pred, y=df_test_y_actual, s=15)
plt.plot(x, y, color='red')  # y = x
plt.xlabel('y_predicted')
plt.ylabel('y_actual')
plt.title('y_predicted vs y_actual')
plt.show();
```



[42]:
```
residuals = df_test_y_actual - y_pred

x_min = np.min(y_pred)
x_max = np.max(y_pred)
y_min = np.min(residuals)
y_max = np.max(residuals)

x = np.arange(x_min, x_max + 1)
y = np.zeros(len(x))
```

```
plt.scatter(x=y_pred, y=residuals, s=15)
plt.plot(x, y, color='red')  # y = x
plt.xlabel('y_predicted')
plt.ylabel('residuals')
plt.title('y_predicted vs residuals')
plt.show();
```



[43]:
```
error = df_test_y_actual - y_pred
```