REGRESSION ANALYSIS

Aim: To develop a predictive model for data science salaries using multiple regression analysis, examining factors such as experience level, employment type, and work model.

IMPORTING LIBRARIES

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
import seaborn as sns
```

LOADING THE DATASET

```
data = pd.read_csv('data_science_salaries.csv')
data
```

```
                     job_title experience_level employment_type work_models  \
0                Data Engineer        Mid-level       Full-time      Remote
1                Data Engineer        Mid-level       Full-time      Remote
2               Data Scientist     Senior-level       Full-time      Remote
3               Data Scientist     Senior-level       Full-time      Remote
4                 BI Developer        Mid-level       Full-time     On-site
...                        ...              ...             ...         ...
6594        Staff Data Analyst      Entry-level        Contract      Hybrid
6595        Staff Data Analyst  Executive-level       Full-time     On-site
6596  Machine Learning Manager     Senior-level       Full-time      Hybrid
6597             Data Engineer        Mid-level       Full-time      Hybrid
6598            Data Scientist     Senior-level       Full-time     On-site

      work_year employee_residence  salary salary_currency  salary_in_usd  \
0          2024      United States  148100             USD         148100
1          2024      United States   98700             USD          98700
2          2024      United States  140032             USD         140032
3          2024      United States  100022             USD         100022
4          2024      United States  120000             USD         120000
...         ...                ...     ...             ...            ...
6594       2020             Canada   60000             CAD          44753
6595       2020            Nigeria   15000             USD          15000
6596       2020             Canada  157000             CAD         117104
6597       2020            Austria   65000             EUR          74130
6598       2020            Austria   80000             EUR          91237
```

REGRESSION ANALYSIS

```
company_location company_size
0       United States       Medium
1       United States       Medium
2       United States       Medium
3       United States       Medium
4       United States       Medium
...               ...          ...
6594           Canada        Large
6595           Canada       Medium
6596           Canada        Large
6597          Austria        Large
6598          Austria        Small

[6599 rows x 11 columns]
```

BASIC INFORMATION ABOUT THE DATASET

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6599 entries, 0 to 6598
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   job_title           6599 non-null   object
 1   experience_level    6599 non-null   object
 2   employment_type     6599 non-null   object
 3   work_models         6599 non-null   object
 4   work_year           6599 non-null   int64
 5   employee_residence  6599 non-null   object
 6   salary              6599 non-null   int64
 7   salary_currency     6599 non-null   object
 8   salary_in_usd       6599 non-null   int64
 9   company_location    6599 non-null   object
 10  company_size        6599 non-null   object
dtypes: int64(3), object(8)
memory usage: 567.2+ KB
```

REGRESSION ANALYSIS

UNDERSTANDING THE STRUCTURE AND FORMAT OF THE DATA

```
data.head()
```

```
        job_title experience_level employment_type work_models  work_year  \
0   Data Engineer        Mid-level       Full-time      Remote       2024
1   Data Engineer        Mid-level       Full-time      Remote       2024
2  Data Scientist     Senior-level       Full-time      Remote       2024
3  Data Scientist     Senior-level       Full-time      Remote       2024
4    BI Developer        Mid-level       Full-time     On-site       2024

  employee_residence  salary salary_currency  salary_in_usd company_location
\
0      United States  148100             USD         148100    United States
1      United States   98700             USD          98700    United States
2      United States  140032             USD         140032    United States
3      United States  100022             USD         100022    United States
4      United States  120000             USD         120000    United States

  company_size
0       Medium
1       Medium
2       Medium
3       Medium
4       Medium
```

UNIVARIATE ANALYSIS

NUMERICAL VARIABLES:

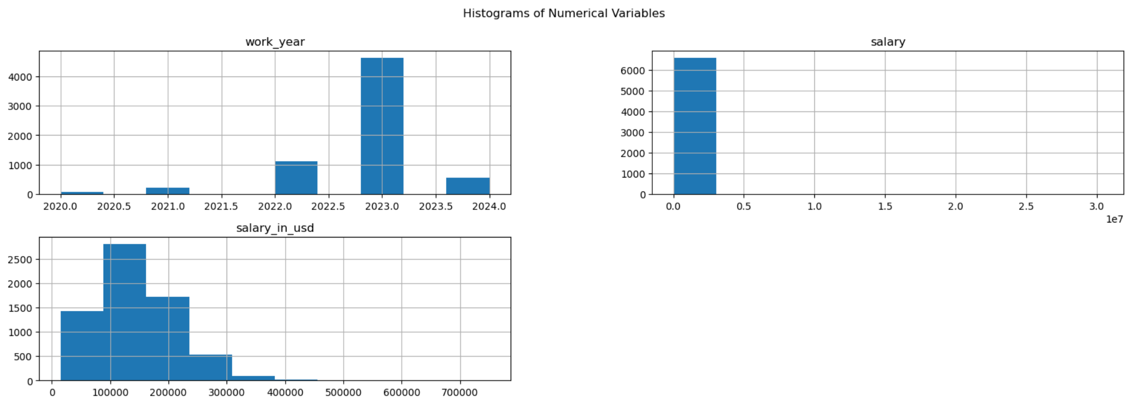
```
data.describe()
```

```
         work_year        salary  salary_in_usd
count  6599.000000  6.599000e+03    6599.000000
mean   2022.818457  1.792833e+05  145560.558569
std       0.674809  5.263722e+05   70946.838070
min    2020.000000  1.400000e+04   15000.000000
25%    2023.000000  9.600000e+04   95000.000000
50%    2023.000000  1.400000e+05  138666.000000
75%    2023.000000  1.875000e+05  185000.000000
max    2024.000000  3.040000e+07  750000.000000
```

REGRESSION ANALYSIS

VISUALIZATION:

```
data.hist(figsize=(20, 6))
plt.suptitle('Histograms of Numerical Variables')
plt.show()
```



CATEGORICAL VARIABLES:

FREQUENCY TABLES SHOWING COUNTS AND PERCENTAGES

```
for column in data.select_dtypes(include='object').columns:
    print(f"Frequency table for {column}:")
    print(data[column].value_counts())
    print("\nPercentage table for {column}:")
    print(data[column].value_counts(normalize=True) * 100)
    print("\n")
```

```
Frequency table for job_title:
job_title
Data Engineer                1307
Data Scientist               1243
Data Analyst                  910
Machine Learning Engineer     629
Analytics Engineer            246
                             ...
Deep Learning Researcher        1
Power BI Developer              1
Marketing Data Scientist        1
AI Product Manager              1
Sales Data Analyst              1
Name: count, Length: 132, dtype: int64

Percentage table for {column}:
job_title
Data Engineer                19.806031
Data Scientist               18.836187
Data Analyst                 13.789968
```

REGRESSION ANALYSIS

```
Machine Learning Engineer     9.531747
Analytics Engineer            3.727838
                                ...
Deep Learning Researcher      0.015154
Power BI Developer            0.015154
Marketing Data Scientist      0.015154
AI Product Manager            0.015154
Sales Data Analyst            0.015154
Name: proportion, Length: 132, dtype: float64


Frequency table for experience_level:
experience_level
Senior-level       4105
Mid-level          1675
Entry-level         565
Executive-level     254
Name: count, dtype: int64

Percentage table for {column}:
experience_level
Senior-level       62.206395
Mid-level          25.382634
Entry-level         8.561903
Executive-level     3.849068
Name: proportion, dtype: float64


Frequency table for employment_type:
employment_type
Full-time    6552
Contract       19
Part-time      16
Freelance      12
Name: count, dtype: int64

Percentage table for {column}:
employment_type
Full-time    99.287771
Contract      0.287922
Part-time     0.242461
Freelance     0.181846
Name: proportion, dtype: float64


Frequency table for work_models:
work_models
On-site    3813
Remote     2561
```

REGRESSION ANALYSIS

```
Hybrid      225
Name: count, dtype: int64

Percentage table for {column}:
work_models
On-site    57.781482
Remote     38.808910
Hybrid      3.409608
Name: proportion, dtype: float64


Frequency table for employee_residence:
employee_residence
United States     5305
United Kingdom     401
Canada             241
Germany             71
India               70
                  ...
Georgia              1
Israel               1
Qatar                1
Peru                 1
Honduras             1
Name: count, Length: 87, dtype: int64

Percentage table for {column}:
employee_residence
United States     80.390968
United Kingdom     6.076678
Canada             3.652068
Germany            1.075921
India              1.060767
                     ...
Georgia            0.015154
Israel             0.015154
Qatar              0.015154
Peru               0.015154
Honduras           0.015154
Name: proportion, Length: 87, dtype: float64


Frequency table for salary_currency:
salary_currency
USD    5827
GBP     334
EUR     292
INR      51
CAD      39
```

REGRESSION ANALYSIS

```
AUD      11
PLN       7
SGD       6
CHF       5
JPY       4
BRL       4
DKK       3
HUF       3
TRY       3
NOK       2
THB       2
CLP       1
ILS       1
HKD       1
PHP       1
ZAR       1
MXN       1
Name: count, dtype: int64

Percentage table for {column}:
salary_currency
USD    88.301258
GBP     5.061373
EUR     4.424913
INR     0.772844
CAD     0.590999
AUD     0.166692
PLN     0.106077
SGD     0.090923
CHF     0.075769
JPY     0.060615
BRL     0.060615
DKK     0.045461
HUF     0.045461
TRY     0.045461
NOK     0.030308
THB     0.030308
CLP     0.015154
ILS     0.015154
HKD     0.015154
PHP     0.015154
ZAR     0.015154
MXN     0.015154
Name: proportion, dtype: float64


Frequency table for company_location:
company_location
United States             5354
United Kingdom             408
```

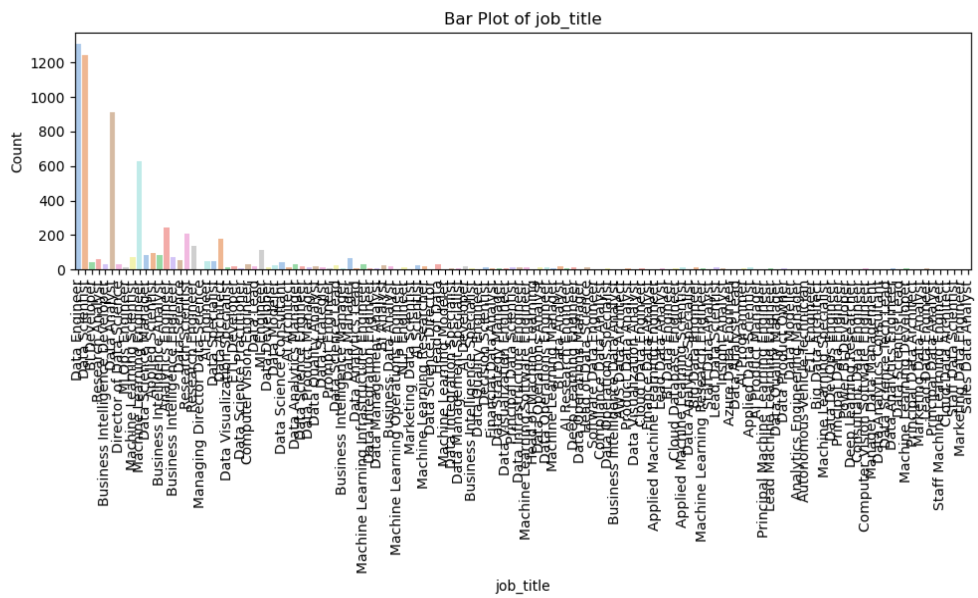
```
Canada                       243
Germany                       78
Spain                         63
                           ...
Armenia                        1
Bosnia and Herzegovina         1
Qatar                          1
Ecuador                        1
Honduras                       1
Name: count, Length: 75, dtype: int64

Percentage table for {column}:
company_location
United States              81.133505
United Kingdom              6.182755
Canada                      3.682376
Germany                     1.181997
Spain                       0.954690
                             ...
Armenia                     0.015154
Bosnia and Herzegovina      0.015154
Qatar                       0.015154
Ecuador                     0.015154
Honduras                    0.015154
Name: proportion, Length: 75, dtype: float64


Frequency table for company_size:
company_size
Medium    5860
Large      569
Small      170
Name: count, dtype: int64

Percentage table for {column}:
company_size
Medium    88.801334
Large      8.622519
Small      2.576148
Name: proportion, dtype: float64
```

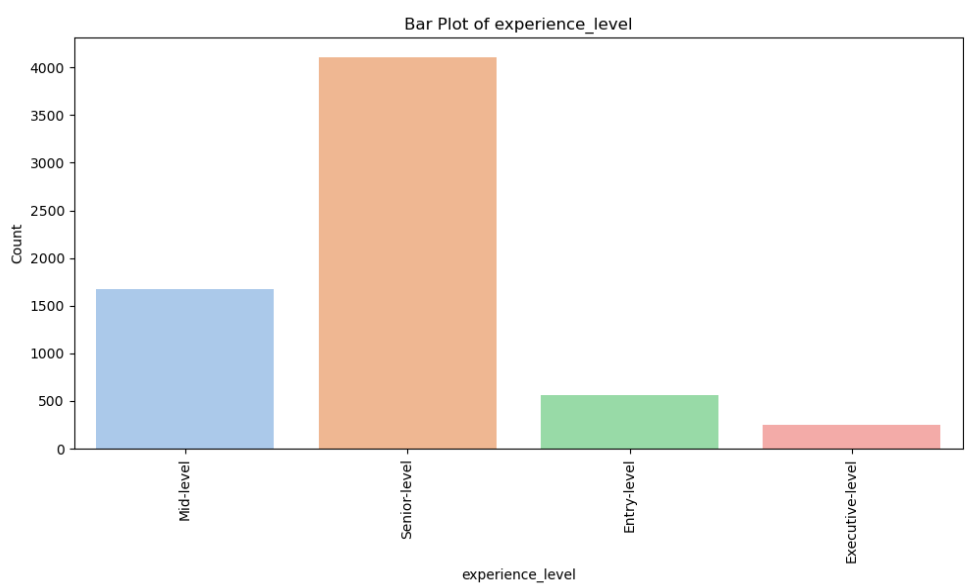
REGRESSION ANALYSIS

VISUALIZATION:

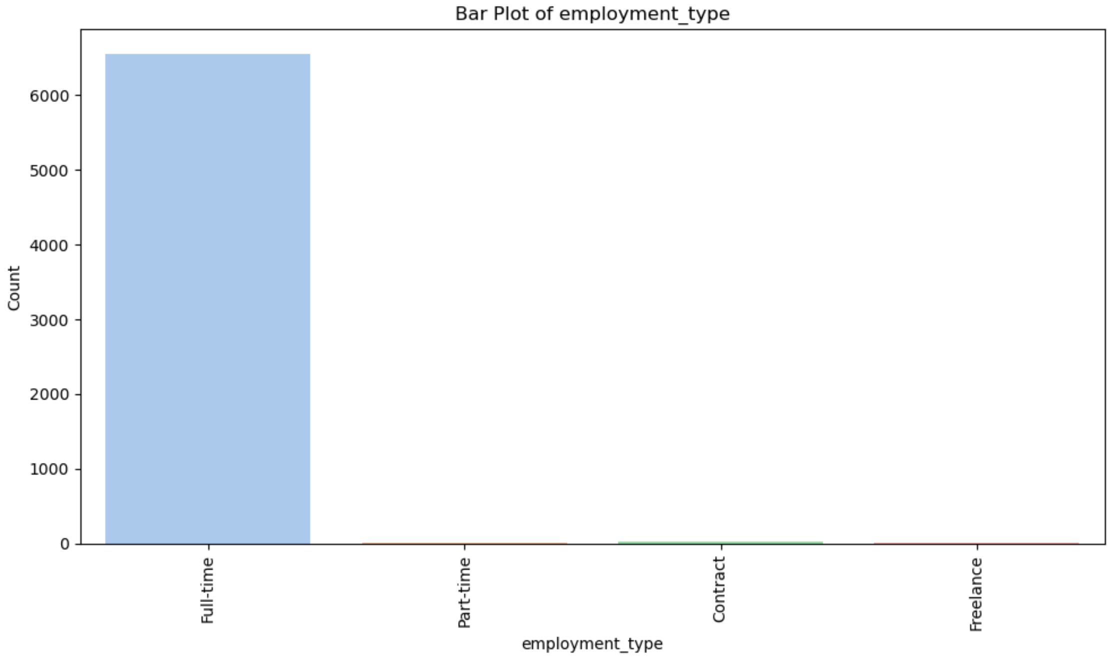
```
plt.figure(figsize=(100, 20))
for column in data.select_dtypes(include='object').columns:
    plt.figure(figsize=(10, 6))
    sns.countplot(data=data, x=column, palette='pastel')
    plt.xlabel(column)
    plt.ylabel('Count')
    plt.title(f'Bar Plot of {column}')
    plt.xticks(rotation=90)
    plt.tight_layout()
    plt.show()
```

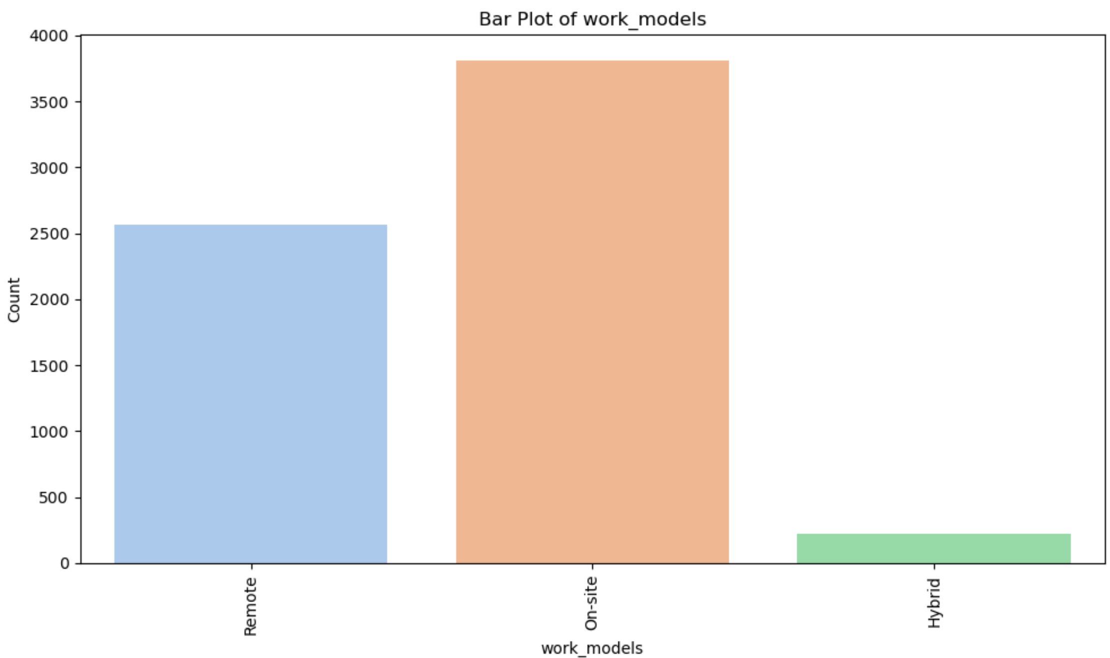
```
<Figure size 10000x2000 with 0 Axes>
```



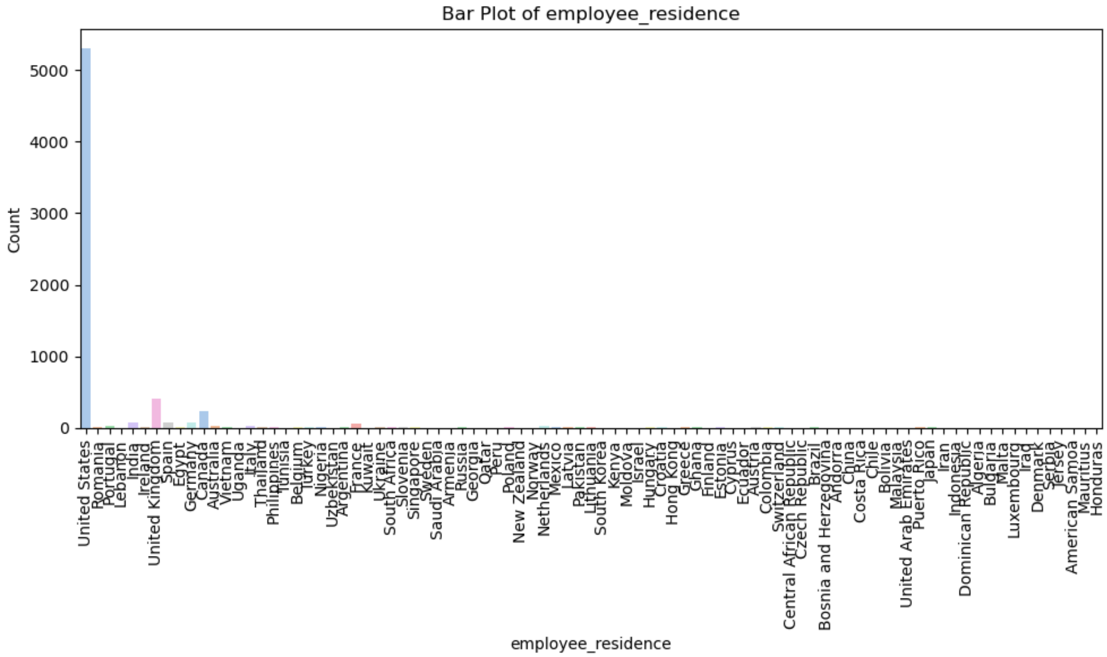


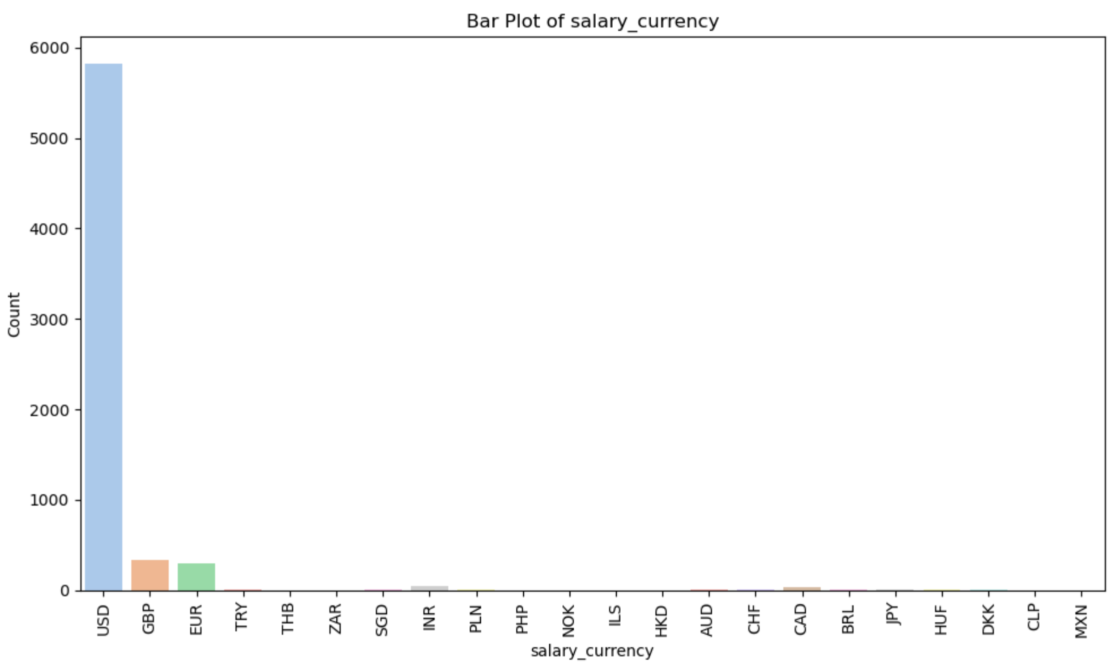
REGRESSION ANALYSIS



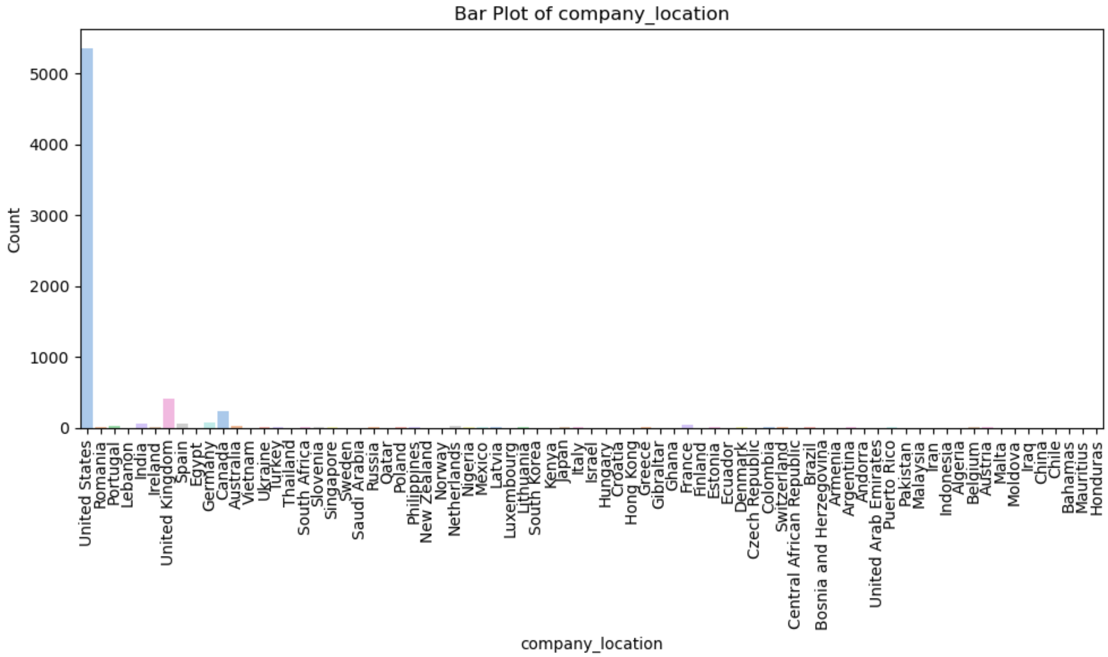


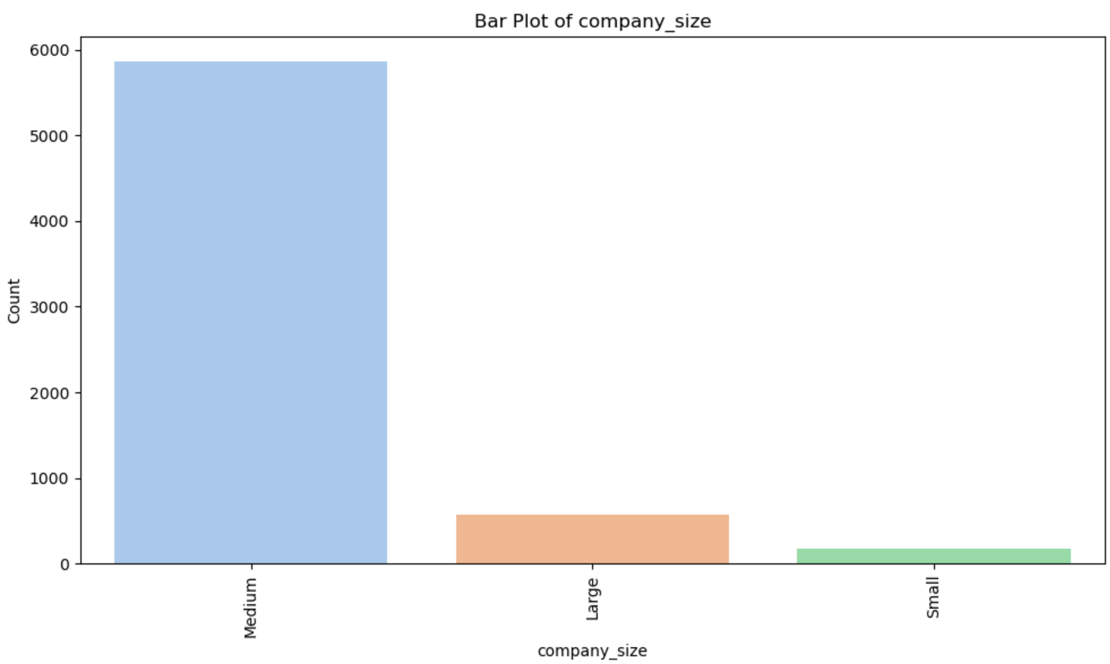
REGRESSION ANALYSIS





REGRESSION ANALYSIS



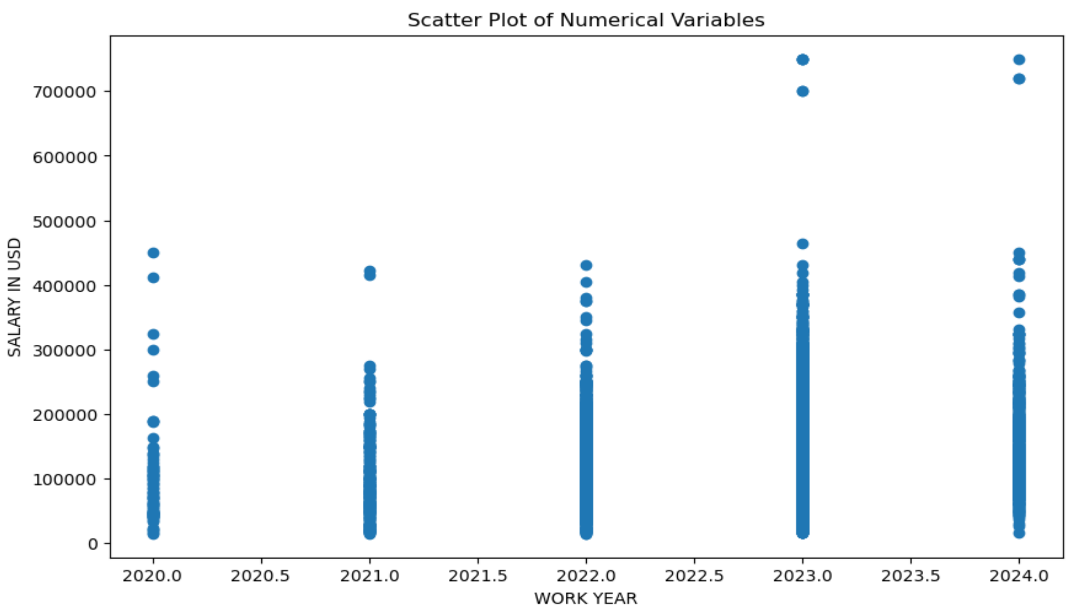


REGRESSION ANALYSIS

BIVARIATE ANALYSIS

RELATIONSHIPS BETWEEN PAIRS OF NUMERICAL VARIABLES USING SCATTER PLOTS:

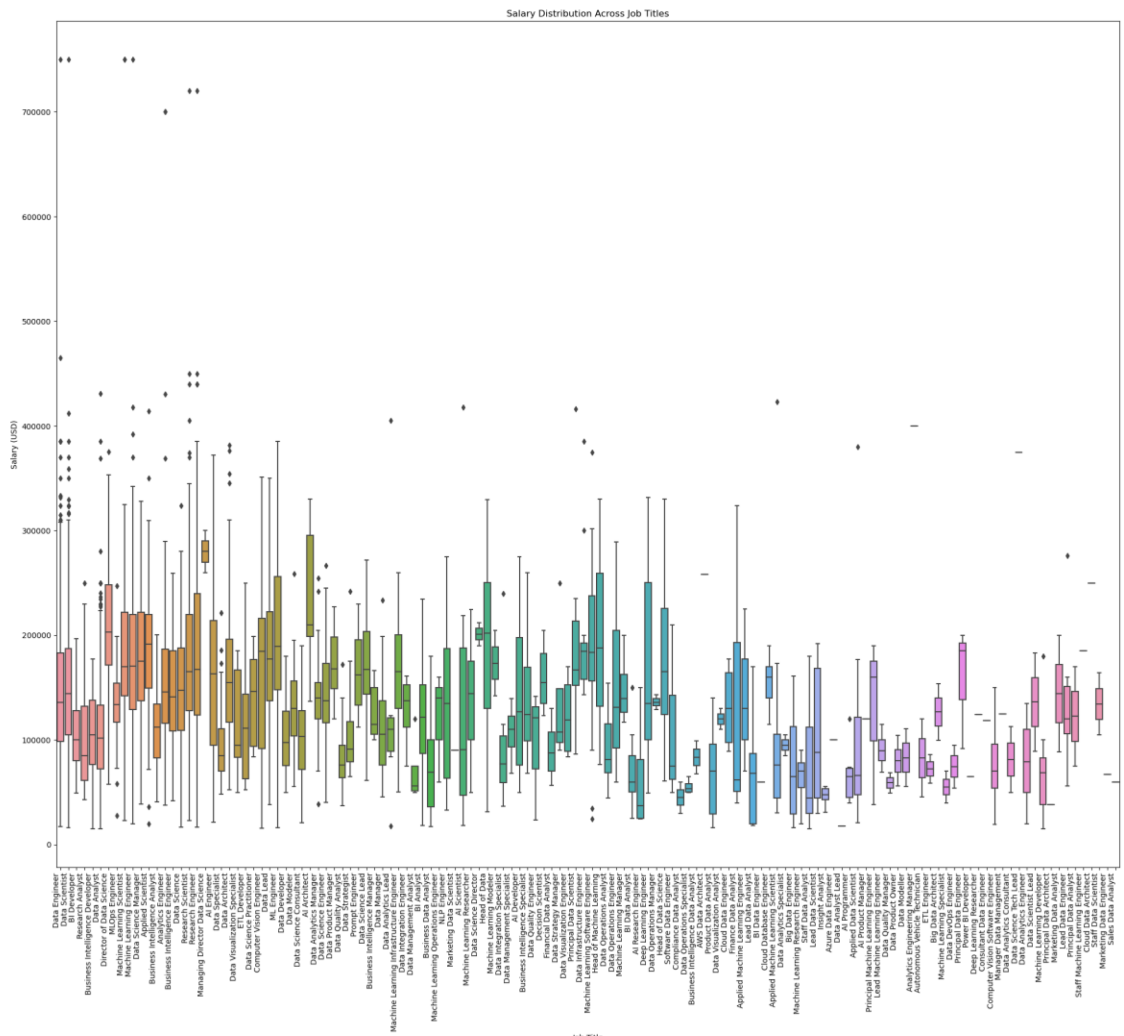
```
plt.figure(figsize=(10, 6))
plt.scatter(data['work_year'], data['salary_in_usd'])
plt.xlabel('WORK YEAR')
plt.ylabel('SALARY IN USD')
plt.title('Scatter Plot of Numerical Variables')
plt.show()
```



REGRESSION ANALYSIS

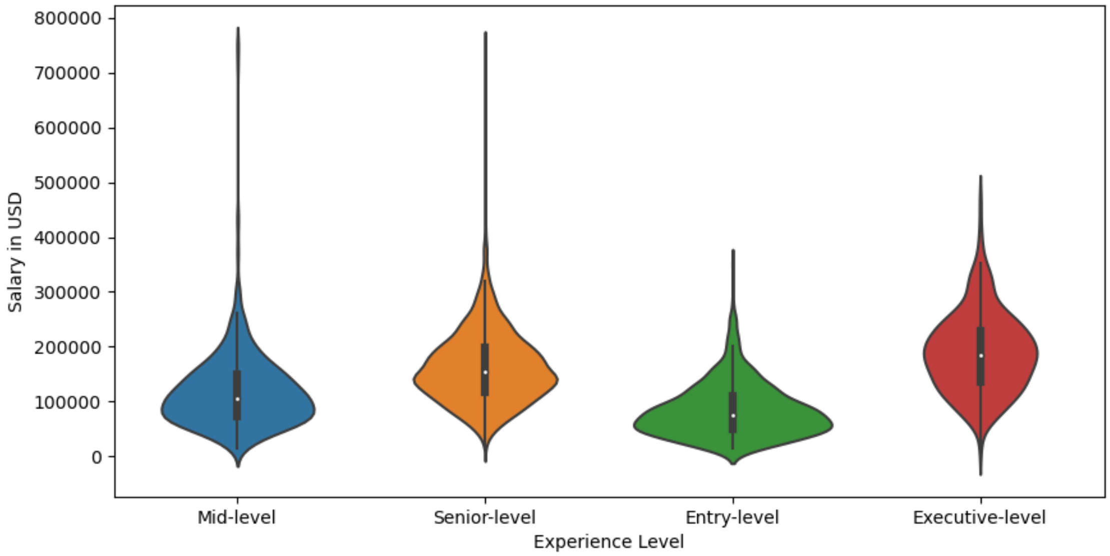
RELATIONSHIPS BETWEEN PAIRS OF NUMERICAL AND CATEGORICAL VARIABLES USING VIOLIN PLOTS:

```
plt.figure(figsize=(25, 20))
sns.boxplot(x='job_title', y='salary_in_usd', data=data)
plt.title('Salary Distribution Across Job Titles')
plt.xlabel('Job Title')
plt.ylabel('Salary (USD)')
plt.xticks(rotation=90, ha='right')
plt.show()
```



REGRESSION ANALYSIS

```
plt.figure(figsize=(10, 5))
sns.violinplot(data=data, x='experience_level', y='salary_in_usd')
plt.xlabel('Experience Level')
plt.ylabel('Salary in USD')
plt.show()
```



REPLACING STRING VALUES WITH FLOAT

```
data['experience_level'] = data['experience_level'].replace({'Mid-level':
1.0, 'Senior-level': 2.0, 'Entry-level': 3.0, 'Executive-level': 4.0})
data['employment_type'] = data['employment_type'].replace({'Full-time':
1.0,'Part-time':2.0,'Contract':3.0,'Freelance':4.0})
data['work_models'] = data['work_models'].replace({
    'Remote': 1.0,
    'Hybrid': 2.0,
    'On-site': 3.0
})
```

CORRELATION COEFFICIENTS

```
correlation_matrix = data[['experience_level', 'salary_in_usd']].corr()
print("Correlation Matrix:", correlation_matrix)
```

```
Correlation Matrix:                   experience_level  salary_in_usd
experience_level          1.000000       0.099199
salary_in_usd             0.099199       1.000000
```

REGRESSION ANALYSIS

MULTIPLE REGRESSION

```
X = data[['experience_level','employment_type','work_models']]
y = data['salary_in_usd']
```

SPLITTING THE DATA INTO TRAINING AND TESTING SETS

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

CREATING A LINEAR REGRESSION MODEL

```
model = LinearRegression()
```

TRAINING THE MODEL

```
model.fit(X_train, y_train)
```

```
LinearRegression()
```

MAKING PREDICTIONS

```
y_pred = model.predict(X_test)
```

CALCULATING MEAN SQUARED ERROR

```
mse = mean_squared_error(y_test, y_pred)
mse
```

```
5712527799.890261
```

CALCULATING RMSE

```
rmse = np.sqrt(mse)
```

CALCULATING R-SQUARED

```
r2 = r2_score(y_test, y_pred)
r2
```

```
0.02798586412343318
```

COEFFICIENTS AND INTERCEPT

```
print("Coefficients:", model.coef_)
print("Intercept:", model.intercept_)
```

```
Coefficients: [  9260.0493725  -30368.28929475   5679.30589811]
Intercept: 146201.23742183286
```

REGRESSION ANALYSIS

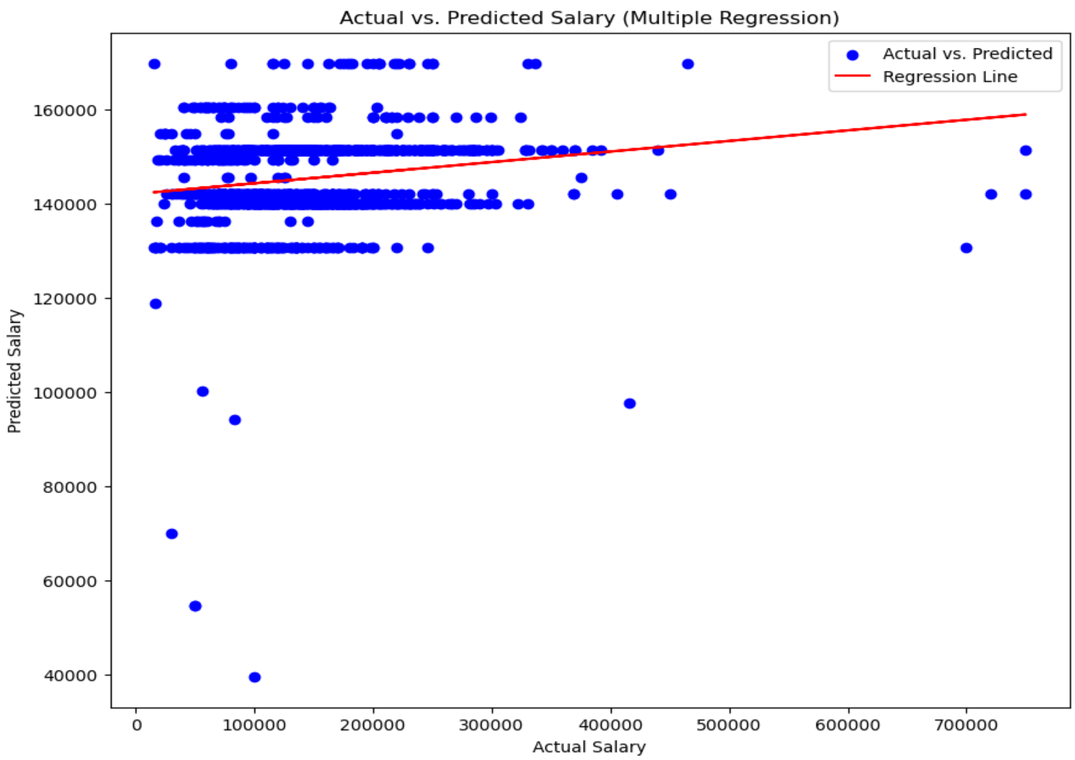
```
print("The predicted values for",X_test," are:",y_pred)
```

```
The predicted values for       experience_level  employment_type  work_models
2098                 2.0              1.0             1.0
5338                 2.0              1.0             1.0
4407                 2.0              1.0             3.0
6108                 1.0              1.0             1.0
4025                 2.0              1.0             1.0
...                  ...              ...             ...
3268                 4.0              1.0             1.0
3949                 3.0              1.0             2.0
3077                 1.0              1.0             3.0
3767                 2.0              1.0             3.0
4942                 2.0              1.0             3.0

[1320 rows x 3 columns]  are: [140032.35277019 140032.35277019
151390.96456642 ... 142130.91519392
 151390.96456642 151390.96456642]
```

PLOTTING THE DATA AND THE REGRESSION CURVE

```
# Plot the regression curve
plt.figure(figsize=(10, 8))
plt.scatter(y_test, y_pred, color='blue', label='Actual vs. Predicted')

# Add a trendline
z = np.polyfit(y_test, y_pred, 1)
p = np.poly1d(z)
plt.plot(y_test, p(y_test), color='red', label='Regression Line')

plt.title('Actual vs. Predicted Salary (Multiple Regression)')
plt.xlabel('Actual Salary')
plt.ylabel('Predicted Salary')
plt.legend()
plt.show()
```

REGRESSION ANALYSIS



Conclusion: We used multiple regression model in the dataset to see how different things like experience level, employment type, and work models affect salaries in data science. The study successfully establishes key associations between experience level, employment type, work model, and salary levels, highlighting the significance of multiple regression techniques in understanding salary dynamics within the data science field.