```
## TASK 5 EMPLOYEE SALARIES FOR DIFFERENT JOB ROLS
```

```
import pandas as pd

# Load the dataset
salary_data = pd.read_excel('/content/ds_salaries.xlsx')

# Display basic information about the dataset
salary_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 607 entries, 0 to 606
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Unnamed: 0          607 non-null    int64 
 1   work_year           607 non-null    int64 
 2   experience_level    607 non-null    object
 3   employment_type     607 non-null    object
 4   job_title           607 non-null    object
 5   salary              607 non-null    int64 
 6   salary_currency     607 non-null    object
 7   salary_in_usd       607 non-null    int64 
 8   employee_residence  607 non-null    object
 9   remote_ratio        607 non-null    int64 
 10  company_location    607 non-null    object
 11  company_size        607 non-null    object
dtypes: int64(5), object(7)
memory usage: 57.0+ KB
```

```
# Display basic statistics for numerical columns
salary_data.describe()
```

	Unnamed: 0	work_year	salary	salary_in_usd	remote_ratio
count	607.000000	607.000000	6.070000e+02	607.000000	607.00000
mean	303.000000	2021.405272	3.240001e+05	112297.869852	70.92257
std	175.370085	0.692133	1.544357e+06	70957.259411	40.70913
min	0.000000	2020.000000	4.000000e+03	2859.000000	0.00000
25%	151.500000	2021.000000	7.000000e+04	62726.000000	50.00000
50%	303.000000	2022.000000	1.150000e+05	101570.000000	100.00000
75%	454.500000	2022.000000	1.650000e+05	150000.000000	100.00000
max	606.000000	2022.000000	3.040000e+07	600000.000000	100.00000

```
numerical_columns = salary_data.select_dtypes(include=['float64', 'int64']).columns
descriptive_stats = salary_data[numerical_columns].agg(['mean', 'median', 'std'])
print(descriptive_stats)
```

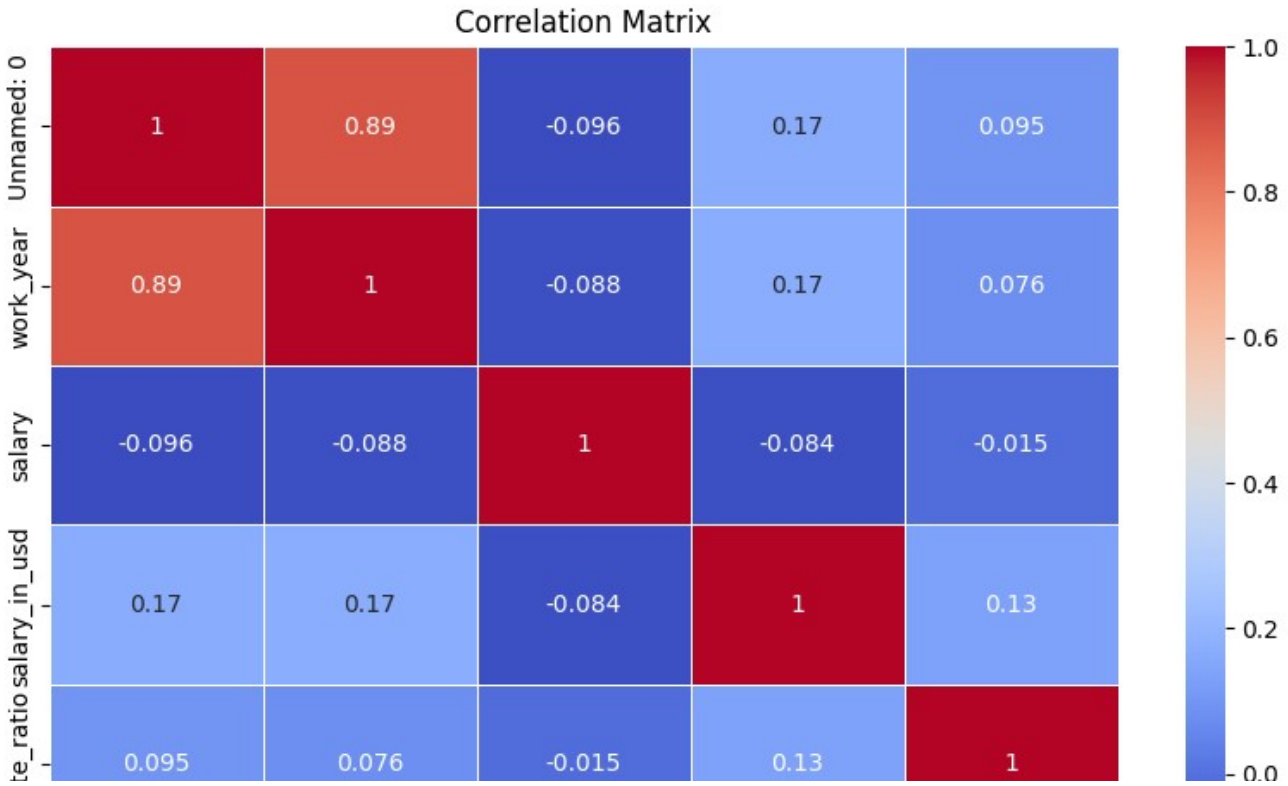
```
        Unnamed: 0    work_year        salary  salary_in_usd  remote_ratio
mean    303.000000  2021.405272  3.240001e+05  112297.869852      70.92257
median  303.000000  2022.000000  1.150000e+05  101570.000000     100.00000
std     175.370085     0.692133  1.544357e+06   70957.259411      40.70913
```

```
import pandas as pd

# Load the dataset
salary_data = pd.read_excel('/content/ds_salaries.xlsx')
import seaborn as sns
import matplotlib.pyplot as plt

# Compute the correlation matrix
correlation_matrix = salary_data[numerical_columns].corr()

# Plot the heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Matrix')
plt.show()
```




```
import pandas as pd

# Load the dataset
file_path = 'your_file_path_here.xlsx'
salary_data = pd.read_excel('/content/ds_salaries.xlsx')

# Find the highest salary
highest_salary = salary_data['salary'].max()
highest_salary_details = salary_data[salary_data['salary'] == highest_salary] # Changed '

# Find the lowest salary
lowest_salary = salary_data['salary'].min()
lowest_salary_details = salary_data[salary_data['salary'] == lowest_salary] # Changed 'Sa

# Print the results
print("Highest salary:")
print(highest_salary_details)

print("\nLowest salary:")
print(lowest_salary_details)
```

```
Highest salary:
     Unnamed: 0  work_year experience_level employment_type       job_title  \
177         177       2021               MI              FT  Data Scientist

       salary salary_currency  salary_in_usd employee_residence  remote_ratio  \
177  30400000             CLP          40038                 CL           100

    company_location company_size
177               CL            L

Lowest salary:
     Unnamed: 0  work_year experience_level employment_type       job_title  \
185         185       2021               MI              FT   Data Engineer
238         238       2021               EN              FT  Data Scientist

     salary salary_currency  salary_in_usd employee_residence  remote_ratio  \
185    4000             USD           4000                 IR           100
238    4000             USD           4000                 VN             0

    company_location company_size
185               IR            M
```

```
     238              VN            M
```

```
import pandas as pd

# Load the dataset

salary_data = pd.read_excel('/content/ds_salaries.xlsx')

# Check for the correct column name (case-sensitive)
print(salary_data.columns)

# Assuming the correct column name is 'work_year', proceed with the conversion:
salary_data['work_year'] = pd.to_datetime(salary_data['work_year']) # Changed 'Date' to '

# Plot salary trends over time
plt.figure(figsize=(14, 8))
sns.lineplot(x='work_year', y='salary', data=salary_data) # Changed 'Date' to 'work_year'
plt.title('Salary Trends Over Time')
plt.xlabel('Year') # Changed label to 'Year'
plt.ylabel('Salary')
plt.show()
```

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