## Data Set

The chosen data set revolves around the salaries of data scientists in various

```
filepath = '/content/ds_salaries.csv'
```

```
import pandas as pd
import numpy as np
import statistics
```

## Original Data Set

```
data = pd.read_csv(filepath)
data
```

| | id | work_year | experience_level | employment_type | job_title | salary | salary_currency | salary_in_usd | employee_residence | remote_ra |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 2020 | MI | FT | Data Scientist | 70000 | EUR | 79833 | DE | |
| **1** | 1 | 2020 | SE | FT | Machine Learning Scientist | 260000 | USD | 260000 | JP | |
| **2** | 2 | 2020 | SE | FT | Big Data Engineer | 85000 | GBP | 109024 | GB | |
| **3** | 3 | 2020 | MI | FT | Product Data Analyst | 20000 | USD | 20000 | HN | |
| **4** | 4 | 2020 | SE | FT | Machine Learning Engineer | 150000 | USD | 150000 | US | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **602** | 602 | 2022 | SE | FT | Data Engineer | 154000 | USD | 154000 | US | |
| **603** | 603 | 2022 | SE | FT | Data Engineer | 126000 | USD | 126000 | US | |
| **604** | 604 | 2022 | SE | FT | Data Analyst | 129000 | USD | 129000 | US | |
| **605** | 605 | 2022 | SE | FT | Data Analyst | 150000 | USD | 150000 | US | |
| **606** | 606 | 2022 | MI | FT | AI Scientist | 200000 | USD | 200000 | IN | |

607 rows × 12 columns

Next steps: View recommended plots

```
column_names = data.columns
print(column_names)
```

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

```
data_types = data.dtypes
print(data_types)
```

```
id                    int64
work_year             int64
experience_level     object
employment_type      object
```

```
job_title             object
salary                 int64
salary_currency       object
salary_in_usd          int64
employee_residence    object
remote_ratio           int64
company_location      object
company_size          object
dtype: object
```

```
total_records = data.shape
rows = data.shape[0]
columns = data.shape[1]
print("Total: ", total_records)
print("Rows: ", rows)
print("Columns: ", columns)
```

```
Total:  (607, 12)
Rows:  607
Columns:  12
```

```
data.describe()
```

| | id | 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 |

## Excluding columns that do not affect salary

```
df = pd.read_csv(filepath, index_col = 'id')
df.drop([ 'salary', 'salary_currency', 'employee_residence'], axis = 1, inplace = True)
df
```

| id | work_year | experience_level | employment_type | job_title | salary_in_usd | remote_rat |
|---|---|---|---|---|---|---|
| **0** | 2020 | MI | FT | Data Scientist | 79833 | |
| **1** | 2020 | SE | FT | Machine Learning Scientist | 260000 | |
| **2** | 2020 | SE | FT | Big Data Engineer | 109024 | |
| **3** | 2020 | MI | FT | Product Data Analyst | 20000 | |
| **4** | 2020 | SE | FT | Machine Learning Engineer | 150000 | |
| **...** | ... | ... | ... | ... | ... | |
| **602** | 2022 | SE | FT | Data Engineer | 154000 | 1 |
| **603** | 2022 | SE | FT | Data Engineer | 126000 | 1 |
| **604** | 2022 | SE | FT | Data Analyst | 129000 | |
| **605** | 2022 | SE | FT | Data Analyst | 150000 | 1 |
| **606** | 2022 | MI | FT | AI Scientist | 200000 | 1 |

607 rows × 8 columns

Next steps: View recommended plots

## Number of employees per experience level

EN = Entry Level

MI = Mid Level

SE = Senior Level

EX = Executive Level

```
df = pd.read_csv(filepath, index_col = 'experience_level')
df.drop([ 'salary', 'salary_currency', 'employee_residence'], axis = 1, inplace = True)
df.sort_values("experience_level")
```

| experience_level | id | work_year | employment_type | job_title | salary_in_usd | remote_rat |
|---|---|---|---|---|---|---|
| **EN** | 113 | 2021 | PT | AI Scientist | 12000 | 1 |
| **EN** | 139 | 2021 | FT | Data Scientist | 80000 | 1 |
| **EN** | 134 | 2021 | FT | Data Scientist | 100000 | |
| **EN** | 131 | 2021 | FT | Data Scientist | 49646 | |
| **EN** | 130 | 2021 | FT | Machine Learning Developer | 100000 | |
| **...** | ... | ... | ... | ... | ... | |
| **SE** | 364 | 2022 | FT | Data Engineer | 160000 | |
| **SE** | 363 | 2022 | FT | Data Analyst | 61300 | 1 |
| **SE** | 362 | 2022 | FT | Data Analyst | 130000 | 1 |
| **SE** | 371 | 2022 | FT | Machine Learning Engineer | 189650 | |
| **SE** | 303 | 2022 | FT | Data Scientist | 123000 | 1 |

607 rows × 8 columns

## Employees based on Job Title

There are 50 unique job titles

```
df = pd.read_csv(filepath, index_col = 'job_title')
df.drop([ 'salary', 'salary_currency', 'employee_residence'], axis = 1, inplace = True)
df.sort_values("job_title")
```

| job_title | id | work_year | experience_level | employment_type | salary_in_usd | remote_rat |
|---|---|---|---|---|---|---|
| **3D Computer Vision Researcher** | 77 | 2021 | MI | PT | 5409 | |
| **AI Scientist** | 606 | 2022 | MI | FT | 200000 | 1 |
| **AI Scientist** | 96 | 2021 | EN | PT | 12000 | 1 |
| **AI Scientist** | 113 | 2021 | EN | PT | 12000 | 1 |
| **AI Scientist** | 244 | 2021 | EN | FT | 18053 | 1 |
| **...** | ... | ... | ... | ... | ... | |
| **Research Scientist** | 508 | 2022 | EN | FT | 120000 | 1 |
| **Research Scientist** | 507 | 2022 | MI | FT | 64849 | |
| **Research Scientist** | 26 | 2020 | EN | FT | 42000 | |
| **Research Scientist** | 236 | 2021 | MI | FT | 63810 | 1 |
| **Staff Data Scientist** | 283 | 2021 | SE | CT | 105000 | 1 |

607 rows × 8 columns

## Employees based on employment type

```
employment_type = pd.DataFrame(df)
employment_type = df[df['employment_type'] == 'CT'].copy()
employment_type # This results into 5 employees with CT
```

| experience_level | id | work_year | employment_type | job_title | salary | salary_currency | salary_in_usd | employee_residence | remote_ratio |
|---|---|---|---|---|---|---|---|---|---|
| **EN** | 28 | 2020 | CT | Business Data Analyst | 100000 | USD | 100000 | US | 100 |
| **MI** | 78 | 2021 | CT | ML Engineer | 270000 | USD | 270000 | US | 100 |
| **EX** | 225 | 2021 | CT | Principal Data Scientist | 416000 | USD | 416000 | US | 100 |
| **SE** | 283 | 2021 | CT | Staff Data Scientist | 105000 | USD | 105000 | US | 100 |
| **EN** | 489 | 2022 | CT | Applied Machine Learning Scientist | 29000 | EUR | 31875 | TN | 100 |

Next steps: View recommended plots

```
employment_type = df[df['employment_type'] == 'PT'].copy()
employment_type # This results into 10 employees with Part Time Employment Type
```

| experience_level | id | work_year | employment_type | job_title | salary | salary_currency |
|---|---|---|---|---|---|---|
| **EN** | 45 | 2020 | PT | ML Engineer | 14000 | EUR |
| **EN** | 62 | 2020 | PT | Data Scientist | 19000 | EUR |
| **MI** | 77 | 2021 | PT | 3D Computer Vision Researcher | 400000 | INR |
| **EN** | 96 | 2021 | PT | AI Scientist | 12000 | USD |