Data Science Salary Analysis

August 31, 2023

[1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
from tabulate import tabulate
import matplotlib.pyplot as plt
```

[2]:
```
df=pd.read_csv(r"C:\Users\kazit\Downloads\Data\Data_Science_Salaries.csv")
df.head()
```

[2]:

	Job Title	Employment Type	Experience Level	Expertise Level	Salary	Salary Currency	Company Location	Salary in USD	Employee Residence	Company Size	Year
0	AI Scientist	Full-Time	Senior	Expert	60000	Euro	Germany	64781	Germany	Large	2023
1	Data Engineer	Full-Time	Mid	Intermediate	160000	United States Dollar	United States	160000	United States	Medium	2023
2	Data Engineer	Full-Time	Mid	Intermediate	140000	United States Dollar	United States	140000	United States	Medium	2023
3	Data Engineer	Full-Time	Mid	Intermediate	139152	United States Dollar	United States	139152	United States	Large	2023
4	Data Engineer	Full-Time	Mid	Intermediate	82452	United States Dollar	United States	82452	United States	Large	2023

0.0.1 Data Info

[3]:
```
df.isnull().sum()
```

[3]:
```
Job Title            0
Employment Type      0
Experience Level     0
Expertise Level      0
```

```
Salary                0
Salary Currency       0
Company Location      0
Salary in USD         0
Employee Residence    0
Company Size          0
Year                  0
dtype: int64
```

[4]:
```
df.columns
```

[4]:
```
Index(['Job Title', 'Employment Type', 'Experience Level', 'Expertise Level',
       'Salary', 'Salary Currency', 'Company Location', 'Salary in USD',
       'Employee Residence', 'Company Size', 'Year'],
      dtype='object')
```

[5]:
```
df.dtypes
```

[5]:
```
Job Title             object
Employment Type       object
Experience Level      object
Expertise Level       object
Salary                 int64
Salary Currency       object
Company Location      object
Salary in USD          int64
Employee Residence    object
Company Size          object
Year                   int64
dtype: object
```

[6]:
```
df.shape
```

[6]:
```
(3683, 11)
```

0.0.2 Data Preprocessing

[7]:
```
df['Job Title'].unique()
```

[7]:
```
array(['AI Scientist', 'Data Engineer', 'Decision Scientist',
       'Data Strategist', 'Data Analyst', 'Research Scientist',
       'Machine Learning Engineer', 'ML Engineer', 'Analytics Engineer',
       'Data Manager', 'MLOps Engineer', 'Data Science Lead',
       'Data Scientist', 'Data Specialist', 'Data Integration Specialist',
       'Data Science Consultant', 'Business Intelligence Analyst',
       'Data Science Practitioner', 'Data Management Specialist',
       'Business Intelligence Engineer', 'AI Architect', 'Head of Data',
       'Business Data Analyst', 'AI Engineer', 'Applied Scientist',
```

```
'Data Architect', 'Applied Machine Learning Scientist',
'AI Research Engineer', 'Data Modeler', 'Research Engineer',
'BI Developer', 'Machine Learning Scientist', 'Research Analyst',
'Data Analytics Lead', 'Data Operations Analyst',
'Data Operations Engineer', 'Machine Learning Manager',
'BI Data Analyst', 'Data Lead', 'Data Visualization Engineer',
'Business Intelligence Developer', 'Lead Data Scientist',
'Director of Data Science', 'Principal Machine Learning Engineer',
'Principal Data Engineer', 'Data Analytics Manager',
'Data Science Manager', 'AI Developer', 'Power BI Developer',
'Data Quality Analyst', 'Applied Data Scientist',
'Head of Data Science', 'Machine Learning Software Engineer',
'BI Analyst', 'AI Programmer', 'Computer Vision Engineer',
'Principal Data Scientist', 'Staff Machine Learning Engineer',
'Staff Data Scientist', 'Consultant Data Engineer',
'Machine Learning Specialist', 'Data Quality Engineer',
'Deep Learning Engineer', 'Data Visualization Specialist',
'Business Intelligence Data Analyst', 'Data Science Engineer',
'Data Operations Manager', 'Lead Machine Learning Engineer',
'Managing Director Data Science', 'Data Modeller',
'Finance Data Analyst', 'Software Data Engineer',
'Machine Learning Research Engineer', 'Compliance Data Analyst',
'Data Operations Specialist', 'Data Engineer 2',
'Cloud Data Engineer', 'Analytics Engineering Manager',
'Machine Learning Infrastructure Engineer', 'Insight Analyst',
'ETL Developer', 'NLP Engineer', 'Staff Data Analyst',
'AWS Data Architect', 'Product Data Analyst',
'Machine Learning Developer', 'Data Visualization Analyst',
'Autonomous Vehicle Technician', 'Sales Data Analyst',
'Applied Machine Learning Engineer', 'ETL Engineer',
'Data DevOps Engineer', 'Machine Learning Researcher',
'Big Data Engineer', 'Lead Data Analyst', 'BI Data Engineer',
'Cloud Database Engineer', 'Financial Data Analyst',
'Data Infrastructure Engineer', 'Deep Learning Researcher',
'Data Analytics Specialist', 'Big Data Architect',
'Computer Vision Software Engineer', 'Azure Data Engineer',
'Marketing Data Engineer', 'Manager Data Management',
'Data Analytics Consultant', 'Data Science Tech Lead',
'Data Scientist Lead', 'Marketing Data Analyst',
'Principal Data Architect', 'Data Analytics Engineer',
'Cloud Data Architect', 'Lead Data Engineer',
'Head of Machine Learning', 'Principal Data Analyst'], dtype=object)
```

[8]:
```
display(df['Job Title'].value_counts())
```

```
Data Engineer                   771
Data Scientist                  697
Data Analyst                    501
```

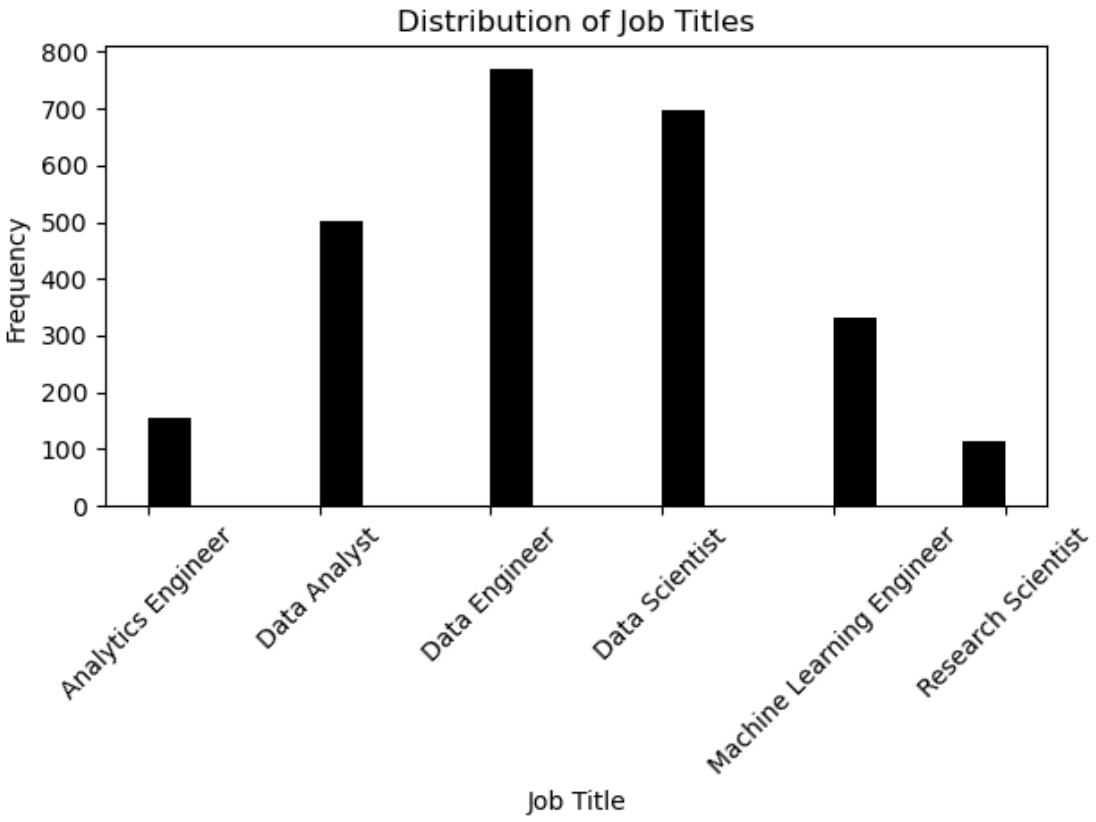
```
Machine Learning Engineer        333
Analytics Engineer               154
                                 …
Data DevOps Engineer               1
Data Engineer 2                    1
Analytics Engineering Manager      1
Sales Data Analyst                 1
Data Quality Engineer              1
Name: Job Title, Length: 116, dtype: int64
```

[9]:
```
#delet which Job Title value count less then 100

df = df[df['Job Title'].map(df['Job Title'].value_counts()) >= 100]


df['Job Title'].value_counts().tolist()
```

[9]:
```
[771, 697, 501, 333, 154, 115]
```

0.0.3 Data Viz

[10]:
```
x = df['Job Title'].sort_values()

plt.hist(x, bins=20,color='black')
plt.xticks(rotation=45)
plt.xlabel('Job Title')
plt.ylabel('Frequency')
plt.title('Distribution of Job Titles')
plt.tight_layout()
plt.show()
```



[11]:
```
employment_counts = df['Employment Type'].value_counts()
employment_counts
```

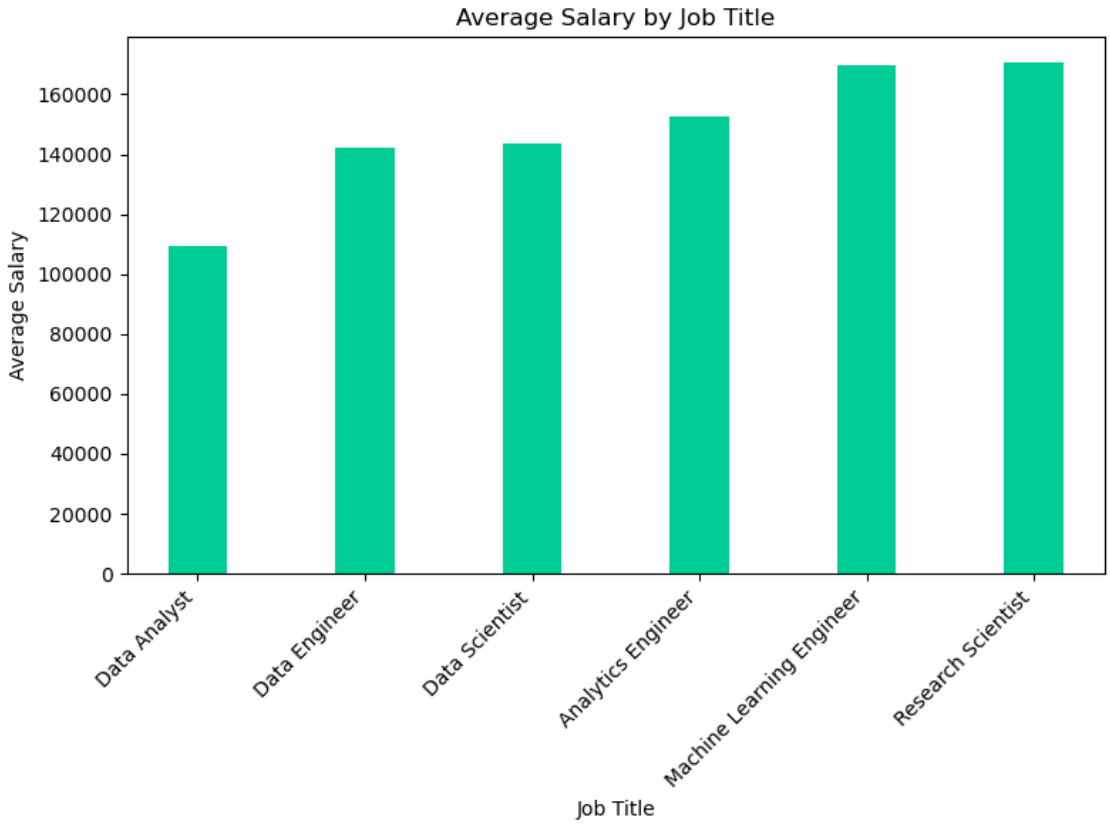
[11]:
```
Full-Time    2556
Part-Time       9
Contract        3
Freelance       3
Name: Employment Type, dtype: int64
```

[12]:
```
import plotly.graph_objects as go

labels = employment_counts.index
values = employment_counts.values

# Use `hole` to create a donut-like pie chart
fig = go.Figure(data=[go.Pie(labels=labels, values=values, hole=0.7)])
fig.show()
```

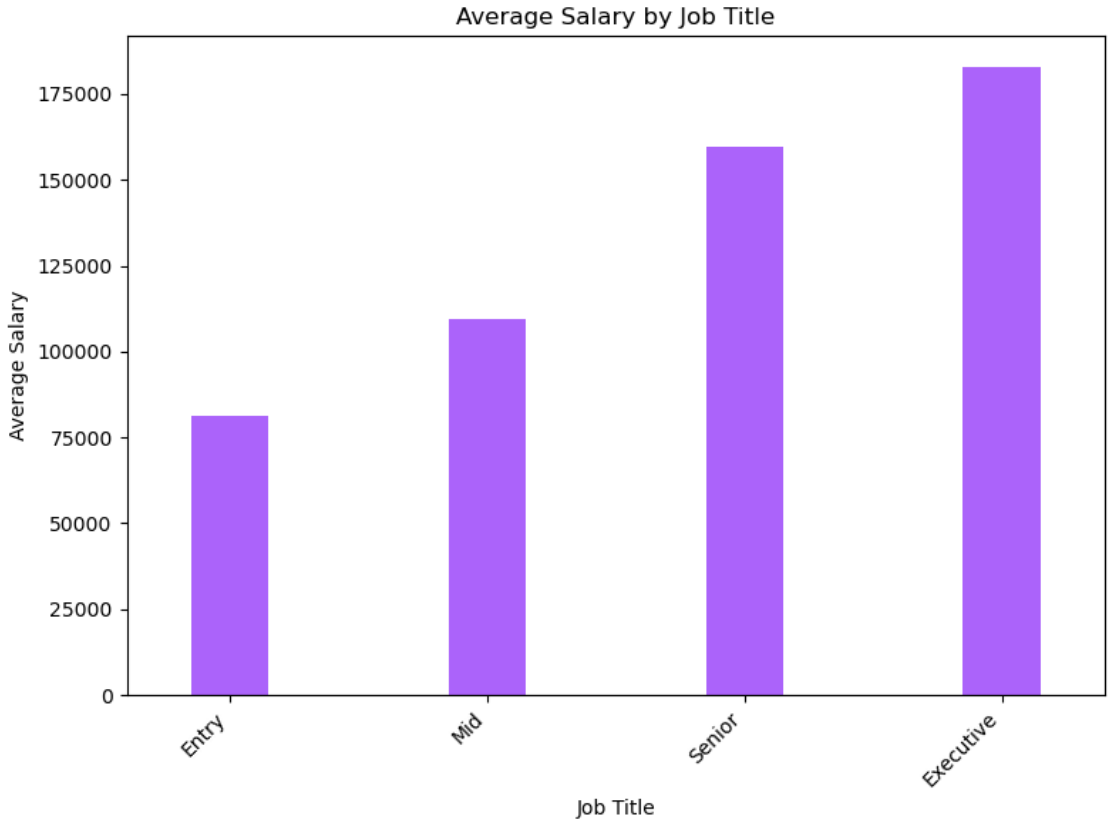
[13]:
```
 Experience_Level=df['Experience Level'].value_counts()
```

[14]:
```
import plotly.graph_objects as go

labels = Experience_Level.index
values = Experience_Level.values

# Use `hole` to create a donut-like pie chart
fig = go.Figure(data=[go.Pie(labels=labels, values=values, hole=0.7)])
fig.show()
```

[15]:
```
# Group by 'Job Title' and calculate the average salary
average_salary_by_title = df.groupby('Job Title')['Salary in USD'].mean().
 ↪sort_values(ascending=True)

# Create a bar plot
plt.figure(figsize=(8, 6))
average_salary_by_title.plot(kind='bar', color='#00cc96',width=0.35)
plt.xlabel('Job Title')
plt.ylabel('Average Salary')
plt.title('Average Salary by Job Title')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



[16]:
```
# Group by 'Job Title' and calculate the average salary
average_salary_by_title = df.groupby('Experience Level')['Salary in USD'].
 ↪mean().sort_values(ascending=True)

# Create a bar plot
plt.figure(figsize=(8, 6))
average_salary_by_title.plot(kind='bar', color='#ab63fa',width=0.3)
plt.xlabel('Job Title')
plt.ylabel('Average Salary')
plt.title('Average Salary by Job Title')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



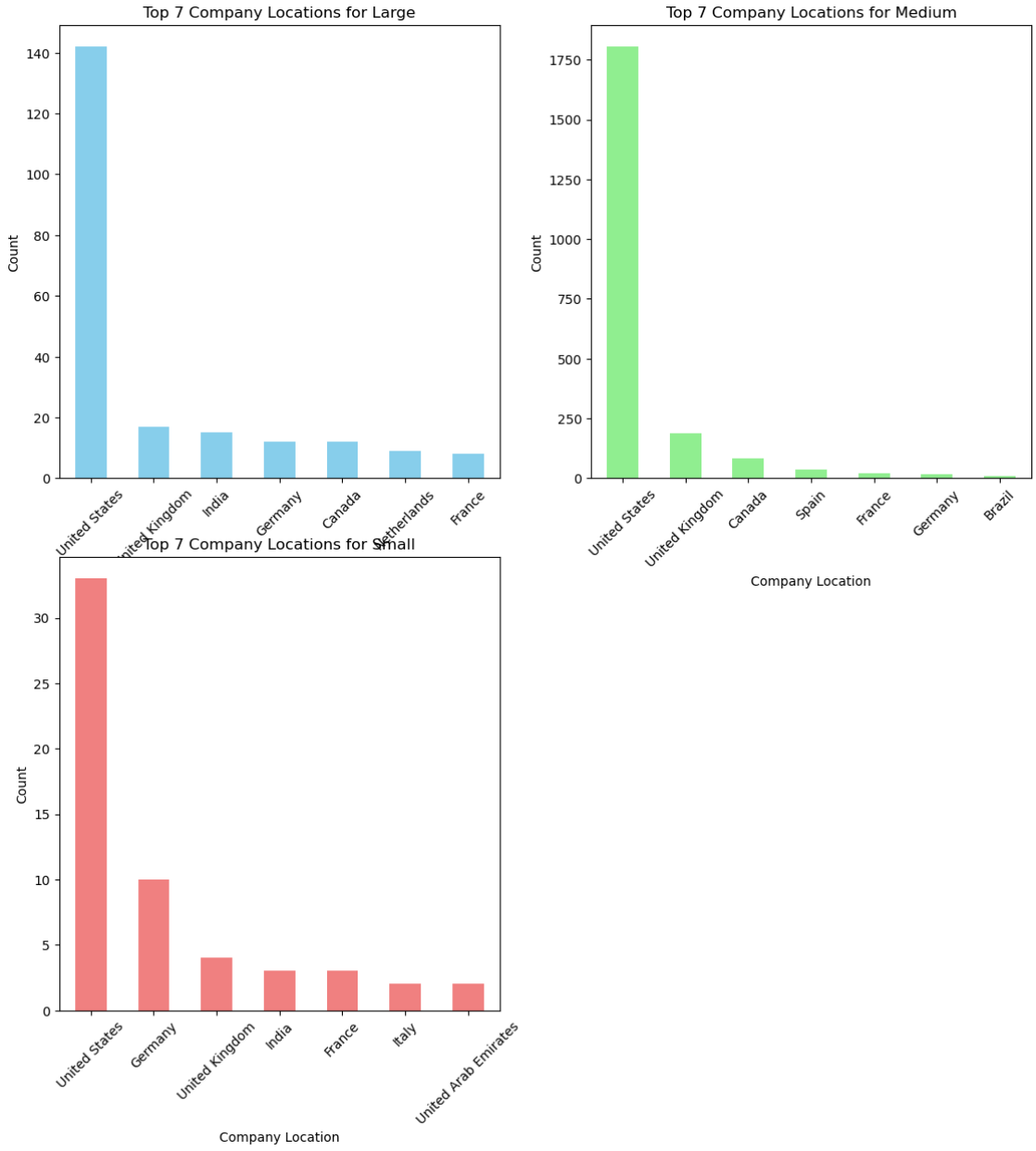
[17]:
```
company_size_by_Company_Location = df.groupby('Company Size')['Company␣
 ↪Location']
for company_size, location_group in company_size_by_Company_Location:
    print(f"Company Size: {company_size}")
    print(location_group.value_counts()[:7])
```

```
    print("\n")
```

```
Company Size: Large
United States     142
United Kingdom     17
India              15
Germany            12
Canada             12
Netherlands         9
France              8
Name: Company Location, dtype: int64


Company Size: Medium
United States     1805
United Kingdom     188
Canada              84
Spain               35
France              20
Germany             17
Brazil               8
Name: Company Location, dtype: int64


Company Size: Small
United States           33
Germany                 10
United Kingdom           4
India                    3
France                   3
Italy                    2
United Arab Emirates     2
Name: Company Location, dtype: int64
```

[18]:
```
company_size_by_Company_Location = df.groupby('Company Size')['Company␣
 ↪Location']

top_countries = 7

num_rows = (len(company_size_by_Company_Location) + 1) // 2
num_cols = 2

# Create subplots
fig, axes = plt.subplots(num_rows, num_cols, figsize=(12, 6 * num_rows))
fig.tight_layout(pad=4.0)
```

```
# Define a list of colors for the bars
colors = ['skyblue', 'lightgreen', 'lightcoral']

for (company_size, location_group), ax, color in␣
 ↪zip(company_size_by_Company_Location, axes.flatten(), colors):
    location_counts = location_group.value_counts().head(top_countries)
    location_counts.plot(kind='bar', color=color, ax=ax)
    ax.set_title(f"Top {top_countries} Company Locations for {company_size}")
    ax.set_xlabel('Company Location')
    ax.set_ylabel('Count')
    ax.tick_params(axis='x', rotation=45)  # Rotate x-axis labels by 45 degrees

# Hide any remaining empty subplots
for ax in axes.flatten()[len(company_size_by_Company_Location):]:
    ax.axis('off')

plt.show()
```



[19]:
```
# Calculate average salary for each combination
average_salary_df = df.groupby(['Job Title', 'Employment Type', 'Experience␣
↪Level'])['Salary in USD'].mean().reset_index()

average_salary_df.rename(columns={"Salary in USD": "Average Salary"},␣
↪inplace=True)

average_salary_df["Average Salary"] = average_salary_df["Average Salary"].
↪astype(int)
```

```
table = tabulate(average_salary_df, headers='keys', tablefmt='pretty',␣
 ↪showindex=False)

print(table)
```

```
+--------------------------+-----------------+------------------+--------------
--+
|         Job Title        | Employment Type | Experience Level | Average
Salary |
+--------------------------+-----------------+------------------+--------------
--+
|    Analytics Engineer    |    Full-Time    |      Entry       |     101333
|
|    Analytics Engineer    |    Full-Time    |    Executive     |     175125
|
|    Analytics Engineer    |    Full-Time    |       Mid        |     127864
|
|    Analytics Engineer    |    Full-Time    |      Senior      |     157525
|
|       Data Analyst       |    Contract     |      Senior      |     90000
|
|       Data Analyst       |    Full-Time    |      Entry       |     68473
|
|       Data Analyst       |    Full-Time    |    Executive     |     100833
|
|       Data Analyst       |    Full-Time    |       Mid        |     98164
|
|       Data Analyst       |    Full-Time    |      Senior      |     124575
|
|       Data Analyst       |    Part-Time    |      Entry       |     50775
|
|       Data Engineer      |    Freelance    |       Mid        |     20000
|
|       Data Engineer      |    Full-Time    |      Entry       |     75702
|
|       Data Engineer      |    Full-Time    |    Executive     |     191050
|
|       Data Engineer      |    Full-Time    |       Mid        |     108985
|
|       Data Engineer      |    Full-Time    |      Senior      |     156034
|
|       Data Engineer      |    Part-Time    |       Mid        |     61137
|
|      Data Scientist      |    Freelance    |       Mid        |     100000
|
|      Data Scientist      |    Full-Time    |      Entry       |     77002
|
```

```
|      Data Scientist       |    Full-Time    |    Executive     |     188429
|
|      Data Scientist       |    Full-Time    |       Mid        |     99513
|
|      Data Scientist       |    Full-Time    |      Senior      |     169058
|
|      Data Scientist       |    Part-Time    |      Entry       |     77223
|
| Machine Learning Engineer |    Contract     |       Mid        |     142500
|
| Machine Learning Engineer |    Freelance    |      Entry       |     100000
|
| Machine Learning Engineer |    Full-Time    |      Entry       |     98304
|
| Machine Learning Engineer |    Full-Time    |    Executive     |     201425
|
| Machine Learning Engineer |    Full-Time    |       Mid        |     131513
|
| Machine Learning Engineer |    Full-Time    |      Senior      |     185805
|
|    Research Scientist     |    Full-Time    |      Entry       |     149845
|
|    Research Scientist     |    Full-Time    |    Executive     |     84053
|
|    Research Scientist     |    Full-Time    |       Mid        |     154460
|
|    Research Scientist     |    Full-Time    |      Senior      |     184129
|
+---------------------------+-----------------+------------------+--------------
--+
```

[ ]: