Global Salaries for Data Science

Data Analysis Final Project

- Caroleen Ataria
- Safa Shehade
- Tasnim Taya

Exploring and Analyzing the Data:
Python, Tableau

[Tableau Visualization - Direct Link]

EDA of "Global Salaries for Data Science" data set

we analysed the data set in Kaggle "https://www.kaggle.com/datasets/lainguyn123/data-science-salary-landscape/data ", using Paython and Tableaue.

We are a group of three Data analysis students, and we are interested in analyzing this dataset to understand the changes in the data science job market and its potential. What are the highest-paying positions and the countries with the best job offers? Such information would help us with our career decisions and could help other individuals seeking such positions

firstly, we import the Pandas and the dataset:

```
import pandas as pd
df=pd.read_csv('/content/salaries.csv')
```

presenting the dataset:

```
df.head()
```

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	co
0	2020	EN	FT	Azure Data Engineer	100000	USD	100000	MU	0	
1	2020	EN	CT	Staff Data Analyst	60000	CAD	44753	CA	50	
2	2020	SE	FT	Staff Data Scientist	164000	USD	164000	US	50	
3	2020	EN	FT	Data Analyst	42000	EUR	47899	DE	0	
4	2020	EX	FT	Data	300000	USD	300000	US	100	

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37234 entries, 0 to 37233
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   work_year           37234 non-null  int64
 1   experience_level    37234 non-null  object
 2   employment_type     37234 non-null  object
 3   job_title           37234 non-null  object
 4   salary              37234 non-null  int64
 5   salary_currency     37234 non-null  object
 6   salary_in_usd       37234 non-null  int64
 7   employee_residence  37234 non-null  object
 8   remote_ratio        37234 non-null  int64
 9   company_location    37234 non-null  object
 10  company_size        37234 non-null  object
dtypes: int64(4), object(7)
memory usage: 3.1+ MB
```

looking at the data we notice that we have 11 columns and 37,234 raws, and that there are no Null values. we also noticed that the data type of the columns are not accurate and usess unnecessarily too much memory. therefore we will be converting the columns with the data type "int64" to "int32" and with the dta type "object" to "str" (looking at the excel file and the describtion of the variables in the dataset kaggel we see that this change is relevant). in addition, we will be checking if there are any dublicates in the data and delet them. so we are entering the **Data Cleaning** step:

```
df['work_year']=df['work_year'].astype('int32')
df['salary']=df['salary'].astype('int32')
df['salary_in_usd']=df['salary_in_usd'].astype('int32')
df['remote_ratio']=df['remote_ratio'].astype('int32')
df['experience_level']=df['experience_level'].astype('str')
df['employment_type']=df['employment_type'].astype('str')
df['job_title']=df['job_title'].astype('str')
df['salary_currency']=df['salary_currency'].astype('str')
df['employee_residence']=df['employee_residence'].astype('str')
df['company_location']=df['company_location'].astype('str')
df['company_size']=df['company_size'].astype('str')
```

```
df.duplicated().any()
```

```
True
```

```
df.drop_duplicates()
```

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio
0	2020	EN	FT	Azure Data Engineer	100000	USD	100000	MU	0
1	2020	EN	CT	Staff Data Analyst	60000	CAD	44753	CA	50
2	2020	SE	FT	Staff Data Scientist	164000	USD	164000	US	50
3	2020	EN	FT	Data Analyst	42000	EUR	47899	DE	0
4	2020	EX	FT	Data Scientist	300000	USD	300000	US	100
...	...	...	...	...	...	...	...	...	...
37227	2024	MI	FT	Data Specialist	79200	USD	79200	US	0
37230	2024	SE	FT	Data Scientist	195500	USD	195500	US	100
37231	2024	SE	FT	Data Scientist	141300	USD	141300	US	100
37232	2024	SE	FT	Data Engineer	139810	USD	139810	US	0
37233	2024	SE	FT	Data Engineer	95325	USD	95325	US	0

after cleaning the data, we are checking the summary of descriptive statistics of the numerical columns in the data:

```
df.describe()
```

	work_year	salary	salary_in_usd	remote_ratio
count	37234.000000	3.723400e+04	37234.000000	37234.000000
mean	2023.656443	1.667366e+05	160540.603105	23.197884
std	0.611469	2.338090e+05	72679.876280	42.005217
min	2020.000000	1.400000e+04	15000.000000	0.000000
25%	2023.000000	1.100000e+05	110000.000000	0.000000
50%	2024.000000	1.500000e+05	150000.000000	0.000000
75%	2024.000000	2.000000e+05	200000.000000	0.000000

and we check the count of each value in the non nomirc columns:

```
df['job_title'].value_counts()
```

job_title	count
Data Scientist	7448
Data Engineer	6103
Data Analyst	4351
Machine Learning Engineer	3990
Software Engineer	2935
...	...
Principal Data Architect	1
Deep Learning Researcher	1
BI Data Engineer	1
AWS Data Architect	1
Power BI Developer	1

215 rows × 1 columns

```
df['experience_level'].value_counts()
```

experience_level	count
SE	22523
MI	10723
EN	3166
EX	822

```
df['employment_type'].value_counts()
```

employment_type	count
FT	37121
PT	52
CT	47
FL	14

```
df['company_location'].value_counts()
```

company_location	count
US	33806
GB	1129
CA	1115
DE	149
ES	139
...	...
EC	1
AD	1
MY	1
QA	1
MU	1

81 rows × 1 columns

```
df['employee_residence'].value_counts()
```

employee_residence	count
US	33755
GB	1121
CA	1113
ES	143
DE	142
...	...
UG	1
DO	1
ID	1
OM	1
MU	1

91 rows × 1 columns

```
df['salary_currency'].value_counts()
```

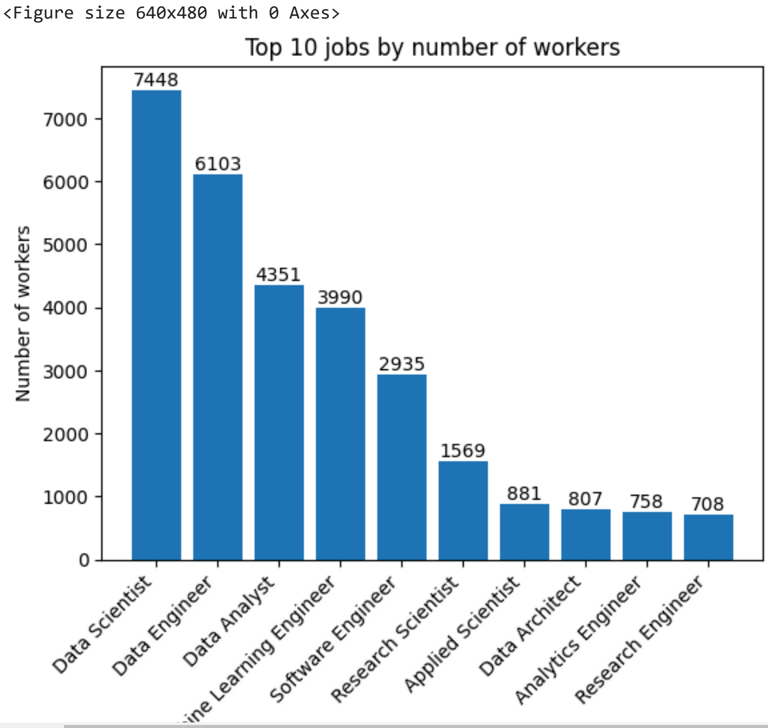
salary_currency	count
USD	35443
GBP	1029
EUR	551
CAD	81
INR	56
AUD	12
PLN	11
CHF	10
SGD	6
BRL	5
TRY	4
DKK	4
JPY	4
HUF	3
ZAR	3
ILS	2
NOK	2
THB	2
CLP	1
MXN	1
PHP	1
HKD	1
SEK	1
NZD	1

The next step is **Data Analyzing**, for that we import the panda libirary "matplotlib".

we are checking what are **the top 10 job positions with the highest workers number**:

```
import matplotlib.pyplot as plt
plt.figure()
jobs=df['job_title'].value_counts()[:10]
fig, ax = plt.subplots()
bar_container=ax.bar(jobs.index,jobs.values)
ax.bar_label(bar_container)
ax.set(ylabel='Number of workers', title='Top 10 jobs by number of workers')
plt.xticks(rotation=45,ha='right')
plt.show()
```

```
<Figure size 640x480 with 0 Axes>
```

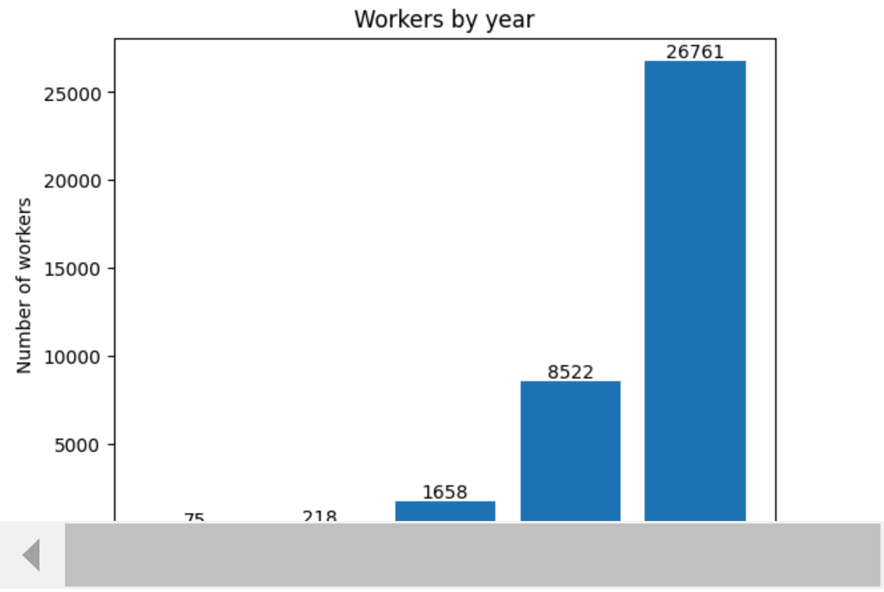


graph with **the number of workers over the years**:

```
import matplotlib.pyplot as plt
plt.figure()
```

```
work_years=df['work_year'].value_counts()
fig, ax = plt.subplots()
bar_container=ax.bar(work_years.index,work_years.values)
ax.bar_label(bar_container)
ax.set(ylabel='Number of workers', title='Workers by year')
plt.show()
```

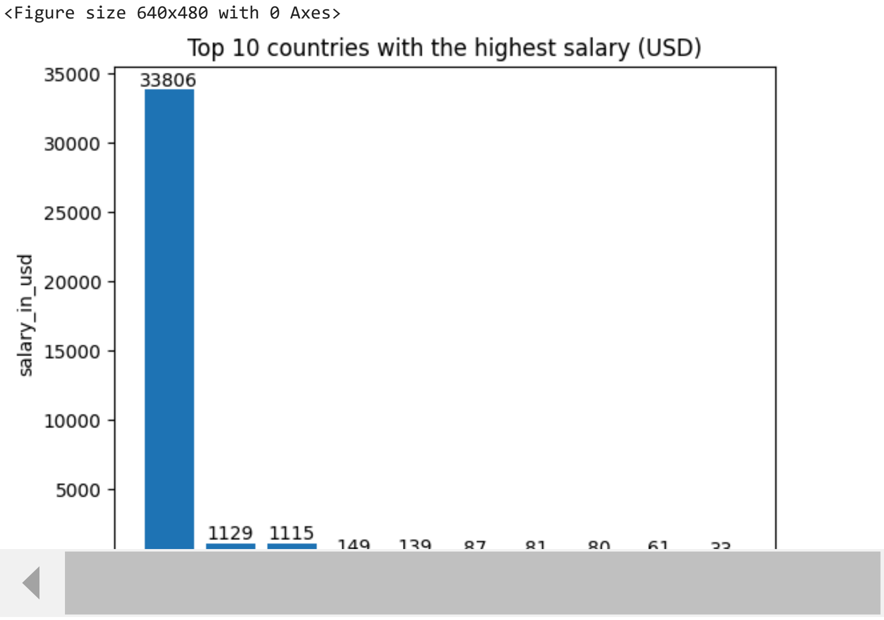
```
<Figure size 640x480 with 0 Axes>
```



graph of **the top 10 countries with the highest average salary**:

```
plt.figure()
company_location=df['company_location'].value_counts()[:10]
fig, ax = plt.subplots()
bar_container=ax.bar(company_location.index,company_location.values)
ax.bar_label(bar_container)
ax.set(ylabel='salary_in_usd', title='Top 10 countries with the highest salary (USD)')
plt.show()
```

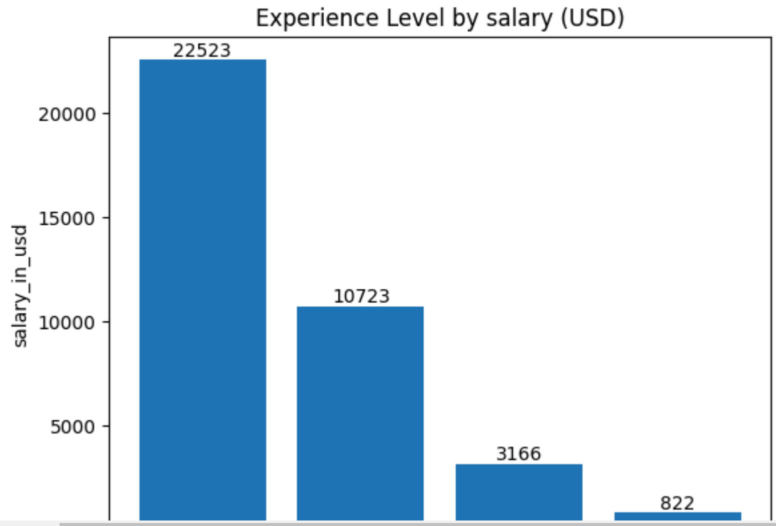
```
<Figure size 640x480 with 0 Axes>
```



graph of **the experince level by salary**:

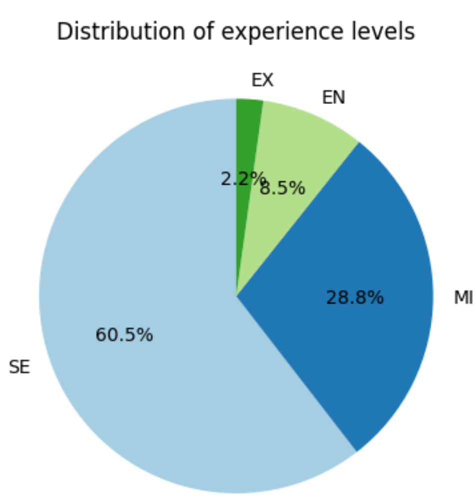
```
plt.figure()
experience_level=df['experience_level'].value_counts()
fig, ax = plt.subplots()
bar_container=ax.bar(experience_level.index,experience_level.values)
ax.bar_label(bar_container)
ax.set(ylabel='salary_in_usd', title='Experience Level by salary (USD)')
plt.show()
```

```
<Figure size 640x480 with 0 Axes>
```



Pie chart of **the distribution of the experince level**:

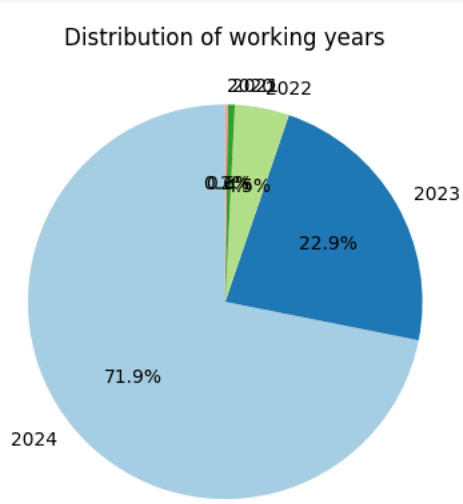
```
plt.figure()
experience_level=df['experience_level'].value_counts()
plt.pie(experience_level, labels=experience_level.index, autopct='%1.1f%%', startangle=90, colors=plt.cm.Paired.colors)
plt.title('Distribution of experience levels')
plt.show()
```



The distribution of the experience level reflects the reality we are facing, most companies are opening jop titles that requier's senior level of experince. while Mid level is the second common level, it is still with very low percintage (28.8%), and even worse the entry level percintage is 8.5%!

Pie chart of **the distribution of Working years**:

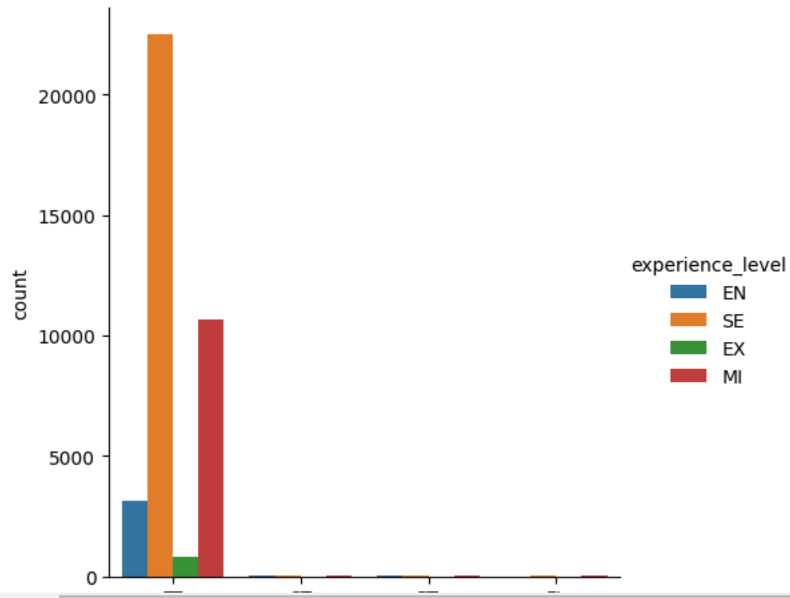
```
plt.figure()
work_year=df['work_year'].value_counts()
plt.pie(work_year, labels=work_year.index, autopct='%1.1f%%', startangle=90, colors=plt.cm.Paired.colors)
plt.title('Distribution of working years')
plt.show()
```



We can see that each year the work flow increases. The low working flow in the 2020-2021 is explained by the pandamic, after 2021 companies started catching up.

Visualize relationships between experince level and employment type.

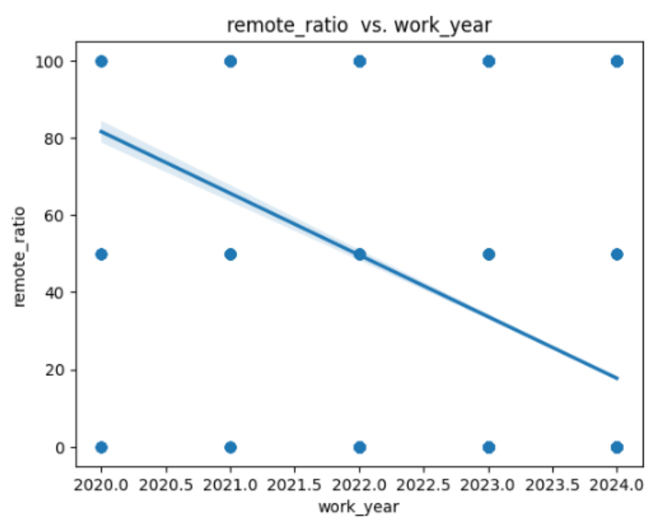
```
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
sns.catplot(x='employment_type', hue='experience_level', kind='count', data=df)
plt.show()
```



We can see that "Full-Time" is the most common employment type. And that "Senior" and "Mid_Level" seem to be more prevalent in "Full-Time" roles. There are fewer "Expert" level individuals in "Full-Time" roles compared to other experience levels.

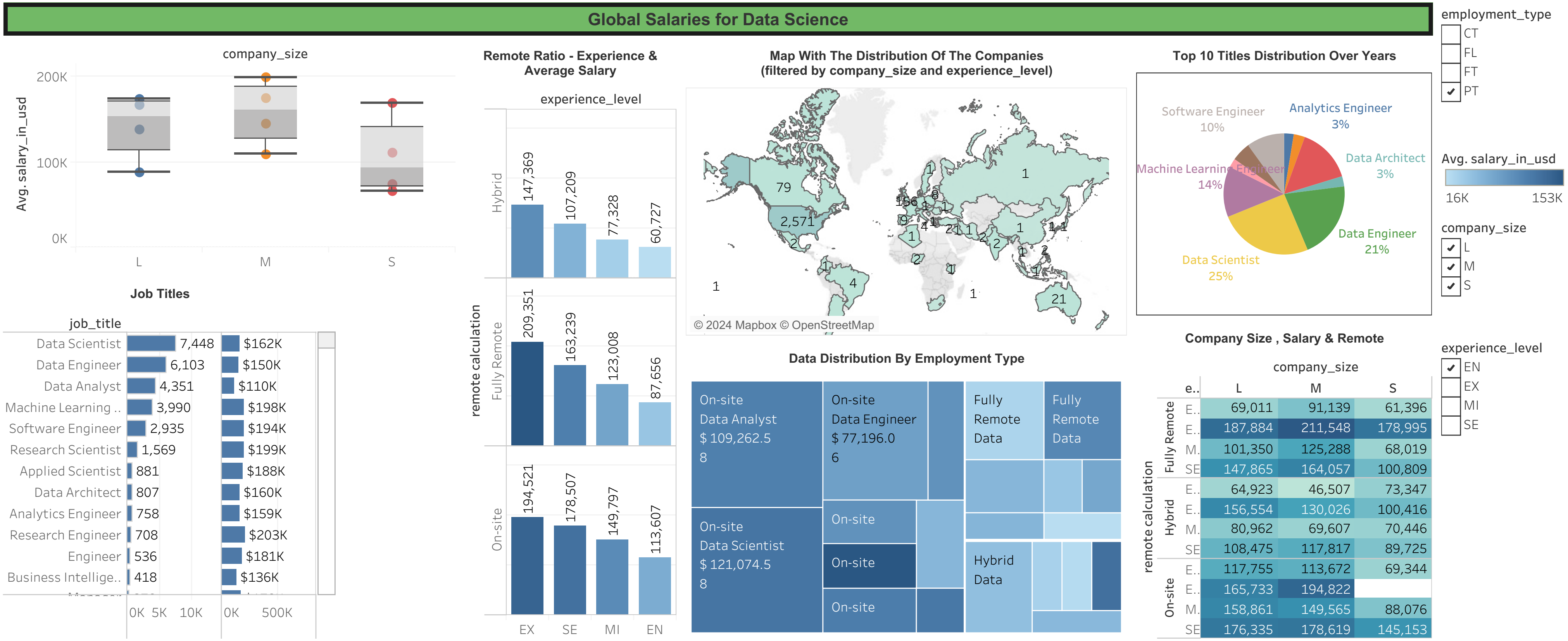
regression line- Remote Ratio VS Working Years

```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
sns.regplot(x='work_year', y='remote_ratio', data=df)
plt.title('remote_ratio  vs. work_year')
plt.xlabel('work_year')
plt.ylabel('remote_ratio')
plt.show()
```

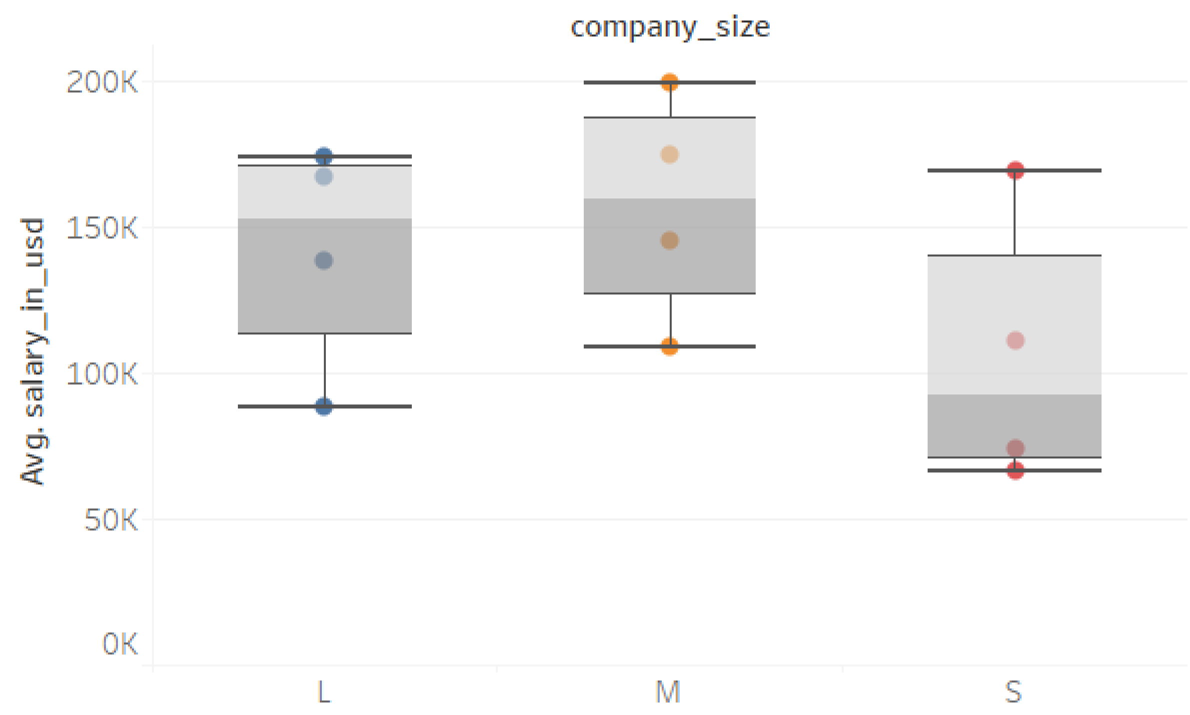


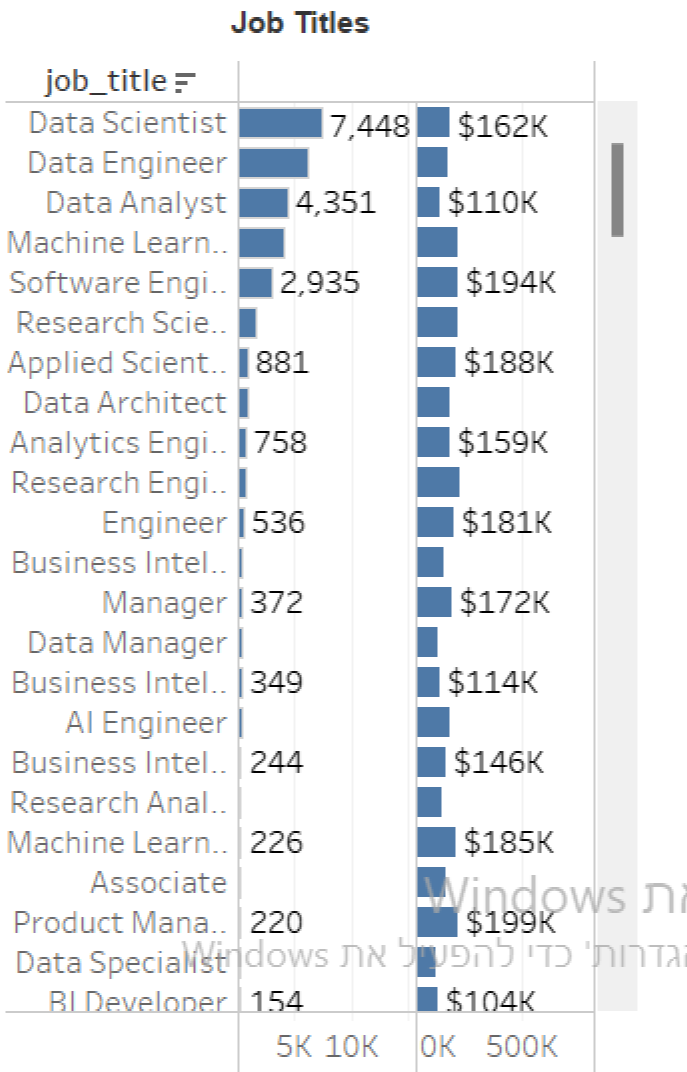
From the line we notice a downward in the remote ratio as years go on. The downward suggest that companies are moving away from remote work and returning to more traditional office-based work arrangements (after the pandamic year).

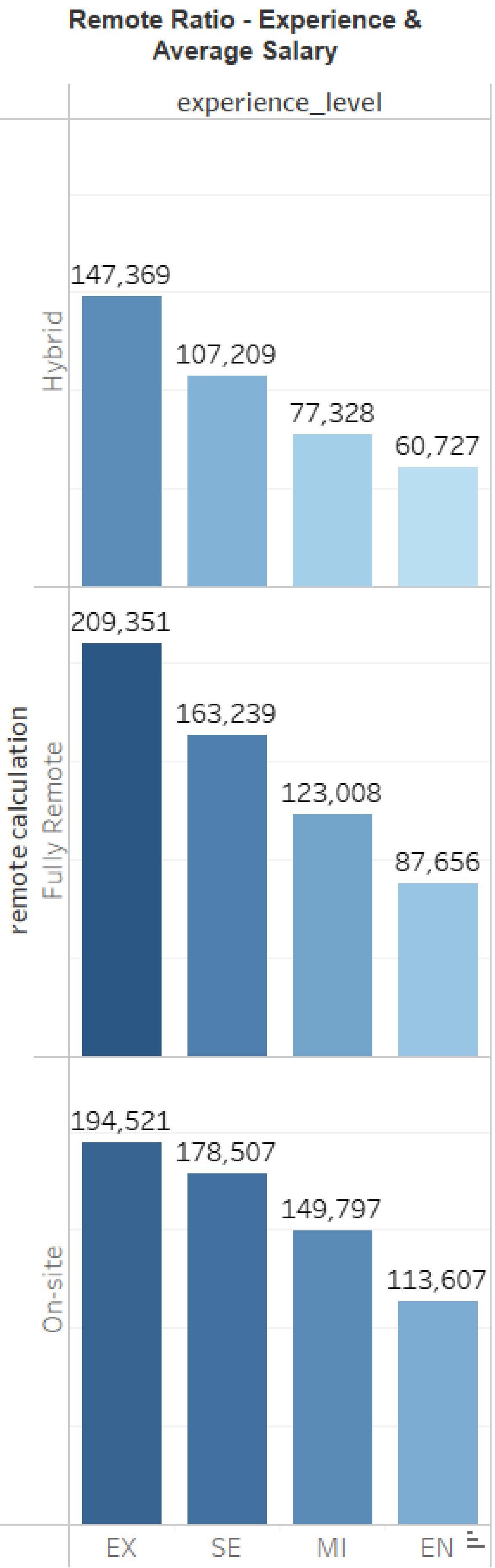
The data points are scattered, indicating some variability in the "remote ratio" for each year and for different companies. While some companies might be more inclined to continue with remote work, others might be transitioning back to in-person work.

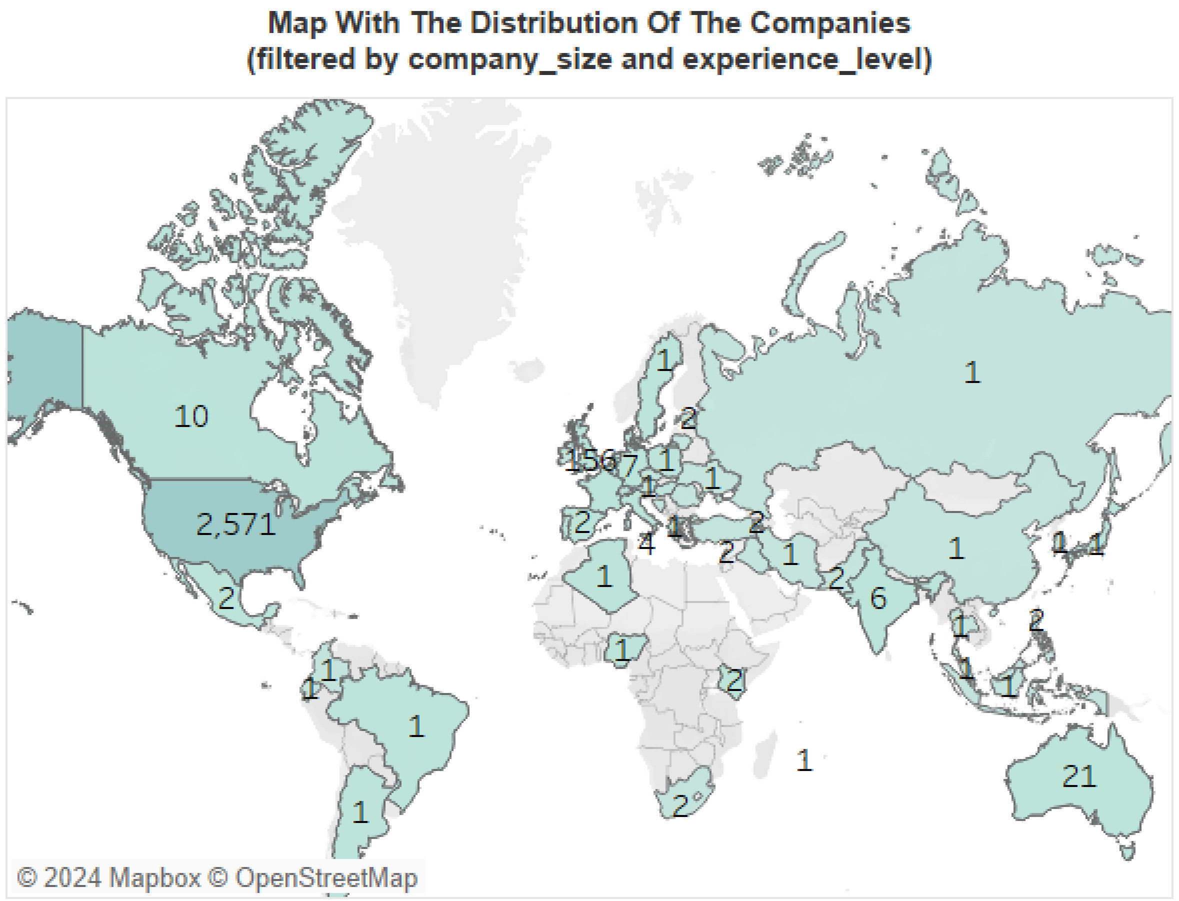

Global Salaries for Data Science
company_size
Avg. salary_in_usd
200K
100K
0K
L
M
S
Job Titles
job_title
Data Scientist 7,448 $162K
Data Engineer 6,103 $150K
Data Analyst 4,351 $110K
Machine Learning .. 3,990 $198K
Software Engineer 2,935 $194K
Research Scientist 1,569 $199K
Applied Scientist 881 $188K
Data Architect 807 $160K
Analytics Engineer 758 $159K
Research Engineer 708 $203K
Engineer 536 $181K
Business Intellige.. 418 $136K
0K 5K 10K
0K 500K
Remote Ratio - Experience & Average Salary
experience_level
remote calculation
Hybrid
147,369
107,209
77,328
60,727
Fully Remote
209,351
163,239
123,008
87,656
On-site
194,521
178,507
149,797
113,607
EX
SE
MI
EN
Map With The Distribution Of The Companies (filtered by company_size and experience_level)
79
2,571
1
4
21
© 2024 Mapbox © OpenStreetMap
Data Distribution By Employment Type
On-site
Data Analyst
$ 109,262.5
8
On-site
Data Engineer
$ 77,196.0
6
Fully Remote Data
Fully Remote Data
On-site
Data Scientist
$ 121,074.5
8
On-site
On-site
On-site
Hybrid Data
Top 10 Titles Distribution Over Years
Software Engineer 10%
Analytics Engineer 3%
Machine Learning Engineer 14%
Data Architect 3%
Data Engineer 21%
Data Scientist 25%
Company Size , Salary & Remote
company_size
remote calculation
e.. L M S
Fully Remote
E.. 69,011 91,139 61,396
E.. 187,884 211,548 178,995
M.. 101,350 125,288 68,019
SE 147,865 164,057 100,809
Hybrid
E.. 64,923 46,507 73,347
E.. 156,554 130,026 100,416
M.. 80,962 69,607 70,446
SE 108,475 117,817 89,725
On-site
E.. 117,755 113,672 69,344
E.. 165,733 194,822
M.. 158,861 149,565 88,076
SE 176,335 178,619 145,153
employment_type
CT
FL
FT
PT
Avg. salary_in_usd
16K
153K
company_size
L
M
S
experience_level
EN
EX
MI
SE

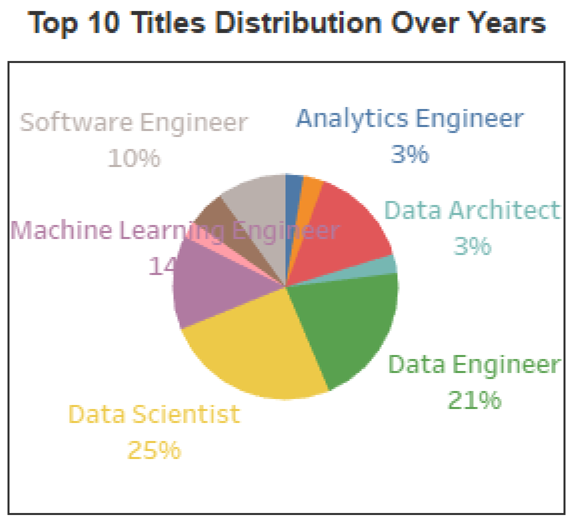
Global Salaries for Data Science









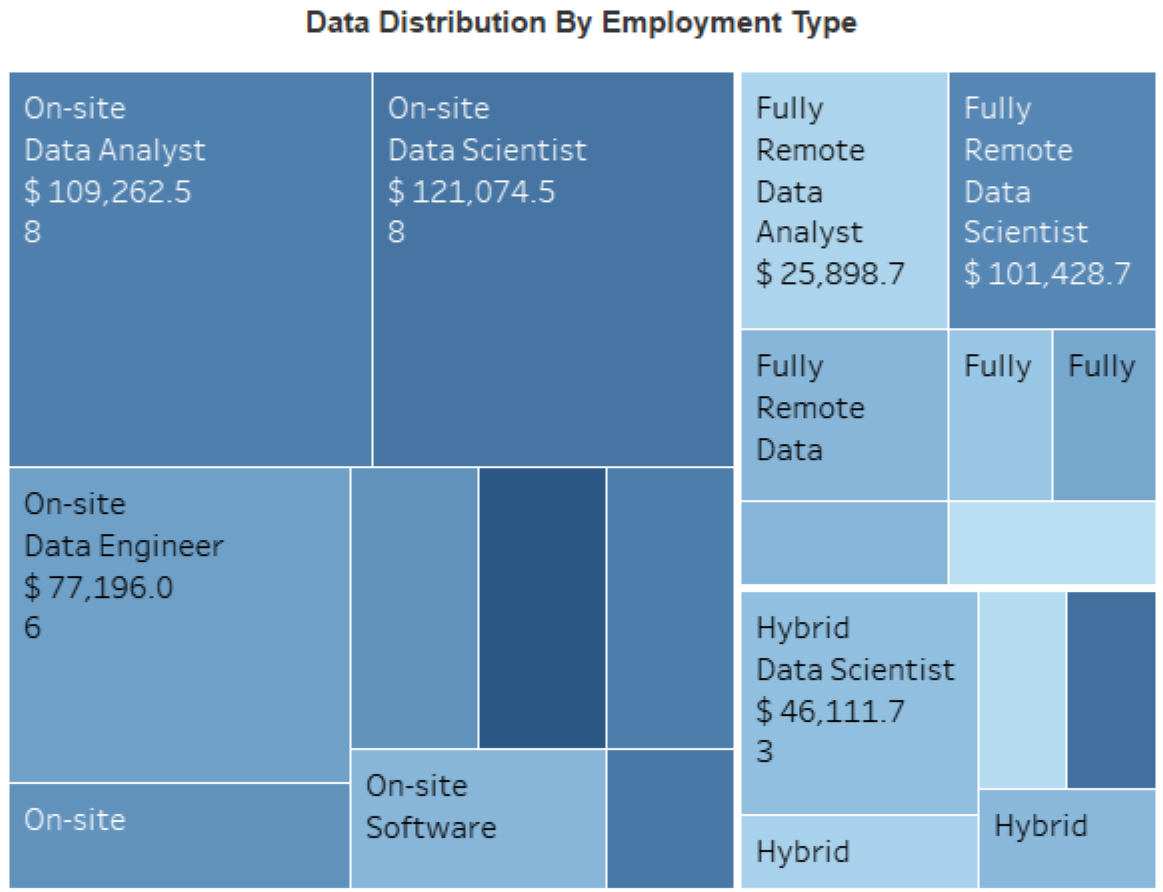


employment_t..
- (All)
- CT
- FL
- FT
- PT

Avg. salary_in_..
16K 153K

company_size
- (All)
- L
- M
- S

experience_level
- (All)
- EN
- EX
- MI
- SE



Company Size , Salary & Remote

remote calculation	e.	L	M	S
Fully Remote	E.	69,011	91,139	61,396
	E.	187,884	211,548	178,995
	.	101,350	125,288	68,019
	S.	147,865	164,057	100,809
Hybrid	E.	64,923	46,507	73,347
	E.	156,554	130,026	100,416
	.	80,962	69,607	70,446
	S.	108,475	117,817	89,725
On-site	E.	117,755	113,672	69,344
	E.	165,733	194,822	
	.	158,861	149,565	88,076
	S.	176,335	178,619	145,153

Summary of Findings

1. Focused on entry-level and junior mid-levels: The US has the most incidence for S, M, L company sizes for hiring entry-level and junior mid-levels among all the companies all over the world.
2. Small companies have no on-site employees at the expert executive level.
3. Expert executive-levels gain the most salary on average among all other employee levels.
4. Apparently, on-site employees tend to earn the highest salary on average compared to other remote types (Hybrid and Fully Remote) in L and M companies.
5. Employees within small companies get paid the lowest of all expertise.
6. Hybrid-hired employees get paid the lowest among all other remote types of employment.
7. Full remote Expert executive levels gain the highest salary on average among other executive remote types
8. The role of “Data Scientist” holds the largest share in the job title distribution over the years, compared to other roles.
9. Top 10 titles are full-time job.
10. Top 10 titles are mostly on-site employees.
11. The top 3 titles who get paid the most are Analytics Engineering manager, Data science Tech Lead and Applied AI ML Lead.
12. Medium-sized companies appear to grant the highest salaries on average compared to small and large companies.
13. On-site employees are offered significantly higher salaries over the years 2020 – 2024, while employees under the hybrid trend do not indicate a change; this finding is surprising, particularly after the COVID-19 pandemic