Labor Market for Data Scientists in Europe: What Factors Influence Salary?

Exploratory Analysis of Data Related to Data Scientist Salaries from 2020 to 2024.

1. Introduction and Objective of the Analysis

Context: The job market in data science is booming due to the exponential growth of data generated and companies' interest in leveraging this data for strategic decision-making. More and more organizations, from startups to large corporations, are seeking professionals who can transform large volumes of data into valuable information, creating a high demand for data scientists, data engineers, and other related roles.

Objective: Identify key patterns and characteristics to understand how each of the studied variables influences the total salary earned by these professionals.

2. Data Description

Data Source: The study is based on data collected from the "Data Science Salaries Dataset" on Kaggle, which provides data of sufficient quality to analyze this issue.

Variable Descriptions:

Variable	**Description**	Type
work_year	The year in which the salary data was collected.	years
experience_level	The employee's experience level (e.g., Junior, Mid-level, Senior, Expert).	Ordinal Categorical
employment_type	The type of employment (e.g., Full-Time, Part-Time, Contract).	Ordinal Categorical
job_title	The title or role of the employee in the data science field.	Nominal Categorical
salary	The employee's salary in the currency specified by salary_currency.	Numeric
salary_currency	The currency in which the salary is denoted.	Nominal Categorical
salary_in_usd	The employee's salary converted to USD for standardization.	Numeric
employee_residence	The location of the employee's residence.	Nominal Categorical
remote_ratio	The percentage of remote work allowed for the position (e.g., 0, 50, 100).	Ordinal Categorical

company_location	The location of the company where the employee works.	Ordinal Categorical
company_size	The size of the company based on employee count (e.g., Small, Medium, Large).	Ordinal Categorical

Original dataset size: 14,838 rows and 11 columns.

3. Data Preprocessing

Missing data: No missing data or duplicate rows were found.

Data type conversion: It was not necessary to convert data types, as the dataset is composed of Int64, object, and float64, making analysis easier.

Transformations: New variables were created in new columns to facilitate the analysis.

New Variable	**Description**	Type
job_category	job title has been distributed into 4 higher categories: Data Analyst , Data Scientist, Data Architect / Enginner and Data Manager.	Nominal Categorical
salary_in_euro	Euro convertion from salary_in_usd	Numeric
country_employee_residence	Country name related to the ISO code	Ordinal Categorical
europe_zone	Country residence of employee have been categorized by regions	Nominal Categorical
experience_level_num	Numerical format of the level of experience	Numeric
salary_category_num	Numerical format of the salary category	Numeric
salary_category	Salary category of employee have been categorized by several groups: ‘under 40.000’, '40.000€-50.000€', 50.000€-60.000€', '60.000€-70.000€', '70.000€-80.000€' and 'More than 80.000€'	Ordinal Categorical
company_size_num	Numerical format of the company size	Numerical

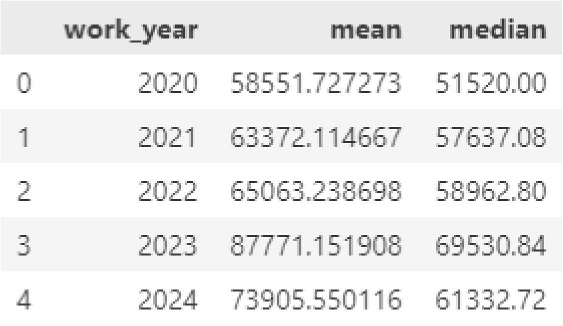
Dataset Filtering: The analysis will be conducted only for Data Scientists residing in Europe.

4. Univariate Analysis

The study begins by individually analyzing each of the variables to draw conclusions regarding their distribution, number of unique values, and the relevance of the variable.

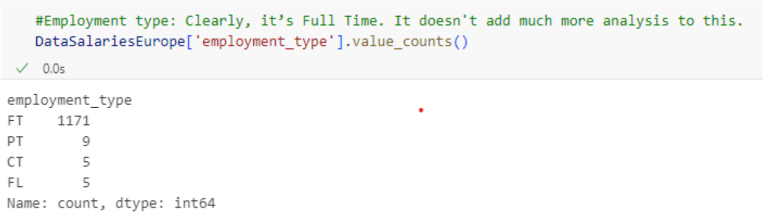
Central tendency measures:
Mean and median of **salary_in_euro**: the trend has been upward until 2023, both in terms of average salary and median.

It seems that this trend is changing in 2024.



	work_year	mean	median
0	2020	58551.727273	51520.00
1	2021	63372.114667	57637.08
2	2022	65063.238698	58962.80
3	2023	87771.151908	69530.84
4	2024	73905.550116	61332.72

The variable called "**employment_type**" will not be considered in the study, as 98.40% of data scientists work "Full Time." There is not enough variability in this category to calculate its influence on salary.

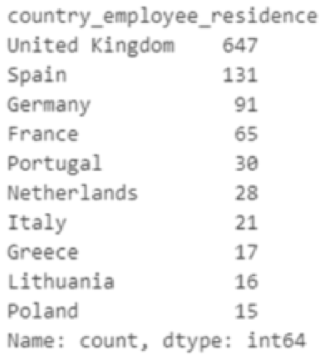


```
#Employment type: Clearly, it’s Full Time. It doesn't add much more analysis to this.
DataSalariesEurope['employment_type'].value_counts()
```

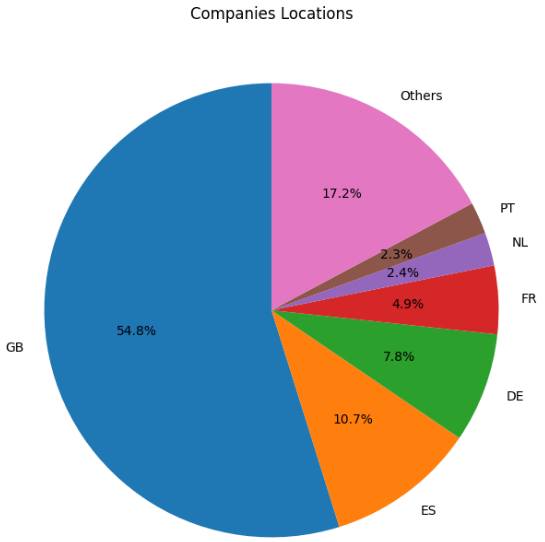
✓ 0.0s

```
employment_type
FT    1171
PT       9
CT       5
FL       5
Name: count, dtype: int64
```

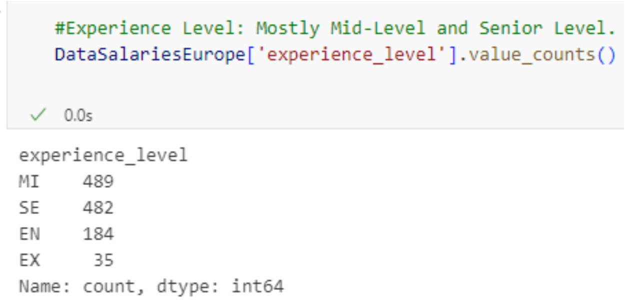
In terms of the **locations** of both professional data science talent and **hiring companies**, the United Kingdom stands out as the leading country.



```
country_employee_residence
United Kingdom    647
Spain             131
Germany            91
France             65
Portugal           30
Netherlands        28
Italy              21
Greece             17
Lithuania          16
Poland             15
Name: count, dtype: int64
```



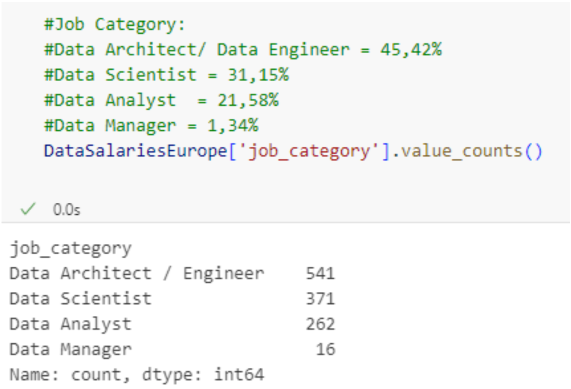
The variables such as "experience_level" and "**job_category**" will be useful to us, as their values are evenly distributed across their different categories.



```
#Experience Level: Mostly Mid-Level and Senior Level.
DataSalariesEurope['experience_level'].value_counts()
```

✓ 0.0s

```
experience_level
MI    489
SE    482
EN    184
EX     35
Name: count, dtype: int64
```

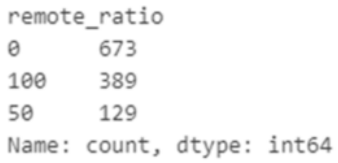


```
#Job Category:
#Data Architect/ Data Engineer = 45,42%
#Data Scientist = 31,15%
#Data Analyst  = 21,58%
#Data Manager = 1,34%
DataSalariesEurope['job_category'].value_counts()
```

✓ 0.0s

```
job_category
Data Architect / Engineer    541
Data Scientist               371
Data Analyst                 262
Data Manager                  16
Name: count, dtype: int64
```

Surprisingly, **remote work** is not as common as one might expect. On-site work is still the most common option for these professionals.

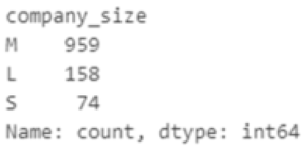


```
remote_ratio
0      673
100    389
50     129
Name: count, dtype: int64
```

Finally, the distribution of the variable **'company_size'** was also surprising. The vast majority (80.58%) of hiring companies are medium-sized.

What do we mean by small, medium, or large company? Convention tells us that the size of the company depends on the level of income and the number of employees.

	Small	Medium	Large
Revenues in Millions Euros	<10 M	10 M < Medium < 50 M	> 50 M
Employees_num	< 50	50 > Medium > 250	> 250 M

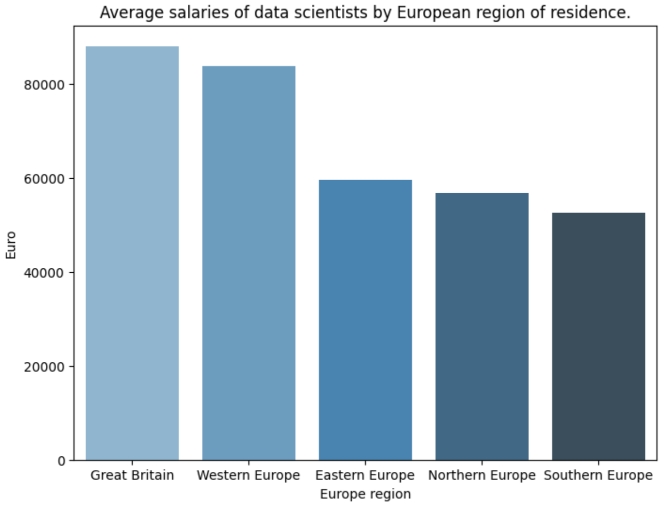


```
company_size
M    959
L    158
S     74
Name: count, dtype: int64
```

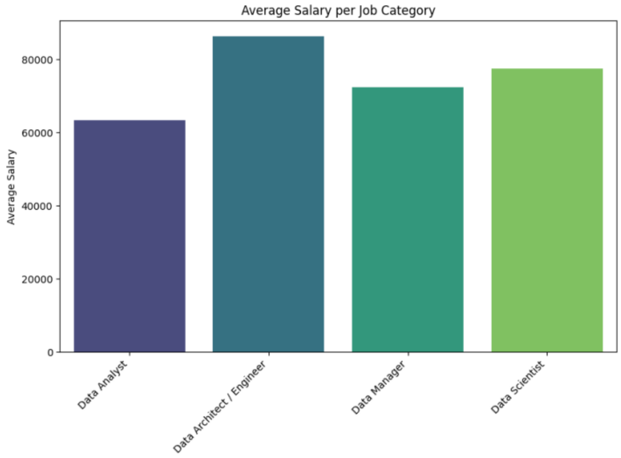
5. Bivariate y Multivariate Analysis

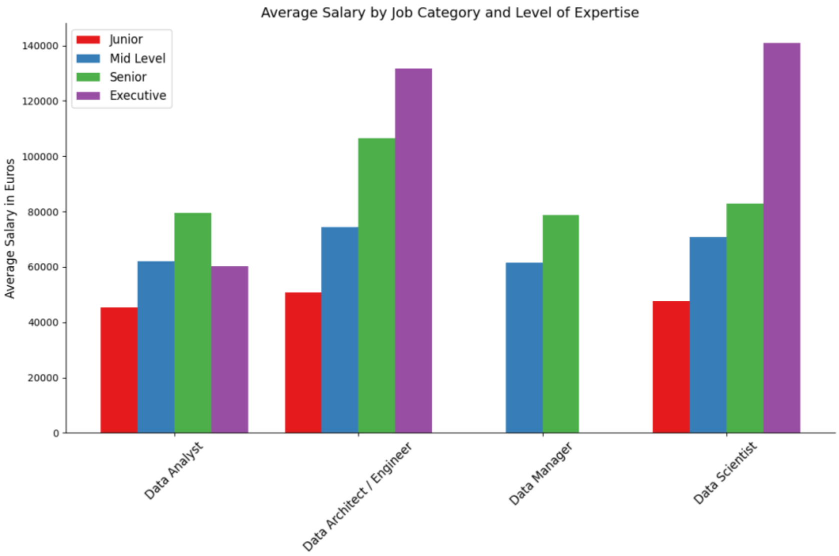
All the variables studied significantly influence salary determination; however, 3 of them seem to be even more relevant:

Employee_residence: Data Scientists residing in Great Britain and Western Europe are clearly better paid than their European neighbors.



Job_category and Experience_level: The job category influences salary determination. In general, Data Analysts earn less on average than Data Architects/Engineers, who are the most valued. It seems obvious to think that the more experience you have, the more you will earn. And this is true for all categories except Data Analyst. From the study, we can infer the cause: There are very few Executive Data Analysts, and the few that exist are not valued as much as their counterparts in other categories. Perhaps this range is mostly occupied by Data Scientists





6. Challenges during the analysis

The main challenge in this study was the CATEGORICAL nature of the variables.

The existence of correlation is not so straightforward to calculate. I tried to create a radar chart that showed the correlation between each variable and salary_in_euro based on the ANOVA analysis. The result was that all the variables had a significant influence, but none stood out particularly. The chart didn't reveal anything clear, as all the values were close to zero. Even after applying the logarithm, this did not improve.

The correlation basically had to be analyzed using histograms and bar charts.

For the nominal qualitative variables, it was necessary to group each of them into more generic categories. For example: Job_title -> Job_category or employee_residence -> europe_zone.

On the other hand, for the ordinal variables (where there is an order from lowest to highest in the values), each of the values was simply converted to numeric. For example: company_size -> Small:1; Medium:2; and Large:3.

7. Conclusions

For data science professionals, focusing on the continuous development of in-demand skills (such as artificial intelligence and cloud computing) relevant to roles like Data Architect/Engineer or Data Scientist, and considering geographic mobility, can be effective strategies to improve salary prospects.

This analysis provides a detailed map of the salary landscape for data scientists in Europe, highlighting how factors like experience, location, and professional category affect compensation. The insights gained can help both employers and candidates make informed decisions in a rapidly evolving job market.

For companies and employers, this analysis suggests that establishing competitive salaries involves considering not only the local cost of living but also the technical specialization and experience these professionals bring. Additionally, to attract and retain talent, investing in professional development programs that strengthen emerging technical skills is recommended.