Just to get your attention






Data Science
Labor Market Research

Years covered:

2023 – 2024

Affected variables:

Position
Qualification
Salary
Work type
Remoteness
Size and Location of the company

Content

🔔 Clarifications

- This is not a 3-slide report, but a clear projection of each step of my work, placed in a compact presentation for easy familiarization

- This work does not oblige the viewer to fully study the entire project

- The main emphasis is on the approach to data, their presentation, and accompanying thinking

- All the extracts from the project are in the sections "The Bottom Line" AND "What I learned from the project"

- The project is my first work
  Many mistakes were realized during or at the end of the project

Goal Setting

Relationships:

- Analyze relationships between categorical variables
- Identify wage relationships between categorical variables

Salary Trends:

- Compare total and particular wages across years
- Model a regression model to determine the most influential variables on wages

Staff Demand:

- Analyze demand for specialists across years
- Consider qualification level
- Consider company size

the detailed course of the investigation can be found in the 4th section

Remote Work:

- Compare trends in remote work adoption

Context & Background

"To know what you know and what you do not know, that is true knowledge."
— Confucius

1. Prior knowledge

2. Data Source & Collection Method

3. Collection Purpose

4. Timeframe

5. Limitations
bias, external influence

Prior knowledge

Number of rows: 27 500

work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
2024	SE	FT	Machine Learning Engineer	150000	USD	150000	US	0	US	M
2024	MI	FT	Data Manager	128000	USD	128000	US	0	US	M
2024	SE	FT	Research Engineer	202000	USD	202000	US	100	US	M
2023	SE	FT	BI Developer	140000	USD	140000	US	100	US	M
2023	SE	FT	Data Engineer	123700	USD	123700	US	0	US	L

```
SALARY_CURRENCY:
['USD' 'GBP' 'CAD' 'EUR' 'TRY' 'PLN' 'ZAR' 'SEK' 'INR' 'DKK' 'ILS' 'BRL'
 'CHF' 'NZD' 'AUD' 'PHP' 'NOK' 'JPY' 'HKD' 'SGD' 'THB' 'HUF' 'MXN' 'CLP']
```

```
WORK_YEAR:
[2024 2022 2023 2020 2021]

EXPERIENCE_LEVEL:
['SE' 'EN' 'MI' 'EX']

EMPLOYMENT_TYPE:
['FT' 'PT' 'CT' 'FL']
```

```
REMOTE_RATIO:
[100   0  50]
```

```
COMPANY_SIZE:
['M' 'L' 'S']
```

```
COMPANY_LOCATION:
['US' 'GB' 'AR' 'CA' 'KR' 'EG' 'AU' 'IN' 'NZ' 'BR' 'IE' 'NL' 'FR' 'ZA'
 'PH' 'IT' 'TR' 'AM' 'MX' 'DE' 'SK' 'PL' 'LT' 'PT' 'CL' 'DZ' 'AS' 'IL'
 'FI' 'LU' 'KE' 'RS' 'GR' 'UA' 'ES' 'SE' 'DK' 'LV' 'AT' 'CH' 'AE' 'SA'
 'OM' 'BA' 'EE' 'MT' 'HU' 'LB' 'RO' 'VN' 'NG' 'GI' 'CO' 'SI' 'MU' 'RU'
 'CZ' 'QA' 'GH' 'AD' 'EC' 'NO' 'JP' 'HK' 'CF' 'SG' 'TH' 'HR' 'PK' 'IR'
 'BS' 'PR' 'BE' 'ID' 'MY' 'HN' 'IQ' 'CN' 'MD']
```

```
JOB_TITLE:
['Software Engineer' 'Data Architect' 'AI Developer' 'Data Analyst'
 'Data Scientist' 'Data Product Owner' 'Data Engineer'
 'Machine Learning Engineer' 'Engineering Manager' 'Software Developer'
 'Machine Learning Scientist' 'Solutions Architect' 'Research Scientist'
 'Applied Scientist' 'Analytics Engineer' 'DevOps Engineer'
 'Data Specialist' 'Business Intelligence Engineer'
 'Data Operations Analyst' 'Frontend Engineer' 'AI Engineer'
 'Data Analytics Specialist' 'Site Reliability Engineer'
 'Solutions Engineer' 'AI Architect' 'Research Engineer'
```

```
'Business Intelligence Developer' 'Data Modeler' 'ETL Developer'
'Business Analyst' 'Business Intelligence' 'Data Product Manager'
'Decision Scientist' 'Data Visualization Engineer'
'Encounter Data Management Professional' 'Data Strategist'
'Data Quality Analyst' 'Statistical Programmer' 'Systems Engineer'
'Software Development Engineer' 'Data Analytics Manager'
'Full Stack Engineer' 'Data Quality Manager'
'Data Visualization Specialist' 'Research Analyst' 'Head of Data'
'Data Team Lead' 'Robotics Engineer' 'Data Governance Lead'
```

```
'Data Governance Specialist' 'Head of AI' 'Data Management Manager'
'Data Governance Engineer' 'Data Management Specialist'
'Machine Learning Model Engineer' 'Artificial Intelligence Engineer'
'Machine Learning Manager' 'MLOps Engineer' 'ML Infrastructure Engineer'
'Cloud Database Administrator' 'Computer Vision Engineer'
'Data Integration Engineer' 'Data Quality Engineer' 'Python Developer'
'Data Operations Engineer' 'Data Infrastructure Engineer'
'Data Operations Specialist' 'Data Visualization Analyst'
'Machine Learning Architect' 'Machine Learning Modeler'
```

Data Source & Collection Method

Source link:
https://ai-jobs.net/

Collection method:
...based on internal data (survey submissions (form) + jobs with open salaries)

Despite the public availability of the electronic form (bias, intentionality), the number of observations in the data is greater than 25 000, thus mitigating this vulnerability

Collection Purpose

Our goal is to have open salary data for everyone.
So newbies, experienced pros, hiring managers, recruiters and also startup founders or people wanting to switch careers can make better decisions

Timeframe

2020 – 2024

The data is processed and updated on a weekly basis

Limitations

(bias, external influence)

Missing data:

- Job location
  (city, state/province)
  may influence salary levels

- Industry or sector
  may influence salary levels

- Requirements for each job title
  may influence salary levels

- Education & Qualifications
  (educational background, degrees, certifications)
  may influence salary levels

- Employee demographics
  (age, gender, or ethnicity)
  potential biases or disparities in compensation

Yes, these points can affect results, but we're looking for the general trend, not strict perfection

Merging data sources can be valuable

Exploratory Data Analysis

1. Any incorrect types?
2. Any missing values?
3. Any duplicates?
4. Any inconsistencies?
5. Any outliers?
6. Feature engineering (1/2)
7. Analyze variable distribution

initial mental model of the dataset

Any incorrect types?

Minor corrections are available

Insignificant conversion of
textual representation to
categorical representation:

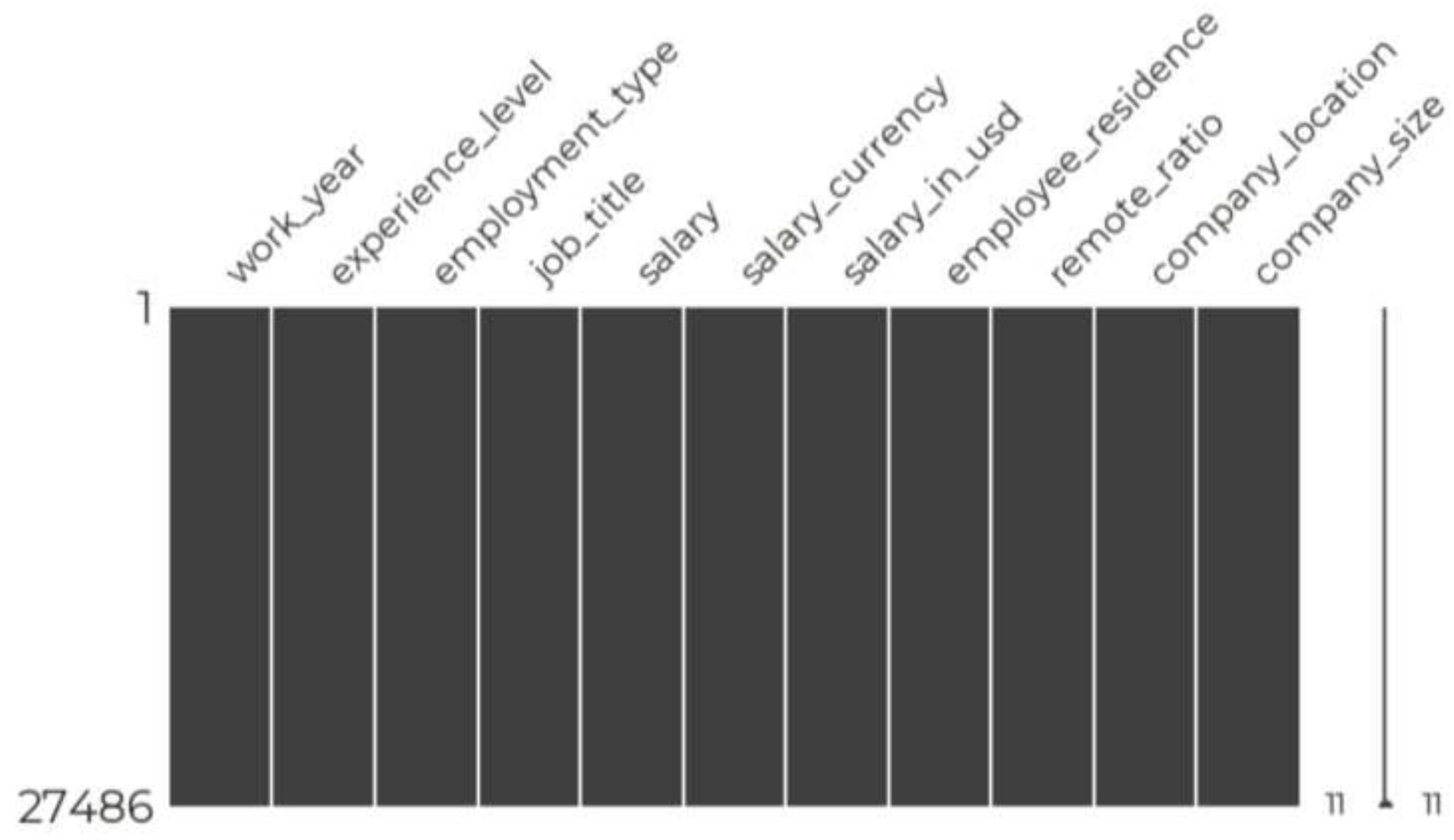
`experience_level`
`employment_type`
`company_size`
`remote_ratio`

```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   work_year           27486 non-null  int64
 1   experience_level    27486 non-null  object
 2   employment_type     27486 non-null  object
 3   job_title           27486 non-null  object
 4   salary              27486 non-null  int64
 5   salary_currency     27486 non-null  object
 6   salary_in_usd       27486 non-null  int64
 7   employee_residence  27486 non-null  object
 8   remote_ratio        27486 non-null  int64
 9   company_location    27486 non-null  object
 10  company_size        27486 non-null  object
dtypes: int64(4), object(7)
```

```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   work_year           27486 non-null  int64
 1   experience_level    27486 non-null  category
 2   employment_type     27486 non-null  category
 3   job_title           27486 non-null  object
 4   salary              27486 non-null  int64
 5   salary_currency     27486 non-null  object
 6   salary_in_usd       27486 non-null  int64
 7   employee_residence  27486 non-null  object
 8   remote_ratio        27486 non-null  category
 9   company_location    27486 non-null  object
 10  company_size        27486 non-null  category
dtypes: category(4), int64(3), object(4)
```

Any missing values?

No. There is no missing data
column homogeneity indicates that
there are no NULL values



Any duplicates?

Ratio:

- 45% of the data in the **horizontal** view (row) is repeated
- There is an overall extremely low uniqueness of values in the **vertical** relationship (column)

Assumed origin of repetitions:

Identical overlaps between variables arise due to (a) lack of unique identifiers and (b) frequently occurring values in each individual variable (popular combinations such as [2024, US, Data Analyst, EN])

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size	count
0	2023	SE	FT	Applied Scientist	136000	USD	136000	US	0	US	L	53
1	2024	MI	FT	Applied Scientist	222200	USD	222200	US	0	US	L	42
2	2024	MI	FT	Applied Scientist	136000	USD	136000	US	0	US	L	42
3	2023	SE	FT	Machine Learning Engineer	204500	USD	204500	US	0	US	M	34

Any inconsistencies?

Identical positions have different lengths for their titles, ignoring which can significantly affect the interpretation of the results

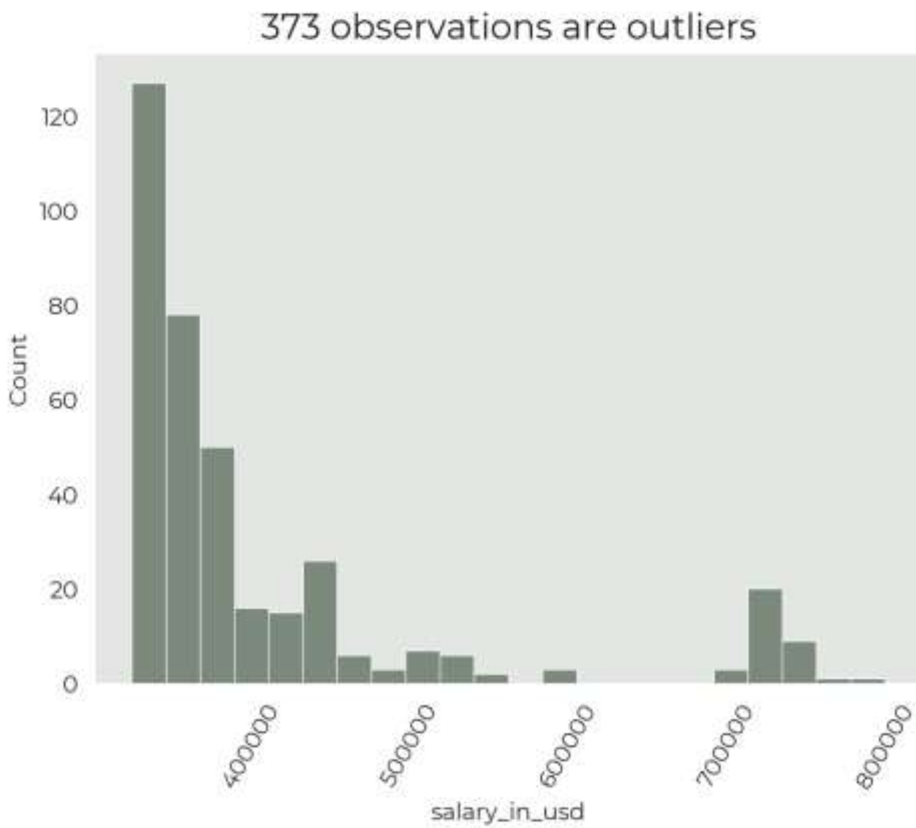
ML Ops = ML Operations

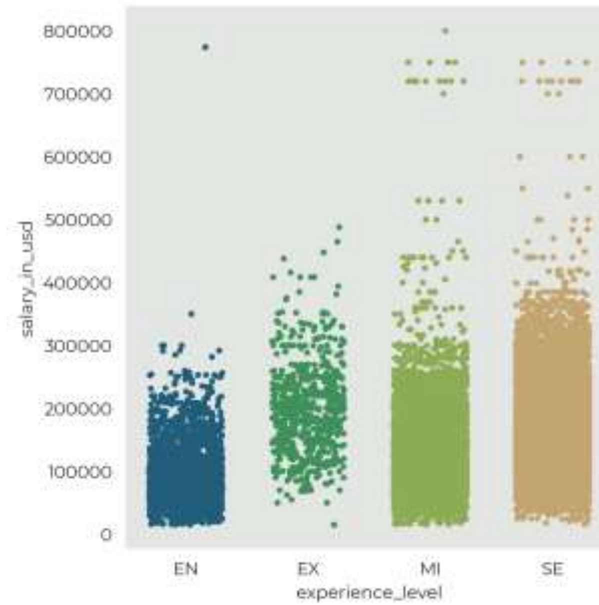
BI Analyst = Business Intelligence Analyst

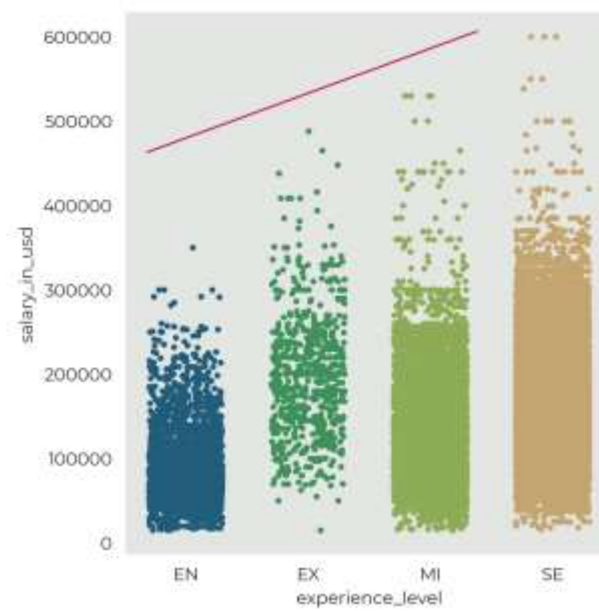
```
ds_copy.job_title.replace({'^Data Science Engineer$': 'Data Engineer',
                           '^Data Science Analyst$': 'Data Analyst',
                           '^Data Science Manager$': 'Data Manager',
                           'Machine Learning': 'ML',
                           '^ML Ops$|^ML Operations': 'MLOps',
                           'Artificial Intelligence': 'AI',
                           'Business Intelligence': 'BI',
                           '^BI Data Analyst$': 'BI Analyst',
                           '^Finance Data Analyst$': 'Financial Data Analyst',
                           '^Head of Data Science$': 'Head of Data',
                           '^Data Lead$': 'Head of Data',
                           '^Data Science Lead$': 'Head of Data Science',
                           '^Bear Robotics$': 'Robotics Engineer',
                           '^Data Analyst Lead$': 'Lead Data Analyst',
                           '^Data Scientist Lead$': 'Lead Data Scientist',
                           '^Data Science Tech Lead$': 'Tech Lead Data Science',
                           '^Data Science Director$': 'Director of Data Science',
                           '^Data Analytics Engineer$': 'Analytics Engineer',
                           '^Applied Data Scientist$': 'Applied Scientist'},
                           inplace=True, regex=True)
```

Any outliers?

Salaries are skewed to the right, so robust IQR would be an appropriate method for identifying outliers







Feature engineering (1/2)

It would be great to categorize `job_title` into major groups for deeper analysis

`job_title`:

- Data
- Engineer
- Scientist
- Analyst
- ML
- AI
- BI
- Manager
- Research
- Applied

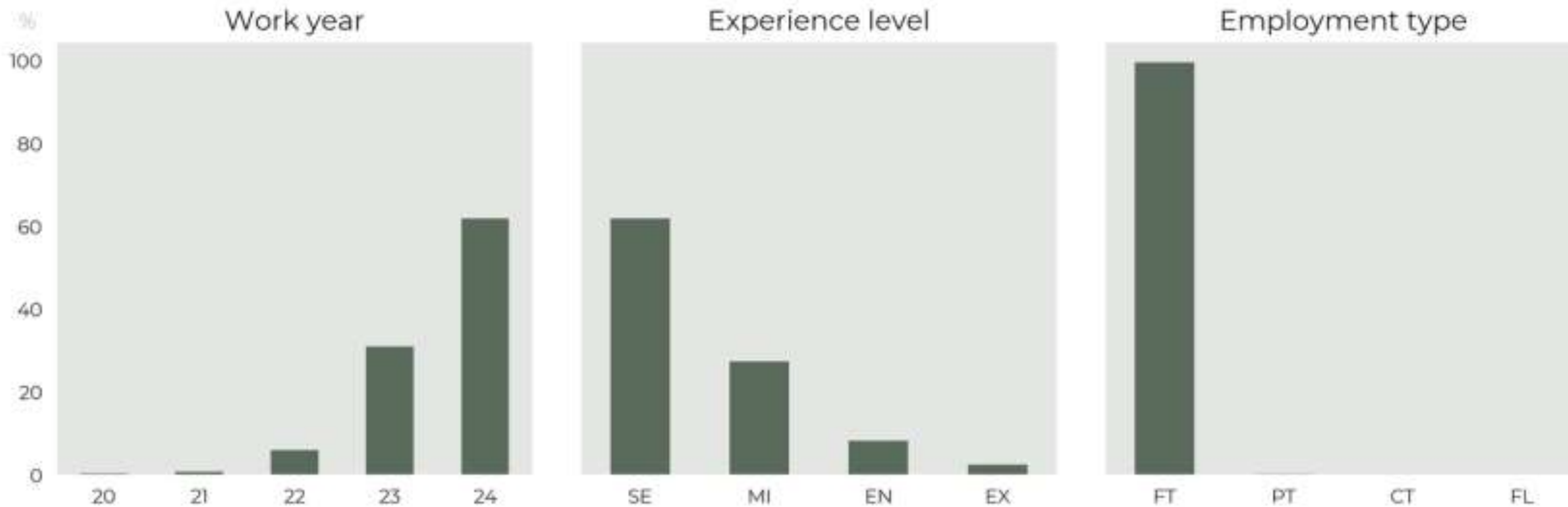
It would also be nice to segment dollar salaries into groups

`salary_in_usd`:
(in thousands)

- 0-50
- 50-100
- 100-150
- 150-200
- 200-250
- 250-300
- 300-350
- 350-400
- 400+

Analyze variable distribution

the initial mental model
of the dataset



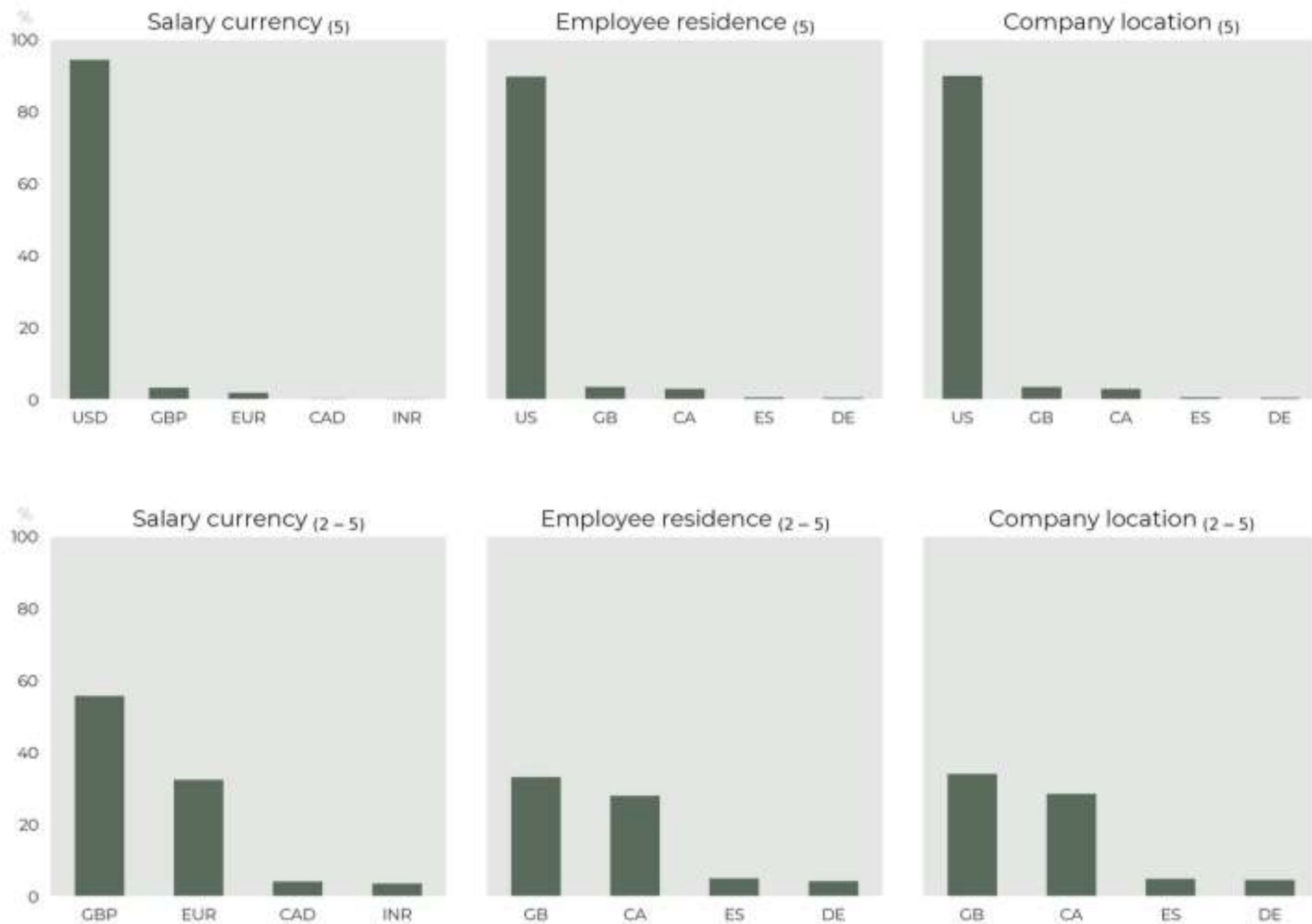
SE – Senior
MI – Middle
EN – Entry-level
EX – Executive-level

FT – Full-time
PT – Part-time
CT – Contract
FL – Freelance

The market demands highly skilled professionals (SE) totally immersed in business issues (FT)

Dependencies between variables will be considered later to identify a more explainable labor market

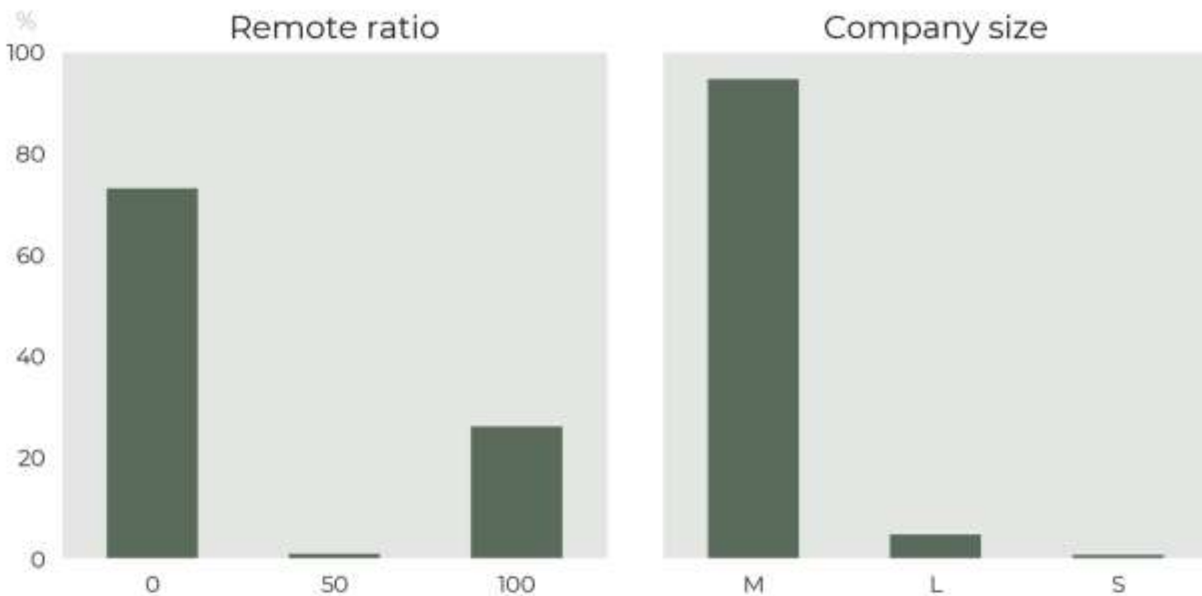
It's not a surprising fact that the US and the Dollar supply the bulk of the IT market



0 – No remote work
(less than 20%)
50 – Hybird
100 – Fully remote
(more than 80%)

Working from home has become a new reality that you don't want to let go of...

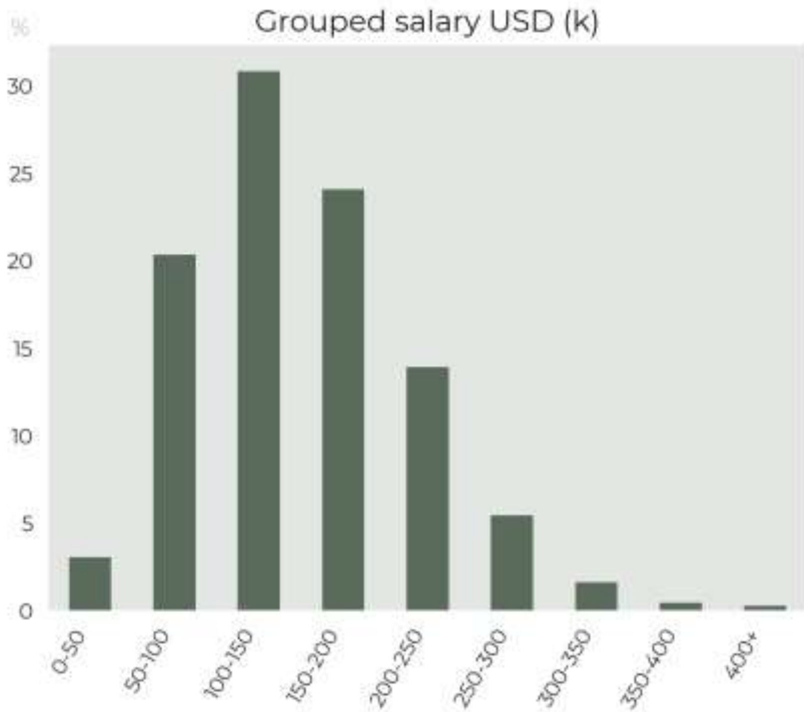
Either way, today's employers want to see employees **in person**



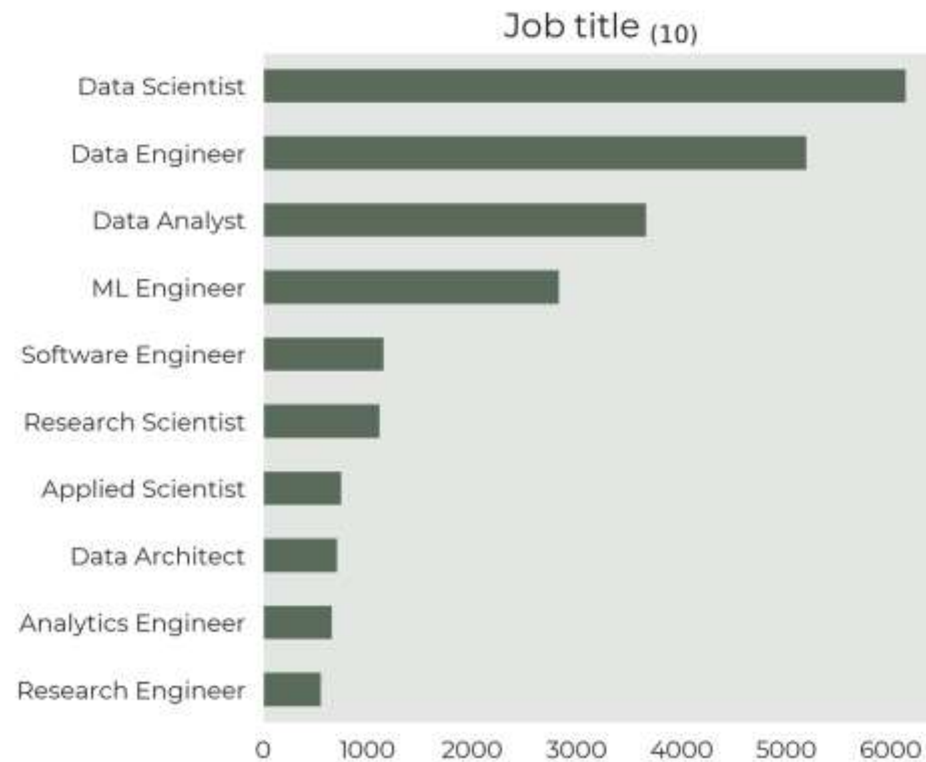
Stanford study: +13% productivity gains from working remotely

Employees who switched to telecommuting showed a productivity gain of 13%. Reduced breaks and sick days, as well as a quieter and more comfortable work environment, contributed to this increase

Salary distributions in relation to other variables will be covered later

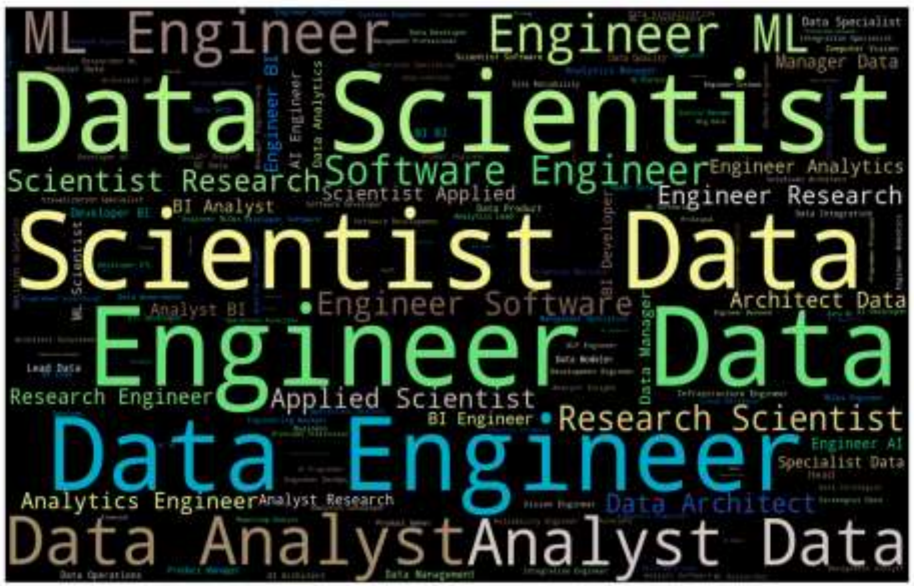


The clear demand for Data Scientists is also seen in the Research Scientist and Research Engineer pairing, as well as in the more practical occupation of Applied Scientist



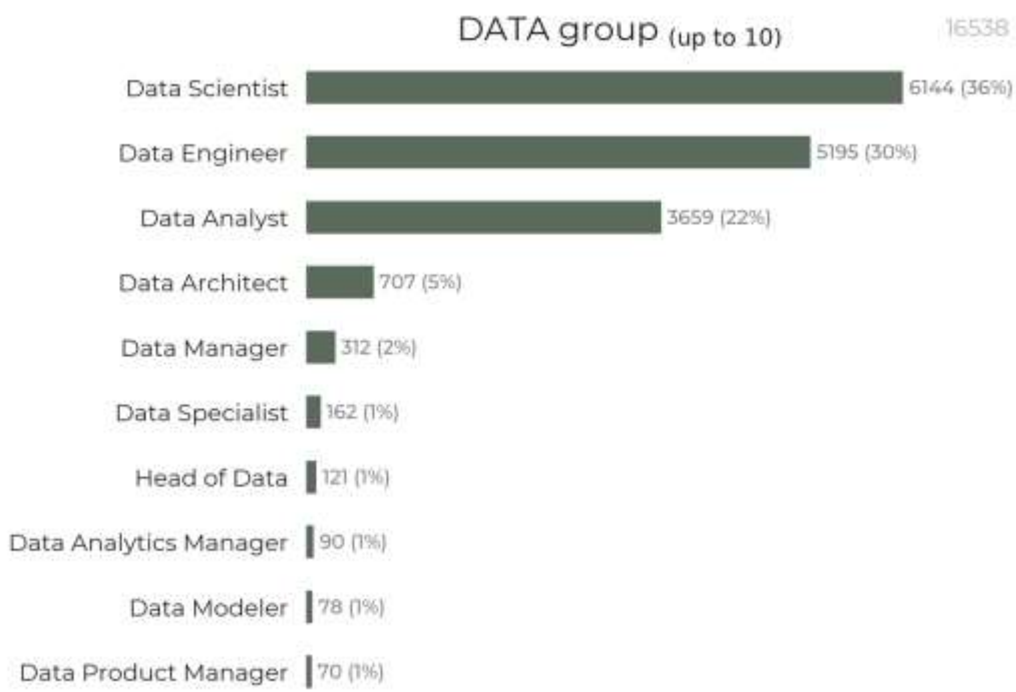
A visual representation of the demand volume

Considering the most encompassing first 4 position



Not including the first 4 positions

Analyzing job distribution across predetermined categories

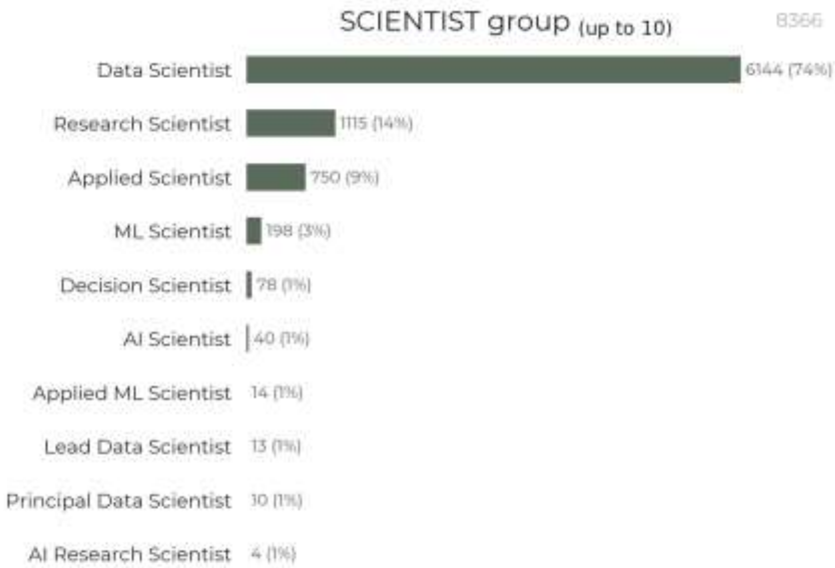


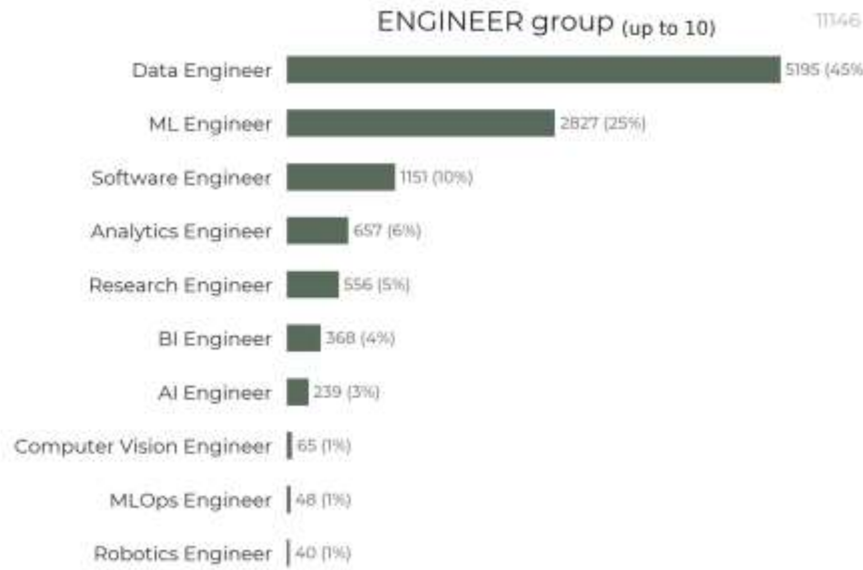
Based on this, we can assume the **immaturity** of companies in adapting data for data-driven decision making

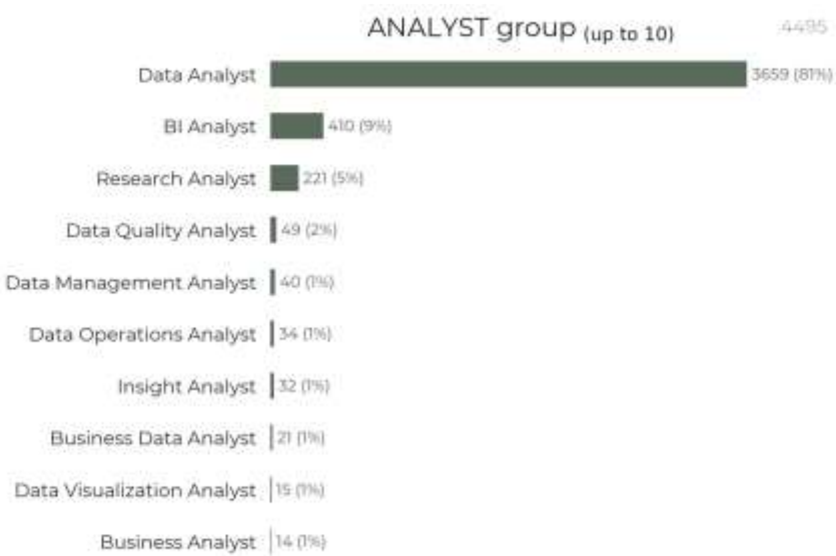
- There is a clear increase in demand for Data Architects
  a more fundamental and advanced level of work with structures, types and stores

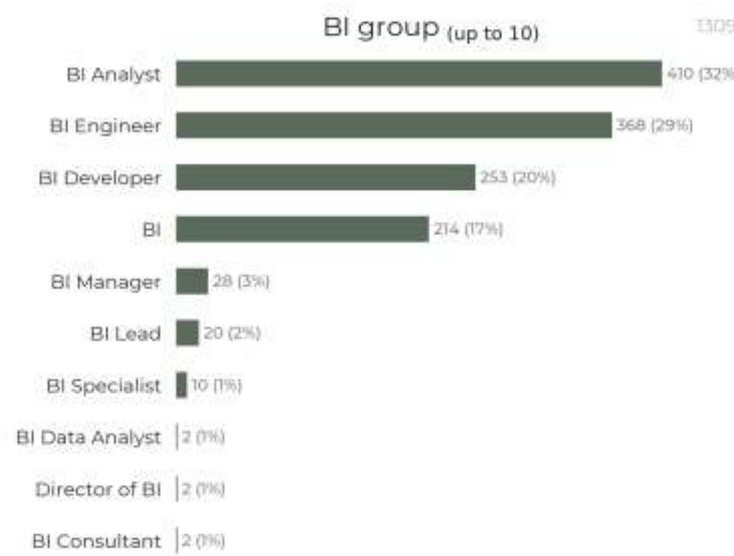
  The work of Data Architects lays the foundation for other data professionals such as analysts and scientists

- The Data Manager position lays the foundation for a managerial approach in working with data

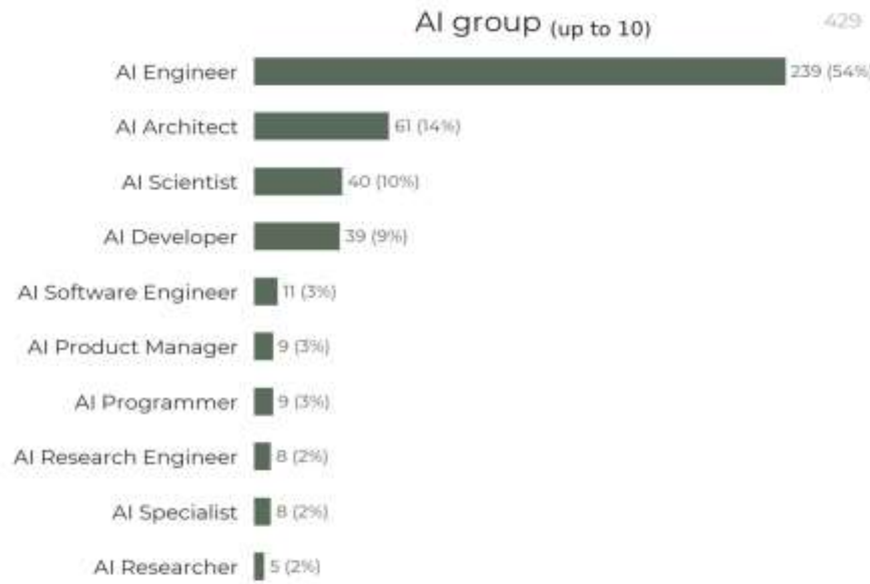

SCIENTIST group (up to 10)
8366
Data Scientist 6144 (74%)
Research Scientist 1115 (14%)
Applied Scientist 750 (9%)
ML Scientist 198 (3%)
Decision Scientist 78 (1%)
AI Scientist 40 (1%)
Applied ML Scientist 14 (1%)
Lead Data Scientist 13 (1%)
Principal Data Scientist 10 (1%)
AI Research Scientist 4 (1%)

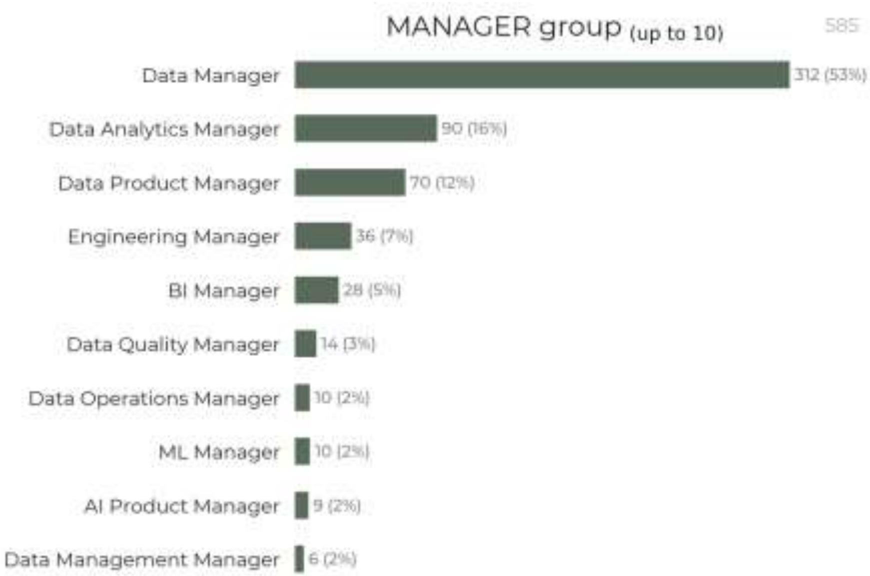

ENGINEER group (up to 10)
11146
Data Engineer 5195 (45%)
ML Engineer 2827 (25%)
Software Engineer 1151 (10%)
Analytics Engineer 657 (6%)
Research Engineer 556 (5%)
BI Engineer 368 (4%)
AI Engineer 239 (3%)
Computer Vision Engineer 65 (1%)
MLOps Engineer 48 (1%)
Robotics Engineer 40 (1%)

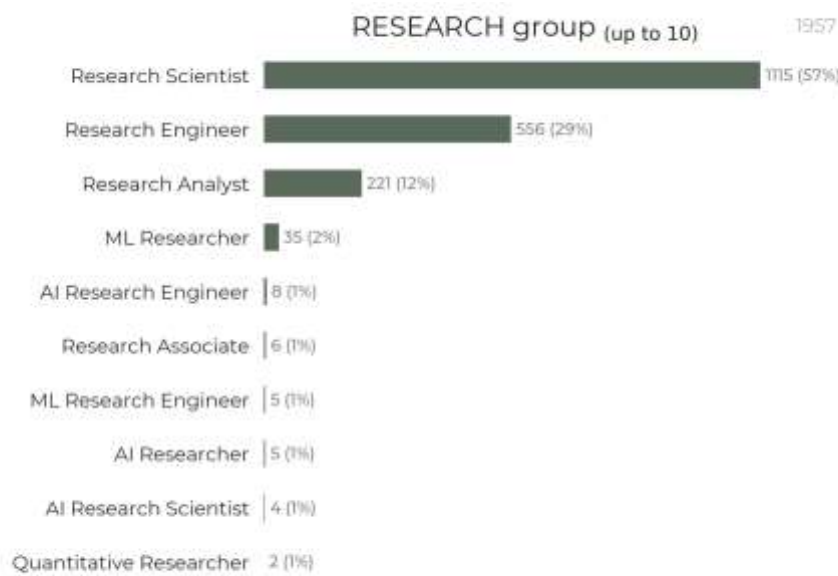

ANALYST group (up to 10)
4495
Data Analyst 3659 (81%)
BI Analyst 410 (9%)
Research Analyst 221 (5%)
Data Quality Analyst 49 (2%)
Data Management Analyst 40 (1%)
Data Operations Analyst 34 (1%)
Insight Analyst 32 (1%)
Business Data Analyst 21 (1%)
Data Visualization Analyst 15 (1%)
Business Analyst 14 (1%)


BI group (up to 10)
1309
BI Analyst 410 (32%)
BI Engineer 368 (29%)
BI Developer 253 (20%)
BI 214 (17%)
BI Manager 28 (3%)
BI Lead 20 (2%)
BI Specialist 10 (1%)
BI Data Analyst 2 (1%)
Director of BI 2 (1%)
BI Consultant 2 (1%)


ML group (up to 10)
3214
ML Engineer 2827 (87%)
ML Scientist 198 (7%)
MLOps Engineer 48 (2%)
ML Infrastructure Engineer 39 (2%)
ML Researcher 35 (2%)
ML Developer 18 (1%)
ML Software Engineer 15 (1%)
Applied ML Scientist 14 (1%)
ML Manager 10 (1%)
ML Modeler 10 (1%)


AI group (up to 10)
429
AI Engineer 239 (54%)
AI Architect 61 (14%)
AI Scientist 40 (10%)
AI Developer 39 (9%)
AI Software Engineer 11 (3%)
AI Product Manager 9 (3%)
AI Programmer 9 (3%)
AI Research Engineer 8 (2%)
AI Specialist 8 (2%)
AI Researcher 5 (2%)


MANAGER group (up to 10)
585
Data Manager 312 (53%)
Data Analytics Manager 90 (16%)
Data Product Manager 70 (12%)
Engineering Manager 36 (7%)
BI Manager 28 (5%)
Data Quality Manager 14 (3%)
Data Operations Manager 10 (2%)
ML Manager 10 (2%)
AI Product Manager 9 (2%)
Data Management Manager 6 (2%)


RESEARCH group (up to 10)
1957
Research Scientist 1115 (57%)
Research Engineer 556 (29%)
Research Analyst 221 (12%)
ML Researcher 35 (2%)
AI Research Engineer 8 (1%)
Research Associate 6 (1%)
ML Research Engineer 5 (1%)
AI Researcher 5 (1%)
AI Research Scientist 4 (1%)
Quantitative Researcher 2 (1%)

General observation

The market is actively saturating with **new specializations** such as:

Applied Scientist, Data Architect, AI Engineer, Research Analyst, ML Scientist, Data Manager and others

forming a multifaceted foundation around data science, bringing more opportunities for human expression in solving business problems

Less-developed groups (groups where the first position is extremely dominant):

APPLIED group, MANAGER group, AI and ML groups, ANALYST and SCIENTIST groups

Brainstorm & Goal Strategies

"Chance favors the connected mind."

— Steven Johnson

1. Brainstorming based on the goals

- Relationships
- Staff Demand
- Salary Trends
- Remote Work

2. Identifying goal achievement methods

Brainstorming based on the goals

Salary Trends

1. Compare overall and specific wages across years (2024 vs 2023)

Overall

- Compare Q1, Q2, and Q3 ratios
- Compare ratios of cumulative sums for each quartile range (Q0-Q1 … Q3-Q4), using samples with equal N
- Compare overall and sub-range standard deviations

Specific

1 variable

Non-working positions

- Top 5 countries
- Qualification
- Company size

>1 variable

- Qualification
  Top 5 most common
- Company size
  Top 5 most common

Working positions

- Top 5 most common
- Top 5 most common by salary groups
- Top 5 positive / negative ratios
- Top 5 common highest-paid (>250k) / lowest-paid (<70k) job titles
- Top 3 most common by group

2. Modeling a regression model to determine the most influential variables on wages

Relationships

Specific relationships will be considered after the tests have been performed

Staff Demand

Analyze demand for specialists across years (2024 vs 2023)

Overall

- Top 10

Specific

- Qualification
- Company size

Remote Work

Compare trends in remote work adoption

Identifying goal achievement methods

Relationships

1. Analyze relationships between categorical variables

- **Pearson correlation** for binarized variables for a clear comparison
- **V Cramer's Coefficient** measures the strength of the association

2. Explore how salaries relate to categorical variables

- **Variance homogeneity** (box-plots or Levene's test)
- **Kruskal-Wallis test**
- Post-hoc: **Dunn's test** (if KW significant)

Salary Trends

1. Compare overall and specific wages across years (2024 vs 2023)

There are no specific methods. Simple data filter and output

2. Modeling a regression model to determine the most influential variables on wages

- Use the most appropriate and efficient model (**lazypredict** library will help decide)
- Use **Partial Dependence Plots** (PDPs)

Staff Demand & Remote Work

There are no specific methods. Simple data filter and output

Realization

1. Feature engineering (2/2)
2. Relationships
3. Salary Trends
4. Staff Demand
5. Remote Work

Feature engineering (2/2)

For a machine learning model to effectively explain the target variable (salary), comprehensive data covering true variable relationships is essential

Outliers, unprocessed, and missing values can significantly distort the model's results

Creating new features from existing data can enhance the explainability of salary data, potentially including external sources

Often, simply removing unnecessary information (possible noise) is sufficient

In my case, I did it this way:

I got rid of noise and imbalanced features (overweight/underweight by number of observations) in such variables as `job_title` and `company_location` by defining features with number of observations ≤ 15 as "Other", while the ranges from 15 to 29 were increased to 30 by bootstrap without replacement

The above variables had many values in the ranges from 2 to 10

Extracted such new data from `company_location`, `experience_level` and `job_title` variables as:

- Continent + Its Directions
  Asia | West, Europe | East ...

- Conditional median by country
  Salary ranges: 0-30k, 30-60k ...

- Salary above/below the median of the [qualification + position]
  For example, a [Junior + Analyst] might earn above the median in one country but below it in another

- Grouping positions into their origin group
  ML Engineer = "Engineer",
  Applied Scientist = "Scientist"

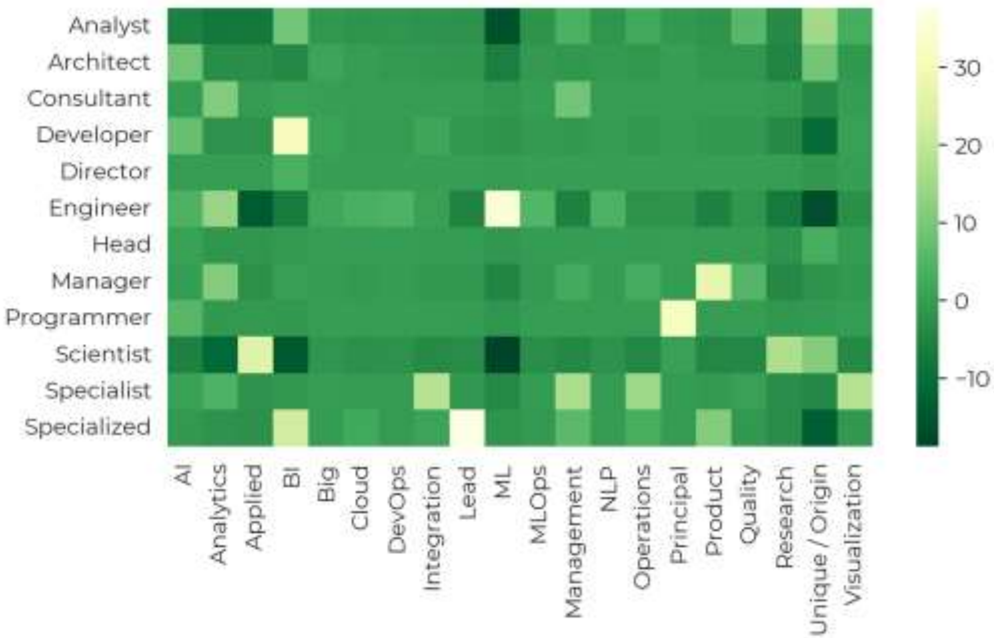
- Highly specialized positions
  ML Engineer = "ML",
  Applied Scientist = "Applied"

country_continent	origin_group	proportion
Australia	Analyst	0.34
	Engineer	0.32
	Scientist	0.21
	Architect	0.04
	Specialist	0.02
	Specialized	0.02

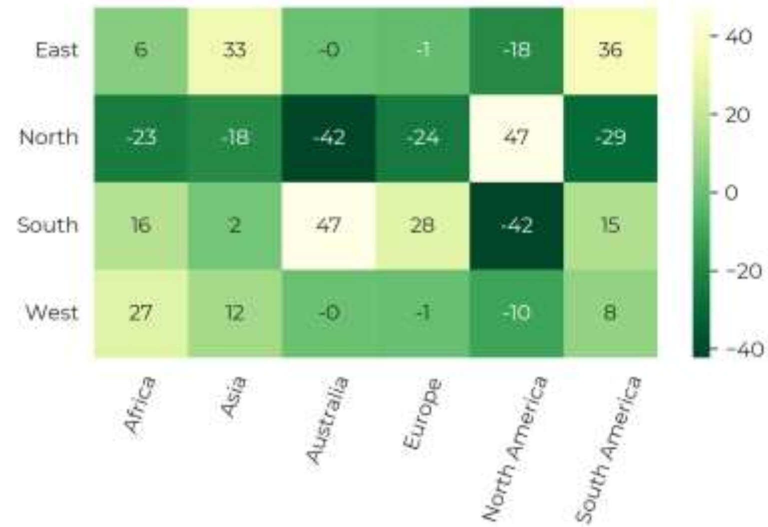
Interestingly, analysts are preferred more in Australia than in other countries

country_continent	country_direction	country_median	count
Africa	East	(n<30)	5
	North	>120k	30
		(n<30)	4
	South	30-60k	30
	West	(n<30)	12
Asia	East	(n<30)	40
	North	30-60k	76
		<30k	30
		(n<30)	14
	South	(n<30)	8
	West	(n<30)	8
Australia	South	90-120k	72
		(n<30)	19
	North	60-90k	1216
		(n<30)	93
		30-60k	90

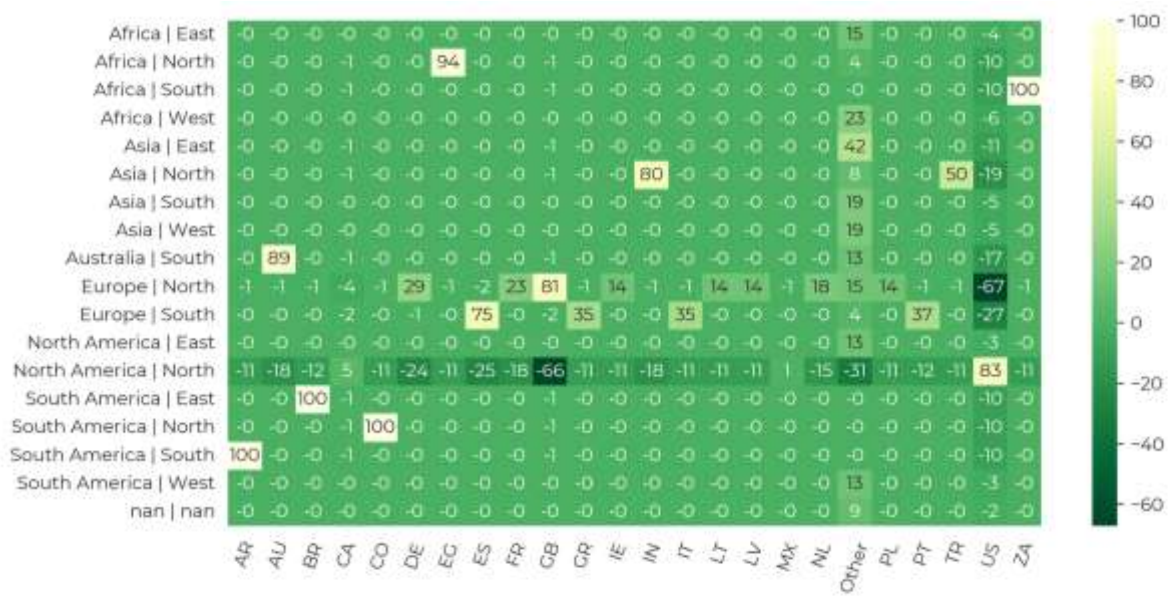
	origin_group	specialization	experience_level	company_location	salary_in_usd	jobexp_median
21425	Analyst	Unique / Origin	MI	US	70000	Below
23837	Engineer	Research	SE	US	189110	Above
20334	Scientist	Unique / Origin	SE	US	204500	Above
21765	Scientist	Unique / Origin	MI	US	115000	Below
27621	Engineer	NLP	MI	IN	32834	(n<15)



Analyst
Architect
Consultant
Developer
Director
Engineer
Head
Manager
Programmer
Scientist
Specialist
Specialized
AI
Analytics
Applied
BI
Big
Cloud
DevOps
Integration
Lead
ML
MLOps
Management
NLP
Operations
Principal
Product
Quality
Research
Unique / Origin
Visualization
30
20
10
0
-10



East
North
South
West
6 33 -0 -1 -18 36
-23 -18 -42 -24 47 -29
16 2 47 28 -42 15
27 12 -0 -1 -10 8
Africa
Asia
Australia
Europe
North America
South America
40
20
0
-20
-40



Africa | East
Africa | North
Africa | South
Africa | West
Asia | East
Asia | North
Asia | South
Asia | West
Australia | South
Europe | North
Europe | South
North America | East
North America | North
South America | East
South America | North
South America | South
South America | West
nan | nan
AR
AU
BR
CA
CO
DE
EC
ES
FR
GB
GR
IE
IN
IT
LT
LV
MX
NL
Other
PL
PT
TR
US
ZA
100
80
60
40
20
0
-20
-40
-60

Relationships

Analyze relationships between categorical variables

- **Pearson correlation** shows the level of linear dependence (relationship) between variables
- **Cramer's V coefficient** measures the strength of this relationship

Both methods give results ranging from 0 to 1, where 0 means no relationship / strength

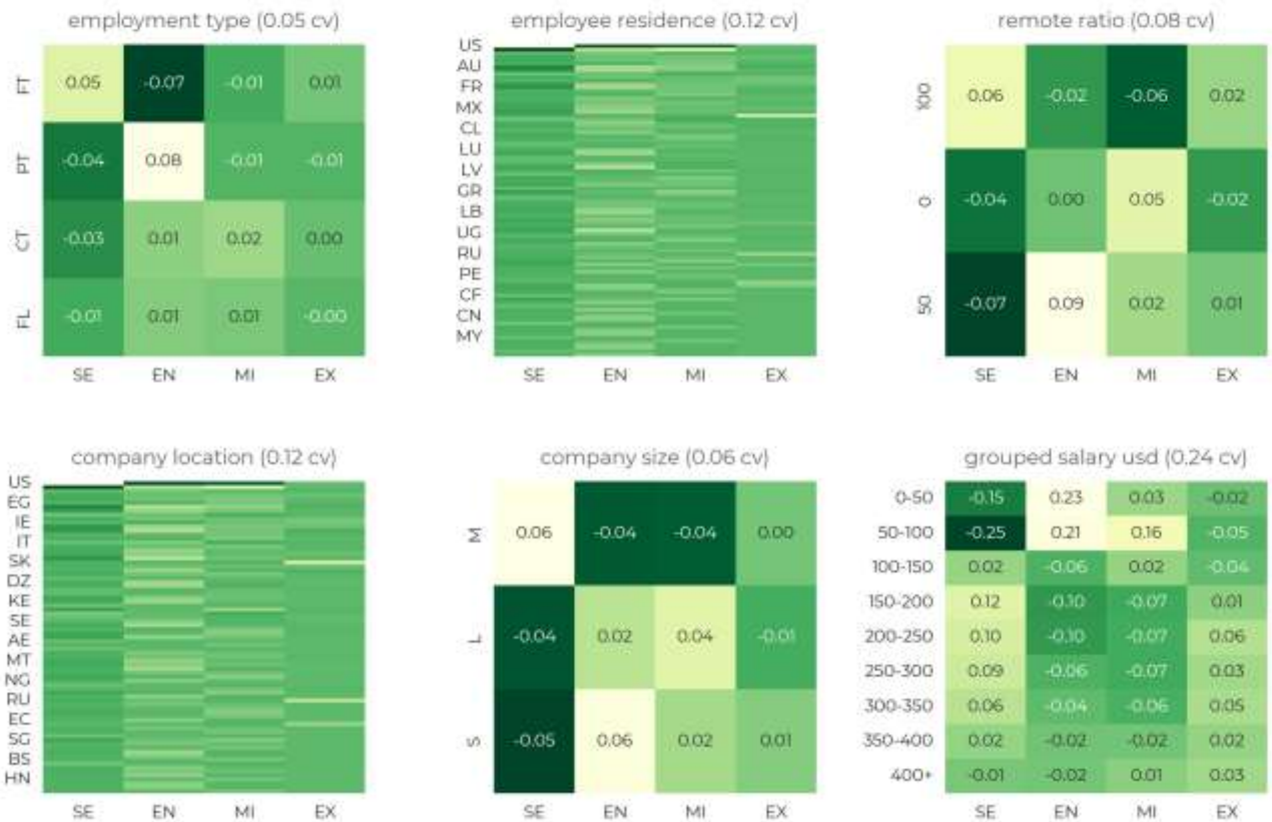
		cramerv
xx	xy	
work_year	experience_level	0.10
	employment_type	0.06
	salary_currency	0.18
	employee_residence	0.24
	remote_ratio	0.29
	company_location	0.23
	company_size	0.28
	grouped_salary_usd	0.09

Example: Age and income

- Correlation: moderate/high (income increases with age)
- Correlation coefficient: 0.6 (approximately)
- Weak correlation: many exceptions (education, industry, decisions)
- Cramer coefficient: 0.3 (approximately)

Experience level

cramer_v mean – 0.11 (weak)

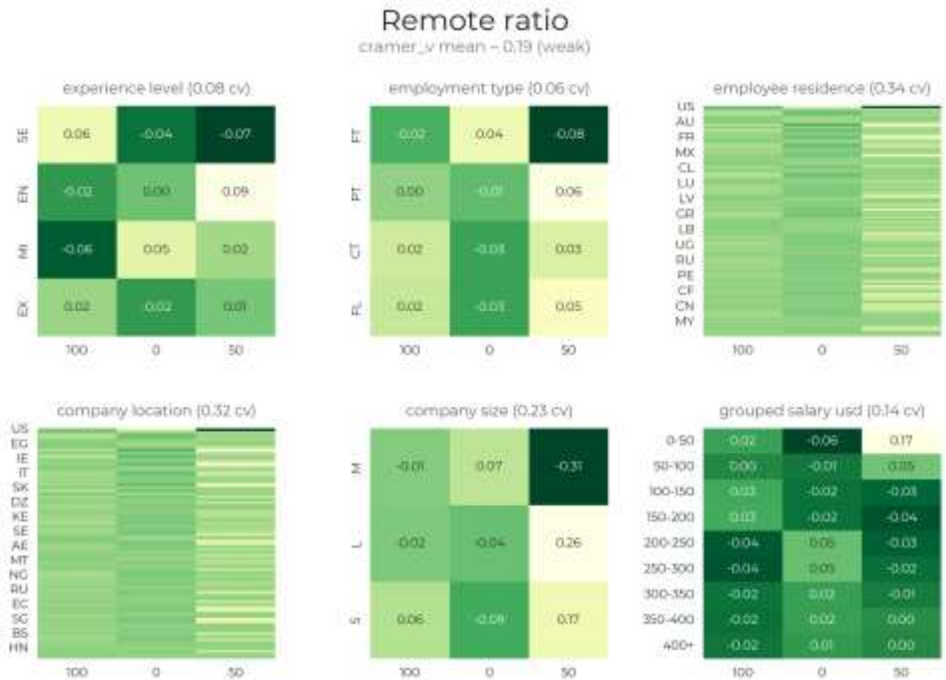


- Entry-level positions **tend to be part-time and hybrid**
- Senior experience levels are the **U.S. standard**; juniors and mid-levels are **less focused**
- Entry-level experience has **weak correlation** with companies / employees, except in North America
- Salary distribution by experience level is self-explanatory

Employment type

cramer_v mean – 0.16 (weak)

- Full-time employment in Data Science is **distinct** from other types (PT, CT, FL) globally
- Medium-sized companies (50-250 employees) **prefer full-time** staff
- Small companies (<50 employees) **favor contract and freelance** workers
- Salaries up to $50,000 / year are **common in part-time and freelance** roles



- Companies **outside America favor hybrid work**, while the U.S. prefers office work
- Hybrid work is **common in large and small companies**, but less so in medium-sized ones

  This trend impacts flexibility and satisfaction in large companies, and cost optimization and scaling in small companies

Frequent and apparent similarities between the locations of companies and employees suggest **the current level of remote work**, where employees work from one country for companies in another

Ultimately

the most desirable conditions (SE experience, M / L company size, FL employment, 0 / 100 remoteness, high salaries) are **satisfied in the US**

Others often use **less attractive, but business-friendly approaches** (EN experience, S company size, PT/FL employment, 50 remoteness, low salaries) based on correlations between countries and these conditions

Relationships

Identify wage relationships between variables

To decide on a test measuring the relationship between categorical and numerical data, certain conditions must be met for reliable interpretation

The choice is between parametric one-factor ANOVA and non-parametric Kruskal-Wallis tests

ANOVA is preferred due to its higher sensitivity, allowing it to detect deeper and more accurate relationships compared to non-parametric tests

To engage this power, it is necessary to ensure that these conditions are met:

- Independence of observations
- Normality of distribution
- Homogeneity of variances

```
P-value by Levene:
   (if n > 30)

0.210 - work_year (5/5)
  Skewness: [2.31, 1.52, 0.53, 0.6, 0.85]
0.000 - origin_group (10/12)
0.000 - specialization (18/20)
0.008 - experience_level (4/4)
0.009 - company_size (3/3)
0.136 - employment_type (3/4)
  Skewness: [2.12, 0.77, 1.48]
0.549 - remote_ratio (3/3)
  Skewness: [0.8, 2.47, 0.64]
0.000 - employee_residence (22/22)
0.000 - company_location (24/24)
0.000 - country_continent (6/6)
0.000 - country_direction (3/4)
0.000 - country_median (6/6)
0.000 - jobexp_median (3/3)
0.000 - salary_currency (6/6)
0.000 - grouped_salary_usd (9/9)
```

Levene's test determines the uniformity of variance among variables (homogeneity condition). Here, only three variables have similar variance among their categories, but the level of skewness (distribution condition) **prevents ANOVA application** due to diverse distributions

Categories with ≥ 30 observations were considered for accuracy

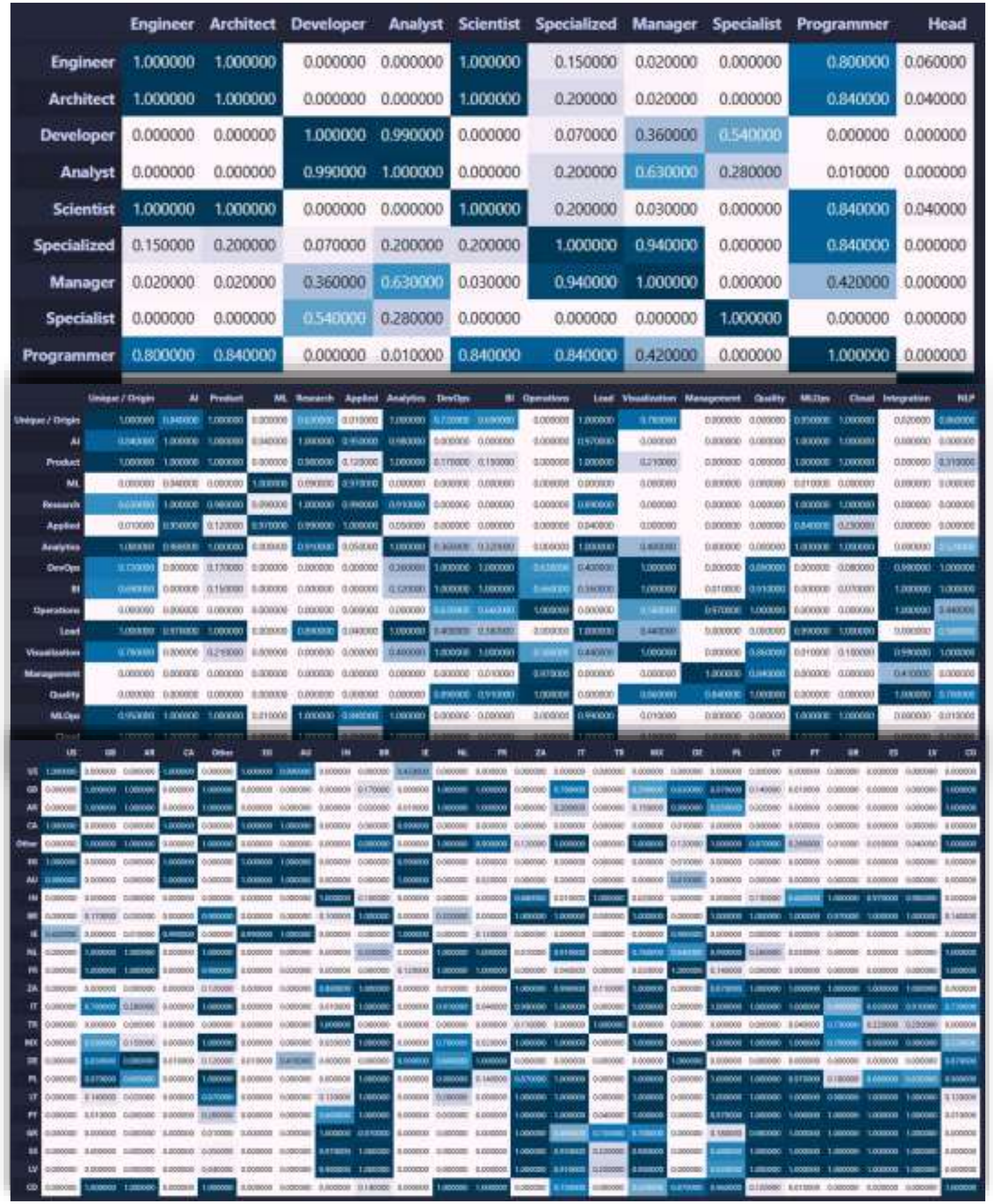
Logarithmizing salaries for three variables could achieve a more normal distribution for ANOVA, but this wasn't done due to the total number of variables not meeting the condition

```
P-value by Kruskal-Wallis:
      (if n > 30)

0.000 - work_year (5/5)
0.000 - origin_group (10/12)
0.000 - specialization (18/20)
0.000 - experience_level (4/4)
0.000 - company_size (3/3)
0.000 - employment_type (3/4)
0.000 - remote_ratio (3/3)
0.000 - employee_residence (22/22)
0.000 - company_location (24/24)
0.000 - country_continent (6/6)
0.000 - country_direction (3/4)
0.000 - country_median (6/6)
0.000 - jobexp_median (3/3)
0.000 - salary_currency (6/6)
0.000 - grouped_salary_usd (9/9)
```

The variable categories clearly contain different median salaries, which is not surprising

We are interested in specifics, particularly the **post-hoc Dunn's test**, which is applied after the Kruskal-Wallis H-test. This test would clarify the relationship of median salaries between categories, as **Kruskal-Wallis does not provide this information**



	Engineer	Architect	Developer	Analyst	Scientist	Specialized	Manager	Specialist	Programmer	Head
Engineer	1.000000	1.000000	0.000000	0.000000	1.000000	0.150000	0.020000	0.000000	0.800000	0.060000
Architect	1.000000	1.000000	0.000000	0.000000	1.000000	0.200000	0.020000	0.000000	0.840000	0.040000
Developer	0.000000	0.000000	1.000000	0.990000	0.000000	0.070000	0.360000	0.540000	0.000000	0.000000
Analyst	0.000000	0.000000	0.990000	1.000000	0.000000	0.200000	0.630000	0.280000	0.010000	0.000000
Scientist	1.000000	1.000000	0.000000	0.000000	1.000000	0.200000	0.030000	0.000000	0.840000	0.040000
Specialized	0.150000	0.200000	0.070000	0.200000	0.200000	1.000000	0.940000	0.000000	0.840000	0.000000
Manager	0.020000	0.020000	0.360000	0.630000	0.030000	0.940000	1.000000	0.000000	0.420000	0.000000
Specialist	0.000000	0.000000	0.540000	0.280000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000
Programmer	0.800000	0.840000	0.000000	0.010000	0.840000	0.840000	0.420000	0.000000	1.000000	0.000000

In this presentation, the results of the previous test have more weight, but **not significantly**, based on the analysis of the variables. **I did not find any obvious clues for further investigation** of patterns that would not be covered in further stages of analysis, **only clarification of the full picture** regarding salary relations of the groups

1.00 – identity / similarity of salaries

Salary Trends

Compare overall and specific wages across years (2024 vs 2023)

Overall

Compare Q1, Q2, and Q3 ratios

	2023	2024	Ratios (%)
Q1	108000	105000	2.9
Q2	145000	146100	-0.8
Q3	190000	198000	-4.0

The displayed averages of quartile intervals are bootstrap results because different years contain different numbers of observations within these intervals and a single common value is needed for accuracy

Compare ratios of means for each quartile range (Q0-Q1 ... Q3-Q4) using samples with equal N

	2023	2024	Ratios (%)
Q0-Q1	77865	78524	-0.8
Q1-Q2	128231	126171	1.6
Q2-Q3	166163	170482	-2.5
Q3-Q4	237929	245914	-3.2

Compare overall and sub-range standard deviations

- Most "noise" comes from the extreme high-salary range, while the Q0-Q1 range is more concentrated
- In 2024, **the labor market's volatility is 5% higher**, indicating more opportunities and greater competition
- The Q2-Q3 range is less volatile due to the large number of positions accumulated in this interval

Overall ratio (23 to 24): -5.22%

	2023	2024	Ratios (%)
Q0-Q1	21087	18631	13.2
Q1-Q2	10890	12166	-10.5
Q2-Q3	13555	14349	-5.5
Q3-Q4	41334	47513	-13.0

There is no downward trend in salaries in 2024

None of the recent negative global phenomena have significantly affected the financial position of the market to date; this reflects **its dominance over external factors**, which is a reference point for further interpretations

The increased variation in the different salary groups in the 24th year compared to the 23rd year suggests **new directions are emerging, possible salary shifts**, and in general the adaptation of companies to new labor conditions

Specific (1 variable)
Non-working positions

Top 5 countries (company location)

- Germany's salary growth in 2024 was 4.7%, nearly double that of 2023
  IT job openings in Germany increased by 50% year-on-year in 2024
  Germany is heavily investing in digitalization and technological innovation
- Interpreting "Other" is difficult due to insufficient country data, but DS wages are expected to rise in 2024
- Canada's economy is slowing in 2024, leading to lower wages in some sectors, including IT
  Many companies have cut IT and Data Science budgets, impacting salaries
- The large gap in UK DS salaries is due to revised budgets and market overheating
- The US remains stable and attractive

company_location	23_(mean)	24_(mean)	Ratio_(%)
DE	111648.0	129452.0	15.9
Other	76699.0	86609.0	12.9
US	159093.0	159838.0	0.5
CA	150155.0	137494.0	-8.4
GB	105356.0	76425.0	-27.5

Despite the Data Science hype, **the market is leveling off**. Previously inflated salaries are being re-evaluated, and **hiring is shifting towards employer terms due to oversupply**, even with a general upward salary trend

Qualification

experience_level	23_(mean)	24_(mean)	Ratio_(%)
MI	121826	140214	15.1
EN	90094	99370	10.3
EX	190775	202469	6.1
SE	164797	171738	4.2

- **Without a strong US influence** on the results of this comparison, SE level specialists started to get paid less (113k → 107k = -5%), while in the US the difference is 8000 in the positive direction (160k → 168k = 5%)
- EN (especially), MI and EX are experiencing wage increases across the board

  EN (43k → 56k = 23%)
  MI (70k → 83k = 15%)
  EX (135k → 160k = 15%)

Company size

company_size	23_(mean)	24_(mean)	Ratio_(%)
S	85336	97575	14.3
L	148482	160398	8.0
M	153616	155750	1.4

- **Excluding the US impact,** salaries in companies of sizes L and M were virtually unchanged (L = +0.1% and M = -3.5%), **while S found a significant increase** of 25% (61k → 81k)
- Smaller companies (<50 employees) are actively involved in modern solutions, at least to keep up with the mainstream

Specific (1 variable)
Working positions

Top 5 most common

job_title	23_(mean)	24_(mean)	Ratio_(%)
ML Engineer	191246.0	196827.0	2.9
Research Scientist	188938.0	193053.0	2.2
Data Engineer	148978.0	150146.0	0.8
Data Analyst	108751.0	107567.0	-1.1
Data Scientist	163205.0	160040.0	-1.9

origin_group	23_(mean)	24_(mean)	Ratio_(%)
Specialized	120587.0	142640.0	18.3
Manager	121358.0	139126.0	14.6
Specialist	93371.0	101612.0	8.8
Developer	114432.0	121243.0	6.0
Head	201580.0	209048.0	3.7

Top 5 most common by salary groups

- up to 100k
- 100k – 200k
- 200k – 300k
- 300k+

(to_100k) job_title	23_(mean)	24_(mean)	Ratio_(%)
Data Scientist	70831.0	78231	10.4
ML Engineer	75348.0	82738	9.8
Data Engineer	77659.0	78704	1.3
Data Analyst	74395.0	73470	-1.2
Research Scientist	nan	83285	nan

(100-200k) job_title	23_(mean)	24_(mean)	Ratio_(%)
Data Analyst	132795.0	135881	2.3
Data Engineer	145299.0	146956	1.1
ML Engineer	157369.0	157277	-0.1
Research Scientist	155059.0	154970	-0.1
Data Scientist	150664.0	149553	-0.7

(200-300k) job_title	23_(mean)	24_(mean)	Ratio_(%)
ML Engineer	237350.0	239057	0.7
Research Scientist	234648.0	235228	0.2
Data Scientist	235999.0	234168	-0.8
Data Analyst	220299.0	218327	-0.9
Data Engineer	235438.0	231706	-1.6

(300k+) job_title	23_(mean)	24_(mean)	Ratio_(%)
ML Engineer	328675.0	339803.0	3.4
Data Scientist	331735.0	335854.0	1.2
Data Engineer	nan	323778.0	nan
Research Scientist	nan	364601.0	nan
Software Engineer	nan	332625.0	nan

Top 5 positive / negative ratios

job_title	23_(mean)	24_(mean)	Ratio_(%)
Research Analyst	95890.0	127986.0	33.5
Other	118177.0	139246.0	17.8
Research Engineer	178581.0	188586.0	5.6
Data Modeler	128412.0	133538.0	4.0
ML Engineer	191358.0	197090.0	3.0

job_title	23_(mean)	24_(mean)	Ratio_(%)
Decision Scientist	165926.0	128623.0	-22.5
BI Engineer	149933.0	126790.0	-15.4
ML Scientist	182209.0	165633.0	-9.1
Data Manager	109211.0	99659.0	-8.7
Data Analytics Manager	148530.0	137380.0	-7.5

- A Research Analyst focuses on qualitative data using surveys and interviews
- A Data Analyst works with quantitative data and large datasets using statistical analysis and data mining
- Research Analysts are common in marketing, social sciences, and politics, while Data Analysts are prevalent in IT, finance, and e-commerce

Given the high salaries among Research Analysts, this position is likely to see **further financial growth**

Top 5 common highest-paid (>250k) / lowest-paid (<70k) job titles

(>250k) job_title	23_(mean)	24_(mean)	Ratio_(%)
Research Scientist	291086.0	300924.0	3.4
ML Engineer	288651.0	294605.0	2.1
Data Engineer	277725.0	282739.0	1.8
Applied Scientist	269105.0	268370.0	-0.3
Data Scientist	288717.0	285361.0	-1.2

(<70k) job_title	23_(mean)	24_(mean)	Ratio_(%)
Other	48030.0	55771.0	16.1
Data Scientist	48983.0	56071.0	14.5
Data Analyst	54918.0	55430.0	0.9
Data Engineer	57340.0	55861.0	-2.6
BI Analyst	nan	59487.0	nan

Top 3 most common by group

Data

job_title	23_(mean)	24_(mean)	Ratio_(%)
Data Engineer	149535.0	150360.0	0.6
Data Analyst	108828.0	107776.0	-1.0
Data Scientist	163665.0	160225.0	-2.1

Engineer

job_title	23_(mean)	24_(mean)	Ratio_(%)
ML Engineer	191331.0	197649.0	3.3
Data Engineer	149789.0	150203.0	0.3
Software Engineer	nan	187688.0	nan

Scientist

job_title	23_(mean)	24_(mean)	Ratio_(%)
Research Scientist	188776.0	192937.0	2.2
Applied Scientist	187435.0	186880.0	-0.3
Data Scientist	164014.0	160307.0	-2.3

Analyst

job_title	23_(mean)	24_(mean)	Ratio_(%)
Research Analyst	96683.0	127916.0	32.3
Data Analyst	108862.0	107950.0	-0.8
BI Analyst	118272.0	110619.0	-6.5

ML

job_title	23_(mean)	24_(mean)	Ratio_(%)
ML Engineer	191659.0	196998.0	2.8
ML Scientist	182407.0	166217.0	-8.9
MLOps Engineer	nan	172217.0	nan

AI

job_title	23_(mean)	24_(mean)	Ratio_(%)
AI Engineer	161401.0	160402.0	-0.6
AI Architect	nan	215027.0	nan

BI

job_title	23_(mean)	24_(mean)	Ratio_(%)
BI Developer	111749.0	105350.0	-5.7
BI Analyst	118169.0	110552.0	-6.4
BI Engineer	149768.0	127295.0	-15.0

Manager

job_title	23_(mean)	24_(mean)	Ratio_(%)
Data Analytics Manager	148542.0	137319.0	-7.6
Data Manager	110122.0	99907.0	-9.3
Data Product Manager	nan	150456.0	nan

Research

job_title	23_(mean)	24_(mean)	Ratio_(%)
Research Analyst	96543.0	128199.0	32.8
Research Engineer	178307.0	188744.0	5.9
Research Scientist	188858.0	192484.0	1.9

Specific (>1 variable)

Qualification

Top 5 most common

Entry-level qualifications **massively adjust the salary cap** for Data Science majors

EN + job_title	23_(mean)	24_(mean)	Ratio_(%)
Research Analyst	77519.0	114651.0	47.9
Data Analyst	76655.0	92331.0	20.5
Data Engineer	90758.0	107279.0	18.2
Data Scientist	91558.0	104428.0	14.1
BI Analyst	nan	102264.0	nan

MI + job_title	23_(mean)	24_(mean)	Ratio_(%)
Data Scientist	126798.0	142633.0	12.5
ML Engineer	154840.0	169816.0	9.7
Data Analyst	95188.0	101394.0	6.5
Data Engineer	122805.0	129489.0	5.4
Research Scientist	177057.0	174655.0	-1.4

SE + job_title	23_(mean)	24_(mean)	Ratio_(%)
Data Analyst	120762.0	126987.0	5.2
ML Engineer	198075.0	206007.0	4.0
Data Engineer	157891.0	161414.0	2.2
Data Scientist	172087.0	171976.0	-0.1
Software Engineer	nan	196676.0	nan

EX + job_title	23_(mean)	24_(mean)	Ratio_(%)
Head of Data	217456.0	230296.0	5.9
Data Engineer	189250.0	189091.0	-0.1
Data Scientist	209776.0	209270.0	-0.2
BI	nan	204868.0	nan
ML Engineer	nan	237093.0	nan

Company size

Top 5 most common

M + job_title	23_(mean)	24_(mean)	Ratio_(%)
ML Engineer	192292.0	197642.0	2.8
Data Engineer	149486.0	150320.0	0.6
Data Analyst	109592.0	107424.0	-2.0
Data Scientist	164688.0	159772.0	-3.0
Software Engineer	nan	187717.0	nan

L + job_title	23_(mean)	24_(mean)	Ratio_(%)
Data Scientist	109381.0	168196.0	53.8
Data Engineer	138904.0	140045.0	0.8
Applied Scientist	185006.0	186294.0	0.7
BI Engineer	139382.0	134198.0	-3.7
Other	102378.0	nan	nan

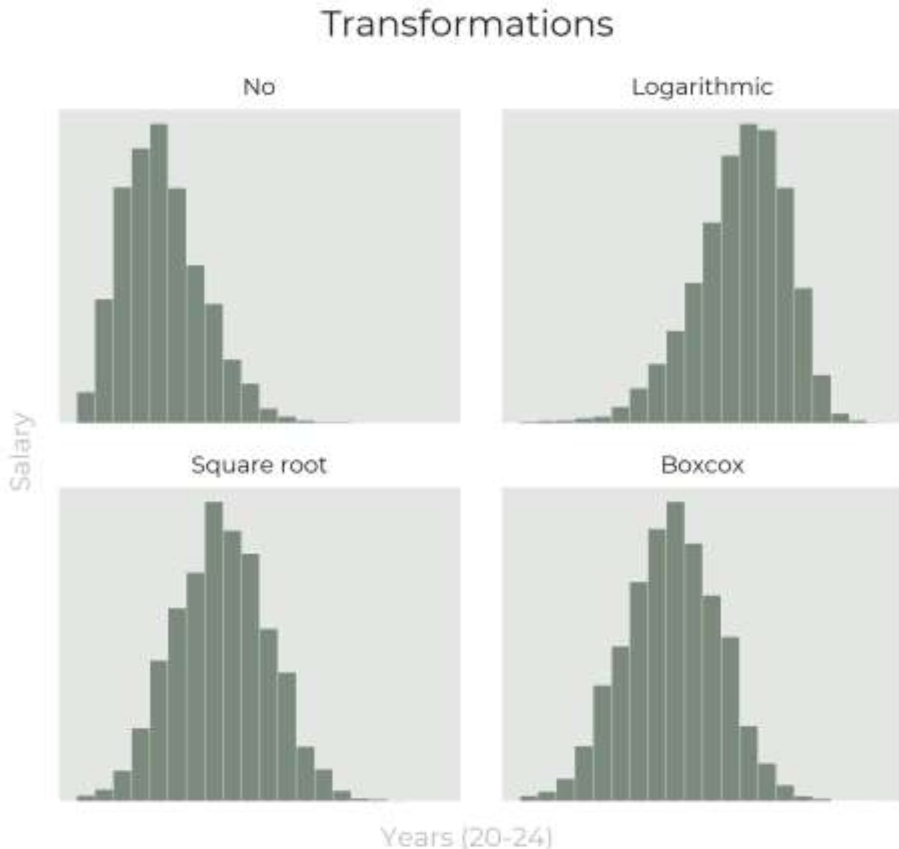
An unexpected observation.
Data Scientist, a particularly **preferred position by large companies**, is, according to demand, the only specialty with a significant **shift in salary by an average of 54%** (a clear shift is observed both in America and Europe)

Salary Trends

Modeling a regression model to determine the most influential variables on wages

The **libraries** involved at this stage of the study:

- **joblib** - to save/use the saved model after a long iteration over the hyperparameter grid in search of the most efficient one, so as not to return to this step
- **sklearn** - for building and evaluating model performance, for processing categorical data, for determining the best performing model parameters
- **lazypredict** - for determining the most efficient model and the resulting time savings
- **xgboost** - machine learning regression algorithm



Square root / Box-Cox transformation:

- Compresses large values, stretches small values (info loss, noise amplification)
- More sensitive to outliers than logarithmic transformation
- Logarithmic transformation better for salary prediction (better accounts for the typical income distribution)

I use the salary transformation here to (a) reduce the impact of high salary ranges, (b) improve interpretability, and (c) make the model more stable

After dividing the data into training (0.7) and test (0.3) data, converting the categorical data into numerical representation using OneHotEncoder, the final data was mass processed using various regression algorithms

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
LGBMRegressor	0.76	0.76	0.23	0.46
XGBRegressor	0.76	0.76	0.23	0.56
HistGradientBoostingRegressor	0.75	0.76	0.23	1.83
RandomForestRegressor	0.75	0.76	0.23	7.06
BaggingRegressor	0.75	0.75	0.24	0.94
GradientBoostingRegressor	0.74	0.75	0.24	7.82
NuSVR	0.74	0.74	0.24	67.24
SVR	0.74	0.74	0.24	44.20
ElasticNetCV	0.73	0.74	0.24	6.28
LassoCV	0.73	0.74	0.24	4.38

The most efficient in performance are algorithms such as: LightGBM and XGBoost. My choice fell on **the second option**

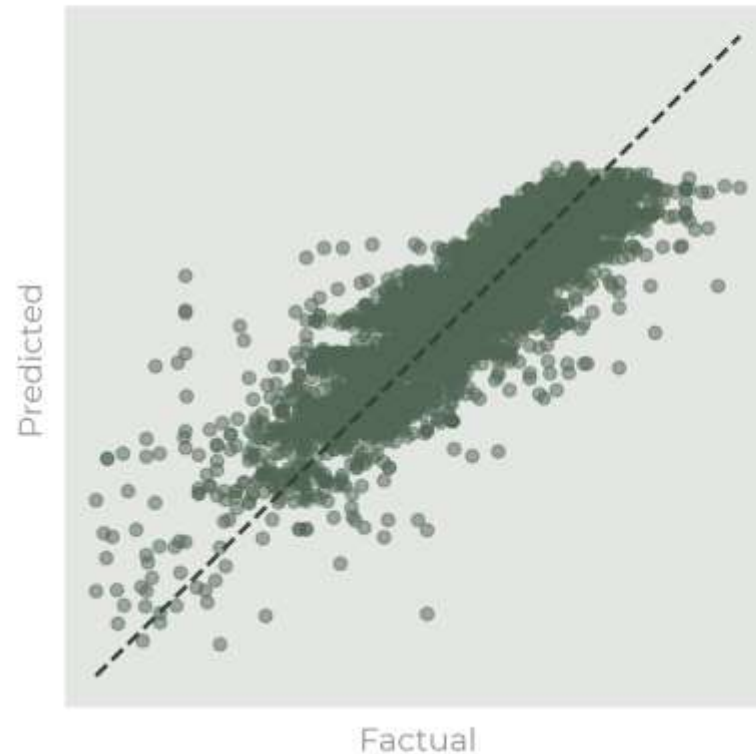
After building the model and finding the optimal parameters, the generalizability of the data by the model (R^2) was **0.78 for the training data and 0.76 for the test data**, where 1.00 = full explainability.
Whereas without the US these values are 0.87 and 0.73 respectively

```
preprocessor_xgb = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features)
    ])

model = xgb.XGBRegressor(
    n_estimators=400, max_depth=6,
    learning_rate=0.05, subsample=0.8,
    colsample_bytree=0.6, reg_alpha=0.5,
    reg_lambda=3, random_state=42
)

model_pipeline = Pipeline([('preprocessor_xgb', preprocessor_xgb),
                           ('regressor', model)])

model_pipeline.fit(X_train, y_train)
```



The importance of the features that influenced the prediction result the most

With **the absence of the U.S.** in the data:

- The importance of the dollar increases
- Importance of North America (including Canada) decreases
- Importance of Analysts decreases
- Importance of entry level experience increases
- The influence of salaries below the median for specialty and experience (Below) decreases
- Importance of countries with median salaries in the 30-60k range increases

	category	importance
0	jobexp_median_Above	9.80
1	country_median_>120k	9.45
2	salary_currency_USD	7.72
3	jobexp_median_Below	5.31
4	experience_level_EN	4.01
5	country_median_30-60k	3.65
6	jobexp_median_(n<15)	2.91
7	cont_direct_Asia | North	2.83
8	cont_direct_North America | North	2.26
9	experience_level_SE	1.99

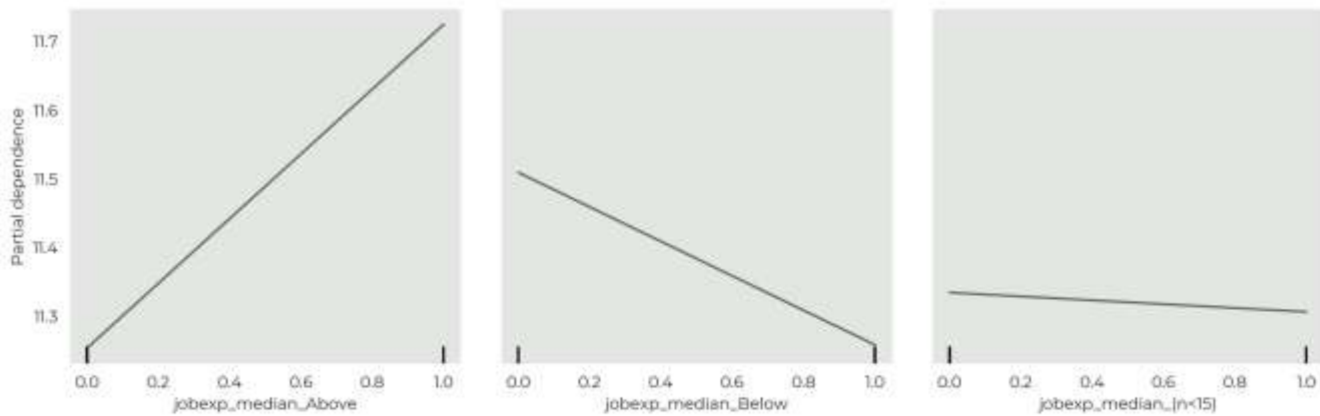
With the absence

	category	importance
0	jobexp_median_Above	18.54
1	jobexp_median_Below	17.23
2	country_median_>120k	13.10
3	origin_group_Analyst	6.24
4	cont_direct_North America | North	5.84
5	salary_currency_USD	3.73
6	experience_level_EN	3.48
7	experience_level_SE	2.90
8	jobexp_median_(n<15)	2.26
9	experience_level_EX	1.52

With the presence

To assess how specific traits influence salary predictions, we used **Partial Dependence Plots (PDPs) from scikit-learn**. PDPs help isolate the effect of individual variables in complex models by averaging out the impact of other features. This technique allows us to understand how a particular trait affects predicted salaries without interference from other factors

Without US in the data:



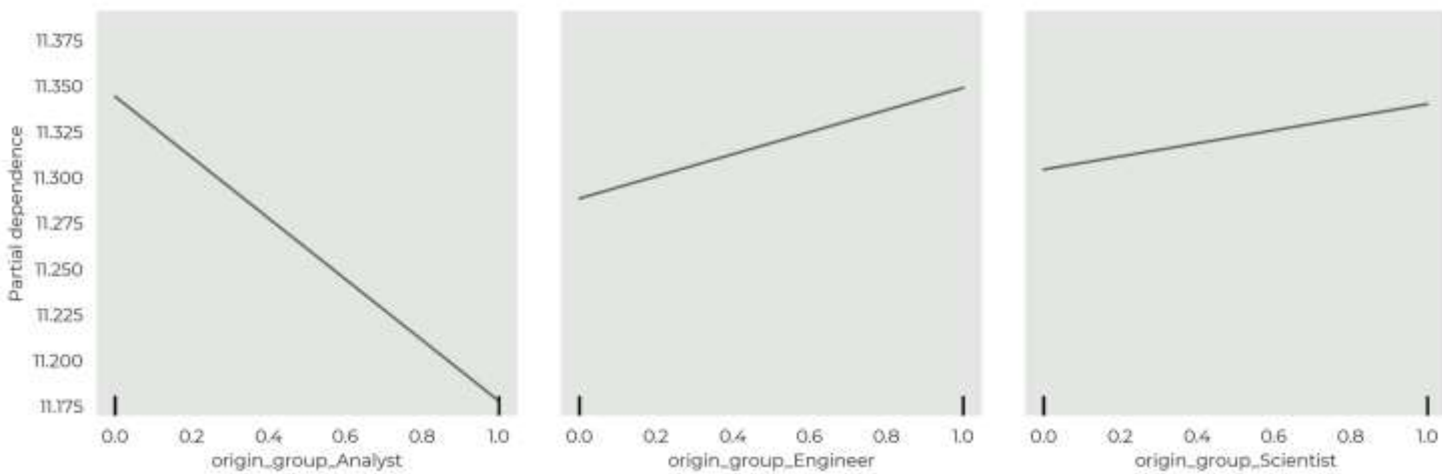
0 – trait absence
1 – presence

The relationship is mostly linear

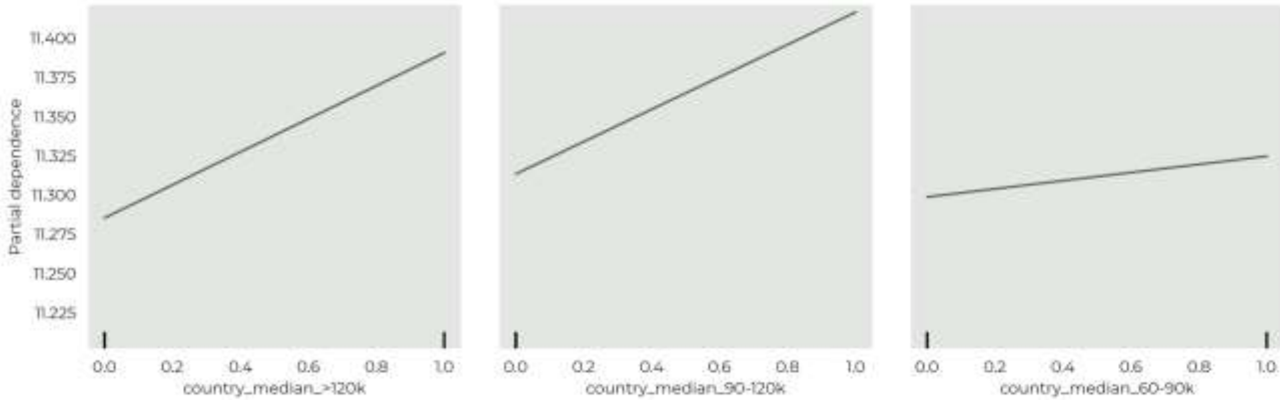
- Above-median salaries (e.g., Junior Analyst at 110k vs. median 100k) **show stronger positive influence** on predictions. This suggests that more frequent occurrences of above-median salaries could indicate broader salary growth trends
- Below-median salaries (e.g., Junior Analyst at 85k vs. median 100k) have **a weaker negative influence**, about half as strong as the positive effect of above-median salaries

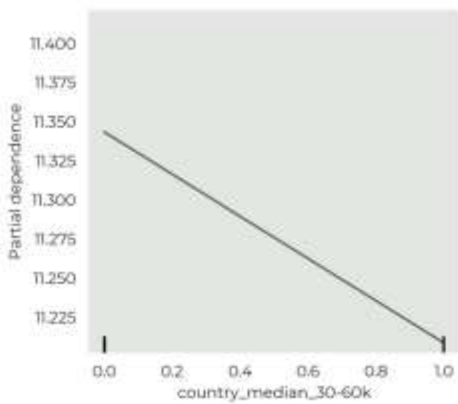
Analysts have a slight negative effect on wage predictions, likely due to:

1. The position's prevalence
2. Its tendency towards lower salaries given certain responsibilities



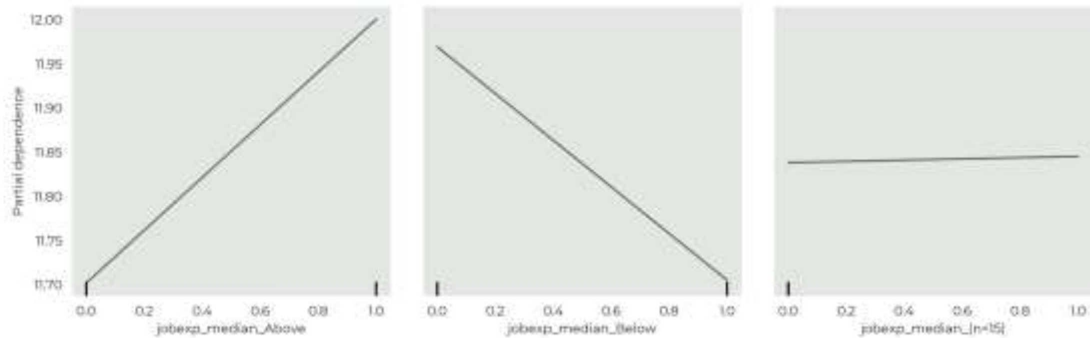
Each trait's impact on salary predictions is isolated, showing near-pure dependence (though not perfect) between the trait and salaries





- High wage medians (90-120k) **positively influence** high salary predictions, but not absolutely
- Mid-range salaries (60-90k) have minimal impact, considered '**stable**'
- Lower medians (30-60k) **significantly influence** the regression model"

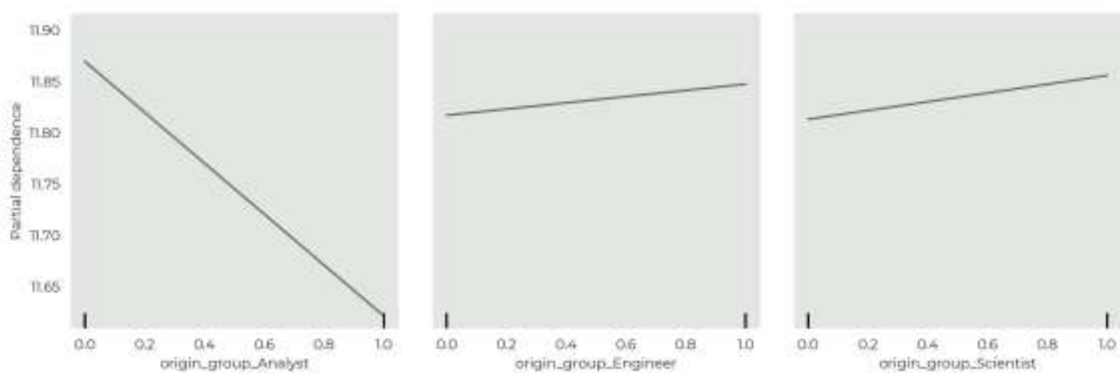
With the US presence:

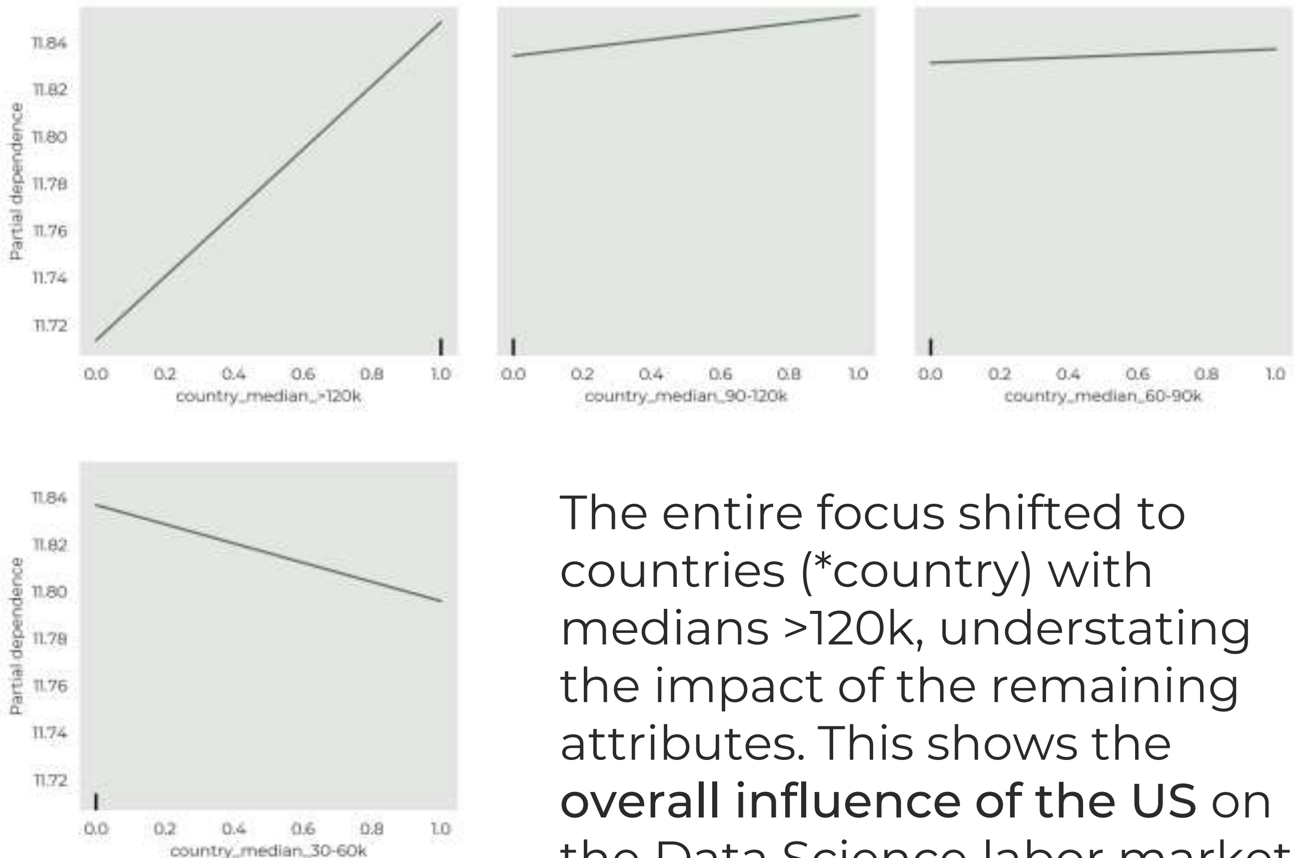


"Below" has become **more significant** in influence

Analysts have higher influence in this mapping due to:

- Greater importance in the initial array
- Larger numbers in the USA
- Consistently lower salaries compared to other common roles (DE, DS)





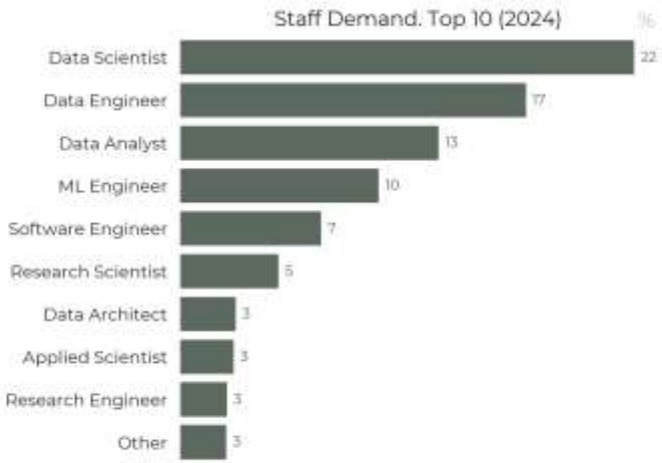
The entire focus shifted to countries (*country) with medians >120k, understating the impact of the remaining attributes. This shows the **overall influence of the US** on the Data Science labor market

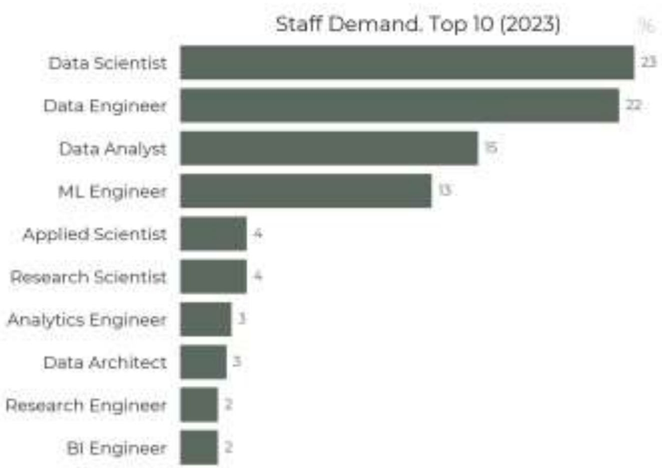
'Above' and 'Below' median salary positions **effectively isolate groups with specific salary features**, as shown by their importance in feature ranking

- High-median countries ($120k+, US and Canada) maintain stable distinctiveness:
  - **Without US**: 2nd in importance
  - **With US**: 3rd in importance This suggests separate analysis is needed for extreme salary ranges
- Analysts impact on overall salary median:
  - **Without US**: 83k / 86k (-3%)
  - **With US**: 145k / 150k (-3%)

Staff Demand

Analyze demand for specialists across years
(2024 vs 2023)

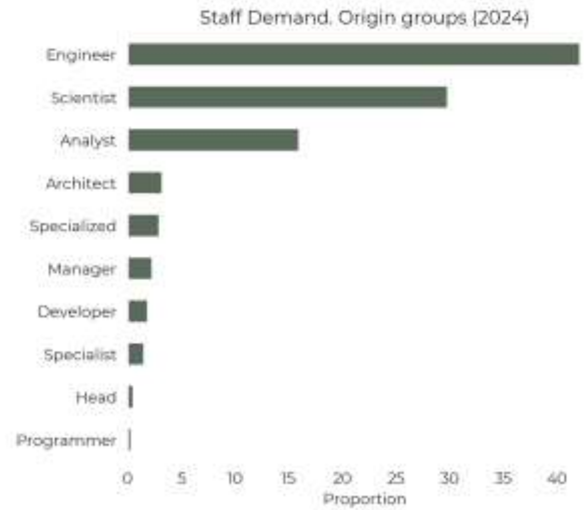


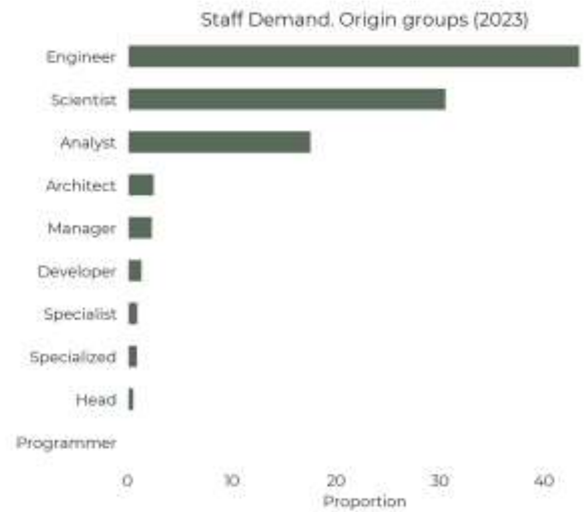


Overall

job_title	diff_2024_%
Other	87
Research Scientist	40
Research Engineer	17
Data Architect	14
Data Scientist	-4
Analytics Engineer	-15
Data Analyst	-17
Applied Scientist	-22
Data Engineer	-24
ML Engineer	-24

What is presented here is not the difference, but the ratio of the two ratios, 24th to 23rd year





Data Scientist demand shows consolidation in 2024, separating from Data Engineer despite a slight overall decline

Downward trends:

- Data Analysts
- Data Engineers
- ML Engineers

"Engineer" branch demand stable over 2 years (includes ML, AI, Data, BI Engineers)

Increasing demand:

- Research Scientist and Research Analyst: +40% each
- AI Engineer: +100% (2x FY23)
- Data Architect: +14%
- BI Analyst: +26%
- Diverse Professionals ("Other"): ~90% increase

Decreasing demand:

- Applied Scientist: -22%
- Analytics Engineer: -15% (out of top)
- BI Engineer: -37% (out of top)

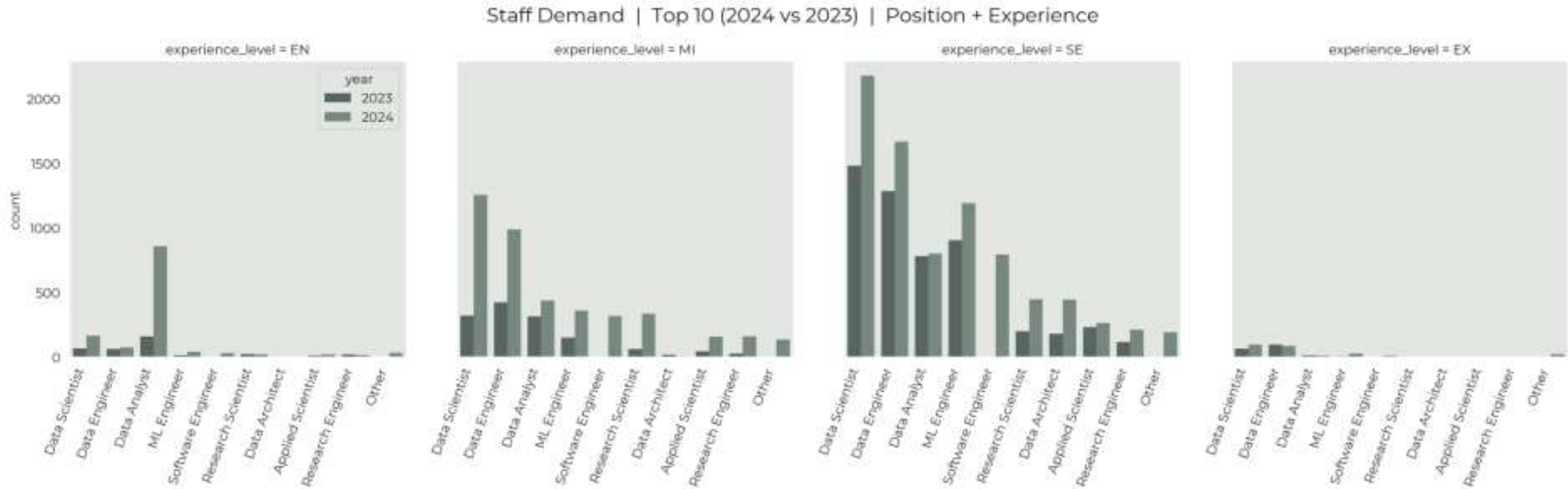
"Other" 2024 demand order:

1. Unique categories (new diverse roles)
2. AI (not only AI Engineer): +300%
3. ML (not only ML Engineer): +160%

4-7. Lead, Management, Operations, Analytics (insufficient FY23 data)

Specific

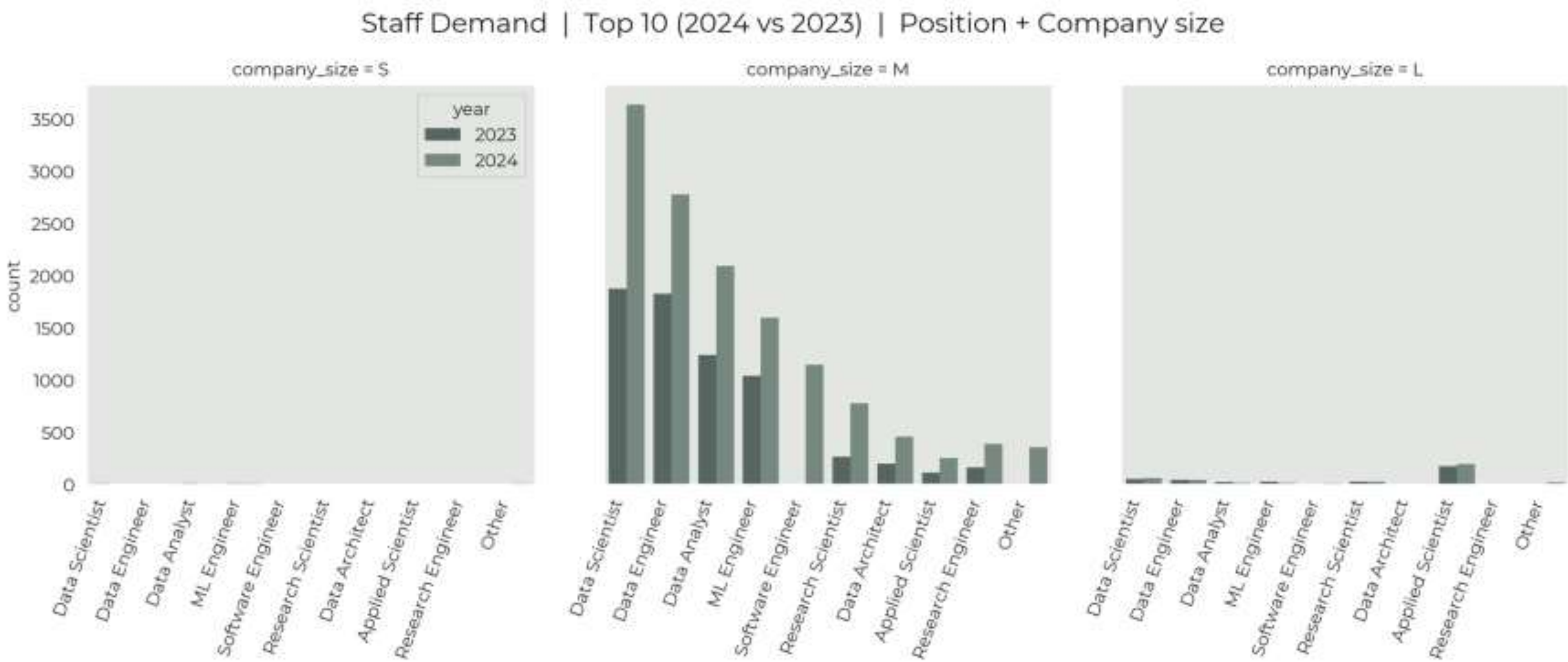
Qualification



- Despite a general decline in analyst demand, **entry-level professionals are leading in FY24**, with mid- and high-level professionals less considered
- Demand for **mid-level** Data Scientists and Data Engineers is **significantly higher than last year**, **especially for Data Scientists**, with a similar increase at the senior level

Specific

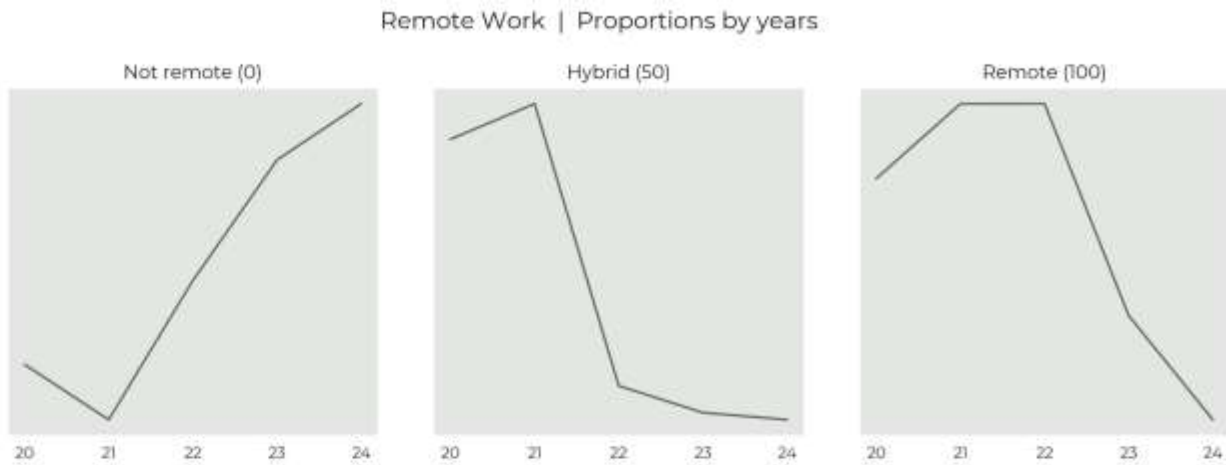
Company size

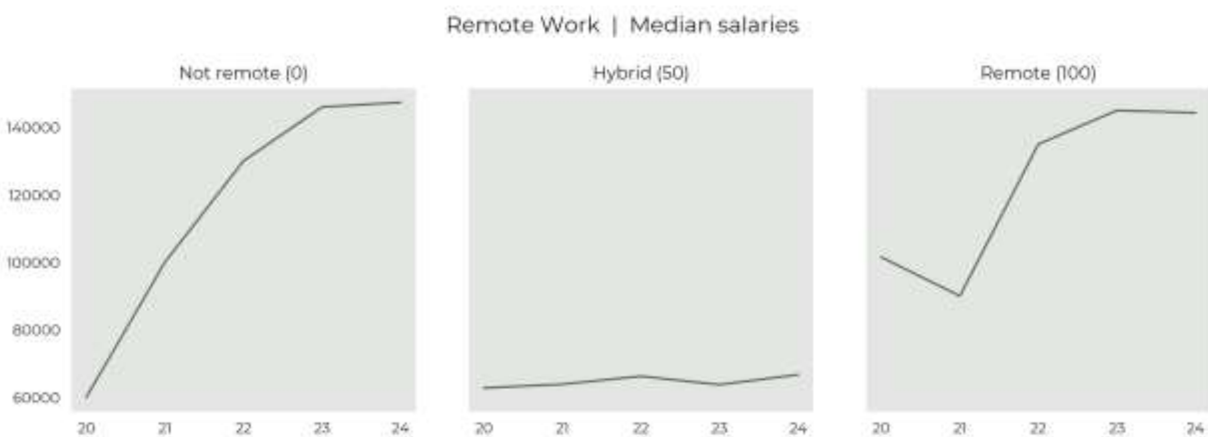


The same trend holds true for
a mid-size company

Remote Work

Compare trends in remote work adoption





- **Office work dominates,** but a balance with remote work is expected soon
- Office and remote work **similarly impact salaries positively,** despite office work's prevalence
- **The most stable** remote work type is hybrid

The bottom line

1. Overall impressions
2. Salaries
3. Demand
4. Remoteness

Overall impressions

- The current state of the market, which is at **an active stage of formation**, provides opportunities for experienced and inexperienced professionals
- The requirements for many specialties are **still not definitively defined**, which leaves **less limited trade-offs** and **consequent uncertainty**
- The growth in competitor volume is explained and at the same time **offset by the overall demand** for specialists
- Collecting, storing, processing and analyzing data is more **actively becoming a business standard**, which is **generating demand** and with it more sophisticated ways to solve business questions about data **without predominantly involving specialists**

  Improved automation is part of the demand, which is also actively changing the shape of labor market requirements

Salaries

- **Salaries are growing**, though not equally or everywhere
- The fluid market causes spikes in salary adjustments, introducing **uncertainty and risk**
- North America, especially the US, is **the most generous and demanding** (64% density of senior professionals).
  Its market impact needs separate analysis
- Germany's average Data Science salary increased by 15% since FY23 (top growth)
- **Entry-level salaries** grew from 43k to 56k (23%), with MI and EX at 15%, and SE with no clear advantage

- **Entry-level salary** increases:
  Data Analyst (+20%), Data Engineer (+18%), Data Scientist (+14%)
- **Mid-level salary** increases:
  Data Scientist (+12%),
  ML Engineer (+10%)
- **Largest salary increase**:
  Research Analyst (+33%),
  especially EN level (+48%);
  biggest declines: Decision Scientist (-22%), BI Engineer (-15%)
- **Shocking growth**:
  Large Company + Data Scientist (+54%, from 110k to 168k)
- **Small company salaries**
  (under 50 people) grew 15-25%

Demand

- Top positions like Data Analysts, Data Engineers, and ML Engineers see a slight decline in demand, **while new specialists** in AI, Management, and Research **rise**
- **Demand for Data Analysts declines** (except entry-level), but BI Analyst demand increases (+26%)
- **Specialist diversity increased significantly** in 2024 (+90%), led by AI and ML branches
- Despite high demand for mid-level Data Scientists, **engineers** (ML, AI, Data, BI) **are the leading group**, with scientists at a lower position
- AI Engineer demand **doubled** in Year 23 (+100%)
- **Research positions** like Research Scientist and Research Analyst see a 40% **increase in demand**
- Data Architect demand increased by 14%, indicating **a need for advanced data structures**
  Data Architects are mostly senior (95%) in mid-sized companies (98%)

Remoteness

- Office and remote work are **virtually identically reflected on salaries in a positive way**, despite the quantitative superiority of office work
- Observing this phenomenon, I consider myself to be in a position where **a balance between remote and office work will be formed over time**, introducing more flexibility. Research shows that remote work only has a positive impact on productivity for a number of simple reasons.
  Stanford study: +13% to productivity when working from home
- The most **stable type of remote work** is the hybrid type

What I learned from the project

Weeks of work in two slides

- **Efficiently utilize key combinations to manipulate code and text.** This allows you to save significant time and keep your focus on what is important
- **Structurally work with the project.** The most fundamental thing that every analyst must be able to do, otherwise (**a**) uncertainty will overwhelm the mind, (**b**) the project results will probably not justify themselves, (**c**) the project may go down the wrong path. Waste of resources as a consequence. **This project started as a thought, not an action**
- **Be clearer about what is important and what is not**
- **Respect uncertainty.** The one thing that stood out in particular was the realization that working with data, in general, is often about **the cruelest feeling of all – the feeling of uncertainty**. I darted from loathing to self-esteem, from discouragement to indescribable joy. A mix of emotions, simply put
- **Provide further clarification as questions arise.** Initially experience a sense of laziness when building this skill, afterward a sense of urgency due to possible incorrect interpretations and important omissions (a sense of excitement comes with this)

- **Suppress perfectionism (in development).** Challenging skill
- **Being honest with oneself.** Unsuspected self-deception is worse than conscious self-deception. It's easier to fix in the second case
- **Writing functions with complex logic.** At least that's what Claude and GPT told me
- **Use the LLM to find better solutions to problems.** The most difficult points to overcome are to accept your mistakes during clearly stated arguments as opposed to your expectations. Thereby deriving the following skill...
- **...to be more cool and judicious with criticism.** Resentment is replaced by a sense of acceptance of missed opportunities
- **Work with basic libraries for processing, analysis, and output.** LeetCode rests in comparison to this experience
- **Build simple machine learning models (but it's not automatic yet).** I've tried combining 2 completely different models (XGBoost and CTGAN), but haven't quite figured out how to turn this experience into something positive
- **Overcome the urge to "leave it alone" when there is something to fix.**
- **Spend less time on super details.** So many unnecessary actions during the project took up my time and nerves

A lot of time was spent on this very first project in my life, so the experience gained is too much and invaluable

Initiatives to improve research approach

- Use **less complex approaches** at the beginning of the project
- Set specific **deadlines** for completing project tasks
- Spend more time determining the actual **benefits of the project**
- Review as many third-party sources as possible to **confirm / refute assumptions**
- Be **less wasteful of time with respect to generating visualizations**, especially in the beginning when getting to know the data
- Engage more frequently with the LLM **to find gaps in thinking** about the project