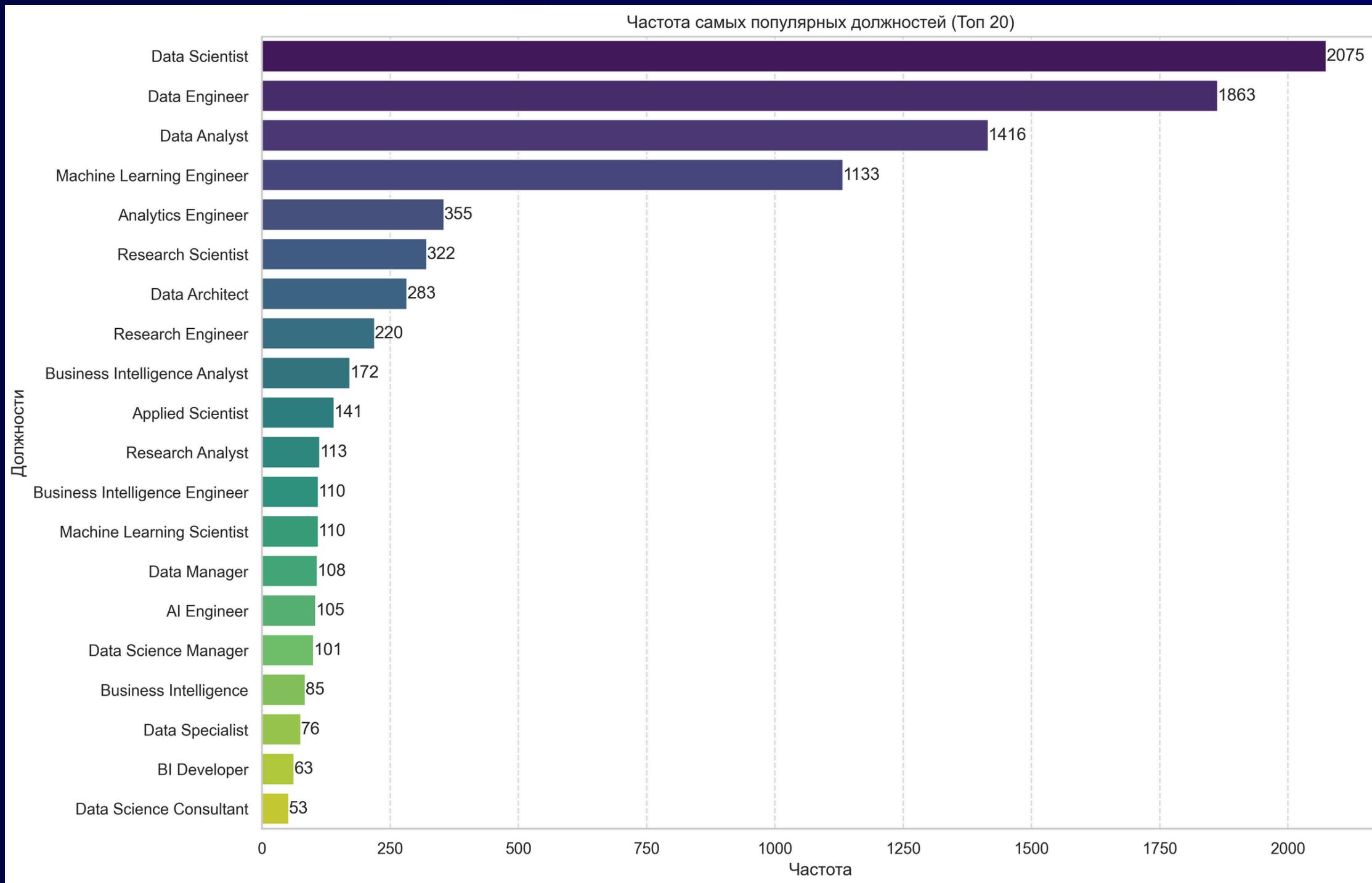


# АНАЛИЗ ЗАРПЛАТ В DATA SCIENCE

---

[Streamlit](#) проекта

[Github](#) проекта



# ВВЕДЕНИЕ И ЦЕЛЬ ПРОЕКТА

Почему выбрана эта тема?

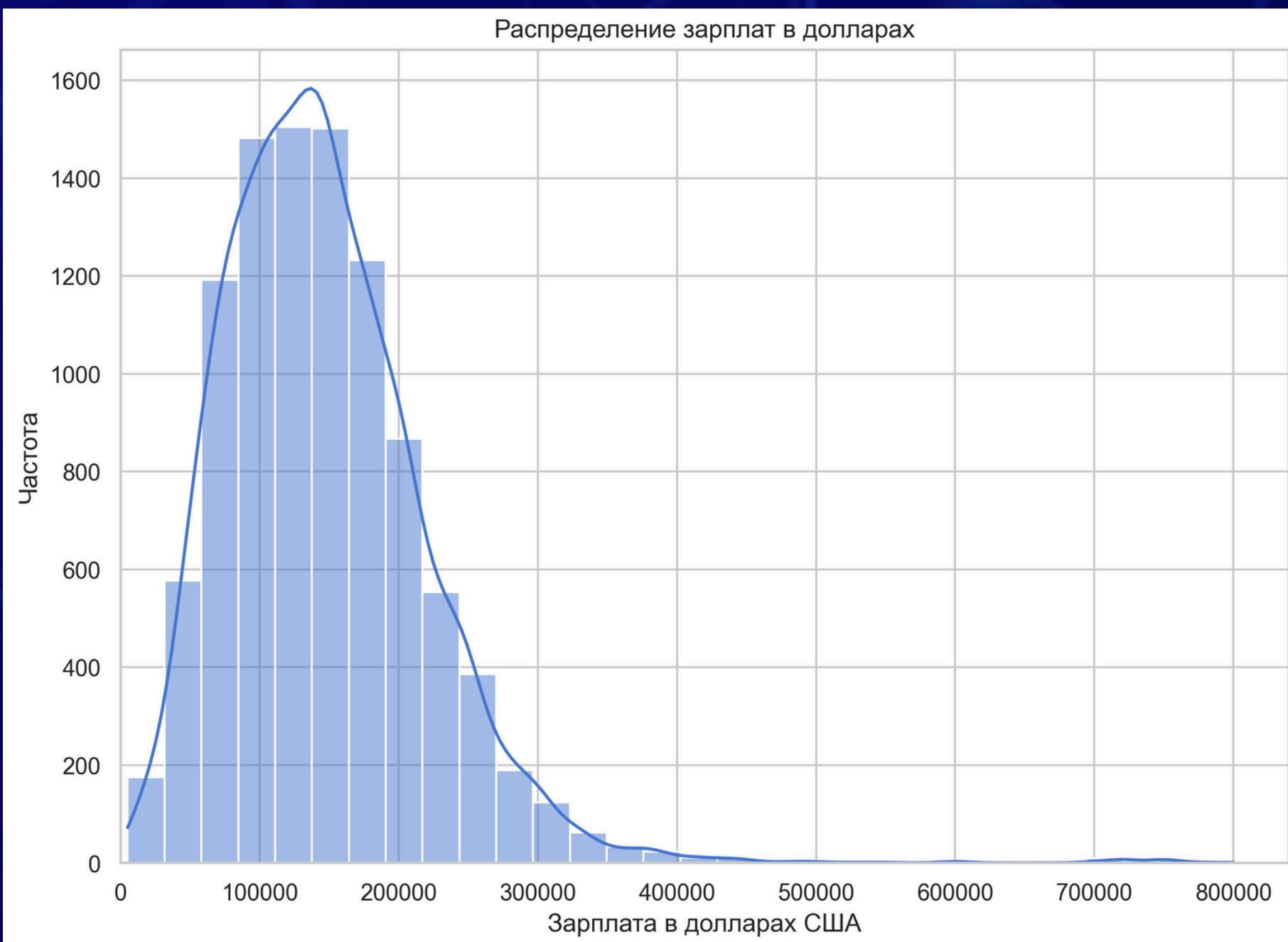
- **Data Science** – перспективная сфера.
- Интересно понять, какие факторы влияют на зарплаты.
- Полезность для начинающих DS-специалистов при выборе карьеры.

# ДАННЫЕ И ИХ ОПИСАНИЕ

index	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
94	2023	SE	FT	Data Scientist	70000	USD	70000	US	0	US	M
150	2023	SE	FT	Analytics Engineer	143200	USD	143200	US	0	US	M
343	2023	EN	FT	Research Engineer	125000	USD	125000	US	0	US	M
920	2023	MI	FT	Data Scientist	115360	USD	115360	US	100	US	M
2153	2022	MI	FT	Data Analyst	106260	USD	106260	US	0	US	M
3776	2024	EN	FT	Data Analyst	66000	USD	66000	US	100	US	M
4176	2024	MI	FT	Data Management Analyst	76050	USD	76050	US	0	US	M
4565	2024	MI	FT	Data Engineer	123000	USD	123000	US	0	US	M
4658	2024	SE	FT	Data Scientist	160000	USD	160000	US	0	US	M
5192	2023	SE	FT	Data Analyst	93918	USD	93918	US	0	US	M
6465	2023	SE	FT	Business Intelligence Analyst	115600	USD	115600	US	0	US	M
7130	2023	MI	FT	Data Scientist	131200	USD	131200	US	100	US	M
8155	2023	SE	FT	Data Scientist	142436	USD	142436	GB	0	GB	M
8654	2024	MI	FT	Business Intelligence	87000	USD	87000	US	0	US	M
8773	2024	EN	FT	Data Analyst	94000	USD	94000	US	0	US	M
8823	2024	SE	FT	Data Scientist	252000	USD	252000	US	0	US	M
9614	2024	MI	FT	Machine Learning Scientist	165000	USD	165000	US	0	US	M

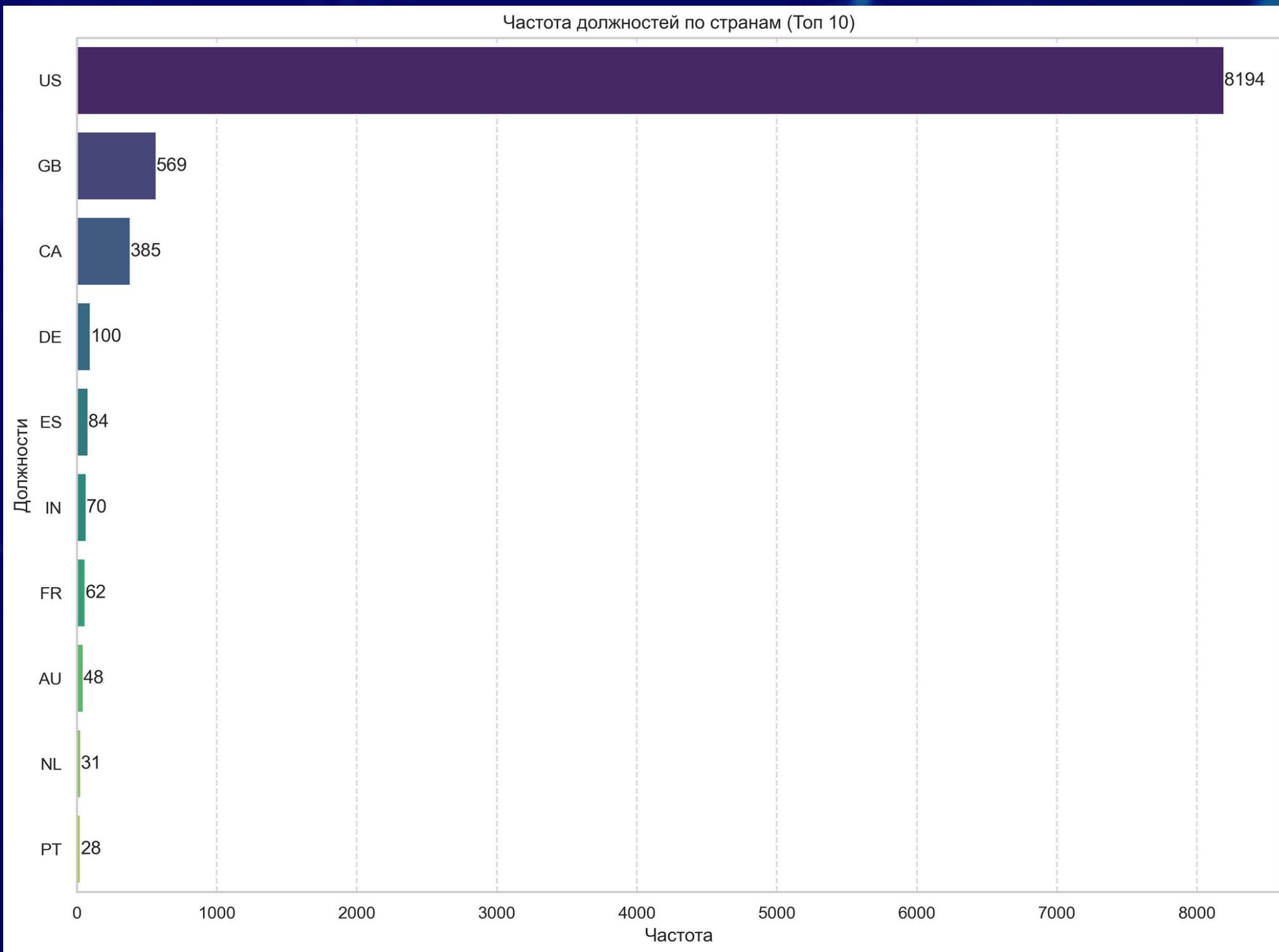
- **work\_year:** Год, в котором была выплачена зарплата.
- **experience\_level:** Уровень опыта на работе в течение данного года.
- **employment\_type:** Тип занятости для данной роли.
- **job\_title:** Название должности, на которой работали в течение года.
- **salary:** Общая сумма выплаченной брутто-зарплаты.
- **salary\_currency:** Валюта, в которой выплачивалась зарплата, в формате ISO 4217.
- **salary\_in\_usd:** Зарплата, переведённая в доллары США.
- **employee\_residence:** Основная страна проживания сотрудника в течение рабочего года в формате **ISO 3166**.
- **remote\_ratio:** Общая доля работы, выполненной удалённо.
- **company\_location:** Страна, где находится главный офис работодателя или его филиал.
- **company\_size:** Среднее количество сотрудников, работавших в компании в течение года.
- **index:** индекс строки в наборе данных

# EDA – ИССЛЕДОВАНИЕ ЗАРПЛАТ



Выводы:

- Распределение сдвинуто вправо → больше низких зарплат.
- Есть выбросы с очень высокими значениями.



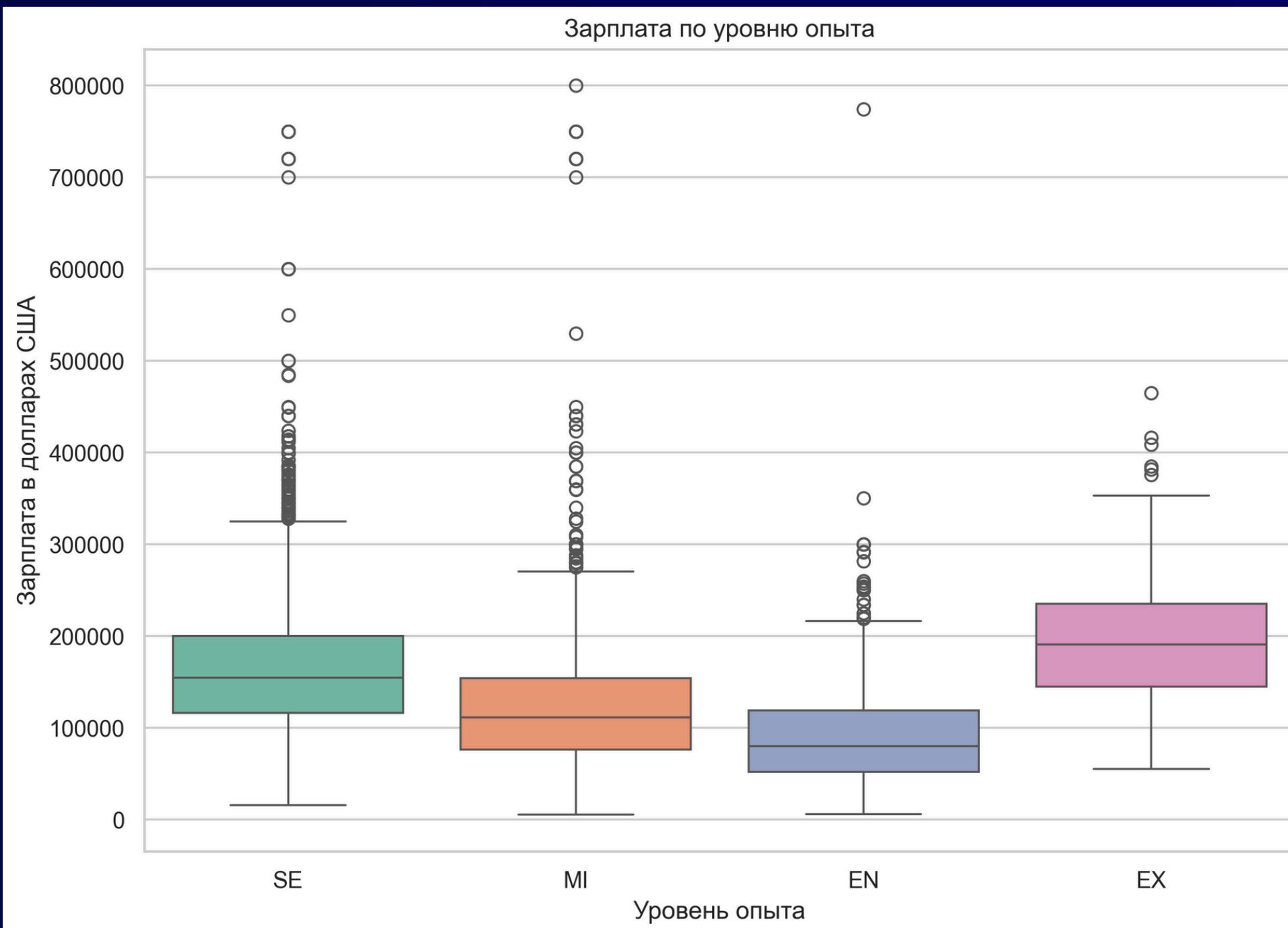
Страна и его ISO код

## СТРАНЫ И КОЛИЧЕСТВО ДОЛЖНОСТЕЙ

США – лидер (8194 вакансии), за ним Великобритания (GB) и Канада (CA).

- Европа (Германия, Испания, Франция) активно развивает технологический сектор.
- Индия и Австралия – ключевые центры IT и аутсорсинга.
- Рост вакансий в Нидерландах и Португалии – новые технологические хабы.

**Вывод:** США доминирует, но Европа и Азия усиливают позиции.

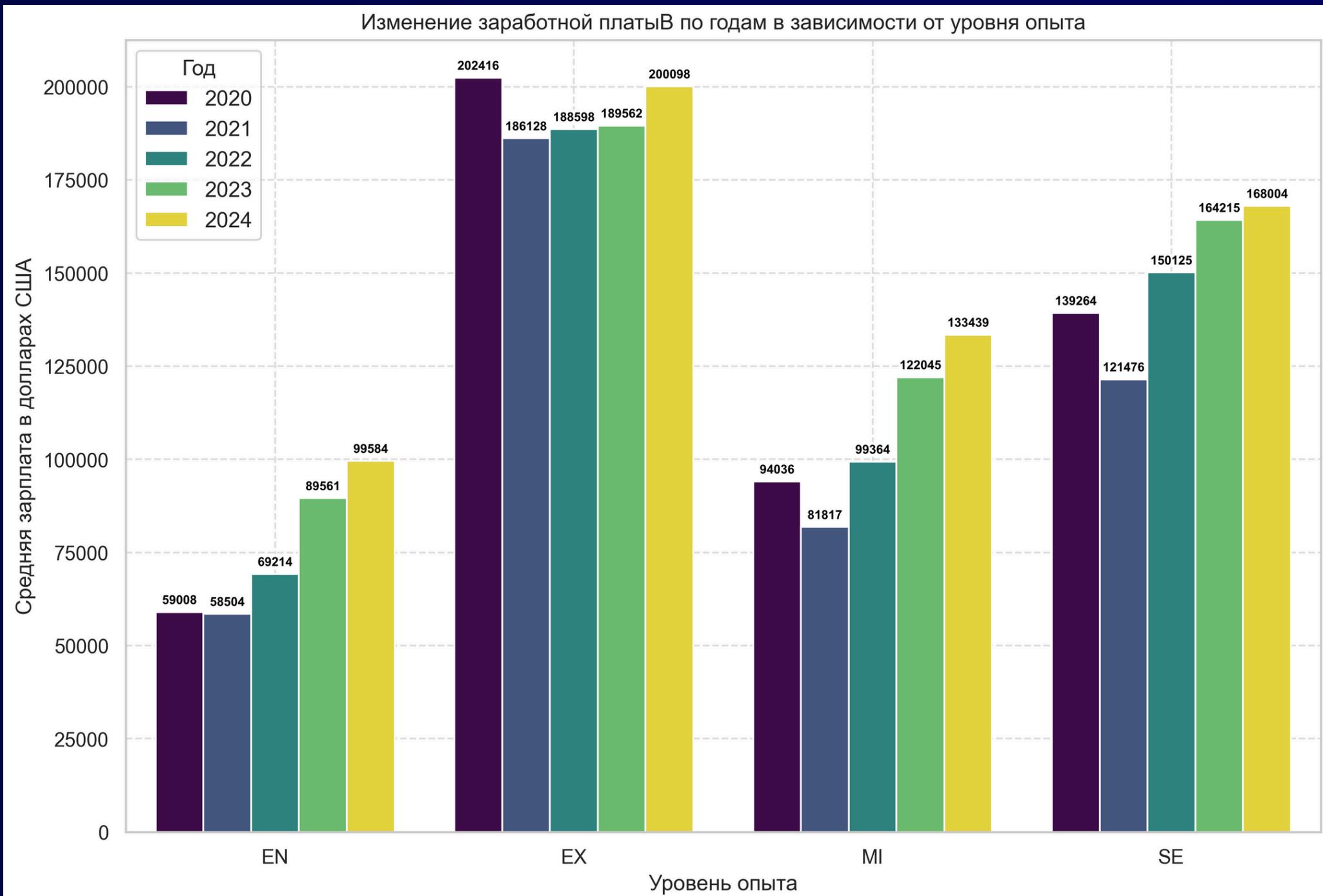


## ВЛИЯНИЕ ОПЫТА НА ЗАРПЛАТУ

Выводы:

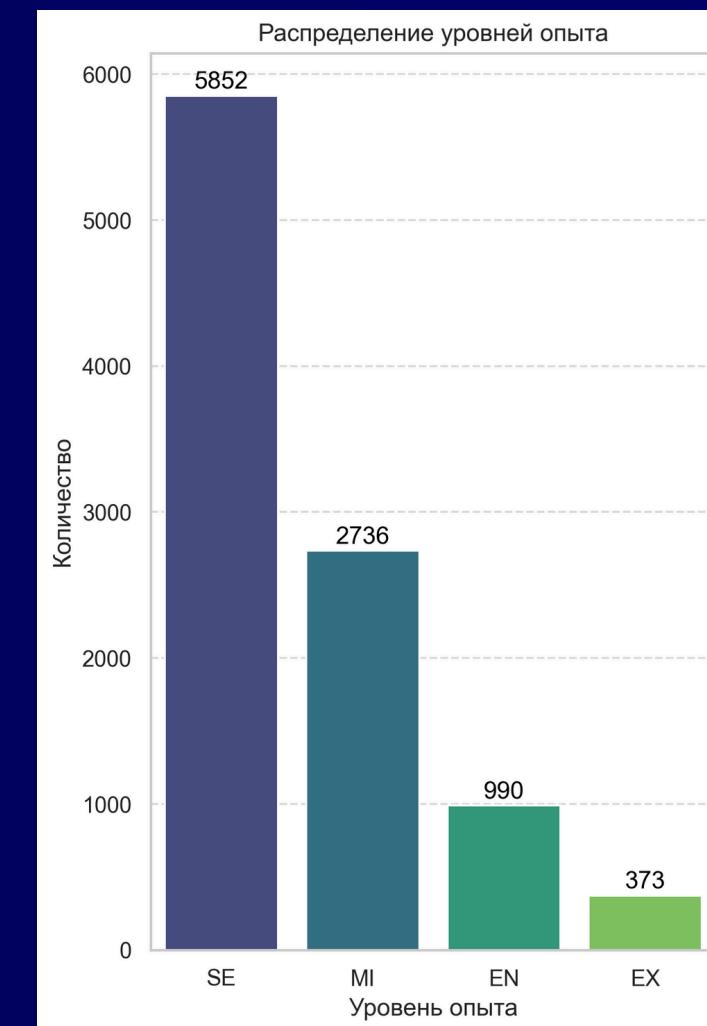
- Опыт сильно влияет на зарплату.
- Junior (EN) получает меньше всех.
- Senior (SE) и Expert (EX) зарабатывают больше.

# ДИНАМИКА ЗАРПЛАТ ПО ГОДАМ



## Выводы:

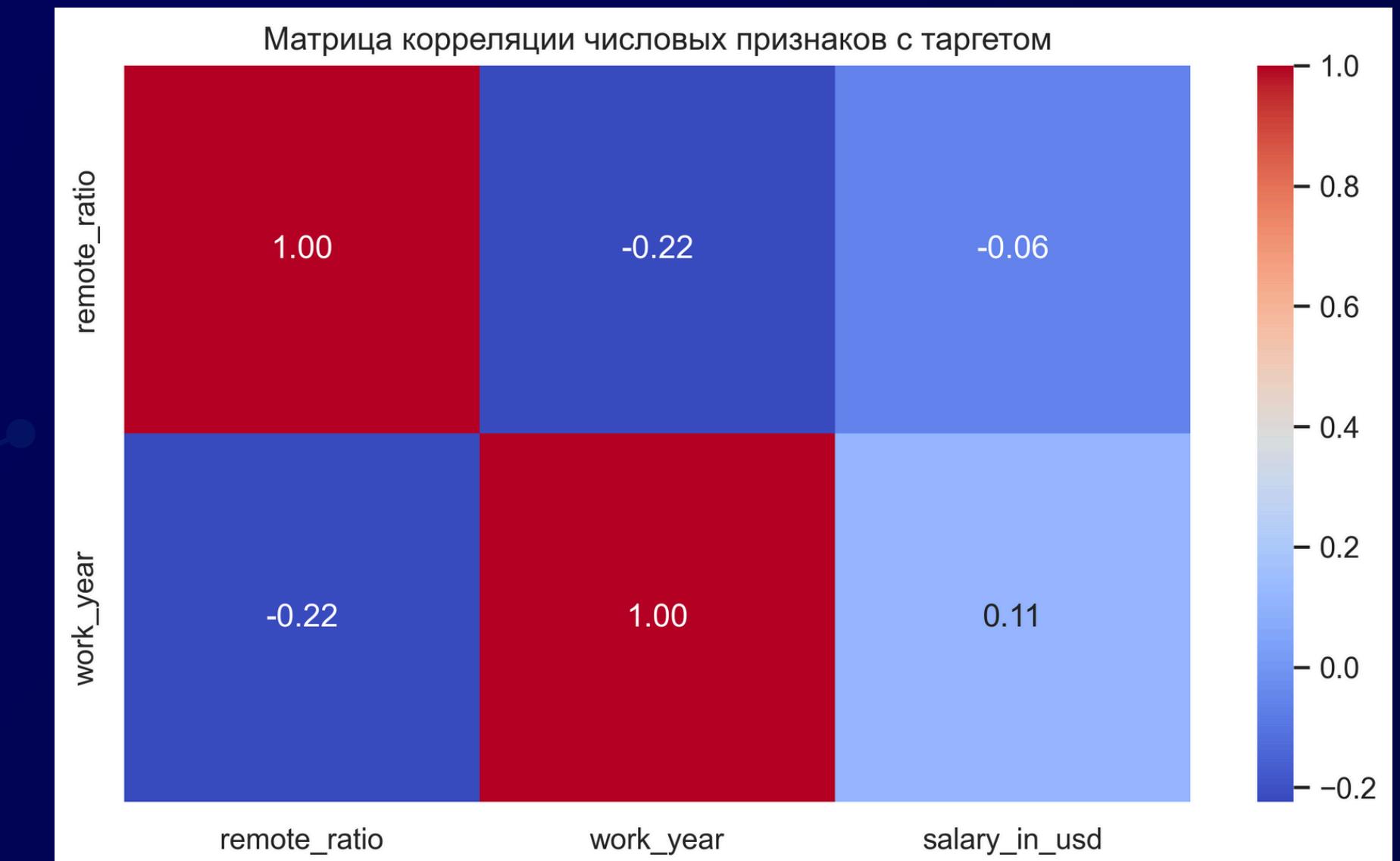
- В 2022–2024 зарплаты значительно выросли.
- Вероятно, связано с ростом спроса на специалистов.



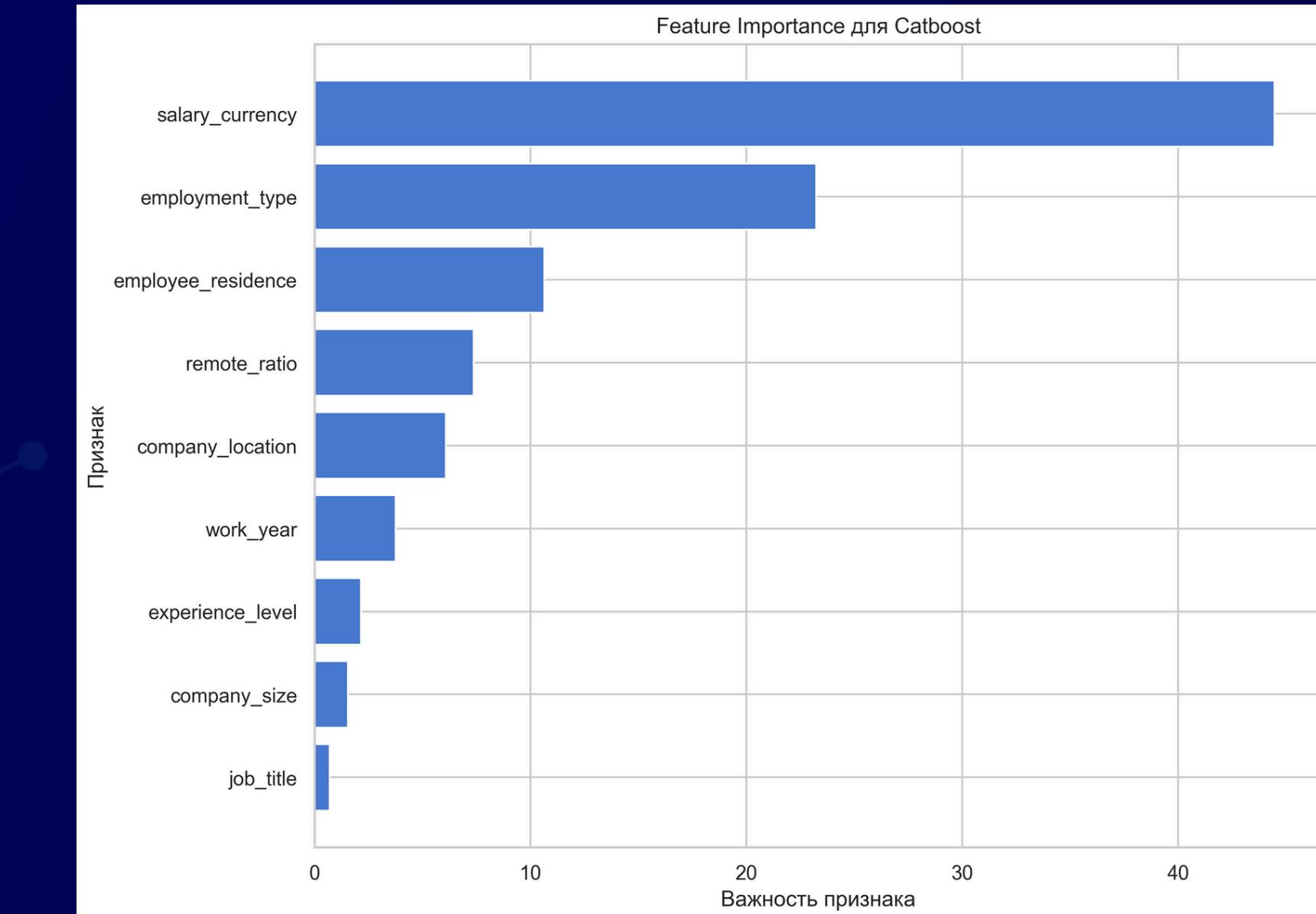
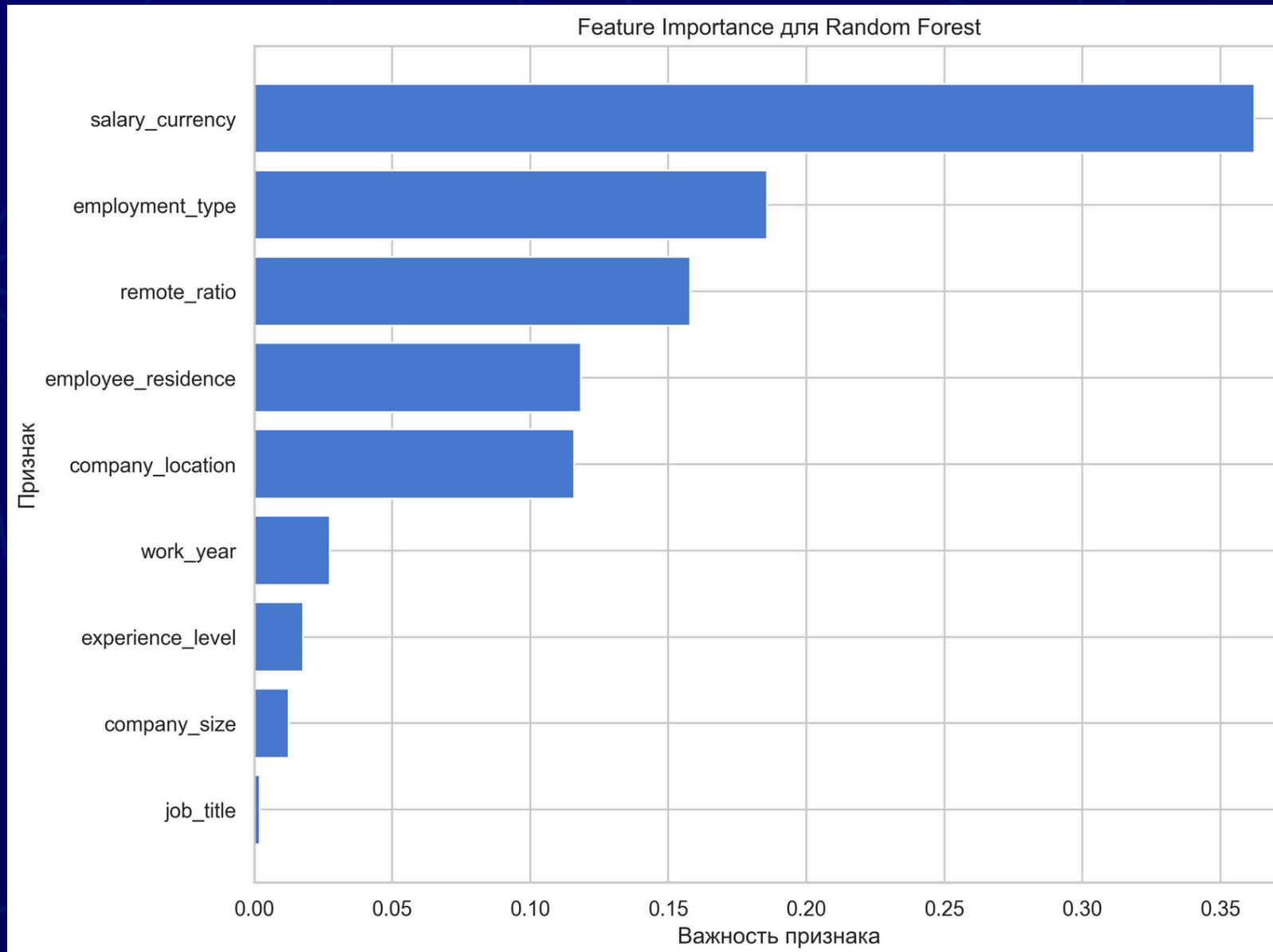
# КОРРЕЛЯЦИЯ ПРИЗНАКОВ

**Целевой признак:**  
• зарплата в USD

**Выводы:**  
• Год работы (*work\_year*) слабо коррелирует с зарплатой.  
• Доля удалённой работы (*remote\_ratio*) почти не влияет.



# ВАЖНОСТЬ ПРИЗНАКОВ



# РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ

Model	MAPE Train (%)	MAPE Test (%)	R2 Train (%)	R2 Test (%)	CV R2 Mean (%)
Mean Predictor	58.95	61.78	0.0	-0.06	-0.05
Median Predictor	55.53	58.32	-1.15	-1.67	-1.22
Linear Regression	37.72	38.68	31.6	33.88	30.43
Polynomial Regression	37.71	38.64	31.62	33.89	30.42
Random Forest	32.6	35.41	38.53	36.48	32.48
CatBoost	32.41	34.84	38.55	36.69	33.47

Лучшая модель: CatBoost

- MAPE: 34.84%
- R<sup>2</sup>: 36.69%
- CV R<sup>2</sup>: 33.47%

Вывод:

CatBoost оказался лучшей моделью, но  
качество можно улучшить

# БЕЙЗЛАЙН И МЕТРИКИ

## Метрики оценки моделей

- MAPE – средняя ошибка в процентах
- $R^2$  – насколько хорошо модель объясняет зависимость

## Вывод:

- Простые модели (Mean, Median) показывают слабые результаты, лучшие – Random Forest и CatBoost.





## БУДУЩЕЕ ПРОЕКТА

Как можно улучшить модель?

- Добавить новые признаки.
- Провести более тщательный feature engineering.
- Использовать более тонкую настройку гиперпараметров модели.

# THANK YOU

for your  
attention