SQL проект

Data jobs salaries Analysis

1 Выбор источника данных, постановка цели и задач проекта

Используемый датасет был взят с международного открытого сайта вакансий по дата профессиям. Данный сайт анонимно собирает информацию о зарплатах от специалистов со всего мира в сферах искусственного интеллекта, машинного обучения, науки о данных. Информация представлена за период с 2020 по 2023 включительно. Преимущественно данные за 2023 год. Данный анализ поможет разобраться в зарплатах дата профессий.

АТРИБУТЫ

Название колонки	Значение
work_year	Год выплаты зарплаты
experience_level	Грейд работника на должности в течение года: Entry-level / Junior, Mid-level / Intermediate, Senior-level / Expert, Executive-level / Director
employment_type	Тип занятости на должности: Part-time, Full-time, Contract, Freelance
job_title	Должность работника в течение года
salary	Выплаченная зарплата
salary_currency	Валюта выплаченной зарплаты
salary_in_usd	Выплаченная зарплата в долларах (средний курс валют за год)
employee_residence	Основная страна проживания работника в течение года
remote_ratio	Общий объём работы, выполненной удалённо: 0: No remote work (less than 20%), 50: Partially remote/hybrid, 100: Fully remote (more than 80%)
company_location	Страна головного офиса компании
company_size	Среднее количество человек, отработавших в компании в течение года: S: less than 50 employees (small), M: 50 to 250 employees (medium), L: more than 250 employees (large)

Цель проекта: проанализировать зарплаты вакансий в сферах AI, ML и BIG DATA с открытого международного банка данных ai-jobs.net

Задачи проекта:

1) Определить количество вакансий с предоставленной информацией по зарплате на сайте по годам в порядке убывания
2) Разбить вакансии по грейду работников за предоставленный период в датасете (experience_level)
3) Разбить вакансии по их типу (employment_type) в 2023 году. Какой тип преобладает у Senior-level?
4) Определить топ 10 самых высокооплачиваемых вакансий за весь период. Определить топ 10 самых низкооплачиваемых вакансий за 2023 год в среднем.
5) Распределить вакансии по регионам стран с головными офисами (company_location) от большего к меньшему. В каком регионе отмечено больше всего Entry-level?
6) Определить количество вакансий с головным офисом в России за весь период со средней зарплатой по годам.
7) Определить топ 10 стран(company_location) с самой высокой средней зарплатой вакансий, представленных на сайте, отобразить количество вакансий по странам.
8) Определить топ 10 самых популярных профессий. Определить страну(company_location), в которой вакансия по рейтингу встречается чаще всего, вывести количество вакансий на страну от общего.
9) Определить вакансии с 'удалёнкой' в США за 2023 год со средней зарплатой более 200 тысяч. Распределить полученную выборку по грейду работника (experience_level) и определить средний заработок. Сортировка в порядке возрастания.

2 Первичная обработка данных и загрузка в БД

Датасет был загружен с открытого сайта ai-jobs.net в формате csv. Далее он был отредактирован, а затем загружен в БД.

Также был загружен вспомогательный датасет с кодами (ISO 3166) по странам и их регионам. Данный датасет был взят с открытого github в формате csv. С его помощью были переименованы значения в атрибутах (employee_residence) и (company_location). Также он был использован для применения оператора JOIN.

Далее была создана пустая БД в DBeaver для последующей работы. К ней подключился через python при помощи метода .connect() библиотеки sqlite3. Далее в пустую БД были загружены для датасета и преобразованы в датафреймы через библиотеку pandas. Преобразование датафреймов осуществлялось с помощью метода .read_csv. Запуск датафреймов в БД был осуществлён с помощью метода .to_sql.

Далее выполнялось решение поставленных задач проекта в DBeaver с помощью языка SQL.

3 Исследование данных с помощью SQL

1) Определить количество вакансий с предоставленной информацией по зарплате на сайте по годам в порядке убывания

```
SELECT
work_year,
COUNT(*) AS amount_of_vacancies
FROM data_salaries ds
GROUP BY work_year
ORDER BY amount_of_vacancies DESC;
```

	work_year	amount_of_vacancies
1	2 023	6 747
2	2 022	1 651
3	2 021	218
4	2 020	75

2) Отсортировать вакансии по грейду работников за предоставленный период в датасете (experience_level)

```
SELECT
experience_level,
COUNT(*) AS amount_of_vacancies_by_grade
FROM data_salaries ds
GROUP BY experience_level
ORDER BY amount_of_vacancies_by_grade DESC;
```

	experience_level	amount_of_vacancies_by_grade
1	Senior-level	6 254
2	Mid-level	1 706
3	Entry-level	466
4	Executive-level	265

3) Отсортировать вакансии по их типу (employment_type) в 2023 году.

```
SELECT
employment_type,
COUNT(*) AS amount_of_vacancies_by_type
FROM data_salaries ds
WHERE work_year = '2023'
GROUP BY employment_type
ORDER BY amount_of_vacancies_by_type DESC;
```

	employment_type	amount_of_vacancies_by_type
1	Full-time	6 735
2	Contract	8
3	Freelance	3
4	Part-time	1

3.1) Какой тип преобладает у Senior-level?

```
SELECT
employment_type,
COUNT(*) AS amount_of_vacancies_by_type
FROM data_salaries ds
WHERE work_year = '2023' AND experience_level = 'Senior-level'
GROUP BY employment_type
ORDER BY amount_of_vacancies_by_type DESC
LIMIT 1;
```

	employment_type	amount_of_vacancies_by_type
1	Full-time	5 020

4) Определить топ 10 самых высокооплачиваемых вакансий в долларах в среднем за весь период.

```
SELECT
job_title,
ROUND(AVG(salary_in_usd)) AS average_top_salary
FROM data_salaries ds
GROUP BY job_title
ORDER BY average_top_salary  DESC
LIMIT 10;
```

	job_title	average_top_salary
1	Analytics Engineering Manager	399 880
2	Data Science Tech Lead	375 000
3	Managing Director Data Science	300 000
4	AWS Data Architect	258 000
5	AI Architect	250 328
6	Cloud Data Architect	250 000
7	Director of Data Science	221 365
8	Head of Data	209 166
9	Data Infrastructure Engineer	201 375
10	Head of Machine Learning	198 103

4.1) Определить топ 10 самых низкооплачиваемых вакансий за 2023 год в среднем.

```
SELECT
job_title,
employee_residence,
ROUND(AVG(salary_in_usd)) AS low_salary
FROM data_salaries ds
WHERE work_year = '2023'
GROUP BY job_title, employee_residence
ORDER BY low_salary  ASC
LIMIT 10;
```

	job_title	employee_residence	low_salary
1	Data Analyst	Philippines	15 680
2	Data Scientist	Ecuador	16 000
3	Product Data Analyst	India	16 417
4	Business Data Analyst	Armenia	17 000
5	Data Analytics Lead	India	17 511
6	Data Engineer	India	17 513
7	Data Analyst	Poland	18 160
8	Lead Data Analyst	India	18 241
9	BI Analyst	Turkey	18 381
10	Data Scientist	Greece	19 434

5) Распределить вакансии по регионам стран с головными офисами(company_location) от большего к меньшему.

```
SELECT
ic.region,
COUNT(*) AS amount_of_vacancies
FROM data_salaries ds
INNER JOIN iso_code ic
ON ds.company_location = ic.name
GROUP BY ic.region
ORDER BY amount_of_vacancies DESC;
```

	region	amount_of_vacancies
1	Americas	7 728
2	Europe	822
3	Asia	95
4	Oceania	27
5	Africa	19

5.1) В каком регионе отмечено больше всего Entry-level?

```
SELECT
ic.region,
COUNT(*) AS amount_of_vacancies
FROM data_salaries ds
INNER JOIN iso_code ic
ON ds.company_location = ic.name
WHERE experience_level = 'Entry-level'
GROUP BY ic.region
ORDER BY amount_of_vacancies DESC
LIMIT 1;
```

	region	amount_of_vacancies
1	Americas	329

6) Определить количество вакансий с головным офисом в России за весь период со средней зарплатой по годам.

```
SELECT
work_year,
COUNT(*) AS vacancies_in_Russia,
ROUND(AVG(salary_in_usd)) AS average_salary_in_Russia
FROM data_salaries ds
WHERE company_location = 'Russian Federation'
GROUP BY work_year;
```

	work_year	vacancies_in_Russia	average_salary_in_Russia
1	2 021	2	157 500
2	2 022	2	61 228
3	2 023	3	36 667

7) Определить топ 10 стран(company_location) с самой высокой средней зарплатой вакансий, представленных на сайте, отобразить количество вакансий по странам.

```
SELECT
company_location,
ROUND(AVG(salary_in_usd)) AS average_salary,
COUNT(salary_in_usd) AS quantity_of_vacancies
FROM data_salaries ds
GROUP BY company_location
ORDER BY average_salary DESC
LIMIT 10;
```

	company_location	average_salary	quantity_of_vacancies
1	Qatar	300 000	1
2	Israel	217 332	3
3	Puerto Rico	167 500	4
4	United States of America	158 348	7 474
5	Canada	141 456	197
6	Saudi Arabia	134 999	2
7	Australia	132 283	24
8	New Zealand	125 000	1
9	Ukraine	121 333	6
10	Bosnia and Herzegovina	120 000	1

8) Определить топ 10 самых популярных профессий, представленных на сайте.

```
SELECT
job_title,
COUNT(*) AS quantity_of_vacancies
FROM data_salaries ds
GROUP BY job_title
ORDER BY quantity_of_vacancies DESC
LIMIT 10;
```

	job_title	quantity_of_vacancies
1	Data Engineer	2 046
2	Data Scientist	1 822
3	Data Analyst	1 296
4	Machine Learning Engineer	890
5	Applied Scientist	258
6	Research Scientist	243
7	Analytics Engineer	237
8	Data Architect	195
9	Research Engineer	133
10	Data Manager	130

8.1) Определить страну(company_location), в которой вакансия по рейтингу встречается чаще всего, вывести количество вакансий на страну от общего.

```
SELECT
job_title,
quantity_of_vacancies,
company_location AS beloved_country,
quantity_of_vacancies_in_beloved_country
FROM (
        SELECT
        job_title,
        company_location,
        COUNT(*) AS quantity_of_vacancies_in_beloved_country,
        ROW_NUMBER() OVER (PARTITION BY job_title order by
        count(*) DESC) AS rating,
        sum(count(*)) OVER (PARTITION BY job_title) AS
        quantity_of_vacancies
        FROM data_salaries ds
        GROUP BY job_title, company_location)
WHERE rating = 1
ORDER BY quantity_of_vacancies DESC
LIMIT 10;
```

	job_title	quantity_of_vacancies	beloved_country	quantity_of_vacancies_in_beloved_country
1	Data Engineer	2 046	United States of America	1 796
2	Data Scientist	1 822	United States of America	1 554
3	Data Analyst	1 296	United States of America	1 136
4	Machine Learning Engineer	890	United States of America	775
5	Applied Scientist	258	United States of America	257
6	Research Scientist	243	United States of America	218
7	Analytics Engineer	237	United States of America	205
8	Data Architect	195	United States of America	183
9	Research Engineer	133	United States of America	119
10	Business Intelligence Engineer	130	United States of America	128

9) Определить вакансии с 'удалёнкой' в США за 2023 год со средней зарплатой более 200 тысяч

```
SELECT
job_title,
experience_level,
employment_type,
remote_ratio,
ROUND(AVG(salary_in_usd)) AS average_salary
FROM data_salaries
WHERE work_year = '2023'
AND company_location = 'United States of America'
AND remote_ratio = 'Fully Remote'
GROUP BY job_title, experience_level, employment_type,
remote_ratio
HAVING average_salary > 200000
ORDER BY average_salary DESC;
```

	job_title	experience_level	employment_type	remote_ratio	average_salary
1	Finance Data Analyst	Senior-level	Contract	Fully Remote	323 905
2	Head of Data	Executive-level	Full-time	Fully Remote	279 743
3	AWS Data Architect	Mid-level	Full-time	Fully Remote	258 000
4	Director of Data Science	Executive-level	Full-time	Fully Remote	230 405
5	Applied Machine Learning Engineer	Executive-level	Full-time	Fully Remote	225 000
6	ML Engineer	Senior-level	Full-time	Fully Remote	223 545
7	Data Engineer	Executive-level	Full-time	Fully Remote	219 794
8	AI Architect	Senior-level	Full-time	Fully Remote	217 333
9	AI Architect	Executive-level	Full-time	Fully Remote	215 936
10	Software Data Engineer	Senior-level	Full-time	Fully Remote	210 000
11	Research Engineer	Senior-level	Full-time	Fully Remote	204 058

9.1) Распределить полученную выборку по грейду работника (experience_level) и определить средний заработок. Сортировка в порядке возрастания.

```
SELECT
experience_level,
COUNT(*) AS quantity_of_vacancies,
ROUND(AVG(average_salary)) AS average_salary_by_exp_level
FROM
    (SELECT
      job_title,
      experience_level,
      employment_type,
      remote_ratio,
      ROUND(AVG(salary_in_usd)) AS average_salary
      FROM data_salaries
      WHERE work_year = '2023'
      AND company_location = 'United States of America'
      AND remote_ratio = 'Fully Remote'
      GROUP BY job_title, experience_level, employment_type,
      remote_ratio
      HAVING average_salary > 200000
      ORDER BY average_salary DESC)
GROUP BY experience_level
ORDER BY average_salary_by_exp_level ASC;
```

	experience_level	quantity_of_vacancies	average_salary_by_exp_level
1	Executive-level	5	234 176
2	Senior-level	5	235 768
3	Mid-level	1	258 000

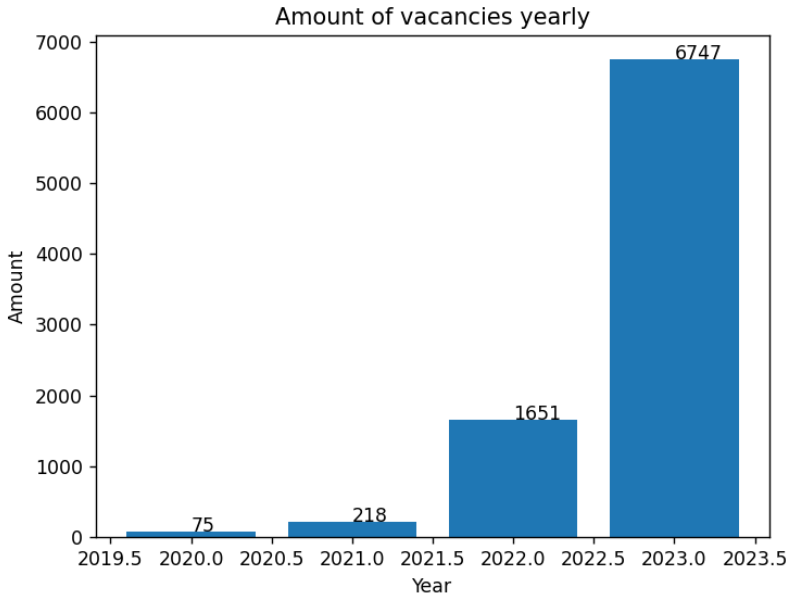
4 Визуализация

Для визуализации данных, полученных SQL запросов, использовался язык python. В качестве IDE был выбран PyCharm. Необходимо было представить

каждый запрос через датафрейм (через библиотеку pandas) с помощью метода .read_sql_qeury('запрос'). Для построения графиков использовалась библиотека matplotlib.

1. Визуализация задачи 1

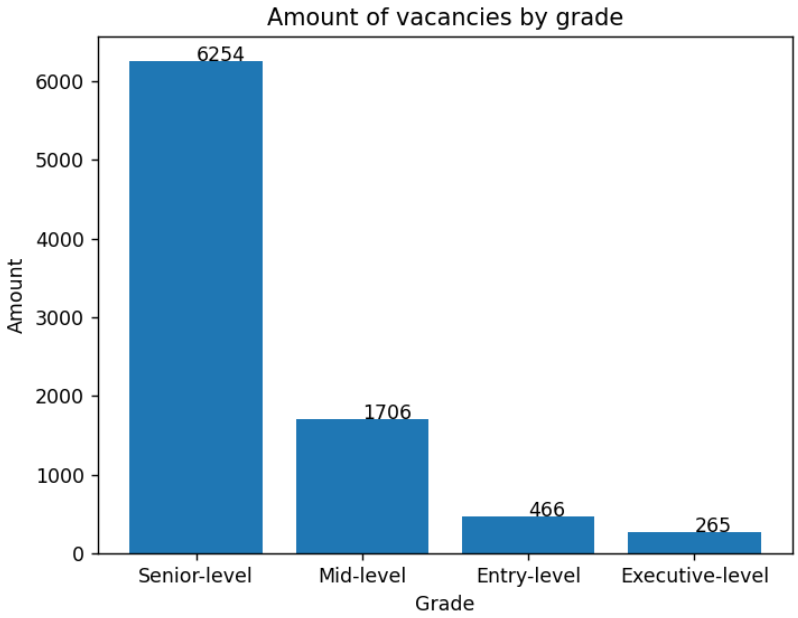
```
#визуаллизация задачи 1
sql_amount_of_vacancies_yearly = pd.read_sql_query( sql: '''SELECT
                                                    work_year,
                                                    COUNT(*) AS amount_of_vacancies
                                                    FROM data_salaries ds
                                                    GROUP BY work_year
                                                    ORDER BY amount_of_vacancies DESC''', con)
plt.bar(sql_amount_of_vacancies_yearly['work_year'], sql_amount_of_vacancies_yearly['amount_of_vacancies'])
for key, value in zip(sql_amount_of_vacancies_yearly['work_year'], sql_amount_of_vacancies_yearly['amount_of_vacancies']):
    plt.text(key, value, str(value))
plt.xlabel('Year')
plt.ylabel('Amount')
plt.title('Amount of vacancies yearly')
plt.show()
```



Вывод: по диаграмме видно, что данные о вакансиях на сайте начинали появляться с 2020 года. С каждым годом количество вакансий растёт. Заметный скачок в 2023 году.

2. Визуализация задачи 2

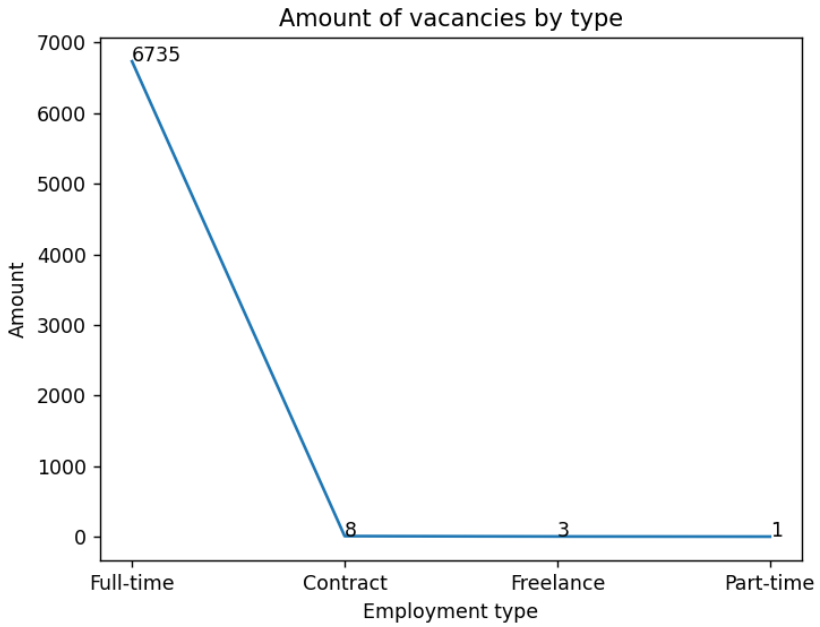
```
#визуализация задачи 2
sql_amount_of_vacancies_by_experience_level = pd.read_sql_query( sql: '''SELECT
                                                    experience_level,
                                                    COUNT(*) AS amount_of_vacancies_by_grade
                                                    FROM data_salaries ds
                                                    GROUP BY experience_level
                                                    ORDER BY amount_of_vacancies_by_grade DESC''', con)
plt.bar(sql_amount_of_vacancies_by_experience_level['experience_level'], sql_amount_of_vacancies_by_experience_level['amount_of_vacancies_by_grade'])
for key, value in zip(sql_amount_of_vacancies_by_experience_level['experience_level'], sql_amount_of_vacancies_by_experience_level['amount_of_vacancies_by_grade']):
    plt.text(key, value, str(value))
plt.xlabel('Grade')
plt.ylabel('Amount')
plt.title('Amount of vacancies by grade')
plt.show()
```



Вывод: по диаграмме видно, что на сайте преимущественно представлена информация по вакансиям уровня Сеньор.

3. Визуализация задачи 3

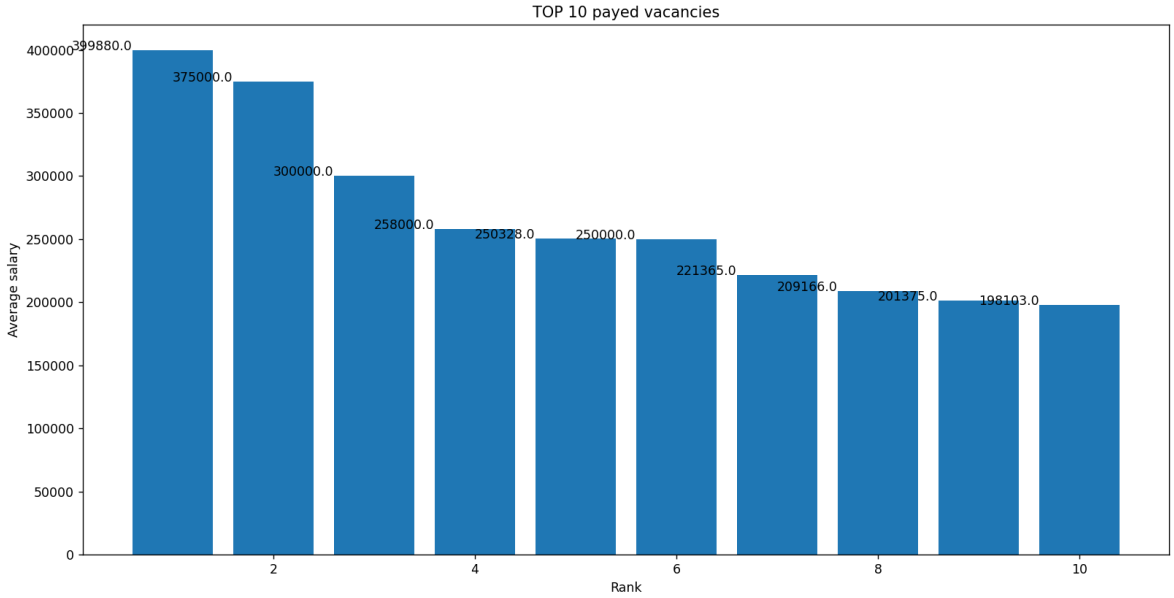
```
#Визуализация задачи 3
sql_amount_of_vacancies_by_type = pd.read_sql_query( sql: '''SELECT
                                                    employment_type,
                                                    COUNT(*) AS amount_of_vacancies_by_type
                                                    FROM data_salaries ds
                                                    WHERE work_year = '2023'
                                                    GROUP BY employment_type
                                                    ORDER BY amount_of_vacancies_by_type DESC''', con)
plt.plot( *args: sql_amount_of_vacancies_by_type['employment_type'], sql_amount_of_vacancies_by_type['amount_of_vacancies_by_type'])
for key, value in zip(sql_amount_of_vacancies_by_type['employment_type'], sql_amount_of_vacancies_by_type['amount_of_vacancies_by_type']):
    plt.text(key, value, str(value))
plt.xlabel('Employment type')
plt.ylabel('Amount')
plt.title('Amount of vacancies by type')
plt.show()

```



Вывод: ПО графику видно, что на сайте преобладают вакансии с полной занятостью

4. Визуализация задачи 4

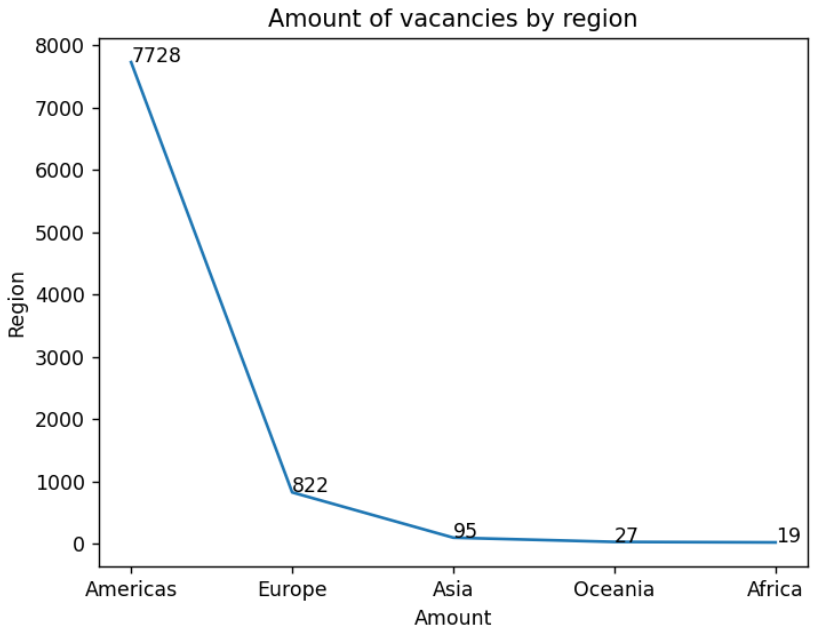
```
#Визуализация задачи 4
sql_top_10_most_payed_vacancies = pd.read_sql_query( sql: '''SELECT
                                                    job_title,
                                                    ROUND(AVG(salary_in_usd)) AS average_top_salary
                                                    FROM data_salaries ds
                                                    GROUP BY job_title
                                                    ORDER BY average_top_salary  DESC
                                                    LIMIT 10''', con)
plt.bar(range(1,11), sql_top_10_most_payed_vacancies['average_top_salary'])
for index, value in enumerate(sql_top_10_most_payed_vacancies['average_top_salary']):
    plt.text(index, value, str(value))
plt.xlabel('Rank')
plt.ylabel('Average salary')
plt.title('TOP 10 payed vacancies')
plt.show()

```



Вывод: по диаграмме видно, что разница меду средней максимальной зп и средней минимальной из выборки составляет около двухсот тысяч долларов

5. Визуализация задачи 5

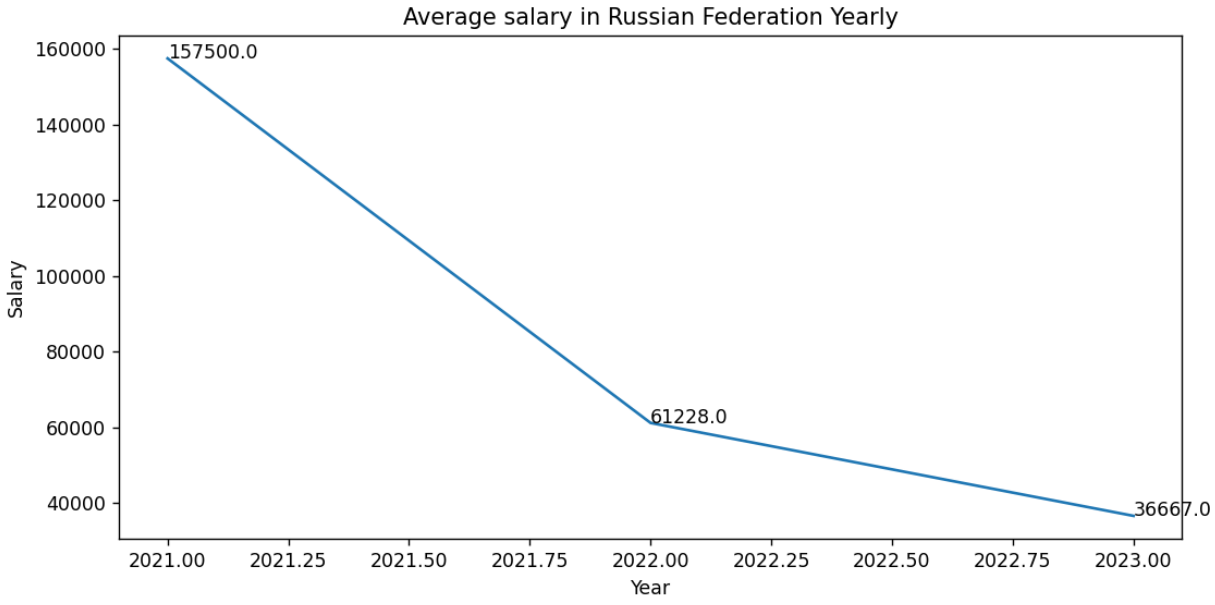
```
#Визуализация задачи 5
sql_amount_of_vacancies_by_region = pd.read_sql_query( sql: '''SELECT
                                                    ic.region,
                                                    COUNT(*) AS amount_of_vacancies
                                                    FROM data_salaries ds
                                                    INNER JOIN iso_code ic
                                                    ON ds.company_location = ic.name
                                                    GROUP BY ic.region
                                                    ORDER BY amount_of_vacancies DESC''', con)
plt.plot( *args: sql_amount_of_vacancies_by_region['region'], sql_amount_of_vacancies_by_region['amount_of_vacancies'])
for key, value in zip(sql_amount_of_vacancies_by_region['region'], sql_amount_of_vacancies_by_region['amount_of_vacancies']):
    plt.text(key, value, str(value))
plt.xlabel('Amount')
plt.ylabel('Region')
plt.title('Amount of vacancies by region')
plt.show()
```



Вывод: по графику видно, что на сайте преимущественно представлены вакансии из Америки. Европа на втором месте. Азия, Океания и Африка не набирают даже по 100 вакансий

6. Визуализация задачи 6

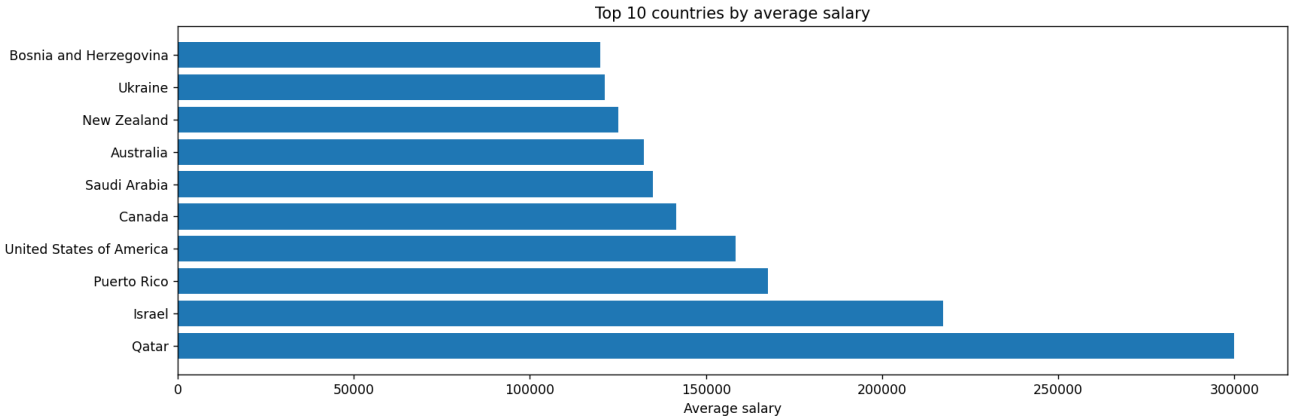
```
#Визуализация задачи 6
sql_average_salary_in_russia_yearly = pd.read_sql_query( sql: '''SELECT
                                                                work_year,
                                                                COUNT(*) AS vacancies_in_Russia,
                                                                ROUND(AVG(salary_in_usd)) AS average_salary_in_Russia
                                                                FROM data_salaries ds
                                                                WHERE company_location = 'Russian Federation'
                                                                GROUP BY work_year''', con)
plt.plot( *args: sql_average_salary_in_russia_yearly['work_year'], sql_average_salary_in_russia_yearly['average_salary_in_Russia'])
for key, value in zip(sql_average_salary_in_russia_yearly['work_year'], sql_average_salary_in_russia_yearly['average_salary_in_Russia']):
    plt.text(key, value, str(value))
plt.xlabel('Year')
plt.ylabel('Salary')
plt.title('Average salary in Russian Federation Yearly')
plt.show()
```



Вывод: по графику видно, что средняя зарплата представленных российских вакансий на сайте падает из года в год. Отмечу, что данный сайт не имеет большую популярность у пользователей из РФ

7. Визуализация задачи 7

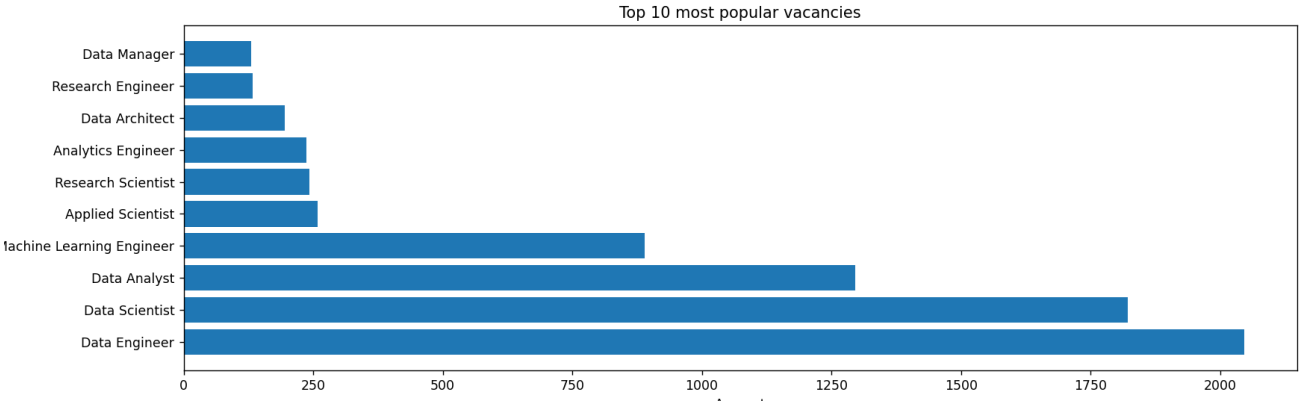
```
#Визуализация задачи 7
sql_top_10_countries_by_average_salary = pd.read_sql_query( sql: '''SELECT
                                                                company_location,
                                                                ROUND(AVG(salary_in_usd)) AS average_salary,
                                                                COUNT(salary_in_usd) AS quantity_of_vacancies
                                                                FROM data_salaries ds
                                                                GROUP BY company_location
                                                                ORDER BY average_salary DESC
                                                                LIMIT 10''', con)
plt.barh(sql_top_10_countries_by_average_salary['company_location'], sql_top_10_countries_by_average_salary['average_salary'])
plt.xlabel('Average salary')
plt.ylabel('Country')
plt.title('Top 10 countries by average salary')
plt.show()
```



Вывод: по диаграмме видно, что, если не учитывать количество представленных на сайте вакансий, то Катар занимает первое место по средней зарплате.

8. Визуализация задачи 8

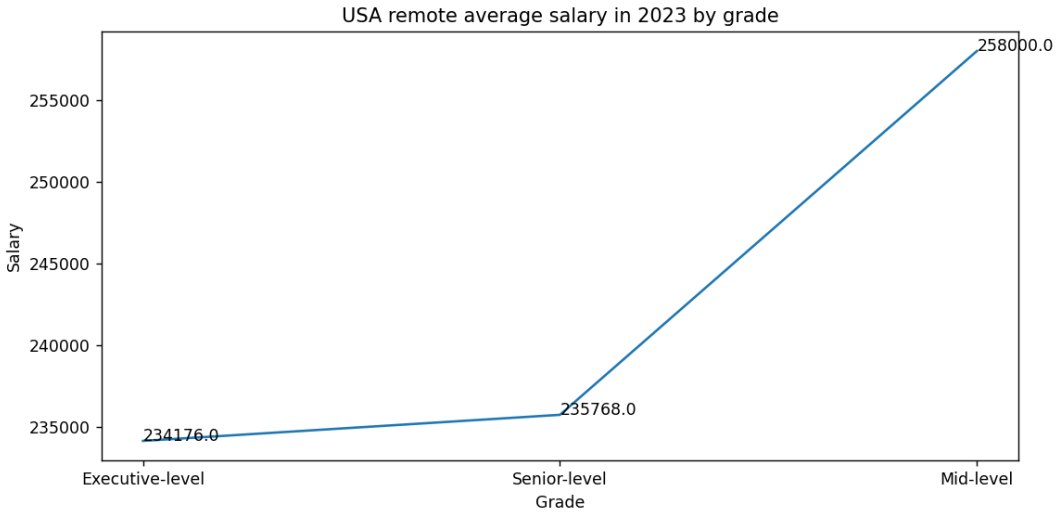
```
#Визуализация задачи 8
sql_top_10_most_popular_vacancies = pd.read_sql_query( sql: '''SELECT
                                                            job_title,
                                                            COUNT(*) AS quantity_of_vacancies
                                                            FROM data_salaries ds
                                                            GROUP BY job_title
                                                            ORDER BY quantity_of_vacancies DESC
                                                            LIMIT 10''', con)
plt.barh(sql_top_10_most_popular_vacancies['job_title'], sql_top_10_most_popular_vacancies['quantity_of_vacancies'])
plt.xlabel('Amount')
plt.ylabel('Job title')
plt.title('Top 10 most popular vacancies')
plt.show()
```



Вывод: В топ 3 самых популярных вакансий на сайте входят: Дата Инженер, Дата Саентист и Дата Аналитик

9. Визуализация задачи 9

```
#Визуализация задачи 9
sql_usa_remote_average_salary_in_2023_by_grade = pd.read_sql_query( sql: '''SELECT
                                                                        experience_level,
                                                                        COUNT(*) AS quantity_of_vacancies,
                                                                        ROUND(AVG(average_salary)) AS average_salary_by_exp_level
                                                                        FROM
                                                                            (SELECT
                                                                             job_title,
                                                                             experience_level,
                                                                             employment_type,
                                                                             remote_ratio,
                                                                             ROUND(AVG(salary_in_usd)) AS average_salary
                                                                             FROM data_salaries
                                                                             WHERE work_year = '2023'
                                                                             AND company_location = 'United States of America'
                                                                             AND remote_ratio = 'Fully Remote'
                                                                             GROUP BY job_title, experience_level, employment_type, remote_ratio
                                                                             HAVING average_salary > 200000
                                                                             ORDER BY average_salary DESC)
                                                                             GROUP BY experience_level
                                                                             ORDER BY average_salary_by_exp_level ASC''', con)
plt.plot( *args: sql_usa_remote_average_salary_in_2023_by_grade['experience_level'], sql_usa_remote_average_salary_in_2023_by_grade['average_salary_by_exp_level'])
for key, value in zip(sql_usa_remote_average_salary_in_2023_by_grade['experience_level'], sql_usa_remote_average_salary_in_2023_by_grade['average_salary_by_exp_level']):
  plt.text(key, value, str(value))
plt.xlabel('Grade')
plt.ylabel('Salary')
plt.title('USA remote average salary in 2023 by grade')
plt.show()
```



Вывод: исходя из графика кому-то в 2023 году в США очень повезло иметь удалённую работу на позиции мидла с зарплатой выше, чем у сеньоров и тех. диров.