Проект: Анализ рынка вакансий аналитиков данных и системных аналитиков

Автор: Алексей Котов

Почта: alexkotov1001@yandex.ru

Telegram: @kotov1001

Цель

Проанализировать базу данных вакансий, чтобы понять текущие тенденции на рынке труда для аналитиков данных и системных аналитиков.

Задачи

- Определить диапазон заработных плат: среднее значение, медиану, минимумы и максимумы нижних и верхних порогов зарплаты.
- Выявить регионы и компании, в которых сосредоточено наибольшее количество вакансий.
- Проанализировать, какие преобладают типы занятости, а также графики работы.
- Изучить распределение грейдов среди аналитиков данных и системных аналитиков.
- Выявить основных работодателей, предлагаемые зарплаты и условия труда для аналитиков.
- Определить наиболее востребованные навыки (как жёсткие, так и мягкие).

Загрузка библиотек и подключение к базе данных

In [23]:
```python
# Загрузка необходимых библиотек
# для работы с SQL
from sqlalchemy import create_engine, text
# для работы с таблицами
import pandas as pd
```

In [24]:
```python
# Создаем движок и подключаемся к базе данных
engine = create_engine("postgresql://<...>")

# Выполняем проверку соединения
with engine.connect() as conn: # гарантирует автоматическое закрытие соединения после выполнения блока (даже при ошибке)
    conn.execute(text("SELECT 1"))
    print("Подключение успешно!")
```

```
Подключение успешно!
```

In [25]:
```python
# Определим функцию для вывода запроса
def query_to_df(query, engine):
    with engine.connect() as conn: # гарантируем закрытие соединения после выполнения
        try: # Обработка ошибки
            return pd.read_sql(text(query), conn) # text() для явного указания, что это SQL-код
        except Exception as e: # Вывод ошибки
            print(f'Ошибка SQL-запроса:\n{e}')
            return None
```

Знакомство с данными

Проверим первые 10 записей базы данных.

In [26]:
```python
# Составляем SQL-запрос
query = """
--sql
SELECT *
FROM public.parcing_table
LIMIT 10;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[26]:

	id	name	published_at	employer	department	area	experience	schedule	employment	salary_from	salary_to	salary_bin	key_skills_1	key_skills_2	key_skill
0	100069131	Дата аналитик	2024-05-24 13:05:01	СБЕР	Сбер для экспертов	Санкт-Петербург	Junior+ (1-3 years)	Полный день	Полная занятость	NaN	NaN	ЗП не указана	Документация	Проактивность	Коммуника
1	100069821	Аналитик данных	2024-06-10 16:49:49	МТС	«МТС»	Казань	Junior+ (1-3 years)	Полный день	Полная занятость	72000.0	NaN	ЗП не указана		None	N
2	100071014	Аналитик данных	2024-06-07 11:08:22	Россети Урал	None	Екатеринбург	Junior+ (1-3 years)	Полный день	Полная занятость	51000.0	NaN	ЗП не указана	Аналитическое мышление	None	N
3	100077503	Data Analyst	2024-05-24 14:14:00	СБЕР	Сбер для экспертов	Москва	Middle (3-6 years)	Полный день	Полная занятость	NaN	NaN	ЗП не указана	Pandas	None	N
4	100077910	Data Analyst / Data Scientist	2024-06-11 14:17:47	Итсен	None	Москва	Middle (3-6 years)	Полный день	Полная занятость	350000.0	NaN	ЗП не указана	Linux	SQL	Бизн ана
5	100080002	Продуктовый аналитик / Data Analyst	2024-06-11 14:34:15	Photo Lab	None	Москва	Junior+ (1-3 years)	Удаленная работа	Полная занятость	135000.0	185000.0	От 100 тысяч до 200 тысяч	Linux	SQL	поведе пользовате
6	100080293	Data analyst / Аналитик данных	2024-05-24 14:36:52	Колл Солюшенс	None	Санкт-Петербург	Middle (3-6 years)	Полный день	Полная занятость	NaN	NaN	ЗП не указана	Анализ данных	Проактивность	
7	100082545	Аналитик данных (Отдел по анализу конкурентов)	2024-06-05 11:38:18	Ozon	Ozon Офис и Коммерция	Москва	Junior+ (1-3 years)	Полный день	Полная занятость	NaN	NaN	ЗП не указана		None	N
8	100084753	Data Analyst	2024-05-24 15:13:27	Самокат (ООО Умный ритейл)	None	Москва	Middle (3-6 years)	Полный день	Полная занятость	NaN	NaN	ЗП не указана		None	N
9	100087368	Data Analyst	2024-05-24 15:42:07	ЮТИМ	None	Москва	Junior+ (1-3 years)	Полный день	Полная занятость	NaN	200000.0	От 200 тысяч до 300 тысяч	Английский язык	MS SQL	

Данные соответсвуют описанию.

Проверим, какие вакансии представлены в базе.

In [27]:
```python
# Составляем SQL-запрос
query = """
--sql
SELECT DISTINCT name,
       COUNT(name)
FROM public.parcing_table
GROUP BY name
ORDER BY COUNT(name) DESC
LIMIT 10;
```

```
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[27]:

	name	count
0	Аналитик данных	463
1	Data Analyst	84
2	Ведущий аналитик данных	30
3	Data analyst	25
4	Старший аналитик данных	23
5	Младший аналитик данных	20
6	Аналитик данных IVR	14
7	Senior Data Analyst	14
8	Data analyst / Аналитик данных	12
9	Аналитик данных / Data Analyst	11

В базе данных представлены только релевантные вакансии аналитиков с различными формулировками. В дальнейшем будем рассматривать все вакансии из данной базы.

Дополнительно найдем количество вакансий, где в названии только строгие формулировки, и общее количество вакансий.

In [28]:
```
# Составляем SQL-запрос
query = """
--sql
SELECT
    COUNT(*) FILTER (
        WHERE name LIKE '%Аналитик данных%'
        OR name LIKE '%аналитик данных%'
                OR name LIKE '%Системный аналитик%'
                OR name LIKE '%системный аналитик%') AS count_after_filter,
    COUNT(*) AS count_before_filter
FROM public.parcing_table;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[28]:

	count_after_filter	count_before_filter
0	1326	1801

Вакансий со строгой формулировкой 1326 из 1801. Но остальные вакансии релевантны, поэтому будет рассматривать их все.

Диапазон заработных плат

Определим диапазон нижнего порога заработных плат: среднее значение, медиану, минимум и максимум.

In [29]:
```
# Составляем SQL-запрос
query = """
--sql
SELECT ROUND(MIN(salary_from)) AS salary_from_min,
       ROUND(MAX(salary_from)) AS salary_from_max,
           ROUND(AVG(salary_from)) AS salary_from_avg,
       PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY salary_from) AS salary_from_med
FROM public.parcing_table
-- отсекаем аномально низкие (например, в тысячах руб.)
WHERE salary_from IS NOT NULL AND salary_from > 1000;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[29]:

	salary_from_min	salary_from_max	salary_from_avg	salary_from_med
0	25000.0	398000.0	109841.0	90000.0

- Минимум нижнего порога: 25 000
- Максимум нижнего порога: 398 000
- Среднее значение нижнего порога: 109 841
- Медиана нижнего порога: 90 000

Определим диапазон верхнего порога заработных плат: среднее значение, медиану, минимум и максимум.

In [30]:
```
# Составляем SQL-запрос
query = """
--sql
SELECT ROUND(MIN(salary_to)) AS salary_to_min,
          ROUND(MAX(salary_to)) AS salary_to_max,
          ROUND(AVG(salary_to)) AS salary_to_ang,
          PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY salary_to) AS salary_to_med
FROM public.parcing_table
-- отсекаем аномально низкие (например, в тысячах руб.)
WHERE salary_to IS NOT NULL AND salary_to > 1000;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[30]:

	salary_to_min	salary_to_max	salary_to_ang	salary_to_med
0	25000.0	497500.0	153847.0	120000.0

- Минимум верхнего порога: 25 000
- Максимум верхнего порога: 497 500
- Среднее значение верхнего порога: 153 847
- Медиана верхнего порога: 120 000

Распределение вакансий по зарплатной вилке.

In [31]:
```
# Составляем SQL-запрос
query = """
--sql
SELECT salary_bin, COUNT(salary_bin) AS count_salary_bin
FROM public.parcing_table
```

```
GROUP BY salary_bin
ORDER BY COUNT(salary_bin) DESC
LIMIT 10;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[31]:

	salary_bin	count_salary_bin
0	ЗП не указана	1573
1	От 100 тысяч до 200 тысяч	96
2	Меньше 100 тысяч	68
3	От 200 тысяч до 300 тысяч	36
4	Больше 300 тысяч	28

В большинстве вакансий не указана отдельно зарплатная вилка.

Регионы и компании

Выявим регионы, в которых сосредоточено наибольшее количество вакансий.

In [32]:
```
# Составляем SQL-запрос
query = """
--sql
SELECT area,
          COUNT(area) AS count_area
FROM public.parcing_table
GROUP BY area
ORDER BY COUNT(area) DESC
LIMIT 10;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[32]:

	area	count_area
0	Москва	1247
1	Санкт-Петербург	181
2	Екатеринбург	51
3	Нижний Новгород	33
4	Новосибирск	33
5	Владивосток	31
6	Казань	29
7	Краснодар	22
8	Самара	11
9	Ростов-на-Дону	10

Число вакансий в Москве больше остальных, вместе взятых. Причина - большое число Бигтех компаний имеет головные офисы именно в Москве.

Выявим компании, в которых сосредоточено наибольшее количество вакансий.

In [33]:
```
# Составляем SQL-запрос
query = """
--sql
SELECT employer,
          COUNT(employer) AS count_employer
FROM public.parcing_table
GROUP BY employer
ORDER BY COUNT(employer) DESC
LIMIT 10;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[33]:

	employer	count_employer
0	СБЕР	243
1	WILDBERRIES	43
2	Ozon	34
3	Банк ВТБ (ПАО)	28
4	Т1	26
5	МАГНИТ, Розничная сеть	24
6	МТС	22
7	Okko	19
8	Центральный банк Российской Федерации	16
9	Правительство Москвы	15

Лидеры - это банки и маркетплейсы, которые у всех на слуху. Стоит отметить, что на 9 и 10 месте государственные структуры - это подчеркивает важность анализа данных в управленческих решениях.

Тип занятости и график работы

Проанализируем, какие преобладают типы занятости.

In [34]:
```
# Составляем SQL-запрос
query = """
--sql
SELECT employment,
          COUNT(employment) AS count_employment
FROM public.parcing_table
GROUP BY employment
ORDER BY COUNT(employment) DESC;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[34]:

	employment	count_employment
0	Полная занятость	1764
1	Частичная занятость	16
2	Стажировка	16
3	Проектная работа	5

Почти все вакансии предполагают полную занятость. Аналитику данных часто приходится решать ad hoc задачи для принятия оперативного управленческого решения.

Проанализируем, какие преобладают графики работы.

In [35]:
```
# Составляем SQL-запрос
query = """
--sql
SELECT schedule,
           COUNT(schedule) AS count_schedule
FROM public.parcing_table
GROUP BY schedule
ORDER BY COUNT(schedule) DESC;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[35]:

	schedule	count_schedule
0	Полный день	1441
1	Удаленная работа	310
2	Гибкий график	41
3	Сменный график	9

Подавляющее количество вакансий в офисе, но примерно 1/5 часть всех вакансий дает возможность кандидату работать удаленно или по гибкому графику.

Распределение грейдов

In [36]:
```
# Составляем SQL-запрос
query = """
--sql
SELECT experience,
           COUNT(experience) AS count_experience
FROM public.parcing_table
GROUP BY experience
ORDER BY COUNT(experience) DESC;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[36]:

	experience	count_experience
0	Junior+ (1-3 years)	1091
1	Middle (3-6 years)	555
2	Junior (no experince)	142
3	Senior (6+ years)	13

Больше половины вакансий для кандидатов с опытом 1-3 года (Junior+) - такие специалисты уже могут решать самостоятельно большинство рабочих задач.

Значительная часть для кандидатов с опытом 3-6 лет (Middle) - такие специалисты могут уже обладать смежными компетенциями, такими как машинное обучение и/или инженерия данных.

Новички без опыта (Junior) или опытные специалисты с опытом больше 6 лет (Senior) востребованы меньше всего. Но важно понимать, что Senior-аналитиков могут чаще искать по другим каналам.

Условия основных работодателей

Выявим основных работодателей, предлагаемые, зарплаты и условия труда для аналитиков.

In [37]:
```
# Составляем SQL-запрос
query = """
--sql
SELECT employer,
       employment,
        schedule,
       ROUND(AVG(salary_from)) AS salary_from_AVG,
       ROUND(AVG(salary_to)) AS salary_to_AVG,
       COUNT(employer)
FROM public.parcing_table
WHERE name LIKE '%Аналитик данных%' OR name LIKE '%Системный аналитик%'
GROUP BY employer,
         employment,
         schedule
ORDER BY COUNT(employer) DESC
LIMIT 10;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[37]:

	employer	employment	schedule	salary_from_avg	salary_to_avg	count
0	СБЕР	Полная занятость	Полный день	112000.0	73333.0	103
1	Банк ВТБ (ПАО)	Полная занятость	Полный день	NaN	NaN	25
2	Ozon	Полная занятость	Полный день	NaN	NaN	18
3	T1	Полная занятость	Полный день	NaN	NaN	17
4	Правительство Москвы	Полная занятость	Полный день	NaN	NaN	15
5	Яндекс	Полная занятость	Полный день	NaN	NaN	14
6	WILDBERRIES	Полная занятость	Полный день	NaN	NaN	12
7	ГКУ Центр занятости населения города Москвы	Полная занятость	Полный день	101667.0	108000.0	12
8	Центральный банк Российской Федерации	Полная занятость	Полный день	NaN	NaN	12
9	Ростелеком	Полная занятость	Удаленная работа	NaN	100000.0	11

Подавляющее большинство вакансий у Сбера. Средняя зарплата «до» ниже средней зарплаты «от» — скорее всего, часто пишут сумму «до», не указывая сумму «от». Но следует отметит, что у многих вакансий вообще не указана зарплата. Основной тип занятости - полная, и график работы - полный день.

Востребованные навыки

Определим сочетание двух самых востребованных жестких навыков у специалистов.

In [38]:
```sql
# Составляем SQL-запрос
query = """
--sql
SELECT key_skills_1,
       key_skills_2,
           COUNT(*)
FROM public.parcing_table
WHERE key_skills_1 != ''
GROUP BY key_skills_1,
         key_skills_2
ORDER BY COUNT(*) DESC
LIMIT 10;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[38]:

	key_skills_1	key_skills_2	count
0	Анализ данных	SQL	136
1	Pandas	None	84
2	Документация	None	54
3	SQL	Python	49
4	Аналитическое мышление	None	45
5	Python	SQL	45
6	Коммуникация	None	39
7	SQL	Power BI	28
8	MS SQL	Анализ данных	27
9	Анализ данных	Коммуникация	26

Для вакансий требуется уметь проводить анализ данных, знать SQL и Python (в частности библиотеку Pandas). Надо обратить внимание, что "коммуникация" выделена для многих вакансий как основной жесткий навык.

Определим сочетание двух самых востребованных мягких навыков у специалистов.

In [39]:
```sql
# Составляем SQL-запрос
query = """
--sql
SELECT soft_skills_1,
       soft_skills_2,
           COUNT(*)
FROM public.parcing_table
WHERE soft_skills_1 != ''
GROUP BY soft_skills_1,
         soft_skills_2
ORDER BY COUNT(*) DESC
LIMIT 10;
"""
# Выполняем запрос
df = query_to_df(query, engine)
# Выводим результат
df
```

Out[39]:

	soft_skills_1	soft_skills_2	count
0	Документация	None	214
1	Аналитическое мышление	None	109
2	Коммуникация	None	105
3	Коммуникация	Документация	46
4	Проактивность	None	30
5	Документация	Аналитическое мышление	12
6	Коммуникация	Проактивность	8
7	Проактивность	Аналитическое мышление	7
8	Коммуникация	Аналитическое мышление	7
9	Креативность	Аналитическое мышление	6

Работодатели в качестве мягких навыков выделяют: умение работать с документами (видимо, подразумевается подготовка отчетов и презентаций), аналитическое мышление, коммуникацию и проактивность. Следует отметить, что подобные мягкие навыки традиционно требуются для большинства вакансий.

Выводы

1. **Объём базы данных и релевантность вакансий:**

- в базе данных 1801 вакансия, из них 1326 содержат строгие формулировки («Аналитик данных», «Системный аналитик»);
- остальные вакансии также релевантны и учтены в анализе.

2. **Диапазон заработных плат:**

- **нижний порог зарплат:**
  - минимум — 25 000 руб.;
  - максимум — 398 000 руб.;
  - среднее значение — 109 841 руб.;
  - медиана — 90 000 руб.;
- **верхний порог зарплат:**
  - минимум — 25 000 руб.;
  - максимум — 497 500 руб.;
  - среднее значение — 153 847 руб.;
  - медиана — 120 000 руб.
- наблюдается значительный разброс зарплат, что говорит о разнообразии предложений на рынке.

3. **География вакансий:**

- лидирует **Москва** — здесь сосредоточено больше всего вакансий (больше, чем в остальных регионах вместе взятых);
- причина — наличие головных офисов крупных IT- и финансовых компаний в столице.

4. **Основные работодатели:**

- лидируют **банки и маркетплейсы** — они активно нанимают аналитиков данных;
- присутствуют и **государственные структуры** (9–10 место в рейтинге), что подчёркивает важность аналитики данных в госуправлении.

5. **Тип занятости и график работы:**

- **преобладает полная занятость** (почти все вакансии);
- **график работы:** большинство вакансий — офис, но около 1/5 предлагают удалённую работу или гибкий график.

6. **Распределение по грейдам (опыту работы):**

- **наиболее востребованы специалисты с опытом 1–3 года (Junior+)** — могут самостоятельно решать большинство рабочих задач;
- **значительная доля вакансий для специалистов с опытом 3–6 лет (Middle)** — ожидается владение смежными компетенциями (машинное обучение, инженерия данных);
- **меньше всего востребованы:**
  - новички без опыта (Junior);
  - опытные специалисты с опытом более 6 лет (Senior) — возможно, таких специалистов чаще ищут по другим каналам.

7. **Условия труда у основных работодателей:**

- **основной тип занятости** — полная;
- **график работы** — полный день;
- средние зарплаты («от» и «до») различаются, часто указывается только верхняя граница зарплаты;
- **Сбер** — доминирующий работодатель в исследуемой выборке.

8. **Востребованные навыки:**

- **жёсткие навыки (hard skills):**
  - анализ данных;
  - знание SQL;
  - владение Python (в частности, библиотекой Pandas);
  - коммуникация (выделена как ключевой навык для многих вакансий).
- **мягкие навыки (soft skills):**
  - работа с документами (подготовка отчётов и презентаций);
  - аналитическое мышление;
  - коммуникация;
  - проактивность.

9. **Общие тенденции рынка:**

- высокий спрос на аналитиков данных и системных аналитиков;
- баланс между офисной и удалённой работой (с преобладанием офисной);
- акцент на специалистов среднего уровня (Junior+ и Middle);
- важность как технических, так и коммуникативных навыков.

Итог: рынок вакансий для аналитиков данных и системных аналитиков активен, с широким диапазоном зарплат и разнообразием условий труда. Наибольший спрос наблюдается в Москве, среди крупных банков, маркетплейсов и госструктур. Работодатели ищут специалистов с опытом 1–6 лет, владеющих ключевыми инструментами аналитики и обладающих сильными коммуникативными навыками.