

# АНАЛИЗ ВАКАНСИЙ с HEADHUNTER

*исполнитель: ФИЛОНЕНКО АЛЕКСАНДР*

# ЦЕЛЬ:

**провести анализ вакансии по специализации аналитика данных  
на основе файлов с выгрузкой данных с сайта headhunter**

**Файлы для анализа:**

`vacancies_data_for_analysis.xlsx`  
`area.xlsx`  
`employers.xlsx`

## ПОЛУЧИТЬ ОТВЕТЫ НА СЛЕДУЮЩИЕ ЗАДАЧИ(вопросы):

1. **Определить общее кол-во вакансий.**
2. **Найти дату последней опубликованной вакансии**
3. **Найти вакансию с максимально предлагаемой зарплатой.**
4. **Определить среднюю вилку зарплат.**
5. **Определить, сколько вакансий предлагают работу без опыта.**
6. **Определить, сколько вакансий относятся к разным значениям.**
7. **Определить средние вилки зарплат в разрезе опыта работы.**
8. **Определить количество вакансий по городам.**
9. **Определить топ - 10 рейтинг работодателей.**

# ЭТАП(Задача) 1

*Прочитать файлы excel в датафреймы и провести первоначальный анализ и очистку данных:*

- 1)Сделать отчет с помощью библиотеки pandas\_profiling;**
- 2)Удалить дубликаты строк данных при необходимости;**
- 3)Оставить в датафрейме по вакансиям только нужные столбцы:**

Список столбцов из датафрейма вакансий:

'id', 'name', 'area.id', 'address.metro.station\_name', 'salary.from', 'salary.to',  
'address.raw', 'experience.name', 'schedule.name', 'employment.name',  
'employer.id', 'alternate\_url', 'created\_at'

FILONENKO\_PROJECT\_SQL.ipynb

ФайлИзменитьВидВставкаСреда выполненияИнструментыСправкаИзменения сохранены

КомментироватьПоделиться

ОЗУДиск

Файлы

sample\_dataarea.xlsxemployer.xlsxvacancies\_data\_for\_analysis.xlsx

+ Код+ Текст

0сек.

[23] #импортируем pandas\_profiling  
import pandas\_profiling

0сек.

[24] #готовим датафрейм 'area' с данными  
df1=pd.read\_excel('/content/area.xlsx')

df1


[16] #проводим анализ-отчет датафрейма 'area' с помощью pandas\_profiling  
df1.profile\_report()

[18] #готовим датафрейм 'employer' с данными  
df2=pd.read\_excel('/content/employer.xlsx')

[19] df2

[20] #проводим анализ-отчет датафрейма 'employer' с помощью pandas\_profiling  
df2.profile\_report()





[ ] # готовим датафрейм 'vacancies\_data\_for\_analysis' с данными  
df3=pd.read\_excel('/content/vacancies\_data\_for\_analysis.xlsx')

 **FILONENKO\_PROJECT\_SQL.ipynb** ☆

ФайлИзменитьВидВставкаСреда выполненияИнструментыСправкаИзменения сохранены

КомментироватьПоделиться⚙️

Файлы



..  
sample\_data  
area.xlsx  
employer.xlsx  
vacancies\_data\_for\_analysis.xlsx

+ Код + Текст

ОЗУ  
Диск

✓  
0 сек.

[10] # импортируем библиотеки sqlite3 и pandas  
import sqlite3  
import pandas as pd

✓

[11] # устанавливаем pandas\_profiling  
!pip install pandas\_profiling[notebook]


✓  
0 сек.

[12] # импортируем pandas\_profiling  
import pandas\_profiling

✓  
0 сек.

[13] #готовим таблицу 'area'с данными  
df1=pd.read\_excel('/content/area.xlsx')

✓  
0 сек.

 df1

1 to 6 of 6 entries Filter

index	area_id	area_name
0	1	Москва
1	2	Санкт-Петербург
2	3	Екатеринбург

FILONENKO\_PROJECT\_SQL.ipynb ☆

КомментироватьПоделиться

ФайлИзменитьВидВставкаСреда выполненияИнструментыСправкаИзменения сохранены

2 сек.

↑↓↻💬⚙

sample\_data  
area.xlsx  
employer.xlsx  
vacancies\_data\_for\_analysis.xlsx

OverviewAlerts 5Reproduction

Dataset statistics

Number of variables	2
Number of observations	6
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	224.0 B
Average record size in memory	37.3 B

Variable types

Numeric	1
Categorical	1

Доступно: 81.88 GB

# Отчет по датафрейму 'area':

Число столбцов(атрибутов)(Number of variables)-2

Число строк(Number of observations)-6

Пустые(отсутствующие) значения(клетки)(Missing cells)-0

Дубликаты(Duplicate rows)-0



FILONENKO\_PROJECT\_SQL.ipynb ☆

Файл Изменить Вид Вставка Среда выполнения Инструменты Справка Изменения сохранены

Комментировать Поделить

Файлы

- ..
- sample\_data
- area.xlsx
- employer.xlsx
- vacancies\_data\_for\_analysis.xlsx

+ Код + Текст

4 сек.

# Overview

Overview Alerts 6 Reproduction

## Dataset statistics

Number of variables	3
Number of observations	123
Missing cells	2
Missing cells (%)	0.5%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	3.0 KiB
Average record size in memory	25.0 B

## Variable types

Numeric	1
Categorical	2

Ресурсы: 91.88 GB

# Отчет по датафрейму 'employer':

Число столбцов(атрибутов)(Number of variables)-3

Число строк(Number of observations)-123

Пустые(отсутствующие) значения(клетки)(Missing cells)-2

Дубликаты(Duplicate rows)-0

FILONENKO\_PROJECT\_SQL.ipynb ☆

КомментироватьПоделиться⚙️

ФайлИзменитьВидВставкаСреда выполненияИнструментыСправкаИзменения сохранены

Файлы

+

sample\_dataarea.xlsxemployer.xlsxvacancies\_data\_for\_analysis.xlsx

+ Код+ Текст

df2.profile\_report()

[29] # готовим датафрейм 'vacancies\_data\_for\_analysis' с данными  
df3=pd.read\_excel('/content/vacancies\_data\_for\_analysis.xlsx')

df3

[ ] # проводим анализ-отчет датафрейма 'vacancies\_data\_for\_analysis' с помощью pandas\_profiling  
df3.profile\_report()

[ ] # удаление дубликатов из датафрейма 'vacancies\_data\_for\_analysis'  
df3 = df3.drop\_duplicates()

[ ] df3

[ ] # Определение нужных столбцов и сохранение результатов  
df3 = df3[['id', 'name', 'area.id','address.metro.station\_name','salary.from','salary.to','address.raw','experien

[ ] df3

ФайлИзменитьВидВставкаСреда выполненияИнструментыСправкаИзменения сохранены

Файлы

sample\_dataarea.xlsxemployer.xlsxvacancies\_data\_for\_analysis.xlsx

ДискДоступно: 81.88 GB

ОЗУДиск

↑↓↻⌨⚙

Overview

Alerts121Reproduction

Dataset statistics

Number of variables	88
Number of observations	174
Missing cells	5232
Missing cells (%)	34.2%
Duplicate rows	16
Duplicate rows (%)	9.2%
Total size in memory	105.5 KiB
Average record size in memory	620.7 B

Variable types

Numeric	11
Boolean	12
Categorical	51
Unsupported	14

# Отчет по датафрейму 'vacancies\_data\_for\_analisy':

Число столбцов(атрибутов)(Number of variables)-88

Число строк(Number of observations)-174

Пустые(отсутствующие) значения(клетки)(Missing  
cells)-5232

Дубликаты(Duplicate rows)-16

Файл Изменить Вид Вставка Среда выполнения Инструменты Справка Изменения сохранены Комментировать Поделиться ОЗУ Диск

Файлы

- ..
- sample\_data
- area.xlsx
- employer.xlsx
- vacancies\_data\_for\_analysis.xlsx

+ Код + Текст

```
[31] df3.profile_report()
```

```
[32] # удаление дубликатов из датафрейма 'vacancies_data_for_analysis'
df3 = df3.drop_duplicates()
```

```
[33] df3
```

```
[34] # Определение нужных столбцов и сохранение результатов
df3 = df3[['id', 'name', 'area.id', 'address.metro.station_name', 'salary.from', 'salary.to', 'address.raw', 'experience.name', 'schedule.name', 'employment.name', 'employer.id', 'alternate_url', 'created_at']]
```

```
[35] df3
```

```
df3.columns
```

```
Index(['id', 'name', 'area.id', 'address.metro.station_name', 'salary.from',
       'salary.to', 'address.raw', 'experience.name', 'schedule.name',
       'employment.name', 'employer.id', 'alternate_url', 'created_at'],
      dtype='object')
```

17сек.

Render HTML: 100%

1/1 [00:00<00:00, 1.23it/s]

ОЗУ

Диск

Pandas Profiling Report

OverviewVariablesInteractionsCorrelationsMissing values

Dataset statistics

Number of variables	13
Number of observations	158
Missing cells	450
Missing cells (%)	21.9%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	21.4 KiB
Average record size in memory	138.4 B

Variable types

Numeric	5
Categorical	8

15

# Отчет по датафрейму 'vacancies\_data\_for\_analisys' после очистки и преобразования:

Число столбцов(атрибутов)(Number of variables)-13

Число строк(Number of observations)-158

Пустые(отсутствующие) значения(клетки)(Missing cells)-450

Дубликаты(Duplicate rows)-0



## ЭТАП(Задача) 2

*Загрузить с помощью Python данные в таблицы базы данных  
(использовал библиотеку pandas)*

*Таблицы следующего формата:*

Таблица vacancies

Column Name
123 id
ABC name
123 area.id
ABC address.metro.station_name
123 salary.from
123 salary.to
ABC address.raw
ABC experience.name
ABC schedule.name
ABC employment.name
123 employer.id
ABC alternate_url
ABC created_at

Таблица employer

Column Name
123 employer_id
ABC employer_name
ABC employer_url

Таблица area

Column Name
123 area_id
ABC area_name



+ Код + Текст

```
[ ] # подключение к базе данных, создание новой базы 'vacancies_data_analysis.db'  
con=sqlite3.connect('vacancies_data_analysis.db', timeout=10)  
cur=con.cursor()
```

```
[ ] # загружаем таблицу 'area' в базу данных  
df1.to_sql(con=con, name='area', index=False)
```










6

```
[ ] # загружаем таблицу 'employer' в базу данных  
df2.to_sql(con=con, name='employer', index=False)
```

123

```
[ ] # загружаем таблицу 'vacancies' в базу данных  
df3.to_sql(con=con, name='vacancies', index=False)
```

158

-  coffe\_shop\_practice
-  sales
-  vacancies\_data\_analysis.db
  -  Таблицы
    -  area
    -  employer
    -  vacancies
      -  Колонки
        - 123 id (INTEGER)
        - ABC name (TEXT)
        - 123 area.id (INTEGER)
        - ABC address.metro.station\_name (TEXT)
        - 123 salary.from (REAL)
        - 123 salary.to (REAL)
        - ABC address.raw (TEXT)
        - ABC experience.name (TEXT)
        - ABC schedule.name (TEXT)
        - ABC employment.name (TEXT)
        - 123 employer.id (REAL)
        - ABC alternate\_url (TEXT)
        - ABC created\_at (TEXT)
      -  Ключи

Файл Редактирование Навигация Search Редактор SQL База данных Окна Справка

SQL Commit Rollback Auto Auto vacancies\_data\_analysis.db < N/A >

Базы данных × Проекты

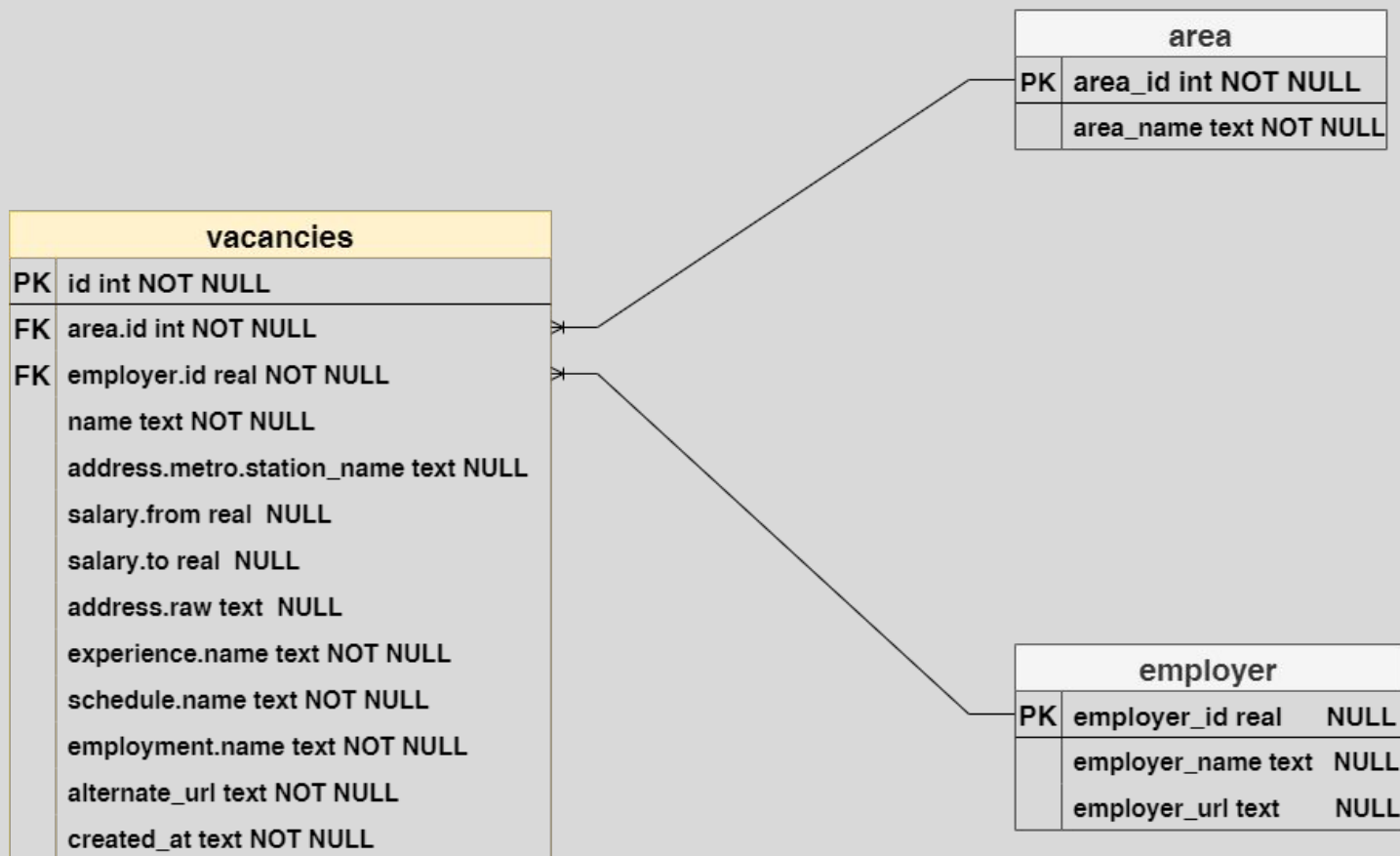
Введите часть имени объекта для поиска

- > coffe\_shop\_practice
- > sales
- ▼ vacancies\_data\_analysis.db
  - ▼ Таблицы
    - ▼ area
      - ▼ Колонки
        - 123 area\_id (INTEGER)
        - ABC area\_name (TEXT)
      - > Ключи
      - > Внешние ключи
      - > Индексы
      - > Ссылки
      - > Триггеры
    - ▼ employer
      - ▼ Колонки
        - 123 employer\_id (REAL)
        - ABC employer\_name (TEXT)
        - ABC employer\_url (TEXT)
      - > Ключи
      - > Внешние ключи
      - > Индексы
      - > Ссылки
      - > Триггеры
    - > vacancies
  - > Представления

## ЭТАП(Задание) 3

*Описать физическую модель данных для 3 таблиц,  
загруженных в базу данных  
(использовал инструмент draw.io)*

## ФИЗИЧЕСКАЯ МОДЕЛЬ



## ЭТАП(Задача) 4

*С помощью SQL-запросов решить поставленные задачи  
(получить ответы)  
(скриншоты с DBeaver)*



--ЗАДАЧА 1. ПОСЧИТАТЬ ОБЩЕЕ КОЛИЧЕСТВО ВАКАНСИЙ.

```
SELECT
    COUNT(*) AS count_vac
FROM
    vacancies v
```

Результат 1 ✕

SELECT COUNT(\*) AS count\_vac FROM vacancies v | Введите SQL выражение чтобы отфильтровать результаты

[illegible]

```
SELECT
    MAX(created_at)
FROM
    vacancies v
```

`SELECT MAX(created_at) FROM vacancies v`  Введите SQL выражение чтобы отфильтровать результаты

26

--ЗАДАЧА 3. НАЙТИ ВАКАНСИЮ С МАКСИМАЛЬНО ПРЕДЛАГАЕМОЙ ЗАРПЛАТОЙ ПО ВИЛКЕ.

```
--id,  
name,  
MAX("salary.to" ) AS max_salary  
FROM  
vacancies v
```

vacancies 1 ×

SELECT name, MAX("salary.to" ) AS max\_salary FROM vaca

[illegible]

--ЗАДАЧА 4. ПОСЧИТАТЬ СРЕДНЮЮ ВИЛКУ ЗАРПЛАТ.  
-- Нижнюю границу зарплаты посчитать как среднее по всем указанным в вакансиях salary\_from.  
--Верхнюю границу вилки посчитать аналогично, только по полю salary.to.

```
SELECT  
    ROUND(AVG("salary.from")) AS avg_salary_from,  
    ROUND(AVG("salary.to")) AS avg_salary_to  
FROM  
    vacancies v
```

Результат 1 X

SELECT ROUND(AVG("salary.from")) AS avg\_salary\_from, R Введите SQL выражение чтобы отфильтровать результаты

	123 avg_salary_from	123 avg_salary_to
1	138 697	179 783

--ЗАДАЧА 5. ПОСЧИТАТЬ СКОЛЬКО ВАКАНСИЙ ПРЕДЛАГАЕТ РАБОТУ БЕЗ ОПЫТА.

```
SELECT "experience.name",  
       COUNT(*)  
FROM   vacancies v  
WHERE  "experience.name" = 'Нет опыта'
```


vacancies 1 X

SELECT "experience.name", COUNT(\*) FROM vacancies v | Введите SQL выражение чтобы отфильтровать результаты

	ABC experience.name	123 COUNT(*)
1	Нет опыта	6

--ЗАДАЧА 6. ПОСЧИТАТЬ СКОЛЬКО ВАКАНСИЙ ОТНОСИТСЯ К РАЗНЫМ ЗНАЧЕНИЯМ `schedule.name` ('Полный день', 'Удаленная работа', 'Гибкий график').

```
SELECT
    "schedule.name",
    COUNT(*)
FROM
    vacancies v
WHERE
    "schedule.name" IN ('Полный день', 'Удаленная работа', 'Гибкий график')
GROUP BY
    "schedule.name";
```

vacancies 1 

```
'SELECT "schedule.name", COUNT(*) FROM vacancies v WHERE
```

[illegible]



area 1 X

PHILIPPO I  1752 I 1753

831110



--Задача 9. Сделать топ-10 рейтинг работодателей с их названиями по числу опубликованных вакансий.

```
SELECT
    e.employer_name,
    COUNT(*) AS vacancy_count
FROM
    vacancies v
JOIN employer e
    ON e.employer_id = v."employer.id"
GROUP BY
    e.employer_name
ORDER BY
    vacancy_count DESC
LIMIT 10;
```

employer 1 X

SELECT e.employer\_name, COUNT(\*) AS vacancy\_count FF Введите SQL выражение чтобы отфильтровать результаты

	Таблица	123	
		ABC employer_name	vacancy_count
1		Ozon	6
2		Газпромбанк	5
3		Тинькофф	4
4		РОСБАНК	4
5		Альфа-Банк	4
6		билайн	3
7		Центральный банк Российской Федерации	3
8		СБЕР	3
9		Банк ВТБ (ПАО)	3
10		Яндекс	2

# РЕЗУЛЬТАТЫ:

1. *Общее количество вакансий-158.*
2. *Дата последней опубликованной вакансии: 2023-06-22T21:45:35+0300*
3. *Вакансия с максимально предлагаемой зарплатой -  
Аналитик DWH (350000 руб.).*
4. *Средняя вилку зарплат: от 138697 руб. до 179783 руб.*
5. *Количество вакансий, где предлагают работу без опыта-6.*
6. *Количество вакансий относительно режима работы:  
Гибкий график-6, Полный день-118, Удаленная работа-34.*

# РЕЗУЛЬТАТЫ:

*7. Средние вилки зарплат в разрезе опыта работы:*

**Более 6 лет- от 300 000 до 300 000,  
Нет опыта - от 45 000 до 70 000,  
От 1 года до 3 лет от 111667 до 141250,  
От 3 до 6 лет от 158 000 до 230 000.**

*8. Количество вакансий по городам:*

**Воронеж-1, Екатеринбург-2, Казань-1, Москва-146, Новосибирск-1,  
Санкт-Петербург-7.**

*9. Топ- 10 рейтинг работодателей:*

**1.Ozon, 2.Газпромбанк, 3.Тинькофф, 4.РОСБАНК, 5.Альфа-Банк, 6.  
Билайн,  
7.Центральный банк, 8.СБЕР, 9.Банк ВТБ (ПАО), 10.Яндекс.**

# ВЫВОДЫ:

1. На данный момент в общем вакансия Аналитика является востребованной.
2. Специальность наиболее востребована в г. Москва.
3. Специальность наиболее востребована в банках и в Ozon.
4. Предпочтительный режим работы: полный день или удаленная работа.
5. Зарплата в среднем варьируется: от 138697 руб. до 179783 руб.