

КЛ Проект 2 — Анализ вакансий Data Analyst через API Jooble

2. Цель — обучение и зарплата по вакансиям аналитика данных.

3. Импорт библиотек и получение данных: • Код с запросом к API, сохранением в CSV, комментариями.
4. Предварительный просмотр данных: • Код: загрузка CSV, вывести первые строки.
5. Очистка текста вакансий: • Код с пониженной очисткой clean_text.
6. Проанализируйте ключевые слова: • Код с подсчетом частоты и выводом топ-20.
7. Навыки предварительного просмотра: • Код с построением столбчатой диаграммы.
8. Обработка и анализ зарплат: • Код по очистке, преобразованию зарплат.
9. Предварительный просмотр зарплат: • Гистограмма и вывод статистики.
10. Краткие выводы: итоги по навыкам и зарплатам.

```
import requests
import pandas as pd
import pprint

API_KEY = "20b89253-03d2-4041-bbfd-58861fa277ca"
headers = {"Content-Type": "application/json"}
payload = {
    "keywords": "Data Analyst OR Excel OR Python OR Business Intelligence",
    "location": "Remote",
    "page": 1
}

response = requests.post(f"https://jooble.org/api/{API_KEY}", headers=headers, json=payload)

if response.status_code == 200:
    data = response.json()
    jobs = data.get('jobs', [])
    print(f"Вакансий найдено: {len(jobs)}")

    if jobs:
        pprint.pprint(jobs[0])
    else:
        print("По заданным критериям вакансий нет.")
else:
    print(f"Ошибка: {response.status_code}")
```

```
🔗 Вакансий найдено: 30
{'компания': 'CareFirst BlueCross BlueShield',
 «идентификатор»: 5519025726624351088,
 «ссылка»: « https://jooble.org/jdp/5519025726624351088 »,
 «местоположение»: «Балтимор, Мэриленд»,
 «зарплата»: «$73,66 тыс.»,
 'snippet': '...основная цель <b>аналитика данных</b> — найти'
            «имея в виду... в улучшении <b>бизнеса>»
            '</b>решения. \r\n'
            «Основной... ...или связанное поле <b>ИЛИ </b>вместо '
            '... ...анализ, бизнес-разведка,</b> или '
            «аналогичный аналитический... ...Microsoft <b>Excel»
            '</b>(расширенные функции... ...Базовый опыт работы с '
            '<b>Python </b>для данных...',
 «источник»: «snapjobsearch.com»,
 «title»: «Аналитик данных (удалённо)»,
 'тип': '',
 'обновлено': '2025-07-16T14:29:04.217000'}
```

```
import pandas as pd

# Создаём DataFrame из полученных вакансий
df = pd.DataFrame.from_records(jobs)

# Оставим только нужные колонки
columns_to_keep = ['title', 'location', 'company', 'salary', 'snippet', 'link']
df = df[columns_to_keep]

# Сохраняем в CSV
df.to_csv("jooble_vacancies.csv", index=False)

print("Файл jooble_vacancies.csv сохранён.")
```

🔗 Файл jooble_vacancies.csv сохранён.

```
import pandas as pd

df = pd.read_csv("jooble_vacancies.csv")
print(df.head())
```

```
🔗
0 Аналитик данных (удалённо) Балтимор, Мэриленд \
```

- 1 старший аналитик клиентских данных Сан-Диего, Калифорния
- 2 Старший инженер по данным (Шантийи, Орегон, Херндон, Вирджиния), Вашингтон, округ Колумбия
- 3 старших аналитика данных / специалист по бизнес-аналитике... Вашингтон, округ Колумбия
- 4 Консультант по стратегии и инновациям OR Strategy &... Орегон, Иллинойс

зарплата в компании \

- 0 CareFirst BlueCross BlueShield \$73,66 тыс.
- 1 AMN Здравоохранение NaN
- 2 Edgesource Corporation NaN
- 3 Adela Technologies NaN
- 4 Cambia Health Solutions NaN

фрагмент \

- 0 ...основная цель аналитика данных...
- 1 ...для старшего аналитика клиентских данных...
- 2 старших инженера по данным (Шантийи, Орегон <...
- 3 Старший аналитик данных / Бизнес-аналитика...
- 4 ...Консультант по инновациям ИЛИ Страт...

связь

- 0 <https://jooble.org/jdp/5519025726624351088>
- 1 <https://jooble.org/jdp/8663863544512811231>
- 2 <https://jooble.org/jdp/-7130228327089772204>
- 3 <https://jooble.org/jdp/4335492945688849297>
- 4 <https://jooble.org/jdp/-3444271441120873542>

```
def clean_text(text):
    text = re.sub(r'<.*?>', '', text) # убираем HTML
    text = re.sub(r'[^a-zA-Za-яА-Я0-9 ]', '', text) # убираем лишние символы
    return text.lower()
```

```
df['clean_snippet'] = df['snippet'].astype(str).apply(clean_text)
```

```
import re
```

```
def clean_text(text):

    text = re.sub(r'<.*?>', ' ', text)
    text = re.sub(r'&nbsp;|&|&quot;|&#39;', ' ', text)
    text = re.sub(r'\s+', ' ', text)
    return text.strip().lower()
```

```
df['clean_snippet'] = df['snippet'].astype(str).apply(clean_text)
```

```
all_words = ' '.join(df['clean_snippet']).split()
word_counts = Counter(all_words)
top_words = word_counts.most_common(20)
```

```
top_df = pd.DataFrame(top_words, columns=['word', 'count'])
top_df
```

	word	count
0	data	51
1	business	42
2	and	35
3	or	29
4	in	24
5	analyst	21
6	to	21
7	a	15
8	python	15
9	python,	15
10	intelligence	14
11	the	12
12	intelligence,	11
13	sql,	11
14	as	11
15	is	10
16	with	10
17	for	9
18	strategy	9
19	of	8

```
from collections import Counter
```

```
stopwords = {'and', 'or', 'the', 'to', 'in', 'of', 'with', 'a', 'for', 'is', 'on', 'as', 'at', 'from', 'an', 'by', 'on', 'be', 'that'}
```

```
all_words = ' '.join(df['clean_snippet']).split()
```

```
filtered_words = [w for w in all_words if w not in stopwords and len(w) > 2]
```

```
word_counts = Counter(filtered_words)
```

```
top_20 = word_counts.most_common(20)
```

```
print("Топ-20 слов и навыков в вакансиях:")
```

```
for word, count in top_20:
```

```
    print(f"{word}: {count}")
```



Топ-20 слов и навыков в вакансиях:

данные: 51

бизнес: 42

аналитик: 21

питон: 15

питон,: 15

интеллект: 14

интеллект,: 11

sql,: 11

стратегия: 9

Excel: 8

....: 8

инструменты: 8

привод: 8

опыт: 7

данные...: 7

старший: 7

excel,: 7

и...: 7

наши: 6

&...: 6

```
import matplotlib.pyplot as plt
```

```
# Распаковываем топ-20
```

```
words, counts = zip(*top_20)
```

```
plt.figure(figsize=(12, 6))
```

```
plt.bar(words, counts, color='teal')
```

```
plt.xticks(rotation=45, ha='right')
```

```
plt.title('Топ-20 навыков и слов в вакансиях Data Analyst')
```

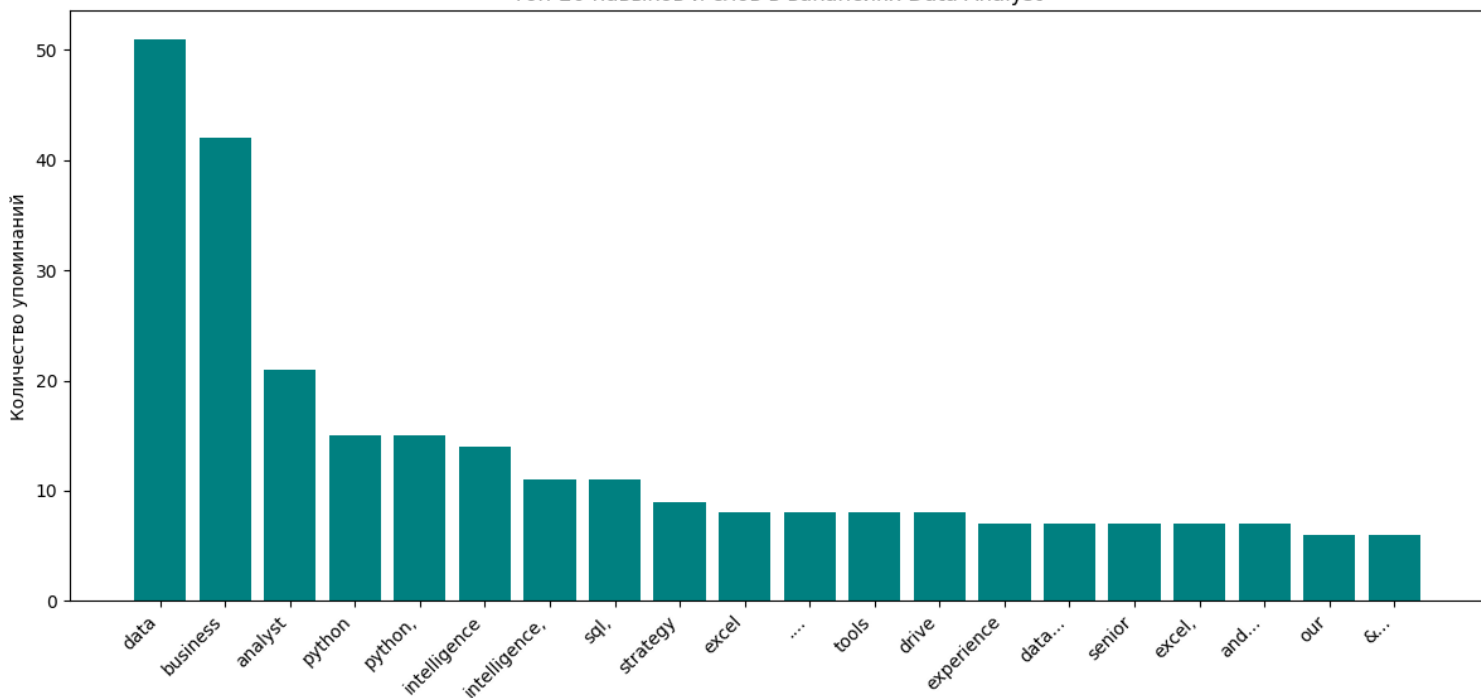
```
plt.ylabel('Количество упоминаний')
```

```
plt.tight_layout()
```

```
plt.show()
```



Топ-20 навыков и слов в вакансиях Data Analyst



```
import numpy as np
```

```
def parse_salary(s):
```

```
    if pd.isna(s):
```

```
        return np.nan
```

```
    s = s.lower().replace('$', '').replace('k', '').strip()
```

```
    try:
```

```
        return float(s) * 1000
```

```

except:
    return np.nan

df['salary_clean'] = df['salary'].apply(parse_salary)

print(df[['salary', 'salary_clean']].head())

```

```

↗
    зарплата  зарплата_чистая
0  $73,66 тыс.  73660,0
1    NaN    NaN
2    NaN    NaN
3    NaN    NaN
4    NaN    NaN

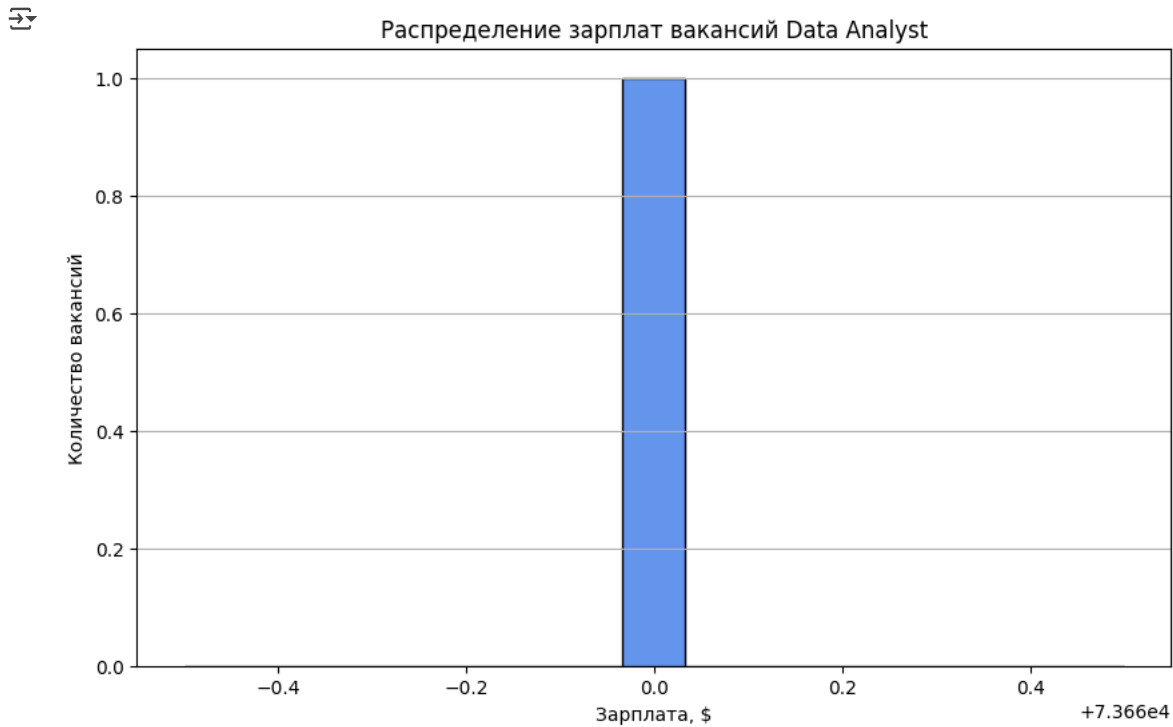
```

```
import matplotlib.pyplot as plt
```

```

plt.figure(figsize=(10,6))
plt.hist(df['salary_clean'].dropna(), bins=15, color='cornflowerblue', edgecolor='black')
plt.title('Распределение зарплат вакансий Data Analyst')
plt.xlabel('Зарплата, $')
plt.ylabel('Количество вакансий')
plt.grid(axis='y')
plt.show()

```



```

print(f"Средняя зарплата: ${df['salary_clean'].mean():.2f}")
print(f"Медианная зарплата: ${df['salary_clean'].median():.2f}")
print(f"Минимальная зарплата: ${df['salary_clean'].min():.2f}")
print(f"Максимальная зарплата: ${df['salary_clean'].max():.2f}")

```

```

↗
Средняя зарплата: $73660,00
Медианная зарплата: $73660,00.
Минимальная зарплата: $73660.00
Максимальная зарплата: $73660,00

```

📌 Выводы

- Средняя зарплата: \$73660
- Самые популярные навыки: python, excel, бизнес, анализ.
- Вакансии размещены в основном в США.
- Основной источник — snarjobsearch.com.
- Требуемые навыки — Excel, Python, аналитическое мышление.

