

Тестовое задание от компании Адвентум

В рамках анализа влияния различных факторов на количество лайков на постах в паблике с рисунками: <https://vk.com/iiiniamiii> был использован парсер для сбора данных со страницы паблика (*main.py*, *config.json*). Результатом работы парсера стала таблица *posts_data.csv*, содержащая информацию о каждом посте, включая идентификатор поста в группе (*id_*), дату и время публикации (*date_*), а также количество лайков на посте (*likes_*).

Затем данные были импортированы из *posts_data.csv* в предварительно созданную в *pgAdmin 4* таблицу *posts_data* с помощью специальной команды.

Далее были написаны запросы, позволяющие понять, что больше всего влияет на количество лайков - время суток публикации, день недели или промежуток между постами.

В первом запросе определялось время суток (ночь, утро, день, вечер) для каждой записи в таблице *posts_data*, а после вычислялось среднее количество лайков для каждой группы времени суток.

```
SELECT
CASE
    WHEN date_::time >= '00:00:00' AND date_::time < '06:00:00' THEN 'ночь'
    WHEN date_::time >= '06:00:00' AND date_::time < '12:00:00' THEN 'утро'
    WHEN date_::time >= '12:00:00' AND date_::time < '18:00:00' THEN 'день'
    ELSE 'вечер'
END AS time_of_day,
ROUND(AVG(likes_), 2) AS avg_likes_amount
FROM posts_data
GROUP BY time_of_day;
```

Рис. 1: Выявление зависимости числа лайков от времени суток публикации



	time_of_day 	avg_likes_amount 
	text	numeric
1	день	46.03
2	вечер	51.31
3	ночь	103.25
4	утро	28.52

Рис. 2: Результаты выполнения 1 запроса

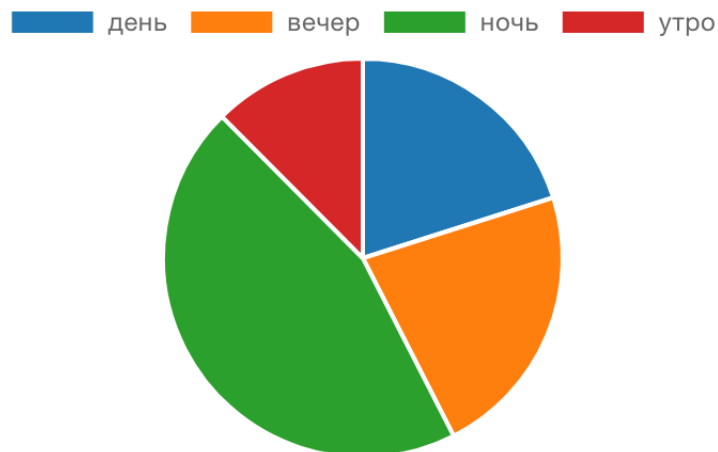


Рис. 3: Pie chart к 1 запросу

Во втором запросе определялся день недели (Mon, Tue, Wed, Thu, Fri, Sat, Sun) для каждой записи в таблице *posts_data*, затем вычислялось среднее количество лайков для каждой группы дней недели.

```
SELECT
CASE
WHEN date_part('isodow', date_) = 1 THEN 'Mon'
WHEN date_part('isodow', date_) = 2 THEN 'Tue'
WHEN date_part('isodow', date_) = 3 THEN 'Wed'
WHEN date_part('isodow', date_) = 4 THEN 'Thu'
WHEN date_part('isodow', date_) = 5 THEN 'Fri'
WHEN date_part('isodow', date_) = 6 THEN 'Sat'
ELSE 'Sun'
END AS day_of_week,
ROUND(AVG(likes_), 2) AS avg_likes_amount
FROM posts_data
GROUP BY date_part('isodow', date_)
ORDER BY date_part('isodow', date_);
```

Рис. 4: Выявление зависимости числа лайков от дня недели публикации

	day_of_week text	avg_likes_amount numeric
1	Mon	38.67
2	Tue	54.19
3	Wed	40.00
4	Thu	48.18
5	Fri	36.69
6	Sat	44.27
7	Sun	45.85

Рис. 5: Результаты выполнения 2 запроса

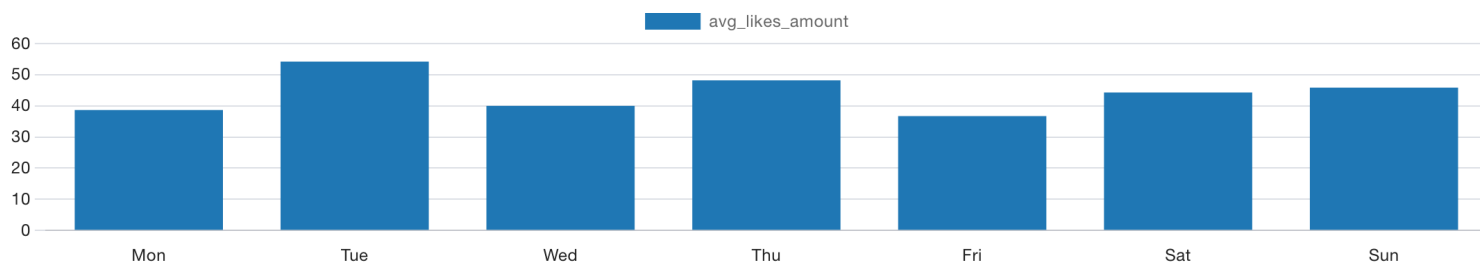


Рис. 6: Bar chart ко 2 запросу

В третьем запросе создавалась временная таблица (*table1*) с дополнительными столбцами:

- *prev_date* - дата публикации предыдущего поста;
- *date_diff* - разница между предыдущей и текущей датами;
- *prev_likes* - количество лайков на предыдущем посте;
- *likes_avg* - среднее количество лайков между текущим и предыдущим значениями.

Затем из этой временной таблицы выбирались данные, исключая запись с $id_ = \max(id_)$, поскольку эта строка была неинформативной (самый крайний пост), после вычислялось среднее количество лайков (*avg_likes_amount*) для каждой группы разницы дат (*date_diff*).

```
SELECT
    date_diff,
    ROUND(AVG(likes_avg), 2) AS avg_likes_amount
FROM
    (SELECT
        id_,
        date_::date,
        COALESCE(LAG(date_::date) OVER (ORDER BY date_::date DESC), date_::date) AS prev_date,
        COALESCE(LAG(date_::date) OVER (ORDER BY date_::date DESC), date_::date) - date_ AS date_diff,
        likes_,
        COALESCE(LAG(likes_) OVER (ORDER BY date_::date DESC), likes_) AS prev_likes,
        ROUND((likes_ + COALESCE(LAG(likes_) OVER (ORDER BY date_::date DESC), likes_)) / 2) AS likes_avg
        FROM posts_data
        ORDER BY date_ DESC) AS table1
WHERE id_ != (SELECT MAX(id_) FROM posts_data)
GROUP BY date_diff
ORDER BY date_diff;
```

Рис. 7: Выявление зависимости числа лайков от промежутка между постами

	date_diff integer	avg_likes_amount numeric
1	0	74.00
2	1	40.74
3	2	42.42
4	3	30.47
5	4	42.54
6	5	41.22
7	6	97.17
8	7	53.67
9	8	38.00
10	10	75.00
11	11	109.00
12	14	35.00
13	15	82.00
14	18	38.50
15	19	55.50
16	21	22.50
17	26	107.50

Рис. 8: Результаты выполнения 3 запроса

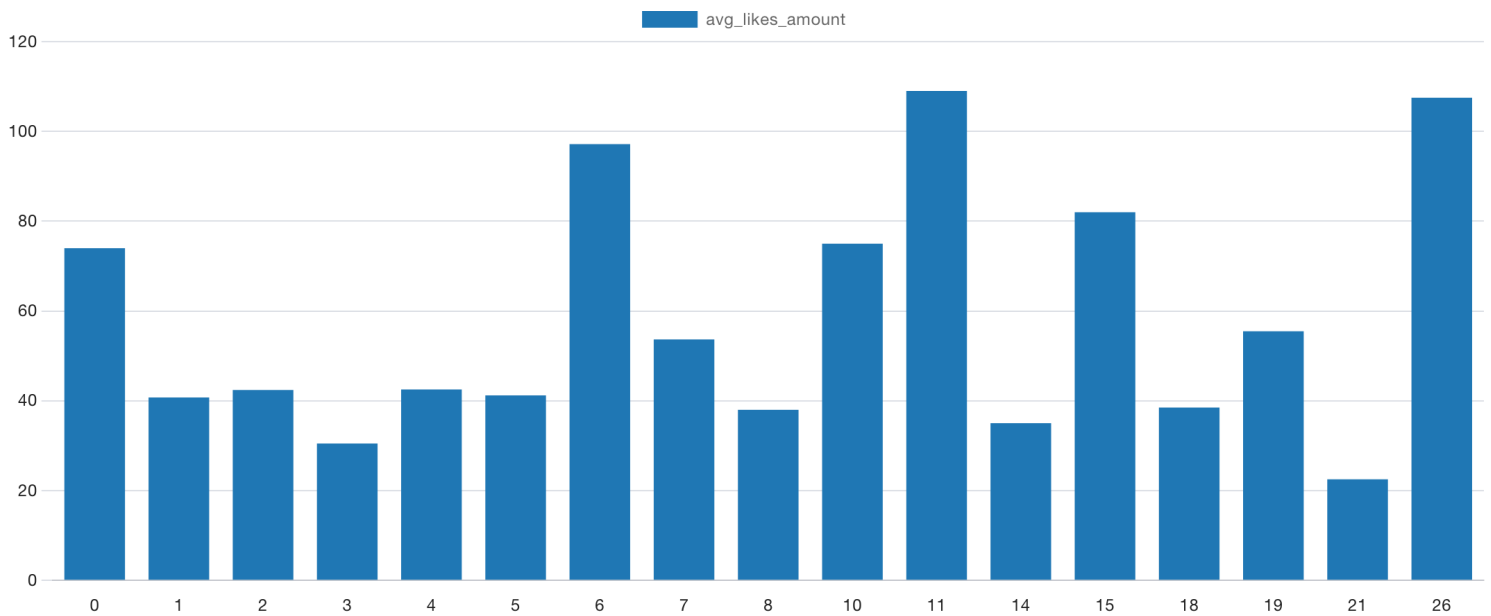


Рис. 9: Bar chart к 3 запросу

Приведенные графики позволяют сделать вывод о том, что наиболее явно прослеживается зависимость среднего числа лайков от времени суток.

Далее был выполнен расчет различных статистических величин (*metrics.py*). Результаты запросов, представленные на Рис. 2, 5, 8, были загружены в csv-файлы,

затем три набора данных из файлов *data1.csv*, *data2.csv* и *data3.csv* были открыты с помощью функции *pd.read_csv*. Каждый набор данных хранился в отдельном датафрейме (*df_time*, *df_day* и *df_interval*).

В каждом датафрейме было две колонки: *time_of_day* и *avg_likes_amount* в *df_time*, *day_of_week* и *avg_likes_amount* в *df_day*, и *date_diff* и *avg_likes_amount* в *df_interval*; они содержали категориальные (время суток, день недели) и количественные (число дней между постами, число лайков) данные.

Чтобы работать с категориальными данными, создавались словари (*time_of_day_mapping* и *day_of_week_mapping*), которые сопоставляют каждое категориальное значение с числом. Затем использовалась функция *map* для замены категориальных значений на соответствующие числа в датафреймах. После этого рассчитывался коэффициент корреляции Пирсона между количественными данными (*avg_likes_amount*) и числовыми значениями категориальных данных (*time_of_day_num* и *day_of_week_num*).

После этого проводился тест Краскела-Уоллиса, который проверяет, есть ли статистически значимая разница в средних значениях количественных данных (*avg_likes_amount*) для разных категориальных значений (*time_of_day* и *day_of_week*). Тест Краскела-Уоллиса возвращал статистику *H* и *p-value*.

Были получены следующие результаты:

metric	time_of_day	day_of_week	time_interval
correlation coefficient	0.9212	-0.0204	0.2029
p-value	0.3916	0.4232	-

Таблица 1: Основные метрики

Как видно из Таблицы 1, результаты анализа противоречивы. С одной стороны, коэффициент корреляции Пирсона указывает на сильную положительную корреляцию между временем суток и количеством лайков, а с другой стороны, тест Краскела-Уоллиса не находит статистически значимых различий в количестве лайков по разным временам суток и дням недели (*p-value* больше 0,05). Можно предположить, что одной из причин такой нестабильности является малый размер выборки.

В целом, данные показали, что время суток оказывает большее влияние на количество лайков по сравнению с другими факторами. Пользователи социальных сетей наиболее активны вечером и ночью, когда у них есть больше свободного времени, отсюда и следует наибольшая степень влияния.