Задание

Взять страницу ВК, собрать по ней таблицу с датой постов и количеством лайков и написать SQL-запросы, которые позволят ответить на вопрос: что больше всего влияет на количество лайков: время суток публикации, день недели или промежуток между постами.

Шаги

Для написания запросов использовал MySQL. Для визуализации использовал Apache Superset и Python.

Написал скрипт на Python и собрал данные о последних постах со страницы группы Павла Воли. В базу данных включил следующие поля:

ИМЯ ПОЛЯ	ПОЯСНЕНИЕ
post_id	Номер поста (с конца)
post_date	Дата и время поста
attachment_type	Тип прилагаемых файлов (фото, видео, репост, другое)
text_size	Размер текста в посте
like_count	Количество лайков
comment_count	Количество комментариев
repost_count	Количество репостов
view_count	Количество просмотров
duration	Длительность (если приложено видео, в противном случае – пустое значение)

Помимо дня недели, времени суток и интервалом между постами, хотел посмотреть на влияние других факторов на количество лайков (вид поста, размер описания, количество комментариев и т.д.)

В таблицу включил данные о последних 550 постах со страницы группы.

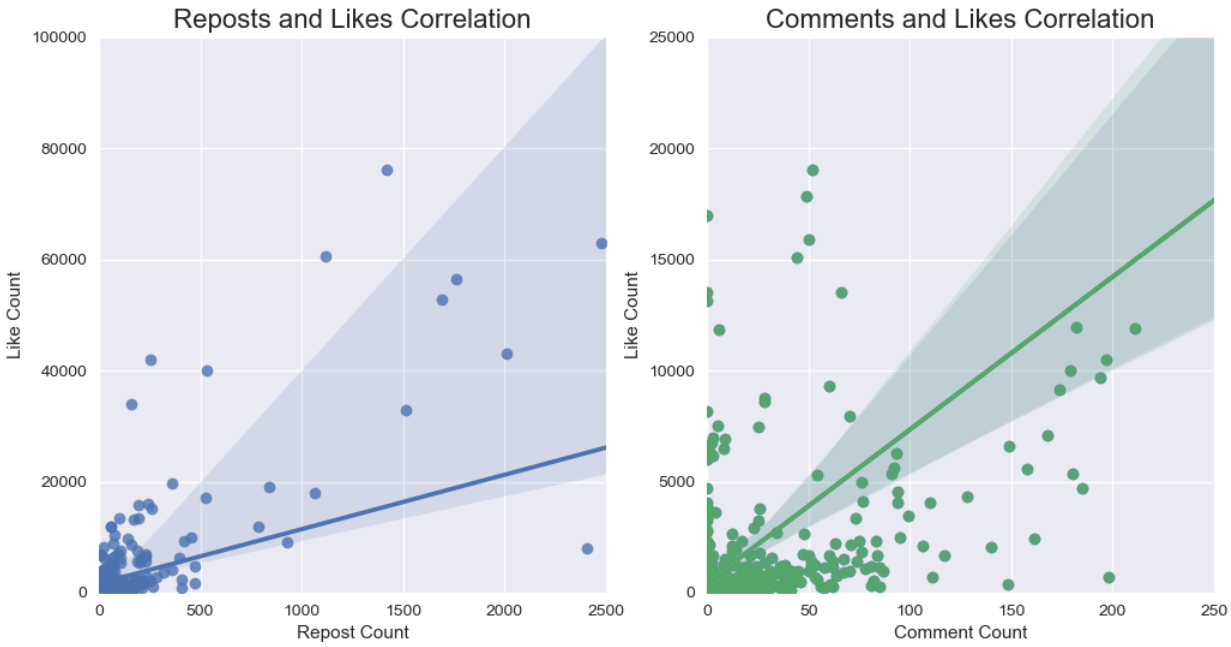
Используя библиотеку Pandas в Python, я рассчитал коэффициенты корреляции для количества лайков и следующих полей: количество репостов, комментариев, просмотров, размер текста и длительность видео. Результаты свел в таблицу:

ПОЛЕ	КОЛИЧЕСТВО ЛАЙКОВ
Количество репостов	0,705
Количество комментариев	0,659
Количество просмотров	0,404

Оказалось, что в наибольшей степени на количество лайков под постами влияют количество репостов, комментариев и просмотров (что, вообще говоря, является очевидным, поскольку с ростом

одного показателя пост попадается большему количеству пользователей, откуда растут и последующие показатели).

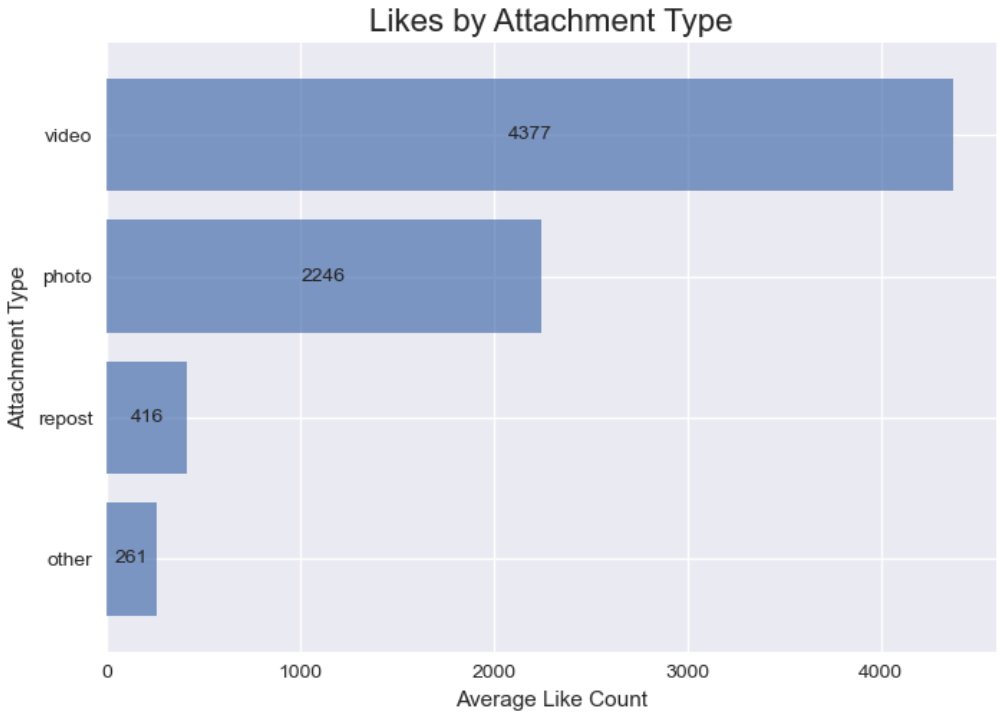
Более детально данные зависимости можно посмотреть на графиках ниже:



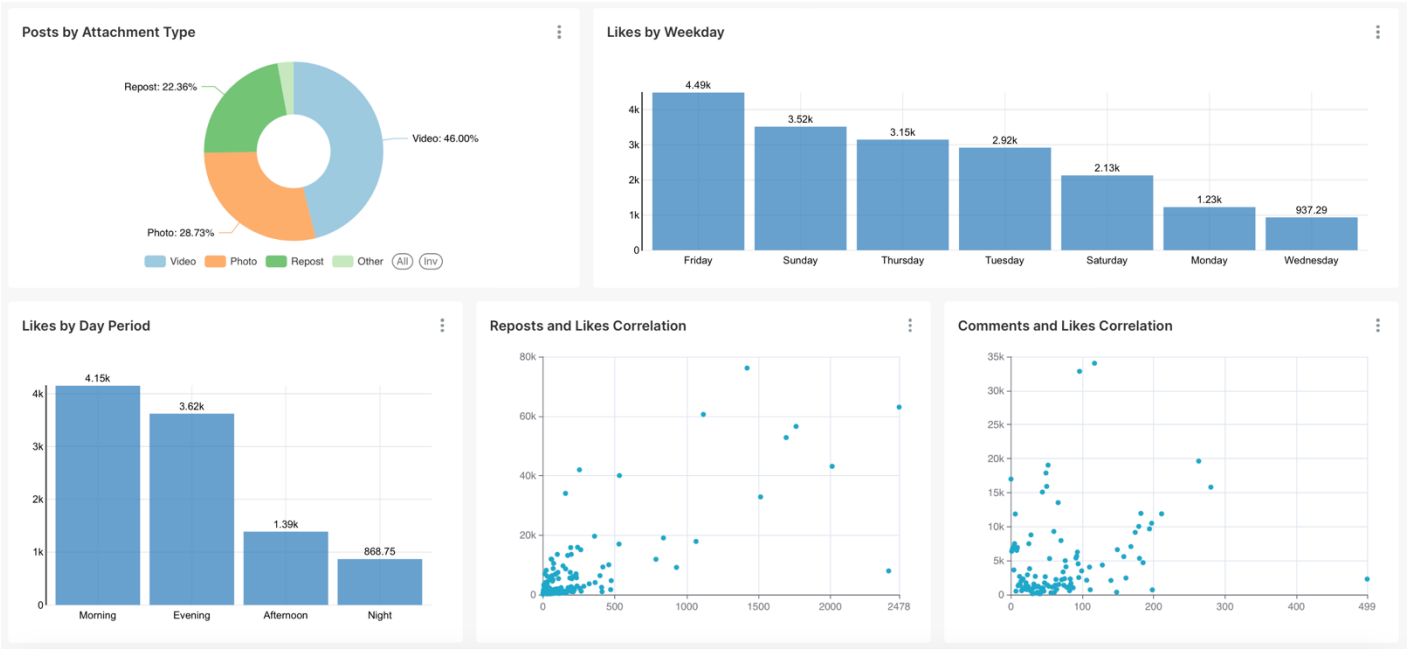
Слева – зависимость количества репостов от лайков, справа – количество комментариев от лайков.

Здесь более детально видна зависимость между двумя показателями. Она в особенности заметна на правом графике.

Также важно отметить, что в среднем количество лайков больше на постах с видео. Вероятнее всего, что данный показатель зависит от типа контента на определенной странице. В данном случае полученный результат имеет смысл, учитывая то, какой контент размещается на выбранной мною странице.

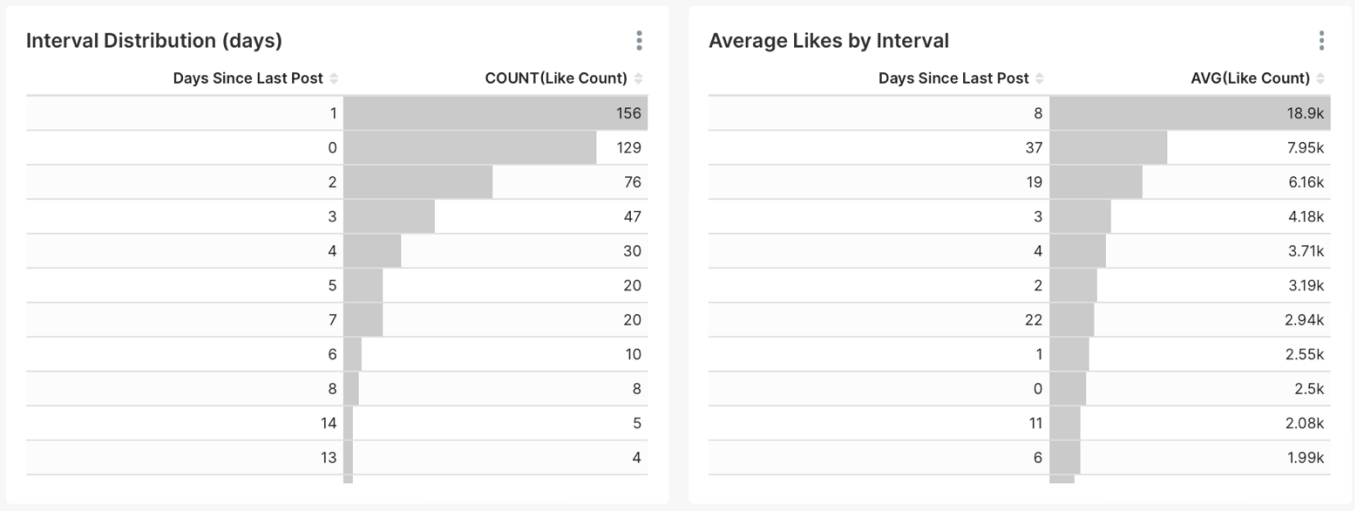


Среднее число лайков по типу постов



Для визуализации остальных показателей я построил дашборд в Superset. Из него можно сделать следующие выводы:

- Виды постов: видео (46%), фото (28,73%), репосты (22,36%), другие (2,91%)
- В среднем больше лайков набирают посты, выложенные в пятницу и воскресенье утром (с 6 до 11) и вечером (с 18 до 23)
- Также включены графики зависимости количества комментариев и репостов от лайков.



Interval Distribution (days)

Days Since Last Post	COUNT(Like Count)
1	156
0	129
2	76
3	47
4	30
5	20
7	20
6	10
8	8
14	5
13	4

Average Likes by Interval

Days Since Last Post	AVG(Like Count)
8	18.9k
37	7.95k
19	6.16k
3	4.18k
4	3.71k
2	3.19k
22	2.94k
1	2.55k
0	2.5k
11	2.08k
6	1.99k

Слева – распределение постов по интервалам, справа – среднее количество лайков под постом, опубликованный с определенным интервалом

При исследовании зависимости интервалов между постами и количеством лайков можно сделать следующие выводы:

- За все время большинство постов выкладывались с интервалом 0 (несколько постов в один и тот же день) – 4 дня
- Больше всего лайков набрали посты с интервалом в 8 дней.

Основываясь на полученных данных, сложно сказать о зависимости между частотой публикаций и количеством лайков. Тем более в данную выборку с интервалом 8 дней входят только 8 постов. Возможно, для более точного результата понадобилась бы выборка большего размера, однако вероятность того, что результат будет другим – мала.

ВЫВОД

Для анализа использовалась выборка с последними постами со страницы группы Павла Воли. В итоге в таблице собралось 550 записей.

На основе полученных данных можно сделать следующие выводы:

1. количество лайков напрямую зависит от количества репостов, комментариев и просмотров. С увеличением данных показателей растет популярность поста и, соответственно, количество лайков под ним;
2. большую активность пользователи проявляют в пятницу и воскресенье в утреннее и вечернее время;
3. интервал между постами, как оказалось, не играет большой роли в количестве лайков под постом.

SQL-ЗАПРОСЫ:

```sql
/*
  Основной SQL-запрос
  Использовал оператор CASE, чтобы из всех типов постов
  выбрать наиболее популярные (фото, видео, репост), вывел их
  с заглавной буквы
  Оставшиеся пометил как "другие"
*/

SELECT
  post_id,
  post_date,
  CASE
  WHEN attachment_type IN ('photo', 'video', 'repost') THEN
CONCAT(UPPER(SUBSTR(attachment_type, 1, 1)), SUBSTR(attachment_type, 2))
  ELSE 'Other'
  END AS 'attachment_type',
  text_size,
  like_count,
  comment_count,
  repost_count,
  view_count,
  duration
FROM posts;
```

```sql
/*
  Данный запрос выводит информацию
  о среднем количестве лайков в разное время суток
  Группируется по времени суток и сортируется по
  среднему количеству лайков в порядке убывания
*/

SELECT
  CASE
  WHEN HOUR(post_date) between 0 AND 5 THEN 'Night'
  WHEN HOUR(post_date) between 6 AND 11 THEN 'Morning'
  WHEN HOUR(post_date) between 12 AND 17 THEN 'Afternoon'
  WHEN HOUR(post_date) between 18 AND 23 THEN 'Evening'
  END AS 'post_hour',
  ROUND(AVG(like_count), 2) as avg_likes
FROM
  posts
GROUP BY
  post_hour
ORDER BY
  avg_likes DESC
```

```sql
/*
  Данный запрос выводит информацию
  о среднем количестве лайков в разные дни недели
  Группируется по дню недели и сортируется по
  среднему количеству лайков в порядке убывания
*/

SELECT
  DAYNAME(post_date) as day_of_week,
  ROUND(AVG(like_count), 2) as avg_likes
FROM
  posts
GROUP BY
  day_of_week
ORDER BY
  avg_likes DESC;
```

```sql
/*
  Данный запрос выводит информацию о постах и количестве
  дней, прошедших с момента публикации предыдущего поста.
  Чтобы это реализовать, я использовал CROSS JOIN. Я
  соединил таблицы таким образом, чтобы каждый пост в таблице
  "a" сопоставлялся с предыдущим по счету постом в таблице
  "b". Это условие я реализовал в WHERE (поскольку посты в
  таблице находятся в порядке убывания по дате,
  ID таблицы "a" должно быть на 1 меньше ID таблицы "b").
  В запрос включено количество постов для каждого интервала
  (post_count) и среднее количество лайков для каждого
  интервала (avg_likes)
*/

SELECT
    DATEDIFF(a.post_date, b.post_date) AS days_since_last_post,
    COUNT(a.like_count) AS post_count,
    AVG(a.like_count) as avg_likes
FROM
    posts a, posts b
WHERE
    a.post_id = b.post_id - 1
GROUP BY
    days_since_last_post;
```