**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](https://vk.com/pavelvolyaofficial). 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.)

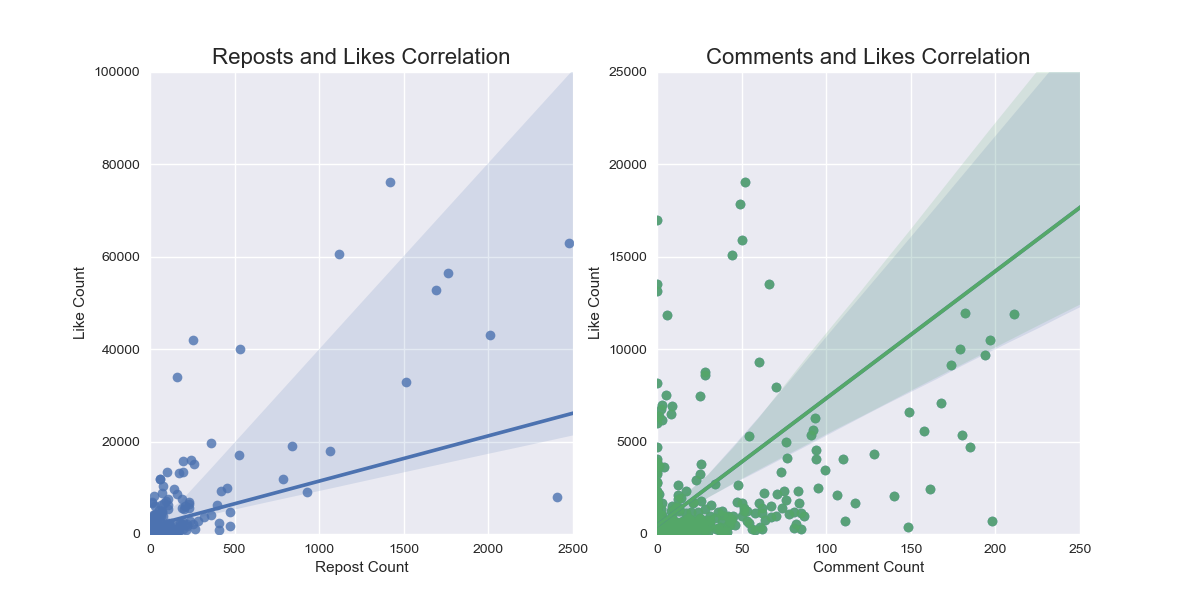
В таблицу включил данные о последних 550 постах со страницы группы.

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).

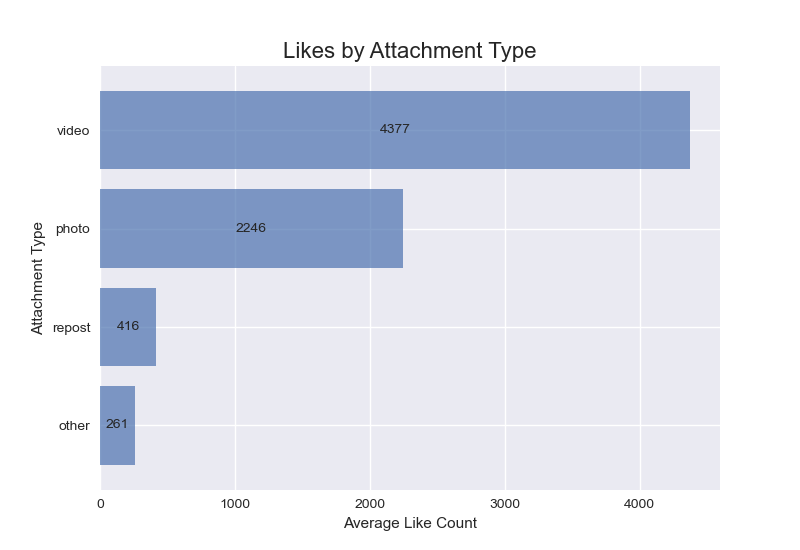
Более детально данные зависимости можно посмотреть на графиках ниже:



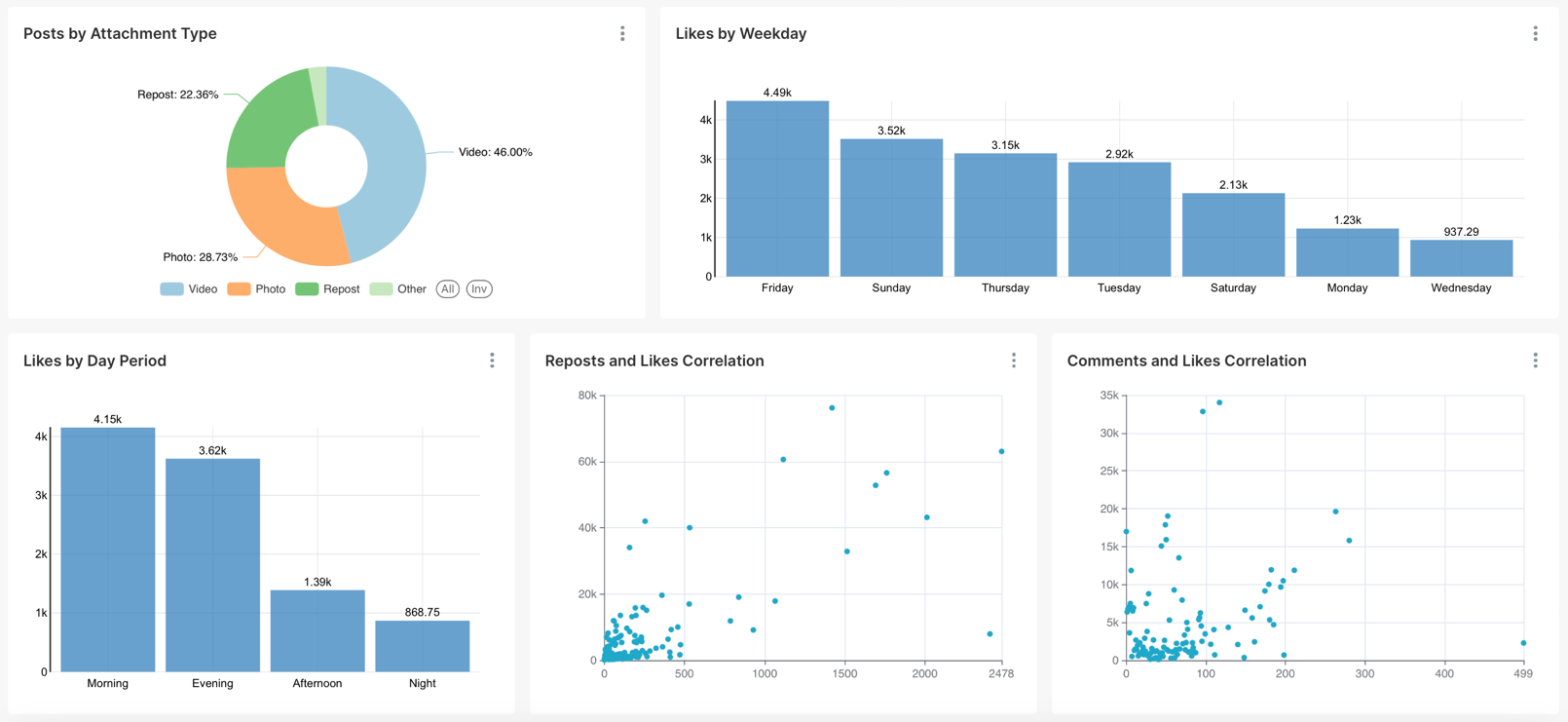
Слева – зависимость количества репостов от лайков, справа – количество комментариев от лайков.

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.

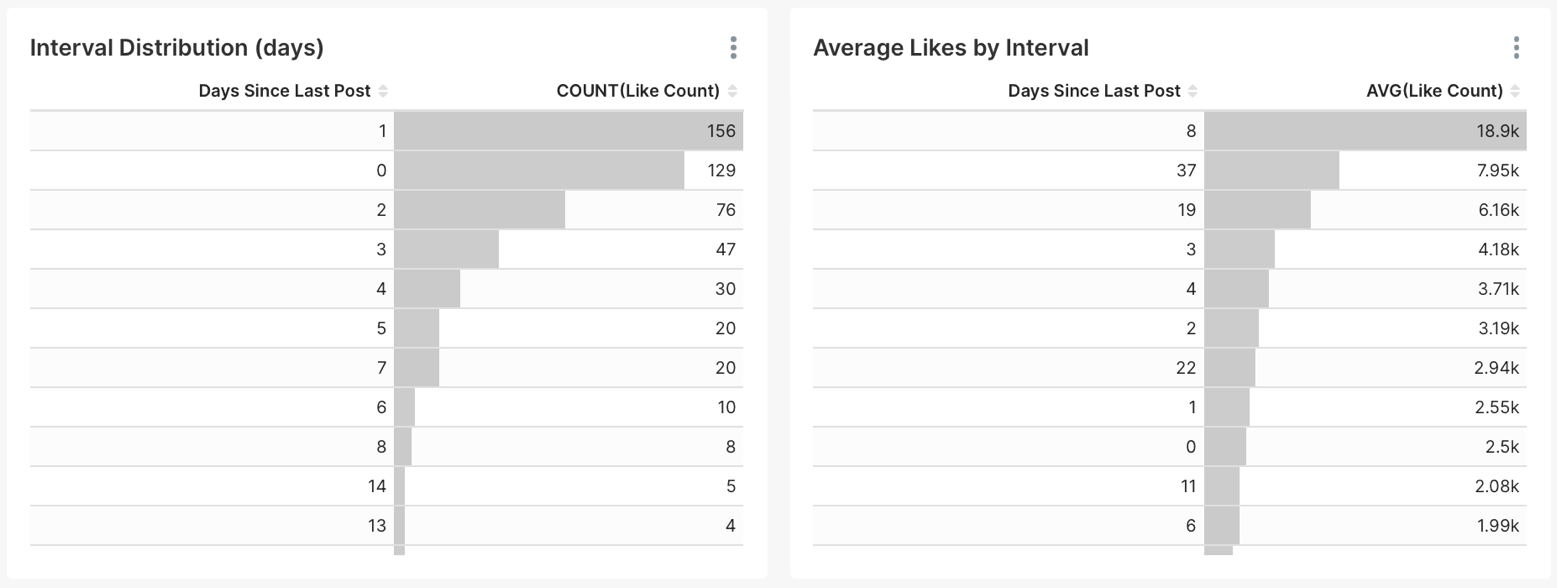


Среднее число лайков по типу постов



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.



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:**



