facebook

# Facebook Comments Volume Dataset

This dataset is composed of data about Facebook posts. It provides 53 variables, such as the Page Category or Popularity, or the number of comments a post has received short after its publication. The objective is to predict the number of comments the post will receive in the future.

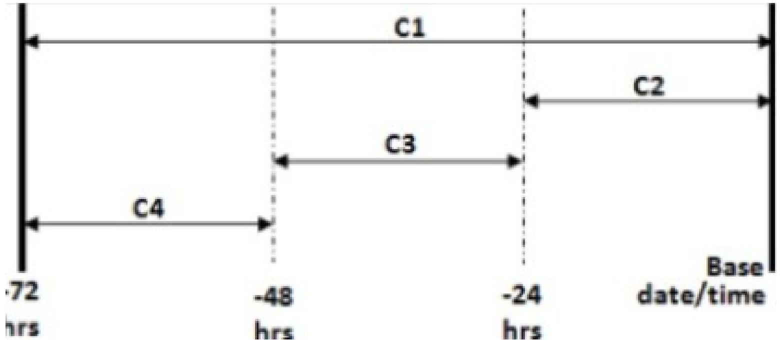
The different variables are divided in 5 categories, according to the responsible of the dataset:

- Page Features

Those variables do not describe the post in itself, but the page that posted it, such as the Page Category, or its total number of likes.

- Essential Features

The aim of this dataset is to predict the number of comments a post will receive in the future. The dataset contains the pattern of comments on the post in various time intervals (first 24 hours, first 24 to first 48h...). It also contains 'derived' features, where the previously discussed features are aggregated by page.



- Weekday Features

Two variables describe the weekday a post was published, and the weekday the web crawler that built the dataset got inspected the post.

- Other Variables

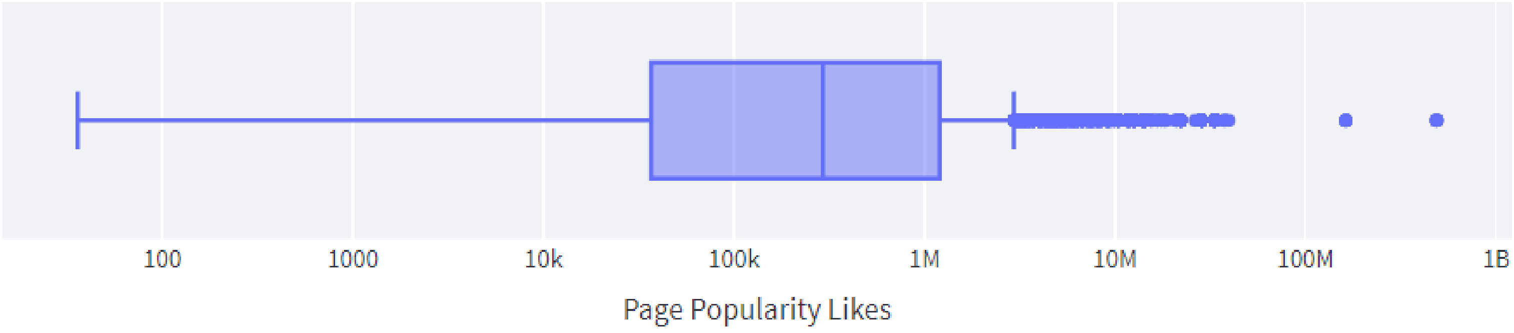
Some variables do not fit our typology, but may still be important. One variable contains the time between the post publication, and the time the crawler inspected it. We also have information about the post length, and the post share count.

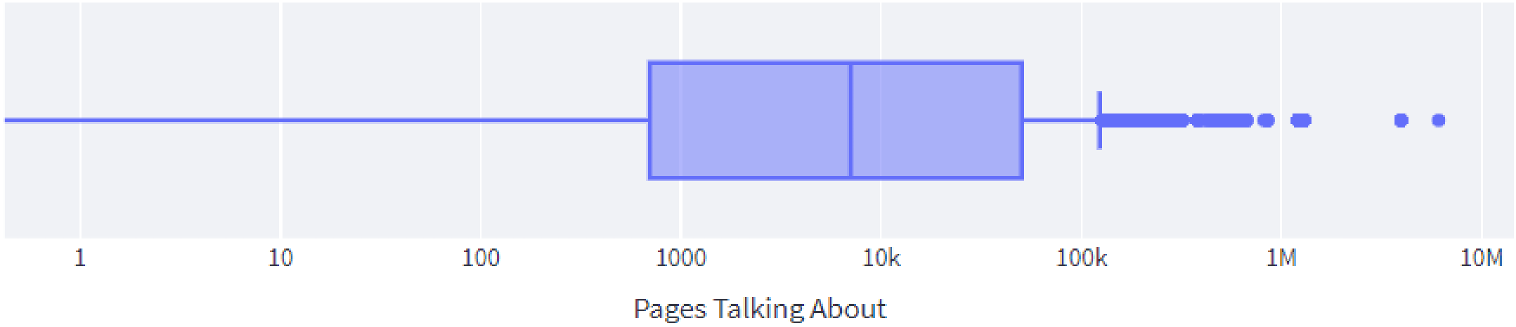
- Target Variable

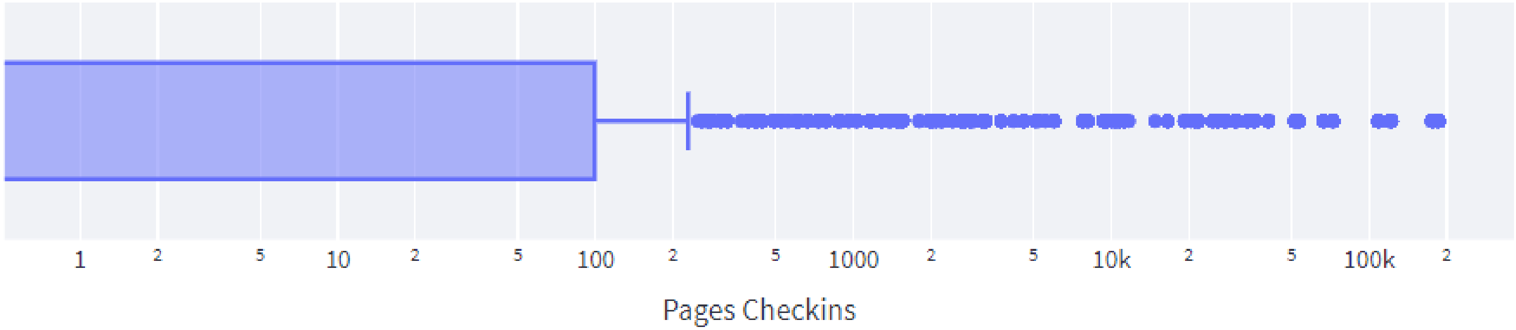
Finally, this dataset it built to predict the number of comments a post will get in the future. Our target variable stores the number of comments received after the crawler firt inspected it. The delay is also stored.

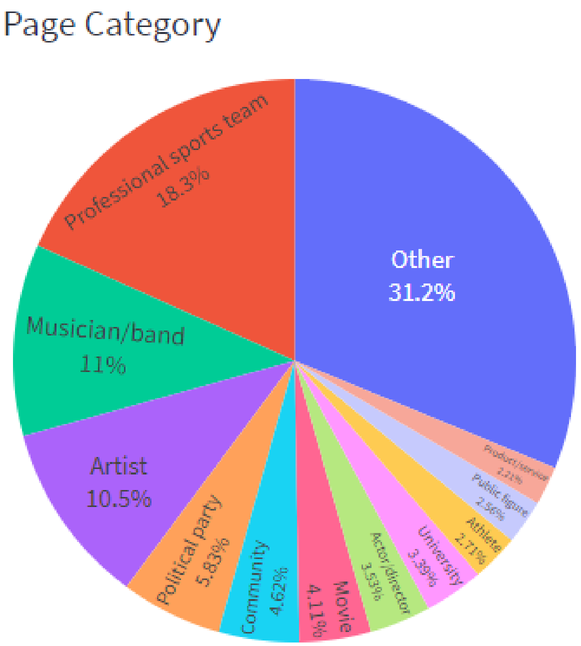
# Univariate Analysis

**Page Features** | Essential Features | Weekday Features | Other Features | Target Variable

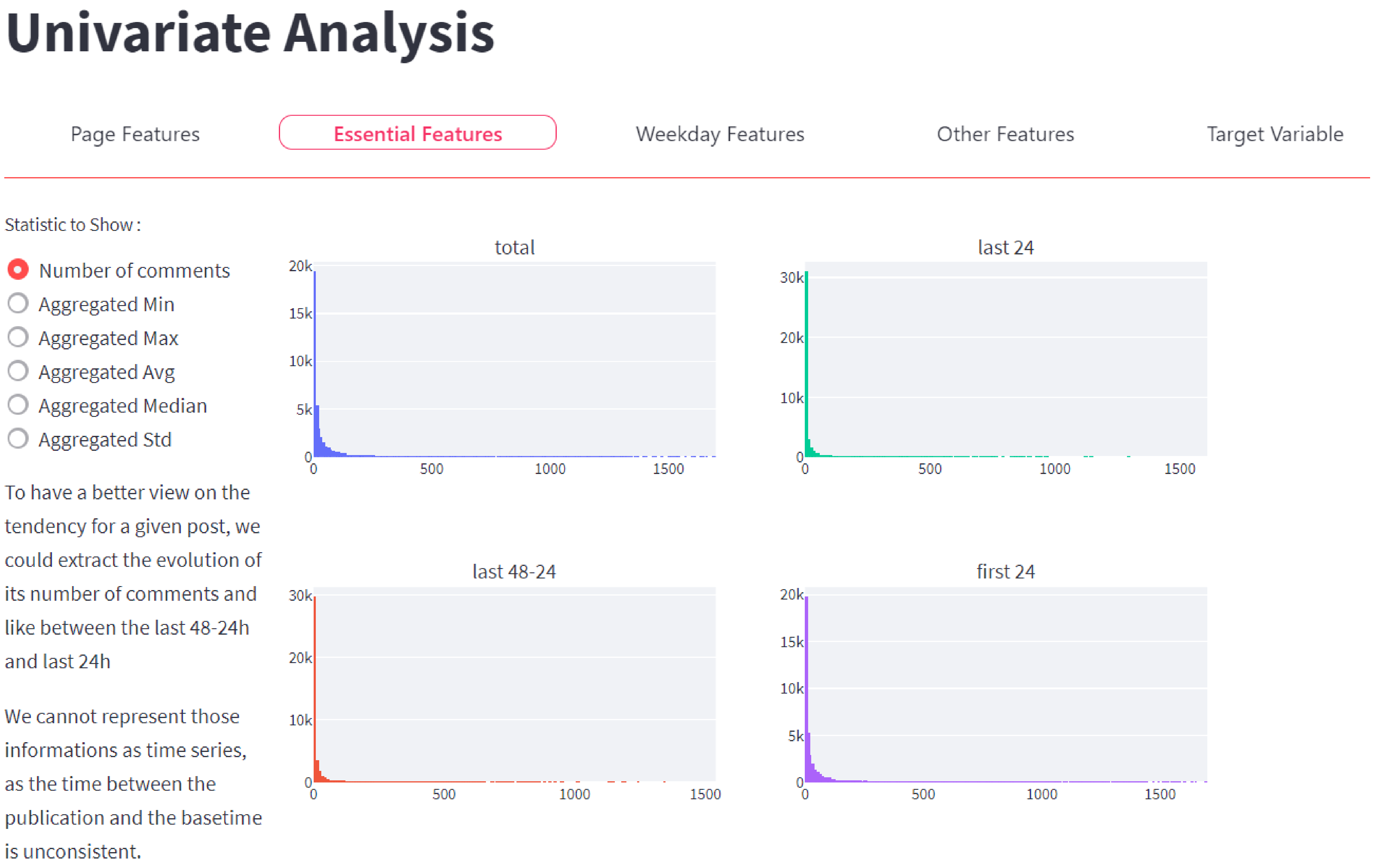








To have a better understanding of the consequences of the page category, we can group the categories that have a similar meaning.



# Univariate Analysis

Page Features | **Essential Features** | Weekday Features | Other Features | Target Variable

Statistic to Show :

● Number of comments
○ Aggregated Min
○ Aggregated Max
○ Aggregated Avg
○ Aggregated Median
○ Aggregated Std

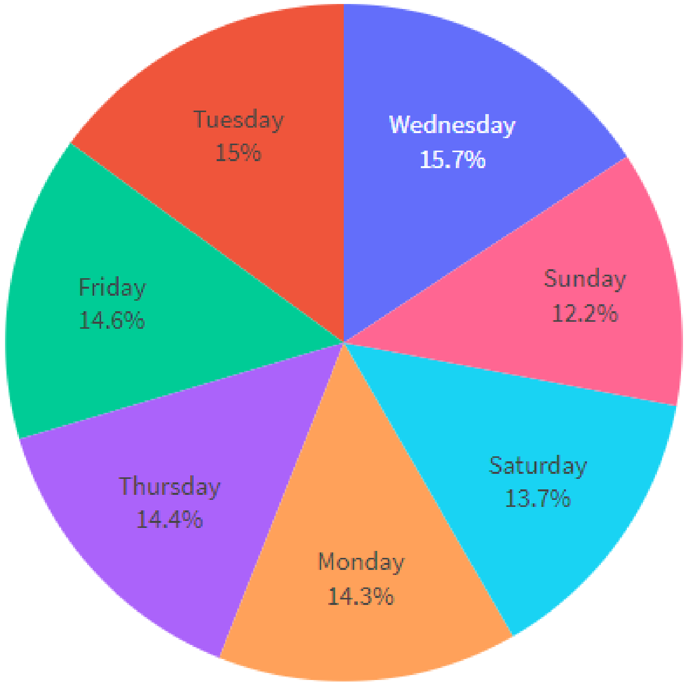
To have a better view on the tendency for a given post, we could extract the evolution of its number of comments and like between the last 48-24h and last 24h

We cannot represent those informations as time series, as the time between the publication and the basetime is unconsistent.

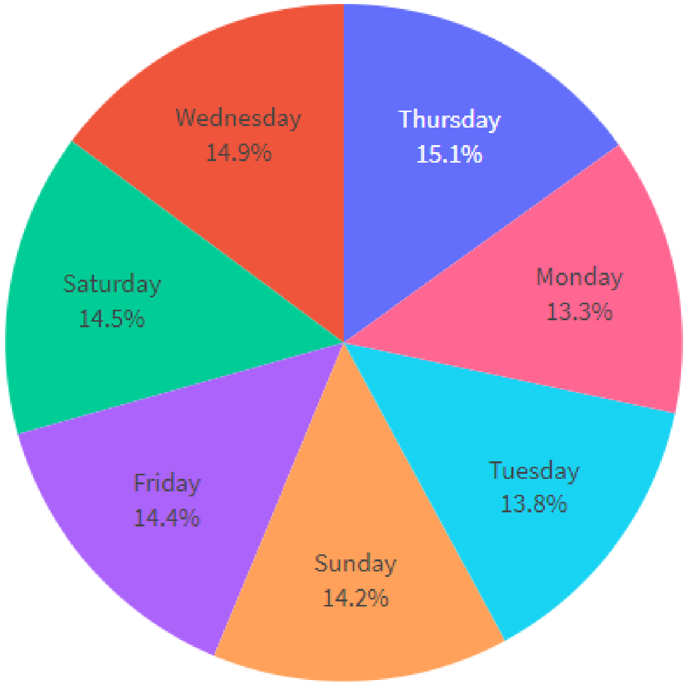
# Univariate Analysis

Page Features | Essential Features | **Weekday Features** | Other Features | Target Variable

Post Published Weekday



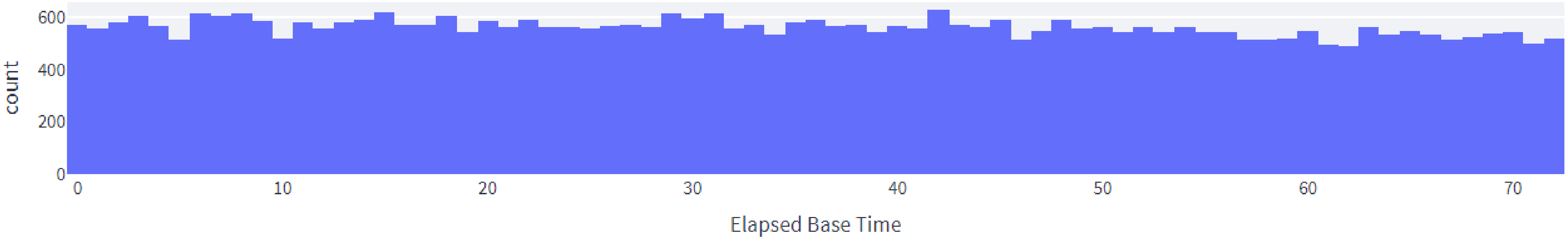
Base DateTime Weekday

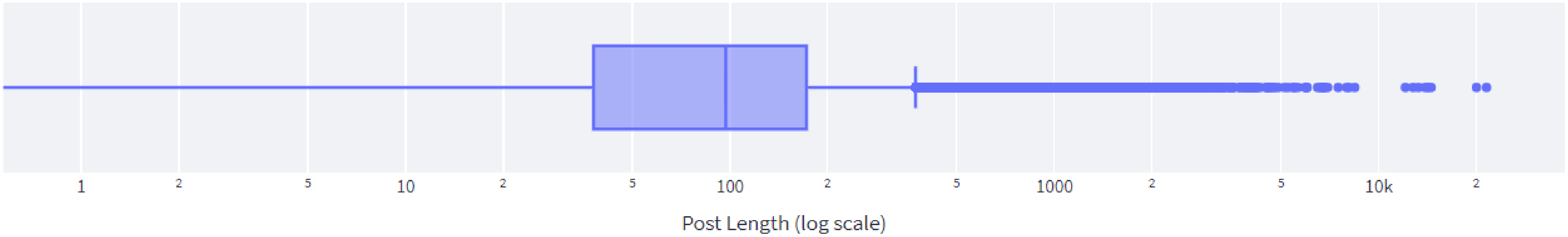


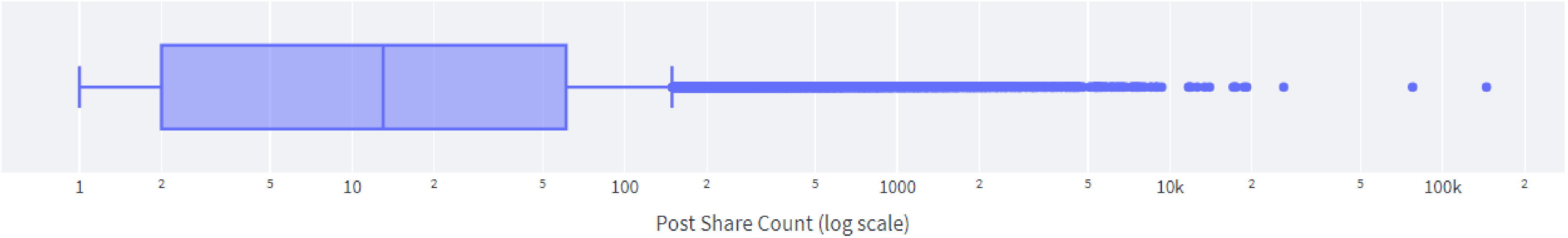
The weekday post were published, and the web crawler inspected them are almost evenly distributed

# Univariate Analysis

Page Features | Essential Features | Weekday Features | **Other Features** | Target Variable



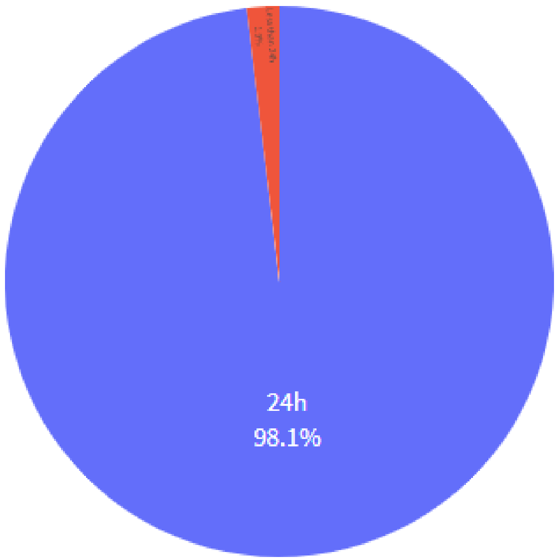




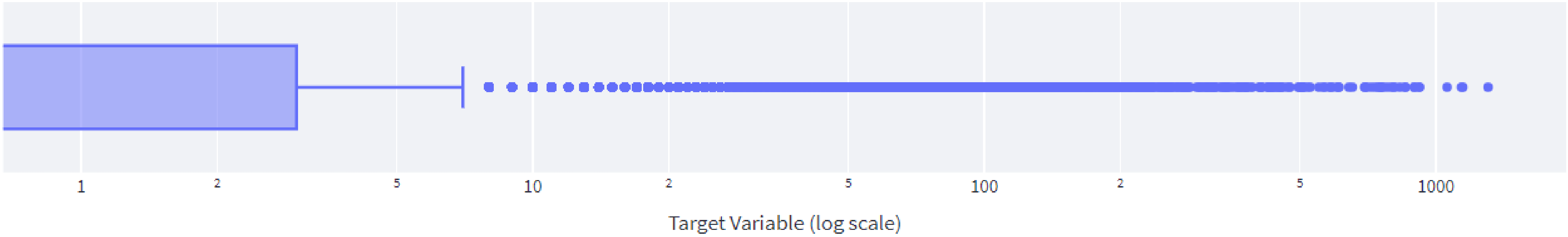
# Univariate Analysis

Page Features | Essential Features | Weekday Features | Other Features | Target Variable

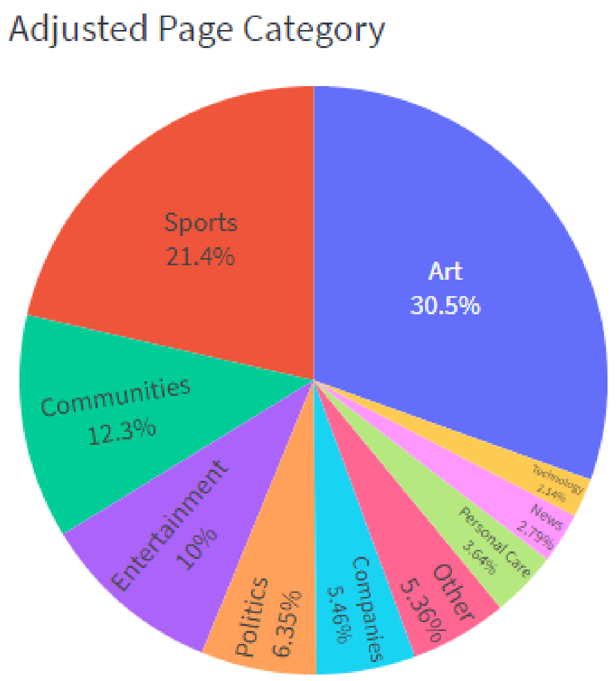
Target Measurement Time

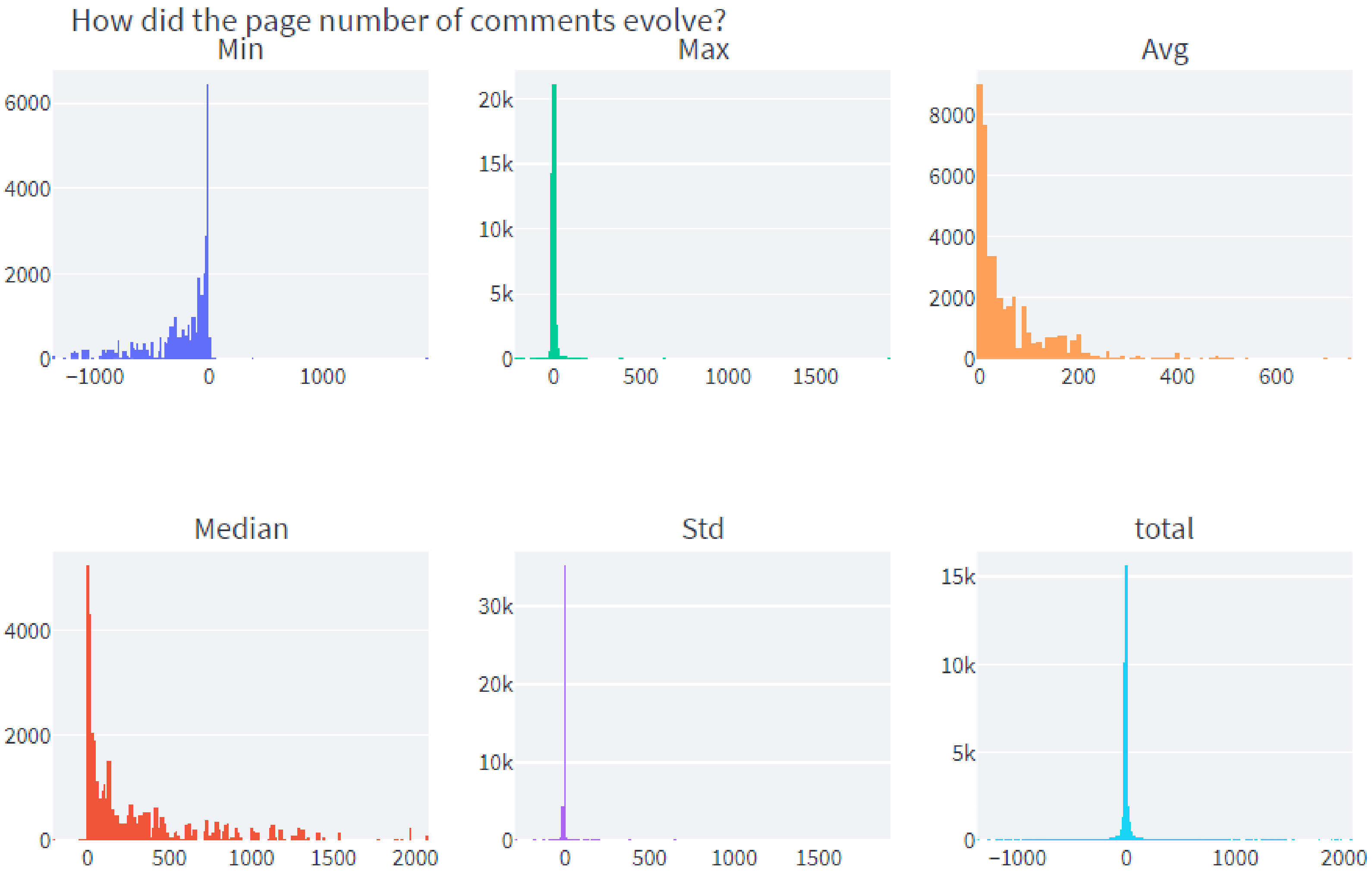


The vast majority of the Target Variable (number of comments) measurement occure 24h after the base time. Also, more than half of the posts of the dataset get no comments during this interval and 95% get less than 8, while very rare posts get +1000 comments.



# Created Variables



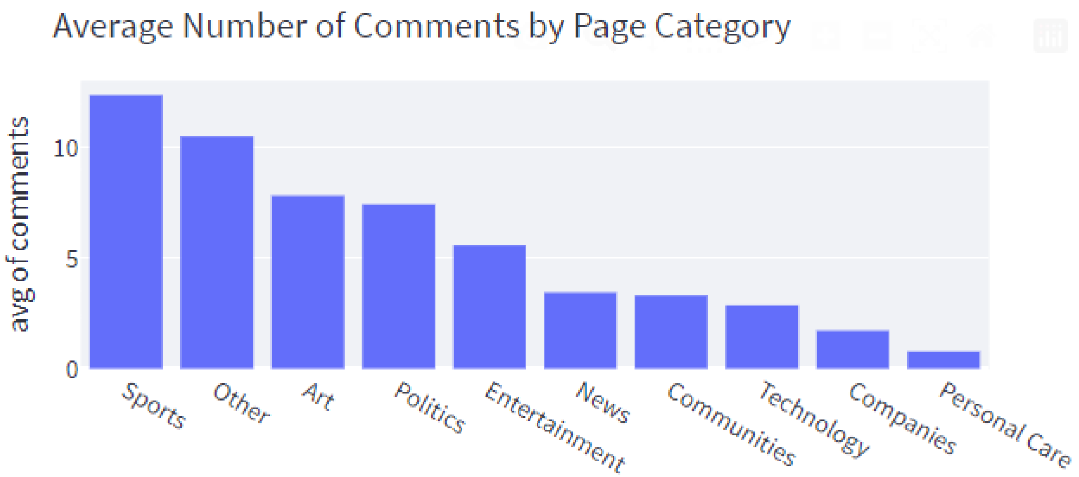


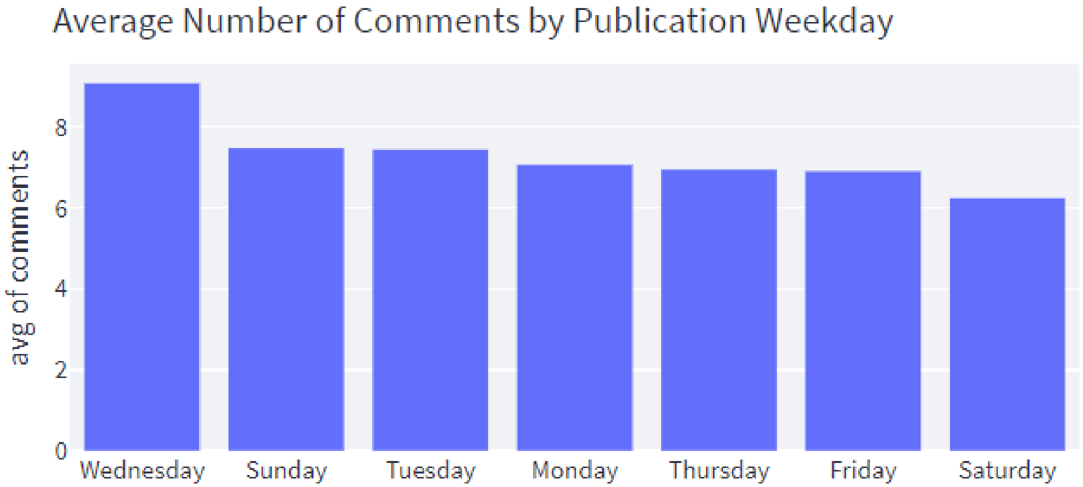
Merging very similar categories (like 'sports league' and 'sports team') makes 'Page Category' more readable.

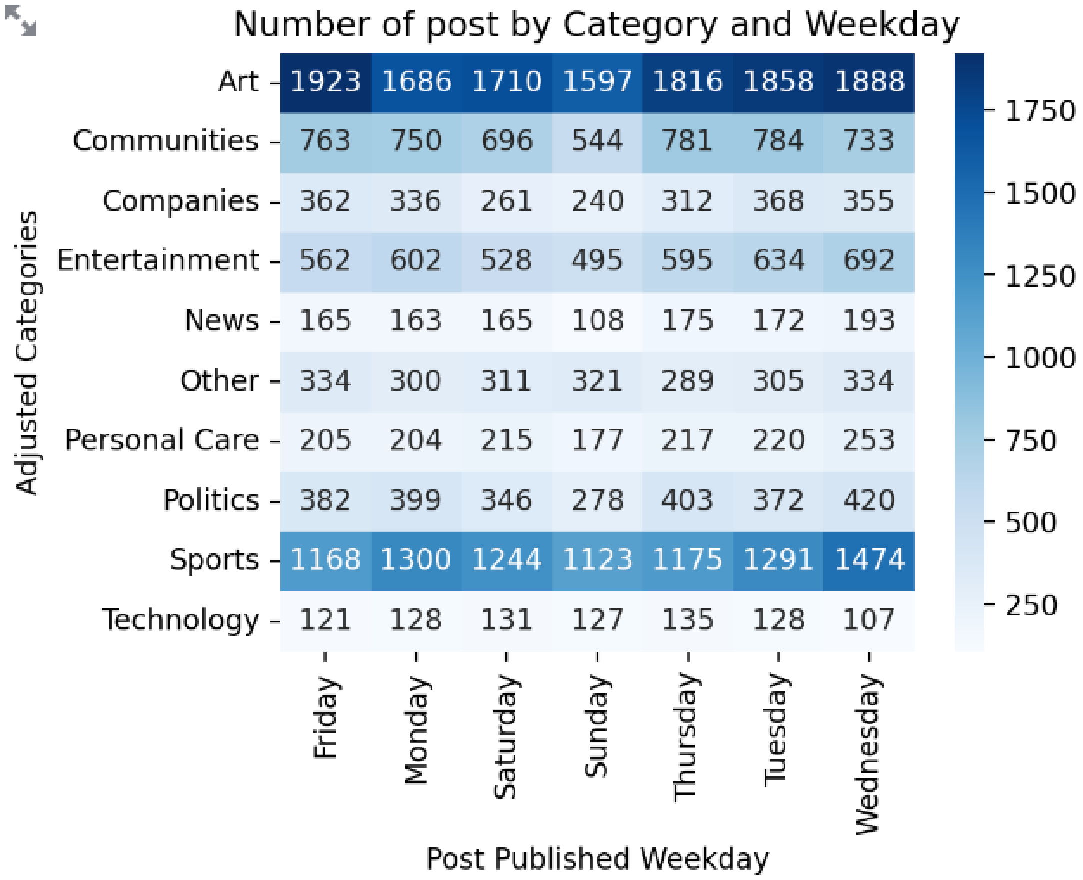
Also, we represented the evolution of the essential values between the last 48 to 24 hours and the last 24 hours to have a better understanding on the tendency.

# Bivariate Analysis

Weekday and Category | Correlations | Essential Features


Average Number of Comments by Page Category


Average Number of Comments by Publication Weekday

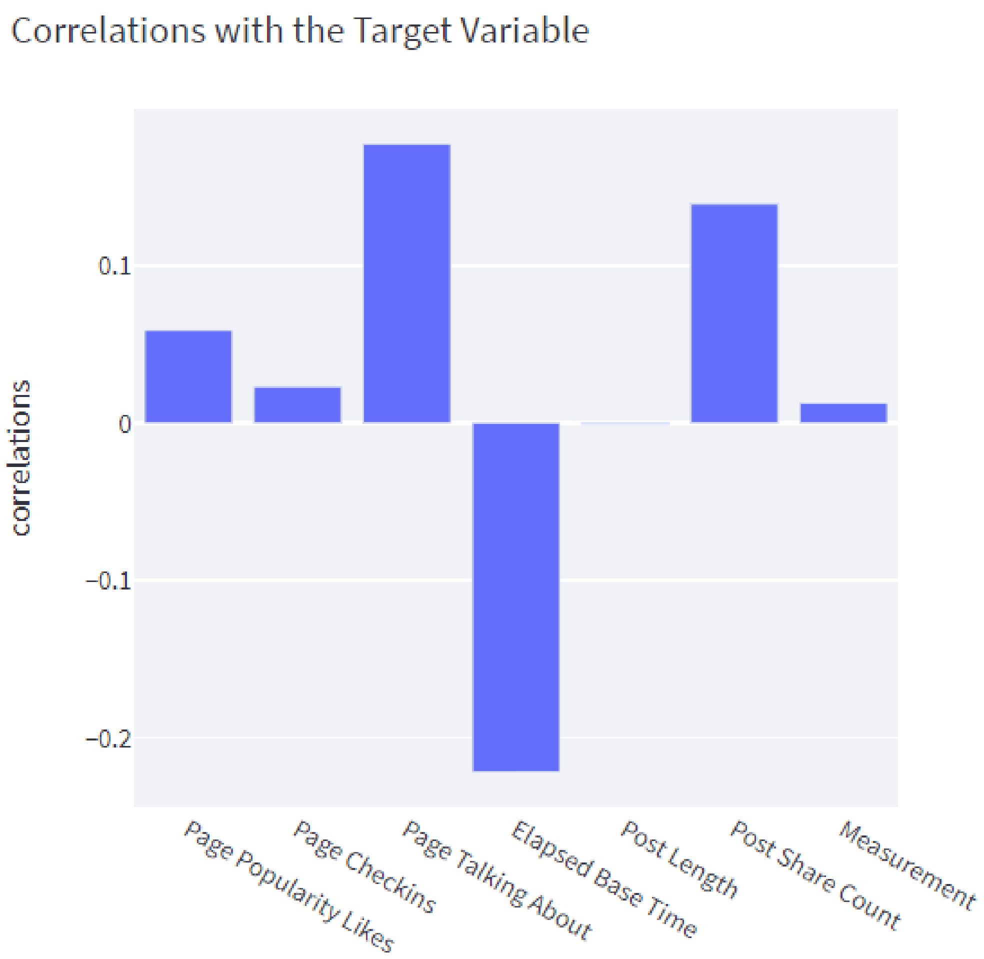

Number of post by Category and Weekday

| Adjusted Categories | Friday | Monday | Saturday | Sunday | Thursday | Tuesday | Wednesday |
|---|---|---|---|---|---|---|---|
| Art | 1923 | 1686 | 1710 | 1597 | 1816 | 1858 | 1888 |
| Communities | 763 | 750 | 696 | 544 | 781 | 784 | 733 |
| Companies | 362 | 336 | 261 | 240 | 312 | 368 | 355 |
| Entertainment | 562 | 602 | 528 | 495 | 595 | 634 | 692 |
| News | 165 | 163 | 165 | 108 | 175 | 172 | 193 |
| Other | 334 | 300 | 311 | 321 | 289 | 305 | 334 |
| Personal Care | 205 | 204 | 215 | 177 | 217 | 220 | 253 |
| Politics | 382 | 399 | 346 | 278 | 403 | 372 | 420 |
| Sports | 1168 | 1300 | 1244 | 1123 | 1175 | 1291 | 1474 |
| Technology | 121 | 128 | 131 | 127 | 135 | 128 | 107 |

Wednesday is the day the posts get the more comments. This also is the day where the more popular categories are the more active.

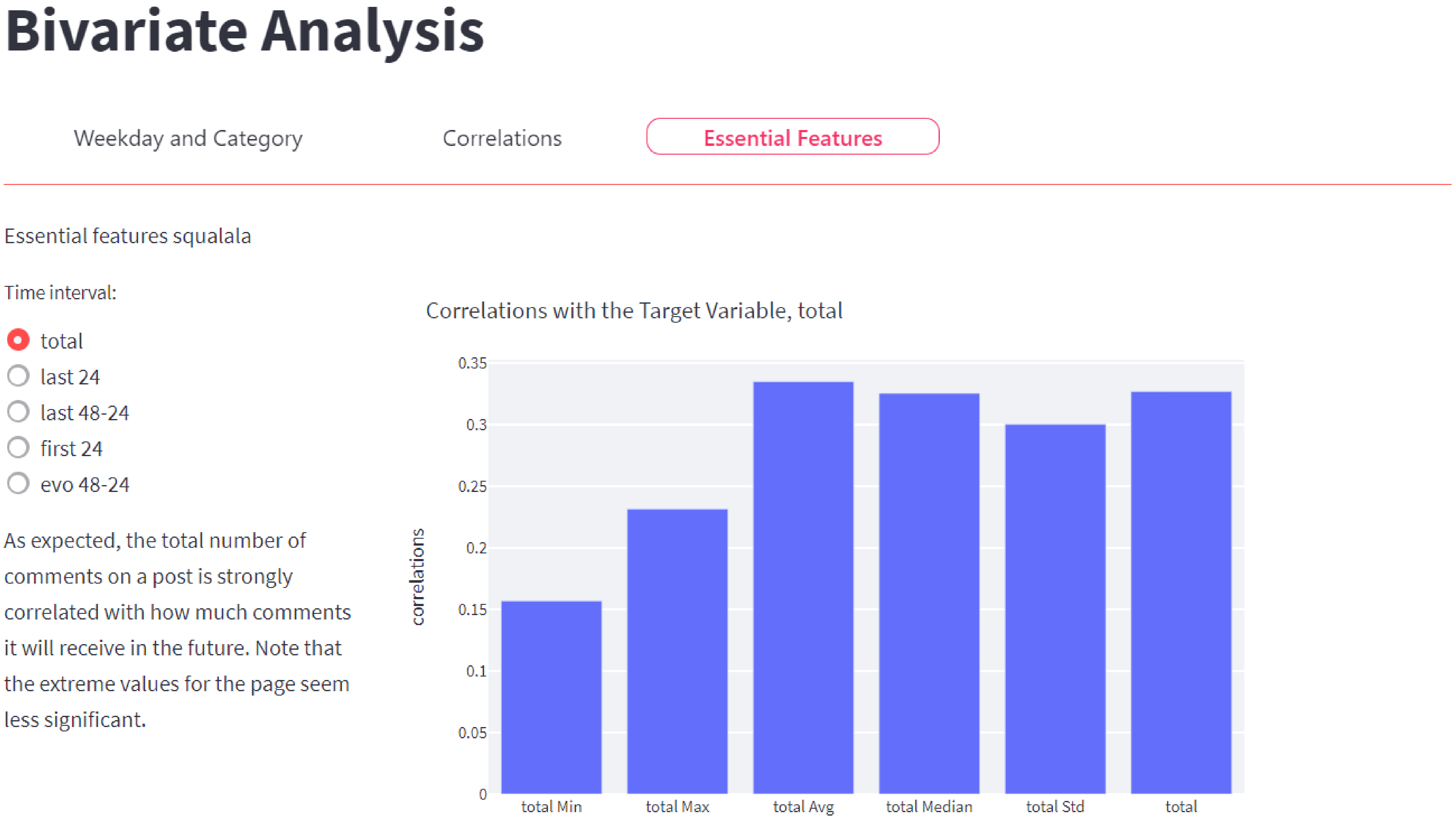
# Bivariate Analysis

Weekday and Category | Correlations | Essential Features



The time between the post publication and the web crawler inspection is strongly anti-correlated to the number of comments the post will get in the future. It means that a post generally gets a lot more comments shortly after its publication, but it quickly slows down.

Note that from the page features, the most important one in our case seems to be the number pages talking about the page that made the post. Also, the Length of a post seems to have almost no correlation to its number of comments.



# Bivariate Analysis

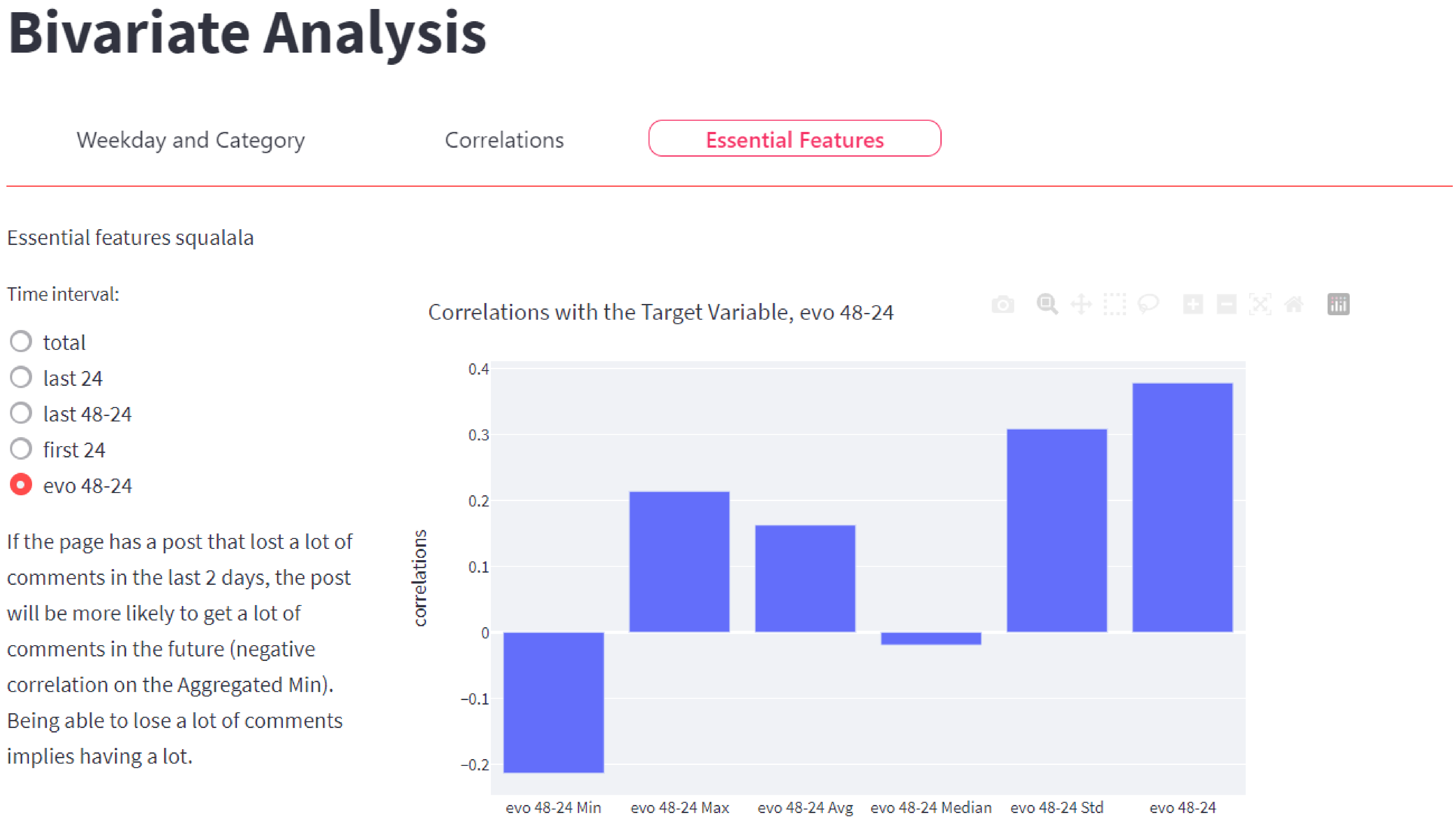
Essential features squalala

Time interval:

● total
○ last 24
○ last 48-24
○ first 24
○ evo 48-24

As expected, the total number of comments on a post is strongly correlated with how much comments it will receive in the future. Note that the extreme values for the page seem less significant.

Correlations with the Target Variable, total



# Bivariate Analysis

Essential features squalala

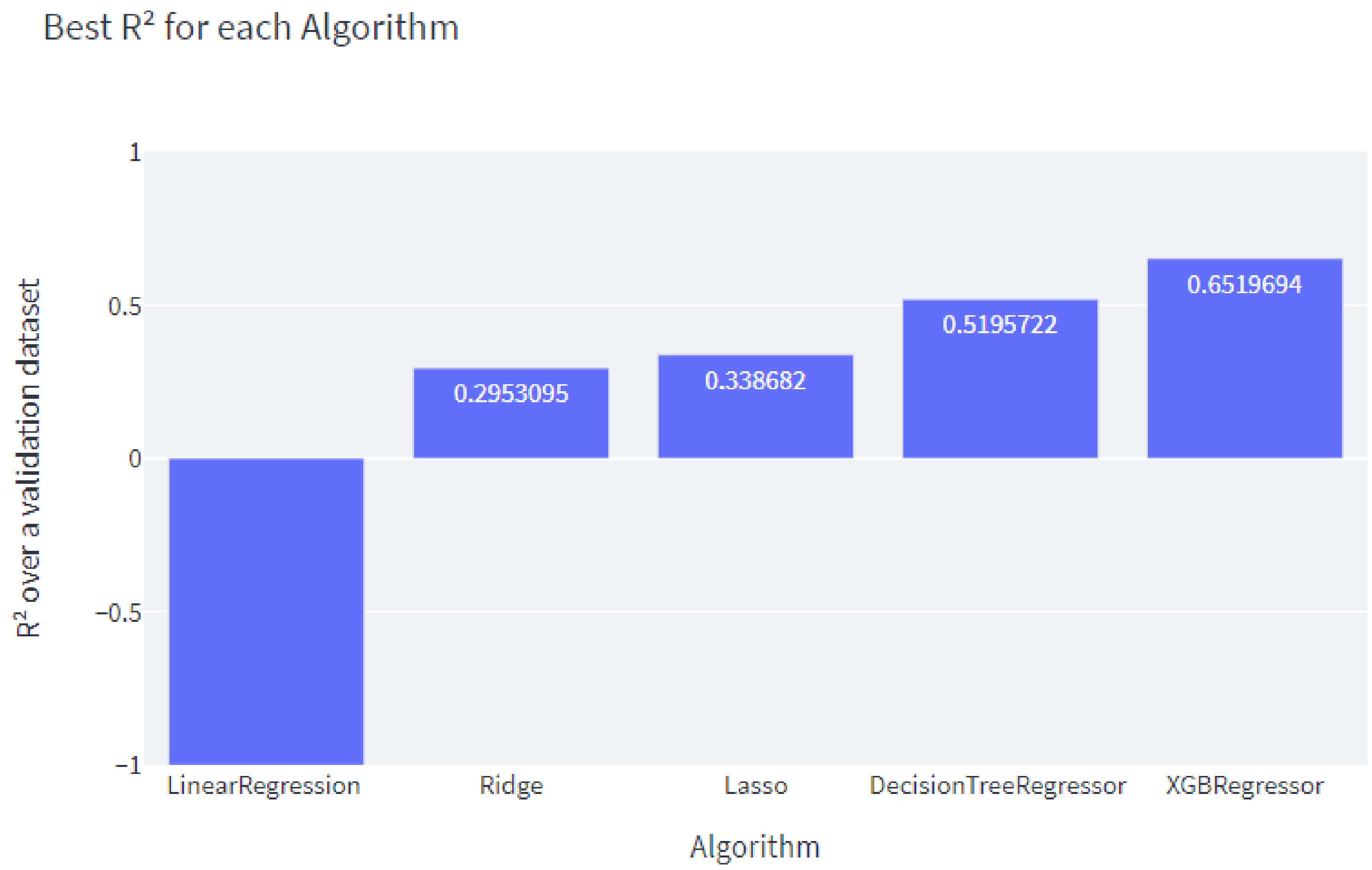
Time interval:

○ total
○ last 24
○ last 48-24
○ first 24
● evo 48-24

If the page has a post that lost a lot of comments in the last 2 days, the post will be more likely to get a lot of comments in the future (negative correlation on the Aggregated Min). Being able to lose a lot of comments implies having a lot.

Correlations with the Target Variable, evo 48-24

# Predictions



As we want to predict the number of comments a post will get in the future, we have a regression problem. Knowing that, we test multiple regression algorithm with multiple hyperparameters, using a grid search,to obtain the best possible model.

Our data cannot fit a linear regression at all. Although, XGBoost Regressor is the best performing, as we could expect, especially on large and heterogenous data. It performs quite well, but our data (or the social medias) does not seem totally predictable