EAS 4/587 – Data Intensive Computing
Phase 1 Report

Kumar Priyansh
UBIT Name: kumarpri

Ritu Dimri
UBIT Name: ritudimr

1 Motivation & Problem Statement

1.1 Motivation

- Reddit is a social media website where users can post links to articles, images, videos, etc. and other users can comment on them. Authors of the posts, generally look to drive maximum engagement from their posts.
- Unlike other social media websites, Reddit has a unique feature of upvoting and downvoting the posts. Also, since the posts are publicly visible, factors like the time of posting, the number of upvotes, the number of comments, etc. matter a lot.
- Since there are a lot of posts being made every minute, significant posts can get lost in the crowd. Also, the posts that are made at a particular time of the day, may not be visible to the users who are active at a different time of the day.

1.2 Problem Statement

- Fetch the data using the Reddit Developer API from different programming related subreddits (communities). Since, there are a lot of subreddits on Reddit; we will keep the scope of the project limited.
- Analyze the data and find relevant insights after cleaning and preprocessing the data.

- Build a model to predict the engagement a post will likely receive, given the time of posting, the number of upvotes, the number of comments, and other factors.

2 Data Collection

The data was collected from the Reddit Developer API[1]. The API provides access to the data of the posts made on Reddit. The data was collected from the following subreddits:

`python`, `datascience`, `javascript`, `linux`, `opensource`, `node`, `programming`, `computerscience`, `webdev`, `statistics`, `machinelearning`, `compsci`, `java`, `rust`, `typescript`.

3 Data Cleaning & Preprocessing

Note: Since the data is coming from a live API, the data is not static. Hence, the plots and the results may vary from the ones shown in the report.

After collecting the data, we got more than 15,700 rows and 118 columns. The data was cleaned and preprocessed to remove the unnecessary columns and to make the data more readable. The data was cleaned and preprocessed using the following steps:

1. **Dropping Saturated Columns**: Saturated columns are the columns that have a single value for all the rows. These columns do not provide any useful information and hence, they were dropped.

2. **Fixing Data Types**: Some columns can have incorrect data type when converted into a Pandas DataFrame. All data was converted to `string` type, except numeric (`int` & `float`) and `boolean` columns. Some boolean colums like `distinguished` and `author_premium` required forced type conversion to `boolean` type.

3. **Drop Duplicates**: Sometimes, the same post is copied and pasted in multiple subreddits. These posts are duplicates and hence, they were dropped, keeping only one copy of the post.

[1] https://www.reddit.com/dev/api/

4. **Check for Missing and Null Values**: The data was checked for missing and null values. After the above steps, there were no missing or null values in the data.

5. **Handling Flairs**: Flairs are the tags that are assigned to the posts. They are used to categorize the posts. For example, a post can be tagged as `help`, `discussion`, `news`, etc. Only the relevant flair columns were kept. The flairs might also contain emojis, which were also removed in a later step.

6. **Handling Deleted Posts & Posts by Deleted Users**: Posts that were deleted by the author or the moderators were removed from the data. Also, posts by deleted users were removed.

7. **Handling Polls**: Polls are posts that are used to collect votes from the users. These posts were removed from the data as they do not provide any useful information for text analysis.

8. **Dropping Unnecessary Columns**: Some columns were dropped as they were not relevant for the analysis. For example, columns like `thumbnail_height` and `thumbnail_width` were dropped as they do not provide any useful information.

9. **Lowecasing Text Columns**: All text columns were converted to lowercase.

10. **Removing HTML Entities, Escape Characters and URLs**: HTML entities, escape characters and URLs were removed from the text columns.

11. **Removing Punctuations, Numbers and Emojis**: Punctuations, numbers and emojis were removed from the text columns.

12. **Removing Stopwords & Contractions**: Stopwords and contractions were removed from the text columns.

13. **Lemmatization & Stemming**: Lemmatization and stemming were performed on the text columns to reduce the words to their root form.

4 Expolaratory Data Analysis

After cleaning and preprocessing the data, we were left with around 8,000+ rows and 28 columns. The data was analyzed to find relevant insights. The data was analyzed using the following steps:

Number of Posts Fetched per Subreddit

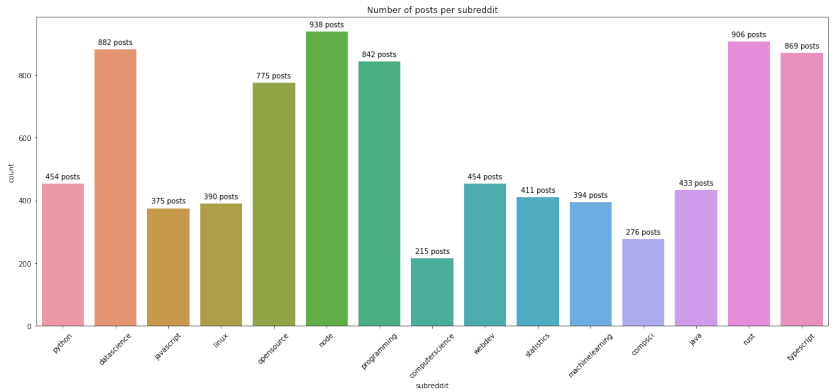

Figure 1: Number of Posts per Subreddit

Number of Subscribers per Subreddit

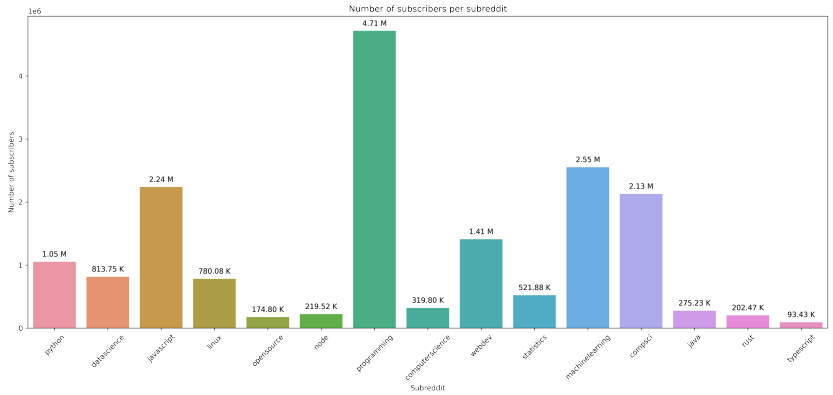

Figure 2: Number of Subscribers per Subreddit

Number of Authors who post in multiple subreddits

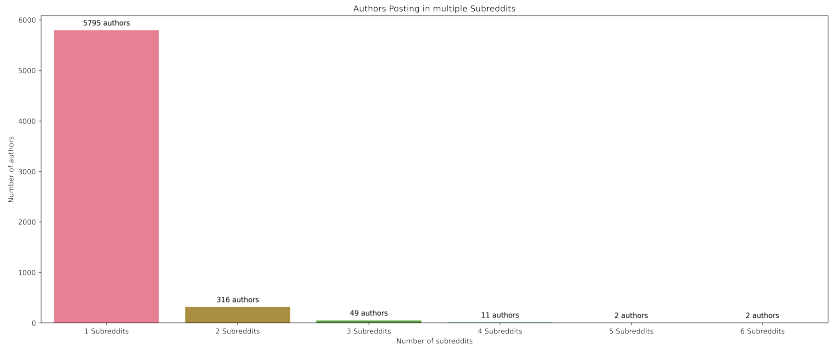

Figure 3: Number of Authors who post in multiple subreddits

From the above figure, it seems like most of the authors post in only one subreddit.

Does posting in multiple subreddits drives more upvotes?

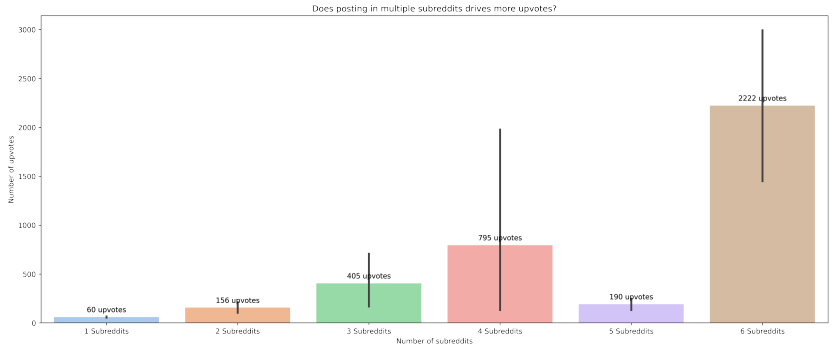

Figure 4: Does posting in multiple subreddits drives more upvotes?

From the above figure, it seems like authors who post in multiple subreddits get more upvotes, generally.

History of posts per day per subreddit (Past 6 months)

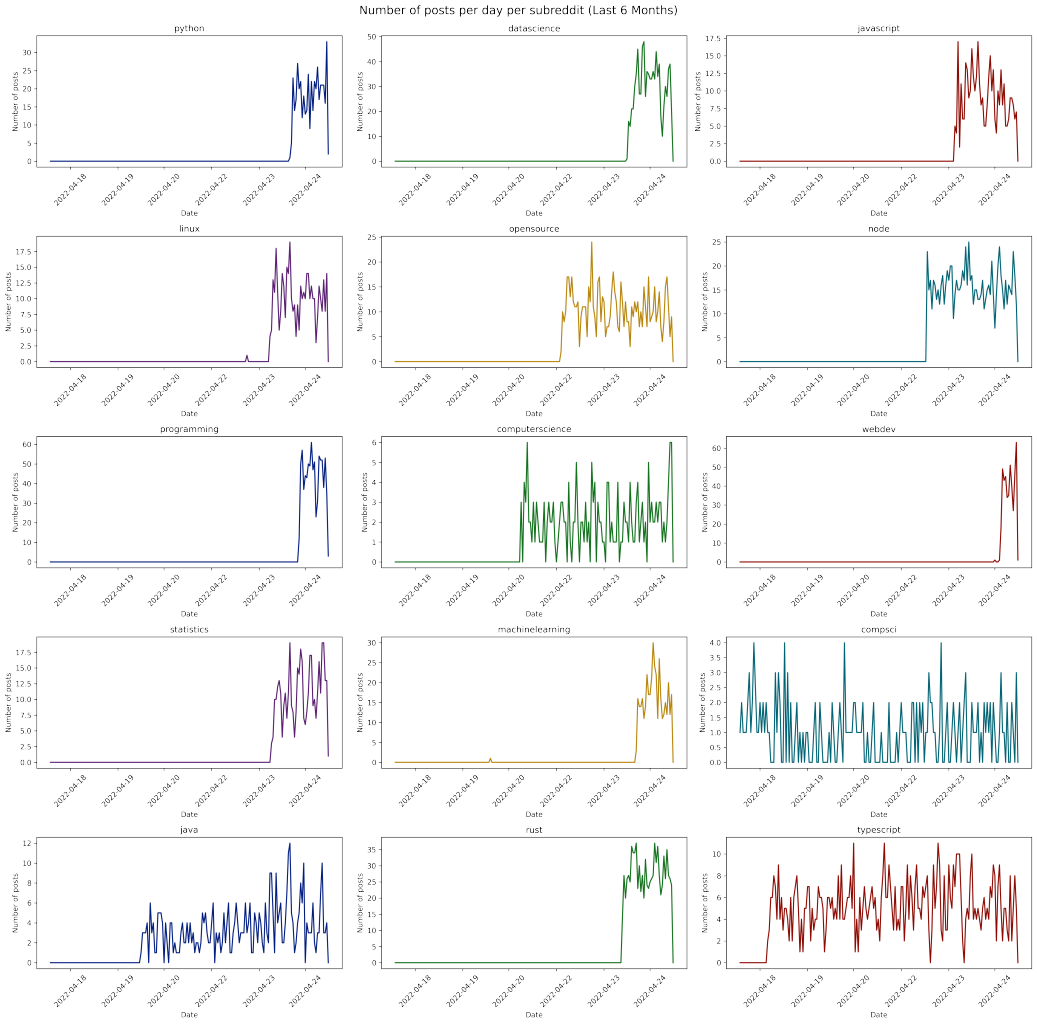

Figure 5: History of posts per day per subreddit

From the above figure, it seems like the number of posts per day is increasing for most of the subreddits.

Number of posts per subreddit, categorized by days of the week

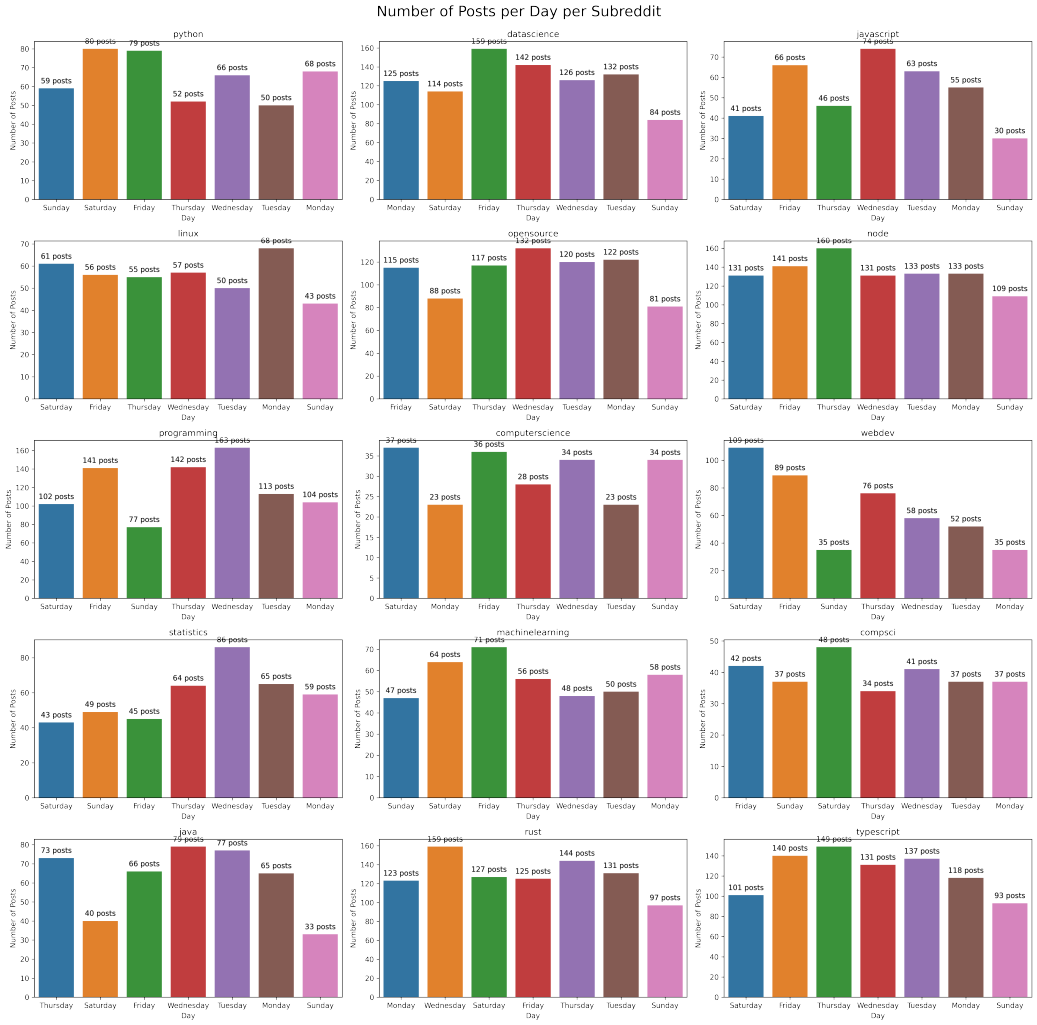

Figure 6: Number of posts per day per subreddit

The above figure presents total number of posts per subreddit, categorized by the days of the week.

Top 10 Authors Per Subreddit

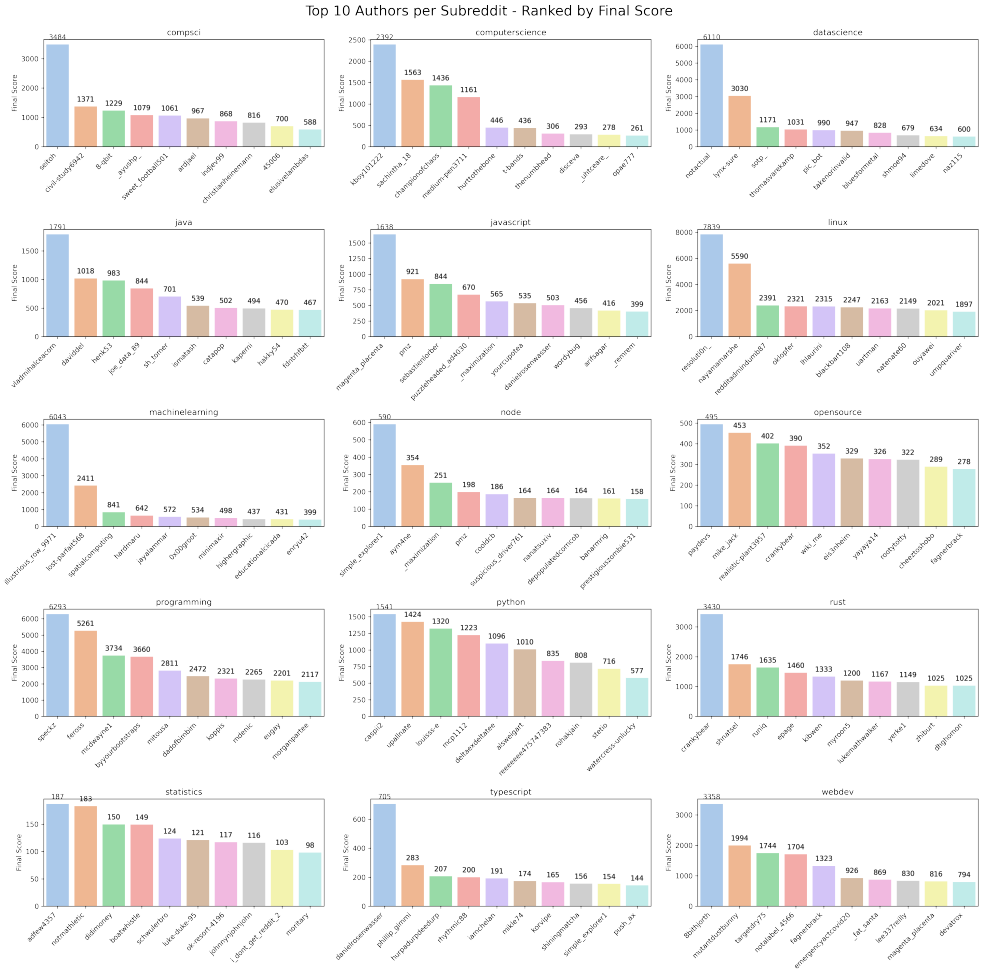

Figure 7: Top 10 Authors Per Subreddit

The above plot was generated by creating an author ranking system. The authors were then sorted in descending order of their rank. The top 10 authors were then plotted. The ranking system uses the following simple formula to rank the authors:

$$\text{Rank} = \text{Post Score} \times \text{Upvote Ratio} + \text{Number of Comments} \tag{1}$$

Finding the best time to post on each subreddit

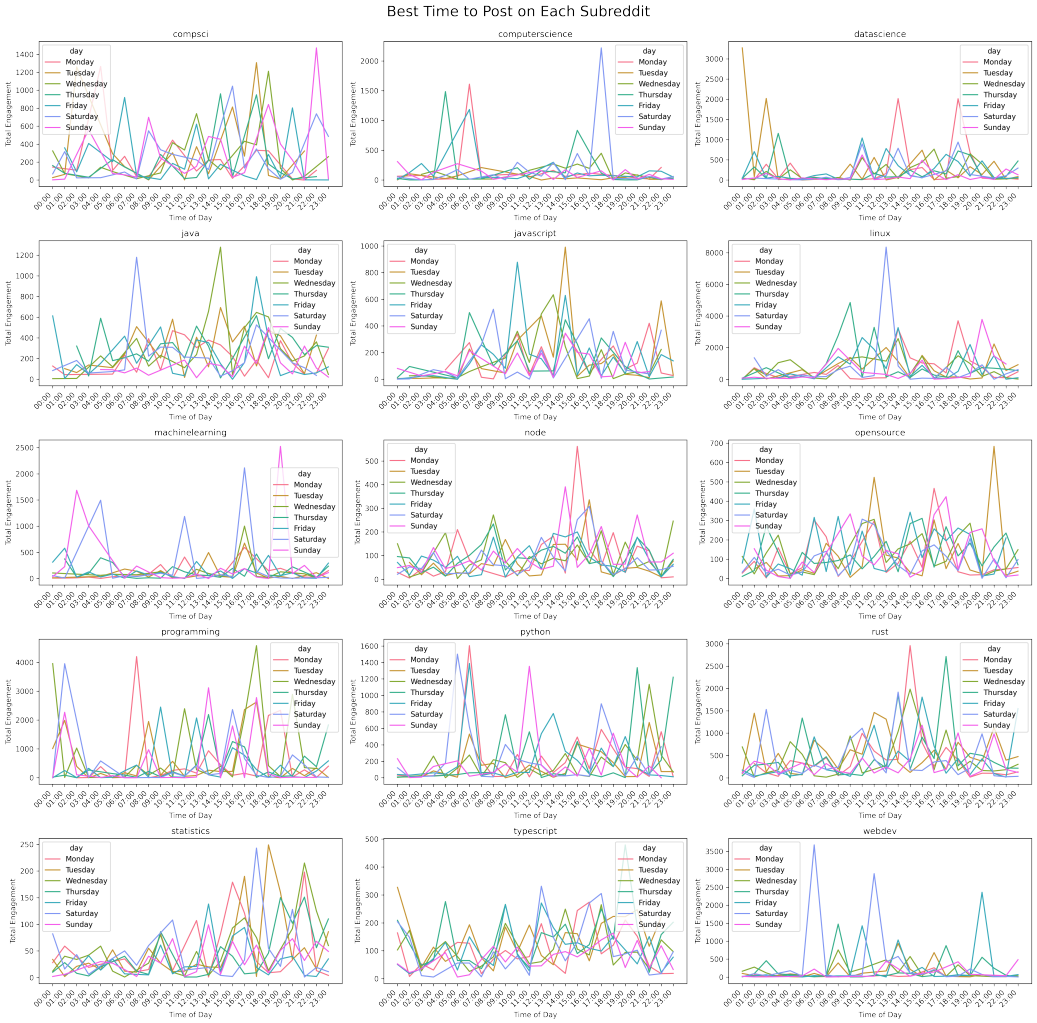

Figure 8: Best Time to Post on each Subreddit

The above plot was generated by finding the total hourly engagement on each subreddit, categorized by each day. It seems like the best time to post on each subreddit is different and there happens to be a peak in the engagement at different times of the day.

Scatterplot of the scores vs number of comments in each subreddit

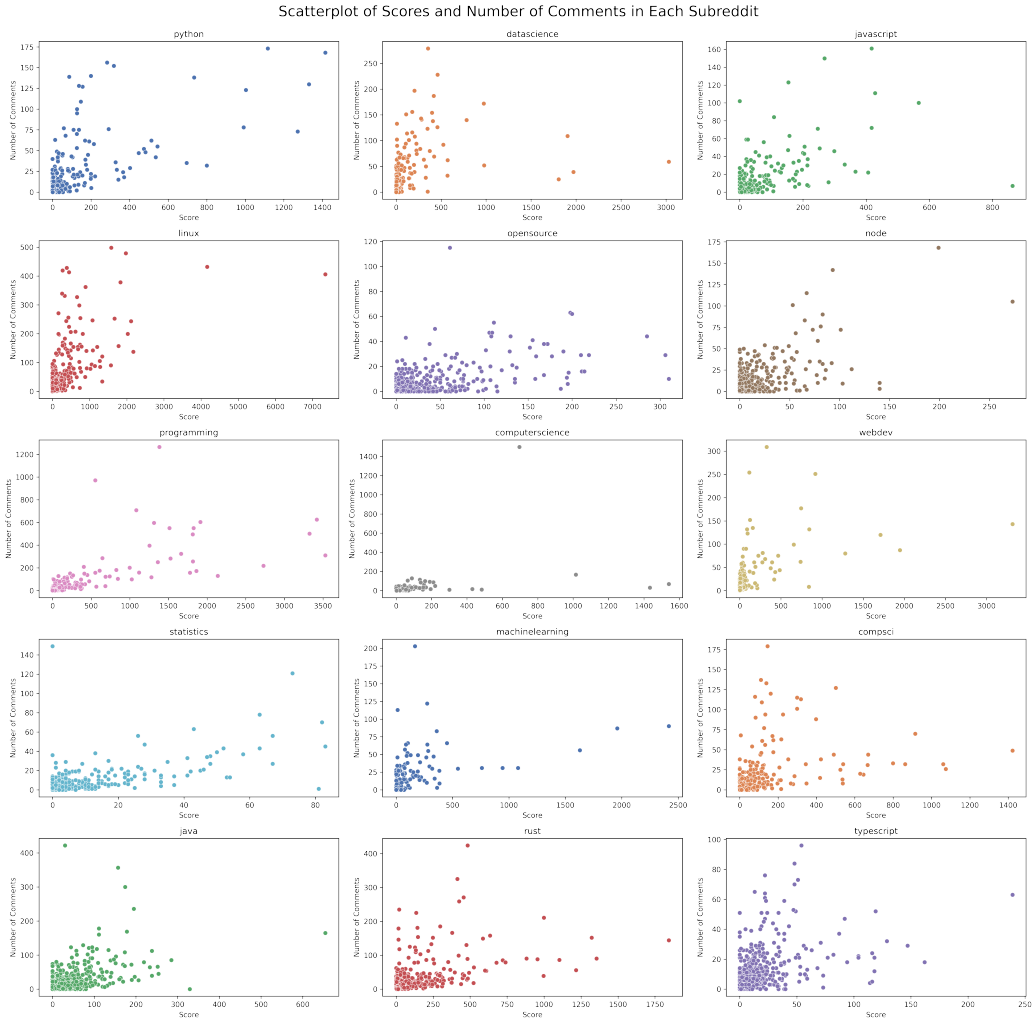

Figure 9: Scatterplot of the scores vs number of comments in each subreddit

The above plot was generated by plotting the scores vs number of comments in each subreddit. It seems like there is a positive correlation between the scores and the number of comments, except for some controversial posts, which tend to get more comments than the scores.

% of Authors opted to receive comment notifications per Subreddit

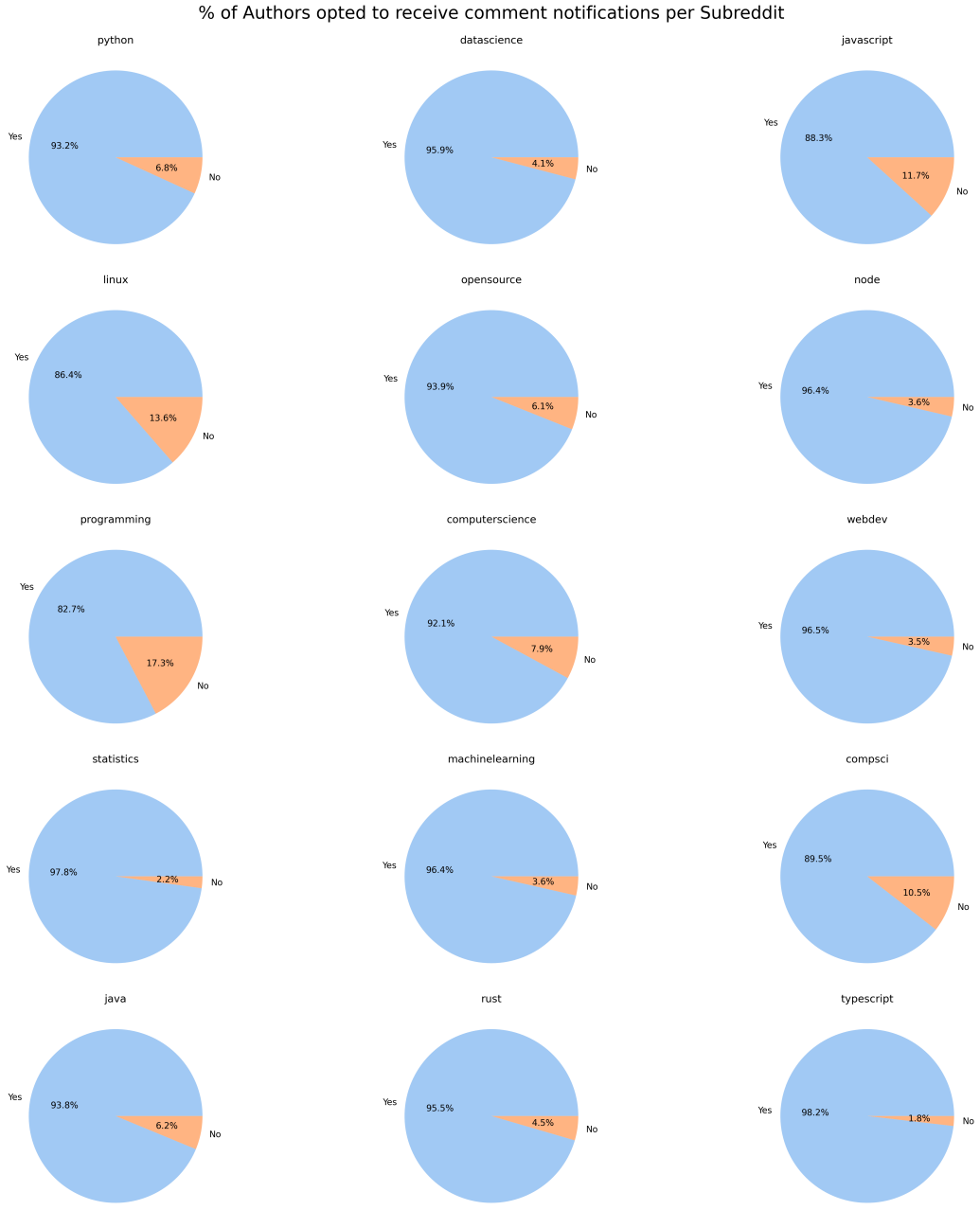

Figure 10: % of Authors opted to receive comment notifications per Subreddit

The above plot was generated by finding the percentage of authors who opted to receive comment notifications in each subreddit. It seems like most of the authors opted to receive comment notifications.

Total Awards Received vs Upvotes per Subreddit

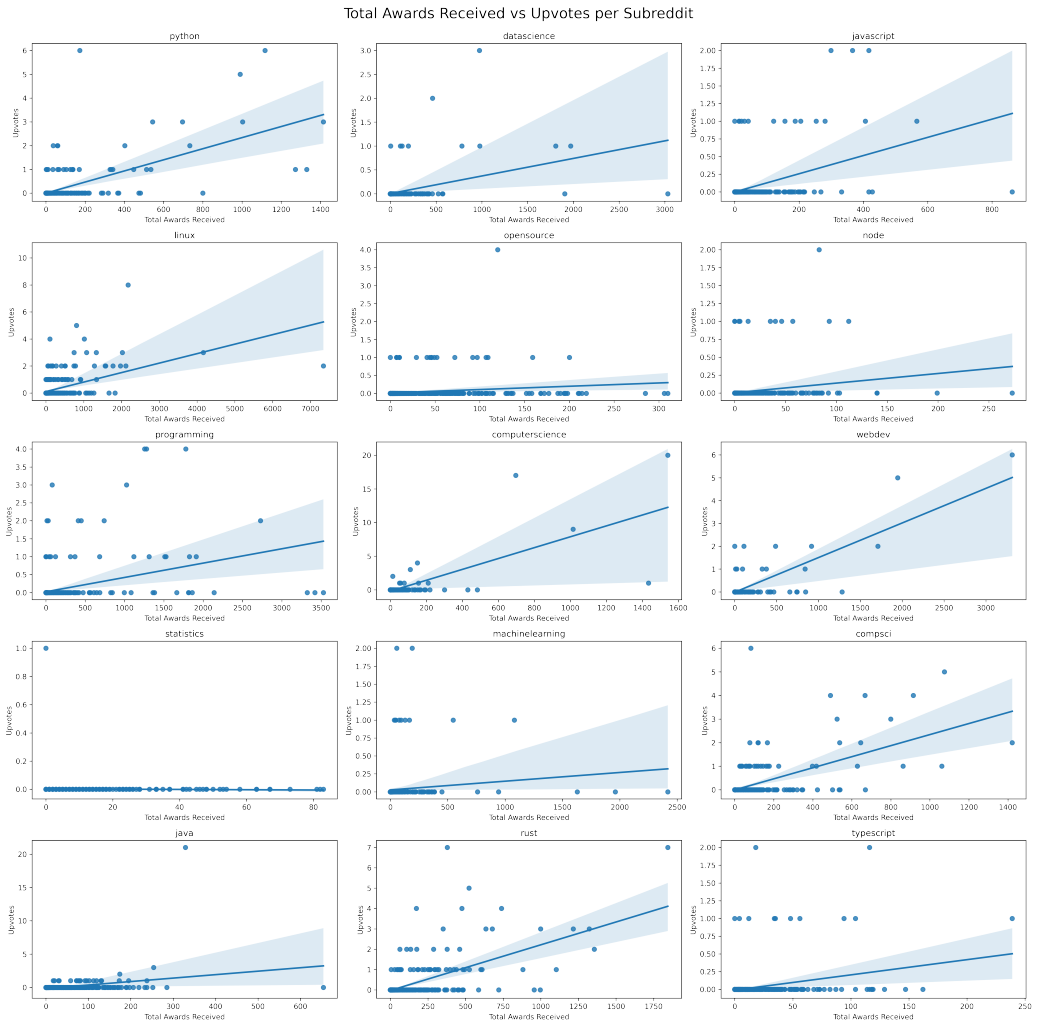

Figure 11: Total Awards Received vs Upvotes per Subreddit

Good posts are usually awarded with "awards". The above plot was generated by plotting the total awards received vs upvotes in each subreddit. It

seems like there is a positive correlation between the total awards received and the upvotes.

Boxplot of the scores in each subreddit

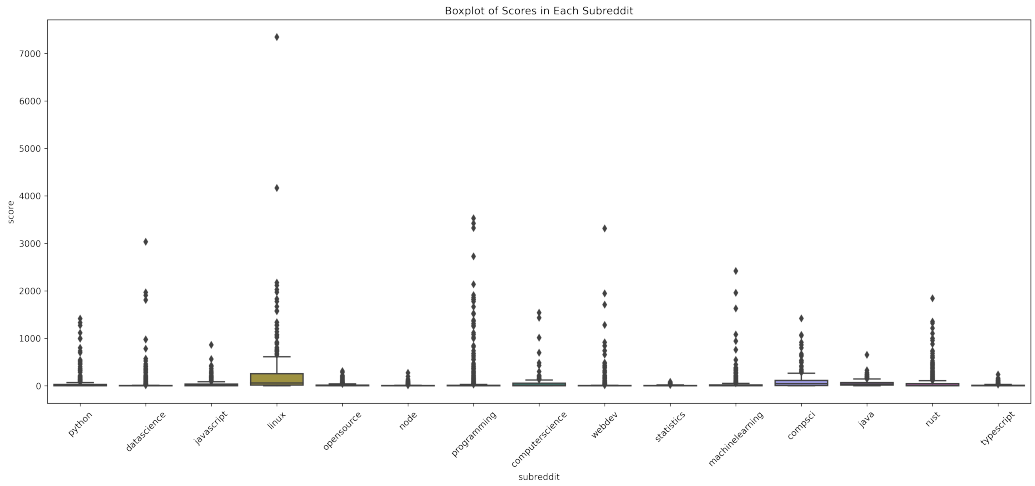

Figure 12: Boxplot of the scores in each subreddit

The above plot was generated by plotting the boxplot of the scores in each subreddit. It seems like most of the scores are concentrated in the lower range, except for a few outliers – they are the posts that went viral. It means that most of the posts are not very popular.