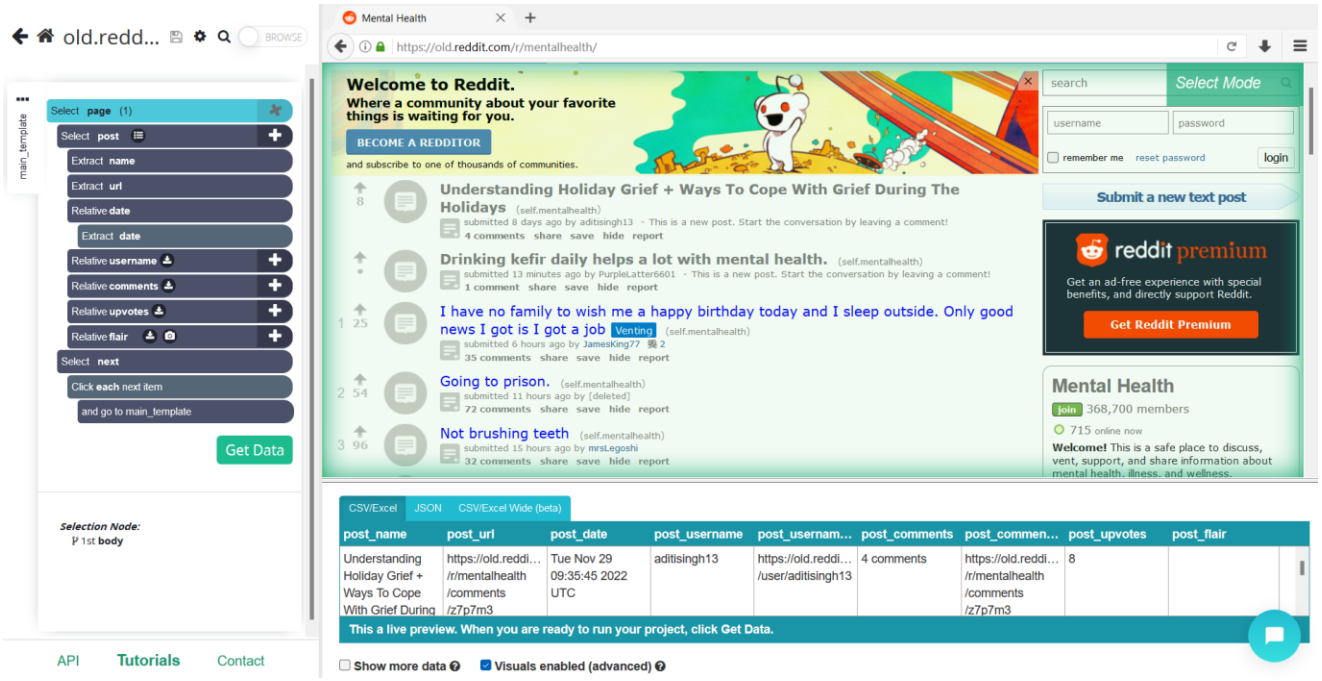
***Site:*** Reddit; specifically, a subreddit dedicated to mental health.

***r/mentalhealth***

## ***Approach/Objective:***

To find out what kind of posts and keywords are prevalent in a forum dedicated to mental health, and a peak into the demographic of the subreddit. It was sorted by "Hot" to remove any kind of bias that would be prevalent in "Most Upvoted", "Most Recent", "Most Controversial", so this category would bring a mix of recent posts and also a few trending posts that would also be relatively recent and not too old.




All variables were pretty straight-forward to fetch, but if I were to fetch how long ago each post was made as it was, it would be redundant since it would not change dynamically with the passage of time, so I altered the Date attribute by fetching the exact date and time the post was made. I extracted several pages (probably up to 17), and the number of rows fetched were about 402.

I have used R for data wrangling. There is a duplicate column for post upvotes which I have only created for testing purposes, and it serves no real purpose, therefore, the attributes that were originally extracted will only be used for my analysis.

Upon attempting visualizations, I discovered that post_upvotes was a Character column and had a peculiar string in the place of 0, so I replaced that and also converted the class type to Numeric.

Additionally, post_comments had the word comment(s) attached to the number and I figured that this could potentially interfere with my analysis, so I replaced the comment with blank in Excel, and the entirely blank cells with 0. Just like above, I have created a duplicate column for comments for testing purposes, and nothing more.

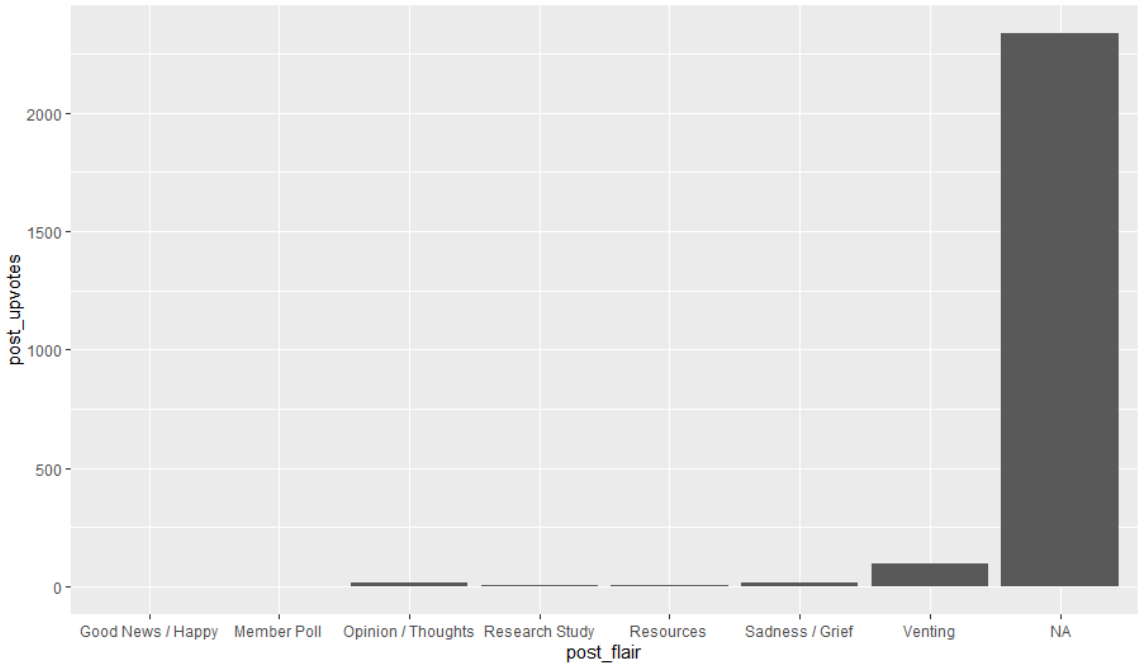
Furthermore, I replaced the null spaces in post_flairs to NA.

***Analysis***

1) I wanted to find out what words appeared more frequently in all the posts I have extracted, to attain insight into the psyche of people who frequently post on the forum and what kind of issues are most prevalent among them, which could be figured out by key-word usage. So, after bringing this data into R Studio, I started with mapping out visualizations that would be useful pertaining to the site and my analysis approach. So, I made a WordCloud befitting the scenario of finding out the word frequency usage for the posts made at the forum. Here are the results:


The results in this are pretty self-explanatory as a cry for help is more evident which would obviously be on top of any mental health forum, followed by some existential crisis, anxiety, depression, psychosis, and other mental illnesses. Other issues like relationships, familial bonds, loneliness, and emotions relating to prolonged monotony in life is also evident in this visualization.

2) Next, I wanted to find out what flairs garnered most attention in this subreddit. For this I created a Bar Plot to find out the most prevalent flairs attached to the post.



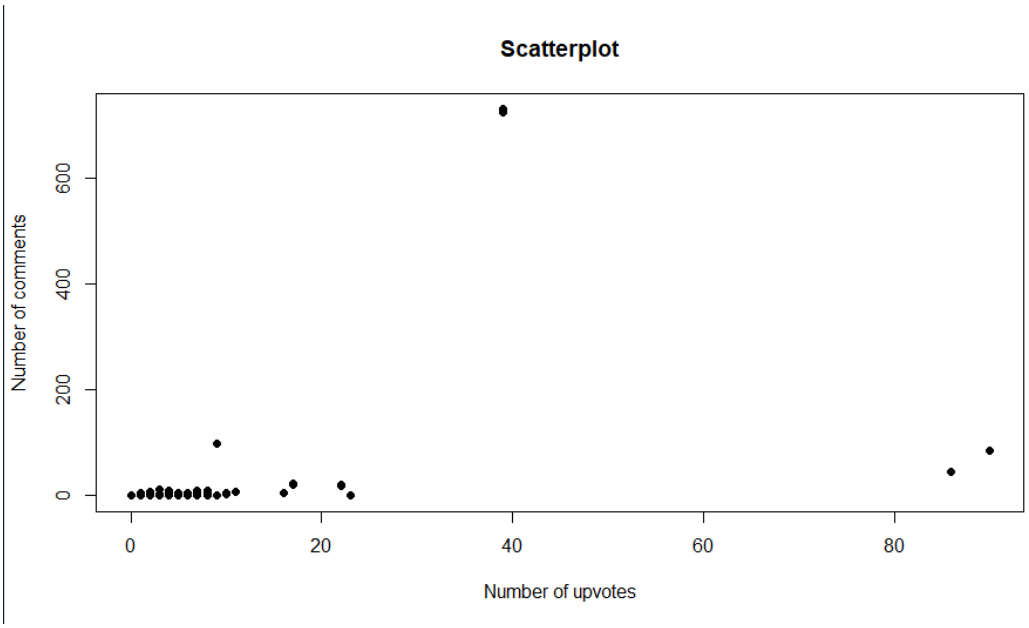

As we can see here, most users did not even bother attaching a flair on their posts, yet it garnered most of the upvotes. That is because, statistically, most users do not attach a flair on their post on Reddit unless if it is made mandatory by a particular subreddit, in this case it was not a requirement, as such most posts are without a flair, so obviously it will, by default contain the highest number of votes. However, if we consider that it is loosely followed by Venting flair, we can see that a lot of people may relate to or provide some kind of support or advice as it seems to be the highest upvoted among actual flairs. Or perhaps this bias occurs due to more users posting on this subreddit to actually vent their frustrations than other kind of posts?

3) Upon finding out the number of posts attached to the flair or not, there were about 309/402 posts that did not have a flair. As we can see here, these are the number of posts that did have flairs each.

```
> length(which(data$post_flair == "Venting"))
[1] 58
> length(which(data$post_flair == "Resources"))
[1] 6
> length(which(data$post_flair == "Research Study"))
[1] 2
> length(which(data$post_flair == "Sadness / Grief"))
[1] 13
> length(which(data$post_flair == "Opinion / Thoughts"))
[1] 11
> length(which(data$post_flair == "Member Poll"))
[1] 2
> length(which(data$post_flair == "Good News / Happy"))
[1] 1
```

Adding them all up gives a total of 93/402 posts that did have a flair attached. So, this supports my explanation above regarding the number of upvotes posts garnered is due to the number of posts that have a particular flair attached. We can see here that Venting kind of posts are still most prevalent, followed by Sadness/Grief.

4) Does a post garnering attention through upvotes necessarily mean that users are also more likely to comment on those posts? We need to check the relationship between these two variables, and I believe a ScatterPlot would do just the trick.



Here we can see that this is not necessarily true, as posts relatively lesser number of upvotes may also have a lot of comments, and posts with high number of upvotes may or may not have the same number of comments as those posts with fewer number of upvotes. Therefore, post upvotes does not necessarily equate further engagement, as it would depend on the quality or nature of post, what it is insinuating, and what kind of audience it is attracting.

5) Linking to our post engagement analysis above, with addition to our first analysis (Word Cloud), it would be essential to perform a sentiment analysis derived from the posts in the r/mentalhealth forum. So, after attaining sentiment scores, it would be more efficient to fetch the scores for the posts that are at the front page of the subreddit.

```
> head(snt)
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     0            1       0    0   1       1        0     1        1        2
2     0            0       0    0   1       1        0     1        0        1
3     3            1       1    1   1       2        1     1        3        1
4     0            0       0    0   0       0        0     0        0        0
5     0            0       0    0   0       0        0     0        0        0
6     1            0       1    1   1       2        0     1        1        1
>
```

Among these posts, we can see that the third one appears to have greater scores in all categories and carry mixed emotions. Let's see what it says.

```
> postname[3]
[1] "how can i stop feeling guilty about self harming after an argument with my s/o?"
>
```

Okay, even I was surprised by this. I am at a loss for words. That is, for this person, I can still go on with my analysis.

Feelings of guilt and self-harm does factor into a lot of these scores like sadness, anger, negative. Additionally, argument may also play a part in this, and I do have a feeling perhaps s/o which stands for significant other may factor into positive and joy? I certainly can't speak for certainty, but we'll let data speak for itself.

```
> get_nrc_sentiment('guilty')
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     1            0       0    0   0       1        0     0        1        0
```

```
> get_nrc_sentiment('feeling')
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     1            1       1    1   1       1        1     1        1        1
```

I certainly didn't realize how the word "feeling" could factor into all categories, which kind of makes sense but falls short in terms of context, but without context it can be pretty ambiguous and therefore factored into all categories.

```
> get_nrc_sentiment('argument')
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     1            0       0    0   0       0        0     0        1        0
```

As discussed above, argument has factored into the scores we've predicted. However, it returned no score for s/o so it does not factor that. Even self harming returned nil.

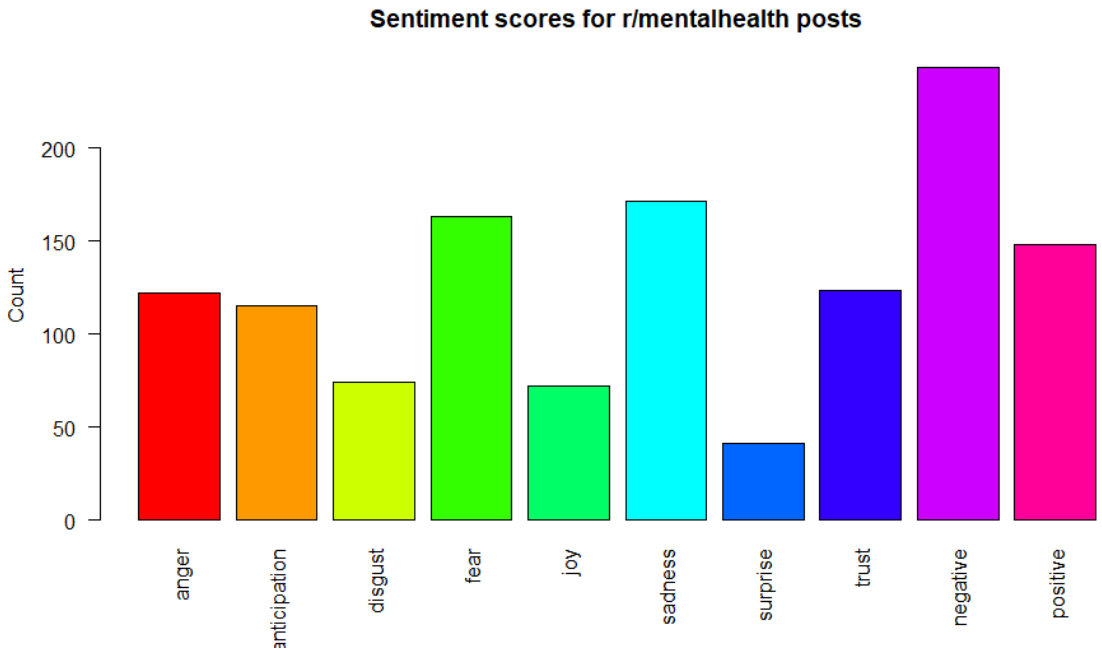
Now, let's find out the person behind this post.

```
> data$post_username[3]
[1] "Fun_Distribution_143"
>
```

With this, we shall fetch all the posts made by this person. Apparently, this was the only one.

```
3 how can i stop feeling guilty about self harming after an argument with my s/o?
                                                                                post_
url
3 https://old.reddit.com/r/mentalhealth/comments/ze7p22/how_can_i_stop_feeling_guilty_about_self_harmi
ng/
                  post_date        post_username
3 Tue Dec 6 14:07:32 2022 UTC Fun_Distribution_143
                              post_username_url post_comments
3 https://old.reddit.com/user/Fun_Distribution_143             1
                                                                          post_comments_
url
3 https://old.reddit.com/r/mentalhealth/comments/ze7p22/how_can_i_stop_feeling_guilty_about_self_harmi
ng/
  post_upvotes post_flair post_votes post_comment
3            0       <NA>          0            1
```

There’s not much to see here, so I made a Bar Plot for visualizing the sentiment scores for the mental health subreddit posts.




As expected, negative sentiment heavily overshadows positive ones, where sadness, fear, and anger factor in the most of it. Additionally, we can see that trust is there up quite high as well, slightly above anger itself, so that’s an interesting aspect to look at.

Author - Obaidullah