Analysis

Identify and analyze problems people on India talk about on social media. The social media platform chosen for this task is Reddit.

Reddit is an American social news aggregation, content rating, and discussion website. Registered users submit content to the site such as links, text posts, images, and videos, which are then voted up or down by other members.
The site is composed of hundreds of subcommunities, known as subreddits. Each subreddit has a specific topic, such as technology, politics or music. Reddit's homepage, or the front page, as it is often called, is composed of the most popular posts from each default subreddit.

The subreddit chosen for analysis is r/India, which is a subreddit dedicated to discuss about problems/events concurring to India & Indian people.

Contents :

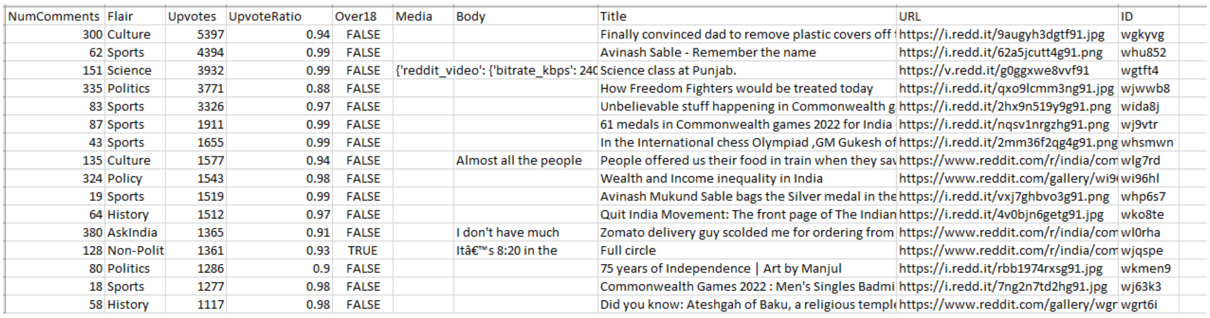
Introduction

While making a submission to r/India, the user has to tag *flairs* which appropriately associates the post. This helps in identifying and interacting with the people interested in the same domain. Different type of flairs in this subreddit are :

- Politics
- Sports
- Rant
- Non Political
- Law
- Crime
- Health
- Science
- And many more

The data collected was for the top posts spanning for a week. In total there are 937 different submissions (& 10 columns) collected spanning across different flairs.

A look into the data

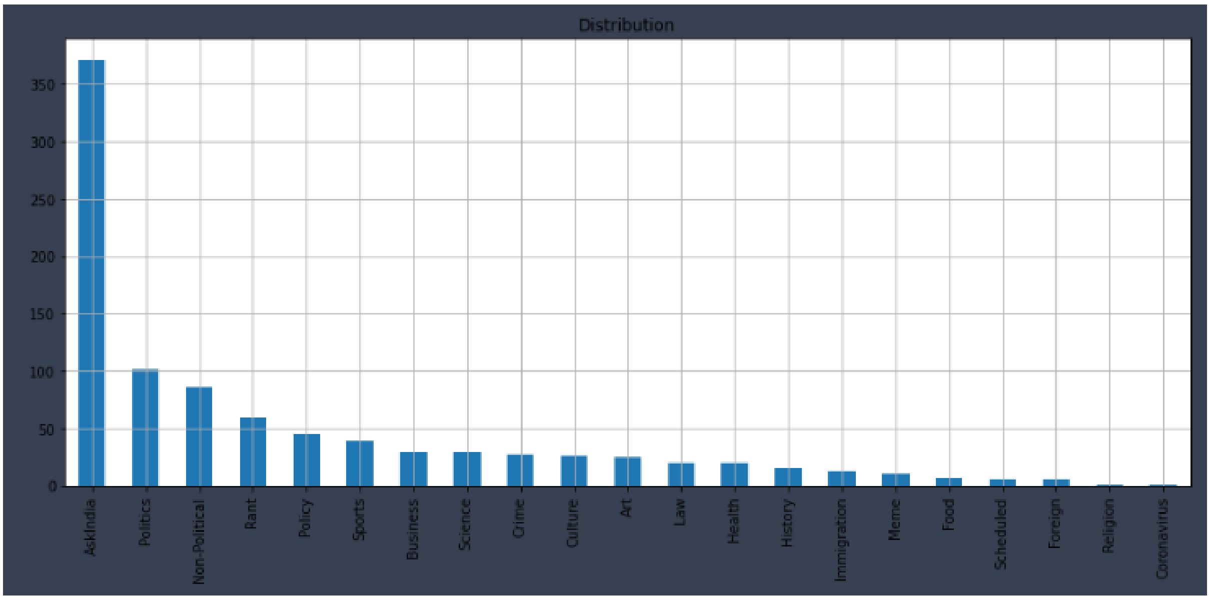


NumComments	Flair	Upvotes	UpvoteRatio	Over18	Media	Body	Title	URL	ID
300	Culture	5397	0.94	FALSE			Finally convinced dad to remove plastic covers off	https://i.redd.it/9augyh3dgtf91.jpg	wgkyvg
62	Sports	4394	0.99	FALSE			Avinash Sable - Remember the name	https://i.redd.it/62a5jcutt4g91.png	whu852
151	Science	3932	0.99	FALSE	{'reddit_video': {'bitrate_kbps': 240		Science class at Punjab.	https://v.redd.it/g0ggxwe8vvf91	wgtft4
335	Politics	3771	0.88	FALSE			How Freedom Fighters would be treated today	https://i.redd.it/qxo9lcmm3ng91.jpg	wjwwb8
83	Sports	3326	0.97	FALSE			Unbelievable stuff happening in Commonwealth g	https://i.redd.it/2hx9n519y9g91.png	wida8j
87	Sports	1911	0.99	FALSE			61 medals in Commonwealth games 2022 for India	https://i.redd.it/nqsv1nrgzhg91.png	wj9vtr
43	Sports	1655	0.99	FALSE			In the International chess Olympiad ,GM Gukesh of	https://i.redd.it/2mm36f2qg4g91.png	whsmwn
135	Culture	1577	0.94	FALSE		Almost all the people	People offered us their food in train when they sa	https://www.reddit.com/r/india/com	wlg7rd
324	Policy	1543	0.98	FALSE			Wealth and Income inequality in India	https://www.reddit.com/gallery/wi9	wi96hl
19	Sports	1519	0.99	FALSE			Avinash Mukund Sable bags the Silver medal in the	https://i.redd.it/vxj7ghbvo3g91.png	whp6s7
64	History	1512	0.97	FALSE			Quit India Movement: The front page of The Indian	https://i.redd.it/4v0bjn6getg91.jpg	wko8te
380	AskIndia	1365	0.91	FALSE		I don't have much	Zomato delivery guy scolded me for ordering from	https://www.reddit.com/r/india/com	wl0rha
128	Non-Polit	1361	0.93	TRUE		Itâ€™s 8:20 in the	Full circle	https://www.reddit.com/r/india/com	wjqspe
80	Politics	1286	0.9	FALSE			75 years of Independence | Art by Manjul	https://i.redd.it/rbb1974rxsg91.jpg	wkmen9
18	Sports	1277	0.98	FALSE			Commonwealth Games 2022 : Men's Singles Badmi	https://i.redd.it/7ng2n7td2hg91.jpg	wj63k3
58	History	1117	0.98	FALSE			Did you know: Ateshgah of Baku, a religious templ	https://www.reddit.com/gallery/wgr	wgrt6i

Overall Analysis

Distribution of Flairs:

Not displaying flairs with posts < 15.

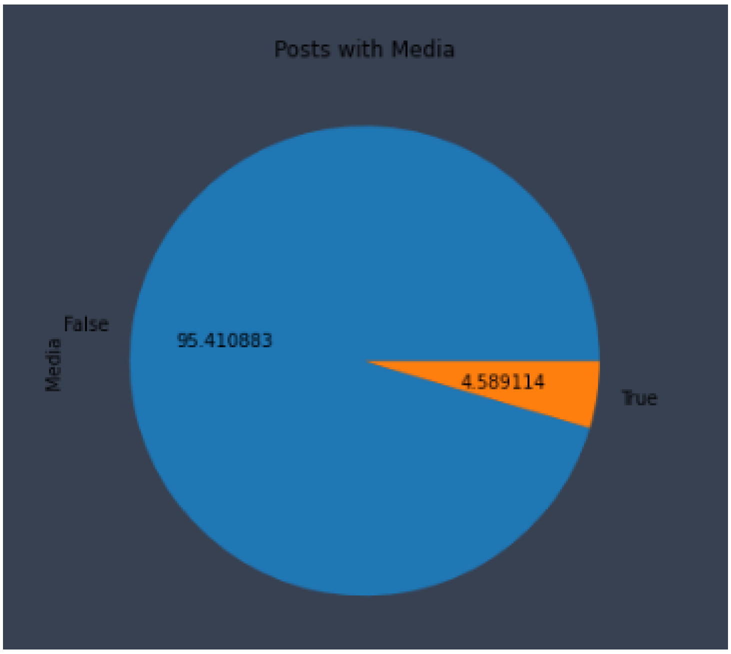


```
AskIndia          371
Politics          102
Non-Political      86
Rant               60
Policy             45
Sports             39
Business           29
Science            29
Crime              27
Culture            26
Art                25
Law                20
Health             20
```

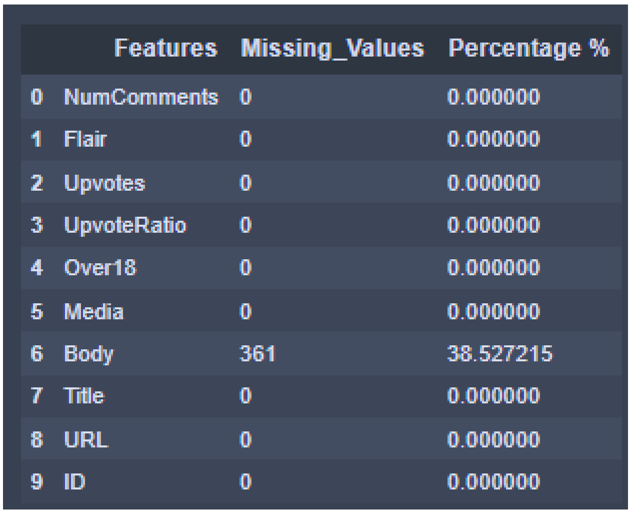
When making a post to reddit, the user has to choose a flair which suits his/her/their submission so that it is easier for the community to interact. These flairs can be a good starting point in investigating the problems people of India face and talk about online.

Posts containing Media

```
No Media    894
Media      43
```

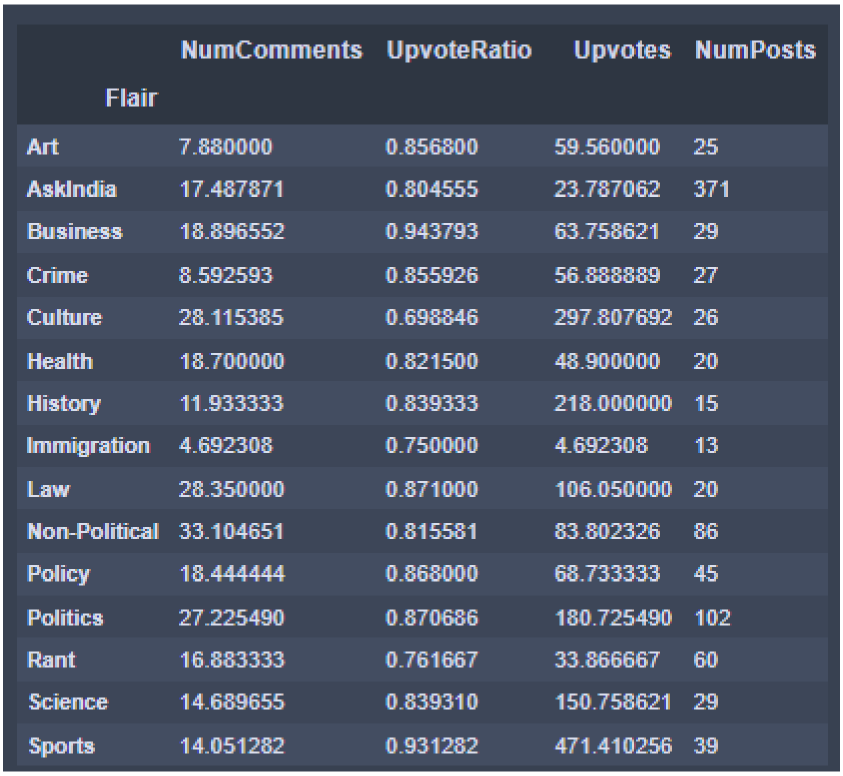


Checking for NULL values



	Features	Missing_Values	Percentage %
0	NumComments	0	0.000000
1	Flair	0	0.000000
2	Upvotes	0	0.000000
3	UpvoteRatio	0	0.000000
4	Over18	0	0.000000
5	Media	0	0.000000
6	Body	361	38.527215
7	Title	0	0.000000
8	URL	0	0.000000
9	ID	0	0.000000

Comparing features across different flairs

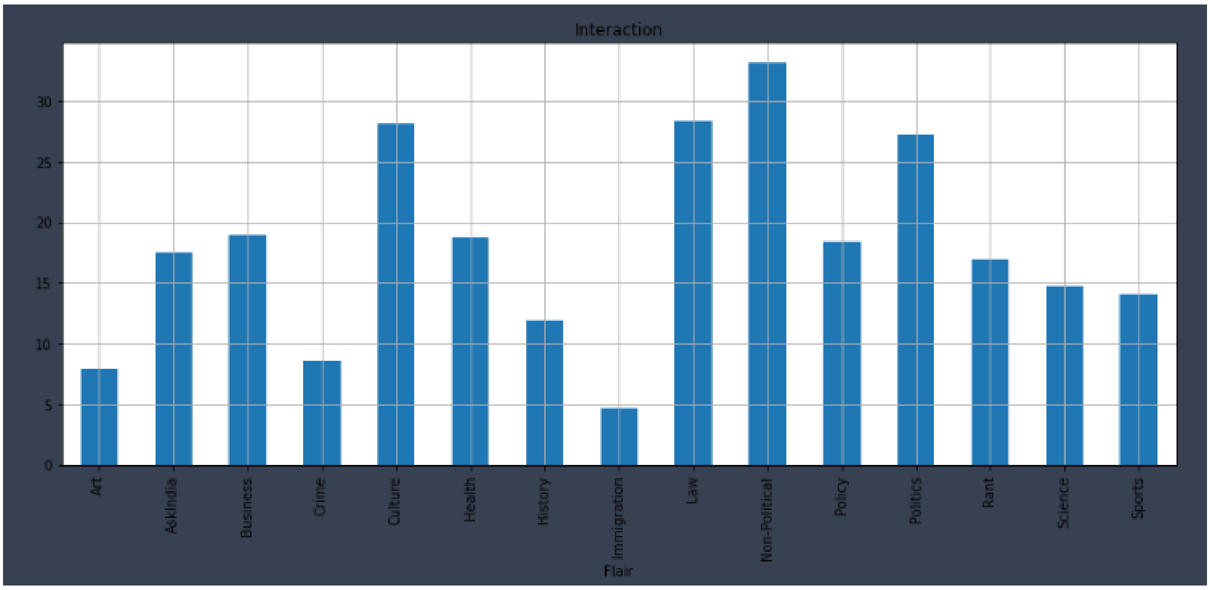


	NumComments	UpvoteRatio	Upvotes	NumPosts
Flair				
Art	7.880000	0.856800	59.560000	25
AskIndia	17.487871	0.804555	23.787062	371
Business	18.896552	0.943793	63.758621	29
Crime	8.592593	0.855926	56.888889	27
Culture	28.115385	0.698846	297.807692	26
Health	18.700000	0.821500	48.900000	20
History	11.933333	0.839333	218.000000	15
Immigration	4.692308	0.750000	4.692308	13
Law	28.350000	0.871000	106.050000	20
Non-Political	33.104651	0.815581	83.802326	86
Policy	18.444444	0.868000	68.733333	45
Politics	27.225490	0.870686	180.725490	102
Rant	16.883333	0.761667	33.866667	60
Science	14.689655	0.839310	150.758621	29
Sports	14.051282	0.931282	471.410256	39

Insights

Most Discussed Topics

Comparing flairs based on the number of comments, number of comments (specifically -> average comments for a flair) is a good feature to determine whether a problem is talked about or not since it shows the number of interactions the community has made with the post // flair.

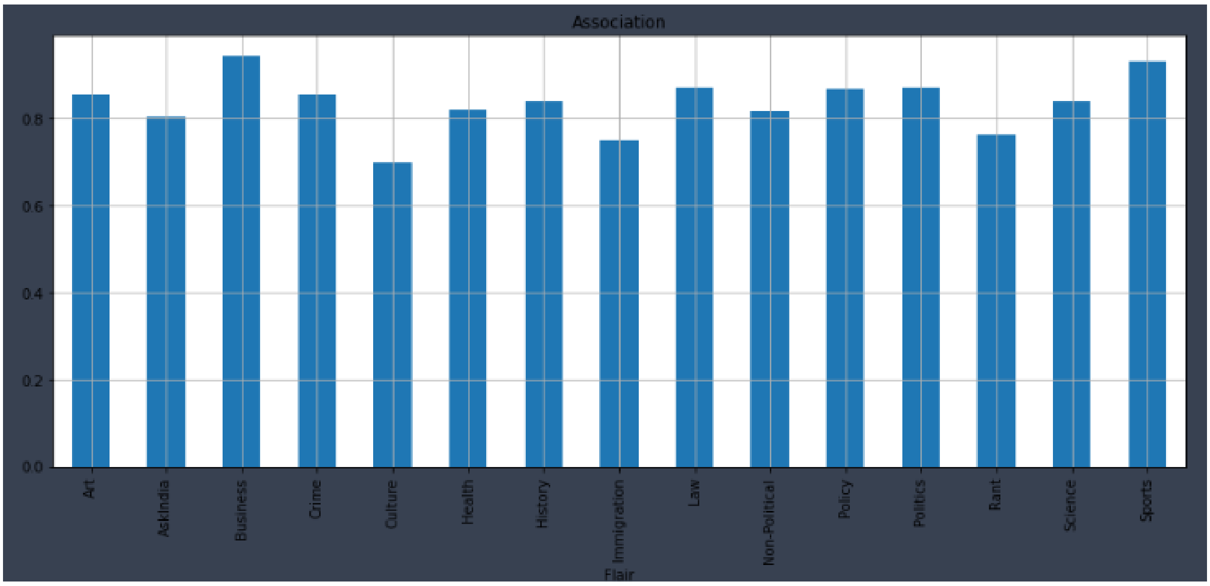


- Highly Discussed Topics : **Politics, Culture, Law, Non-Political**
- Moderately Discussed Topics : Business, Health, Policy, Rant, Science, Sports
- Less Discussed Topics : Art, Crime, Immigration

Problems : Politics, Culture, Law, Non-Politics

Most Common Problems

Upvote ratio is a good metric in determining whether the community associates with the problem others are facing. A good ratio signifies the community agrees with the flair// submission made by a certain individual.

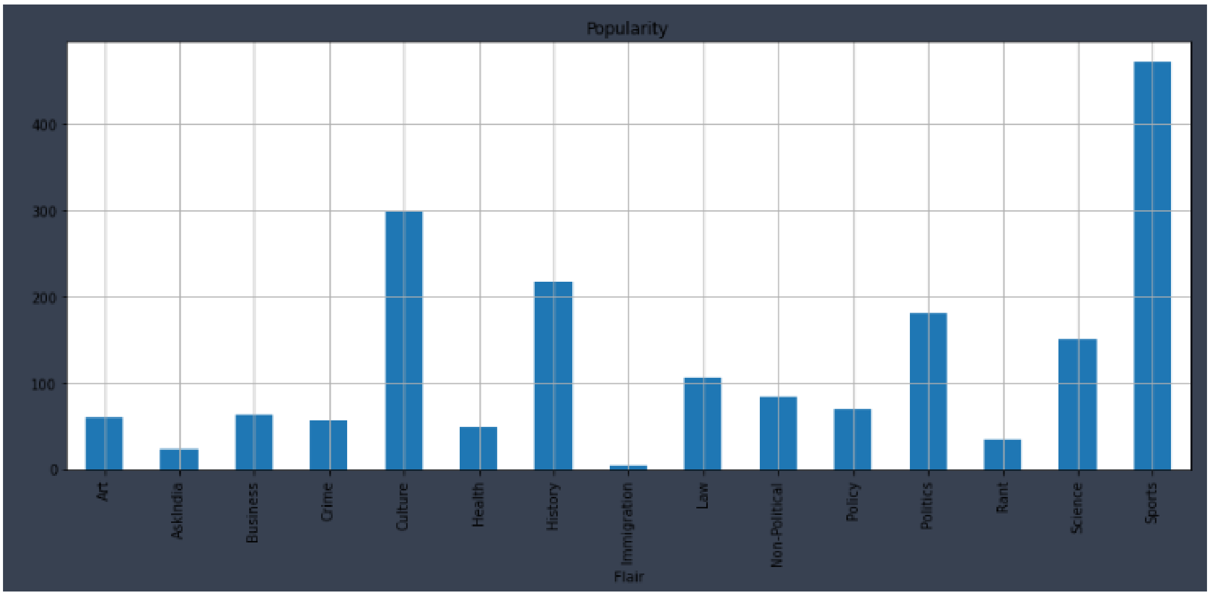


- Topics with High Association : **Business, Sports**
- Topics with Moderate Association : Art, AskIndia, Crime, Health, History, Policy, Non Political, Political, Science
- Topics with Low Association : Culture, Immigration, Rant

Problems : Business, Sports

Most Popular Problems

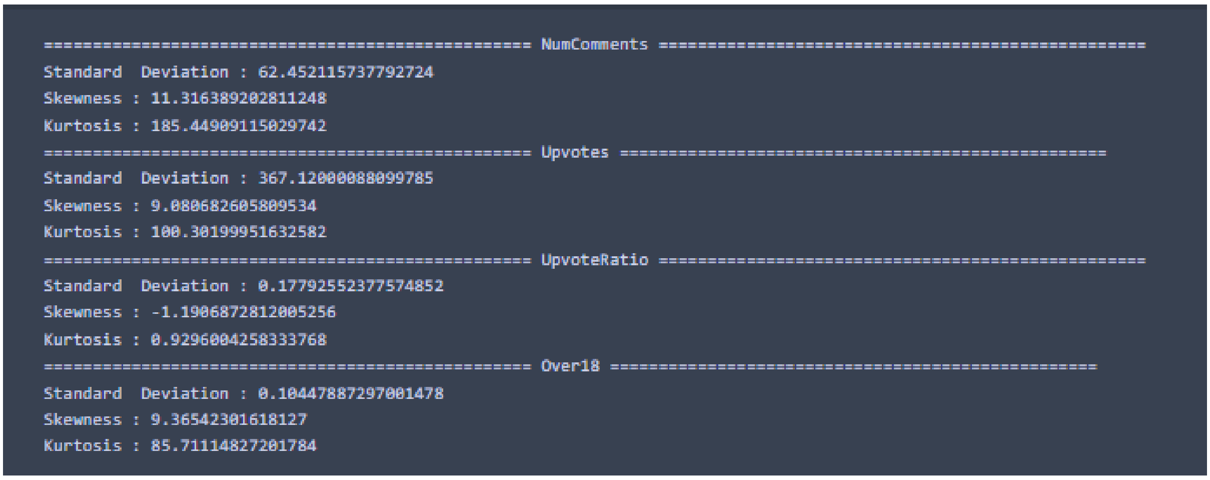
Upvote is a method of showing appreciation to a particular submission. It works similar to how likes work on facebook. A high upvote number signifies a larger population appreciating the submission. Hence it is a good metric to evaluate the popularity of submissions across different flairs.



- Most Popular Topic : **Sports**
- Moderately Popular Topics : Culture, History, Politics, Science
- Least Popular Topics : Art, AskIndia, Business, Crime, Health, Immigration, Law, Non Political, Policy, Rant

Problem : Sports

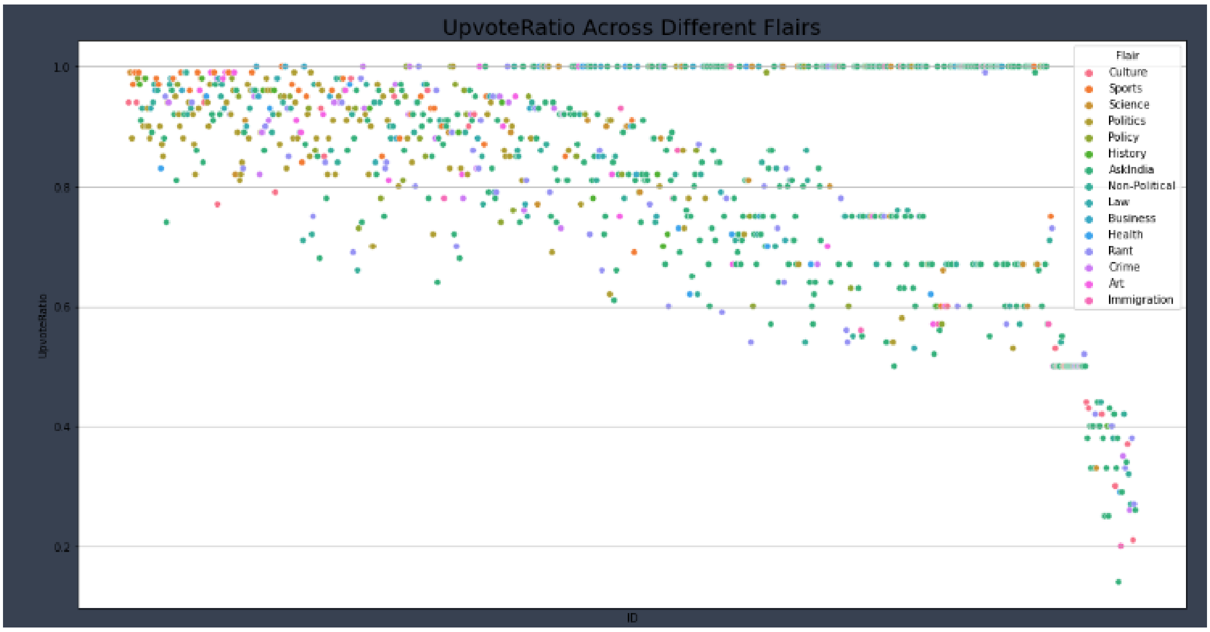
Checking outliers

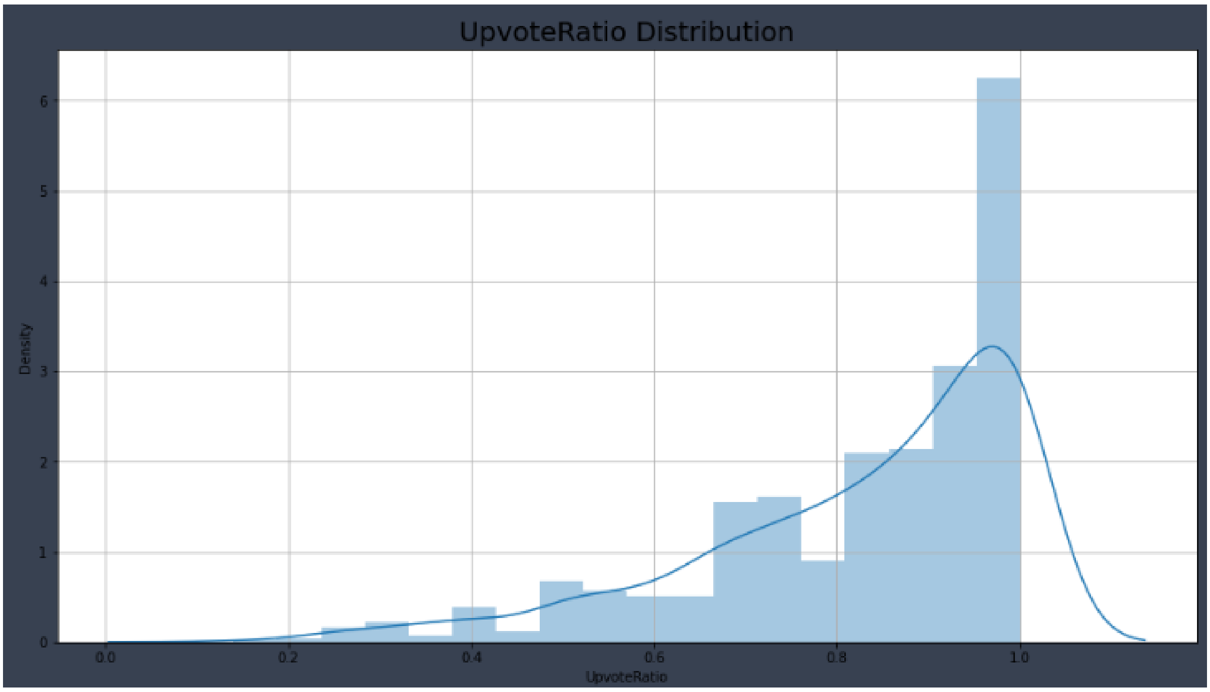


```
================================================== NumComments ==================================================
Standard  Deviation : 62.452115737792724
Skewness : 11.316389202811248
Kurtosis : 185.44909115029742
================================================== Upvotes ==================================================
Standard  Deviation : 367.12000088099785
Skewness : 9.080682605809534
Kurtosis : 100.30199951632582
================================================== UpvoteRatio ==================================================
Standard  Deviation : 0.17792552377574852
Skewness : -1.1906872812005256
Kurtosis : 0.9296004258333768
================================================== Over18 ==================================================
Standard  Deviation : 0.10447887297001478
Skewness : 9.36542301618127
Kurtosis : 85.71114827201784
```

- Number of Comments has the highest kurtosis, which means that the number of outliers present is huge, or there might less outliers but the magnitude of that is huge in comparison.
- Upvote ratio most closely resembles a normal distribution.

Analyzing Upvote Ratio





Hypothesis Testing

Students t-test

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

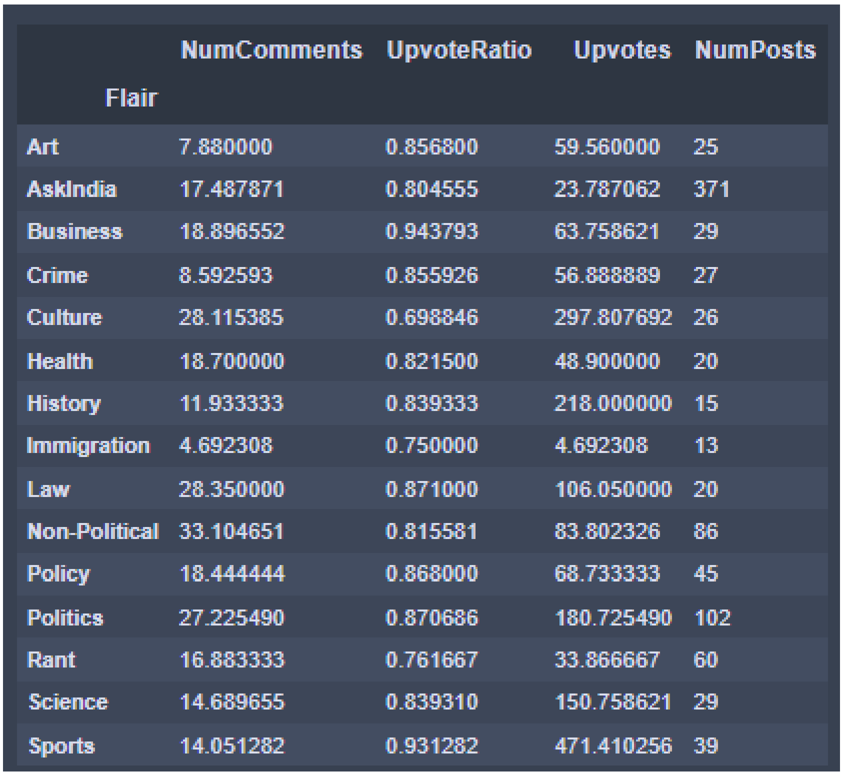
A t-value is calculated and compared to t-critical (which is different for different data). If the t value is greater than the critical value, the null hypothesis is accepted. If the critical value is greater, the null hypothesis is rejected and alternate hypothesis is accepted.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Where x → population mean
u → sample mean
s → standard Deviation
n → Number of Observations

The criteria for performing a t-test is that the sample distribution must be normal.

Using the t-test for hypothesis testing, elaborated below.



	NumComments	UpvoteRatio	Upvotes	NumPosts
Flair				
Art	7.880000	0.856800	59.560000	25
AskIndia	17.487871	0.804555	23.787062	371
Business	18.896552	0.943793	63.758621	29
Crime	8.592593	0.855926	56.888889	27
Culture	28.115385	0.698846	297.807692	26
Health	18.700000	0.821500	48.900000	20
History	11.933333	0.839333	218.000000	15
Immigration	4.692308	0.750000	4.692308	13
Law	28.350000	0.871000	106.050000	20
Non-Political	33.104651	0.815581	83.802326	86
Policy	18.444444	0.868000	68.733333	45
Politics	27.225490	0.870686	180.725490	102
Rant	16.883333	0.761667	33.866667	60
Science	14.689655	0.839310	150.758621	29
Sports	14.051282	0.931282	471.410256	39

Health has a mean of upvotes ~ 49 whereas the mean of upvotes for entire data is ~90. Since Health only has 20 observations, drawing inference for all submissions made under flair 'Health' using these 20 observations. Using student's t-test to determine and accept hypothesis.

- Null Hypothesis : the average upvotes among all posts tagged flair 'Health' is = 90.
- Alternate Hypothesis : the average upvotes for submissions tagged flair 'Health' is != 90.

Checking the type of distribution for Upvotes :

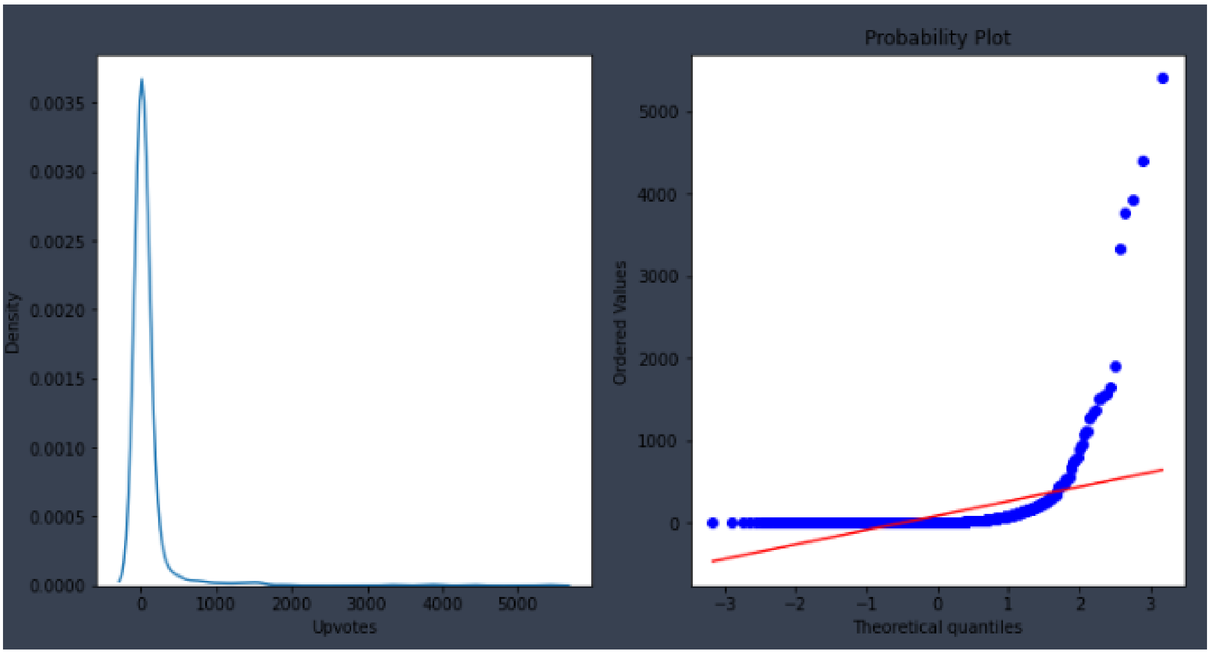


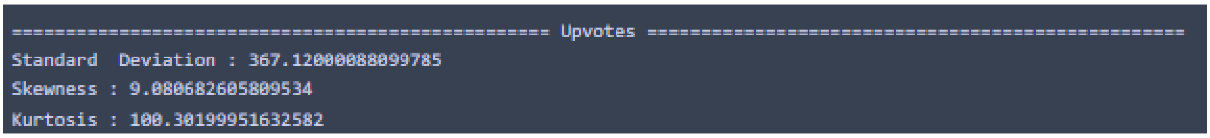
Fig a,b

Fig a is a distribution plot of all the datapoints, and Fig b is the Q-Q plot.
Both of these graphs help in concluding that the data is not normally distributed, hence performing t test on such data will yield in inaccurate results.

Normalizing the data using a $log(1+x)$ transform. The reason for choosing this over the conventional $log(x)$ transformation is due to the presence of 0 in the dataset. There are several submissions which have 0 upvotes, on applying a $log(x)$ transform, these values will get updated to NaN, which will generate issues on :

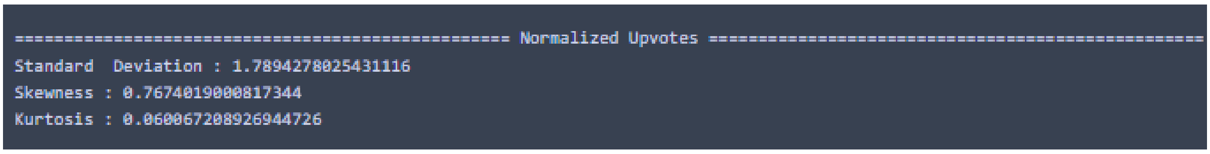
- Inverse transform
- Conducting the t test

Applying the $log(1+x)$ transform, results:



```
================================================== Upvotes ==================================================
Standard  Deviation : 367.12000088099785
Skewness : 9.080682605809534
Kurtosis : 100.30199951632582
```

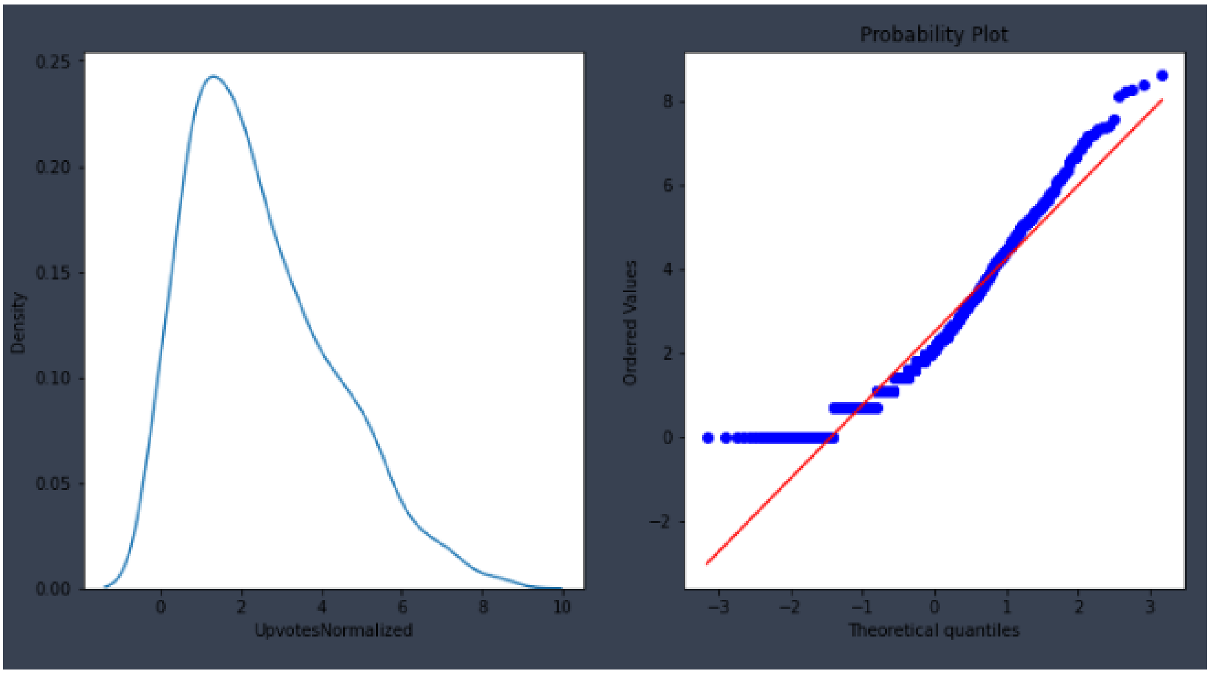
Before



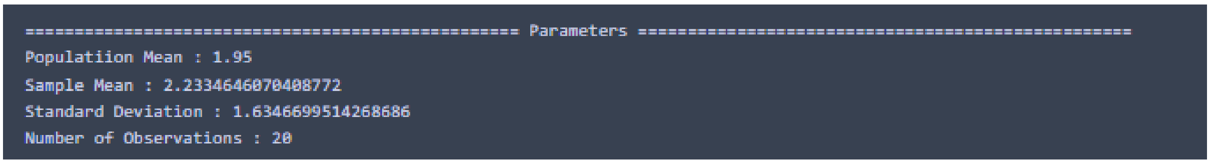
```
================================================== Normalized Upvotes ==================================================
Standard  Deviation : 1.7894278025431116
Skewness : 0.7674019000817344
Kurtosis : 0.060067208926944726
```

After

Checking distribution :



The data is now normally distributed. It is ready for performing statistical tests.



```
================================================== Parameters ==================================================
Populatiion Mean : 1.95
Sample Mean : 2.2334646070408772
Standard Deviation : 1.6346699514268686
Number of Observations : 20
```

Parameters for T-Test

Values obtained after performing calculations :

T - value : `-6.936314129795976`

Critical Value : ` -1.7291328115213678`

Since the t value of the sample is less than the t critical value , we can **reject the null hypothesis** with a 95% confidence.
Hence overall submissions which are tagged with flair 'Health' do not have an average upvote of ~90 (inverse transform from 1.95).

This type of test can be performed on each flair to check which have greater significance, even if the number of submissions is less.