Reddit posts - A Data Analysis

Joshua Sung, Data Scientist

Guiding questions

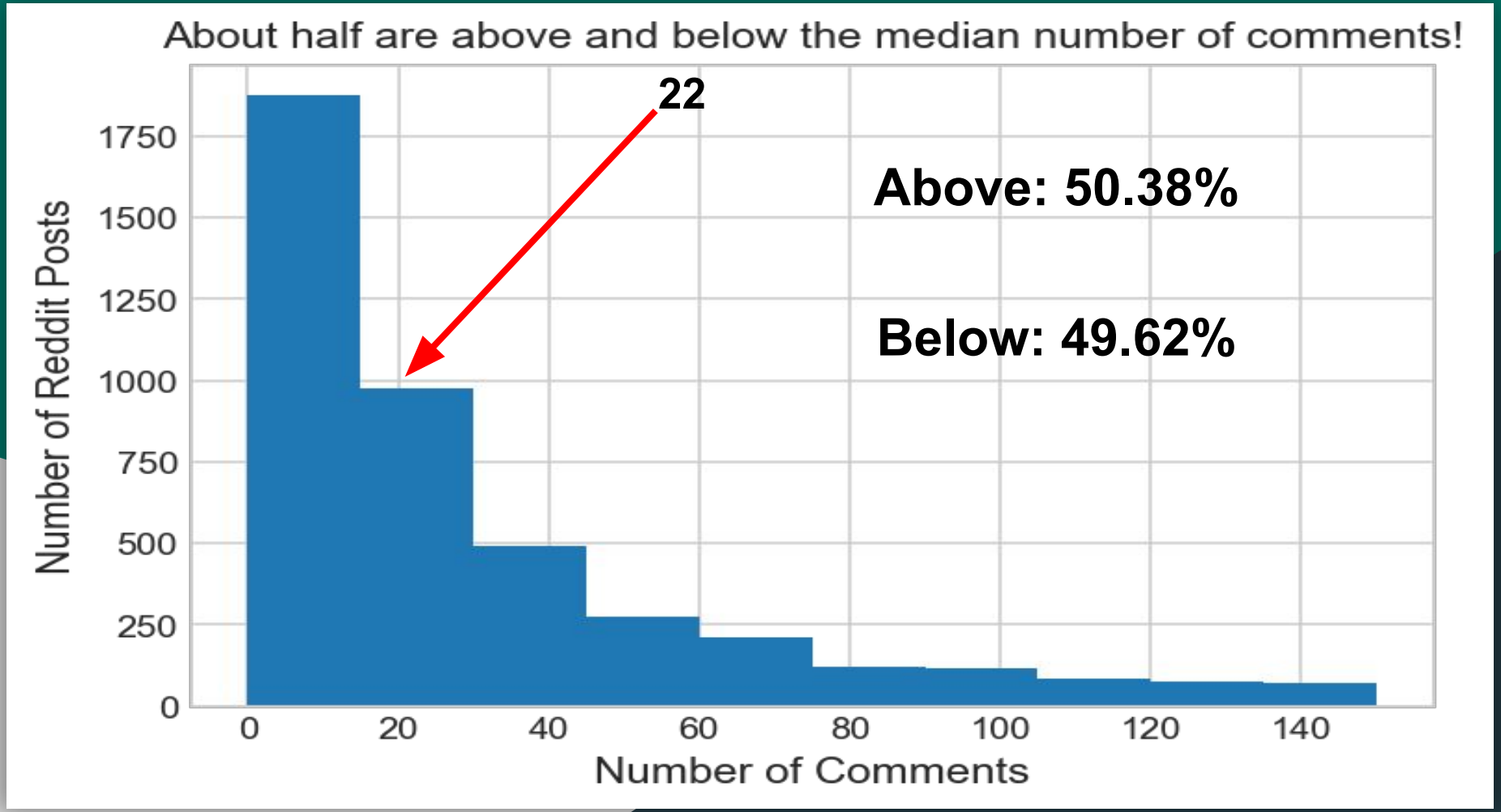
- What characteristics of a post on Reddit are most predictive of the overall interaction on a thread (as measured by number of comments)?
- Of those, which of them predict a post to be above or below the median number of comments the best?

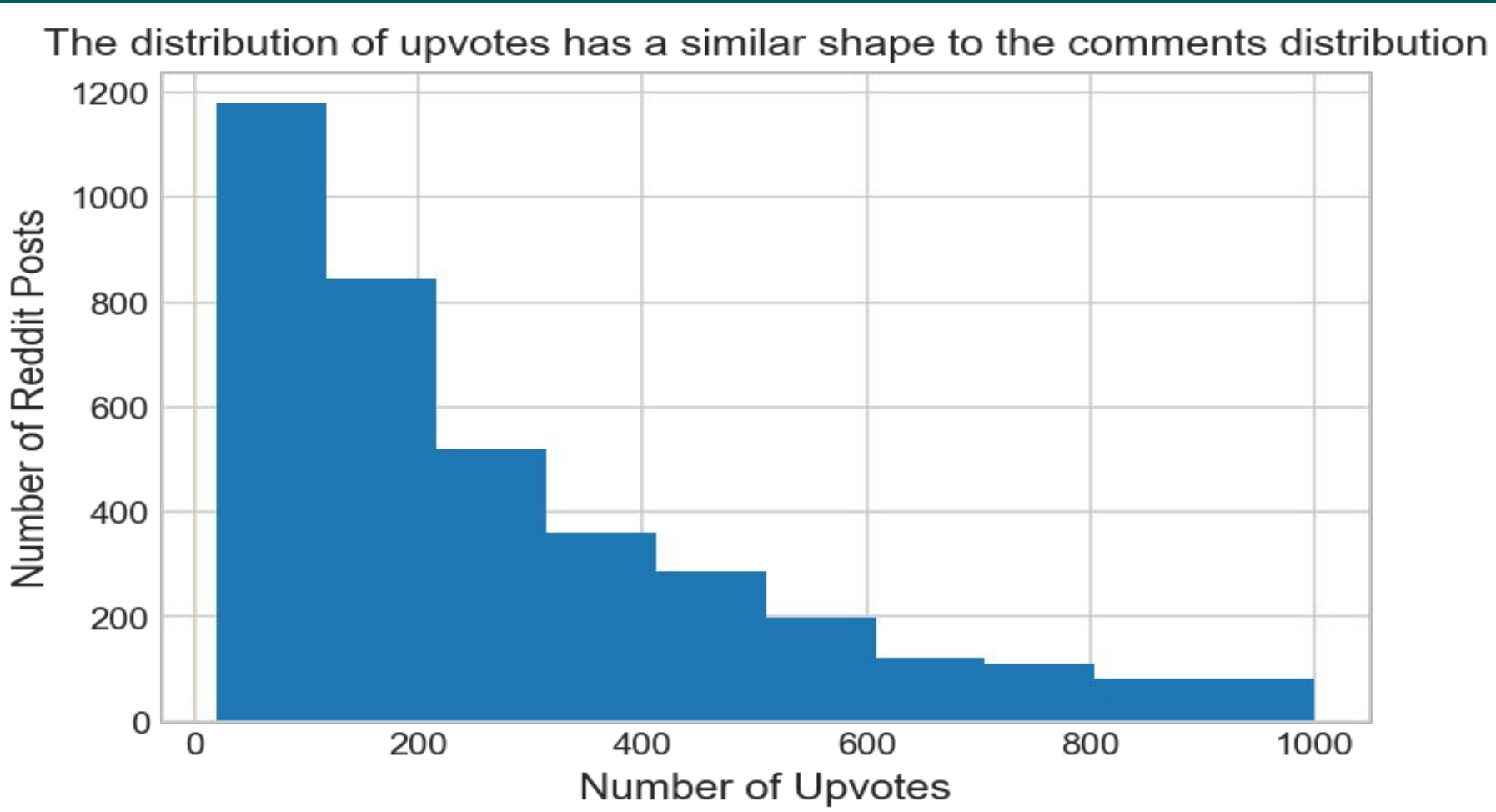
What’s in a Reddit post?

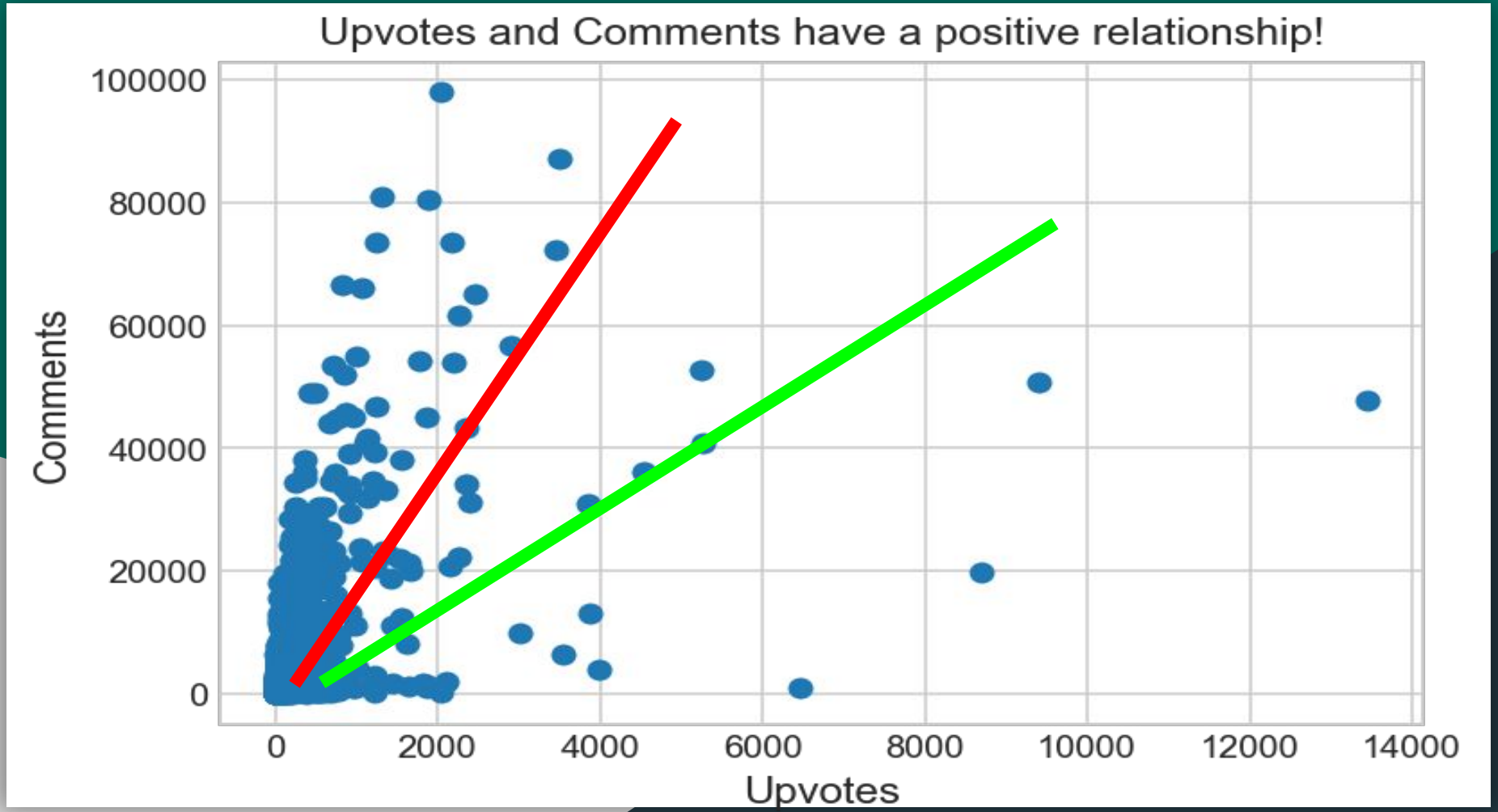
- Title
- Source (i.e. URL)
- Author
- Upvotes
- Subreddit
- Time posted

What did I collect?

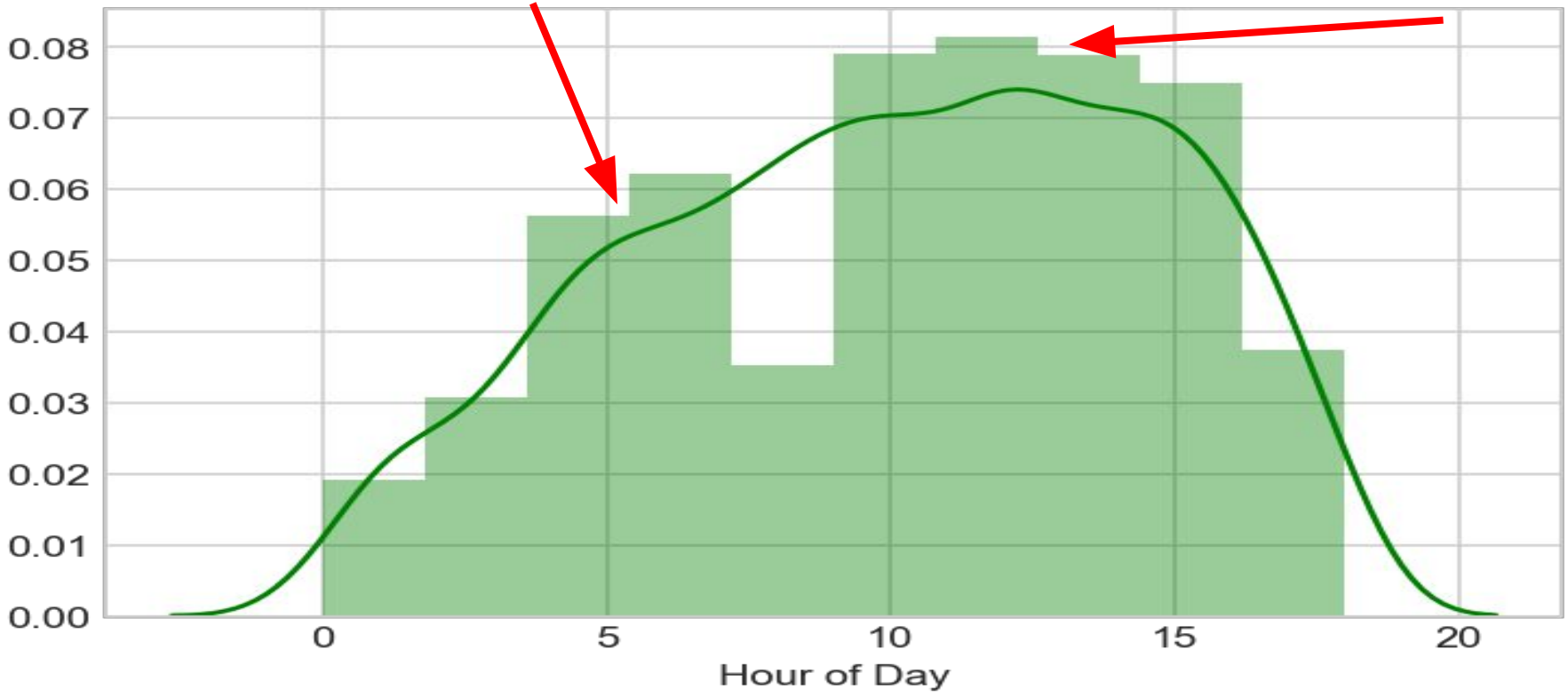
- Data scraped from Reddit.com
  - Titles
  - Number of comments
  - Time posted
  - Number of upvotes
  - Subreddit


About half are above and below the median number of comments!
22
Above: 50.38%
Below: 49.62%
Number of Reddit Posts
1750
1500
1250
1000
750
500
250
0
0
20
40
60
80
100
120
140
Number of Comments


The distribution of upvotes has a similar shape to the comments distribution
Number of Reddit Posts
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
Number of Upvotes


Upvotes and Comments have a positive relationship!
Comments
100000
80000
60000
40000
20000
0
0
2000
4000
6000
8000
10000
12000
14000
Upvotes

Most people post on Reddit during the work hours and in the early morning from 4am to 7am.



Findings

- Posting on Reddit during ‘work day hours’ and ‘morning hours’ along with the number of upvotes → 70% accuracy
- Subreddits only and words in a title not so much → 51% accuracy

Conclusion

- Collected data at a specific point in time.
  - Future data collection: over a longer period of time, different times of day, scrape less and more frequent
- Distinguishing between posts based on author, the source?
- Post during the work day, early in the morning