**Analyzing Hacker News Dataset**

In this dataset we will be analyzing the dataset from Hacker News and will try to find out the (1) frequency of published articles based on all the months of a year and also the hour of the day (2) The most used words in the headlines of he articles (3) the variation of upvotes on word count of headlines and hour of the day (4) the number of articles published for each domain. This dataset consists of four columns, the description of which is given below

* submission\_time: when the story was published
* upvotes: number of upvotes the submission got
* url: the base domain of the submission
* headline: the headline of the submission. Users can edit this, and it doesn’t have to match the headline of the article

My submission consists of five ?.py files. The description of each file is given below:

1. read.py – Contains (load\_data) function to read he .csv file (which is without header). I have imported this file in rest of thee files to build the dataframe using load\_data function
2. domains.py – File to explore the domains where most articles were submitted.
3. count.py – contains code to find out which words (& how many time) were used the most in the headlines
4. time.py – From this file I have analyzed at what hour of the day and the month of the year most articles are published
5. upvotes.py – Here I have tried to find out the variation of upvotes on the posts depending on their hour of publication and also length (number counts) of the headlines

Findings:

* **The 10 most used word in the headlines were:**

to 1448

the 1190

of 1091

for 1019

in 916

a 883

and 871

The 852

734

HN: 537

* **Top 10 domains were:**

github.com 174

techcrunch.com 172

youtube.com 142

nytimes.com 109

medium.com 88

wired.com 76

arstechnica.com 73

bbc.co.uk 53

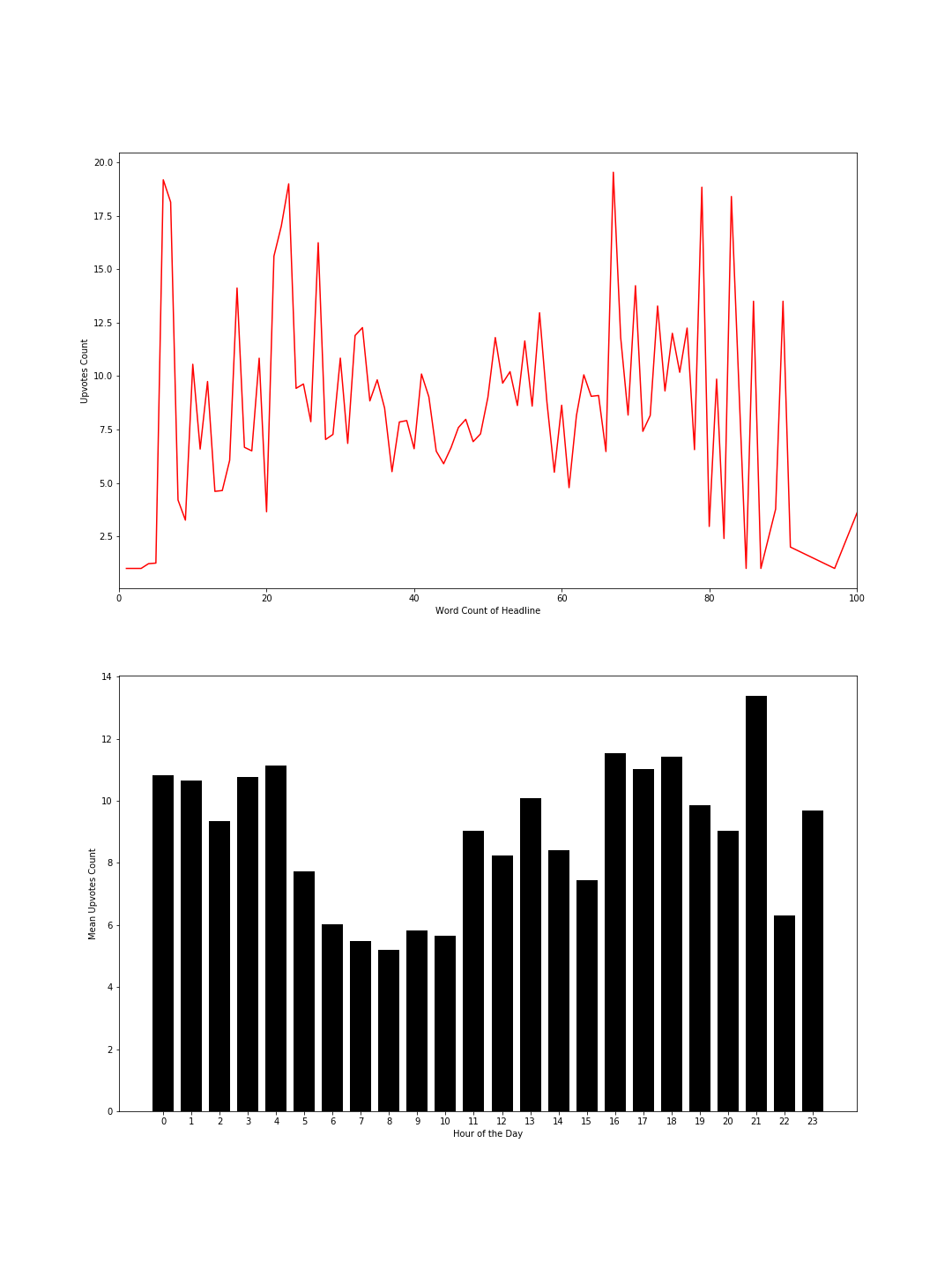
en.wikipedia.org 49

online.wsj.com 41

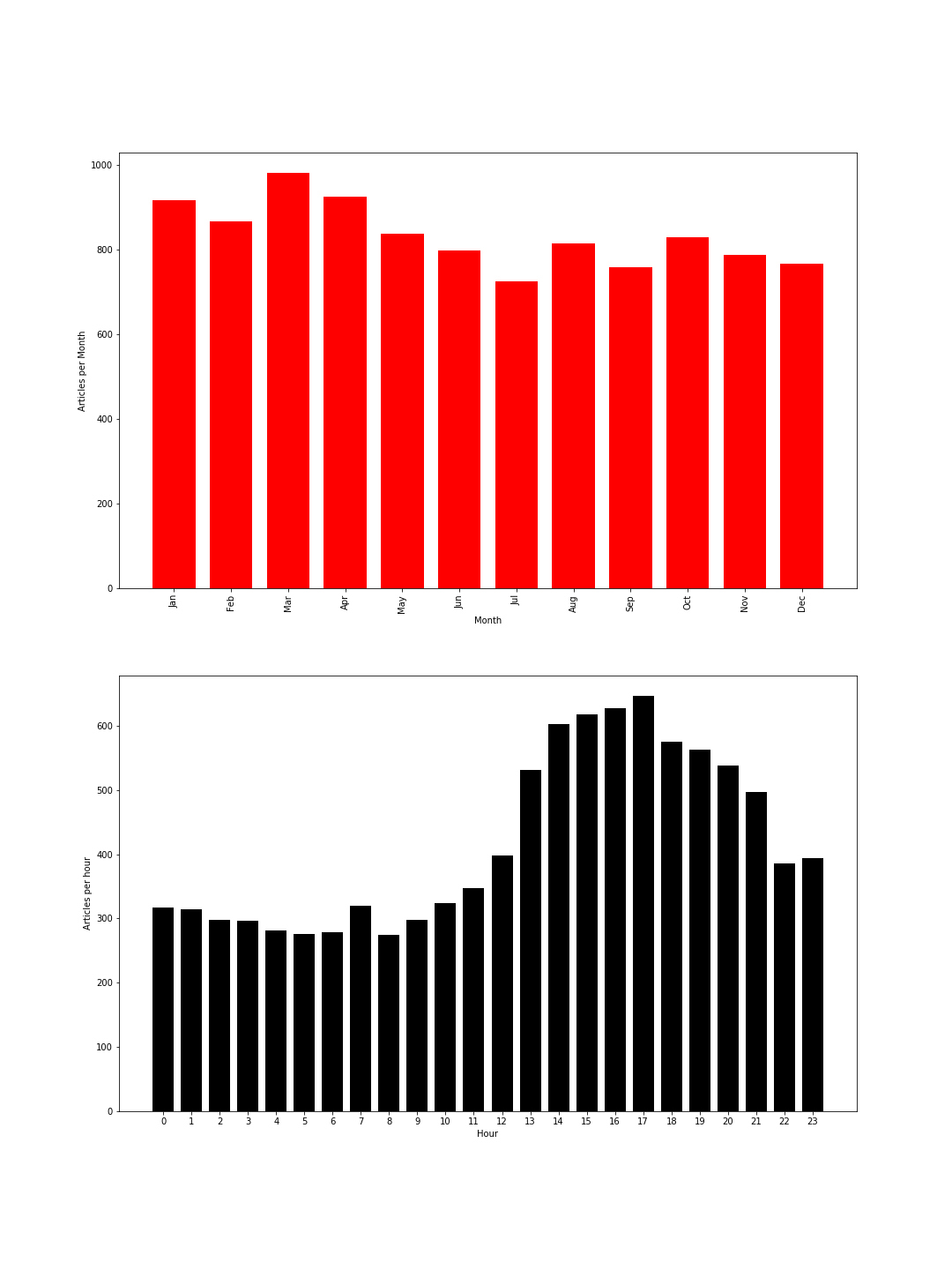
* **Variation of Upvotes:**

- The articles which got minimum amount of average likes where published at the start of the day i.e. window from 6 AM to 10 AM. I guess people published articles in that period would have office commitments or pulled an all nighter which might have impacted the quality of their articles.

- The is not much effect of the word count in articles headlines on the average upvotes



* Variation of articles submitted every month and each hour of the day:



- As you can see the most articles were submitted in the months of March;

- And From 5-6 PM of the day most articles were published