CS 363L – DATABASE SYSTEMS (LAB)

PROJECT REPORT

Presented to

Miss Darakhshan Abdul Ghaffar
University of Engineering and Technology
Lahore, Pakistan

Prepared by

2019-CE-03 (Sundas Noreen)
2019-CE-04 (Saba)

Department of Computer Engineering
University of Engineering and Technology
Lahore, Pakistan

May 09, 2022

[SPRING 2022]

Project 3 - Hacker News

PROJECT OVERVIEW

About Hacker News:

Hacker News (sometimes abbreviated as HN) is a social news website focusing on computer science and entrepreneurship. It is run by the investment fund and startup incubator Y Combinator. In general, content that can be submitted is defined as "anything that gratifies one's intellectual curiosity." The word hacker in "Hacker News" is used in its original meaning and refers to the hacker culture which consists of people who enjoy tinkering with technology.

The Hacker News (THN) is a leading, trusted, and widely recognized cybersecurity news platform that attracts over 8 million readers monthly, including IT professionals, researchers, hackers, technologists, and enthusiasts.

About the Dataset:

This dataset contains all stories and comments from Hacker News from its launch in 2006. Each story contains a story id, the author that made the post, when it was written, and the number of points the story received. The dataset is available on big query with the name *bigquery-public-data.hacker_news* and is of 19.2132 GB.

Questions:

1. Hacker News would like to send awards to prolific commentors i.e., everyone who has written more than 10,000 posts. How to find that?
2. Recent studies have found that many forums tend to be dominated by a very small fraction of users. Is this true of Hacker News?
3. Hacker News has received complaints that the site is biased towards Y Combinator startups. Do the data support this?

ANALYSIS OF DATASET

This dataset contains all stories and comments from Hacker News from its launch in 2006. Each story contains a story id, the author that made the post, when it was written, and the number of points the story received.

There are 4 tables in this public dataset:

- stories (a.k.a posts)
  Primary Key: ID
- comments
  Primary Key: ID
  Foreign Key from the same table: Parent

- full
  Primary Key: ID
- full_2015
  Primary Key: ID

Other basic points are:

- A post can have many comments.
- A comment can have many comments.
- Posts can have scores.
- Posts have rankings showing their popularity.

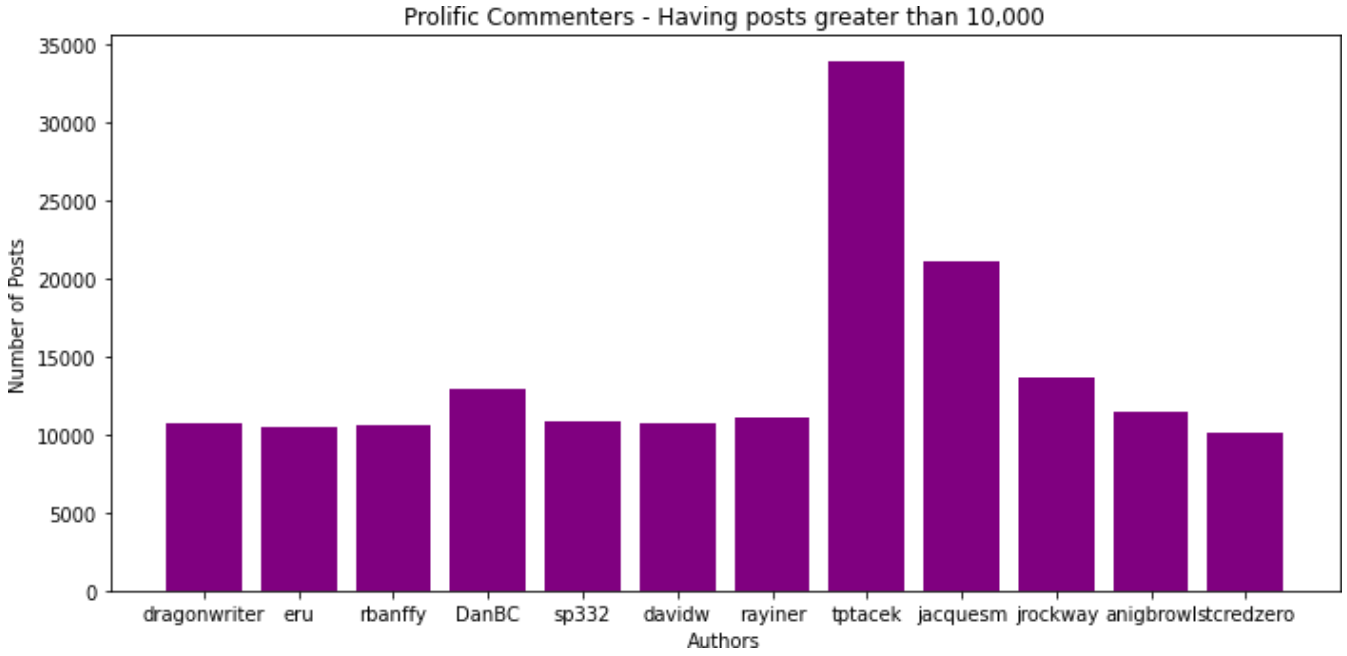
EXPLORING QUESTIONS WITH VISUALIZATIONS

Question 1:

Hacker News would like to send awards to prolific commentors i.e., everyone who has written more than 10,000 posts. How to find that?

Visualizations:

For this, lets query all authors with more than 10,000 posts as well as their post counts. Here are the visualizations of the results:



So, see how easy it is for Hacker News now to maintain it's success and appreciate the prolific commentors.

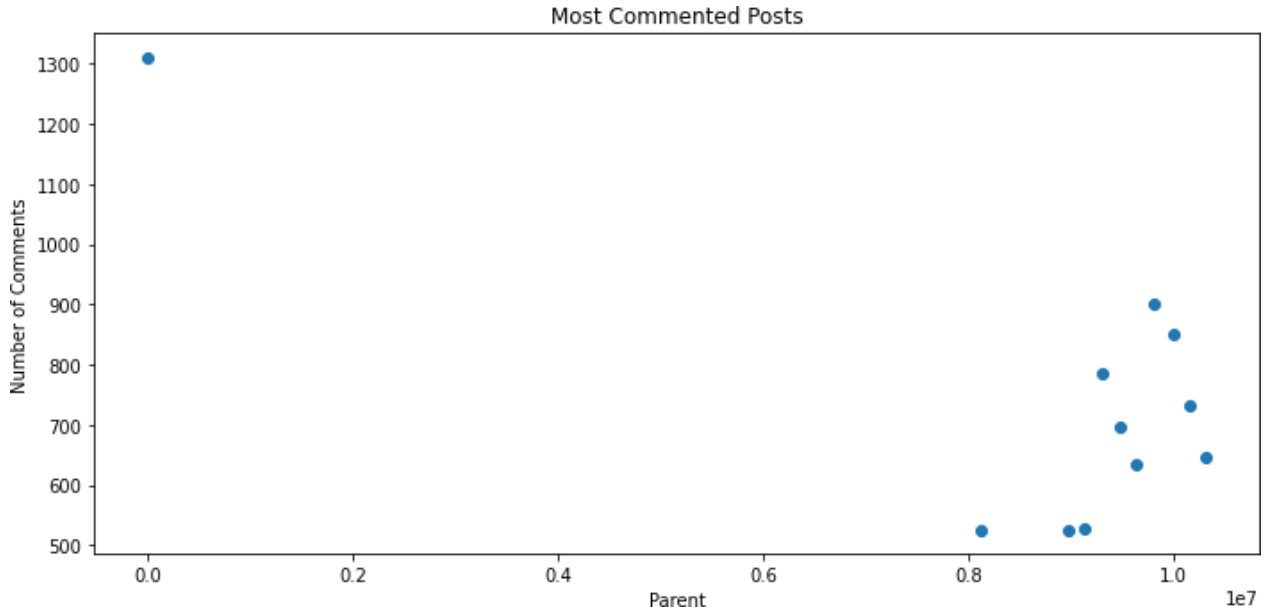
Question 2:

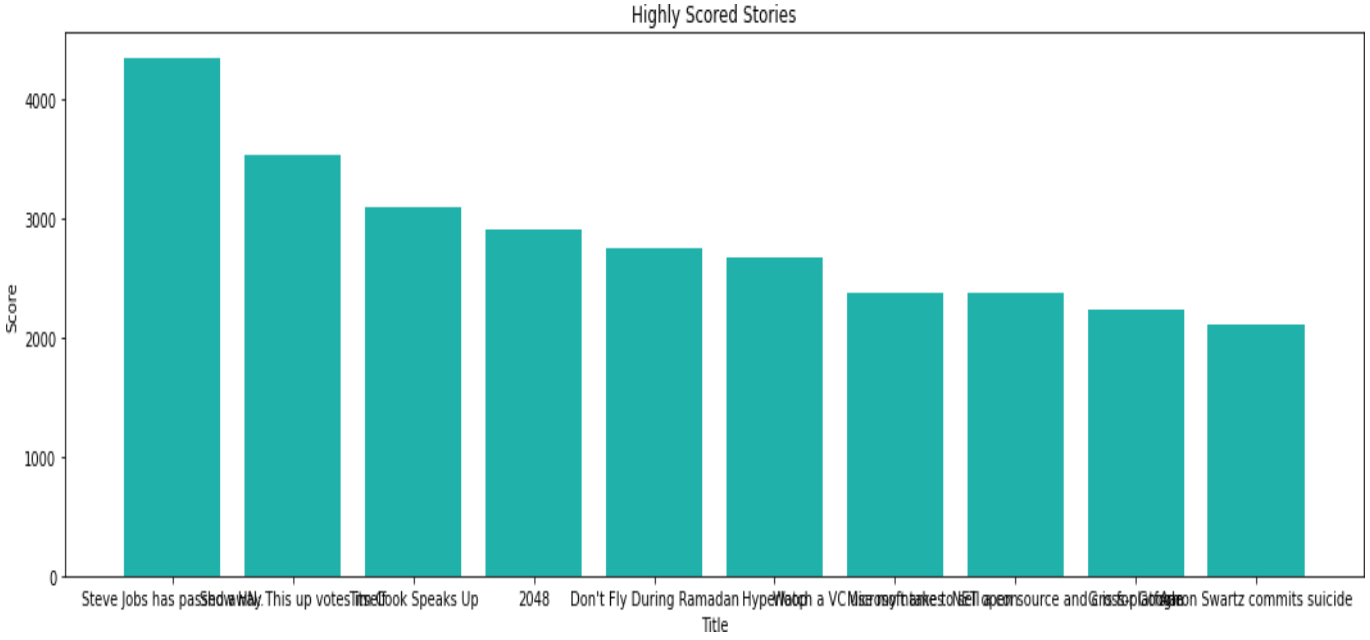
Recent studies have found that many forums tend to be dominated by a very small fraction of users. Is this true of Hacker News?

Visualizations:

For this, lets query the most commented posts and highly scored stories. It will help in finding the actual number of scores and comments on a story.

Here are the visualizations of the results:





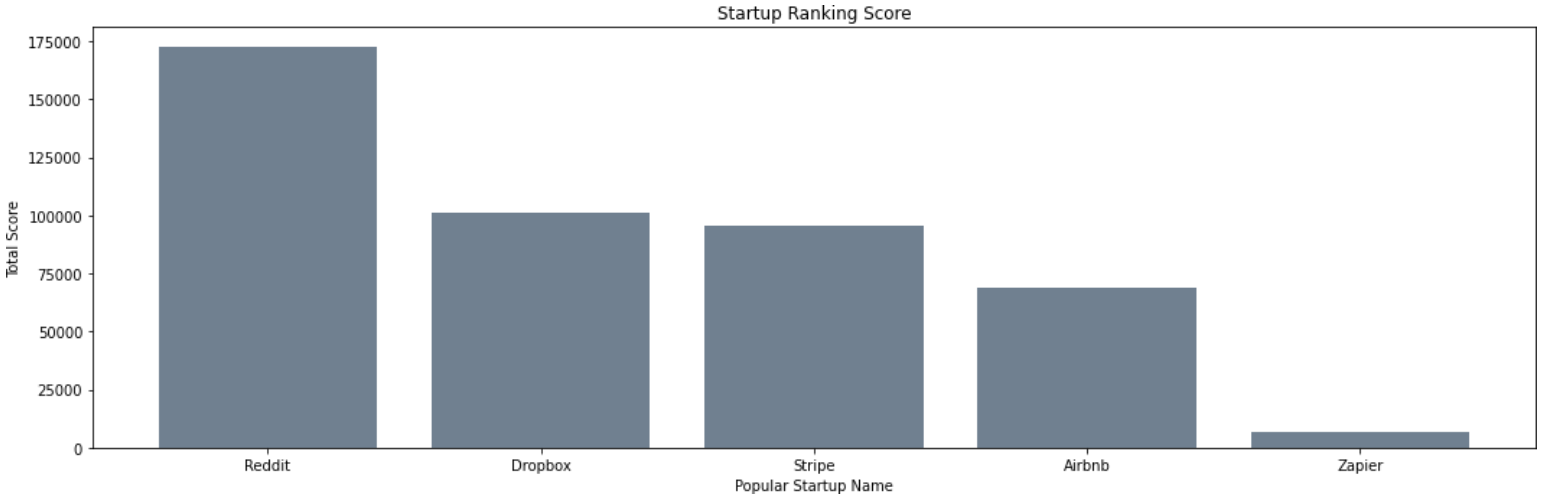
The first chart shows that the maximum number of comments on a post are 1300, and the maximum score of a story is 4000. These are quite a large number proving that Hacker News is not dominated by a small fraction of users.

Question 3:

Hacker News has received complaints that the site is biased towards Y Combinator startups. Do the data support this?

Visualizations:

For this, lets query the most popular Y combinator startup's' score to check their pattern. Here are the visualizations of the results:



The above chart shows that for some Y combinator startups, the score is very high while for some the score is very low. If Hacker News is biased towards Y combinator startups, the score should have been high for all startups while this is not the case here.

DATA PREDICTION

In Hacker News, there are different classes of readers. All readers are there to read some specific type of stories relating to some particular topics and trends.

So, we have designed a machine learning model that will predict the score from the title of the story. The title of the story may have some specific words and terms that attracts the users which increases the score of the story. Hence, attractive title enhances a story.

With our model, we have predicted the score of the stories that contain the word *security* in it.

	predicted_label	score	feature
0	3.215563	3	Ask HN: Need your input on an idea in the real...
1	4.710915	4	Tell HN: Unfuddle (svn hosting) has bad passwo...
2	3.215563	5	Download HD Instagram image bypassing instagra...
3	3.215563	3	A Civil Perspective on Cybersecurity
4	4.190915	3	Introduction to API(REST) security
5	4.190915	3	Research on poker a good deal for airport secu...
6	4.190915	3	Hacker news for security
7	4.190915	3	8-Year-Old CEO Reuben Paul Proves That Kids Ar...
8	3.670915	3	Beginner's Guide to API(REST) security
9	4.190915	3	Test/grade your email server security
10	4.710915	4	NASA surplus computer sale compromised data se...
11	4.970915	5	How hackers gave Subway a $30 million lesson i...
12	5.230915	5	Experts on Mac vs. PC security
13	3.215563	3	Going to the doctor and worrying about cyberse...
14	3.215563	3	The terahertz revolution and local security
15	4.190915	3	Keystroke fingerprinting - using Javascript to...
16	3.215563	3	Cyber War: Microsoft a weak link in national s...
17	4.190915	3	Booby-trapping PDF files: A new how-to; Built-...
18	3.215563	3	Schneier hacking airport security
19	3.215563	3	Central Asian Cybersecurity

CONCLUSION

With our data exploration, we have shown that the Hacker News can easily appreciate the prolific commentors and that Hacker News is not dominated by a small fraction of users. It has a huge audience who read its stories and comments on them. Moreover, we have predicted the score of the story from its title indicating that a particular class of readers is there to read a particular sort of stories. Also, we have explored that Hacker News is not biased towards Y combinator startups.