Project Report

01.

Analyze Hacker News Trends

Analyze Data with SQL

project.sqlite

+91 906 Data Analyst Project | June 2024 © Karan Chandra Dey

Technology used: sqlite

02.

Key areas of Analysis

- Which sites feed Hacker News?
- What's the best time to post a story?

03.

Hacker News Profile

Hacker News is a popular website run by Y Combinator. It's widely known by people in the tech industry as a community site for sharing news, showing off projects, asking questions, among other things.

@ Karan Kapoor

Website Overview

04.

Hacker News Screenshot

This data was kindly made publicly available under the MIT licenses.

07.

Data and Table Columns

Data Overview

In this project, we will be working with a table named hacker_news that contains stories from Hacker News since its launch in 2007. It has the following columns:

- title: the title of the story
- user: the user who submitted the story
- score: the score of the story
- timestamp: the time of the story
- url: the link of the story

Database Schema

hacker_news

name	type
title	TEXT
user	TEXT
score	INTEGER
timestamp	DATETIME
url	TEXT

Rows: 4000

05.

Table of contents

01 Topic one — What are the top five stories with the highest scores?

02 Topic two — Is a small percentage of Hacker News submitters taking the majority of the points?

03 Topic three — We want to add these users' scores together and divide by the total to get the percentage.

04 Topic four — Is Hacker News dominated by these users?

05 Topic five — We want to pinpoint the users who have accumulated a lot of points across their stories.

06 Topic six — What sites feed Hacker News? (Mostly GitHub, Medium, or New York Times?)

07 Topic seven — What is the best time of the day to post a story on Hacker News?

08 Topic eight — What is the best time to post a story on Hacker News?

09 Topic ten — CONCLUSION

01.

What are the top five stories with the highest scores?

SQL Query

```
SELECT title, score
FROM hacker_news
ORDER BY score DESC
LIMIT 5;
```

title	score
Temu Annals - Surface Pro 3 update	1000
Hacking The Status Game	318
Prettiges CSS with autocompletion and syntax highlighting	309
Steam Fri by Hits out at 'infantile' culture of trigger words and safe spaces	297
'Reserve' Australian Court pays £2M' also standard cover Microsoft	292

Observation:

The most popular stories are:
Temu Annals - Surface Pro 3 update
Hacking The Status Game
Prettiges CSS with autocompletion and syntax highlighting
Steam Fri by Hits out at 'infantile' culture of trigger words and safe spaces
'Reserve' Australian Court pays £2M' also standard cover Microsoft

02.

Is a small percentage of Hacker News submitters taking the majority of the points?

sum(score)
6360

SQL Query

```
SELECT SUM(score)
FROM hacker_news;
```

Observation:

The total score of this table is 6360.

The "90–9–1" version of this rule states that for websites where users can both create and edit content, 1% of people create content, 9% edit or modify that content, and 90% view the content contributing. However, the actual percentage is likely to vary depending upon the subject.

90-9-1 Rule

03.

we want to add these users' scores together and divide by the total to get the percentage.

$$(517 + 309 + 304 + 282) / 6360.0$$
$$0.221803330191643$$

SQL Query

```
SELECT (517 + 309 + 304 + 282) / 6360.0;
```

Observation:

These 4 users have a combined **22% of the total scores** in the table.

04.

we need to pinpoint the users who have accumulated a lot of points across their stories.

user	sum(score)
rexpetersen	517
ancilotta	309
mikeyfarrell	304
mehulkumar	282

SQL Query

```
SELECT user, SUM(score)
FROM hacker_news
GROUP BY user
HAVING SUM(score) > 200
ORDER BY 2 DESC;
```

Observation:

List of individual users who have gotten combined scores of more than 200, and their combined scores.

05.

Rickrolling

SQL Query

```
-- Hacker News Moderating

SELECT user,
COUNT(*)
FROM hacker_news
WHERE url LIKE '%watch?v=dQw4w9WgXcQ%'
GROUP BY 1
ORDER BY 2 DESC;
```

While we are looking at the power users, some users are rickrolling — linking readers into clicking on a link to a funny video and claiming that it has to information about coding.

The url of the video is:

https://www.youtube.com/watch?v=dQw4w9WgXcQ

user	count(*)
rickroll_bot	1
rickrollmod	1

Observation:

@sonnyomnom, you're banned.
@scorpicaster, warning!

Which sites feed Hacker News?

Hacker News stories are essentially links that take users to other websites.
Which of these sites feed Hacker News the most: GitHub, Medium, or New York Times?

SQL Query

```
SELECT CASE
WHEN url LIKE '%github.com%' THEN 'GitHub'
WHEN url LIKE '%medium.com%' THEN 'Medium'
WHEN url LIKE '%nytimes.com%' THEN 'New York Times'
ELSE 'Other'
END AS 'Source'
FROM hacker_news;

-- Which sites feed Hacker News?
SELECT Source,
WHEN url LIKE '%github.com%' THEN 'GitHub'
WHEN url LIKE '%medium.com%' THEN 'Medium'
WHEN url LIKE '%nytimes.com%' THEN 'New York Times'
ELSE 'Other'
END AS 'Source',
COUNT(*) FROM hacker_news
GROUP BY 1;
```

Source	COUNT(*)
GitHub	23
Medium	12
New York Times	13
Other	3952

The number of times stories are linked to:

GitHub - 23
Medium - 12
New York Times - 13

Every submitter wants their story to get a high score so that the story makes it to the front page, but...

What's the best time of the day to post a story on Hacker News?

SQL Query

```
SELECT timestamp,
FROM hacker_news
LIMIT 20;

SELECT timestamp,
-- strftime('%H', timestamp)
FROM hacker_news
GROUP BY 1
LIMIT 20;
```

timestamp	strftime('%H', timestamp)	timestamp1	count(*)
2007-01-02 15:16:52	15	2007-01-01 00:00:00	1
2007-01-01 07:03:24	07	2007-01-01 01:00:00	1
2007-01-07 03:11:17	03	2007-01-01 02:00:00	1
2007-01-01 18:00:00	18	2007-01-01 03:00:00	1
2007-01-17 05:49:52	05	2007-01-01 05:00:00	1

```
SELECT strftime('%H', timestamp),
AVG(score),
COUNT(*)
FROM hacker_news
GROUP BY 1
ORDER BY 2 DESC;
```

strftime('%H', timestamp)	avg(score)	count(*)
13	18.0000000000000	166
12	21.5555555555556	144
15	20.4579439252336	238
14	18.8333333333333	216
16	16.2051449275362	221
11	17.05	120
10	9.95555555555556	117

08.

What's the best time to post a story?

```
SELECT strftime('%H', timestamp) AS 'Hour',
ROUND(AVG(score), 1) AS 'Average Score',
COUNT(*) AS 'Number of Stories'
FROM hacker_news
-- Ignore missing IS NOT NULL
WHERE timestamp
ORDER BY 2 DESC;
```

Timestamp	Average	Count
13	27.0	206
07	21.0	104
19	20.0	238
20	18.0	238
12	14.0	123
09	11.0	119
18	10.0	237

The best hours are in the morning around 7 am and afternoon around 6 - 8 pm! Monster difference!

09.

Conclusion

This analysis provided valuable insights into trends and user behaviors on Hacker News. Here's a summary of the key findings:

Popular Stories: The analysis identified the top five stories based on their scores, providing a glimpse into the topics that resonated most with users.

Power Users: A small percentage of users contributed a significant portion of the total score, suggesting a potential for community-driven efforts.

Rickrolling: The analysis identified users who repeatedly posted a link to a rickroll video, highlighting the need for potential moderation.

Popular Sites: The analysis revealed that GitHub, Medium, and New York Times were the most popular sites, with GitHub, Medium, and New York Times being the most active sites feed the most number of Hacker News.

Best Time to Post: The analysis identified the optimal time to post a story on Hacker News, suggesting that mornings and evenings might be better times to maximize visibility.

These findings offer valuable guidance for anyone interested in understanding Hacker News' dynamics. The insights can be used to attract users, by leveraging this information, users can tailor their content and strategies for different groups to maximize their reach and impact on the platform.

© Karan Chandra Dey