Exploring Hacker News Posts:

Do Ask HN or Show HN receive more comments?

Hacker News (https://google.at) is a site where users can submitt posts which are voted on and commented upon by other users. I will focus on two genres of posts from Hacker News: Ask HN and Show HN. Ask HN are posts which ask the Hacker News community specific questions while Show HN submissions display user projects to the community. The top Hacker News posts which have the most comments can get hundreds of thousands of visitors to their sites.

In this project I will analayze a data set (https://www.kaggle.com/hacker-news/hacker-news-posts) from kaggle to determine whether Ask HN or Show HN posts recieve more comments on average. The data set has been reduced from ~3000,000 rows to ~20,000 rows by removing all submissions which did not receive any comments.

In [1]:

```
import csv
from csv import reader
```

In [2]:

```
opened_file=open('hacker_news.csv')
read_file=reader(opened_file)
hn=list(read_file)
```

Data Cleaning:

In [3]:

```
for row in hn[:6]:
    print(row)
    print("\n")
```

```
['id', 'title', 'url', 'num_points', 'num_comments', 'author', 'crea
ted_at']


['12224879', 'Interactive Dynamic Video', 'http://www.interactivedyn
amicvideo.com/', '386', '52', 'ne0phyte', '8/4/2016 11:52']


['10975351', 'How to Use Open Source and Shut the Fuck Up at the Sam
e Time', 'http://hueniverse.com/2016/01/26/how-to-use-open-source-an
d-shut-the-fuck-up-at-the-same-time/', '39', '10', 'josep2', '1/26/2
016 19:30']


['11964716', "Florida DJs May Face Felony for April Fools' Water Jok
e", 'http://www.thewire.com/entertainment/2013/04/florida-djs-april-
fools-water-joke/63798/', '2', '1', 'vezycash', '6/23/2016 22:20']


['11919867', 'Technology ventures: From Idea to Enterprise', 'https:
//www.amazon.com/Technology-Ventures-Enterprise-Thomas-Byers/dp/0073
523429', '3', '1', 'hswarna', '6/17/2016 0:01']


['10301696', 'Note by Note: The Making of Steinway L1037 (2007)', 'h
ttp://www.nytimes.com/2007/11/07/movies/07stein.html?_r=0', '8', '2'
, 'walterbell', '9/30/2015 4:12']
```

In [4]:

```
#remove header row from dataset

headers=hn[0]
hn=hn[1:]

print(headers)
```

```
['id', 'title', 'url', 'num_points', 'num_comments', 'author', 'crea
ted_at']
```

In [5]:

```
for row in hn[:5]:
    print(row)
    print("\n")
```

```
['12224879', 'Interactive Dynamic Video', 'http://www.interactivedyn
amicvideo.com/', '386', '52', 'ne0phyte', '8/4/2016 11:52']


['10975351', 'How to Use Open Source and Shut the Fuck Up at the Sam
e Time', 'http://hueniverse.com/2016/01/26/how-to-use-open-source-an
d-shut-the-fuck-up-at-the-same-time/', '39', '10', 'josep2', '1/26/2
016 19:30']


['11964716', "Florida DJs May Face Felony for April Fools' Water Jok
e", 'http://www.thewire.com/entertainment/2013/04/florida-djs-april-
fools-water-joke/63798/', '2', '1', 'vezycash', '6/23/2016 22:20']


['11919867', 'Technology ventures: From Idea to Enterprise', 'https:
//www.amazon.com/Technology-Ventures-Enterprise-Thomas-Byers/dp/0073
523429', '3', '1', 'hswarna', '6/17/2016 0:01']


['10301696', 'Note by Note: The Making of Steinway L1037 (2007)', 'h
ttp://www.nytimes.com/2007/11/07/movies/07stein.html?_r=0', '8', '2'
, 'walterbell', '9/30/2015 4:12']
```

In [12]:

```
ask_posts=[]
show_posts=[]
other_posts=[]

for row in hn:
    title=row[1]
    title_lower= title.lower()
    if title_lower.startswith('ask hn'):
        ask_posts.append(row)
    elif title_lower.startswith('show hn'):
        show_posts.append(row)
    else:
        other_posts.append(row)

print("length Ask HN: ", len(ask_posts))
print("length Show HN: ", len(show_posts))
print("length other: ", len(other_posts))
```

```
length Ask HN:  1744
length Show HN:  1162
length other:  17194
```

The code above shows that there are more posts with the Ask HN title than there are with the Show HN title.
We will have to use this information to create an average number of comments.

In [9]:

```
for row in ask_posts[:5]:
    print(row)
    print("\n")
```

```
['12296411', 'Ask HN: How to improve my personal website?', '', '2',
'6', 'ahmedbaracat', '8/16/2016 9:55']


['10610020', 'Ask HN: Am I the only one outraged by Twitter shutting
down share counts?', '', '28', '29', 'tkfx', '11/22/2015 13:43']


['11610310', 'Ask HN: Aby recent changes to CSS that broke mobile?',
'', '1', '1', 'polskibus', '5/2/2016 10:14']


['12210105', 'Ask HN: Looking for Employee #3 How do I do it?', '',
'1', '3', 'sph130', '8/2/2016 14:20']


['10394168', 'Ask HN: Someone offered to buy my browser extension fr
om me. What now?', '', '28', '17', 'roykolak', '10/15/2015 16:38']
```

In [10]:

```
for row in show_posts[:5]:
    print(row)
    print("\n")
```

```
['10627194', 'Show HN: Wio Link  ESP8266 Based Web of Things Hardwar
e Development Platform', 'https://iot.seeed.cc', '26', '22', 'kfihih
c', '11/25/2015 14:03']


['10646440', 'Show HN: Something pointless I made', 'http://dn.ht/pi
cklecat/', '747', '102', 'dhotson', '11/29/2015 22:46']


['11590768', 'Show HN: Shanhu.io, a programming playground powered b
y e8vm', 'https://shanhu.io', '1', '1', 'h8liu', '4/28/2016 18:05']


['12178806', 'Show HN: Webscope  Easy way for web developers to comm
unicate with Clients', 'http://webscopeapp.com', '3', '3', 'fastbric
k', '7/28/2016 7:11']


['10872799', 'Show HN: GeoScreenshot  Easily test Geo-IP based web p
ages', 'https://www.geoscreenshot.com/', '1', '9', 'kpsychwave', '1/
9/2016 20:45']
```

Data Analysis:

Total number of comments in ask posts:

In [13]:

```
total_ask_comments= 0

for row in ask_posts:
    num_comments=int(row[4])
    total_ask_comments += num_comments

avg_ask_comments = total_ask_comments/len(ask_posts)
print(avg_ask_comments)
```

```
14.038417431192661
```

Total number of show posts:

In [14]:

```
total_show_comments= 0

for row in show_posts:
    num_comments=int(row[4])
    total_show_comments += num_comments

avg_show_comments = total_show_comments/len(show_posts)
print(avg_show_comments)
```

```
10.31669535283993
```

Conclusion:

Posts whose title begins with 'Ask HN' have more average comments than posts whose title begins with 'Show HN'.

This conclusion makes logical sence since posts who are prompting a discussion would be more likely to receive comments than posts whose goal is to display user work.

Further Analysis: Are ask posts created at a certain time more likely to attract comments?

To asses this, I will first calculate the amount of ask posts created and the number of comments received in each hour of the day. Then I will calcualte the average number of comments ask posts receive by the hour created.

In [15]:

```
import datetime as dt
```

In [29]:

```
result_list= []

for row in ask_posts:
    created_at = row[6]
    num_comments = int(row[4])
    result=[created_at, num_comments]
    result_list.append(result)

counts_by_hour= {}
comments_by_hour= {}

date_format = "%m/%d/%Y %H:%M"

for date, comments in result_list:
    date_dt= dt.datetime.strptime(date, date_format)
    hour=date_dt.hour
    if hour not in counts_by_hour:
        counts_by_hour[hour]=1
        comments_by_hour[hour]= comments
    else:
        counts_by_hour[hour]+=1
        comments_by_hour[hour]+= comments
```

counts_by_hour: contains the number of ask posts created during each hour of the day

comments_by_hour: contains the corresponding number of comments ask posts created at each hour received

In [31]:

```
print('Counts by Hour:', counts_by_hour)
print('\n')
print('Comments by Hour:', comments_by_hour)
```

```
Counts by Hour: {0: 55, 1: 60, 2: 58, 3: 54, 4: 47, 5: 46, 6: 44, 7:
34, 8: 48, 9: 45, 10: 59, 11: 58, 12: 73, 13: 85, 14: 107, 15: 116,
16: 108, 17: 100, 18: 109, 19: 110, 20: 80, 21: 109, 22: 71, 23: 68}


Comments by Hour: {0: 447, 1: 683, 2: 1381, 3: 421, 4: 337, 5: 464,
6: 397, 7: 267, 8: 492, 9: 251, 10: 793, 11: 641, 12: 687, 13: 1253,
14: 1416, 15: 4477, 16: 1814, 17: 1146, 18: 1439, 19: 1188, 20: 1722
, 21: 1745, 22: 479, 23: 543}
```

Next, I will use these two dictionaries to calculate the average number of comments for posts created during each hour of the day.

In [38]:

```
avg_by_hour = [[h, comments_by_hour[h]/counts_by_hour[h]] for h in counts_by_hou
r]

print(avg_by_hour)
```

```
[[0, 8.127272727272727], [1, 11.383333333333333], [2, 23.81034482758
6206], [3, 7.796296296296297], [4, 7.170212765957447], [5, 10.086956
52173913], [6, 9.022727272727273], [7, 7.852941176470588], [8, 10.25
], [9, 5.5777777777777775], [10, 13.440677966101696], [11, 11.051724
137931034], [12, 9.41095890410959], [13, 14.741176470588234], [14, 1
3.233644859813085], [15, 38.5948275862069], [16, 16.796296296296298]
, [17, 11.46], [18, 13.201834862385321], [19, 10.8], [20, 21.525], [2
1, 16.009174311926607], [22, 6.746478873239437], [23, 7.985294117647
059]]
```

In [42]:

```
swap_avg_by_hour=[[comments_by_hour[h]/counts_by_hour[h], h] for h in counts_by_
hour]

print(swap_avg_by_hour)
```

```
[[8.127272727272727, 0], [11.383333333333333, 1], [23.810344827586206
6, 2], [7.796296296296297, 3], [7.170212765957447, 4], [10.086956521
73913, 5], [9.022727272727273, 6], [7.852941176470588, 7], [10.25, 8
], [5.5777777777777775, 9], [13.440677966101696, 10], [11.0517241379
31034, 11], [9.41095890410959, 12], [14.741176470588234, 13], [13.23
3644859813085, 14], [38.5948275862069, 15], [16.796296296296298, 16]
, [11.46, 17], [13.201834862385321, 18], [10.8, 19], [21.525, 20], [1
6.009174311926607, 21], [6.746478873239437, 22], [7.985294117647059,
23]]
```

In [44]:

```
sorted_swap= sorted(swap_avg_by_hour, reverse=True)
print(sorted_swap)
```

```
[[38.5948275862069, 15], [23.810344827586206, 2], [21.525, 20], [16.
796296296296298, 16], [16.009174311926607, 21], [14.741176470588234,
13], [13.440677966101696, 10], [13.233644859813085, 14], [13.2018348
6238532, 18], [11.46, 17], [11.383333333333333, 1], [11.051724137931
034, 11], [10.8, 19], [10.25, 8], [10.08695652173913, 5], [9.4109589
0410959, 12], [9.022727272727273, 6], [8.127272727272727, 0], [7.985
294117647059, 23], [7.852941176470588, 7], [7.796296296296297, 3], [
7.170212765957447, 4], [6.746478873239437, 22], [5.5777777777777775,
9]]
```

In [47]:

```
print("Top 5 Hours for Ask Posts Comments: ")

for average, hr in sorted_swap[:5]:
    hr_dt = dt.datetime.strptime(str(hr),'%H')
    hr_str = hr_dt.strftime("%H:%M")
    print("{}: {:.2f} average comments per post".format(hr_str, comments))
```

```
Top 5 Hours for Ask Posts Comments:
15:00: 2.00 average comments per post
02:00: 2.00 average comments per post
20:00: 2.00 average comments per post
16:00: 2.00 average comments per post
21:00: 2.00 average comments per post
```

Project Conclusion:

Earlier in this project it was established that Hacker News posts with the Ask HN posts have more average comments than the Post Hn posts. Therefore, a user is more likely to get their post to the top of the Hacker News listing if they create Ask HN post types.

After analyzing the time of each posting and averaging the number of comments per hour, the data suggests that the top 5 hours for Ask Posts to get comments are: 15, 2, 20, 16, and 21 UTC.

In [ ]: