Hacker News Project

June 18, 2024

1 Hacker News Project

We're specifically interested in posts with titles that begin with either Ask HN or Show HN. Users submit Ask HN posts to ask the Hacker News community a specific question.
Likewise, users submit Show HN posts to show the Hacker News community a project, product, or just something interesting.
We'll compare these two types of posts to determine the following:
-Do Ask HN or Show HN receive more comments on average? -Do posts created at a certain time receive more comments on average?

[6]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
```

[4]:

```
df = pd.read_csv('HN_posts_year_to_Sep_26_2016.csv')
df.head(20)
```

[4]:

```
          id                                              title  \
0   12579008  You have two days to comment if you want stem …
1   12579005                       SQLAR  the SQLite Archiver
2   12578997  What if we just printed a flatscreen televisio…
3   12578989                                algorithmic music
4   12578979  How the Data Vault Enables the Next-Gen Data W…
5   12578975                    Saving the Hassle of Shopping
6   12578954  Macalifa  A new open-source music app for UWP …
7   12578942  GitHub  theweavrs/Macalifa: A music player wri…
8   12578919                    Google Allo  first Impression
9   12578918    Advanced Multimedia on the Linux Command Line
10  12578908  Ask HN: What TLD do you use for local developm…
11  12578893                                       Muroc Maru
12  12578879          Why companies make their products worse
13  12578866                            Tuning AWS SQS Queues
14  12578857                            The Promise of GitHub
15  12578834                  Joint R&D Has Its Ups and Downs
16  12578831  IBM announces next implementation of Apples Sw…
17  12578822  Amazons Algorithms Dont Find You the Best Deals
```

```
18  12578816                               Ruffled Feathers
19  12578806     The Veil of Ignorance  Design and Accessbility

                                                  url  num_points  \
0   http://www.regulations.gov/document?D=FDA-2015…           1
1    https://www.sqlite.org/sqlar/doc/trunk/README.md           1
2   https://medium.com/vanmoof/our-secrets-out-f21…           1
3   http://cacm.acm.org/magazines/2011/7/109891-al…           1
4   https://www.talend.com/blog/2016/05/12/talend-…           1
5   https://blog.menswr.com/2016/09/07/whats-new-w…           1
6   http://forums.windowscentral.com/windows-phone…           1
7               https://github.com/theweavrs/Macalifa           1
8   http://prodissues.com/2016/09/google-allo-firs…           3
9   https://avi.alkalay.net/2016/09/multimedia-lin…           1
10                                                NaN           4
11   http://www.weirdca.com/location.php?location=511           1
12  https://www.1843magazine.com/ideas/the-daily/w…           4
13  http://blog.simontaranto.com/post/2016-09-25-t…           3
14  http://constantbetasoftware.com/2016/09/26/git…           2
15  http://semiengineering.com/joint-rd-has-its-up…           1
16  https://9to5mac.com/2016/09/25/ibm-announces-n…           2
17  https://www.technologyreview.com/s/602442/amaz…           1
18  http://www.texasmonthly.com/articles/whooping-…           1
19  https://blog.marvelapp.com/the-veil-of-ignorance/           3

    num_comments          author      created_at
0              0         altstar  9/26/2016 3:26
1              0        blacksqr  9/26/2016 3:24
2              0    pavel_lishin  9/26/2016 3:19
3              0    poindontcare  9/26/2016 3:16
4              0     markgainor1  9/26/2016 3:14
5              1           bdoux  9/26/2016 3:13
6              0        thecodrr  9/26/2016 3:06
7              0        thecodrr  9/26/2016 3:04
8              0          jandll  9/26/2016 2:57
9              0  mynameislegion  9/26/2016 2:56
10             7         Sevrene  9/26/2016 2:53
11             0            x43b  9/26/2016 2:46
12             0         RachelF  9/26/2016 2:40
13             0           srt32  9/26/2016 2:37
14             0            ttam  9/26/2016 2:34
15             0           Lind5  9/26/2016 2:28
16             0           phodo  9/26/2016 2:28
17             1       yarapavan  9/26/2016 2:26
18             0          Thevet  9/26/2016 2:23
19             0      muratmutlu  9/26/2016 2:21
```

2 Mean Analysis

[19]:

```
#get all comments that have Ask HN and Show HN
model = df[df['title'].str.contains('Ask HN|Show HN')].reset_index()
```

[20]:

```
model
```

[20]:

```
        index        id                                              title  \
0          10  12578908  Ask HN: What TLD do you use for local developm…
1          42  12578522  Ask HN: How do you pass on your work when you …
2          52  12578335              Show HN: Finding puns computationally
3          58  12578182  Show HN: A simple library for complicated anim…
4          64  12578098      Show HN: WebGL visualization of DNA sequences
…       …       …                                                …
19287  293047  10177359  Ask HN: Is coursera specialization in product …
19288  293052  10177317  Ask HN: Any meteor devs out there who could sp…
19289  293055  10177309  Ask HN: Any recommendations for books about ra…
19290  293073  10177200  Ask HN: Where do you look for work if you need…
19291  293114  10176919        Ask HN: What is/are your favorite quote(s)?

                                                    url  num_points  \
0                                                   NaN           4
1                                                   NaN           6
2                         http://puns.samueltaylor.org/           2
3       https://christinecha.github.io/choreographer-js/           1
4                    http://grondilu.github.io/dna.html           1
…                                                 …         …
19287                                               NaN           1
19288                                               NaN           2
19289                                               NaN           2
19290                                               NaN          14
19291                                               NaN          15

       num_comments          author       created_at
0                 7         Sevrene   9/26/2016 2:53
1                 3      PascLeRasc   9/26/2016 1:17
2                 0           saamm   9/26/2016 0:36
3                 0    christinecha   9/26/2016 0:01
4                 0        grondilu  9/25/2016 23:44
…             …             …              …
19287             0         pipipzz   9/6/2015 11:27
19288             1      louisswiss   9/6/2015 10:52
19289             4  rationalthrowa   9/6/2015 10:46
19290            20      coroutines    9/6/2015 9:36
19291            20        kumarski    9/6/2015 6:02

[19292 rows x 8 columns]
```

[23]:

```
#Just leave Ask HN and Show HN in title
model['title'] = model.title.str.extract('(Ask HN|Show HN)')
```

[24]:

```
model
```

[24]:

```
        index        id    title  \
0          10  12578908   Ask HN
1          42  12578522   Ask HN
2          52  12578335  Show HN
3          58  12578182  Show HN
4          64  12578098  Show HN
…       …       …      …
19287  293047  10177359   Ask HN
19288  293052  10177317   Ask HN
19289  293055  10177309   Ask HN
19290  293073  10177200   Ask HN
19291  293114  10176919   Ask HN

                                                    url  num_points  \
0                                                   NaN           4
1                                                   NaN           6
2                          http://puns.samueltaylor.org/           2
3      https://christinecha.github.io/choreographer-js/           1
4                     http://grondilu.github.io/dna.html           1
…                                                 …         …
19287                                               NaN           1
19288                                               NaN           2
19289                                               NaN           2
19290                                               NaN          14
19291                                               NaN          15

       num_comments          author       created_at
0                 7         Sevrene   9/26/2016 2:53
1                 3      PascLeRasc   9/26/2016 1:17
2                 0           saamm   9/26/2016 0:36
3                 0    christinecha   9/26/2016 0:01
4                 0        grondilu  9/25/2016 23:44
…             …             …              …
19287             0         pipipzz   9/6/2015 11:27
19288             1      louisswiss   9/6/2015 10:52
19289             4  rationalthrowa   9/6/2015 10:46
19290            20      coroutines    9/6/2015 9:36
19291            20        kumarski    9/6/2015 6:02

[19292 rows x 8 columns]
```

[28]:

```
model = model.drop(['index','id','url'],axis=1)
```

```
[29]: #find any null values
      model.isna().sum()
```

```
[29]: title           0
      num_points      0
      num_comments    0
      author          0
      created_at      0
      dtype: int64
```

```
[31]: #find any duplicates
      model.duplicated().sum()
```

```
[31]: 1
```

```
[33]: # drop duplicates
      model = model.drop_duplicates()
```

```
[34]: model.head()
```

```
[34]:      title  num_points  num_comments        author       created_at
      0   Ask HN           4             7       Sevrene   9/26/2016 2:53
      1   Ask HN           6             3    PascLeRasc   9/26/2016 1:17
      2  Show HN           2             0         saamm   9/26/2016 0:36
      3  Show HN           1             0  christinecha   9/26/2016 0:01
      4  Show HN           1             0      grondilu  9/25/2016 23:44
```

```
[43]: #group by sum and average for each title
      summary = model.groupby('title').agg(comm_sum␣
       ↪=('num_comments','sum'),comm_avg=('num_comments','mean')).reset_index()
```

```
[44]: summary
```

```
[44]:      title  comm_sum   comm_avg
      0   Ask HN     94940  10.402104
      1  Show HN     49678   4.887643
```

```
[51]: summary.iloc[0,1]/sum(summary.comm_sum)*100
```

```
[51]: 65.64881273423778
```

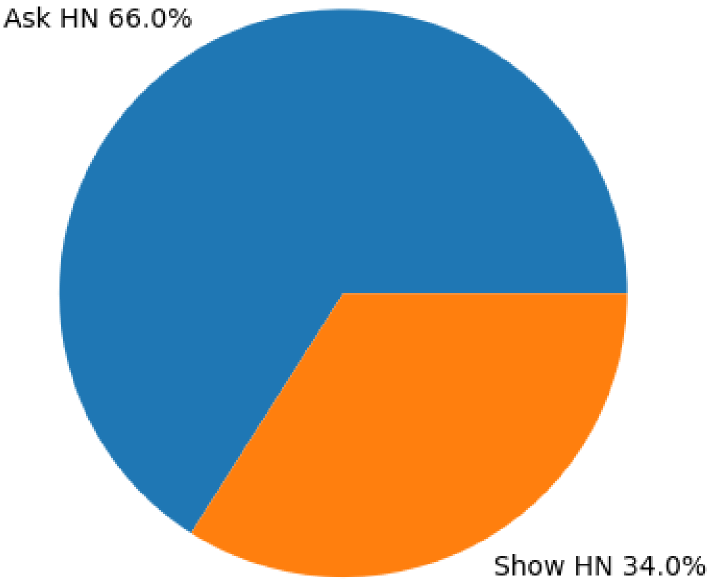
```
[54]: #add % columns
      summary['comm %']=round(summary.iloc[:,1]/sum(summary.comm_sum)*100)
```

```
[55]: summary
```

```
[55]:      title  comm_sum   comm_avg  comm %
      0   Ask HN     94940  10.402104    66.0
```

```
1  Show HN      49678   4.887643    34.0
```

[66]:
```
#pie chart to visualize comparison
plt.pie('comm %', data=summary, labels= summary['title']+' '+ summary['comm %'].
 ↪astype(str)+'%')
plt.show()
```



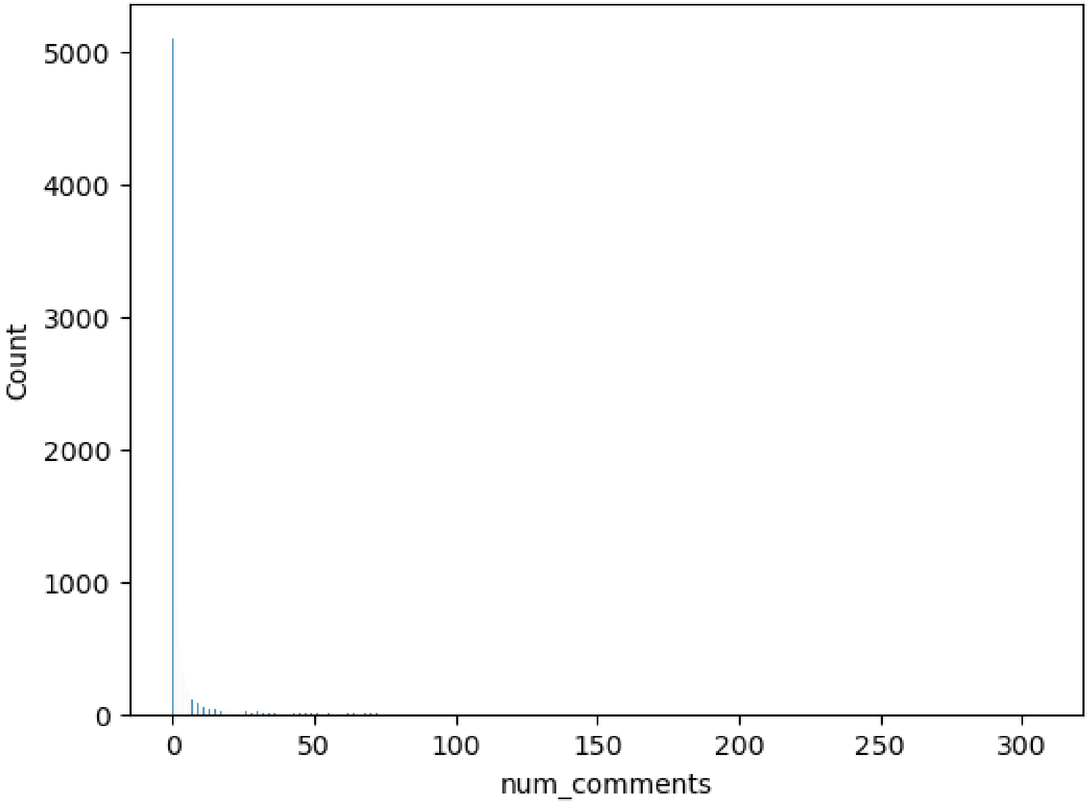
[77]:
```
#get the array of all comment numbers for Ask HN
a = model[model.title.str.contains('Ask HN')].num_comments
```

[78]:
```
#get the array of all comment numbers for Show HN
b = model[model.title.str.contains('Show HN')].num_comments
```

[81]:
```
#histplot of b to visualise distribution
sns.histplot(b)
```

```
C:\Users\natha\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

[81]:
```
<Axes: xlabel='num_comments', ylabel='Count'>
```



[82]:
```
#appears to be outliers lets investigate
model.describe()
```

[82]:
```
         num_points    num_comments
count  19291.000000  19291.000000
mean      13.182780      7.496656
std       46.997198     32.275180
min        1.000000      0.000000
25%        2.000000      0.000000
50%        3.000000      1.000000
75%        7.000000      4.000000
max     1624.000000   1007.000000
```

[84]:
```
#See how many comments are above 100
model[model.num_comments>100]
```

[84]:
```
         title  num_points  num_comments         author      created_at
58     Show HN         301           102  natashabaker  9/24/2016 15:06
81      Ask HN         171           477         prmph  9/23/2016 20:18
161     Ask HN         442           266    curiousgal  9/22/2016 11:52
177    Show HN         184           167      gilsadis  9/21/2016 21:55
```

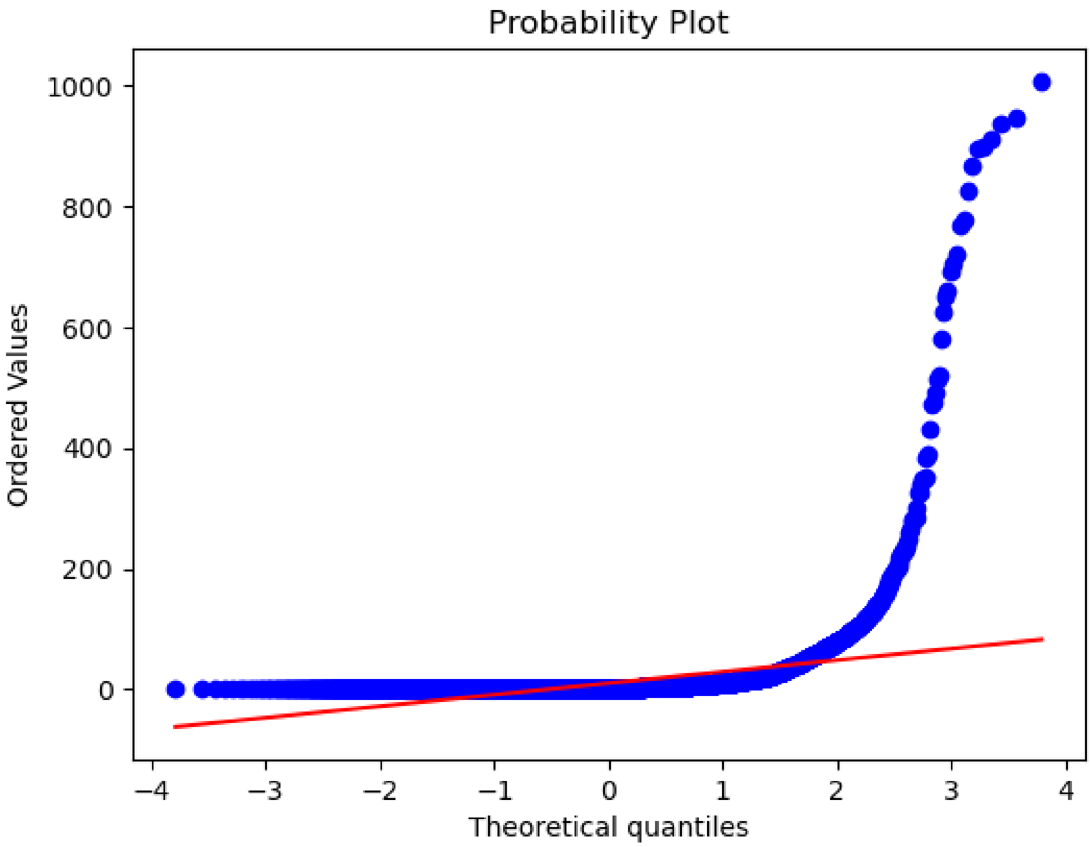
```
285     Show HN         893           169         fogleman  9/20/2016 12:55
…         …         …           …              …              …
18160    Ask HN         157           205        networked  9/30/2015 10:26
18313   Show HN         681           103  damjanstankovic  9/26/2015 20:29
18616   Show HN         572           163              orf  9/20/2015 19:50
18981   Show HN         134           103           navlio  9/12/2015 15:37
19037   Show HN        1172           136       hannahmitt  9/11/2015 14:58

[200 rows x 5 columns]
```

[86]:

```python
#Q-Q plot for a
import scipy.stats as stats
stats.probplot(a, dist="norm", plot=plt)
```

[86]:

```
((array([-3.7879693 , -3.56155022, -3.43718551, …,  3.43718551,
          3.56155022,  3.7879693 ]),
  array([   0,    0,    0, …,  937,  947, 1007], dtype=int64)),
 (19.072485682545683, 10.40210364851539, 0.4379412453296621))
```



Normal assumption is violated. Data is heavy right skew. The tail end comments appear to be popular posts so their value is important and can't be taken away. Therefore, I wil perform a non-parametric t-test that doesn't require normality assumption.

```
[87]: # Perform Mann-Whitney U test
      u_stat, p_value = stats.mannwhitneyu(a, b)

      print(f"U-statistic: {u_stat}, p-value: {p_value}")
```

```
U-statistic: 60916618.0, p-value: 0.0
```

I perfomed a Mann-Whitnet U test as data failed normality assumption as seen from the Q-Q plot and histplots. P-value is 0.00 so we reject null hypothesis and hence, means are different. We can conclude Ask HN have more comments on average.

```
[88]: model.head()
```

```
[88]:     title  num_points  num_comments        author       created_at
      0   Ask HN           4             7       Sevrene   9/26/2016 2:53
      1   Ask HN           6             3     PascLeRasc   9/26/2016 1:17
      2  Show HN           2             0         saamm   9/26/2016 0:36
      3  Show HN           1             0   christinecha   9/26/2016 0:01
      4  Show HN           1             0      grondilu  9/25/2016 23:44
```

3 Time Analysis

```
[89]: #check data types
      model.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 19291 entries, 0 to 19291
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   title         19291 non-null  object
 1   num_points    19291 non-null  int64
 2   num_comments  19291 non-null  int64
 3   author        19291 non-null  object
 4   created_at    19291 non-null  object
dtypes: int64(2), object(3)
memory usage: 904.3+ KB
```

```
[95]: #datetime is object so i will extract hour for analysis
      model['time']=pd.to_datetime(model.created_at).dt.hour
```

```
[96]: model
```

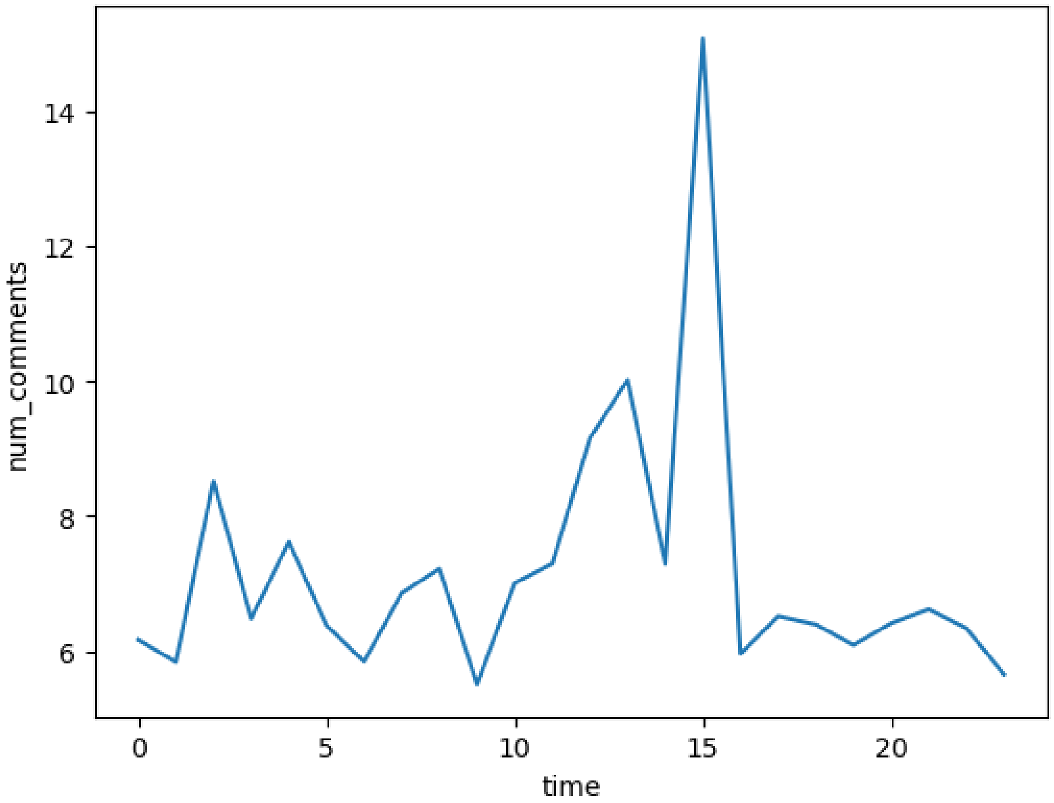
```
[96]:        title  num_points  num_comments          author       created_at  \
      0     Ask HN           4             7         Sevrene   9/26/2016 2:53
```

```
1        Ask HN            6             3      PascLeRasc   9/26/2016 1:17
2       Show HN            2             0           saamm   9/26/2016 0:36
3       Show HN            1             0     christinecha   9/26/2016 0:01
4       Show HN            1             0         grondilu  9/25/2016 23:44
…          …         …            …            …              …
19287    Ask HN            1             0          pipipzz   9/6/2015 11:27
19288    Ask HN            2             1       louisswiss   9/6/2015 10:52
19289    Ask HN            2             4  rationalthrowa   9/6/2015 10:46
19290    Ask HN           14            20       coroutines    9/6/2015 9:36
19291    Ask HN           15            20         kumarski    9/6/2015 6:02

       time
0         2
1         1
2         0
3         0
4        23
…      …
19287    11
19288    10
19289    10
19290     9
19291     6

[19291 rows x 6 columns]
```

[130]:

```
#Lineplot to visualise if any times stand out
sns.lineplot(x='time',y='num_comments',data=model, ci=None)
plt.show()
```

```
C:\Users\natha\AppData\Local\Temp\ipykernel_12968\21469759.py:1: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.

  sns.lineplot(x='time',y='num_comments',data=model, ci=None)
C:\Users\natha\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\natha\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



[106]:
```
#Perform Anova test to compare each hour of day to see if statistically␣
 ↪different
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.multicomp import pairwise_tukeyhsd
#turn time column into category for anova
model['time'] = pd.Categorical(model['time'])

# Fit ANOVA model
anova = ols('num_comments ~ C(time)', data=model).fit()

# Perform ANOVA (Type 1)
anova_table = sm.stats.anova_lm(anova, typ=1)
```

[107]:
```
# Print ANOVA table
print(anova_table)
```

```
                df        sum_sq      mean_sq         F        PR(>F)
C(time)       23.0  1.152029e+05  5008.820974  4.830333  1.793287e-13
Residual   19267.0  1.997894e+07  1036.951453       NaN           NaN
```

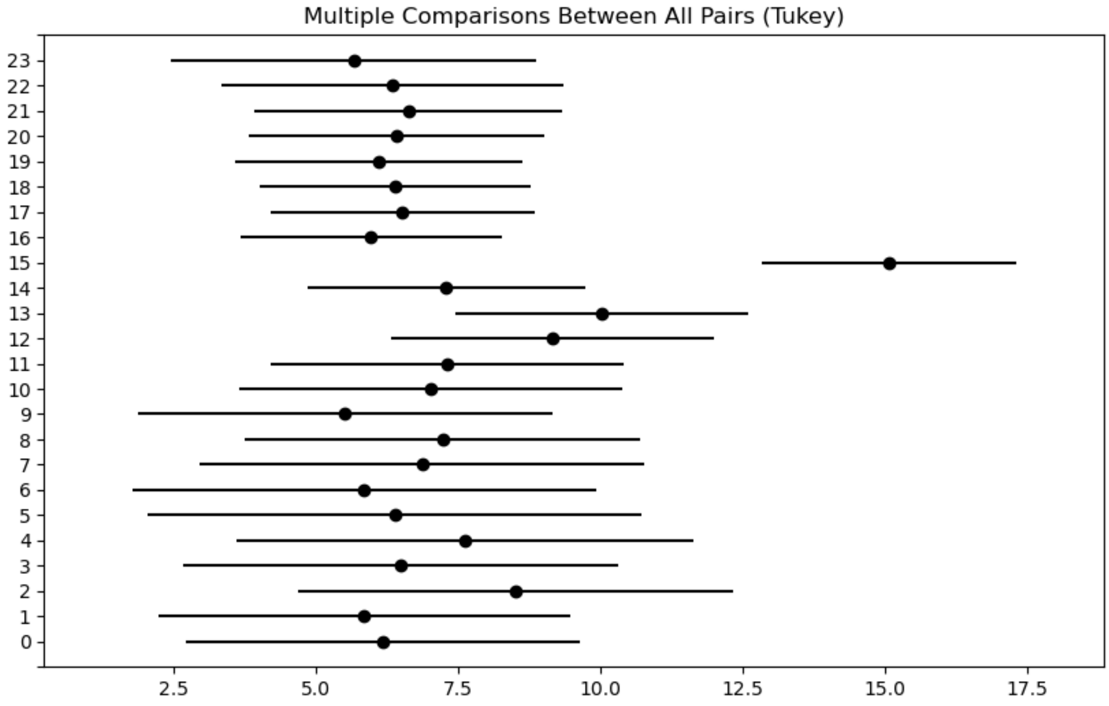
[119]:

```
# Perform the Tukey's HSD post hoc test
tukey_results = pairwise_tukeyhsd(endog=model['num_comments'],␣
 ↪groups=model['time'], alpha=0.05)
a = tukey_results.summary()
```

```
C:\Users\natha\anaconda3\Lib\site-packages\scipy\integrate\_quadpack_py.py:1233:
IntegrationWarning: The integral is probably divergent, or slowly convergent.
  quad_r = quad(f, low, high, args=args, full_output=self.full_output,
```

[114]:

```
tukey_results.plot_simultaneous()
plt.show()
```



4 Summary

The analysis perfomed can now answer the original questions: 1. From non-parametric t-test we see means are statistically different and Ask HN has more comments on average. 2. From anova and tukey test we see most comments are posted at 3pm.

[ ]:

```
```