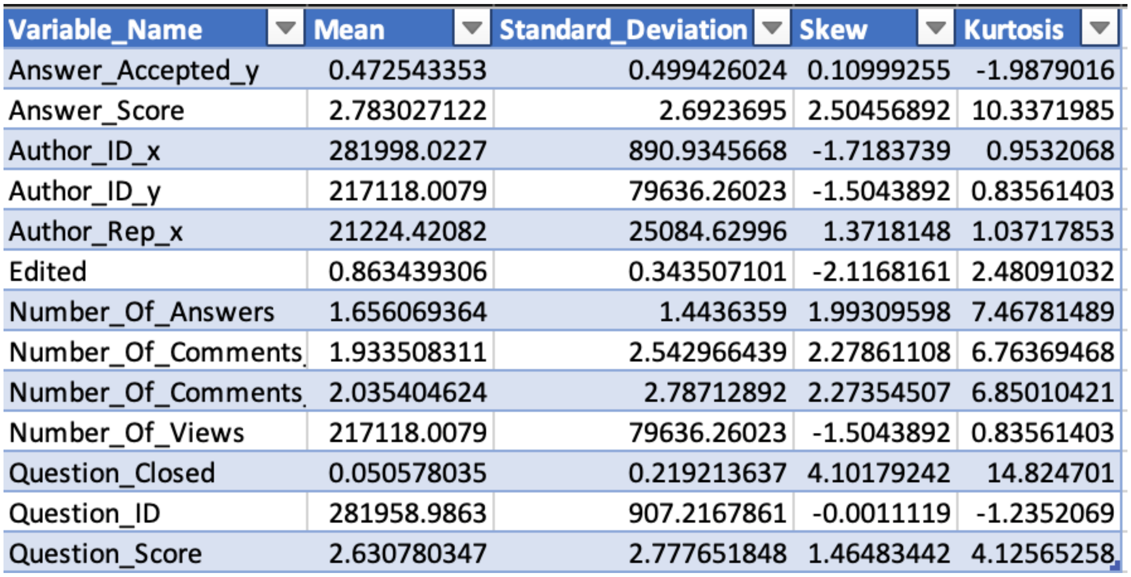
James Clarke
Troy Ladka
David Weinstein
CMPSC-310-A
April 25th, 2023

Exploring Data Report

Moments:

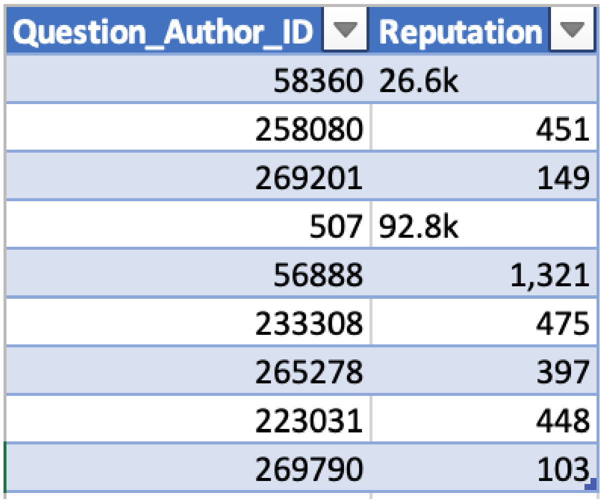


Variable_Name	Mean	Standard_Deviation	Skew	Kurtosis
Answer_Accepted_y	0.472543353	0.499426024	0.10999255	-1.9879016
Answer_Score	2.783027122	2.6923695	2.50456892	10.3371985
Author_ID_x	281998.0227	890.9345668	-1.7183739	0.9532068
Author_ID_y	217118.0079	79636.26023	-1.5043892	0.83561403
Author_Rep_x	21224.42082	25084.62996	1.3718148	1.03717853
Edited	0.863439306	0.343507101	-2.1168161	2.48091032
Number_Of_Answers	1.656069364	1.4436359	1.99309598	7.46781489
Number_Of_Comments	1.933508311	2.542966439	2.27861108	6.76369468
Number_Of_Comments	2.035404624	2.78712892	2.27354507	6.85010421
Number_Of_Views	217118.0079	79636.26023	-1.5043892	0.83561403
Question_Closed	0.050578035	0.219213637	4.10179242	14.824701
Question_ID	281958.9863	907.2167861	-0.0011119	-1.2352069
Question_Score	2.630780347	2.777651848	1.46483442	4.12565258

Interpretation:

The acceptance rate for answers is close to 50%, meaning that many answers on the platform receive at least one correct answer. However, the questions' closed rate is low, meaning that many questions are left open or unanswered. This may indicate that although a correct answer may exist, the forum is left open to continue discussion on the topic.

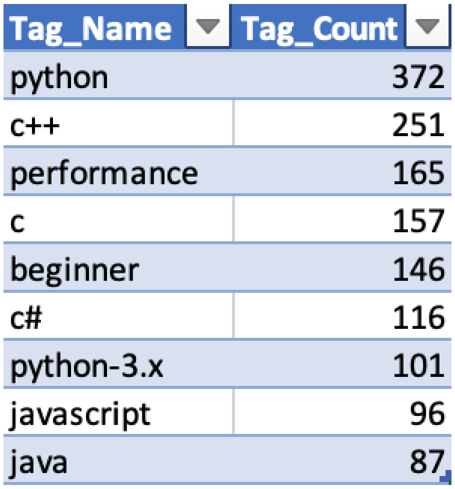
Top Authors:



Question_Author_ID	Reputation
58360	26.6k
258080	451
269201	149
507	92.8k
56888	1,321
233308	475
265278	397
223031	448
269790	103

Interpretation:
It is clear that there is no direct correlation between author's reputation score and their productivity.

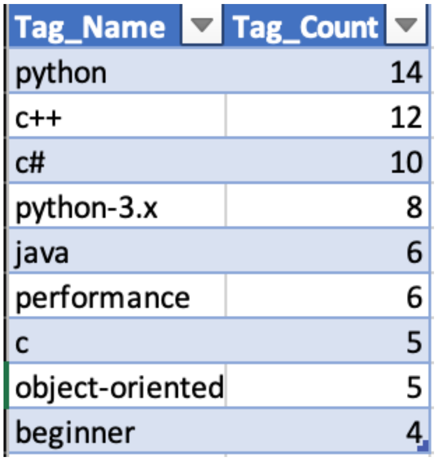
Unfiltered Tags:



Tag_Name	Tag_Count
python	372
c++	251
performance	165
c	157
beginner	146
c#	116
python-3.x	101
javascript	96
java	87

Interpretation:
Python, C-Plus-Plus, C, C-Sharp, JavaScript and Java are all extremely popular programming languages, which may explain why they are frequently used as tags. Python and C-Plus-Plus, which are commonly taught at the university level, appear at the top of the list.

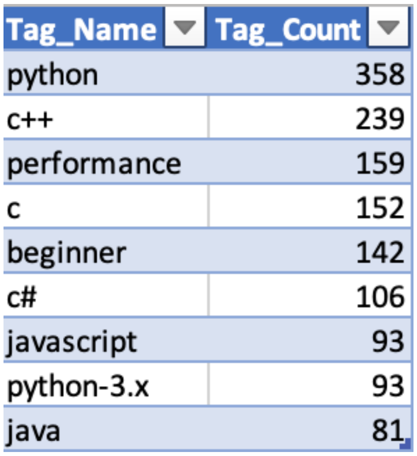
Closed Question Tags:



Tag_Name	Tag_Count
python	14
c++	12
c#	10
python-3.x	8
java	6
performance	6
c	5
object-oriented	5
beginner	4

vs.

Not Closed Question Tags:

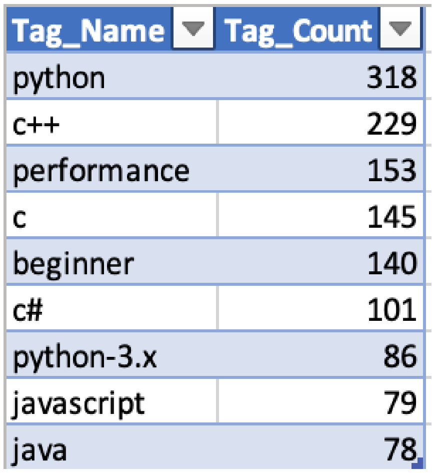


Tag_Name	Tag_Count
python	358
c++	239
performance	159
c	152
beginner	142
c#	106
javascript	93
python-3.x	93
java	81

Interpretation:

There are little differences in the rankings of tags used in closed vs. non-closed questions. What is more evident is the large difference in the number of tags that appear in non-closed questions as opposed to closed questions. As aforementioned, most questions on the platform are left open, which may be why we are seeing these results.

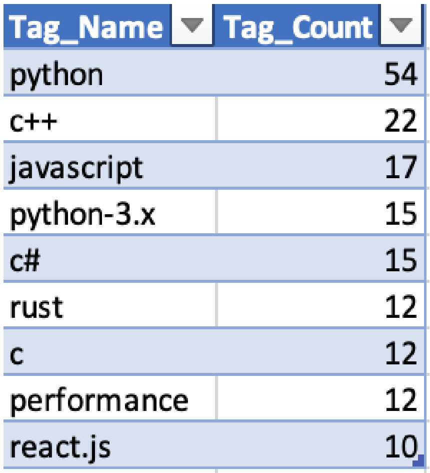
Edited Question Tags:



Tag_Name	Tag_Count
python	318
c++	229
performance	153
c	145
beginner	140
c#	101
python-3.x	86
javascript	79
java	78

vs.

Not Edited Question Tags:

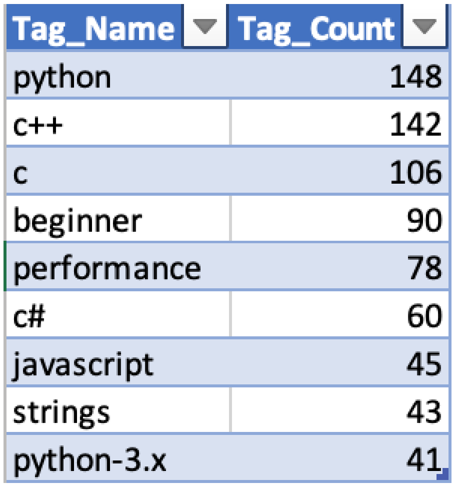


Tag_Name	Tag_Count
python	54
c++	22
javascript	17
python-3.x	15
c#	15
rust	12
c	12
performance	12
react.js	10

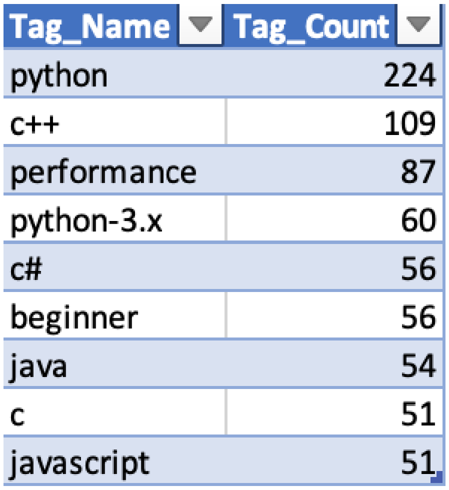
Interpretation:

It is clear that both lists contain similar tags, suggesting both edited and unedited questions contain similar subject queries. There is a large difference in the number of tags present in edited questions as opposed to unedited questions.

Has Accepted Answer Tags: vs. Does Not Have Accepted Answer Tags:



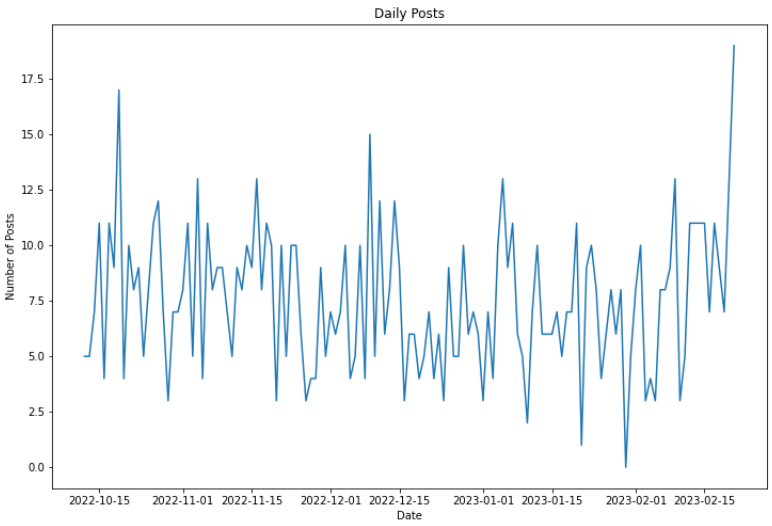
Tag_Name	Tag_Count
python	148
c++	142
c	106
beginner	90
performance	78
c#	60
javascript	45
strings	43
python-3.x	41



Tag_Name	Tag_Count
python	224
c++	109
performance	87
python-3.x	60
c#	56
beginner	56
java	54
c	51
javascript	51

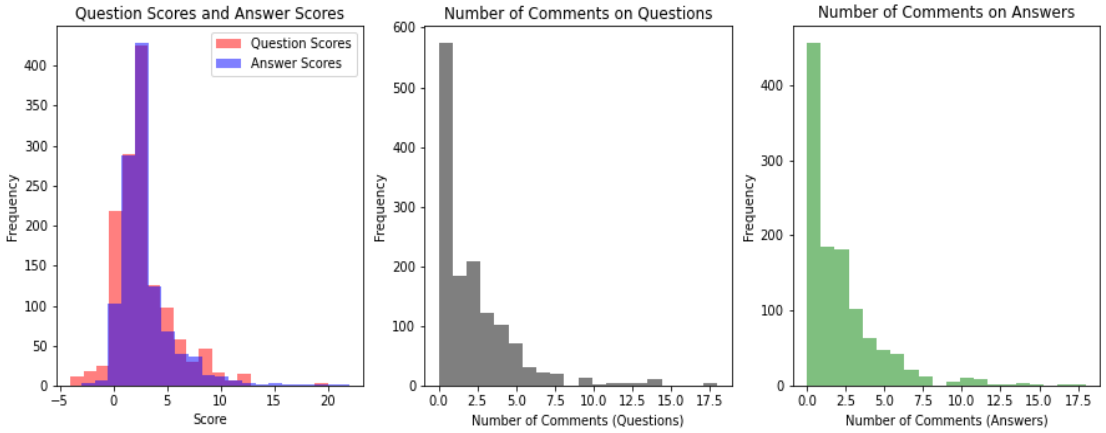
Interpretation:
Both lists of most commonly used tags appear to be similar in content. It is evident that the number tags present in accepted answers is similar to the number of tags present in non-accepted answers. This can confirm some of the data we found in our moments.csv which clearly shows acceptance rate for answers is close to 50%.

Number of Daily Posts:



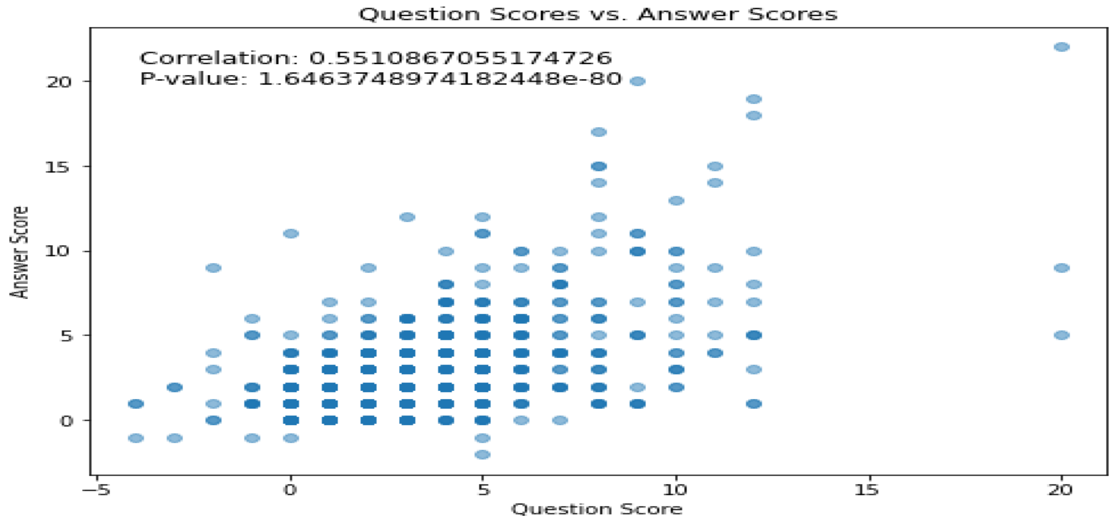
Interpretation:
The number of daily posts on the platform is evenly distributed across the given time period. The graph is showing highs of around 17.5, and lows of 0 - 0.5. Towards the end of February, 2023, there is a dramatic increase in the number of daily posts that may continue to climb if we explore more data.

Histograms:



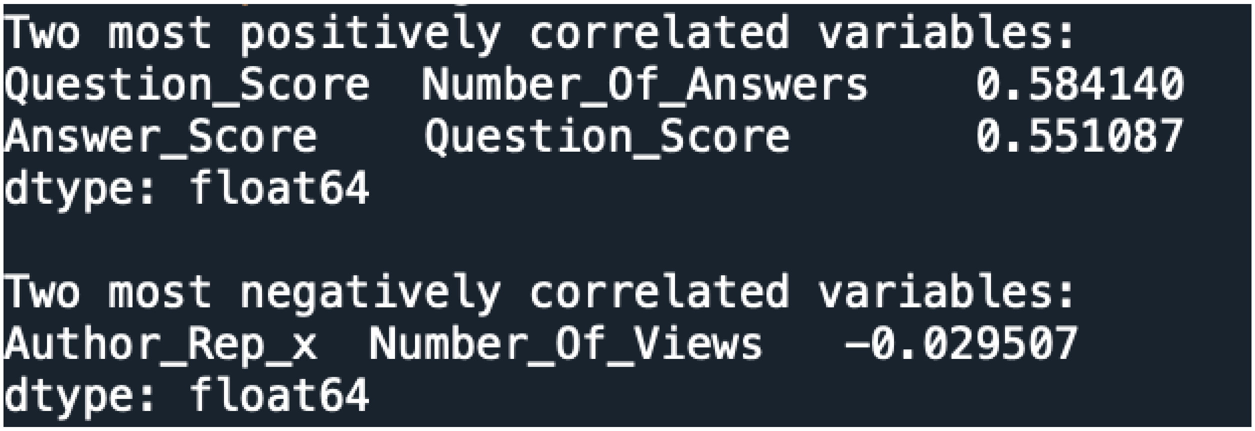
Interpretation:
It's clear the number of question scores and answer scores take a similar shape on the first histogram as they could be highly correlated. The highest frequency of answer/question scores is between 0 and 5. The highest frequency of comments on questions is low, around 0 - 0.5, which would make sense as most questions are not expected to receive a large number of comments. The number of comments on answers shows very similar results.

Scatter Plot:



Interpretation:
This correlation coefficient is the second strongest we found between variables in our dataframe. This shows a moderate relationship. The small p-value indicates a statistically significant correlation.

Most Positively and Negatively Correlated Variables:



```
Two most positively correlated variables:
Question_Score  Number_Of_Answers    0.584140
Answer_Score    Question_Score       0.551087
dtype: float64

Two most negatively correlated variables:
Author_Rep_x  Number_Of_Views   -0.029507
dtype: float64
```

Interpretation:

The two most positively correlated variables in our dataframe were Question_Score vs. Number_Of_Answers (0.584140) and Answer_Score vs. Question_Score (0.551087). It makes sense that if the question score is high, the answer score and number of answers might also be high. However, this correlation is not very strong. The most negatively correlated variables in our dataframe were author reputation and number of views (-0.02957). This value indicates these values are seemingly uncorrelated.