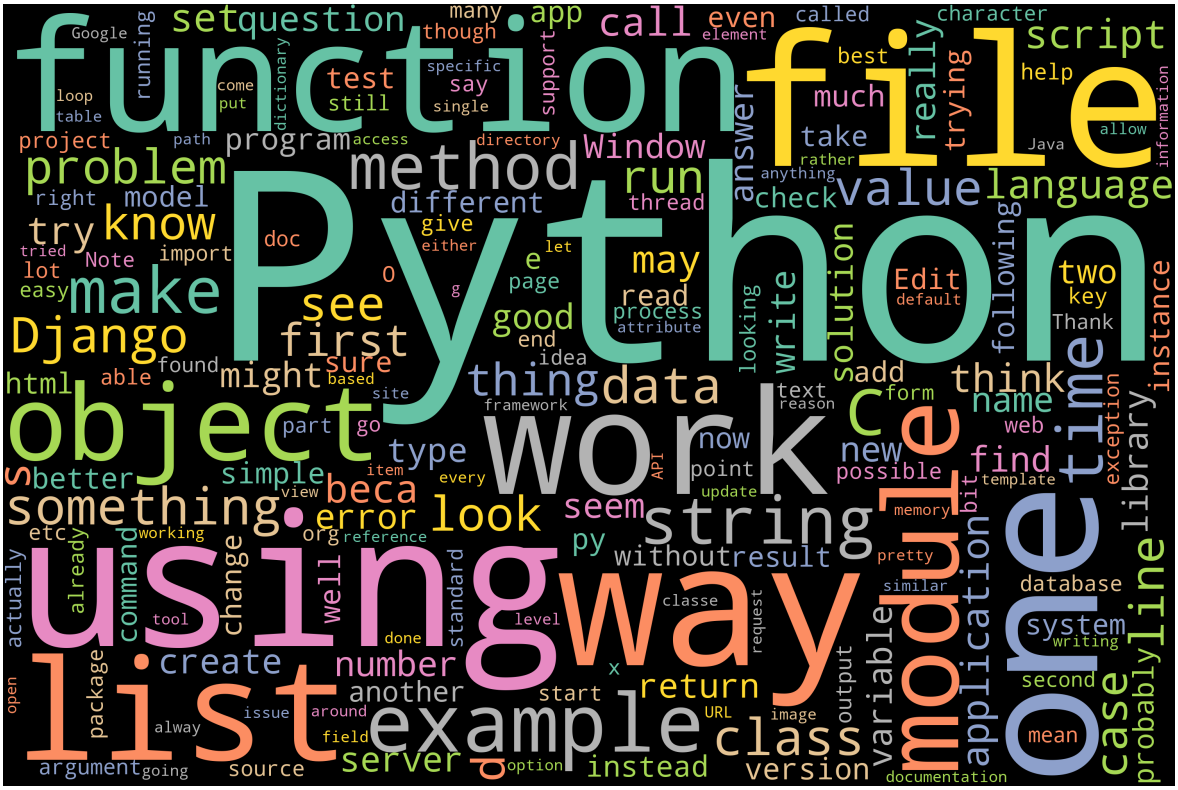
Criteria for subsampling

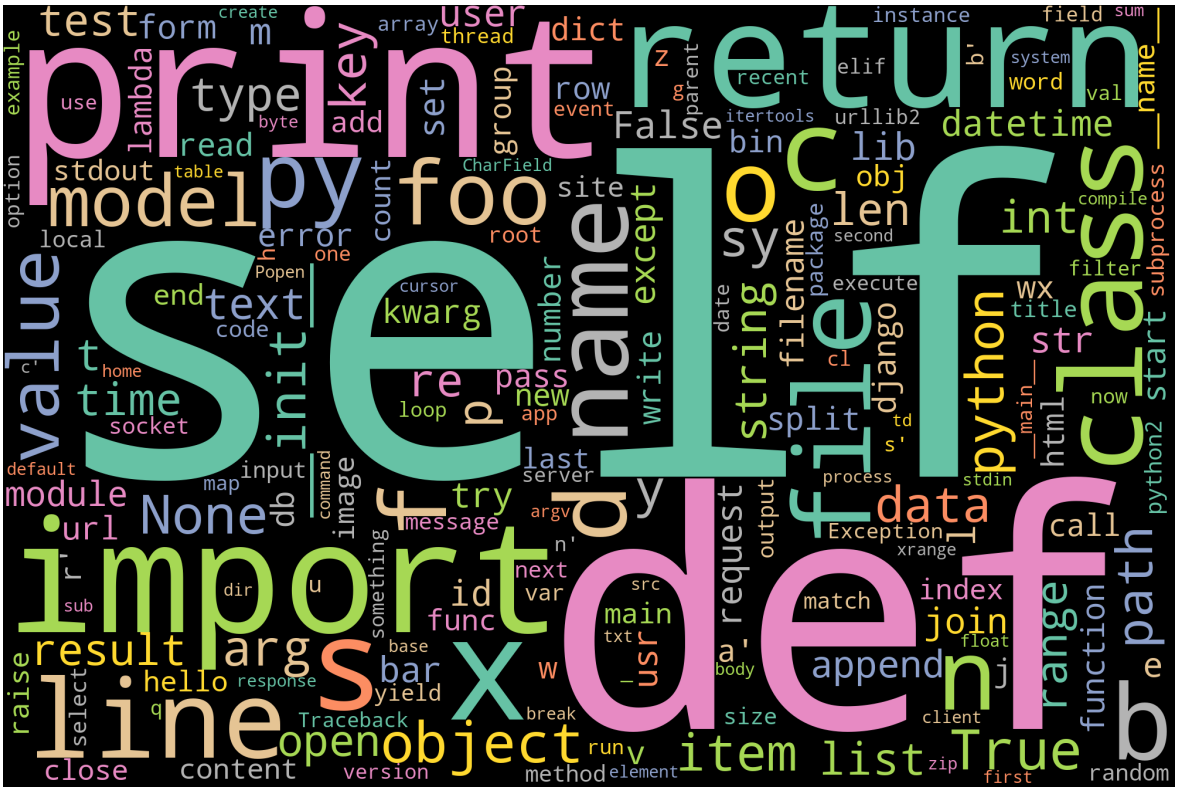
All the questions in the posts have “Python” Tag. All users, votes, answers are related to these questions with “Python” Tag

Word Cloud

Word Cloud for posts text (except the code)

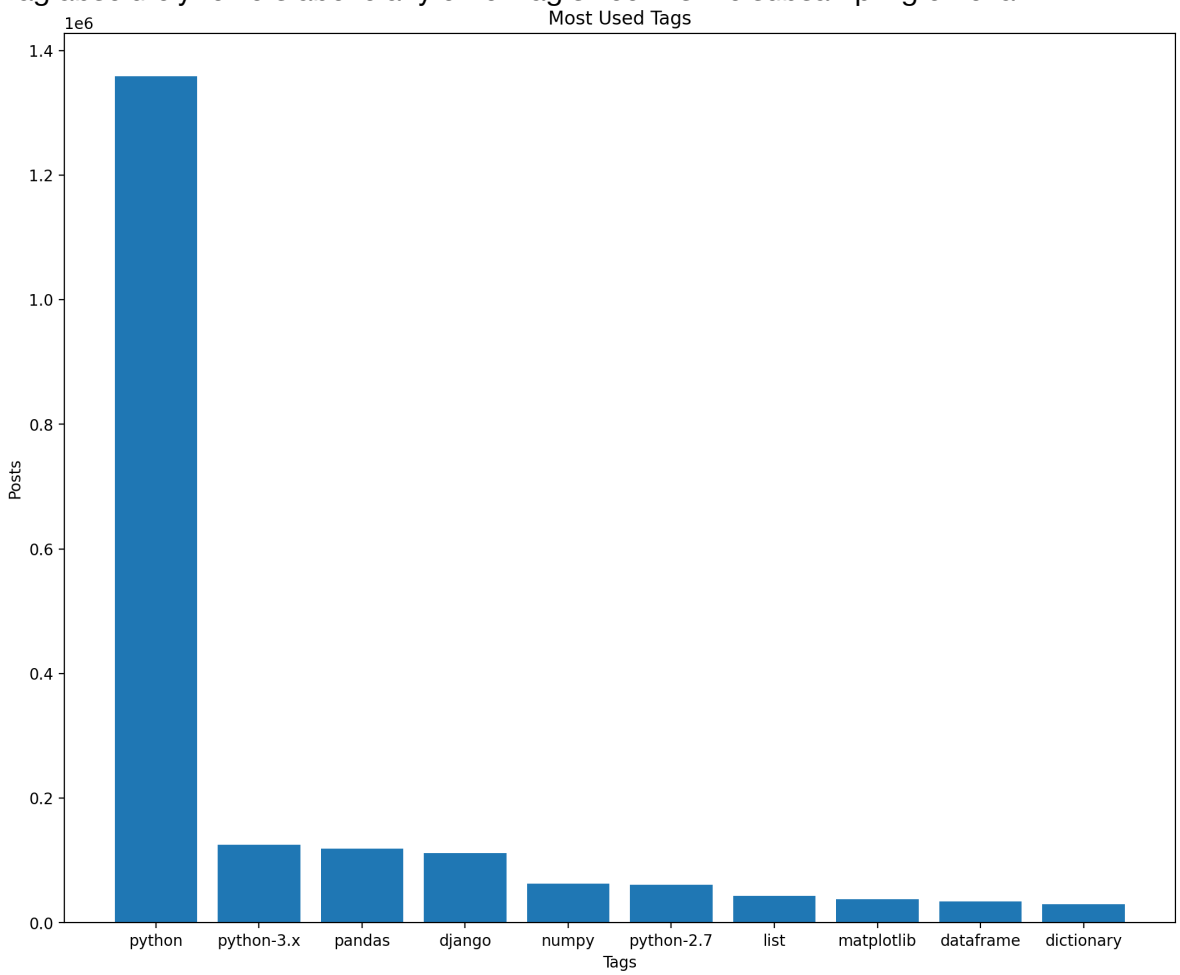


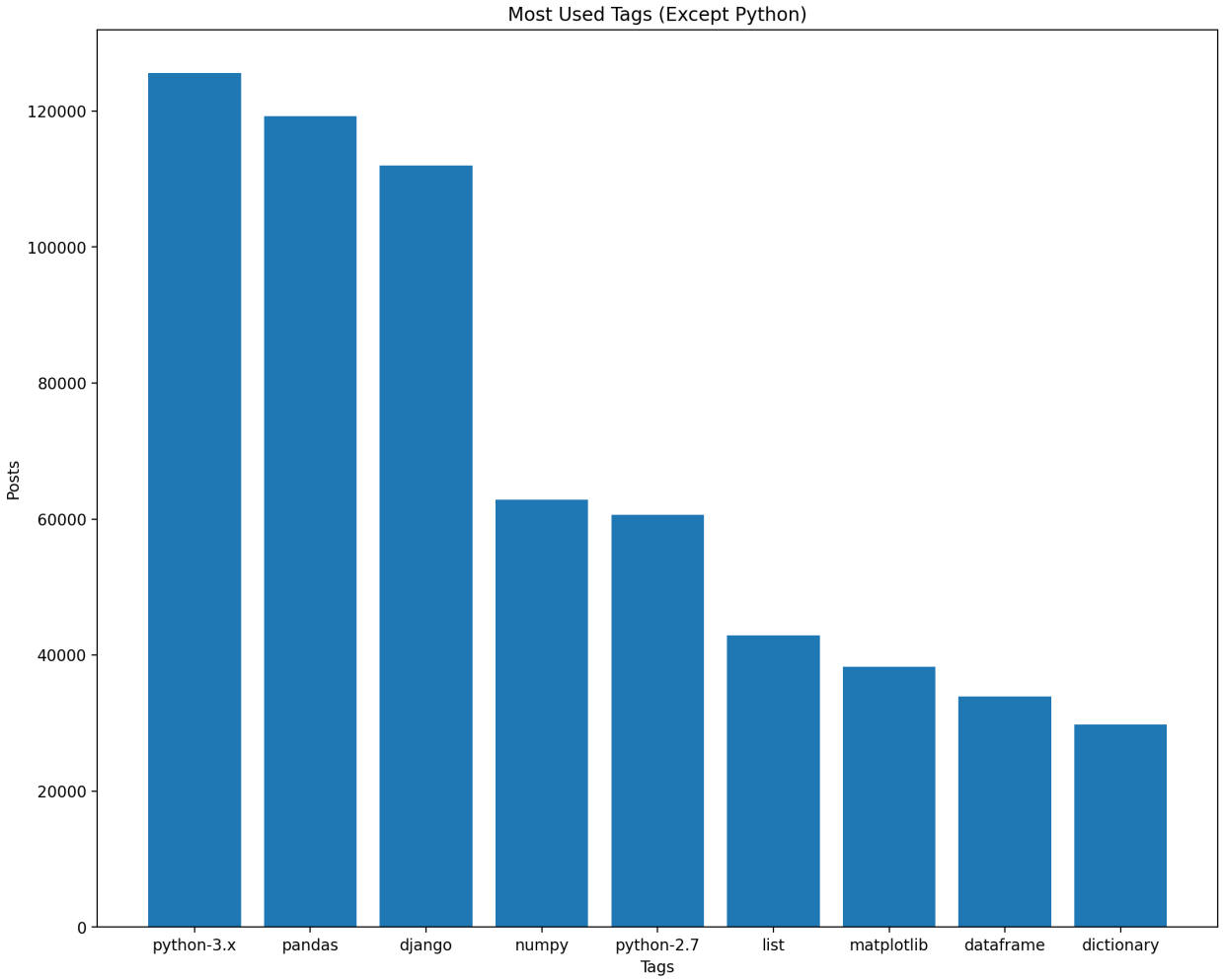
Word Cloud for the code in Posts, these are mostly Python keywords and common functions. Also ‘foo’, ‘x’ etc show up which are commonly used as variable names



Bar Plots for commonly used tags

Python Tag absolutely towers above any other Tag since it is the subsampling criteria





Without the Python Tag, the most posts have pandas framework related questions closely followed by Django framework

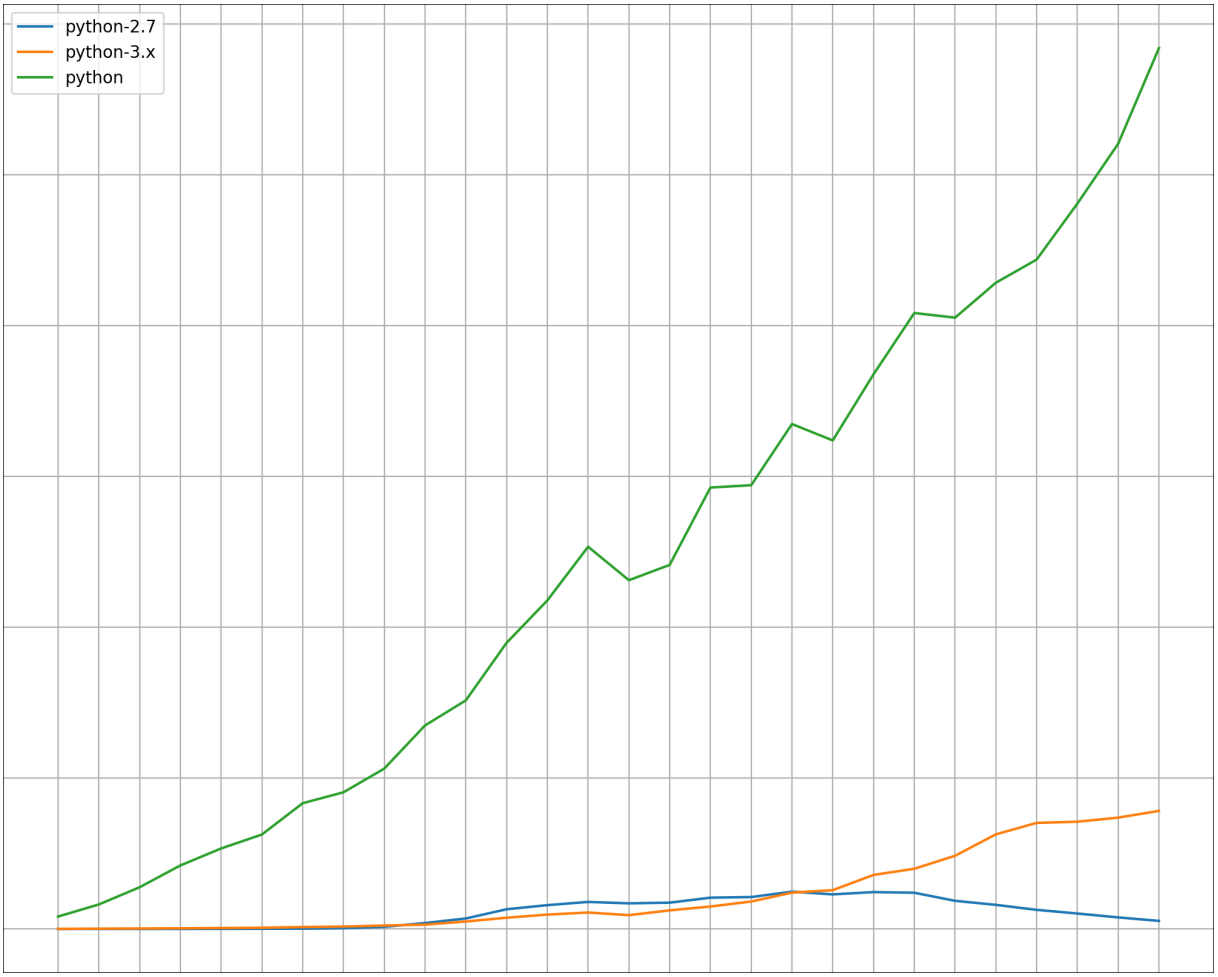
Exploratory Analysis

Registered Users	670395
Total Questions	1358860
Total Answers	2021741
Total Posts	3380601
Total Votes	13516747
Answers/Questions	1.48782140912235

- 3.05% of users make up of 50% of the posts
- 20.28% of users make up 80% of posts, which surprisingly lines up with the Pareto Principle
- Average posts which has been closed takes more than 3 months (101 days) to be closed

---

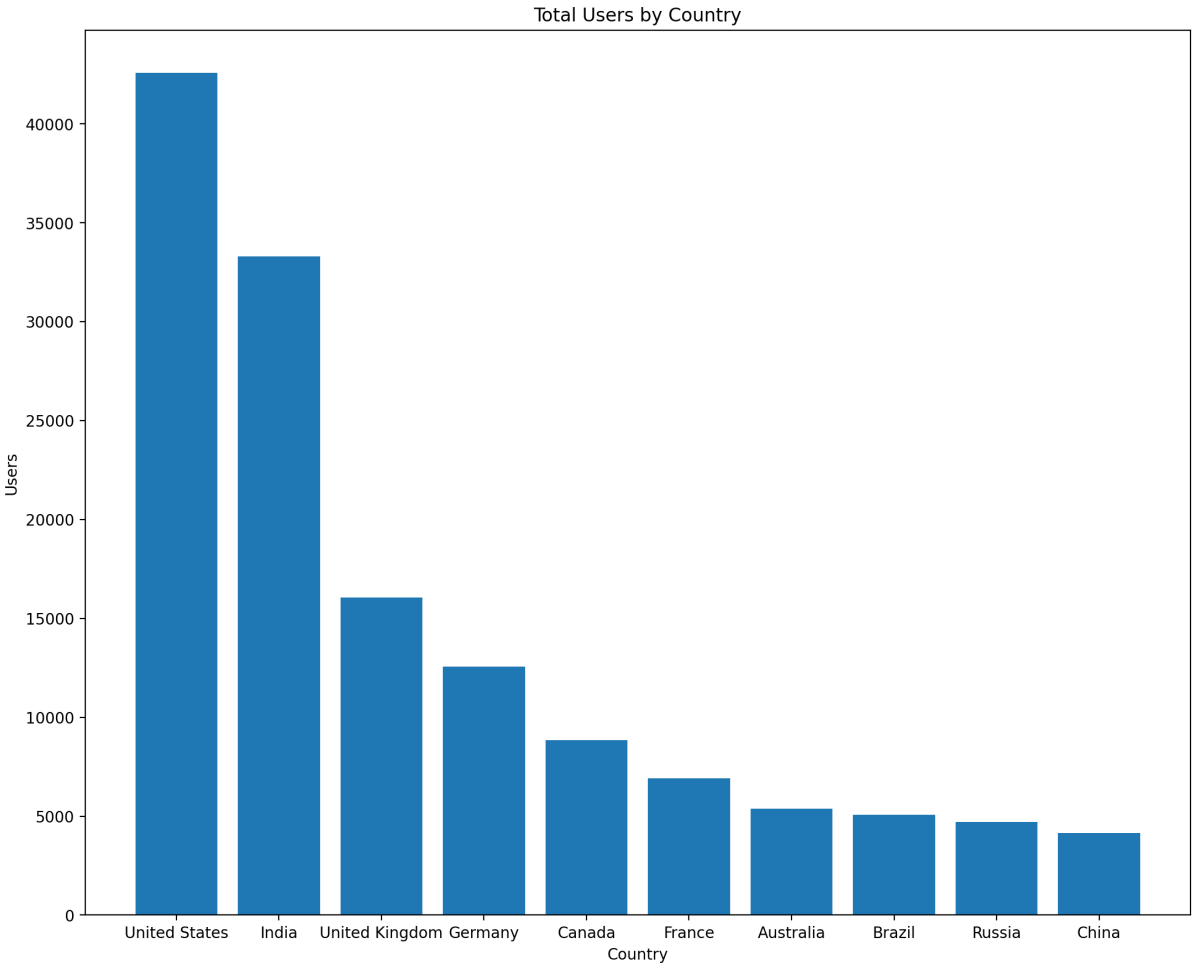
Usage of Different Python Tags Overtime

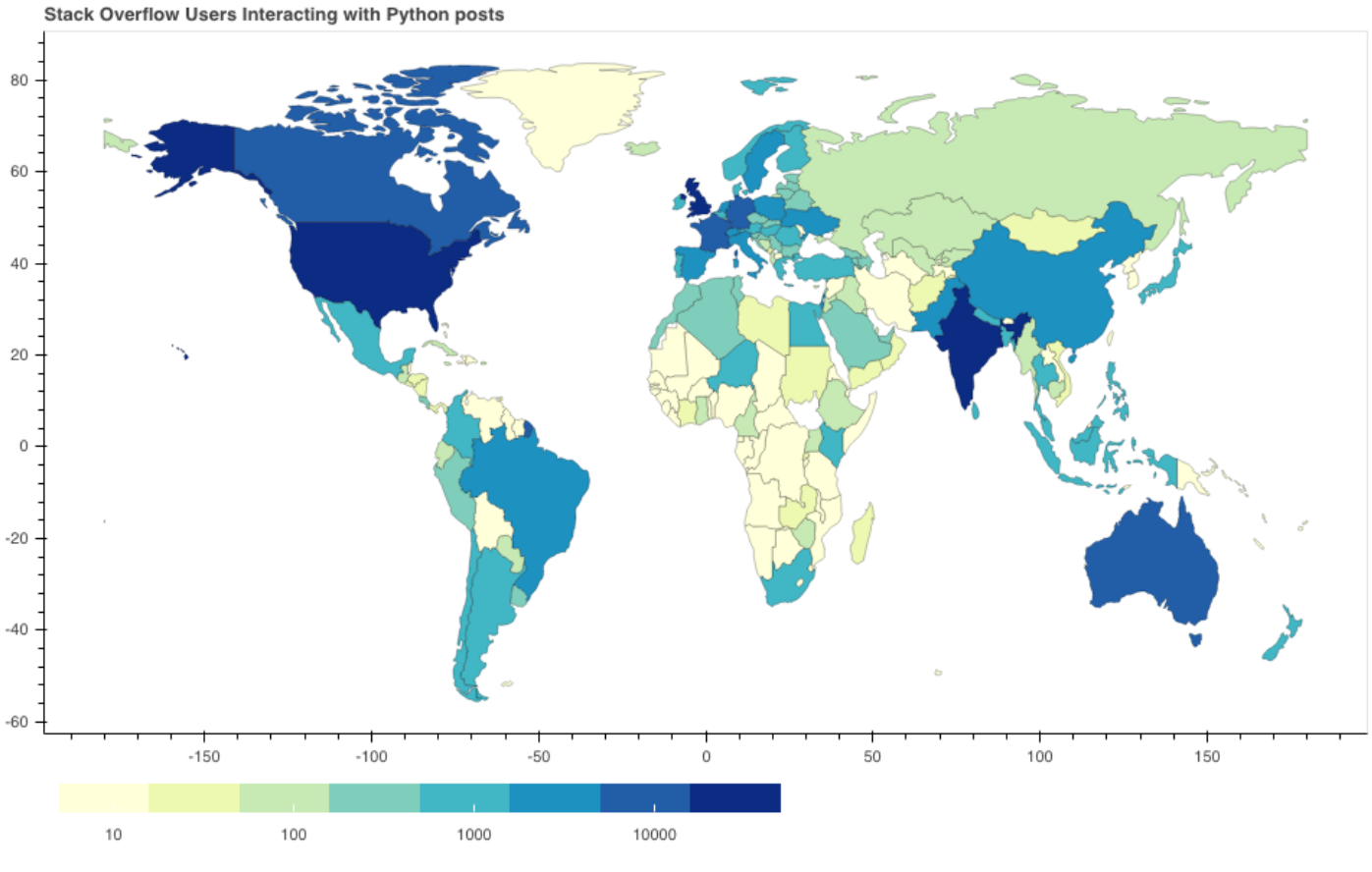


The Python tag in new posts on the website reaching over 100000 posts in the last 5 month interval and it does not seem to slow down anytime soon.
Python 2.7 had more new posts than Python 3.x till mid 2016 after which Python 3 has become the dominant tag.

Location Based Analysis

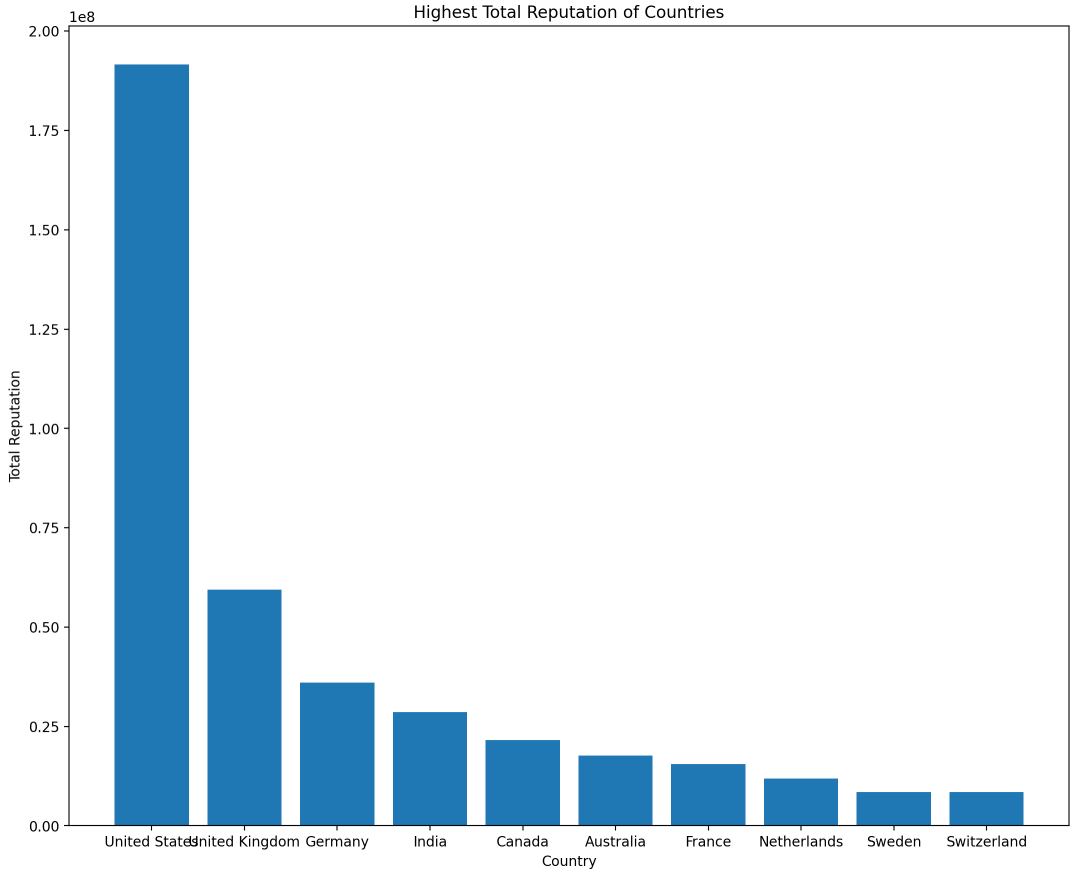
262,141 users had shared their location. By using pattern matching I was able to classify 90% of them into the country from which they belong to. Interestingly 37% of users of USA only mention their State name in their location unlike most users of rest of the world which mention their country. I suspect this has a lot of reasons including USA having a really strong federal structure, and USA being the biggest hub of IT



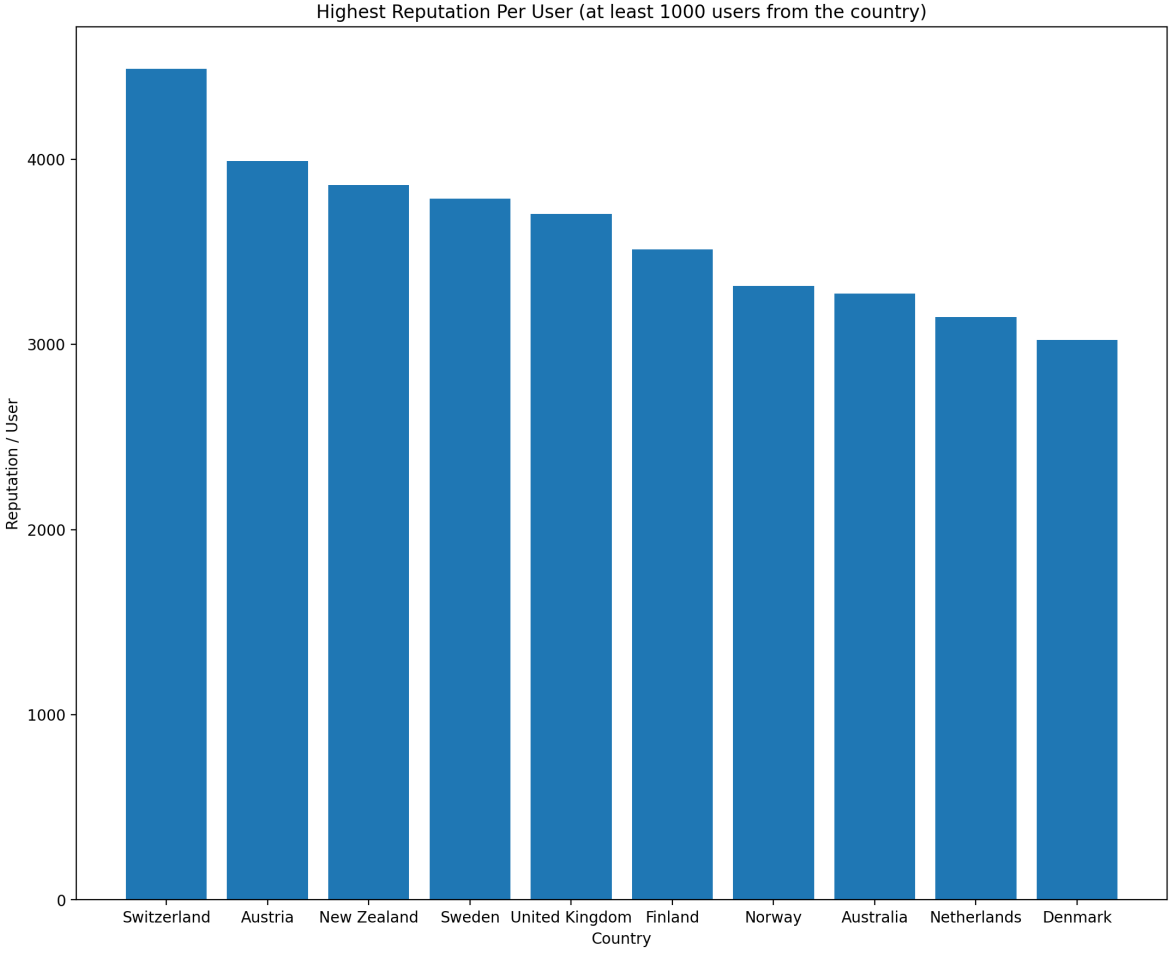


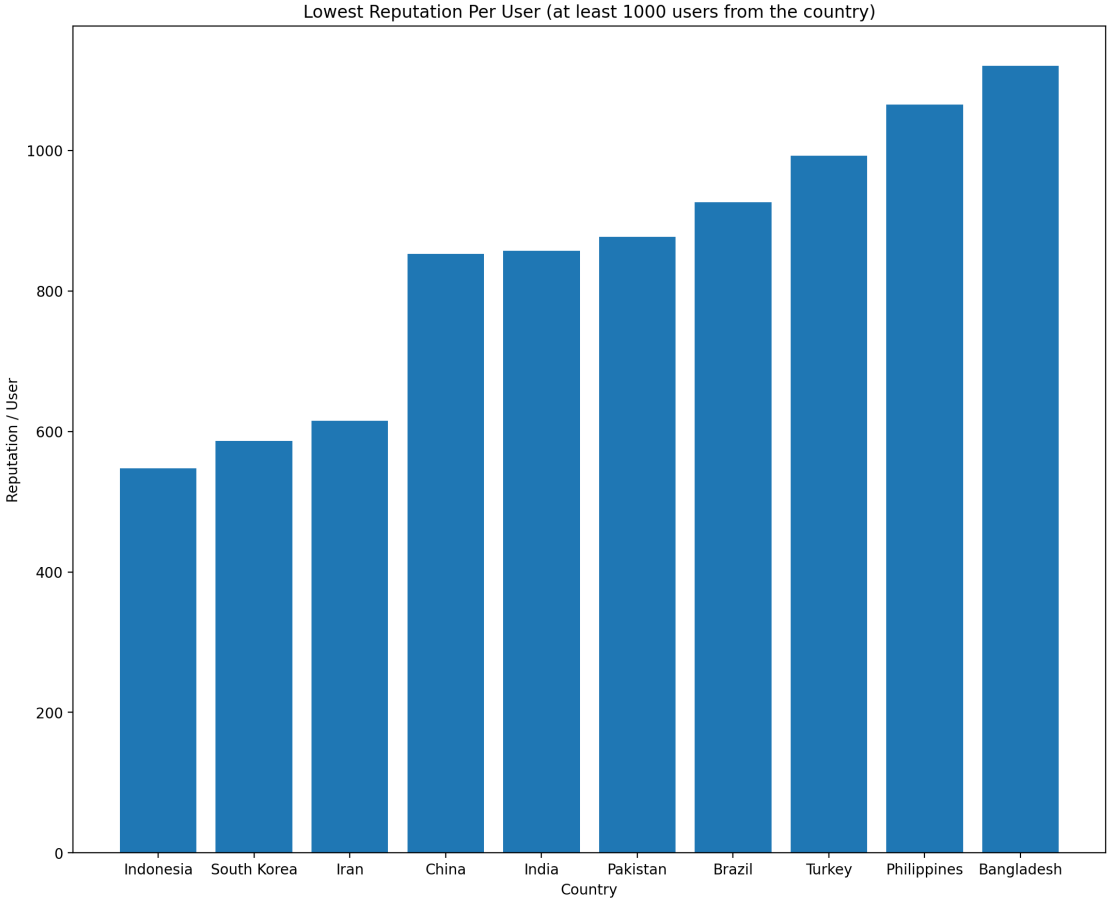
Combining Reputation and Location Analysis

Reputation is a point system on Stackoverflow which indicates how much helpful has a user has been.
Using the previous data based on location we can visualise the most helpful countries on stack overflow.



I also used the number of users per country to calculate most and least helpful country per user.
I have excluded countries with less than 1000 users since they skew the data.





I also analyzed the growth of users from different countries overtime. These growth trends are only for Python related posts but we can clearly see the boom in Indian users over the past decade as India becomes an IT hub.