Analysing StackOverflow Data

Stack Overflow is not just a Q&A site but has become an important global social community. Millions of developers use it to interact and exchange knowledge. It has become an indispensable part of life of any programmer. The site offers a platform to learn and discuss all kinds of programming doubts. The moderators are doing a great job in maintaining the quality of content on Stack Overflow. The site also encourages user participation by rewarding them with points and badges. The diverse set of users and content on Stack Overflow offer great opportunities for data analytics.

For the purpose of this task, I used the Stack Overflow data shared by PreCog. In order to study the data, I first had to parse it and store it in a MongoDB database. I have done the same in `Parse_XML.py`. After successfully storing the data in a MongoDB database, I conducted Exploratory Data Analysis on the data as found in the notebook `StackOverflowEDA.ipynb`.

Subsampling

To identify the subsampling method used, I first tried to count unique instances of each attribute in the Posts collection. Ideally, if a subset of data was chosen by applying condition on some attribute, that attribute would have some pattern in the unique values or just one unique value. However, there was not any such observable pattern. Almost all attributes were found to have a good diversity.

```
PostTypeId 2
AnswerCount 53
CommentCount 53
FavoriteCount 396
Score 1230
CommunityOwnedDate 5534
LastEditorDisplayName 5733
OwnerDisplayName 15482
ViewCount 40082
ClosedDate 84382
LastEditorUserId 262918
Tags 467829
OwnerUserId 670395
AcceptedAnswerId 730440
ParentId 1158038
Title 1358340
LastEditDate 1410408
LastActivityDate 2501726
CreationDate 3377143
Body 3380053
_id 3380601
```

I made some other observations (as given in subsequent pages). The top tags in Tags collection were Javascript, Java, C#,Python and PHP. However in the Posts collection, the only top tag was Python. This seems to be a subsampling criteria.

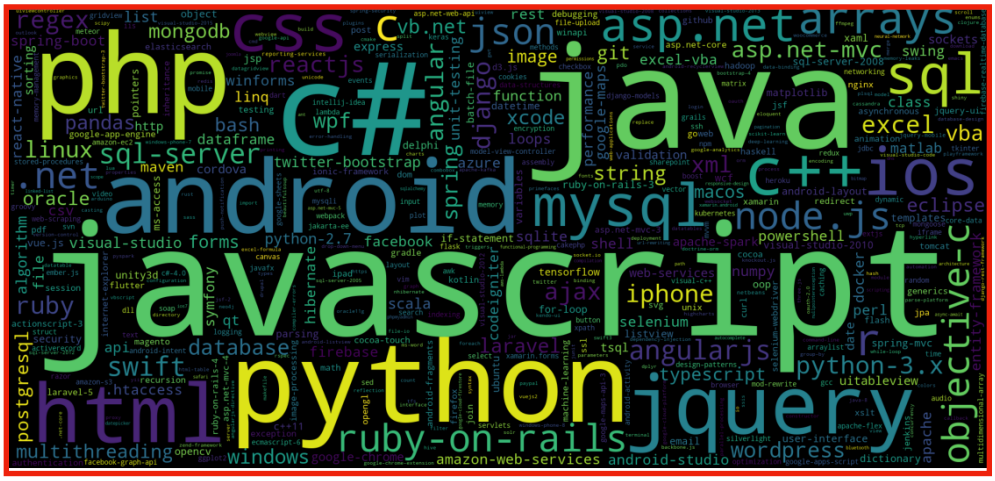
	TagName	Count
2	javascript	1955557
11	java	1641102
6	c#	1385220
10	python	1359126
4	php	1335050

My another guess at subsampling is creation date. I am not sure when the data given to use was collected but the last date in it is around March, 2020 and currently it is December, 2020. [LastModified Date on official site is of December]. This is not good parameter to judge but felt like mentioning it.

Anuneet Anand
2018022

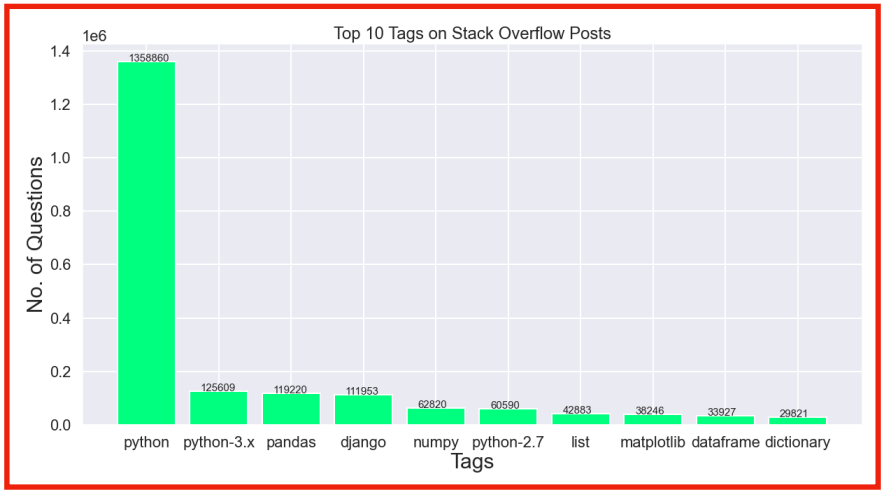
Insightful Word Cloud

I used the frequency of different tags used on Stack Overflow from Tags collection (Not Post Collection) to generate a word cloud. The word cloud beautifully captures what's trending on Stack Overflow.



Top 10 Tags In Posts

I counted number of occurrences of unique in all posts from the Posts collection and plotted a bar plot to showcase the top 10 tags.



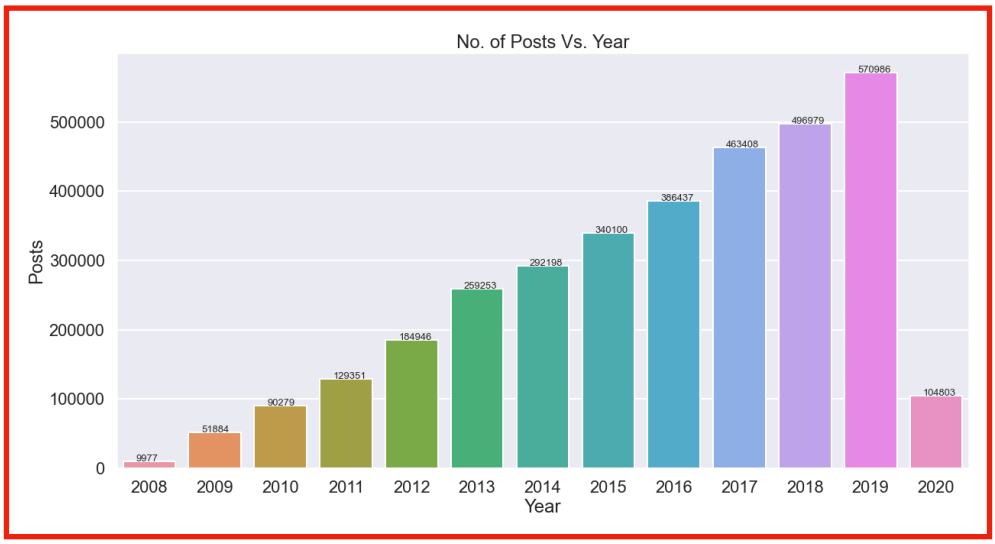

```
Total Posts : 3380601
Tagged Posts : 1358860
Unique Tags : 22512
```

Python and its associated modules seems to dominate in this dataset. No wonder since python has become one of most widely used language.

Anuneet Anand
2018022

Activity On Stack Overflow Over The Years

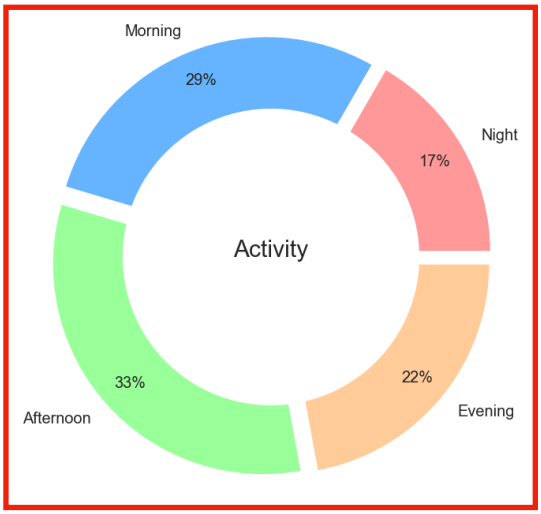
I counted the number of posts posted on Stack Overflow in each year and the results were something which we can all agree with. Stack Overflow community is growing!



Note : The bar of 2020 is short as the dataset did not have Posts data of the entire year.

When Are Developers Most Active?

I analysed the time at which the posts were made and gained some interesting insights. Maximum number of questions were asked between 12:00 and 18:00.



Bucket	Period
Morning	06:00 - 12:00
Afternoon	12:00 - 18:00
Evening	18:00 - 00:00
Night	00:00 - 06:00

Anuneet Anand
2018022

Distribution Of Scores

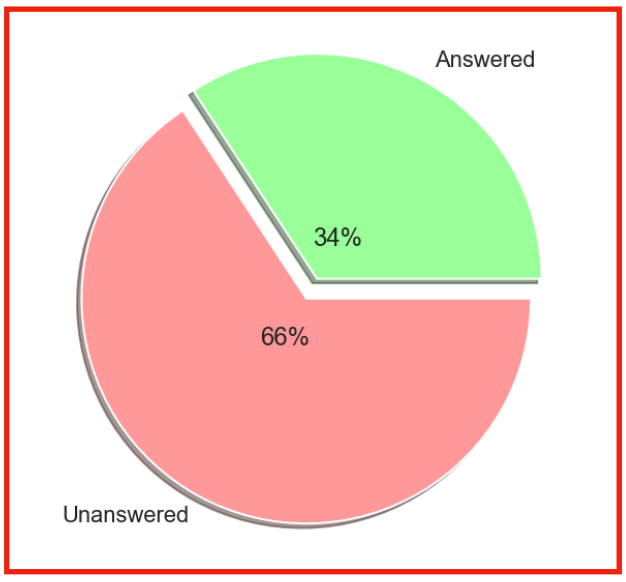
I analysed the statistics and distribution of scores from Posts and noted the following insights.

- The scores on the post ranged from -64 to 14282, showing a big difference.
- The mean score was merely 3 and the 75th percentile was also as low as 2.
- The standard deviation was found to be around 30.
- This shows that most of scores are concentrated around 0 to 3 and only few posts achieve really high scores.

Statistic	Value
count	3380601.00
mean	3.04
std	30.27
min	-64.00
25%	0.00
50%	1.00
75%	2.00
max	14282.00

How Many Questions Are Left Unanswered?

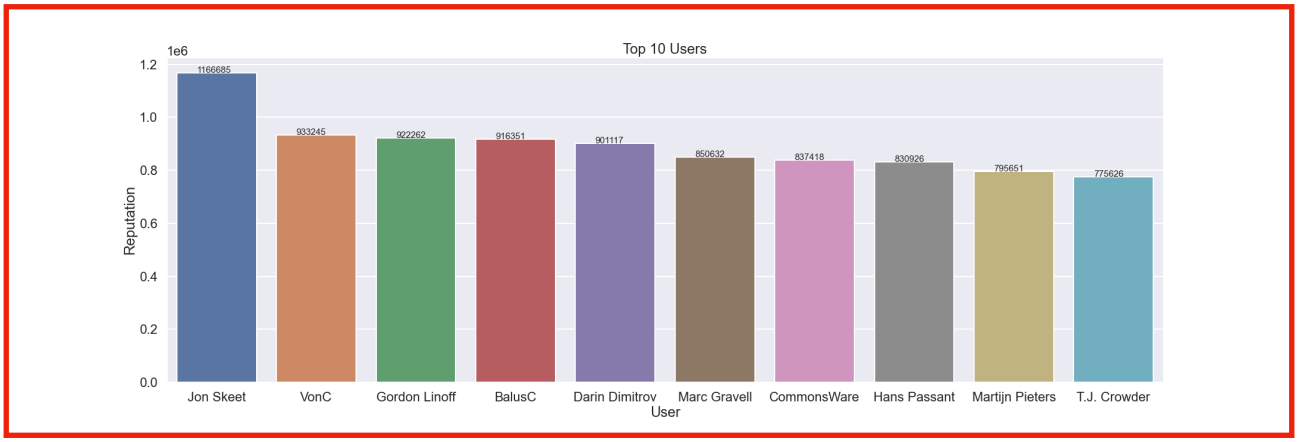
I counted the number of posts which had non-zero answer count. Almost 66% of the questions were found to have no accepted answers. More dedication is required from the user side to improve this ratio.



Anuneet Anand
2018022

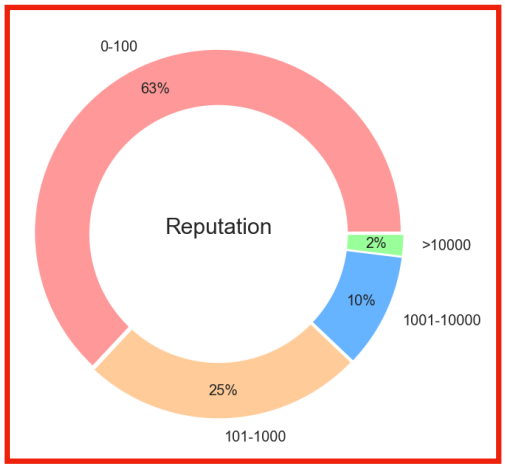
Top 10 Reputed Users

I sorted the users based on their reputation scores. Jon Skeet top the charts with an amazing reputation of 1166685, followed by VonC and Gordon Linoff. The reputation feature introduces the gamification factor on Stack Overflow and provide incentives to the users to contribute to the welfare of the community.



Distribution Of Reputation

I analysed the distribution of reputation scores of users and arranged them in buckets. Majority of users (63%) were found to have a reputation score less than 100.



Anuneet Anand
2018022