# Exploratory Data Analysis

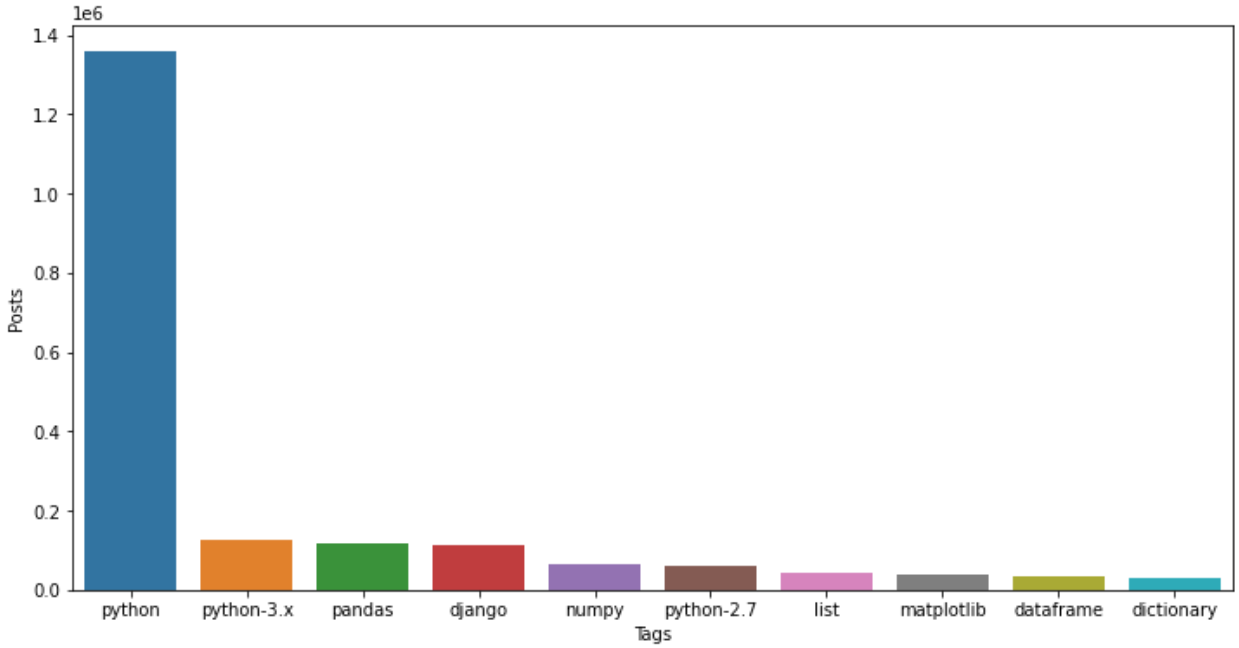
Before writing this report, I would like to thank the team at PreCog for setting up these tasks. I had no prior experience in most of the tools required to complete them and slowly learning them over the week has been an amazing experience.
I would also like to add there is a very good chance that some of my observations and analysis might be way off base 😅.

**Figure out what criteria did we use for subsampling (i.e. what is common factor amongst the whole dataset)**
The extremely large size of the dataset makes it rather difficult to manually explore and find any patterns in the data.
To remedy this one approach is to use just a small random subset of the data.
We begin by trying this approach on the tags collection. No success was achieved here because no common theme appeared in the rather small number of fields. Neither in name or any of the ids.
We next move our attention to the posts collection. Here we found a possible point of commonality in the data. Printing the first 100 or so documents. It seems that any post that is tagged (i.e. have the Tags field), has a python tag. Increasing the count of documents to 500, we see that the hypothesis holds.
Next we write some code to iterate over all the posts and if they have the Tag attribute check for the python tag. The code seems to confirm our conjecture.
We conclude that posts were sampled by filtering out those whose tags did not contain python.

**Include one meaningful Word Cloud**


This word cloud shows the display names for the users with the sizes corresponding to the views received by them.

**Draw a barplot of the top-10 occuring tags. The bar plot will show the number of questions on the y axis, and the name of each tag on the x axis**



**Some observations**
Similar to the twitter ecosystem, multiple metrics are related here. A user with a high reputation is correlated with the number of upvotes they have received.

The upvotes, downvotes, views and reputation lie closer to the lower end and are close to zero for most of the users.

Note:
Due to the extremely large size of the data, it is rather difficult to explore the data by trying different subsets of fields and creating any meaningful visualisation.
It is easy to create trivial charts for example the distribution of posts over months, account creation over years, etc. However to create something that tells the viewer some useful information requires the use of techniques and tools which I need to learn more about.

I would love links to a few resources that teach fast data exploration on large datasets that don't just boil down to get better hardware.