***Data Processing in R and Python***

***Third Assignment Report***

Introduction

*The objective of this assignment is to create and plot fascinating results from the data downloaded from the Stackexchange website. By using different data analyzing and processing techniques.*

*I have imported five different datasets from the website related to the topics:*

* *Computer Science*
* *Data Science*
* *Cryptocurrency*
* *Money*
* *Bitcoin*
* *Ehtereum*

*This assignment is divided into two parts lets first start by importing the libraries and data.*

*Importing the libraries :*

Graphical user interface

Description automatically generated

Matplotlib:

* Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications.

Seaborn:

* Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Numpy:

* NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Pandas:

* pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Importing the required data:

The downloaded data was originally in the XML format, but has been converted to CSV using python script.

Text

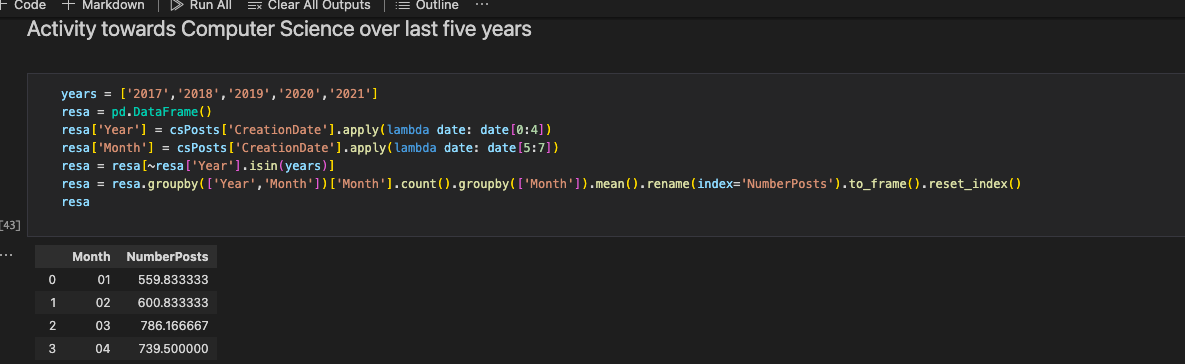
Description automatically generated

***First part:***

In this section, I have used the two datasets Computer and Data science and

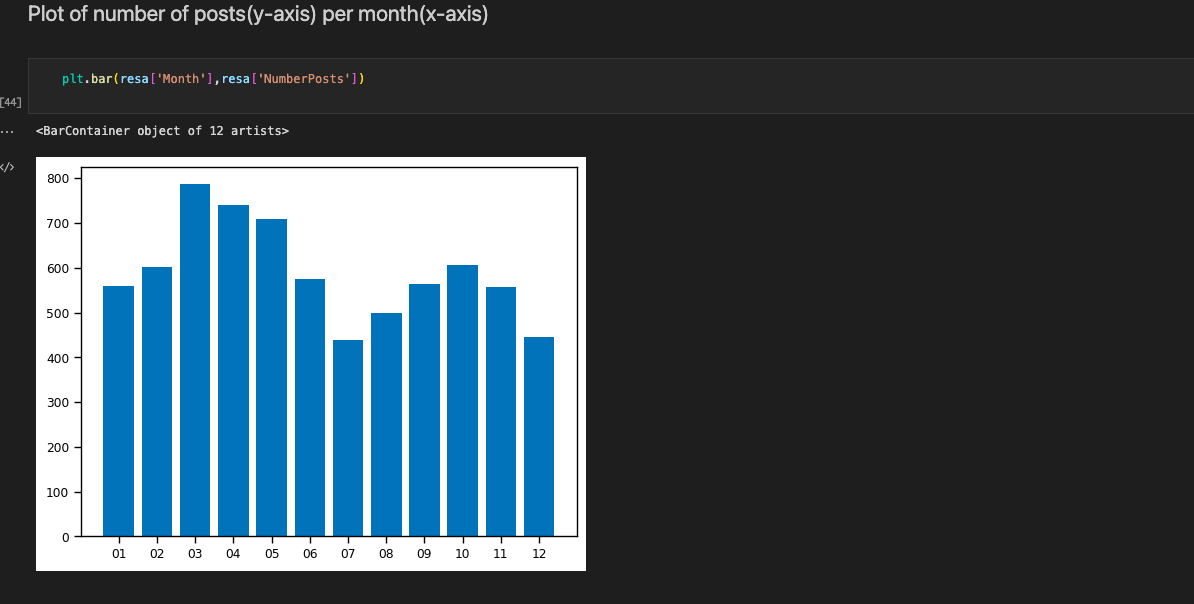
tried to find out the most popular topic among the two in the previous five years, by

using the following method:

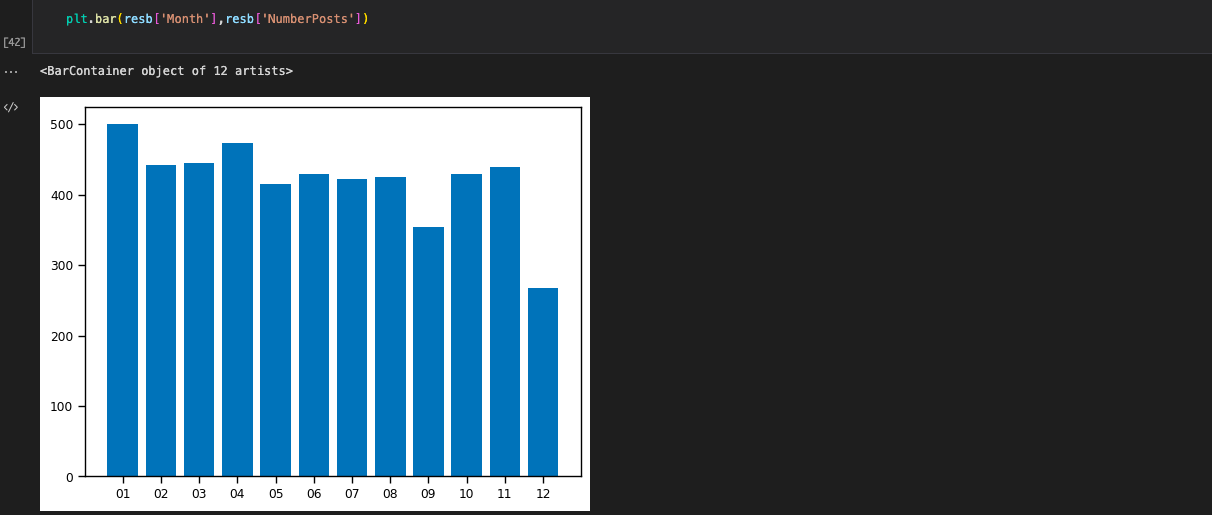


I tried to compare the audience activity towards these topics by counting the number of posts for each topic in different months of a year.

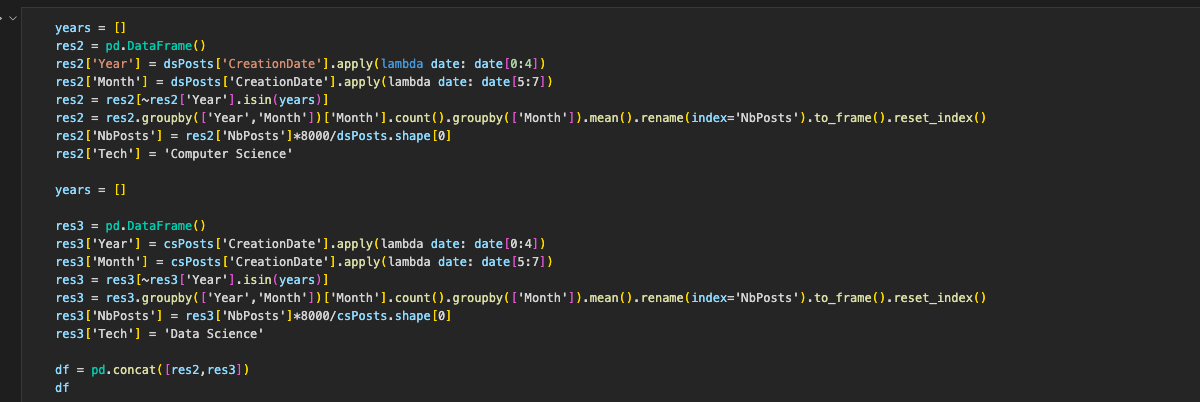
Using the matplot library I plotted the graph for number of posts in different months for the Computer science dataset.



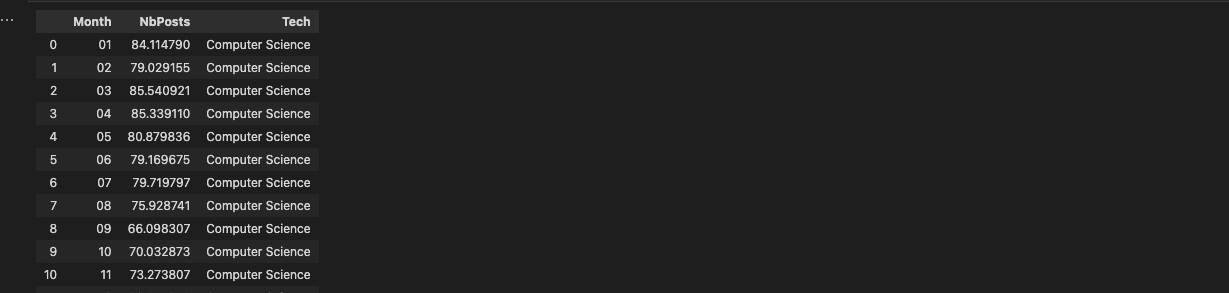
The same methods have been applied for Data science dataset and the following graph has been generated:



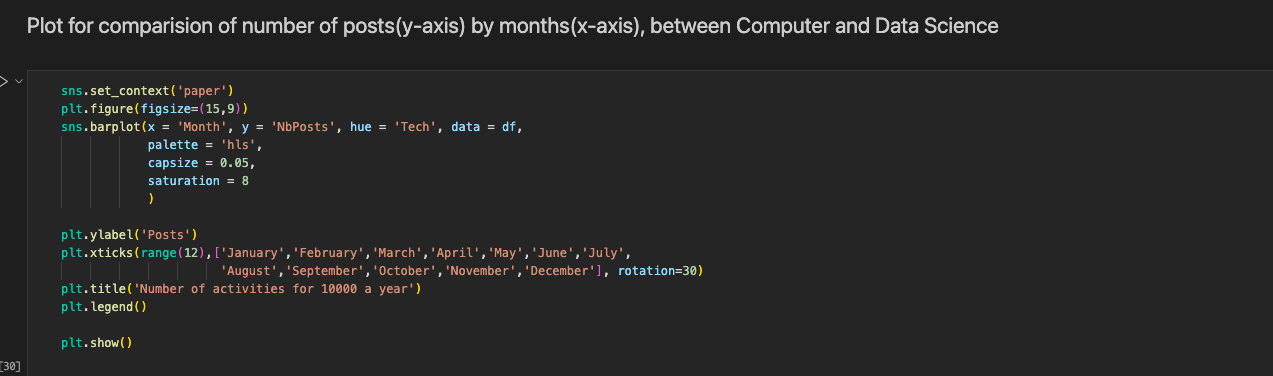
After that using “concat function” from the “pandas” library, I have joined the two comparisons:



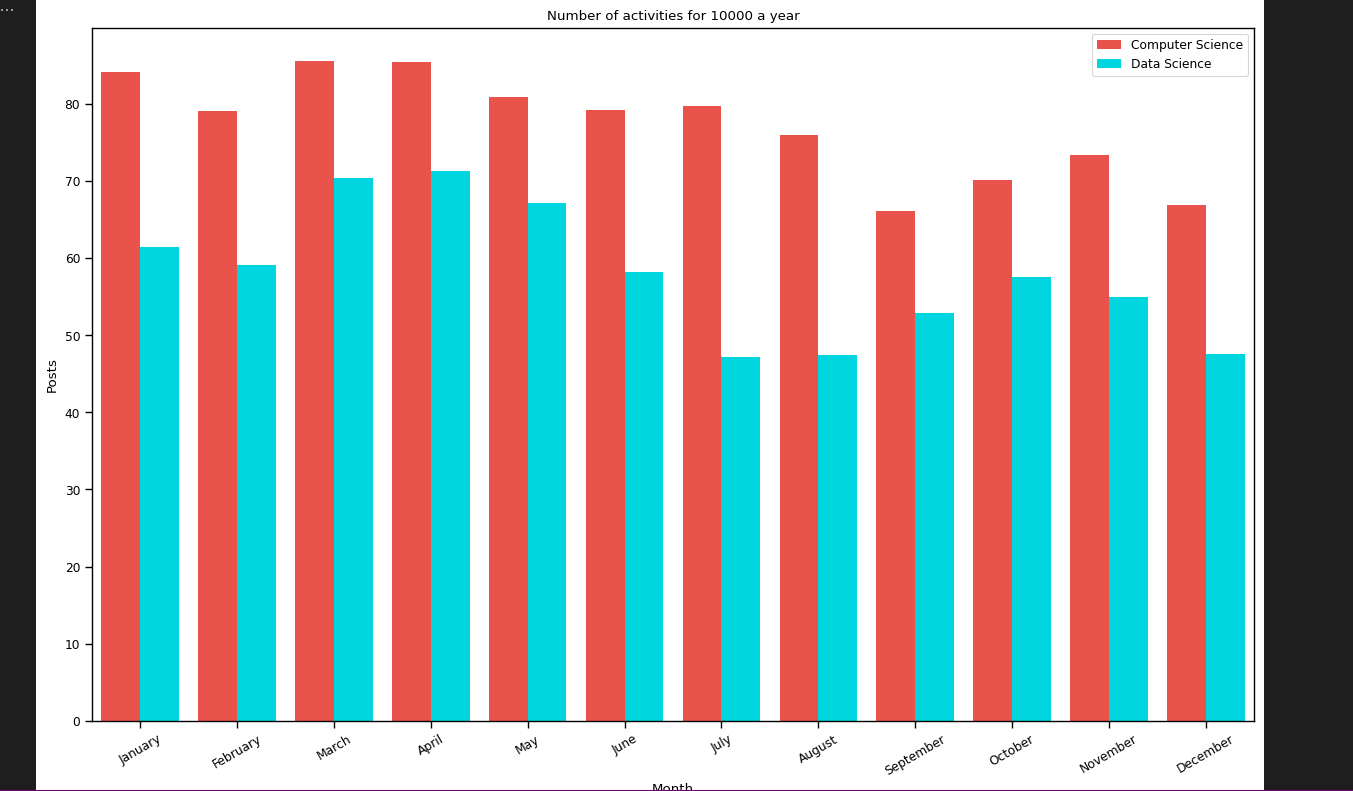
The obtained result is as follows:



By using seaborn and matplotlib library the plot has been generated which compares the activities of these topics in different months of the last five years. The code used for this is as follows:



The resulted plot is as follows:



- Conclusion:

We can clearly see that altough Data science has created so much hype and it is very emerging then also stats for Computer science is more in comparision. Which clearly states the interest of the audience.

***Second part:***

In this section, I have used four datasets Bitcoin, Ethereum, Cryptocurrency and Money. The approach was to know how frequently these cryptocurrencies (bitcoin and Ethereum) have been discussed in the “Money market and Crypto ecosystem,

By counting the number of times these tags have been used.

After that comparing these two cryptocurrencies to support the resulting outcomes.

The function used for counting the tags namely “ethereum and bitcoin”.

Using the pandas library the following tags were searched and concatenated from the two datasets:

Text

Description automatically generated

The desired outcome counting the repetition of each tag.

Graphical user interface

Description automatically generated with medium confidence

The following code was used in order to get the final plot for the searched tags:



The final plot:

Chart, bar chart

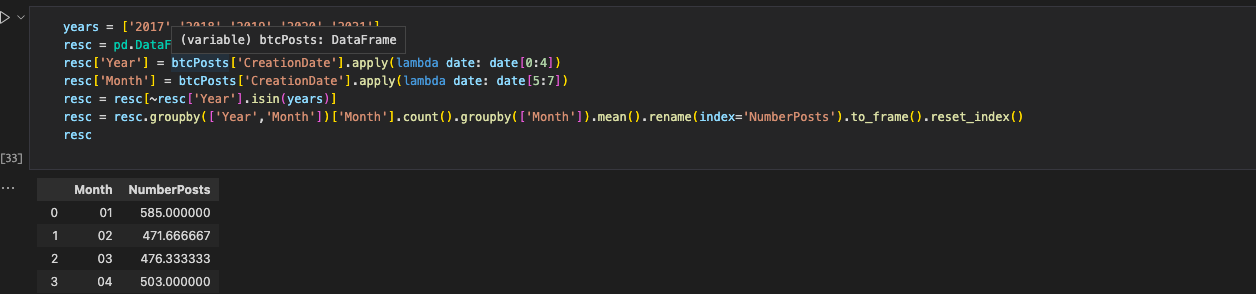
Description automatically generated

Conclusion:

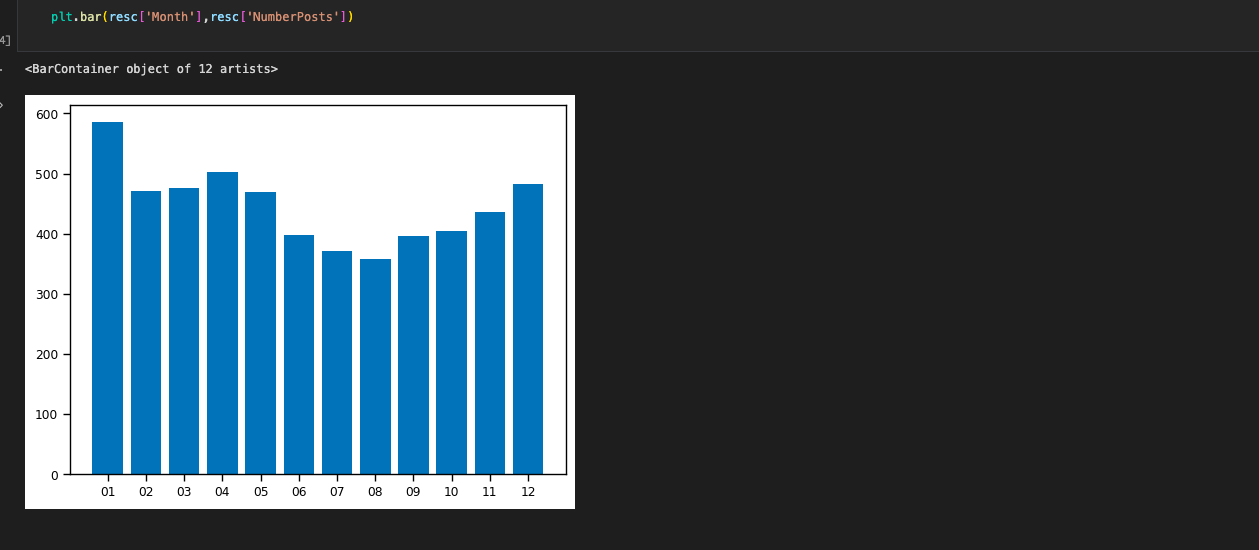
From the plots it’s clearly visible that bitcoin is tag which is more frequently used which means this topic is in heat now.

To prove this contradiction lets count the number of posts in last five years for both Bitcoin and Ethereum.

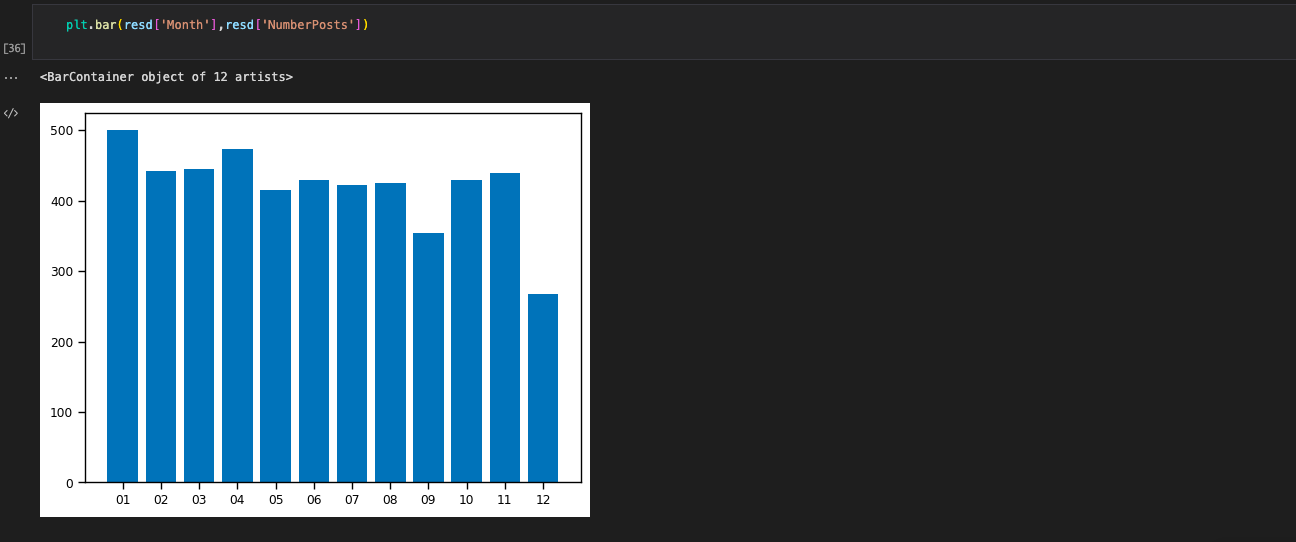
The following code was used to count the activity towards Bitcoin:



The resulting plot:



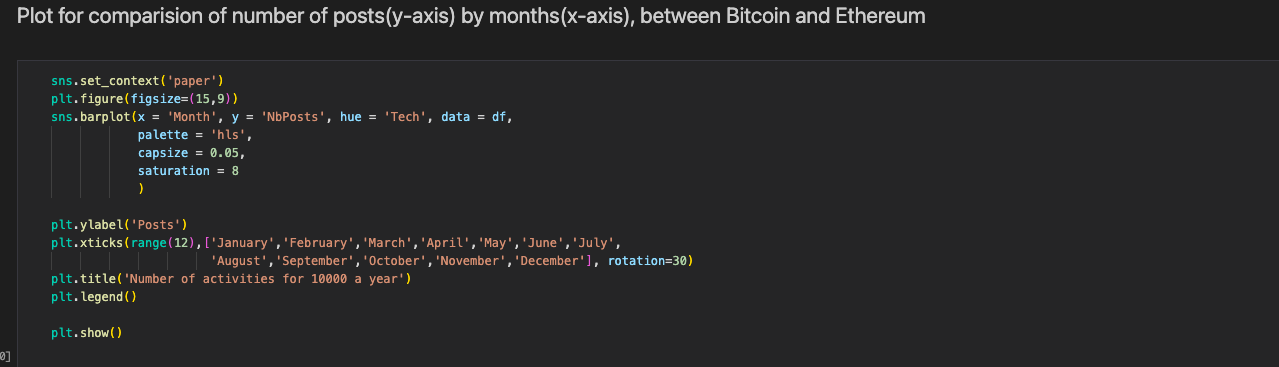
The same for Ethereum and here is the resulting plot:



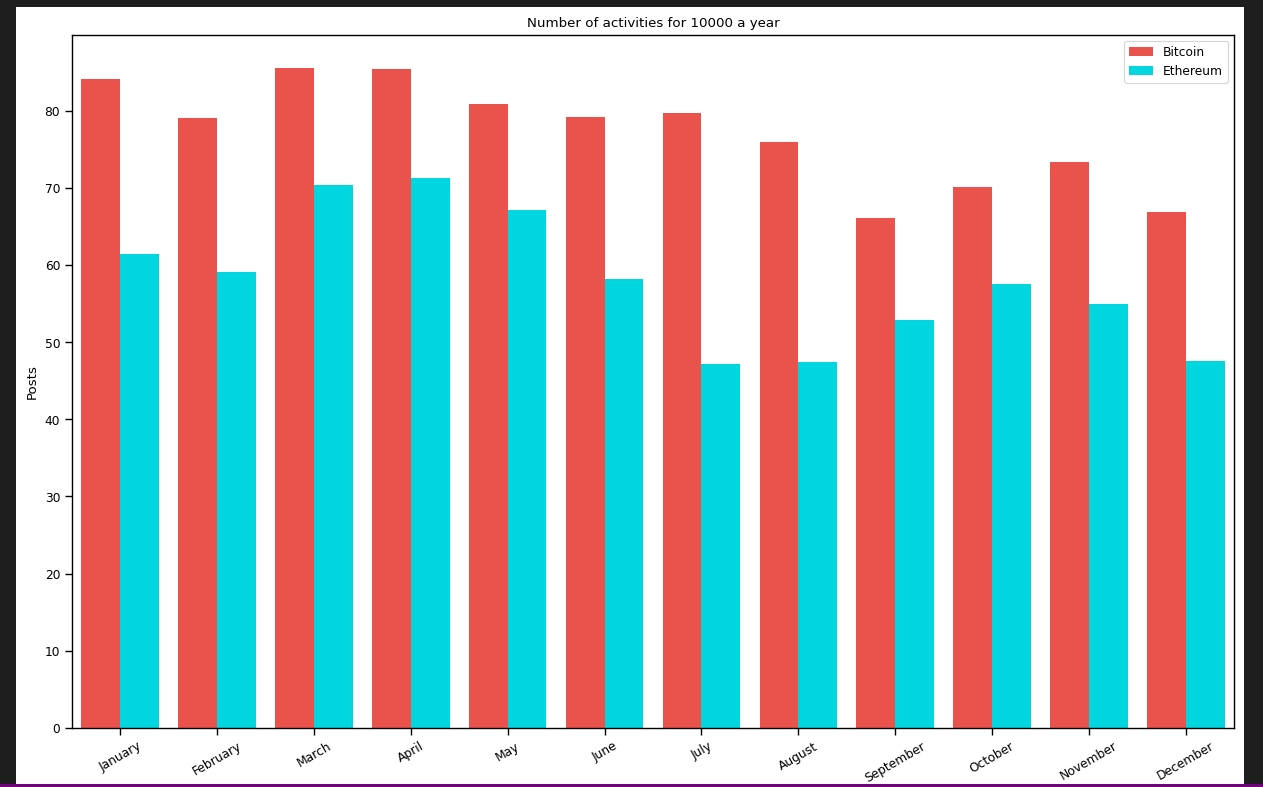
After that using “concat function” from the “pandas” library, I have joined the two comparisons:



By using seaborn and matplotlib library the plot has been generated which compares the activities of these topics in different months of the last five years. The code used for this is as follows:



The resulting plot:



Conclusion:

It’s true from the above plot that it is Bitcoin who has more number of posts in comparison with Ethereum, no matter what month or year it is.

This proves the previous contradiction that as in the Money and Crypto market Bitcoin is more discussed that’s why it has more number of posts also.

Which leads me to this conclusion that the rates of Bitcoin are always higher than Ethereum and also it is more consistent.

Here is the graph from google containing some stats which proves this point.

Graphical user interface, application

Description automatically generated

By: Naveen Tiwari

K6971