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 :

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
```

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.

```
# Data Science
dsPosts = pd.read_csv(r'/Users/naveen/Downloads/dprpy3/ds/Posts.csv')
dsComments = pd.read_csv(r'/Users/naveen/Downloads/dprpy3/ds/Comments.csv')

# Computer Science
csPosts = pd.read_csv(r'/Users/naveen/Downloads/dprpy3/cs/Posts.csv')
csComments = pd.read_csv(r'/Users/naveen/Downloads/dprpy3/cs/Comments.csv')

# Bitcoin
btcPosts = pd.read_csv(r'/Users/naveen/Downloads/dprpy3/btc/Posts.csv')
btcComments = pd.read_csv(r'/Users/naveen/Downloads/dprpy3/btc/Comments.csv')

# Ethereum
ethPosts = pd.read_csv(r'/Users/naveen/Downloads/dprpy3/eth/Posts.csv')
ethComments = pd.read_csv(r'/Users/naveen/Downloads/dprpy3/eth/Comments.csv')

# Crypto
cryPosts = pd.read_csv(r'/Users/naveen/Downloads/dprpy3/cry/Posts.csv')

#Money
monPosts = pd.read_csv(r'/Users/naveen/Downloads/dprpy3/mon/Posts.csv')
```

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:

Activity towards Computer Science over last five years

```
years = ['2017','2018','2019','2020','2021']
resa = pd.DataFrame()
resa['Year'] = csPosts['CreationDate'].apply(lambda date: date[0:4])
resa['Month'] = csPosts['CreationDate'].apply(lambda date: date[5:7])
resa = resa[~resa['Year'].isin(years)]
resa = resa.groupby(['Year','Month'])['Month'].count().groupby(['Month']).mean().rename(index='NumberPosts').to_frame().reset_index()
resa
```

[43]

	Month	NumberPosts
0	01	559.833333
1	02	600.833333
2	03	786.166667
3	04	739.500000

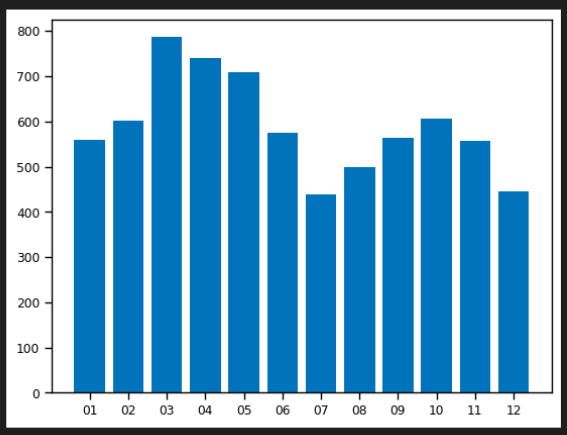
I tried to compare the audience activity towards these topics by counting the number of posts for each topic in different months of a year.
Using the matplot library I plotted the graph for number of posts in different months for the Computer science dataset.

Plot of number of posts(y-axis) per month(x-axis)

```
plt.bar(resa['Month'],resa['NumberPosts'])
```

[44]

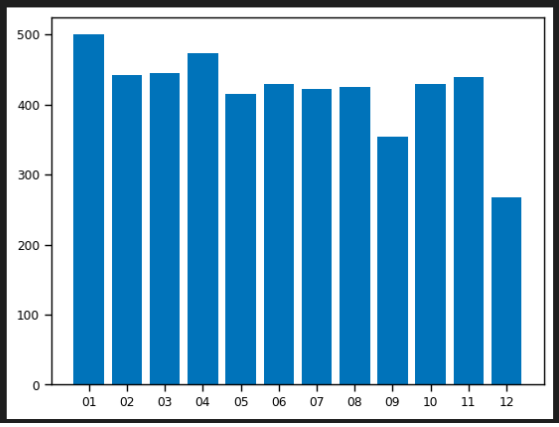
```
<BarContainer object of 12 artists>
```



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

```
[42]  plt.bar(resb['Month'],resb['NumberPosts'])
```

```
<BarContainer object of 12 artists>
```



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

```
years = []
res2 = pd.DataFrame()
res2['Year'] = dsPosts['CreationDate'].apply(lambda date: date[0:4])
res2['Month'] = dsPosts['CreationDate'].apply(lambda date: date[5:7])
res2 = res2[~res2['Year'].isin(years)]
res2 = res2.groupby(['Year','Month'])['Month'].count().groupby(['Month']).mean().rename(index='NbPosts').to_frame().reset_index()
res2['NbPosts'] = res2['NbPosts']*8000/dsPosts.shape[0]
res2['Tech'] = 'Computer Science'

years = []

res3 = pd.DataFrame()
res3['Year'] = csPosts['CreationDate'].apply(lambda date: date[0:4])
res3['Month'] = csPosts['CreationDate'].apply(lambda date: date[5:7])
res3 = res3[~res3['Year'].isin(years)]
res3 = res3.groupby(['Year','Month'])['Month'].count().groupby(['Month']).mean().rename(index='NbPosts').to_frame().reset_index()
res3['NbPosts'] = res3['NbPosts']*8000/csPosts.shape[0]
res3['Tech'] = 'Data Science'

df = pd.concat([res2,res3])
df
```

The obtained result is as follows:

	Month	NbPosts	Tech
0	01	84.114790	Computer Science
1	02	79.029155	Computer Science
2	03	85.540921	Computer Science
3	04	85.339110	Computer Science
4	05	80.879836	Computer Science
5	06	79.169675	Computer Science
6	07	79.719797	Computer Science
7	08	75.928741	Computer Science
8	09	66.098307	Computer Science
9	10	70.032873	Computer Science
10	11	73.273807	Computer Science

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:

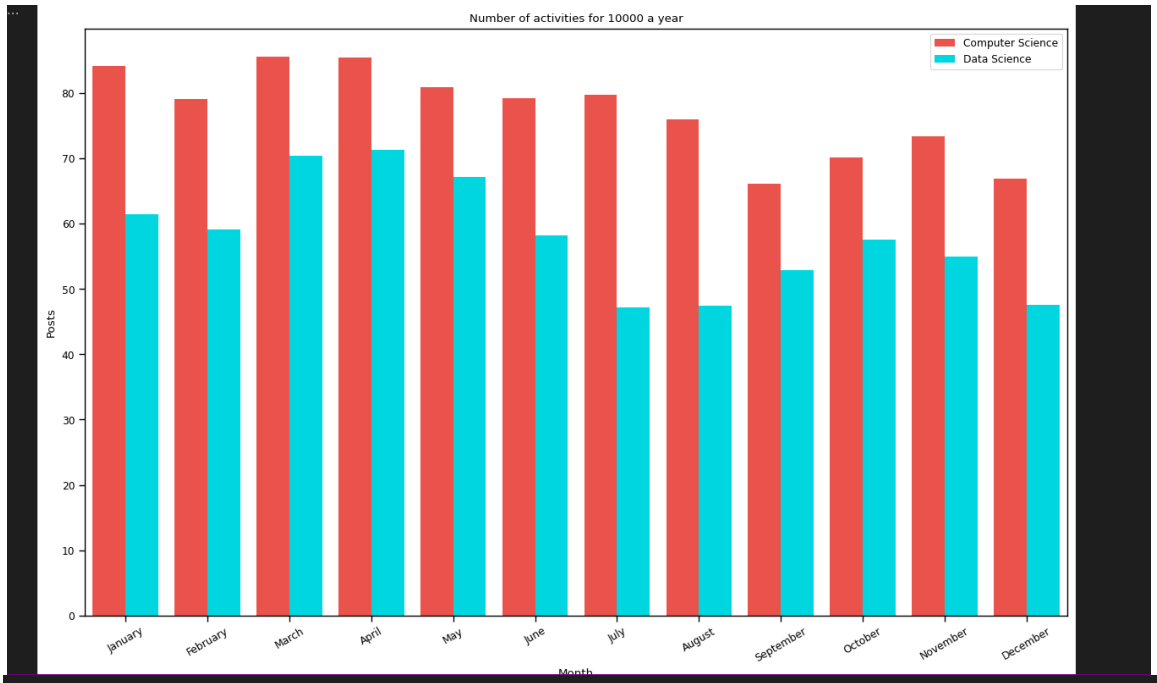
Plot for comparision of number of posts(y-axis) by months(x-axis), between Computer and Data Science

```
sns.set_context('paper')
plt.figure(figsize=(15,9))
sns.barplot(x = 'Month', y = 'NbPosts', hue = 'Tech', data = df,
            palette = 'hls',
            capsize = 0.05,
            saturation = 8
            )

plt.ylabel('Posts')
plt.xticks(range(12),['January','February','March','April','May','June','July',
                      'August','September','October','November','December'], rotation=30)
plt.title('Number of activities for 10000 a year')
plt.legend()

plt.show()
```

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:

```
import re
def detect_words(titles,words):
    lst = [
        int(titles.str.count(w).dropna().sum())
        for w in words
    ]
    return np.array(lst)
```

[35] ✓ 0.0s Python

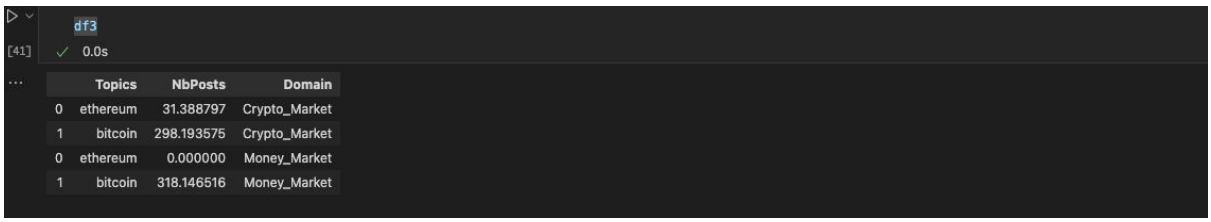
```
df = pd.DataFrame()
df['Topics'] = ['ethereum','bitcoin']
df['NbPosts'] = detect_words(cryPosts['Title'], df['Topics'])*1000000/cryPosts.shape[0]
df['Domain'] = 'Crypto_Market'

df2 = pd.DataFrame()
df2['Topics'] = ['ethereum','bitcoin']
df2['NbPosts'] = detect_words(monPosts['Title'], df['Topics'])*1000000/monPosts.shape[0]
df2['Domain'] = 'Money_Market'

df3 = pd.concat([df,df2])
```

[40] ✓ 0.2s Python

The desired outcome counting the repetition of each tag.



```
df3
```

[41] ✓ 0.0s

	Topics	NbPosts	Domain
0	ethereum	31.388797	Crypto_Market
1	bitcoin	298.193575	Crypto_Market
0	ethereum	0.000000	Money_Market
1	bitcoin	318.146516	Money_Market

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

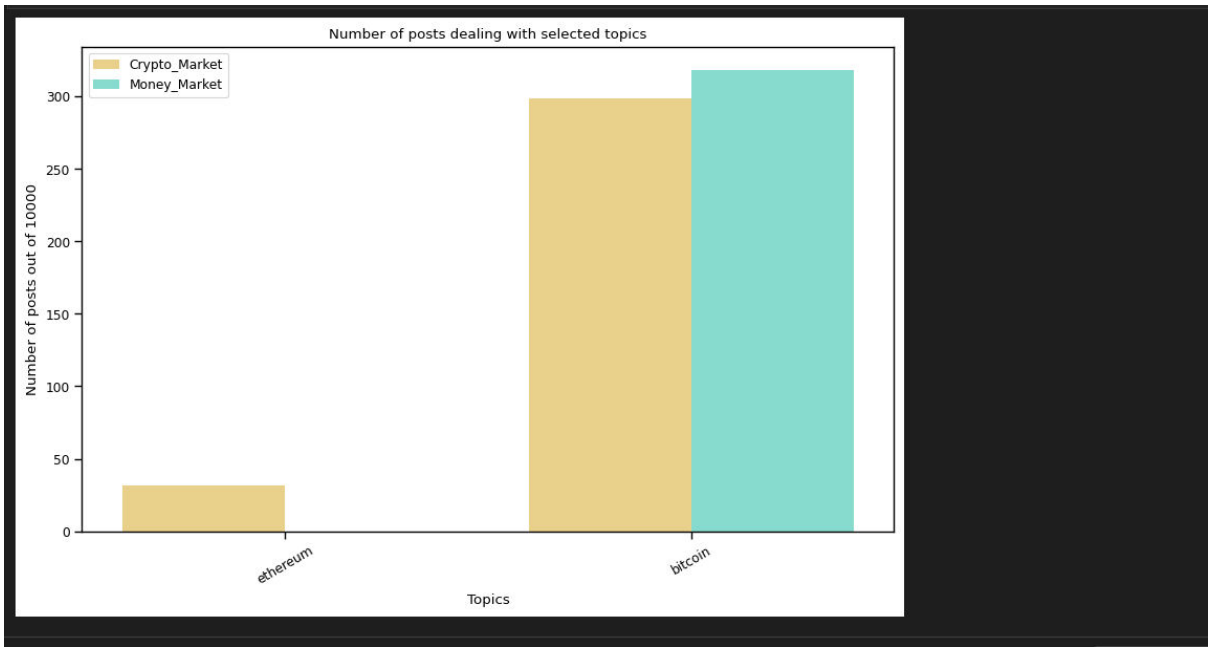
```
sns.set_context('paper')

# create plot
plt.figure(figsize=(10,6))
sns.barplot(x = 'Topics', y = 'NbPosts', hue = 'Domain', data = df3,
            palette = 'BrBG',
            capsize = 0.05,
            saturation = 8
            )

plt.ylabel('Number of posts out of 10000')
plt.xticks(range(2),['ethereum','bitcoin'], rotation=30)
plt.title('Number of posts dealing with selected topics')
plt.legend()

plt.show()
```

The final plot:



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:

```
years = ['2017', '2018', '2019', '2020', '2021']
resc = pd.DataF (variable) btcPosts: DataFrame
resc['Year'] = btcPosts['CreationDate'].apply(lambda date: date[0:4])
resc['Month'] = btcPosts['CreationDate'].apply(lambda date: date[5:7])
resc = resc[~resc['Year'].isin(years)]
resc = resc.groupby(['Year','Month'])['Month'].count().groupby(['Month']).mean().rename(index='NumberPosts').to_frame().reset_index()
resc
```

[33]

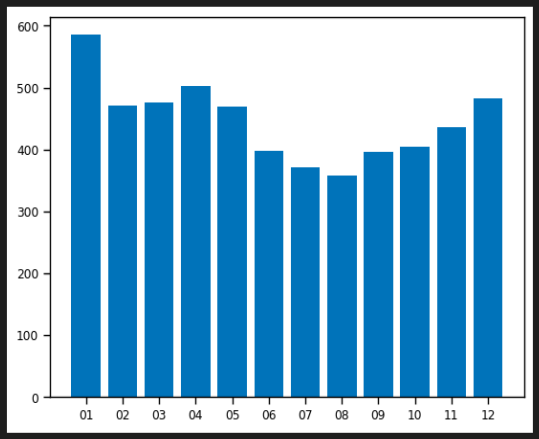
	Month	NumberPosts
0	01	585.000000
1	02	471.666667
2	03	476.333333
3	04	503.000000

The resulting plot:

```
plt.bar(resc['Month'],resc['NumberPosts'])
```

[4]

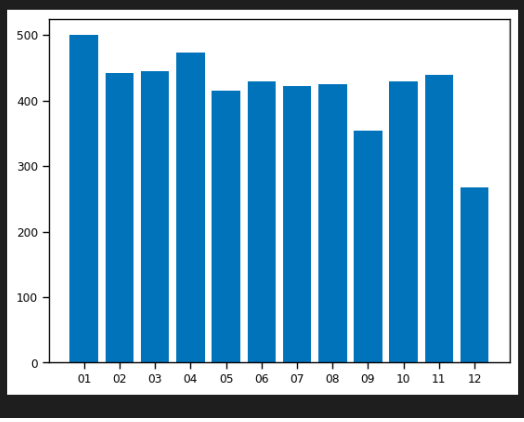
```
<BarContainer object of 12 artists>
```



The same for Ethereum and here is the resulting plot:

```
[36]  plt.bar(resd['Month'],resd['NumberPosts'])

...   <BarContainer object of 12 artists>
```



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

```
[39]  years = []
      res4 = pd.DataFrame()
      res4['Year'] = dsPosts['CreationDate'].apply(lambda date: date[0:4])
      res4['Month'] = dsPosts['CreationDate'].apply(lambda date: date[5:7])
      res4 = res4[~res4['Year'].isin(years)]
      res4 = res4.groupby(['Year','Month'])['Month'].count().groupby(['Month']).mean().rename(index='NbPosts').to_frame().reset_index()
      res4['NbPosts'] = res4['NbPosts']*8000/dsPosts.shape[0]
      res4['Tech'] = 'Bitcoin'

      years = []

      res5 = pd.DataFrame()
      res5['Year'] = csPosts['CreationDate'].apply(lambda date: date[0:4])
      res5['Month'] = csPosts['CreationDate'].apply(lambda date: date[5:7])
      res5 = res5[~res5['Year'].isin(years)]
      res5 = res5.groupby(['Year','Month'])['Month'].count().groupby(['Month']).mean().rename(index='NbPosts').to_frame().reset_index()
      res5['NbPosts'] = res5['NbPosts']*8000/csPosts.shape[0]
      res5['Tech'] = 'Ethereum'

      df = pd.concat([res4,res5])
      df
```

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:

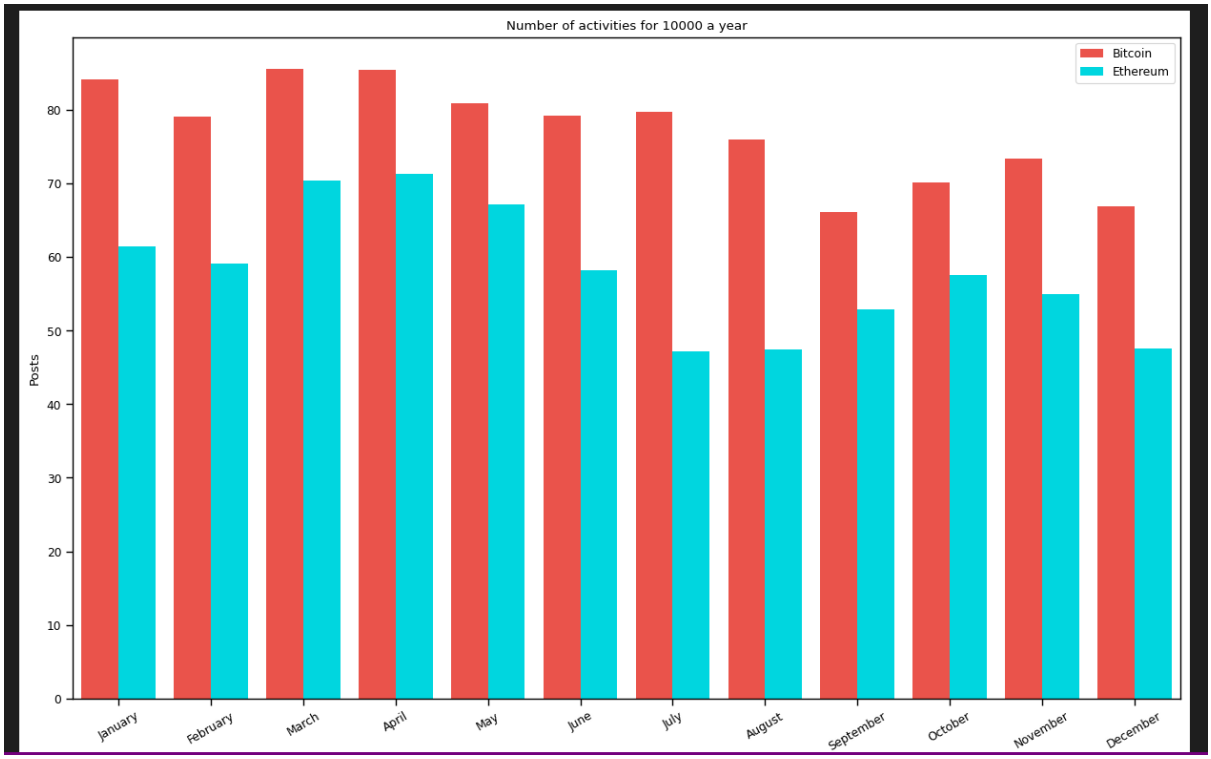
Plot for comparision of number of posts(y-axis) by months(x-axis), between Bitcoin and Ethereum

```
      sns.set_context('paper')
      plt.figure(figsize=(15,9))
      sns.barplot(x = 'Month', y = 'NbPosts', hue = 'Tech', data = df,
                  palette = 'hls',
                  capsize = 0.05,
                  saturation = 8
                  )

      plt.ylabel('Posts')
      plt.xticks(range(12),['January','February','March','April','May','June','July',
                            'August','September','October','November','December'], rotation=30)
      plt.title('Number of activities for 10000 a year')
      plt.legend()

      plt.show()
```

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.

By: Naveen Tiwari
K6971