```
!pip install gdown
!gdown --id 1eHChVwaAwG66p9eDhISOROPIvKXnMg9W
```

```
    Requirement already satisfied: gdown in /usr/local/lib/python3.10/dist-packages (4.7.3)
    Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from gdown) (3.13.3)
    Requirement already satisfied: requests[socks] in /usr/local/lib/python3.10/dist-packages (from gdown) (2.31.0)
    Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages (from gdown) (1.16.0)
    Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from gdown) (4.66.2)
    Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.10/dist-packages (from gdown) (4.12.3)
    Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.10/dist-packages (from beautifulsoup4->gdown) (2.5)
    Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests[socks]->gdown) (3
    Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests[socks]->gdown) (3.6)
    Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests[socks]->gdown) (2.0.7)
    Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests[socks]->gdown) (2024.2.
    Requirement already satisfied: PySocks!=1.5.7,>=1.5.6 in /usr/local/lib/python3.10/dist-packages (from requests[socks]->gdown) (1.7
    /usr/local/lib/python3.10/dist-packages/gdown/cli.py:138: FutureWarning: Option `--id` was deprecated in version 4.3.1 and will be
      warnings.warn(
    Downloading...
    From (original): https://drive.google.com/uc?id=1eHChVwaAwG66p9eDhISOROPIvKXnMg9W
    From (redirected): https://drive.google.com/uc?id=1eHChVwaAwG66p9eDhISOROPIvKXnMg9W&confirm=t&uuid=8bcb1e52-ced8-4f6b-a40b-05992da2
    To: /content/Arjun_Assignment_data-20220427T165022Z-002.zip
    100% 1.67G/1.67G [00:20<00:00, 79.4MB/s]
```

```
from zipfile import ZipFile
with ZipFile("/content/Arjun_Assignment_data-20220427T165022Z-002.zip", 'r') as zObject:
    zObject.extractall(
        path="/content/Arjun_Assignment_data-20220427T165022Z-002")
```

```
from zipfile import ZipFile
with ZipFile("/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-001.zip", 'r') as zObje
    zObject.extractall(
        path="/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-001")
```

```
from zipfile import ZipFile
with ZipFile("/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-002.zip", 'r') as zObje
    zObject.extractall(
        path="/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-002")
```

```
from zipfile import ZipFile
with ZipFile("/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-003.zip", 'r') as zObje
    zObject.extractall(
        path="/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-003")
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
from collections import Counter
```

```
D1_postlinks = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-001/data
print(D1_postlinks)
```

```
              Unnamed: 0          id            creation_date   post_id  \
    0                  0          19  2010-04-26T02:59:48.130       109   
    1                  1          37  2010-04-26T02:59:48.600      1970   
    2                  2          42  2010-04-26T02:59:48.647      2154   
    3                  3          48  2010-04-26T02:59:48.740      2483   
    4                  4          52  2010-04-26T02:59:48.757      2572   
    ...              ...         ...                      ...       ...   
    5292619      5292619  1624278139  2018-09-02T08:09:41.520  52133002   
    5292620      5292620  1624278147  2018-09-02T08:10:50.820  52134991   
    5292621      5292621  1624278315  2018-09-02T08:14:26.470  52135049   
    5292622      5292622  1624278337  2018-09-02T08:15:36.387  52135007   
    5292623      5292623  1624278449  2018-09-02T08:17:32.137  52135049   

              related_post_id  link_type_id  
    0                   32412             1  
    1                  617600             1  
    2                 2451138             1  
    3                  496096             1  
    4                  209329             1  
    ...                   ...           ...  
    5292619          31486547             1  
    5292620           5500805             1  
    5292621          30461565             1  
    5292622           1761051             1  
    5292623           3127429             1  
    [5292624 rows x 6 columns]
```

```
D1_postlongs = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-001/datase
print(D1_postlongs.columns)
```

```
    <ipython-input-22-c1cb7ddbc190>:1: DtypeWarning: Columns (13) have mixed types. Specify dtype option on import or set low_memory=Fal
      D1_postlongs = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-00
    Index(['Unnamed: 0', 'id', 'post_type_id', 'accepted_answer_id', 'parent_id',
           'creation_date', 'score', 'view_count', 'owner_user_id', 'tags',
           'answer_count', 'comment_count', 'favorite_count',
           'community_owned_date', 'title', 'body'],
          dtype='object')
```

```
D1_postshort = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-001/datas
print(D1_postshort.columns)
```

```
    <ipython-input-6-a0f67573e964>:1: DtypeWarning: Columns (13) have mixed types. Specify dtype option on import or set low_memory=Fals
      D1_postshort = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-00
    Index(['Unnamed: 0', 'id', 'post_type_id', 'accepted_answer_id', 'parent_id',
           'creation_date', 'score', 'view_count', 'owner_user_id', 'tags',
           'answer_count', 'comment_count', 'favorite_count',
           'community_owned_date'],
          dtype='object')
```

```
D1_postlinks_json = pd.read_json('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-00
print(D1_postlinks_json.columns)
```

```
    Index(['id', 'creation_date', 'post_id', 'related_post_id', 'link_type_id'], dtype='object')
```

```
D2_User = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-002/dataset/us
print(D2_User.columns)
```

```
    Index(['Unnamed: 0', 'id', 'reputation', 'creation_date', 'display_name',
           'views', 'upvotes', 'downvotes', 'account_id'],
          dtype='object')
```

```
D2_postslong_json = pd.read_json('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-002/
print(D2_postslong_json.columns)
```

```
    Index(['id', 'post_type_id', 'accepted_answer_id', 'parent_id',
           'creation_date', 'score', 'view_count', 'owner_user_id', 'tags',
           'answer_count', 'comment_count', 'favorite_count',
           'community_owned_date', 'title', 'body'],
          dtype='object')
```

```
D3_post_history = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-003/dat
print(D3_post_history.columns)
print(D3_post_history)
```

```
    Index(['Unnamed: 0', 'id', 'ph_type_id', 'post_id', 'revision_guid',
           'creation_date', 'user_id', 'user_display_name', 'comment', 'text'],
          dtype='object')
             Unnamed: 0         id  ph_type_id   post_id  \
    0                 0         12           1        17
    1                 1         13           3        17
    2                 2         14           2        17
    3                 3        219           1       123
    4                 4        220           3       123
    ...             ...        ...         ...       ...
    3640996     3640996  180953899           1  52134558
    3640997     3640997  180953900           3  52134558
    3640998     3640998  180953916           5  52134558
    3640999     3640999  180954028          10  52134558
    3641000     3641000  180954083           5  52134558

                                    revision_guid            creation_date  \
    0        0421fb42-a29a-4cb2-84ba-a828725410f8  2008-08-01T05:09:55.993
    1        0421fb42-a29a-4cb2-84ba-a828725410f8  2008-08-01T05:09:55.993
    2        0421fb42-a29a-4cb2-84ba-a828725410f8  2008-08-01T05:09:55.993
    3        5dc36325-a80d-4ef2-8bd6-fde1720d7e7a  2008-08-01T16:08:52.353
    4        5dc36325-a80d-4ef2-8bd6-fde1720d7e7a  2008-08-01T16:08:52.353
    ...                                       ...                      ...
    3640996  52c126dc-b431-4980-9ba4-f844d639bcc7  2018-09-02T06:30:21.917
    3640997  52c126dc-b431-4980-9ba4-f844d639bcc7  2018-09-02T06:30:21.917
    3640998  7688ee34-445b-48e7-8a90-4be78833db93  2018-09-02T06:31:22.833
    3640999  3c161d4d-2735-485a-9216-1228d4515e58  2018-09-02T06:36:50.840
    3641000  20c390d6-8ea4-4a48-8a91-8d3d045c9286  2018-09-02T06:39:58.227

              user_id user_display_name                               comment  \
    0               2               NaN                                   NaN
```

```
1               2               NaN                                       NaN
2               2               NaN                                       NaN
3              78               NaN                                       NaN
4              78               NaN                                       NaN
...           ...               ...                                       ...
3640996  10305684               NaN                                       NaN
3640997  10305684               NaN                                       NaN
3640998  10305684               NaN               added 2 characters in body
3640999   9515207               NaN                                       101
3641000   3589092               NaN  added image instead of just the link

                                                    text
0                                    Binary Data in MYSQL
1                                       <database><mysql>
2                        How do I store binary data in mysql?
3                                         CSV File to XML
4                                      <csv><xml><java><>
...                                                   ...
3640996  What are the 3 dots the JavaScript console ret...
3640997                            <javascript><console>
3640998  I was messing around with the Javascript conso...
3640999  {"OriginalQuestionIds":[21997803],"Voters":[{"...
3641000  I was messing around with the Javascript conso...

[3641001 rows x 10 columns]
```

```
D1_postlinks = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-001/data
D1_postlongs = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-001/data
D1_postshort = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-001/data
D1_postlinks_json = pd.read_json('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-00
D2_postslong_json = pd.read_json('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-00
D2_User = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-002/dataset/u
D3_post_history = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-003/d
```

```
<ipython-input-3-80c054f5ad6a>:2: DtypeWarning: Columns (13) have mixed types. Specif
  D1_postlongs = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arj
<ipython-input-3-80c054f5ad6a>:3: DtypeWarning: Columns (13) have mixed types. Specif
  D1_postshort = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-002/Arj
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
<ipython-input-3-80c054f5ad6a> in <cell line: 4>()
      2 D1_postlongs = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-
002/Arjun_Assignment_data/dataset-20210607T020316Z-001/dataset/posts_long.csv')
      3 D1_postshort = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-
002/Arjun_Assignment_data/dataset-20210607T020316Z-001/dataset/posts_short.csv')
----> 4 D1_postlinks_json = pd.read_json('/content/Arjun_Assignment_data-
20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-
001/dataset/postLinks.json')
      5 D2_postslong_json = pd.read_json('/content/Arjun_Assignment_data-
20220427T165022Z-002/Arjun_Assignment_data/dataset-20210607T020316Z-
002/dataset/posts_long.json')
      6 D2_User = pd.read_csv('/content/Arjun_Assignment_data-20220427T165022Z-
002/Arjun_Assignment_data/dataset-20210607T020316Z-002/dataset/users.csv')

                              10 frames
/usr/local/lib/python3.10/dist-packages/pandas/core/internals/construction.py in
_homogenize(data, index, dtype)
    613                         # see test_constructor_subclass_dict
    614                         val = dict(val)
--> 615                     val = lib.fast_multiget(val, oindex._values, default=np.nan)
    616 
    617                 val = sanitize_array(
```

```
# Question 1 :- Determine the Number of Tags Per Question

Q1_DS = pd.concat([D1_postlongs[['id' , 'tags']] , D1_postshort[['id' , 'tags']] , D2_postslong_json[['id' , 'tags']]])
no_of_tags = []
for tag in Q1_DS['tags'] :
  tags = re.findall(r'<.*?>', tag)
  no_of_tags.append(len(tags))
Q1_DS['No_of_Tags'] = no_of_tags
Q1_ANS = Q1_DS[['id' , 'No_of_Tags']]
print(Q1_DS)
```

```
              id                                               tags  \
0              4  <c#><floating-point><type-conversion><double><...
1              6              <html><css><css3><internet-explorer-7>
2              9                                <c#><.net><datetime>
3             11  <c#><datetime><time><datediff><relative-time-s...
4             13  <javascript><html><browser><timezone><timezone...
...          ...                                                ...
676199  52133457                                     <python><import>
```

```
676200  52133674                   <java><firebase><android-studio>
676201  52133700                                                <c>
676202  52133880                         <angularjs><node.js><ajax>
676203  52134121                                        <php><html>

        No_of_Tags
0                5
1                4
2                3
3                5
4                5
...            ...
676199           2
676200           3
676201           1
676202           3
676203           2

[2028612 rows x 3 columns]
```

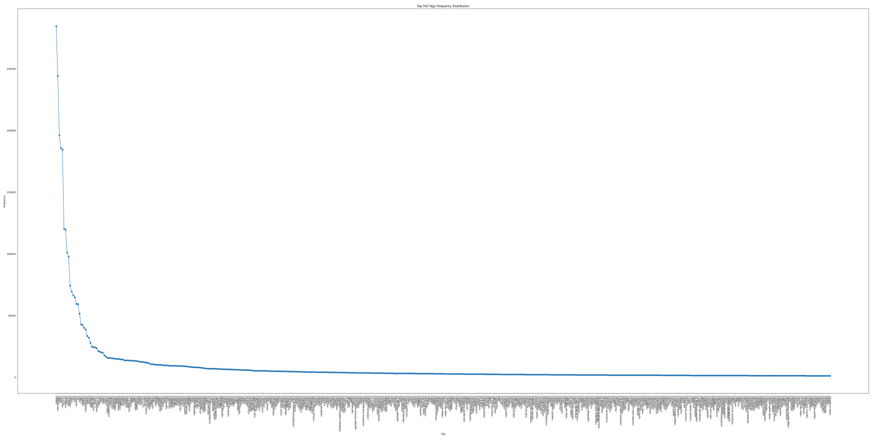
```
Q2_DS = pd.concat([D1_postlongs[['id' , 'tags']] , D1_postshort[['id' , 'tags']] , D2_postslong_json[['id' , 'tags']]])
total_tags = []
for tag in Q2_DS['tags'] :
  tags = re.findall(r'<.*?>', tag)
  total_tags.extend(tags)

total_tags = pd.Series(total_tags)
print(total_tags.nunique())
```

```
25310
```

```
# Question - 3 :-Determine the top-25 Tags appearing frequently

Q3_DS = pd.concat([D1_postlongs[['id', 'tags']], D1_postshort[['id', 'tags']], D2_postslong_json[['id', 'tags']]])

total_tags = []

for tag in Q3_DS['tags']:
    tags = re.findall(r'<.*?>', tag)
    total_tags.extend(tags)

total_tags_series = pd.Series(total_tags)
tag_frequency = total_tags_series.value_counts()
frequency_df = pd.DataFrame({'Tag': tag_frequency.index, 'Frequency': tag_frequency.values})

Q3_Ans = frequency_df.head(25)
print(Q3_Ans)
```

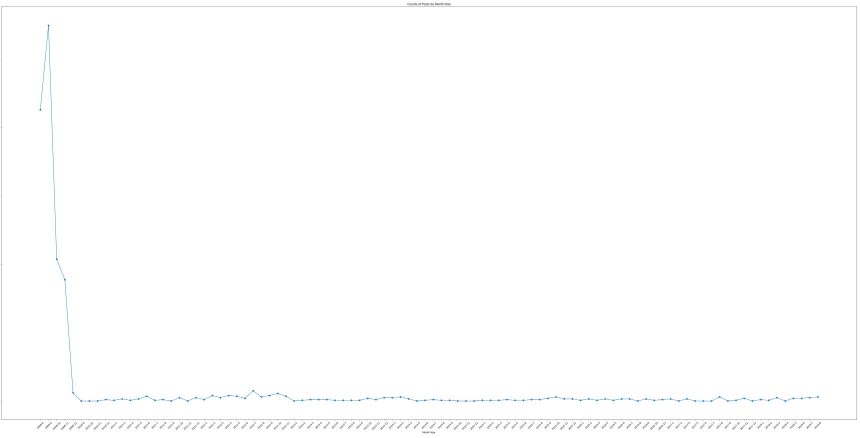
```
              Tag  Frequency
0          <java>     284538
1    <javascript>     244038
2           <php>     195999
3            <c#>     185697
4        <python>     184515
5           <c++>     120474
6       <android>     119730
7          <html>     101208
8        <jquery>      97803
9         <mysql>      74295
10          <css>      69483
11            <c>      66498
12          <ios>      64749
13       <arrays>      59601
14            <r>      59232
15          <sql>      51681
16         <.net>      42870
17        <regex>      42321
18       <string>      40191
19  <objective-c>      38847
20        <swift>      33591
21         <json>      32115
22       <iphone>      27885
23      <asp.net>      24828
24   <sql-server>      24495
```

```
Q4_Ans = frequency_df.head(500)
plt.figure(figsize=(50, 25))
plt.plot(Q4_Ans['Tag'], Q4_Ans['Frequency'], marker='o', linestyle='-')
plt.xlabel('Tag')
plt.ylabel('Frequency')
plt.title('Top 500 Tags Frequency Distribution')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```



```
Q_5_DS = pd.merge(D3_post_history, D1_postlinks, on='id')
filtered_Q_5_DS = Q_5_DS[Q_5_DS['ph_type_id'] == 3]
filtered_Q_5_DS['creation_date_x'] = pd.to_datetime(filtered_Q_5_DS['creation_date_x'])
filtered_Q_5_DS["Year"] = filtered_Q_5_DS["creation_date_x"].dt.year
filtered_Q_5_DS["Month"] = filtered_Q_5_DS["creation_date_x"].dt.month
Q5_Ans = filtered_Q_5_DS.groupby(['Year', 'Month'])['creation_date_x'].count().reset_index(name='counts')
total_counts = Q5_Ans['counts'].sum()
Q5_Ans['relative_frequency'] = (Q5_Ans['counts'] / total_counts) * 100
plt.figure(figsize=(50, 25))
plt.plot(Q5_Ans['Year'].astype(str) + '-' + Q5_Ans['Month'].astype(str), Q5_Ans['counts'], marker='o', linestyle='-')
plt.xlabel('Month-Year')
plt.ylabel('Counts')
plt.title('Counts of Posts by Month-Year')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

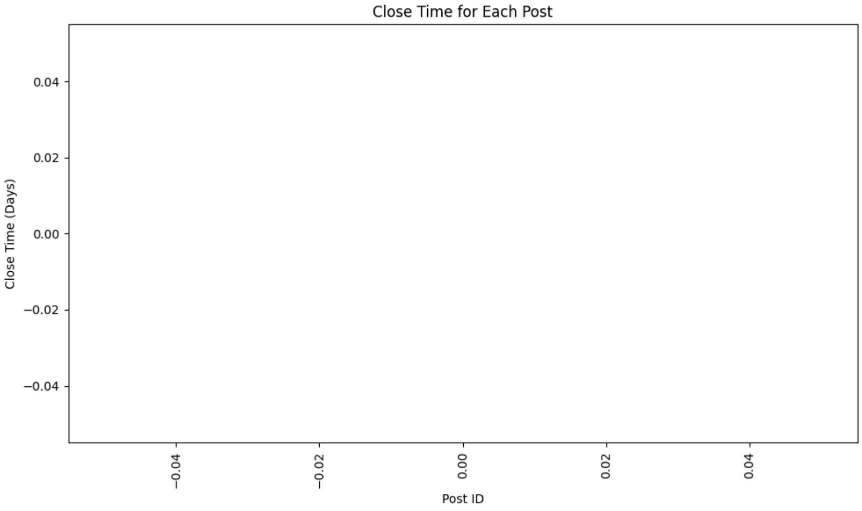
```
<ipython-input-49-67379cf232f2>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
  filtered_Q_5_DS['creation_date_x'] = pd.to_datetime(filtered_Q_5_DS['creation_date_
<ipython-input-49-67379cf232f2>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
  filtered_Q_5_DS["Year"] = filtered_Q_5_DS["creation_date_x"].dt.year
<ipython-input-49-67379cf232f2>:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
  filtered_Q_5_DS["Month"] = filtered_Q_5_DS["creation_date_x"].dt.month
```



```
# Question 6

Q6_DS = pd.concat([D1_postlongs[['id', 'tags']], D1_postshort[['id', 'tags']], D2_postslong_json[['id', 'tags']]])
Q6_DS = pd.merge(Q6_DS , D3_post_history , on = 'id')
total_tags = []

for tag in Q6_DS['tags']:
    tags = re.findall(r'<.*?>', tag)
    total_tags.append(tags)

Q6_DS['tags_diff'] = total_tags
Q6_DS = Q6_DS[Q6_DS['ph_type_id'] == 3]
print(Q6_DS.columns)
flattened_tags = [tag for sublist in Q6_DS['tags_diff'] for tag in sublist]
tag_counts = Counter(flattened_tags)
tag_counts_df = pd.DataFrame(tag_counts.items(), columns=['Tag', 'Count'])
total_tags_count = tag_counts_df['Count'].sum()
tag_counts_df['Percentage'] = (tag_counts_df['Count'] / total_tags_count) * 100
tag_counts_df = tag_counts_df.sort_values(by='Count', ascending=False)
top_20 = tag_counts_df[:20]

# Plotting
plt.figure(figsize=(50, 25))
plt.bar(top_20['Tag'], top_20['Percentage'], color='blue', alpha=0.7)

for i, value in enumerate(top_20['Percentage']):
    plt.text(i, value, str(value), ha='center', va='bottom', weight='bold')

plt.xlabel('Tag')
plt.ylabel('Percentage')
plt.title('Tag Percentage')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```

```
     Index(['id', 'tags', 'Unnamed: 0', 'ph_type_id', 'post_id', 'revision_guid',
            'creation_date', 'user_id', 'user_display_name', 'comment', 'text',
            'tags_diff'],
           dtype='object')
```

```
# Question - 7

Q7_DS = pd.concat([D1_postlongs, D1_postshort])
Q7_DS = pd.merge(Q7_DS , D3_post_history , on = 'id')
Q7_DS = pd.merge(Q7_DS, D1_postlinks , on = 'id')
Q7_DS = Q7_DS[Q7_DS['link_type_id'] == 3]
Q7_DS['creation_date'] = pd.to_datetime(Q7_DS['creation_date'])
Q7_DS['community_owned_date'] = pd.to_datetime(Q7_DS['community_owned_date'])
Q7_DS['close_time'] = Q7_DS['community_owned_date'] - Q7_DS['creation_date']
Q7_DS_Plot = Q7_DS[['id' , 'close_time']]
plt.figure(figsize=(10, 6))
bars = plt.bar(Q7_DS_Plot['id'], Q7_DS_Plot['close_time'].dt.days, color='blue', alpha=0.7)
plt.xlabel('Post ID')
plt.ylabel('Close Time (Days)')
plt.title('Close Time for Each Post')
plt.xticks(rotation=90)

for bar, time in zip(bars, Q7_DS_Plot['close_time'].dt.days):
    plt.text(bar.get_x() + bar.get_width() / 2, bar.get_height(), str(time), ha='center', va='bottom')

plt.tight_layout()
plt.show()
```



```
# Question - 8
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
<ipython-input-17-8c89cfa1dd47> in <cell line: 3>()
      1 # Question - 8
```

```
#Q8
merge = pd.merge(D2_User, Q6_DS, on='id', how='inner')
dele = [ 'creation_date','display_name','views','upvotes','downvotes', 'tags', 'Unnamed: 0', 'ph_type_id', 'post_id', 'revision_guid',
merge = merge.drop(columns=dele)
print(merge)
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
<ipython-input-19-b0ac734ae968> in <cell line: 3>()
      1 merge = pd.merge(D2_User, Q6_DS, on='id', how='inner')
      2 dele = [ 'creation_date','display_name','views','upvotes','downvotes',
'tags', 'Unnamed: 0', 'ph_type_id', 'post_id', 'revision_guid', 'creation_date',
'comment', 'text', 'tags_diff']
----> 3 merge = merge.drop(columns=dele)
      4 print(merge)

                              5 frames

/usr/local/lib/python3.10/dist-packages/pandas/core/indexes/base.py in drop(self,
labels, errors)
   6932         if mask.any():
   6933             if errors != "ignore":
-> 6934                 raise KeyError(f"{list(labels[mask])} not found in axis")
   6935             indexer = indexer[~mask]
   6936         return self.delete(indexer)

KevError: "['creation date'. 'Unnamed: 0'. 'creation date'] not found in axis"
```

```
merge = pd.merge(D2_User, Q6_DS, on='id', how='inner')
dele = [ 'display_name','views','upvotes','downvotes', 'tags', 'ph_type_id', 'post_id', 'revision_guid', 'creation_date', 'comment', 'tex
merge = merge.drop(columns=dele)
print(merge)
```

```
Q9_DS = D1_postlongs['community_owned_date'][:500]
Percentage = Q9_DS.isnull().sum().sum() / Q9_DS.shape[0] * 100
print(Percentage)
```