In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
df=pd.read_csv("D:/covid-variants.csv")
df.head()
```

Out[1]:

	location	date	variant	num_sequences	perc_sequences	num_sequences_total
0	Angola	2020-07-06	Alpha	0	0.0	3
1	Angola	2020-07-06	B.1.1.277	0	0.0	3
2	Angola	2020-07-06	B.1.1.302	0	0.0	3
3	Angola	2020-07-06	B.1.1.519	0	0.0	3
4	Angola	2020-07-06	B.1.160	0	0.0	3

In [2]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100416 entries, 0 to 100415
Data columns (total 6 columns):
 #   Column               Non-Null Count   Dtype  
---  ------               --------------   -----  
 0   location             100416 non-null  object 
 1   date                 100416 non-null  object 
 2   variant              100416 non-null  object 
 3   num_sequences        100416 non-null  int64  
 4   perc_sequences       100416 non-null  float64
 5   num_sequences_total  100416 non-null  int64  
dtypes: float64(1), int64(2), object(3)
memory usage: 4.6+ MB
```

In [15]:

```
df.isnull().sum()
```

Out[15]:

```
location              0
date                  0
variant               0
num_sequences         0
perc_sequences        0
num_sequences_total   0
year                  0
month                 0
dtype: int64
```

In [3]:

```
df.describe()
```

Out[3]:

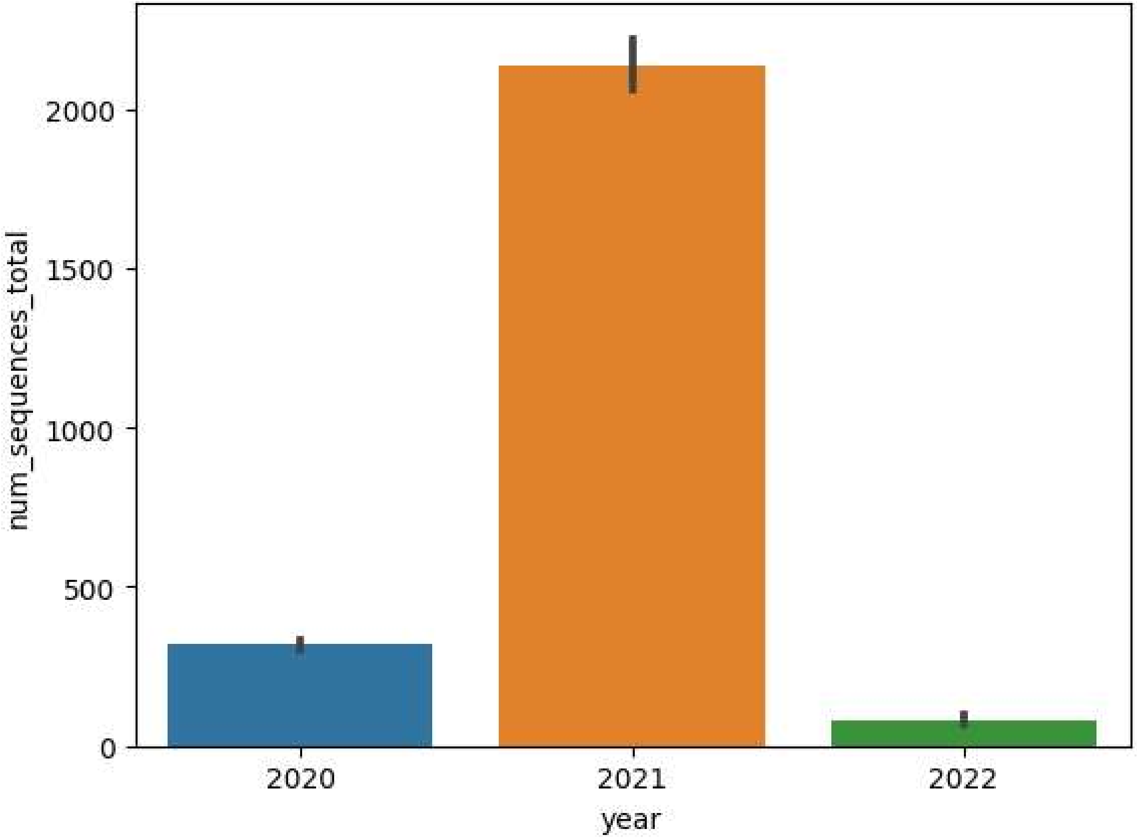
	num_sequences	perc_sequences	num_sequences_total
count	100416.000000	100416.000000	100416.000000
mean	72.171676	6.154355	1509.582457
std	1669.262169	21.898989	8445.291772
min	0.000000	-0.010000	1.000000
25%	0.000000	0.000000	12.000000
50%	0.000000	0.000000	59.000000
75%	0.000000	0.000000	394.000000
max	142280.000000	100.000000	146170.000000

```
In [3]: df['year'] = pd.DatetimeIndex(df['date']).year
        df['month'] = pd.DatetimeIndex(df['date']).month
        df.head()
```

```
Out[3]:
     location        date   variant  num_sequences  perc_sequences  num_sequences_total  year  month
  0    Angola  2020-07-06     Alpha              0             0.0                    3  2020      7
  1    Angola  2020-07-06  B.1.1.277             0             0.0                    3  2020      7
  2    Angola  2020-07-06  B.1.1.302             0             0.0                    3  2020      7
  3    Angola  2020-07-06  B.1.1.519             0             0.0                    3  2020      7
  4    Angola  2020-07-06   B.1.160              0             0.0                    3  2020      7
```

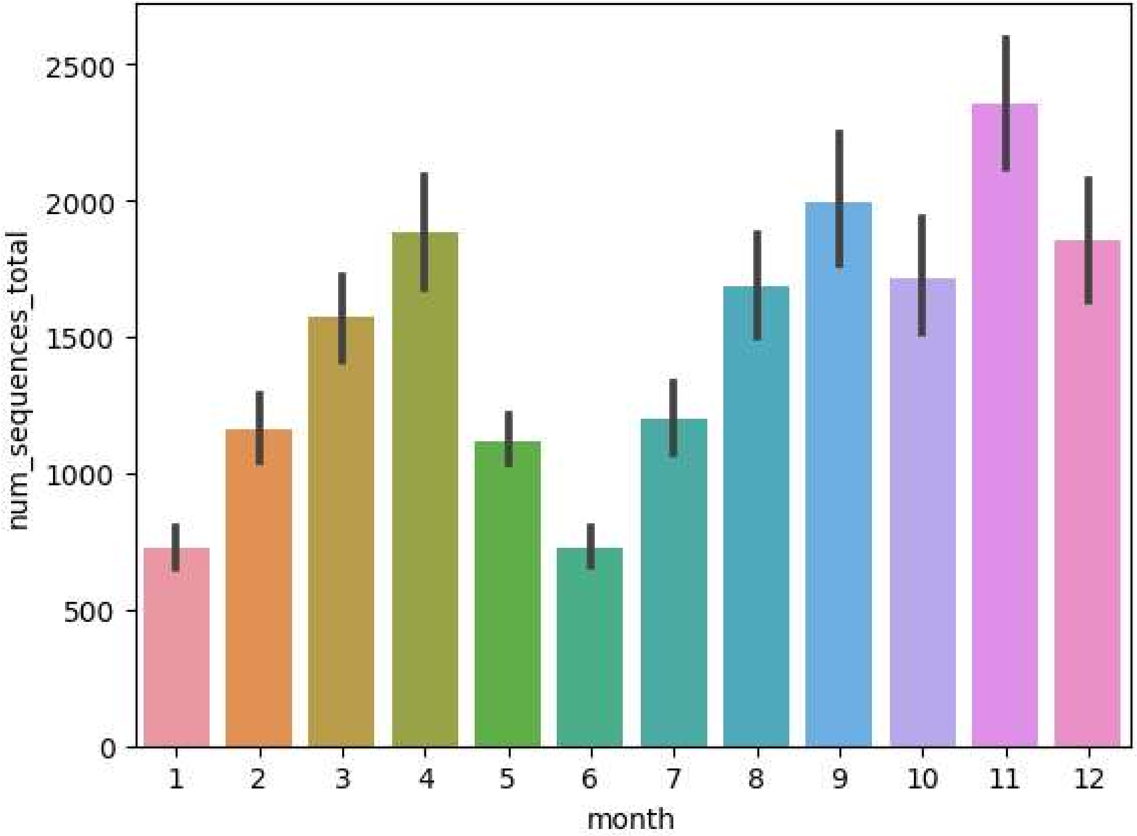
```
In [12]: sns.barplot(x = df['year'], y = df['num_sequences_total'])
```

```
Out[12]: <AxesSubplot: xlabel='year', ylabel='num_sequences_total'>
```



```
In [26]: sns.barplot(x = df['month'], y = df['num_sequences_total'])
```

```
Out[26]: <AxesSubplot: xlabel='month', ylabel='num_sequences_total'>
```



In [37]:
```
df.groupby(df['variant'])['num_sequences_total'].sum()
```

Out[37]:
```
variant
Alpha             6316093
B.1.1.277         6316093
B.1.1.302         6316093
B.1.1.519         6316093
B.1.160           6316093
B.1.177           6316093
B.1.221           6316093
B.1.258           6316093
B.1.367           6316093
B.1.620           6316093
Beta              6316093
Delta             6316093
Epsilon           6316093
Eta               6316093
Gamma             6316093
Iota              6316093
Kappa             6316093
Lambda            6316093
Mu                6316093
Omicron           6316093
S:677H.Robin1     6316093
S:677P.Pelican    6316093
non_who           6316093
others            6316093
Name: num_sequences_total, dtype: int64
```

In [14]:

```
var_of_india=df[df['location']=='India']['variant'].value_counts()
var_of_india
```

Out[14]:

```
Alpha             44
B.1.1.277         44
others            44
S:677P.Pelican    44
S:677H.Robin1     44
Omicron           44
Mu                44
Lambda            44
Kappa             44
Iota              44
Gamma             44
Eta               44
Epsilon           44
Delta             44
Beta              44
B.1.620           44
B.1.367           44
B.1.258           44
B.1.221           44
B.1.177           44
```

```
In [22]: df['location'].unique()
```

```
Out[22]: array(['Angola', 'Argentina', 'Aruba', 'Australia', 'Austria', 'Bahrain',
                'Bangladesh', 'Belgium', 'Belize', 'Benin',
                'Bosnia and Herzegovina', 'Botswana', 'Brazil', 'Brunei',
                'Bulgaria', 'Cambodia', 'Cameroon', 'Canada', 'Chile', 'Colombia',
                'Costa Rica', 'Croatia', 'Curacao', 'Cyprus', 'Czechia', 'Denmark',
                'Djibouti', 'Dominican Republic', 'Ecuador', 'Egypt', 'Estonia',
                'Ethiopia', 'Fiji', 'Finland', 'France', 'Gambia', 'Georgia',
                'Germany', 'Ghana', 'Greece', 'Guatemala', 'Hong Kong', 'Hungary',
                'Iceland', 'India', 'Indonesia', 'Iran', 'Iraq', 'Ireland',
                'Israel', 'Italy', 'Jamaica', 'Japan', 'Jordan', 'Kazakhstan',
                'Kenya', 'Kosovo', 'Kuwait', 'Latvia', 'Lebanon', 'Liechtenstein',
                'Lithuania', 'Luxembourg', 'Madagascar', 'Malawi', 'Malaysia',
                'Maldives', 'Malta', 'Mauritius', 'Mexico', 'Moldova', 'Monaco',
                'Mongolia', 'Montenegro', 'Morocco', 'Mozambique', 'Nepal',
                'Netherlands', 'New Zealand', 'Nigeria', 'North Macedonia',
                'Norway', 'Oman', 'Pakistan', 'Papua New Guinea', 'Paraguay',
                'Peru', 'Philippines', 'Poland', 'Portugal', 'Qatar', 'Romania',
                'Russia', 'Rwanda', 'Senegal', 'Serbia', 'Seychelles', 'Singapore',
                'Sint Maarten (Dutch part)', 'Slovakia', 'Slovenia',
                'South Africa', 'South Korea', 'Spain', 'Sri Lanka', 'Suriname',
                'Sweden', 'Switzerland', 'Thailand', 'Togo', 'Trinidad and Tobago',
                'Turkey', 'Uganda', 'Ukraine', 'United Arab Emirates',
                'United Kingdom', 'United States', 'Uruguay', 'Vietnam', 'Zambia',
                'Zimbabwe'], dtype=object)
```

```
In [35]: df['variant'].unique()
```

```
Out[35]: array(['Alpha', 'B.1.1.277', 'B.1.1.302', 'B.1.1.519', 'B.1.160',
                'B.1.177', 'B.1.221', 'B.1.258', 'B.1.367', 'B.1.620', 'Beta',
                'Delta', 'Epsilon', 'Eta', 'Gamma', 'Iota', 'Kappa', 'Lambda',
                'Mu', 'Omicron', 'S:677H.Robin1', 'S:677P.Pelican', 'others',
                'non_who'], dtype=object)
```

In [8]:

```
df['location'].value_counts()
```

Out[8]:

```
Bangladesh        1080
Belgium           1080
United States     1080
United Kingdom    1080
France            1080
                  ... 
Montenegro         384
Monaco             360
Fiji               336
Benin              336
Brunei             240
Name: location, Length: 121, dtype: int64
```

In [23]:

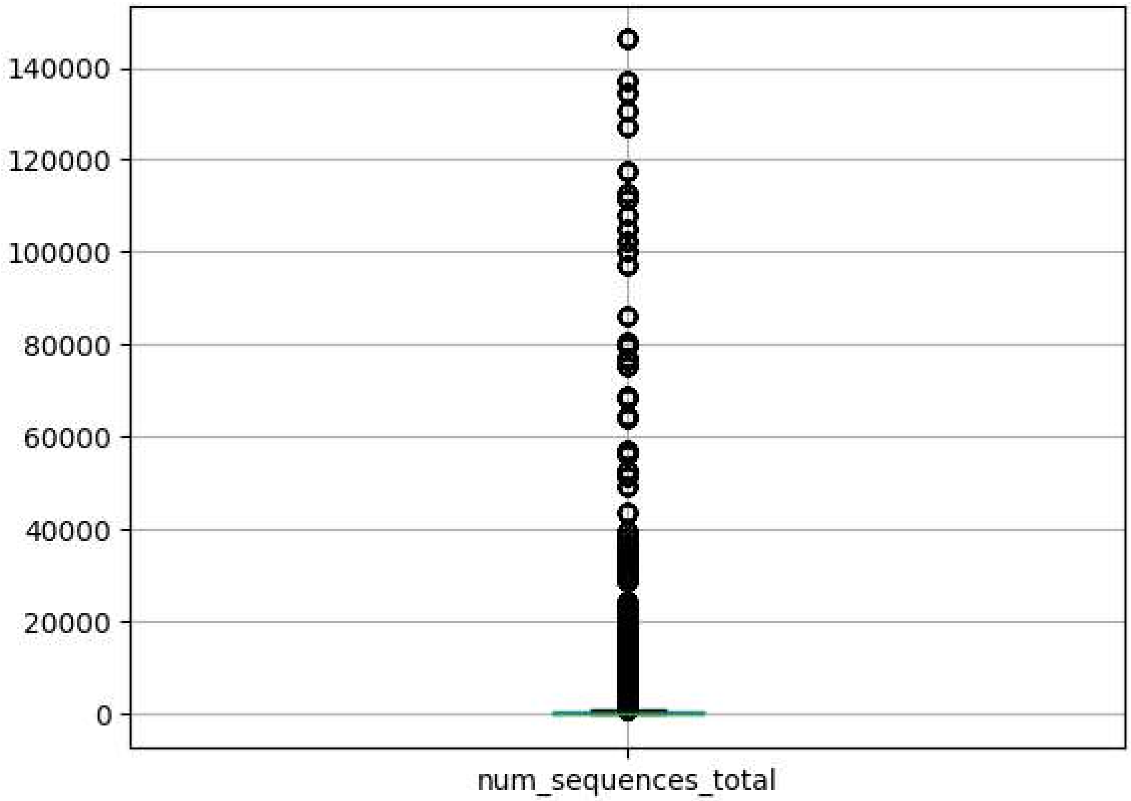
```
new_df=df.groupby(df['month'])['variant'].value_counts()
new_df
```

Out[23]:

```
month  variant        
1      Alpha              332
       B.1.1.277          332
       B.1.1.302          332
       B.1.1.519          332
       B.1.160            332
                         ... 
12     Omicron            344
       S:677H.Robin1      344
       S:677P.Pelican     344
       non_who            344
       others             344
Name: variant, Length: 288, dtype: int64
```

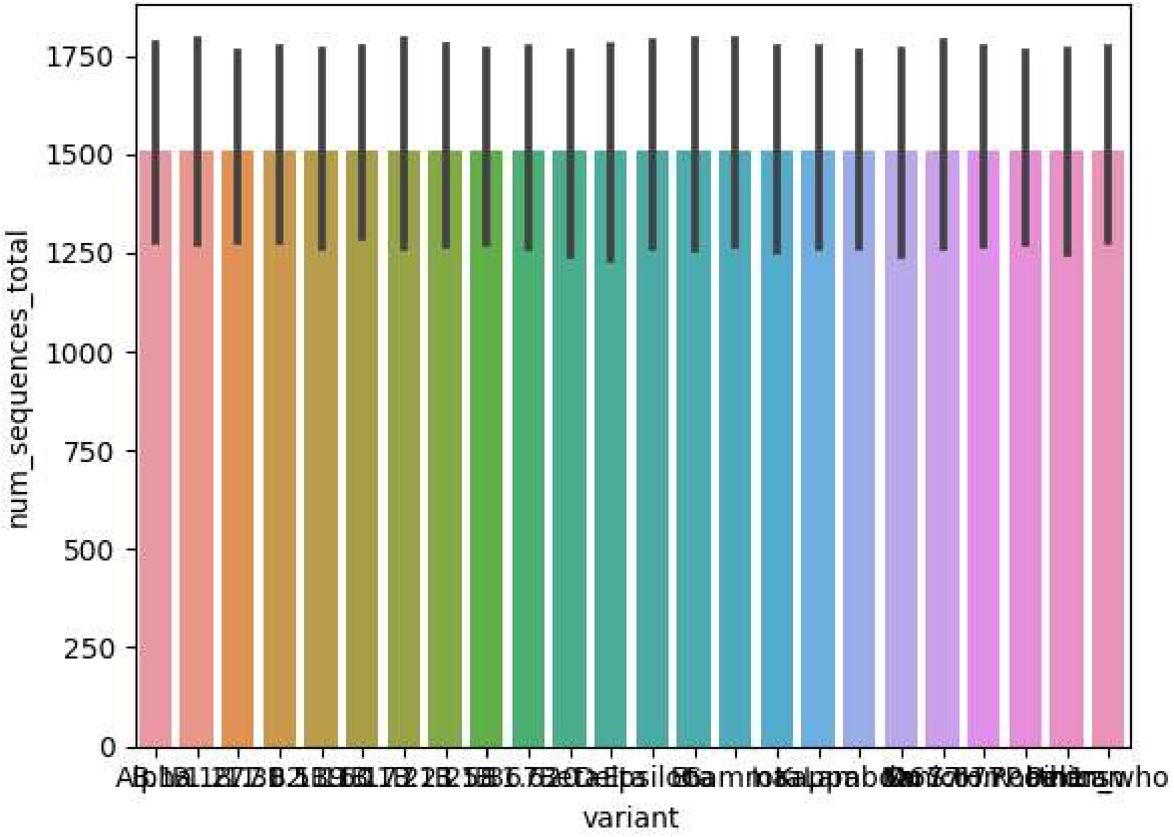
```
In [40]: df.boxplot(column=['num_sequences_total'])
```

```
Out[40]: <AxesSubplot: >
```



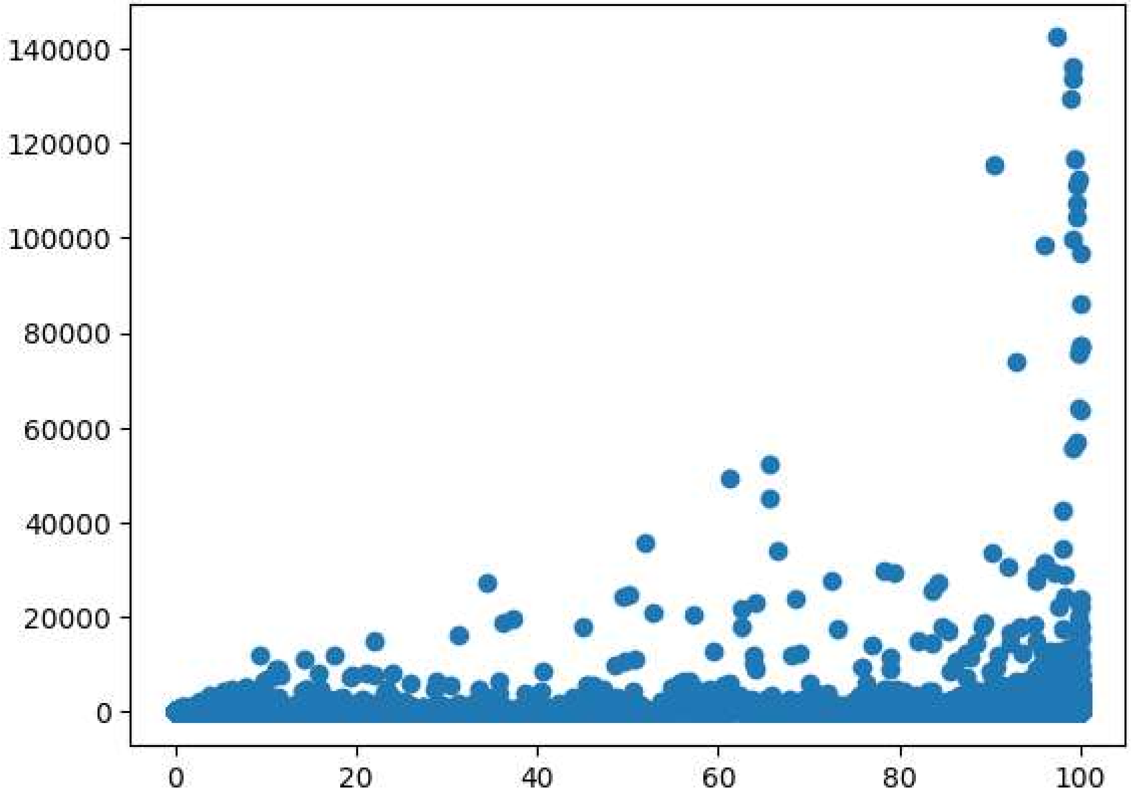
```
In [38]: sns.barplot(x = df['variant'], y = df['num_sequences_total'])
```

```
Out[38]: <AxesSubplot: xlabel='variant', ylabel='num_sequences_total'>
```



```
In [39]: plt.scatter(df['perc_sequences'],df['num_sequences'])

Out[39]: <matplotlib.collections.PathCollection at 0x17fdf48c8d0>
```



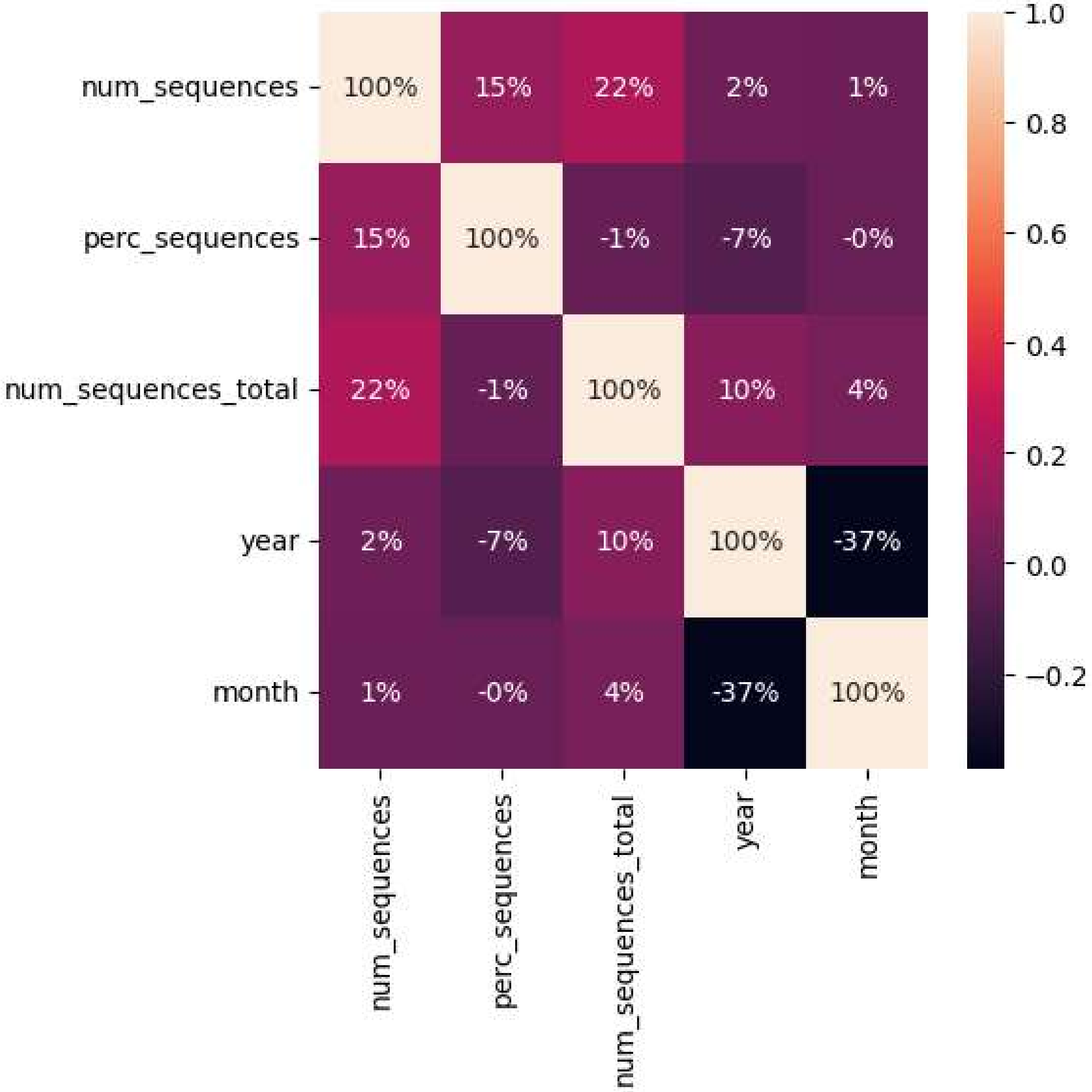
In [16]:

```
plt.figure(figsize=(5,5))
sns.heatmap(df.corr(),annot=True,fmt=".0%")
```

```
C:\Users\kunal vashistha\AppData\Local\Temp\ipykernel_2764\3529239389.py:2: FutureWarning: The default value of nume
ric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns o
r specify the value of numeric_only to silence this warning.
  sns.heatmap(df.corr(),annot=True,fmt=".0%")
```

Out[16]:

```
<AxesSubplot: >
```


num_sequences
perc_sequences
num_sequences_total
year
month
100%
15%
22%
2%
1%
15%
100%
-1%
-7%
-0%
22%
-1%
100%
10%
4%
2%
-7%
10%
100%
-37%
1%
-0%
4%
-37%
100%
num_sequences
perc_sequences
num_sequences_total
year
month
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

In [ ]: