In [1]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
```

In [2]:
```
data = pd.read_csv("articles.csv")
```

In [3]:
```
data
```

Out[3]:

	title	author_id	last_updated	link	category
0	5 Best Practices For Writing SQL Joins	priyankab14	21-Feb-22	https://www.geeksforgeeks.org/5-best-practices...	easy
1	Foundation CSS Dropdown Menu	ishankhandelwals	20-Feb-22	https://www.geeksforgeeks.org/foundation-css-d...	easy
2	Top 20 Excel Shortcuts That You Need To Know	priyankab14	17-Feb-22	https://www.geeksforgeeks.org/top-20-excel-sho...	easy
3	Servlet – Fetching Result	nishatiwari1719	17-Feb-22	https://www.geeksforgeeks.org/servlet-fetching...	easy
4	Suffix Sum Array	rohit768	21-Feb-22	https://www.geeksforgeeks.org/suffix-sum-array/	easy
...	...	...	...	...	...
34569	Data Structures | Queue | Question 11	GeeksforGeeks	28-Jun-21	https://www.geeksforgeeks.org/data-structures-...	expert
34570	Data Structures | Binary Trees | Question 1	GeeksforGeeks	28-Jun-21	https://www.geeksforgeeks.org/data-structures-...	expert
34571	Amazon Interview | Set 9	GeeksforGeeks	28-Apr-17	https://www.geeksforgeeks.org/amazon-interview...	expert
34572	Python Program for Rat in a Maze | Backtracking-2	GeeksforGeeks	02-Aug-21	https://www.geeksforgeeks.org/python-program-f...	expert
34573	Data Structures and Algorithms | Set 21	GeeksforGeeks	27-Mar-17	https://www.geeksforgeeks.org/data-structures-...	expert

34574 rows × 5 columns

In [4]:
```
data.head()
```

Out[4]:

	title	author_id	last_updated	link	category
0	5 Best Practices For Writing SQL Joins	priyankab14	21-Feb-22	https://www.geeksforgeeks.org/5-best-practices...	easy
1	Foundation CSS Dropdown Menu	ishankhandelwals	20-Feb-22	https://www.geeksforgeeks.org/foundation-css-d...	easy
2	Top 20 Excel Shortcuts That You Need To Know	priyankab14	17-Feb-22	https://www.geeksforgeeks.org/top-20-excel-sho...	easy
3	Servlet – Fetching Result	nishatiwari1719	17-Feb-22	https://www.geeksforgeeks.org/servlet-fetching...	easy
4	Suffix Sum Array	rohit768	21-Feb-22	https://www.geeksforgeeks.org/suffix-sum-array/	easy

In [5]:
```
data.tail()
```

Out[5]:

	title	author_id	last_updated	link	category
34569	Data Structures | Queue | Question 11	GeeksforGeeks	28-Jun-21	https://www.geeksforgeeks.org/data-structures-...	expert
34570	Data Structures | Binary Trees | Question 1	GeeksforGeeks	28-Jun-21	https://www.geeksforgeeks.org/data-structures-...	expert
34571	Amazon Interview | Set 9	GeeksforGeeks	28-Apr-17	https://www.geeksforgeeks.org/amazon-interview...	expert
34572	Python Program for Rat in a Maze | Backtracking-2	GeeksforGeeks	02-Aug-21	https://www.geeksforgeeks.org/python-program-f...	expert
34573	Data Structures and Algorithms | Set 21	GeeksforGeeks	27-Mar-17	https://www.geeksforgeeks.org/data-structures-...	expert

In [6]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34574 entries, 0 to 34573
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   title         34574 non-null  object
 1   author_id     34555 non-null  object
 2   last_updated  34556 non-null  object
 3   link          34574 non-null  object
 4   category      34574 non-null  object
dtypes: object(5)
memory usage: 1.3+ MB
```

In [7]:
```
data.isnull().sum()
```

Out[7]:
```
title            0
author_id       19
last_updated    18
link             0
category         0
dtype: int64
```

In [8]:

```
# There are 19 null vlaues in author_id and 18 in last_update
# so lets delete these rows.

data.dropna(inplace = True)
```

In [9]:

```
data.isnull().any().any()
```

Out[9]:

```
False
```

In [10]:

```
# Most popular author in terms of the number of articles.

data.groupby('author_id').size().sort_values(ascending = False).head()
```

Out[10]:

```
author_id
GeeksforGeeks    11957
ManasChhabra2      317
Striver            265
manjeet_04         246
Chinmoy Lenka      192
dtype: int64
```

In [12]:

```
import matplotlib.pyplot as plt
```

In [16]:

```
data['last_updated'] = pd.to_datetime(data['last_updated'])
```

In [18]:

```
print(data['last_updated'])
```

```
0       2022-02-21
1       2022-02-20
2       2022-02-17
3       2022-02-17
4       2022-02-21
           ...
34569   2021-06-28
34570   2021-06-28
34571   2017-04-28
34572   2021-08-02
34573   2017-03-27
Name: last_updated, Length: 34455, dtype: datetime64[ns]
```

In [21]:

```
data['last_updated'] = data['last_updated'].astype(str)
```

In [25]:

```
data[["year", "month", "day"]] = data["last_updated"].str.split("-", expand = True)
```

In [26]:

```
data['year']
```

Out[26]:

```
0        2022
1        2022
2        2022
3        2022
4        2022
         ...
34569    2021
34570    2021
34571    2017
34572    2021
34573    2017
Name: year, Length: 34455, dtype: object
```

In [27]:

```
data['month']
```

Out[27]:

```
0        02
1        02
2        02
3        02
4        02
         ..
34569    06
34570    06
34571    04
34572    08
34573    03
Name: month, Length: 34455, dtype: object
```

In [28]:
```
data['day']
```

Out[28]:
```
0        21
1        20
2        17
3        17
4        21
         ..
34569    28
34570    28
34571    28
34572    02
34573    27
Name: day, Length: 34455, dtype: object
```

In [29]:
```
data.head()
```

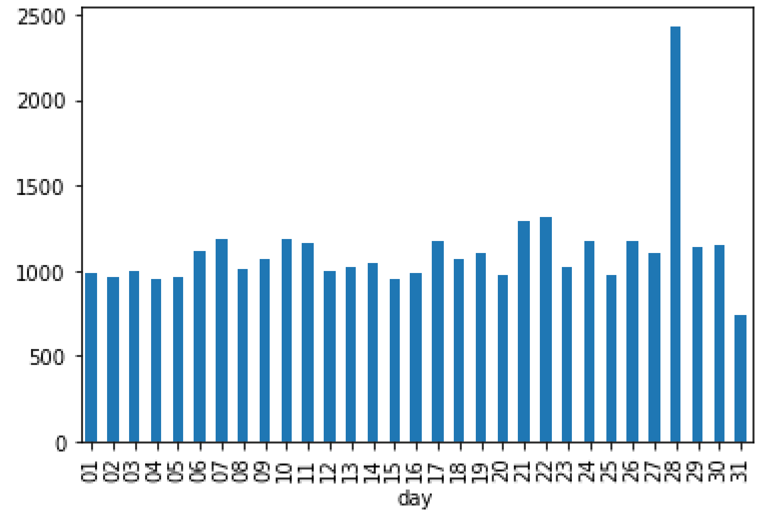
Out[29]:

	title	author_id	last_updated	link	category	day	month	year
0	5 Best Practices For Writing SQL Joins	priyankab14	2022-02-21	https://www.geeksforgeeks.org/5-best-practices...	easy	21	02	2022
1	Foundation CSS Dropdown Menu	ishankhandelwals	2022-02-20	https://www.geeksforgeeks.org/foundation-css-d...	easy	20	02	2022
2	Top 20 Excel Shortcuts That You Need To Know	priyankab14	2022-02-17	https://www.geeksforgeeks.org/top-20-excel-sho...	easy	17	02	2022
3	Servlet – Fetching Result	nishatiwari1719	2022-02-17	https://www.geeksforgeeks.org/servlet-fetching...	easy	17	02	2022
4	Suffix Sum Array	rohit768	2022-02-21	https://www.geeksforgeeks.org/suffix-sum-array/	easy	21	02	2022

In [30]:
```
# Day-wise Analysis of Articles Frequency

data.groupby('day').size().plot(kind = 'bar')
```

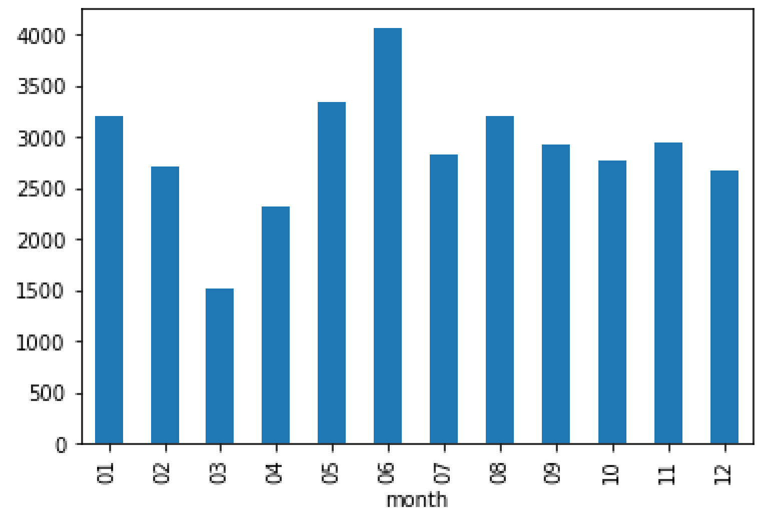
Out[30]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x829e793f70>
```



In [32]:
```
# Month-wise Analysis of Articles Frequency

data.groupby('month').size().plot(kind = 'bar')
```

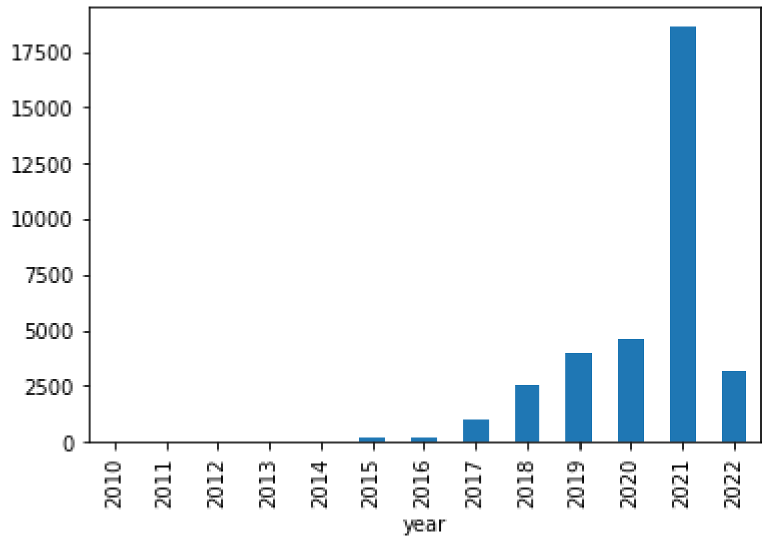
Out[32]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x829e8a69a0>
```



```
In [33]:  # Year-wise Analysis of Articles Frequency

          data.groupby('year').size().plot(kind = 'bar')
```

```
Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x829e936b80>
```



```
In [35]:  # Finding articles of your favorite Author

          data[data['author_id'] == 'GeeksforGeeks']
```

Out[35]:

	title	author_id	last_updated	link	category	day	month	year
8	Free Resume Builder By GeeksforGeeks – Create ...	GeeksforGeeks	2022-02-08	https://www.geeksforgeeks.org/free-resume-buil...	easy	08	02	2022
23	FREE GATE CS 2022 Mock Test – All India Mock B...	GeeksforGeeks	2022-02-16	https://www.geeksforgeeks.org/free-gate-cs-202...	easy	16	02	2022
25	Amazon WOW Internship Interview Experience 2021	GeeksforGeeks	2022-01-24	https://www.geeksforgeeks.org/amazon-wow-inter...	easy	24	01	2022
31	Bi-Wizard School Coding Tournament By Geeksfor...	GeeksforGeeks	2022-02-02	https://www.geeksforgeeks.org/bi-wizard-school...	easy	02	02	2022
43	FREE Online Courses By GeeksforGeeks – Learn N...	GeeksforGeeks	2022-01-06	https://www.geeksforgeeks.org/free-online-cour...	easy	06	01	2022
...	...	...	...	...	...	...	...	...
34569	Data Structures | Queue | Question 11	GeeksforGeeks	2021-06-28	https://www.geeksforgeeks.org/data-structures-...	expert	28	06	2021
34570	Data Structures | Binary Trees | Question 1	GeeksforGeeks	2021-06-28	https://www.geeksforgeeks.org/data-structures-...	expert	28	06	2021
34571	Amazon Interview | Set 9	GeeksforGeeks	2017-04-28	https://www.geeksforgeeks.org/amazon-interview...	expert	28	04	2017
34572	Python Program for Rat in a Maze | Backtracking-2	GeeksforGeeks	2021-08-02	https://www.geeksforgeeks.org/python-program-f...	expert	02	08	2021
34573	Data Structures and Algorithms | Set 21	GeeksforGeeks	2017-03-27	https://www.geeksforgeeks.org/data-structures-...	expert	27	03	2017

11932 rows × 8 columns

In [36]:

```
# Finding articles based on tags

tag = 'Algorithm'.lower()
data1 = data.values
for i in range(len(data1)):
    if tag in data1[i][0].lower():
        print(data1[i][0], data1[i][3])
```

```
Boyer-Moore Majority Voting Algorithm https://www.geeksforgeeks.org/boyer-moore-majority-voting-algorithm/ (h
ttps://www.geeksforgeeks.org/boyer-moore-majority-voting-algorithm/)
Java Program to Implement CAS (Compare and Swap) Algorithm https://www.geeksforgeeks.org/java-program-to-impl
ement-cas-compare-and-swap-algorithm/ (https://www.geeksforgeeks.org/java-program-to-implement-cas-compare-an
d-swap-algorithm/)
Java Program to Implement the RSA Algorithm https://www.geeksforgeeks.org/java-program-to-implement-the-rsa-a
lgorithm/ (https://www.geeksforgeeks.org/java-program-to-implement-the-rsa-algorithm/)
Java Program to Implement Shunting Yard Algorithm https://www.geeksforgeeks.org/java-program-to-implement-shu
nting-yard-algorithm/ (https://www.geeksforgeeks.org/java-program-to-implement-shunting-yard-algorithm/)
The Slowest Sorting Algorithms https://www.geeksforgeeks.org/the-slowest-sorting-algorithms/ (https://www.gee
ksforgeeks.org/the-slowest-sorting-algorithms/)
Comparison between Tarjan’s and Kosaraju’s Algorithm https://www.geeksforgeeks.org/comparision-between-tarjan
s-and-kosarajus-algorithm/ (https://www.geeksforgeeks.org/comparision-between-tarjans-and-kosarajus-algorith
m/)
Time-Space Trade-Off in Algorithms https://www.geeksforgeeks.org/time-space-trade-off-in-algorithms/ (http
s://www.geeksforgeeks.org/time-space-trade-off-in-algorithms/)
Basic understanding of Jarvis-Patrick Clustering Algorithm https://www.geeksforgeeks.org/basic-understanding-
of-jarvis-patrick-clustering-algorithm/ (https://www.geeksforgeeks.org/basic-understanding-of-jarvis-patrick-
clustering-algorithm/)
```

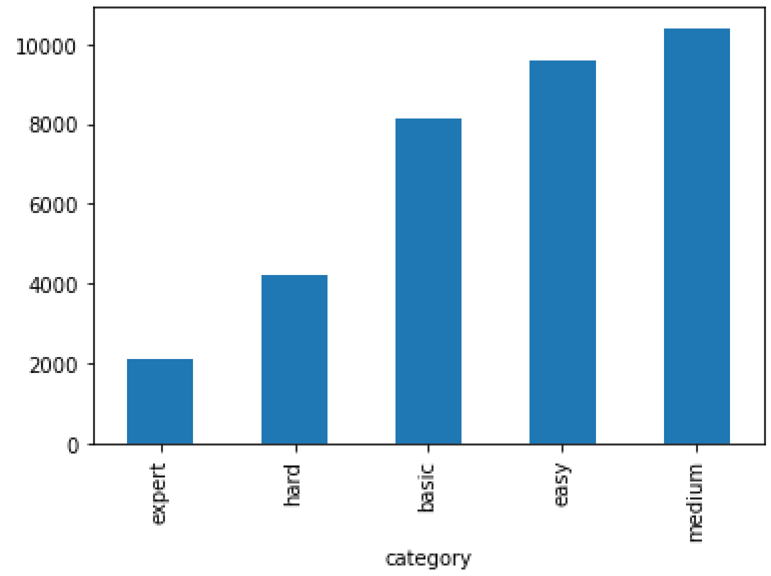
In [37]:

```
# Category Analysis Distribution

data.groupby('category').size().sort_values(ascending = True).plot(kind = 'bar')
```

Out[37]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x829d5f63d0>
```



In [46]:

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

In [47]:

```
data.author_id = le.fit_transform(data.author_id)
data.category = le.fit_transform(data.category)
```

In [48]:

```
data.head()
```

Out[48]:

	title	author_id	last_updated	link	category	day	month	year
0	5 Best Practices For Writing SQL Joins	4016	2022-02-21	https://www.geeksforgeeks.org/5-best-practices...	1	21	02	2022
1	Foundation CSS Dropdown Menu	2925	2022-02-20	https://www.geeksforgeeks.org/foundation-css-d...	1	20	02	2022
2	Top 20 Excel Shortcuts That You Need To Know	4016	2022-02-17	https://www.geeksforgeeks.org/top-20-excel-sho...	1	17	02	2022
3	Servlet – Fetching Result	3700	2022-02-17	https://www.geeksforgeeks.org/servlet-fetching...	1	17	02	2022
4	Suffix Sum Array	4298	2022-02-21	https://www.geeksforgeeks.org/suffix-sum-array/	1	21	02	2022

In [55]:

```
x = data.drop(['category', 'title', 'last_updated', 'link'], axis = 1)
```

In [56]:

```
y = data.category
```

In [57]:

```
x.shape
```

Out[57]:

```
(34455, 4)
```

In [58]:

```
y.shape
```

Out[58]:

```
(34455,)
```

In [63]:

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.2)
```

In [64]:

```
classifier= DecisionTreeClassifier(criterion='entropy', random_state=0)
classifier.fit(X_train, y_train)
```

Out[64]:

```
DecisionTreeClassifier(criterion='entropy', random_state=0)
```

In [65]:

```
y_pred = classifier.predict(X_test)
```

In [66]:

```
print("Training Accuracy :", classifier.score(X_train, y_train))
print("Testing Accuracy :", classifier.score(X_test, y_test))
```

```
Training Accuracy : 0.7794224350602235
Testing Accuracy : 0.30474531998258597
```