Week 7 Assignment

This week we are collecting some data from Reddit and doing some basic EDA on it. You should:

- create your Reddit account and API keys
- collect data from a subreddit of your choice
  - at a minimum, collect the posts from the subreddit; optionally collect comments on the posts
- save the data to a SQLite3 database
- perform some basic EDA on the data
  - create at least 2 plots
- write a short analysis at the end describing the process and results
- turn in the Jupyter Notebook and PDF printout or export to the week 7 dropbox

***Optional* advanced section**

- Practice SQL queries and select a subsection of the posts you collected
- Modify your code to collect data beyond the 1000 item limit
- Collect comments from the posts for analysis next week and do some EDA on the comments (e.g. who is the top commenter, which commenters have the most up and down votes or most controversial posts, etc)
- examine n-grams (bigrams, trigrams) or collocations

Note: There is no solution file for this week.

Collecting data from Reddit

In [7]:

```
#import praw
#import pandas as pd

#import credentials
```

In [8]:

```
reddit = praw.Reddit(client_id=credentials.client_id,
                     client_secret=credentials.client_secret,
                     user_agent=credentials.user_agent)
```

In [9]:

```
co_subreddit = reddit.subreddit('Python').hot(limit=10)
```

In [12]:

```
reddit_data = {'title': [],
             'link': [],
             'author': [],
             'n_comments': [],
             'score': [],
              'text': []}


co_subreddit = reddit.subreddit('colorado').hot(limit=None)

for post in list(co_subreddit):
    reddit_data['title'].append(post.title)
    reddit_data['link'].append(post.permalink)
    if post.author is None:
        reddit_data['author'].append('')
    else:
        reddit_data['author'].append(post.author.name)

    reddit_data['n_comments'].append(post.num_comments)
    reddit_data['score'].append(post.score)
    reddit_data['text'].append(post.selftext)
```

In [13]:

```
co_df = pd.DataFrame(reddit_data)
```

In [14]:

```
co_df
```

Out[14]:

	title	link	author
0	Wind power has gone from just an idea to one o...	/r/Colorado/comments/1dp2ozn/wind_power_has_go...	thecoloradosun
1	Last light turns dunes purple in Great Sand Du...	/r/Colorado/comments/1dp3y93/last_light_turns_...	_raidboss
2	Stunning sunset last night. Longmont, CO.	/r/Colorado/comments/1doxkkg/stunning_sunset_l...	razzledazzle125
3	Mysterious monolith appears in Northern Colorado	/r/Colorado/comments/1dpc62f/mysterious_monoli...	Knightbear49
4	From a hike in Woodland Park	/r/Colorado/comments/1dp66i5/from_a_hike_in_wo...	invincible789
...	...	...	...
637	Calling all CO musicians. What's been your exp...	/r/Colorado/comments/1b88yav/calling_all_co_mu...	J8R9L
638	Colorado grandmother awarded $3.76M after bung...	/r/Colorado/comments/1b7kc2k/colorado_grandmot...	nbcnews
639	State lawmakers introduce bill to reintroduce ...	/r/Colorado/comments/1b7ailg/state_lawmakers_i...	ButterscotchEmpty535
640	Congressional Candidate Now Supports a Nationa...	/r/Colorado/comments/1b7guy3/congressional_can...	Odd_Cranberry_8059
641	Red Rocks Through the	/r/Colorado/comments/1b7f2ix/red_rocks_through...	vegandread

title	link	author
Ruins on Mt. Falcon		

642 rows × 6 columns

Save data to sqlite

In [15]:

```
#import sqlite3

con = sqlite3.connect("data/co_reddit.sqlite")
co_df.to_sql('posts', con, if_exists='replace', index=False)
```

Out[15]:

```
642
```

In [16]:

```
co_df_check = pd.read_sql_query('SELECT * FROM posts;', con)
# it's best to close the connection when finished
con.close()
co_df_check
```

Out[16]:

	title	link	author
0	Wind power has gone from just an idea to one o...	/r/Colorado/comments/1dp2ozn/wind_power_has_go...	thecoloradosun
1	Last light turns dunes purple in Great Sand Du...	/r/Colorado/comments/1dp3y93/last_light_turns_...	_raidboss
2	Stunning sunset last night. Longmont, CO.	/r/Colorado/comments/1doxkkg/stunning_sunset_l...	razzledazzle125
3	Mysterious monolith appears in Northern Colorado	/r/Colorado/comments/1dpc62f/mysterious_monoli...	Knightbear49
4	From a hike in Woodland Park	/r/Colorado/comments/1dp66i5/from_a_hike_in_wo...	invincible789
...	...	...	...
637	Calling all CO musicians. What's been your exp...	/r/Colorado/comments/1b88yav/calling_all_co_mu...	J8R9L
638	Colorado grandmother awarded $3.76M after bung...	/r/Colorado/comments/1b7kc2k/colorado_grandmot...	nbcnews
639	State lawmakers introduce bill to reintroduce ...	/r/Colorado/comments/1b7ailg/state_lawmakers_i...	ButterscotchEmpty535
640	Congressional Candidate Now Supports a Nationa...	/r/Colorado/comments/1b7guy3/congressional_can...	Odd_Cranberry_8059
641	Red Rocks Through the	/r/Colorado/comments/1b7f2ix/red_rocks_through...	vegandread

title	link	author
Ruins on Mt. Falcon		

642 rows × 6 columns

Basic EDA

In [17]:

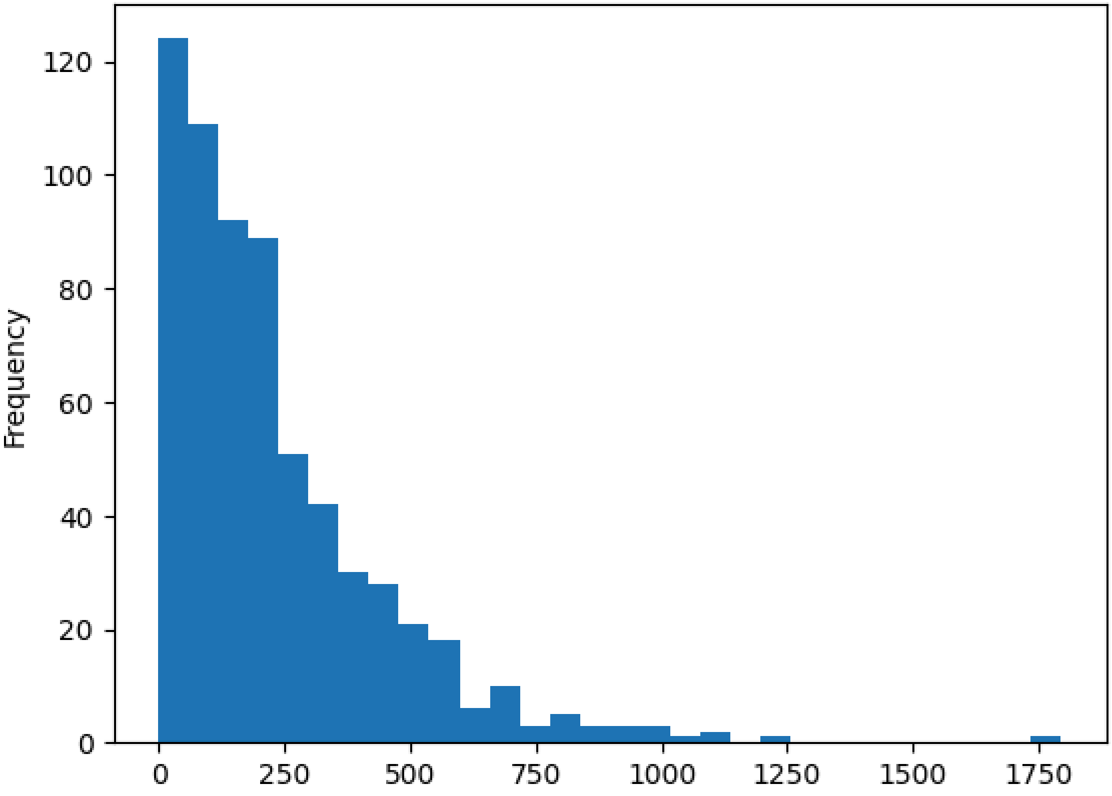
```
import matplotlib.pyplot as plt
```

In [18]:

```
co_df['score'].plot.hist(bins=30)
```

Out[18]:

```
<Axes: ylabel='Frequency'>
```



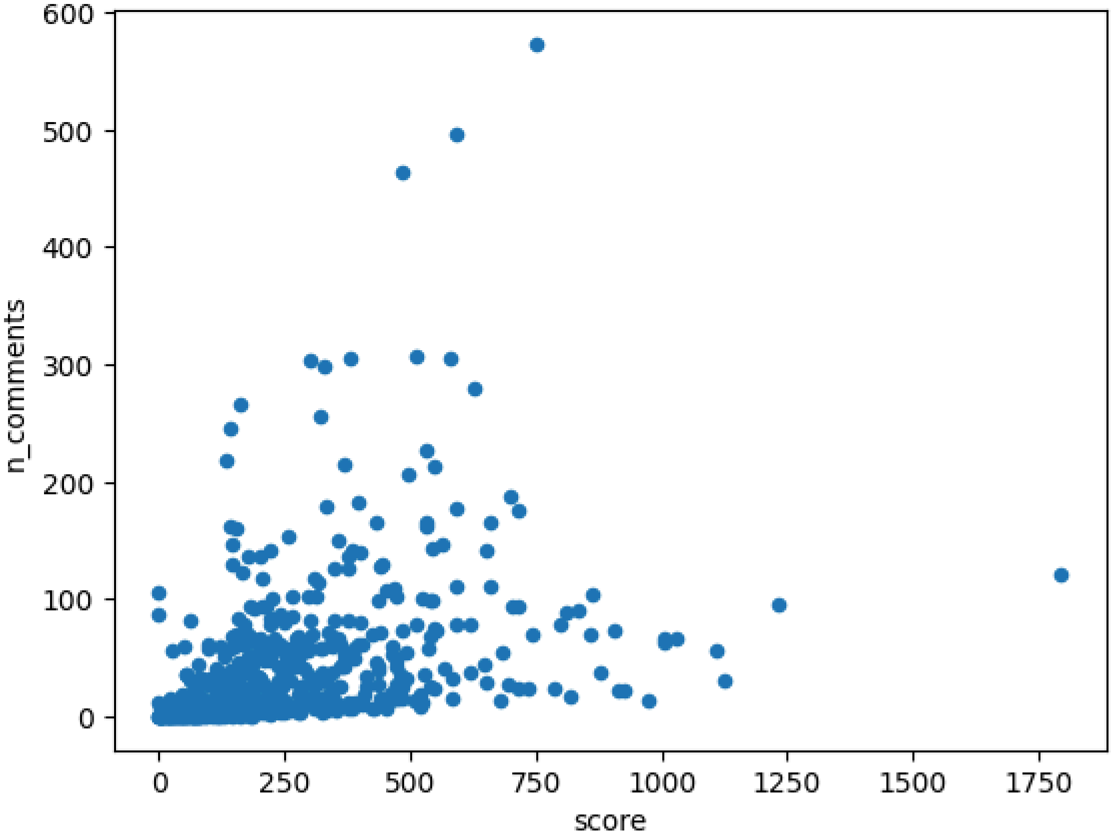
High number of posts close to a zero score is likely due to new posts

In [20]:

```
co_df.plot.scatter(x='score', y='n_comments')
```

Out[20]:

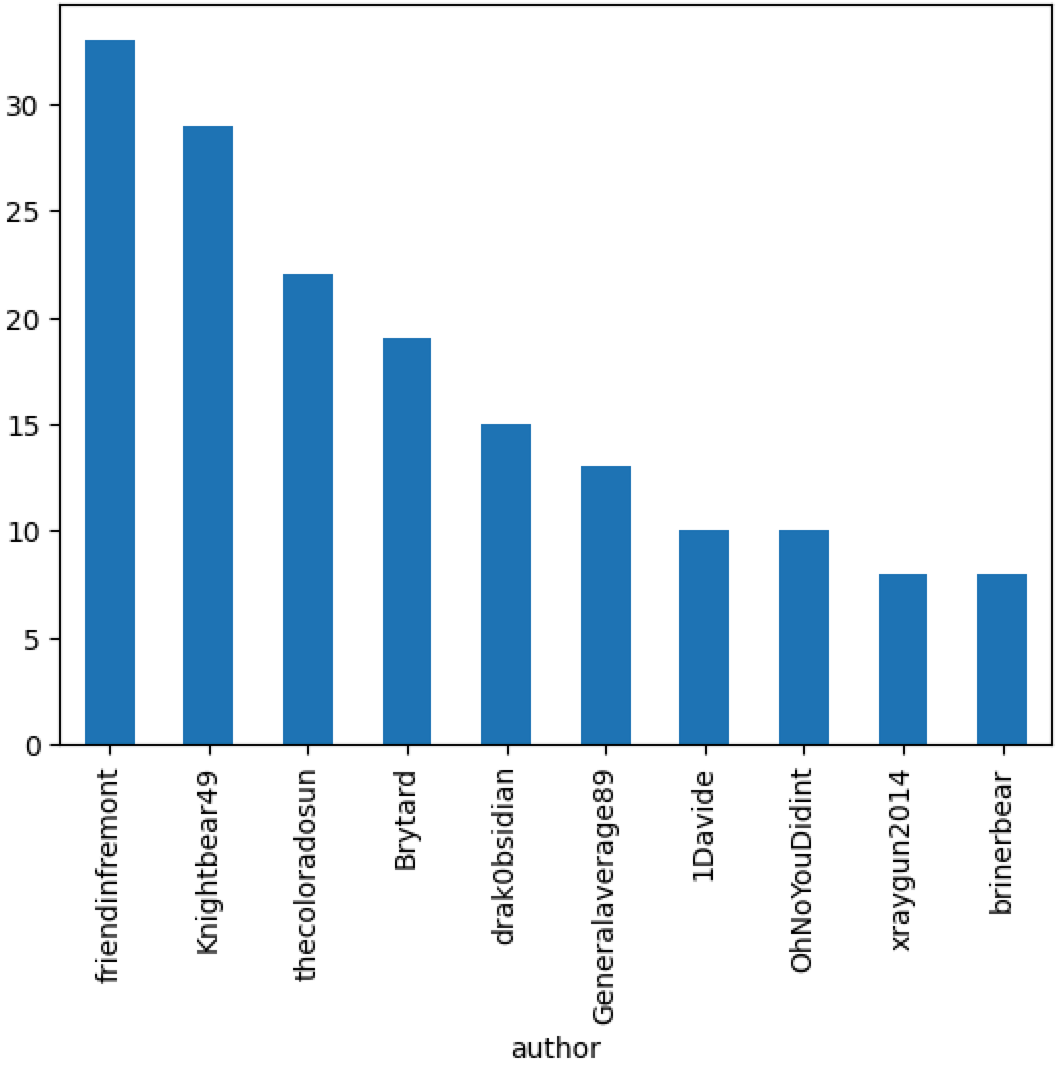
```
<Axes: xlabel='score', ylabel='n_comments'>
```



Roughly positive relationship between the score and number of comments

```
In [21]: co_df['author'].value_counts()[:10].plot.bar()
```

```
Out[21]: <Axes: xlabel='author'>
```



Frequency of users who post. friendinfremont has been busy, let's see what they have been posting.

In [23]:
```
co_df[co_df['author'] == 'friendinfremont']
```

Out[23]:

	title	link	author	n_com
16	Adaptive mountain bikes allow for a return to ...	/r/Colorado/comments/1do71lj/adaptive_mountain...	friendinfremont	
22	There's good news about Colorado youth mental ...	/r/Colorado/comments/1dngvpw/theres_good_news_...	friendinfremont	
23	First Creek renovations to bring back what was...	/r/Colorado/comments/1dnexma/first_creek_renov...	friendinfremont	
86	Coloradans from every political party, age gro...	/r/Colorado/comments/1dedeyc/coloradans_from_e...	friendinfremont	
106	'Cheesy, fungus-y, death-y': The science behin...	/r/Colorado/comments/1ddju6a/cheesy_fungusy_de...	friendinfremont	
131	Trial by fire: A new generation of wildland fi...	/r/Colorado/comments/1d9j4wv/trial_by_fire_a_n...	friendinfremont	
135	Book bans have become a powerful censorship to...	/r/Colorado/comments/1d8w0lc/book_bans_have_be...	friendinfremont	
147	The main highway between Gunnison and Montrose...	/r/Colorado/comments/1d7zndt/the_main_highway_...	friendinfremont	
152	Eldora withdraws objections,	/r/Colorado/comments/1d7dut5/eldora_withdraws_...	friendinfremont	

	title	link	author	n_com
	formalizing ski p...			
167	The past, present and future of windmills conv...	/r/Colorado/comments/1d4vs7m/the_past_present_...	friendinfremont	
223	Wolves killing livestock was expected, but is ...	/r/Colorado/comments/1cwfj80/wolves_killing_li...	friendinfremont	
225	‘Plovers’ are for Lovers: How the Mountain Plo...	/r/Colorado/comments/1cwfcod/plovers_are_for_l...	friendinfremont	
238	Newcomers Ski Group brings Black immigrants to...	/r/Colorado/comments/1cu76fg/newcomers_ski_gro...	friendinfremont	
248	9 common myths about newcomers to Colorado	/r/Colorado/comments/1csm2di/9_common_myths_ab...	friendinfremont	
258	Facing climate change in one of Colorado’s mos...	/r/Colorado/comments/1crtl0m/facing_climate_ch...	friendinfremont	
264	Dearfield once thrived in the plains of Colora...	/r/Colorado/comments/1cr1m48/dearfield_once_th...	friendinfremont	
325	Colorado Natural Heritage Program	/r/Colorado/comments/1cn6ytd/colorado_natural_...	friendinfremont	

	title	link	author	n_com
	launches his...			
337	Last night at Nimo's	/r/Colorado/comments/1cllldr/last_night_at_nimos/	friendinfremont	
338	Inside Denver's only micro-community for trans...	/r/Colorado/comments/1clllo1/inside_denvers_on...	friendinfremont	
397	How advocates around Colorado are fighting per...	/r/Colorado/comments/1cb67mq/how_advocates_aro...	friendinfremont	
406	Colorado UPK: With reduced funding for three-y...	/r/Colorado/comments/1caeubp/colorado_upk_with...	friendinfremont	
420	Eckert, Colorado becomes Sandhill Crane rest-s...	/r/Colorado/comments/1c80r6t/eckert_colorado_b...	friendinfremont	
438	Colorado HOAs can no longer stop you from hard...	/r/Colorado/comments/1c4p8vj/colorado_hoas_can...	friendinfremont	
469	Conductor Brad's last ride	/r/Colorado/comments/1c0mk39/conductor_brads_l...	friendinfremont	
485	Colorado Science: Researchers discover secret ...	/r/Colorado/comments/1bwnif9/colorado_science_...	friendinfremont	
498	A secretary drives the school bus and the supe...	/r/Colorado/comments/1bu0ssp/a_secretary_drive...	friendinfremont	
506	Eldora ski patrollers	/r/Colorado/comments/1bt2gdn/eldora_ski_patrol...	friendinfremont	

	title	link	author	n_com
	vote to form union			
510	Resign and advance: How Colorado police office...	/r/Colorado/comments/1brwx05/resign_and_advanc...	friendinfremont	
523	'The Inside Report' may have published its fin...	/r/Colorado/comments/1bpypu2/the_inside_report...	friendinfremont	
549	Resource Center opens for unhoused people who ...	/r/Colorado/comments/1bl0xly/resource_center_o...	friendinfremont	
564	A Boulder bistro brings unity and reprieve	/r/Colorado/comments/1bjg9qz/a_boulder_bistro_...	friendinfremont	
573	Colorado’s new Universal Preschool program rel...	/r/Colorado/comments/1bhtzbm/colorados_new_uni...	friendinfremont	
631	Wolf myth-busting with wildlife biologist Kevi...	/r/Colorado/comments/1b8zec3/wolf_mythbusting_...	friendinfremont	

Analysis

We started by importing praw and pandas to collect data from Reddit, specifically the 'Python' subreddit. Using our Reddit credentials, we gathered key details like titles, links, authors, comments, scores, and text from all the hot posts. We stored this info in a dictionary and then turned it into a DataFrame.

Next, we saved this DataFrame to an SQLite database and checked to make sure everything was saved correctly. After that, we did some basic exploratory data analysis (EDA) with matplotlib. We created a histogram of post scores and saw most scores were close to zero, likely because the posts were new. A scatter plot showed a positive correlation between scores and the number of comments. We also made a bar plot of the top 10 most active authors and found that 'friendinfremont' was super busy posting.

In the end, we successfully collected, stored, and did some initial analysis of the Reddit data, giving us a good foundation for deeper analysis later on.

In [ ]: