import, preparation

In [1]:
```
import numpy as np
import pandas as pd
%matplotlib inline
%config Completer.use_jedi = False
```

In [2]:
```
!ls
!echo ----
!cd ..
!pwd
```

```
Untitled.ipynb   Untitled4.ipynb             data36_popup_analysis.ipynb
Untitled1.ipynb  best_bet.ipynb              yahoo_finance.ipynb
Untitled2.ipynb  coinbase_api_example.ipynb
Untitled3.ipynb  data36_popup.ipynb
----
/home/slackroo/JDS/data_practice/API_practice
```

1. opening files

In [3]:
```
#each and every pageview in a log
pageviews = pd.read_csv('/home/slackroo/JDS/data_practice/pageviews.tsv', sep='\t
            names = ['date', 'time', 'country', 'user_id', 'event', 'source', 'pa
```

In [5]:
```
pageviews.head(5)
```

Out[5]:

	date	time	country	user_id	event	source	
0	2021-02-01	00:00:19.679	MY	u8515925	b'pageview_blog	NaN	https://data36.co tutorial-1-b
1	2021-02-01	00:00:31.810	US	u8544901	b'pageview_blog	https://www.google.com	https://data36.cc nested-loo
2	2021-02-01	00:00:57.138	NaN	u8535534	b'pageview_blog	https://www.google.com	https://data36.c
3	2021-02-01	00:01:30.771	MY	u8594125	b'pageview_blog	https://www.google.com/	https://data36.cc import-data
4	2021-02-01	00:02:31.284	NaN	u8564427	b'pageview_blog	https://www.google.com/	https://data36.com/ bias-ty

In [4]:
```
#each and every newsletter subscription in a log
newsletters = pd.read_csv('/home/slackroo/JDS/data_practice/newsletter.tsv', sep=
            names = ['date', 'time', 'country', 'user_id', 'event', 'button', 'pa
```

In [5]:
```
#each and every click on the JDS site in a log
jds = pd.read_csv('/home/slackroo/JDS/data_practice/jds_site_click.tsv', sep='\t'
            names = ['date', 'time', 'country', 'user_id1', 'event', 'button', 'p
```

2. data discovery

toplist of visited pages

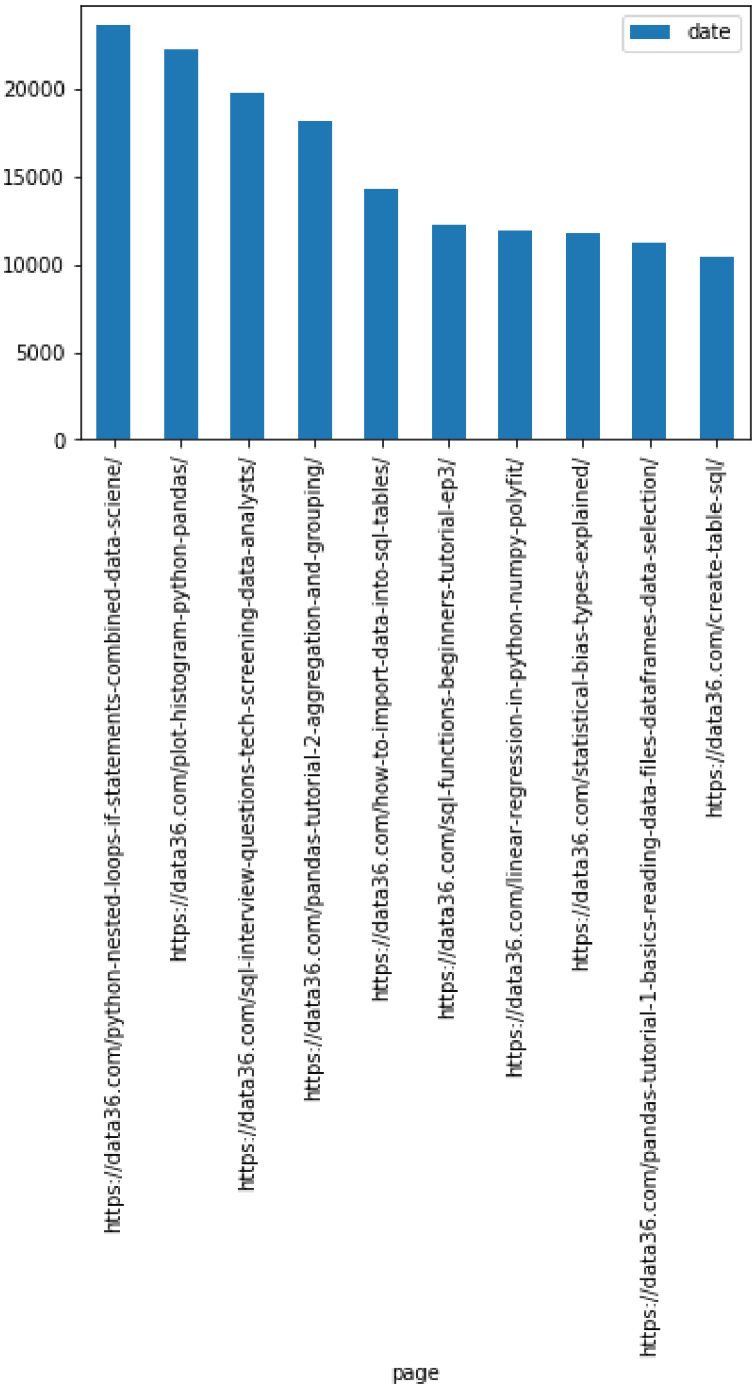
In [50]:
```
pageview_count = pageviews.groupby('page').count().sort_values(by='date', ascendi
pageview_count.head(80)[['date']]
```

Out[50]:

page	date
https://data36.com/python-nested-loops-if-statements-combined-data-sciene/	23634
https://data36.com/plot-histogram-python-pandas/	22262
https://data36.com/sql-interview-questions-tech-screening-data-analysts/	19844
https://data36.com/pandas-tutorial-2-aggregation-and-grouping/	18238
https://data36.com/how-to-import-data-into-sql-tables/	14367
https://data36.com/sql-functions-beginners-tutorial-ep3/	12272
https://data36.com/linear-regression-in-python-numpy-polyfit/	11913
https://data36.com/statistical-bias-types-explained/	11808
https://data36.com/pandas-tutorial-1-basics-reading-data-files-dataframes-data-selection/	11278
https://data36.com/create-table-sql/	10480

In [7]:
```
#the same thing on a bar chart
pageview_count.head(10)[['date']].plot.bar()
```

Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x7f85ae7e1100>



In [8]:
```
#saving top pages into a Python list for later
#top_pages = list(pageview_count.head(20).index)
```

toplist: on which pages do people subscribe

Newsletter subscriptions gives interested users

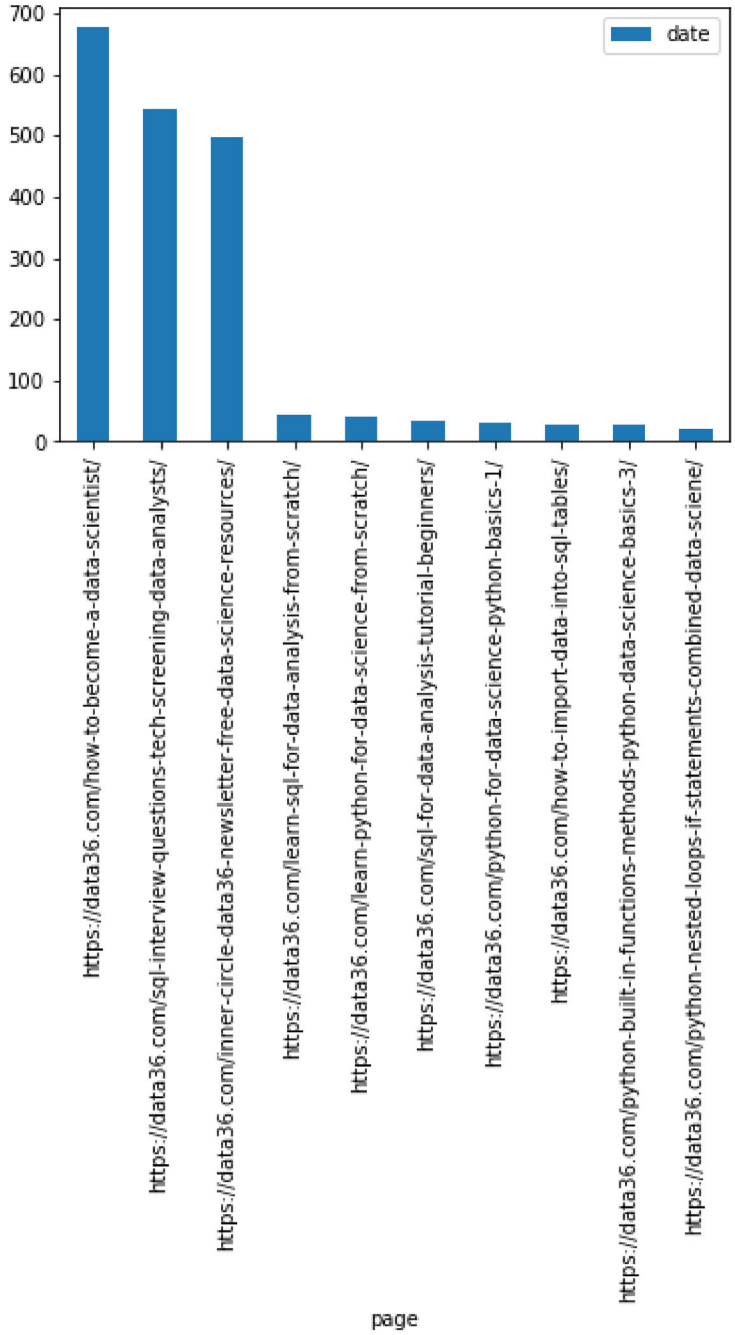
In [8]:
```
newsletters.groupby('page').count()[['date']].sort_values(by='date',ascending = F
```

Out[8]:

page	date
https://data36.com/how-to-become-a-data-scientist/	676
https://data36.com/sql-interview-questions-tech-screening-data-analysts/	543
https://data36.com/inner-circle-data36-newsletter-free-data-science-resources/	499
https://data36.com/learn-sql-for-data-analysis-from-scratch/	43
https://data36.com/learn-python-for-data-science-from-scratch/	41
https://data36.com/sql-for-data-analysis-tutorial-beginners/	35
https://data36.com/python-for-data-science-python-basics-1/	32
https://data36.com/how-to-import-data-into-sql-tables/	27
https://data36.com/python-built-in-functions-methods-python-data-science-basics-3/	26
https://data36.com/python-nested-loops-if-statements-combined-data-sciene/	22

In [9]:
```
#same thing on a bar chart
newsletters.groupby('page').count()[['date']].sort_values(by='date', ascending =
```

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x7f85ae7229d0>



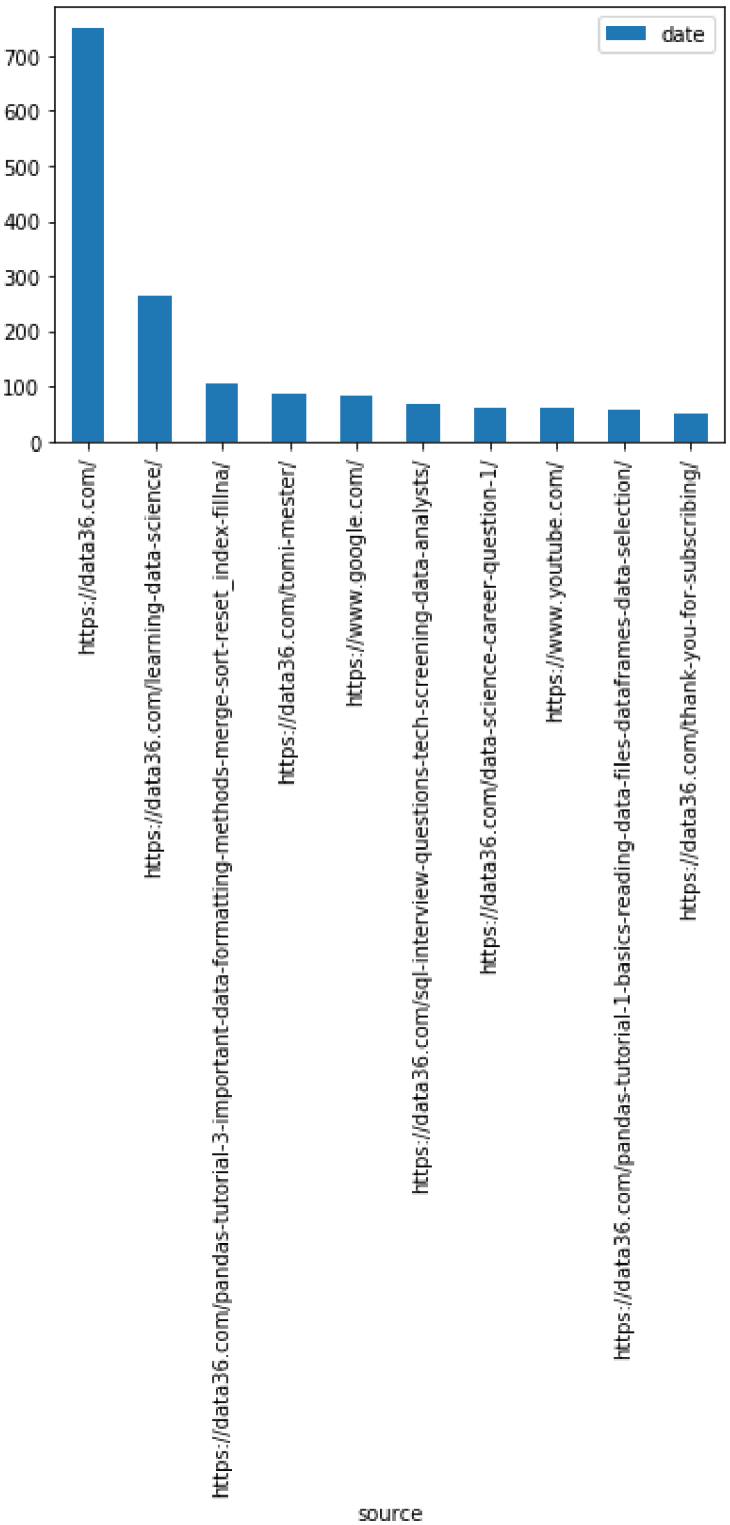
Finding where people come in for the main course selling

pages

In [10]:
```
jds_sources = pageviews[pageviews.page == 'https://data36.com/the-junior-data-sci
sql_sources = pageviews[pageviews.page == 'https://data36.com/sql-for-aspiring-da
htb_sources = pageviews[pageviews.page == 'https://data36.com/how-to-become-a-dat
```

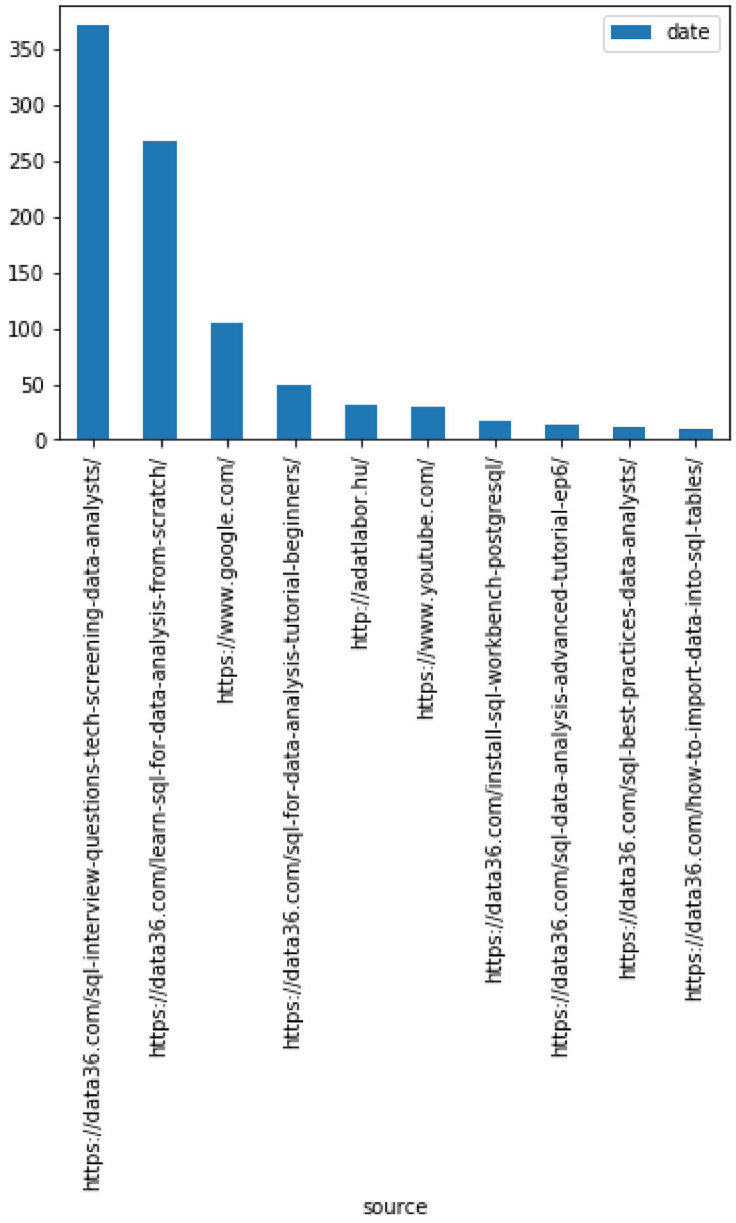
In [11]:
```
jds_sources.groupby('source').count()[['date']].sort_values(by='date', ascending
```

Out[11]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f85ae6b9310>
```



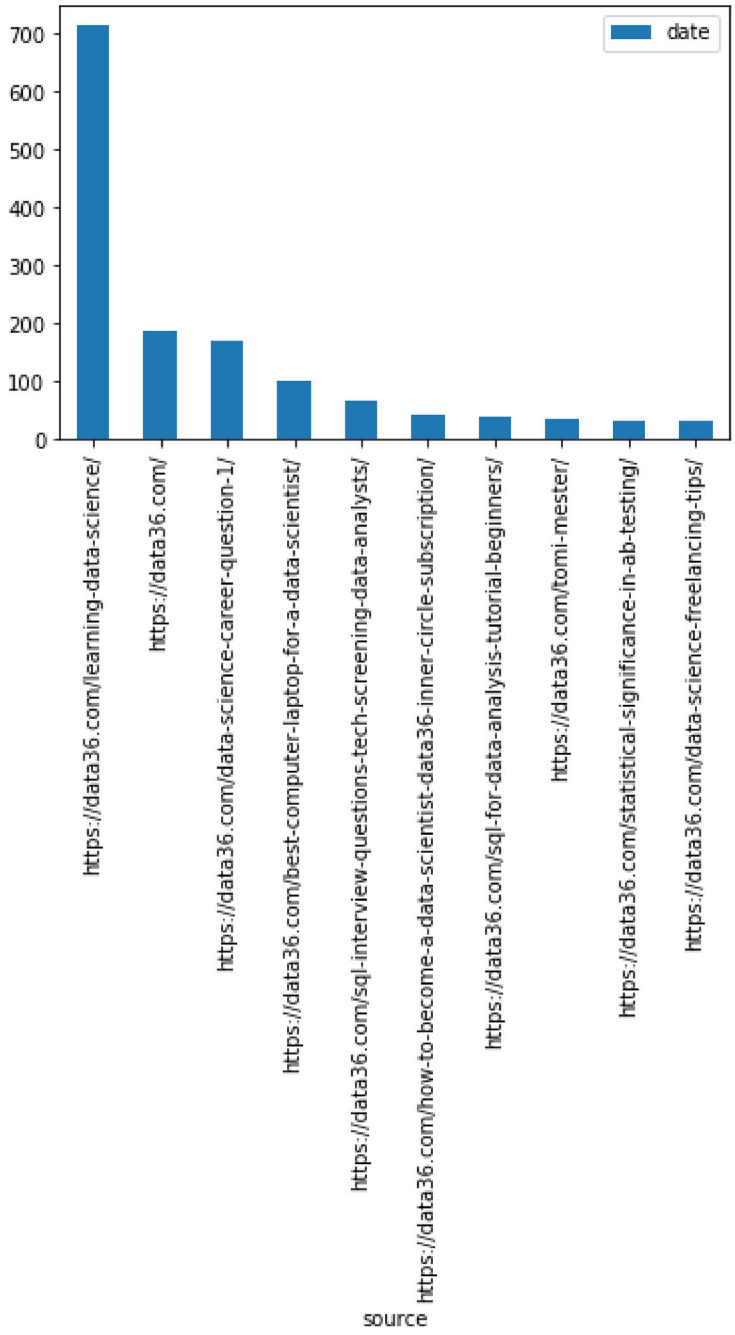
In [12]:
```
sql_sources.groupby('source').count()[['date']].sort_values(by='date', ascending
```

Out[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f85ae680310>
```



In [13]:
```
htb_sources.groupby('source').count()[['date']].sort_values(by='date', ascending
```

Out[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f85ae5516a0>
```



error! click through rate in most read articles

In [14]:
```
page = 'https://data36.com/learning-data-science/'
print(pageviews[pageviews.page == page].count().date)
newsletters[newsletters.page == page].date.count()
```

```
5956
```

Out[14]:
```
3
```

In [15]:
```
pages = ['https://data36.com/learning-data-science/',
        'https://data36.com/become-data-scientist-7-plus-1-selfish-reasons/',
        'https://data36.com/data-science-career-question-1/',
        'https://data36.com/presentation-tips-for-data-professionals/']

for i in pages:
    page = i
    print((pageviews[pageviews.page == page].count().date), newsletters[newslette
```

```
5956 3
544 0
2090 1
727 0
```

conclusion: the conversion rate of the popup is terrible!

3. DEFINING CONVERSION EVENTS

how to become a DS micro conversion

In [16]:
```
howto_visited = pageviews[pageviews.page == 'https://data36.com/how-to-become-a-d
howto_visited = howto_visited.drop_duplicates('user_id', keep='last').groupby('so
howto_visited = howto_visited[howto_visited.date > 5]
howto_visited = howto_visited.reset_index()[['source','date']]
```

how to become a DS macro conversion

In [17]:
```
howto_converted = newsletters[newsletters.page == 'https://data36.com/how-to-becc
    'user_id', 'date', 'time']]
howto_converted = howto_converted.drop_duplicates('user_id')
```

In [18]:
```
howto_converted
```

Out[18]:

	user_id	date	time
0	u8446607	2021-02-01	02:35:44.972
12	u8528818	2021-02-01	12:29:09.329
13	u8440763	2021-02-01	12:56:49.099
18	u8572725	2021-02-01	14:23:07.309
23	u8430160	2021-02-01	20:50:30.942
...	...	...	...
2133	u8554835	2021-04-29	14:26:12.328
2138	u8603988	2021-04-29	16:23:22.549
2142	u8601163	2021-04-30	08:51:38.771
2150	u8496510	2021-04-30	15:31:07.538
2155	u8420346	2021-04-30	18:50:06.163

411 rows × 3 columns

JDS micro conversion

In [19]:
```
jds_visited = pageviews[pageviews.page == 'https://data36.com/the-junior-data-sci
jds_visited = jds_visited.drop_duplicates('user_id', keep='last').groupby('source
jds_visited = jds_visited[jds_visited.date > 5]
jds_visited = jds_visited.reset_index()[['source','date']]
```

JDS macro conversion

In [20]:
```
jds_converted = jds[jds.button == "Subscribe!"][['user_id1', 'date', 'time']].dro
```

JDS vs HOWTO converted people

is there an overlap? how big?

In [21]:
```
howto_converted.count()
```

Out[21]:
```
user_id    411
date       411
time       411
dtype: int64
```

In [22]:
```
jds_converted.count()
```

Out[22]:
```
user_id1    216
date        216
time        216
dtype: int64
```

how many people joined both JDS and how to

In [23]:
```
jds_converted.merge(howto_converted, left_on = 'user_id1', right_on = 'user_id',
```

Out[23]:
```
user_id1    42
date_x      42
time_x      42
user_id     42
date_y      42
time_y      42
dtype: int64
```

hypothesis #1 is disproven!!!

4. CREATING A "SUPER TABLE" WITH: VISITS, MICRO- & MACRO-CONVERSIONS

this is the tricky part -- creating last click attribution

In [25]:
```
#merging the visited pages to the macro conversion event
howto_converted_source = howto_converted.merge(pageviews, left_on = 'user_id', ri
```

In [26]:
```
#filtering for only those events that happened BEFORE the actual conversion
howto_converted_source = howto_converted_source[
    (howto_converted_source.time_x > howto_converted_source.time_y) & (
        howto_converted_source.date_x == howto_converted_source.date_y)]
```

In [27]:
```
#filtering for the LAST event before the conversion, remove duplicates, group, co
howto_converted_source = howto_converted_source[
    howto_converted_source.page == 'https://data36.com/how-to-become-a-data-scier
                  ].drop_duplicates('user_id', keep='last').groupby('source').cou
    by='date_x', ascending = False)
```

In [28]:
```
#formatting
howto_converted_source = howto_converted_source.reset_index()
```

In [29]:
```
###this is the list of the last articles people visited before coming
###to the https://data36.com/how-to-become-a-data-scientist/ and actually subscri

howto_converted_source.head(3)
```

Out[29]:

	source	date_x
0	https://data36.com/learning-data-science/	151
1	https://data36.com/data-science-career-questio...	39
2	https://data36.com/	27

What converts JDS people

this is the tricky part -- creating last click attribution

In [30]:
```
#merging the visited pages to the macro conversion event
jds_converted_source = jds_converted.merge(pageviews, left_on = 'user_id1', right
```

In [31]:
```
#filtering for only those events that happened BEFORE the actual conversion
jds_converted_source = jds_converted_source[
    (jds_converted_source.time_x > jds_converted_source.time_y) & (
        jds_converted_source.date_x == jds_converted_source.date_y)]
```

In [32]:
```
#filtering for the LAST event before the conversion, remove duplicates, group, co
jds_converted_source = jds_converted_source[jds_converted_source.page == 'https:/
                    ].drop_duplicates('user_id1', keep='last').groupby('source').co
    by='date_x', ascending = False)
```

In [33]:
```
#formatting
jds_converted_source = jds_converted_source.reset_index() #source
```

In [34]:
```
###this is the list of the last articles people visited before coming
###to the https://data36.com/how-to-become-a-data-scientist/ and actually subscri
jds_converted_source
```

Out[34]:

	source	date_x
0	https://data36.com/	39
1	https://data36.com/learning-data-science/	26
2	https://data36.com/how-to-become-a-data-scient...	11
3	https://data36.com/sql-interview-questions-tec...	7
4	https://data36.com/data-science-career-questio...	7
5	https://data36.com/plot-histogram-python-pandas/	4
6	https://tomimester.medium.com/how-to-break-int...	4
7	https://data36.com/tomi-mester/	4
8	https://data36.com/thank-you-for-subscribing/	4
9	https://data36.com/sql-data-analysis-advanced-...	4
10	https://data36.com/find-data-science-mentor/	3
11	https://www.youtube.com/	3
12	https://data36.com/data36-inner-circle-subscri...	2
13	https://data36.com/page/2/	2
14	https://data36.com/funnel-analysis/	2
15	https://data36.com/python-libraries-packages-d...	2
16	https://data36.com/python-for-data-science-pyt...	2
17	https://data36.com/best-computer-laptop-for-a-...	2
18	https://data36.com/data-coding-101-install-pyt...	2
19	https://data36.com/become-data-scientist-7-plu...	2
20	https://data36.com/pandas-tutorial-1-basics-re...	2
21	https://data36.com/get-job-data-science-analyt...	2
22	https://data36.com/what-is-data-science/	1
23	https://mailchi.mp/	1
24	https://data36.com/computer-setup-data-science/	1
25	https://t.co/	1
26	https://tomimester.medium.com/aspiring-data-sc...	1
27	https://data36.com/beautiful-soup-tutorial-web...	1
28	https://www.google.com/	1
29	https://www.linkedin.com/	1
30	https://data36.com/sublime-text-data-science-r...	1
31	https://data36.com/statistical-significance-in...	1
32	https://data36.com/statistical-bias-types-expl...	1

	source	date_x
33	https://data36.com/statistical-bias-types-exam...	1
34	https://data36.com/statistical-averages-mean-m...	1
35	https://data36.com/sql-best-practices-data-ana...	1
36	https://data36.com/sql-for-data-analysis-tutor...	1
37	https://data36.com/create-table-sql/	1
38	https://data36.com/sql-data-analysis-advanced-...	1
39	https://data36.com/scraping-multiple-web-pages...	1
40	https://data36.com/python-nested-loops-if-stat...	1
41	https://data36.com/python-for-data-science-and...	1
42	https://data36.com/pandas-tutorial-3-important...	1
43	https://data36.com/linear-regression-in-python...	1
44	https://data36.com/data-coding-101-introductio...	1
45	https://data36.com/learn-python-for-data-scien...	1
46	https://data36.com/learn-data-analytics-bash-s...	1
47	https://data36.com/install-sql-workbench-postg...	1
48	https://data36.com/how-to-import-data-into-sql...	1
49	https://data36.com/data-collection/	1
50	android-app://com.google.android.gm/	1

CREATING THE SUPER TABLE

In [35]:
```
howto_visited.head(3)
```

Out[35]:

	source	date
0	https://data36.com/	147
1	https://data36.com/become-data-scientist-7-plu...	7
2	https://data36.com/best-computer-laptop-for-a-...	73

In [36]:
```
howto_converted_source.head(3)
```

Out[36]:

	source	date_x
0	https://data36.com/learning-data-science/	151
1	https://data36.com/data-science-career-questio...	39
2	https://data36.com/	27

In [37]:
```
jds_visited.head(3)
```

Out[37]:

	source	date
0	http://m.facebook.com/	11
1	https://data36.com/	516
2	https://data36.com/beautiful-soup-tutorial-web...	7

In [38]:
```
jds_converted_source.head(3)
```

Out[38]:

	source	date_x
0	https://data36.com/	39
1	https://data36.com/learning-data-science/	26
2	https://data36.com/how-to-become-a-data-scient...	11

In [39]:
```
#formatting the pageviews dataset, then counting the number of pageviews for each
numb_of_pageviews = pageviews.drop_duplicates(subset=['user_id', 'page']).groupby

#removing articles with less than 300 views -- and some formatting
numb_of_pageviews = numb_of_pageviews[numb_of_pageviews.date > 300].reset_index()
```

In [40]:
```
######################
# MERGING EVERYTHING #
######################

super_table = numb_of_pageviews.merge(
    howto_visited, how='left', left_on = 'page', right_on = 'source').merge(
    howto_converted_source, how='left', left_on = 'page', right_on = 'source').me
    jds_visited, how='left', left_on = 'page', right_on = 'source').merge(
    jds_converted_source, how='left', left_on = 'page', right_on = 'source')

super_table = super_table[['page', 'date_x_x', 'date_y', 'date_x_y', 'date', 'dat
super_table.columns = ['page', 'article_reads', 'howto_visited', 'howto_conv', 'j

super_table = super_table.fillna(0)
```

In [41]:
```
super_table.sort_values(by = 'article_reads', ascending = False).to_csv('super_ta
```

In [42]:
```
super_table.sort_values(by = 'howto_conv', ascending = False).head(10)
```

Out[42]:

	page	article_reads	howto_visited	howto_conv	jds_visited	jds_conv
33	https://data36.com/learning-data-science/	4384	521.0	151.0	175.0	26.0
10	https://data36.com/data-science-career-questio...	1628	130.0	39.0	47.0	7.0
0	https://data36.com/	4916	147.0	27.0	516.0	39.0
2	https://data36.com/best-computer-laptop-for-a-...	8878	73.0	13.0	28.0	2.0
59	https://data36.com/sql-interview-questions-tec...	14389	53.0	12.0	58.0	7.0
32	https://data36.com/learn-sql-for-data-analysis...	1363	21.0	7.0	15.0	0.0
14	https://data36.com/data-scientists-day/	330	13.0	6.0	6.0	0.0
70	https://data36.com/tomi-mester/	435	31.0	6.0	49.0	4.0
65	https://data36.com/statistical-significance-in...	2350	23.0	5.0	9.0	1.0
49	https://data36.com/python-nested-loops-if-stat...	20802	6.0	4.0	17.0	1.0

In [55]:
```
pd.set_option('display.max_rows', 500)
pd.set_option('display.max_colwidth', -1)

super_table.sort_values(by = 'jds_conv', ascending = False).reset_index(drop= Tru
```

35	https://data36.com/create-table-sql/	9091	0.0	0.0	0.0	1.(
36	https://data36.com/sql-for-aspiring-data-scientists-7-day-online-course/	1007	0.0	0.0	0.0	0.(
37	https://data36.com/data-science-projects-for-boosting-your-resume/	618	0.0	0.0	0.0	0.(
38	https://data36.com/sql-functions-beginners-tutorial-ep3/	10633	0.0	0.0	0.0	0.(
39	https://data36.com/data-science-freelancing-tips/	476	18.0	3.0	7.0	0.(
40	https://data36.com/sql-join-data-analysis-tutorial-ep5/	327	0.0	0.0	0.0	0.(
41	https://data36.com/sql-where-clause-tutorial-beginners-ep2/	382	0.0	0.0	0.0	0.(
42	https://data36.com/data-science-cv-resume-cover-letter-github/	368	0.0	0.0	0.0	0.(

5. What people do before they subscribe to

JDS

In [44]:
```
#let's see hypothesis #2!
```

how many articles people read before they subscribe to JDS

In [45]:
```
#jds conversions
jds_converted_before = jds[jds.button == "Subscribe!"][['user_id1', 'date', 'time
```

In [46]:
```
#merging the visited pages to the macro conversion event
jds_converted_before = jds_converted_before.merge(
    pageviews, left_on = 'user_id1', right_on = 'user_id', how = 'inner')
```

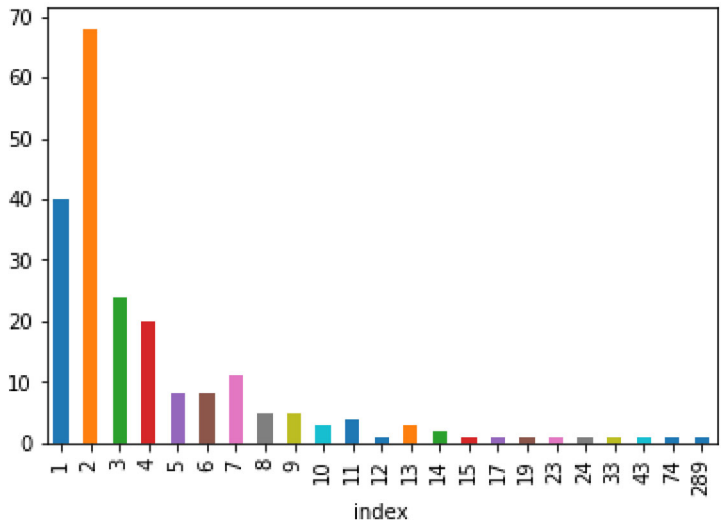
In [47]:
```
#filtering for only those events that happened BEFORE the actual conversion
jds_converted_before = jds_converted_before[(
    jds_converted_before.time_x > jds_converted_before.time_y) & (
        jds_converted_before.date_x >= jds_converted_before.date_y)]
```

In [48]:
```
#formatting, counting, then counting again (How many articles people read before
jds_converted_before.reset_index().groupby('user_id1').count().groupby('index').c
```

Out[48]:
```
index
1    40
2    68
3    24
Name: date_x, dtype: int64
```

In [49]:
```
jds_converted_before.reset_index().groupby('user_id1').count().groupby('index').c
```

Out[49]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fed2e68c8d0>
```



6. What people do before they subscribe to JDS -- for specific articles

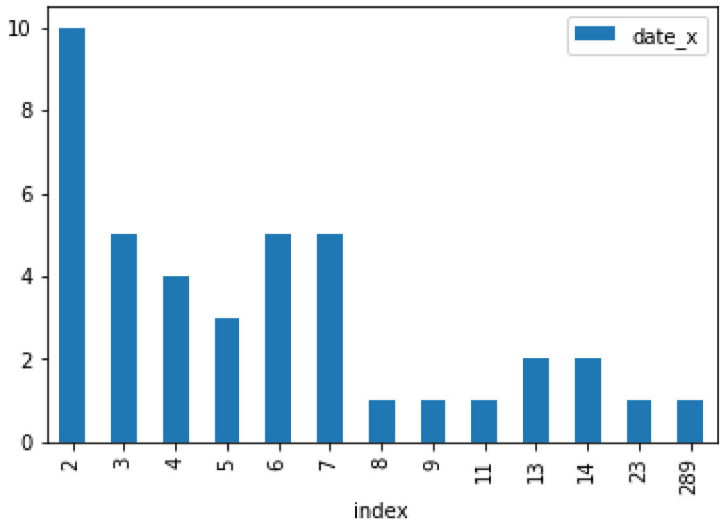
In [53]:
```
spec_article = 'https://data36.com/learning-data-science/'

converted_read_spec_art = jds_converted_before[jds_converted_before.page == spec_
print(spec_article)

converted_read_spec_art.merge(
    jds_converted_before, how = 'inner').reset_index(
    ).groupby('user_id1').count().groupby('index').count()[['date_x']].plot.bar()
```

```
https://data36.com/learning-data-science/ (https://data36.com/learning-data-sci
ence/)
```

Out[53]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fed2e737518>
```



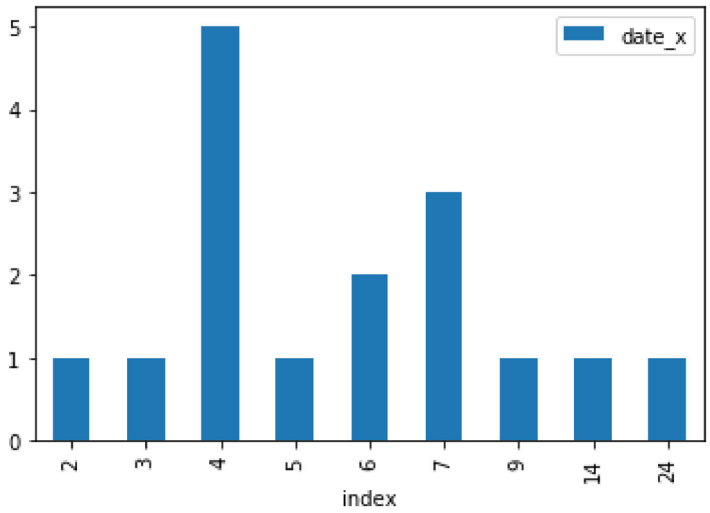
In [55]:
```
spec_article = 'https://data36.com/data-science-career-question-1/'

converted_read_spec_art = jds_converted_before[jds_converted_before.page == spec_
print(spec_article)

converted_read_spec_art.merge(
    jds_converted_before, how = 'inner').reset_index(
    ).groupby('user_id1').count().groupby('index').count()[['date_x']].plot.bar()
```

https://data36.com/data-science-career-question-1/ (https://data36.com/data-science-career-question-1/)

Out[55]: <matplotlib.axes._subplots.AxesSubplot at 0x7fed2e413898>



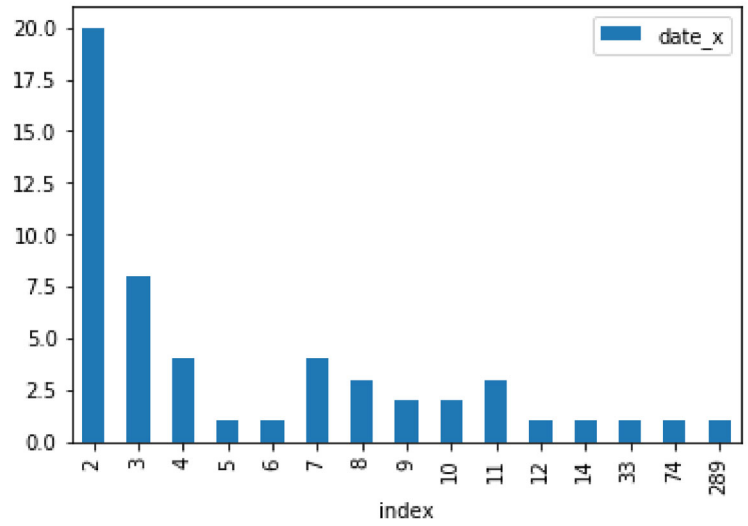
In [54]:
```
spec_article = 'https://data36.com/'

converted_read_spec_art = jds_converted_before[jds_converted_before.page == spec_
print(spec_article)

converted_read_spec_art.merge(
    jds_converted_before, how = 'inner').reset_index(
    ).groupby('user_id1').count().groupby('index').count()[['date_x']].plot.bar()
```

```
https://data36.com/ (https://data36.com/)
```

Out[54]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fed2e67ccf8>
```



In [ ]:
```

```