Homework 5

Load package

In [2]:

```
# packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline

# you may add more if you need
```

Skewed data

In [3]:

```
# load the online user followers data
user_follower = pd.read_csv("./online_user_followers.csv")
print(user_follower.shape)
user_follower.head()
```

```
(215464, 4)
```

Out[3]:

	user_name	user_created	user_followers	user_friends
0	MyNewsNE	24-05-2020 10:18	64.0	11.0
1	Shubham Gupta	14-08-2020 16:42	1.0	17.0
2	Journal of Infectiology	14-12-2017 07:07	143.0	566.0
3	Zane	18-09-2019 11:01	29.0	25.0
4	Ann-Maree O'Connor	24-01-2013 14:53	83.0	497.0

Question: The `user_followers` and `user_friends` have missing values:

At first, try to remove the missing rows for `user_followers` ,

Then, try to fill the missing values with 0 for `user_friends` .

In [4]:

```
user_follower = user_follower.loc[user_follower.user_followers.isnull() == False]
user_follower["user_friends"].fillna(0,inplace = True)
```

Question: Show the min, 25% percentile, median, 75% percentile, max, mean and the stardard deviations of `user_followers` .

In [5]:

```
print("Min:",user_follower["user_followers"].min())
print("25% percentile:",user_follower["user_followers"].quantile(.25))
print("Median",user_follower["user_followers"].median())
print("75% percentile:",user_follower["user_followers"].quantile(.75))
print("Max:",user_follower["user_followers"].max())
print("Mean:",user_follower["user_followers"].mean())
print("Standard Deviation:",user_follower["user_followers"].std())
```

```
Min: 0.0
25% percentile: 121.0
Median 509.0
75% percentile: 2123.0
Max: 16270203.0
Mean: 44859.586354310755
Standard Deviation: 609132.4231951021
```

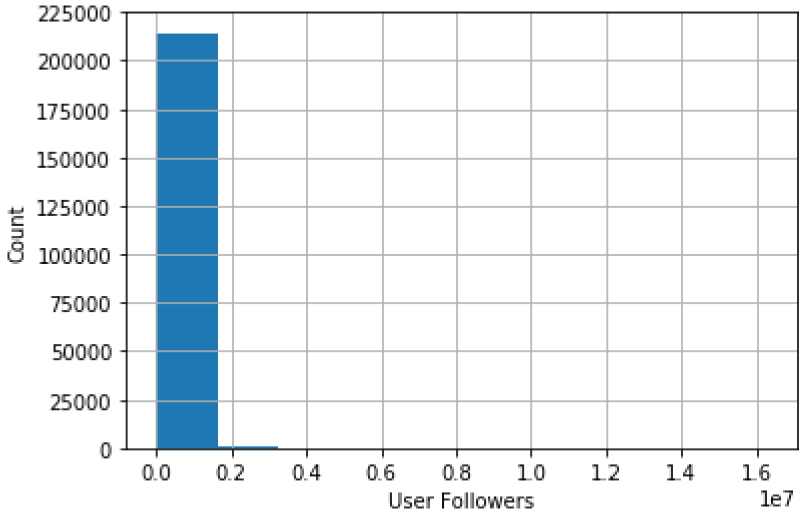
Question: Make a histogram with linear binning for `user_followers` , try to choose a proper number of bins.

In [6]:

```
a = user_follower["user_followers"].hist(bins = 10)
a.set_xlabel("User Followers")
a.set_ylabel("Count")
```

Out[6]:

```
Text(0, 0.5,'Count')
```



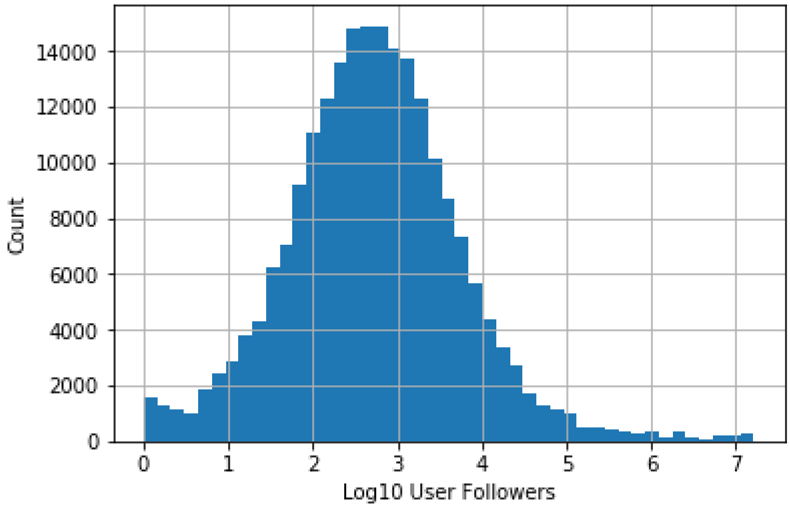
Question: Make a histogram with logarithmic binning for `user_followers` . Because the values has lots of zeros, we need to make a translation by adding 1 for each value, i.e., make the histogram for `user_followers` +1.

In [7]:

```
a = np.log10(user_follower["user_followers"] + 1).hist(bins = 45)
a.set_xlabel("Log10 User Followers")
a.set_ylabel("Count")
```

Out[7]:

```
Text(0,0.5,'Count')
```



Question: Show the PDF, CDF and CCDF for `user_followers` +1 using the logarithmic binning.

In [8]:

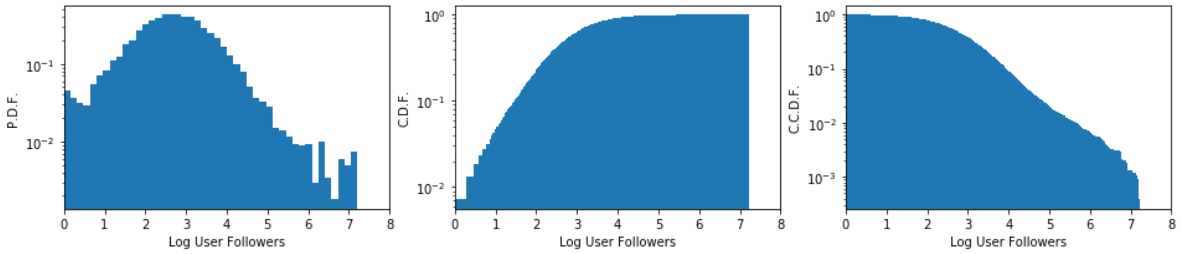
```
fig, axes = plt.subplots(nrows=1, ncols=3, figsize=(16, 3))
ax1 = axes[0]
ax1.hist(np.log10(user_follower["user_followers"] + 1), 45, density=True, cumulative=False,
ax1.set_xlim(0,8)
ax1.set_xlabel("Log User Followers")
ax1.set_ylabel("P.D.F.")

ax2 = axes[1]
ax2.hist(np.log10(user_follower["user_followers"] + 1), 256, density=True, cumulative=True,
ax2.set_xlim(0,8)
ax2.set_xlabel("Log User Followers")
ax2.set_ylabel("C.D.F.")

ax3 = axes[2]
ax3.hist(np.log10(user_follower["user_followers"] + 1), 256, density=True, cumulative=-1, l
ax3.set_xlim(0,8)
ax3.set_xlabel("Log User Followers")
ax3.set_ylabel("C.C.D.F.")
```

Out[8]:

```
Text(0, 0.5, 'C.C.D.F.')
```



Question: For the `user_friends` values, try to convert it to float values.

Hint: the values may contain strings that are not numbers, you need to convert them to `NAN`, then drop the missing values.

In [9]:

```
user_follower["user_friends"] = user_follower["user_friends"].apply(pd.to_numeric, errors='
user_follower = user_follower.loc[~user_follower.user_friends.isna()].copy()

# Check whether the conversion is successful
user_follower.loc[user_follower.user_followers.map(lambda x: type(x) == type(1.0))]
```

Out[9]:

	user_name	user_created	user_followers	user_friends
0	MyNewsNE	24-05-2020 10:18	64.0	11.0
1	Shubham Gupta	14-08-2020 16:42	1.0	17.0
2	Journal of Infectiology	14-12-2017 07:07	143.0	566.0
3	Zane	18-09-2019 11:01	29.0	25.0
4	Ann-Maree O'Connor	24-01-2013 14:53	83.0	497.0
5	Raunak Scherbatsky DankWorth	03-08-2020 13:39	3.0	27.0
6	Rajesh Tadepalli	07-05-2013 03:57	918.0	2561.0
7	AKisASocialisolationist wash yer damn hands	07-02-2015 07:24	2321.0	3236.0
8	Dr. Joseph Santoro	17-01-2009 21:10	19091.0	20986.0
9	VUMC OAP	16-03-2017 20:22	282.0	96.0
10	HrNxt.com	25-03-2011 13:46	87.0	21.0
11	Mohammadali Naseri	02-02-2018 17:20	4.0	88.0
12	LabTwin - Voice & AI-powered digital lab assis...	05-11-2018 21:14	239.0	693.0
13	BioDrivers	30-09-2015 11:32	50.0	609.0
14	Live sport is back!	31-03-2009 08:14	147.0	351.0
15	moneycontrol	26-08-2009 07:55	904607.0	288.0
16	Ravi Prakash Singh	04-06-2019 01:56	12.0	161.0
17	Duvar English	18-10-2019 09:17	17401.0	1.0
18	neonatal2k20	21-12-2017 17:22	591.0	1464.0
19	Kumar Yuvraj	13-04-2020 05:29	19.0	28.0
20	Imran	14-03-2009 09:07	584.0	651.0
21	JPIMedia Design Hub	14-11-2014 13:38	638.0	681.0
22	Philip John Brown	16-11-2017 17:19	2435.0	5001.0
23	See Latest	18-12-2017 03:27	230.0	0.0
24	Oneindia News	16-12-2008 09:44	63491.0	665.0
25	Yash Tiwari Speaks	27-05-2020 12:40	38.0	13.0
26	MK Mania Social News Tv	14-07-2020 14:23	2.0	7.0
27	World School of Bangladesh	10-08-2020 03:43	9.0	61.0
28	Dorjay Namgial	27-06-2020 10:58	17.0	25.0

	user_name	user_created	user_followers	user_friends
29	traceydoesrhymetime	25-02-2020 20:28	1039.0	2117.0
...	...	...	...	...
215434	FactPro	2017-11-12 20:58:17	828.0	675.0
215435	Hi I'm Gabe	2019-11-27 20:50:46	428.0	1191.0
215436	Tasha Sturm	2015-08-21 02:53:51	1479.0	1279.0
215437	Kristin Sommers	2011-01-27 17:41:56	341.0	349.0
215438	Aviation Data Corp - (ASLN.TV)	2014-06-30 02:15:31	584.0	855.0
215439	Steven Chen 陳持威	2018-04-06 14:42:13	11023.0	379.0
215440	Jerry Macdonald (fully vaxxed)	2012-01-27 04:11:34	437.0	263.0
215441	Erik Sandvick	2011-01-16 01:17:59	113.0	163.0
215442	SonofaMitch	2009-08-10 14:46:28	991.0	4143.0
215443	Texan Mama	2009-08-08 04:48:21	225.0	1226.0
215444	Perry Como Chinguss	2016-01-20 21:23:14	387.0	1279.0
215445	United News of India	2015-06-01 14:00:31	3995.0	0.0
215446	syd piper	2009-04-09 04:38:08	207.0	532.0
215447	#Resistance is not futile = Support voting rig...	2009-03-12 15:50:46	758.0	2060.0
215448	I Know 🤔	2010-03-03 05:58:31	523.0	3105.0
215449	Robbie ✊🏻🌹♿💉💉	2010-06-01 06:29:30	168.0	152.0
215450	Antonio Giordano, MD PhD	2010-05-31 15:44:18	1416.0	892.0
215451	Jamie Kay	2018-02-08 20:44:58	43185.0	39002.0
215452	nawialgnehc	2013-03-18 02:51:15	405.0	420.0
215453	Larissa Andrade, MD (She/Her/Hers)	2020-08-27 02:29:50	3028.0	2268.0
215454	13thGenAmerican 👩us	2013-08-31 17:59:52	605.0	6.0
215455	Stephen	2015-02-14 23:30:47	4174.0	4090.0
215456	Dr. Mira Maximos PharmD, MSc, ACPR	2019-04-24 22:05:48	1362.0	695.0
215457	Island Girl - Listen Up, Get Your Covid-Vaccine	2008-02-18 01:52:57	4825.0	3411.0

	user_name	user_created	user_followers	user_friends
215458	Bane's the name	2018-07-03 11:31:26	319.0	225.0
215459	Marcela Ulate	2012-05-05 19:32:50	67.0	196.0
215460	Dr. Alison Obr	2009-01-26 21:43:44	806.0	2360.0
215461	Hillary Hoffmann	2019-05-07 19:30:49	348.0	588.0
215462	JudeME	2018-04-19 00:26:25	1439.0	951.0
215463	hobbitseeker 🦃 Get Vaccinated!	2008-03-15 01:30:45	146.0	371.0

215429 rows × 4 columns

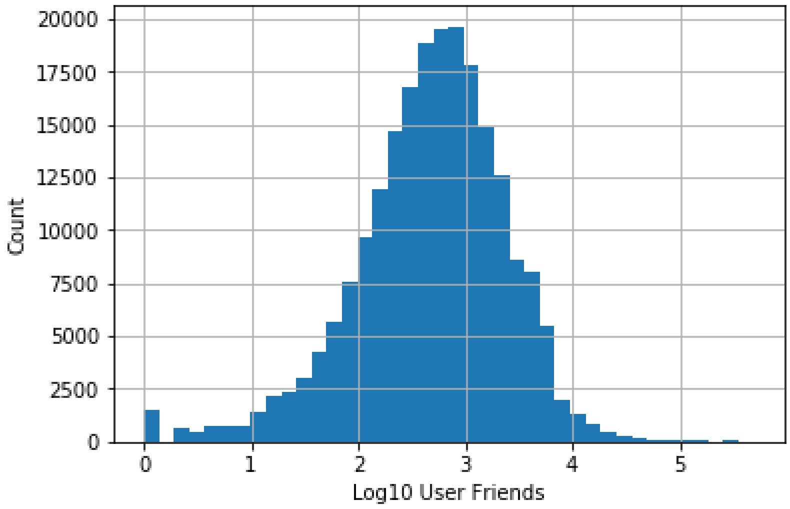
Question: Make a histogram with logarithmic binning for `user_friends` +1.

In [10]:

```
a = np.log10(np.abs(user_follower["user_friends"]) + 1).hist(bins = 40)
a.set_xlabel("Log10 User Friends")
a.set_ylabel("Count")
```

Out[10]:

```
Text(0, 0.5, 'Count')
```



Network data

Data Description:

Character Interaction Networks for the HBO Series "Game of Thrones"

The network data contains the nodes (got-s1-nodes.csv) and edges (got-s1-edges.csv) for the season one series.

Pairs of characters are connected by (undirected) edges weighted by the number of interactions.

There are five interaction types. Character A and Character B are connected whenever:

1. Character A speaks directly after Character B
2. Character A speaks about Character B
3. Character C speaks about Character A and Character B
4. Character A and Character B are mentioned in the same stage direction
5. Character A and Character B appear in a scene together

Question: Create the `Adjacency List` for this network using the edge data: `got-s1-edges.csv`. Print the adjacency list for `Ned` and `Cersei`.

In [11]:

```python
# your code here
edge = pd.read_csv('got-s1-edges.csv')
node = pd.read_csv('got-s1-nodes.csv')

df = np.array(edge)
adjacencylist = {}

for source,target,weight,season in df:
    if(source not in adjacencylist):
        adjacencylist[source] = set()
    if(target not in adjacencylist):
        adjacencylist[target] = set()
    adjacencylist[source].add(target)
    adjacencylist[target].add(source)

print(adjacencylist["NED"])
print(adjacencylist["CERSEI"])
```

```
{'ROOSE_BOLTON', 'VARYS', 'CATELYN', 'BRAN', 'RODRIK', 'BENJEN', 'MYCAH', 'LYANNA',
'MOUNTAIN', 'TYWIN', 'PYP', 'WILL', 'LANCEL', 'AERYS', 'TYRION', 'BRANDON_STARK', 'V
ARLY', 'BARRISTAN', 'SANSA', 'LITTLEFINGER', 'BERIC', 'JANOS', 'GENDRY', 'JAIME', 'R
ENLY', 'ROS', 'CERSEI', 'BAELOR', 'MHAEGEN', 'RICKARD_STARK', 'ILYN_PAYNE', 'DAENERY
S', 'THEON', 'ROBERT', 'TOMARD', 'SYRIO_FOREL', 'LORAS', 'GREATJON_UMBER', 'STANNI
S', 'PYCELLE', 'JOFFREY', 'JON', 'ROBB', 'TOBHO_MOTT', 'JON_ARRYN', 'JEOR', 'JORY_CA
SSEL', 'ARYA', 'HOUND', 'JORAH', 'HIGH_SEPTON', 'SEPTA_MORDANE', 'MAESTER_LUWIN', 'M
ERYN_TRANT', 'HUGH_OF_THE_VALE', 'YOREN', 'STEFFON'}
{'CATELYN', 'VARYS', 'BRAN', 'BENJEN', 'LYANNA', 'TYWIN', 'LANCEL', 'TYRION', 'BARRI
STAN', 'SANSA', 'LITTLEFINGER', 'JAIME', 'RENLY', 'ROS', 'BAELOR', 'ILYN_PAYNE', 'RO
BERT', 'STANNIS', 'PYCELLE', 'JOFFREY', 'JON', 'ROBB', 'JON_ARRYN', 'JEOR', 'HOUND',
'ARYA', 'MERYN_TRANT', 'YOREN', 'NED'}
```

Question: For each character (node) i, calculate the number of nodes connected to it, denote as k_i. Then make a histgram of the distribution of k_i for all nodes.

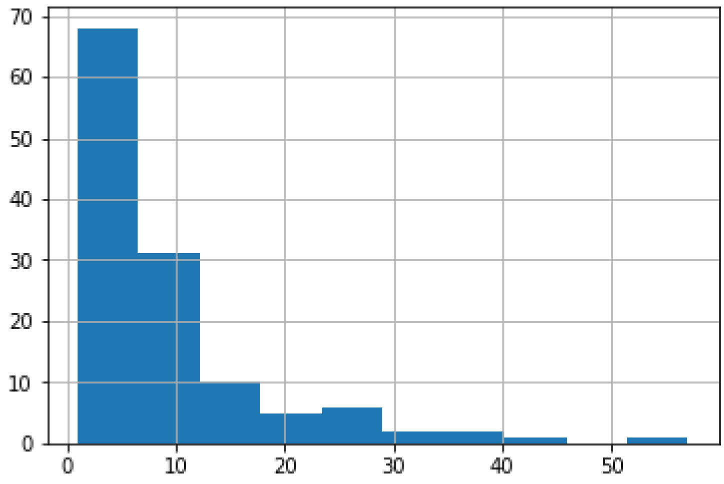
In [17]:

```
# your code here
k = []
for i in adjacencylist:
    k.append(len(adjacencylist[i]))

pd.Series(k).hist()
```

Out[17]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x167e33d00f0>
```



The End