In [1]:

```
import pandas as pd
import numpy as np
import scipy.stats as sps
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt
%matplotlib inline
```

In [2]:

```
%%time
fa_dir = '/Users/stevecoggeshall/Documents/Teaching/Fraud Analytics/2018 USC fraud
mydata = pd.read_csv(fa_dir + '/data/product applications/applications.csv')
```

```
CPU times: user 313 ms, sys: 63 ms, total: 376 ms
Wall time: 391 ms
```

In [3]:

```
mydata.dtypes
```

Out[3]:

```
record        int64
date         object
ssn           int64
firstname    object
lastname     object
address      object
zip5          int64
dob          object
homephone     int64
fraud         int64
dtype: object
```

In [4]:

```
mydata.head(10)
```

Out[4]:

	record	date	ssn	firstname	lastname	address	zip5	dob	hom
0	1	1/1/16	509998359	XRAAAXUAM	SMTAAXRS	4168 XEMMZ PL 19304	19304	11/3/30	6387
						8409			

1	2	1/1/16	615509747	SSXTUJSJM	UTUREERX	ASUZ ST 03563	3563	4/10/21	1069
2	3	1/1/16	532801671	SZMMUJEZS	EZJEAZ	9782 UMSME LN 42178	42178	9/11/13	8719
3	4	1/1/16	302334738	EAZSRMZXZ	SMSMJMMT	2687 XRXAX DR 34631	34631	6/26/07	6314
4	5	1/1/16	737610282	SMRAUMMMZ	MEAXJUX	4775 ETRXZ BLVD 88175	88175	6/26/07	9105
5	6	1/1/16	915986896	SUXEEAZJX	SZEJSXZU	2713 UJZJ ST 09310	9310	5/16/23	9177
6	7	1/1/16	896738279	XSJZEXRZJ	TATMSSJ	8261 TSSJ CT 83503	83503	11/19/72	6497
7	8	1/1/16	601993774	XJZAUEZTX	USSMTRX	3535 RMSJU RD 95839	95839	10/17/95	4809
8	9	1/1/16	131340674	TZERZRXZ	USZMSMEZ	3307 SUZXR ST 04362	4362	3/14/15	3501
9	10	1/1/16	888484341	EAXRRUMUX	RAUZRMEA	508 UMJXM BLVD 67490	67490	6/28/86	1557

Summary statistics

```
In [5]:
```

```
mydata.shape
```

```
Out[5]:

(94866, 10)
```

```
In [6]:
```

```
mydata.describe(include = 'all')
```

```
Out[6]:
```

	record	date	ssn	firstname	lastname	address	zip5
count	94866.000000	94866	9.486600e+04	94866	94866	94866	94866.0000
unique	NaN	365	NaN	14626	31513	88167	NaN
top	NaN	6/9/16	NaN	EASEXMJAT	ERJSAXA	8911 MZSU DR 43516	NaN
freq	NaN	329	NaN	1414	1515	57	NaN
mean	47433.500000	NaN	5.039438e+08	NaN	NaN	NaN	49848.4566
std	27385.599656	NaN	2.879555e+08	NaN	NaN	NaN	28889.4208
min	1.000000	NaN	3.600000e+01	NaN	NaN	NaN	2.000000
25%	23717.250000	NaN	2.532461e+08	NaN	NaN	NaN	24782.0000
50%	47433.500000	NaN	5.102548e+08	NaN	NaN	NaN	50190.5000
75%	71149.750000	NaN	7.469134e+08	NaN	NaN	NaN	74192.0000
max	94866.000000	NaN	9.999946e+08	NaN	NaN	NaN	99999.0000

In [7]:

```
mydata.count()
```

Out[7]:

```
record      94866
date         94866
ssn          94866
firstname    94866
lastname     94866
address      94866
zip5         94866
dob          94866
homephone    94866
fraud        94866
dtype: int64
```

Field by field statistics

In [8]:

```
# len(mydata['record'].unique())
```

In [9]:

```
mydata['date'].value_counts()
```

Out[9]:

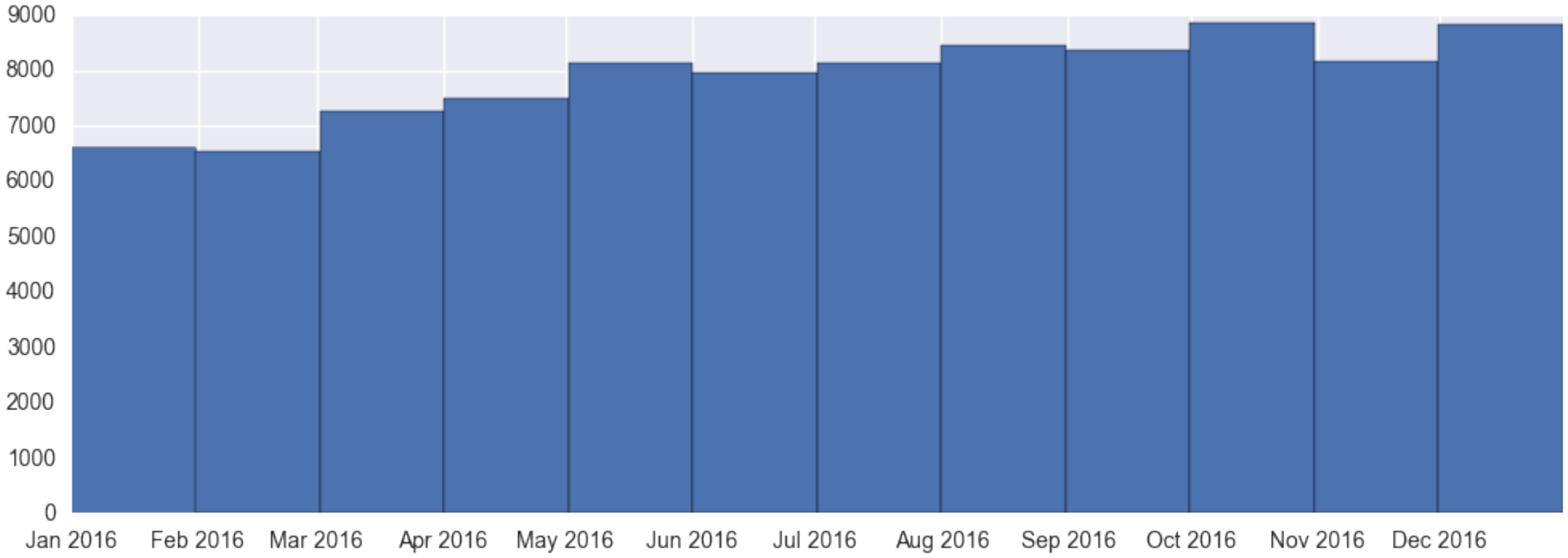
```
6/9/16      329
12/29/16    328
11/19/16    325
9/18/16     324
10/18/16    324
10/2/16     320
12/10/16    320
12/8/16     320
10/7/16     320
12/30/16    319
8/27/16     315
12/31/16    307
9/25/16     306
10/21/16    305
9/15/16     305
9/20/16     304
8/18/16     303
10/12/16    303
```

In [10]:

```
mydata['date'] = pd.to_datetime(mydata['date'])
```

In [11]:

```
fig=plt.figure(figsize = (12,4))
fig = mydata['date'].hist(bins=12)
```



In [12]:

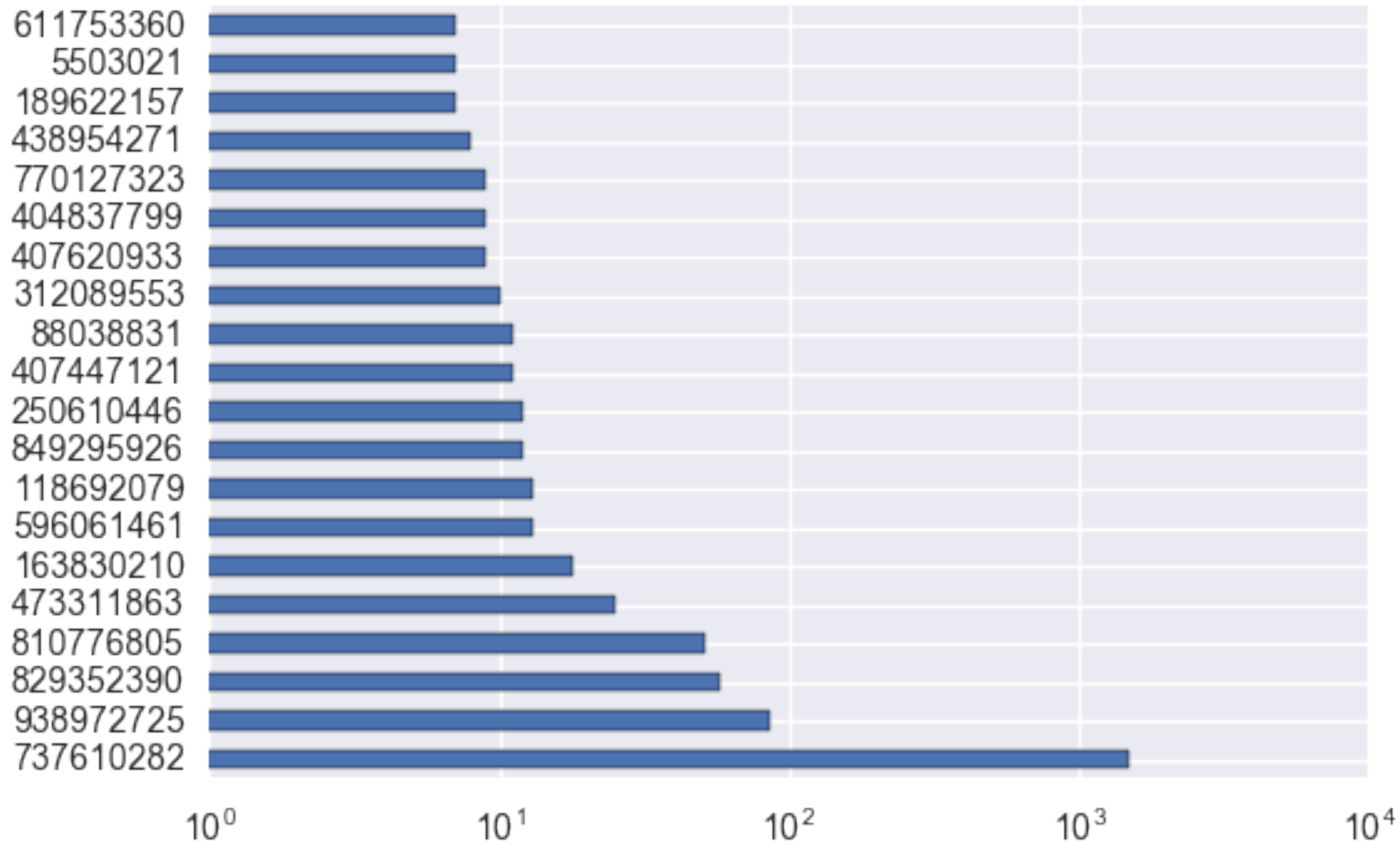
```
len(mydata['ssn'].unique())
```

Out[12]:

```
86771
```

In [13]:

```
mydata['ssn'].value_counts().head(20).plot(kind = 'barh')
plt.xscale('log')
```



In [14]:

```
len(mydata['firstname'].unique())
```

Out[14]:

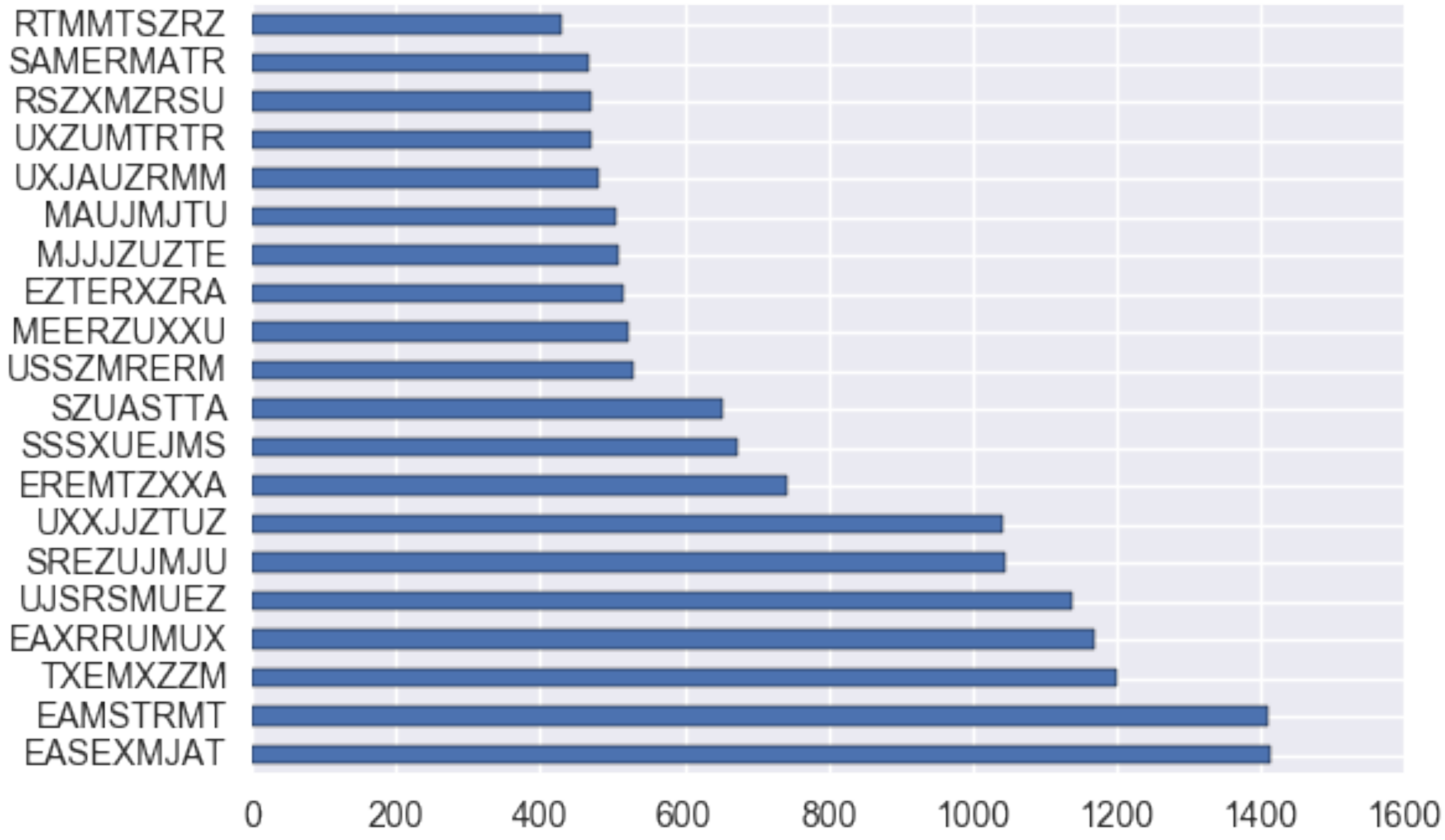
```
14626
```

In [15]:

```
mydata['firstname'].value_counts().head(20).plot(kind = 'barh')
```

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x118e32358>
```



In [16]:

```
len(mydata['lastname'].unique())
```

Out[16]:

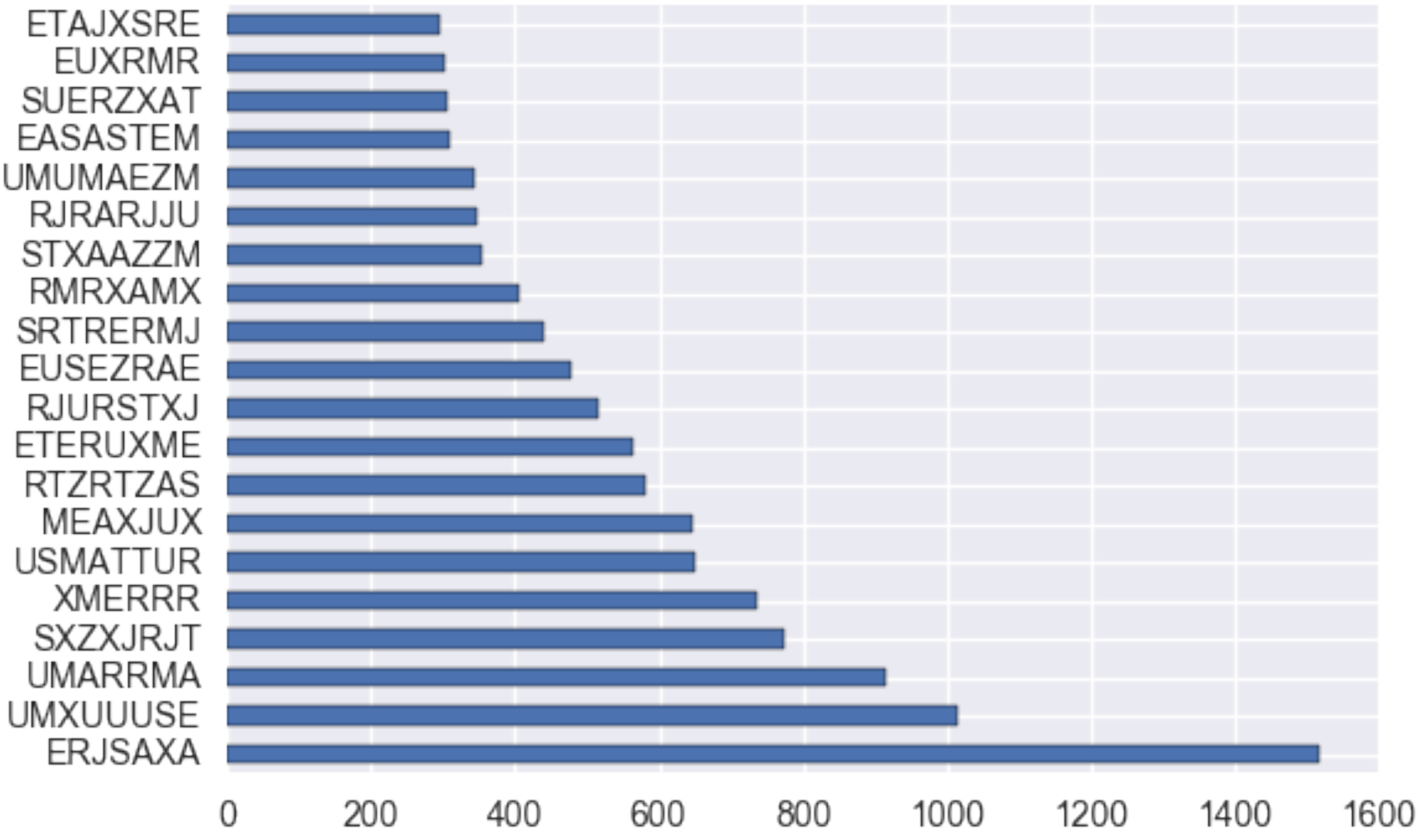
```
31513
```

In [17]:

```
mydata['lastname'].value_counts().head(20).plot(kind = 'barh')
```

Out[17]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11b43dd68>
```



In [18]:

```
mydata['lastname'].value_counts()
```

Out[18]:

```
ERJSAXA     1515
UMXUUUSE    1013
UMARRMA      913
SXZXJRJT     775
XMERRR       737
USMATTUR     649
MEAXJUX      645
RTZRTZAS     582
ETERUXME     562
RJURSTXJ     515
EUSEZRAE     476
SRTRERMJ     438
RMRXAMX      405
STXAAZZM     352
RJRARJJU     348
UMUMAEZM     342
EASASTEM     310
```

```
SUERZXAT     306
EUXRMR       302
ETAJXSRE     295
UXJEXUJR     271
ARUZTZM      270
SMTTZJJX     267
SJURETUX     267
MZRUMMJ      266
STZRUXZM     252
SRRTAZTX     250
ERXSZZMA     230
RMXAUUA      219
EMRSJTXE     217
            ...
ETJMTMUS       1
ESURSUZZ       1
UURUTJTR       1
SRMUUXSJ       1
TMAXJTT        1
ETETJEUT       1
ERZASZU        1
SRJURERJ       1
UZJTSRMZ       1
ZTZEMAA        1
SAREEJAM       1
SZZRTAUE       1
RZSERJMJ       1
ETMZUUTX       1
EESZXJMU       1
TTXZXZZ        1
EUUERMSU       1
EJAJURZA       1
UARZSETZ       1
SUMRMMZS       1
UATMRRJ        1
SXMUXSSE       1
RRREARXZ       1
RTESXRXX       1
EJREZXUE       1
EAAZRJAJ       1
RAXSZTZ        1
MUXERMR        1
RUEJATXU       1
EUZJJEME       1
Name: lastname, dtype: int64
```

In [19]:

```
len(mydata['address'].unique())
```

Out[19]:

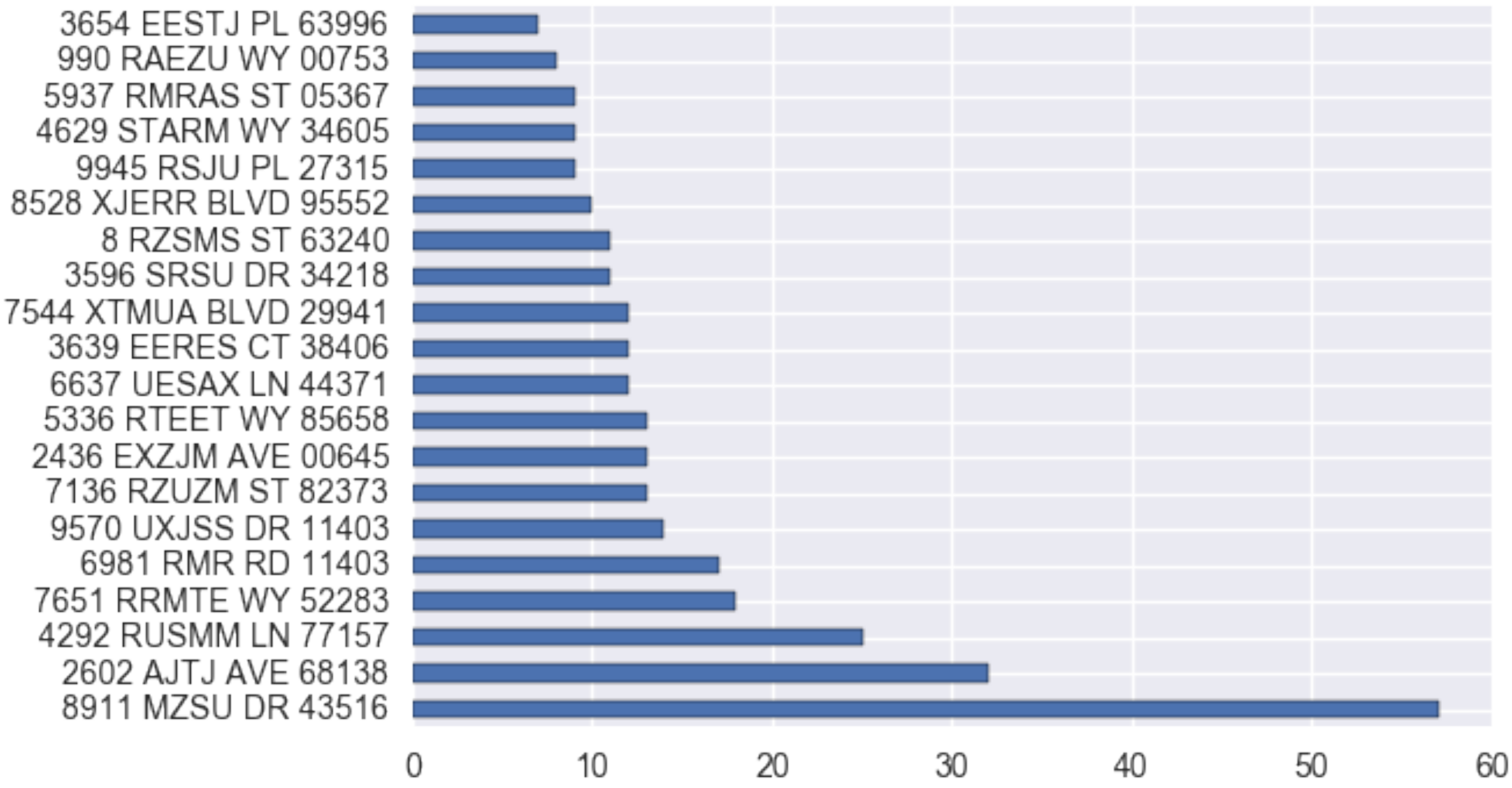
```
88167
```

In [20]:

```
mydata['address'].value_counts().head(20).plot(kind = 'barh')
```

Out[20]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x118e16b38>
```



In [21]:

```
mydata['address'].value_counts()
```

Out[21]:

```
8911 MZSU DR 43516       57
2602 AJTJ AVE 68138      32
4292 RUSMM LN 77157      25
7651 RRMTE WY 52283      18
6981 RMR RD 11403        17
9570 UXJSS DR 11403      14
7136 RZUZM ST 82373      13
2436 EXZJM AVE 00645     13
5336 RTEET WY 85658      13
6637 UESAX LN 44371      12
3639 EERES CT 38406      12
```

```
7544 XTMUA BLVD 29941    12
3596 SRSU DR 34218       11
8 RZSMS ST 63240         11
8528 XJERR BLVD 95552    10
9945 RSJU PL 27315        9
4629 STARM WY 34605       9
5937 RMRAS ST 05367       9
990 RAEZU WY 00753        8
3654 EESTJ PL 63996       7
9638 UTMZS ST 20059       7
8154 AMTX ST 59902        7
1735 UTXEZ ST 78226       7
1149 SJZSR BLVD 30425     7
9074 UUJJE ST 25894       6
2581 RREJ BLVD 83426      6
4467 SSZSM PL 28822       6
1690 STAMJ WY 76516       6
8730 RTMMX CT 01365       6
5267 EXAZZ BLVD 65709     6
                         ..
4004 XZEUM WY 97080       1
4126 EXZEE ST 99056       1
4400 ESRXM ST 58251       1
4920 RMZEU BLVD 01718     1
698 UAEUU ST 25691        1
7452 XARU ST 84699        1
3836 EAUJR ST 00298       1
6771 XJJUT DR 95981       1
1499 SSATA PL 06076       1
8341 MAJU ST 62007        1
7324 EZRRZ RD 33633       1
5200 RAMJR BLVD 92171     1
3338 RUUUE AVE 90170      1
4763 XZAUJ BLVD 40507     1
7654 XZJET AVE 77876      1
4405 ESETZ DR 97224       1
2097 MTZR RD 71565        1
1979 ARJJ ST 74238        1
9784 XURUS LN 73578       1
9640 RRAA ST 30116        1
6112 SAZMU LN 88891       1
7407 SJZTE LN 36951       1
5842 XZJMX ST 37900       1
4022 ZTZA DR 57807        1
7433 RAEZA ST 01151       1
771 XREXX ST 05196        1
617 RRAU RD 88751         1
168 ESSMR AVE 86997       1
9358 EUREE PL 04685       1
9427 ZSSE RD 83264        1
Name: address, dtype: int64
```

In [22]:

```
len(mydata['zip5'].unique())
```

Out[22]:

```
15855
```

In [23]:

```
mydata['zip5'].value_counts()
```

Out[23]:

```
43516    64
1362     53
80692    51
84983    49
14931    47
94992    46
86500    46
10664    45
47208    44
89835    44
66474    44
34031    44
59066    43
90042    43
33768    43
13440    43
57682    43
52317    42
12700    42
1097     42
27132    42
73686    42
66902    41
56155    41
35227    41
23582    41
72192    41
53182    40
30136    40
49129    40
         ..
65849     1
345       1
12639     1
28410     1
59115     1
26644     1
3993      1
97655     1
```

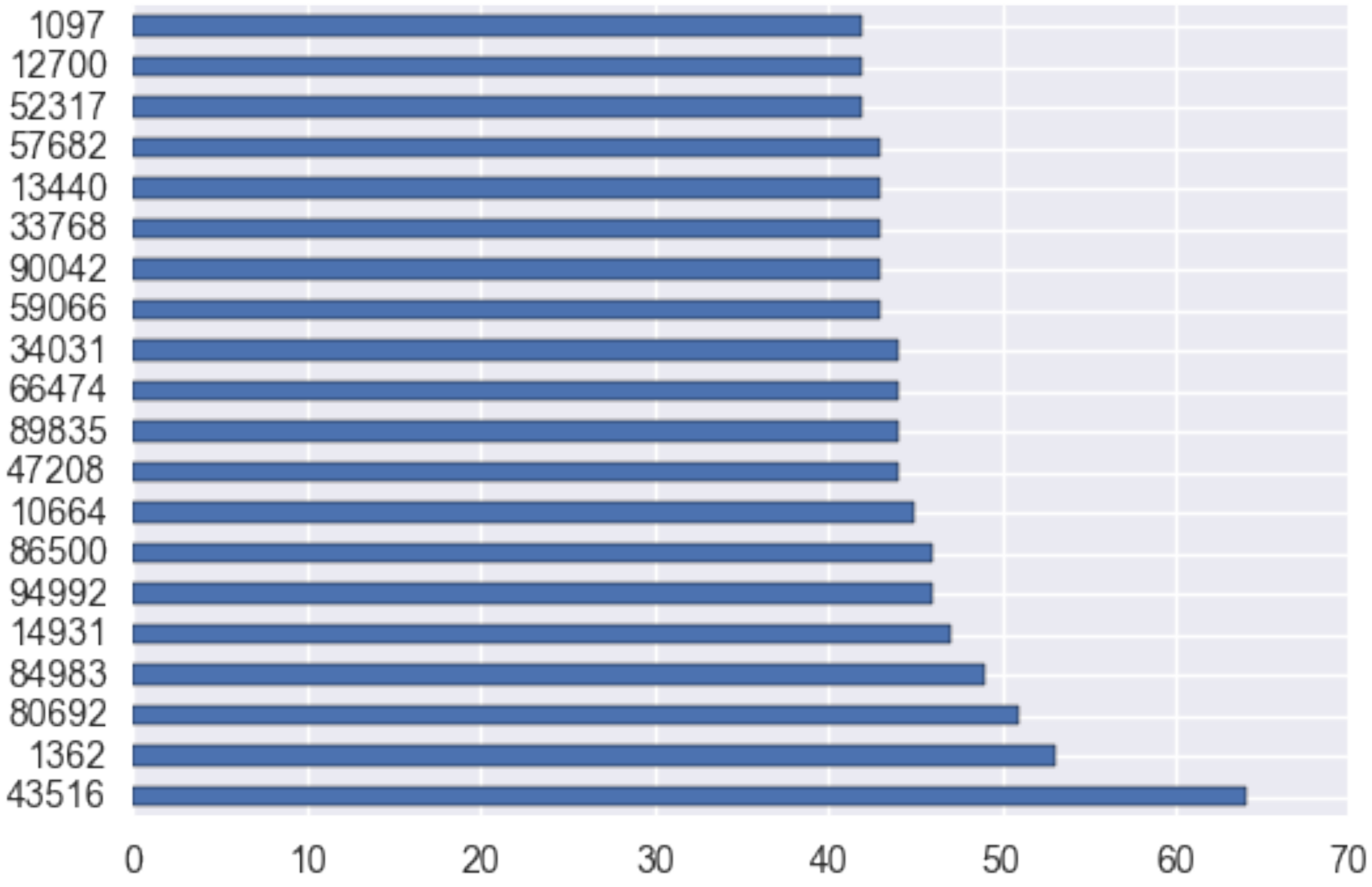
```
38474     1
75132     1
11677     1
97719     1
30166     1
34312     1
28181     1
85553     1
34376     1
36425     1
87666     1
12146     1
75388     1
77437     1
81535     1
98264     1
6069      1
55042     1
12125     1
75667     1
81791     1
23379     1
Name: zip5, dtype: int64
```

In [24]:

```
mydata['zip5'].value_counts().head(20).plot(kind = 'barh')
```

Out[24]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11aadfa20>
```



In [25]:

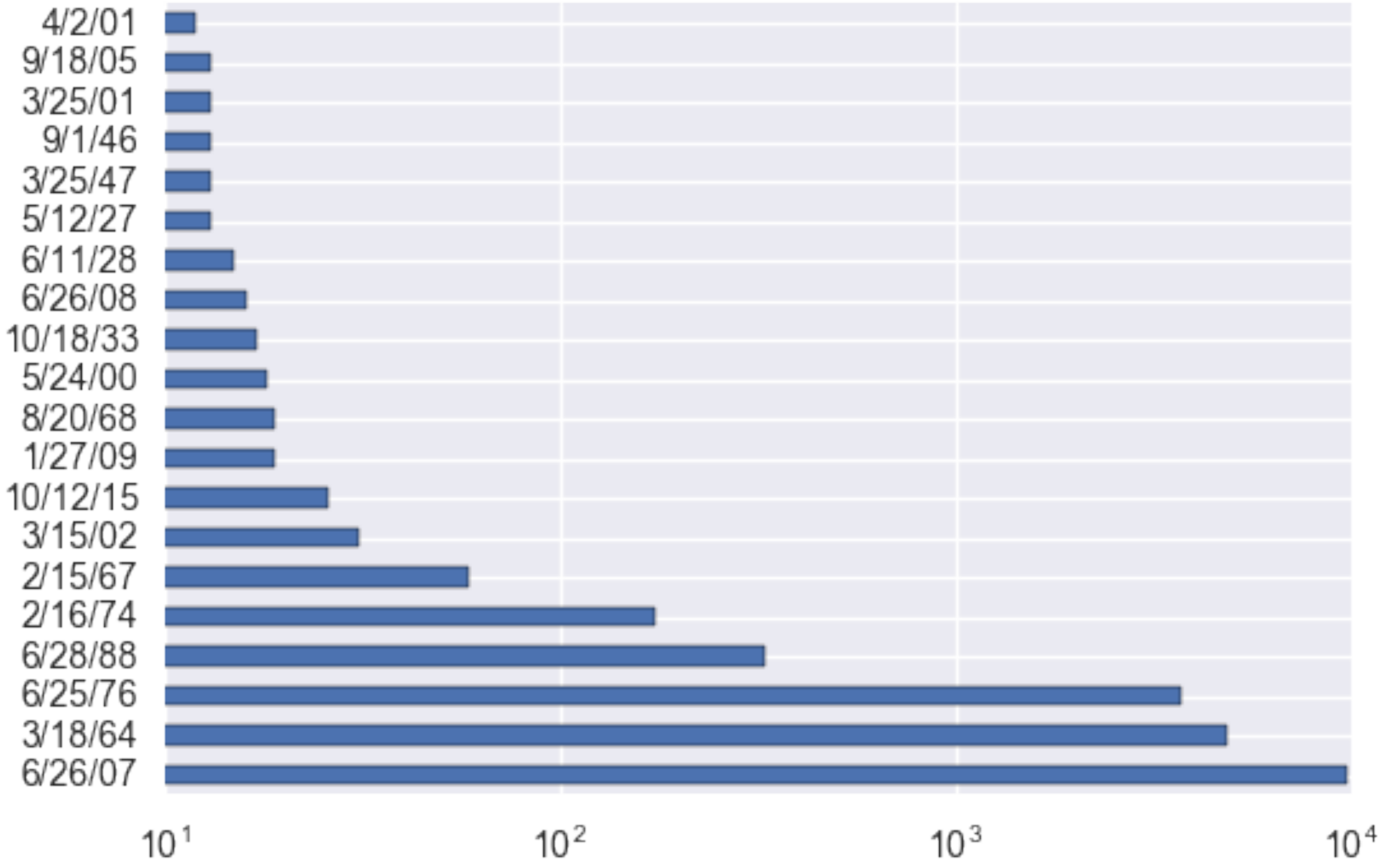
```
len(mydata['dob'].unique())
```

Out[25]:

```
30599
```

In [26]:

```
mydata['dob'].value_counts().head(20).plot(kind = 'barh')
plt.xscale('log')
```



In [27]:

```
mydata['dob'].value_counts()
```

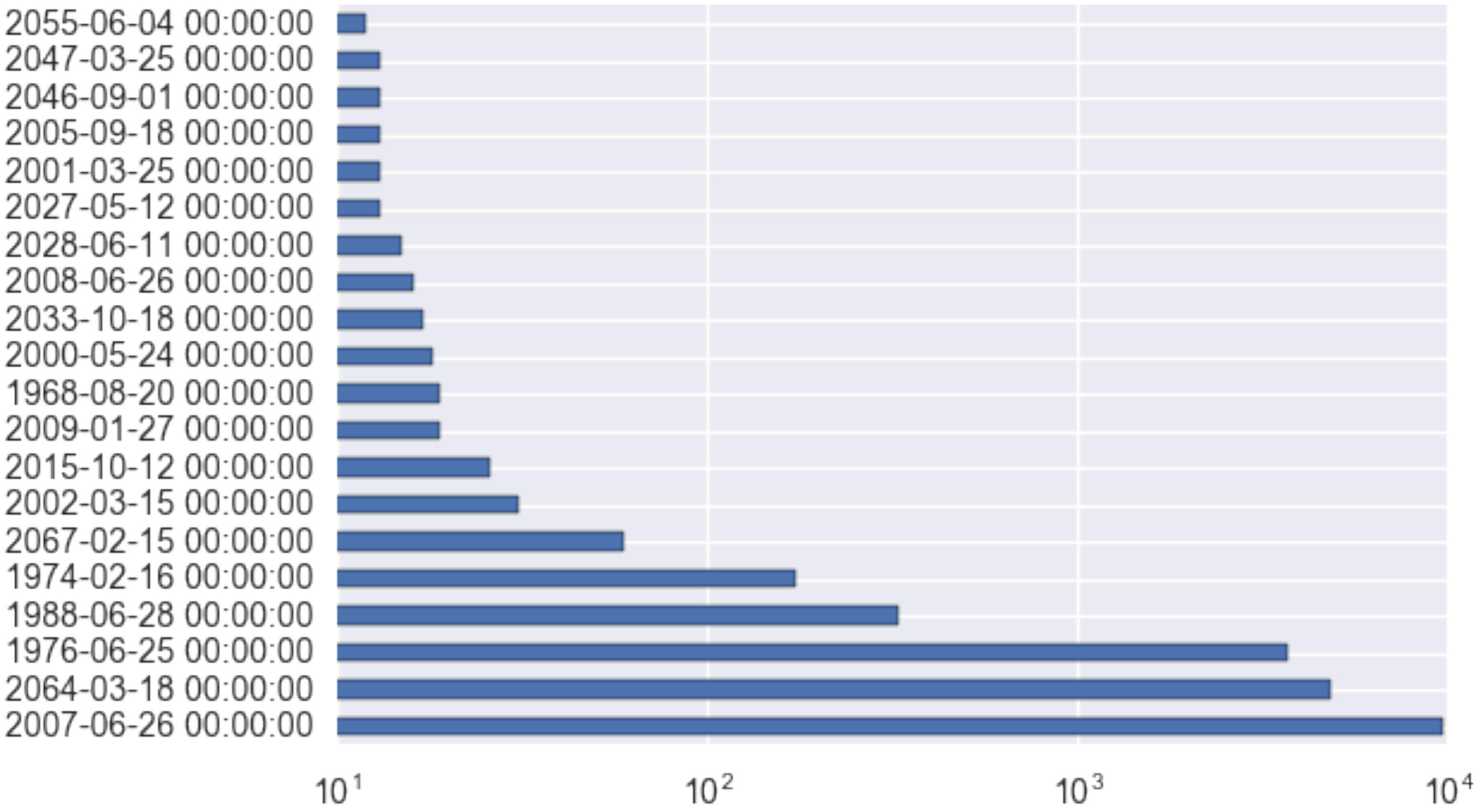
Out[27]:

```
6/26/07     9681
3/18/64     4808
6/25/76     3698
6/28/88      330
2/16/74      173
2/15/67       59
3/15/02       31
10/12/15      26
1/27/09       19
8/20/68       19
5/24/00       18
10/18/33      17
6/26/08       16
6/11/28       15
5/12/27       13
3/25/47       13
9/1/46        13
3/25/01       13
9/18/05       13
4/2/01        12
```

```
6/18/06       12
2/10/25       12
6/4/55        12
3/14/00       12
7/28/00       11
8/20/59       11
5/6/03        11
9/9/34        11
4/13/91       11
6/26/16       11
            ...
10/8/63        1
7/15/44        1
3/29/42        1
6/24/26        1
4/29/27        1
1/22/28        1
3/31/44        1
4/24/21        1
12/1/90        1
7/29/44        1
6/3/89         1
8/18/53        1
12/11/75       1
4/2/55         1
6/9/74         1
8/10/00        1
8/27/70        1
9/1/20         1
11/16/34       1
8/30/70        1
3/1/58         1
3/23/19        1
9/16/19        1
5/7/87         1
4/29/40        1
5/28/54        1
7/6/53         1
8/8/39         1
7/1/33         1
7/28/27        1
Name: dob, dtype: int64
```

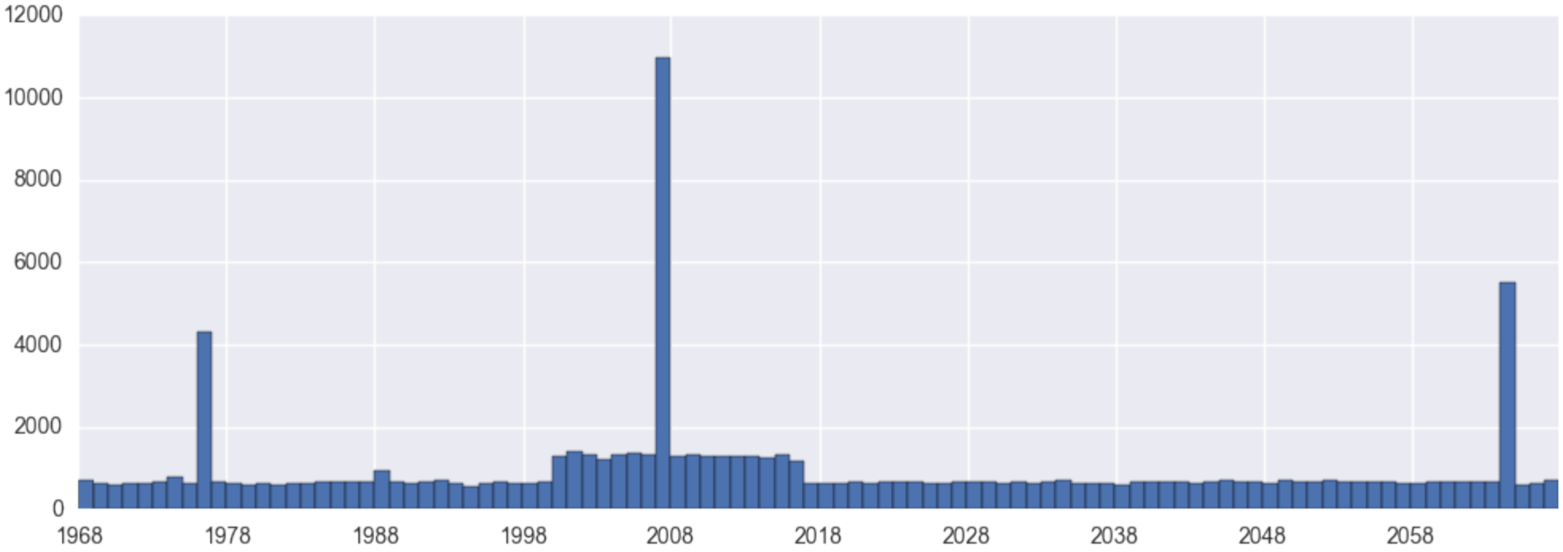
In [28]:

```
mydata['dob'] = pd.to_datetime(mydata['dob'])
mydata['dob'].value_counts().head(20).plot(kind = 'barh')
plt.xscale('log')
```



In [29]:

```
fig=plt.figure(figsize = (12,4))
fig = mydata['dob'].hist(bins=100)
```



In [30]:

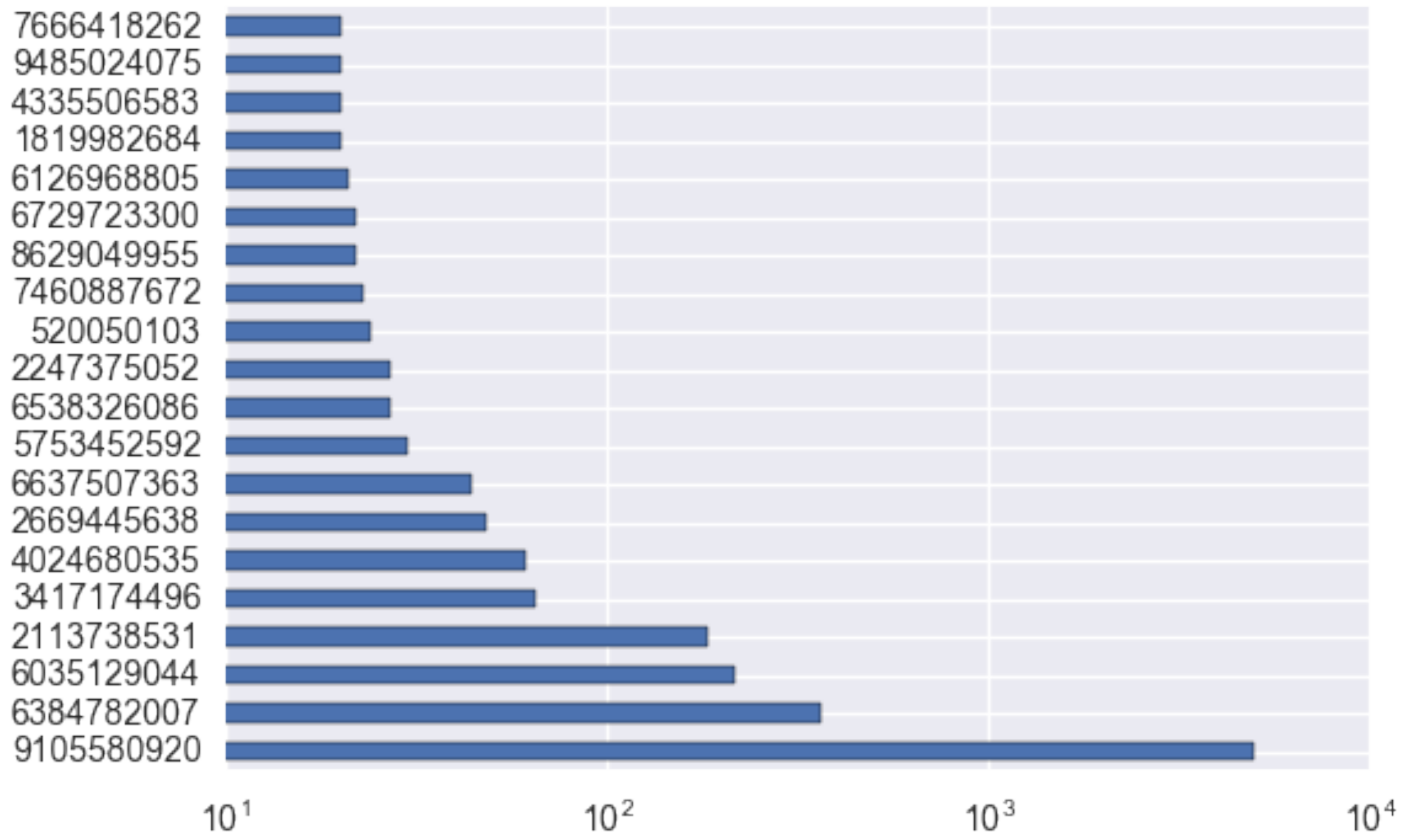
```
len(mydata['homephone'].unique())
```

Out[30]:

```
20762
```

In [31]:

```
mydata['homephone'].value_counts().head(20).plot(kind = 'barh')
plt.xscale('log')
```



In [32]:

```
mydata['homephone'].value_counts()
```

Out[32]:

```
9105580920    4974
6384782007     364
6035129044     215
2113738531     184
3417174496      65
4024680535      61
2669445638      48
6637507363      44
5753452592      30
6538326086      27
2247375052      27
520050103       24
```

```
7460887672     23
8629049955     22
6729723300     22
6126968805     21
1819982684     20
4335506583     20
9485024075     20
7666418262     20
1584890200     19
8940354172     19
7802891638     19
8880326532     19
8803722913     19
1648678851     19
8678041990     19
9537440042     18
1907432097     18
8293886748     18
               ..
7880961013      1
7625970404      1
2342158500      1
2018277563      1
2875087939      1
3377611840      1
9392725051      1
6978713460      1
1737980295      1
8455904275      1
2781316946      1
5560281297      1
4258466005      1
8565614398      1
9815979434      1
7379071298      1
7556880623      1
5130690301      1
5035796262      1
2233783250      1
908798225       1
6036446671      1
6900134846      1
6050656416      1
2053192623      1
1713366814      1
7924239023      1
9713124248      1
5651886998      1
2019168330      1
Name: homephone, dtype: int64
```

In [33]:

```
mydata['fraud'].value_counts()
```

Out[33]:

```
0    74702
1    20164
Name: fraud, dtype: int64
```

Plot number of transactions per day

In [34]:

```
mydata.head().transpose()
```

Out[34]:

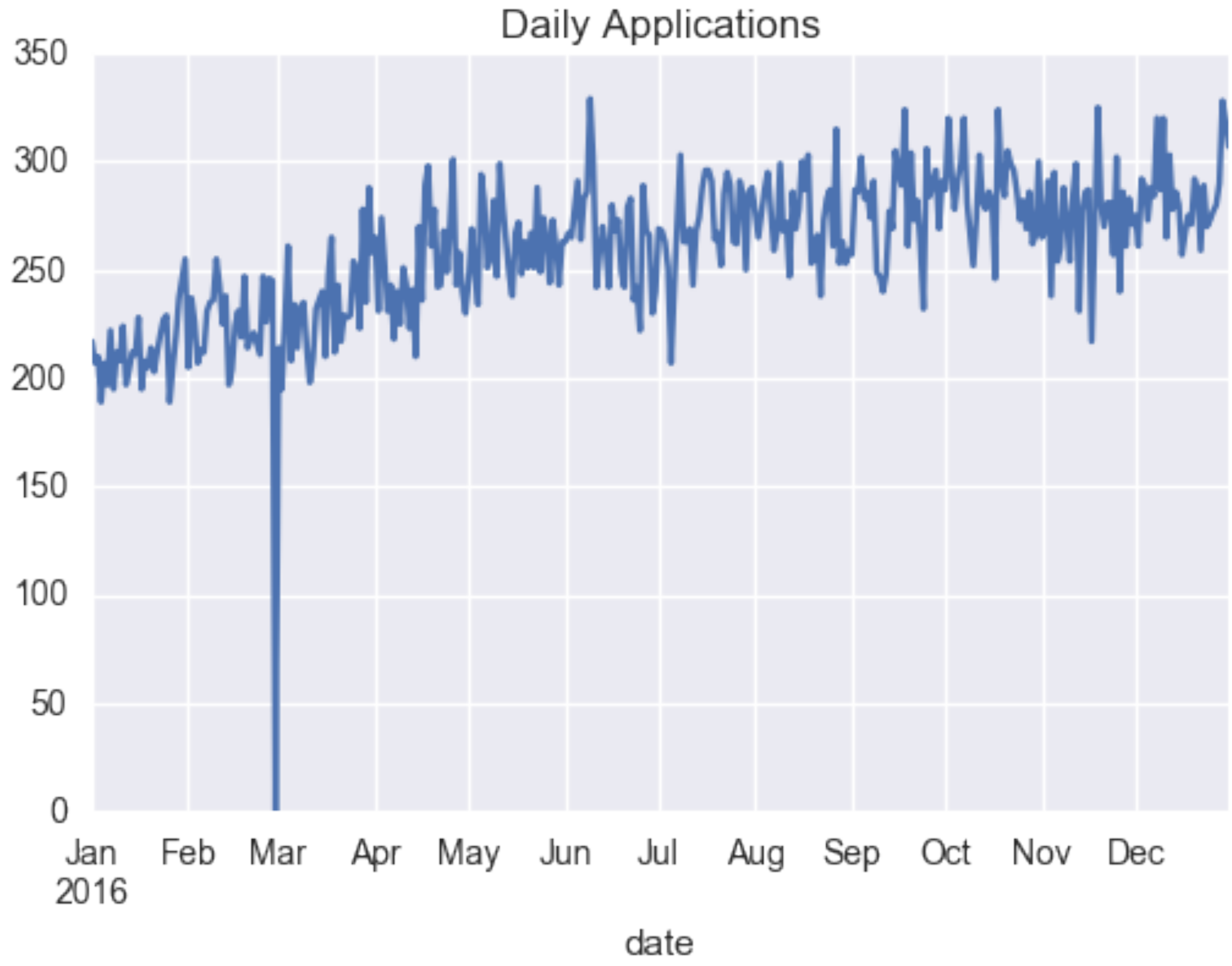
	0	**1**	**2**	**3**	**4**
record	1	2	3	4	5
date	2016-01-01 00:00:00	2016-01-01 00:00:00	2016-01-01 00:00:00	2016-01-01 00:00:00	2016-01-01 00:00:00
ssn	509998359	615509747	532801671	302334738	737610282
firstname	XRAAAXUAM	SSXTUJSJM	SZMMUJEZS	EAZSRMZXZ	SMRAUMMMZ
lastname	SMTAAXRS	UTUREERX	EZJEAZ	SMSMJMMT	MEAXJUX
address	4168 XEMMZ PL 19304	8409 ASUZ ST 03563	9782 UMSME LN 42178	2687 XRXAX DR 34631	4775 ETRXZ BLVD 88175
zip5	19304	3563	42178	34631	88175
dob	2030-11-03 00:00:00	2021-04-10 00:00:00	2013-09-11 00:00:00	2007-06-26 00:00:00	2007-06-26 00:00:00
homephone	6387900398	1069037699	8719510343	6314026324	9105580920
fraud	1	0	1	1	0

In [35]:

```
mydata.assign(trx = np.ones(len(mydata.index)))\
    .set_index(mydata['date'].astype(dt.datetime))\
    .resample(dt.timedelta(days = 1))\
    .count()\
    .trx\
    .plot(title = 'Daily Applications')
```

Out[35]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11abb0f28>
```



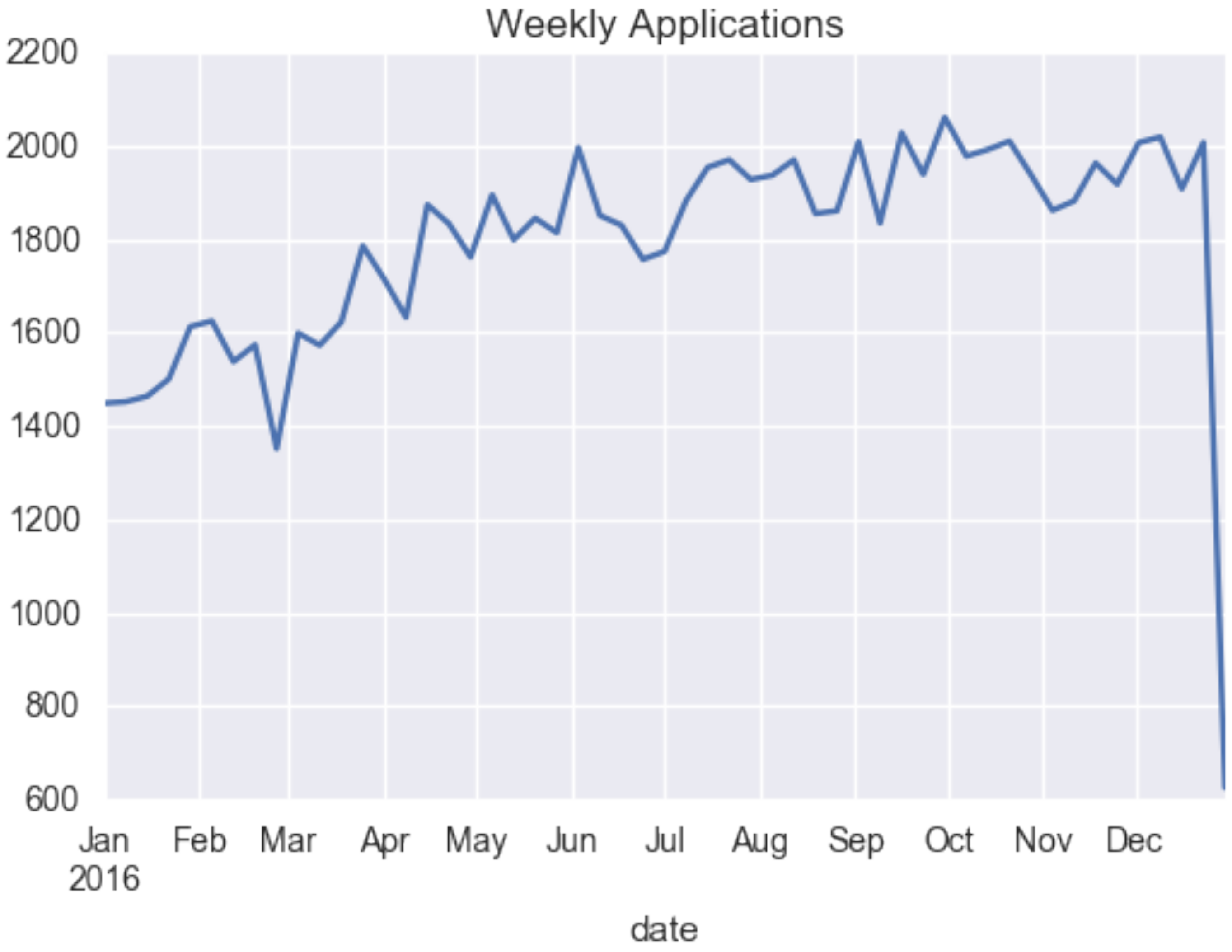
Count # transactions for the next week, 30 days

In [36]:

```
mydata.assign(trx = np.ones(len(mydata.index)))\
    .set_index(mydata['date'].astype(dt.datetime))\
    .resample(dt.timedelta(days = 7))\
    .count()\
    .trx\
    .plot(title = 'Weekly Applications')
```

Out[36]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11ef05588>
```

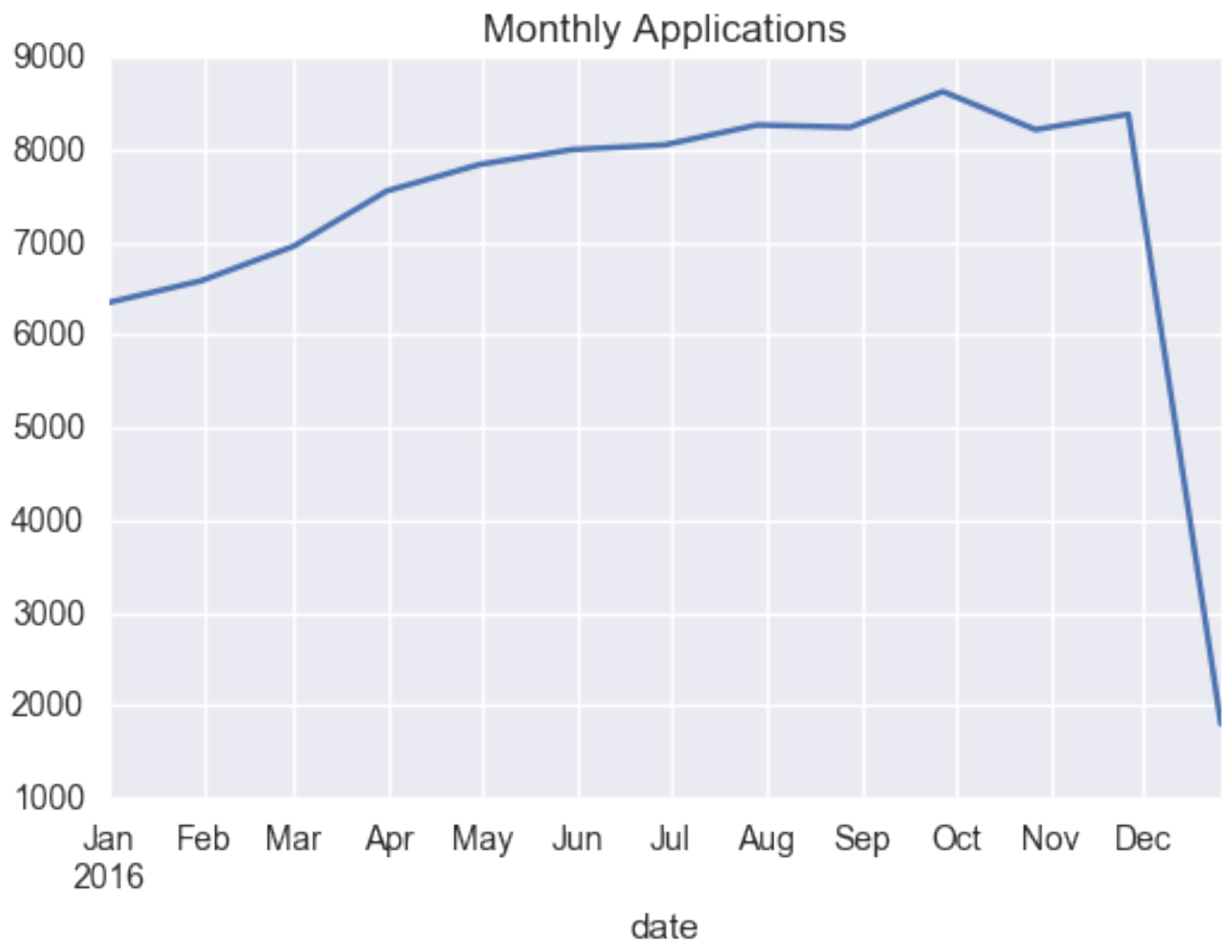


In [37]:

```
mydata.assign(trx = np.ones(len(mydata.index)))\
    .set_index(mydata['date'].astype(dt.datetime))\
    .resample(dt.timedelta(days = 30))\
    .count()\
    .trx\
    .plot(title = 'Monthly Applications')
```

Out[37]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11c5025f8>
```



In [ ]:

```

```