Final Project Data Analysis

Topic: EdTech Startup Landscape

For my final project, I will be diving into data on education technology startup companies. Specifically, I will be looking for statistics on the company such as location, date founded, total funding, number of funding rounds, number of employees, and other, similar data points.

By gathering this data I will be trying to measure the level of a company's success in comparison to others in the EdTech field. For example, I will be able to find the speed of a company's rise to success through the date it was founded to the date it had an IPO, if it did go public. My hypotheses on what I expect to find will be detailed after a thorough review of the data collected.

Dataset: CrunchBase API

Sibling to the popular startup news tech blog 'TechCrunch', CrunchBase has a wide range of data on companies. For each organization listed, it is possible to get data on individual employees, funding rounds, acquisitions, and a variety of other details. As such, one of the reasons I chose EdTech was the relatively small quantity of companies with this tag on CrunchBase, compared to the hundreds of thousands under 'education'.

Due to API rate limit restrictions I contacted CrunchBase and informed them of my idea for the project, and they were happy to accomodate me with a standard, non-commercial lisence, essentially giving me free reign to the API.

Data Analysis

In [39]:
```
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data = pd.read_csv('organizationDF.tsv', sep='\t')
data = data.drop('Unnamed: 0', 1)
data.head(1)
```

Out[39]:

	name	location	description	founded	funding	funding_rounds	ipo	acquisitions	n
0	general-assembly	New York, New York, United States	General Assembly is a global educational insti...	2011-01-01	49500000	4	0	0	0

Funding

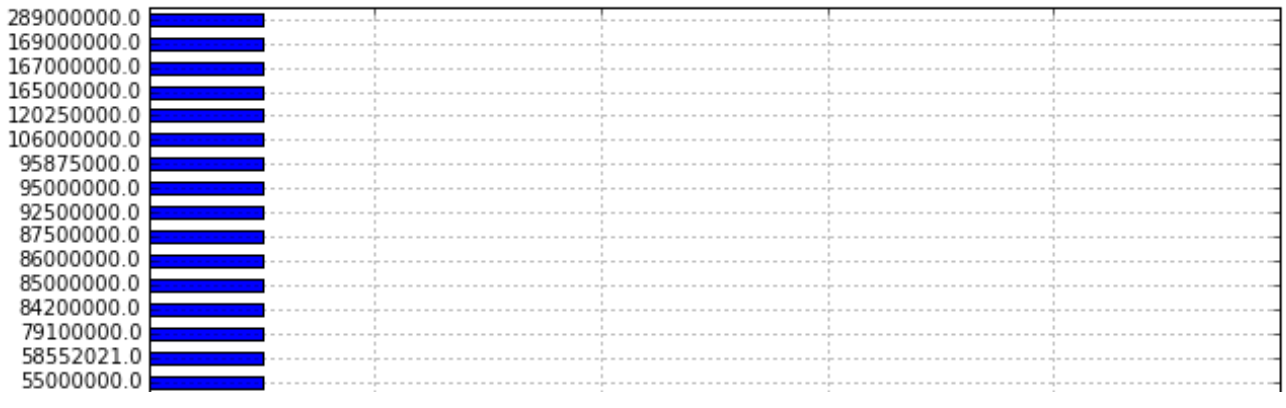
To my dismay, it appears that a key statistic, the quantity in USD that a company received in funding, has many holes. Almost 1000 companies have a count of 0 for funding.

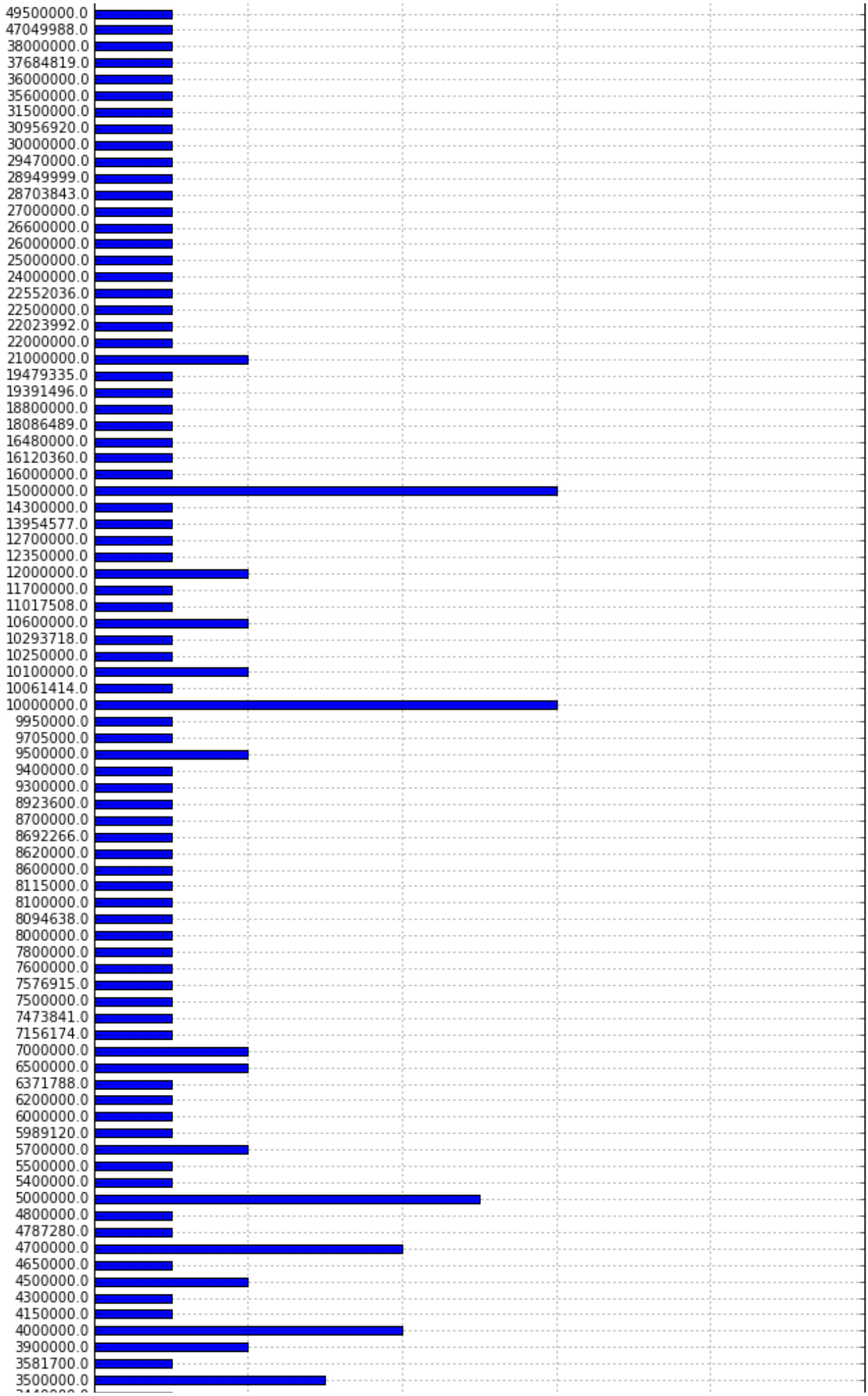
Although I would rather use a tag such as 'education', which would undoubtedly provide better results, with tens of thousands of education companies I wouldn't be able to process the data via the API rate limits. 'EdTech' still provides a focused view of hundreds of companies that received funding, and ultimately gives a more intimate perspective of this market.

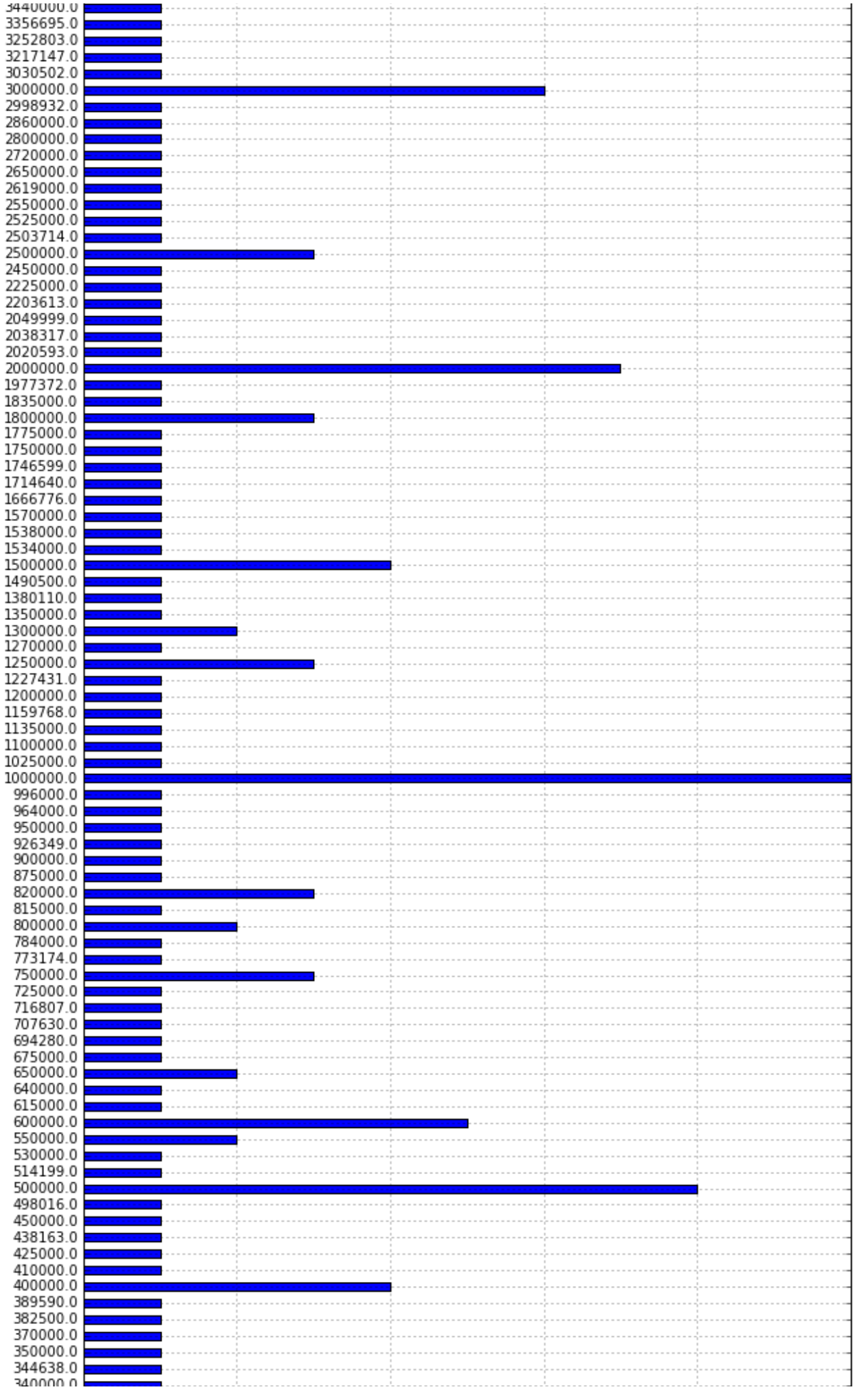
Despite the fact that many companies don't have funding information, I will work around this problem by focusing my analysis upon the companies that do have this information. And in the same method of focusing my approach to the companies with funding information, I will compare this subset of companies on other measures such as date founded, number of employees, number of acquisitions, and ipo.

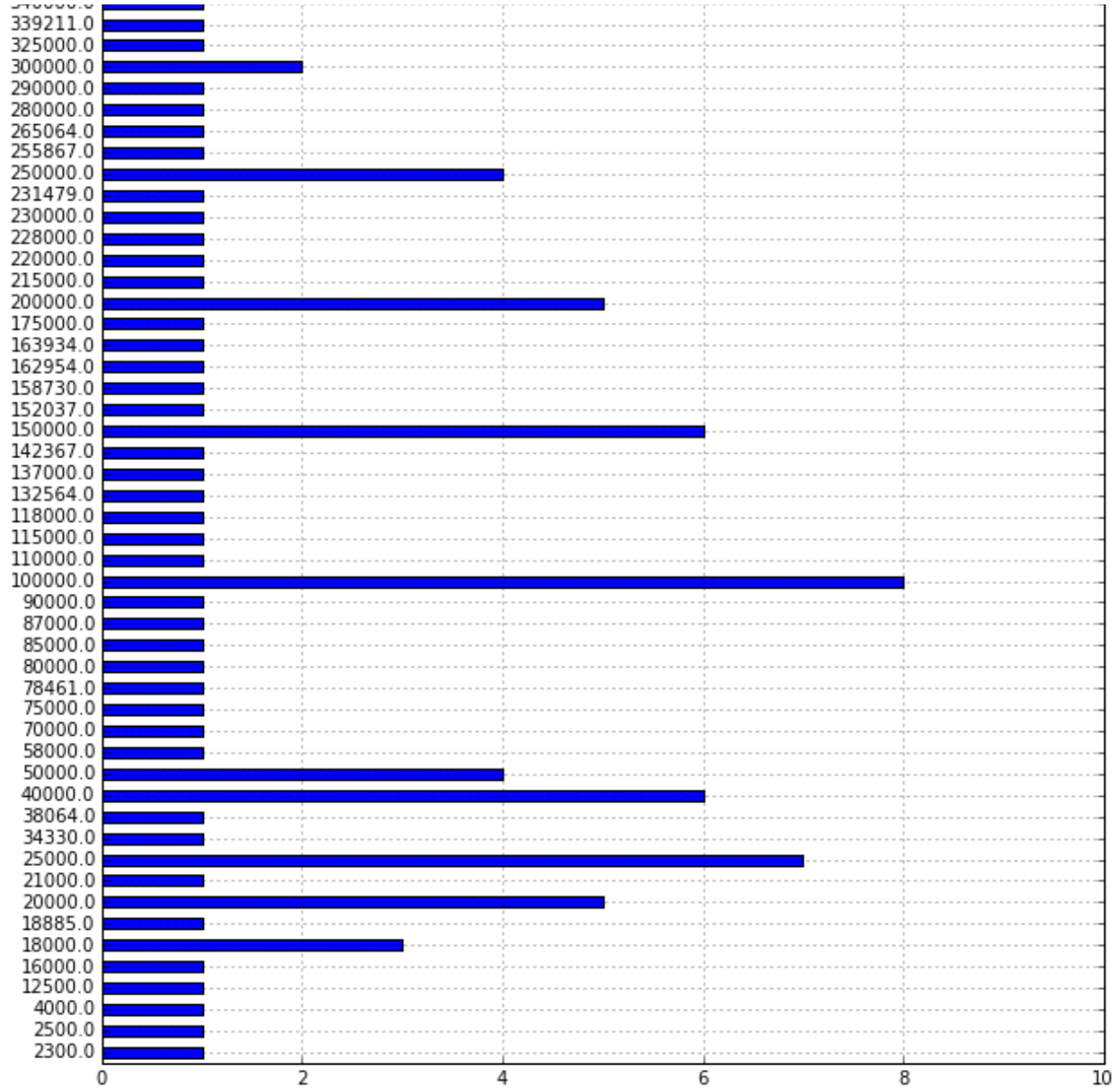
In [76]:
```
data = data[data.funding != 0]
funding = data['funding'].value_counts(ascending=True)
funding = funding.sort_index()
funding.plot(kind='barh', figsize=(10,50), use_index=True)
plt.show()

# Total Funding (in USD)
```



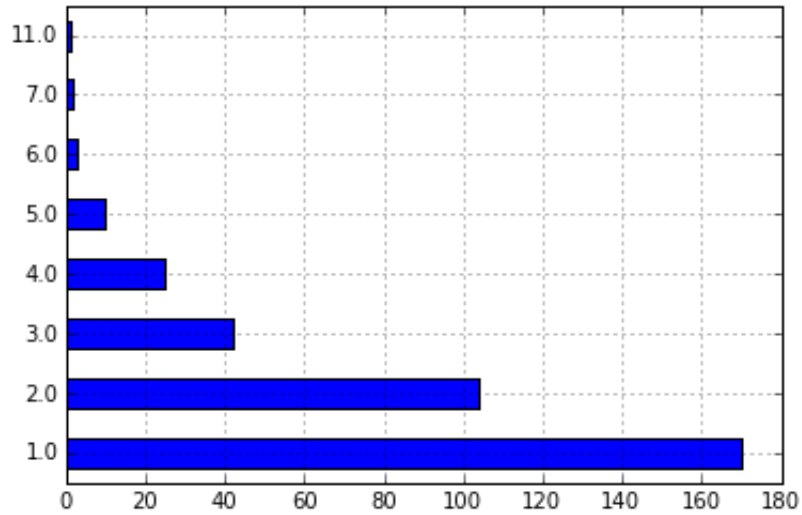

49500000.0
47049988.0
38000000.0
37684819.0
36000000.0
35600000.0
31500000.0
30956920.0
30000000.0
29470000.0
28949999.0
28703843.0
27000000.0
26600000.0
26000000.0
25000000.0
24000000.0
22552036.0
22500000.0
22023992.0
22000000.0
21000000.0
19479335.0
19391496.0
18800000.0
18086489.0
16480000.0
16120360.0
16000000.0
15000000.0
14300000.0
13954577.0
12700000.0
12350000.0
12000000.0
11700000.0
11017508.0
10600000.0
10293718.0
10250000.0
10100000.0
10061414.0
10000000.0
9950000.0
9705000.0
9500000.0
9400000.0
9300000.0
8923600.0
8700000.0
8692266.0
8620000.0
8600000.0
8115000.0
8100000.0
8094638.0
8000000.0
7800000.0
7600000.0
7576915.0
7500000.0
7473841.0
7156174.0
7000000.0
6500000.0
6371788.0
6200000.0
6000000.0
5989120.0
5700000.0
5500000.0
5400000.0
5000000.0
4800000.0
4787280.0
4700000.0
4650000.0
4500000.0
4300000.0
4150000.0
4000000.0
3900000.0
3581700.0
3500000.0


3440000.0
3356695.0
3252803.0
3217147.0
3030502.0
3000000.0
2998932.0
2860000.0
2800000.0
2720000.0
2650000.0
2619000.0
2550000.0
2525000.0
2503714.0
2500000.0
2450000.0
2225000.0
2203613.0
2049999.0
2038317.0
2020593.0
2000000.0
1977372.0
1835000.0
1800000.0
1775000.0
1750000.0
1746599.0
1714640.0
1666776.0
1570000.0
1538000.0
1534000.0
1500000.0
1490500.0
1380110.0
1350000.0
1300000.0
1270000.0
1250000.0
1227431.0
1200000.0
1159768.0
1135000.0
1100000.0
1025000.0
1000000.0
996000.0
964000.0
950000.0
926349.0
900000.0
875000.0
820000.0
815000.0
800000.0
784000.0
773174.0
750000.0
725000.0
716807.0
707630.0
694280.0
675000.0
650000.0
640000.0
615000.0
600000.0
550000.0
530000.0
514199.0
500000.0
498016.0
450000.0
438163.0
425000.0
410000.0
400000.0
389590.0
382500.0
370000.0
350000.0
344638.0
340000.0



In [51]:
```
funding_rounds = data['funding_rounds'].value_counts()
funding_rounds.plot(kind='barh')
plt.show()

# Number of Funding Rounds
```

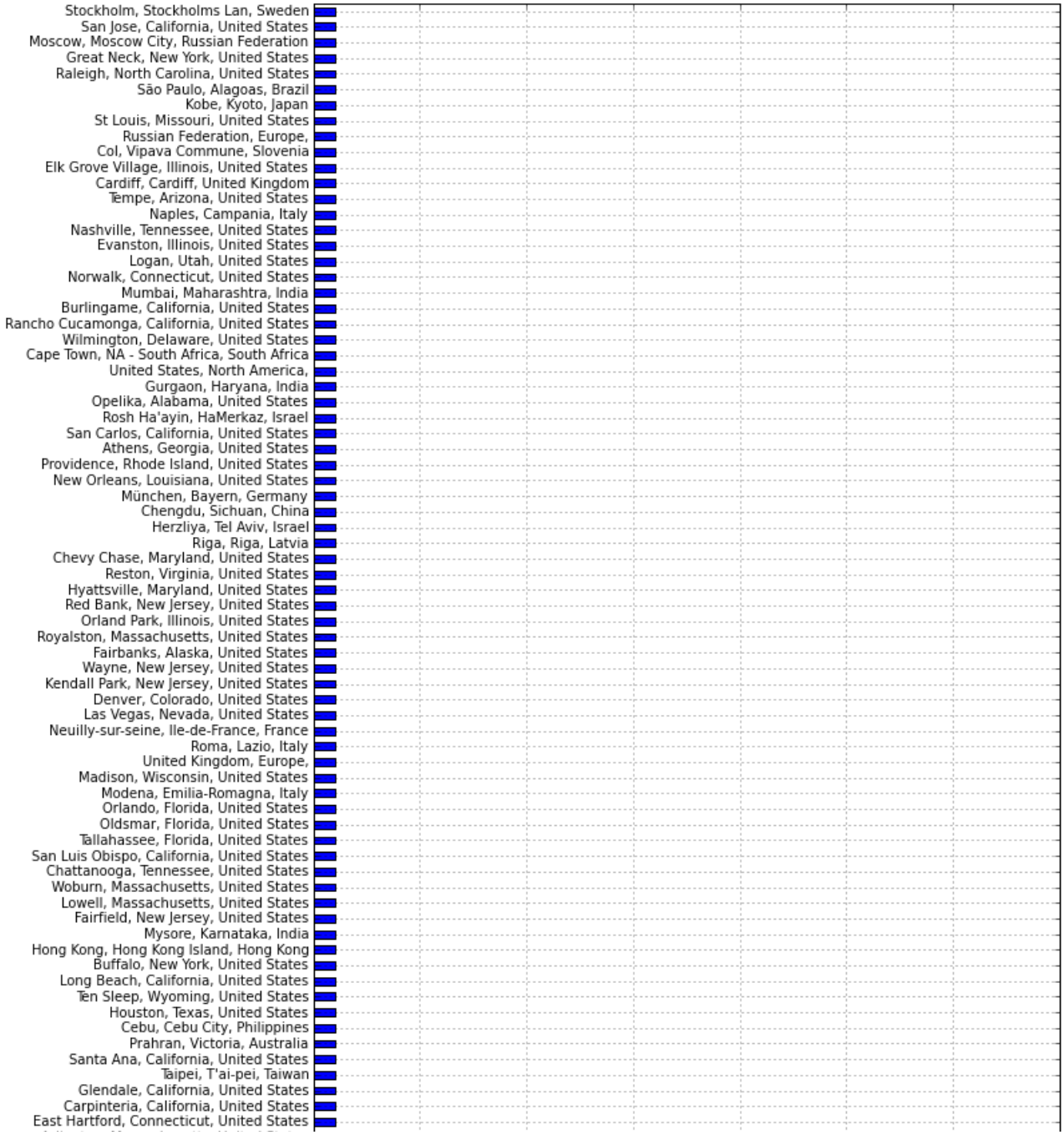


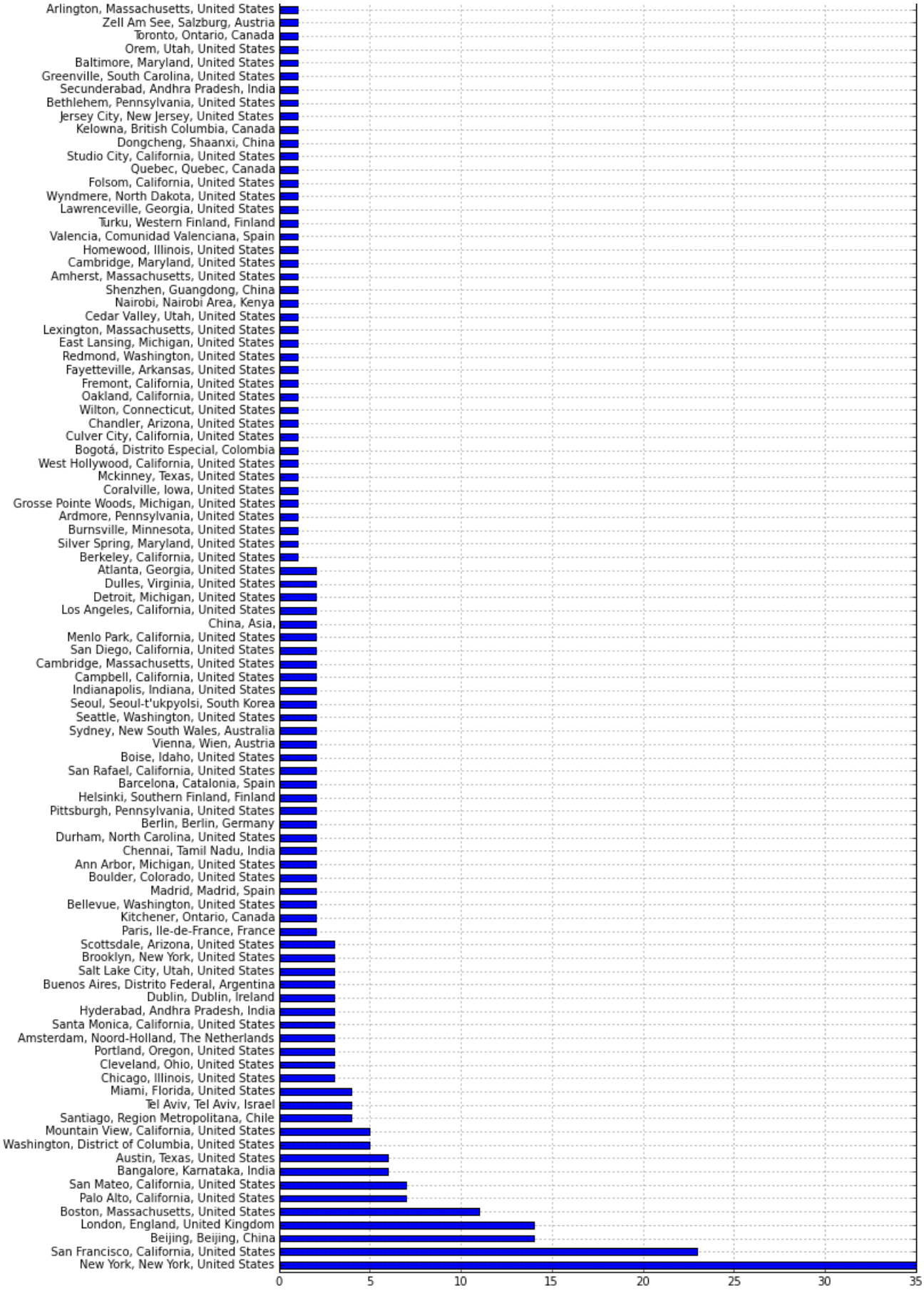
Location

Our dataset is international, and as such, we get a snapshot of the internation EdTech startup scene. Surprisingly, New York City represents the largest number of funded EdTech startups, with San Francisco in a close second. But the Sillicon Valley area encompasses not only San Francisco, but San Mateo and Palo Alto. Combining these 3 areas together we can see that, as expected, Sillicon Valley hosts the largest number of funded EdTech startups in the world.

After Sillicon Valley and New York, Beijing and London host an impressive number of funded EdTech companies as well. Boston, MA comes in next to round out the top 5 funded EdTech startup hubs of the world.

In [90]:
```
location = data.location.value_counts()
location.sort_index()
location.plot(kind='barh', figsize=(10,35))
plt.show()
```




Arlington, Massachusetts, United States
Zell Am See, Salzburg, Austria
Toronto, Ontario, Canada
Orem, Utah, United States
Baltimore, Maryland, United States
Greenville, South Carolina, United States
Secunderabad, Andhra Pradesh, India
Bethlehem, Pennsylvania, United States
Jersey City, New Jersey, United States
Kelowna, British Columbia, Canada
Dongcheng, Shaanxi, China
Studio City, California, United States
Quebec, Quebec, Canada
Folsom, California, United States
Wyndmere, North Dakota, United States
Lawrenceville, Georgia, United States
Turku, Western Finland, Finland
Valencia, Comunidad Valenciana, Spain
Homewood, Illinois, United States
Cambridge, Maryland, United States
Amherst, Massachusetts, United States
Shenzhen, Guangdong, China
Nairobi, Nairobi Area, Kenya
Cedar Valley, Utah, United States
Lexington, Massachusetts, United States
East Lansing, Michigan, United States
Redmond, Washington, United States
Fayetteville, Arkansas, United States
Fremont, California, United States
Oakland, California, United States
Wilton, Connecticut, United States
Chandler, Arizona, United States
Culver City, California, United States
Bogotá, Distrito Especial, Colombia
West Hollywood, California, United States
Mckinney, Texas, United States
Coralville, Iowa, United States
Grosse Pointe Woods, Michigan, United States
Ardmore, Pennsylvania, United States
Burnsville, Minnesota, United States
Silver Spring, Maryland, United States
Berkeley, California, United States
Atlanta, Georgia, United States
Dulles, Virginia, United States
Detroit, Michigan, United States
Los Angeles, California, United States
China, Asia,
Menlo Park, California, United States
San Diego, California, United States
Cambridge, Massachusetts, United States
Campbell, California, United States
Indianapolis, Indiana, United States
Seoul, Seoul-t'ukpyolsi, South Korea
Seattle, Washington, United States
Sydney, New South Wales, Australia
Vienna, Wien, Austria
Boise, Idaho, United States
San Rafael, California, United States
Barcelona, Catalonia, Spain
Helsinki, Southern Finland, Finland
Pittsburgh, Pennsylvania, United States
Berlin, Berlin, Germany
Durham, North Carolina, United States
Chennai, Tamil Nadu, India
Ann Arbor, Michigan, United States
Boulder, Colorado, United States
Madrid, Madrid, Spain
Bellevue, Washington, United States
Kitchener, Ontario, Canada
Paris, Ile-de-France, France
Scottsdale, Arizona, United States
Brooklyn, New York, United States
Salt Lake City, Utah, United States
Buenos Aires, Distrito Federal, Argentina
Dublin, Dublin, Ireland
Hyderabad, Andhra Pradesh, India
Santa Monica, California, United States
Amsterdam, Noord-Holland, The Netherlands
Portland, Oregon, United States
Cleveland, Ohio, United States
Chicago, Illinois, United States
Miami, Florida, United States
Tel Aviv, Tel Aviv, Israel
Santiago, Region Metropolitana, Chile
Mountain View, California, United States
Washington, District of Columbia, United States
Austin, Texas, United States
Bangalore, Karnataka, India
San Mateo, California, United States
Palo Alto, California, United States
Boston, Massachusetts, United States
London, England, United Kingdom
Beijing, Beijing, China
San Francisco, California, United States
New York, New York, United States
0
5
10
15
20
25
30
35

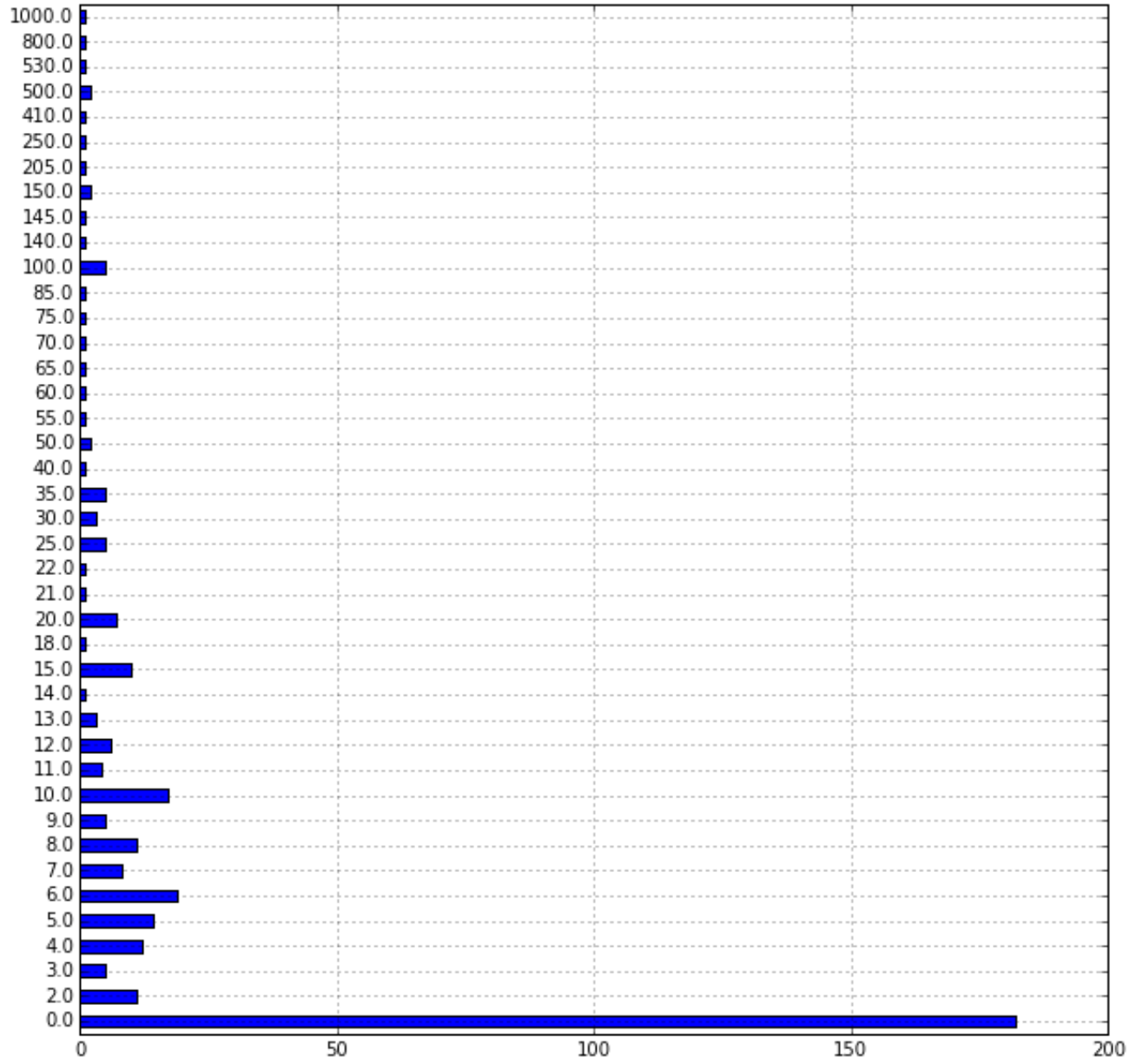
Number of Employees

To keep the type of company that is representative of a funded EdTech startup in perspective, it is useful to look at the number of employees. A large quantity of these companies don't report the number of employees - and we can assume these companies employee only a handful of people.

Most of these companies employ 10 or less people, although there are some relative outliers, with one company employing 1000 people and a few other employing in the 100's.

In [97]:
```
num_employees = data.num_employees.value_counts()
num_employees = num_employees.sort_index()
num_employees.plot(kind='barh', figsize=(10,10))
```

Out[97]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x108dc6d10>
```



Summary of Dataset

Given our analysis, we can see that many of the companies found in the CrunchBase API on EdTech companies don't have information on funding. Because this could point to companies that aren't startups, I made the decision to instead focus on **funded EdTech startups** from across the globe.

Hypotheses

With a more focused dataset, I feel that I will be able to provide deeper, more valuable insights. Given the nature of the data we are working with, I believe I will find a few things.

1. Sillicon Valley not only hosts the most EdTech startups, but their startups also receive the most funding, number of employees, have the most longevity and overall success.
2. Despite a large quantity of startups, New York and Boston don't possess startups of the magnitude of Sillicon Valley - typically only those of 10 or less employees and lesser funding.
3. Although this may be a result of the CrunchBase API, the international EdTech startup scene hasn't created a significantly large, succseful EdTech startup (for instance, Chegg).

In [ ]: