Startup funding pattern analysis

Prepared by Mid Eum Lee

Summary

Objective

The objective of this analysis is to investigate the history of investment in different categories of startups and get an insight on the significant factors for funding.

Goal

The goal is to overview the patterns of startup investment and provide the useful advice or insight for new entrepreneurs when they launch a startup.

Solution

Exploratory data analysis and inferential statistics

Project outline

The analysis is performed in three steps:

- Dataset was explored and analyzed.
- Dataset was cleaned and modified.
- Data visualization and statistical analysis were applied on the dataset.

Importing of packages

```
import pandas as pd
import numpy as np
from scipy import stats
import statistics
import matplotlib
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure
import scipy.stats as st


from IPython import display
import seaborn as sns
import csv
from statistics import mean
```

Exploration of the data

First , the dataset is loaded using pandas and this dataframe contains 20 columns and 13732 rows.

```
df = pd.read_csv('startup.csv')
df.head()
```

	name	category_list	funding_total_usd	status	country_code	state_code	city	funding_rounds	
0	H2O.ai	Software	33.600000	operating	USA	CA	Mountain View	4.0	Capital
1	One Inc.	Mobile	1.150050	operating	USA	CA	San Francisco	3.0	Ventures:
2	1000 Corks	Software	0.040000	operating	USA	OR	Lake Oswego	1.0	
3	1000museums.com	Software	6.795451	operating	USA	MA	Lenox	9.0	Alliance c
4	Redox	Health	4.000000	operating	USA	WI	Madison	2.0	.4

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13732 entries, 0 to 13731
Data columns (total 20 columns):
name                   13712 non-null object
category_list          13713 non-null object
funding_total_usd      13713 non-null float64
status                 13713 non-null object
country_code           13713 non-null object
state_code             13713 non-null object
city                   13713 non-null object
funding_rounds         13713 non-null float64
Investors              13713 non-null object
Number_of_Investors    13713 non-null float64
Acquirer               13713 non-null object
Acquirer_Category      13713 non-null object
Acquirer_Country       13713 non-null object
Acquirer_State         13713 non-null object
Acquirer_City          13713 non-null object
Acquired_Price         13713 non-null object
Acquired Currency      13713 non-null object
county                 13713 non-null object
founded_at             13713 non-null object
Coordinates            13732 non-null object
dtypes: float64(3), object(17)
memory usage: 2.1+ MB
```

```
df.isnull().any()
```

```
name                   True
category_list          True
funding_total_usd      True
status                 True
country_code           True
state_code             True
city                   True
funding_rounds         True
Investors              True
Number_of_Investors    True
Acquirer               True
Acquirer_Category      True
Acquirer_Country       True
Acquirer_State         True
Acquirer_City          True
Acquired_Price         True
Acquired Currency      True
county                 True
founded_at             True
Coordinates           False
dtype: bool
```

This dataset contains null values in most columns. This needs to be cleaned.

Data cleaning

All null values were dropped and the new dataset named df_dropped was created. Total rows changed from 13732 rows to 13712 rows.

All null values were removed.

Data distributions

The unique values in ‘founded_at’ were explored.

The funding_total_usd for all available years were navigated using a scatter plot.

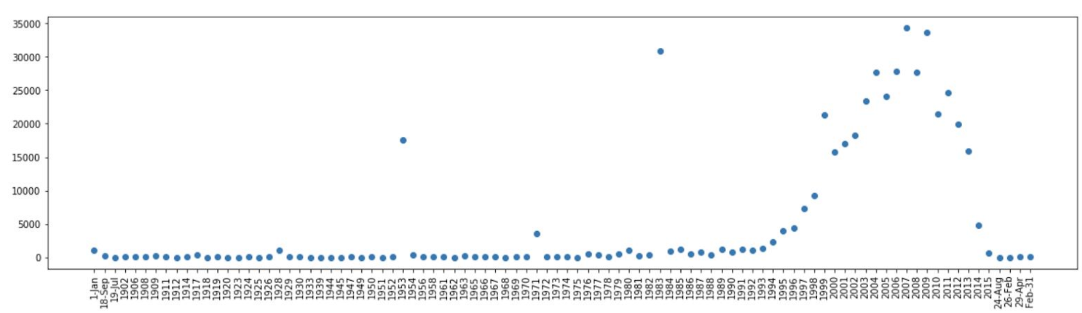
```
ls = []
for i in list(df_dropped['founded_at'].sort_values().unique()):
    ls.append(df_dropped[df_dropped['founded_at'] == i]['funding_total_usd'].sum())

print(ls)
```

```
[987.853094, 250.0, 9.0, 20.0, 16.0, 16.5, 245.0, 19.33, 2.0, 18.0, 300.0, 0.157, 120.0, 2.5,
1.75, 81.35, 0.15432, 16.600216, 1000.0, 59.929933, 25.65, 0.150768, 7.5, 5.0, 7.4, 10.0, 2.
0, 28.5, 6.0, 96.30000000000001, 17600.0, 331.0, 60.0, 19.6, 23.5, 6.0, 207.259114, 25.0, 10.
0, 90.8, 4.67321, 38.65, 13.378196, 3591.0, 21.36, 57.0726, 52.0, 0.75, 506.2, 334.096, 97.41
040000000001, 489.73600000000005, 1109.7, 203.2, 402.561365, 30826.27123, 969.485851, 1235.96
4973, 430.713798, 716.3299959999999, 296.064354, 1181.0722830000002, 792.1911379999999, 1247.
632421, 1073.55674, 1374.605991, 2250.795642, 3907.8827180000003, 4378.504615, 7232.24658342,
9221.717702399, 21293.925880151, 15731.084839000001, 16978.641015884, 18256.383254, 23352.259
765249997, 27726.36569675, 24054.835289000002, 27880.752675999996, 34298.53740258, 27645.4010
71570002, 33681.691192, 21523.016813984003, 24591.242369436, 19868.646866529, 15863.696440848
998, 4790.459357878, 672.336355021, 5.5, 1.0, 26.72, 18.5]
```

For ‘Month-Date’ values in the founding date(x-axis), this data may belong to years near the end of graphs based on its low funding total in overall trend.

```
plt.figure(figsize = (20,5), frameon = False)
plt.xticks(rotation=90)
plt.plot(list(df_dropped['founded_at'].sort_values().unique()), ls,'o')
```

```
[<matplotlib.lines.Line2D at 0x1241952d0>]
```

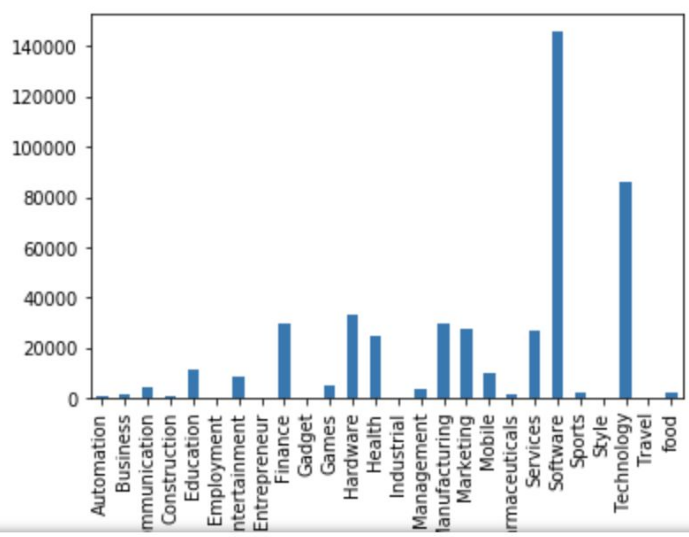


Data navigation

The funding_total_usd was plotted using category_list variable.

```
df_dropped.groupby('category_list')['funding_total_usd'].sum().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x10ae16fd0>
```



Another way of data visualization is performed using WordCloud.

```
from wordcloud import WordCloud

names = df_dropped["category_list"][~pd.isnull(df_dropped["category_list"])]
#print(names)
wordcloud = WordCloud(max_font_size=50, width=600, height=300).generate(' '.join(names))
plt.figure(figsize=(15,8))
plt.imshow(wordcloud)
plt.title("Wordcloud for category_list", fontsize=35)
plt.axis("off")
plt.show()
```

Data distributions

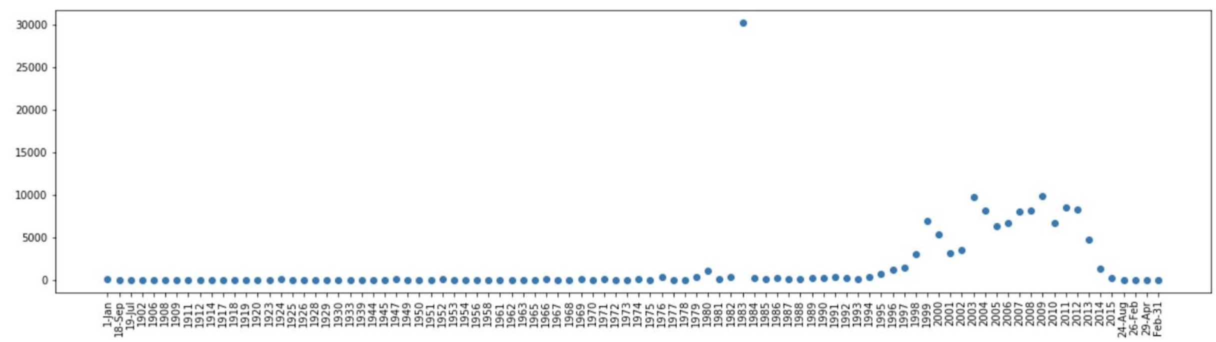
The ’funding_total_usd’ distribution was closely examined in the software category.
The unexpected peak was observed in 1983.

```
ls_software = []
for i in list(df_dropped['founded_at'].sort_values().unique()):
    ls_software.append(df_dropped[(df_dropped['founded_at'] == i) & (df_dropped['category_list'] == 'Software')]['fundi

print(ls_software)
```

```
[68.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 35.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
5.0, 0.0, 10.0, 0.0, 0.0, 0.0, 11.4, 0.0, 0.0, 0.0, 0.0, 0.5, 6.0, 0.0, 0.0, 10.0, 0.0, 4.5, 25.0, 0.0, 13.0, 0.0, 2.
5726, 12.0, 0.75, 330.2, 0.0, 0.0, 288.3, 1030.5, 104.0, 346.0, 30127.201999999997, 200.26204800000002, 101.729971, 2
36.38709300000002, 66.680511, 114.068238, 246.649674, 190.7, 344.546725, 157.02749999999997, 123.93339999999999, 284.
2, 722.272222, 1218.3065430000001, 1403.4045914199999, 2995.954001399, 6851.748235581001, 5274.368905, 3170.16211188
4, 3520.6371440000003, 9759.921296249999, 8078.887584, 6230.730723000001, 6664.218580000001, 8051.633184, 8089.57401
5, 9883.793102, 6659.123789, 8531.500703976, 8183.62137026, 4727.556383104, 1324.721404466, 159.1, 0.0, 0.0, 0.0, 0.
0]
```

```
plt.figure(figsize = (20,5), frameon = False)
plt.xticks(rotation=90)
plt.plot(list(df_dropped['founded_at'].sort_values().unique()), ls_software,'o')
```

```
[<matplotlib.lines.Line2D at 0x123afa750>]
```



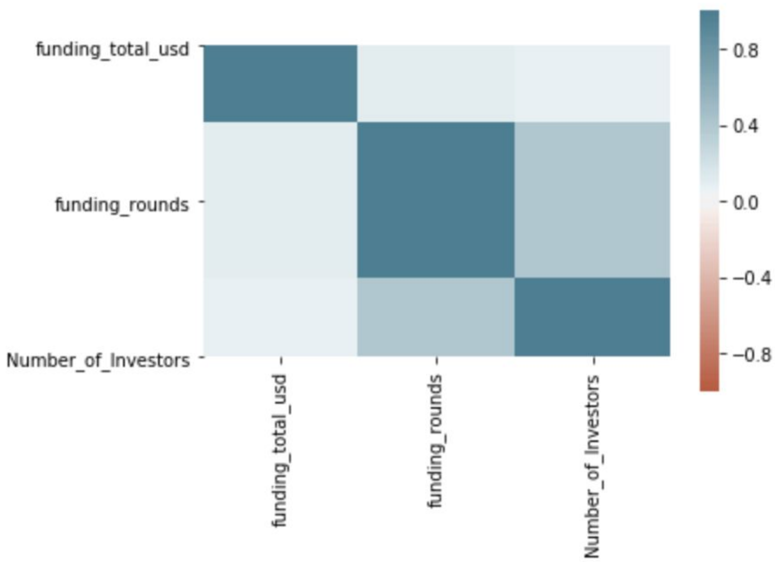
Data distributions

The ’funding_total_usd’ distribution was closely examined in the hardware category.
The unexpected peak was observed in 1953.

Data correlation

The data was analyzed to find the correlation between features.

```
corr = df_dropped.corr()
ax = sns.heatmap(
    corr,
    vmin=-1, vmax=1, center=0,
    cmap=sns.diverging_palette(20, 220, n=200),
    square=True
)
ax.set_xticklabels(
    ax.get_xticklabels(),
    rotation=90,
)
ax.set_yticklabels(
    ax.get_yticklabels(),
    rotation=0,
    horizontalalignment='right'
);
```



There is a bigger correlation between ‘number_of_investors’ and ‘funding_rounds’ compared to the ones : ‘funding_total_usd’ vs. ‘funding_rounds’ or ‘Number_of_Investors’.

Data correlation

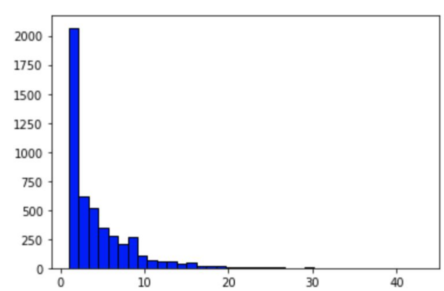
Pearson correlation between ‘number of investors’ and ‘funding total’ was calculated in order to test the hypothesis.

Null hypothesis: ‘number of investors’ is correlated with ‘funding total’.

In fact, the number of investors in the software industry (most dominant category) is mostly less than 2.

```
# matplotlib histogram
plt.hist(software_funding['Number_of_Investors'], color = 'blue', edgecolor = 'black',
         bins = int(180/5))
```

```
(array([2.069e+03, 6.180e+02, 5.150e+02, 3.470e+02, 2.790e+02, 2.110e+02,
        2.690e+02, 1.040e+02, 6.400e+01, 5.400e+01, 5.400e+01, 3.300e+01,
        4.700e+01, 1.900e+01, 1.600e+01, 1.600e+01, 8.000e+00, 1.000e+01,
        1.200e+01, 5.000e+00, 4.000e+00, 3.000e+00, 1.000e+00, 1.000e+00,
        7.000e+00, 2.000e+00, 0.000e+00, 1.000e+00, 2.000e+00, 0.000e+00,
        1.000e+00, 0.000e+00, 0.000e+00, 0.000e+00, 0.000e+00, 1.000e+00]),
 array([ 1.        ,  2.16666667,  3.33333333,  4.5       ,  5.66666667,
         6.83333333,  8.        ,  9.16666667, 10.33333333, 11.5       ,
        12.66666667, 13.83333333, 15.        , 16.16666667, 17.33333333,
        18.5       , 19.66666667, 20.83333333, 22.        , 23.16666667,
        24.33333333, 25.5       , 26.66666667, 27.83333333, 29.        ,
        30.16666667, 31.33333333, 32.5       , 33.66666667, 34.83333333,
        36.        , 37.16666667, 38.33333333, 39.5       , 40.66666667,
        41.83333333, 43.        ]),
 <a list of 36 Patch objects>)
```

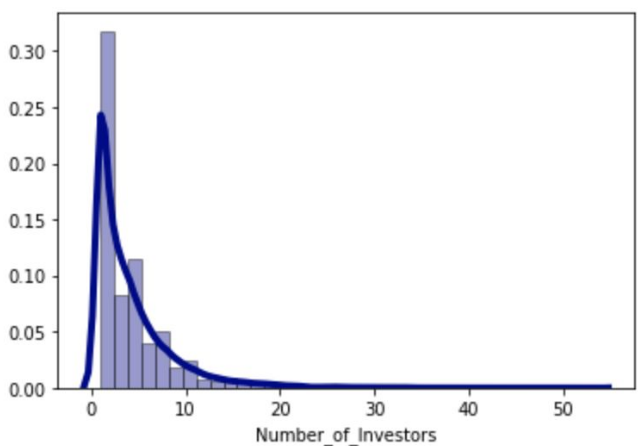


Data distributions

The distribution of the number of investors in the startup industry was plotted using density plot.

```
sns.distplot(df_dropped['Number_of_Investors'], hist=True, kde=True,
             bins=int(180/5), color = 'darkblue',
             hist_kws={'edgecolor':'black'},
             kde_kws={'linewidth': 4})
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1264e3610>
```



Data distributions

The funding total distribution grouped by the state is plotted. Most startups are located in CA.

Conclusion

Patterns of downgrowth

- After years of growth since 2007, investment has continued to underperform.
- From 2010 to 2013, investment declined significantly.

Patterns of upgrowth

- In 1983, massive funding was made in the software industry.
- In the Hardware field, the biggest investment was made only in 1953. This might be due to the fact that IBM effectively created the computer market in 1953 with the IBM 650.

Over 30% startups got 1~2 investors, and the funding round and number of investors showed a mild correlation.