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 useful advice or insight to new entrepreneurs and investors.

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.

```
df_dropped = df.dropna()
```

```
df_dropped.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13712 entries, 0 to 13712
Data columns (total 20 columns):
name                    13712 non-null object
category_list           13712 non-null object
funding_total_usd       13712 non-null float64
status                  13712 non-null object
country_code            13712 non-null object
state_code              13712 non-null object
city                    13712 non-null object
funding_rounds          13712 non-null float64
Investors               13712 non-null object
Number_of_Investors     13712 non-null float64
Acquirer                13712 non-null object
Acquirer_Category       13712 non-null object
Acquirer_Country        13712 non-null object
Acquirer_State          13712 non-null object
Acquirer_City           13712 non-null object
Acquired_Price          13712 non-null object
Acquired Currency       13712 non-null object
county                  13712 non-null object
founded_at              13712 non-null object
Coordinates             13712 non-null object
dtypes: float64(3), object(17)
memory usage: 2.2+ MB
```

```
df_dropped.isnull().any()
```

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

All null values were removed.

Data distributions

The unique values in ‘founded_at’ were explored.

```
df_dropped['founded_at'].unique()
```

```
array(['2011', '2008', '2014', '2000', '2013', '2012', '2007', '2010',
       '1990', '2002', '2001', '2009', '2006', '1999', '1998', '2004',
       '2015', '2005', '1986', '2003', '1971', '1989', '1997', '1979',
       '1993', '1987', '1961', '1996', '1947', '1994', '1992', '1984',
       '1995', '1949', '1985', '1974', '1969', '1975', '1918', '1980',
       '1944', '1972', '1983', '1982', '1976', '1988', '1991', '1977',
       '26-Feb', '1973', '1981', '1917', 'Feb-31', '1-Jan', '1970',
       '1906', '1978', '19-Jul', '1958', '1966', '1953', '1952', '1956',
       '1908', '1962', '1945', '1967', '1909', '1954', '1968', '1965',
       '1928', '1963', '18-Sep', '24-Aug', '1920', '1925', '1930', '1926',
       '1919', '1924', '1951', '29-Apr', '1923', '1914', '1929', '1902',
       '1939', '1912', '1911', '1933', '1950'], dtype=object)
```

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

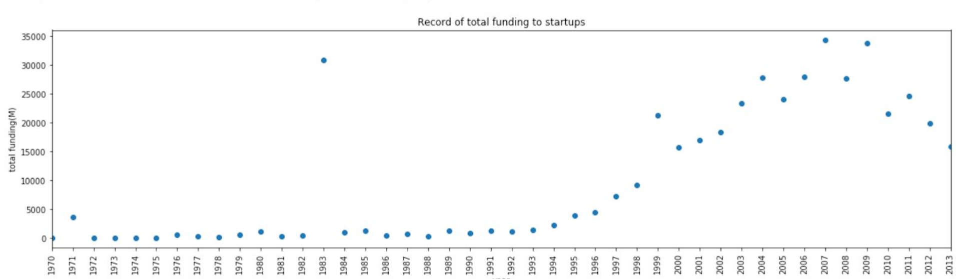
```
ls = []
for i in list(df_dropped['founded_at'].sort_values().unique()):
    ls.append(df_dropped[df_dropped['founded_at'] == i]['funding_total_usd'].sum())

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]
```

The startup funding before 1970 is not relevant to anything today. So the dataset was truncated to help bring out more important points. It seems very odd that the funding falls so dramatically from 2008. From 2014, funding steadily increased based on other data sources so this dataset is not enough to see the recent trend.

```
plt.figure(figsize = (20,5), frameon = False)
plt.xticks(rotation=90)
plt.plot(list(df_dropped['founded_at'].sort_values().unique()), ls,'o')
plt.xlim('1970','2013')
plt.ylabel('total funding(M)')
plt.xlabel('year')
plt.title ('Record of total funding to startups')
```

```
Text(0.5, 1.0, 'Record of total funding to startups')
```

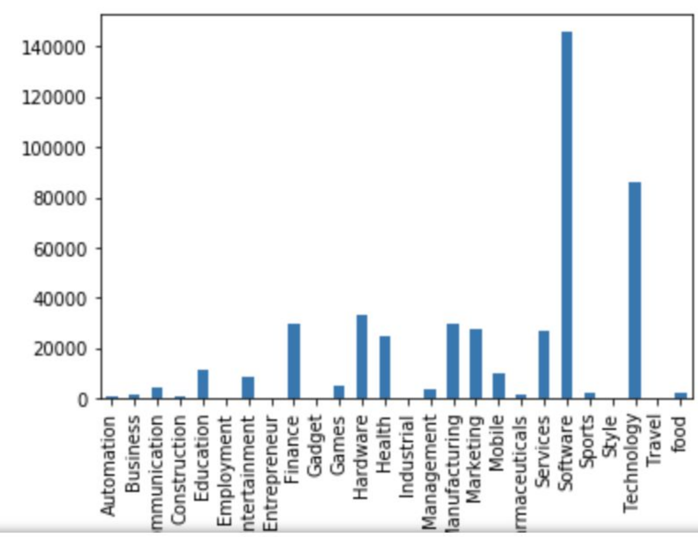


Data navigation

The funding_total_usd was plotted using category_list variable.

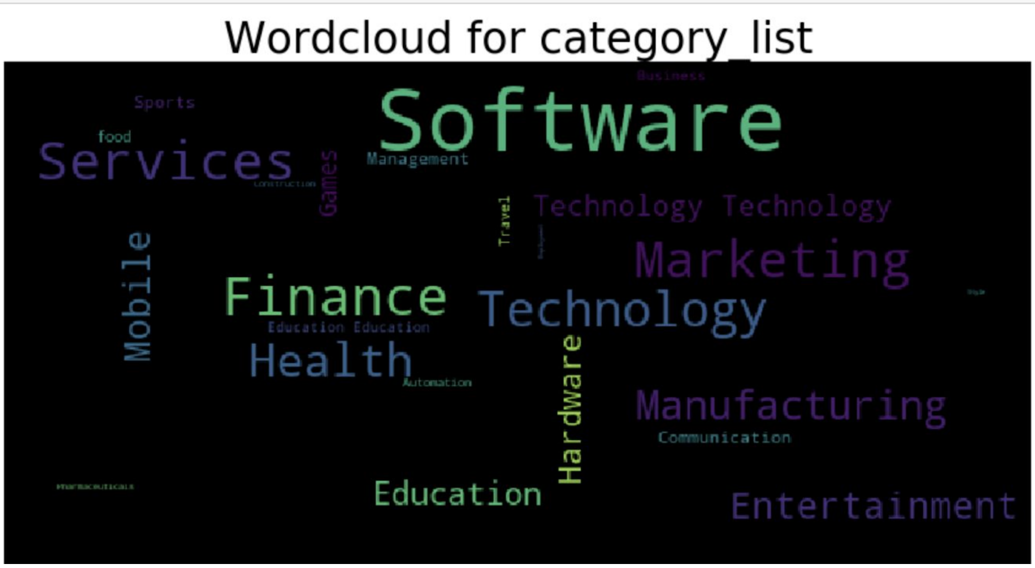
```
df_dropped.groupby('category_list')['funding_total_usd'].sum().plot(kind='bar')
```

```
<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()
```

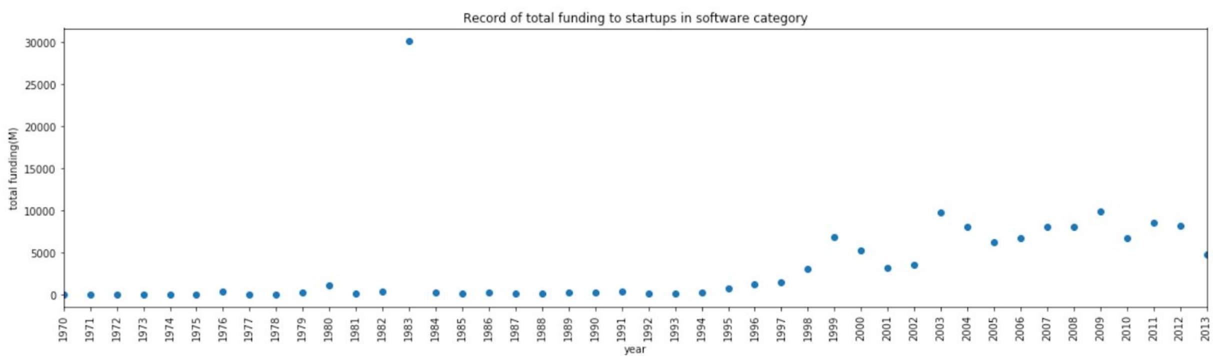


Data distributions

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

```
plt.figure(figsize = (20,5), frameon = False)
plt.xticks(rotation=90)
plt.plot(list(df_dropped['founded_at'].sort_values().unique()), ls_software,'o')
plt.xlim('1970','2013')
plt.ylabel('total funding(M)')
plt.xlabel('year')
plt.title ('Record of total funding to startups in software category')
```

```
Text(0.5, 1.0, 'Record of total funding to startups in software category')
```

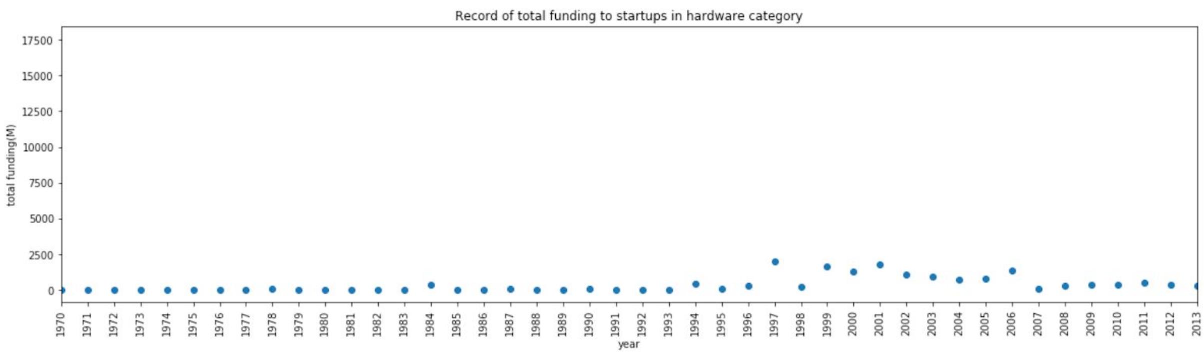


Data distributions

The 'funding_total_usd' distribution was closely examined in the hardware category.

```
plt.figure(figsize = (20,5), frameon = False)
plt.xticks(rotation=90)
plt.plot(list(df_dropped['founded_at'].sort_values().unique()), ls_Hardware,'o')
plt.xlim('1970','2013')
plt.ylabel('total funding(M)')
plt.xlabel('year')
plt.title ('Record of total funding to startups in hardware category')
```

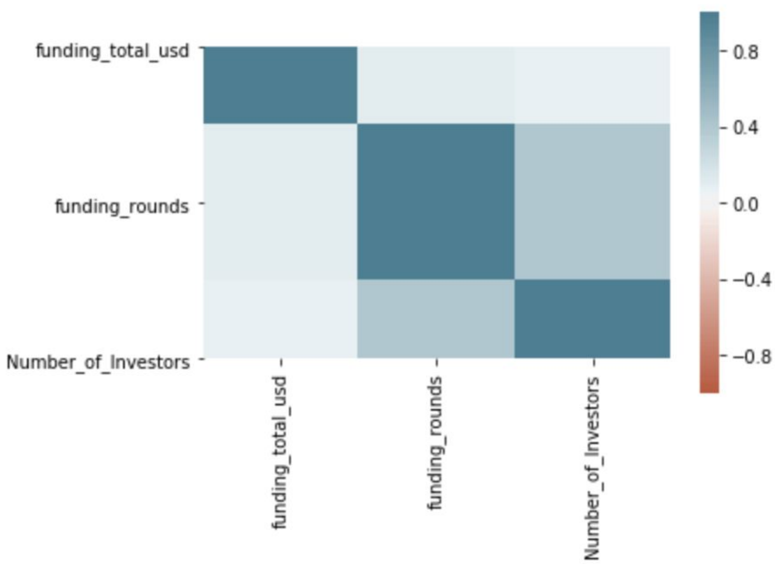
```
Text(0.5, 1.0, 'Record of total funding to startups in hardware category')
```



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’.

```
X = df_dropped.Number_of_Investors.values
Y = df_dropped.funding_total_usd.values
```

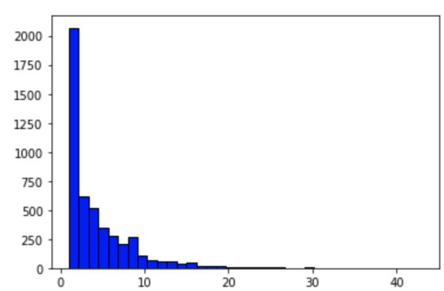
```
def pearson_r(X, Y):
    corr_mat=np.corrcoef(X,Y)
    return corr_mat[0,1]
r_obs = pearson_r(X,Y)
print('Observed significance value=',r_obs)
```

```
Observed significance value= 0.08024147738834027
```

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>)
```

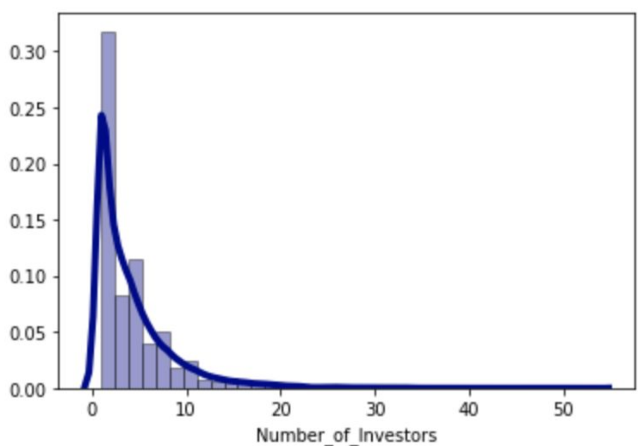


Data distributions

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>
```

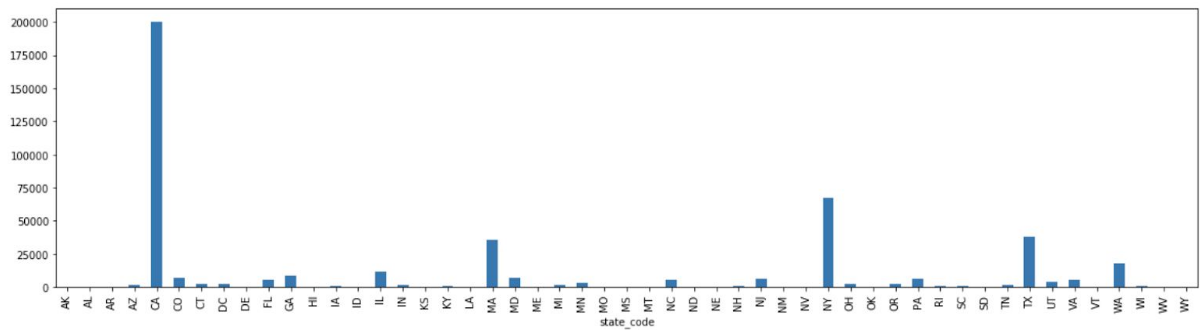


Data distributions

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

```
plt.figure(figsize = (20,5), frameon = False)
df_dropped.groupby('state_code')['funding_total_usd'].sum().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1250e6d10>
```



Conclusion

Patterns of downgrowth

- After years of growth since 2007, investment has continued to underperform.
- From 2010 to 2013, investment declined significantly. From 2014 forward, data from another data source (https://nvca.org/8-takeaways-8-graphics-historic-2018-venture-capital/) says funding steadily increased, which is not shown from this dataset

Patterns of upgrowth

- In 1983, massive funding was made in the software industry.
- In the Hardware field, we do not see any peak in startup investment.

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