analyze “dead company dataset” by python

Summary of dataset:

Dataset download link: [6271家死亡公墓公司 - Heywhale.com](https://www.heywhale.com/mw/dataset/5e023cd12823a10036af49b4)

It is

1. full year data from 2019
2. 6,271 companies in various industries are included
3. Contains 15 feature dimensions

This is a data set of information of 6,271 dead companies.

A certain row shows that some attributes of a company.

Eg. If open dataset(by csv editor),we can find out the company’s the name, address, category, subdivision of the category, description, birthday, dead day, survive time, financing way, the total money of financing, dead reason

We want to analyze this data set to get some caveats if you want to invest in companies

Analysis steps:

1. Dataset cleaning
2. make a histogram / Pie Chart of the data
3. Draw a heat map and analyze the correlation coefficient of each attribute in the dataset
4. get conclusion

**Dataset cleaning**

It is observed that the data set has empty columns, and a round of data cleaning is required first. In fact, fill in the blank with “ ‘unknow “

If there is no data cleaning, the length of the matrix is different. According to the code applying this program, it will eventually fail to run.( For example: the rows and columns of the two matrices do not correspond, and the multiplication calculation cannot be performed.)

**Draw a histogram by survival time**

We can observe that at the Years 3(1000 day) and 7 (2500 day)is the hard time for numbers of company .

**Draw a Industry Distribution Pie Chart**

We can observe that E-commerce and enterprise services are the industries with the most distribution among dead companies

**Draw a heatmap**

#Heatmap

Analyze the relationship between survival time and other characteristics (province, territory, financing stage)

Heatmaps are often used in practice to display the correlation coefficient matrix of a set of variables,

It also has a great use in displaying the data distribution of the contingency table. Through the heat map, we can intuitively feel the difference in numerical value.

We are going to perform one-hot encoding on three of the label variables (category, financing way, comany\_address

) int to digital data, then calculate the Correlation coefficient and draw a heat map

#One-Hot Encoding,

also known as one-bit effective encoding, is to use an N-bit state register to encode N states,

# Each state has its own register bits, and only one of them is valid at any time. That is, only one bit is 1, and the rest are zero values.

According to the program for making the heat map, each number corresponds to an attribute through one-hot encoding, which is divided into four categories, the total is 73

# 区间

# 0~18 领域

# 19~34 融资方法

# 35~71 公司地点

# 72 存活天数

There are indeed some dark spots on the heat map (the correlation coefficient is relatively large), The darker the heatmap, the stronger the correlation, we print out them directly

# 20 0.054765

# 39 0.055830

# 23 0.056662

# 22 0.077564

# 31 0.094593

# live\_days 1.000000

print(data.financing.value\_counts().index[31-19])

print(data.financing.value\_counts().index[22-19])

print(data.financing.value\_counts().index[23-19])

print(data.com\_addr.value\_counts().index[39-35])

print(data.financing.value\_counts().index[20-19])

These attributes are positively correlated with the company's survival time:

B+輪 A輪 B輪 四川 未獲得投資

can also draw some interesting conclusions from the same principle…

corr[num] query by column (column) as key

According to Wikipedia's definition of correlation coefficient

相关性 负 正

无 −0.09 to 0.0 0.0 to 0.09

弱 −0.3 to −0.1 0.1 to 0.3

中 −0.5 to −0.3 0.3 to 0.5

强 −1.0 to −0.5 0.5 to 1.0

These properties are correlated (weakly correlated):

# 企業服務: 香港 = 0.1335995260407172

# A輪 : 福建 = 0.12507250066324943

# 種子: 福建 = 0.14665307301514435

# B+輪 : 福建 = 0.16248528147042204

Conclusion:

* Through the above analysis, we can draw a rough conclusion：
* If you‘re going to invest in a company, maybe…
* Try to avoid companies that are in their 3th and 7th years
* E-commerce and enterprise services are popular, but they are also really prone to failure
* Seeing that companies have these attributes, maybe should be careful
* B+輪,A輪,B輪, 未獲得投資,四川
* These attributes are related in pairs in dead companies, and may represent the economic composition of the region, which is worth further studying
* # 企業服務: 香港
* # A輪 : 福建
* # 種子: 福建
* # B+輪 : 福建