CS838 Data Science project progress report - Stage 5

Jinman Zhao jzhao237@wisc.edu

Bin Guo bguo23@wisc.edu

Di Wu dwu73@wisc.edu

1. Data description

- **Statistics on Table E: specifically, what is the schema of Table E, how many tuples are in Table E? Give at least four sample tuples from Table E.**

The combined table contains 702 matched tuples from the two table. 2407 unmatched tuples from FORBES500 table and 4011 unmatched tuples from NASDAQ table.

- The schema of the table E is:
  - ltable_id: 3110 valid data in integer type
    table id from the left table (FORBES500 table)
  - rtable_id: 4714 valid data in integer type
    table id from the right table (NASDAQ table)
  - Name: 7121 String
    The name of the company, using the name from the NASDAQ table
  - Country: 7121 String
    The country or area in which the company is located, extracted from the NASDAQ table
  - Industry: 6224 valid data in string type
    The industry the company belongs to, extracted from NASDAQ table
  - Market Value: 6576 valid data in float type
    Market value of the company extracted from NASDAQ table
  - Assets: 2598 valid data in float type
    Assets value of the company extracted from NASDAQ table
  - Employee: 983 valid data in float type
    Employee number extracted from the FORBES500 table
  - Sales: 3105 valid data in float type
    Sales value extracted from the NASDAQ table
  - Profits 2796 valid data in float type
    Profits value of the company from the FORBES500 table
  - IPOyear 2225 valid data in float type
    The first year the company went to public, extracted from NASDAQ table
  - Symbol 4714 valid data in string type
    Stock symbol from NASDAQ table
  - Last Sale 4714 valid data in float type
    Latest stock price extracted from NASDAQ table
  - Summary Quote 4714 valid data in string type
    Introduction page of the company
  - Sector 4012 valid data in string type
    The sector to which the company belongs, extracted from NASDAQ table

Some examples are:

	ltable_id	rtable_id	Name	Country	Industry	MarketValue	Assets	Employee	Sales	Profits	IPOyear	Symbol	LastSale	Summary Qu	Sector
0	2051	3485	Principal Fin	United State	Accident &H	18142.66	218700.00		11900.00	1200.00	2001.00	PFG	63.11	http://www.	Finance
1	2053	3517	Prudential Fi	United State	Life Insuranc	45911.88	757400.00		53200.00	5600.00	2001.00	PRU	106.68	http://www.	Finance
2	684	1103	Constellatior	United State	Beverages (P	31775.39	17000.00		6500.00	1100.00		STZ	162.07	http://www.	Consumer Non-Durables
3	2061	4234	Torchmark C	United State	Life Insuranc	9082.82	19900.00		3800.00	527.00		TMK	77.04	http://www.	Finance
7117		4711	Zumiez Inc.	United State	Clothing/Sho	456.46					2005.00	ZUMZ	18.30	http://www.	Consumer Services
7118		4712	Zweig Fund,	United States		180.27					1986.00	ZF	11.12	http://www.nasdaq.com/symbol/zf	
7119		4713	Zynerba Pha	United State	Major Pharm	265.62					2015.00	ZYNE	20.10	http://www.	Health Care
7120		4714	Zynga Inc.	United State	EDP Services	2474.38					2011.00	ZNGA	2.85	http://www.	Technology

2. Results

- **What was the data analysis task that you wanted to do? For that task, describe in detail the data analysis process that you went through.**

We decided to perform some correlation discovery type of data analysis. For this task, we consider the relationship between each pair of the numeric type schemas. We use *"seaborn"* package of Python to draw the correlation graph of each pair and got the following results.

Figure 1. Histograms and pairwise correlation of 7 attributes.

According to Fig. 1, there are only one pair of attributes ("MarketValue" and "Profits") may have correlation. The distribution of "IPOyear" seems to have two significant increasing periods with a huge downturn in between. Thus, we further investigate the relationship between "MarketValue" and "Profits" and the distribution feature of "IPOyear".

2.1. "MarketValue" distribution

As shown in Fig. 2, the majority of companies in our data have market value below 1,000,000 million dollars. And there are very few companies that have large market values up to about 6,000,000 million dollars.

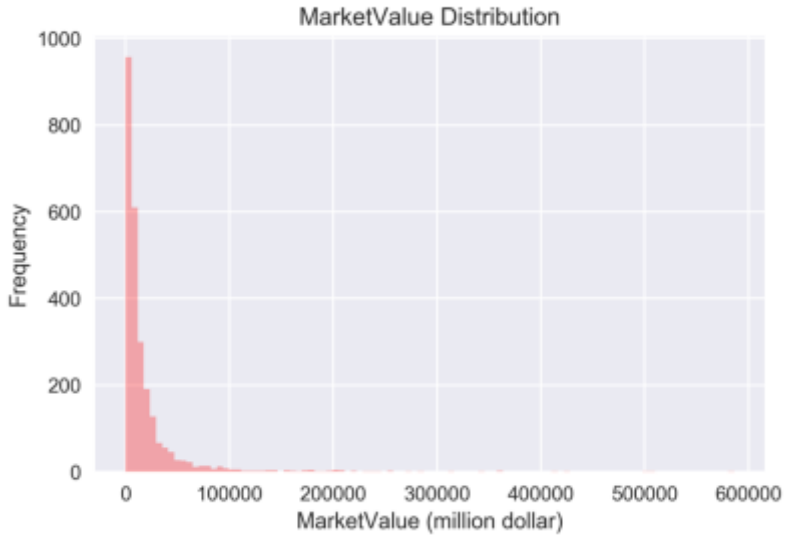

Figure 2. Histogram of market values.

The number of companies and companies' market values have negative correlation, which is within our expectation. In reality, most of the companies that are on "forbes.com" ranking lists or registered in "NASDAQ" have similar market values and should be close to each other. Some "star" companies, which are on the top of "Forbes.com" ranking list, have large market values that are almost with no upper limit. The distribution with a peak in the beginning and a very long tail can be expressed as a probability density function (PDF). We estimated the PDF using "seaborn" package (Fig. 3).

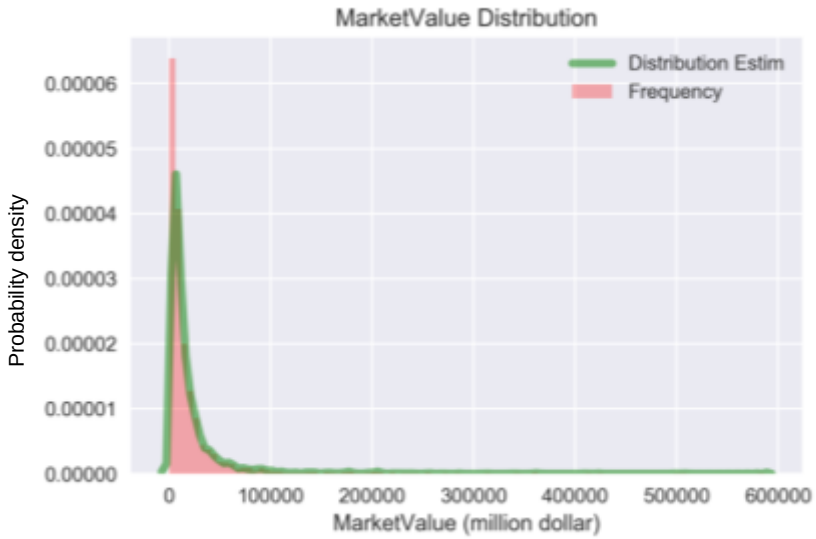

Figure 3. PDF of market values.

2.2. Relationship between "MarketValue" and "Profits"

Fig. 4 shows the relationship between the profits and the market value and used cubic regression to fit the data.

Figure 4. Profits and market value distribution.
Green line is the cubic regression fitting line of the data.

Considering the between Profits value (-5000, +30000), the regression line shows a positive correlation between profits and market value. The market values increase as the profits increase from negative to positive.

2.3. “IPOyear” distribution

From Fig. 1, we observe the gap around year 2000 between the increasing periods in 1980s and 2000s. The first guess would be there might be an economy drop to slow down the number of companies that go public.

We first drew the IPO year distribution histogram in a mid-coarse level (Fig. 5). Surprisingly, when the histogram is plotted in a regular scale, the trend of companies go public each year is increasing since 1970. This is different from the observation in Fig. 1 due to different scale on frequency axis.

To further investigate the distribution of “IPOyear”, we use more bins to plot the histogram (Fig. 6). The number of companies go public on “NASDAQ” increases since 1970. However, there are some decreases embedded between increases. After some research on [website], we find the correlation between financial crises and the number of companies go public. There are 6 financial crises between 1980 and 2010, and they align well with the decreases in the “IPOyear” histogram. The effect of each financial crisis leaded to the drop-in number of companies go public in the flowing years. For the “Asian crisis” in the 1997 – 1998, the drop-in number of companies go public is not as significant as the others. The reason is in our data, the number of Asian companies registered on “NASDAQ” is very limited.

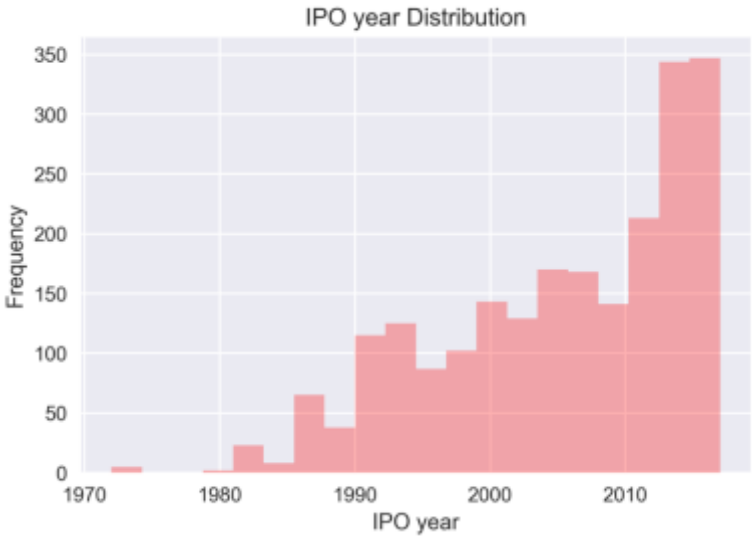

Figure 5. “IPOyear” distribution.

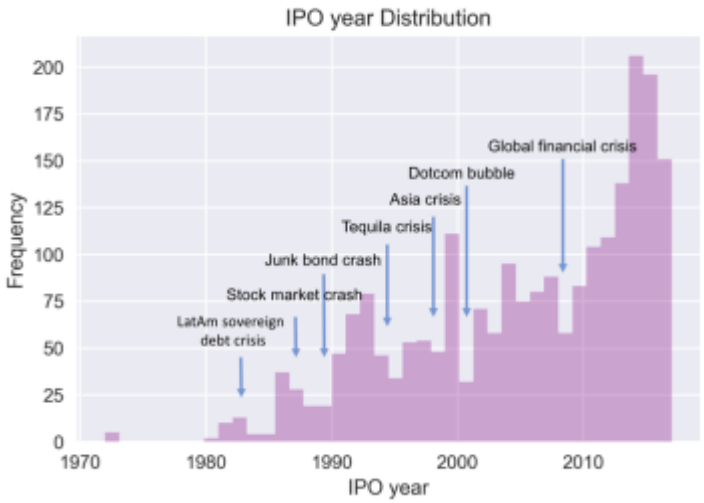

Figure 6. "IPOyear" distribution and relationship with 6 financial crises.

2.4. Other results

- **Give any accuracy numbers that you have obtained (such as precision and recall for your classification scheme).**

  We performed correlation discovery type of data analysis, so no accuracy numbers can be provided.

3. Lessons learnt

- **What did you learn/conclude from your data analysis? Were there any problems with the analysis process and with the data?**

  From the discussion above, we conclude that:
  - The relationship between the profits value and the market value is positively related. This is something we expected. However, there are still some "outliers" shown on the map.
  - The relationship between the IPO year and the financial crises are discovered by accident. As we don't have too much knowledge about the finance, it is a surprise for us to find out that the IPO number of the year can really reflect the state of financial market.

  When analyzing the data, we encountered some errors. The errors are mainly due to the following two reasons:

  - The data is not correct, like in the market value column, there was value that is "Bank", which seems to be the industry value. This may be caused by the bug when scraping the data from the website.
  - The numeric data not in perfect format, like the number "1,234" which will not be correctly recognized by the program as number.

We can conclude that, doing data clean before the data analysis is very important.

- **If you have more time, what would you propose you can do next?**

  For the IPO year distribution analysis, we can further discover whether the financial crises are related to some specific industry by performing some group-by query on the industry column. The summary count the industry property will give us the detailed information about which kind of industries are related to the financial crises.