Fortune 100

GitHub Repsitory: <https://github.com/NuooXu/INFO6210>

### AUDIT VALIDITY/ACCURACY

The definition of data accuracy is whether the data values stored for an object are the correct values. To be correct, a data values must be the right value and must be represented in a consistent and unambiguous form.

### AUDIT COMPLETNESS

In real world, when a list of stock information of companies is requested, a list of particular information will be provided. Also, we can compare different information of different companies and calculate the difference between them. Then we can compare the percentage and similar data of that.

### AUDIT CONSISTENCY/UNIFORMITY

The dataset in this assignment display a uniform relationship between all the dataset since they are lined one by one through some attributes

### REPORT

fies used : data.csv, Fortune500\_01.html

files genearted: data.csv, Fotun100.csv,Fortune500.csv

Code used:

Step 1. Extraction of Data

3 main methods were used for the extraction of data:

Using the API:

From the public api, i found the api provided by some producer and got

pandas to create data frames from the raw data

Using the website to scrap the data

Here the data was extracted using the sites data directly using the libraries like:

request to access the website using the URL (https://fortune.com/fortune500/2019/search/)

BeautifulSoup to scrape the contents of the website

find() and find\_all() methods were used to find the desired content in the system

By loading the csv file:

Download the data information from the Internet related to financial information of some companies

Code used for Merge

### CONCLUSION

From this project, I understand how to get data source from the Internet, including three ways.

First, using beautifulsoup to scrape website. By analyzing tags, we can extract the information that we need. And I have to say Beautifulsoup is a very convient tool, for it very easy to use. Second, using API to get the data. I believe that find a suitable website is important. At the beginning, we try to find Fortune's API, but it dosen't privode it. So we spend a very long time to search suitable data source. Finally, we decide use companies' stock data as the API source. Through the url the website provided, we get the data successfully. Third, we download the row data of the Fortune 500 companies from the Internet.

Also, we learn how to visualize the data that we get from the website to make it easier to understand,by using seaborn, matplotlib and scipy. From the graphs we make, others can have a Intuitive understanding of what we doing and what is the result.

Finally, by using SQL workbanch, we made the E-R model to show the relations between each data source. From the model, we can see what data we have and what type it is.

### CONTRIBUTION

I contributed By Own: 50%

By External source: 50%

### CITATIONS

https://fortune.com/fortune500/2019/search/please

https://www.someka.net/excel-template/fortune-500-excel-list/

https://financialmodelingprep.com/api/v3/company/profile/MSFT

https://financialmodelingprep.com/api/v3/company/profile/AAPL

https://financialmodelingprep.com/api/v3/company/profile/FB

https://financialmodelingprep.com/api/v3/company/profile/NVDA https://financialmodelingprep.com/api/v3/company/profile/ZNGA

https://financialmodelingprep.com/api/v3/company/profile/AMZN

https://financialmodelingprep.com/api/v3/company/profile/NKE