

FE-520 Assignment 4

Dan Wang, Zhiyuan Yao

April 10, 2020

Submission Requirement:

For all the problems in this assignment you need to design and use Python 3, output and present the results in nicely format.

Please submit a written report (pdf), where you detail your results and copy your code into an Appendix. You are required to submit a single python file and a brief report. Your grade will be evaluated by combination of report and code.

You are strongly encouraged to write comment for your code, because it is a convention to have your code documented all the time.

Python script must be a '.py' script, Jupyter notebook '.ipynb' is not allowed.

Do NOT copy and paste from others, all homework will be firstly checked by plagiarism detection tool.

1 Credit Transaction data(40 points)

This dataset is simulated individual credit card transactions by one company. Please use this dataset to answer following question. Please notice that you may need to observe the dataset and clean it before answering the following question.

1. What is total amount spending captured in this dataset?
2. How much was spend at WW GRAINGER?
3. How much was spend at WM SUPERCENTER?
4. How much was spend at GROCERY STORES?

2 Data Processing with Pandas (60 points)

In this practice, you are expected to play around Pandas and get familiar with it. The dataset is quarterly dataset downloading from WRDS. Please remember that you need to do data transformation based on the new dataset generated by previous step. Do not using other package other than numpy and pandas.

1. Read 'Energy.xlsx' and 'EnergyRating.xlsx' as BalanceSheet and Ratings(dataframe).

2. drop the column if more than 90% value in this column is 0 (or missing value).
3. replace all None or NaN with average value of each column.
4. Normalize the table (Only need to normalize numerical parts)

Using `pd.apply()` to normalize the table, in this table, you need to implement follow formula to calculate the normalized value:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

(Do not using any function like `MinMax()`, you need to write it by yourself)

5. Calculate the correlation matrix for variables = ['Current Assets - Other - Total', 'Current Assets - Total', 'Other Long-term Assets', 'Assets Netting Other Adjustments'].
6. If you look at column ('Company Name'), you will find some company name end with 'CORP', 'CO' or 'INC'. Create a new column (Name: 'CO') to store the last word of company name. (For example: 'CORP' or 'CO' or 'INC') (Hint: using map function)
7. Merge (inner) Ratings and BalanceSheet based on 'datadate' and 'Global Company Key', and name merged dataset 'Matched'.
8. Mapping
For dataset 'Matched', we have following mapping:
 AAA = 0
 AA+ = 1
 AA = 2
 AA- = 3
 A+ = 4
 A = 5
 A- = 6
 BBB+ = 7
 BBB = 8
 BBB- = 9
 BB+ = 10
 BB = 11
 others = 12
 Using map function to create a new variable = 'Rate', which maps ratings to numerical ratings.
9. Calculate the rating frequency of company whose name end with 'CO'. (Calculate the distribution of rating given the company name ending with 'CO', Hint, use map function)
10. Output your final dataset as 'HW4.csv'.