FE520 Assignment 4

Dan Wang, Zhiyuan Yao

October 2018

1 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.

1. Read 'AAPL BS.csv' and 'AAPL Ratings.csv' as BalanceSheet and Ratings(dataframe). (5 points)

2. Dealing with missing value (15 points)

   1. Drop the variable = 'aqepsq' and 'gdwlamq'
   2. replace all 0 in the column = 'dlttq' into the mean value of this column.
   3. Fill the missing value in the column = 'intanq' using linear interpolation.

3. Filter outliers (5points)

   For the column 'niq' and 'citotalq', adjust the value equal to 15000 if the original value greater than 15000.

4. Summary the dataset (10 points)

   Using pd.apply() to return a summary table, in this table, you need to calculate the second largest value in each variables, the range for each variable. (in this practice, you only need to focus on variables = ['acoq', 'actq', 'apq', 'chq', 'citotalq'].

5. Calculate the correlation matrix for variables = ['acoq', 'actq', 'apq', 'chq', 'citotalq', cogsq]. (5 points)

6. Merge (inner) Ratings and BalanceSheet based on 'datadate', and name merged dataset 'Matched'. (5 points)

7. Mapping (5 Points)

   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
Using map function to create a new varible = 'Rate', which maps ratings to numerical ratings.

8. Random sample your new dataset to double its size. (5 points)

9. Output your final dataset as 'HW4.csv'. (5 points)

2 Practice on Numpy II (40pt)

1. (10pt) Create a function to replace infinite number in a numpy matrix (2D-array) with 2^8;
   Example:

   ```
   >>> a = np.array([[1,2],[4,0]])
   >>> 1/a
   ... array([[1.  , 0.5 ],
              [0.25,  inf]])
   >>> myFun1(1/a)
   ... array([[1.00e+00, 5.00e-01],
              [2.50e-01, 2.56e+02]])
   ```

2. (10pt) Create a function which take a 2-D numpy array as input, replace all the numbers which is greater than 0 with 1, and those smaller than 0 with -1.
   Example:

   ```
   >>> a = np.array([[1,2],[-4,0]])
   >>> myFun2(a)
   ... array([[1, 1],
              [-1, 0]])
   ```

3. (10pt) Given an arbitrary data matrix ($n \times m$), and a weight 1-D array ($n,$), write a function to return an 1D-array ($m,$) consist of weighted average of each column of data matrix.
   Example:

```
>>> data_matrix = np.array([[1,2,3],[-4,0,3]])
>>> weight = np.array([0.2,0.8])
>>> myFun3(data_matrix, weight)
... array([-3. ,  0.4,  3. ])
```

4. (10pt) Generate two 1D arrays (100,) which consist of 100 normal distributed random numbers by Numpy, write a function which can reshape these two 1D arrays to two 2D arrays with shape (4,25) , then concatenate them to a matrix with shape (8,25).

3 Bonus Question (10pt)

Finish Array Partition I, and Transpose Matrix on Leetcode. Note that you can not use packages like Numpy, Pandas, etc in the solutions for these two questions.

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 and the output as csv format. 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. In your python file, you need contain both function and test part of function. 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.