IS4151/IS5451 – AIoT Solutions and Development

AY 2024/25 Semester 2

Practical Lab 06 – IoT Data Preprocessing

Part 1 – Basic Programming

PE06-1-1 – Basic Data Importing and Understanding with Pandas

This question is based on the occupation dataset by Kevin Markham hosted here – https://github.com/justmarkham/DAT8/blob/master/data/u.user. The actual data file is provided as occupation.csv.

You may refer to the Pandas Documentation's API Reference here – https://pandas.pydata.org/pandas-docs/stable/reference/index.html

Perform each of the following tasks and report the answer, if applicable:

a) Create a new Jupyter notebook with the file extension .ipynb and import the necessary libraries.

b) Import the occupation dataset from the occupation.csv file provided as a DataFrame with user_id as the index and assign it to a suitably named variable.

c) Print out the first 25 rows.

d) Print out the last 10 rows.

e) What is the number of observations in the dataset?

f) What is the number of columns in the dataset?

g) Print out the name of all the columns.

h) How is the dataset indexed?

i) What is the data type of each column?

j) Print out only the occupation column.

k) How many different occupations there are in this dataset?

l) Print out the list of users aged 50 and above. How many of such users are there?

m) Print out the descriptive statistics of the DataFrame.

n) Print out the descriptive statistics for all columns in the DataFrame.

o) Print out the descriptive statistics only for the occupation column.

p) What is the mean age of users?

q) What is the age with least occurrence?

r) Add a new salary column to the DataFrame with an initial value of 0.

s) Set the salary of each user to 100 multiplied by the user's age. For example, the salary for observation 1 would be 24 * 100 = 2400.

Print out the DataFrame to show the computed salary for all users.

Part 2 – Advanced Programming

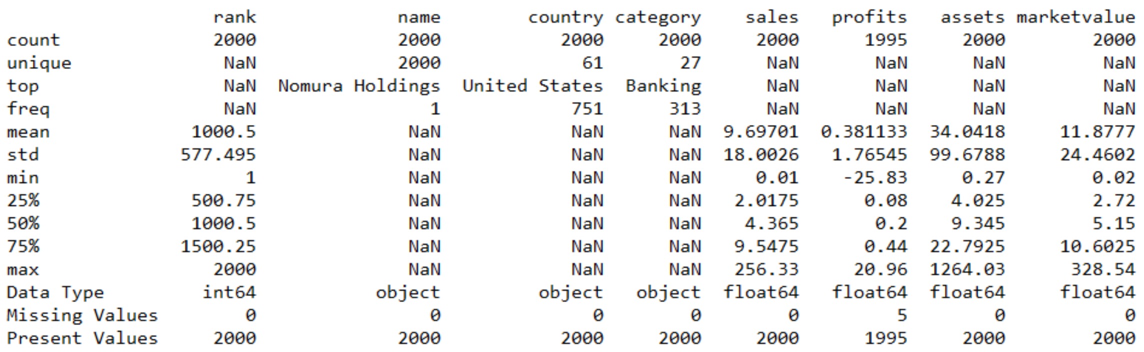
PE06-2-1 –Advanced Data Preprocessing and Visualisation

The data file Forbes2000.csv contains the Forbes 2000 list for the year 2004, i.e., the list of 2,000 world leading companies in 2004 collected by Forbes Magazine. Each of the 2,000 observations contains eight variables:

Variable	**Description**
rank	The ranking of the company.
name	The name of the company.
country	The country the company is situated in.
category	A category describing the products the company produces.
sales	The amount of sales of the company in billion US dollars.
profits	The profit of the company in billion US dollars.
assets	The assets of the company in billion US dollars.
marketvalue	The market value of the company in billion US dollars.

Perform each of the following tasks and report the answer:

a) Generate a data quality report as shown in the figure below:



```
                  rank             name       country category   sales  profits  assets marketvalue
count             2000             2000          2000     2000    2000     1995    2000        2000
unique             NaN             2000            61       27     NaN      NaN     NaN         NaN
top                NaN  Nomura Holdings United States  Banking     NaN      NaN     NaN         NaN
freq               NaN                1           751      313     NaN      NaN     NaN         NaN
mean            1000.5              NaN           NaN      NaN 9.69701 0.381133 34.0418     11.8777
std            577.495              NaN           NaN      NaN 18.0026  1.76545 99.6788     24.4602
min                  1              NaN           NaN      NaN    0.01   -25.83    0.27        0.02
25%             500.75              NaN           NaN      NaN  2.0175     0.08   4.025        2.72
50%             1000.5              NaN           NaN      NaN   4.365      0.2   9.345        5.15
75%            1500.25              NaN           NaN      NaN  9.5475     0.44 22.7925     10.6025
max               2000              NaN           NaN      NaN  256.33    20.96 1264.03      328.54
Data Type        int64           object        object   object float64  float64 float64     float64
Missing Values       0                0             0        0       0        5       0           0
Present Values    2000             2000          2000     2000    2000     1995    2000        2000
```

b) Drop all rows with missing values in the dataset and regenerate the data quality report.

c) Plot a histogram of marketvalue and describe the skewness of marketvalue. Calculate a suitable measure of skewness with Pandas and interpret the result. In particular, state whether the calculated measure is congruent with the histogram.

d) Transform the data with Pandas to resolve the skewness problem in (c). Thereafter, replot the histogram and recalculate the measure of skewness.

e) Seaborn (https://seaborn.pydata.org/) is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. One useful visualization supported by Seaborn is heat map. Seaborn can be installed with the following command using pip:

```
python -m pip install seaborn
```

Generate a heat map for marketvalue using country as the row header and category as the column header.

What information can you extract from this heat map?

f) Suppose we want to predict marketvalue using linear regression analysis. Determine which other numerical variables (i.e., sales, profits and assets) are useful independent variables for predicting marketvalue.

g) Suppose we want to convert the regression problem in (f) into a binary classification problem, i.e., predict whether marketvalue is low or high. Make the necessary changes to the dataset to enable this classification task.