COMP40370 Practical 1

Data Exploration

Prof. Tahar Kechadi

Academic year 2020-2021

Assignment files

- `./practical1.pdf`: this PDF file
- `./specs/AutoMpg_question1.csv`: data file
- `./specs/AutoMpg_question2_a.csv`: data file
- `./specs/AutoMpg_question2_b.csv`: data file
- `./specs/test_practical1.py`: Python test file

Expected output files

- `./run.py`: main Python script
- `./output/question1_out.csv`: data file for first question
- `./output/question2_out.csv`: data file for second question

Requirements

- Python 3.7+
- `pandas` 1.1+
- `numpy` 1.19+

Question 1: Data Cleaning

The file `AutoMpg_question1.csv` contains data related to cars, such as horsepower, weigth, car name, and so on. Unfortunately, some of the values for the *horsepower* and *origin* columns were not properly recorded. Can you tell how many missing values are there for each one of these columns? Write the answer in your report.

1. Replace the missing *horsepower* values with the **average** of this column

2. Replace the missing *origin* values with the **minimum** of this column
3. Save the generated data file to `./output/question1_out.csv`

When saving the generated data, pay extra attention to the columns included in the file (hint: if you are using `pandas`, take a look at the arguments of the `to_csv` function).

Question 2: Data Integration

The files `AutoMpg_question2_a.csv` and `AutoMpg_question2_b.csv` contain similar pieces of information about car models. There are some differences between the 2 files. What you need to do is:

1. The dataset A has an attribute called *car name*, whereas the dataset B has an attribute called *name*. Rename the *name* attribute to *car name* (unintended tongue twister!).
2. The dataset B has an attribute called *other*, which is not present in the dataset A. Create an attribute called *other* in the dataset A, and assign it a default value of 1.
3. Concatenate dataset A and B together, and just like in question 1, save the resulting file to `./output/question2_out.csv`.