RMIT Vietnam University
School of Science and Technology

COSC2789 | Practical Data Science

Assignment 1: Data Cleaning and Summarising

Due: 23:59, Thursday (the 19th, November, 2020) in Week 4
This assignment is worth 30% of your overall mark.

Introduction

In this assignment, you will examine a data file and carry out the first steps of the data science process, including the cleaning and exploring of data. You will need to develop and implement appropriate steps, in IPython, to load a data file into memory, clean, process, and analyse it. This assignment is intended to give you practical experience with the typical first steps of the data science process.

The "Practical Data Science" Canvas contains further announcements and a discussion board for this assignment. Please be sure to check these on a regular basis - it is your responsibility to stay informed with regards to any announcements or changes.

Where to Develop Your Code

You are encouraged to develop and test your code in two environments: Jupyter Notebook (or Jupyter Lab) on Lab PCs or your laptop.

Plagiarism

RMIT University takes plagiarism very seriously. All assignments will be checked with plagiarism-detection software; any student found to have plagiarised will be subject to disciplinary action as described in the course guide. Plagiarism includes submitting code that is not your own or submitting text that is not your own. Allowing others to copy your work is also plagiarism. All plagiarism will be penalised; there are no exceptions and no excuses. More information on Academic Integrity is available at https://www.rmit.edu.vn/students/my-studies/assessment-and-exams/academic-integrity

General Requirements

This section contains information about the general requirements that your assignment must meet. Please read all requirements carefully before you start.
_ You must do the analysis in Python Jupyter Notebook/Jupyter Lab.
_ Please ensure that your submission follows the file naming rules specified in the tasks below. File names are case sensitive, i.e. if it is specified that the file name is gryphon, then that is exactly the file name you should submit; Gryphon, GRYPHON, griffin, and anything else but gryphon will be rejected.

Task 1: Data Preparation (5%)

Have a look at the file bank.csv, which is available in Canvas under the Assignments / Assignment 1 section of the course Canvas.

The Bank Marketing dataset is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The original dataset was created/donated to UCI repository by Moro et al. (2014) (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). A quick description of the attributes is given below.

Input variables:

# bank client data:
1 - age (numeric)
2 - job : type of job (categorical: "admin", "blue-collar", "entrepreneur", "housemaid", "management", "retired", "self-employed", "services", "student", "technician", "unemployed", "unknown")
3 - marital : marital status (categorical: "divorced", "married", "single", "unknown"; note: "divorced" means divorced or widowed)
4 - education (categorical: "basic.4y", "basic.6y", "basic.9y", "high.school", "illiterate", "professional.course", "university.degree", "unknown")
5 - default: has credit in default? (categorical: "no", "yes", "unknown")
6 - housing: has housing loan? (categorical: "no", "yes", "unknown")
7 - loan: has personal loan? (categorical: "no", "yes", "unknown")

# related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: "cellular", "telephone")
9 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
10 - day_of_week: last contact day of the week (categorical: "mon", "tue", "wed", "thu", "fri")
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

# other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: "failure", "nonexistent", "success")

# social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: "yes", "no")

Being a careful data scientist, you know that it is vital to carefully check any available data before starting to analyse it. Your task is to prepare the provided data for analysis. You will start by loading the CSV data from the file (using appropriate pandas functions) and checking whether the loaded data is equivalent to the data in the source CSV file. Then, you need to clean the data by using the knowledge we taught in the lectures. You need to deal with all the potential issues/errors in the data appropriately (such as: typos, extra whitespaces, sanity checks for impossible values, and missing values etc).

Task 2: Data Exploration (10%)

Explore the provided data based on the following steps:

1. Choose 1 column with nominal values, 1 column with ordinal Values, and 1 column with numerical values. (Please try to explore the columns/attributes of potential importance to the analysis, not just a random choice). Then, create a visualization for each of them.
2. Explore the relationships between columns. You need to choose 3 pairs of columns to focus on, and you need to generate 1 visualisation for each pair. Each pair of columns that you choose should address a plausible hypothesis for the data concerned.
3. Build a scatter matrix for all numerical columns. Note, each visualization (graph) should be complete and informative in itself, and should be clear for readers to read and obtain information.

Task 3: Report (5%)

Write your report within the Jupyter notebook using proper format either with HTML or with Jupyter Notebook formatting guide. All the text within the Jupyter Notebook must not be in red colour as the lecturer will use it for marking. Penalties will apply if the report does not satisfy the requirement. Moreover, the quality of the report will be considered, e.g. clarity, grammar mistakes, etc. Remember to clearly cite any sources (including books, research papers, course notes, etc.) that you referred to while designing aspects of your solution.

- Create a heading called “Data Preparation” in your report. Provide a brief explanation of how you addressed the task. For the steps of dealing with the potential issues/errors, please create a sub-section for each type of errors you dealt with (e.g. typos, extra whitespaces, sanity checks for impossible values, and missing values etc), and also explain and justify how you dealt with each kind of errors.

- Create a heading called “Data Exploration" in your report. For each numbered step in Task 2 above, create a sub-section with corresponding numbering.
  - In subsection 1, include all of your graphs from Task 2, Step 1. Under each graph, include a brief explanation of why you chose this graph type(s) to represent the data in a particular column.
  - In subsection 2, include your plots from Task 2, Step 2. With each plot, state the

hypothesis that you are investigating. Then, briefly discuss any interesting relationships (or lack of relationships) that you can observe from your visualisation.

- In subsection 3, present your scatter matrix and analyze what you observe from the graph.

Task 4: Visualisation Dashboard (10%)

Create a visualization dashboard using the provided dataset using Dash framework or similar ones. The visualization dashboard should be able to run on a local machine.

1. There should be at least 2 "meaningful" plots. The plots must be self-explanatory and can provide some insights on the data.
2. The plots should be interactive, which allows users to customise them to some extents.
3. The plots on the visualization dashboard are automatically updated based on the source data. The data and data retrieving code must be accessible for testing this function.

What to Submit, When, and How

The assignment is due at 23:59, Thursday (the 19th, November, 2020) in Week 4.

Assignments submitted after this time will be subject to standard late submission penalties. You need to submit the following files:

- Notebook file containing your python commands for Task 1 and Task, "assignment1.ipynb". Please use the provided solution template to organise your solutions: assignment1_TEMPLATE.ipynb
  - For the notebook files, please make sure to clean them and remove any unnecessary lines of code (cells). Follow these steps before submission:
    - Main menu → Kernel → Restart & Run All
    - Wait till you see the output displayed properly. You should see all the data printed and graphs displayed.

- The dataset and python file(s) of the visualisation dashboard must be included in the submission. A README file must be included with a clear instruction on how to run the visualization dashboard. If it requires any Python library, a "requirements.txt" must be provided with specific version number.

They must be submitted as ONE single zip file, named as your student number (for example, 1234567.zip if your student ID is s1234567). The zip file must be submitted in Canvas: Assignments/Assignment 1. Please do NOT submit other unnecessary files.