Assignments 2: Pandas

Submission deadline: 23:59, Monday, 30 Sep. 2019

Assignment description:

In this assignment, we will use the Pandas package to examine 'Restaurant and consumer' dataset. The dataset is obtained from a recommender system prototype and includes nine files in total: five CSV files from the Restaurant category, one is from the user-item-rating category, another three are from the Consumers category.

- More information about the data set is available at the following link: https://archive.ics.uci.edu/ml/datasets/Restaurant+%26+consumer+data

- The dataset folder, containing all nine CSV files, is available in the assignment folder.

Part 1:

In this part, you will load the datasets, clean them and answer the related questions through writing appropriate codes using Pandas package.

2.1.1 Read the CSV files from all three categories (Restaurant, Rating and Consumers) and store them separately in Pandas data frames.
Note: You can use 'latin-1' encoding method to read the 'geoplaces.csv' file.

2.1. 2 Clean each data frame and provide a brief description of your cleaning tasks for each data frame which required cleaning.

Suggestions:

- Drop duplicate rows if any.
- If there is a row or column where all or most values are missing, drop them.
- Drop ('fax','zip','url') columns from 'geoplaces2.csv'.
- Find the missing values (i.e. '?','None', etc.). First replace them with *nan* in Numpy and then fill in them with appropriate values (methods).
- Fill in missing values with the appropriate methods. For example, for some numeric values, you can calculate the mean or mode value of that column. For string values, it is your choice to find a reasonable way. For example, finding correlation with other columns, etc. (Refer to Lecture 9.)
- Find the words with the same meaning and different spellings (e.g. san luis potosi and San Luis Potosi), then replace them with only one unique format.

In the next step, provide answers to the following questions through writing appropriate queries on Pandas data frames:

2.1.3. What are the names of different restaurants in the state of ‘tamaulipas’?

2.1.4. How many different customers used public transport for going to the restaurants?

2.1.5. What is the least popular payment method among customers?

2.1.6. How many (different) restaurants work until 19:00 in the evenings?

2.1.7. Which type of cooking practice (rcuisine) is the most common among restaurants?

2.1.8. What is the percentage of customers who were born between 1980 and 1990?

2.1.9. What is the percentage of students with a medium budget preferring walking to the restaurants?

Part 2:

In this part, in order to answer the questions, you will require to merge two or multiple datasets depending on the question requirements.

2.2.1. What are the names of restaurants that do not have public parking lots?

2.2.2. What are the addresses of restaurants which only accept ‘cash’?

2.2.3. Name the cities where the restaurants cook and serve ‘fast food’?

2.2.4. What is the most common ‘rating’ among customers with family?

2.2.5. What types of ‘rcuisine’ received the highest rank from people with ‘low’ budget?

2.2.6. What is the average of ‘service rating’, received from ‘social drinkers’ about restaurants which just served ‘Wine-Beer’?

2.2.7. How many smokers gave zero ‘service rating’ to the restaurants without an open area?

2.2.8. Find the correlation between different rating categories ((general) rating, food_rating, service_rating) with the price levels of the restaurants.

Hint: To answer the last question, you need to change the type of attribute values from categorical to numerical.

Grading:

• The result of grading will be either 'Passed' or 'Failed'. Your grade will be assessed on the performance of your code. The accepted performance means that the cleaning decisions should be reasonable and at least 80 percent of queries should be written correctly and should generate meaningful answers.

Note: As the cleaning process is to some extent personalized, there may be no unique correct answers for some questions.

Submission guideline:

• Deadline: 23:59, Monday, Sep. 30, 2019 (submit your assignment through Canvas).

• Source code submitted for the assignment should be your own code. If you use the codes from the internet or if you use someone's code without referencing, that will be treated as plagiarism and you will fail the assignment.

• Source code should be written in one Jupyter notebook file. Filename should be the student's first name, last name and name of the assignment (firstname.lastname.Pandas).

• The assignment is individual and can NOT be solved in groups.

• The deadline cannot be extended.

• If you have any questions, please send an email to this address: aida.mehdipourpirbazari@uis.no.