IS5312 Analytical Programming with Python
Project Description

Updated on October 17, 2022

1 Project Objectives

In this project, you will practice manipulating data files, processing data, conducting exploratory data analysis, and making predictions based on data. This project has two objectives:

1. By conducting this project, students can review and comprehensively practice most Python programming skills learned from this course:
    - Numbers and variables
    - Input and output
    - Relational operations
    - Strings
    - List and tuple
    - Set and dictionary
    - If-else control flow
    - For and while loop control flow
    - File processing
    - Functions
    - Module/Package
    - Class
    - Numpy
    - Pandas
    - Visualization/Matplotlib
2. We also include tasks a little beyond the above listed points but still with reasonable difficulty level, specifically, Task 3.5. The rationale is that when programming, it is very usual to come across new problems you have never seen before, especially considering the rapid development of programming technology and tools. Hence, it is necessary to train students to solve new problems creatively. Task 3.5 is designed to induce students to train their creative problem solving skills when facing new programming tasks. With the help of references from books, papers, and Internet, students can solve these tasks successfully.

2 Data Description

This project has two data files.

1. The first data file named `cadata.txt` is on California housing prices in 1990 (Kelley Pace and Barry, 1997). This dataset includes variables such as the population, median income, and median house value for each block group in California. Block groups are the smallest geographical unit for which the US Census Bureau publishes sample data, and each block typically has between 600 to 3000 people. For each row after the brief introduction, this file contains variables in this order[1]:

   (a) block group ID
   (b) median house value
   (c) median income
   (d) housing median age
   (e) total rooms
   (f) total bedrooms
   (g) population
   (h) households
   (i) latitude
   (j) longitude

   These values are stored with the scientific notation.

2. The second data file named `ocean_proximity.csv` describes how far the houses in these block groups are from the ocean (adapted from Sepulveda (2022)). There are five types of labels, namely `<1H Ocean`, `Inland`, `Island`, `Near Bay`, and `Near Ocean`. For each row except for the first row, this file contains variables:

   (a) block group ID
   (b) ocean proximity

3 Tasks

This project consists of five sequentially connected tasks. Each task is based on the result of the preceding task and lays foundation for the succeeding task, as shown in Figure 1. Each task counts 20% of the total score. You need to write Python programs to finish these tasks and manual manipulations do not count in your score.

For those tasks labeled by **Optional**, students can freely decide to do or not. The optional tasks do not count in the total score. But successfully finishing the optional tasks will obtain a grade bonus of 10% for each task.

[1] We add the variable `block group ID` to both data files so that students can use such IDs to combine entries in this file with those in the second data file.

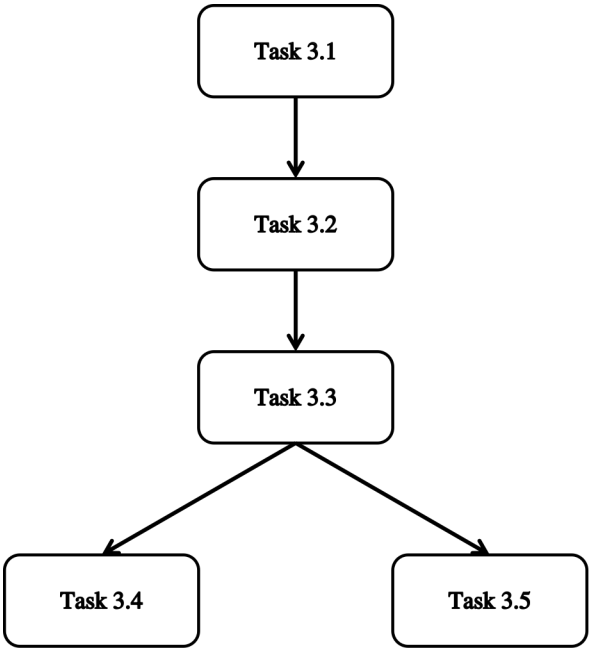

Figure 1: Connection between Tasks

3.1 Read data from txt file (20%)

1. (6%) Please write code to read data stored in the file named `cadata.txt`.
2. (8%) Please delete the brief introduction content at the first several lines of the file and only keep those data for the variables.
3. (8%) Please store the remained data into a list with each row being an element.

3.2 Pre-processing data and store new data into CSV file (20%)

1. (5%) In the file `cadata.txt`, for each line, variables are separated by `Tabs`. Now please replace `Tabs` with commas for each line. As a result, the data in each line will become comma-separated.
2. (5%) Please create a new CSV file named `cadata.csv`, with the first line being the names of these variables in the file `cadata.csv`. These variable names are separated by commas as all. For the names, please see the data description of the first file named `cadata.txt`.
3. (5%) Please convert the data read from `cadata.txt` from the scientific notation to normal numbers or fractions with four decimals after the point. For example, the number in the scientific notation `3.2200000000000e+002` should be converted to `322.0000`.
4. (5%) Please store these lines into file `cadata.csv`. The lines in the file `cadata.csv` should be corresponding to the lines in `cadata.txt` except for the brief introduction part.

3.3 Combining data from two sources (20%)

1. (6%) Please read data from both files named `cadata.csv` and `ocean_proximity.csv`.
2. (8%) Please combine the two datasets based on matching the variable `block group ID`. This means all variables with the same `block group ID` are combined together.
3. (6%) Please store the combined dataset into a new CSV file named `data_for_analysis.csv`. The first line of the new CSV file contains variable names. In each line, data or names are separated by commas.

3.4 Exploratory data analysis (30%)

1. (7%) For variables `median house value`, `median income`, and `households`, conduct the descriptive statistics. In detail, the `mean`, the `variance`, the `mode`, and the `skew` of those variables are calculated.
2. (8%) For the three variables `housing median age`, `total rooms`, and `population`, please draw a histogram for each variable. These histograms are to show how the values distribute among their value intervals. Hint: please refer to the Python package `matplotlib`. Here is a tutorial link from W3Schools: `matplotlib histogram tutorial`.
3. (8%) Please draw scatter plots to preliminarily explore the relationship between variables and `median house value`. Specifically, you should draw these scatter plots:
   - `median income` and `median house value`
   - `total rooms` and `median house value`
   - `population` and `median house value`

   Hint: please refer to the Python package `matplotlib`. Here is a tutorial link from W3Schools: `matplotlib scatter tutorial`.
4. (7%) The above results should be shown in table in pretty print style.

3.5 Partitioning data and predicting housing prices (10%+20% optional)

1. (10%) Partitioning data set into train data set and test data set. To be objective, systematic sampling should be used when partitioning data. The train data set should be about 80% of all data points and the test data set should be 20% of them.
2. **(Optional, 10%)** Students should program to predict the `median house value` based on other variables in the file `data_for_analysis.csv`. Please construct the prediction model as a function/functions and call the function(s) when make prediction. The prediction model can be `linear regression`, `logistic regression` or any other feasible one. The students are required to utilize at lease one self-constructed model for prediction.

3. **(Optional, 10%)** Evaluating the prediction results. The student should compare the prediction value and the real value, and then calculate the `mean squared error (MSE)` of the prediction. Here is the formula of `MSE`:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(f(\boldsymbol{x}_i) - y_i)^2,$$

where m is the number of cases in the test set, $f(\boldsymbol{x}_i)$ is the predicted value of case i in the test set, and y_i is the real value of `median house value` of case i in the test case.
Hint: you may want to seek help from various online resources on Python packages regarding to regressions.

4 Submission Files

To obtain scores of the project, you need to submit these files:

1. A `.py` file containing all your source codes by the order of tasks. Or A Jupyter file containing all your source codes by the order of tasks. To name this file, please follow this format: `studentName_StudentNumber_project_code.py`. For example, if you are CHAN Wai Ting and your student No. is 55664332, then your submitted source code file should be named as `CHANWaiTing_55664332_project_code.py`. Please put all source codes into **one** file for the ease of grading.

2. The two CSV files generated during Tasks 3.2 and 3.3.

3. A report on the results of exploratory data analysis (Task 3.4) and `median house value` prediction (Task 3.5), with four pages at most. To name this file, please follow this format: `studentName_StudentNumber_project_report.pdf`. For example, if you are CHAN Wai Ting and your student No. is 55664332, then your submitted report file should be named as `CHANWaiTing_55664332_project_report.pdf`.

References

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, May 1997. ISSN 0167-7152. doi: 10.1016/S0167-7152(96)00140-X. URL https://www.sciencedirect.com/science/article/pii/S016771529600140X.

Daniel Sepulveda. Housing. *Kaggle*, September 2022. URL https://www.kaggle.com/datasets/kathuman/housing.