IE 496: Independent Studies
Eduardo Lopez Tasks for Jan 20–Feb 17

Instructor: Anirudh Subramanyam

January 19, 2023
Updated: February 3, 2023

1 Introduction

This project aims to provide you with an understanding of the role computation can play in data science and engineering. This understanding shall be aided by specific examples in field of transportation engineering. Specifically, the project will enable you to

1. code basic computational machine learning models
2. analyze the output of such models to make inferences
3. understand some fundamental steps involved in computational research
4. excel in subjects with programming components

The project is designed assuming that you have little or no programming experience At the end of the semester, you should feel confident of your ability to write computer programs that allow you to accomplish the above goals. We will use the Python programming language.

1.1 Task structure

Tasks will be split in time periods of roughly 4 weeks each. Each task will involve some combination of:

- A reading assignment
- A programming assignment
- A presentation assignment, based on the reading and programming tasks: this will be due after roughly 2 weeks

1.2 References

Instructions for downloading all relevant material will be provided along with the weekly tasks. Contact the instructor if you have trouble accessing any of the material.

2 Tasks for Weeks 1–4

Complete the tasks in this order. If you are finished with these tasks, let the instructor know so we can chart out next steps.

1. [Reading] Read Chapter 1 of the book [1]
   (a) Understand the concepts of 'algorithm', 'flowchart' and 'programming language'
   (b) Understand the consequences of programming errors
2. [Reading] Read the introduction to Chapter 2 of the book [1]
3. [Programming] Follow the instructions in Section 2.1 of [1] to install Python on your computer
4. [Programming] Play with Python:
   (a) Complete all sections of 'Learn the Basics' until 'Loops' in this website: https://www.learnpython.org/
   (b) This material is also covered in the remainder of Chapter 2 of the textbook, but the above website allows you to run code within your web browser
5. [Reading] Read this page about travel forecasting: Big Picture
6. [Reading] Read this page about travel forecasting: Travel Behavior
7. [Reading] Read Chapter 1 of this PDF: NHTS-2017 Dataset
8. [Programming] Load the NHTS-2017 dataset in Python
   (a) Download the CSV survey data from this link: https://nhts.ornl.gov/downloads
   (b) Follow instructions in Section 23.1 of the textbook [1] to load the file: 'vehpub.csv'
   (c) Complete both sections of 'Data Science Tutorials' in this website: https://www.learnpython.org/
9. [Reading] Read this page about regression analysis: Regression Analysis
10. [Presentation (Due in 2 weeks)] Make a roughly 10-minute powerpoint presentation summarizing the following:
    (a) What is the basic premise of travel forecasting?
    (b) What are some common assumptions in travel behavior modeling?
    (c) What are some applications of the NHTS Survey dataset?
    (d) The basic idea of regression analysis and multiple regression
    (e) Among all the data columns in the dataset for individual vehicles ('vehpub.csv'), what sets of variables would it make sense to do regression analysis?

2.1 Tasks added on February 3, 2023

1. [Reading] Read Chapter 5 of the book 'Learning Statistics with Python'.
   - You can skip Section 5.3 and Section 5.5
   - As you are working your way through the chapter examples, you may want to directly experiment with data columns of the 'vehpub.csv' file, since analyzing this data will be part of your programming tasks described next. Pay particular attention to the `VEHAGE`, `OD_READ`, `GSTOTCST` and `FEGEMPG` columns in this file.
2. [Programming] For each of the four variables mentioned above:
   (a) Calculate the mean, median, and mode
   (b) Calculate the interquartile range, mean absolute deviation, and standard deviation
   (c) Calculate the correlation coefficient between all pairs of variables (see Section 5.6.4)

   You may use either functions from the `statistics` module or the `pandas` function `describe()` for the first two tasks. Both approaches are described in Chapter 5.
3. [Reading] Read Chapter 6 of the book 'Learning Statistics with Python'.
   - You can skip Sections 6.3.3, 6.3.4 and 6.5
4. [Programming] Plot side-by-side histograms of `GSTOTCST` for:
   (a) the entire U.S. dataset versus only PA data (filter using column ID `HHSTATE`)
   (b) `VEHTYPE` $= 1, 2$ (fossil fuel) versus `VEHTYPE` $= 3$ (hybrid/electric vehicles)
   (c) Urban versus rural house holds (filter using column ID `URBRUR`)
5. [Programming] A scatter matrix is a type of plot that combines histograms and scatter plots in a single graphic. Make a scatter matrix of each of the 4 data columns mentioned above.
6. [Presentation (Due on Feb 17)] Make a roughly 15-minute presentation summarizing your results from the programming tasks. In particular, answer the following research questions:
   (a) Based on the results in task 2(a), were there any differences in the mean and median of any variable? What conclusions can be drawn based on the values of these differences?
   (b) Based on the results in task 2(b), provide an intuitive interpretation of the interquartile ranges and standard deviation you obtained for each of the variables.
   (c) What conclusions can you draw based on the correlations you obtained in task 2(c)?
   (d) What conclusions can you draw based on the histograms you obtained in task 4?
   (e) What conclusions can you draw based on the scatter matrix you obtained in task 5?

2.2 Data analysis tips

1. The meaning of any column can be found in the NHTS Codebook. Press `cmd + F` to quickly search for particular keywords.
2. Pay particular attention to "special entries" like -8 which stands for "I don't know". They are essentially equivalent to missing entries. Make sure to ignore these special entries from your `pandas` dataframe when calculating descriptive statistics or making plots.

2.3 Programming Tips

1. If you don't understand something, google it. A very useful internet forum for all programming-related concepts (including Python) is https://stackoverflow.com/
2. You can also ask questions on Slack - you will hear from the instructor or one of the grad students
3. Check the Style guide for Python
4. The ultimate resource for any programming language is its official documentation

2.4 Presentation Tips

1. Use as many pictures/graphs/plots as possible – humans are visual learners and concepts are best communicated graphically
2. Convey only ONE message per slide

References

[1] John V. Guttag. *Introduction to Computation and Programming Using Python, third edition: With Application to Computational Modeling.* MIT Press, London, England, 2021.