

AML-2203

Advanced Python AI and ML Tools

Assignment

(20% of Final grade)

|  |  |
| --- | --- |
| **Instructor:** | Vahid Hadavi, PhD, P.Eng |
| **Class:** | **AML-2203** |

# Description

Dear Students,

This assignment requires more individual learning than previous assignments in the program. Overall, you are encouraged to check out the [pandas documentation](http://pandas.pydata.org/pandas-docs/stable/) to find functions or methods you might not have used yet, or search/ask questions on [Stack Overflow](http://stackoverflow.com/) and tag them as pandas and python related (we do that quite often at work!). And of course, the discussion forums are open for interaction with your peers and the course instructor.

Please read the following guidelines carefully and make sure to specify the steps mentioned below in your final submissions (ipynb file and pdf report, etc). Overall, understanding the questions is an important skillset which will help you a lot in your future career (as we usually emphasize in the class). Read the questions a couple of times and discuss them with your group members. However, feel free to reach out to your instructor if there are important doubts.

In the future, your employer will ask you to do something very similar. Hence, students are expected to build a complete, deliverable product and present it to the client in a professional manner. The overall product is judged in a competitive setting. Your final mark depends on the quality of your project as well as the quality of your project-report.

You are encouraged to create a free account on GitHub and update your profile to look as professional  
as possible, and push your project code to the GitHub repository. While this is optional, it’s highly  
recommended as in your future; your GitHub profile plays a significant role in your employment  
opportunities. Many organizations will look at your Git profile before they hire you!

**Assignment description and guidelines:**

Step1: for this assignment you can provide your own data. The data must be pulled/scraped from online resources (as covered in other courses) to obtain full credit of this step. For example, you can scrape tweeter data or Kijiji data etc.

Step 2: proceed with initial steps of data wrangling and properly name/label/explain the activities in this (and all other) steps. Use the data you provided in step 1 and apply data wrangling steps including but not limited to data cleaning, formatting and structuring, validating etc.

Step 3: use proper plotting methods to demonstrate the distribution of various attributed (columns) in the data.

Step 4: pandas profiling is one of the useful tools you can use for exploratory data analysis (EDA) in future projects at work! Learning pandas profiling is a quite helpful for your future career. Please use the following link and use pandas profiling and generate a report in .html format (submit that along with your codes and pdf report)

https://pandas-profiling.ydata.ai/docs/master/pages/getting\_started/overview.html

Step 5: use proper encoding methods and create new numerical columns that represent the categorical column in the original data (and/or vice versa)

Step 6: use proper techniques (boxplots AND IQR) to identify the outliers on each column (attribute) and explain your approach/results.

Step 7: use all the following techniques to address the outliers and explain/discuss the differences in the outcome of each of the methods

* + Quantile-based Flooring and Capping
  + Trimming
  + Log Transformation

Step 8: Use the unsupervised learning method of your choice (e.g: K-Means ) on the data you have for this assignment and discuss the outcomes.

Step 9 (Bonus mark): if your input includes txt data, please apply proper NLP techniques (data cleaning, stop word/punctuation removal, tokenizing, stemming/lemmatizing etc) and bonus mark might be granted.

Deliverables and submission instructions for this assignment:

* This is a group assignment and you should do it in teams of 4-5 ppl.
* Only one person per team should submit the deliverables (same person should submit all the following parts:

1. The python code in ipynb format including proper markdowns and explanations (submit into Moodle)
2. The source data (this can be shared using your college one drive and sharing that with your instructor)
3. Pandas profiling report in .html format (submit through Moodle or One drive)
4. Complete professional report including graphs/tables etc and proper explanation of observations and conclusion (preferably ‘avoid’ including lots of code snapshot, please). Your pdf report should be under 15 pages and use college cover page and make sure to include the complete name and student number of all team members on the front page.