MONASH University

Faculty of Information Technology
Summer Semester B, 2020

FIT5145 Introduction to Data Science

Assignment 1 – Exploratory Data Analysis in Python

Submission:	Submission requirements discussed below.
Value:	The assignment is worth 20% of the total marks for the unit.
Due Date:	Sunday 19 January 2020 at 11:59 pm.

Exploratory data analysis (EDA) is one of the most important skills for a data scientist. This assessment aims to guide you in your investigation of a data set via EDA, primarily through visualisation of that data using Python's pandas. You will need to draw on what you have learnt and will continue to learn, in class. You are also encouraged to seek out alternative information from reputable sources. If you use or are 'inspired' by any source code from one of these sources, you must reference this. There is an example at the bottom of this section.

Learning outcomes

You will learn the following by completing this assessment:
1. Read in files and extract data from them into a data frame.
2. Wrangle and process data.
3. Use graphical and non-graphical tools to perform EDA.
4. Use basic tools for managing and processing big data.
5. Determine informati9on
6. Communicate your findings in your report.

Task

In this course, you have learned that the diversity of definitions, skill sets, tools, applications and knowledge domains, can make data science challenging to define. By completing this EDA, we hope you can get a clearer understanding of how a career in data science compares to others in the IT industry.

The Data

In late 2018, a survey was conducted for a large Australian collective of IT professionals. The survey received 7000 responses, and the resulting dataset was made public. We have heavily modified the data set, both to clean it for analysis and to ensure original assignment submission. The data is stored in the file called ***assignment1_dataset.csv***. Each column contains the answers of one respondent to a specific question. Please do not alter the dataset.

Submission

You will need to submit two files, a Jupyter notebook and a PDF of your Jupyter notebook containing your answers (code, figures and answers to all the questions). Make sure to include your visualisations to justify your answers to all the questions.

When printing your code to PDF, make sure that it isn't cut off due it being out of view in the notebook. The best way to do this is to properly format your code so that a horizontal scroll bar is not required. Marks will be assigned to PDFs based presentation, as well as correctness and clarity of your answers and code. The PDF should not contain an excessive number of pages. You should not print the data frames in your PDF (comment out the code that prints these). Screenshots of codes are not acceptable.

You are required to submit your assignment online through Moodle. This will automatically generate a Turnitin report (usually within 2 hours). You can review your assignment and revise it to address any issues raised by the Turnitin report or any other issue you need to address to improve your assignment. You can upload a number of draft versions of your assignment. Once you are happy with your assignment, you should upload it and then press the submit button. Moodle will then require you to agree to an online statement that is the equivalent of the coversheet.

Please do not change any of the directions or answer boxes, the order of questions, order of code entry cells or the name of the input files, in your notebook. Actual Python code must be present in your Jupyter notebook file. This code must match what you have submitted in the PDF of your Jupyter notebook.

General notes

Late submission

Late submissions will attract a deduction of **5% per day**, including weekends and public holidays. You will be unable to submit your assignment for marking after seven days.

Special consideration

Applications for an extension must be submitted and approved by your tutor at least three days before the due date. All requests for extensions must be made in writing (email application is acceptable) and provide legitimate reasons for the extension. Where appropriate, supporting material should also be given. Your tutor could request a draft of your assignment to consider your progress when they address your application. You must negotiate any extensions formally via the in-semester special consideration process:http://www.monash.edu/__data/assets/pdf_file/0006/277152/in-semester.pdf

Please note that requests for extensions on or after the due date will only be considered in exceptional circumstances and are you should apply directly to your lecturer in writing, making sure to cc your tutor.

Zip file submission

Zip file submission will attract a **penalty of 10%**. Do not submit separate files together in one zip file or another compressed format.

Drafts (not submitted)

Please make sure to ***submit*** your assignments and not leave them draft mode. We cannot accept assignments that are not yet submitted, and you may attract a late penalty.

Academic integrity

Monash University takes plagiarism and collusion extremely seriously. You must adhere to the Monash academic integrity policy in completing your assignment. If you are in doubt, please read the policy or see your tutor:
https://www.monash.edu/students/admin/policies/academic-integrity

Turnitin

All assignments must be submitted to Turnitin. As always, there is no minimum or maximum Turnitin score; these are all relative and are individually evaluated by your tutor. Your turn it in score will be higher than normal due to the notebook containing questions and instructions. Please do not be concerned about this. Sources that are poorly paraphrased, not referenced or not made clear are quotes are treated as direct plagiarism.

The best way to avoid this is to make sure your references are correct, and in APA, you have paraphrased and formatted your work appropriately. Make sure that you comply with academic requirements relating to citation of all your sources. Your references must adhere to the FIT Style Guide for citations. The guide can be accessed at http://www.monash.edu/it/current-students/resources-and-support/style-guide If you are unsure of how to properly reference your sources, please consult this or the Monash APA guide https://guides.lib.monash.edu/citing-referencing/apa, your tutor, your lecturer or the Research and Learning skills advisors in the library https://www.monash.edu/library/skills

How to cite from forums like stack overflow

You must attribute any code you have referred to. Without this, close similarity will appear in Turnitin and be treated as plagiarism.

@Root. (Jan 16, 2013). *Answer to question: Import CSV file as a pandas Data frame*. Retrieved from: https://stackoverflow.com/questions/14365542/import-csv-file-as-a-pandas-dataframe. Date accessed: Dec 12, 2019.

@username of the user whose code you adapted. (month day, year of the answer). *Answer to question: The title of the original question*. URL of thread. Date accessed: month day, year of when you accessed the post.

Marking allocation

You will only be able to achieve an HD if you attempt all questions. There are many ways to construct your code. We are looking for the most efficient and concise. However, your tutor will accept all reasonable answers, i.e., there is no one **correct** answer.

This rubric is to be considered as a guide. Your tutor is not bound by the rubric and may allocate marks based on the completeness of each criterion in respect to other criteria in that range. This is their prerogative. You will receive a letter grade (HD, D, C, P or N) and not a numerical grade.

Task — Exploratory Data Analysis

Weighing (90%) [135 marks]	You are currently working in this range		
	Unacceptable	Needs improvement	Excellent
Accuracy of code The code provides the correct output meeting all possible specifications of the question.	• Many errors. • Unanswered questions. • The output is difficult to interpret.	• Some errors. • Incomplete explanation of the output. • Minor errors in output formatting.	• No errors and all outputs are correct. • Well formatted output.
Quality of the code The code is both space and time efficient and does not use unnecessarily longwinded methods to achieve the aim.	• Missing or incomplete code. • Errors not handled properly. • Code does not compile. • Code is not efficient.	• Most of the code works as expected. • Mostly reasonable choices in methods. • Efforts to write efficient code. • Used depreciated methods.	• All code works as expected. • Efficient and clear choices in methods. • No depreciate methods used.
Visualisation The visualisations are informative, well laid out and correctly labelled.	• Misleading and inaccurate. • Poorly formatted, not clear. • No labelling. • Poor scaling.	• Some minor inaccuracies in code formatting. • Efforts made to provide format visualisation (axes, legends, labels etc.)	• Correctly and clearly shows information. • Labelled axes, legend and title. • Visually appealing with efforts made to highlight certain aspects.
Answers to questions The questions have been answered with reference to the external context of the data, visualisations, and the output of the code.	• No or incorrect answers. • No or little justification.	• Mostly Correct answers. • Basic justification.	• Correct answer. • Strong justification. • Answers refer to the code output with logical and clear explanations.
Documentation The code uses appropriately named variables and informative commenting.	• Poor use of variable naming conventions. • Few or no comments used. • Poor spacing.	• Variables are mostly well named. • Few or no comments used. • Commenting notparticuarly informative.	• Excellent use of variable naming conventions. • Inline comments are clear, concise and explain the functions and packages appropriately..

General

Weighing (10%) **[15 marks]**	**You are currently working in this range**		
	Unacceptable	**Needs improvement**	**Excellent**
Referencing All external sources are correctly referenced in APA.	• No, or few references where necessary. AUTOMATIC FAIL	• Complete references. • Poorly or inconsistently formatted. • References not in APA.	• Excellent references following APA with links to code-based references and placed in the markdown near the relevant code.
Presentation of PDF The notebook and PDF are presented in a logical and consistent format.	• Poor syntax. • Grammatical and spelling errors. • Code was cut off. • An excessive number of pages. • Confusing layout.	• Logical layout. • No cut off code. • Good use of space. • Some spelling or grammatical errors. • Attempts to be concise.	• Layout makes for easy reading. • No cut off code. • Good use of space. • Little of no spelling or grammatical errors.

Grade:

Feedback: