Manchester Metropolitan University

Department of Computing and Mathematics

ASSIGNMENT COVER SHEET

Unit code and title:	6G7V0026 Principles of Data Science
Assignment set by:	Luciano Gerber
Assignment ID: *(e.g. 1CWK50)*	1CWK100
Assignment weighting:	100%
Assignment title:	Data Science Project
Type: (Group/Individual)	Individual
Hand-in deadline:	Fri, 13 Jan 23 21:00
Hand-in format and mechanism:	Moodle; PDF, Jupyter Notebook

Learning outcomes being assessed:

LO1. Prepare data for use in a data science scenario.
LO2. Gain insights from data and communicate findings effectively to stakeholders using suitable reporting tools and techniques.
LO3. Develop solutions to real-world data science tasks.
LO4. Apply a wide range of transferable skills and attributes applicable to industry and research.

Note: it is your responsibility to make sure that your work is complete and available for marking by the deadline. Make sure that you have followed the submission instructions carefully, and your work is submitted in the correct format, using the correct hand-in mechanism (e.g., Moodle upload). If submitting via Moodle, you are advised to check your work after upload, to make sure it has uploaded properly. Do not alter your work after the deadline. You should make at least one full backup copy of your work.

Penalties for late hand-in: see Assessment Regulations for Undergraduate/Postgraduate Programmes of Study on the Student Life web pages. The timeliness of submissions is strictly monitored and enforced.

All coursework has a late submission window of 7 days (i.e., 5 working days), but any work submitted within the late window will be capped at 50%, unless you have an agreed extension. Work submitted after the 7-day late window will be capped at zero, unless you have an agreed extension. See below for further information on extensions.

Please note that individual tutors are unable to grant extensions to coursework.

Extensions: For most coursework assessments, you can request a 7-day extension through Moodle. This will not apply to in-class tests, presentations, interviews, etc., that take place at a specific time. If you need a longer extension you can apply for an Evidenced Extension. For an Evidenced Extension you **MUST** submit 3rd party evidence of the condition or situation which has negatively impacted on your ability to submit or perform in an assessment.

Plagiarism: Plagiarism is the unacknowledged representation of another person's work, or use of their ideas, as one's own. Manchester Metropolitan University takes care to detect plagiarism, employs plagiarism detection software, and imposes severe penalties, as outlined in the Student Handbook (http://www.mmu.ac.uk/academic/casqe/regulations/docs/policies_regulations.pdf and Regulations for Postgraduate Programmes (https://www.mmu.ac.uk/academic/casqe/regulations/assessment/docs/pg-regs.pdf). Bad referencing or submitting the wrong assignment may still be treated as plagiarism. If in doubt, seek advice from your tutor.

If you are unable to upload your work to Moodle: If you have problems submitting your work through Moodle you can email it to the Assessment Team's Contingency Submission Inbox using the email address submit@mmu.ac.uk. You should say in your email which unit the work is for, and ideally provide the name of the Unit Leader. The Assessment team will then forward your work to the appropriate person. If you use this submission method, your work must be emailed by the published deadline, or it will be logged as a late submission. Alternatively, you can save your work to your university OneDrive and submit a Word document to Moodle which includes a link to the folder. It is your responsibility to make sure you share the OneDrive folder with the Unit Leader.

As part of a plagiarism check, you may be asked to attend a meeting with the Unit Leader, or another member of the unit delivery team, where you will be asked to explain your work (e.g., explain the code in a programming assignment). If you are called to one of these meetings, it is very important that you attend.

Assessment Criteria:	Indicated in the attached assignment specification.
Formative Feedback:	Formative feedback on provisional work will be given primarily in lab sessions.
Summative Feedback Format:	Summative feedback is provided with a breakdown of marks for the assessment criteria and a sample solution.

Coursework Specification (1CWK100)

6G7V0026 Principles of Data Science

Luciano Gerber

22/23, Semester 1

Overview

In this assignment students have the opportunity to consolidate the learning in the Principles of Data Science unit by creating a **working solution** for a **practical data science problem** that includes typical tasks in data **understanding**, **exploration**, and **preparation**, and in identifying/hypothesising interesting **associations** and **group differences** in the data.

More specifically, you will work with a **Car Sale Adverts dataset** provided by **AutoTrader**, one of our industry partners. The dataset contains an anonymised collection of adverts with information on vehicles such as brand, type, colour, mileage, as well as the **selling price**. You are asked to perform a structured set of tasks with the ultimate goal of learning about associations and group differences that have a significant effect on the valuation of vehicles.

You are expected to produce, as **deliverable**, a **short, structured report** containing fragments of **code** and **explanations** for addressing the set of data science tasks on the dataset. In addition, you are asked to hand in a **fully-documented**, **reproducible Python notebook** with your **rough work**.

The **assessment criteria** of code/explanations such as involves aspects such as usefulness, correctness, clarity, conciseness, among others.

Obtaining and Exploring the Dataset

Please note the following **license information** before downloading and using the dataset:

- *this dataset provided by AutoTrader is to be used only by students enrolled on 6G7V0026 (22/23). Students can place it in their personal storage spaces (e.g., MMU OneDrive) for carrying out their practical data science and machine learning work (e.g., with Jupyter Lab). Please do not redistribute the dataset.*

It is available in our Moodle area here as a `.csv` file, and it has around 400K rows. For reducing computational time, specially during prototyping/experimental phases, you might want:

- Avoid plotting a large number of individual data points; you could turn to heatmaps and hexbins instead of scatterplots, for example.

- Take/work with a small sample of the dataset.

Data Science Tasks to Perform

The work on this assignment is driven by two main, broad data science questions:

1. What are the **best predictors** of the **price** of a vehicle? In other words, of the **individual** and **interactions of pairs of features** in the dataset, what are those that seem to have to strongest association with the feature `price`, and what are the explanations and insights behind those findings?
2. What are **interesting** groupings of the data, involving one or two features, that show significant differences (e.g., trends, averages) in `price`? What can we learn from them and how useful could these findings be for the business?

In order to address the above, you need to implement and run the following tasks that are typical of a Data Science pipeline, using our usual Data Science environment (i.e., Python ecosystem/notebooks):

1. **Data/Domain Understanding and Exploration** (e.g., load the data, sample observations, check correct parsing of data, identify quantitative and qualitative features, analyse data distributions (e.g., range, centrality, dispersion, shape))
2. **Data Processing** (e.g., detect and deal with noise (i.e., erroneous values), missing values, and outliers; subset, reshape, and engineer features improved analysis)
3. **Association and Group Differences Analysis** (e.g., based on correlations, conditioning, group differences)

Marking Criteria

The **assessment criteria** is based on the University's PGT Assessment Criteria and stepped marking, and includes aspects of code/explanations such as :

- clarity, conciseness, style, correctness
- usefulness, challenge, creativity, initiative
- efficiency, reusability, and generality

Report Components and Weights

Please structure your report with the components below including, for each, a small but representative and interesting portion of your work in completing the tasks above. These should consist of a relevant fragment of code, a corresponding output (e.g., table, plot), and a brief explanation (one or two paragraphs) of findings. Next to each subcomponent below you will see a range (e.g., for *Dealing with Missing Values, Outliers, and Noise*, it is *1-2*) which indicates the suggest number of chosen examples to report on.

Component	Weight
1. **Data Understanding and Exploration**	**30%**
1.1. Meaning and Type of Features; Analysis of Distributions (3-4)	10%
1.2. Identification/Commenting on Missing Values (2-3)	10%
1.3. Identification/Commenting on Outliers and Noise (1-2)	10%
2. **Data Processing**	**30%**
2.1. Dealing with Missing Values, Outliers, and Noise (1-2)	10%
2.2. Feature Engineering, Data Transformations (2-3)	10%
2.3. Subsetting (e.g., Feature Selection, Data Sampling) (1-2)	10%
3. **Association and Group Differences Analysis**	**40%**
3.1. Quantitative-Quantitative (1-2)	13%
3.2. Quantitative-Categorical (1-2)	13%
3.3. Categorical-Categorical (1-2)	14%

Submission

Your submission should be a `.zip` file containing (1) a PDF for your report and (2) a `.ipynb` **documented**, **reproducible Python notebook** as **rough work**. There is no need to submit the dataset, but it is expected that one is be able to reproduce your work on the copy of the dataset available on Moodle. You will find the submission link in our unit's Moodle area. Please be careful **not to leave it to the last minute**; internet issues and similar are unlikely to count as exceptional factors.

How to Pass This Assignment

One pathway (narrow/deep) to obtaining at least a pass in the assignment would be to focus on and do a reasonable job on components 1 and 2. Say that you obtain 72% for (1), 62% for (2), and 42% for (3) - that would result in a combined mark of 57%.

Another (broad/shallow) is to have a genuine, but limited attempt at each component. One possible scenario would be 62% for (1), 52% for (2), and 52% for (3) - that would give you 55%.

Regulations and Code of Conduct

Please make sure you are familiar with the Taught Postgraduate Assessment Regulations. As mentioned in the induction week, the pass mark is 50%, and one has a single reassessment opportunity (capped at 50%). If there are mitigating factors, please visit and follow the guidance at Exceptional Factors, such as that on self-certification.

Importantly, please also make sure you are fully aware of the regulations on Academic Misconduct, particularly if this is your first experience with the UK's higher-education system.

One important aspect to highlight is that this is an individual assignment, and the **submission has to be your own**. It is absolutely fine for people to work in groups and collaborate; in fact, this is something that I encourage. It is also fine to be inspired by existing code snippets created by colleagues, contributors at StackOverflow, and the like. But, importantly, you have to **own it** in the end. That is, **you must be able to explain, customise, and apply it** in a similar, but separate context; also, **cite/reference** the sources.

A couple of scenarios of when things **are not fine** are:

- *the representation of another person's work, without acknowledgement of the source, as one's own*: Say, student A wishes to help and shares their solution to the assignment with student B. The latter submits what was shared as their own work (fully or partially, identically or with little modification). This characterises **collusion** and would implicate both A and B.
- *the use of third parties and/or websites to attempt to buy assessments or answers to questions set*: this characterises **contract cheating**.

All cases of Academic Misconduct (e.g., plagiarism) will be reported to and investigated by Student Case Management Team.