Applied Data Science: S1 Coursework

Data Recording, Cleaning, Analysis and Visualisation.

When undertaking the data gathering exercise, please make sure you are suitably dressed for the weather conditions and remain well clear of the traffic.

This is *fieldwork*, not the typical office environment of a developer. Do not do anything that puts you, your partner or others at risk.

In this lab you will gather vehicle arrival and departure data for the main entrance to the university campus. You will then use this and other data to generate an infographic.

As a data scientist you may not always be provided with ready-made datasets. This lab exercise will require you to consider how to collect data efficiently and effectively.

You will design a **data collection form** for gathering your own data. After you have used your form to collect some data, you will convert the data into a csv format and submit this on Moodle.

You will then wrangle the uploaded data (from all of your peers) into a single dataset for analysis. In addition to the data recorded this semester, you will also be given the data recorded during a similar exercise last year.

By considering all the data available to you, your final task will be to construct an informative infographic that conveys a clear message.

The primary output of this coursework is a single informative infographic that adheres to effective infographic design principles along with a brief justification of some additional data gathering.

In support of this you will also submit the data you've collected and the forms and processing scripts you've created. The detailed requirements are given in this document.

You should work in pairs for Task 1 and the collection of data in task 2. For the remainder of task 2 and for task 3 you should work individually.

Deadlines

Task 1 & 2:	**Friday 27th October 8pm.**
Task 3:	**Friday 24th Nov 8pm**.
Peer Review Forms:	**Friday 15th December 8pm**

Submission via Moodle.

Task 1 (22 marks)

In this task you will need to consider effective approaches for manually recording data.

Your task is to design a **data capture form** that will allow you to manually gather “traffic” data.

Your data collection form must allow you to record the required information (below) **accurately, comprehensively and efficiently**.

Recording data **accurately** means errors introduced by the collection process are minimised.

Recording data **comprehensively** means that there should not be gaps in the data during the recording period.

Recording data **efficiently** means that your form should be suitable for recording activities in real-time, without requiring excessive or unnecessary effort. For example, the form should avoid unnecessary writing and repetition. It is also not efficient to require a user to enter information that they would have to ‘look up’ or calculate each time they enter data. You should consider using visual marks / tick boxes to speed up data collection and use table rows/columns effectively in the form to help to organise data and minimise effort for data entry.

You are expected to design a **<u>paper form</u>** that will be more efficient to use than entering data directly into the .csv file that you will upload. The data files that a data scientist uses for analysis (e.g. csv files) are not always effective as data collection forms. You should therefore attempt to optimise your form to allow the user to record data quickly and easily.

Requirements for data collection form:

- The data you collect must be **fully anonymised** at the point of data collection. Specifically, you cannot record any registration number/identification.

- Each activity record must include:
  - The date and time at which each vehicle arrived or left (to nearest 5 mins).
  - The category of vehicle.
  - The occupancy of the vehicle.
  - The direction of travel.

- The data collection process may require more than one sheet but should not be unduly wasteful.

- You should specify an interesting research/analysis question that provides motivation and justification for one or more additional pieces of data that you can collect. You will need to clearly state the research/analysis question that you have identified and provide some justification for why it is a worthwhile question to explore. You should consider the practicality of collecting this data, such that the data collection you propose can feasibly be undertaken as part of the collection process.

You should submit a short text document that explains **(in maximum of 500 words)**, the research question(s) you considered, justifies why it is worthwhile to explore and which explains the types of data that you propose that could help to answer this question. You must provide references if appropriate (references will not count towards the maximum 500 words).

You may work in pairs to design the research question and design the form. You should both submit this work. Any instructions required for completing the form should be included as part of the form.

Your data collection form for Task 1 will be assessed according to the following criteria:

Assessment Criteria for Task 1 Marks should be awarded for each of the following criteria that are satisfied. There are no partial marks. Either award the full mark for each question or zero.	**Marks Awarded**
Does the design of the data collection form allow the user to record **ALL** of the required information (as per the requirements set out in Task 1).	5
Is the design of the data collection form optimised for efficient, accurate and comprehensive data collection? The mark will not be awarded if clear ways in which the data collection could be improved are identified.	4
Has a research/analysis questions been specified, which motivates the collection of additional data not listed in the requirements? AND Is there a convincing explanation for why these are worthwhile questions to explore?	5
Is the additional data to be collected appropriate for answering the research question(s) stated?	4
Has an effective method for capturing the additional data been incorporated into the design of the form AND is the data feasible/practical to collect for a person completing the form?	4

Task 2 (23 marks)

You must now use your form to collect **a 30 minute** period of arrivals or departures from the main campus. You should work in pairs. With one person recording traffic arriving and the other recording traffic leaving campus during the same period.

You should do this some time between 8am and 11am in the week of Monday the 24rd of October to Friday 27th October. (Please don't all do it at 10.30am on Friday.)

Each pair should collect data at the same time whenever possible. Once you have collected the data you should work independently on the remainder of the coursework.

Once you have collected the data using your form, you will be required to transfer data from your form into a .csv file. The .csv must match the column format specified in this. You are allowed to rename the last two columns (AdditionalData1, AdditionalData2) to reflect the additional data you have gathered.

```
Date, Time, Direction, Type, Occupancy, AdditonalData1, AdditionalData2
20-10-2022, 8.05am, in, car, 1, null, null
20-10-2022, 9.00am, out, bus, 10%, null, null
20-10-2022, 10.10am, in, bicycle, 1, null, null
```

Figure 1: Format of csv_example.csv

```
car, bus, bicycle, motorbike, taxi, van, lorry, scooter
```

Figure 2: Vehicle types

Figures 1 and 2 show expected content of the traffic data csv file.

Completing the csv data entry manually may be burdensome, therefore you are permitted/encouraged to perform post-processing on your csv file if it will make completing the csv more efficient. For example, rather than fully, manually completing data entry on all columns, you should consider whether data can be added in some columns using a Python script. You will need to determine for yourself which parts of the csv file may benefit from post-processing to increase efficiency of data entry. You should submit the notebook that you create for post-processing and complete the csv files.

Your completed data collection form and .csv file for Task 2 will be assessed according to the following criteria:

Assessment Criteria for Task 2 **A mark should be awarded for each of the following criteria that are satisfied.**	Marks Awarded
Does the uploaded .csv data file conform to the format of the example .csv provided?	6
Does the uploaded .csv data file include all of the data that was recorded in the completed data collection form?	5
Does the .csv data file contain data for an approximately 30 minute period during the collection window.	7
Has the data provided in the completed data collection form been fully anonymised? I.e., it should not be possible to identify any individual vehicle from the information provided.	5

If you have a DAP in place and are unable to collect data, please contact the lecturer.

Task 1 + 2 Submission

You must submit both Task 1 and Task 2 via Moodle by the same deadline.

You must submit:

1. A blank version of the data collection form that you have designed for Task 1, such that it could be used by your peers for recording their own data. Please ensure that the data collection form is a pdf document. Any instructions should be included as part of the form. You should name the file that you upload '**blank_form.pdf**'.
2. You must also submit a completed version of your data collection form, which you have recorded traffic for 30 minutes. The photocopiers on campus allow the scanning of documents to pdf. You should name the file that you upload '**completed_form.pdf**'.
3. A text document (.txt or .pdf) that explains in a maximum of 500 words, the research question(s) you considered for additional data collection, justifies why these are worthwhile question to explore, providing references if appropriate, and explains the types of data that could be collected to help answer this question. You should name the file that you upload **'additional_data.txt'** or **'additional_data.pdf'**.
4. You must also submit a .csv file containing all of the data that you have recorded, also fully anonymised. This must adhere to the format described in this document. You must name your file '**complete_data.csv**'.
5. You should submit any **Jupyter Notebook** that supports any post-processing of the csv file, which makes entering data into the csv file more efficient. This file should be called **'process.iypnb'**. When uploading a notebook, you should also upload any unprocessed version of the csv file called '**raw_data.csv'.**

Please upload these as separate files. Do not include them in a .zip. **Please use the filenames indicated.** Failing to name the files as specified may result in a ten mark penalty.

Task 3 (45 marks)

After the Task 1 + 2 deadline, the individual datasets that you have each provided will be made available in a single download on Moodle. You can then use the data that has been collected and submitted by all students to conduct an analysis and create a single infographic that conveys an interesting finding. You will also be given the same data gathered by students last year.

You will be expected to compile the numerous individual .csv files that you are provided with into a single dataset. You will be expected to either use a Jupyter Notebook to integrate the datasets and dealt with any issues that arise, e.g. differences in date/time formats. And/or do any manual editing required.

You should formulate a question about the data and then use appropriate analysis and visualisation techniques to attempt to answer this. This does not need to be the same questions as in Task 1. Any Jupyter Notebook should include comments that explain the analysis you are performing and provide outputs such as data visualisations that achieve your analysis results.

From the analysis that you have performed, you should decide on an interesting or compelling message that you want to convey about the data and produce a single infographic that conveys this message clearly. Your infographic should contain visualisation(s), integrate text effectively and conform to appropriate infographic design principles that have been discussed in this unit.

Your infographic should reference at least one external data source allowing the reader to compare and contrast a finding from the dataset that you have analysed. You must cite the source you use in your infographic.

Your Jupyter Notebook and infographic for Task 3 will be assessed according to the following criteria:

Assessment Criteria for Task 3	Marks Awarded
Does the code process the majority of the 2023 .csv files provided, combining the individual datasets selected into a single dataset for analysis?	5
Does the code/markup clearly demonstrate how you have chosen to deal with any data wrangling issues (e.g. missing data, incompatible date/time formats)?	5
Are these issues dealt with sensibly? (e.g. ensuring that different date/time formats are resolved, returning error messages or handling exceptions if incompatible .csv files are provided)?	5
Does the Jupyter Notebook contain comments/markup that clearly explains questions or hypotheses about the data, which were used to inform the analysis?	4
Does the code demonstrate that appropriate analysis steps have been taken in order to answer questions or test hypotheses?	4
Does the infographic convey a clear and informative message?	4
Do the visualisations included in the infographic have a clear purpose that relates to the message?	4
Does the infographic follow the effective design principles discussed within this unit?	6

Does the infographic include at least one visualisation that allows the reader to compare the results from the two years of collected data.	4
Is the additional source integrated effectively into the infographic to support the comparison AND is it properly referenced.	4

Task 3 Submission

You must submit via Moodle:

1. A Jupyter Notebook that contains code for analysing the complete dataset, with outputs such as data visualisations that show the results of your analysis.
2. An infographic in .pdf format.

You are free to name these files as you wish but must submit them separately and not within an archive file such as zip. Do not upload more than one of each.

Peer Review (10 marks)

The lecturers will assess your work according to the criteria given in this specification. This will account for 90% of the mark awarded.

This unit will also make use of peer review. This means that after the deadline for the coursework you will be allocated the work of three other students to review and assign a mark. This will allow you to see how others have tackled the same problem. The purpose of this is to expose you to issues you may not have identified for yourself and improve your understanding of the problem being tackled.

You will be provided details of how to download the three submissions. You are expected to examine these and compare them to the assessment specification given in this document. Each of the criteria is designed to be a pass/fail decision where the submission either meets the requirement or it does not. Where any criteria are not met, you should indicate why you have reached this conclusion.

You will be given a pdf form that you can use to submit an entry for each submission you examine. You must submit all the forms by the peer assessment deadline.

All work you submit for this coursework should be anonymous and not include your name or userid.

To assist you in this exercise, one lecture will be used to present some examples and explain their assessed mark.

Peer Assessment Mark Calculation

The marks your peers give you will not affect your mark for this coursework. Or vice versa. Instead, the 10% of your mark awarded for peer review will be based on how close your assessment is to the mark assigned by the lecturers.

To gain these marks your reviews must be complete. A complete review entry means you will have completed a form for each submission allocated to you and provided a justification for each of the criteria you have labelled as not met. The mark awarded will be based on the sum of the absolute difference in marks of each of the three items reviewed, as shown in Figure 3. This means the absolute difference of each of the three marks will be added together not that the three assigned marks will be added together.

Sum of absolute differences	Mark assigned.
< = 5	10
<= 10	8
< = 15	6
<= 20	4
<= 25	2
30 +	0

Figure 3: Marks for Peer Review

Extensions

Only the Director of Studies can grant extensions in receipt of a completed request form. The unit lecturers are not able to grant extensions.

Appealing your Mark

If you disagree with the mark awarded, you can email the lecturers to ask for a review. The request must be specific. This means that you need to indicate each mark you dispute and justify why you consider it should have been awarded with respect to the criteria specified in this document. Vague demands for more marks for effort will be rejected without review.

KMC 2023