Applied Data Science Lab 3: Time Diary Data Capture, Analysis and Reporting

Deadlines

Task 1 + Task 2	Friday 25th February 8pm
Peer Assessment for Task 1+2	Friday 11th March 8pm
Task 3	Friday 1st April 8pm
Peer Assessment for Task 3	Tuesday 3rd May 8pm
Marks	/20 (20% of overall unit mark)

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 'self-report data', efficiently and effectively.

Many researchers are interested in analysing and understanding patterns of human activity and behaviour. To achieve this they often use "time diaries", sequential and comprehensive accounts of a person's daily life, documenting their time spent on different activities. This is considered to be one of the most reliable and accurate data collection instruments to understand activity patterns of populations and to reveal the relationships between activities and other factors such as health, wellbeing, energy usage, productivity, culture, etc.

You will design a **data collection form** for gathering your own "time diary" data, documenting how you spend your time over the course of three (or more) days. After you have used your form to collect data, you will convert the data into an anonymised .csv format and submit this on Moodle.

You will then wrangle all of the uploaded data (from all of your peers) into a single dataset for analysis. You will be required to devise interesting questions, which will guide your exploration and analysis of patterns, trends and insights in the data. You will then report the results of your analysis using an infographic that conveys a clear and interesting message.

For example, what can the data tell us about how our activities vary on different days of the week? How varied are the working patterns of Data Science students? What activities do we tend to find most or least enjoyable? Does the amount of sleep we get have an impact on our activities throughout the day?

You will produce a blank version of the data collection form, a completed version of the form containing data, as well as .csv file containing the data, Jupyter Notebooks with Python code for data wrangling, analysis and visualisation, as well as a single informative infographic that adheres to effective infographic design principles and conveys a clear and interesting message about what the data reveals.

Task 1 (5 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 "time diary" data. The time diary data should capture information about how you spend your time on different activities throughout at least **three** days (24hr periods from midnight to midnight). These do not need to be consecutive days.

The data capture form that you design can be a digital document (e.g. PDF, spreadsheet, text file etc.), a printable or hand-written worksheet, or any other form that you deem appropriate. You must design your data capture form such that a blank version can be shared with your peers, that would allow them to use it to capture data of their own.

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

Recording data **accurately** means that all activities lasting 15 mins or longer should be documented and logged to the nearest 15 minute time intervals. Activities should have all required labels and meta-data provided (see requirements below).

Recording data **comprehensively** means that there should not be significant gaps in the data throughout each day, for at least three days.

Recording data **efficiently** means that your form should be suitable for recording activities throughout the day, 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' 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 form 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 for allowing the user to record data quickly and easily.

Further requirements for data collection form:

- The data you collect must be **fully anonymised** at the point of data collection.
- Each activity record must include:
    - The date and time at which each activity started and ended (to nearest 15 mins).
    - Assignment to one of the following primary activity codes (and an optional secondary activity code of your choice):
        - Eating / Drinking (Code: ED145)
        - Education / Lectures (Code: EL642)
        - Exercise (Code: X893)
        - Housework (Code: H179)
        - Leisure (Code: L418)
        - Reading (Code: R523)
        - Sleep (Code: S801)
        - Travelling (Code: T695)
        - Using Devices (Code: UD415)
        - Paid Work (Code: PW101)
        - Coursework (Code: CW982)
        - Other (Code: O733)
    - An *optional* short name or descriptive label (e.g. "Reading a book", "Watching Television", "Using mobile", "Travelling to shop")
- The activities each day should include a subjective 'Enjoyment' rating score (-2 = very unenjoyable, -1 = somewhat unenjoyable, 0 = neither enjoyable nor unenjoyable, 1 = somewhat enjoyable, 2 = very enjoyable).

You must also identify two pieces of additional data that can be recorded alongside some or all activities. You should specify one or more interesting research/analysis question(s) that provide motivation and justification for the two additional pieces of data that you would collect. You will need to clearly state the research/analysis question(s) 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 could feasibly be undertaken by yourself or one of your peers, and such that obtaining the data would not require excessive effort.

For example:
Prior research (e.g. Barton et al., 2020)[1] has investigated the relationship between listening to music and people's *productivity* during work activities (such as software development), as well as the relationship between the genres of music that people listen to during different activities (e.g. exercise vs. work). Understanding these relationships through data may help to identify ways to boost productivity and inform automated/intelligent playlist generation for music services, respectively. Therefore, one might propose recording additional data on *productivity* and whether music is being listened to during an activity, and if so the *genre of music*.

Your proposal for two pieces of additional data should differ from the example given.

You should submit a short text document (e.g. .txt, .rtf, .doc, .pdf) that explains **(in maximum of 300 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 can provide references if appropriate (references will not count towards the maximum 300 words).

You must design your data collection form independently. Do not work with others to design your form. 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 **A mark should be awarded for each of the following criteria that are satisfied.**	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).	1
Is the design of the data collection form optimised for efficient, accurate and comprehensive data collection? When peer marking, you should attempt to use the form to capture data yourself in order to see how well it supports efficient, accurate and comprehensive recording of your activities. You should not award the mark if you can identify any clear ways in which the data collection could be improved. You should discuss these in your peer marking feedback.	1
Have research/analysis questions been specified, which motivate the collection of additional data not listed in the requirements? AND Is there a convincing explanation for why these are worthwhile questions to explore?	1
Are the two pieces of additional data to be collected appropriate for answering the research question(s) stated?	1
Have effective methods for capturing the additional pieces of data been incorporated into the design of the form AND are these pieces of data feasible/practical to collect for a person completing the form?	1

[1] L. Barton, G. Candan, T. Fritz, T. Zimmermann and G. C. Murphy, "The Sound of Software Development: Music Listening Among Software Engineers," in *IEEE Software*, vol. 37, no. 2, pp. 78-85, March-April 2020, doi: 10.1109/MS.2019.2906312.

Task 2 (5 marks)

You must now use **your own form** to collect **at least three days** of anonymised data, documenting your activities. If your data has significant gaps during any of the three days for which you have collected data, you should attempt to 'backfill' the data before submission, with a best estimate of how you spent your time.

This data **must not** contain any personally identifiable information. Please note that the data that you submit will only be used for this coursework assignment.

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 in the **'csv_example.csv'** file provided on Moodle. You are allowed to rename the last two columns (AdditionalData1, AdditionalData2) to reflect the additional data you have gathered.

PrimaryActivityCode	SecondaryActivityCode	DescriptiveLabel	StartDate	UniBathWeekNo	DayOfWeek	StartTime	EndTime	DurationMins	EnjoymentScore	AdditionalData1	AdditionalData2
S801		Sleeping	17/02/21	21	Wednesday	23:15	08:00	525			
O733		Getting dressed	18/02/21	21	Thursday	08:00	08:15	15	0		
ED145		Breakfast	18/02/21	21	Thursday	08:15	09:00	45	1		
UD415	L418	Browsing web on laptop	18/02/21	21	Friday	09:00	09:15	15	1		
EL642	UD415	Online lecture	18/02/21	21	Thursday	09:15	10:15	60	-1		

Figure 1: Format of csv_example.csv

Completing all of the csv data entry manually may be burdensome, therefore you are expected to perform some post-processing on your csv file in order to 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 any scripts that you create for post-processing and completing 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?	1
Does the uploaded .csv data file include all of the data that was recorded in the completed data collection form?	1
Does the .csv data file contain data for three or more days AND are there limited gaps in the data for each of these days? (Specifically, no more than 20% of each 24hr period should be unaccounted for).	1
Has the data provided in the completed data collection form been fully anonymised? i.e. it should not be possible to identify any individual from the information provided.	1
Does the Jupyter Notebook clearly demonstrate that post-processing of the .csv file was performed AND Is the post-processing effective in reducing effort of manually entering all data into the .csv file?	1

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 provided in an accessible and common file format (e.g. .pdf, .doc, .rtf, .txt, . xls). Any instructions should be included as part of the form. You should name the file that you upload '**blank_form**'.

2. You must also submit a completed version of your data collection form, which you have used to record your activities for three or more days. This must be fully anonymised. You should name the file that you upload '**completed_form**'. Please ensure that the completed data collection form is provided in an accessible and common file format.
3. A text document (e.g. .txt, .rtf, .doc, .pdf) that explains (maximum of 300 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'**.
4. You must also submit a .csv file containing all of the time diary data that you have recorded, also fully anonymised. This must adhere to the column format of the 'csv_example.csv' file provided on Moodle. You must name your file '**complete_data.csv**'.
5. You should submit a **Jupyter Notebook** called **process.ipynb** file that supports any post-processing of the csv file, which makes entering data into the csv file more efficient.

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

Task 3 (10 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 of the time diary data and create a single infographic that conveys some interesting findings. You will also be given the same data gathered by the previous years students.

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 use a Jupyter Notebook to integrate the datasets and dealt with any issues that arise, e.g. differences in date/time formats, missing values, excluding any personally identifiable information.

You should formulate a set of questions about the data and then use appropriate analysis and visualisation techniques to attempt to answer these questions. Your 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 various analyses 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 multiple visualisations, integrate text effectively and conform to appropriate infographic design principles that have been discussed on this unit.

Your infographic should include at least one visualisation that allows the reader to compare and contrast a finding from the dataset that you have analysed, with a finding from a different publicly available time diary dataset, such as those available at:

https://ec.europa.eu/eurostat/data/database?node_code=tus_00age and
https://www.kaggle.com/bls/american-time-use-survey

You must cite the source of the dataset that 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 all of the .csv files provided, combining all of the individual datasets into a single dataset for analysis AND Does the code/markup clearly demonstrate how you have chosen to deal with any data wrangling issues (e.g. missing data, unrecognised activity codes, incompatible date/time formats). AND 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).	2 (No partial credit)
Does the Jupyter Notebook contain comments/markup that clearly explains questions or hypotheses about the data, which were used to inform the analysis?	1
Does the code demonstrate that appropriate analysis steps have been taken in order to answer these questions or test hypotheses?	1
Does the infographic convey a clear and informative message? AND Do the visualisations included in the infographic have a clear purpose that relates to the message?	1
Does the infographic follow the effective design principles discussed within this unit?	3
Does the infographic include at least one visualisation that allows the reader to compare and contrast findings from a different publicly available time diary dataset? AND Is the source of the data cited in the infographic?	1
Is the data from the additional dataset appropriate for the comparison being made? AND is the additional data integrated effectively into the infographic to support the comparison?	1

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 submit them separately and not within an archive file such as zip.

Peer Assessment

This unit will make use of peer assessment. This means that after the initial deadline for a piece of coursework you will be allocated the work of three other students to examine 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 simple pass/fail assessment where the submission either meets the requirement or it does not. Where any criteria are not met, you must indicate why you have reached this conclusion.

You will be given a link to an online form where you can submit an entry for each submission you examine. You should also submit an entry for your own work. You are strongly recommended to assess your own work after you have reviewed the work of the other students. You must submit all the forms by the peer assessment deadline.

There are no additional marks for completing the peer assessment. However, a penalty of up to 50% will be applied to your lab mark should you fail to complete the peer assessment satisfactorily.

A satisfactory assessment entry means you will have completed a form for each submission allocated to you and provided a valid justification for each of the criteria you have labelled as not met.

The work you submit should be anonymous and not include your name or userid. You should remove any reference to your username in any pathname in your code. Replace it with 'username'. You must not engage in discussion of your mark or the marks you will allocate to your peers with your peers. You should report any attempt by others to influence the marking process.

Mark Calculation

Your mark will be calculated in the following way:

1. The two closest peer marks given will be used. If the three marks are equally spaced, the pair closest to your own estimate will be used.

2. If your own mark estimate lies above the peer marks you will receive the mean of the peer marks.

3. Otherwise, you will receive the mean of the two highest marks. (The two peer marks and you own estimate.)

Your mark will be returned to you once this processing has been done. You will also receive the details of the marks allocated by your peers. This will include their reasoning. This is a provisional mark. If you do not consider the mark to be fair, you can contact the lecturer and ask for it to be reviewed. Your work will be re-marked and where the lecturer determines a different mark, the peer marking will be checked and any unsatisfactory marking will have the penalty applied. Should your request for a review not be justified, a penalty may be applied to your mark as you will have further demonstrated that you have not properly understood the material or the feedback you have received.

After the review period the coursework mark will be finalised. To maximise your marks, you should attempt to be as accurate in the marking of both the peer work and your own.

Extensions

If any student is granted an extension, they will still have to undertake peer marking of others work after their updated deadline, with appropriate extensions. Their own work may be peer marked or assessed by the lecturers/tutors depending on the availability of peer markers at that time.

SJ/KMC 2021