COMP60711 - DATA ENGINEERING
Coursework Description

Coursework Description

1. Data Description

Each student has access to two traffic data files in csv format, **'rawpvr_2018-02-01_28d_1083 TueFri.csv'** and '**rawpvr_2018-02-01_28d_1415 TueFri.csv**'. Each file contains observations collected via Inductive Loops sensors planted on a particular site of Chester Road in the city of Manchester. For example, file **'rawpvr_2018-02-01_28d_1083 TueFri.csv'** contains observations collected from site 1083, while file '**rawpvr_2018-02-01_28d_1415 TueFri.csv**' contains observations collected from site 1415. The observations were collected on all Tuesdays and Fridays of the month of February 2018. Note that each row of the file contains one observation, i.e., the properties associated with one detected vehicle. As a consequence, if you count the total number of records with the following timestamp (06/02/2018), you are able to estimate the total volume of traffic on the 6th of February.

Both data files present the same structure, composed of the following attributes or properties:

- ***Date*** is a timestamp containing day of the month and year and time of the day when a vehicle was detected, with the following format: dd/mm/yyyy HH:MM:SS.
- ***Lane*** is an identifier of a given lane of the road (a road may have multiple lanes and each lane has a unique identifier).
- ***Lane Name*** is the name given to a particular lane of the road. Each lane has a unique name.
- ***Direction*** identifies the direction followed by a road lane (e.g., North, South, etc.). Different lanes may follow the same direction.
- ***Direction Name*** is the name of the direction followed by a lane of the road.
- ***Speed (mph)*** is the speed with which the detected vehicle was moving at the time it was detected.
- ***Headway (s)*** is the time distance between two consecutive vehicles following the same route. More precisely, it is the time distance between the front bumper of one vehicle and the front bumper of the vehicle behind it.
- ***Gap (s)*** is also a time distance between two consecutive vehicles following the same route, but it indicates the time distance between the rear bumper of one vehicle and the front bumper of the vehicle behind it.
- ***Flags*** is a number that identifies the day of the week when a vehicle was detected.
- ***Flag Text*** is the text description of the day of the week when a vehicle was detected.

2. When Developing the Coursework

The coursework is composed of a list of tasks, divided into three sub-lists, as follows:

Sub-list 1: This is the first sub-list and contains a single task, which does not carry any marks, and so, it is NOT MANDATORY. However, you are free to develop it and submit it as a single PDF file.
The main aim of this task is to allow each student to become familiar with the dataset, use case and at least one data preparation tool AND one Programming Language (PL) of his/her choice. We suggest a few tools and a couple of PLs for the student to choose from which are mostly available from the Department's machines and that can be remotely accessed from the comfort of the student's home, for his/her convenience. Please, refer to Section 5 to see the list of suggested Data Preparation tools and PLs. However, we are aware there are other tools and PLs that a student might want to use, and s/he is free to. Note that, for this task, we ask each student to develop two versions of the same solution: one solution developed using a PL, and the second solution, developed using a tool. The deadline for the completion of this task is the end of the first week of the course unit. More specifically, **one hour before** the timetabled **lecture-related** Q&A activity of the following week (Week 2 of the course).

Sub-list 2: This is the second sub-list of tasks and contains three tasks. This sub-list of tasks carries 100 marks and the deadline for its submission is the end of the second week of the course unit. More specifically, **one hour before** the timetabled **lecture-related** Q&A activity of the following week (Week 3) of the course unit. A more detailed description of this task will be made available to you soon.

Sub-list 3: This is the third sub-list of tasks and contains three tasks. This sub-list of tasks carries 100 marks and the deadline for its submission is the end of the second week of the course unit. More specifically, **one hour before** the timetabled **lecture-related** Q&A activity of the following week (Week 4) of the course unit. A more detailed description of this task will be made available to you soon.

When developing the coursework, each student should do the following:

1. Use file '**rawpvr_2018-02-01_28d_1083 TueFri.csv**' for all tasks, unless they are explicitly told in the task to use both files, **'rawpvr_2018-02-01_28d_1083 TueFri.csv'** and '**rawpvr_2018-02-01_28d_1415 TueFri.csv**'.
2. Ideally, aim to choose one Graphical User Interface (GUI) based tool and one Programming Language (PL) to perform all tasks. So, once the tool and PL are chosen, the same tool and PL should be used to perform each of the tasks. The recommended PLs are the following: Python and R; and the recommended GUI-based tools are the following: Knime, OpenRefine, Talend and Microsoft Excel. All of these are available from the machines in the M.Sc. lab (with the exception of Knime), but all can be easily installed on your personal computer.
3. Use the tool/PL you are most familiar with, to avoid steep learning curves.
4. Each week, each student should submit one pdf file containing the solution to the task(s) of the previous week.

3. Each Week's Tasks

Week 1 Task Sub-list 1
Deadline: End of the second week of the course (end of Week 2).

This is a "warm up" task, therefore, there is no discussion session associated with this particular task and it should be developed during Week 1 and the part of Week 2 preceding the week's lecture-related timetabled activity. Since it does not incur marks, it is not a mandatory task, but, if you decided to submit it, then you will receive some rubric-based feedback for it. However, it is an important task, because it allows each student to familiarise him/herself with the data, languages/tools enabling the student to choose his/her preferred languages/tools.

Task 1: (no marks)

Week 2 Tasks Sub-list 2
Three tasks to be developed during the course of Week 2, carrying 100 marks.
Deadline: **one hour before** the timetabled **lecture-related** Q&A activity of the following week (Week 3) of the course unit.
Marking deadline: End of Week 4 of the course unit.

Task 2 (30 marks)
Task 3 (35 marks)
Task 4 (35 marks)

Week 3 Tasks Sub-list 3
Three tasks to be developed during the course of Week 3, carrying 100 marks.
Deadline: **one hour before** the timetabled **lecture-related** Q&A activity of the following week (Week 4) of the course unit.
Marking deadline: End of Week 5 of the course unit.

Task 5 (30 marks)
Task 6 (35 marks)
Task 7 (35 marks)

4. Tasks Description

Week2	
Task2	Provide a simple profile of the hourly traffic volume of the North lanes (considering all the North lanes together) of site 1083 using the following descriptive data summarization measures and focusing only on Tuesdays between 09:00 am and 10:00 am [exclusive]: *Range, 1st Quartile, 2nd Quartile, 3rd Quartile, Interquartile range*. **Output**: Five values, one per required data summarization measure. A step-by-step description of the development of the task, emphasising the features of the tool/language that you used. In other words, **make sure you include:** **(1)** screenshots of intermediate steps you had to carry out while using the technology you chose to use to develop the task (especially where no coding was necessary), **(2)** the final recipe/result, **(3)** an explanation of each step, and **(4)** an interpretation of the results.

	For this task, you should develop one solution, using the technology (tool or PL) of your choice. **Marking scheme:** 5 marks for correct values of all five measures (1 mark for each). 25 marks for a clear, correct and complete step-by-step description, which should include not only the complete code you wrote to prepare the data from its original form to the point the analysis was made, but also an explanation of each step in text and your interpretation of the obtained results/analysis.
Task3	Choose a day of the week, e.g., Tuesday, and use bar plots to visualise the average traffic volume for each hour of the day. To obtain an accurate average traffic volume for a given week day, for example Tuesday, consider all Tuesday records in the file and consider all lanes associated with the North direction, and later (and separately) considering all lanes associated with the South direction. You should generate a separate bar plot for each traffic direction (North and South). **Output**: Two bar plots, one for the North direction and one for South direction, for a weekday of your choice (Tue or Fri). Each bar plot should show the average traffic volume for each hour of the day. A step-by-step description of the development of the task, emphasising the features of the tool/language that you used. In other words, **make sure you include:** **(1)** screenshots of intermediate steps you had to carry out while using the technology you chose to use to develop the task (especially where no coding was necessary), **(2)** the final recipe/result, **(3)** an explanation of each step, and **(4)** an interpretation of the results. **Marking scheme:** 8 marks for correct barplots (5 marks for each). 27 marks for a clear, correct and complete step-by-step description, which should include not only the complete code you wrote to prepare the data from its original form to the point the analysis was made, but also an explanation of each step in text and your interpretation of the obtained results/analysis.
Task4	Develop the same solution for **Task 3** described above using a second technology of your choice (a tool or PL), which must be different to the technology you used to develop **Task 3** for the first time. You can copy and paste here and/or make simple adaptations to the comments and step-by-step description of the solution you provided for Task 3. Considering (i) **Task 3** of this week, described above (which, by now, you should have done) and (ii) the technology you used to develop **Task 4**, compare the two technologies used in the development of these tasks, discussing advantages and disadvantages of each, taking into account not only any extra work (manual or not) you had to do because of absence of features/facilities/functions or limitations of the existing features/facilities/functions associated with the given technology, but also the time it took for you to perform a particular action. **Output**: Two bar plots, one for the North direction and one for South direction, for a weekday of your choice (Tue or Fri). Each bar plot should show the average traffic volume for each hour of the day. A step-by-step description of the development of the task, emphasising the features

	of the technology you used to develop this task. In other words, **make sure you include:** **(1)** screenshots of intermediate steps you had to carry out while using the technology you chose to use to develop the task (especially where coding was not necessary), **(2)** the final recipe/result, **(3)** an explanation of each step A detailed discussion of the two different technologies you used to develop this week's Task 3. **Marking scheme:** 35 marks for the second solution for Task 3, as well as a clear and complete assessment of the two technologies you used to develop this task, which should include functionality (i.e., functions, facilities, features) **similarities, differences, advantages and limitations, all considering to the relevant task**.

Note: When using a GUI-based tool, **make sure you include (1)** screenshots of intermediate steps you had to carry out while using the tool to complete a task (especially where no coding was necessary), **(2)** the final recipe/result, **(3)** an explanation of each step and **(4)** an interpretation of the results.

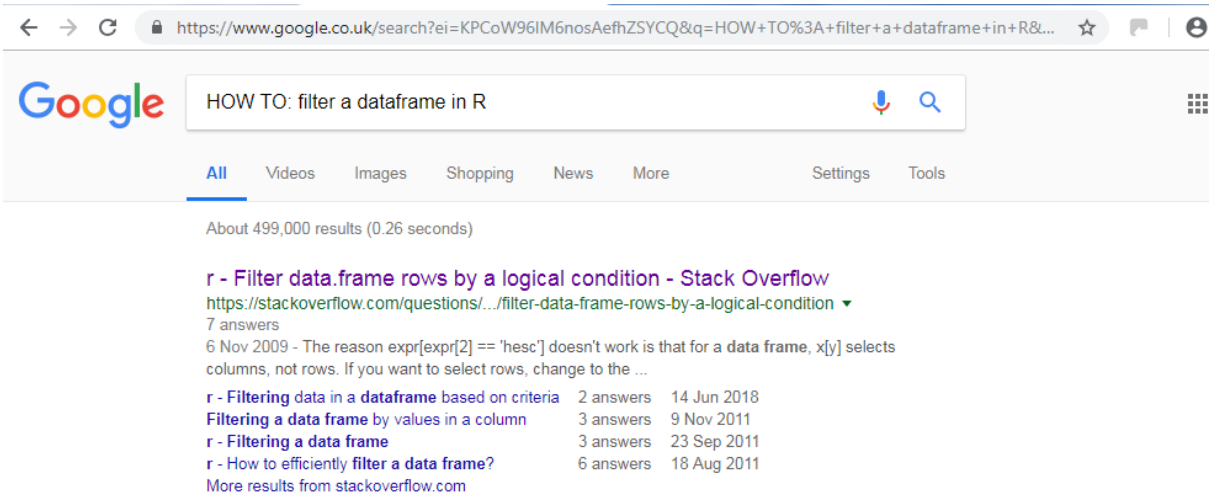
A description of the Weeks 3 tasks will be made available to you soon.

5. Further Advice

The following list contains the PLs and Data Manipulation/Preparation tools that we suggest you use in the development of your coursework: *R, Python, Knime, Excel, OpenRefine and Talend Data Preparation*.

All of these are of easy access and free installation. You can have remote access to them (with the exception of Knime) by connecting to the machines in the Department, or you can just download and install them in your machine (all of them). Note that if you have a Windows machine, Excel will already be in your machine.

If you are not familiar with any of the programming languages and/or tools, then you can search for commands in the Web, as shown below:



If using Python, for example, you will be interested in using commands from packages such as *pandas*, *numpy*, *datetime*, *os* and *calendar*, to handle *Date* related data types. If using R, will be interested in using commands from packages such as *lubridate*, *plyr* and *dplyr*.

General tutorials for each of these can be found from the following links:

- **R**
  An Introduction to R (https://cran.r-project.org/doc/manuals/R-intro.pdf)

- **Knime (https://www.knime.com/downloads/download-knime)**
  Documentation: (https://docs.knime.com/)
  Tutorials: (https://www.youtube.com/watch?v=HEp9Cbql2hs, https://www.youtube.com/watch?v=5WAyOiIfHPg)

- **Talend Data preparation** (you get the free trial software from: (https://www.talend.com/products/data-preparation/?utm_medium=bloglink&utm_source=bloglink&utm_campaign=datapreppage&utm_content=dataprep))
  Introduction to Talend Data Preparation (video) (https://www.youtube.com/watch?v=HmjrnRvJqKU ; https://www.youtube.com/watch?v=lE7P2qG3dB0 ; https://www.youtube.com/watch?v=AAPf3cWoOYk)
  Data Prep 101: Getting Started with Talend Data Preparation (https://www.talend.com/blog/2016/02/10/data-prep-101-getting-started-with-talend-data-preparation/)
  Etc.

- **Excel**
  Excel Tutorials:
  (https://digital.com/excel-tutorials/),
  (https://edu.gcfglobal.org/en/excel2016/),
  (https://trumpexcel.com/learn-excel/)
  Etc.

- **OpenRefine** (https://openrefine.org/download.html)

(https://www.youtube.com/watch?v=WCRexQXYFrI),
(https://www.youtube.com/watch?v=wGVtycv3SS0),
(https://www.youtube.com/watch?v=wfS1qTKFQoI)