

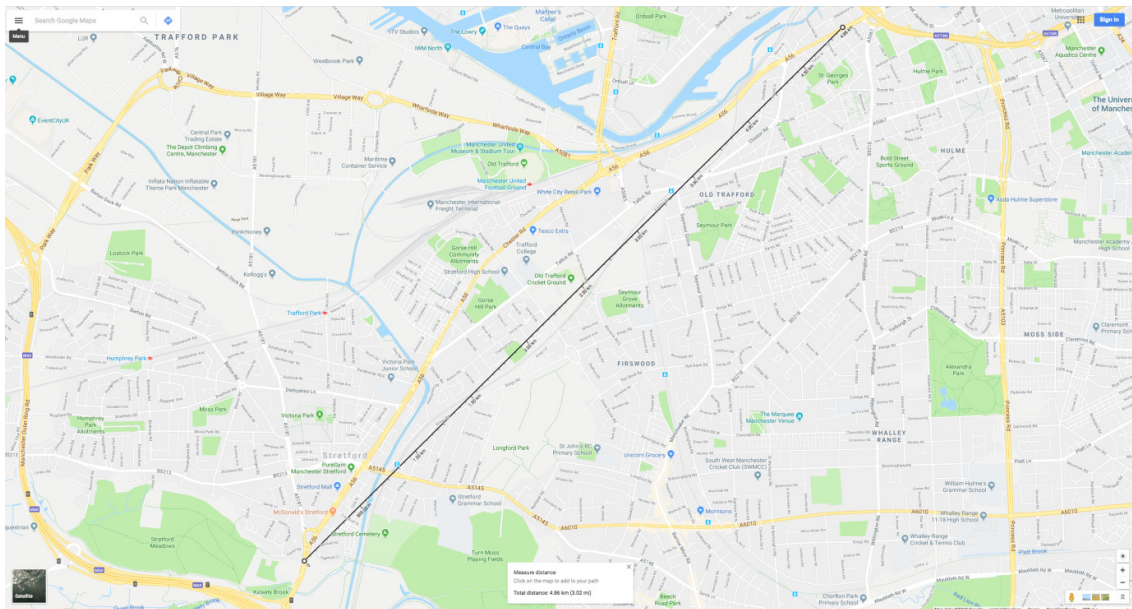
COMP60711 - DATA ENGINEERING
Data (Pre-)Processing Coursework
Description

Coursework Description

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.

Note that sites 1083 and 1415 are consecutive and are all plated on Chester road. The figure below shows the location of the sites on a map.



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 an unique identifier).
- **Lane Name** is the name given to a particular lane of the road. Each lane has an 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.

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 you 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. Develop each task twice, once using a tool or Programming Language (PL), and a second time using a different tool or Programming Language (PL). Aim to choose one Graphical User Interface (GUI) based tool and one Programming Language (PL) to perform all tasks. Once your two methods are chosen (the tool and PL), 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: Trifacta, Talend, and Microsoft Excel. All of these are available from the machines in the M.Sc. lab, and can also 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 (from Week 2), each student should submit one pdf file containing the solution to the task(s) of the previous week.

Tasks for each Week

Week 1 Task

Deadline: Tuesday of Week 2 at 09:00 am.

This is a "warm up" task, therefore, there is no lab session for this particular task and it should be developed during Week 1. Also, it does not incur any marks and so it is not marked by the TAs. 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

Three tasks to be developed during the course of Week 2, carrying 100 marks.

Deadline: Tuesday of Week 3 at 09:00 am.
Marking deadline: Tuesday of Week 4 at 09:00 am.

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

Week 3 Tasks

Three tasks to be developed during the course of Week 3, carrying 100 marks.
Deadline: Tuesday of Week 4 at 09:00 am.
Marking deadline: Tuesday of Week 5 at 09:00 am.

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

Tasks Description

Task 1: Update the 'Flag' and 'Flag Text' columns (for the entire file) by creating an index for each day of the week in the dataset. Rows where the date falls on a Tuesday should have the 'Flags' entry as 2 and 'Flag Text' as 'Tuesday', while Fridays should have 'Flags' 5 and 'Flag Text' 'Friday'. Also, calculate the total traffic volume for each day of the week (i.e., the total traffic volume considering all Tuesdays of the month, and the total traffic volume considering all Fridays of the month). Note that the total traffic volume should not be placed anywhere in the data file, but merely described in the pdf file to be submitted.

Output: Two values corresponding to the total traffic volume of each day of the week + a screenshot of the updated dataset. Provide a step-by-step description of the development of the task, emphasising the features of each of the tools/languages that you used.

Note that the following tasks carry marks. When you are using a GUI-based tool, make sure you include screenshots of intermediate steps you had to carry out while using the tool to complete a task (especially where no coding was necessary), as well as the final recipe/result, and explanations of each step.

Task 2: Provide a simple profile of the traffic volume of the North lanes on site 1083 using the following descriptive data summarization measures and focusing only on Tuesdays at 09:00 am: Range, 1st Quartile, 2nd Quartile, 3rd Quartile, Interquartile range.

Output: Five values, one per required data summarization measure. Provide a step-by-step description of the development of the task, emphasising the features of each of the tools/languages that you used.

Marking scheme:

10 marks for correct values of all five measures (2 marks for each).

20 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.

Task 3: 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, for a weekday of your choice (Tue or Fri). Each bar plot should show the average traffic volume for each hour of the day. Provide a step-by-step description of the development of the task, emphasising the features of each of the tools/languages that you used.

Marking scheme:

10 marks for correct barplots (5 marks 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.

Task 4: Fill the missing values of columns 'Gap (s)' and 'Headway (s)' for all records associated with the NB_MID lane (North direction), considering any Tuesday between 7:00 and 19:00 for which values for one of both of these columns are missing. To fill the missing values, you should use the median calculated for the particular hour of the day when the missing value occurs as replacement value. For example, if missing values are found on Tuesday 06/02/2018 - 10:00 and Tuesday 20/02/2018 - 15:00, then you should calculate the median of gap (or headway) considering all Tuesdays at 10:00 and all Tuesdays at 15:00 to obtain two values, `median_at_10:00_allTuesdays` and `median_at_15:00_allTuesdays`. These are to be used as replacement values. To calculate these values you can do the following:

- sort the values (gap or headway) inside each time interval, e.g. from 10:00 to 11:00; and
- get the value in the middle. If there are two values in the middle, then take the average of both.

Output: X numerical values (X depends on the number of missing values) representing medians of the Gap and Headway columns for all Tuesdays between 7:00 and 19:00. A screenshot of the updated dataset should be included as well. Provide a step-by-step description of the development of the task, emphasising the features of each of the tools/languages that you used.

Marking scheme:

10 marks for correct results and screenshot (5 marks for results and 5 marks for the screenshot).

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.

Task 5: Create bar plots of the average speeds for Tuesdays and Fridays in both traffic directions. Discuss whether the average speed of vehicles is lower in the evening rush hour than that observed in the morning rush hour. You can re-use the results obtained in Task 3. The rush hours are as follows:

- Tuesdays:
 - Morning 7:00-10:00
 - Evening: 16:00-19:00
- Fridays:
 - Morning: 07:00-10:00
 - Evening: 16:00-18:00

Output: Four bar plots with the averages speeds per hour of the day, two per day, one in each traffic direction. Provide a step-by-step description of the development of the task, emphasising the features of each of the tools/languages that you used.

Marking scheme:

10 marks for correct bar plots (5 marks for each).

20 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.

Task 6: Estimate the typical Friday journey time for the road fragment between site 1083 and site 1415 between 17:00 and 18:00, using only the North direction lanes. To make this estimation, you first need to find the average speed at 17:00 on Fridays for the relevant lanes and divide this by the distance between the two sites, which is 4.86km. Multiply the result by 60 to get the result in minutes. To calculate the average speed between 17:00 and 18:00 of the all north lanes, consider not only the three average speeds associated with the three North lanes found on site 1083, but also the two average speeds of the two North lanes found on site 1415 (use file rawpvr_2018-02-01_28d_1415 TueFri.csv for that), since these can be quite different. For example, if the values you found are {15837, 14777, 18000, 13222, 14995}, then take the average of these values.

Output: A value in minutes for the JT. Provide a step-by-step description of the development of the task, emphasising the features of each of the tools/languages that you used.

Marking scheme:

3 marks for correct result.

32 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.

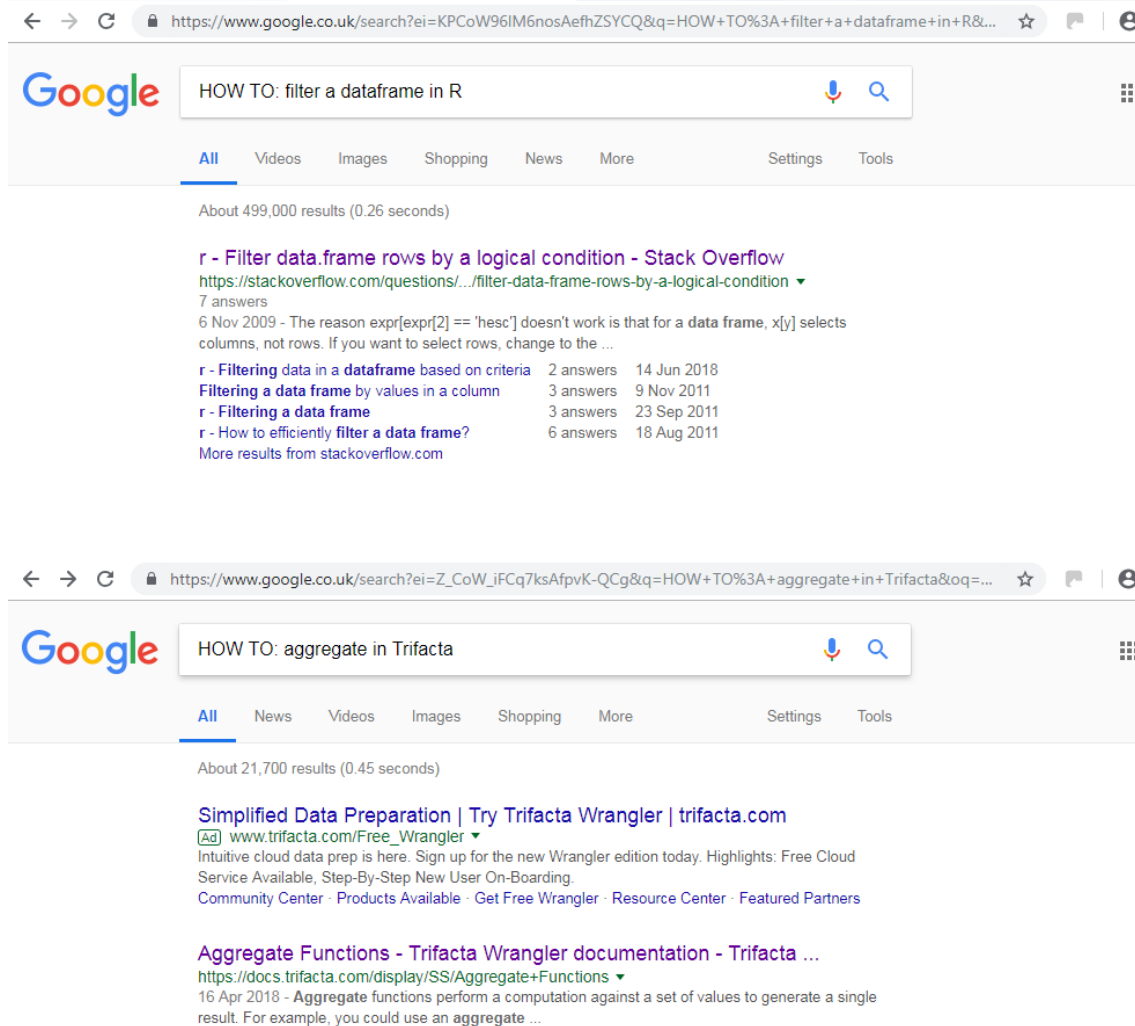
Task 7: Considering the six tasks described above, which by now you should have done and the two tools you have used, compare the tools, discussing advantages and disadvantages of each tool for each task, taking into account not only any extra work (manual or not) you had to do for absence of facilities, or limitations of existing facilities, but also the time it took for you to perform a particular action.

Marking scheme:

35 marks for a clear and complete assessment of the tools, which should include advantages, limitations relating each to a task you have performed using the tool, but also a comparison between the two where they are contrasted.

Getting Help from Web-Available Material

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>)

Talend Data preparation

Introduction to Talend Data Preparation (video)

(<https://www.youtube.com/watch?v=kcU6g4C8Fuk>)

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>)

Trifacta Wrangler

Essentials (<https://www.trifacta.com/support/essentials/>)

Trifacta Tutorial (videos)

(https://www.youtube.com/results?search_query=trifacta+wrangler)

Excel

Excel Tutorial (<https://www.tutorialspoint.com/excel/>)