Big Data and Data Mining.

Workshop Number 1.

(3 & 4/Feb/2022).

Aims of the workshop.

This workshop will review some aspects of the SQL learning from the past week before progressing on into introducing the main PROJECT for this trimester.

Workshop Timetable.

The timetable below should be taken as indicative only. We will modify the times according to how quickly things progress.

|  |  |
| --- | --- |
| **Time** | **Activity** |
| 13:00-13:05 | Welcome and introductions. |
| 13:05-13:20 | Overview of module and workshops. Databases and SQL group discussion. |
| 13:20-13:30 | Introduction to the PROJECT. |
| 13:30-14:30 | Afternoon Exercises 1. |
| 14:30-14:45 | Discussion / Tutorial on Afternoon Exercises. |
| 14:45-15:45 | Afternoon Exercises 2. |
| 15:45-16:00 | Wrap up Discussion. |

Useful Information.  
Throughout this workshop you may find the following useful.

Python Documentation

https://docs.python.org/3.8/

This allows you to lookup core language features of Python 3.8 as well as tangential information about the Python Language. We will refer you back to the Programming module by Dr Brian Tompsett for more notes on this.

Jupyter Notebook Basics

Jupyter itself offers some basic documentation for people new to the editor. These can be found on https://jupyter-notebook.readthedocs.io/en/stable/examples/Notebook/Notebook%20Basics. html

Jupyter can also use markdown cells for text input to describe things. If you wish to annotate each cell as to which Exercise it belongs to, you may find https://jupyter-notebook.readthedocs.io/en/stable/examples/Notebook/Working%20With%20 Markdown%20Cells.html​ useful.

Reminder.

We encourage you to discuss the contents of the workshop with the delivery team, and any findings you gather from the session.

Workshops are not isolated, if you have questions from previous weeks, or lecture content, please talk to us.

The contents of this workshop are ​not​ intended to be 100% complete within the session; as such it’s expected that some of this work be completed outside of the session. Exercises herein represent an example of what to do; feel free to expand upon this.

**13:00. Introductions.**Module Lead:

Dr. Kevin Pimbblet. ([k.pimbblet@hull.ac.uk](mailto:k.pimbblet@hull.ac.uk))

Demonstrator Team.

**13:05. Module overview.**  
We will spend few minutes outlining the major features of the module, including assessments, and its relative timetabling.

**Discussion (mini-tutorial / study group) on Week 1’s learning.**

The past weeks has introduced the concepts of big data, databases, relational databases and how to access information from databases via a new language: SQL.

Prompter Questions for Discussion:   
(a) What advantages does an SQL database hold over something like Microsoft excel?  
What issues could arise by using excel instead? Can you cite examples?

(b) What are some practical limits and(or) considerations for interacting with databases and “big data” in terms of hardware, software and processing times?

**13:20. Introduction to the PROJECT.**

In this slot, we want to spend a short amount of time introducing the project.  
For formal links to the data we have gathered for the project, please follow:

<https://www.gov.uk/government/collections/road-accidents-and-safety-statistics>

This data has already been retrieved from a large database and (happily) is ready to work with without needing to use SQL.

**13:30. Afternoon Exercises, Part 1.**

Exercise 1.

**Understanding the data.**   
Our first challenge is to understand the datasets that we are dealing with. You may do this part in excel, but python is preferred.

What links are there between:

1. **Road Safety Data - Accidents 2019**
2. **Road Safety Data - Casualties 2019**
3. **Road Safety Data- Vehicles 2019**

Explicitly note down the connections and relationships between these datasets. What are the key features that will enable us to investigate the data? Are there any “foreign keys” present?

Exercise 2.

**First Look with EDAs.**  
Once we are happy that we have a feeling for these data, we need to ensure that they are moderately clean.

Using your preferred EDA (e.g., sweetviz, pandas profiling, etc.), please open up each of these datasets and check what each of the columns contains and if there are any possible gaps or problems with the data.

Note down any problems. We will need to deal with each in turn.

Exercise 3: OPTIONAL.

**DTYPE problems.**

Hopefully(!) many of you will encounter something like the following warning when using pandas with these large files:

DtypeWarning: Columns (0,31) have mixed types.Specify dtype option on import or set low\_memory=False.

interactivity=interactivity, compiler=compiler, result=result)

**There is no need to do anything about this – a warning is not an error after all.**

If you wish to understand this warning is a bit more detail, then the answer is that python is trying to guess the dtypes, and this is VERY demanding.

We could help python out!

We can modify the pandas csv read command by adding:

**dtype={'variable\_ID': int}**

inside the brackets. Go ahead and add this to your pd.read\_csv command.

Note that this can generate more errors. These are likely caused by problems with the csv file at this point. Make a copy of the csv file to actively work with – we will need to do some cleaning.

Exercise 4.

**Cleaning**.   
Hopefully you have seen that there are some NaN entries (etc). We wish to RETAIN as much information in the dataset as possible and not throw any data away if avoidable.

In Fundamentals of Data Science, we considered multiple ways of handling this problem.

For the moment, I would like you to insert “unknown” or a numeric flag such as “-9999” or similar values to replace the NaN’s with.

This will suffice for today, but for your final project report, you WILL need to return to clean the data more thoroughly. You should decide how you will accomplish this, and the team will discuss this in further detail with you.

**14:30. Afternoon Discussion.**  
We will start our mid-afternoon discussion by thinking about one of the first tasks: identifying hours of the day, and days of the week, that road traffic accidents occur on.

1. What are our expectations? Should there be particular hours of the day that feature more accidents? Justify your answer.
2. In regards to days of the week, are there going to be particular days with higher rates of accidents? Justify your answer.
3. Can you think of any “special” days of the year when accidents might be higher or lower?
4. What effect (if any) would large sporting events (or similar) have? Justify your answer.

**14:45. Afternoon Exercises.**

In the second session of the afternoon, we want to refresh our memories about data visualization approaches, how to represent the data, and how to quantify some relationships. As we go through the coming weeks, we will add to our analysis here by adding on our new skills to analyse the dataset(s) in fresh ways. For this week, we want to construct the basis for our analysis.

Exercise A.

In order to create a histogram of accidents binned by hour of the day, it might be necessary to convert the hours of the day in to decimal format for ease of use with seaborn or plt.

The problem here is that sns will not like the time stamp format that comes with this data set (i.e., HH:MM). Have a go to try to convert this format into decimal time (i.e. hours + minutes/60.). The answer to this challenge is on the next page, so don’t look until you’ve had a good wrangle with the datetime command.

Here is an illustrative solution to the datetime conundrum.

df['converted\_time'] = pd.DatetimeIndex(df['Time'])

df['decimal\_time'] = time.hour + time.minute/60.

We now have the decimal time stored as df['decimal\_time']. From here, it is simple to make a histogram of the time of day. See if you can reproduce a diagram that looks similar like this.

Chart, histogram

Description automatically generated

Is this shape what you thought it might look like? Explain your answer.

Quantify (in terms of standard deviations) just how much more likely accidents are at evening rush hours compared to the rest of the hours of the day.

Exercise B.

Repeat the above, but for day of the week – this should be much easier. Make sure you can recover a graph that looks similar to the one below.

Extension: Repeat this for day of the year.

Chart, bar chart

Description automatically generated

Exercise C.

On Sunday 24/Feb/2019, Manchester United played against Liverpool (in Old Trafford, Manchester) in a well-attended football match.

Was there an excess of accidents around Old Trafford and the Manchester region due to this match that were in excess of what might have been expected?

To address this, we will need to think about what the “typical” accident rate is for the Manchester region on Sundays. You will need to therefore extract the geographic coordinates of Old Trafford from the internet, and draw an imaginary circle with some radius (5km? a bit more maybe?) around the stadium. Determine the number of accidents within this radius, on average, over every Sunday. Is the number of accidents on 24/Feb/2019 any different to this baseline? Does it stick out as obviously anomalous? Justify your answer.

Take your time with this one.

Exercise D (an optional extension to exercise C).

Using geopandas, try to plot where accidents took place in Manchester on this date.

Is there any obvious concentration of accidents near the stadium or along major roads leading around Manchester on that day or near to the kick-off time of the match?

Exercise E. (Optional, if time permits).

Are pedestrians more safe, or less safe, on football match days in and around Old Trafford?

Justify this using a hypothesis test.

**15:45. Wrapping Up.**

Today we have made a start on the project for the semester. But it is only a small start.

Over the next few weeks, we will be learning about rule generation, clustering, outliers, classifications, and more. We will aim to apply our learning with each fortnight to this particular dataset to obtain useful information out of it.

It must be emphasized that although we are “learning by doing” in these workshops by using the new techniques each fortnight, to achieve the higher grades, you will have to really go beyond what we do in these workshops.

You may like to consider having a first look at the next questions before the next workshop:

1. For motorbikes, are there significant hours of the day, and days of the week, on which accidents occur?
2. For pedestrians involved in accidents, are there significant hours of the day, and days of the week, on which they are more likely to be involved?

For now, well done, and we will continue next time.