001_CWRK:Project Report

29/04/2025

Attempt 1 ⌄

In Progress

NEXT UP: Submit assignment

Add comment

Unlimited Attempts Allowed

06/05/2025

⌄ **Details**

Report Assignment

Assignment (70 %)

Maximum Word Count: 2000 words.
Dataset: Data for the Workshop & Assignment (https://canvas.hull.ac.uk/courses/74993/pages/data-for-the-workshop-and-assignment)
Deadline: 29 April, 2025, 14:00

Context.

This assignment is based on real world data: specifically, road traffic accidents **in 2020.** This assignment is a chance to test your skills against such
real-world data in order to produce meaningful outputs.

Project Background Information.

All road traffic accidents involving casualties are logged and reported in Great Britain, along with (probably) a majority of other non-fatal road traffic accidents. Every year, the government releases a large batch of data associated with these reports. In this assignment we will be **using the data from 2020.**

We have uploaded the relevant data to Canvas, you can access it from the link at the top, or from these links below:

I. accident_data_v1.0.0_2023.db : (https://canvas.hull.ac.uk/courses/74993/files/5776642?wrap=1) ↓ (https://canvas.hull.ac.uk/courses/74993/files/5776642/download?download_frd=1) an sqlite database containing the accident data. You should extract data from 2020 from this database.

II. Reported road casualties in Great Britain: notes, definitions, symbols and conventions: (https://www.gov.uk/government/publications/road-accidents-and-safety-statistics-notes-and-definitions/reported-road-casualties-in-great-britain-notes-definitions-symbols-and-conventions) government guidance on the data set.

III. stats20-2011.pdf: (https://canvas.hull.ac.uk/courses/74993/files/5776643?wrap=1) ↓ (https://canvas.hull.ac.uk/courses/74993/files/5776643/download?download_frd=1) Detailed guidance on how to complete accident reporting forms.

IV. Road Traffic Accidents Statistics Form: (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995422/stats19.pdf) the form used to report road traffic accidents.

V. dft-road-casualty-statistics-road-safety-open-dataset-data-guide-2023-1.xlsx : (https://canvas.hull.ac.uk/courses/74993/files/5776681?wrap=1) ↓ (https://canvas.hull.ac.uk/courses/74993/files/5776681/download?download_frd=1) This form states the meaning of the numerical values in each column.

VI. facebook_combined.txt: (https://canvas.hull.ac.uk/courses/74993/files/5776682?wrap=1) ↓ (https://canvas.hull.ac.uk/courses/74993/files/5776682/download?download_frd=1) This is the social media dataset.

Task

Imagine that you are a data scientist confronted with this data (this is not far from the truth!). Your task is to advise government agencies about how to improve road safety and create a model that would predict such accidents and the injuries that they incur.

Importantly, we have used and will be using time within our workshops to help with this assignment, and it doesn't have to be all completed at once.

The questions (at minimum) that the assignment should address are as follows:

1. Are there significant hours of the day, and days of the week, on which accidents occur?
2. For motorbikes, are there significant hours of the day, and days of the week, on which accidents occur? We suggest a focus on: Motorcycle 125cc and under, Motorcycle over 125cc and up to 500cc, and Motorcycle over 500cc.
3. 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?
4. Using the apriori algorithm, explore the impact of selected variables on accident severity.
5. Identify accidents in our region: Kingston upon Hull, Humberside, and the East Riding of Yorkshire etc. You can do this by filtering on LSOA, or police region or another method if you can find one. Run clustering on this data. What do these clusters reveal about the distribution of accidents in our region?
6. Choose three policing areas by filtering the data using the "police_force" column, then create time series models to predict weekly accident counts for the upcoming year based on historical data from 2017 to 2019.
7. Identify the three Local Super Output Areas (LSOAs) of Hull city that recorded the highest number of road accidents in the first three months of 2020, then employ a time series model to forecast daily accident occurrences for the upcoming month (e.g., July), leveraging data from the preceding six months (e.g., January to June) for these high-incident areas.
8. Construct a social network using the provided data and visualise the network, then provide the basic network characteristics, including numbers of nodes and edges, network density, average degree.
9. Calculate the edge centrality of this network and plot the distribution of the edge centrality values.
10. Use two community detection algorithms to detect the clusters/community within this social network, then compare the difference of results (the number of clusters and numbers of nodes in each cluster).

Your Report (Suggestive).

Please endeavour to structure your report as much as possible using the following outline.

1. **Short introduction.** No more than a few sentences introducing the dataset and the problems that you seek to solve using it.
2. **Analysis.** Present an analysis of the data, including any visualizations, that address the questions above. This should be broken down in to analysing when, where, and under what conditions accidents happen, as per the questions above. For questions 8 to 10, do not forget to justify any method or algorithm you use for those questions. Document any data cleaning relevant to the analysis here.
3. **Predictions**. Discuss the results of any of your predictions and what you learned from them.
4. **Recommendations**. What recommendations can be made to government agencies based on this data and your analysis to improve safety? Keep this to your top 4 or 5 bullet points.

Please upload:

You should upload any Python code you have worked on alongside your written submission.

1. Your written address to the assignment, including visualisations.
2. The code you wrote to produce the results and/or visualisations used in the assignment as **a separate Jupyter notebook** file.

Important notes:

- Upload these files **separately** in the same submission. Do not use ZIP files, the allowed submission file types have been limited.
- **Do not** include Python code in the written report. Python snippets are not allowed in the written report. The Python code should all be in the notebook. We will disregard Python code included in the written report.

Given the word count, it is essential to be concise in your answers. It is strongly suggested that you illustrate your answers with appropriate diagrams (i.e. visualisations) or appendices of example calculations. Further, you might need to read around the topic and undertake library/online research to help with this assignment to achieve the highest grades.

Grading.

The following grading rubric will be applied to your supplied answers. The total number of marks available for this assignment is 70%. Please note that submitting lots of data is unlikely to attract many marks. Instead, we want to see fully reasoned analyses supported by evidence derived from the data supplied.

Supporting Teaching Materials

The activities in the labs support the analysis of this dataset. Please see the lab scripts for examples of analysis that may be relevant to solving this assignment.

Extensions and Additional Considerations

If you need more time and have a good reason, you should apply for an extension or additional consideration via **the portal (https://hull.service-now.com/student)** . You need to do this as early as you can as there are deadlines, please don't delay. DAIM cannot grant extensions, you need to complete this form. If this sounds relevant to you, we highly recommend the **My Journey course (https://myjourney.hull.ac.uk/learner/course)** on Additional Considerations and Extensions. If you want help with mental health or well-being, **Student Support (https://www.hull.ac.uk/choose-hull/student-life/student-support)** can offer you a variety of options. Please also reach out to your personal supervisor for help with anything impacting your academic performance.

Academic integrity

Please note that **all** work that you submit **must be your own** or properly referenced. See the **academic integrity course (https://canvas.hull.ac.uk/courses/67857)** or see your personal/project supervisor for more information.

View Rubric

Big Data and Data Mining

Criteria						Points
	Full marks	**Merit**	**Pass**	**Fail**	**No marks**	
Coding quality and Use of Jupyter Notebook (10 Marks)	Code is well-structured, modular (with use of functions), commented, and easy to understand. Jupyter notebooks are used effectively, with clear headings, explanations, and markdown cells that provide context for the code. The code is efficient and reproducible (easy to run without errors). Appropriate error handling is implemented. Solution relies minimally on code duplication, and additional I/O opening/closing. **9.1 to 10 pts**	Code is generally well-structured and commented. Jupyter notebooks are used effectively. Code is mostly reproducible, but there may be minor issues. Error handling is implemented in most of the cases. **7.6 to 9 pts**	Code structure and comments are adequate, but could be improved. Jupyter notebooks are used, but the organization could be better. Code may require some effort to reproduce. Limited error handling is implemented. **5.1 to 7.5 pts**	Code structure is poor, and comments are lacking. Jupyter notebooks are poorly organized or misused. Code is difficult to reproduce or contains errors. No error handling is implemented. **2.6 to 5 pts**	Code is missing or is completely unreadable. Jupyter notebooks are not used or are used incorrectly. Code is not reproducible. **2.5 pts**	/10 pts

Overall Understanding and Application of Big Data & Data Mining Concepts (15 Marks)	**Full marks** Demonstrates a comprehensive understanding of time series analysis, association rule mining, clustering, social network analysis, and spatiotemporal data analysis techniques. Concepts are applied correctly and insightfully across all questions, showing a strong understanding of when each technique is appropriate and its limitations. **12.1 to 15 pts**	**Merit** Demonstrates a strong understanding of most techniques, with minor gaps in either time series, association rules, clustering, social network analysis or spatiotemporal data analysis. Concepts are generally applied correctly, but the application could be more insightful in some areas. **9.1 to 12 pts**	**Pass** Demonstrates a basic understanding of several techniques, but with significant gaps or misunderstandings Application is attempted but may contain errors or lack depth. Demonstrates a limited understanding of spatiotemporal data analysis, association rules, clustering, time series and social network analysis. **6.1 to 9 pts**	**Fail** Demonstrates a limited understanding of the core techniques. Application is weak, flawed, or demonstrates a significant misunderstanding in multiple areas. Struggles with choosing appropriate techniques. **3.1 to 6 pts**	**No marks** Demonstrates little to no understanding of the core techniques. Fails to apply concepts or provides completely irrelevant information. **3 pts**	/15 pts
Overall Data Analysis and Interpretation (25 Marks)	**Full marks** Conducts thorough and insightful analysis across the report, using appropriate visualization techniques to reveal patterns in accident timing (hourly, daily), geographic distribution, and relationships between variables. Interpretation is accurate, well-supported by data, and demonstrates a clear understanding of the implications of each analysis. **20.1 to 25 pts**	**Merit** Conducts good data analysis across most questions, identifying most key patterns. Interpretation is generally accurate and supported by data, but some nuances may be missed. For social network analysis the analysis could have some gaps to reach perfect interpretation. **15.1 to 20 pts**	**pass** Conducts adequate data analysis. May miss important patterns, lack sufficient support, or contain minor inaccuracies. Interpretation may be superficial. Social network analysis might be less effective. **10.1 to 15 pts**	**fail** Conducts limited or flawed data analysis. Interpretation is weak, unsupported, inaccurate, or demonstrates a misunderstanding of the data. Social network analysis is significantly flawed. **5.1 to 10 pts**	**Low or No marks** Fails to adequately analyze the data or provide any meaningful interpretation. Presents irrelevant or incorrect conclusions. **5 pts**	/25 pts

Overall Methodology and Justification (20 Marks)	**Full marks** Methodologies (Apriori, Clustering, Time Series Models, Social Network Analysis algorithms) are highly appropriate, clearly explained, and thoroughly justified with strong reasoning. The rationale for parameter selection is clear. Alternative approaches are considered and a clear rationale is provided for the chosen methods. **16.1 to 20 pts**	**Merit** Methodologies are appropriate, explained clearly, and justified with reasonable reasoning. Justification could be more thorough or consider alternative approaches more explicitly. The rationale for parameter selection might be less detailed. **12.1 to 16 pts**	**Pass** Methodologies are generally appropriate, but the explanation or justification is weak or incomplete. The connection between the methodologies and the questions may not be entirely clear. The rationale for parameter selection might be missing. **8.1 to 12 pts**	**Fail** Methodologies are generally appropriate, but the explanation or justification is weak or incomplete. The connection between the methodologies and the questions may not be entirely clear. The rationale for parameter selection might be missing. **4.1 to 8 pts**	**No marks** No methodologies are presented, or the chosen methodologies are completely irrelevant. No justification is provided. **4 pts**	/20 pts
Completeness (25 Marks)	**Full marks** All 10 questions are answered completely and thoroughly. The answers demonstrate a strong understanding of the data and the application of appropriate techniques. There are no significant gaps or omissions. **20.1 to 25 pts**	**Merit** 9 questions are answered completely and thoroughly. Minor omissions or superficiality in the answer to one question. The overall understanding and application of techniques are still strong, but one answer may lack some depth or detail. **15.1 to 20 pts**	**Pass** 7 questions are answered adequately. Noticeable omissions or superficiality in the answers to 3 questions. The quality of the answers is generally acceptable, but there are significant gaps or areas where the analysis is lacking in depth or rigor. **10.1 to 15 pts**	**Fail** 4 or fewer questions are answered adequately. Significant omissions or superficiality across multiple questions. The overall quality of the answers is poor, with significant gaps in understanding, methodology, or interpretation. The submission shows a lack of effort and attention to detail. **5.1 to 10 pts**	**No marks** Substantial portions of the assignment are incomplete. Many questions are not addressed, or are addressed with only minimal effort. The submission demonstrates a lack of understanding and effort. **5 pts**	/25 pts

Overall Clarity, Communication and Formatting (5 Marks)	**Full marks** The report is exceptionally clear, concise, and well-organized. Visualizations are clear, effective and properly labeled. The findings from different analyses are synthesized effectively to provide a cohesive narrative. Language is precise, professional, and free of errors. Arguments are logically structured and easy to follow. Formatting is consistent and professional. All references are cited correctly and completely, following the Harvard citation style. **4.1 to 5 pts**	**Merit** The report is clear, concise, and well-organized. Visualizations are effective and labeled. Findings are generally synthesized well. Language is generally precise and professional, with few errors. Arguments are logically structured and easy to follow. Formatting is generally good with minor inconsistencies. Most references are cited correctly, but there may be minor errors or omissions. **3.1 to 4 pts**	**Pass** The report is generally understandable, but may suffer from some lack of clarity, organization, or conciseness. Visualizations may lack detail or clarity. Synthesis of findings may be limited. Language may be imprecise or contain some errors. Formatting is inconsistent. There are noticeable errors or omissions in the citation of references. **2.1 to 3 pts**	**Fail** The report is difficult to understand due to a lack of clarity, poor organization, or significant errors in language. Visualizations are poorly designed or missing labels. Synthesis of findings is weak or missing. Formatting is poor and distracting. There are significant errors or omissions in the citation of references. **1.1 to 2 pts**	**No marks** The report is incomprehensible due to extreme lack of clarity, organization, and numerous errors. No coherent arguments or synthesis of findings are presented. Formatting is absent or completely inappropriate. References are missing or entirely incorrect. **1 pts**	/5 pts

Choose a file to upload

File permitted: IPYNB, PDF, DOC, DOCX

or

Canvas Files

□ I agree to the tool's **End-User License Agreement (https://api.turnitinuk.com/api/lti/1p0/user/static_eula)**
I take full responsibility for my submitted piece of work and have acknowledged all sources of material in accordance with the required referencing style (e.g. Harvard). The required referencing style has been communicated to me in the assessment brief or by other means. My work is not copied and I have acknowledged work and ideas of others by correct referencing in all relevant sections. I understand improper referencing places me at risk of a reduced mark or the penalties of the University's academic misconduct regulations.
*