MSc in Artificial Intelligence and Data Science

**Assignment – 771762 Big Data and Data Mining PROJECT**

*(This assignment is worth 70% of the total marks for this module)*

This assignment should be submitted **via turnitin on CANVAS**

using the specific coversheet for this assignment, which can also be found on CANVAS by

**Friday 29th April 2022 at 23:59 (i.e., teaching week 11) *at the latest.***

**Maximum Word Count for the written part of this assignment = 2000 words.**

**Context.**

Unlike in our prior module, Fundamentals of Data Science, this assignment is based on real world data: specifically, road traffic accidents in 2019. 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 the United Kingdom, 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 2019 as it represents a very complete sample with a lot of ancillary data available. We have uploaded the relevant data to Canvas and it contains the following files:

1. **Brief-guide-to road-accidents-and-safety-data**. This file contains a very short introduction to the data set produced by the government.
2. **variable lookup**. This spreadsheet details the large amount of variables present in each of the datasets.
3. **Road Safety Data - Accidents 2019**. This spreadsheet contains the details of the accidents logged in 2019.
4. **Road Safety Data - Casualties 2019**. This spreadsheet details the casualties contained in the logged accidents.
5. **Road Safety Data- Vehicles 2019**. This spreadsheet contains the details of the vehicles involved in the accidents.
6. **Adjustment Files**. These files contain government modelling for the probabilities of different injuries occurring for each accident.
7. **adjustment figure guidance**. This provides a short amount of government guidance on the figures included.
8. **cas\_adjustment\_lookup\_2019**. This spreadsheet contains the tabulated adjusted probabilities of injuries.

**Please note**: some of these files are simply too large to open even within excel. Therefore, it will be necessary to undertake analysis within Python.

**The 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 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?
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. What impact, if any, does daylight savings have on road traffic accidents in the week after it starts and stops?
5. What impact, if any, does sunrise and sunset times have on road traffic accidents?
6. Are there particular types of vehicles (engine capacity, age of vehicle, etc.) that are more frequently involved in road traffic accidents?
7. Are there particular conditions (weather, geographic location, situations) that generate more road traffic accidents?
8. How does driver related variables affect the outcome (e.g., age of the driver, and the purpose of the journey)?
9. Can we make predictions about when and where accidents will occur, and the severity of the injuries sustained from the data supplied to improve road safety? How well do our models compare to government models?

**Your Report.**

Please structure your report as follows.

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 (a)-(h), above. This should be broken down in to analysing when, where, and under what conditions accidents happen, as per the questions above.
3. **Predictions**. This should be a working model to address point (i), above, that can predict the conditions under which accidents are most likely to occur in, **and** the severity of injuries sustained given the conditions they happen under.
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.

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

**Grading.**

The following grading rubric will be applied to your supplied answers. The total number of marks available for this assignment is 40. 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.

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.

Please upload:

1. Your cover sheet.
2. Your written address to the assignment, including visualisations.
3. The code you wrote to produce the results and/or visualisations used in the assignment.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criteria** | **0 Marks** | **Up to 6 marks** | **Up to 12 marks** | **Up to 20 marks** |
| **Coding in Python.** | No example of code has been uploaded, or what has been uploaded is of very low quality or undertaken exclusively in excel (or similar). | Example of code has been uploaded. The code might not be complete, or it might have some obvious omissions or errors inside it.  Commenting is poor to mediocre. | The code supplied is complete and it is able to generate all of the important items pertinent to the final report.  A variety of appropriate methods have been used in the code. Commenting is extensive and of a good quality. | The code demonstrates superior insight and coding ability that makes it highly efficient and goes beyond simply being functional – it is a highly refined piece of work and at the highest end would be suitable for inclusion in official repositories. |
| **Analysis, Modelling, and Interpretations.** | Little to no attempt has been made at analysing the data or modelling. If an attempt has been made, it is very poor quality or fundamentally flawed. | An attempt has been made to determine the answers to questions (a)-(i). The answers may be poor, in part, or the interpretation wrong in parts. A model is present but may be poorly executed. Some limited visualizations are present. | The model built is successful in predicting the conditions under which accidents happen and the severity of injuries sustained. Nearly all of the analysis and interpretation is wholly correct and supported by the data and appropriate visualizations. Correct recommendations have been made. Citations have been provided to official sources. | The analysis, modelling and interpretations are of a clearly superior nature and demonstrate significant insight to the brief.  At the highest end of the mark scheme, the work could be considered for publication. |