**Task**

**Imagine you are a data scientist recently hired by the government to demonstrate the new SQL database you have created. Imagine the government is also planning on storing names of people involved in accidents. You produced a relational database in SQL. In your presentation you should**

1. **Introduce the potential technical, ethical and legal issues of storing the data. How might an SQL relational database address some of these?**
2. **Present and describe an entity-relationship diagram (ER diagram) for the database. There are a lot of columns in each table, so do not show them all. Just show primary keys, any foreign keys and a representative sample of 2 or 3 of the other columns.**
3. **Present SQL commands for the following, using the full data set from all years**
4. 1. **The age of the oldest driver/rider in the casualty table**
   2. **The total number of vehicle\_type = 19 vehicles in the vehicle table**
   3. **The sex of driver, sex of casualty, speed limit and age of vehicle for accidents in all the lower layer super output area (LSOA) regions of Kingston Upon Hull. You will need to do a JOIN on the lsoa table and some of the other tables. In the LSOA table, the codes are in the column lsoa01cd and the place names are in column lsoa01nm. Put the results in a pandas data frame.**

**You should show the SQL commands you ran and the results. For part ‘c’ just show a sample of the results and the total number of rows.  You should also give some explanation for the more complex parts of the SQL commands you present.**

**We have included a ready-made sqlite database containing the required tables. It is the same data you have been using in the workshops. This presentation is a demonstration of your knowledge of SQL and databases.**

**Information about the accident data:**

**The accident, vehicle and casualty tables contain data described in these webpages:**

[**https://www.gov.uk/government/publications/road-accidents-and-safety-statistics-notes-and-definitions/reported-road-casualties-in-great-britain-notes-definitions-symbols-and-conventionsLinks to an external site.**](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)

[**https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\_data/file/995422/stats19.pdfLinks to an external site.**](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995422/stats19.pdf)

**Information about the LSOA regions (including a map)**

**The lsoa table contains data described here:**

[**Lower layer Super Output Areas (December 2001) EW Population Weighted Centroids | Lower layer Super Output Areas (December 2001) EW Population Weighted Centroids | Open Geography Portal (statistics.gov.uk)**](https://geoportal.statistics.gov.uk/datasets/ons::lower-layer-super-output-areas-december-2001-ew-population-weighted-centroids/explore?location=52.644390%2C-2.431471%2C6.97&showTable=true)

**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.