Introduction to Data Science

Lesson 1: Introduction to Data Science
Activity 2 – Suggested student responses
Cars data (AQA data set)

Answer the following questions:

1. The sheet or sheets in the data set are **rectangular**: they contain columns of data for the **fields** (properties or attributes) and rows for the **records** (each row represents a single instance that the data is being recorded for).

    a. How many fields are there in each sheet?

    There are 15 fields; however, two of these are a reference number and a random number so there are 13 fields with information about each car.

    b. How many records are there in each sheet?

    1 sheet with 3827 records.

2. Which of the fields are numerical (showing measures or quantities) and which are categorical (text or options from a list of items)?

    Numerical: EngineSize, Mass, CO2, CO, NOX and part hc. ReferenceNumber and Random number are also numerical.

    Categorical: Make, PropulsionTypeId, BodyTypeId, GovRegion, KeeperTitleId

    YearRegistered is harder to classify. Year is a function of time, which is a numerical variable; however, it is in this dataset YearRegistered behaves like a label and so can be considered a categorical variable.

    A useful way to decide is to consider whether numerical or statistical operations, such as multiplying or finding the average of two values, would make sense. If they would then the field can be classified as numerical; if not it can be classified as categorical. Note that PropulsionTypeId, BodyTypeId and KeeperTitleId are categorical data here even though they are recorded as numbers.

3. Why has this data been stored?

    This data could be used for a number of purposes. It gives the emissions for various cars so could be used by local/national government organisations who are wanting to assess the environmental impact of different types of cars. It could also be used by individuals who are buying cars and are interested in the environmental impact of different types of cars.

    There are many other possible reasons for storing this data and ways it could be used.

4. The video contained some questions that you could answer using this data. What other questions could you answer with this data.

    Some possible questions:
    - Are there any differences in emissions between different makes of cars?
    - Are the emissions for cars registered in 2002 worse than cars registered in 2016?

- Is there a correlation between mass and emissions?
- Is there a correlation between engine size and emissions?
- Is there a correlation between the CO2 and NOX emissions?
- Is there a link between the engine type and the emissions?
- Are any particular makes of cars more likely to be owned by companies?
- Do different genders have different sized cars?

5. This data is a subset of a much bigger data set. How big is the data set it is taken from? How many fields and records could be involved?

   Approximately 2.9 million vehicles were registered in Great Britain in 2018: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/800502/vehicle-licensing-statistics-2018.pdf

   A full data set for the past 20 years could have over 50 million records.

   There are many other fields that could be recorded for cars such as model, number of owners, fuel consumption, power, … The data recorded would affect the analysis that could be performed.