Introduction to Data Science

Lesson 1: Introduction to Data Science
Activity 2 – Suggested student responses
Methods of travel (OCR 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 4 sheets of data: Method of Travel by LA 2011 has 16 fields, Method of Travel by LA 2001 has 15 fields, Age Structure by LA 2011 has 22 fields and Age Structure by LA 2001 has 20 fields.

    b. How many records are there in each sheet?

    All 4 sheets have 348 records.

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

    In all sheets the categorical fields are: geography code; Region; local authority.

    All the other fields are numerical.

    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.

3. Why has this data been stored?

    This data could be used for a number of purposes. The data could be used explore the differences between travel patterns in different regions. This could be used to plan the provision of public transport or investment in other infrastructure such as roads.

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

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

    Some possible questions:
    - Are people in London more likely to cycle to work than people in regions?
    - Which areas of the country have the highest use of public transport?
    - Have there been any changes between 2001 and 2011?
    - Do areas of the country with older populations drive more?

- Is there a correlation between cycling to work and walking to work?
- Is there a correlation between travelling by underground and travelling by train?
- Are people in any parts of the country more likely to work from home?
- Are there any other regional differences in any of the categories?

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?

   A dataset like this could be recorded for every year or more frequently, such as data recorded for different days of the week which would show different patterns. The data could be broken into smaller geographical areas such as districts or "wards". There are approximately 10,000 wards in the UK. As an example a dataset that recorded this for every ward for each month in the last 20 years would have over 200,000 records and separating this into the patterns for each day of the week would have over 1,000,000 records.

   There are many other fields that could be recorded for this such the time taken to get to work, whether multiple modes of transport are used, the amount of time taken or distance to travel to work as well as additional demographic data such as gender or income. The data recorded would affect the analysis that could be performed.