**(a) What is the information you can obtain from the data set/ data sets?**

The data we choose to work with consists of three separate date sets, namely:  
“dft-road-casualty-statistics-accident-last-5-years”,  
“dft-road-casualty-statistics-vehicle-last-5-years” and;  
“dft-road-casualty-statistics-casualty-last-5-years”.

The datasets contain information concerning accidents, the vehicles (and the drivers) involved in the accidents; and the casualties caused by the accidents, respectively, that took place in Great-Britain, between (and including) the 1st of November, 2016 and the 25th of August, 2020. The accidents in the dataset relate only to personal injury accidents on public roads. For an accident to be included in the dataset it must have been reported to the police using the STATS19 accident reporting form.

**(b) What are the attributes in the data and what is their meaning?**

An attribute is some specific property that can be measured, observed, or logged. For example, attributes could be salary, price, number of sales, protein expression levels, or temperature.

Synonyms for attribute are variable and data dimension, or just dimension for short. Since dimension has many meanings, in the book by Munzner it is reserved for the visual channels of spatial position as discussed in Section 6.3.

The attributes in the accident data are:

|  |  |
| --- | --- |
| accident\_index | A unique value for each accident (can be used to link to vehicle as well as to casualty) |
| accident\_year | The year in which the accident took place |
| accident\_reference | ID used by the police to reference an accident (within a year)  (cannot be used to link datasets together) |
| location\_easting\_osgr | Location of the accident according to the Ordnance Survey National Grid |
| location\_northing\_osgr | Location of the accident according to the Ordnance Survey National Grid |
| longitude | Location of the accident according to longitude |
| latitude | Location of the accident according to latitude |
| police\_force | Police force to whom the accident was reported |
| accident\_severity | Severity of the accident |
| number\_of\_vehicles | Number of vehicles involved in the accident |
| number\_of\_casualties | Number of people injured (or dead) in the accident |
| date | Exact date the accident took place (DD/MM/YYYY) |
| day\_of\_week | Day of the week the accident took place |
| time | Time at which the accident took place |
| local\_authority\_district | Authority district in which the accident took place |
| local\_authority\_ons\_district | Authority district (according to the Office for National Statistics) in which the accident took place |
| local\_authority\_highway | Authority district that is responsible for the maintenance of the highway on which the accident took place |
| first\_road\_class | Class of (one of) the road(s) on which the accident took place (e.g. motorway, A(M) etc.) |
| first\_road\_number | Number of (one of) the road(s) on which the accident took place |
| road\_type | Type of road (e.g. roundabout, one way street etc.) |
| speed\_limit | Speed limit on the road in milers per hour (20, 30, 40, 50, 60 or 70) |
| junction\_detail | Type of junction (e.g. roundabout, T-junction etc.) |
| junction\_control | Type of control present at junction (e.g. authorised person, auto traffic signal etc.) |
| second\_road\_class | Class of the other road on which the accident took place (if the accident took place on a junction) (e.g. motorway, A(M) etc.) |
| second\_road\_number | Number of the other road on which the accident took place (if the accident took place on a junction) |
| pedestrian\_crossing\_human\_control | Type of human control present at pedestrian crossing (e.g. school crossing patrol or another authorised person) |
| pedestrian\_crossing\_physical\_facilities | Type of physical facilities present at pedestrian crossing (e.g. zebra, pelican (crossing where pedestrians press a button that operates the traffic lights to stop the traffic) etc.) |
| light\_conditions | Light conditions at the time of the accident (e.g. daylight, darkness – lights lit etc.) |
| weather\_conditions | Weather conditions at the time of the accident (e.g. fine no high winds, raining no high winds etc.) |
| road\_surface\_conditions | Condition of the surface of the road at the time of the accident (e.g. dry, wet/damp etc.) |
| special\_conditions\_at\_site | Special conditions present on site at the time of the accident (e.g. auto traffic signal out, auto signal part defective etc.) |
| carriageway\_hazards | Type of hazard present on the carriageway (i.e. part of the road that carries traffic) at the time of the accident (e.g. vehicle load on carriageway, other object on carriageway etc.) |
| urban\_or\_rural\_area | Whether the accident took place in an urban or rural area |
| did\_police\_officer\_attend\_scene\_of\_accident | Whether a police officer attended to the scene of the accident or not (yes, no or no – accident was reported using self-completion form) |
| trunk\_road\_flag | Whether the road is managed by Highways England or not |
| lsoa\_of\_accident\_location | Lower-layer Super Output Area of the location of the accident (restricted to England and Wales; thus, Scotland and Northern Ireland are not included)  Super output areas (SOAs) were designed to improve the reporting of small area statistics and are built up from groups of output areas (OAs). The LSOA is one of those groups. In 2011 there were 7201 LSOAs (in England and Wales). |

The attributes in the vehicle data are:

|  |  |
| --- | --- |
| accident\_index | A unique value for each accident (can be used to link to vehicle as well as to casualty) |
| accident\_year | The year in which the accident took place |
| accident\_reference | ID used by the police to reference an accident (within a year) (cannot be used to link datasets together) |
| vehicle\_reference | A unique value for each vehicle in a singular accident (i.e. two vehicles involved in the same accident will have the same accident\_index, but a different vehicle\_reference (as far as I can tell the vehicle\_reference is just the count of the vehicles: if two vehicles are involved, one gets vehicle\_reference = 1 and the other gets vehicle\_reference = 2)) (can be used to link to casualty, but not to accident) |
| vehicle\_type | The type of vehicle involved in the accident (e.g. pedal cycle, motorcycle 50cc etc.) |
| towing\_and\_articulation | Type of tow/articulation (i.e. a permanent or semi-permanent pivot join in a vehicle its construction) (e.g. articulated vehicle, double or multiple trailer etc.) |
| vehicle\_manoeuvre | Type of manoeuvre that the vehicle was making at the time of the accident (e.g. reversing, parked etc.) |
| vehicle\_direction\_from | Direction the vehicle was coming from at the time of the accident (e.g. parked, north etc.) |
| vehicle\_direction\_to | Direction the vehicle was heading in at the time of the accident (e.g. parked, north etc.) (if vehicle\_direction\_from is either parked or unknown, then vehicle\_direction\_to should also be parked or unknown, respectively) |
| vehicle\_location\_restricted\_lane | The type of restricted lane that the vehicle was in at the time of the accident (e.g. tram/light rail track, bus lane etc.) |
| junction\_location | Location of the accident ‘within’ the junction (e.g. approaching junction or waiting/parked at junction approach, cleared junction or waiting/parked at junction exit etc.) |
| skidding\_and\_overturning | Whether the vehicle in the accident skidded (slide), skidded and overturned, jackknifed (articulated vehicle bending into a V-shape in a sliding motion), jackknifed and overturned; or overturned (rolled over) |
| hit\_object\_in\_carriageway | Type of object the vehicle hit in the carriageway (e.g. previous accident, road works etc.) |
| vehicle\_leaving\_carriageway | Location where the vehicle moved off the carriageway in the accident (e.g. nearside, nearside and rebounded etc.) |
| hit\_object\_off\_carriageway | Type of object that the vehicle hit not in the carriageway (e.g, road sign or traffic signal, lamp post etc.) |
| first\_point\_of\_impact | First point of impact on the vehicle in the accident (e.g. front, back etc.) |
| vehicle\_left\_hand\_drive | Whether the vehicle in the accident had its steering wheel (and other controls) on the left hand side or not |
| journey\_purpose\_of\_driver | Purpose of the journey of the driver in the accident (e.g. journey as part of work, commuting to/from work etc.) |
| sex\_of\_driver | Sex of the driver in the accident (male, female) |
| age\_of\_driver | Exact age of the driver in the accident |
| age\_band\_of\_driver | Age band of the driver in the accident (starts at 0 with a width of 5 (i.e. 0-5, 6-10 etc.) up until and including 75, last band is ‘over 75’ |
| engine\_capacity\_cc | Engine capacity in cubic capacity (cc) of the vehicle in the accident (usually ranges between 50cc to 1500cc) |
| propulsion\_code | Way in which the vehicle in the accident is “pushed forward” (petrol, heavy oil etc.) |
| age\_of\_vehicle | Exact age of the vehicle in the accident |
| generic\_make\_model | Model name of the vehicle in the accident (character string) |
| driver\_imd\_decile | The Index of Multiple Deprivation (IMD) measures how deprived the LSOA of the driver is; the IMD score is calculated using 7 domains (income, employment, education, health, crime, barriers to housing and services; and living environment);  (most deprived 10%, more deprived 10-20%, more deprived 20-30%, more deprived 30-40%, more deprived 40-50%, less deprived 40-50%, less deprived 30-40%, less deprived 20-30%, less deprived 10-20%, least deprived 10%) |
| driver\_home\_area\_type | Type of area that the home of the driver in the accident is in (urban area, small town or rural) |

The attributes in the casualty data are:

|  |  |
| --- | --- |
| accident\_index | A unique value for each accident |
| accident\_year | The year in which the accident took place |
| accident\_reference | ID used by the police to reference an accident (within a year) |
| vehicle\_reference | A unique value for each vehicle in a singular accident (i.e. two vehicles involved in the same accident will have the same accident\_index, but a different vehicle\_reference (as far as I can tell the vehicle\_reference is just the count of the vehicles: if two vehicles are involved, one gets vehicle\_reference = 1 and the other gets vehicle\_reference = 2)) |
| casualty\_reference | A unique value for each casualty in a singular accident (i.e. two people involved in the same accident will have the same accident\_index, but a different casualty\_reference (as far as I can tell the casualty\_reference is just the count of the casualties: if two people are injured (or dead), one gets casualty\_reference = 1 and the other gets casualty\_reference = 2)) |
| casualty\_class | ‘Role’ of the person injured (or dead) in the accident (driver or rider, passenger or pedestrian) |
| sex\_of\_casualty | Sex of the person injured (or dead) in the accident (male, female) |
| age\_of\_casualty | Exact age of person injured (or dead) in the accident |
| age\_band\_of\_casualty | Age band of the person injured (or dead) in the accident (starts at 0 with a width of 5 (i.e. 0-5, 6-10 etc.) up until and including 75, last band is ‘over 75’ |
| casualty\_severity | Severity of the injuries of the person in the accident (fatal, serious or slight) |
| pedestrian\_location | Location of the pedestrian (injured (or dead)) in the accident (e.g. crossing on pedestrian crossing facility, crossing in zig-zag approach lines |
| pedestrian\_movement | Movement that the pedestrian injured (or dead) was making at the time of the accident (e.g. crossing from driver’s nearside, crossing from nearside - masked by parked or stationary vehicle etc.) |
| car\_passenger | Location of the passenger (injured (or dead)) in the vehicle in the accident (not car passenger, front seat passenger or rear seat passenger) |
| bus\_or\_coach\_passenger | Movement/location of the passenger (injured (or dead)) in the bus or coach in the accident (boarding, alighting, standing passenger, seated passenger) |
| pedestrian\_road\_maintenance\_worker | ? |
| casualty\_type | Mode of transportation in the accident (e.g. pedestrian, cyclist etc.) |
| casualty\_home\_area\_type | Type of area that the home of the person injured (or dead) in the accident is in (urban area, small town or rural) |
| casualty\_imd\_decile | Decile of how deprived the LSOA of the person injured (or dead) in the accident is |

(e) Try to describe the data set in just few sentences! How is the data provided? Which kind of attributes are contained in the dataset? How large is the dataset in terms of the number of those elements (person, vehicles , geographic regions and locations, extra records and so on)?

Nearly all attributes are either numeric or coded to be numeric. The Road Safety Open Dataset Data Guide provides the character string label for each variable. Missing data or data that is out of range is coded to be -1. Attributes that are not applicable to every instance start at 0, whereas attributes that are applicable to every instance start at 1. Data that is not missing or out of range, but is unknown for some (other) reason is coded to be 9 if the number of possible values does not exceed 9; if the number of possible values does exceed 9, the data is coded to be 99. The accident dataset contains 36 attributes, the vehicle dataset contains 27 attributes and the casualty dataset contains 18 attributes. Together the three datasets consist of (36 + 27 + 18 – vehicle\_reference – accident\_reference (2x) – accident\_year (2x) – accident\_index (2x) =) 74 unique attributes. Possibly, when linked together, the combination of the three datasets will contain (81 – 3 =) 78 attributes, because accident will join vehicle on accident\_index (-1), then accident will join casualty on accident\_index (-1) and vehicle will join casualty on vehicle\_reference (-1).