Data Preparation
OpenRefine

Practical

Before starting the exercises below, make sure you have revised what has been covered so far.

1. Watch the three videos at http://openrefine.org/index.html

2. The file titanicwithErrors.xlsx contains part of modified data about the sinking of the `Titanic' on the night of April 15th, 1912. There is one line per person and are 5 attributes involved:

1) whether the person was travelling first, second or third class
2) whether the person was an adult or a child
3) whether the person was male or female
4) whether they survived or not.
5) the number of members travelled in a family

Examine/Understand the dataset. You are expected to use the knowledge about Understanding data to have a first look at the dataset. For example, for each attribute, find out the frequency of each value; from the data analytics point of view, what do you expect from the dataset?

3. Use Weka to repeat the examination in 2.

4. Clean the titanic dataset using OpenRefine, and export a "cleaner" dataset.

5. The file weatherbad.xlsx contains data about whether people played tennis or not under a given weather condition. There are 5 attributes involved:

1) whether the weather condition was sunny, overcast or rainy
2) the temperature
3) the humidity
4) whether it was windy or not
5) whether they played or not

Clean the data, using OpenRefine and export a "cleaner" dataset.

6. Convert the weather data you cleaned in question 5 to ARFF format by the user generated method, load it to Weka and examine the dataset.

7. Convert the titanic data you cleaned in question 4 to ARFF format by the user generated method, and load it to Weka and examine the dataset.

8. Study relevant references about OpenRefine and Weka.

9. Clean the Traffic dataset: roadAccidentShort - with errors.xlsx. See the data description in Appendix. Values might need to be checked together with other attributes/ For example, Column causal class indicates whether the person is a driver/rider or not, while column age of casualty shows the age. By law, age 1 is not allowed to drive or ride. Therefore if there is any instance with age as 1 and a driver/rider, one of these two values need to be changed. For example, the age can be changed to 18, which is the most common value.

For the following 3 exercises, Python is needed within OpenRefine.

10. Transform all values in the cleaned Titanic Revised dataset into nominal values
11. Transform all values in the cleaned Titanic Revised dataset into numeric values
12. On the cleaned Traffic dataset, transform values on casualty class (CASU_CLS), speed limit (SP_LIM) and Age of Casualty (AGE_CASU) from numeric to nominal values. (**Note**: values **might** need to be categorised).
13. Convert the datasets you cleaned/transformed in questions 10, 11, 12 and 13 to ARFF format by the system generated method, load it to Weka and examine the dataset.

Appendix: DATA DESCRIPTION

The file roadAccident - with errors.xlsx contains data about road accident recorded from 2000 to 2005 in UK. There are 10 variables (attributes) involved, and 20220 records (instances). The table below shows the metadata about the dataset.

Table: Variables and Values for Road Accident Data

Attribute Name	Value	Code description
ACCYEAR		Accident Year
	2000	Year 2000
	2001	Year 2001
	2002	Year 2002
	2003	Year 2003
	2004	Year 2004
	2005	Year 2005
RD_CLS		Road Class
	1	Motorway
	2	A(M)
	3	A
	4	B
	5	C
	6	Unclassified
SP_LIM		Speed Limit
	numeric	Miles/hour, e.g., 30, 60
JUNC_DET		Junction Detail
	0	Not at junction or within 20 metres
	1	Roundabout
	2	Mini-roundabout
	3	T, Y or staggered road
	5	Slip road
	6	Crossroads
	7	Multiple junction
	8	Private drive or entrance
	9	Other junction
LIGHT_COND		Light Condition
	1	Daylight - lights present
	2	Daylight – no lighting
	3	Daylight – lighting unknown
	4	Darkness – lights lit
	5	Darkness – lights unlit
	6	Darkness – no lighting
	7	Darkness – lighting unknown
WEATH_COND		Weather Condition
	1	Fine no high winds
	2	Raining no high winds
	3	Snowing no high winds
	4	Fine + high winds
	5	Raining + high winds

	6	Snowing + high winds
	7	Fog or mist
	8	Other
	9	Unknown
CASU_CLS		Casualty Class
	1	Driver or rider
	2	Passenger
	3	Pedestrian
SEX_CASU		Sex of Casualty
	1	Male
	2	Female
AGE_CASU		Age of Casualty
	numeric	e.g., 18, 25
SEVE_CASU		Severity of Casualty
	1	Fatal
	2	Serious

Like much of the data that companies store in data warehouses, this is genuinely historical data recorded by government, and much of the interest lies in trying to discover patterns within it. For example, under which conditions, an accidence would likely be a Fatal one. Unfortunately, there are a number of errors in the dataset. Before any kind of analysing, the dataset has to be cleansed fist.