MEI Introduction to Data Science

Lesson 2: Pre-processing and cleaning data
Activity 1 – Suggested student responses

This activity explores the main differences in weather between 1987 and 2015 at different the locations in the dataset.

Checkpoint 1

How many records are there in the Heathrow 2015 data set? How many fields are there in the Heathrow 2015 data set?

> *The data shape command returns the result (184,15) showing that there are 184 records (rows of data not including headers) and 15 fields (columns).*

Which numerical fields have been imported as *floating point* numbers and which have been imported as integers?

> *The output from the data.dtypes command lists the data types. The fields corresponding to Daily Mean Temperature and Daily Total Sunshine have been imported as floating point numbers. Daily Mean Windspeed, Daily Maximum Gust, Daily Maximum Relative Humidity, Daily Mean Total Cloud, Daily Mean Visibility, Daily Mean Pressure, Daily Mean Wind Direction and Daily Max Gust Direction have been imported as integers.*

Checkpoint 2

The command `heathrow_2015_data['Daily Mean Temperature'].describe()` gave you some summary values.

The summary value 'count' indicates the number of valid data records in the data set. What about the summary value 'mean'?

> *This calculates the arithmetic mean of the data. It is not possible to do this for Mean Cardinal Direction as this is categorical data.*

Go through all the summary values and say what you think each one tells you about the field 'Daily Mean Temperature'.

> *The values are:*

count	*184.000000*	*The number data items: temperature recorded on 184 days*
mean	*15.658696*	*The arithmetic mean of the daily mean temperature*
std	*3.156489*	*The standard deviation of the temperatures (measures spread)*
min	*8.000000*	*The smallest daily mean temperature*
25%	*13.275000*	*The lower quartile of the daily mean temperature*
50%	*15.300000*	*The median daily mean temperature*
75%	*18.100000*	*The upper quartile of the daily mean temperature*
max	*28.700000*	*The largest data daily mean temperature*

MEI Introduction to Data Science

Checkpoint 3

What differences can you identify between the temperature for the two years? Is this what you expected to see?

The values are:

count	*184.000000*
mean	*14.613587*
std	*3.575108*
min	*6.800000*
25%	*12.050000*
50%	*14.550000*
75%	*16.925000*
max	*23.500000*

Points you may have noticed:

- *Mean and Median for 2015 is about a degree higher than 1987. The quartiles are also higher in 2015.*
- *2015 had higher minimum and maximum daily mean temperatures*
- *2015 had slightly less variable temperatures than 1987*

All suggesting that 2015 was warmer on average and had more warmer days.

Checkpoint 4

Which value would you use to replace the string 'tr' in the rainfall field? Give your reasons.

You could comment on the impact of replacing it by different values. There is no single correct answer to this but you should be aware that removing it completely would distort the dataset.

Some appropriate (though possibly mutually exclusive) comments:

- *The midpoint for the range of values is 0.025 here so this is sensible.*
- *The rainfall has not been recorded to this degree of accuracy so you could argue that you should use zero as that is sufficiently accurate.*
- *There are already lots of days with zero rainfall and it is useful to be able to distinguish between days with no rainfall and others with a very small amount of rain.*

Looking at the summary statistics for 2015, the mean is 1.81 (3sf) and the SD is 5.43 (3sf) regardless of which value we use. The only statistic affected is the median, which unsurprisingly changes from 0.25 to zero.

Checkpoint 5

Use the statistics and charts produced to answer the initial problem:
What were the main differences in weather between 1987 and 2015 at the locations in the dataset?

You should comment on the two years for different weather stations using charts or statistics from multiple fields. The averages and measures of spread should be used to justify whether a change has been noticed and how variable the weather was in that year. Do the differences suggest a marked difference in the weather or could they be due to natural variation (e.g. would other years in the 2010s be different to years in the 1980s)?