Statistics for Data Analysis

josullivan@cct.ie

CCT College Dublin

Feedback on assignment

- Everybody has now received specific individual feedback regarding their assignments
- These slides are to provide some general feedback too, in order to help you as you begin to work on CA2

General feedback

- There were some nice plots produced by many students - pay attention to labels, legends, titles etc.
- Many had a clear layout/presentation, but some didn't - ensure to use a clear structure with relevant sections and sub-sections etc.
- It's essential that you comment on specific code output - for example, if you produce a correlation matrix, ensure that you interpret it. Say what pairs of variables are most highly correlated etc.
- Ensure that you proofread your assignment before submission.

General feedback

Remember to start simple:

- Provide a clear list of the variables and the data type of each variable - bullet points can be used here
- Then provide an appropriate univariate analysis of all relevant variables - this analysis depends on the data type
- Then you can proceed to bivariate analyses such as correlation matrices and visualisations of multiple scatterplots etc.

General feedback

Description alone is not enough.

- Important verbs to think about as you perform data analysis is to **explore**, **interpret**, **justify**, **analyse**, **evaluate**, and **critique**
- Explore the data thoroughly; interpret the plots and data summaries; justify any modelling decisions made; analyse the output from different model runs; evaluate and critique your results and your overall approach

Things to avoid

Some frequent errors or omissions that I noticed:

- Be very clear on the data type of each variable - for example, for ID variables (including latitutde and longitude, which are essentially IDs of specific locations), it doesn't make sense to find variances, produce histograms, etc.
- Don't explain in detail what something is (such as a correlation matrix), but then fail to interpret the specific correlation matrix for your data.
- Don't do *any* machine learning/modelling etc. before a full EDA, data summary, data visualisation etc.

Some useful links

Here are some useful links which help show how to clearly and logically explore a dataset before modelling it.

- This link is good - note how every plot has some bullet pointed observations following it. This is important to include.
- There is some good exploratory data analysis here too. Again, note how plots are commented on after being produced, and, as above, univariate plots are examined first.
- There are some useful pieces here, if you skip to the *Analyzing the data* tab.
- There are some useful steps and functions here, though this example doesn't actually comment on the output (and so is not a good guide for content).