Lab 3: Multivariate Exploratory Data Analysis

M. Muldoon <mark.muldoon@manchester.ac.uk>

Week of Monday, 7 October 2024

Supporting materials

To support you in your multivariate analyses, I've provided the following files, which you can use the lab session to discuss

- The auto data, `auto.txt`, and associated metadata `auto-mpg.names.txt`.
- R code, `auto-eda.R`, to produce some of the results shown in lecture.
- A jupyter notebook, `auto-python.ipynb`, that produced some of the other results shown in the lecture.
- A jupyter notebook, `deviate.ipynb`, that shows how to produce the multivariate datasets that deviate from normality as shown in the lecture.

The main lab assignment asks you to find and explore some data that interests you. Good places to look for data sets include:

- The University of California, Irvine, Machine Learning Repository:
  https://archive.ics.uci.edu/ml/index.php
- Kaggle, a company that organises data science competitions:
  https://www.kaggle.com/
- Information is Beautiful: a company that specialises in visualisation:
  https://informationisbeautiful.net/data/
- Data is Plural: A weekly newsletter of useful/curious datasets curated by Jeremy Singer-Vine:
  https://www.data-is-plural.com/
- Open data published by central government, local authorities and public bodies:
  https://data.gov.uk/
- The Greater Manchester Combined Authority Open Data pages:
  https://greatermanchester-ca.gov.uk/what-we-do/research/open-data/

while interesting code, discussions and examples of figures can be found at, for example, https://www.r-bloggers.com/.

Lab assignment

Find and a multivariate dataset that: (a) interests you, (b) was **not** presented in the lectures and (c) consists of real measurements (as opposed to being synthetic data), then do an exploratory analysis of it. As an exercise, you may prepare a 500-word report, which should contain the following sections:

1. A brief description of the data, including its origin and quality issues. You should imagine you are writing for a group who have no idea what this dataset is about!
2. A univariate exploratory data analysis of one of the columns in the dataset.
3. A multivariate exploratory data analysis, involving at least 2 columns.
4. R or Python code used to produce the analysis (don't count this against the word limit).

Illustrate your analysis with appropriate figures. If you wish, you can submit the report (by email, to `mark.muldoon@manchester.ac.uk`) for *informal* feedback, but this is **optional** and **will not affect your grade in the module**. If you wish to do this, please submit your report by **5:00 on Friday 11 October**.

A model solution will be published on Blackboard after this week's lab sessions have finished.