Exploratory Data Analysis

How do we get from data to answers? Exploratory data analysis is a process for exploring datasets, answering questions, and visualising results.

This week we will look to bring together your learning on this module so far and perform a full exploratory analysis for a specific case study. The tasks will be to clean and validate data, to visualise distributions and relationships between variables, and to use models to explain insight from the data.

Task 1 - Investigation into the Salary of Workers in the UK

A government research group has been asked to investigate the salaries of workers in the UK. They have collected information on each participant of the survey to determine whether they earn a "higher" or "lower" salary.

The information in the dataset is as follows:

- Workclass – Whether the worker is self-employed, private or another form of working category.
- Background_diversity – An algorithm that ensures that the members chosen in the come from diverse backgrounds.
- Qualification_Level – Specifies their qualification, and in some cases their resulting grade.
- Qualification_Level_ID – gives an associated number that directly correlates with Qualification_Level
- civil_status – The participant civil status.
- Profession – The area in which the participant works in (specific understanding is not important).
- Relationship – Their relationship status.
- Race – The participant's ethnicity.
- Sex – Is the participant male or female.
- Investment-profit – the participants profits from investments.
- Investment-losses – The participant's losses from investment.
- Hours-worked-per-week – Hours each participant works in their profession.
- Residency – Where the participant lives.
- Salary – Whether the participant earns a higher salary (greater than £40,000) or a lower salary (less than or equal to £40,000).

Task 2 – House Prices

The Housing dataset was compiled for use in data science education.

With 79 explanatory variables describing (almost) every aspect of residential homes in California this original Kaggle competition challenged you to predict the final price of each home.

https://www.kaggle.com/harrywang/housing

https://github.com/ageron/handson-ml/tree/master/datasets/housing