Unit 2: Exploratory Data Analysis (EDA)

This unit’s focus was on EDA. EDA is a critical initial step in research that involves examining data sets to summarise their main characteristics often with visual methods. It facilitates the detection of distribution, anomalies, or patterns to inform subsequent analysis and hypothesis testing. The essence of EDA, grounded in John Tukey's work from 1977, is to maximise insight into a dataset, uncover underlying structure, extract important variables, detect outliers and anomalies, and test assumptions (Mosteller and Tukey, 1977; Tukey, 1977) . The use of both graphical and non-graphical methods, ranging from histograms and scatter plots to summary statistics and correlation measures, helps in understanding both the breadth and depth of data. EDA's flexible approach allows data professionals to navigate data in a robust explorative manner to formulate clearer, insightful hypotheses and modeling strategies, enhancing the comprehension of complex data sets (Komorowski et al., 2016).

The knowledge acquired in this unit was reinforced through the practical application of EDA techniques on the Auto-mpg dataset. This dataset serves as a rich repository of automotive data encompassing various car models and attributes (Quinlan, 1993). EDA activities conducted on this dataset included essential checks such as assessing its size, identifying missing values, and evaluating skewness and kurtosis to understand the distributional characteristics of variables. To streamline these processes, the pandas profiling report library was utilised. This versatile tool consolidates EDA findings into a comprehensive summary, facilitating further exploration and interpretation of the dataset's nuances.

See Jupyter notebook and pandas profiling report below:

**References**

Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y. (2016). Exploratory Data Analysis. In: Secondary Analysis of Electronic Health Records. Springer, Cham. 185–203. <https://doi.org/10.1007/978-3-319-43742-2_15>

Mosteller, F. and Tukey, J.W. (1977) Data analysis and regression. A second course in statistics. Addison-Wesley series in behavioral science: quantitative methods.

Quinlan, R., 1993. UCI Machine Learning Repository. Auto MPG.

Tukey, J.W., 1977. Exploratory data analysis (Vol. 2, pp. 131-160). Reading, MA: Addison-wesley.