**COSC2753 – Machine Learning**

Individual Assessment 1

**Statement of Originality:**

We declare that the solution and work presented in this report are the results of our collective efforts and represent our original work. Any sources or references used have been properly acknowledged.

**Author Information:** Pham Phuoc Sang - s3975979

**Affiliations**: RMIT University, School of Science, Engineering & Technology

**Contact Details:** Pham Phuoc Sang - [s3975979@rmit.edu.vn](mailto:s3975979@rmit.edu.vn)

**Date of Report:** April 4, 2025

# Introduction

# Objectives

# Exploratory Data Analysis (EDA)

In the Exploratory Data Analysis (EDA) phase, our primary goal is to develop a thorough understanding of the data's structure, identify underlying patterns, and ensure that the data meets the necessary assumptions for linear regression, as highlighted by [1]. We will start with data preprocessing, which entails loading the data from the file and examining the dataset's size and any missing values. Next, we will leverage our domain knowledge to guide variable selection, supplemented by correlation analysis, as suggested in [2]. Following this, we will systematically assess the key assumptions of regression, including linearity, normality of residuals, homoscedasticity, independence, and the absence of multicollinearity, as discussed in [3]. To validate these assumptions, we will employ specific visualizations: QQ plots will evaluate the normality of residuals, boxplots will identify outliers that may influence normality and variance, distribution plots will illustrate variable distributions for normality checks, and heatmaps will reveal correlations to detect multicollinearity, in accordance with the outlined methodologies.

## Assumption of regression

Regression assumptions play a crucial role in linear regression analyses, as they ensure that the model's predictions and statistical inferences are both valid and reliable. Violating these assumptions can lead to biased estimates and misleading conclusions, which may adversely affect decision-making. Bakker and Wicherts (2011) emphasize that such violations can result in misreported findings in research [1]. The key assumptions include:

**- Linearity**: This assumption requires a linear relationship between independent and dependent variables to achieve accurate predictions. Violations of this assumption can introduce bias (Towards Data Science, 2021) [2].

**- Normality of Residuals**: This is essential for conducting valid hypothesis tests, particularly in small sample sizes. Deviations from normality can invalidate the results (Statistics Solutions, 2010) [3].

**- Homoscedasticity**: This assumption demands that the residuals have constant variance. When heteroscedasticity is present, it can lead to inefficient estimates (GeeksforGeeks, 2024) [4].

**- Independence**: Observations must be uncorrelated to avoid biased predictions (Statology, 2020) [5].

**- No Multicollinearity**: This requires low correlation among predictor variables to ensure stable coefficient estimates. High correlations can lead to overfitting (Statistics Solutions, 2010) [3].

# References

[1] C. Chatfield, “Exploratory data analysis,” *European Journal of Operational Research*, vol. 23, no. 1, pp. 5–13, Jan. 1986, doi: doi.org/10.1016/0377-2217(86)90209-2.