# Data Manipulation and Detection

Tuvshin Selenge
Immatrikulationsnummer: 11815791
e11815791@student.tuwien.ac.at

**Main supervisors:**
Assoz.Prof. PD Dr. Jakob Müllner, Wirtschaftsuniversität Wien
**Co-supervisor:**
(Not decided yet), Technische Universität Wien

**Domain-specific lecture:** 4366 Foundations of International Business (WU Wien)



# 1 Project Title

**Tracing the Mechanisms of Data Falsification: An Empirical Study on Data Manipulation Techniques and their Detection**

# 2 Project Overview

## 2.1 Data

The data for this empirical study will be the well-known cars dataset covering the period from 1970 to 1982, which is now considered a classic for exploratory analysis and hypothesis testing. The dataset includes 11 variables and 398 observations, making it a decent size for both data manipulation and detection purposes.

## 2.2 Planned Workflow and Methodology

### 2.2.1 Statistical Analysis and Hypothesis Creation

First, we will perform an exploratory data analysis to visualize and understand all the relevant variables. Based on these variables, three hypotheses will be formulated.

For each hypothesis, if the observed outcome does not match the desired result (for example, whether the hypothesis is true or false), the data will be deliberately manipulated to force the outcome to the desired direction. This manipulation will be applied consistently across all hypotheses.

This process will form the basis for our further analysis, which focuses on detecting manipulated data. In a later stage, once we successfully detect the falsified data, we will investigate how it was manipulated.

#### 2.2.2 Data Manipulation

The dataset will be manipulated as described above to alter the ground truth. This new dataset will also serve as the basis for analysis by a colleague who is undertaking a similar project with different or similar methods to identify the parts of the data that were falsified. This approach will help determine whether the methods developed individually are applicable to unseen data, for example data on which no machine learning model was pre-trained. This leads to the next part of the study, identifying and applying methods to detect data manipulation and determine how it was accomplished.

# 3 Methods to Detect Data Manipulation and Determine How It Was Manipulated

## 3.1 Detecting data manipulation

Since we are working with unseen data, we will rely on unsupervised models and statistical approaches. Detecting data manipulation is as challenging as identifying outliers, because manipulated values deviate from the true ones. As described by Raymaekers and Rousseeuw in their paper [1], detecting outliers at the cell level is not trivial. Therefore, we will explore the following methods:

1. **Statistical Approaches:**
   By using z-scores and distributional analysis, this method aims to flag marginal outliers or manipulations in multivariate numerical data.

2. **Cellwise Detection:**
   This method treats each cell in the data matrix as potentially contaminated. The model is represented as:
   $$X = (I - B)\,Y + B\,Z,$$
   where:

   - $X$ is the observed value in a cell,
   - $I$ is the identity matrix,
   - $Y$ is the true (uncontaminated) value,
   - $Z$ represents the contaminating (noise) value, and
   - $B$ is an indicator drawn from a Bernoulli distribution (with $B = 1$ indicating a contaminated cell and $B = 0$ otherwise).

   This framework allows individual cells to be contaminated without necessarily affecting entire rows. An R package, `cellWise`, has been developed to implement this model [1].

3. **Isolation Forest:**
   Isolation Forest is an ensemble method for anomaly detection [2]. The algorithm builds trees by randomly selecting a feature and a split value, partitioning the data into subgroups. Anomalies are typically isolated quickly because they do not belong to dense clusters. In other words, if an anomalous point is separated from the majority within just a few splits (resulting in a short path length), it indicates abnormality. Normal points, which are part of larger, denser clusters, require more splits to isolate, resulting in longer path lengths.

4. **Outlier Detection with LOF (Local Outlier Factor):**
   LOF is an unsupervised anomaly detection method that computes the local density deviation of a given data point with respect to its neighbors [3]. Samples with substantially lower density than their neighbors are flagged as outliers.

### 3.2 Determining How It Was Manipulated

To determine how the data was manipulated, the first step is to accurately identify which values have been altered. Once these manipulated cells are detected, we will conduct a statistical analysis to reverse-engineer the manipulation process. This involves comparing the flagged values with the majority of data in the specific variables to assess their deviations, and to infer the modifications applied.

## 4 End Results of the Data Manipulation Detection

By the end of this project, we aim to develop a robust method that can not only pinpoint the manipulated values but also uncover the processes used in the manipulation. The objective is to design an approach capable of detecting anomalies, outliers, or data manipulations in an untrained dataset, mimicking real-world scenarios where verifiers do not have access to the underlying true data. Such a method could prove invaluable for identifying p-hacking or other similar attempts at data manipulation.

# References


[1] C. Raymaekers and P. Rousseeuw, *Challenges of Cellwise Outliers*, ScienceDirect, https://www.sciencedirect.com/science/article/pii/S2452306224000078.

[2] scikit-learn, *Isolation Forest*, Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html.

[3] scikit-learn, *Local Outlier Factor (LOF)*, Available at: https://scikit-learn.org/dev/auto_examples/neighbors/plot_lof_outlier_detection.html.