# Title: Analysis of Employee Dataset

Prepared by:

Ziad Ahmed Mohamed 42010617

Abdelrahman Ahmed 42010611

Abanoub Ragheb 42010601

Ahmed Gamal Abdelwahed 42010555

Mohamed Mohamed 42010249

Mostafa Hamza 42010521

Date: 5 / 30 / 2023

# Problem Definition

The purpose of this analysis is to explore and gain insights from a dataset containing information about employees, including their income and various demographic factors. The dataset includes attributes such as age, marital status, education level, occupation, commute distance, and the number of cars owned. The main objectives of this analysis are as follows:

* Determine the average income of the employees in the dataset.
* Calculate the percentage of employees earning more than $50,000.
* Identify the percentage of employees who have purchased a bike.
* Determine the most common occupation among the employees.
* Calculate the number of employees in the dataset who have no children.
* Determine the average number of cars owned by the employees.
* Identify the number of employees living in the Pacific region.
* Determine the average age of the employees.
* Calculate the percentage of employees who have a commute distance of 5-10 miles and own a home.
* Identify the most common commute distance in the dataset.
* Determine the most common gender among the employees.
* Calculate the average income of male employees.
* Calculate the average income of female employees.
* Determine the percentage of male employees in the dataset.
* Calculate the percentage of female employees in the dataset.
* Determine the number of employees who have 2 or more cars and an income less than $50,000.
* Calculate the percentage of individuals who are homeowners and have purchased a bike.
* Determine the highest income in the dataset.
* Calculate the number of employees with a partial college education.
* Determine the number of employees over 50 years old.
* Calculate the percentage of male employees over 50 years old.
* Determine the number of employees with a skilled manual occupation.
* Identify the ID of all rows with the highest income.
* Calculate the number of employees with a graduate degree.
* Determine the average income in Europe.
* Determine the average income in the Pacific region.
* Identify the most common marital status in the dataset.
* Calculate the average income for single and married employees.

# Method

To address the problem statement and achieve the objectives, the following methods were employed:

Data Preprocessing:

* The dataset was loaded using the Pandas library in Python.
* Various data cleaning steps were performed, including handling missing values, converting data types, and standardizing categorical values.
* Data exploration and descriptive statistics were conducted to gain a better understanding of the dataset.

Data Analysis:

* Key metrics and insights were extracted from the dataset using pandas functionalities.
* Summary statistics were calculated to answer specific questions about income, occupation, age, gender, etc.
* Visualizations, including bar plots, box plots, heatmaps, and scatter plots, were created to illustrate patterns and relationships in the data.

Feature Engineering:

* Additional features were created based on existing variables, such as age squared, income multiplied by the number of cars, and children multiplied by the number of cars.
* Binning was performed on the income variable to create income categories for classification.

Machine Learning:

* The dataset was split into training and testing sets.
* Feature selection was conducted using the ANOVA F-test to select the most relevant features for classification.
* Class imbalance was addressed using random oversampling.
* A logistic regression model was trained on the resampled data to predict income categories.
* The trained model was evaluated using accuracy as the performance metric.

# Experiment

The analysis was performed on the provided employee dataset using Python programming language and its libraries like “pandas, numpy, seaborn, matplotlib, mpl\_toolkits, sklearn, imblearn ”.

# Reference

Kaggle:

<https://www.kaggle.com/datasets/heeraldedhia/bike-buyers?resource=download>