Week 3: Data Manipulation and Preparation with Pandas

## Assignment Description

## Objective

## This assignment will test your ability to manipulate and prepare data using Pandas. You will work with the “netherlands\_IT\_salaries.xlsx” Dataset, cleaning and transforming the data step by step. The goal is to practice data sorting, handling missing values, removing duplicates, detecting outliers, merging, applying functions, and feature engineering.

## Dataset: IT Salaries

## You will use a dataset containing information about IT professionals in the Netherlands, including salary, experience, role, education, and location. The dataset has missing values and can be used to demonstrate various data manipulation techniques.

## Dataset Columns:

## Year: The year of data collection.

## Age: The age of the employee.

## Role: The job title.

## City: The city where the employee works.

## Education: The education level.

## Experience: Years of experience in the IT field.

## Salary: Monthly salary.

## Task

**1. Sorting Data**

* Sort the dataset by Salary in descending order.
* Sort by City and then by Role, both in ascending order.

**2. Handling Missing Data**

* Check if there are any missing values in each column.
* Display only the rows with missing values.
* Fill missing Education values with 'Unknown'.
* Replace missing Salary values (if any) with the average salary.
* Drop rows where City is missing.
* Decide what to do with missing values in Experience.

**3. Removing Duplicates**

* Check if there are any duplicate rows.
* If duplicates exist, remove them while keeping only the first occurrence.

**4. Detecting Outliers**

* Use the Interquartile Range (IQR) method to identify potential outliers in Salary.
* Display the rows containing salary outliers.
* Remove detected outliers from the dataset.

**5. GroupBy Operations**

* Group the data by City and calculate:
  + The average salary in each city.
  + The number of roles in each city.
* Group the data by Education and find the maximum salary for each education level.
* Group the data by Seniority and find the median salary for each seniority level.

**6. Apply and Map Functions**

* Add a new column Annual Salary by applying a function to calculate Salary x 12.
* Use the .map() function to categorize education levels into 'Low', 'Medium', and 'High':
  + MBO, Mavo, and Unknown → 'Low'
  + HBO → 'Medium'
  + WO → 'High'
  + Store the new values in a column called Education Level.

**7. Feature Engineering**

* Create a new column Year of Birth using the formula: Year - Age.
* Create a new column Seniority based on experience:
  + Less than 5 years → 'Junior'
  + Between 5 and 10 years → 'Mid-level'
  + More than 10 years → 'Senior'
* Create a new column City Category:
  + If the city is 'Randstad', 'Amsterdam', or 'Utrecht', categorize it as 'Urban'.
  + Otherwise, categorize it as 'Non-Urban'.
* Create a new column Salary Bracket based on salary:
  + Less than €3000 → 'Low'
  + Between €3000 and €5000 → 'Medium'
  + More than €5000 → 'High'.

**8. Advanced Analysis**

* Find the top 5 highest-paying roles.
* Find the most common role in the dataset.

**9. Save the Cleaned Data**

* Save the final cleaned dataset to a CSV file named cleaned\_IT\_salaries.csv.

## Submission Instructions

* Submit your Jupyter Notebook (.ipynb) file with the naming format: M2\_A3\_groupX.ipynb (replace X with your group number).
* Ensure all tasks are clearly labeled with comments explaining each step.
* Include the final cleaned dataset (cleaned\_IT\_salaries.csv) in your submission.
* **Note**: Remember that only the team leader of each group will submit the solution on behalf of the group.