OLA 1- Data Wrangling and Descriptive Statistics

The following OLA gives you an opportunity to work on the first stages of a machine learning / AI project. It is a group project and is intended to give you a chance to practice using typical Numpy/ Pandas/ matplotlib and Seaborn functions to prepare and visualize data using simple plots.

DUE DATE: Week 9 (26 February) - hand in on Moodle

Task 1: Data Exploration and Cleaning

Objective: Understand the dataset's structure, clean the data, and handle missing values.

Tasks:

1. **Find and Download a Dataset:** Use Kaggle to find a dataset of interest. It should have both numerical and categorical data and some missing values.

2. **Data Exploration:**
   - Load the dataset using pandas.
   - Use .describe(), .info(), and .head() to explore the dataset's structure, summary statistics, and first few rows. Use comments or markdown cells to explain the data.

3. **Data Cleaning:**
   - Identify columns with missing values.
   - For numerical columns, interpolate missing values.
   - For categorical columns, replace missing values with the mode or another standard technique (such as the mean of the two adjacent data points)
   - Drop columns with more than 50% missing values.
   - Drop extreme outliers and explain why they were considered outliers

4. **Data Visualization:**
   - Use matplotlib or seaborn to visualize the distribution of variables both quantitative and categorical (parametric and non-parametric)

Task 2: Feature Engineering and Descriptive Statistics

Objective: Enhance the dataset with new features and then use descriptive statistics to explain the distribution of the data.

Tasks:

1. **Feature Engineering:**
    - Create a new feature by binning a numerical variable into categories (e.g., low, medium, high). Put ranges (eg age, into three or four groups rather than a continuous distribution)
    - Implement one-hot encoding for a categorical variable.

2. **Descriptive Statistics:**
    - Calculate the mean, median, and standard deviation for numerical features.
    - For categorical features, count the frequency of each category.

3. **Visualization:**
    - Use seaborn to create box plots for numerical features to identify outliers.
    - Visualize the distribution of categorical features using bar plots.

Task 3: Data Wrangling and Analysis

Objective: Perform data wrangling to prepare data for analysis and conduct simple analysis to extract stories about the data - what can we say about this data?.

Tasks:

1. **Data Selection and Wrangling:**
    - Select a subset of columns relevant to a hypothetical question of interest (e.g., predicting a target variable).
    - Use .groupby() to aggregate data and calculate mean values for each category of a selected categorical variable.
2. **Data Analysis:**
    - Use seaborn to create scatter plots to visualize relationships between pairs of numerical variables.(X an Y axis are used for the variables)
    - Create a pairplot to visualize the pairwise relationships in the dataset.

      See https://seaborn.pydata.org/generated/seaborn.pairplot.html

3. **Insights:**
    - Based on the visualizations and descriptive statistics, write down 3 insights about the dataset.

Deliverables

- A Jupyter notebook containing all the code used.
- A short report (2-3 pages) summarizing the findings and insights from the exercises, including figures and tables as necessary. (This can but does not have to be done in LaTeX! It could be done in Markdown format inside the Jupyter Notebook)
- Submit work to the hand-in folder via the Moodle course - it is a group report but please hand in individually with all group member names and group name easy to find.