Final project

For the final project, you may choose **ONE dataset** from the list provided below or select your own dataset, preferably containing at least 50,000 observations. Please ensure that the selected dataset contains sufficient data and appropriate variables to successfully complete all required tasks in this project, including exploratory data analysis, hypothesis testing, and regression analysis.

The project may be completed **individually or in groups of up to three (3) students**.

If you choose to work in a group:

- all group members must be clearly listed on the title page of the report;
- all members share equal responsibility for the quality of the analysis, code, and interpretation.

Dataset links:

- Melbourne House Prices: https://www.kaggle.com/datasets/thedrzee/melbourne-house-prices-dataset
- Global AI & Data Science Job Market (2020–2026): https://www.kaggle.com/datasets/mann14/global-ai-and-data-science-job-market-20202026
- Hotel Booking Demand: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand
- Employee Performance and Productivity Data: https://www.kaggle.com/datasets/mexwell/employee-performance-and-productivity-data
- Student Depression and Lifestyle (100k): https://www.kaggle.com/datasets/aldinwhyudii/student-depression-and-lifestyle-100k-data

Academic Integrity:

- You may use AI tools for the coding part only (e.g., data loading, preprocessing, and implementation of statistical methods).
- The analytical report must be written by the student(s) and reflect their own understanding of the analysis.
- The use of AI-generated text in the report is not permitted. If AI-generated content is detected, the final grade will be reduced proportionally to the estimated percentage of such content.

Project Structure and Requirements

1. Data Preparation and Exploratory Data Analysis (EDA)— 10 points

Your exploratory analysis must include the following components.

a. Data loading and structure

- load the dataset;
- report the number of observations and variables;
- identify data types;
- check for missing values;
- provide a brief description of the dataset and its context.

b. Data cleaning and preparation

Describe and justify any preprocessing steps applied to the data.
This may include, but is not limited to:

- handling missing values;
- renaming or recoding variables;
- splitting or merging columns;
- converting units of measurement;
- filtering observations;
- creating new variables if necessary.

*All data preparation steps must be clearly explained and justified.

c. Visualization (at least three plots)

- choose appropriate visualizations based on variable types;
- include at least three different plots (e.g., histogram, boxplot, scatterplot, grouped bar chart);
- all plots must have titles and labeled axes.

d. Initial insight

Briefly comment on which variables appear to be most strongly associated with the main outcome variable.
Justify your reasoning using descriptive statistics and visualizations.

2. Hypothesis Testing — 30 points

All hypothesis tests must be conducted using a **significance level of $\alpha = 0.05$**.

You must perform and report the following statistical tests:

a. One-sample t-test
b. Two-sample t-test

- check whether the variances are equal;
- conduct the appropriate version of the test and justify your choice.

c. Paired t-test

- clearly explain why the observations are paired.

d. Correlation analysis

- Pearson correlation;
- Spearman correlation;
- explain why the results may differ.

e. One-way ANOVA

- compare means across three or more groups.

For **every statistical test**, you must clearly state:

- the research question;
- null hypothesis (H_0);
- alternative hypothesis (H_1);
- justification for using the test;
- interpretation of the results in context.

3. Regression Analysis — 20 points

a. Simple Linear Regression

- fit a regression model with one predictor;
- interpret the slope and intercept in the context of the data;
- report R^2 and comment on goodness-of-fit.

b. Multiple Linear Regression

- fit a regression model using **at least three predictors**;
- interpret all regression coefficients;
- report R^2 and adjusted R^2;
- briefly discuss the limitations of the model.

4. Analytical Report — 30 points

You must submit a written analytical report that clearly explains the analysis and results in your own words and demonstrates a solid understanding of statistical concepts.
The provided report template should be used as a guideline. The content of each section may be adapted based on the chosen dataset and analysis; however, the overall structure of the report is recommended.

Please ensure that the report is clearly written and well-formatted. Avoid directly copying or pasting AI-generated text, do not remove paragraph indentation, and avoid excessive use of bold formatting.

5. Defense — 10 points

During the project defense, each student (or group) will be asked **one question by the instructor** related to their analysis.

Submission requirements

Please submit ONE ZIP or folder named as *Name_Surname_Assignment2.zip* containing:

1. Jupyter Notebook (.ipynb)
   - all Python code,
   - outputs,
   - brief comments explaining steps.
2. Report (PDF-formatted)