Home Assignment 02

Marco Nieddu

Data Analysis for Economics – Fall 2025

Who are we building on?

This lecture builds

- on the materials of the book by Békés & Kézdi, *Data Analysis for Business, Economics, and Policy* (2021);
- on the lecture slides from Pedro H.C. Sant'Anna: https://psantanna.com/Econ520/index.html.

Background

- The data were obtained by web scraping a price comparison website.
- The dataset reports, for each accommodation, prices and other features such as reviews, stars, and distance from the city center.
- The data were collected on a weekday in November 2017.
- Our goal is to explore this dataset and understand the relationship between price and other variables.

Part I: Getting started

Folder structure

- Ensure that you have a well-structured folder for your project.
- Suggested structure:
  - `class04/source`
  - `class04/clean`
  - `class04/codes`
  - `class04/output`

Getting the data in

- You can access the data on the course DB folder.
  - It is a read-only folder, so copy the data into **your** folder.
- Start a Python script and set the working directory.

- Load the data from `hotelbookingdata-vienna.csv` into Python as a dataframe.
- How many observations and variables are in the dataset?

AI prompt: *"How can I check the number of rows and columns of a dataset in pandas, and print its first few lines?"*

Part II: Analyzing the data

Duplicates and basic summaries

- Do we have any accommodation with multiple entries? If so, how many?
- Are they duplicates? If so, keep only one of them.
- What is the average price of the accommodations in the dataset?

AI prompt: *"How can I identify and remove duplicates in a pandas dataframe, keeping only the first occurrence?"*

Types and missingness

- How many observations are hotels, hostels, and other types of accommodations?
- Create a new variable `type` that classifies the accommodations into these three categories.
- What is the proportion of missing average customer ratings in the entire dataset?
- What is the proportion of missing average customer ratings by type of accommodation?

AI prompt: *"How can I calculate the share of missing values overall and by category in pandas?"*

Cleaning text-based numeric variables

- Some variables are stored as text, such as review scores (e.g. "4.3/5") and distances ("1.2 miles").
- Clean and convert them into numeric format to make them usable for descriptive and graphical analysis.

AI prompt: *"How can I extract the numeric part from a text string in pandas and convert it to a float?"*

Prices by type

- How does the average price vary by type of accommodation?
- Produce a table with the average price, standard deviation, median price, and the number of observations by accommodation type.

- Save the table as a CSV file in the `tables` or `output` folder.
- Produce a bar plot with the average price by accommodation type.
- How does the bar plot look? Can you make it more readable (e.g. colors, order, labels)?

AI prompt: *"How can I compute group statistics (mean, median, standard deviation, count) and plot them as a bar chart in pandas or matplotlib?"*

Price vs rating

- How do price and rating vary? Produce a scatter plot of price vs. rating.
- How do they vary by type of accommodation?
- Add a linear fit, legend, title, and axis labels to the scatter plot.
- Use transparency (`alpha`) to reduce overplotting and distinct colors for each type.
- Save the scatter plot as a PNG file in the `output` folder.

AI prompt: *"How can I create a scatter plot with different colors for each category and add a linear trend line in matplotlib?"*

Advanced (optional): regression fit

- Fit a linear regression of price on rating using `statsmodels`.
- Report the estimated relationship and its interpretation.
- Repeat the analysis by accommodation type: does the relationship differ?

AI prompt: *"How can I run a simple linear regression in Python with statsmodels?"*

Part III: More on visualization

Distributions and outliers

- Construct a histogram for the average star rating.
- Construct a histogram for the average number of reviews.
- Construct a histogram for the average distance from the city center.
- Are there any extreme values? Are they outliers?
- Clip or filter extreme distances (e.g. greater than 8 miles) before visualization.
- Construct a density plot for the average price.
- Construct a histogram for the average price among hotels only, and add a density plot.

Comparative distributions

- Construct a box plot for the average price by accommodation type.
- Construct a violin plot for the average price by accommodation type.
- Which type has the largest variation? What might explain it?

Saving outputs

- Save all plots with clear filenames (e.g. `price_by_type.png`, `scatter_price_rating.png`).
- Store summary tables and cleaned data in your `output` folder for reproducibility.

AI prompt: *"How can I save my figures and data tables automatically to a specific folder using Python?"*

Part IV: Conclusion and interpretation of findings

- What have you learned in this exercise?
- Which factors seem most correlated with price?
- What economic mechanisms could explain the observed relationships?
- What would be the next steps in this analysis?
- What are the limitations of your current dataset (e.g. missing amenities, location precision)?

Submission

If you would like feedback, please send your Python script and exported plots to: `mgnieddu@unica.it`