Data Analytics I
Assignment 1

Release: 12 August 2025
Deadline: 23 August 2025 (11:55 pm)

The objective of this assignment is to introduce you to data visualization and develop an understanding of data processing using Python.

For this assignment, you are provided with two comprehensive datasets: one containing information on adult census demographics and income, and the other detailing various real estate properties.

Use visualisation techniques such as histograms, distribution plots, scatter plots, quantile plots, correlation matrices, and categorical groupings as a basis for constructing your inferences and answers. The suggested techniques listed after each question are provided for reference; you are encouraged to go beyond these and incorporate more in-depth, creative, and insightful analyses and plots where appropriate.

Part I [30 Marks]

1. Education Distribution & Grouping [5 Marks]

Identify the unique education levels and their frequencies. Group them into broader, meaningful categories (e.g., elementary, higher education, etc.). Explain the reasoning behind your mapping and support your answer with visual evidence.

2. Age–Work Intensity Relationship and Grouping [10 Marks]

Examine the distributions of age and hours worked per week. Create broader groups for each variable (e.g., part-time, full-time, overtime for work intensity) and compare category counts and distributions before and after grouping. Analyse the relationship between age and work intensity before and after grouping. Comment on whether grouping improves interpretability or reduces useful detail, supported by statistical evidence and visualisations.

3. Capital Gains/Losses and Group Performance [10 Marks]

Analyse the distributions of capital gain and loss, both overall and for non-zero values, and report the proportion of individuals with any capital activity. Using your age and work intensity groupings from the prior question, compare groups on average net capital (Capital Gain - Capital Loss) and proportion with capital activity. Illustrate patterns using visualisations. Discuss whether age or work intensity shows a stronger association with net capital, and how grouping impacts interpretability.

4. Final Dataset Refinement and Structure [5 Marks]

Summarise the final refined dataset structure, noting newly created, transformed, categorised, or grouped features, and any that were removed. Compare category counts before and after grouping for key variables, highlight changes in missing values, and describe how these refinements improve interpretability and readiness for modelling.

Part II [60 Marks]

1. Price Segmentation & Market Overview [7.5 Marks]

Divide the properties into three **price ranges** and analyse their distribution across **cities**. Provide a high-level market summary for each price range, incorporating **property type**, **city**, and key **amenities** to give investors a quick snapshot of the market.

2. City-Level Comparative Analysis (Mumbai vs Thane) [7.5 Marks]

Compare investment opportunities in **Mumbai** and **Thane** by analysing differences in **property types**, **Carpet Area**, and **prices**. Highlight both **residential** and **commercial** segments for a complete picture.

3. Location-Based Premium Analysis [5 Marks]

Within each city, compare **high-budget prime-location** properties to similar **non-prime-location** properties in terms of average **Carpet Area**, **amenities**, and **price per square foot**. Identify whether **location** justifies the price premium.

4. Value-for-Money Opportunities [10 Marks]

Identify properties offering the **highest Carpet Area per unit price** across all cities. Present **city-wise rankings** and insights into which segments deliver the best value for **budget-conscious investors**.

5. Feature & Amenity Impact on Price [5 Marks]

Analyse how various **amenities** (e.g., Swimming Pool, Gymnasium, Club House) and **developer reputation** affect property **prices**. Determine which features add the most value and whether preferences differ across **cities**.

6. Timeline & Readiness Effect on Pricing [10 Marks]

Investigate how **Possession Status** (ready-to-move vs under-construction) and "**Availability Starts From**" dates influence **prices** in Mumbai and Thane. Highlight trends and investor takeaways for **short-term** vs **long-term** planning.

7. Developer Impact on Properties [15 Marks]

Analyse the impact of **developers** on **property prices** and **features**. Determine if certain developers are associated with **higher-end properties** or **better amenities**.

Part III [10 Marks]

1. Code Quality [5 Marks]

Ensure the code is written in a **vectorised** way for **data preprocessing** using **pandas**, instead of writing individual `for` loops (wherever possible).

2. Report [5 Marks]

Based on your analyses from the previous questions, create a **comprehensive report** for both datasets. This should include data cleaning methods used as well as a small summary of findings in each dataset.