Data Analytics Assignment 1 Report

Team Id:12
Aryan Bansal(2021111018)
B V G Poojitha(2024801011)

1 Part I: Adult Census Dataset Analysis

Data Cleaning and Preprocessing

The dataset was cleaned by removing the non-informative `fnlwgt` column and converting `Age` to integer format.The RangeIndex had 48843 entries, but some columns have 48842 non-null values, one row is almost empty and should be removed. Missing values in `Workclass`, `Occupation`, and `Country` were imputed using logical defaults and mode. Float columns were converted to integers for memory efficiency. Two new features were created: `Has_Capital_Activity` and `Net_Capital`. The cleaned dataset was saved as `final_dataset_1.csv`.

Education Distribution and Grouping

The original 16 education levels were grouped into six categories: Compulsory Education, High School, Some College, Associate's, Bachelor's, and Advanced Degree. This reduced dimensionality and improved interpretability. The original Education feature contains 16 categories, and most of them have small, infrequent counts. The new Education Group reduces this complexity by consolidating these categories into broader groups, making it more interpretable. The new Education Group has a more ordered progression from least to most education.

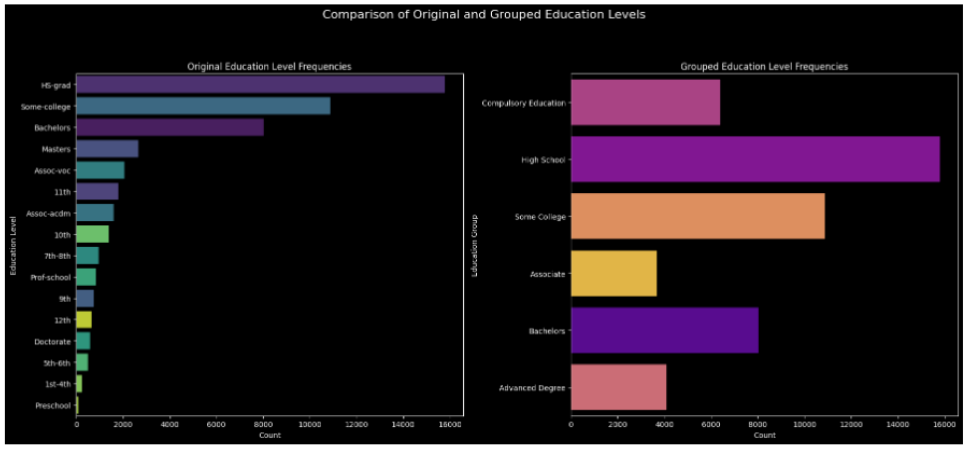

Figure 1: Original vs Grouped Education Level Frequencies

Age–Work Intensity Relationship

Age was grouped into four life stages and work hours into four intensity levels. Heatmaps and hexbin plots revealed that work intensity peaks in middle age and declines in older groups.There is an extremely dense band of people working exactly 40 hours per week, spanning from young adulthood to late middle-age, however, it's difficult to make precise comparisons. For example, it's hard to tell if "Young Adults" work more part-time hours than "Senior Adults".The heatmap on the right simplifies the relationship into discrete categories. The color intensity and the annotated numbers represent the count for each combination.The largest single group is Middle-Aged (31-45) individuals working Full-Time (40), with over 9,000 people. Young Adults (17-30) are the most likely group to work Part-Time.Work intensity appears to peak in middle age. The Middle-Aged and Senior Adult groups have the highest counts for Overtime and High-Intensity work. Elderly (61+) individuals are predominantly in the Part-Time category, which aligns with expectations of retirement and reduced work hours.

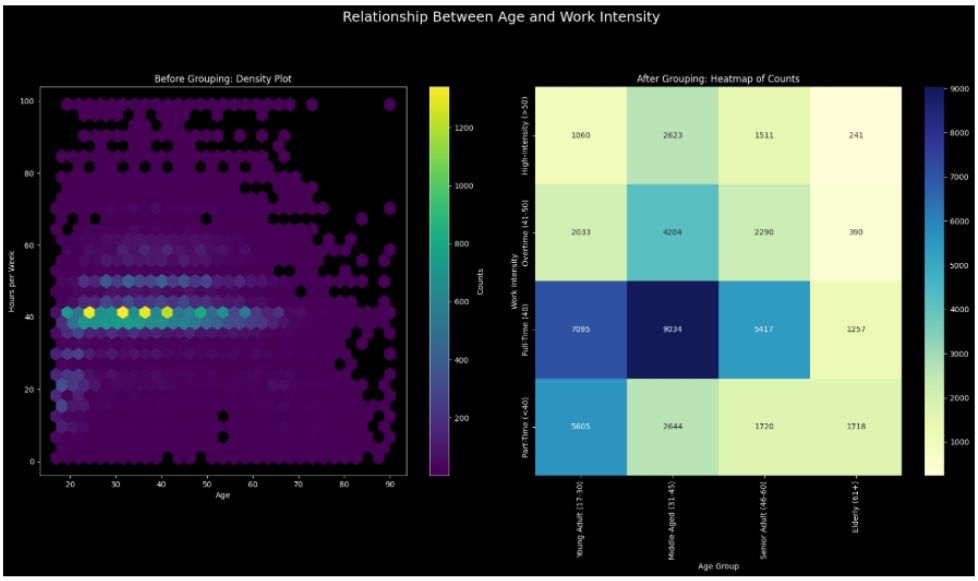

Figure 2: Heatmap of Age Group vs Work Intensity

Capital Gains/Losses and Group Performance

Only 13% of individuals had capital activity. Net capital and activity proportion increased with both age and work intensity.The main drawback is the loss of granularity. The Part-Time (¡40) category, for instance, treats someone working 5 hours a week the same as someone working 35 hours. Similarly, the Age Group Young Adult (17-30) combines people just starting their careers with those who are well-established. The weak Pearson correlation failed to capture the pattern that work hours tend to increase from young adulthood, peak in middle age, and then decrease, a non-linear trend that the grouping helps to reveal.

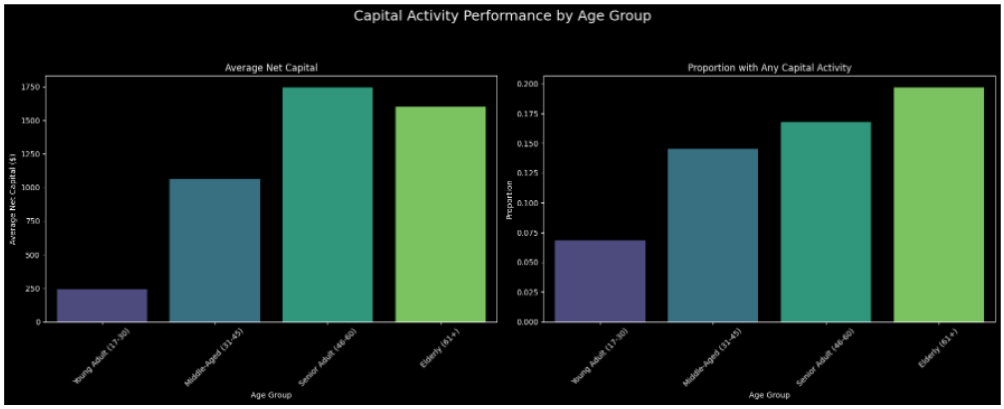

Figure 3: Capital Activity by Age Group

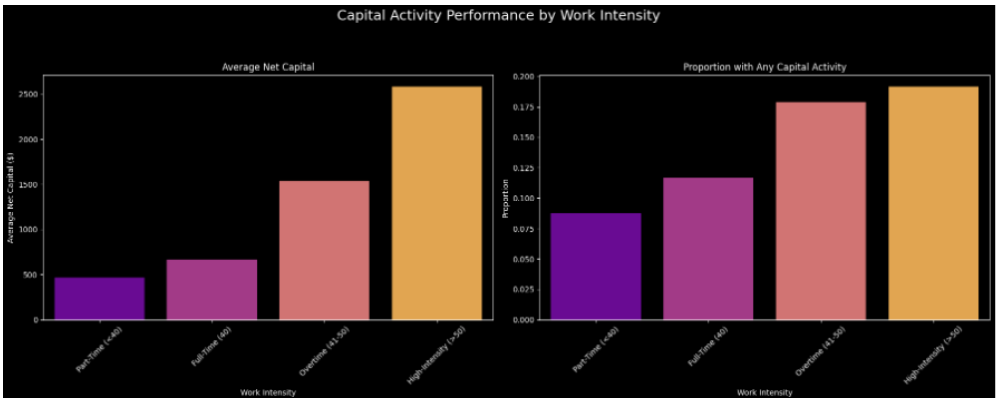

Figure 4: Capital Activity by Work Intensity

Final Dataset Summary

Grouped features revealed patterns that raw data could not. The dataset is now modeling-ready, with improved clarity and reduced dimensionality.

2 Part II: Real Estate Dataset Analysis

Data Preprocessing

The preprocessing phase involved rigorous schema standardisation, where column names were normalised and redundant fields like CovArea, Coverage Area, Carpet area were removed.Price values were converted to lakhs for interpretability, and engineered features such as price per square foot, amenity count, and luxury flags were introduced. Amenities were grouped into logical categories—like kids amenities, water facilities, and food and beverage options—to simplify analysis. Missing values were identified and addressed, and outliers were flagged using z-score thresholds on price, carpet area, and price per square foot. The cleaned dataset was exported for downstream analysis and exploratory data analysis.The cleaned dataset was saved as `cleaned_real_estate_data.csv`

Price Segmentation and Market Overview

Properties were segmented into affordable (0–50L), mid-range (50–100L), and luxury (100L+) bands. Visualisations revealed that Mumbai dominates the luxury segment, while Thane offers a more balanced mix.

Figure 5: Price Segmentation Across Cities

Property type distribution showed that Mumbai features more premium formats like villas and penthouses, whereas Thane is concentrated in apartments and residential houses.

Figure 6: Property Type Distribution

Amenity count analysis indicated that luxury listings tend to offer more features, though some mid-range properties rival them in richness. Violin plots revealed that luxury does not always equate to larger carpet areas—some affordable and mid-range listings offer comparable space.

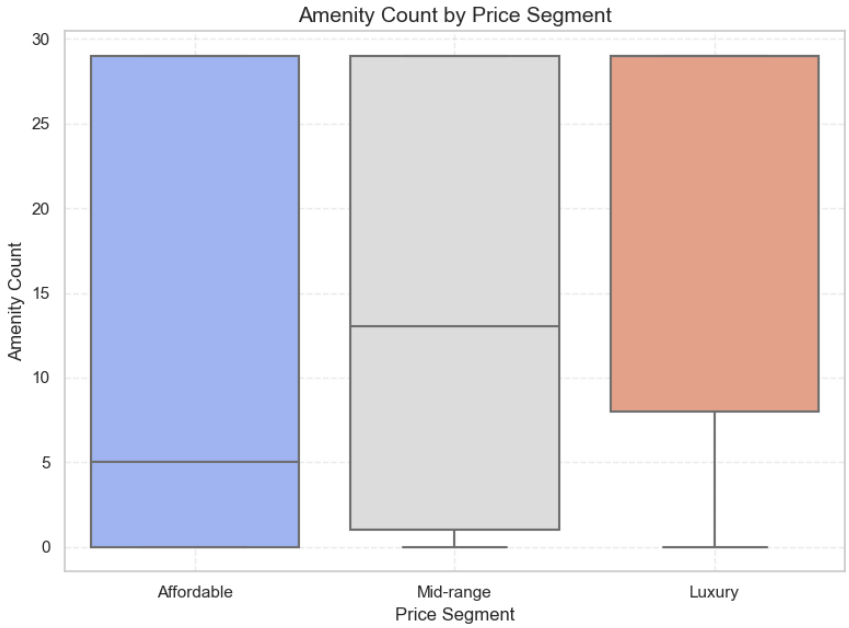

Figure 7: Amenity Count by Price Segment

Mumbai vs Thane Comparison

Thane properties generally offer larger carpet areas and better spatial efficiency, while Mumbai listings are more compact and expensive. Residential and commercial segments were both active, though Mumbai had a slightly higher share of commercial listings. Thane's commercial market is emerging, offering opportunities for expansion.

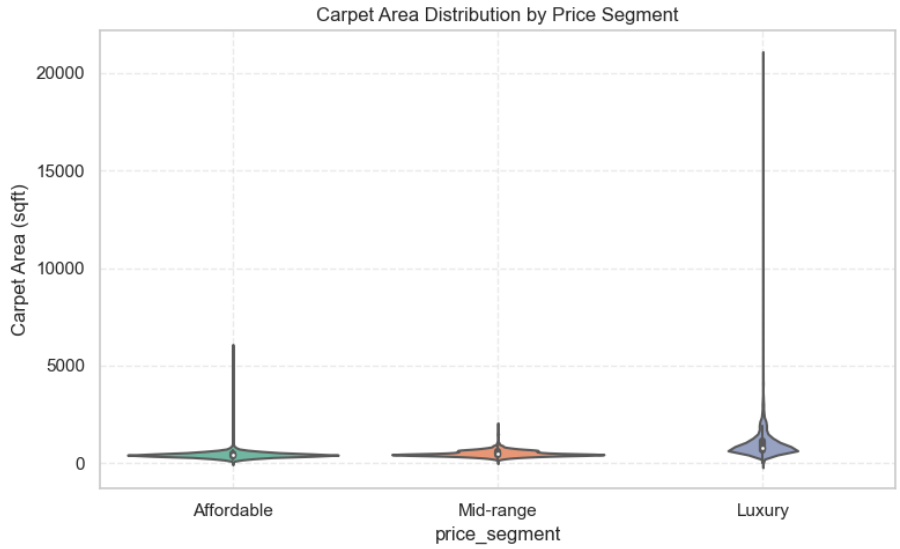

Figure 8: Carpet Area Distribution by City

Bar charts showed that Thane delivers nearly double the carpet area per lakh spent compared

to Mumbai, making it attractive for budget-conscious buyers.

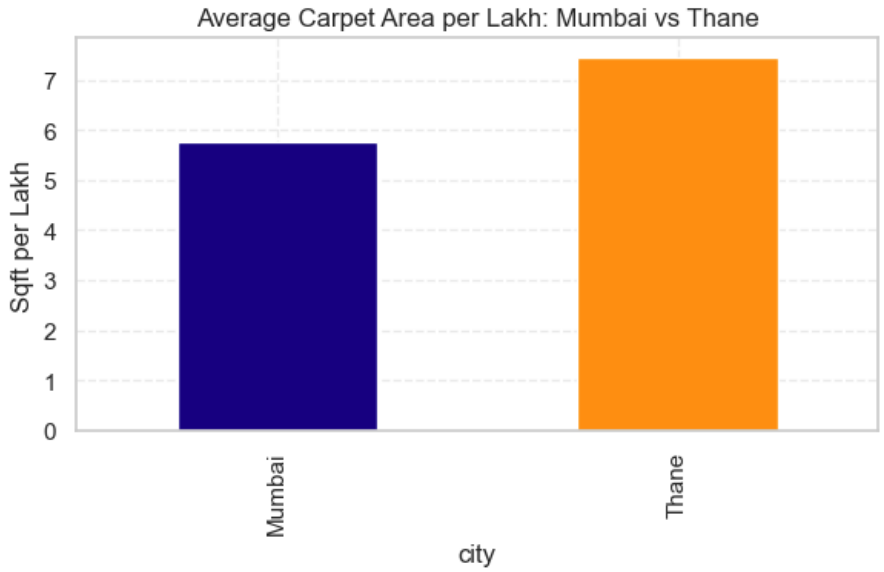

Figure 9: Carpet Area per Lakh Comparison

Location Premium and Value Efficiency

Prime-location properties command significantly higher price per square foot and offer more amenities, though they don't always provide more space. In Mumbai, the uplift is sharper, while in Thane, non-prime listings offer better value for money.

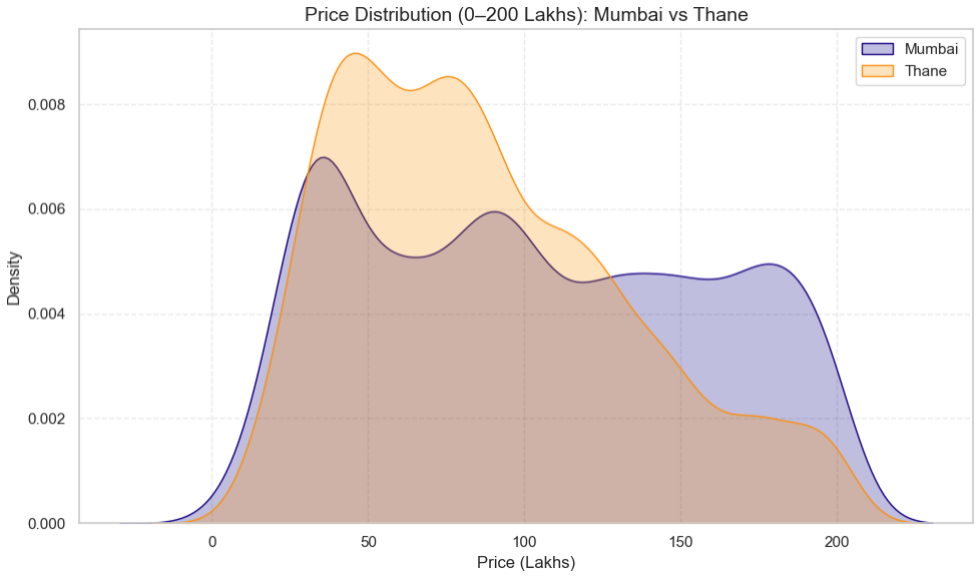

Figure 10: Price per Sqft – Prime vs Non-Prime

Cities like Gurgaon, Nagpur, and Kalyan were identified as top performers in carpet area per

lakh spent. Builder floor apartments and residential houses offered the best spatial efficiency, outperforming high-rise formats.

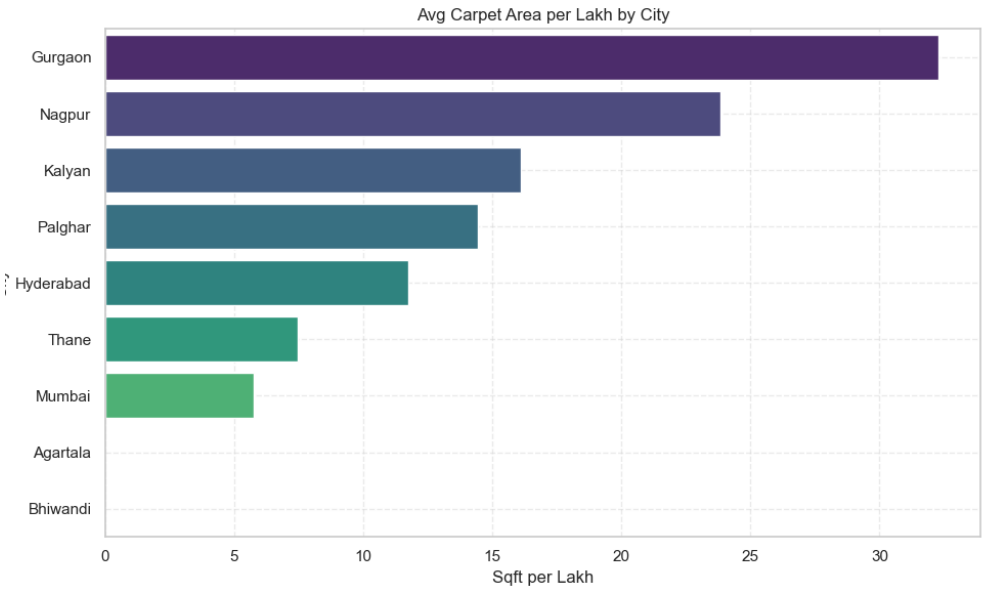

Figure 11: Carpet Area per Lakh by City

Amenity and Developer Impact

Swimming pools and gymnasiums significantly influence pricing, especially in Mumbai. Clubhouse type also added value, with premium clubhouses correlating with higher prices. Listings with all three amenities—pool, gym, and premium clubhouse—commanded the highest prices.

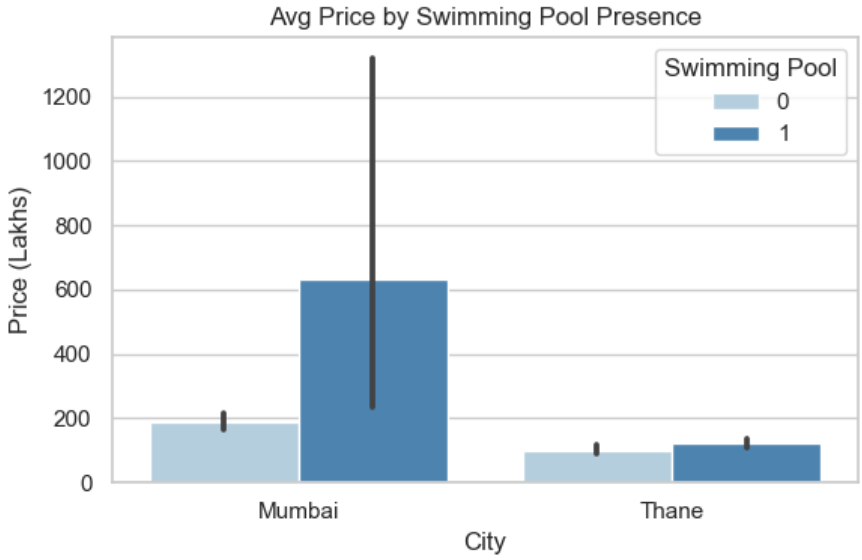

Figure 12: Swimming Pool Impact on Price

Ready-to-move properties are priced higher than under-construction listings, with the price gap more pronounced in Mumbai. Listings with near-term possession dates showed a premium, while longer timelines correlated with lower prices. Developer analysis showed that brands like Rustomjee and Lodha dominate Mumbai's luxury segment, commanding the highest average prices. In Thane, developers such as Kalpataru and Godrej offer more affordable pricing with solid amenity offerings.

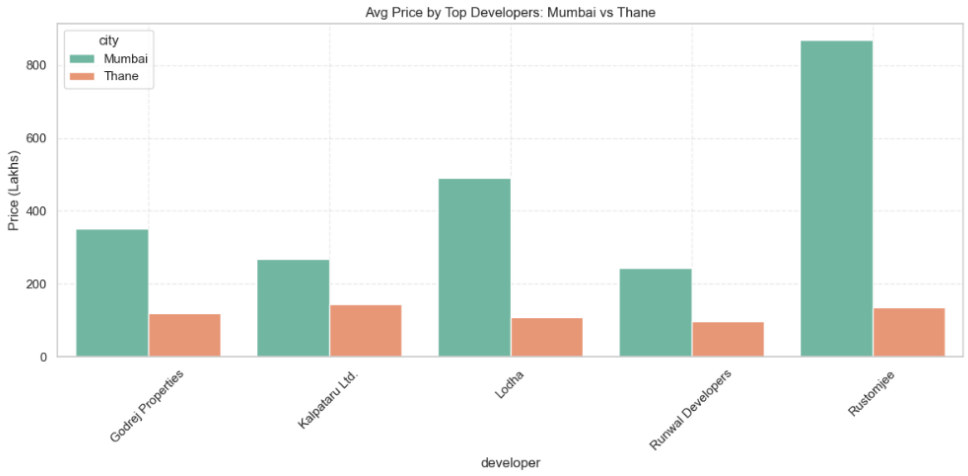

Figure 13: Average Price by Developer

Conclusion

This analysis provides a data-driven guide for investors. Mumbai offers premium branding and lifestyle features, ideal for luxury-focused buyers. Thane delivers better spatial efficiency and affordability, making it attractive for mid-range and value-conscious investors. The insights on location, amenities, possession timelines, and developer influence equip stakeholders to make informed decisions in a dynamic real estate market.