DATA ANALYTICS –1

ASSIGNMENT –1

Team Members:

1. Saloni Goyal (2023115001): DataCleaning.py, Part1, Part2(Q6)
2. Sachi Thonse Rao (2023101120): Part2(Q1-5, Q7)

Data Cleaning (Pre-processing)

1. Checking for missing values (Nan, NULL, "?", " ", etc.): *interactive_next_step*
    a. Delete that row (If substituting some other value in that place can skew the result): *remove_missing_values*
    b. Replace with Mean (If the data is centric in some region): *fill_missing_with_mean*
    c. Replace with Median (If the data has a lot of outliers): *fill_missing_with_median*
2. Detecting the outliers in the code (For ex, if Age has an entry which is 10,000 means this is incorrect data and needs to be removed from the dataset):
    a. Interquartile method (Checks if there are quantities less than 5*the value below which 25% of the data lies or are greater than 5* the value above which 75% of the data lies; chose 5 here to check only for extreme outliers): *detect_outliers_iqr*
    b. Z-Score method (Checks if there are values 3 standard deviations away from the mean value; chose threshold=3 here to check only for extreme outliers): *detect_outliers_zscore*
3. Ensuring the data type (If a column requires to be numeric eg. Age but is string, we need to convert it so that other operations can be performed on it): *ensure_data_type*
    a. Given, col_name and target as parameters, if the column's datatype is same as target skip, else convert to target.
4. Removing duplicates (Necessary in datasets which include columns like index, Aadhar number, etc.): *remove_duplicates*

PART – 1

Question 1: Education distribution and Grouping

Data cleaning for columns: fnlwgt, Education

Unique education levels along with their frequencies:

{'Bachelors': 1511601409.0, 'HS-grad': 2977279353.0, '11th': 353525318.0, 'Masters': 482192559.0, '9th': 150449180.0, 'Some-college': 2066871041.0, 'Assoc-acdm': 310113821.0, 'Assoc-voc': 369786888.0, '7th-8th': 179147014.0, 'Doctorate': 109349744.0, 'Prof-school': 155612621.0, '5th-6th': 116998125.0, '10th': 272983724.0, '1st-4th': 58050070.0, 'Preschool': 19832208.0, '12th': 129782587.0}

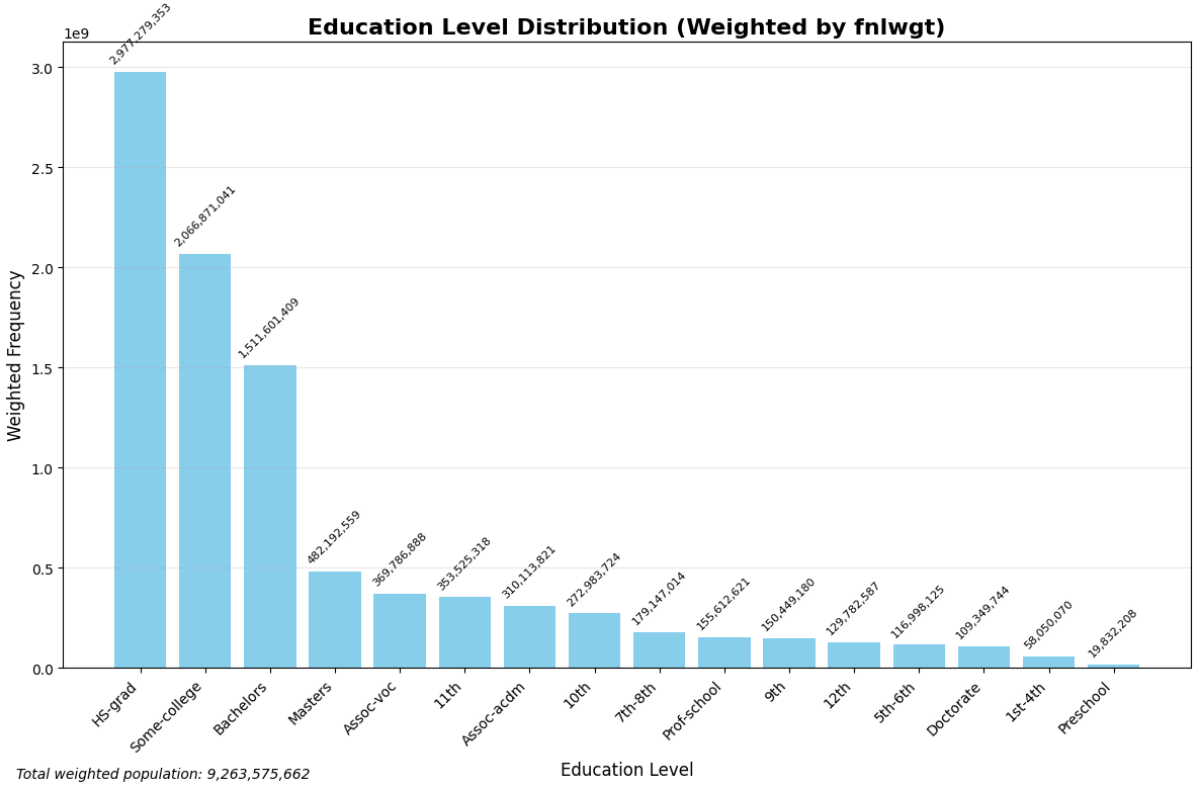


Fig. Bar Graph

Education Groups:

Elementary Education (until class 8th): Preschool, 1st-4th, 5th-6th, 7th-8th
Secondary Education (class 9th-10th): 9th, 10th, HS-grad
Senior Secondary Education (class 11th-12th): 11th, 12th
Associate degree: Assoc-acdm, Assoc-voc
Some college: Some-college
Graduation: Bachelors
Post-graduation: Masters, Doctorate, Prof-school

Reasoning for grouping: Till Class 8th, i grouped the variables in Elementary education as some people leave education after this stage. Then, 9th,10th, ad High-School Graduates in Secondary Education, and 11th and 12th in Senior-secondary education, which are basically the terms given to those classes in general. Then i categorised associate roles different from other grads as the time taken to complete it varies, hence its relationship with other columns will also vary. Similar with Some college, because some college can be any institution providing any course. Then i classified Graduation and Post-graduation as normal Bachelors, and (Masters, doctorate, and prof-school) respectively.

Elementary Education (until class 8th): 374027417.0
Secondary Education (class 9th-10th): 3400712257.0
Senior Secondary Education (class 11th-12th): 483307905.0
Associate degree: 679900709.0
Some college: 2066871041.0
Graduation: 1511601409.0
Post-graduation: 747154924.0

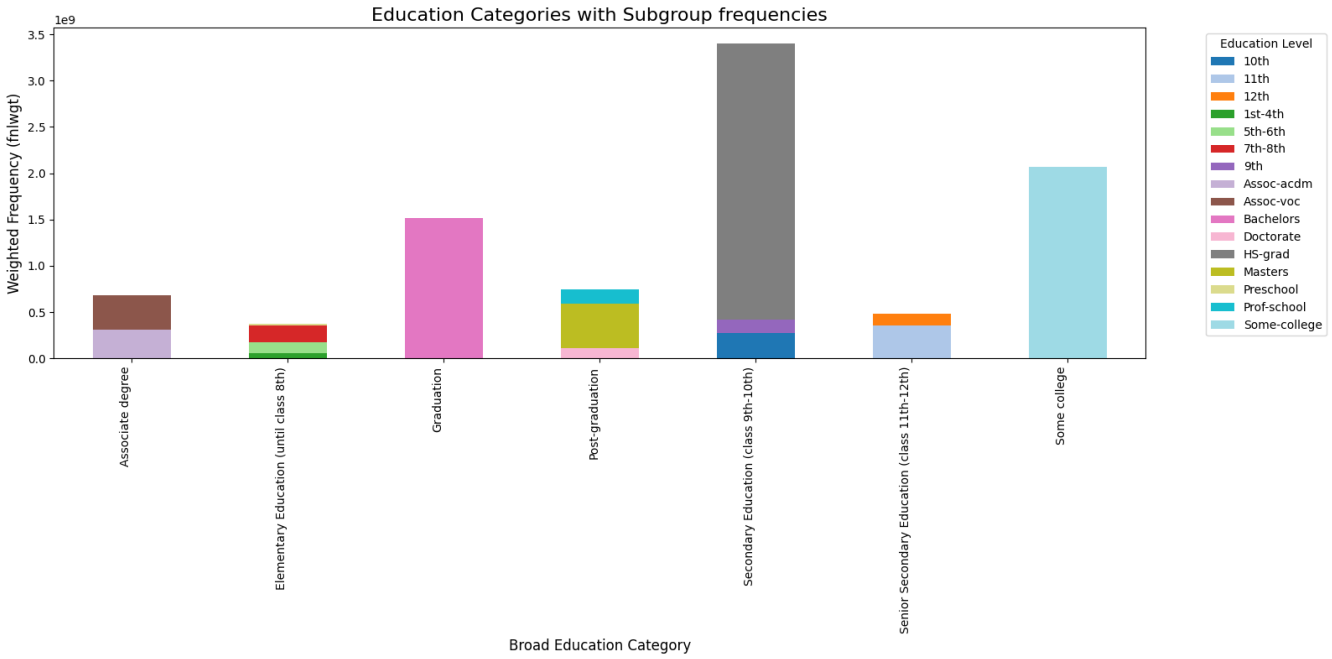


Fig. Stacked Bar Graph

Results Explanation: The tallest bar is for Secondary Education, mainly driven by HS-grad which has a huge frequency. This suggests a large proportion of the population in the dataset completed schooling until 9th–12th grade, but didn't necessarily move on to college. The "Some college" group is the second highest, meaning many individuals started college but didn't complete a degree. This reflects common educational patterns where higher enrollment exists, but not all transition to graduation. The category (Masters, Doctorate, Prof-school) has lower counts compared to Graduation and Secondary Education, but still significant. It shows fewer people pursue advanced degrees.

Question 2: Age-work intensity relationship and grouping

Data cleaning for columns Age, Hours_per_week, fnlwgt

Distribution of Age

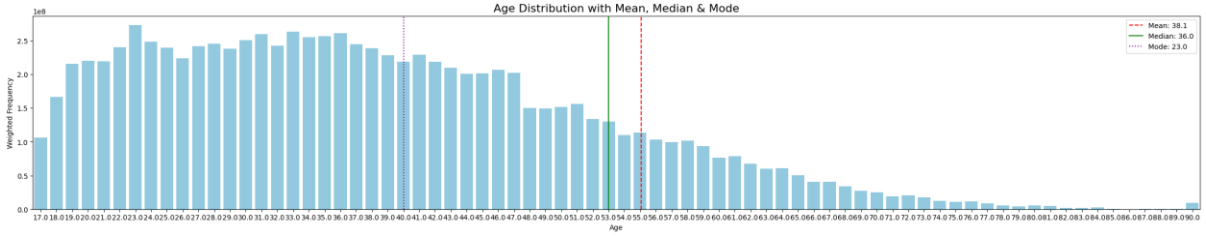


Fig. Histogram

We can see from the graph that dataset has fewer people <22, the major part till 47 and then fewer >48

Mean: 38.1, Median: 26.0, Mode: 23

Distribution of Hours per week

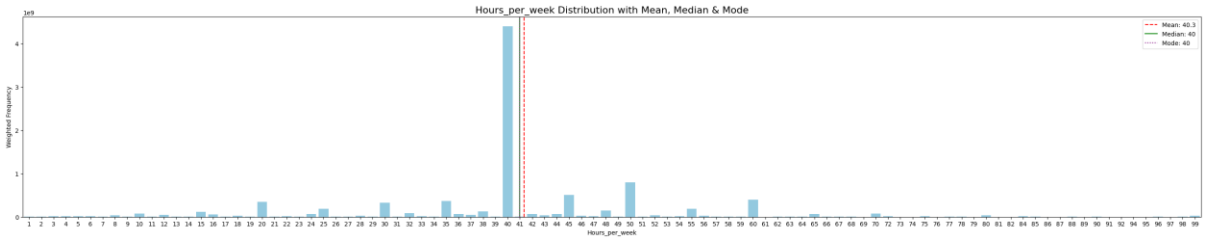


Fig. Histogram

We can see from the graph that major population is working 40 hrs a week.

Mean: 40.3, Median: 40, Mode: 40

Broader Group for Age:

0< Child<18

18<=Youth<30

30<=Adult<50

50<Senior

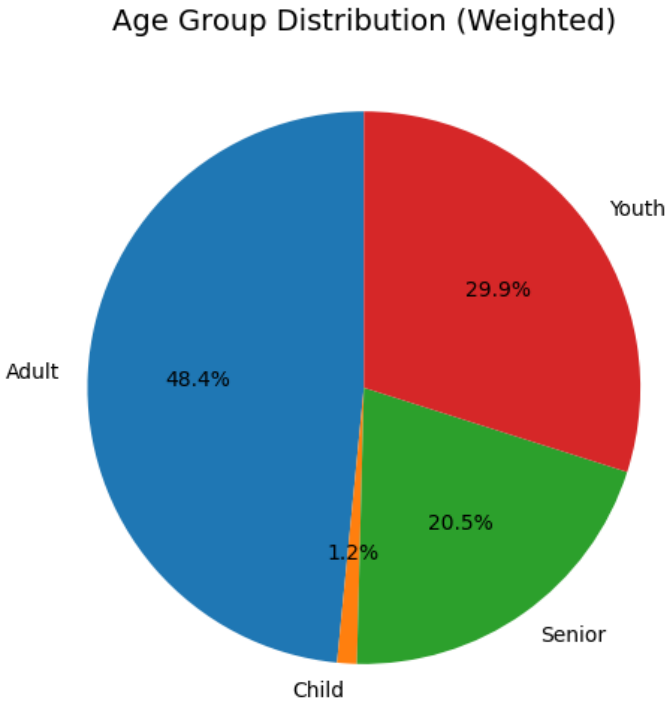


Fig. Pie Chart

Before grouping, I was just able to say that people between age of 22 and 47 are occupying the greatest number of rows in the dataset. But now I can confirm that from the chart as almost 50% of the population is lying in that range. Individual graph was better for visualising the mean, median and mode. But the grouped graph is better when the dataset is huge, and we want to see the statistics clearer.

Broader Group for Hours_per_week:

Part-Time: Hours<20

Full-Time: 20<=Hours<=40

Over-time: <40

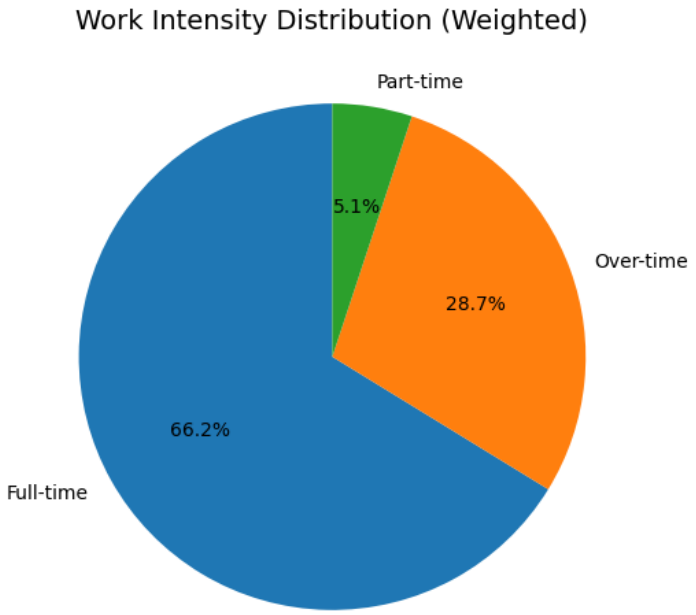


Fig. Pie Chart

Before grouping, I was just able to say that people who are working around 40 hours a day are occupying the greatest number of rows in the dataset. But now I can confirm that from the chart as almost 66% of the population is lying in that range. Individual graph was better for visualising the mean, median and mode. But the grouped graph is better when the dataset is huge, and we want to see the statistics clearer.

Relationship between Age and Work intensity before Grouping:

Fig. Heatmaps

There are a lot of values (fnlwgt on top), hence, the heat map is not giving clear results. The result is a patchy heatmap with small specks, dominated by large fnlwgt outliers. Still, we can see that it's concentrated when work intensity is between 40-50 and age is 20-50.

Relationship between Age and Work intensity after Grouping:

	Age_Group	fnlwgt
0	Adult	4.487723e+09
1	Child	1.065989e+08
2	Senior	1.895944e+09
3	Youth	2.773310e+09

	Hours_Group	fnlwgt
0	Full-time	6.135568e+09
1	Over-time	2.657025e+09
2	Part-time	4.709822e+08

	Age_Group	Hours_Group	fnlwgt
0	Adult	Full-time	2.791471e+09
1	Adult	Over-time	1.613742e+09
2	Adult	Part-time	8.250927e+07
3	Child	Full-time	5.901201e+07
4	Child	Over-time	1.233184e+06
5	Child	Part-time	4.635373e+07
6	Senior	Full-time	1.228654e+09
7	Senior	Over-time	5.176703e+08
8	Senior	Part-time	1.496200e+08
9	Youth	Full-time	2.056432e+09
10	Youth	Over-time	5.243794e+08
11	Youth	Part-time	1.924992e+08

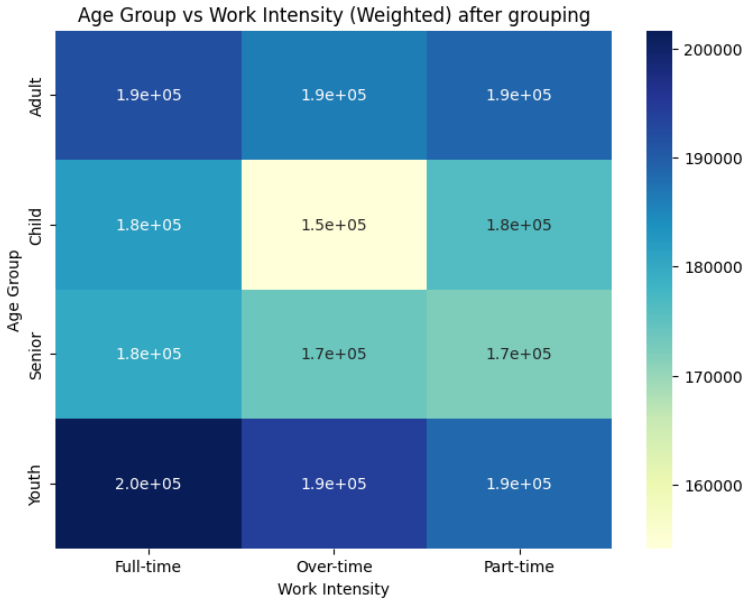


Fig. Heatmaps

Now, as data is grouped, heat maps are working properly. We can see that Full-time youth's and adult's are working the most while Over-time Children are almost negligible compared to other results.

Analysis:

Before Grouping, it Preserved the fine-grained details (exact hours, exact age) but was too noisy and hard to interpret visually (Outliers dominated).

After Grouping, it Improved interpretability by condensing data into broad categories. Heatmap clearly showed the dominant trends. Even Though some detail is lost (e.g., difference between 30 hrs vs 40 hrs/week both fall under "Full-time"), grouping is better as it gives high-level insights.

Question 3: Capital Gains/Losses and Group Performance

Data cleaning of the columns fnlwgt, Capital Gain, Capital Loss

Capital Gain distribution (overall): Log-scaled Histogram on Y-axis because otherwise 0 dominates.

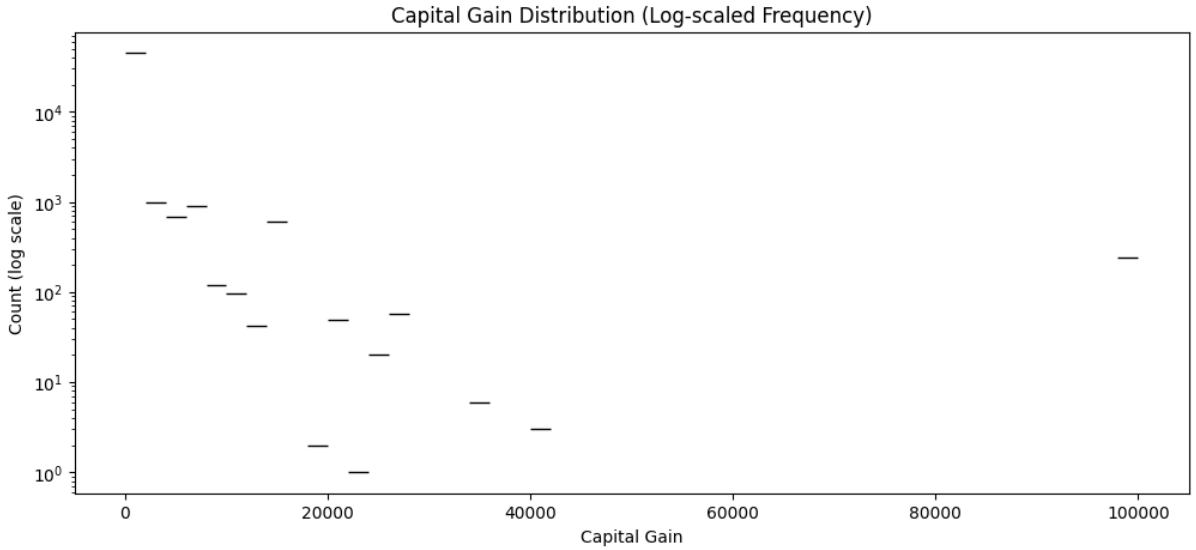


Fig. Log-scaled histogram

We can see that 0 has the highest frequency which means that most of the people in the dataset are not in capital gain.

Capital Gain distribution (non-zero values): Boxplot

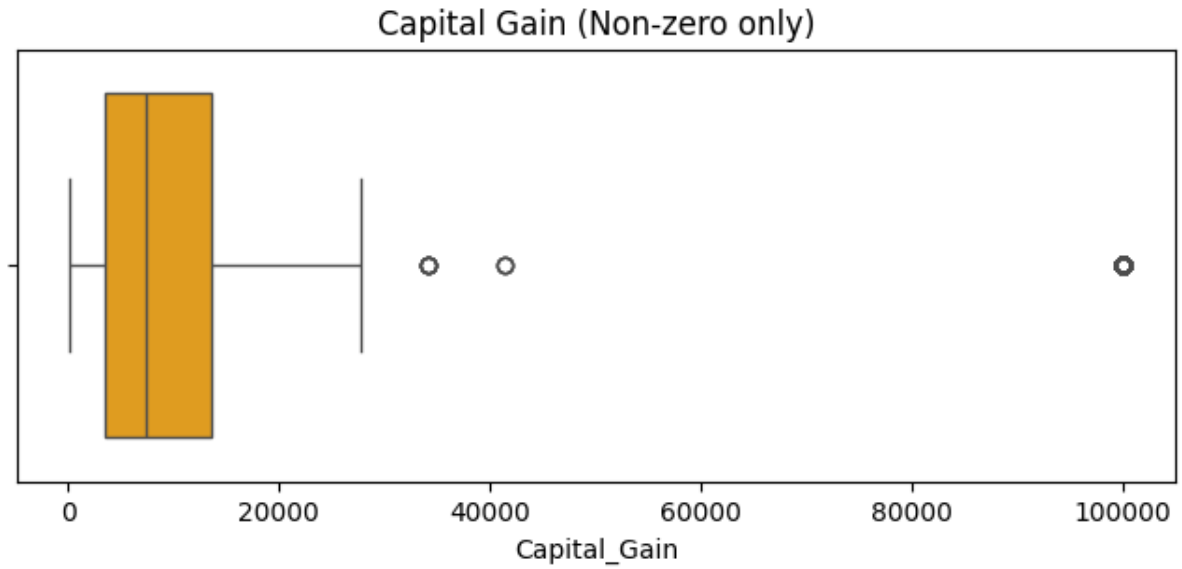


Fig. BoxPlot

The Boxplot works on interquartile range. Box (orange rectangle) represents the middle 50% of the data (Interquartile Range, IQR). Line inside box shows the median (middle value). Whiskers extend to the minimum and maximum values within 1.5 × IQR. And the dots represent outliers (values outside the whisker range).

After 0 values are removed, we can see that the graph has shifted.

Capital Loss distribution (overall): Log-scaled Histogram because otherwise 0 dominates.

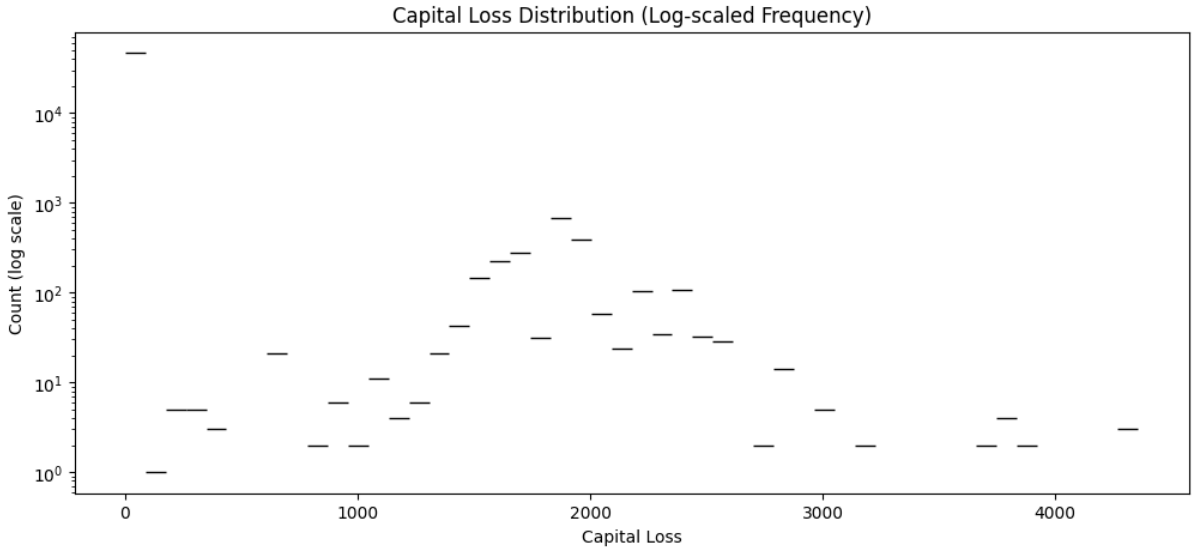


Fig. Log-scaled histogram

We can see that 0 has the highest frequency which means that most of the people in the dataset are not in capital loss.

Capital Loss distribution (non-zero values): Boxplot

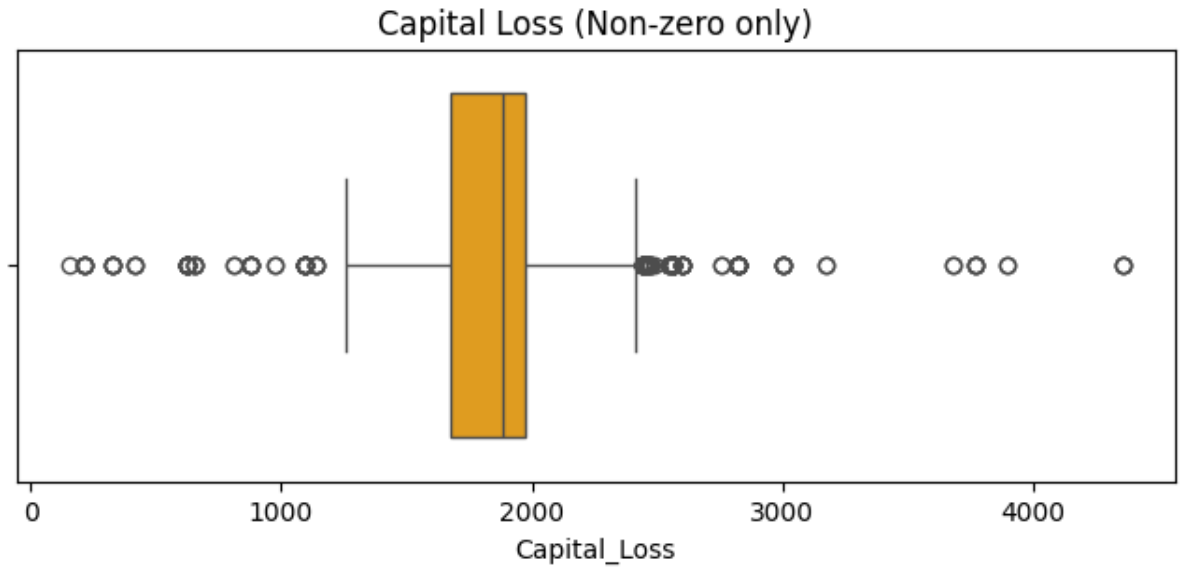


Fig. BoxPlot

After 0 values are removed, we can see that the graph has shifted. The outliers (circles) show some people who are either in very less loss or very great loss compared to the mean range.

Capital Activity:

Capital activity: Capital Loss>0 or Capital Gain>0

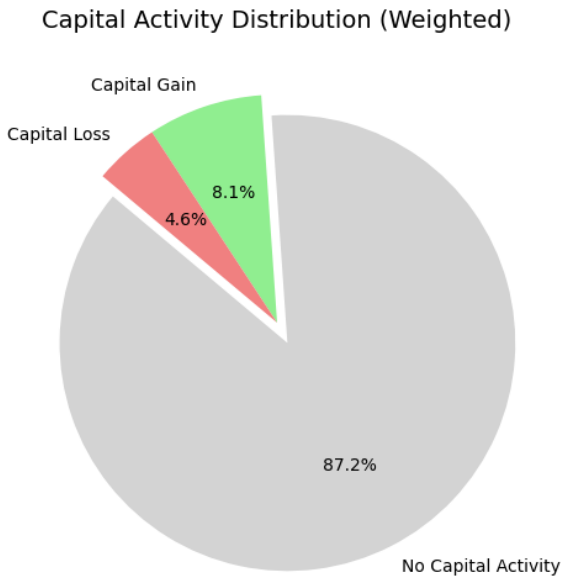


Fig. Pie Chart

We deduce from this that most of the population in the dataset are not doing any transaction, only 8.1% are in profit, 4.6 in loss, and the rest nothing.

Comparing Age groups and Work Intensity with Avg net capital

Average Net Capital = Capital Gain-Capital Loss

	Age_Group	Hours_Group	Avg_Net_Capital	Prop_Capital_Activity
0	Adult	Full-time	702.551359	0.119738
1	Adult	Over-time	2113.244559	0.199754
2	Adult	Part-time	105.143016	0.083641
3	Child	Full-time	190.520395	0.043596
4	Child	Over-time	0.000000	0.000000
5	Child	Part-time	-38.515951	0.053051
6	Senior	Full-time	1162.916228	0.154380
7	Senior	Over-time	3110.515632	0.236522
8	Senior	Part-time	564.510863	0.160983
9	Youth	Full-time	150.075360	0.058188
10	Youth	Over-time	396.041425	0.097185
11	Youth	Part-time	118.127555	0.042948

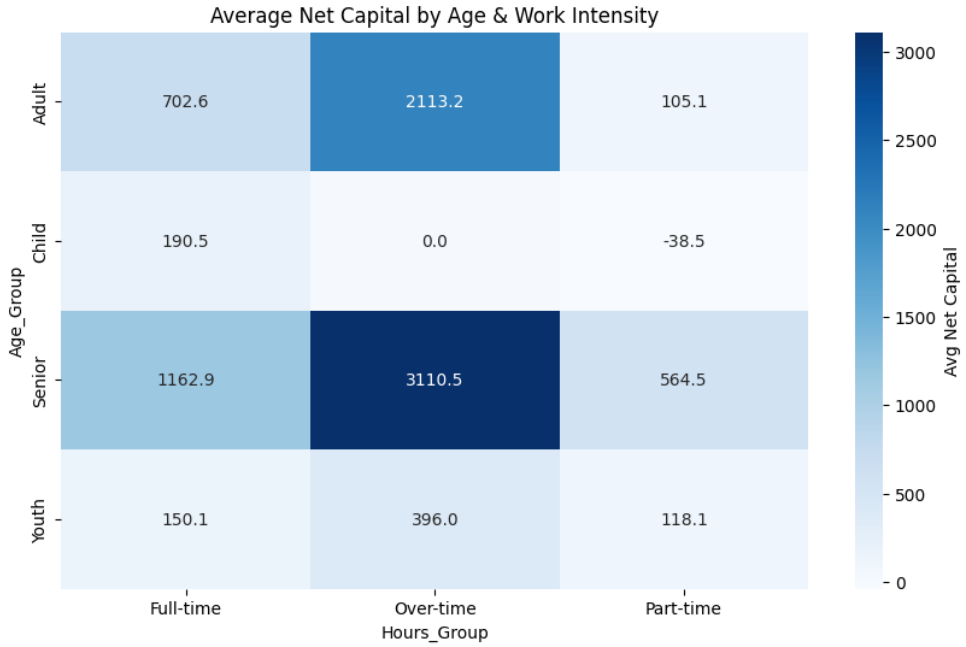


Fig. Heatmap

Seniors working overtime have the highest average net capital (~3110). Even seniors working full-time (~1163) or part-time (~565) still show higher averages than younger groups. Suggests wealth accumulation over life + possibly investment income/capital gains.

Adults working overtime also show high average capital (~2113). Full-time (~703) and part-time (~105) are lower, but still above youth and children. Indicates prime working age with higher productivity & income.

Youth, generally low across all work intensities. Overtime (~396) is better than full-time (~150) or part-time (~118), but still far below adults/seniors. Early career stage, less savings/wealth.

Children (<18) have very small or negative values (e.g., part-time ~ -38.5, overtime = 0). Expected, since most children don't work or contribute capital meaningfully.

<u>Comparing Age groups and Work Intensity with Proportional Capital Activity:</u>

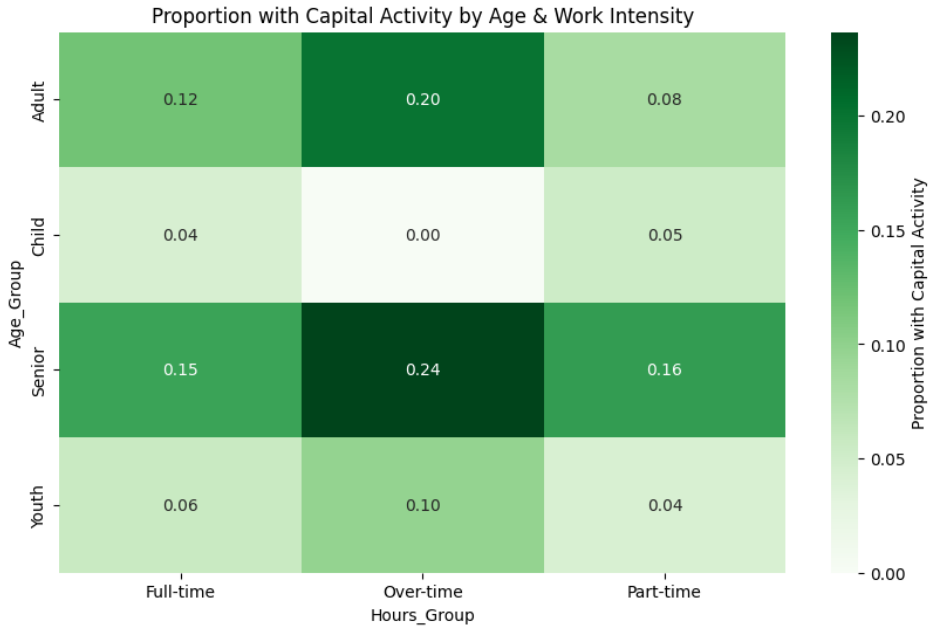


Fig. Heatmap

Capital Activity: The values represent the share of people (proportion) within each group who have non-zero capital activity.

Seniors have the highest engagement in capital activity. Overtime-working seniors: 24% have capital activity (darkest cell). Even full-time (15%) and part-time (16%) seniors are above other groups. Suggests seniors rely more on investments (retirement savings, rental income, etc.).

Adults (30–49) are second-highest. Overtime: 20% with capital activity. Full-time: 12%, part-time: 8%. Prime working age adults often supplement wages with investments.

Youth (18–29) have much lower proportions: 6–10%. Early in career → less investment activity.

Children (<18): Very minimal: 0–5%. As expected, almost no capital activity.

Analysis:

Work intensity shows a stronger association with net capital than age, since higher weekly hours generally align with higher earnings. Age shows some effect but is less direct. Grouping improves interpretability by reducing noise and highlighting clear trends (e.g., part-time vs full-time vs overtime), though it may also mask finer details.

Question 4: Final Dataset Refinement and Structure

Original Shape: 48443 rows × 19 columns

Final Shape: 48442 rows × 19 columns (after cleaning and feature engineering)

Continuous -> Categorical Groups

1. Age: Child, Youth, Adult, Senior
2. Hours_per_week: Part-time, Full-time, Over-time

Newly Created: Net capital (Capital gain-capital loss), Capital Activity (Binary), etc.

Transformed Features:

1. Education (Enhanced Categorization)
   Before: 16 individual education levels
   After: 7 meaningful categories:

Data type corrections:

Age: Object → Numeric (float64)

Capital_Gain: Object → Numeric (float64)

Capital_Loss: Object → Numeric (float64)

Hours_per_week: Object → Numeric (float64)

fnlwgt: Object → Numeric (float64)

Removed features: None

Analysis:

The final refined dataset includes newly created grouped variables, so the noisy categories are removed. Continuous features like fnlwgt are cleaned and outliers treated. Before grouping, variables such as education and age had many fragmented categories, but grouping reduced them into fewer, more interpretable classes with balanced counts. Missing values were reduced through imputation and removal of irrelevant entries. Overall, these refinements improved clarity, reduced sparsity, and made patterns in net capital more interpretable, preparing the dataset for robust modelling.

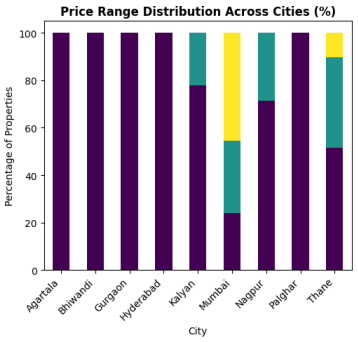
Data Cleaning Techniques (for Part-2)

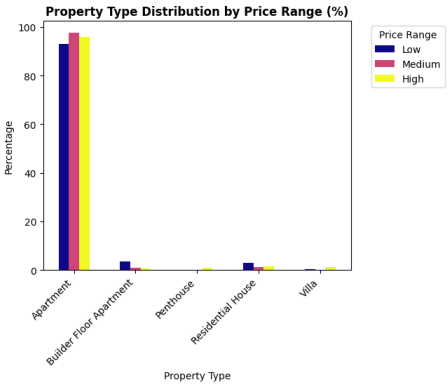
The dataset was first cleaned by converting key columns like 'Price', 'Carpet Area', and amenity fields to numeric types, coercing errors into NaN. Rows with missing or invalid entries in essential fields such as 'Price', 'City', or 'Area Name' were removed. Amenities were standardized by converting them to binary values (1 for available, 0 for not available) and filling missing values with 0. Outliers in columns like 'Carpet Area' and 'Price_per_sqft' were removed using threshold-based filtering or upper quantiles. Developer names were cleaned by stripping whitespace, title-casing, and replacing placeholders like "N/A" with nulls. Properties were categorized into price ranges—Low, Medium, and High—using percentiles, and a value score metric was introduced to compare affordability. High-budget properties were identified based on city-wise price percentiles, and further classified as "Prime" or "Non-Prime" using a curated list of premium localities. The 'Commercial' label was inferred either from a dedicated column or by checking keywords in the property type. Listings were filtered to focus on specific cities like Mumbai and Thane in some analyses. Finally, developers with insufficient data were excluded to ensure reliable insights.

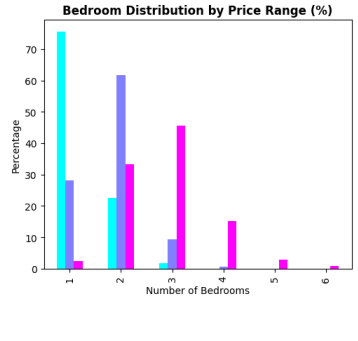
PART – 2

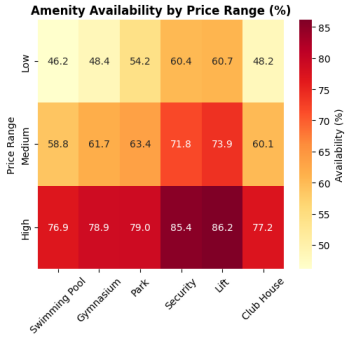
Question 1: Price Segmentation and Market Overview

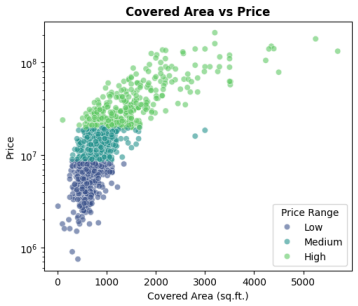
The real estate market is evenly distributed across three price segments: low, medium, and high, each comprising roughly one-third of the total listings. Low-priced properties (₹1L–₹87L) are dominated by 1–2 BHK apartments in Thane and Mumbai, with top amenities like lift, park, and security, and Lodha emerging as the leading developer. Medium-priced properties (₹87.5L–₹1.98Cr) are largely concentrated in Mumbai, offering slightly larger apartments with enhanced amenities such as clubhouses and swimming pools, and developers like Godrej and Kalpataru joining the list. High-priced properties (₹1.99Cr–₹408Cr) are overwhelmingly located in Mumbai, offering spacious 3 BHK+ apartments with luxury features and top-tier amenities like gyms, pools, and security, developed by premium brands like Bombay Realty and Oberoi. Across all segments, apartments dominate the supply, with villas appearing only in the high range. Amenity availability consistently improves with price, particularly for clubhouses, gyms, and swimming pools. Mumbai consistently commands the highest average prices (₹3.39 Cr), followed by Thane (₹1.08 Cr), while cities like Hyderabad and Bhiwandi are the most affordable. Developer concentration is strong, with Lodha leading across all segments. The market demonstrates a clear relationship between price and property size, bedroom count, and luxury offerings.

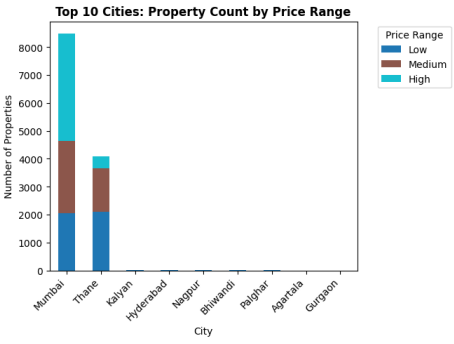


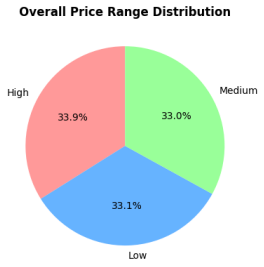






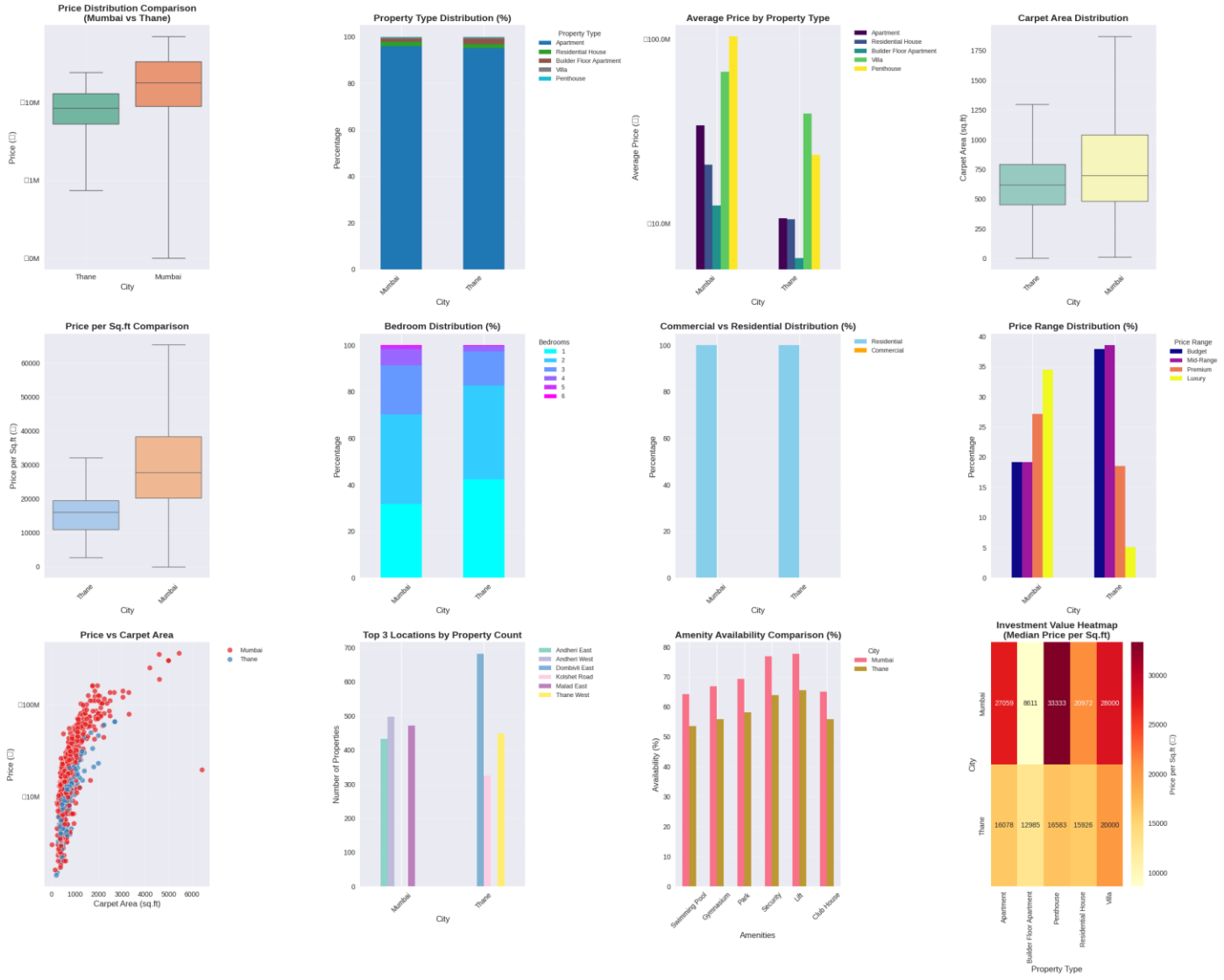






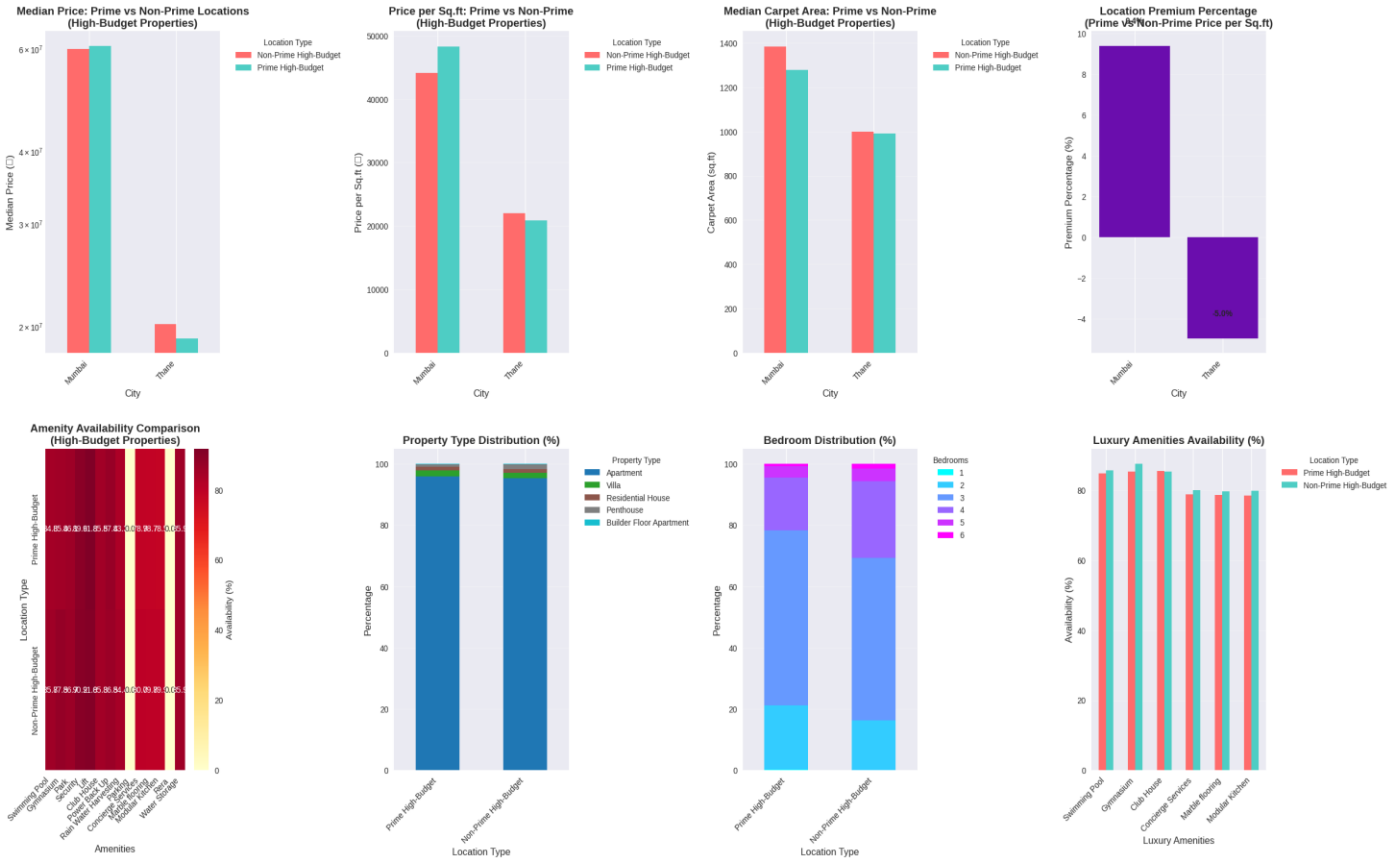
Question 2: City-Level Comparative Analysis

Mumbai offers a significantly higher investment threshold than Thane, with an average property price of ₹3.39 crore compared to Thane's ₹1.08 crore. The median price in Mumbai is ₹1.8 crore, more than double Thane's ₹85 lakh, highlighting the premium nature of the Mumbai market. Mumbai properties also offer larger carpet areas (876 sq.ft.) but at a much higher average price per sq.ft. of ₹42,272, compared to Thane's 685 sq.ft. and ₹16,926 per sq.ft. Apartments dominate both markets, comprising over 95% of listings in each. While Mumbai has more high-end property types like villas and penthouses, they still form a small portion of the overall inventory. Residential use accounts for 100% of the properties in both cities, with no commercial listings reported. The data reflects Mumbai's status as a premium real estate market with dense high-value residential offerings. Thane, in contrast, presents more affordable investment opportunities with smaller units and lower per sq.ft. costs. Both cities show strong residential focus and similar property type distributions, but cater to distinctly different budget segments.



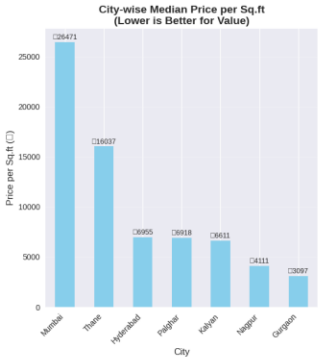
Question 3: Location-Based premium analysis

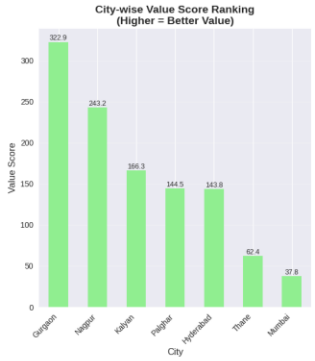
The analysis compares high-budget properties based on whether they are in prime locations, revealing notable differences between Mumbai and Thane. In total, 2,494 high-end properties were analyzed, with 888 classified as prime and 1,606 as non-prime. Mumbai shows a clear location-based premium, where prime properties command a 9.4% higher price per sq.ft. than non-prime ones (₹48,270 vs ₹44,132). This indicates strong investor preference and perceived value in Mumbai's prime areas. In contrast, Thane exhibits a negative premium of -5.0%, where non-prime areas surprisingly have higher price per sq.ft. (₹22,000) than prime locations (₹20,909), suggesting either market mispricing or shifting buyer focus. Mumbai also has a much larger supply of non-prime high-budget properties, highlighting wider availability across zones. Overall, prime location adds substantial value in Mumbai but not in Thane, underscoring the importance of micro-market behavior in investment strategy.

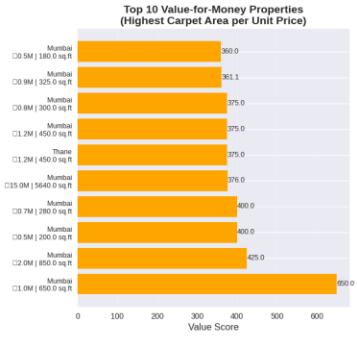


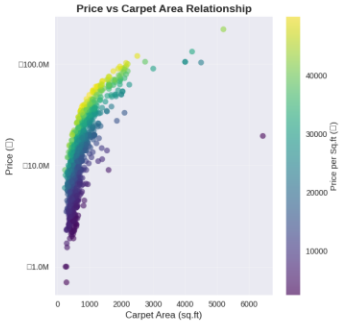
Question 4: Value-for-Money opportunities

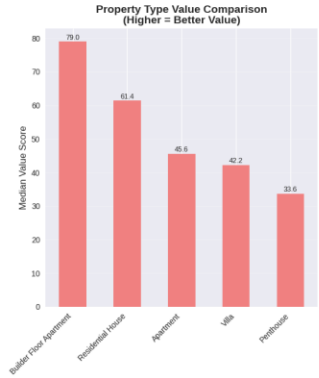
The value-for-money analysis highlights Gurgaon as the best-performing city, offering the lowest price per sq.ft. (₹3,097) and the highest value score (322.9), making it ideal for budget-conscious investors. Cities like Nagpur and Kalyan also deliver strong value due to affordable prices and relatively spacious properties. In contrast, Mumbai ranks lowest in value score (37.8), with high prices (₹26,471 per sq.ft.) and modest average carpet areas, indicating limited value for investment. Builder Floor Apartments emerge as the best value property type, followed by Residential Houses, offering the most space per rupee spent. Among Mumbai and Thane localities, Virar East and Nandivli stand out as the top value zones respectively. The report recommends focusing on ready-to-move-in properties in emerging suburbs with infrastructure growth for optimal returns. Budget-friendly investors should aim for properties priced under ₹2.4 Cr, with price/sq.ft. below ₹21,818 and carpet areas between 650–870 sq.ft.

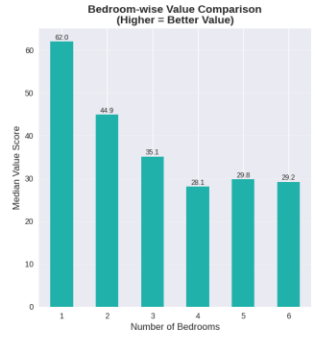


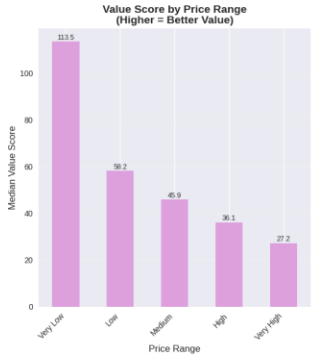


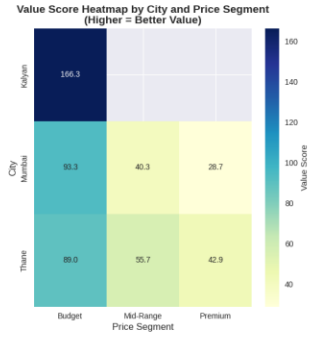


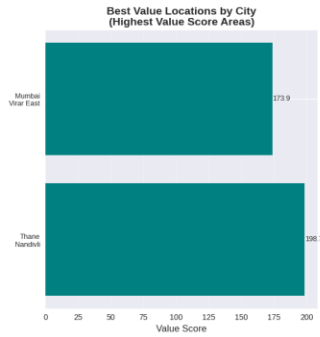






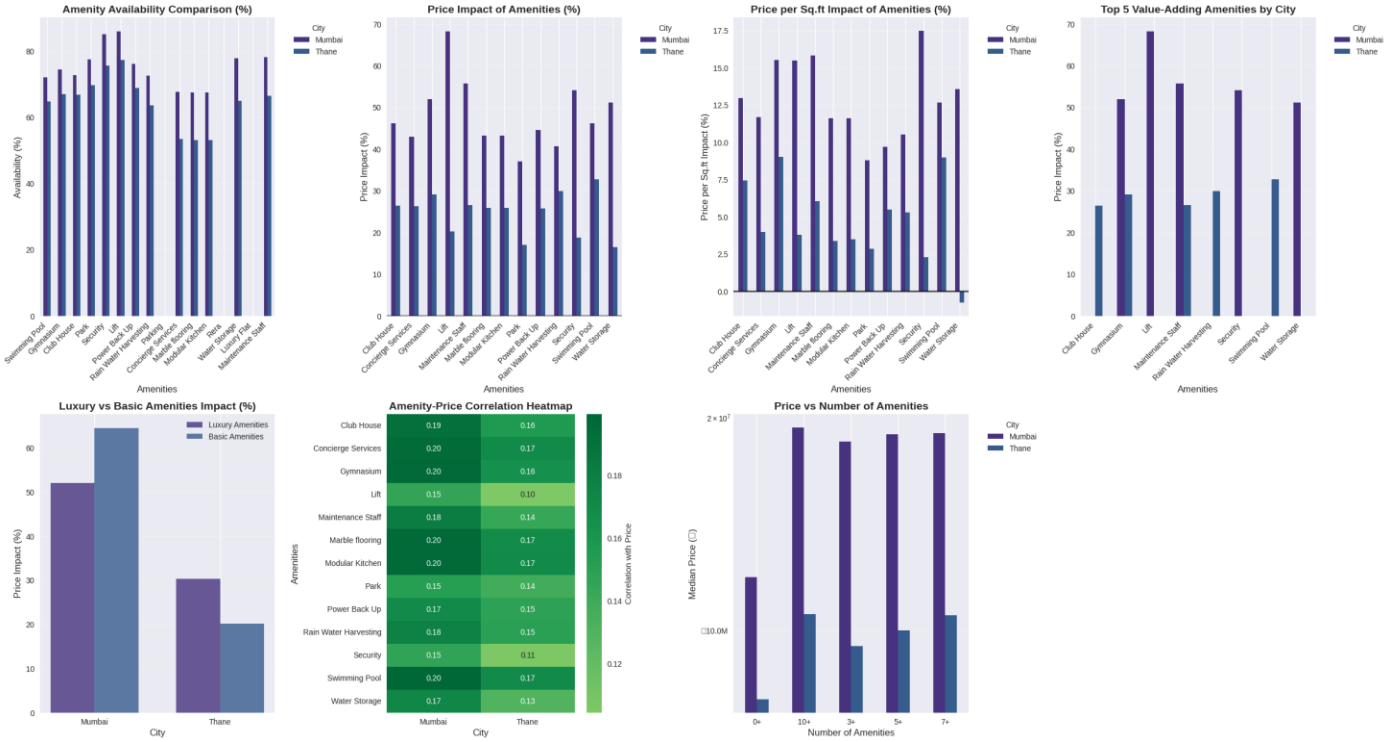






Question 5: Feature and amenity impact on price

The analysis shows that amenities significantly impact property prices, with Mumbai properties showing much higher sensitivity to amenity presence than Thane. In Mumbai, essential amenities like lifts, security, and maintenance staff contribute to over 50% price increases, with lifts topping the list at a 68.2% impact. In contrast, Thane's top value-adding amenity—swimming pools—shows a more modest 32.7% price impact, indicating lower premium placed on amenities in that market. The gap is especially stark for features like lifts (47.9% higher impact in Mumbai) and security (35.4% higher), reflecting Mumbai buyers' stronger preference or expectations for full-service buildings. Developers in Mumbai should prioritize high-ROI amenities like lifts, gyms, and maintenance services when targeting premium segments. In Thane, a balanced mix of functional and lifestyle amenities, such as rainwater harvesting and clubhouses, may be more effective. Budget allocation should reflect local buyer expectations, maximizing value without overspending on underutilized features.



Question 6: Timeline and Readiness effect on pricing

Data cleaning for the columns Possession Status, Availability Starts From, Price, City

Price: the normal range is 30 lakhs to 1.5 crore, but there are 259 rows that exist in outliers and are in range 3 crores. So, we will use the median value for substituting. As the dataset is small and anyway we have to delete the null valued-rows in non-numerical columns, we will substitute the null values in price column with the median value of that column.

Filled 84 rows in 'Price' with mean: 13500000.0000

Then there were no missing values in City column.

The 78 rows having null values in the Possession status column, we deleted them because it's in string format and hence can't be substituted with mean or median values.

Now, we find the subset of columns Price and City where city is either Mumbai or Thane.

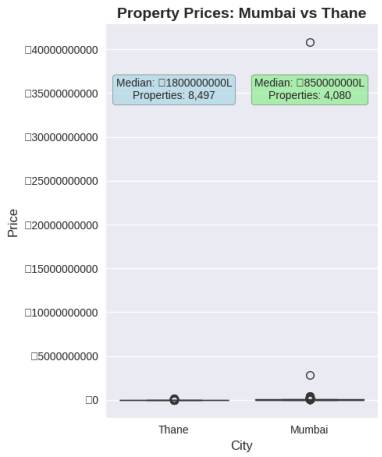
Out of 12607 rows, 12577 rows came into this category.

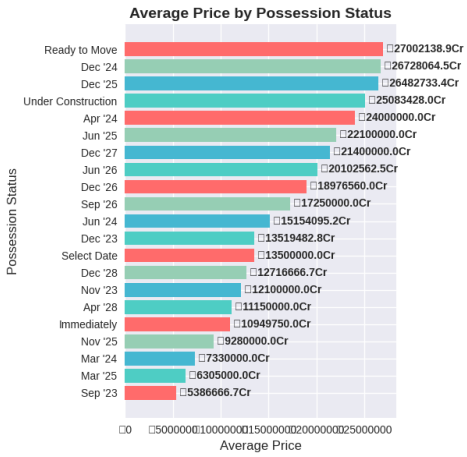
With Posession Status:

City Price Comparison (Left Graph): Mumbai properties are significantly more expensive than Thane properties. The median price in Mumbai is ₹180 crores compared to Thane's ₹85 crores - making Mumbai more than double the price of Thane. Mumbai also shows much greater price variation with several high-value outliers reaching ₹400+ crores, while Thane prices are more consistently clustered in the lower range. Mumbai has 8,497 properties in the dataset versus Thane's 4,080 properties.

Possession Status Impact (Middle Graph): "Ready to Move" properties command the highest average prices at around ₹270 crores, followed closely by "Dec '24" deliveries at ₹267 crores. Properties with longer completion timelines like "Dec '25" and "Under Construction" are priced lower in the ₹250-260 crore range. The most affordable category is "Sep '23" at around ₹54 crores, likely representing older or different types of properties. This shows a clear premium for immediate availability.

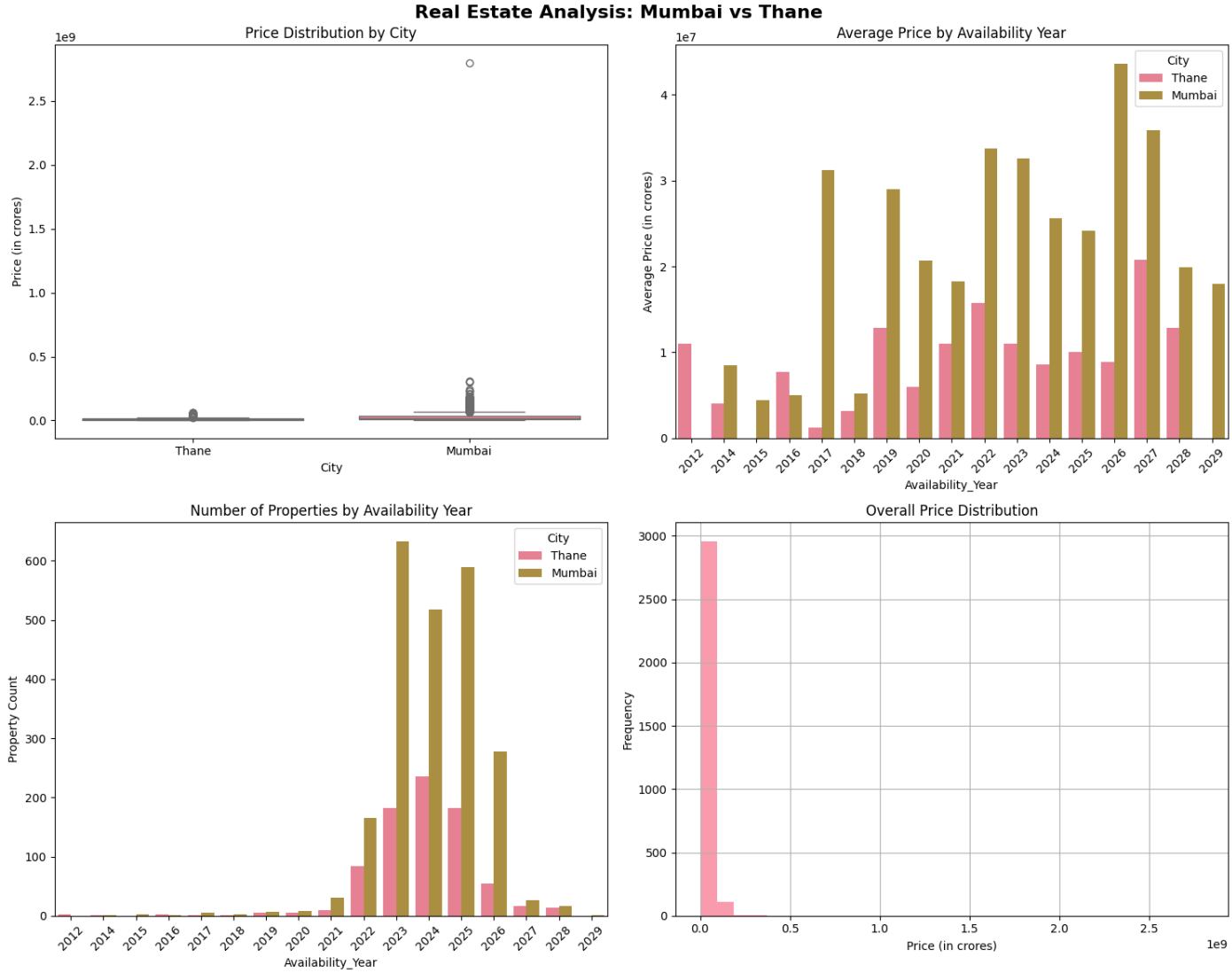
Combined City & Status Analysis (Right Graph): Mumbai consistently shows higher prices across all possession status categories compared to Thane. However, the pattern is interesting - while Mumbai dominates in most categories, Thane shows competitive pricing in certain possession statuses. The graph reveals that possession status impact varies by city, with Mumbai showing more dramatic price differences between different delivery timelines. Both cities show the "Ready to Move" premium, but Mumbai's premium is more pronounced.





With Availability Starts from:

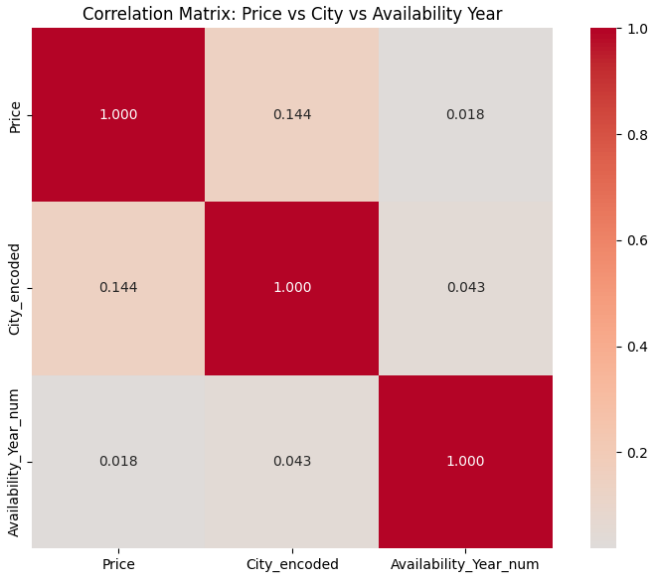
Approx 9k rows with missing availability starts from values deleted.



Price Distribution by City (Top Left- Box Plot): Mumbai has significantly higher property prices with greater variability. Thane shows more concentrated, lower pricing with fewer outliers.

Average Price by Availability Year (Top Right - Bar Chart): 2026 shows the highest average prices for both cities (~₹4.5 crores for Mumbai). Mumbai consistently commands higher prices across all years. Price trends vary by year - some years show dramatic increases (like 2026, 2022-2023). Thane prices are more stable and generally range from ₹0.5-2 crores

Property Count by Availability Year (Bottom Left - Bar Chart): 2023 has the highest inventory (650+ Mumbai properties, 240+ Thane properties): Mumbai dominates the market volume - roughly 2-3x more properties than Thane: Market activity concentrated in 2022-2025 period: Sharp decline in available properties for years beyond 2025.



Analysis:

Mumbai's Strategy: Premium Market, hence focus on ready properties for rental income. Target Areas can be emerging suburbs with metro connectivity. Budget Allocation: Higher investment, lower volume approach for Long-term hold (5-10 years) for maximum appreciation.

Thane's Strategy: Value Market i.e. Mix of ready and under-construction properties. Target Areas can be Infrastructure development zones. Budget Allocation: Volume play with multiple smaller investments for Medium-term (3-5 years) with potential for quicker exits.

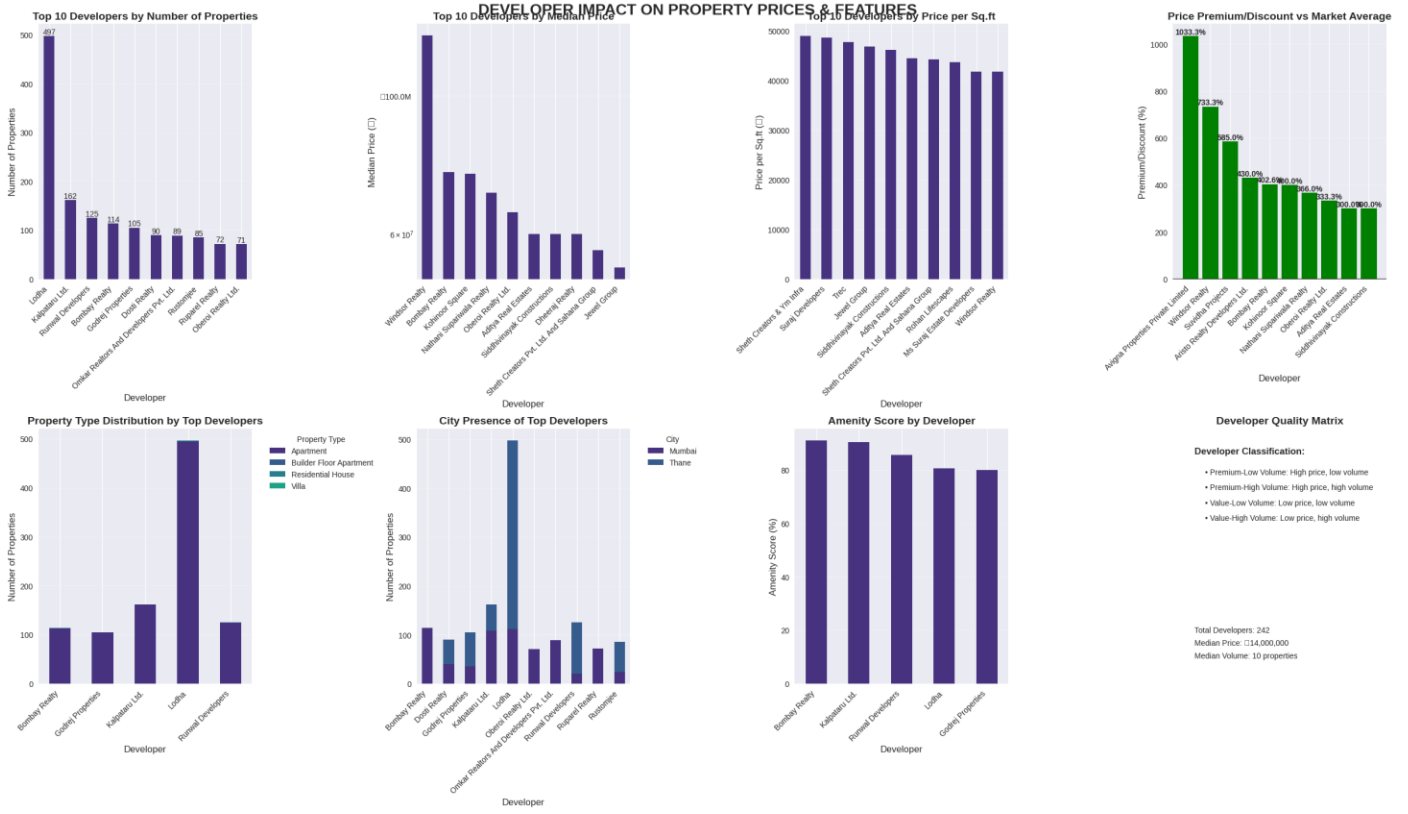
Question 7: Developer impact on properties

The analysis covers 4,537 properties across 255 unique developers, highlighting how developers influence pricing, amenities, and property types. Lodha leads by a wide margin with 497 properties, dominating the high-volume segment, followed by Kalpataru Ltd. (162) and Runwal Developers (125). In terms of median price, premium developers such as Avigna Properties (₹17 Cr) and Windsor Realty (₹12.5 Cr) far outpace others, positioning themselves at the luxury end of the market. Bombay Realty ranks 4th in volume (114 properties) but offers very high-value properties with a median price of ₹7.53 Cr and price per sq.ft. of ₹37,782.

When ranked by price per sq.ft., Oberoi Realty Ltd. leads at ₹40,000/sq.ft, followed by Ruparel Realty (₹34,198/sq.ft) and Bombay Realty, indicating a strong correlation between brand and pricing power. The Price Premium graph reveals that developers like Avigna and Windsor command over 700–1000% premiums compared to the market average, emphasizing exclusivity. Conversely, volume leaders like Lodha and Runwal command relatively modest premiums (under 150%), suggesting a value-for-volume strategy. The Developer Quality Matrix classifies firms based on volume and pricing, distinguishing premium-low volume and value-high volume developers for strategic targeting.

Property type distribution is dominated by apartments, especially for Lodha and Runwal, though developers like Kalpataru and Godrej also offer some Builder Floor and Residential Houses. Lodha has the widest city presence, especially in Mumbai and Thane, reflecting its pan-urban reach. In contrast, many premium developers maintain focused operations, offering exclusivity but limited geographic footprint. Amenity analysis shows Bombay Realty and Kalpataru with the highest amenity scores (~90%), highlighting their focus on high-end offerings. Lodha, despite its volume, has a slightly lower amenity score (~80%), balancing scale with affordability.

Developers with higher amenity scores also tend to be priced higher, suggesting amenities are a key driver of premium positioning. Investors looking for stability and brand reliability may prefer high-volume players like Lodha, while those seeking exclusivity and appreciation might opt for premium developers like Bombay Realty or Avigna. Lastly, the developer landscape is rich and diverse, requiring investors to match budget, location preference, and risk appetite with the appropriate developer classification.



DATASET SUMMARY:

DATASET 1: Contains data regarding the working class and their social status in different societies. It also contains the target capital they want to achieve and their education and population. This is comparatively a better dataset as there are almost 50k entries and null values are also low.

DATASET 2: Contains data regarding the different cities and amenities found under different conditions and varied prices. Requires a lot of pre-processing as there are a lot of NULL values. Comparatively not that good dataset as there are just 12k entries with a lot of null values. In one case, removing null values lead to a subset of just 3k entries which was not good enough for prediction.