Dataset Context

The provided dataset is a **Pakistani real estate listings dataset (Zameen)** containing ~168k property records with attributes such as price, city, location, property type, area, bedrooms, baths, purpose (sale/rent), date added, and geographic coordinates.

Dataset link:
https://drive.google.com/file/d/1RpoJPV7d7qcHmLFHJK0eWXRtBzyeZpoT/view?usp=sharing

Participants will work **entirely in a Jupyter Notebook**. Any visualization tool or library is allowed (Matplotlib, Seaborn, Plotly, Tableau, Power BI, Excel, etc.), as long as outputs are clearly embedded or attached.

Competition Structure (2 Hours)

- **Part A: Data Understanding & Cleaning (Mandatory)**
- **Part B: Scenario-Based EDA & Visualization Tasks (Choose any 3 out of 6)**
- **Part C: Insight Summary (Mandatory)**

Part A – Data Understanding & Cleaning (Baseline Task)

Objective: Ensure the dataset is analysis-ready.

Participants should:

1. Inspect data types and basic statistics.
2. Identify and handle missing values (e.g., bedrooms, baths, area, agency/agent).
3. Detect and treat outliers in price and area (justify chosen method).
4. Standardize or validate area-related fields (Area Type, Area Size, Area Category).
5. Create at least **one derived feature** (e.g., price per marla, price per bedroom, property age proxy using date_added).

Deliverable:

- Cleaned dataframe
- Short markdown explanation of decisions

Part B – Scenario-Based EDA & Visualization Tasks

Scenario 1: Real Estate Investor Decision

Context: An investor wants to maximize return by buying properties in cities with strong value potential.

Tasks:

- Compare **price distributions** across major cities (Islamabad, Lahore, Karachi, Rawalpindi, etc.).
- Analyze **price per unit area** by city and property type.
- Identify **undervalued cities or locations** using visual evidence.

Expected Visualizations:

- Boxplots or violin plots of prices by city
- Bar/line charts for average price per area
- Optional: Geo-visualization using latitude/longitude

Scenario 2: Buyer Affordability Analysis

Context: A middle-income buyer has a limited budget and wants the best living options.

Tasks:

- Define an affordability threshold (justify it).
- Analyze how **bedrooms, baths, and area** vary under this budget.
- Compare **Flats vs Houses** in affordable segments.

Expected Visualizations:

- Scatter plots (price vs area / bedrooms)
- Stacked bars for property type distribution
- Faceted charts by city

Scenario 3: City-Level Market Comparison

Context: A policymaker wants to understand housing disparities across cities.

Tasks:

- Rank cities by **median price**, **price per area**, and **listing volume**.
- Identify cities with high demand but limited supply.
- Highlight inequality or concentration patterns.

Expected Visualizations:

- Ranked bar charts
- Heatmaps (city vs metrics)
- Lorenz-style or cumulative distribution plots (optional)

Scenario 4: Property Type Behavior

Context: A construction company is deciding what type of properties to build next.

Tasks:

- Compare **Houses vs Flats** across price, area, and bedrooms.
- Analyze how property type preference changes by city.
- Identify the most common property configuration.

Expected Visualizations:

- Grouped bar charts
- Boxplots by property type
- Mosaic or percentage plots

Scenario 5: Time-Based Market Trends

Context: Analysts want to understand how the market evolves over time.

Tasks:

- Analyze listings over time using `date_added`.
- Identify seasonal or monthly trends in prices or volume.
- Compare recent vs older listings.

Expected Visualizations:

- Time-series line charts
- Rolling averages
- Monthly/quarterly trend plots

Scenario 6: Location Intelligence & Mapping

Context: A real estate portal wants to enhance its location-based recommendations.

Tasks:

- Use latitude and longitude to visualize spatial price patterns.
- Identify high-price and low-price clusters.
- Compare central vs peripheral locations within a city.

Expected Visualizations:

- Scatter maps
- Density or cluster plots
- Annotated location charts

Part C – Insight Summary (Mandatory)

Participants must provide a **markdown summary** answering:

- What are the **3 most important insights** discovered?
- Which visualization best supports each insight?
- What **real-world decision** could be made using these insights?

Clarity and storytelling matter.

Evaluation Metrics & Scoring Rubric (100 Points)

1. Data Understanding & Cleaning (20 Points)

- Correct handling of missing values (8)
- Outlier detection and justification (6)
- Feature engineering & validation (6)

2. Analysis & Insights (30 Points)

- Relevance to scenario (10)
- Depth of EDA and reasoning (10)

- Correct interpretation of patterns (10)

3. Visualization Quality (25 Points)

- Correct chart choice for data (10)
- Clarity, labeling, and readability (10)
- Effective comparison and storytelling (5)

4. Methodology & Logic (15 Points)

- Structured workflow in notebook (8)
- Clear assumptions and explanations (7)

5. Creativity & Interpretation (10 Points)

- Novel insights or angles (5)
- Innovative or advanced visualizations (5)

Notes for Participants

- No single “correct” answer exists.
- Justification and reasoning are as important as plots.
- Clean, well-documented notebooks score higher than complex but unclear work.