Descriptive Analysis

1. Introduction

The notebook appears to focus on performing **descriptive and exploratory data analysis** on two datasets: geometries and simulations. The purpose is to understand relationships among building layouts, sunlight exposure, noise, and other simulation results.

2. Libraries Imported

The following Python libraries are used in the analysis:

- **pandas**: For data manipulation and analysis.
- **matplotlib.pyplot and seaborn**: For visualizations.
- **numpy**: For numerical computations.
- Additional libraries like scipy.stats and statsmodels might be used for statistical analysis (if included in later cells).

3. Data Loading

- **geometries Dataset**: Contains spatial or geometric data, likely for apartment buildings or units.
- **simulations Dataset**: Provides results of simulations, including metrics like sunlight exposure, noise levels, and accessibility.

4. Exploratory Data Analysis (EDA)

4.1 Dataset Overviews

- The notebook explores basic properties of both datasets:
  - **Shape**: Number of rows and columns.
  - **Columns**: Lists all column names and data types.
  - **Missing Values**: Summarizes null values in each column.

4.2 Statistical Summaries

- Uses describe() to compute basic statistics (mean, median, std dev, etc.) for numeric columns.

5. Data Cleaning

Includes:

- **Handling Missing Values**:
  - Imputing missing data (e.g., replacing NaNs with column means or a default value).
- **Data Type Conversions**:

    - Optimizing memory usage by converting data types (e.g., float to float32).

6. Visualization

6.1 Histograms

- Visualizes the distribution of numerical columns, likely focusing on attributes like layout_area, sun_*, and noise_*.

6.2 Correlation Heatmaps

- Displays relationships between simulation metrics (e.g., view_greenery_mean, sunlight_mean).

6.3 Scatterplots

- Plots relationships such as:
    - layout_area vs. layout_room_count.
    - layout_compactness vs. other geometric properties.

6.4 Spatial Visualizations

- Attempts to visualize spatial relationships between buildings/units using geopandas or scatter plots.

7. Domain-Specific Analysis

7.1 Sustainability

- Investigates energy efficiency by analyzing relationships between:
    - sunlight exposure (sun_* metrics).
    - layout_net_area and view_greenery_mean.

7.2 Accessibility

- Analyzes features like:
    - connectivity_entrance_door_distance_mean.
    - floor_has_elevator.

7.3 Noise Pollution

- Explores:
    - noise_traffic_day.
    - noise_train_day in relation to apartment geometry.

7.4 Building Efficiency

- Examines compactness (layout_compactness) against:
    - layout_room_count.
    - layout_area.

8. Challenges Encountered

The notebook mentions challenges like:

- **Large Dataset Issues**:
    - Loading and merging large datasets causes memory errors.
- **Data Alignment**:
    - Handling mismatched indices when performing operations across datasets.

9. Statistical Models and Predictive Analysis

9.1 Regression Analysis

- Examines relationships between simulation metrics (e.g., view_greenery_mean) and geometric data (e.g., area_id).

9.2 Cluster Analysis

- Performs clustering (e.g., using K-means) to group apartments or buildings based on:
    - Sunlight exposure.
    - Noise levels.
    - Connectivity metrics.

10. Results and Insights

While the file doesn't explicitly show outputs, likely conclusions from this type of analysis include:

- **Sustainability Patterns**:
    - Apartments with higher sunlight exposure have better energy efficiency.
- **Accessibility Metrics**:
    - Buildings with shorter entrance door distances or elevators are more accessible.
- **Noise Trends**:
    - Noise pollution varies with proximity to traffic or train lines.
- **Efficiency Trade-offs**:
    - Compact layouts may sacrifice room count for space efficiency.

11. Final Steps

The notebook likely ends with:

- Summarizing key findings.
- Recommendations for urban planning or apartment design based on insights.