**Potential Scam Detection – Arrondissement Level**

**NOTE: PLEASE READ THE FOLLOWING TO UNDERSTAND THE PROCESS TAKEN TO IDENTIFY THE OUTLIERS IN MERGED DATASET**

Understanding the impact of hedonic characteristics such as the number of bathrooms and rooms is crucial for assessing price sensitivity in housing markets. These characteristics can significantly influence the price of a house, whether for purchase or rental. Although there may be a general correlation of 0.71 between price and sale across the entire dataset, it is essential to recognize that this correlation can vary widely at a more granular level. For example, a house with an area of 100 m² in location A might be priced at 500K, whereas a similar-sized house in location B might be priced at just 10K. Therefore, focusing on each arrondissement individually rather than treating the whole of Paris as a single entity is critical for a more accurate and nuanced understanding of price dynamics.

**The process taken to code the following steps using Python:**

**Data Preprocessing**

1. The merged dataset was cleaned to remove any remaining null values
2. Created a derived variable, *Price per Square Meter (Price/Sqm)*, to assess affluence levels.
3. Added a column to map arrondissement numbers to their corresponding names.
4. Extracted a subset from the merged data frame focused on sales (housing) data to streamline the analysis.

**Percentile Analysis**

1. Categorized properties into percentile ranges (0-10th, 10-20th, ..., 90-100th) based on *Price/Sqm* to understand affluence distribution.
2. Counted the number of properties within each percentile range for each arrondissement to measure distribution by affluence.
3. Created bar charts to visualize the distribution of properties across percentile ranges for each arrondissement.

**Analysing Outliers (Potential Scams) using Confidence Interval**

1. Calculated an 80% confidence interval to determine if a customer's chosen *Price/Sqm* falls within the expected range, considering a margin of error. Only arrondissements with a sample size above 30 were considered for the confidence interval analysis, as smaller sample sizes may not provide reliable results.