RIPHAH
INTERNATIONAL
UNIVERSITY

Lab 2 Assignment:

We have already practiced data cleaning on the Titanic dataset. Now it's your challenge: apply the same techniques on the Airbnb NYC dataset (AB_NYC_2019.csv). Visualize missing values, drop columns with excessive missingness, impute both numeric and categorical values, detect and treat outliers using the IQR method, and finally save your cleaned dataset. Show that you can perform the complete cleaning workflow on your own.

Lab 2 – Tasks (Airbnb NYC Dataset)

1. Visualize missing values with a heatmap. Identify which columns have missing data and which one has the most.
2. Drop the last_review column and explain why dropping is reasonable.
3. Impute reviews_per_month with the median. Plot histograms before and after imputation, and explain why median is preferred.
4. Impute missing values in the name column with the most frequent value. Print that value and discuss whether imputing names is meaningful or biased.
5. Create a boxplot for price and describe its spread and skewness.
6. Detect outliers in price using the IQR method (calculate Q1, Q3, IQR, upper whisker) and count them.
7. Cap price at the upper whisker, create a new column price_capped, and compare old vs new boxplots.
8. Explain why capping outliers may be better than deleting them in this dataset.
9. Compare using **mean** vs **median** for imputing reviews_per_month. How would the choice affect the distribution?
10. Save the cleaned dataset as AB_NYC_2019_cleaned.csv and summarize the changes compared to the raw dataset.