Data Cleaning and Exploratory Data Analysis (EDA) Report

1. Data Loading and Initial Cleaning

- **Library Imports:** The analysis was performed using **pandas, NumPy, seaborn, and matplotlib**.
- **Dataset Loaded:** Read using pd.read_csv().
- **Column Cleanup:**
- Dropped columns with **more than 50% missing values**.
- Removed string-based columns that couldn't be categorized.
- Standardized column names by stripping whitespace and converting them to lowercase.

2. Handling Missing Values

- **Detection Method: Used df.isnull().sum() to identify missing values.**
- **Imputation Strategy:**
- **Backward fill (b_fill) applied for categorical data.**
- **Mean/Median Imputation for numerical columns where applicable.**
- **Dropped Rows if missing values exceeded a threshold.**

3. Duplicate Detection and Removal

- **Checked for duplicate records using df.duplicated().sum().**
- **Removed Exact Duplicates using df.drop_duplicates().**

4. Outlier Detection and Treatment

Outlier Detection Methods:

- **Z-Score Method: Identified outliers using z-score > 3.**
- **Interquartile Range (IQR) Method:**
- **Outliers were flagged if they fell outside Q1 - 1.5*IQR or Q3 + 1.5*IQR.**

Outlier Handling:

- **If Outliers <10% of Data: Rows were dropped.**

- **If Outliers >10% of Data: Winsorization was applied (clip(lower, upper)).**

5. Exploratory Data Analysis (EDA)

Univariate Analysis (Single Variable)

- **Summary Statistics:**
- **Mean, Median, Variance, Skewness calculated using .describe() and .skew().**

Visualizations Used

- **Histograms**: Showed the distribution of numerical features.
- **Box Plots**: Identified data spread and detected outliers.
- **Pie Chart**: Represented the categorical distribution, displaying the proportion of each permit type.

Bivariate Analysis (Two Variables):

- **Correlation Matrix (df.corr()) to analyze relationships.**
- **Scatter Plots: Relationship between continuous variables.**
- **Bar Plots & Violin Plots: Comparison of categorical and numerical variables.**
- **Heatmaps: Visualized relationships between categorical and numerical features.**

6. Multivariate Analysis (Multiple Variables):

- **Categorical vs. Numerical Heatmaps**
- **Example: Neighborhoods vs. Permit Types vs. Street Name**
- **Used pd.crosstab() to create frequency tables.**
- **Grouped Bar Plots**
- **Example: Permit Type vs. Existing Units grouped by Neighborhoods**
- **Swarm Plots & Violin Plots**
- **Example: Number of Stories vs. Street Number categorized by Neighborhoods**

Key Findings and Insights

1. **Permit Volume Trends**

- **The number of permits issued varies yearly, influenced by economic conditions and urban development policies.**

2. **Neighborhood Development Patterns**

- **Certain neighborhoods show significantly higher construction activity, indicating concentrated urban growth.**

3. **Processing Delays**

- **The time from permit application to approval varies, highlighting inefficiencies in the approval process.**

4. **Cost Disparities**

- **Some permit types have substantially higher estimated costs, reflecting differences in project scale and complexity.**