Property Repairs Data Analysis Report

By Dermot Madsen

GitHub Repo: Cluid-Housing-Task

1. Overview

This report presents an analysis of a synthetic property dataset to understand total repair costs, and the key factors influencing this feature. The analysis is divided into two sections:

1. **Descriptive Analysis** – summarising the data and understanding distributions.
2. **Predictive Analysis** – building a decision tree model to identify features that drive total repair costs.

---

2. Objective

- **Descriptive Analysis:** Explore the dataset, examine distributions, and identify patterns in property repairs, costs, and age.
- **Predictive Analysis:** Identify which factors most influence total repair costs and assess predictive performance using a Decision Tree Regressor.

---

3. Data Description

Visualisations

To quickly visualise how each numeric variable is distributed, the code uses a list of histograms that plot the 162,500-row count for each feature – Figure 1.

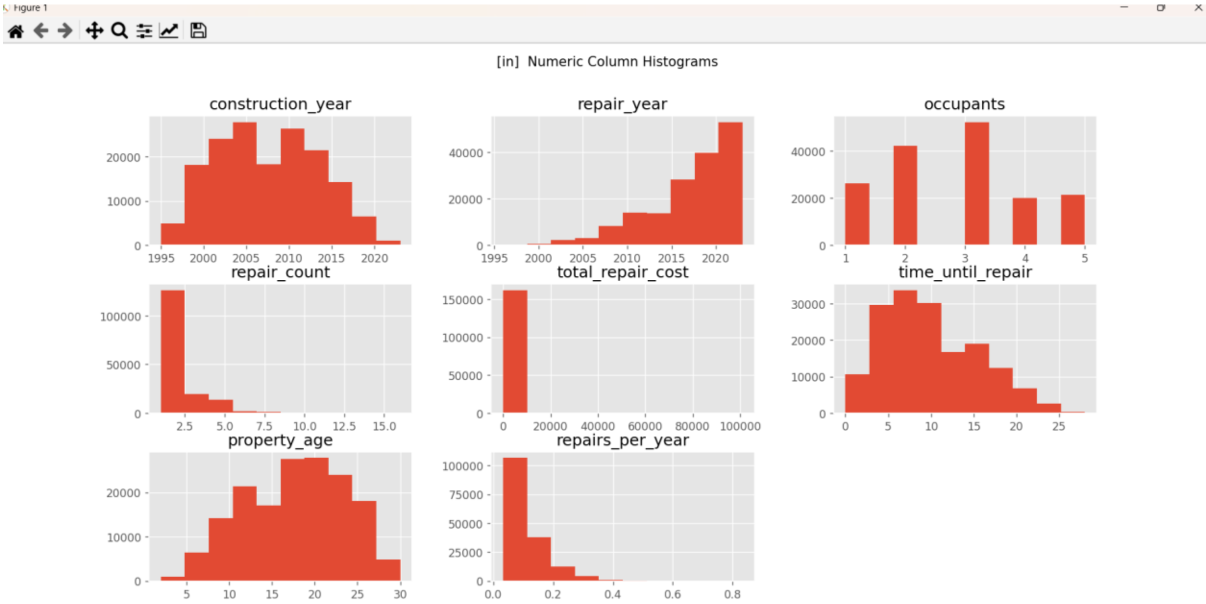


Figure 1- Histogram of variables

Quick Insights:

- **Construction Year:** Evenly distributed
- **Repair Year:** Left-Skewness
- **Occupants:** Evenly distributed
- **Repair Count:** Right-Skewness
- **Time Until Repair**: Right-Skewness
- **Property Age:** Evenly distributed
- **Repairs Per Year:** Right-Skewness

A skewed distribution can be an indication of the presence of outliers in the dataset. Here in Figure 2. there is an obvious cost repair outlier identified at repair year 2007/2008

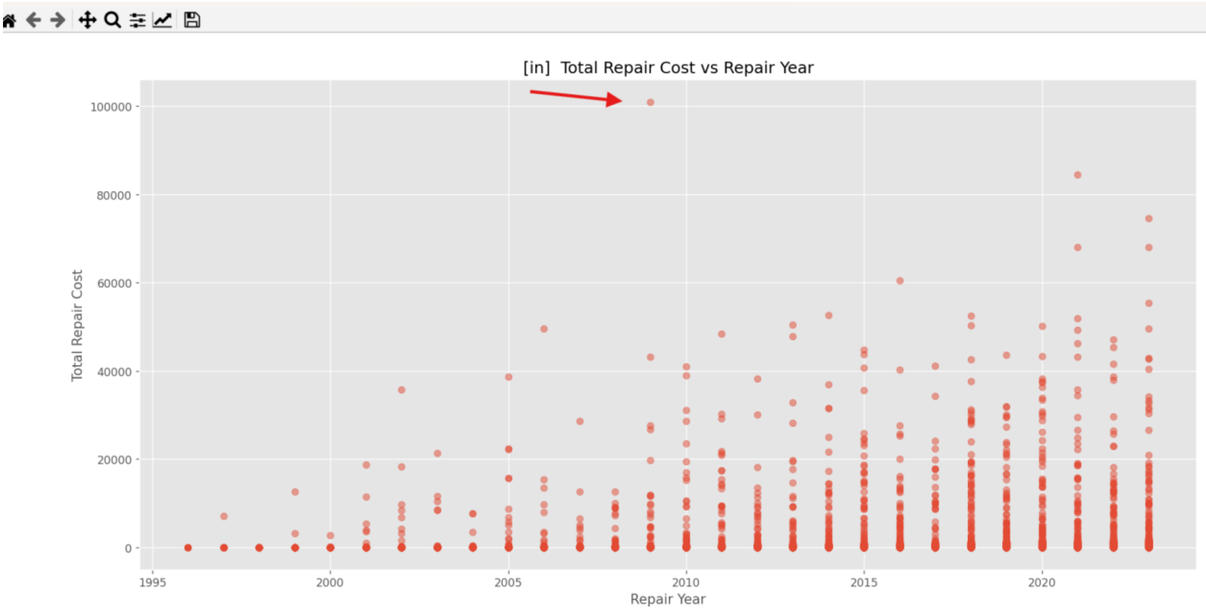
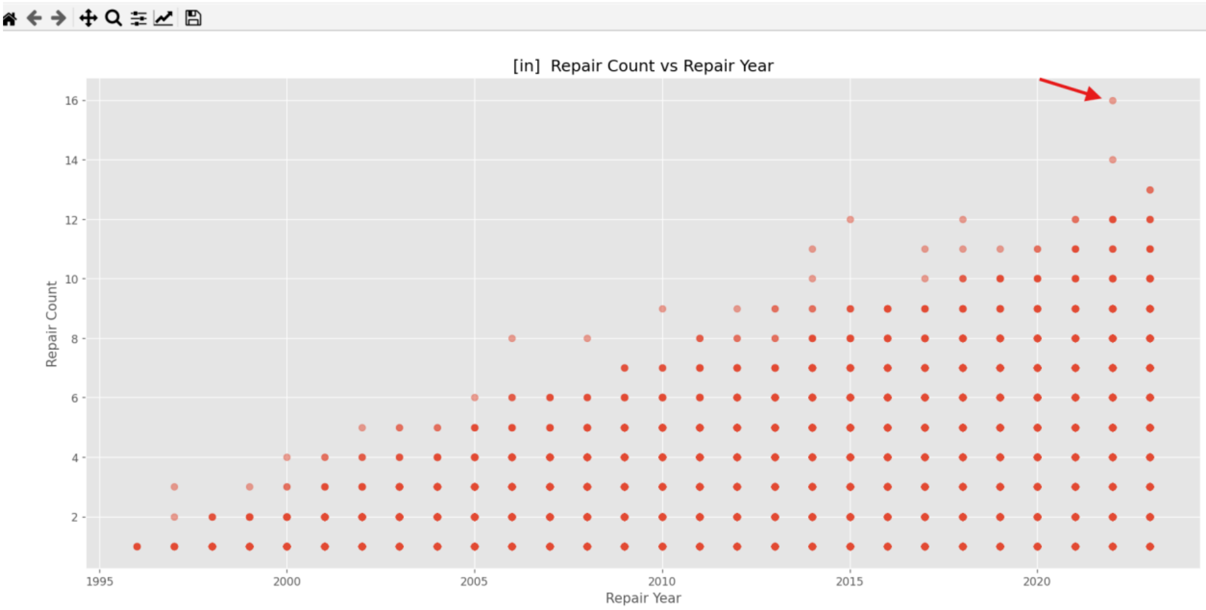


Figure 2 - Scatter Plot Repair Year vs Total Repair Costs

Again, Ther is an obvious outlier visible showing a repair count of 16 repairs in repair year 2016 approximately.



Descriptive Analysis Table

The dataset contains 162,500 properties with the following key numeric variables:

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
construction_year	162500	2007.38	5.90	1995	2003	2007	2012	2023
repair_year	162500	2017.26	4.95	1996	2014	2018	2021	2023
occupants	162500	2.80	1.23	1	2	3	4	5
repair_count	162500	1.88	1.27	1	1	1	2	16
total_repair_cost	162500	134.18	1194.63	0	9.22	30.33	90.06	100988.72
time_until_repair	162500	9.88	5.50	0	6	9	14	28
property_age	162500	17.62	5.90	2	13	18	22	30
repairs_per_year	162500	0.108	0.070	0.032	0.056	0.087	0.138	0.833

Table 1 -Descriptive Analysis Table

Observations from Descriptive Analysis Table:

- **Repair Cost Skew:** total_repair_cost is highly right-skewed:
    - Mean = 134, Median = 30, Max = 100,989
    - Large standard deviation indicates extreme outliers (heavy tails).

- **Repair Frequency Skew:** repair_count is also right-skewed:
  - Most properties have 1 repair, but some have up to 16 repairs.
- **Property Age:** Most properties are 13–22 years old, average 17.6 years.

Correlations between key variables:

Feature	repair_count	total_repair_cost	time_until_repair	occupants	property_age	repairs_per_year
repair_count	1.000	0.690	0.410	0.314	0.253	0.786
total_repair_cost	0.690	1.000	0.402	0.346	0.222	0.487
time_until_repair	0.410	0.402	1.000	-0.099	0.625	0.026
occupants	0.314	0.346	-0.099	1.000	-0.074	0.340
property_age	0.253	0.222	0.625	-0.074	1.000	-0.302
repairs_per_year	0.786	0.487	0.026	0.340	-0.302	1.000

Table 2 - Correlation Table

Key Insights from Correlation Table:

- repair_count and repairs_per_year are strongly correlated with each other and with total_repair_cost.
- time_until_repair and occupants show moderate correlation with costs.
- Property age is moderately correlated with time until repair but has a weaker effect on cost.

---

4. Predictive Analysis

4.1 Objective

To identify which features are most influential in predicting **total repair costs** using a **Decision Tree Regressor**.

4.2 Methodology

- **Features Used:** repair_count, time_until_repair, occupants, property_age, repairs_per_year
- **Target Variable:** total_repair_cost (log-transformed to reduce skew)
- **Model:** Decision Tree Regressor with max_depth=5
- **Evaluation Metric:** Root Mean Squared Error (RMSE) on original cost scale

4.3 Results

Correlation Summary:

- Strongest correlations with repair cost: repair_count (0.69) and repairs_per_year (0.49)
- Moderate: time_until_repair (0.40), occupants (0.35)
- Weak: property_age (0.22)

Decision Tree Performance:

- **RMSE (original scale):** 94.84

Feature Importance:

Feature	Importance
repair_count	0.798
time_until_repair	0.118
occupants	0.084
property_age	0.000
repairs_per_year	0.000

Interpretation:

- Repair count is the primary driver of repair costs.
- time_until_repair and occupants provide minor contributions.
- property_age and repairs_per_year do not improve the model significantly.

---

5. Conclusion

- **Descriptive Analysis:**

    - Repair cost and repair frequency are highly skewed, with most properties having low costs and few repairs, but with heavy right tails.
    - Strong relationships exist between repair_count and total_repair_cost.
- **Predictive Analysis:**
    - A Decision Tree Regressor confirms that the number of repairs is the dominant factor influencing total repair costs.
    - Other variables have minor effects, and some (like property_age or repairs_per_year) do not improve predictive power.

Recommendation:

- For predicting or managing repair budgets, focus primarily on repair frequency.
- Consider handling outliers and skewness in cost data when building more advanced predictive models.