Data Quality Report – Initial Findings

# 1. Overview

Based on the cleaned dataset (ppr 21200542 cleaned.csv), this report will summarize the early findings. It will include a summary of the data, as well as a description of the many data quality concerns that have been identified and how they will be resolved. For further information on this dataset, please consult the appendix. Terminology, assumptions, explanations, and a summary of modifications to the original dataset are included in the appendix. This covers data visualizations such as feature summaries, histograms, and boxplots.

The dataset looks to be quite clean at first glance. Null values, duplicate columns, and columns with unusual cardinalities are not present. Special values for continuous data and logical error in categorical data were the two most common concerns encountered. A considerable number of outliers were also observed.

In addition, the material was subjected to many logical tests, which revealed a substantial number of discrepancies.

# 2. Summary

Several tests were run to ensure that the data was logically sound. This resulted in a substantial number of data failures. There were a total of 33 instances of illogical data found. For example, in 21 cases, the address does not match the county value or the post code does not match the county. This is obviously not doable. This unreasonable data will need to be handled with, and the domain expert should be consulted. For further information, see the logical integrity section.

There are several outliers in the continuous characteristics that will have a substantial impact on the outcome. This unreasonable data will need to be handled with, and the domain expert should be consulted. For further information, see the logical integrity section.

Several adjustments are suggested for categorical values. In the 'Postal Code' and 'Property Size Description' features, there are several missing values. Furthermore, the meanings of "greater than 125 sq metres" and "greater than or equal to 125 sq metres" are interchangeable. As a result, the value should be modified to "more than or equal to 125 sq metres." It's a good idea to double-check this with a domain specialist.

Across the feature set, there were a considerable number of outliers. However, these figures look credible at first glance, but they need be scrutinized carefully.

# 3. Review Logical Integrity

7 tests were carried out. The failures are below;

• Test 1 - Check if Post code is in dublin but County is another county which is impossible.
    o We saw that 21 instances failed this test.

• Test 2 - Check if there are irregular address or ambiguous address.
    o 3 cases found

• Test 3 - Check for cases that are later than 24 January 2022(impossible)
    o 8 cases found

# 4. Review Continuous Features

## 4.1. Descriptive Statistics

There are two continuous features. One is Price (€) and the other is Date of Sale (dd/mm/yyyy).

• Logical error in data time (Count 8)

    o We can see that there are 8 cases which are later than "2022/02/01", which is impossible.

• Outliers in Price (Count 554)

    o We can see that there are 554 cases, this unreasonable data will need to be handled with, and the domain expert should be consulted.

# 5. Examine Categorical Characteristics

## 5.1 Descriptive Statistics

The dataset has seven categorical features: Address, Postal Code, County, Not Full Market Price, VAT Exclusive, Property Description, and Property Size Description. They both logically have a substantial link to the price outcome. However, the Address has a large cardinality, and the postal code and property size descriptions both include several missing entries.

## 6. Action to take

- Logical integrity
    - o Rows failing the logical test will need to be dropped
- Categorical Features
    - o Map both numerical scales to a common scale
    - o Change all “unknown delinquency” values to “current or never delinquent”
- Outliers
    - o Review outliers, checking for validity