Data Quality Plan

17350796

From the financial crash to the crippling housing crisis: What is the future of Ireland's housing nightmare?

2022

Here I provide a plan of the necessary solutions to the problems outlined in the data quality report.

The list of issues and solutions are as follows:

- **Date of Sale (dd/mm/yyyy):**

This data was converted to datetime. The months and the years have been extracted as separate features in order for it to be easier to compare them to other features. As mentioned in the data quality report, tt was discovered that there was dates that indicated that they were sold in the future. All the dates appeared as they were sold later in the year 2022 on the 1st day of the month. It was concluded that this was error that happened when the data was being logged. It is believed that the day and month got mixed up, thus, they were all actually sold in January of 2022. All the dates were corrected.

It is important to note that it is of course possible, that for other years, days and months were mixed. However, if a day does not match a month (e.g. the 31st of February) the datetime type should flag the error. Furthermore, if there is such a mix up, I do not believe it would have a significant impact on our overall aim of predicting price.

- **Postal Code**:

A large proportion of the sales addresses have no postal code attributed to them. I have opted to keep as could provide further geographical information about Dublin which is the most represented county in the data. In the following section I will attempt to parse the address in order to gain more postal codes. Other potential solutions could be incorporate EirCodes or longitude and latitude to replace the postal codes.

- **Price**:

There is a large amount of outliers in this feature. I have discovered that some of the largest outliers are due to the fact they are sales of large blocks of apartments. I have taken the decision to remove these sales as our aim is to predict the price of single dwellings. There was no meaningful way to break up these apartments with accurate prices attributed to them.

- **Description of Property**:

Some property descriptions were in Irish. These Irish descriptions were translated so there were only two unique values for consistency.

- **Property Size Description**:

This feature has a large amount of missing data, nearly 90%. However, this data set lacks unique features regarding the house sales specifications. This features adds something that can distinguish houses. This type of information is very important for machine learning. For this and the overall aim of avoiding deleting information I have opted to keep this information.

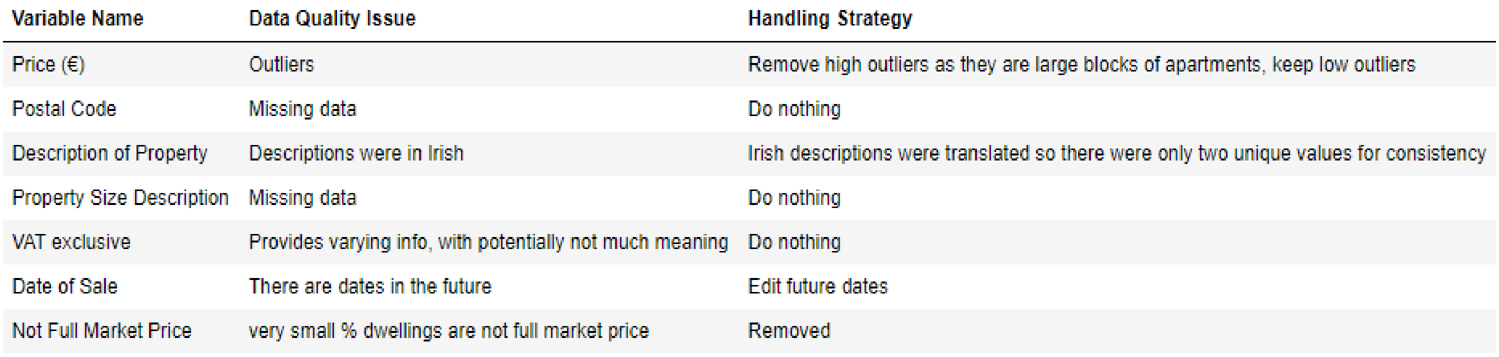
- **Vat Exclusive**:

The issue with this feature is that it is difficult to interpret what this means and how it impacts the feature. Less than 17% of the sales do not include the VAT in the sale price. I have chosen not to separate this data or drop it as the sales are important and

- **Not Full Market Price**:

I have decided to remove those addresses that are not full market price in order to keep as much consistency as possible with the price data. This will only impact a small amount of rows (448 rows only).

Here is the final summary table:



Variable Name	Data Quality Issue	Handling Strategy
Price (€)	Outliers	Remove high outliers as they are large blocks of apartments, keep low outliers
Postal Code	Missing data	Do nothing
Description of Property	Descriptions were in Irish	Irish descriptions were translated so there were only two unique values for consistency
Property Size Description	Missing data	Do nothing
VAT exclusive	Provides varying info, with potentially not much meaning	Do nothing
Date of Sale	There are dates in the future	Edit future dates
Not Full Market Price	very small % dwellings are not full market price	Removed