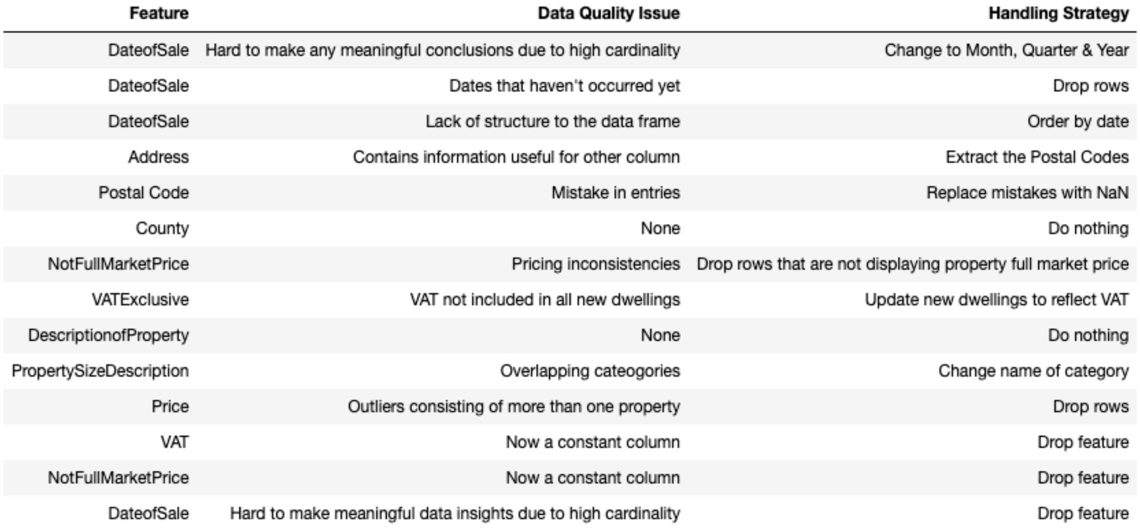
Data Quality Plan



Feature	Data Quality Issue	Handling Strategy
DateofSale	Hard to make any meaningful conclusions due to high cardinality	Change to Month, Quarter & Year
DateofSale	Dates that haven't occurred yet	Drop rows
DateofSale	Lack of structure to the data frame	Order by date
Address	Contains information useful for other column	Extract the Postal Codes
Postal Code	Mistake in entries	Replace mistakes with NaN
County	None	Do nothing
NotFullMarketPrice	Pricing inconsistencies	Drop rows that are not displaying property full market price
VATExclusive	VAT not included in all new dwellings	Update new dwellings to reflect VAT
DescriptionofProperty	None	Do nothing
PropertySizeDescription	Overlapping cateogories	Change name of category
Price	Outliers consisting of more than one property	Drop rows
VAT	Now a constant column	Drop feature
NotFullMarketPrice	Now a constant column	Drop feature
DateofSale	Hard to make meaningful data insights due to high cardinality	Drop feature

Reasoning as described in markdown in notebook:

1) Dates that have not yet occurred:

As mentioned in the logical integrity section, there was a date entered that had not yet occurred and this was to be deleted. Furthermore, as the year is not finished and this data set could only show information up until February/March of 2022, this could cause unreliable conclusions in future sections. For example, in the examination of house sales per month, January and February would have an extra of year of sales in comparision to the other months. Hence, the decision was made to exclude all data from 2022.

2) Ordering by date:

No structure in the csv file.

3) Extracting information from Address column:

The address field appears to contain a large number of Postal Codes for the properties sold in Dublin. This is useful information and therefore has been extracted and placed into the "Postal Code" column. The address field will remain the same otherwise as it will be useful for obtaining coordinates in a later section.

4) Postal Code mistakes

Have been entered for counties other than Dublin

5) Not Full Market Price

As only properties with prices reflecting their full market price are not in the data frame, this column is now constant with only one value. This offers no insight and hence will be dropped.

6) VAT Exclusive

Dealing with the failed findings from test 8 in the logical integrity section including the VAT. This is important as the buyer of the house will consider what he/she paid for the property when they are deciding on a suitable selling price. All property described as "New Dwelling house /Apartment" has been increased by 13.5%. This in accordance with what is said about this type of property on the Property Price Register wesbite (https://www.propertypriceregister.ie/).

7) Property Size Description

There has been a mistake in the 'Property Size Description' column with two categories overlapping. The decision has been made to replace all entries with "greater than 125 sq metres" with "greater than or equal to 125 sq metres". There are now three categories for the "Property Size Description".

8) Outliers in Price

There is a considerable difference between the 75th percentile and max price value in the data frame. This along with the histograms and the box plots from section 1 of the report suggests extreme outliers in the data.

The easiest way to handle outliers is to use a clamp transformation. This clamps all the values above an upper threshold and below a lower threshold to these threshold values, and then removing any outliers:
* lower threshold: 1st quartile value minus (1.5 x interquartile range)
* upper threshold: 3rd quartile value plus (1.5 x interquartile range)

Any values outside these thresholds are then converted to the threshold values. While this is an easy way to deal with the outliers, it is not necessarily the best option for this data set, the reason being that while the outliers are different from the rest, they provide valuable information. For example, in Dublin, on a number of occasions a house sells for €900,000 plus. This should be kept with no change to convey the high prices for certain areas.

In terms of price outliers, the decision has been made to delete data that contains more than one property:

The following is from the Property Price Register website (https://www.propertypriceregister.ie/):

Where a number of apartments are sold for a single price (e.g. Apartments Nos 1 to 15 Oak Drive are sold for €2m), the information on the Register depends on what was filed with the Revenue Commissioners.

For example, the filer may have input the address of - only one of the apartments (e.g. No. 1 Oak Drive) for a price of €2m; each apartment separately and divided the price between each apartment (e.g. Apartment No. 1 for €133,333, Apartment No.2 for €133,333 and so on to Apartment No. 15 for €133,333. all the apartments (e.g. Nos 1 to 15 Oak Drive) for a price of €2m. The Date of Sale is the date input by the filer of the Stamp Duty Return as the date of the Deed transferring ownership of the property."

This is potentially the reason for the large price outliers. Any property that has been inputted into the data set which is an amalgamation of more than one property should be removed as this would otherwise cause unreliable results. For example, in terms of price, this would be increasing the mean for an area. While it was originally thought that the price could be divided by the number of apartments/houses, there was a lack of information in some fields. Furthermore, in the case that it was made clear how many properties it included, this did not mean that each property was priced an equal amount.

9) <u>VAT</u>

Now a constant column

10) Not Full Market Price

Now a constant column

11) Date of Sale

The DateofSale column has already been separated into month, quarter and year. This data will be used from now on. The high cardinality of DateOfSale would make it hard to make any meaningful data insights and hence will be dropped from the data frame.