**Data Quality Report**

**17350796**

This report aims to give an initial discussion of my findings of the cleaned data set. It seeks to describe the data quality issues associated with this file. My initial impression was that the data was relatively clean with good logical integrity. Several tests were performed on the data to check its logical integrity.

**Types**

‘Date of Sale’ was converted to datetime64. ‘Postal Code’, ‘County’, ‘Not Full Market Price’, ‘Vat Exclusive’, ‘Description of property’, and ‘Property Size Description’ were converted to category as they all have a number of fixed possible values.

**Null Values**

‘Property size description’ was dropped as it had over 89% missing values and provided relatively vague information. Currently, postal codes remain in the data set even though they are only relevant to Dublin and only 58% percent have been included. However, Dublin represents over 30% of the data and this information could provide insight into more detailed statistical insight later. Different trends of price and time could be looked at in more detail in different areas of Dublin.

‘Description of Property’ showed 3 unique values. ‘Second-Hand Dwelling house /Apartment’, ‘Second-Hand Dwelling house /Apartment’ and ‘Teach/Árasán Cónaithe Atháimhe’ which is simply the Irish translation for second-hand dwelling house/apartment. While Irish people have the right to use their official language for the purpose of data analysis I have absorbed this value into its English translation leaving only two unique values for this feature.

As mentioned the ‘Date of Sale’ showed no missing data. There was a list of descriptive statistics on this data also, such as frequency, mean, max and min. This data will be used to see trends when compared to the price. If these values are broken up, monthly and yearly trends of price data should be able to be explored. From looking at the histogram you can see that the data is slightly left skewed. With the bulk of the data between 2018 and 2020.

From looking at the ‘Price’ column, this was the only pure numeric data in the csv. There is a large range of data with a large standard deviation. This is visually clear from observing the Price histogram, it is extremely right skewed. This further confirmed when looking at the box plot which shows the outliers. This will have to manipulated in order to make meaningful inferences about this column,

Table of descriptive statistics for all the continuous features:

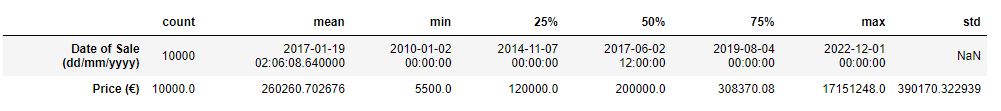


Table of descriptive statistics for all the categorical feature:

Graphical user interface, application

Description automatically generated

Chart

Description automatically generated