**Data Quality Report**

**Overview**

This report will detail my initial findings from the subset of data from the Property Price Register website, which gives information on residential properties sold. The aim of this analysis is to both identify problems with the data that may impinge upon analysis in later sections of this project, as well as to develop a comprehension of what each feature conveys and how it conveys it. This is an essential first step to examining the relationships and interactions between features and will ultimately serve to provide more meaningful analyses.

The dataset consists of 10,000 datapoints and 9 features. The target feature can be considered the Price each property is sold for, and the aim will be price prediction, using the other features in the dataset or datasets found elsewhere. There is 1 numerical column, Price, and 1 datetime column, while the rest are categorical.

**Feature Overview & Logic Checks**

Date of Sale (dd/mm/yyyy)

Provides the date that a property was sold on. Once converted to a pandas datetime data type, the format changed to yyyy-mm-dd and so the name was updated.

A check was conducted to ensure that the timespan was over a sensible period.

Address

Provides a comma separated string of the address of each property. There is no clear pattern to the string, with some containing 2, 3 or 4 fields, some ending with the county, some ending with the postal code etc. Some addresses contain the postal code for data points that do not have the postal code given in the Postal Code feature. This will be addressed in section 2.

Postal Code

Provides a postal code for the properties in Dublin city and so this is empty for many data points. All postal codes in Dublin City are represented, however not all properties in Dublin city have an associated postal code.

A check was conducted to ensure all those entries with postal codes provided were in County Dublin. 2 properties outside of Dublin were identified as having incorrect postal codes in this way. A check was also conducted to ensure that all postal codes were represented, and none were misrepresented.

County

Provides the County each property is located in. Each of the 26 counties in the Republic of Ireland are represented.

A check was conducted to ensure each county was spelled correctly/not duplicated.

Price (€)

Provides the price each property was sold at.

The prices were checked for the minimum and maximum. No negative prices were identified, and the prices ranged from c. €5000 to c. €26,000,000, a reasonable range.

Not Full Market Price

Indicates if the property was sold at its full market value.

A check was conducted to ensure the entries were only Yes/No.

VAT Exclusive

Indicates if the price listed is inclusive of the standard 13.5% rate.

A check was conducted to ensure the entries were only Yes/No. It was identified that only some new build properties were exclusive of VAT, which will need to be addressed later.

Description of Property

Indicates if the property sold is new or second hand.

A check was conducted to ensure all properties fell into one of the 2 categories.

Property Size Description

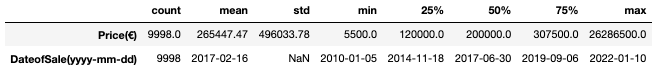
Details the size of the property – limited availability across the dataset.

A check was carried out to see the categories – ‘greater than 125 sq metres’ and ‘greater than or equal to 125 sq metres’ essentially mean the same thing.

2 duplicate rows were also dropped from the data to result in 9998 rows and 9 features being used in the rest of the analysis.

**Continuous Features**

There are two continuous features in the dataset, Price (numeric) and date (datetime).



**Observations:**

* Each row in the dataframe has an associated Price and Date. Therefore, missing data is not a factor for either of these features.
* The price has an extremely large range of €26,281,000. Because of this the mean value is considerably greater than the median as there is a positive skew in the distribution, implying the presence of a few, very large, outliers.
* That being said, the standard deviation is indeed greater than the mean (almost double). This means that there is considerable variation between values, and that the median and interquartile range (IQR) will be more effective in describing the distribution, as they are not affected by outliers.
* The IQR implies that the data is centred around a much smaller range than the standard deviation may imply. The IQR is €187,500, with the prices in this range being more representative of what most of the population are paying for properties.
* The dates range from the beginning of 2010 to the beginning of 2022.

**Shape

Description automatically generated**

*Figure 1: Price distribution*

A picture containing text, screenshot

Description automatically generated

*Figure 2: Boxplot of Price*

Chart, histogram

Description automatically generated

*Figure 3: Price distribution up to 2 standard deviations above the mean*

Chart, box and whisker chart

Description automatically generated

*Figure 4: Boxplot of prices up to 2 standard deviations above the mean*

Chart, histogram

Description automatically generated

*Figure 5: Date distribution for properties sold by year*

Chart, histogram

Description automatically generated

*Figure 6: Date distribution for properties sold by year*

**Observations:**

* Due to the positive skew in the data, the plots for the unchanged Price data is essentially meaningless.
* Introducing a maximum price into the charts of 2 standard deviations above the mean (i.e., excluding the top 2.5% of prices) we can see that more meaningful information is available to us. However, it’s clear positive skew is still a factor, and the prices look to be lognormally distributed, a common trend observed in house prices.
* The distribution of sales seems to increase as time goes by, however it is worth being cautious in interpreting anything from this, as we are dealing with a sample dataset with no knowledge of how the data was chosen. It may be that more datapoints were collected from later years than might otherwise have been in a random sample.

At the end of this document there are histograms of the price distribution by county and by postal code, excluding the top 2.5% of prices. It is interesting to note how the counties with larger populations (Dublin, Galway, Cork & Limerick) have a greater concentration of more expensive properties, than do smaller counties such as Donegal, Laois, and Leitrim. Equally as noteworthy is the histograms by Dublin Postal Code, which show a tendency towards higher prices in places like Dublin 4, widely regarded as an expensive area. These histograms can be seen more clearly in the accompanying notebook.

**Categorical Features**

**A picture containing table

Description automatically generated**

**Observations:**

* There are no constant categorical features. All have at least 2 categories.
* 22 postal codes and 26 counties are represented, what we would expect given that the postal codes are only for Dublin and the counties represent those in the Republic.
* Close to 81% and 90% of rows do not have a Postal Code or Size Description, respectively.
* Chart, bar chart

  Description automatically generatedEach of the categories with only 2 options have significant weightings to one of the options.

*Figure 7: Postal Code Distribution*

Chart, histogram

Description automatically generated

*Figure 8: County Distribution*

Chart, bar chart

Description automatically generated

*Figure 9: Histogram of Not Full Market Price feature*

Chart, bar chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated*Figure 9: Histogram of VAT Exclusive feature*

*Figure 10: Histogram of Description of Property feature*

Chart, box and whisker chart

Description automatically generated

*Figure 11: Property Size Description feature*

**Observations:**

* The dataset has a huge number of datapoints in Dublin, meaning predictions made for this county should be more statistically significant. There are quite a few counties that do not have many observations.
* Likewise for Postal Codes, there is a wide variation in the representation of different areas, likely due to their size and the number of houses sold there (although this is an extract from the actual database).
* Most properties were sold for full price. It may be worthwhile investigating if those not sold for full price should be excluded from the dataset.
* There are many second-hand properties in the dataset (reflecting real life as a new property can only be sold once), meaning it may be more relevant to model these types of properties.

**Feature Discussion**

Date of Sale (dd/mm/yyyy)

This is distributed between 2010 and the beginning of 2022. The dates are not evenly distributed with a far greater proportion coming in the later years of the dataset. This would be interesting to explore, however, as this is a subset of the actual dataset and I do not know how it was sampled, this would not produce meaningful insights and for this reason I will not consider this going forward.

Address

The data provided in the address is inconsistent in format, and liable to spelling mistakes and other errors. Therefore, aside from manual solutions, it would not be possible to extract correct data for every property. It should be possible to extract some meaningful information however, such as towns and postal codes, however these will be liable to be incorrect or missing.

In addition, I discovered that many entries in the dataframe pertain to the purchase of blocks of properties. This distorts the distribution of the data as clearly if more than 1 property is included in a row the price will likely be much higher than that of a single property.

Postal Code

All postal codes in Dublin city are represented, with none provided elsewhere. In addition, not all properties in Dublin city have an associated postal code. In some cases, these are provided in the Address field and as such it will be possible to extract them for some properties.

Most counties are not broken into postal codes like Dublin so the data not being provided for other counties is not too important.

It will be interesting to analyse trends based on Dublin postal codes as there will be a large degree of variation in this data given its size. Outside of Dublin I will focus on County or Town level analysis due to the missing data

County

Each of the 26 counties in the Republic of Ireland are represented for every property sold meaning that an analysis by county will be possible. Unsurprisingly there are far more entries for Dublin, Cork Kildare, and Galway.

Price (€)

The prices look to have a lognormal distribution due to the presence of outliers. This lends itself to a simple log transformation of the data to remove outliers and produce more meaningful analysis that have been stripped of unusual prices.

Data will not be removed however, rather a new column created.

Not Full Market Price

Over 10% of the properties were not sold at Full Market Price. According to the PPR website these can be for a wide variety of reasons which we do not have data for. This will be addressed in the Data Quality Plan. However, it is interesting to note that the average price for these properties is considerably less than for the full dataset.

VAT Exclusive

I note that the only properties listed as VAT Exclusive are new properties, but not all of them. Given that the VAT rate is 13.5% it will be necessary to update the prices to align them to the rest of the data.

Description of Property

This is another feature with a binary choice, indicating if the property is new or second hand. Over 83% of the properties in the dataset are second hand.

Property Size Description

This feature is broken into 4 unique values depending on the size of the property. However, 'greater than 125 sq metres' is essentially the same as 'greater than or equal to 125 sq metres', and so I will be able to merge these. Most properties (74.8%) fall into the category ‘greater than or equal to 38 sq metres and less than 125 sq metres’ and so it may be more difficult to make predictions based on size for the other categories given there are not many observations.

**Planned Actions**

* Addresses in current format have little value
  + Change format to list to allow for information extraction
* Missing Postal Codes
  + Extract Postal Codes from the Address Field using lists
* Some Dublin Postal Codes are not in Dublin
  + Remove Postal Code from the rows
* Missing Property Size Descriptions
  + No action
* Duplicated Property Size Description Category
  + Merge categories
* Not all new build properties are inclusive of VAT
  + Add VAT to properties exclusive of VAT
* Some prices not at market value
  + No action
* Dataset contains block purchases
  + Remove properties identified as duplicates from a new price column
* Prices are not normalised - significant outliers
  + Remove log prices that are 2 standard deviations above the mean of each county

Shape

Description automatically generated with low confidence

Shape

Description automatically generated with low confidence