COMP47350: Homework 1
Student ID: 21200122

DATA ANALYTICS – HOMEWORK 1

DATA QUALITY PLAN

Feature Name	Issue	Handling Strategy
Price	Presence of Entries that Refer to Multiple Properties	Split entries out into multiple rows
Price	Presence of high outliers	Replace values below and above threshold to certain quantile
Address	Data has a high cardinality and non-standardised input	Do nothing
Postal Code	Large number of missing values	Change feature name to Postal Code (Dublin)
Not Full Market Price	Feature framed as a negative	Replace "yes" with "no" and vice-versa
Not Full Market Price	Non-boolean values used for boolean category	Change values to boolean values
VAT Exclusive	Non-boolean values used for boolean category	Change values to boolean values
VAT Exclusive / Description of Property	Potential reference to same facet of data	Do nothing
Description of Property	Values can simplified to a simple boolean in relation to second-hand property	Replace category with boolean category
Property Size	Feature has large amount of missing values	Drop feature

1 OVERVIEW

This data quality plan (the **Plan**) is a summary of why specific handling strategies for data quality issues were chosen. It should be noted that the content in this PDF is largely very similar to what is contained in the corresponding section of the notebook.

2 PRICE

2.1 It was decided that splitting out entries that refer to a single property purchase would have the dual effect of representing the true intent of the entry within the context of our dataset (i.e., one property per row) and would help reduce some of the high outliers contained within the dataset. The fact that our assumption that each property within the group purchase cost the same amount may not be correct was deemed to be a small sacrifice when weighed against fixing logical integrity of the data and removing outliers.

2.2 For the rest of the outliers, it was decided to clamp these at the first and ninety-nineth percentiles. When tested against some of the other options – such as clamping using inter-quartile range, as we did for our histogram, or replacing all values above and below a particular threshold with the median – this led to the least number of rows being affected, while still maintaining our expected histogram tendency.

3 ADDRESS

3.1 Ultimately, it was decided to do nothing with this feature. An attempt was made to use the Folium library to transform the address field - however, the library seemed to have had quite a lot of issues with parsing some of the address strings. At the end of the process, only 15%-20% of strings were successfully parsed - this was decided to be too low a value to be considered satisfactory as a replacement for the "Address" feature. Any attempts to manipulate the address data to see if this would allow Geopy to parse successfully - such as changing the case of the address or adding "Ireland" to the end of the string, proved to be unsuccessful.

3.2 As such, the Address feature was left the way it was. The reason that it was not dropped altogether is that perhaps at a later stage a new library will be released, or even a new update for Folium, that will be able to successfully parse the data in this dataset, in which case we will make use of this feature.

4 POSTAL CODE

It does not make sense to drop the entire Postal Code feature, as the missing values just represent the fact that the property is not located in Dublin. For this reason, the feature was renamed to ensure that it more accurately describes the data contained therein. The non-Dublin entries are left as "NaN", as these will represent the non-Dublin properties.

5 NOT FULL MARKET PRICE / VAT EXCLUSIVE

To make the information easier to work with and to understand, the “Not Full Market Price” feature was reframed as a positive as opposed to a negative (as saying "Not Full Market Price" is less intuitive than just saying "Full Market Price". For similar simplification purposes the values were mapped to Boolean values, and a similar Boolean conversion was carried out in respect of the “VAT Exclusive” feature.

6 DESCRIPTION OF PROPERTY

6.1 As mentioned in the main notebook, the "Description of Property" feature currently contains only three values, one of which is in Irish. These values ultimately state if a property is either new or second-hand. We should flag that the value that is in Irish does not seem to specify whether or not the property is second-hand or not, so we will make the assumption that it is referring to a new property. In any event this assumption relates to one entry only, so even if it turns out to be incorrect, it should not cause any significant distortion of the data.

6.2 For these reasons, we can clean the "Description of Property" feature by transforming it into a simple Boolean "Second-Hand" column which will indicate whether or not the purchase is of a second-hand property. The reason we are using "Second-Hand" as opposed to "New" as our feature name is because there are more second-hand properties being sold in our dataframe than new ones, and as such it makes sense to frame the feature using the majority value.

6.3 We will be replacing the current "Description of Property Feature" when we transform it. This is because it is very difficult to glean any information from the feature as it currently stands, and it requires the user to review the actual text values of the entries, which is unintuitive. For these reasons, there is no point in keeping the original feature - we can replace it with our new one.

7 PROPERTY SIZE

After careful consideration, it was decided that keeping this data would not be beneficial to glean any new information for the data set. While there are a few entries from which potentially specific property sizes could be learned about, there are so few of these that it would be impossible to make any general statement about the relationship of the property size to any other feature in the data set. For these reasons, this feature was dropped.

8 VAT EXCLUSIVE / DESCRIPTION OF PROPERTY

8.1 As mentioned in the Data Quality Report, there is a possibility that both VAT Exclusive and Description of Property features are merely referring to the same facet of the data, namely because VAT is payable mostly on new homes, not second-hand ones. If the Description of Property merely refers to whether the property is second hand, then we can potentially drop this feature for redundancy.

8.2 However, upon further review, it was discovered that there are, in fact certain circumstances in which VAT could be payable on second-hand homes. For example, guidance from the Irish Revenue Commissioners (please see

notebook for link reference) states that a vendor and purchaser may, for instance, decide to include VAT in a second-hand property purchase to avoid clawback.

8.3 After a review of the rows where the VAT Exclusive and Description of Property do not match, it was discovered that there are 26 such instances.

8.4 It should be noted that:

- 8.4.1 Irish tax law is not our area of expertise;
- 8.4.2 it is not possible to contact the domain expert to clarify whether or not the mismatched entries are errors; and
- 8.4.3 we have discovered reliable guidance from Revenue to suggest at least one reason as to why there might be a mismatch between Second-Hand and VAT Exclusive features.

8.5 For these reasons, it was decided to keep both features in, as it would seem that there is ambiguity as to whether or not these two features are one and the same, and we wish to minimise loss of data where possible.

END