Data Quality Report

Data Quality Report - Initial Findings

1. Introduction

The report outlines the initial findings of key issues identified in the data based on the cleaned dataset of 'ppr-21200066-updated.csv'. In this report, I will analyse the characteristics of data, describe the key issues I found in the dataset and provide potential solutions to address them. Please check the appendix for the background of this dataset. Appendix includes terminology, assumptions, explanations, and summary of changes made to the original dataset. This also includes feature summaries, histograms and boxplots used to visualise the data.

The dataset currently contains one datetime feature, one continuous feature and seven categorical features. There was initially one row which was duplicate of other one. And it has been removed from the updated csv file. There are two features contains a huge amount of missing values, which are 'PostalCode' and 'PropertySizeDescription'. Aside from that, other issues include the categorical features 'Address' with high cardinalities, overlap of scales of categories of PropertySizeDescription'. Also, the features describing the property size are currently not in a useful format. Finally, I implemented 8 logical integrity tests, and there are some rows of data failed the logical integrity tests.

2. Summary

Several logical integrity tests were implemented to check the logical integrity of the data. These tests checked out 218 instances of potentially irrational data (Here we neglected the number of instances failed in test 4, reason will be demonstrated later). For example, one of the tests showed that there are 48 new properties which are not exclusive of VAT. This is basically inconsistent with the policy that "If the property is a new property, the price shown should be exclusive of VAT at 13.5%". Refer to section 3 for more details on these integrity tests and results.

For the continuous feature, the only issue detected is the outliners. Thus, I implemented outlier handling process in the solutions which will be introduced below.

For the categorical features, the most troublesome issue is the missing values in 'PostalCode' and 'PropertySizeDescription', which account 80% and 89% of the total number value respectively. Solutions are recommended below to fix this issue.

3. Review Logical Integrity

Eight tests were carried out. The failures are below:

- Test 1: check whether New Dwelling house/Apartment have a VAT exclusive and whether Second-Hand Dwelling house /Apartment have NO VAT exclusive.
  - It is tested that 48 instances with New Dwelling house/Apartment feature don't have VAT exclusive.
- Test 3: check whether 'DateofSale' is in range of 2010-1-1 to present.

    - It is tested that 7 instances having the DateofSale which is in the future.
- Test 4: check if all the deals that happened in Dublin County has PostalCode. Usually speaking, the properties in Dublin County should have postal code, but there might be exceptions. Because there are some areas of Dublin County who don't have the postal code, such as Swords.
    - It is tested that 1309 instances belongs to Dublin County who don't have postal code.
    - Considering that there is a huge amount of instances failed this test, I developed the Test 8, which evaluate PostalCode column in a better way.
- Test 5: check if the information in 'PostalCode' and 'County' match with each other. This could check if the postal code is in Dublin, somehow the County is in another city, which is logically impossible.
    - It is tested that 4 instances failed this test.
- Test 8: check if the information of postal code is hidden in the address. For some instances, the PostalCode column is null, while the real postal code information is written in the Address column.
    - It is tested that 159 rows of instances failed this test.

4. Review Continuous Features

4.1 Descriptive Statistics

'Price' is the only continuous feature in the dataset, with the average 268,348.67. The majority of value fall in the range (120,000 – 310,000). There are several outliers with much higher prices than average, which need to be investigated further.

The feature 'Price' has the range of 99,757,009.37, which represents a huge gap between maximum price and minimum price of the sold property. The value '150,000.0' is the value of highest frequency.

4.2 Histograms

The histogram can be found on the appendix as summary sheet. Overall, the feature showed plausible distribution.

4.3 Box plots

The boxplot can be found on the appendix as summary sheet. The outliers will be investigated further and some of them might need to be dropped for further analysis.

5. Review Categorical Features

5.1Descriptive Statistics

In this part, all the categorical features will be discussed below.

'Address' has an extremely high cardinality of 9985, due to the nature of this feature. Therefore, it is quite rational for this feature to have a high cardinality. On further inspection of the 20 most frequent values of 'Address', it was found that there are 14 addresses that occur twice in the dataset, which means the same properties are sold twice in the period from 2010 to present. In order to make sure that there is no duplicate transaction input, logical integrity Test 6 and Test 7 are carried out to check if there is any property sold on the same date or sold with the same price for more than once (e.g. a property sold twice on the same date, which could be duplicate information input error). No instance fails the tests, which means that those 14 addresses appear twice in the dataset are reasonable and valid transactions.

‘PostalCode’ has 23 values, including 22 values of Dublin postal codes and NaN value. The most frequent value is ‘Dublin 15’ which is plausible. Additionally, ‘PostalCode’ has 80.66% missing values, which is significant enough to immediately drop this feature. However, considering the fact that most counties doesn’t have postal code except Dublin, which explains the reason why there is a high proportion of missing value for this feature. Even though there is a high proportion of missing value in this feature, I would still choose to keep this feature, because this feature can be utilized to evaluate the postal district distribution of properties sold in Dublin.

The categorical feature ‘County’ has 26 values, the most frequent of which is ‘Dublin’, and the less frequent one is ‘Leitrim’. This feature is of proper cardinality and no missing value. Thus, there is no issue with the structure of this feature, and it will be left as it is. However, as we mentioned in the Test 5 in Logical Integrity, there are several instances whose county information doesn’t match with the postal code, which could result from inputting error. Therefore, we might need to correct some of the county names in the ‘County’ column.

The categorical feature ‘NotFullMarketPrice’ and ‘VATExclusive’ both contain 2 values, value of ‘Yes’ and value of ‘No’. There is no obvious issue with the feature ‘NotFullMarketPrice’ as it is expected that the prices for most properties should be full market price. However, this categorical feature couldn’t provide enough helpful information for the analysis, as we have no access to any difference between full market price and non-full market price, and it is hardly to find any logical relationship between ‘NotFullMarketPrice’ and other features.

As to the feature ‘VATExclusive’, there are some instances which failed the logical integrity Test 1, and the solution will be introduced below.

‘DescriptionofProperty’ is another categorical feature that has 2 values, ‘Second-Hand Dwelling house /Apartment’ and ‘New Dwelling house /Apartment’. There is no key issue with the feature itself.

‘PropertySizeDescription’ has 4 values, including ‘greater than 125 sq metres’, ‘greater than or equal to 125 sq metres’, ‘greater than or equal to 38 sq metres and less than 125 sq metres’, and ‘less than 38 sq metres’. However, the category ‘greater than 125 sq metres’ and category ‘greater than or equal to 125 sq metres’ overlap. Also, this is not the most helpful representation of the area of the property, as a better way could be inputting the specific area of the property instead of classifying the area into four categories roughly. Additionally, there are 89.45% missing values in this feature, which makes this categorical feature can only provide limited information.

5.2Bar plots

Refer to the accompanying pdfs to see all bar plots.

6. Actions to take

There are 7 actions to take, which is listed below:

- Duplicate row:

It is detected that there is one row which is duplicate with another one. Therefore, I dropped the duplicate one in the database.

- Instances failed in Test 1:

There are 48 instances failed in the Test 1. And in order to address this issue, I replace the value ‘No’ by ‘Yes’ considering they are all new dwelling. Additionally, I presume that they are manual inputting errors, so I didn’t implement any update on the price for these 48 instances.

- Instances failed in Test 3:

There are 7 instances whose date of sales are in the future, which is logically impossible. Based on my observation, those 7 instances have one common feature, which is that the days are all listed as 01 and the months are all less than 12. Therefore, I suspect that the data was switched the months with the days erroneously. To fix this issue, I simply swapped the values of day and month.

- Instance failed in Test 5:

There are 4 instances whose postal code information don’t match with the county information. After investigating the detailed addresses of these 4 instances, I confirmed that their postal codes are all incorrect, as they are not located in Dublin at all, and the detailed addresses correspond to their counies. Thus, I dropped their postal code.

- Missing values in ‘PostalCode’:

There are 8066 missing values in the feature of ‘PostalCode’. Regardless of the properties that are not in County Dublin, there are still 1309 instances don’t have postal code. In order to fix this issue, I filtered out those instances whose postal code information are written in the column of ‘Address’, and refill them into ‘PostalCode’ column. For those properties located outside the County Dublin, instead of leaving them empty, I fill in the value ‘N/A’ to indicate that those counties don’t have postal code.

- Overlap category in ‘PropertySizeDescription’

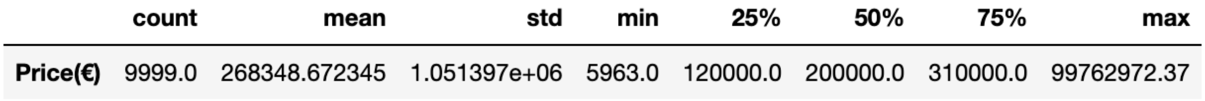
As I mentioned above, the category ‘greater than 125 sq metres’ and category ‘greater than or equal to 125 sq metres’ overlap. Here I will simply combine these two categories (replace ‘greater than 125 sq metres’ by ‘greater than or equal to 125 sq metres’).

- Outlier check

Review all the outliers and check for validity.

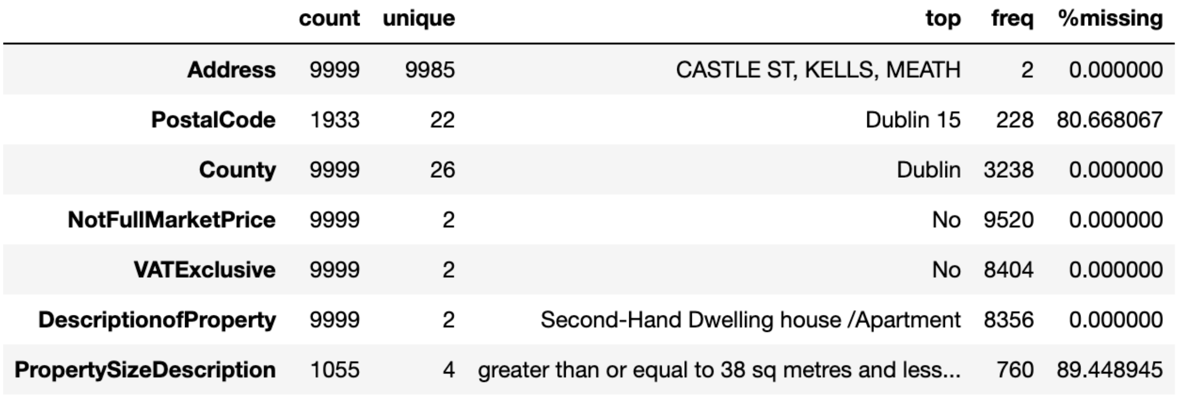
7. Appendix

7.1 Table of descriptive statistics for continuous feature



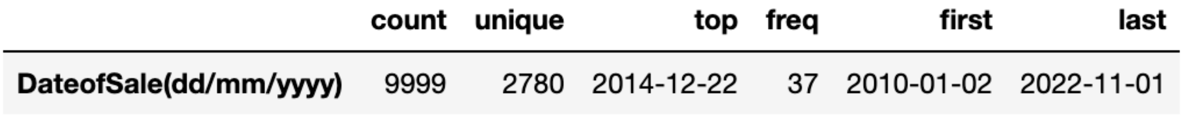
	count	mean	std	min	25%	50%	75%	max
Price(€)	9999.0	268348.672345	1.051397e+06	5963.0	120000.0	200000.0	310000.0	99762972.37

7.2 Table of descriptive statistics for categorical features



	count	unique	top	freq	%missing
Address	9999	9985	CASTLE ST, KELLS, MEATH	2	0.000000
PostalCode	1933	22	Dublin 15	228	80.668067
County	9999	26	Dublin	3238	0.000000
NotFullMarketPrice	9999	2	No	9520	0.000000
VATExclusive	9999	2	No	8404	0.000000
DescriptionofProperty	9999	2	Second-Hand Dwelling house /Apartment	8356	0.000000
PropertySizeDescription	1055	4	greater than or equal to 38 sq metres and less...	760	89.448945

7.3 Table of descriptive statistics for datetime feature



	count	unique	top	freq	first	last
DateofSale(dd/mm/yyyy)	9999	2780	2014-12-22	37	2010-01-02	2022-11-01

7.4 Bar plots & box plots

Check the following accompanying pdfs for plots:

- continuous_histograms_1-1.pdf
- continuous_boxplots_1-1.pdf
- categorical_lowcardinality_barcharts.pdf
- categorical_high_cardinality_Address.pdf
- continuous_boxplots_1-2.pdf