Data Quality Report

Conor Kennedy

16722649

1. Overview

This report will outline the initial findings based on the file 'ppr-16722649.csv'. The document includes information on property purchases in Ireland since 1st January 2010, as declared to the Revenue Commissioners for stamp duty purposes. The data was retrieved from the Residential Property Price Register (RPPR) and is available on their website (https://www.propertypriceregister.ie/). This report gives an overview of the data, delineates the various data quality issues and how they will be addressed to provide higher quality analysis.

The Appendices includes feature summaries, histograms and boxplots.

2. High-Level Summary

- The dataset does not contain a primary key.
- There are 10,000 rows in the csv file.
- There are 9 column headings in the original data set:
    - Date of Sale (dd/mm/yyyy)
    - Address
    - Postal Code
    - County
    - Price (€)
    - Not Full Market Price
    - VAT Exclusive
    - Description of Property
    - Property Size Description
- There are no duplicate rows or columns in the dataset.
- The dataset does not include any constant columns.
- There are a large portion of null values in the columns "Postal Code" and "Property Size Description".

Logical Integrity

A number of tests were carried out to check the logical integrity of the data. While most of the tests passed, there was still significant number of failures. Below is an overview of all the tests carried out:

- Test 1: Checking to see the county names in the dataset
    - Passed – All 26 counties are represented and there were no placenames from outside Ireland.
- Test 2 – Checking if properties with prices reflecting their full market price only are in the dataset.
    - Failed – there are 425 instances were a property's price is not full market price. This would cause inconsistent results that relate to the 'Price' column as it could be significantly higher or lower for each of these 425 properties. This was noted for the Data Quality Plan.
- Test 3 – Examine the earliest date entered.
    - Passed – all data is from after 01/01/2010. No issues here.
- Test 4 – Inspect the latest date.
    - Failed – the latest date is 01/07/2022. This date is in the future and hence does not make sense. Noted for the data quality report.
- Test 5 – Check the cheapest property price
    - Passed after inspection – the lowest price is €5,179 which is very cheap for a property. Further investigation into this property and other low priced properties with online resources assured there was no issues here.
- Test 6 – Checking the most expensive property
    - Passed – the property price is very expensive for Ireland and there was no error in inputting the price (e.g. entering an extra '0'). It is clear that it is an extreme outlier – over €71,000,000 between this property and the second highest property. This was noted for the data quality plan.
- Test 7 – Checking if any property described as "New Dwelling house /Apartment" has been sold twice.
    - Passed – no issues here. There were properties that were purchased more than once over the time frame but none of these were described as "New Dwelling house /Apartment" twice.

- Test 8 – Checking that all new dwellings have VAT included.
    - Failed - According to Residential Property Price Register, "If the property is a new property, the price shown should be exclusive of VAT at 13.5%.". This was not the case on 29 occasions. This has been noted for the Data Quality Report.
- Test 9 – Checking if any second hand property has included VAT.
    - Passed – no issues here.
- Test 10 – Checking if Postal Codes have only been entered for Dublin.
    - Failed – the Postal Codes have been entered for counties other than Dublin. This will be managed with in the data quality plan.

Review Continuous Features:

Please see Appendix A for an overview.

There is only one column that has been labelled as continuous which is "Price" as it can take any value.

- Null entries: None
- Max value: €79,306,927
- Min value: €5,179

Histogram Analysis: The histogram in Appendix D is skewed to the right with the mode price between €150,000 and €200,000. The majority of property that was sold in Ireland between 2010 and 2021 is around the €50,000 - €300,000 price range. Important to note that the bin to the very right in the histogram includes all values that are greater than €900,000. The histogram has been completed in this way so that it gives better insight into the property purchases. The range between the minimum price value and maximum was too large otherwise to view the data in a meaningful way.

Boxplot Analysis: The notebook includes two boxplots. The first includes the outliers and thus provides little insight due to the significant difference between the price values. The second boxplot (included in this document – Appendix E) removes any outliers by capping the price at 1.5 x the interquartile range. This gives a better understanding for the bulk of the

data. From the boxplot, the median is at the €200,000 mark with the interquartile range being €200,000 - €320,000.

Review Categorical Features

Please see Appendix B for an overview.

- Address
    - Null: None
    - Unique Values: 9985 (15 have been sold twice)
    - The data inputted into this column does not follow any structure. For example, in some rows the county name has been inputted and in others there is only the town name. It does offer useful functionality, for example, for retrieving latitude and longitude coordinates and extracting Postal Codes.

- Postal Code
    - Null: 8191
    - Unique Values: 22

All 22 Dublin postal codes are represented in the data frame. There is an issue here with the null values (81.91% missing). Upon further investigation, it is clear that the post codes are only relatable to the county Dublin. Therefore, the percentage of postal codes that were entered for properties in Dublin was 57.56%. Furthermore, there are a number of instances where the postal code has not been inputted into the "Postal Code" column but it is in the "Address" field. This should be extracted and placed into the correct column. Hence, there is enough data here for investigating property prices in Dublin.

- County
    - Null: None
    - Unique Values: 26 (all counties are represented).
    - Top frequency: Dublin

All 26 counties are represented in the data. Unsurprisingly, Dublin has the highest amount of properties sold compared to the other counties.

- Not Full Market Price
    - Null: None
    - Unique values: 2

It is important to note that this column is a double negative. An entry of "no" means that the property price is full market price. 9,575 properties include the full market price. This means that a significant number (425) properties do not display their full price which causes inconsistencies in the data. This could be down to a relative selling property to a relative at a very cheap price to avoid paying tax. These properties will be addressed.

- Vat Exclusive
    - Null: None
    - Unique Values: 2

New dwellings/apartments should include 13.5% tax payment.

- Description of Property
    - Unique: 2
    - Null: None

Each property is described as "New Dwelling house/Apartment" or "Second-Hand Dwelling house /Apartment". 8,408 properties are described as second hand.

- Property Size Description:
    - Null: 9012
    - Unique: 4
- There are four different property descriptions entered:
    - "Greater than or equal to 38 sq metres and less than 125 sq metres"
    - "Greater than 125 sq metres"
    - "Greater than or equal to 125 sq metres"
    - "Less than 38 sq metres"

  There is an issue here with two of the descriptions overlapping. This will be managed with in the Data Quality Report. Furthermore, the description has only been entered for properties that have been described as "New Dwelling house /Apartment". For this type of property, the size has been entered for 62.06% of the properties.

Bar plot Analysis: The bar plots are available in the file "categorical_barplots.pdf".
The following are a list of the key takeaways from each plot:

- Postal Code: Dublin 15 was the area where most property was purchased and Dublin 6W had the least. There is nothing particularly unusual or surprising in the plot.
- County: Dublin, Cork, Galway are the three counties with the highest frequency of houses bought. This could be because each of these are big cities with a large number of opportunities in terms of jobs. On the other hand, Longford, Monaghan and Leitrim all have small towns and thus perhaps less opportunities – people may need to move to bigger cities for opportunities.
- Not Full Market Price: 425 properties purchased were not at full market price.
- Vat Exclusive: The majority of homes sold did not include VAT.
- Description of Property: The majority of houses / apartments purchased were second hand.
- Property Size Description: The most bought property was more than 38 square metres and less than 125 square metres.
- Year: In 2019, the most amount of property was bought. In 2010 and 2011 a noticeable amount less than other years was purchased. While 2022 shows the least amount, this is because it only includes a quarter of the years sales.
- Month: The latter months in a year (October, November and December) are the most popular for buying houses with the opposite true for the early months.

Review DateTime Features

Please see Appendix C for an overview.

- Date of Sale (yyyy/mm/dd)
    - Null: None
    - Unique: 2,749

  As the data contains information from 2010 – 2020, it is unsurprising that there is high cardinality for this column. This does not make this column useful for data analytics as it is hard to draw conclusions from it. Hence, it has been separated into a month, quarter and year column.

Bar plot analysis: Due to the high cardinality in this column, the decision was made not to make a bar plot with this data as it would not provide any meaningful analysis.

Actions to take:

- Check for price outliers.
- Check for rows that include more than one property.
- Make change to the categories in “Property Size Description” column.
- Extract the Postal Code from the Address column.
- Remove any postal codes that are not in Dublin.
- Only include data up to and including 2021 - 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 comparison to the other months. Hence, the decision was made to exclude all data from 2022.
- Drop rows that are not full market price
- Change the price of the new dwelling to include a vat of 13.5%
- Change the date column – instead include a month, year and quarter column.

Appendices

Appendix A

	count	mean	std	min	25%	50%	75%	max
Price	10000	€ 262,592	€ 837,897	€ 5,179	€ 120,000	€ 200,000	€ 307,500	€ 79,306,927

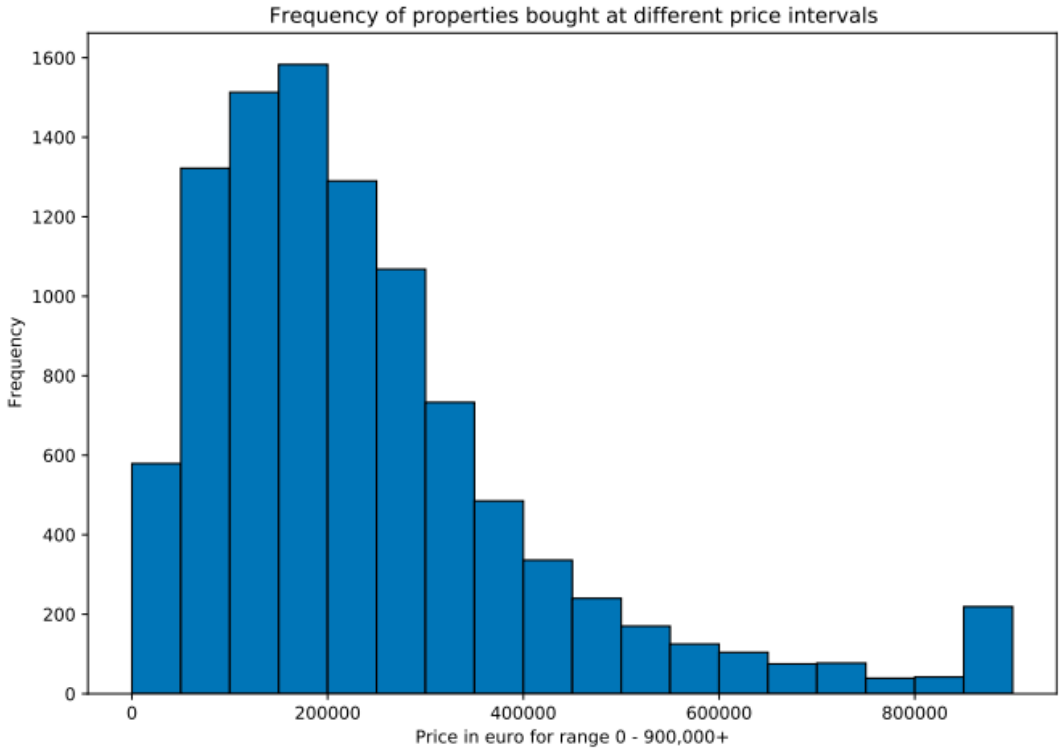
Appendix B

	count	unique	top	freq
Address	10000	9985	APT 1, 20 MARYS STREET, GALWAY	2
PostalCode	1809	22	Dublin 15	232
County	10000	26	Dublin	3143
NotFullMarketPrice	10000	2	No	9575
VATExclusive	10000	2	No	8437
DescriptionofProperty	10000	2	Second-Hand Dwelling house /Apartment	8408
PropertySizeDescription	988	4	greater than or equal to 38 sq metres and less than 125 sq metres	691

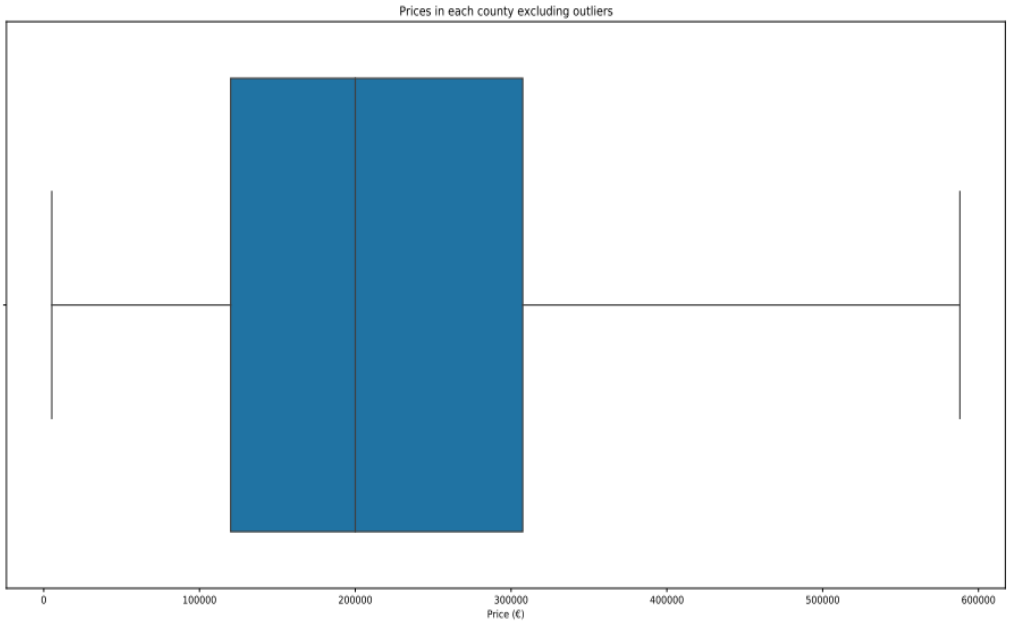
Appendix C

	count	unique	top	freq	first	last
DateofSale	10000	2749	22/12/2014	32	03/01/2010	01/07/2022

Appendix D


Frequency of properties bought at different price intervals
Frequency
1600
1400
1200
1000
800
600
400
200
0
0
200000
400000
600000
800000
Price in euro for range 0 - 900,000+

Appendix E


Prices in each county excluding outliers
0
100000
200000
300000
400000
500000
600000
Price (€)

]