Data Quality Report – Initial Findings

1. Overview

The initial findings from the updated dataset (Residential_Property_Price_Prediction_20211342) will be presented in this report. It will present a summary of the data, as well as a description of the many data quality issues that have been identified and how they will be resolved. For further information on this dataset, please consult the appendix. Terminology, assumptions, explanations, and a summary of changes to the original dataset are included in the appendix. This covers data visualizations such as feature summaries, histograms, and boxplots.
From this updated dataset we can see that there are no duplicate rows and columns. However, when we look for constant columns, all categorial data has unique values is greater than 1 and there are no constant columns, object data (Address) has 9990 unique values and there are no constant columns. In float64 data does not have constant columns and all data standard deviations are non-zero. Thus, all rows do not contain a single constant value and in this case, none of the continuous features are constant. Otherwise, in the logical integrity of this data, I can only find one inconsistency.

2. Summary

From the test for this dataset, we can see there are 40 houses or apartments don not have VAT, because they are new houses or apartments. For the continuous feature we can see

3. Review Logical Integrity

3.1 Test_1 check if VAT Exclusive is not equal to Description of Property and show all differences.
The number of rows failing the test are 40.

3.2 Test_2 checks for null values. Then we can find that PostalCode has a lot of null values. Then we check for the not null values for PostalCode, there are 1918 not null values. When we see the Address and PostalCode, we can find that some addresses contain DUBLIN and some don't. After checking the logical integrity, we can see that there are 1751 houses or apartments with addresses include DUBLIN.
The number of rows failing the test are 1751.

3.3 Test_3 checks DateofSale(dd/mm/yyyy). As can be seen from the rules, starting from 2010-01-01, the date range I set is between 2010-01-01 and 2022-03-07. After checking the logical integrity, we can see there are 10 houses or apartments are over the date range.
The number of rows failing the test are 10.

4. Review Continuous Feature

There is 1 continuous feature in this CSV. As we can see the mean value is 267581.872415. The standard deviation of the observation is 780575.142298. The minimum value is 5252.0. The maximum value is 60000000.0.

Histogram for the continuous feature:
The histogram for the continuous can be found on the appendix. There is only one feature for the price. This feature showed plausible distributions. The outliers will be investigated further, but no action is expected right now.

Box plot for the continuous feature:
The box plot for the continuous can be found on the appendix. Again, the outliers will be investigated further, but no action is expected right now.

5. Review Categorical Features

There are 6 categorical features in this CSV, four of which are County, Not Full Market Price, VAT Exclusive and Description of Property won't be evaluated here. The two remaining are "Postal Code" and "Property Size Description".

Both features contain a lot of null values, which should not exist. This is because the Postal Code and Property Size Description should be provided, not just partially displayed.

Bar plots for the categorical features:
The bar plots can be found on the appendix.

6. Review Objective Feature

There is 1 objective feature in this CSV. There is only one feature which is "Address". Because addresses are essentially unique, in-depth reviews will not be conducted here.

7. Action to take

Logical integrity:
Drop any rows that fail the logical test.

Categorical Features:
Fill in the data that missed.

Outliers:
Examine outliers to see if they are valid.

8. Appendix

Descriptive Statistics:

Continuous Feature

	count	mean	std	min	25%	50%	75%	max
Price(€)	10000.0	267581.872415	780575.142298	5252.0	120000.0	200000.0	308370.04	60000000.0

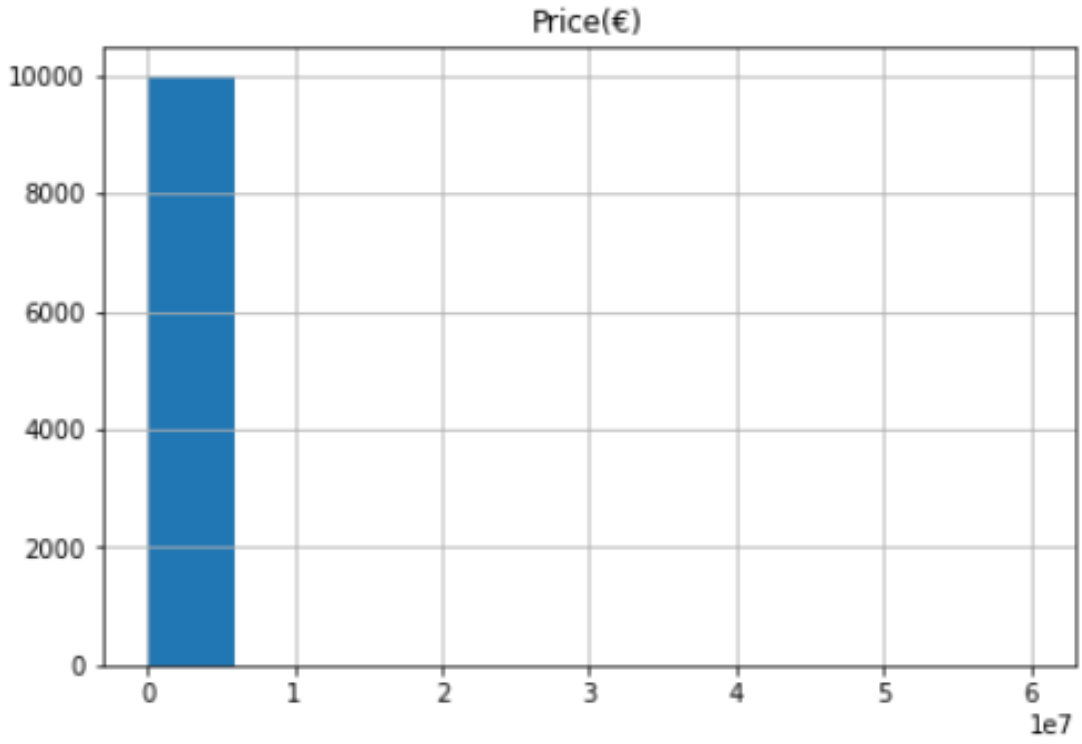
Categorial Features

	count	unique	top	freq
PostalCode	1918	22	Dublin 15	247
County	10000	26	Dublin	3275
NotFullMarketPrice	10000	2	No	9480
VATExclusive	10000	2	No	8440
DescriptionofProperty	10000	2	Second-Hand Dwelling house /Apartment	8400
PropertySizeDescription	1022	4	greater than or equal to 38 sq metres and less...	719

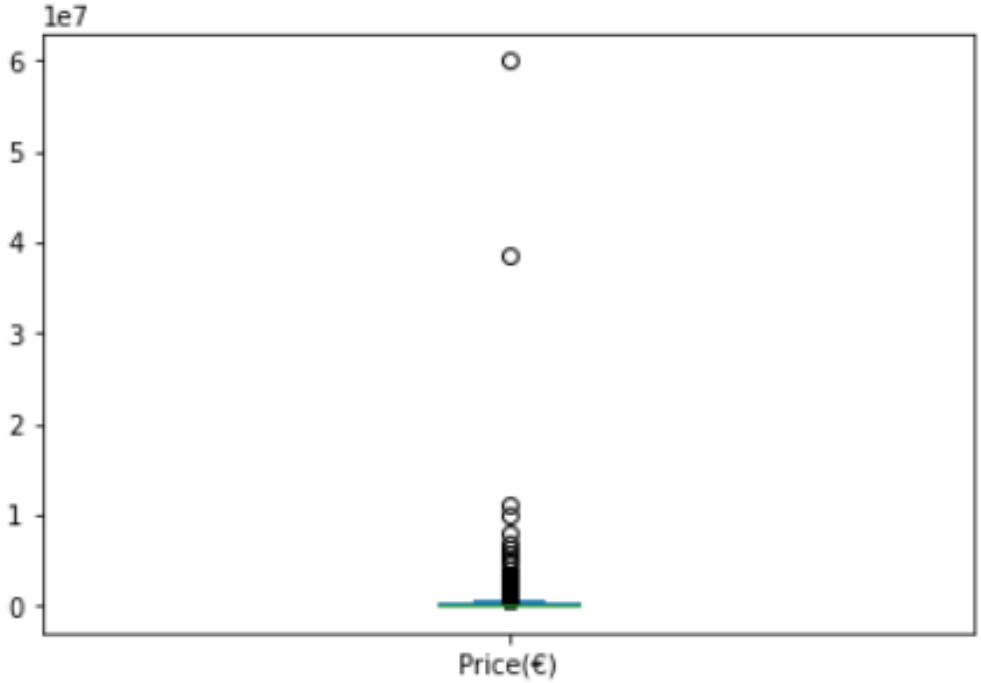
Objective Features

	count	unique	top	freq
Address	10000	9990	CAPPINCUR, TULLAMORE, OFFALY	2

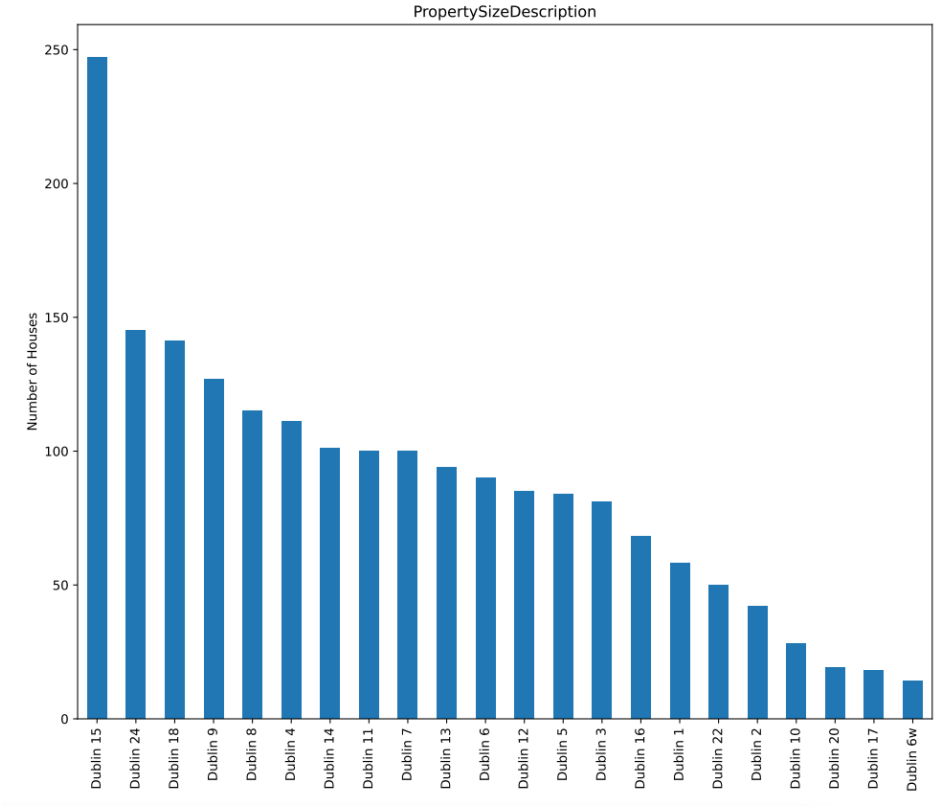
Histograms for Continuous Feature

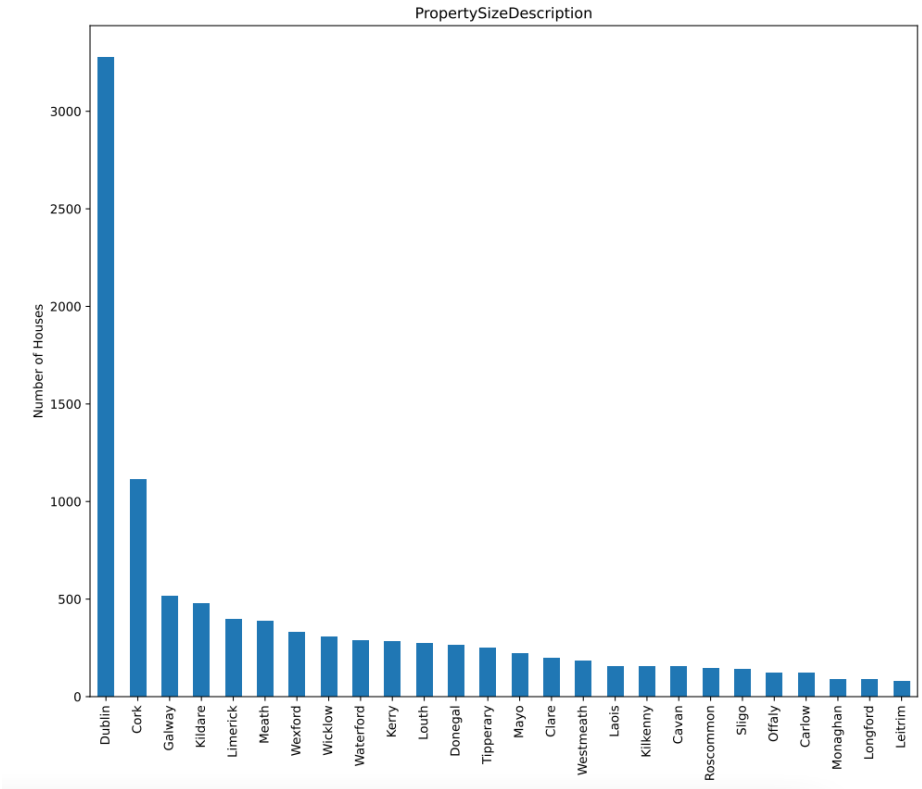


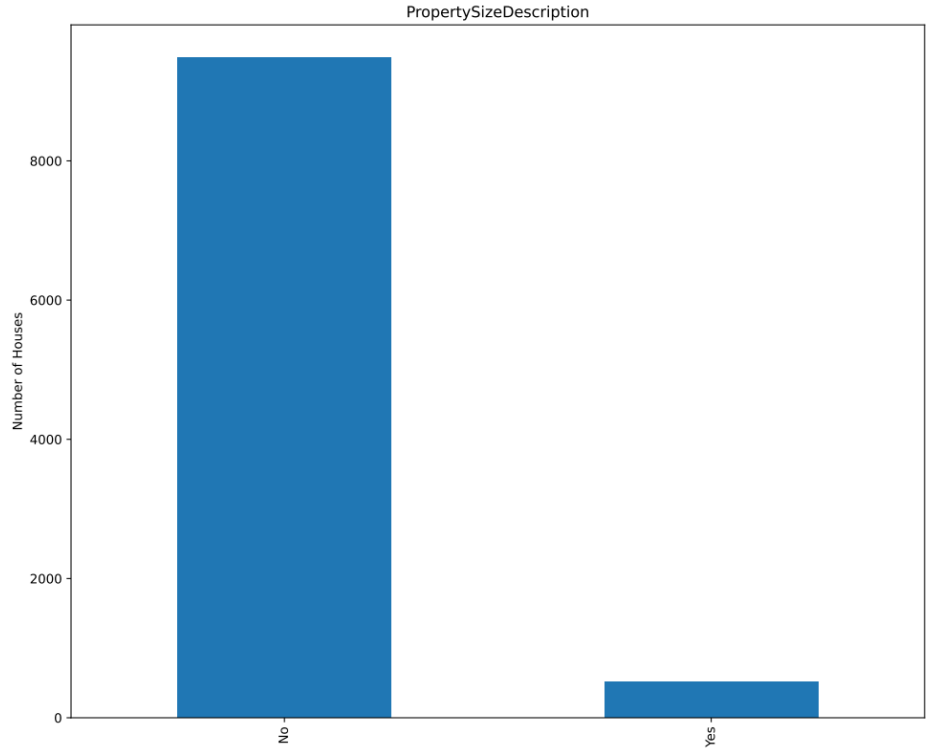
Boxplot for Continuous Feature

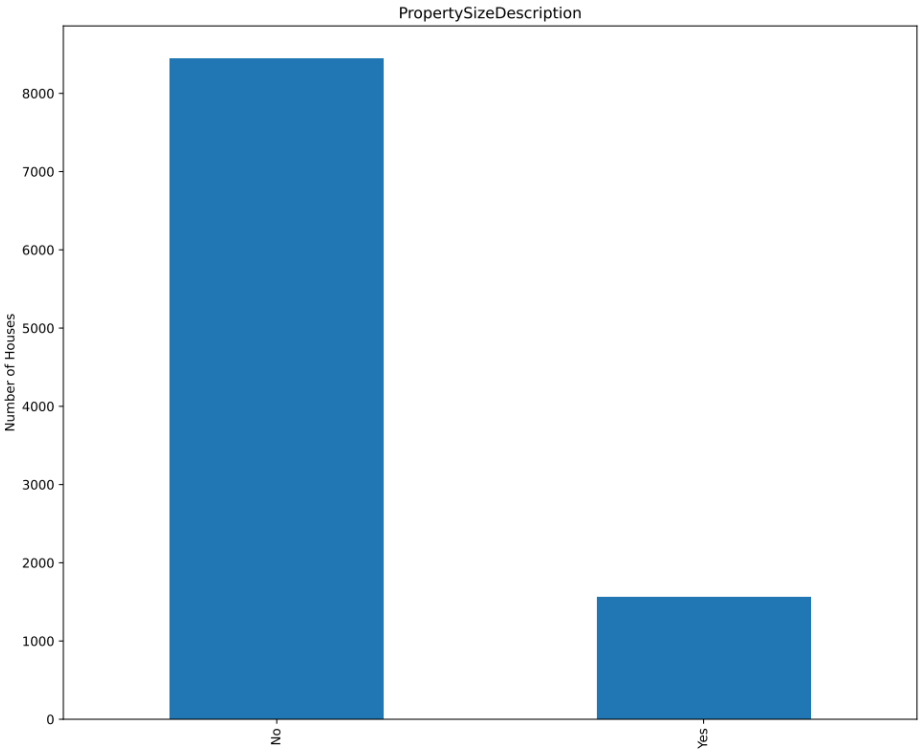


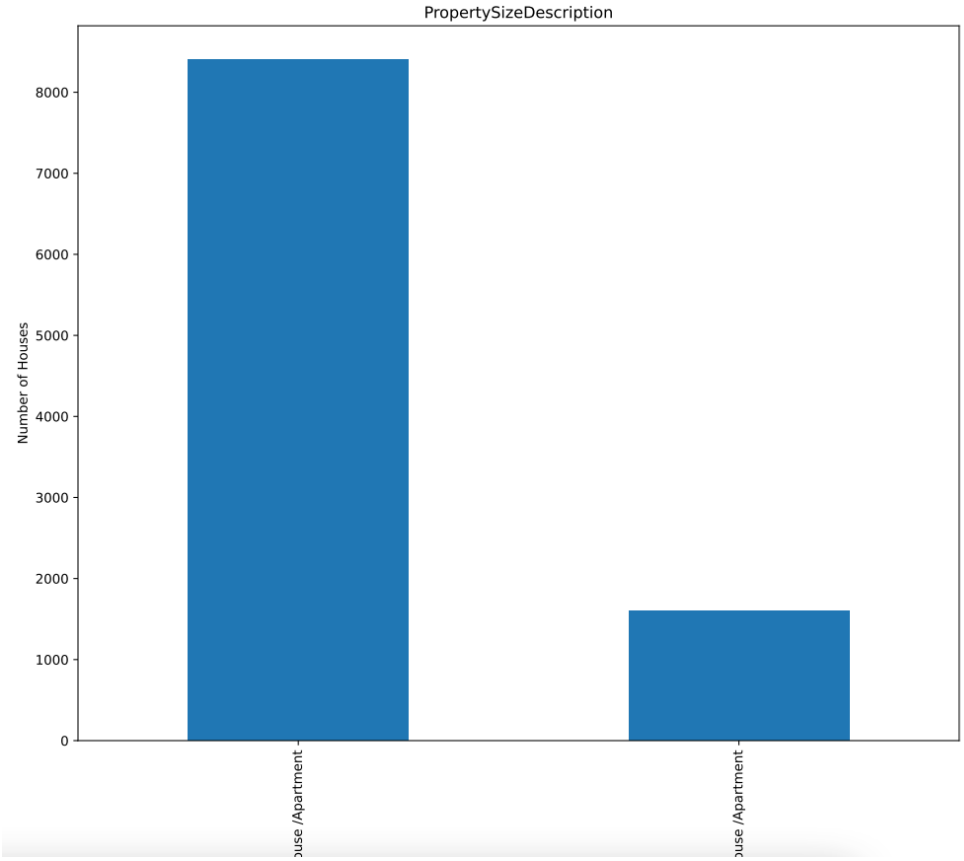
Bar plots for Categorial Features

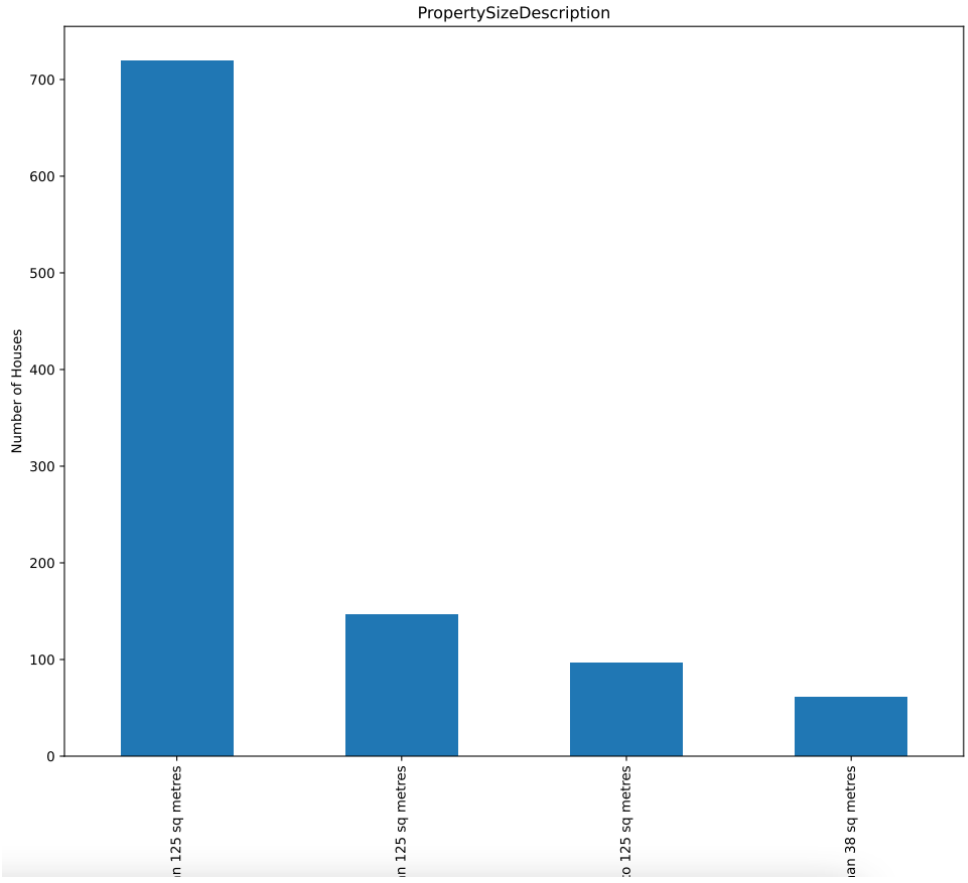





PropertySizeDescription
Number of Houses
0
2000
4000
6000
8000
No
Yes


PropertySizeDescription
Number of Houses
0
1000
2000
3000
4000
5000
6000
7000
8000
No
Yes


PropertySizeDescription
Number of Houses
8000
7000
6000
5000
4000
3000
2000
1000
0
ouse /Apartment
ouse /Apartment


PropertySizeDescription
Number of Houses
700
600
500
400
300
200
100
0
n 125 sq metres
n 125 sq metres
o 125 sq metres
an 38 sq metres