DATA ANALYTICS – HOMEWORK 1

DATA QUALITY REPORT

1 OVERVIEW

1.1 This data quality report (the **Report**) consists of an initial summary of the review carried out on the cleaned dataset located in this directory and entitled “q1_cleaned_dataframe.csv” (the **Dataset**). The Report contains a summary of the Dataset, the various issues encountered with this Dataset, and some proposed solutions for resolving those issues. The appendices to the Report (the **Appendices**) contains the descriptive statistics tables, bar charts, histograms and box plots referred to in the Report.

1.2 The initial review suggests that the Dataset is somewhat clean. While there are no irregular cardinalities or duplicate columns, there were two features with a large proportion of null values. There was also a duplicate row discovered, which has now been removed. It should be noted that several logic tests were carried out in respect of the Dataset, and no inconsistencies were found. There were also a significant number of outliers discovered in connection with one of the features.

2 LOGICAL TESTS

2.1 Several logical tests were conducted in connection with this dataset. There were only two tests that failed, as listed below.

2.2 Test 3: The postal code must be a valid post code from differing counties in Ireland

While all post codes provided were valid, all the postcodes in our dataframe were in Ireland. As such, there is no postal code data provided for any of the other county, which is what we would have expected given the title of the feature.

2.3 Test 6: There must be one property sale recorded per entry in the dataset

There were numerous entries in the dataset that represented a multiple property purchase, leading to significant outliers.

2.4 Strategies for dealing with these test failures will be discussed in subsequent sections of this report.

3 DESCRIPTIVE STATISTICS TABLES

3.1 While some observations have already made in relation to the descriptive statistics tables within the notebook, it is useful to summarise these here for reference. These are set out in Appendix 1 to this Report.

3.2 Continuous Features

While the Date of Sale feature seems largely in order, we can see an incredibly large number of outliers in relation to the Price feature. The minimum and maximum values do not seem to be reasonable values for a property purchase, and the standard deviation is incredibly large. This is something that we will consider later in the Report.

3.3 Categorical Features

The categorical features seem largely in order, except for some irregularities in connection with some of the features. For example, we can see that Postal Code and Property Size Description have a huge number of missing values. We can also tell that some of the features, such as Not Full Market Price and Vat Exclusive, seem to be Boolean in nature. Furthermore, one of the Boolean values seems to dominate over the other, which we can tell by looking at the proportion of the first model. It is also unsurprising to see that the most property sales occurred in Dublin and Cork, given that they are largest and second-largest cities in Ireland respectively.

4 REVIEW CONTINUOUS FEATURES

Please note that the analysis of each feature listed below is based on a review of the statistic tables and the plotted graphs contained in the Appendices. There is a separate section in each paragraph containing analysis that is specific to the graphs.

4.1 Descriptive Statistics

FEATURE NAME	FEATURE DESCRIPTION	OBSERVATIONS
Date of Sale	This feature contains the date of sale of a particular property.	This feature has no missing entries. The dates range from 4 January 2010 to 14 January 2022, which is the time range expected given the source of the Dataset. The distribution shown in the histogram seems to be slightly skewed towards the latter half of the decade. This is in line with expectation, given that the at the start of the decade the Irish property market will still have been suffering from the effects of the 2008 economic crisis and thus less property sales will have taken place at that time.
Price (€)	This feature contains the price for which a particular property was sold.	This feature has no missing entries. The main issue with this feature is the significant number of outliers. Prices range from as low as €5,953 to €139,165,000, which are both implausible as property prices. Upon further investigation, a cause for the high outliers was discovered – it appears that certain group sales of properties were mistakenly entered as a single “entry” in the Dataset, causing an inflated price to be included. While it is plausible that the sale of such properties took place as part of a single transaction – it is common in Ireland for investors to purchase a large portfolio of properties for renting out – it does not fit with the logical integrity of this feature, which states that each row must represent the sale of a single property. It is not clear what the cause of the low outliers is, although it does seem like an entry made in error. A discussion should be had with the domain expert in this regard.

4.2 Histograms

The histograms are in Appendix 2 to the Report. There does not seem to be anything unusual about the Date of Sale histogram, as noted in the section above. The Price (€) histogram, however, is incredibly skewed given the presence of extreme outliers. For this purpose, another histogram is provided with the minimum and maximum outliers clamped using the process more particularly set out in the notebook. The second histogram shows a distribution that is expected, namely the fact that there is a left-skewed distribution (meaning that most sales made were in the average range, with some large infrequent outliers where very expensive properties were purchased.

4.3 Box Plots

The box plots are in Appendix 2 to the Report. The analysis here is largely the same as the analysis located in para.4.2 (*Histograms*) above – it is useful to note, however, that for the purposes of the clamped box plot, the spread between second and fourth quartiles seems to be between 20k-40k, with some large outliers over 56k. For the date of sale box plot, spread between the second and fourth quartiles seems to be between 2015–2020.

4.4 Potential Solutions

ISSUE	POTENTIAL SOLUTIONS	ADVANTAGES	DISADVANTAGES
Presence of high outliers	For rows where several properties were entered as a single entry, split those rows into several entries representing each property, under the assumption that each property cost the same amount.	Very little loss of data, represents the true intent of the entry in the context of the table (a record of the sale of several properties).	Assumption that each property cost the same amount may be incorrect. Also, it is difficult to track down which entries represent multiple properties within the Dataset.
	Replace all high outliers above a certain value with the median.	Simple operation that will lead to a more evenly distributed Dataset.	Large loss of data – some of the high outliers could represent actual properties that were sold for a large value, which is information which would be valuable for the model.
	Clamp the high outliers at a certain value.	Simple operation.	Even larger loss of data, cardinality becomes significantly reduced.

ISSUE	POTENTIAL SOLUTIONS	ADVANTAGES	DISADVANTAGES
Presence of low outliers	Replace all low outliers below a certain value with the median.	Simple operation that will lead to a more evenly distributed Dataset.	Large loss of data – some of the low outliers could represent actual properties that were sold for a low value, which is information which would be valuable for the model.
	Clamp the low outliers at a certain value.	Simple operation.	Even larger loss of data, cardinality becomes significantly reduced.

5 REVIEW CATEGORICAL FEATURES

5.1 Descriptive Statistics

FEATURE NAME	FEATURE DESCRIPTION	OBSERVATIONS
Address	The address of the property that was sold.	There do not seem to be any major anomalies with this feature. However, the extremely large cardinality makes it difficult to see how any meaningful analysis could be carried out in relation to it.
Postal Code	The postal code of the property that was sold.	There are many missing values in respect of this feature. Upon further inspection, the postal codes only appear if the property was sold in Dublin. Furthermore, the postal codes being used are not the more recent Eircodes.
County	The county where the property being sold was located.	There do not seem to be any anomalies with this feature.
Not Full Market Price	Indicates whether the property was sold at full value.	It is a little unclear from the RPPR information note what this feature is supposed to represent. The RPPR suggests that where this feature is set to “True”, “the price shown does not represent the full market price of the property concerned for a variety of reasons”, such as the interest in the property being retained by the previous owner, or the house being purchased under the Affordable Homes Scheme. It is unclear whether this means that: (a) the home was purchased for the full price

FEATURE NAME	FEATURE DESCRIPTION	OBSERVATIONS
		and a lower price is shown in the register or (b) the home was actually purchased at a discount. Because it is not possible to discuss with a domain expert, we will have to make an assumption here. For the purposes of this assignment, we will assume that “Not Full Market Price” means that the relevant property was actually purchased at a price lower than its market value, for whatever reason that may be. There do not seem to be any anomalies with this feature. However, the fact that it is represented through a negative (i.e., “Not Full Market Value” as opposed to “Full Market Value”), makes it a little counter-intuitive to work with. Also, given that these are Boolean values, it might be easier to work with these if they were equal to “True” and “False” as opposed to “Yes” and “No”. As mentioned, it seems that most of the houses sold were sold for full market price.
VAT Exclusive	Indicates if the purchase price includes VAT or not.	There do not seem to be any anomalies with this feature. Also, given that these are Boolean values, it might be easier to work with these if they were equal to “True” and “False” as opposed to “Yes” and “No”. Based on the information note in the RPPR, this feature represents the fact that if VAT was payable on the property, then this would be mentioned here (and the Price displayed would not include this VAT). As mentioned, the majority of the houses sold are marked not VAT Exclusive.
Description of Property	Gives a brief description of the property.	This feature has a cardinality of three, and upon further inspection one of the unique values is simply an Irish translation of one of the other unique values. Thus, the two possible values of this column are “New Dwelling House / Apartment” and “Second-Hand Dwelling House Apartment”. One interesting point here is the fact that this feature merely suggests if the property sold was new and second hand. Given that VAT is payable mostly on new Irish properties, it should be considered whether VAT Exclusive and Description of Property are merely representing the same facet of a property sale, namely whether the property sold was second hand.

FEATURE NAME	FEATURE DESCRIPTION	OBSERVATIONS
Property Size Description	Gives a description of the property size.	This feature has a cardinality of four and seems to have split the properties into four separate ranges. However, due to the enormous number of missing values (over 90%) it is difficult to say whether any useful information could be gleaned from this feature.

5.2 Bar Charts

The bar charts are in Appendix 1 to the Report. There does not seem to be anything unusual about any of the bar charts, and in fact the outcomes are largely as expected. Of particular interest is the fact that the most sales took place in the two largest cities in Ireland i.e., Dublin and Cork. It is also interesting to see that Dublin 15 has the most sales out of any postcode in Dublin. Dublin 15 is a suburb, and it is possible that the large amounts of sales in this area suggests that more people are moving to live there.

An attempt was made to plot the top 15 most frequent addresses, but this proved not to be particularly fruitful, given that: (a) the maximum cardinality of any address is equal to two; and (b) those 15 addresses all seem to be spread all over Ireland, meaning that there does not seem to be any major link between frequency of reselling and location.

We can also see that the Not Full Market Price features and VAT Exclusive features seem to heavily lean on one of their Boolean values ("No" for both), which is something that we should consider when plotting correlations between features. Finally, we can see that most houses sold were second-hand, and were larger than 125 square metres.

5.3 Potential Solutions

Note: Some of the potential solutions for issues that were considered seem to have no disadvantages, or where the alternative is to do nothing. For such issues, only one solution is presented, as the fact that doing nothing as an alternative is implied (unless doing nothing has a disadvantage, as with the first row in the table below, in which case this is presented as a valid option).

ISSUE	POTENTIAL SOLUTIONS	ADVANTAGES	DISADVANTAGES
No real use for address feature	Do nothing	No loss of data.	Potentially keeping "dead weight" in the Dataset, as there is nothing really that can be done with this feature.

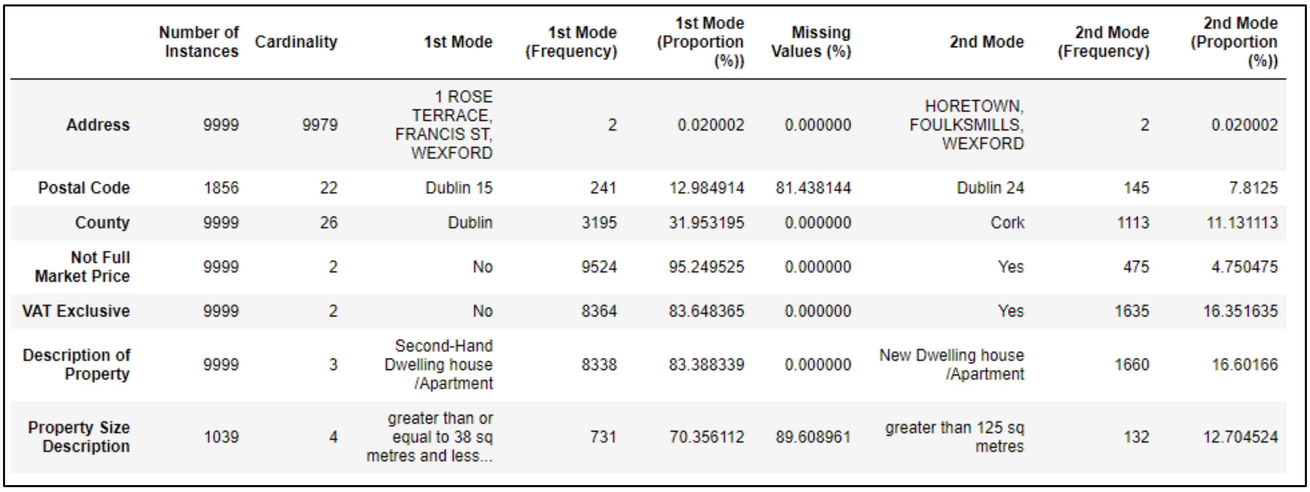
ISSUE	POTENTIAL SOLUTIONS	ADVANTAGES	DISADVANTAGES
	Drop feature	We no longer have a feature in our database that we are not using.	Massive loss of data.
	Use third party libraries to transform address strings into ones that can be more easily worked with	Using a library such as Geopy will allow us to easily work with the address fields, and map them using a library such as Folium	None (potentially takes a long time)
Large number of missing values in postal code	Drop feature	Given the large number of missing values (i.e., 81%) it is difficult to see how this feature could provide us with information about the dataset, and as such may be dropped.	Loss of data.
	Change the title of feature to “Postal Code (Dublin)”	By keeping the postal codes, we may be able to see a correlation between certain areas of Dublin and the property prices in Dublin. The change in feature name will make it a more accurate descriptor of the data contained therein.	Missing values not useful for gleaning any new information about the other features.
Not Full Market Price is framed as a negative and does not contain Boolean values	Transform feature into “Full Market Price”, changing the “Yes” values to “False” and the “No” values to “True”	Data becomes easier to work with.	None.
VAT exclusive does not contain Boolean values	Change the “Yes” values to “True” and the “No” values to “False”	Data becomes easier to work with.	None.

ISSUE	POTENTIAL SOLUTIONS	ADVANTAGES	DISADVANTAGES
VAT Exclusive and Description of Property could potentially be referencing the same facet of the data	Drop Description of Property.	Less redundancy of data, as we will be able to infer from VAT Exclusive whether the property is second hand.	There could be situations where VAT was payable on a second-hand home, and as such, there may be a loss of data if we drop Description of Property.
Description of Property feature is currently difficult to read	Convert into “Second Hand” feature and map the values accordingly.	Data becomes easier to work with.	None.
Property Size Description has large number of missing values	Drop feature	We no longer have a feature that does not provide us with new information about the Dataset.	Loss of data.

APPENDIX 1

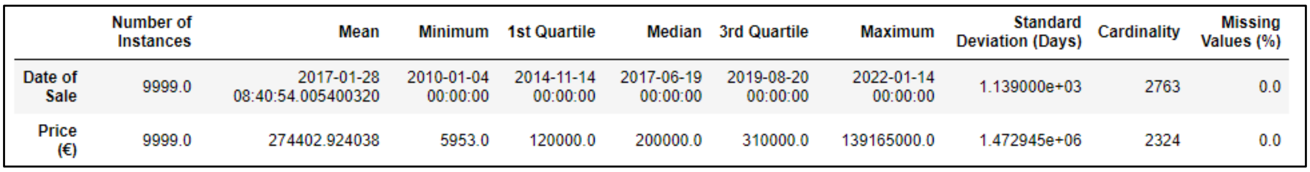
DESCRIPTIVE STATISTICS TABLES

CATEGORICAL FEATURES



	Number of Instances	Cardinality	1st Mode	1st Mode (Frequency)	1st Mode (Proportion (%))	Missing Values (%)	2nd Mode	2nd Mode (Frequency)	2nd Mode (Proportion (%))
Address	9999	9979	1 ROSE TERRACE, FRANCIS ST, WEXFORD	2	0.020002	0.000000	HORETOWN, FOULKSMILLS, WEXFORD	2	0.020002
Postal Code	1856	22	Dublin 15	241	12.984914	81.438144	Dublin 24	145	7.8125
County	9999	26	Dublin	3195	31.953195	0.000000	Cork	1113	11.131113
Not Full Market Price	9999	2	No	9524	95.249525	0.000000	Yes	475	4.750475
VAT Exclusive	9999	2	No	8364	83.648365	0.000000	Yes	1635	16.351635
Description of Property	9999	3	Second-Hand Dwelling house /Apartment	8338	83.388339	0.000000	New Dwelling house /Apartment	1660	16.60166
Property Size Description	1039	4	greater than or equal to 38 sq metres and less...	731	70.356112	89.608961	greater than 125 sq metres	132	12.704524

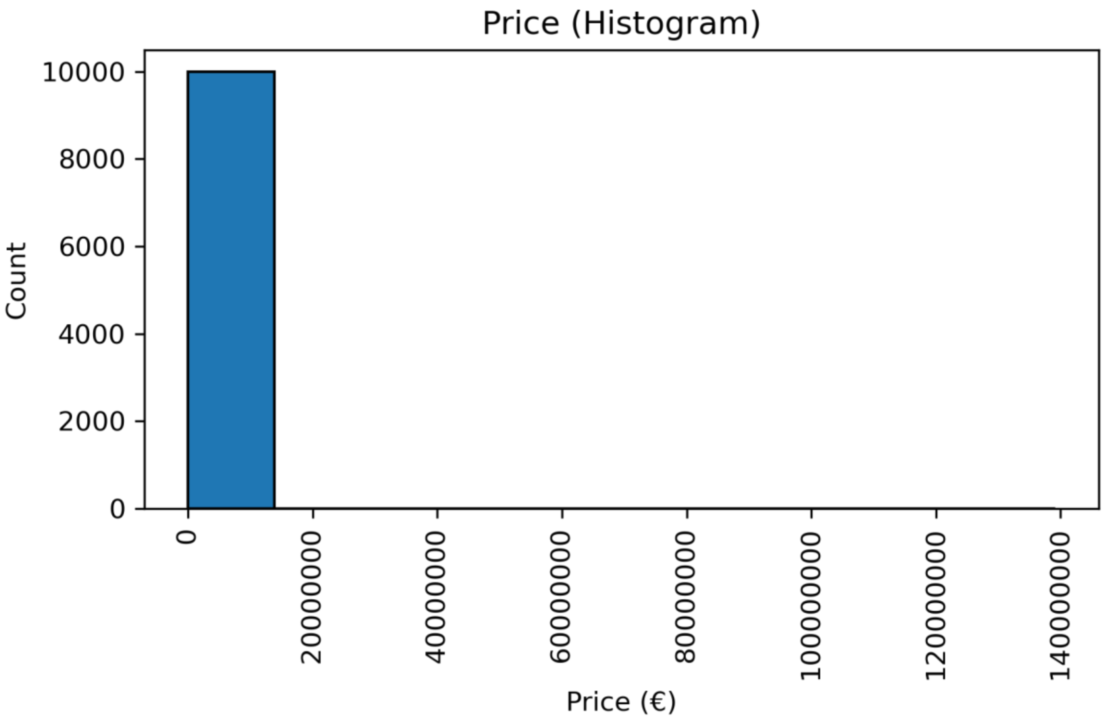
CONTINUOUS FEATURES

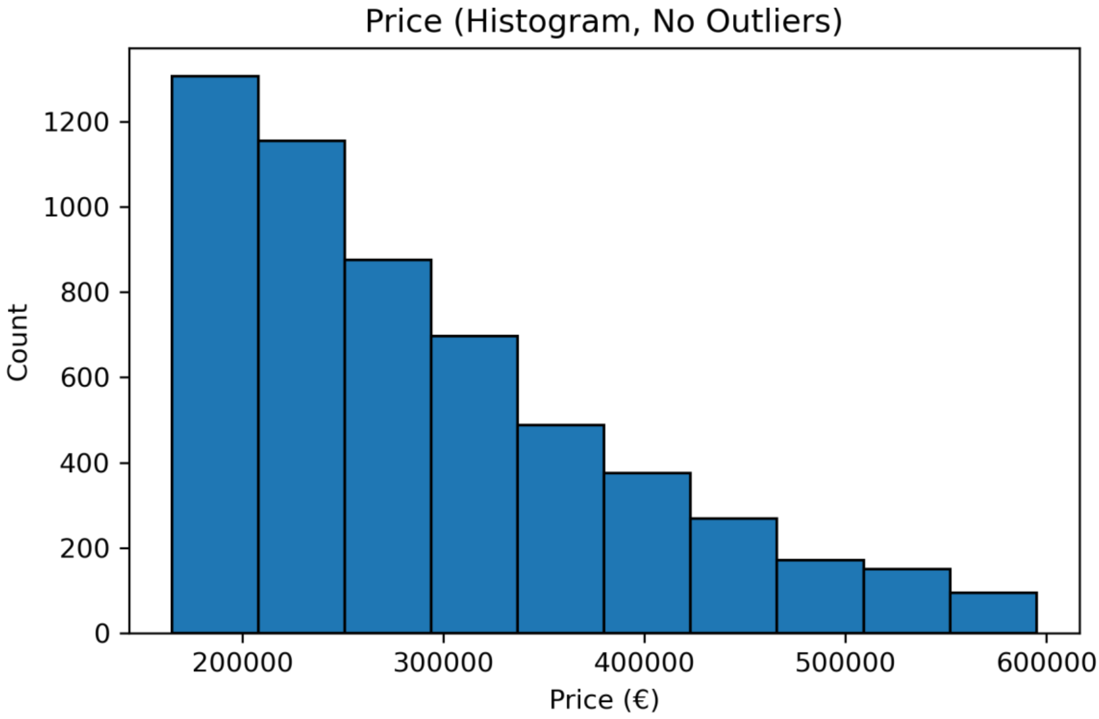


	Number of Instances	Mean	Minimum	1st Quartile	Median	3rd Quartile	Maximum	Standard Deviation (Days)	Cardinality	Missing Values (%)
Date of Sale	9999.0	2017-01-28 08:40:54.005400320	2010-01-04 00:00:00	2014-11-14 00:00:00	2017-06-19 00:00:00	2019-08-20 00:00:00	2022-01-14 00:00:00	1.139000e+03	2763	0.0
Price (€)	9999.0	274402.924038	5953.0	120000.0	200000.0	310000.0	139165000.0	1.472945e+06	2324	0.0

APPENDIX 2

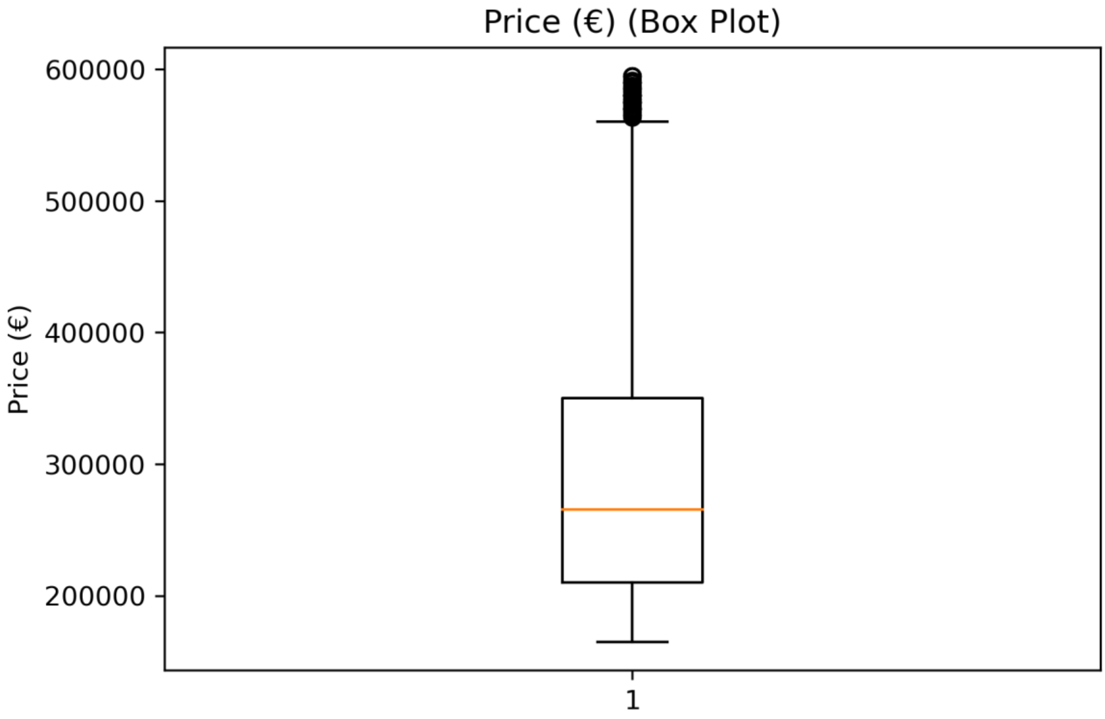
CONTINUOUS FEATURES (HISTOGRAMS)

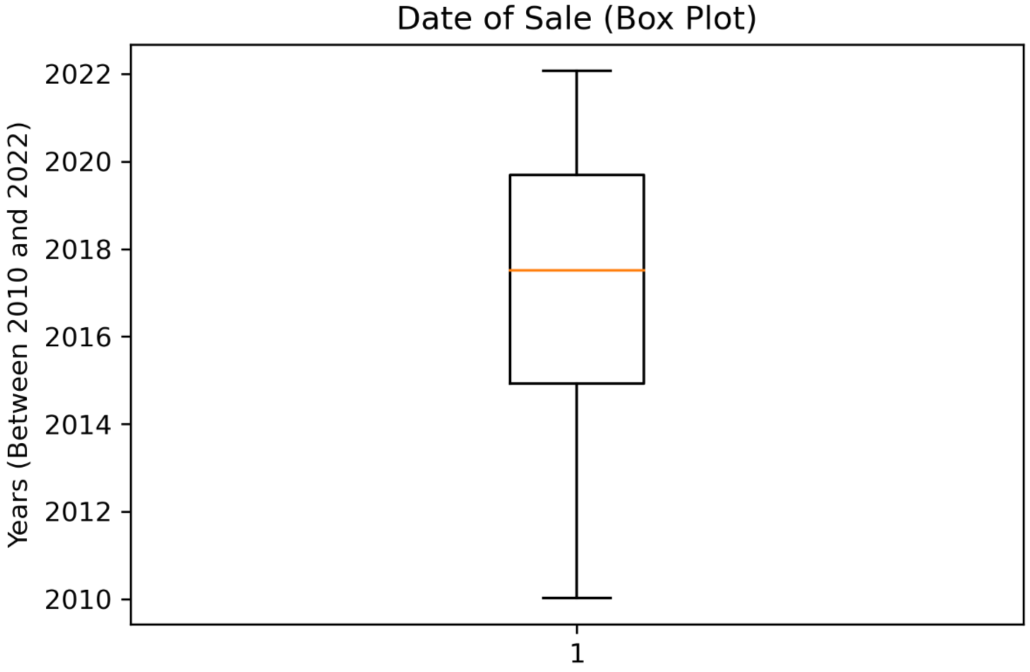




APPENDIX 3

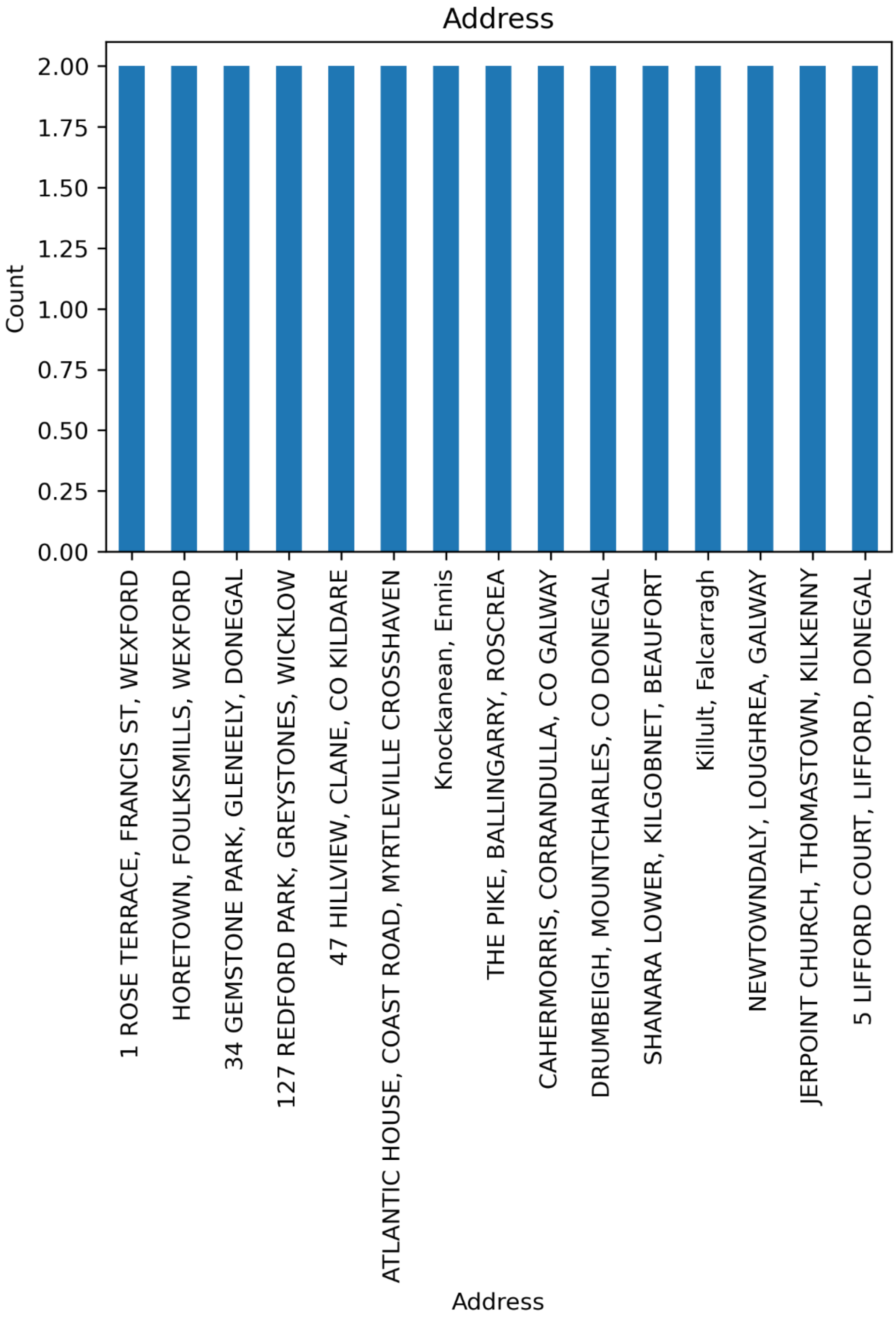
CONTINUOUS FEATURES (BOX PLOTS)

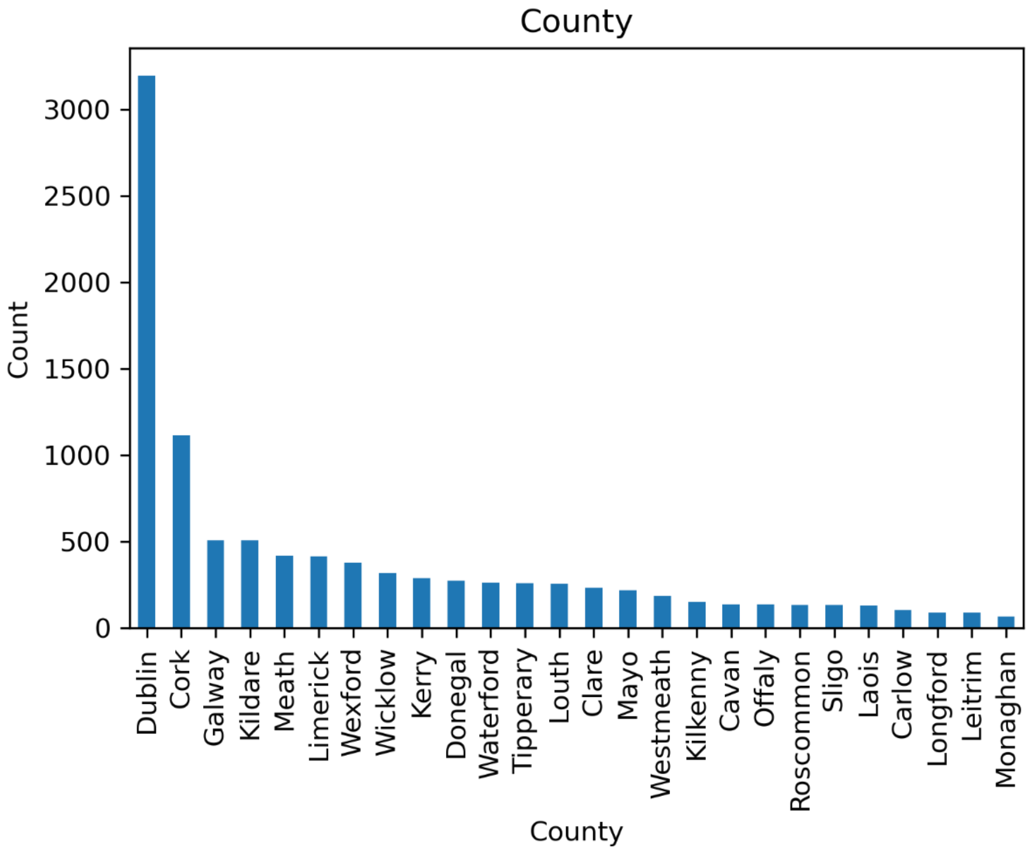


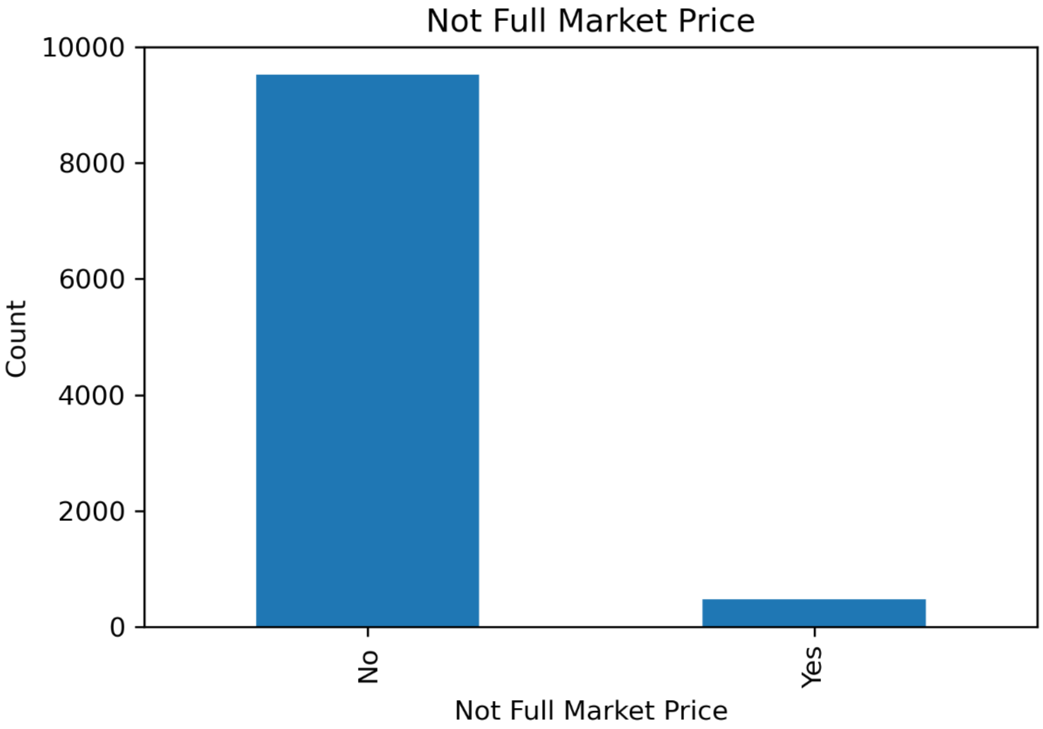


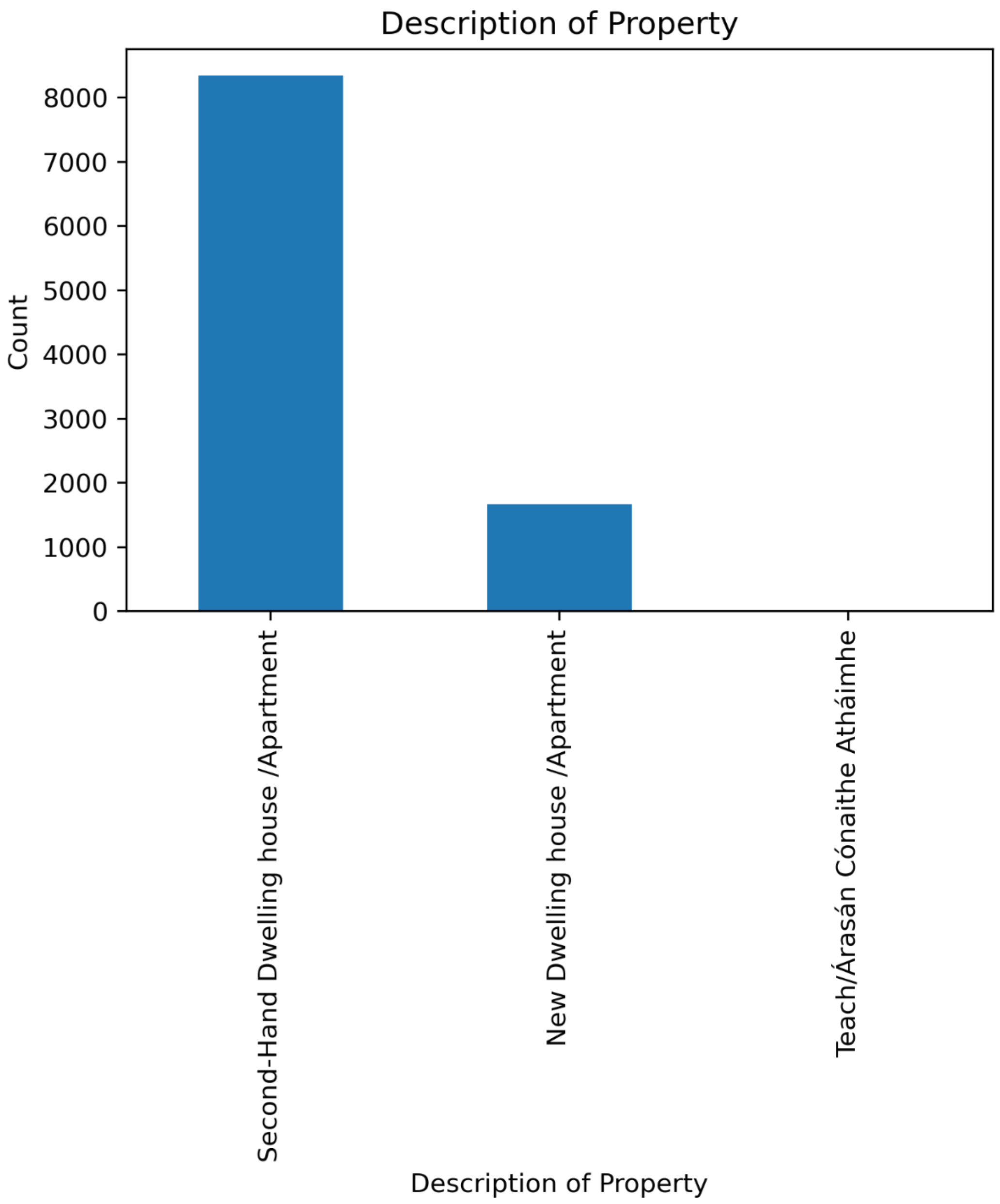
APPENDIX 4

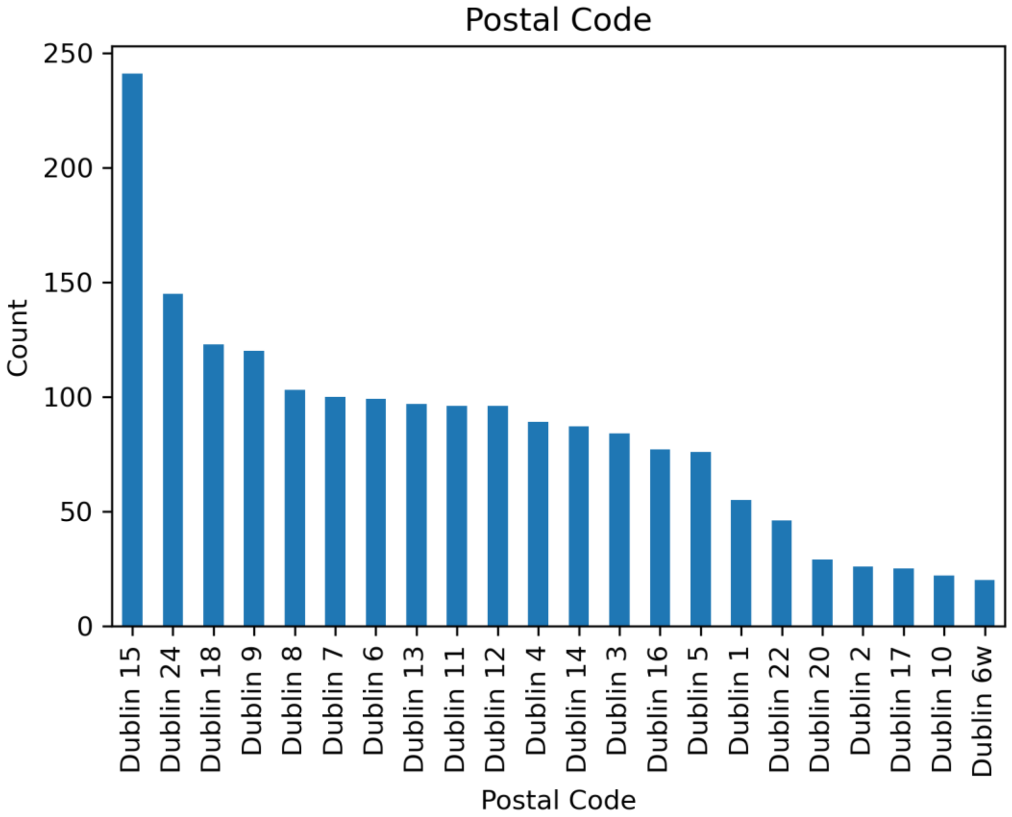
CATEGORICAL FEATURES (BAR PLOTS)

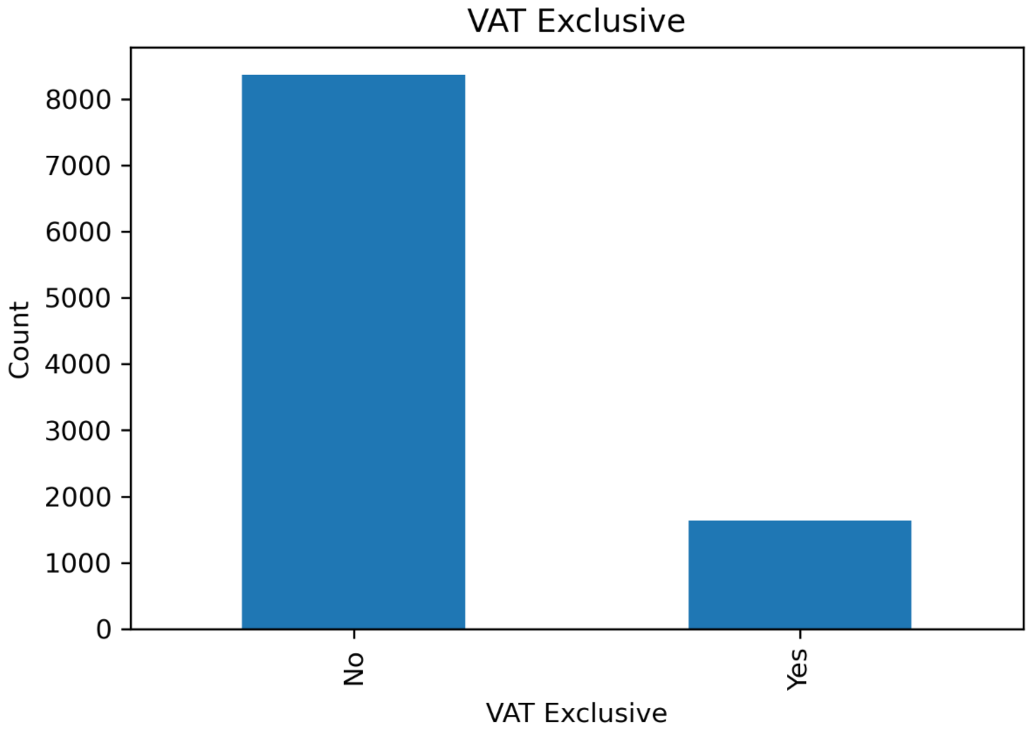


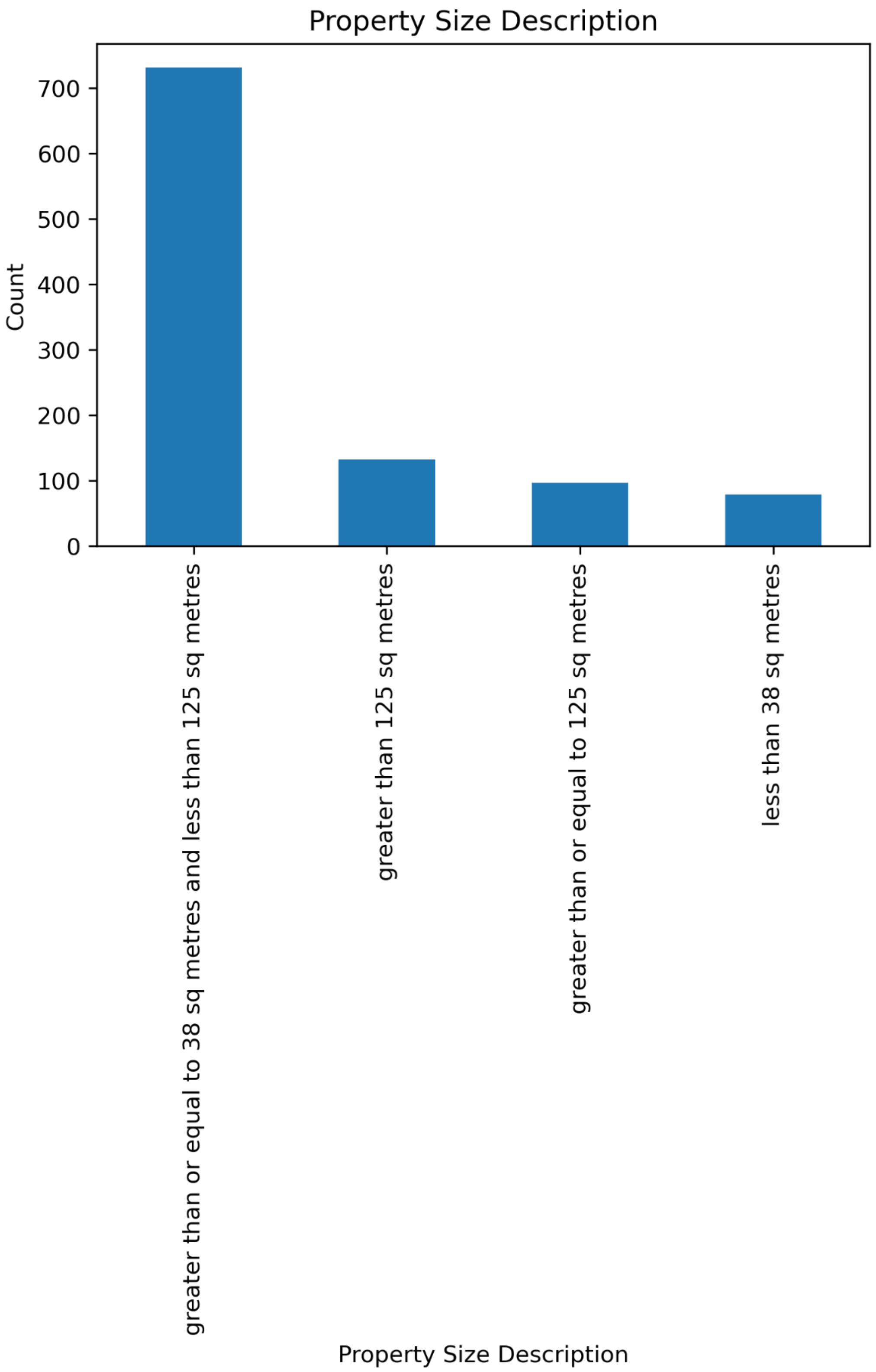

County
Count
3000
2500
2000
1500
1000
500
0
Dublin
Cork
Galway
Kildare
Meath
Limerick
Wexford
Wicklow
Kerry
Donegal
Waterford
Tipperary
Louth
Clare
Mayo
Westmeath
Kilkenny
Cavan
Offaly
Roscommon
Sligo
Laois
Carlow
Longford
Leitrim
Monaghan
County


Not Full Market Price
Count
10000
8000
6000
4000
2000
0
No
Yes
Not Full Market Price


Description of Property
Count
0
1000
2000
3000
4000
5000
6000
7000
8000
Second-Hand Dwelling house /Apartment
New Dwelling house /Apartment
Teach/Árasán Cónaithe Atháimhe
Description of Property


Postal Code
Count
250
200
150
100
50
0
Dublin 15
Dublin 24
Dublin 18
Dublin 9
Dublin 8
Dublin 7
Dublin 6
Dublin 13
Dublin 11
Dublin 12
Dublin 4
Dublin 14
Dublin 3
Dublin 16
Dublin 5
Dublin 1
Dublin 22
Dublin 20
Dublin 2
Dublin 17
Dublin 10
Dublin 6w
Postal Code


VAT Exclusive
Count
8000
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
VAT Exclusive


Property Size Description
Count
0
100
200
300
400
500
600
700
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
Property Size Description

END