Data Quality Report

17350796

From the financial crash to the crippling housing crisis: What is the future of Ireland's housing nightmare?

2022

This report aims to give an initial discussion of my findings of the cleaned data set. Moreover, it seeks to describe the data quality issues associated with this file.

My initial impression was that the data was relatively clean with good logical integrity. However, it lacks numeric and unique data. Several tests and examinations were performed on the data to check its logical integrity.

Types

The data frame was changed to appropriate types so suitable statical tests could be performed on the features. 'Date of Sale' was converted to datetime64. 'Postal Code', 'Address' 'County', 'Not Full Market Price', 'Vat Exclusive', 'Description of property', and 'Property Size Description' were all converted to category. This is because all have a number of fixed possible values. 'Price (€)' is the only continuous data type in the data frame. Thus, it was converted to a float data type. I think the lack of other continuous features is a weakness of this data frame. There is potential for low sensitivity with predictions due to the data frame being comprised of largely categorical data types.

With regards to '**Date of Sale'** both month and year were extracted and placed in separate columns. This is to make to easier to analyse monthly and yearly trends. The dates were checked to see if they fell into range given in the brief. It was found that there were house sales in the future of 2022 (*see sample screenshot below*). It appears that these dates were entered with the

	Date of Sale (yyyy/mm/dd)	Address	Postal Code	County	Price (€)	Not Full Market Price	VAT Exclusive	Description of Property	Property Size Description	Month	Year
1032	2022-04-01	5 STRAFFAN GATE, STRAFFAN, KILDARE	NaN	Kildare	186974.0	No	No	Second-Hand Dwelling house /Apartment	NaN	4	2022
2742	2022-07-01	78 HILL OF DOWN, SPENCER DOCK, MAYOR ST DUBLIN 1	Dublin 1	Dublin	382000.0	No	No	Second-Hand Dwelling house /Apartment	NaN	7	2022
3042	2022-04-01	4 CNOC BEAG, LIS CARA, CARRICK ON SHANNON	NaN	Leitrim	161000.0	Yes	No	Second-Hand Dwelling house /Apartment	NaN	4	2022

month and day swapped. It is assumed they are all sold in January of 2022. This will be corrected.

It is of course possible that other dates throughout the years were entered incorrectly. However, there is no other meaningful way of correcting these errors other than evaluating how they look in the data frame. The datetime will throw an error if a day does not match a month for example.

Null Values

Only two features saw null values:

```
Date of Sale (yyyy/mm/dd)       0
Address                         0
Postal Code                  8140
County                          0
Price (€)                       0
Not Full Market Price           0
VAT Exclusive                   0
Description of Property         0
Property Size Description    8971
Month                           0
Year                            0
dtype: int64
```

'**Property size description'** has over 89% missing values. This is obviously a huge proportion. However, the information provides an indication of the size of properties. This metric could be very powerful in terms of price prediction as there is a lack of unique data about the houses in the data set. For this, it may be valuable in the future if there was an appropriate way to fill in the missing data. All NaN values were changed to none for descriptive purposes.

Currently, '**Postal Codes'** remain in the data set even though they are only relevant to Dublin and only 58% of Dublin in this data frame has them(*see Appendix B*). Dublin represents over 30% of the data and this information could provide insight about the most populated County. For example, different trends of price and time could be looked at in more detail in different areas of Dublin. A way to add more post codes is also worth exploring.

Other Issues

'**Description of Property'** showed 3 unique values. 'Second-Hand Dwelling house /Apartment', 'Second-Hand Dwelling house /Apartment' and 'Teach/Árasán Cónaithe Atháimhe'. The last one is simply the Irish translation for second-hand dwelling house/apartment. While Irish people have the right to use their official language, for the purpose of data analysis I have absorbed this value into its English translation leaving only two unique values for this feature.

As mentioned the '**Date of Sale**' showed no missing data. There was a list of descriptive statistics on this data also, such as frequency, mean, max and min. This data will be used to see trends when compared to the price. From looking at the bar charts for years (*see*

Appendix E) you can see that the number of sales is increasing as the years go by. Only one month of data is supplied for 2022.

From looking at the '**Price'** column, this was the only pure numeric data in the csv. This is also our target feature. There is a large range of data with a large standard deviation(*see Appendix L*). This is visually clear from observing the Price histogram, it is extremely right skewed (*see Appendix H*). This further confirmed when looking at the box plot which shows the outliers (*see Appendix M*). The outliers were temporarily removed by exclusion through percentiles. As you can see from looking at the histogram with the adjusted price (*see Appendix I*) whereby only from the 0.01 percentile to the 0.99 percentile are taken into account, thus, removing outliers you can see the price points in which most of the data lay. We can see that most price reside around the 100,000 mark to the 500,000 mark. The actual outliers will have to be examined in more detail in the data quality plan. For example, the top sale is €17 million which is a group of apartments. This type of data will have to be removed or changed in order to keep consistency as our aim is to predict house prices of single dwellings.

With regards to **'Not full Market',** there is a small percentage of 4.5% of houses that sold not at the full price (*see Appendix F*). This can be due to a number of reasons like the dwelling may have been purchased at a reduced price under the Affordable Homes Scheme or only a fraction of the dwelling was sold. Those sales will be removed for consistency.

With regards to '**Vat Exclusive',** the overwhelming majority of house sales in this data set were not VAT exclusive (*see Appendix D*). Again, this shows good consistency in the price data. Those that were Vat exclusive will not be removed as the difference does not appear to be great enough to warrant the deleting of data.

Unique Values:

As mentioned, all unique values were verified to be valid (*see figure below*).

```
Date of Sale (yyyy/mm/dd)    2759
Address                      9980
Postal Code                    22
County                         26
Price (€)                    2285
Not Full Market Price             2
VAT Exclusive                     2
Description of Property           3
Property Size Description         4
Month                            12
Year                             13
dtype: int64
```

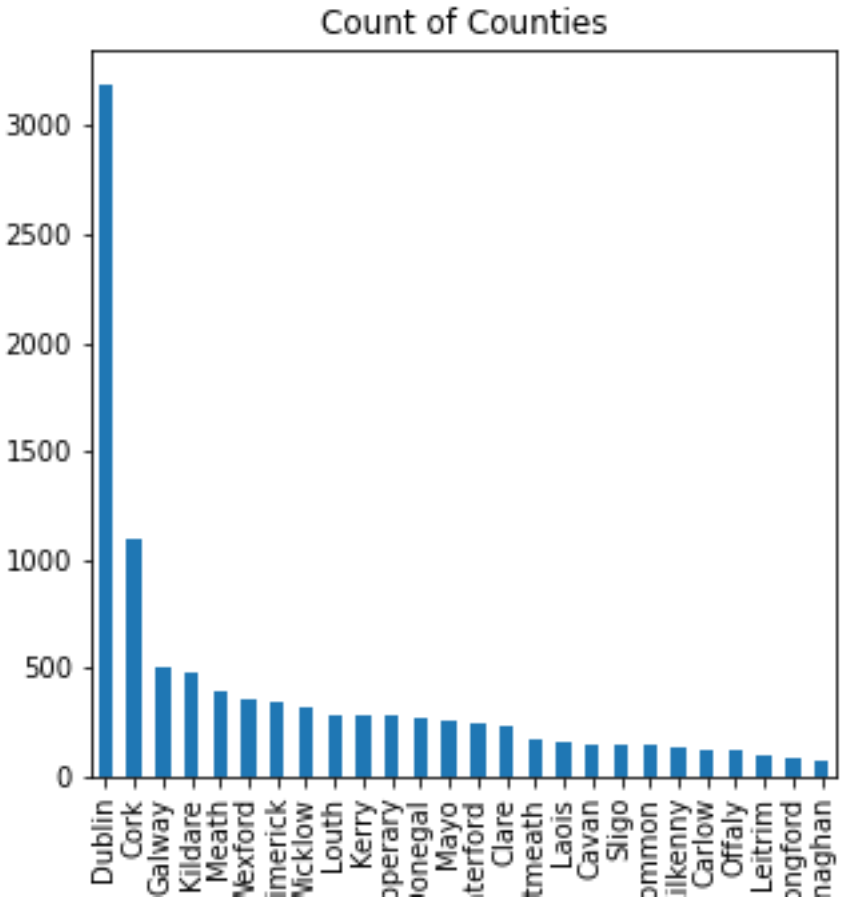
With regards to **County,** we can see from the table above and *Appendix A* we can see that the 26 counties were included. This makes logical sense as the 6 counties in the North of Ireland are not included in Property Price Register. Furthermore, we can see that number of sales appears to be roughly proportional to population of each county. This will be explored in more detail.

From running unique value checks it is clear that there are actually very few unique elements in this dataset. As mentioned, this may be in an issue when it coming to designing a model to predict house prices. For example, a feature like garden size would give us more statistical feature. That is why there is a section in this model whereby new features can be added to strengthen the data frame.

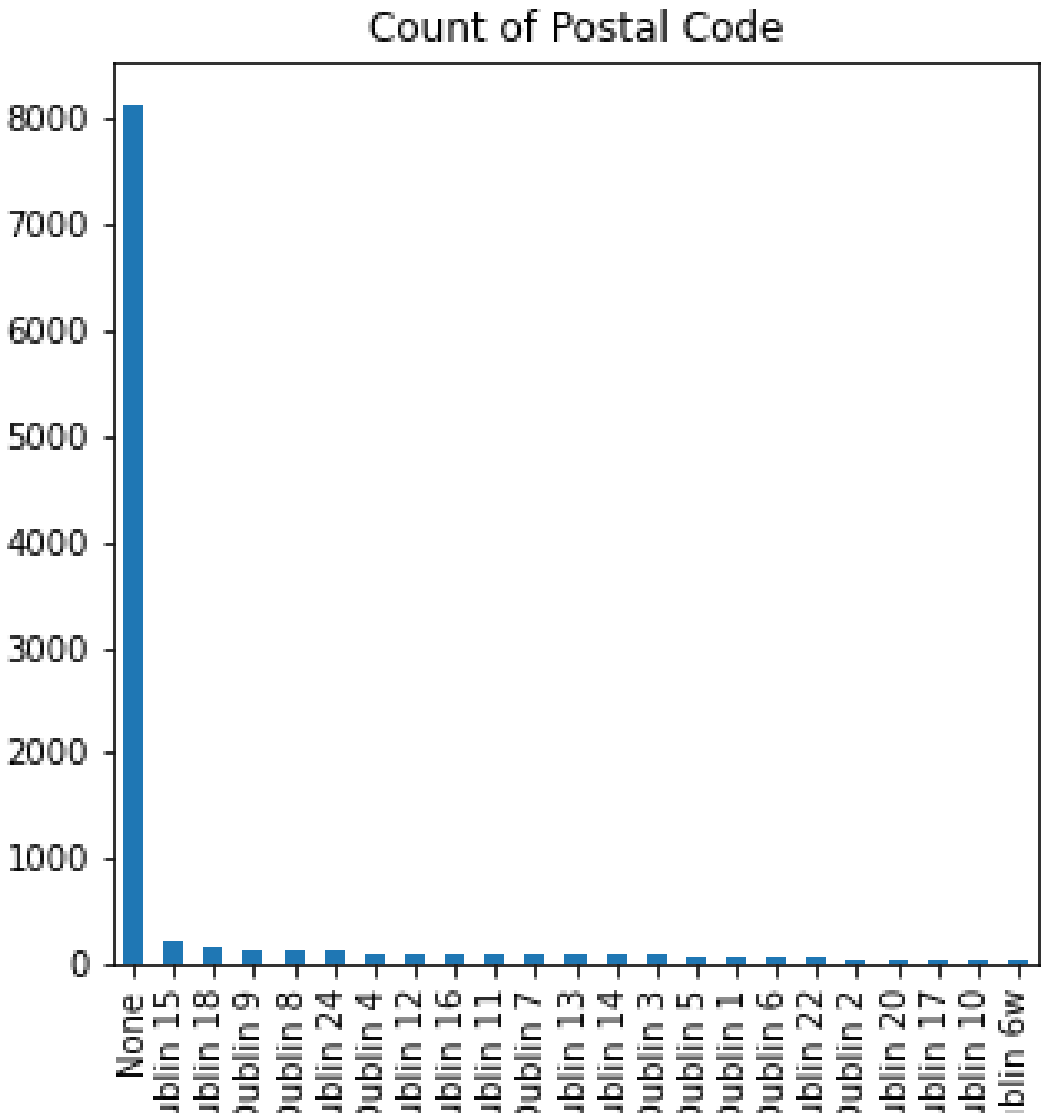
Appendices:

All graphs and descriptives tables:

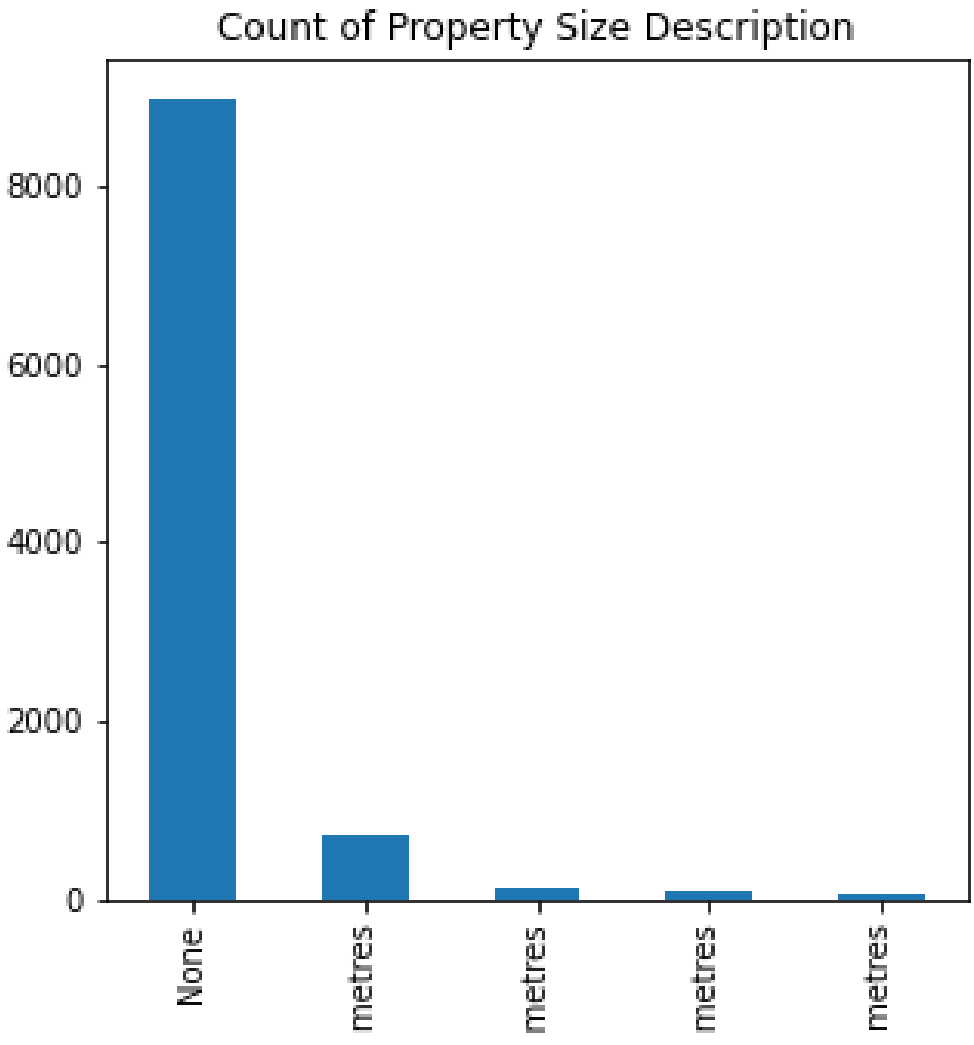
Appendix A



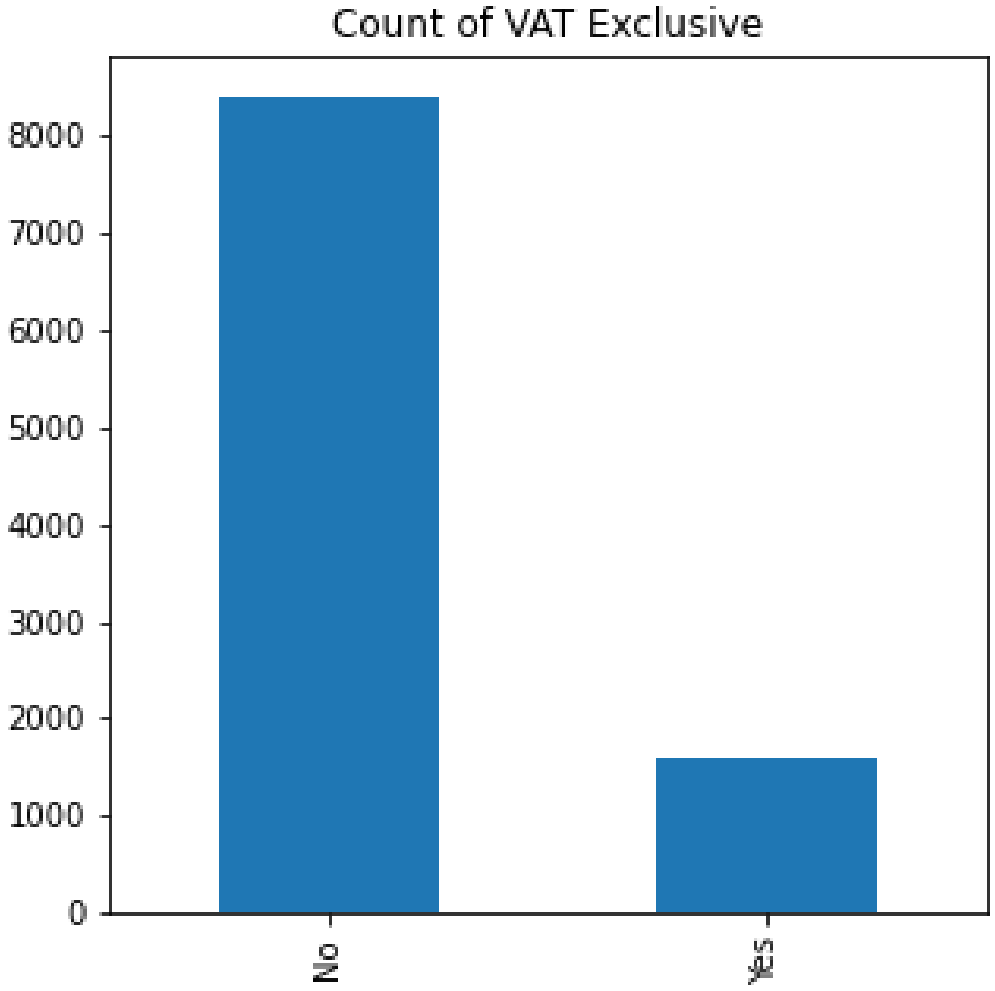
Appendix B


Count of Postal Code
8000
7000
6000
5000
4000
3000
2000
1000
0
None
ublin 15
ublin 18
ublin 9
ublin 8
ublin 24
ublin 4
ublin 12
ublin 16
ublin 11
ublin 7
ublin 13
ublin 14
ublin 3
ublin 5
ublin 1
ublin 6
ublin 22
ublin 2
ublin 20
ublin 17
ublin 10
blin 6w

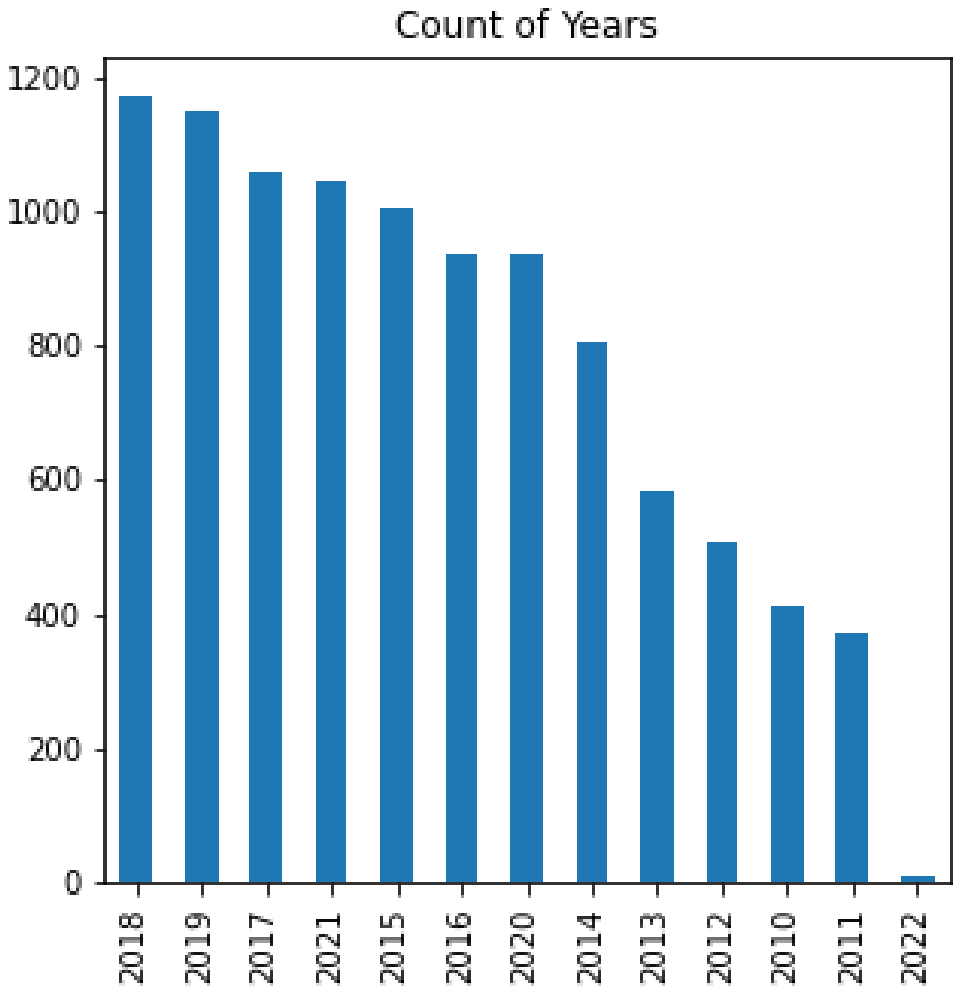
Appendix C


Count of Property Size Description
8000
6000
4000
2000
0
None
metres
metres
metres
metres

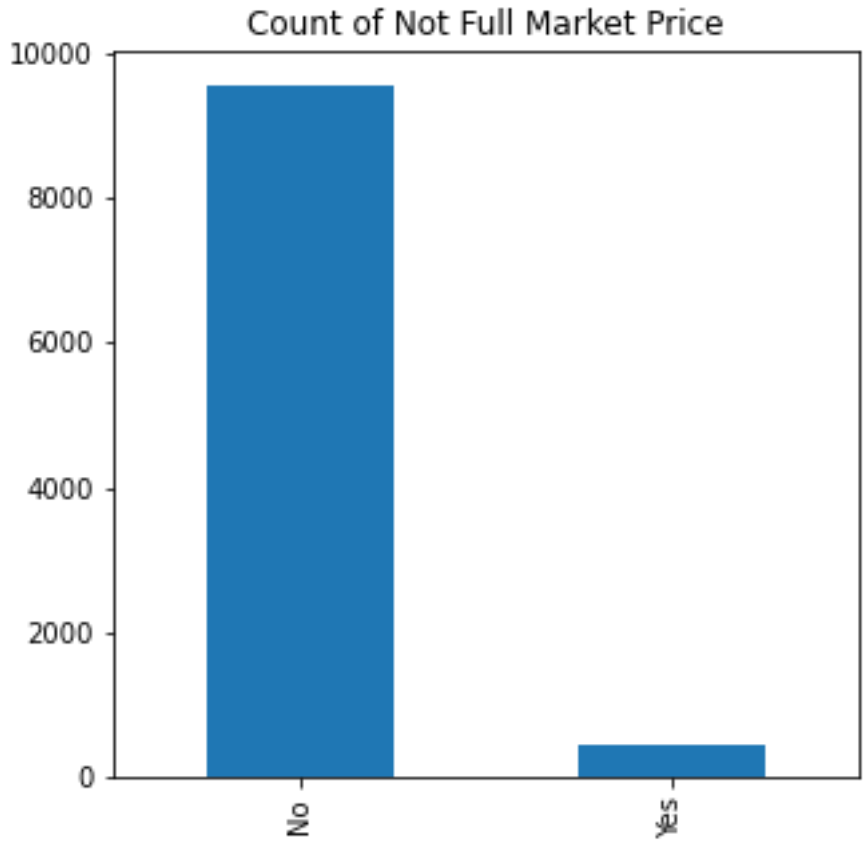
Appendix D


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

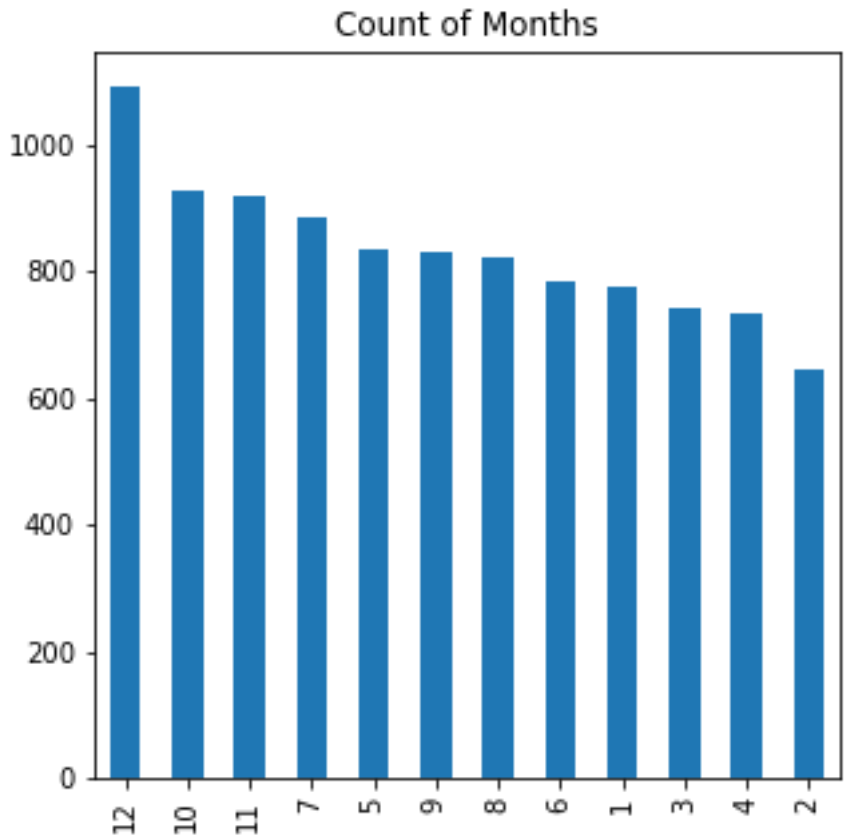
Appendix E


Count of Years
1200
1000
800
600
400
200
0
2018
2019
2017
2021
2015
2016
2020
2014
2013
2012
2010
2011
2022

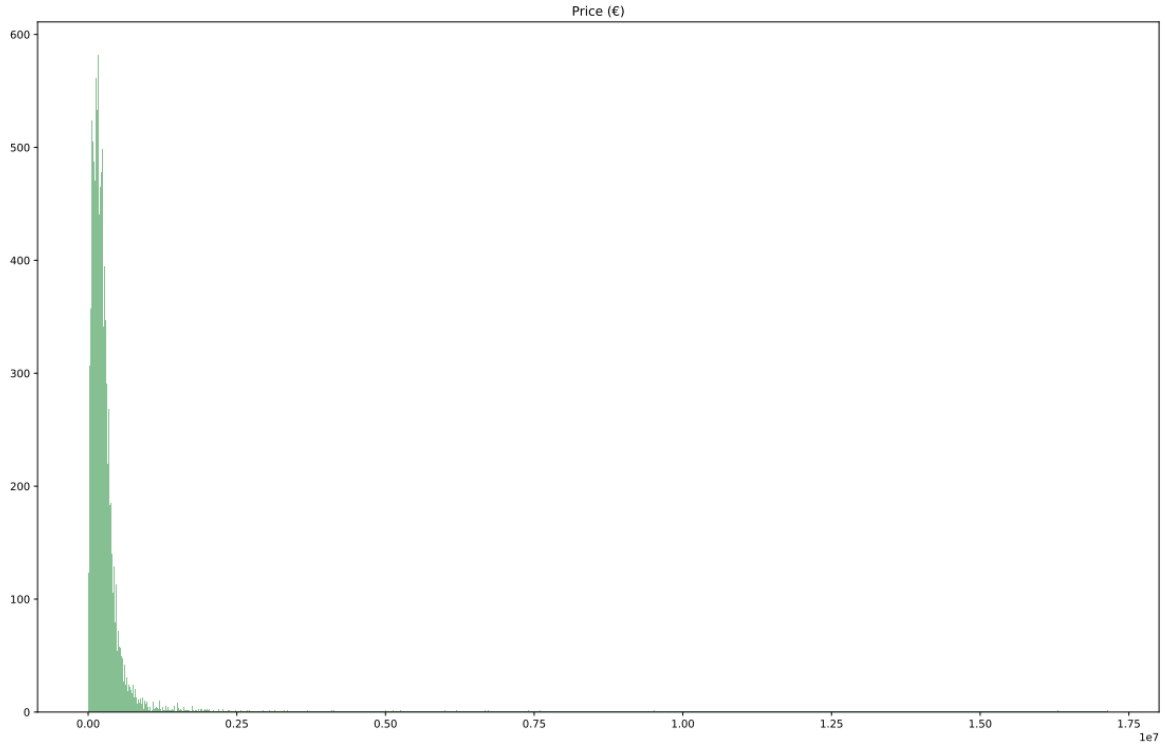
Appendix F


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

Appendix G

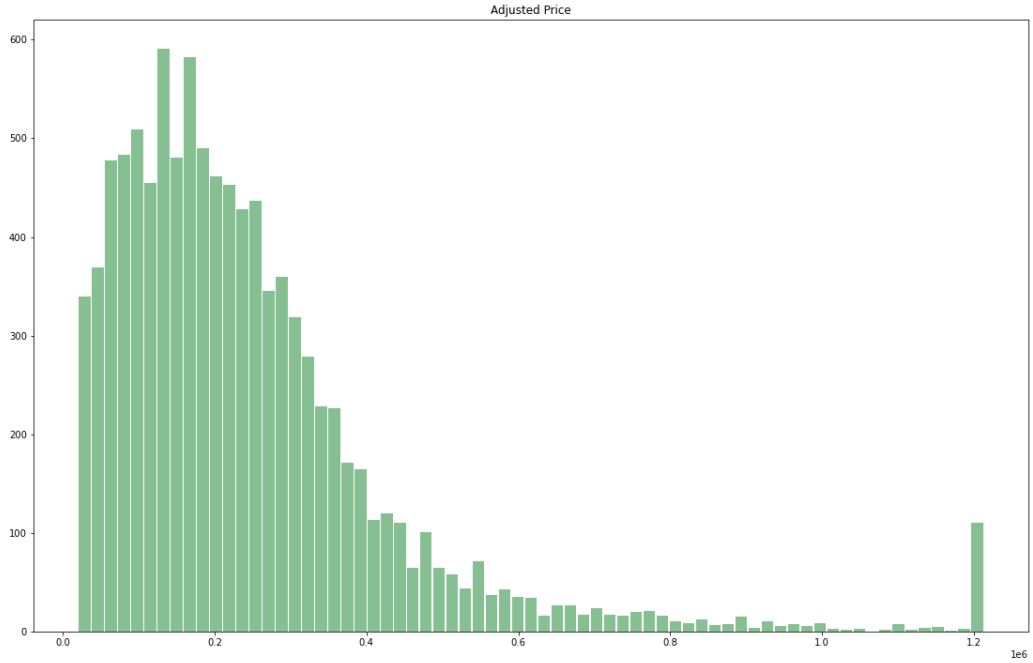


Appendix H


Price (€)
600
500
400
300
200
100
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e7

Appendix I

Clamped Price:



Appendix K

	count	unique	top	freq	mode	freq_mode	%mode	2ndmode	freq_2ndmode	%2ndmode	%missing
Postal Code	1860	22	Dublin 15	232	Dublin 15	232	0.124731	Dublin 18	143	0.076882	81.4
County	10000	26	Dublin	3193	Dublin	3193	0.3193	Cork	1099	0.1099	0.0
Not Full Market Price	10000	2	No	9552	No	9552	0.9552	Yes	448	0.0448	0.0
VAT Exclusive	10000	2	No	8399	No	8399	0.8399	Yes	1601	0.1601	0.0
Description of Property	10000	2	Second-Hand Dwelling house /Apartment	8364	Second-Hand Dwelling house /Apartment	8364	0.8364	New Dwelling house /Apartment	1636	0.1636	0.0

Appendix L

	count	mean	min	25%	50%	75%	max	std
Date of Sale (dd/mm/yyyy)	10000	2017-01-19 02:06:08.640000	2010-01-02 00:00:00	2014-11-07 00:00:00	2017-06-02 12:00:00	2019-08-04 00:00:00	2022-12-01 00:00:00	NaN
Price (€)	10000.0	260260.702676	5500.0	120000.0	200000.0	308370.08	17151248.0	390170.322939

Appendix M