Data Quality Report - Initial Findings

1. Overview

This report will outline my initial findings based on the cleaned dataset (PPR-ALL.csv), giving an overview of the data, while outlining data quality issues found and how they will be addressed. By using a combination of summary statistics and visual data representation such as histograms, box plots and bar charts, I will analyse the data and the relationship between features.

Initially there appeared to be lots of missing data regarding address Eircodes and the description of the property size. Investigating further, I found there were other data quality issues within the dataset which I will outline in this report.

2. Summary

Before I investigated each feature, I checked the column names and cleaned them by removing the whitespace and special characters from each.
From here, I investigated the data where the first data quality issue I came across was the type of data for each feature. I expected there to be a variation in data types since the dataset contained information relating to price and dates, however all features were of the object data type. Firstly I changed the DateofSale column to datetime data type, which would allow me to make a histogram using this data. Next I made Price data continuous, removing the '€' from each row in order to convert the data to a float data type. It was important to have price as a continuous feature as it was the main feature to be analysed in my opinion.

A number of rows where data had been entered more than once. There were 850 of these duplicated rows, and so they were dropped.

Both Eircode and Property Size Description had a large amount of data missing, over 84% and 91% missing respectively for each. I will investigate further and see whether or not it is possible to retrieve information for these features from data in other columns.

3. Logic Check

I performed one logic check on the DateofSale data to ensure that dates were in the specified time range and did not begin before 1/1/2010 or take place after the end date 5/5/2023. It appeared that all dates were entered correctly within the expected timeframe.

From here, I checked the values of features with low cardinality, where I found both PropertySizeDescription and DescriptionofProperty to have extra values that fell under one of the other values. There were five categories for DescriptionofProperty, three of which were in Irish. The English translation for these values are
and so I changed it to the English translation which is the same as 'Second-Hand Dwelling house /Apartment' and ‘New Dwelling house /Apartment’.

For PropertySizeDescription, the values 'greater than 125 sq metres' and 'greater than or equal to 125 sq metres', are the same category so I combined the two. There were two other categories in Irish which were then translated to English.

4. Review Continuous Features

4.1 Descriptive statistics

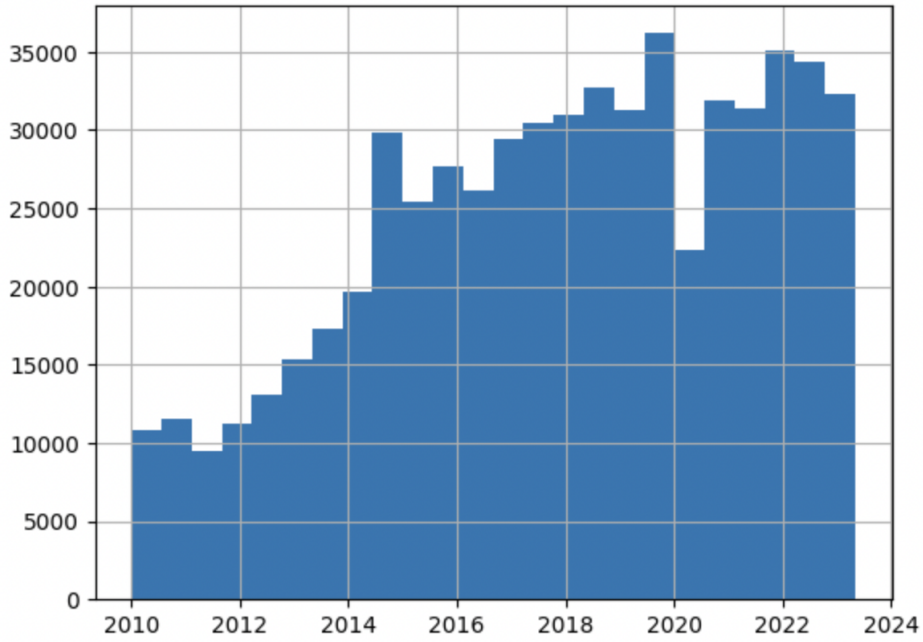
- Numeric: There are two main categories of continuous features, numeric and datetime. The sole numeric feature is the price each property was sold for, while the only datetime feature relates to the date it was sold. I will discuss the numeric and datetime features separately below. The price for properties range from €5,001 - €225,000,000 which seems to be an extremely large range. The mean is €281,362.67 and the third quartile is €322,000 both of which are not even 1% of the max price. This appears to suggest the presence of a number of outliers in the data, which will need to be investigated further. This will become more evident later in the report when I detail some of the graphs I have created.

- Datetime: It appears the timeframe this dataset is focused on is between 01/10/2010 to 05/05/2023, with the first date of sale taking place on 01/01/2010 and the last date of sale on 05/05/2023.

4.2 Histograms
All of the plots I have created are available in the accompanying notebook but I will focus here on a few plots which I believe yielded particularly valuable information which may need to be acted on. The histograms which I believe displayed pertinent informations are the following:

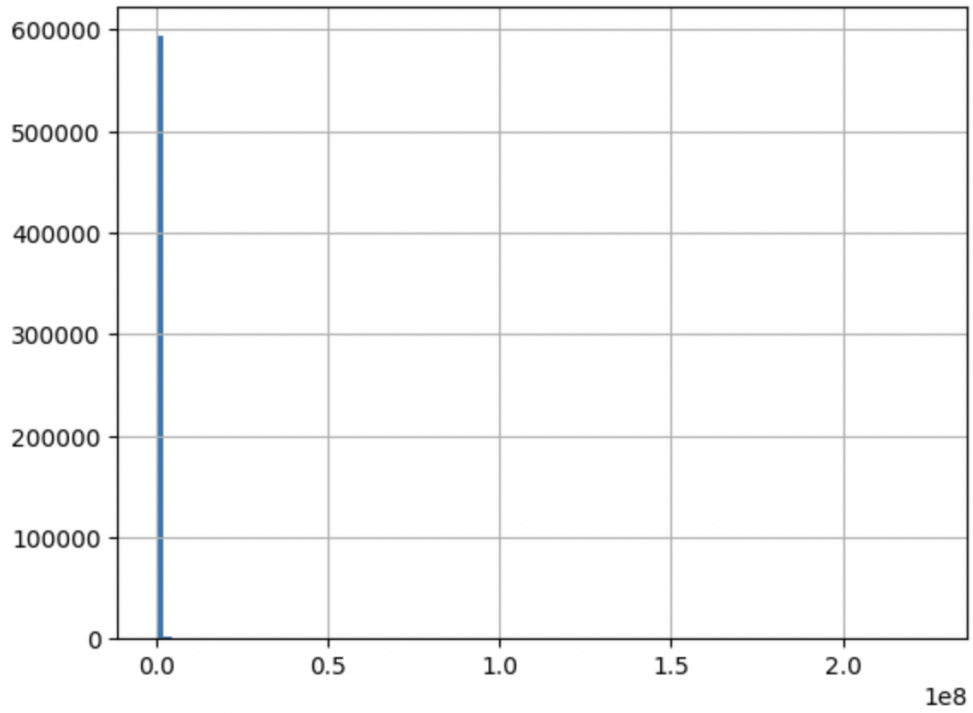
In [62]:
```
#Plot histogram for date of sale
df['DateofSale'].hist(bins=24)
```

Out[62]: <Axes: >



```
In [61]: # Plot histogram for Price
         df['Price'].hist(bins=100)
```

```
Out[61]: <Axes: >
```



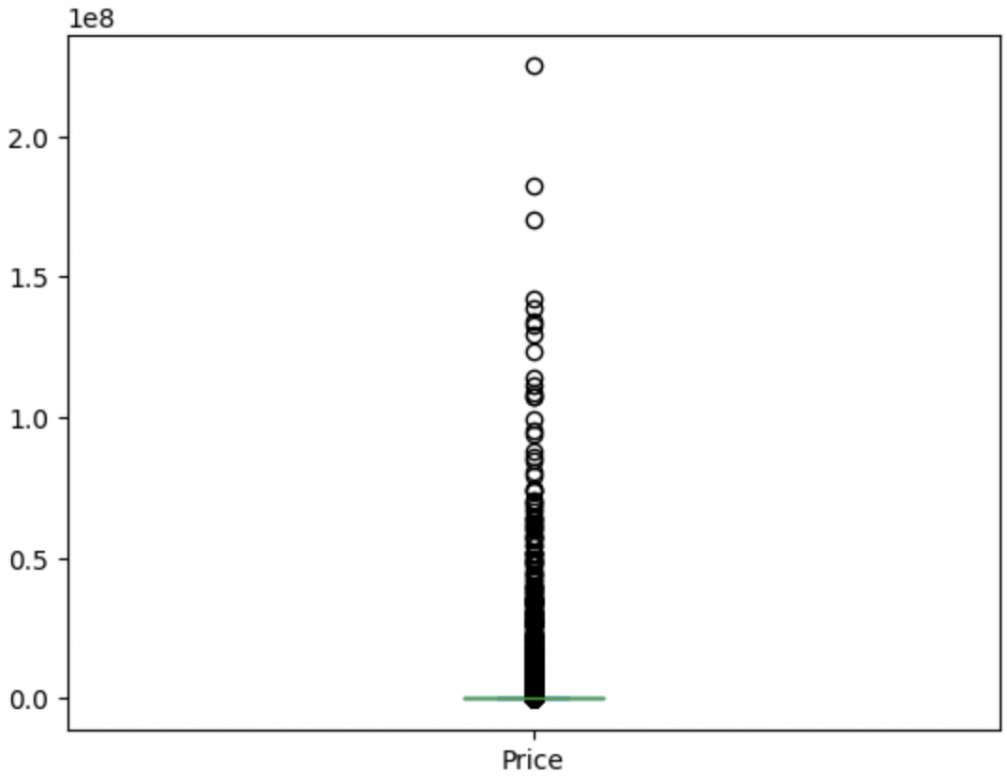
The first histogram shows the date of sale on the x-axis and the number of sales on the y-axis. It is clear to see a positively skewed distribution showing an overall increase in property sales from the years 2010 - 2023.

The second histogram shows the price of property sales on the x-axis and the number of sales on the y-axis. It is hard to make an inference about the data using this histogram as only one bin is visible for the price. This shows the presence of significant outliers over a very large range.

4.3 Box Plots

```
In [63]: # Plot boxplot for Price
         df['Price'].plot(kind='box')
```

```
Out[63]: <Axes: >
```



The box plot clearly shows outliers in price data that make it difficult to evaluate the data. The interquartile range is quite squashed due to the presence of these outliers.

5. Review Categorical Features

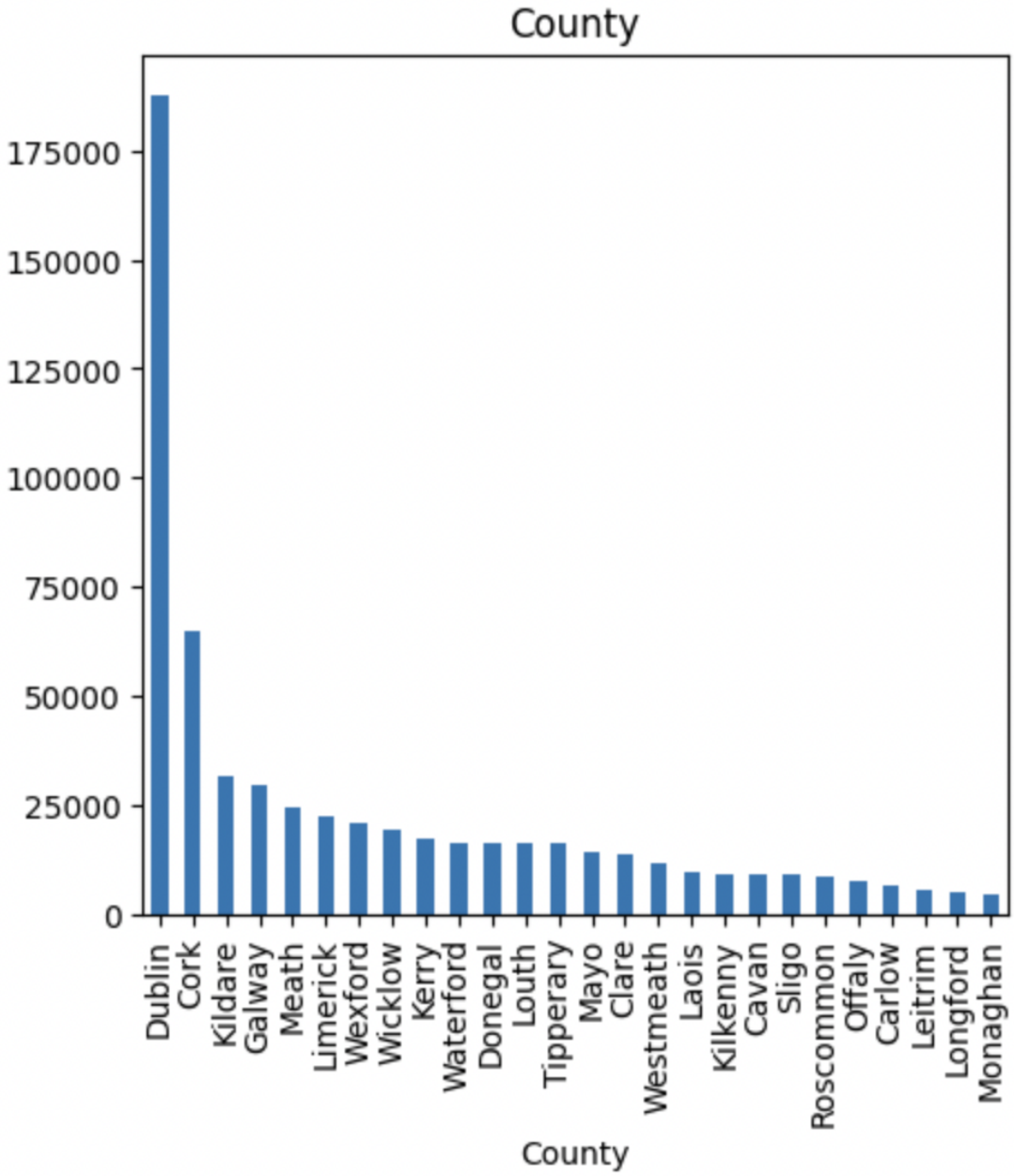
5.1 Descriptive statistics

The table detailing the descriptive statistics of the categorical features which is available in the accompanying notebook clearly shows that the “Address” feature has a high cardinality, which will be difficult to use for analysis.

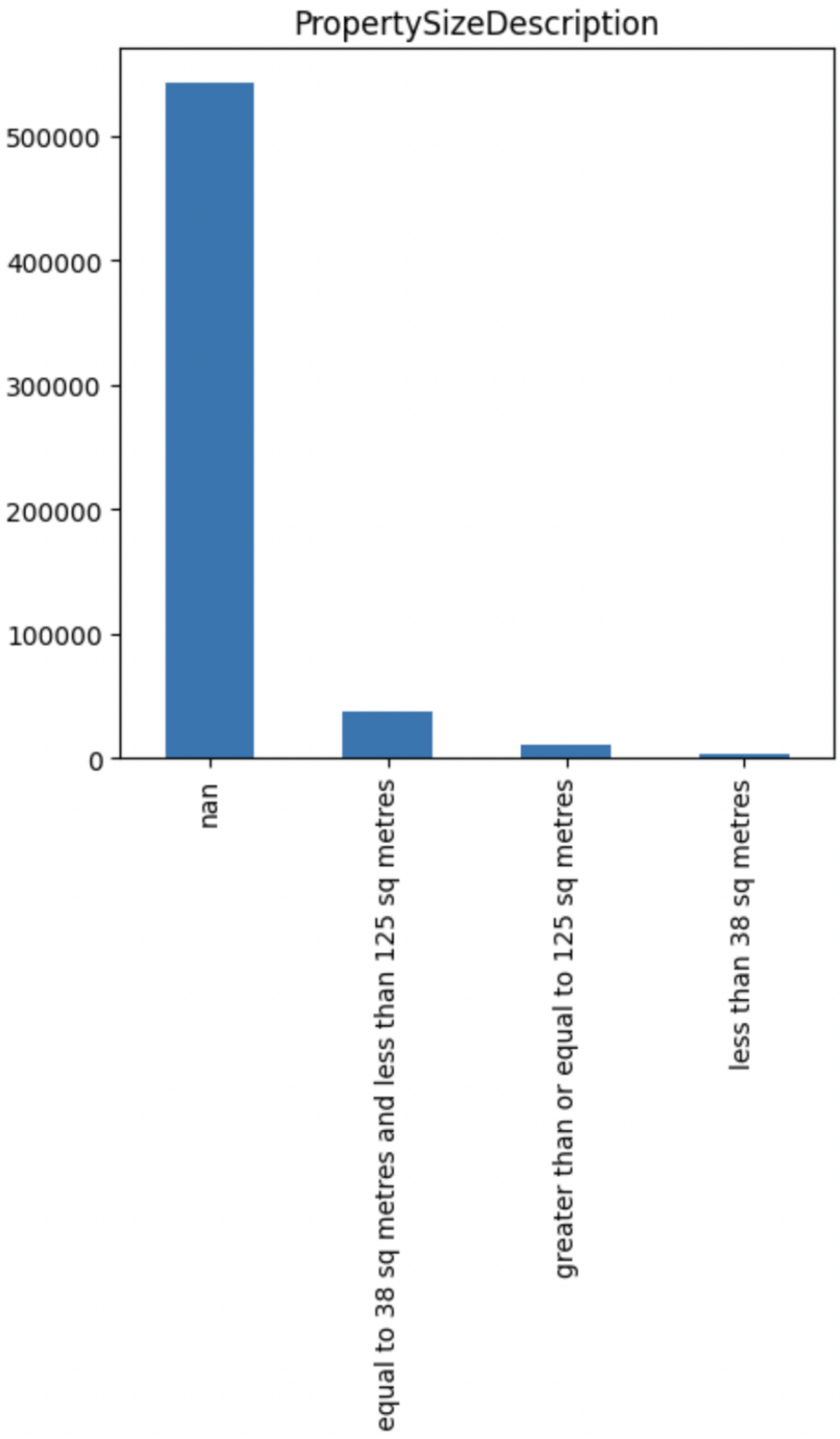
PropertySizeDescription is missing over 91% of data but has a similar count to the frequency of ‘New Dwelling house /Apartment’ in the DescriptionofProperty description and so I will investigate whether the two are connected before I decide to drop the PropertySizeDescription feature.

	count	unique	top	freq	%missing	card
County	595678	26	Dublin	187661	0.000000	26
Eircode	91516	89876	A96WV79	8	84.636666	89876
NotFullMarketPrice	595678	2	No	567023	0.000000	2
VATExclusive	595678	2	No	499505	0.000000	2
DescriptionofProperty	595678	2	Second-Hand Dwelling house /Apartment	497608	0.000000	2
PropertySizeDescription	52579	3	greater than or equal to 38 sq metres and less...	37891	91.173251	3

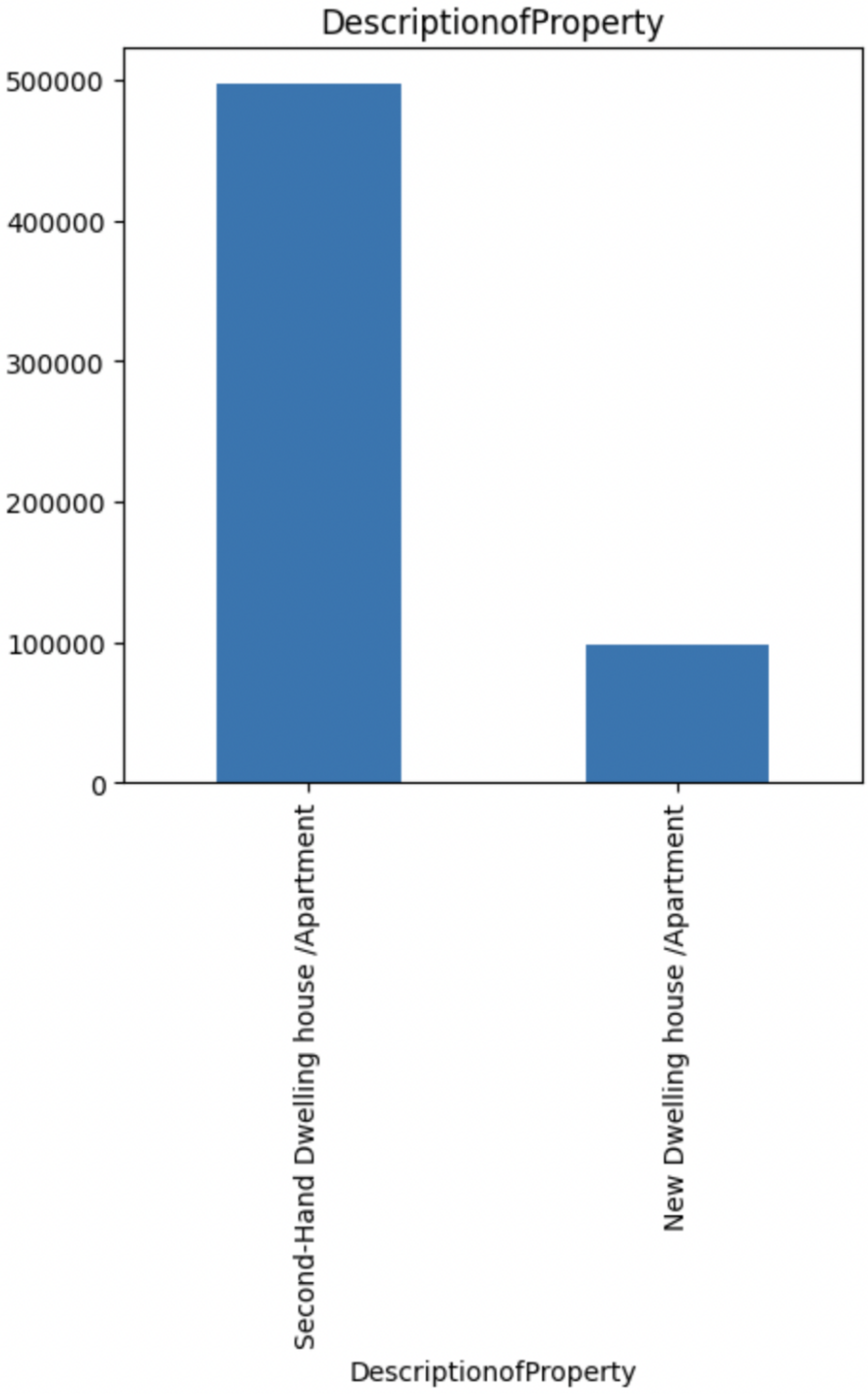
5.2 Bar Charts



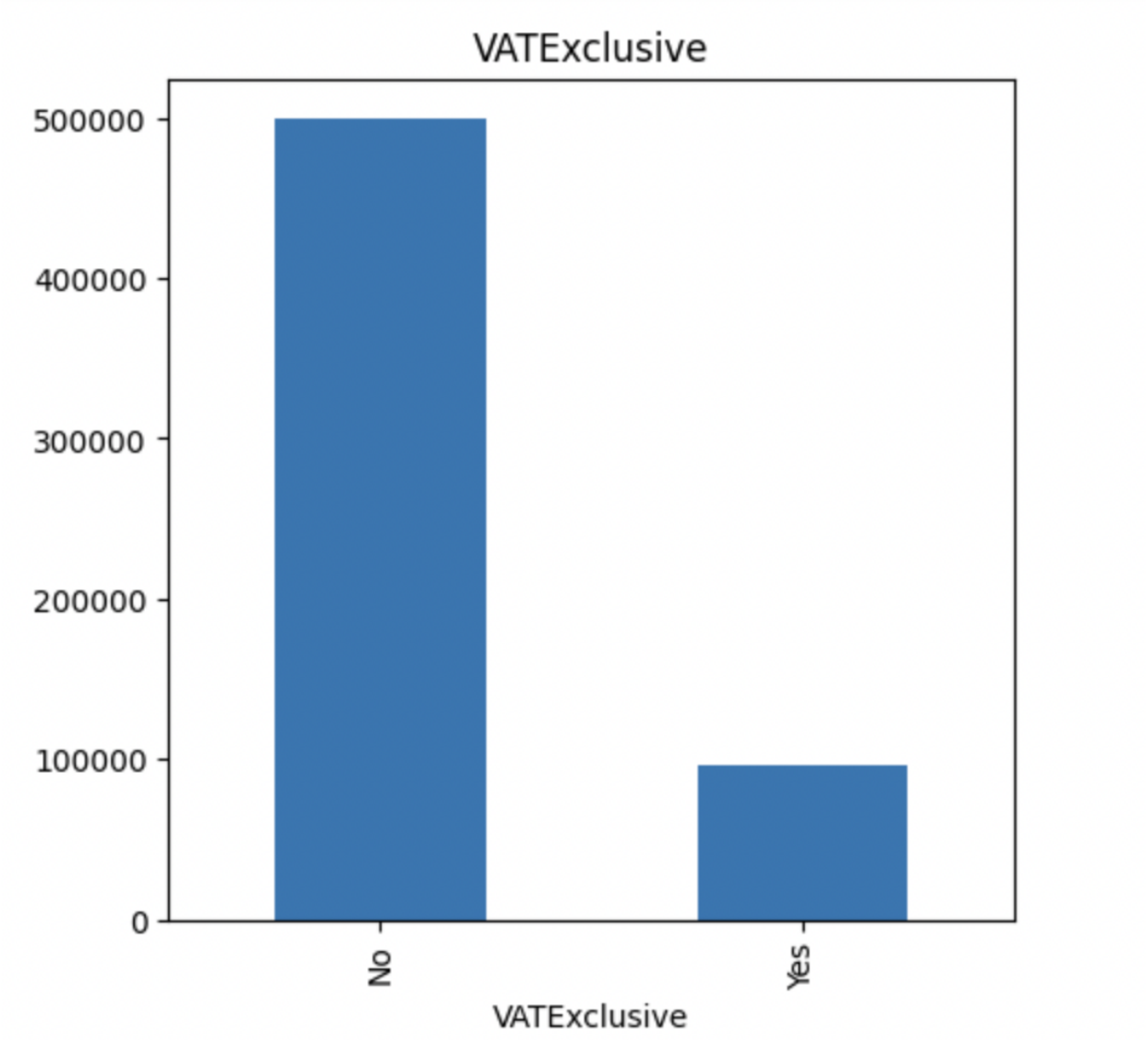
Here we can clearly see that the most common addresses are in Dublin, with the next most common is Cork. These are the two most populous counties in Ireland and it would make sense for them to have the most properties sold, although it is interesting that the number of Dublin sales are over double the amount of Cork sales.



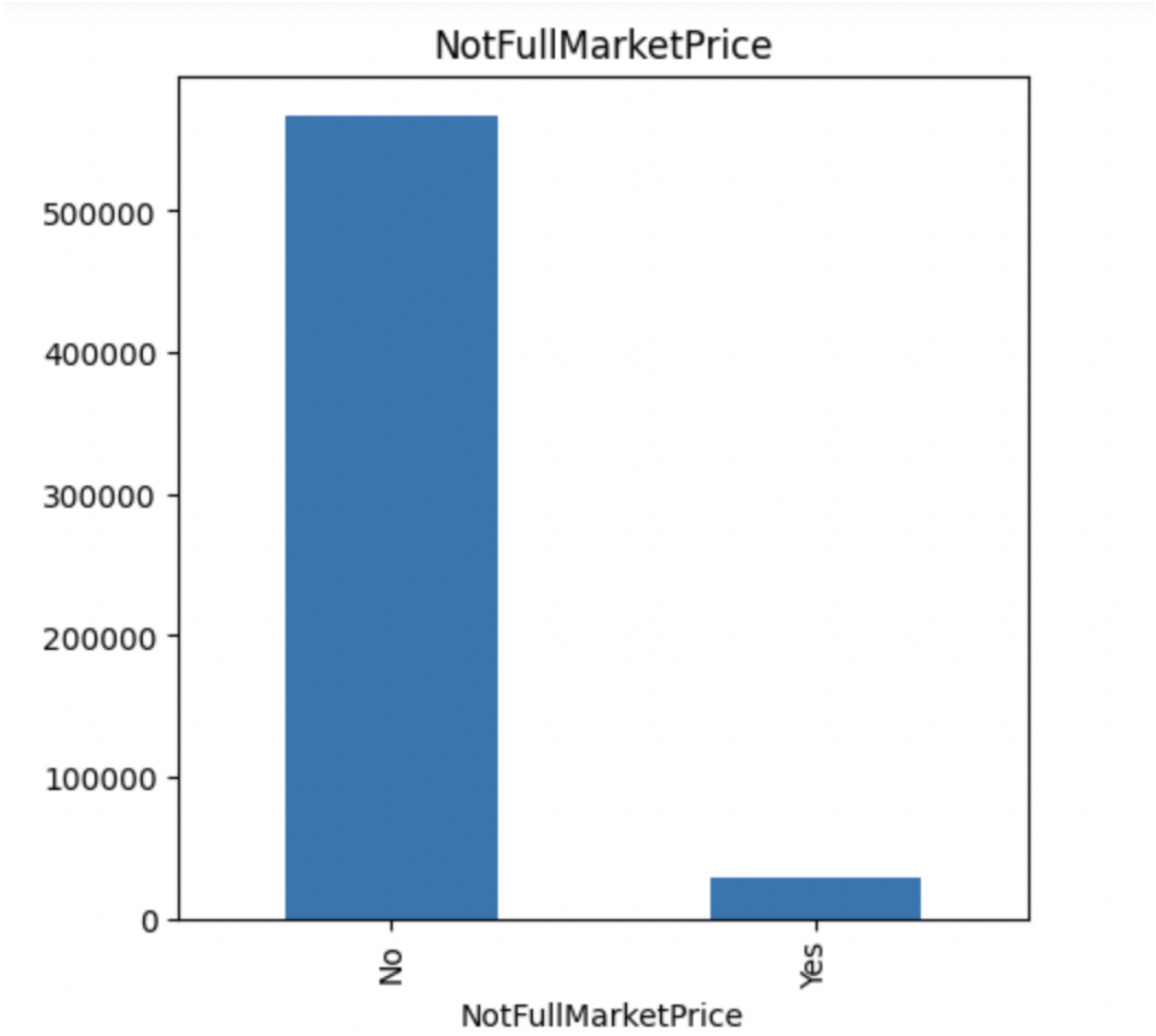
The most common properties have a range of size between 38 sq metres and less than 125 sq metres. However, most of the data is missing from this feature, which will be addressed in the data quality report.



It appears over 83% of properties sold are a Second-Hand Dwelling house/Apartment. This would suggest that the number of new homes being built is quite low compared to available housing.



Roughly 80% of property sales were not VAT exclusive.



Over 95% of properties were sold at their full market price.

6. Planned Actions

- “Missing values in Eircode column"
    - Inspect the Address data further and try to extract Eircode data from the Address column.

- “Missing values in PropertySizeDescription column”
    - Investigate the data further and see if there is a correlation between rows with 'New Dwelling house /Apartment' and non-empty PropertySizeDescription.

- “Outliers in Price data”
    - Investigate the data further.