Data Quality Report

Summary

This report will indicate the initial findings on the cleaned dataset “propertydataSecondClean.csv“ and will summarize the data and issues with this data , alongside solutions that were applied . The dataset is seen to have several issues which were addressed throughout the course of correction. There appeared to be no duplicate columns or rows, however there was many null values present in the postal code and size description columns. The postal code information was only available for to a selection of properties in Dublin, while size description was only available to new homes.

There were no issues found regarding address, not full market price and date however, data was split into day, month, and year columns to allow for greater investigation of the data. A years since purchase column was also added to help with added features that were implemented in the final clean by subtracting the value in the year column form 2022. The description column was largely left unchanged, however there were a small number of properties that were changed from their Irish to English spelling, as was all other data presented in Irish.

The price column presented the most attention required with several outliers as well as several rows where price did not include vat of 13.5% which was indicated through the vat exclusive column.

Review of features

Date of sale

The date of sale was tested by checking for any properties which had been sold on invalid dates. However, when checked it was found that while there were properties that sold in 2022 all of these had been for dates in January of that year and there were also no invalid dates in the past that were present. The date column was split into three columns of day, month and year before being dropped, allowing for properties sold on a certain date to still be found while allowing for more data comparing. The years since purchase column was added to create new features in later cleans.

Address

While there were a total of 20 duplicate addresses , these can be seen to be the same property sold on different occasions. These can be seen in the cases of 5 Marine Terrace, Tramore, Co.Waterford where a specific property number five of that address was sold for 350,000 in 2018 before being sold for 410,000 again in 2021, or in the case of the Highfield, Drynan, swords where there are two sales on different dates in 2017. For these two sales while in the same year and address, the properties are seen to sell for vastly different prices of 216,823 and 900,000 which would lead me to believe that these are not the same property as it is unlikely that the same location would have this large of an increase/ decrease in value in such a short period of time. Therefor all rows were seen as valuable information which were kept. The address column was left unaltered and was used to identify properties rather than to compare them. The data was tested to make sure that there was no null data.

Postal Code

The postal code column was missing a large amount of data in this dataset with this information mainly only available for some of the Dublin properties. While average property value in non-null postal code rows were compared in a bar chart, these values were removed from later cleans due to the amount of null values present. There was one row which broke the trend of being outside of Dublin and having a postal code this was seen as a mistake as this information was also present in the address column and is not a valid postal code. The spelling of the Irish of Dublin 15 was also changed to English.

County

The county values were tested to check if each had a valid entry of correct spelling of one of the 26 counties in the republic of Ireland, which the dataset passed. No further adjustments were made to this column.

Price and vat exclusive

There were several issues addressed with price. Firstly all data was clipped at the .01 quantile and .99 quantile to remove all outliers from skewing the data. Several properties such as 47 Thorndale Donegal sold for 10,570 euro in 2021 when the average sale of a property within the area of Letterkenny was 130,000 euro in the same year or in the case of the range, BallyWilliam, Wexford where the property sold for 7000 when the average price within 10 kilometers of this property was approximately 260,000 euro. These examples lead me to believe that they may have been errors when entering the data of some of these properties with potential for a value to be missing a digit reducing the sale price by up to 90% if present. There were also several sales such as a block of apartments in swords that sold for approximately 57 million and another set of apartments which sold for approximately 38 million in the alliance building which represent multiple properties grouped together and therefore also do not represent the value of each single apartment.

This clipping was done on two occasions before and after the vat was included in several properties which will be discussed. The vat exclusive column was used to see which price columns did not include a 13.5% vat fee in the properties price. For the columns this was indicated as “yes” these columns were multiplied by 1.135 to show their true value. Any column that had a “no” had vat already included in this price and was therefore not multiplied 1.135. the data was clipped for a second time to clip the properties that would have pasted the limit set of 1.25 million after vat was included. The vat exclusive column was then removed as after adjustments all properties would have had a value of “no” for this column and therefore it was not needed.

Not full market price

The not full market price column was tested for any values that were not equal to yes or no, however none were found. This column remained unchanged in the dataset and was plotted as a bar chart. It also had no null values.

Description of property

The description of property indicated whether a property was a new or secondhand dwelling. This remained in the data set largely unchanged except for some Irish to English translations. It was tested with no null values or invalid inputs present.

Property size description

This column only had information present for new dwellings. These were plotted after two values of greater than 125 sq meters and greater than or equal to 125 sq meters were merged. These findings

should only be seen to represent new dwellings as it cannot be said for certain that they represent all dwellings due to changes in popularity of locations and style of homes in new dwellings when compared to those built in the past. This column was also dropped after being plotted due to the large percentage of null values.

Review of histograms and boxplots

Please see attached pdf (categorical_bar, numerical_histograms and numerical_boxplots) files containing relevant histograms, bar charts and box plots for graphical representations.

Numerical histograms

Sales per day - A histogram was made of the total amount of properties sold on each day of the month. This was divided into 31 bins to represent each day. The 31st seems to be the only outlier as each other day is between approximately 275 and 400, however this will be due to the fact that there are not 31 days in each month. This would indicate that there is no particular day of the month in which people decide to buy property.

Sales per month - A histogram with total amount of properties in each month was made. The histogram would indicate that there is a trend of more properties being sold in each much as the year continues with the least amount being sold in January and February with a steady increase as the year progresses. There is also an outlier month of December which would indicate that a larger number of homes sold in this month compared to the trend. This shows that people are less likely to buy property the earlier in the year it is and that there is a spike in buying in December in particular.

Sales per year -The created histogram displaying sales per year shows a gradual increase each year from 2011 to 2019, after a slight drop from 2010 to 2011. 2011 was the year with the least sales and 2019 was the year with the most. There was a dip in 2020 before rebounding in 2021. 2022 can be seen to be an outlier as the year had not been concluded at the time of investigation and not all sales were concluded. The data was divided into 13 bins to represent the 13 years available in the data.

Price - Due to the outliers in the property data as discussed above it was decided to clip the data in this column. It was decided to clip based on the .99 and .01 quantiles removing a small percentage of outliers from each side. The higher values were clipped to 1.25 million and the lower values were clipped to 20,000 in the hope of reducing the impact of the outliers. The value of each bin in the below plot can be found by multiplying the figure on the x axis by 1 million for example a property displayed as having a value of 1 will represent a value of 1 million. The data is skewed to the right as the majority of the properties can be seen to sell for between 50 and 400 thousand with a lower and lower number selling for each value above this 400 thousand. Each bin is also equal to a range of 25000 as the total of 0 to 1.25 million is divided into 50 segments. The outlier of between 1.225 and 1.25 million can also be seen due to the clipping of properties above this value.

Numerical boxplots

Day of sale boxplot - The box plot of days on which properties were sold would seem to back up the idea that the data is evenly distributed. The minimum value is 1, the lower quartile is 9 and the higher is 23 with a maximum of 31 and a median of 16. There are also no outliers present in the diagram as no values are greater than 3/2 times of the higher percentile and none are lower than 3/2 times less than the lower quartile.

Month of sale boxplot - The box plot of which month each property was sold would suggest again that there are more properties sold in the later parts of the year than at the start. There are no outliers with a minimum value of 1 and a maximum of 12. The median value is higher than half way and the higher quartile is at approximately 10 which would again back up this idea of a higher amount of properties sold towards the end of the year.

Year of sale boxplot - The box plot for year in which properties were sold shows that there was an increase in the number of properties sold as time progresses. The median of the data is 2017 with a higher quartile of 2019. This is the case even though the 2022 data is incomplete, and one would expect that both these figures would increase once 2022 has been completed, showing that the data is skewed to the left.

Price boxplot – This boxplot shows again that the majority of were under 300,000 euro and that while there were no outliers less than 3/2 times the lower quartile, even though the higher values were clipped at 1.25 million there were still a significant number of outliers greater than 3/2 times the higher quartile. If the data had not been clipped there would be even more significant outliers that would be seen in the data.

Categorical bar charts

Not full market price bar chart – This bar chart shows that the vast majority of homes did not not sell for their full market value meaning that the vast majority of homes sold for their market value.

Postal code bar chart including nan values - The intention of this bar chart is to show that the majority of properties did not have any information for postal code and that the only properties that had this information where properties listed in Dublin, with several different Dublin area codes.

Postal code bar chart not including nan values - This bar chart shows only the properties sold which have a listed postal code. This plot shows that the area of Dublin 15 sold the most properties by a large amount having nearly double the amount listed compared to Dublin 24 and 8 which came in 2nd and 3rd for most listed with their postal code.

County bar chart - This bar plot shows that Dublin had the most sales of the properties having nearly 1/3 of the total sales occurring in this county. Cork came in second with approx. 1000 sales while Longford and Monaghan were the counties with the least number of sales.

Vat exclusive bar chart - This bar chart shows that most properties had vat included in their sales price with between with 1 and 2 thousand not having included this extra 13.5 percent fee. This column was used to adjust the price of properties in which it had not been included.

Description of property bar chart - This data shows that the majority of homes were secondhand dwellings with just of over 8 thousand of the total ten thousand.

Size of property including nan values - This box plot shows that the majority of homes did not have a listing of their size. These values were only available for new dwelling which were show above to only account for between 10 and 15 percent of the data showing that there is a correlation as between 80 and 90 percent of properties are also missing this data due to being second hand dwellings.

Size of property not including nan values - This box plot shows the sizes of the properties only where this information was given, removing all nan values. This shows that most properties were between 38 and 125 meters in size for new homes only. This data may not be a valid description of all homes

as it cannot be said for certain if older homes were built in similar sizes due to changes in where homes are located and styles of homes changing over time.

Statistics for both categorical and continuous features

Statistics for both categorical and continuous features can be found in the attached csv files categoryDescription and continousDescription.