4) DATA ANALYSIS AND PREPARATION

This chapter contains a data profile and quality assessment, exploratory data analysis, and details the data cleaning and pre-processing procedure.

4.1) Data Profile

Data Description

The dataset consists of 2,982 observations and 16 variables for property sale ads placed in Dublin, Ireland. The dataset is made up of 7 numerical (3 discrete and 4 continuous), 6 categorical (1 ordinal and 5 nominal) and 3 unstructured variables. A description of each variable is given in APPENDIX A.

Data Quality Assessment and Cleaning

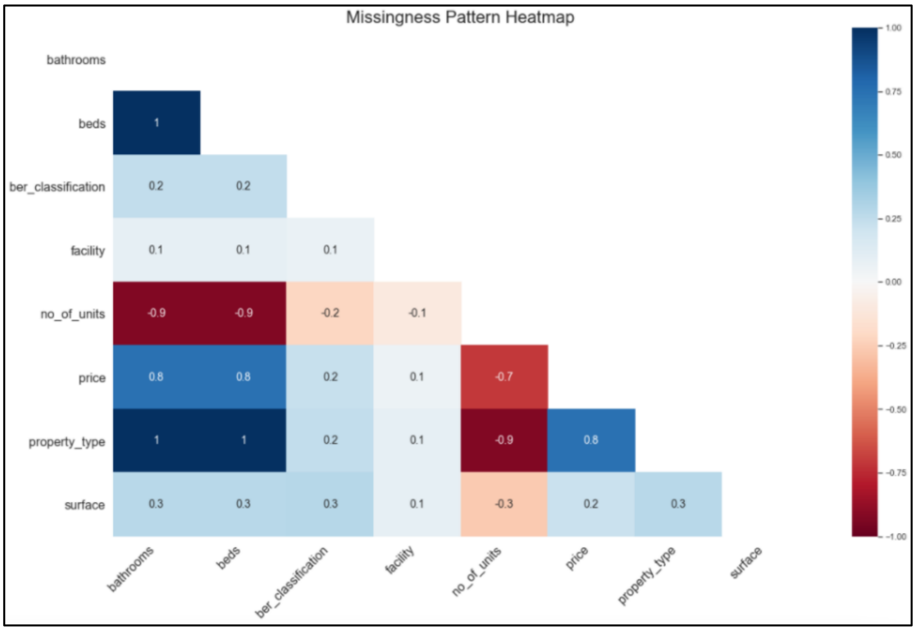
"High-quality datasets are essential for developing machine learning models." (Gudivada, et al., 2017)

The quality of the data that is fed to machine learning models directly impacts the efficiency of the models and the accuracy of their output. Prior to training a model it is important to assess the quality of the data and improve the dataset through data cleaning processes (Gudivada, et al., 2017).

Typical data quality assessments applied to organisations evaluate the data across six dimensions, timeliness, completeness, uniqueness, validity, consistency, and accuracy (CDC, n.d.). As the dataset is static and there is no data warehouse or pipeline, the data quality assessment in this report will consider only the following dimensions, completeness, uniqueness, and validity.

Completeness refers to the number of missing data points in the dataset. Analysing the missingness patterns of data points allows us to assess whether the absence of a value is indicative of an absence of a property feature or whether it constitutes a null value. This will help us determine how best to handle missing data.

There are missing values in 8 of the 16 variables in the datasets, with 4 variables missing more than 10% of their values. The completeness of each variable is shown in APPENDIX B. To assess the type of missingness present in the data, the missingness pattern heatmap below was created, this visualises the correlation between the presence of a value for one variable and the presence of a value for another.



It is evident from the heatmap that the presence of 'no_of_units' negatively impacts the presence of values in other variables. The 'no_of_units' variable is linked to 59 new development parent ads in the data, these sparsely populated observations were removed from the dataset and the 'no_of_units' variable was dropped. This improved completeness and there were now only 4 variables with missing data, 'price', 'surface', 'facility' and 'ber_classification'. Despite improvements in completeness, the 'facility' variable was only populated for 33% of observations, this falls below an acceptable level of completeness (75%) and as such the variable was removed. The other 3 variables had acceptable levels of completeness and remained in the dataset.

Uniqueness measures the number of duplicated rows in the dataset. Thankfully there were no duplicates present in the data. There were however some variables, 'county' and 'environment', that were constant across all observations. These variables add no new information to the model and as such were removed from the dataset.

Validity is the extent to which variables conform to a given format. Checks were made for the presence of any special characters present in the categorical data, any non-numerical values in the numeric data, and for any other aberrations from the format of the structured data. Only one issue was found in the 'ber_classification' variable, this was resolved with a quick change to exclude special characters.

Overall, the quality of the dataset is good, there are some missing values present in the data, but these are handled by removing sparse observations and dropping variables that fall below an acceptable completeness threshold. The absence of duplicates or incorrectly formatted data also enhance the quality of the dataset.

4.2) Exploratory Data Analysis

Relationships between variables

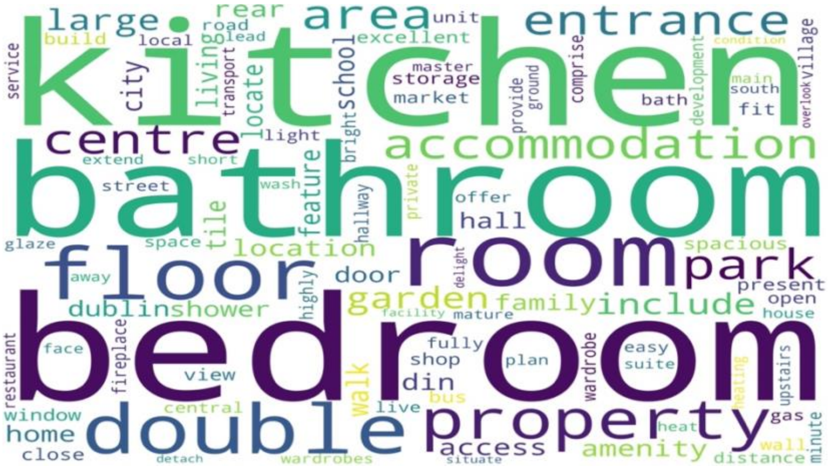
Correlation matrices and variance inflation factors (VIF) were used to identify highly correlated variables and assess whether any variable redundancy exists resulting from information overlap. The correlation matrices and variance inflation factors are shown in APPENDIX C. Analysing solely the numerical values, strong correlation exists between 'surface' and 'beds', this is expected as more bedrooms generally equates to more surface area. The VIF for both variables also indicate a moderate correlation between the two variables. It was decided that neither variable would be removed as their correlations with other variables were not strong and both provide valuable information about a property.

When all variables, categorical and numerical, are analysed we see a strong correlation between 'area' and both geographical coordinate values 'latitude' and 'longitude', again this is an expected relationship but not one that is seen as problematic to our either of our models.

Unstructured Data

Unstructured data is qualitative data that does not follow a set format such as text, images, audio etc. as such it can't be processed by conventional tools and models (IBM Cloud Education, 2021). The unstructured data in the dataset is free text entered by real estate agents describing each property and its features.

Natural language processing (NLP) tools were used to pre-process and analyse the unstructured text data. The text data is by nature noisy and contains many words that won't provide any useful information to the models, the aim of natural language processing in this project is to eliminate any noise and extract informative features that will boost model accuracy.



Points of interest are the frequency of different words and word types in property descriptions. The most frequently used words are shown in the word cloud above. Different word types may also be informative to the models, nouns for example may indicate the presence of a property feature, a garage for example. Meanwhile different adjectives may be used to describe properties of a certain price bracket or location.

Furthermore, the number of nouns or adjectives could be informative to the models, more nouns may mean more features, more adjectives may indicate a nicer property or location. The average length of words was also analysed, longer adjectives could potentially be used in more expensive property descriptions, for example 'pristine' condition in contrast with 'good' condition. Analysis of these features can be found in the tables below.

	Mean Count	Mean Length
All Words	397.49	6.40
Nouns	157.72	6.37
Adjectives	42.14	6.31

	Top 25% (Price)	Bottom 25% (Price)
Mean Noun Count	223.73	113.49
Mean Adjective Count	62.29	28.62
Mean Noun Length	6.03	6.21
Mean Adjective Length	6.29	6.09

Outlier detection and removal

Outliers were detected firstly by visualising the distribution of each numerical variable to identify any skewing. These distributions can be found in APPENDIX D. Positive skew was identified in five of the six numerical variables indicating the presence of 'high outliers' that are greater than most other observations in the dataset. Negative skew is identified in the final variable indicating 'low outliers' that fall below the level of other observations.

Noting the skewed nature of the variables, it was decided that in the absence of normal distributions, using the inter-quartile range (IQR) would be a more suitable method than Z-scores to numerically identify outliers and remove them from the dataset. The formulae below were used to identify both types of outliers, where Q1 is the first quartile and Q3 is the third.

$$Low\ Outliers = Q1 - 1.5 \times IQR$$

$$High\ Outliers = Q3 - 1.5 \times IQR$$

The number of outliers for each variable is shown in the table below. These were removed from the dataset leaving a final number of 2,458 observations.

Variable	Outliers
bathrooms	25
beds	17
latitude	168
longitude	5
price	221
surface	203
Total	**465**

4.3) Data Pre-processing

Before training the models with the dataset, the data needs to be pre-processed. This is done in two stages, the first stage handles the structured data, the second the unstructured data.

Structured Data

After cleaning the data by removing any outliers or sparsely populated observations or variables, the data is sent for pre-processing. Pre-processing ensures that the data is in a format that is ingestible to the models. This involves imputing any remaining missing values and transforming and encoding the data where needed.

The remaining missing numeric values found in the 'price' and 'surface' variables were imputed to have a value of 0. This may seem counterintuitive, as a property can't have a surface area of 0 square metres and generally don't sell for free, however modern machine learning algorithms such as XGBoost recognise these 0 values as missing and treat them accordingly.

The missing categorical values ('ber_classification') are handled in the labelling process. Many regression algorithms don't accept categorical values in their original form as such they must be encoded. Two common methods exist, one-hot encoding and label encoding. One-hot encoding decomposes categorical variables into binary indicator variables or dummy variables. Using one-hot encoding would increase the dimensionality of the dataset by 171 sparsely populated variables and as such it was decided that labelling would be a more appropriate solution. Label encoding assigns the unique values of each variable an integer value. This does not increase the dimensionality of the dataset.

Finally, to improve predictive accuracy for the price predictive model a log transform is made to the independent variable 'price'. This is done as the distribution of 'price' is positively skewed, by using a log-transform this distribution (see APPENDIX E) more closely resembles a normal distribution, additionally the variance of the independent variable is decreased this allows for improved predictive accuracy.

Unstructured Data

The 'description_block' and 'features' columns are combined to create a consolidated unstructured column, named 'desc_feat', containing all the property's descriptive information. This column is then cleaned. All words are converted to lower case and any special characters, punctuation, and numbers are removed. Any 'stop words' are then removed from the text, stop words are commonly used words that don't add any information to the text, for example 'the', 'a', 'I' etc. Any remaining words with less than 3 characters are removed, these were found to be typos in most cases. Finally, any additional whitespace is trimmed from the text.

The cleaned data is then tokenised. Tokenisation is the process of splitting a sentence into individual words and storing them in an array.

"this is a new property" becomes **["this", 'is", "a", "new", "property"]**.

The tokenised text then goes through a process called "Part of Speech (POS) Tagging". This assigns a tag to each tokenised word, the tag indicates the "part of speech" (noun, verb, adjective, adverb, etc.) of the tokenised word. Using the POS tags, new columns containing only words of certain POS can be created. For the purposes of this project new columns are created only for nouns and adjectives. Additional numerical columns for the count and average length of nouns and adjectives are created.

Each of the unstructured columns ('desc_feat', 'nouns', and 'adjectives') is then lemmatised. Lemmatisation is the process of "removing inflectional endings only and returning the base or dictionary form of a word" (Manning, et al., 2008). By removing inflections, "bats" becomes "bat" and "boy's" or "boys' " become "boy" etc. Lemmatisation will also convert more complex cases, for example past tense "bought" will become "buy".

Lemmatisation is helpful in reducing the dimensionality of 'bag of words' matrices, these are binary dummy matrices that indicate the presence of a word or phrase in a text. By removing inflections, lemmatisation ensures there won't be separate matrix columns for 'doors' and 'door' for example. Bag of words matrices were made for each lemmatised unstructured column and are used to create informative features to train the models and improve their accuracy.

To extract the most informative features analysis on word frequency is conducted. As seen in the word cloud earlier the most frequently used words ('kitchen, 'bedroom', 'bathroom' etc.) are not very informative, most if not every property will have these features. The least used words are also uninformative as they are often typos or property specific. As such words with moderate frequencies are seen as potentially the most informative.

The dataset was subset into 3 groups, the top 25% most expensive properties, the middle 50%, and the bottom 25%. Words that featured in 30% to 60% of these properties' descriptions were extracted as keywords and their corresponding columns in the bag of words matrices were added to the dataset. Any duplicated columns were removed.

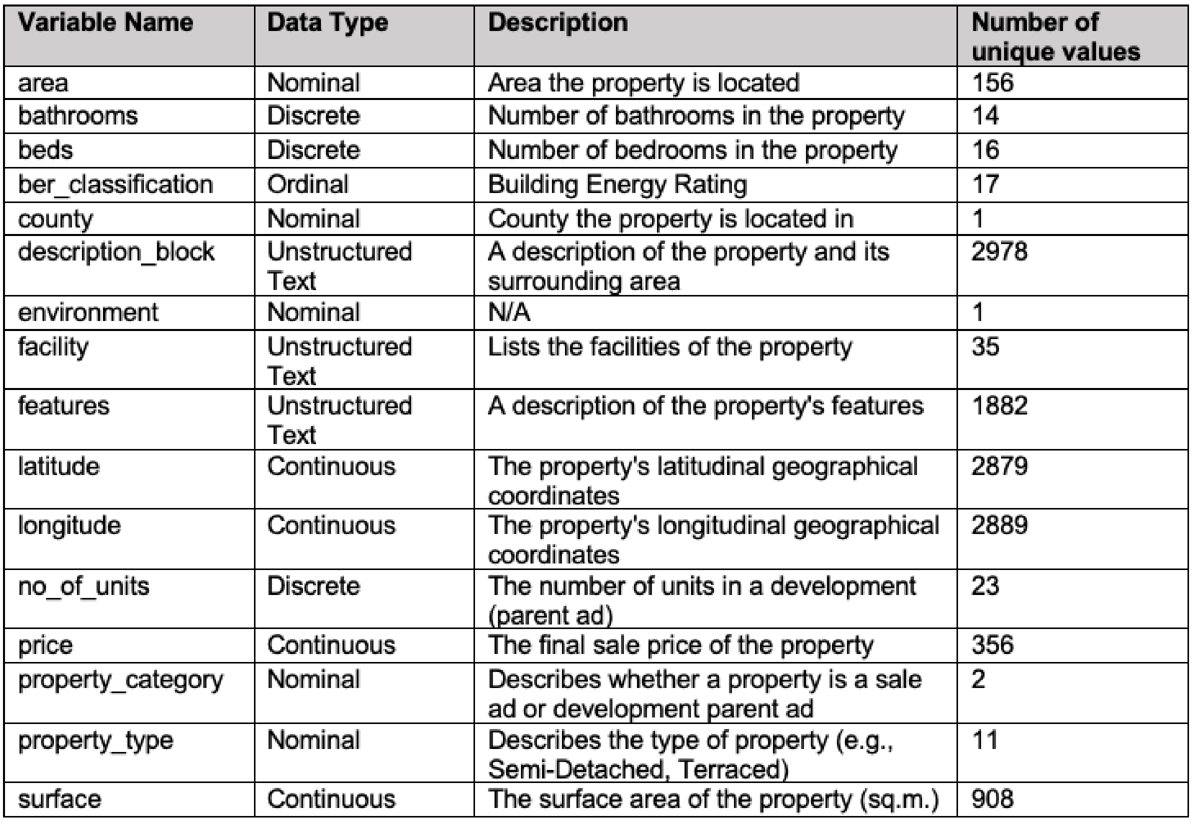
A limitation to the bag of words approach is that it disregards the context or order of the words, for nouns and adjectives this is not a major issue, however when taking the whole descriptions this can cause some very noisy data to enter the bag of words. One solution to this is to use bigrams, bigrams are consecutively written words in the description, this helps add to the context and preserve the order of words in the dataset. Bigrams are used in the bag of words matrix for the lemmatised description.

Final Dataset Description

The final dataset consists of 2,458 observations and 122 variables. 3 categorical variables, 13 numerical variables, and 106 binary dummy variables that indicate the presence of a word in the property's description or features. The dummy variables consist of 69 nouns, 15 adjectives and 22 lemmatised bigrams.

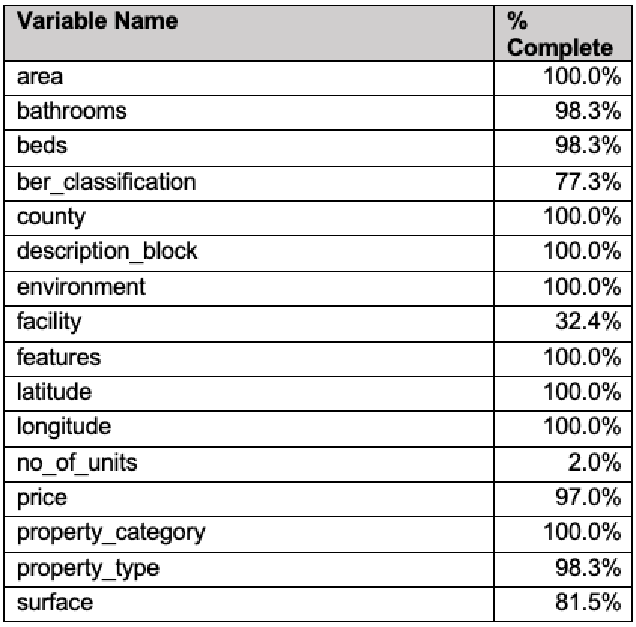
APPENDICES

Appendix A: Description of variables in the dataset.



Variable Name	Data Type	Description	Number of unique values
area	Nominal	Area the property is located	156
bathrooms	Discrete	Number of bathrooms in the property	14
beds	Discrete	Number of bedrooms in the property	16
ber_classification	Ordinal	Building Energy Rating	17
county	Nominal	County the property is located in	1
description_block	Unstructured Text	A description of the property and its surrounding area	2978
environment	Nominal	N/A	1
facility	Unstructured Text	Lists the facilities of the property	35
features	Unstructured Text	A description of the property's features	1882
latitude	Continuous	The property's latitudinal geographical coordinates	2879
longitude	Continuous	The property's longitudinal geographical coordinates	2889
no_of_units	Discrete	The number of units in a development (parent ad)	23
price	Continuous	The final sale price of the property	356
property_category	Nominal	Describes whether a property is a sale ad or development parent ad	2
property_type	Nominal	Describes the type of property (e.g., Semi-Detached, Terraced)	11
surface	Continuous	The surface area of the property (sq.m.)	908

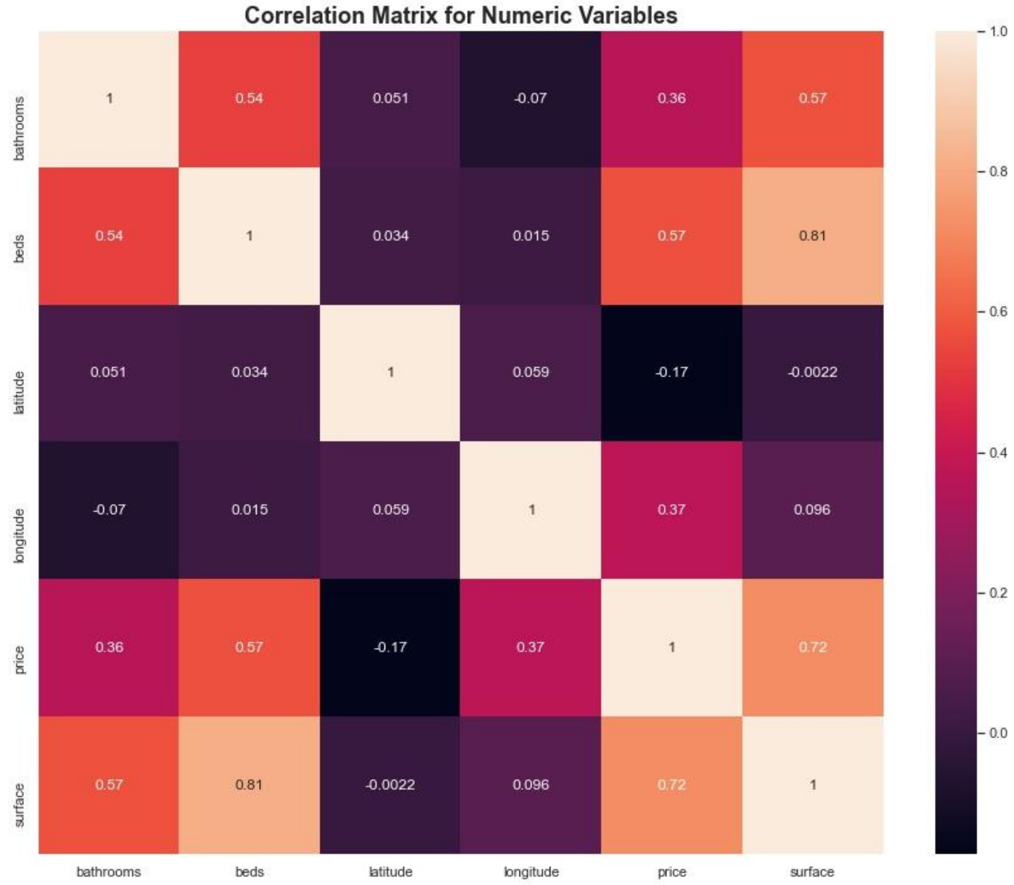
Appendix B: Completeness of each variable.



Variable Name	% Complete
area	100.0%
bathrooms	98.3%
beds	98.3%
ber_classification	77.3%
county	100.0%
description_block	100.0%
environment	100.0%
facility	32.4%
features	100.0%
latitude	100.0%
longitude	100.0%
no_of_units	2.0%
price	97.0%
property_category	100.0%
property_type	98.3%
surface	81.5%

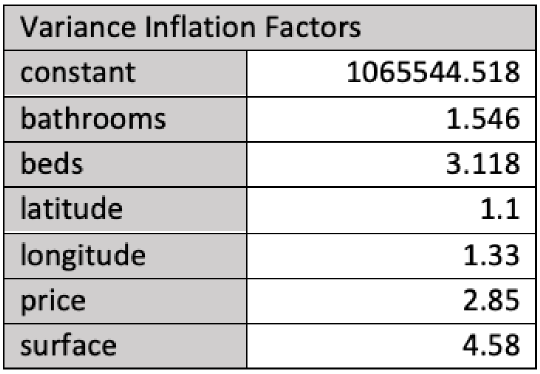
Appendix C:

C.1.: Correlation Matrix for Numeric Variables.



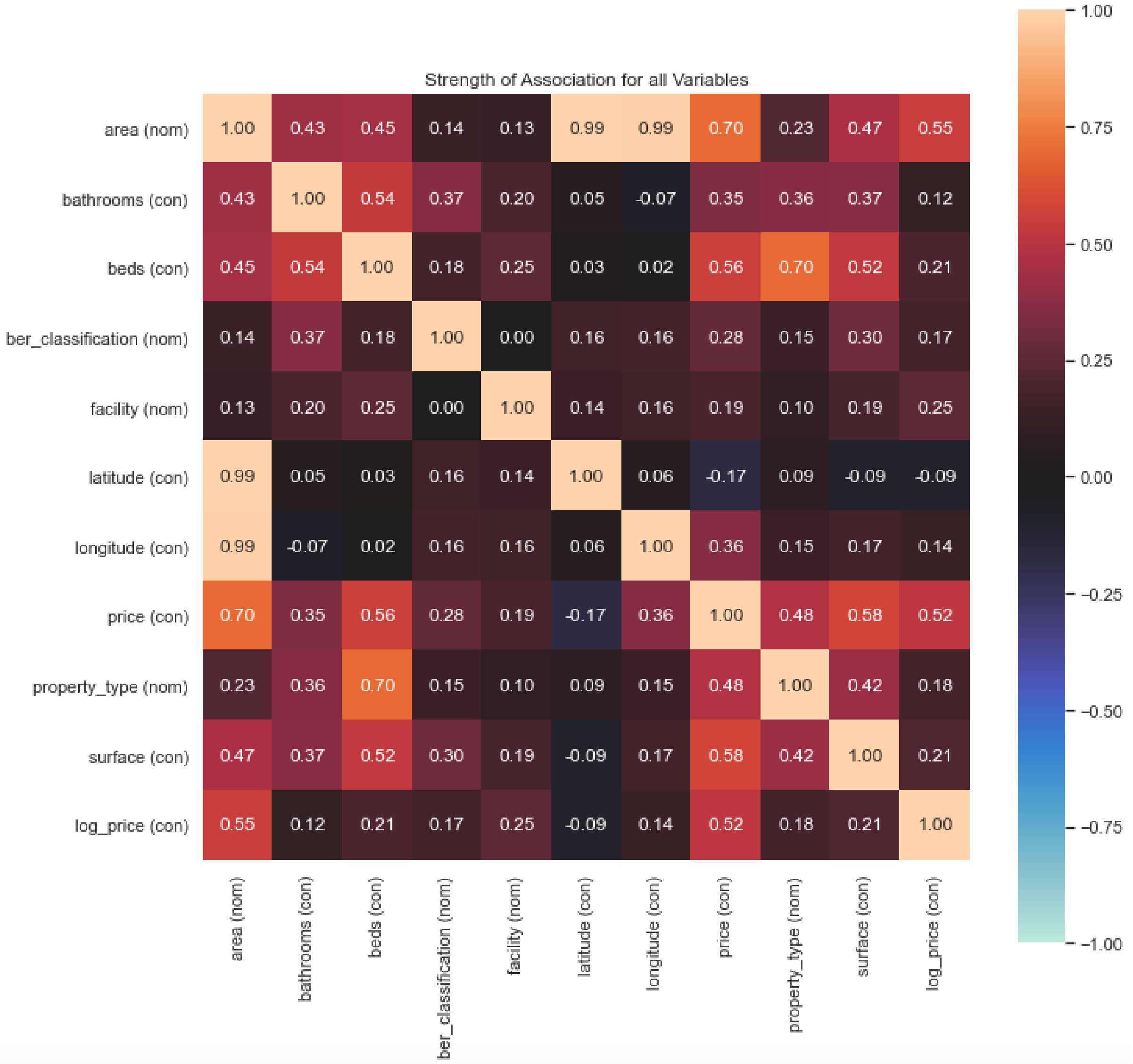
	bathrooms	beds	latitude	longitude	price	surface
bathrooms	1	0.54	0.051	-0.07	0.36	0.57
beds	0.54	1	0.034	0.015	0.57	0.81
latitude	0.051	0.034	1	0.059	-0.17	-0.0022
longitude	-0.07	0.015	0.059	1	0.37	0.096
price	0.36	0.57	-0.17	0.37	1	0.72
surface	0.57	0.81	-0.0022	0.096	0.72	1

C.2.: Variance Inflation Factors for Numeric Variables.



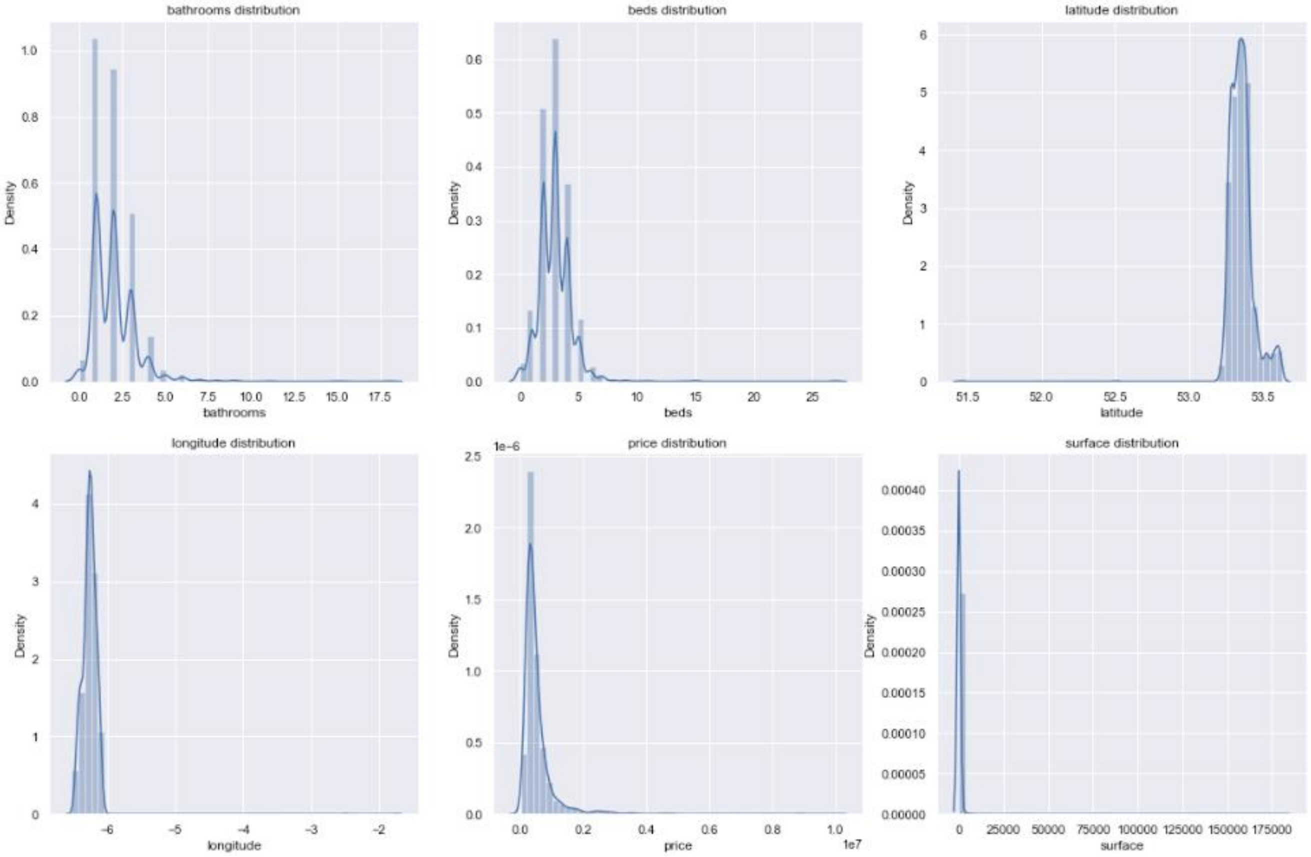
Variance Inflation Factors	
constant	1065544.518
bathrooms	1.546
beds	3.118
latitude	1.1
longitude	1.33
price	2.85
surface	4.58

C.3.: Strength of Association Matrix for All Variables.

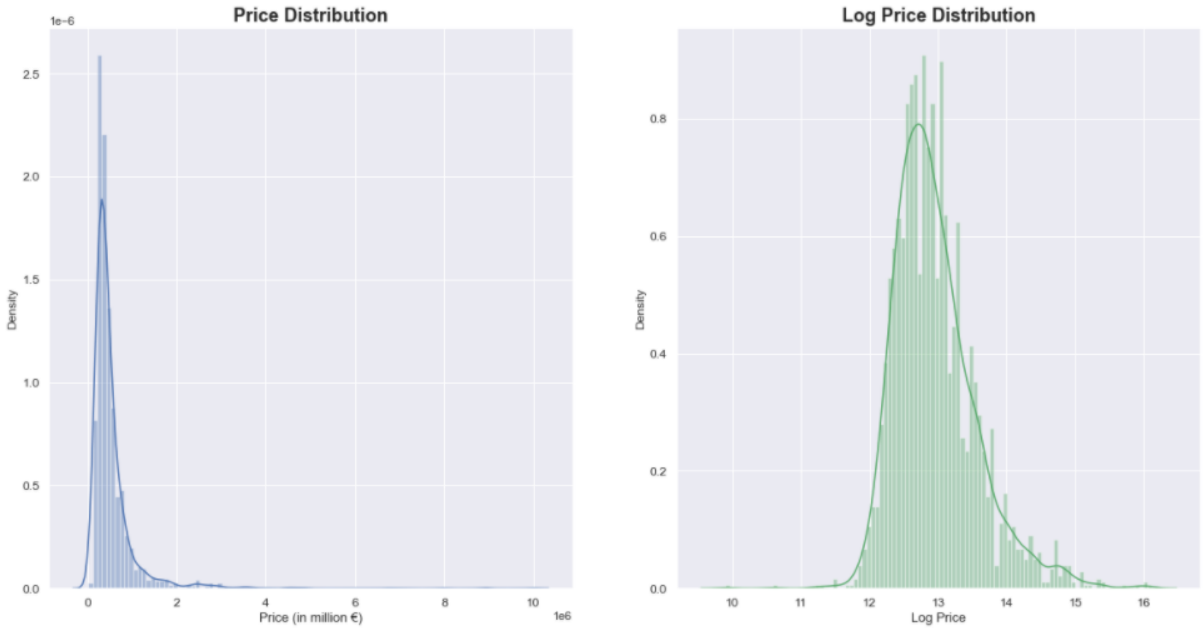


	area (nom)	bathrooms (con)	beds (con)	ber_classification (nom)	facility (nom)	latitude (con)	longitude (con)	price (con)	property_type (nom)	surface (con)	log_price (con)
area (nom)	1.00	0.43	0.45	0.14	0.13	0.99	0.99	0.70	0.23	0.47	0.55
bathrooms (con)	0.43	1.00	0.54	0.37	0.20	0.05	-0.07	0.35	0.36	0.37	0.12
beds (con)	0.45	0.54	1.00	0.18	0.25	0.03	0.02	0.56	0.70	0.52	0.21
ber_classification (nom)	0.14	0.37	0.18	1.00	0.00	0.16	0.16	0.28	0.15	0.30	0.17
facility (nom)	0.13	0.20	0.25	0.00	1.00	0.14	0.16	0.19	0.10	0.19	0.25
latitude (con)	0.99	0.05	0.03	0.16	0.14	1.00	0.06	-0.17	0.09	-0.09	-0.09
longitude (con)	0.99	-0.07	0.02	0.16	0.16	0.06	1.00	0.36	0.15	0.17	0.14
price (con)	0.70	0.35	0.56	0.28	0.19	-0.17	0.36	1.00	0.48	0.58	0.52
property_type (nom)	0.23	0.36	0.70	0.15	0.10	0.09	0.15	0.48	1.00	0.42	0.18
surface (con)	0.47	0.37	0.52	0.30	0.19	-0.09	0.17	0.58	0.42	1.00	0.21
log_price (con)	0.55	0.12	0.21	0.17	0.25	-0.09	0.14	0.52	0.18	0.21	1.00

Appendix D: Distributions of numeric variables



Appendix E: Price Distribution vs. Log Price Distribution


Price Distribution
1e−6
Density
1e6
Price (in million €)
Log Price Distribution
Density
Log Price