What socio-economic and geographical factors influence house prices in Ireland

Marcos Cavalcante Barboza*[1]; Ana Julia Righetto[2]

[1] Software Developer. 26 Church View, BT18 9DP, Holywood, County Down, Northern Ireland – United Kingdom
[2] Alvaz. Head in Statistics & Customer Experience. Av. Ayrton Senna da Silva, 600, Sala 602 – Gleba Fazenda Palhano; 86030440 Londrina, Paraná, Brasil
*author's email: marcos.bcc2011@gmail.com

What socio-economic and geographical factors influence house prices in Ireland

Abstract

This study aimed at examining house prices in Ireland and understand whether socioeconomic variables influence house prices or those are mostly explained by the characteristics of a house along with its spatial location. Exploratory data analysis was applied to understand the spatial relationship between prices in different counties. Regression trees, Random Forest and Extreme Gradient Boosting were used to create regression models to predict house prices on a dataset containing socioeconomic information along with house's characteristics and their location. The study concluded that house characteristics such as size and price per square meter have the most influence on the valuation of a house, however, algorithms that based solely on those characteristics (regression trees) do not perform as well as ensemble algorithms which explore a much wider combination of features.

Keywords: Ireland houses dataset; random forest; socioeconomic; exploratory data analysis; extreme gradient boosting; regression trees; general linear models

Introduction

Housing is publicly considered as the most important social problem in Ireland [3], for which reason it is a topic of focus by policymakers in government and a subject of utmost relevance by the main political parties in the next general election in 2025 [2] [3].

In Ireland, house prices have been on a steady rise for the last 10 years [1]. Amongst countries in the OECD, Ireland saw a 474% price change between 1995 and 2008, the sharpest increase amongst countries in the OECD, reveals a comparative study [4].

Young people and those on low incomes have found it increasingly difficult to be able to afford the current cost of a house in the capital and major cities in the country [14] [3.1] [4.1].

Housing prices have also been a key factor in the calculation of rent prices in Ireland. Studies have shown that during periods of strong economic performance, rent prices reflected the increase on the cost of houses, conversely, in periods of slow or negative economic performance, rents tend to follow the decrease on the cost of houses [4]. Consequently, rental costs play an important role in the cost of living, resulting in even more pressure during times of high inflation [5] [6].

Recent figures show that house prices have been increasing to levels only seen during the 2008 economic crisis [1]. According to the Central Statistics Office (CSO), property prices have risen nationally by 123% since its lowest value in 2013 [7].

Studies have been carried out to investigate what variables influenced and contributed to the recent booms in house prices, so that potential solutions to the problem could be evaluated in Ireland. The analysis of housing prices has been tackled from several different perspectives in the literature. The three main areas of focus are:

- **Macro-economic shocks**: Those macro-economic shocks can be defined as: low interest rates, GDP growth, credit, and exchange rate [8] [9].
- **Regional Factors:** Supply and Demand, scarcity of labour, rent legislation, and household income [2] [10] [11].
- **Geospatial variables:** address, proximity with other cities and towns may help explain house prices in a neighbourhood, city, or country [12] [13].

With regards to the studies that used geospatial variables [12][13], the property valuation was analysed from a local perspective where only county Dublin was considered in [12] and census data from 2016 was used in the geospatial analysis in [13].

The focus of this study, however, is to evaluate whether the presence of socio-economic variables influence on the prediction of house prices. The variables of focus are schools, universities, hospitals, public transport, and garda stations. These socio-economic variables are analysed in a geospatial context by computing the count on the number of schools, garda stations, hospital and public transport within a given radius in km from each house in the dataset, lastly, each of those computations, for every socioeconomic factor, is stored as a new variable against the property from which it got computed.

The analysis also considers the characteristics about the house, such as: number of bedrooms, bathrooms, property size, and type. Having those variables about the property is necessary so that a comparison between a model that takes them in consideration alone is more efficient or not than a model that also contains socio-economic variables.

The goal of this research is, therefore, to understand whether the presence of socioeconomic variables in a housing dataset contribute towards the prediction of the valuation of a house. Some of the questions this study aims at answering are:

- Does the proximity with *hospitals* help in the house valuation model?
- Does the closeness with *schools and universities* contribute to the prediction model?
- Does having *garda stations* near a property affect a house price?
- Does being close to bus stops and train stations influence on the house price?
- Do the house characteristics along with its spatial location are enough to explain its price?

Material and Methods

This research examined the relationship between housing prices in the Republic of Ireland and various socioeconomic factors. Multiple datasets from diverse sources were gathered to provide insights into the characteristics of homes and socioeconomic features. It is worth noting that this study specifically focused on housing prices in the Republic of Ireland, and therefore excluded data from Northern Ireland. Despite being located on the same island and having spatial influence on each other, Northern Ireland belongs to a different jurisdiction, which may impact housing prices due to factors such as currency and differences in interest rates.

The Python programming language along with the Google Geocoordinates API were used in this project for data collection. The R language was used during the data cleaning/preparation, exploration, and model training tasks.

Before covering the datasets below, it is important to highlight that the terms geo coordinates and latitude-longitude are used interchangeably in the text, also when omitted the projection system used is World Geodetic System 1984 (WGS84) format. Furthermore, the only variables used from the socioeconomic datasets were latitude and longitude, no other piece of information was used because the aim is to compute the total number of points of interest within a given radius of the house for each type of socioeconomic variables, this will be discussed in length further. Finally, all datasets were stored in CSV format as that was the format deemed most manageable for dealing with such data.

Datasets

Housing Price dataset

The housing prices dataset used in this analysis was obtained from the biggest property website in Ireland, Daft.ie, via its API by which the company exposes their properties on sale and for rental [18]. A Python script, that uses the property website API, was created to collect the data solely about the properties on sale. The resulting dataset is very rich, there are over 100,327 thousand residential properties with 18 features, it contains detailed information about the house, for example - size, property type, geocoordinates, address, price, energy rating, number of bedrooms and bathrooms.

The data was collected in August 2022 and, it is subject to change as those were the properties on sale in the market at that time. Nevertheless, the original dataset used in this research can be found in the code repository provided in the Appendix section.

Public Transport datasets

Two datasets were collected regarding the public transport in Ireland, one dataset holds information about all bus stops in the island of Ireland, the bus stops dataset was obtained from the National Transport Authority [22]. The bus stop dataset contains over 16351 entries and 19 features, and among the features, it includes each bus stop latitude and longitude in the WGS84 Coordinate System.

The second dataset holds data with the location of 168 train stations in the island of Ireland and 6 feature variables, it also stores the latitude and longitude for each entry. The dataset was obtained from the Irish Rail API [23].

For the two public transports dataset, bus stops and train stations in Northern Ireland were removed as mentioned before because the country belongs to a different jurisdiction.

Education datasets

Two datasets containing information about teaching institutions in Ireland were collected. The first dataset contained data about all the schools in Ireland and was obtained through web scrapping the government directory page on schools [24]. 3594 schools and 7 feature variables were collected. The dataset did not have any variable about the school's geolocation, though the address in plain text was present.

The second dataset containing all the universities in Ireland was manually collected from the Education in Ireland webpage [25]. A total of 22 universities names and their addresses were collected from the government department, Education in Ireland, webpage.

Hospitals dataset

The dataset containing the list of hospitals in Ireland was obtained from the HSE - Health Service Executive via the data.gov.ie portal, which is Ireland's open data portal, and it contains numerous datasets [21].

This dataset contains 38 entries, which represents each hospital in Ireland and the following data is available: hospital name, address, postcode, x and y (coordinates in the Irish projections system).

Garda stations dataset

In this project, all the garda stations in Ireland were collected. The dataset was created by web scrapping [29] the garda stations directory [20]. A total of over 568 garda stations and their geolocation were collected from the webpage and saved in CSV format.

Data Preparation – Socio-economic dataset

The following data preparation tasks were performed on the socio-economic datasets:

- Cleaning: Instances on those datasets with missing data for latitude and longitude were removed, for the bus stops dataset, those marked as inactive got removed.
- Feature Engineering: Hospitals and Education (universities and schools) datasets did not contain latitude and longitude for their observations; therefore, a bespoke Python script was created to query the Google Geocoordinates API [26] to get the geocoordinate of those addresses based on their plain text address and or institution name.
- Further Cleaning: For observations where the query did not find any result, it was decided to remove them. Moreover, any observation located in Northern Ireland got removed as they are not part of this study due to the reasons mentioned previously.

In the end, a total of 126 bus stops and 9 train stations were removed from the public transport datasets. 63 schools got removed from the education datasets. Lastly, only 2 garda station were removed from that dataset.

Data Preparation – Housing prices dataset

The housing dataset contained many duplicated rows, the initial dataset had about 100 thousand observations. The de-duplication process was done by removing rows with the same id and url_link fields as those are deemed to be unique. In the end, the dataset shrank to 9009 thousand observations. There was also data missing on some of the variables collected, such as price, ber_code, size_meters_squared and bathrooms. Overall, the default approach was to simply remove the variable or observation entirely – for some instances i.e., ber_code, there was just too much data missing about 42% and, for other cases, using techniques for inputting data was not considered as it could have added bias to the dataset.

In summary, the steps below were performed to clean the housing dataset:

1. Removal of duplicates: this was done by using the observations id and url_link.
2. Removal of missing values: Observations where the fields size_meters_squared and price were missing got removed as price was the target variable and size_meters_squared was missing on over 26% of the data. Similarly, about 2% of the observations had the variable bathrooms missing and those were removed too.

3. Removal of unnecessary variables: features such as - *url_link*, *id*, *daftShortcode*, *publishDate*, *category* and *propertySize* were dropped from the dataset as they mostly contained information used by Daft.ie to manage those records and for propertySize that was already present in the dataset under size_meters_squared feature.
4. Data type conversion: Appropriate data types were used for the variables in the dataset and some conversions were necessary, for example: from character to numeric or character to factor.
5. Feature Engineering: new features holding the county and neighbourhood were included in the dataset, those were created from the location variable. Additionally, pricePerSqMeter was created from the division between price by size_meters_squared. Lastly, features for each socioeconomic dataset category – hospitals, education, transport, and garda stations were created in the following manner: the count on the number of points of interest under each category was counted within a given radius in km from each house – the distance was calculated using the Haversine method – distHaversine function in R [30], which computes the straight line distance between two geo coordinates taking into consideration the radius of the Earth. For each category, the value for the radius was decided on an empirical manner:
    a. Education institutions and Transport stops: The value of 5 km was chosen as it's within a 45 minutes' walk or 6 minutes' drive within a city and therefore considered to be a close distance.
    b. Hospitals and Garda Stations: The value of 10 km was picked as it is within a short driving distance within a city - 12 minutes.
6. Outliers' removal: Some outliers were removed as those clearly represented error in the data input, for example, 10 houses were under 17 square meters of size, when those were looked up on Google maps, they clearly had far bigger size. In other instances, the presence of outliers was removed after training the machine learning models due to the high error metric values they were causing. To do the latter, a modified version of the z-scores function was used, where values with a z-score over 4 standard-deviations of the mean were removed, 4 standard deviation was used to tune down the sensitivity of the algorithm to outliers and thus, not remove as many observations. This method was applied only against the price and size variables.

Table 1 depicts the dataset produced after the cleaning and preparation stages, the final number of observations is 9009 and the dataset was stored in R data frame format.

Table 1: Description of the dataset created at the end of the cleaning stage.

Variable Name	Description	Data Type
address	A long form of the property address	Character
bathrooms	The number of bathrooms in this property	numeric
bedrooms	The number of bedrooms in this property	numeric
berRating	BER Rating of this property, i.e. A1, B2	factor
county	The county the property is in. Examples: “Co. Wicklow”, “Co. Kerry”	factor
latitude	The latitude of the property	numeric
location	A short form of the property address area, i.e., Dublin 1, Co. Dublin	character
longitude	The longitude of the property	numeric
price	The price of the property, in euro €	numeric
propertyType	The type of the property, i.e., Apartment, End of Terrace, Semi-D, Terrace	factor
size	The size of the property in square meters	numeric
pricePerSqMeter	Ratio between price and size of a given house	numeric
nearestHospitals	Count on the number of hospitals within a 10km radius of the house	numeric
nearestGardaStations	Count on the number of garda stations within a 10km radius of the house	numeric
nearestEducationCentres	Count on the number of schools and universities within a 5km radius of the house	numeric
nearestPublicTransports	Count on the number of bus stops and train stations within a 5km radius of the house	numeric

Source: Original data from the research, it can be found in the repository linked in the Appendix

In Table 2, it is possible to see a summary of the number of observations across all the datasets.

Table 2: Count on the number of observations for each dataset after the preparation stage.

Dataset	Total Observations
Housing Prices	8876
Bus stops - transport	16225
Train stations – transport	159

Schools – education	3531
Universities and Colleges – education	22
Hospitals - health	38
Garda stations - security	566

Source: Original data from the research, it can be found in the repository linked in the Appendix

Exploratory Data Analysis

After performing the data collection and preparation steps, the next stage in the study was to do the exploratory data analysis (EDA). In this stage, the focus was to get insights into the variable's relationships, distributions, and patterns.

To do the EDA, a variety of methods was used - summary statistics, bivariate plots using ggplot2 [31], spatial visualisations with tmap [32] and, other statistical methods such as Pearson correlation, Analysis of Variance (ANOVA) and box-plots were used to explore the dataset.

The summary statistics information used were – mean, median, percentiles and counts for each variable. The correlation between numerical variables and the target variable price was investigated using the Pearson correlation, whilst the relationship between categorical variables and price was examined using ANOVA (Fisher, R.A - 1925) as well as the *aov* function from the *stats* package was used in R to compute it.

Bivariate plots were created between each numerical variable and price to help understand some questions regarding the relationship of those variables. Density plots and histograms were largely used to learn about the variable's distribution in the dataset. Finally, spatial visualisation was applied to understand the connection between some of the categorical and numerical variables with the target variable price.

Finally, although not entirely part of the EDA process, the categorical variables (factor type) – berRating, county and propertyType were transformed into dummy variables so machine learning algorithms can make better use of them. Furthermore, the variables of type character: location and address were removed from the dataset as those cannot be used by machine learning algorithms when performing regression, townOrNeighbourhood was also removed as it has a very high cardinality and did not aid the decision process of the algorithms, though those variables proved to be used to gain a good understanding of the dataset.

Model selection and training

The dataset used in this project was randomly split into 3 subsets called – training, validation, and testing datasets. The training dataset contains 70% of all observations, while validation and testing datasets held 15% of all observations each. The training dataset was used for training the ML models and tuning of hyperparameters, while the validation dataset was used to test the ML models on unseen data and to iteratively tweak the hyperparameters. The testing dataset was the most pristine dataset as it was kept for a final evaluation of the model on completely unseen data.

This study evaluated the performance of 4 different machine learning algorithms: Generalized Linear Models (GLMs) with Linear Regression, Regression trees, Random Forest, and Extreme Gradient Boosting (XGB).

To evaluate GLMs models, Adjusted R-squared and R-squared were used as evaluation metrics and because this is a statistical type of machine learning model, only the training and testing datasets were used. And, to evaluate the regression trees and the tree ensemble models, the metric used was the root mean square error (RMSE) and all the tree datasets – training, testing and validation sets were used as those models are considered as stochastic models.

To perform the training of the regression tree and ensemble algorithms, hyperparameter tuning was done via grid search of the hyper parameters and evaluation of the model with the tuned hyperparameters was done in the validation dataset. Furthermore, during the training and validation process of the ML models, a 10-fold cross validation was used to obtain a more thorough evaluation of model’s performance as well as helping to prevent overfitting of the data.

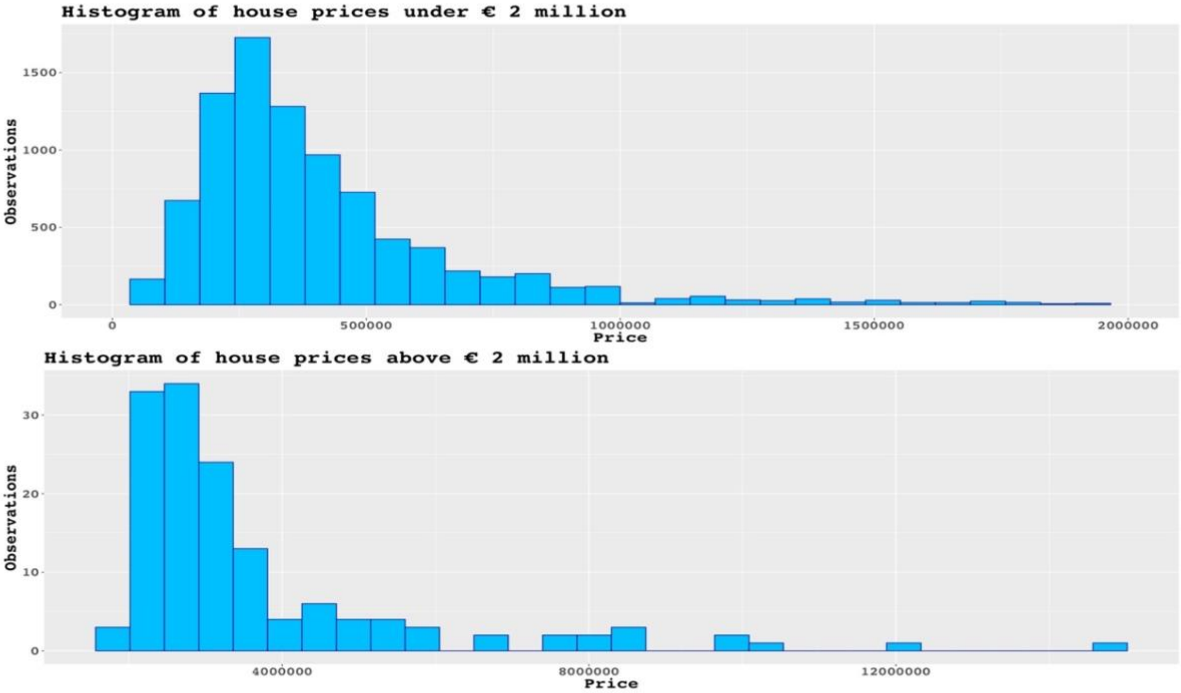
Results and Discussion

Descriptive Analysis and Visualization

Initially, the target variable, price, was carefully examined. Figure 1 presents 2 histograms of the price variable; the price was split into 2 segments at the €2 million euro mark to aid visualisation. What can be inferred from the histogram is that most of the houses, about 75%, are under €500,000.00. This makes sense as few people can afford expensive house, thus making the histogram right skewed.

Moreover, it is also clear that there are some quite expensive houses on the other side of the scale, a few dozen properties valued above € 4 million euro, those can potentially be outliers.

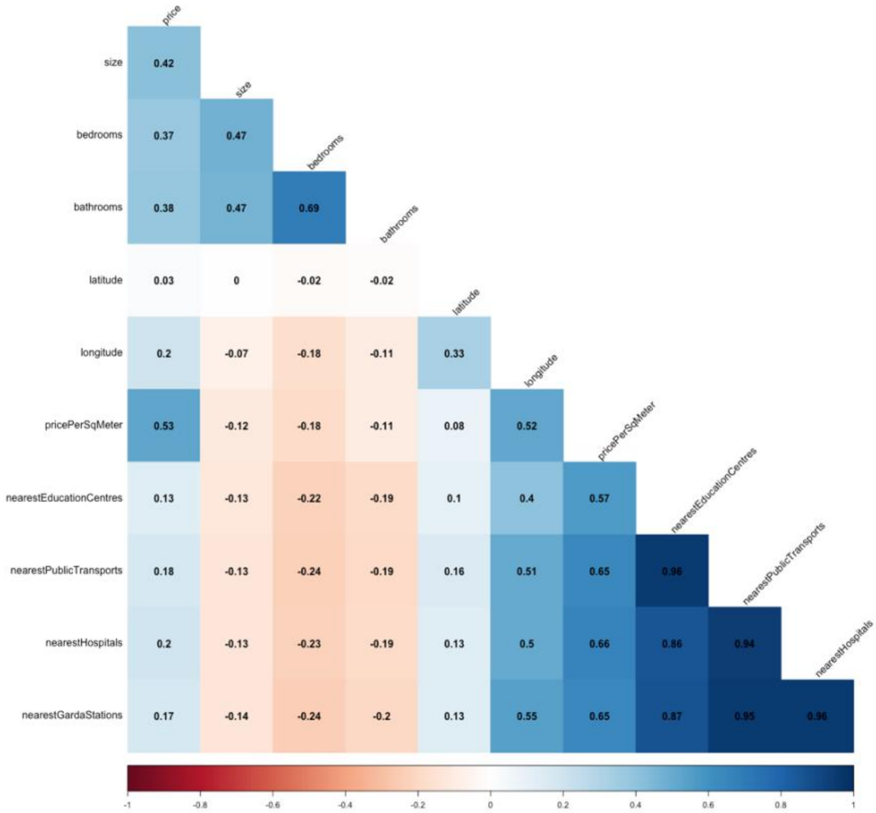
Figure 1: Histograms of the price variable



Source: Original data from the research, it can be found in the repository linked in the Appendix

Next, one way to comprehend the strength or weakness of the relationship among the variables in a dataset is by calculating its correlation matrix. By doing that, it is possible to see how strongly the target variable in the dataset correlates with the other variables and whether there could problems with multicollinearity. Figure 2 shows the correlation matrix computed using the Pearson method and the corrplot package in R.

Figure 2: Correlation matrix of the numeric variables in the dataset



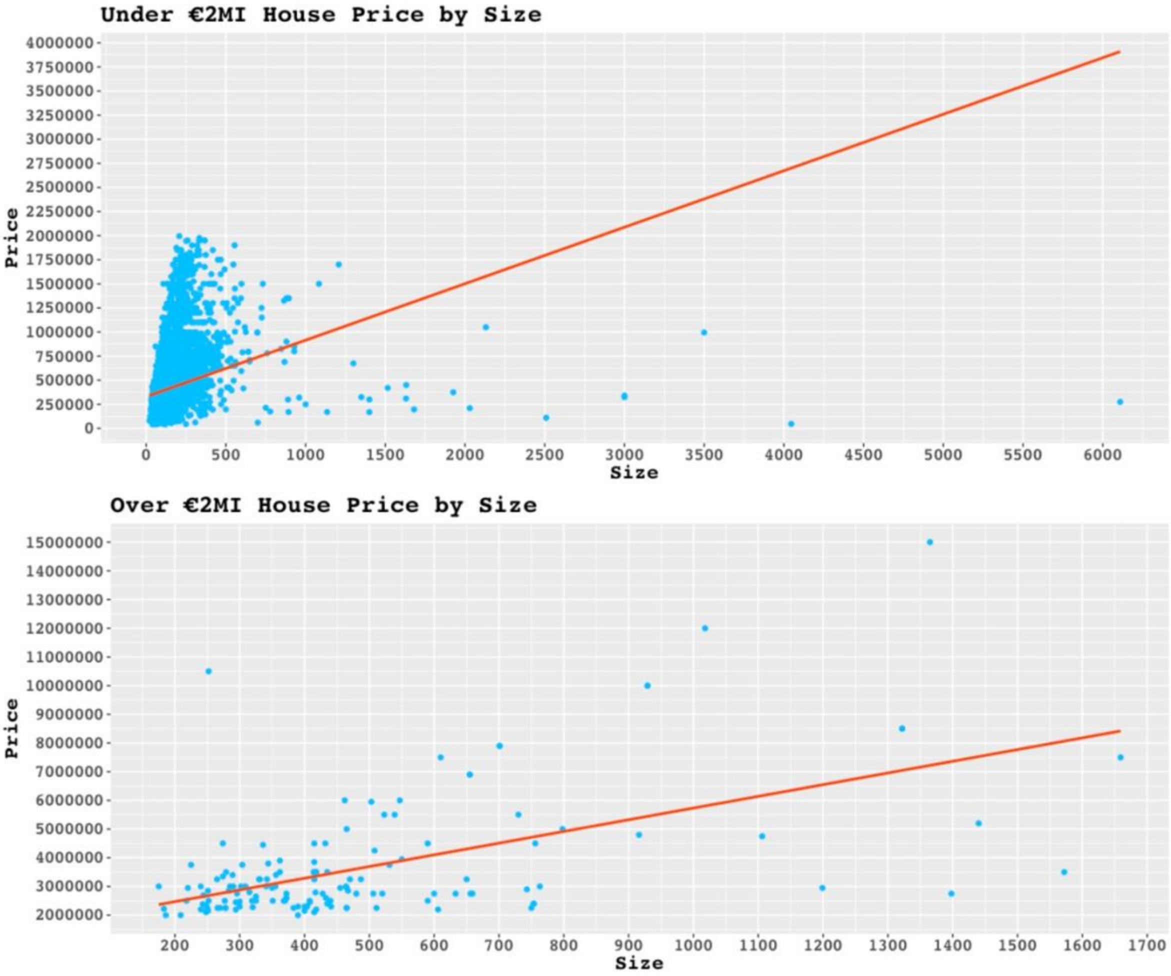
Source: Original data from the research, it can be found in the repository linked in the Appendix

From the correlation matrix above, some important information was inferred:

- There is a **moderate positive correlation** between price and size as well as price and pricePerSqMeter, which indicates that there is some relationship between those variables.
- There is a **weak positive correlation** between price and the variables: bathroom, bedroom, nearestHospitals, nearestGardaStations, nearestEducationCentres and nearestPublicTransports.
- There is also a **strong positive correlation** between bathrooms and bedrooms; and, amongst the socio-economic variables; which might be an early indication of multicollinearity.

The *price* variable was also analysed with other numerical variables in the dataset such as size, bathrooms, and bedrooms. When comparing price per property size, it can be inferred that the bigger the house, the more expensive it gets, however there are some outliers as it can be seen in Figure 3.

Figure 3: Bivariate graphs of houses under and over €2,000,000 by their size

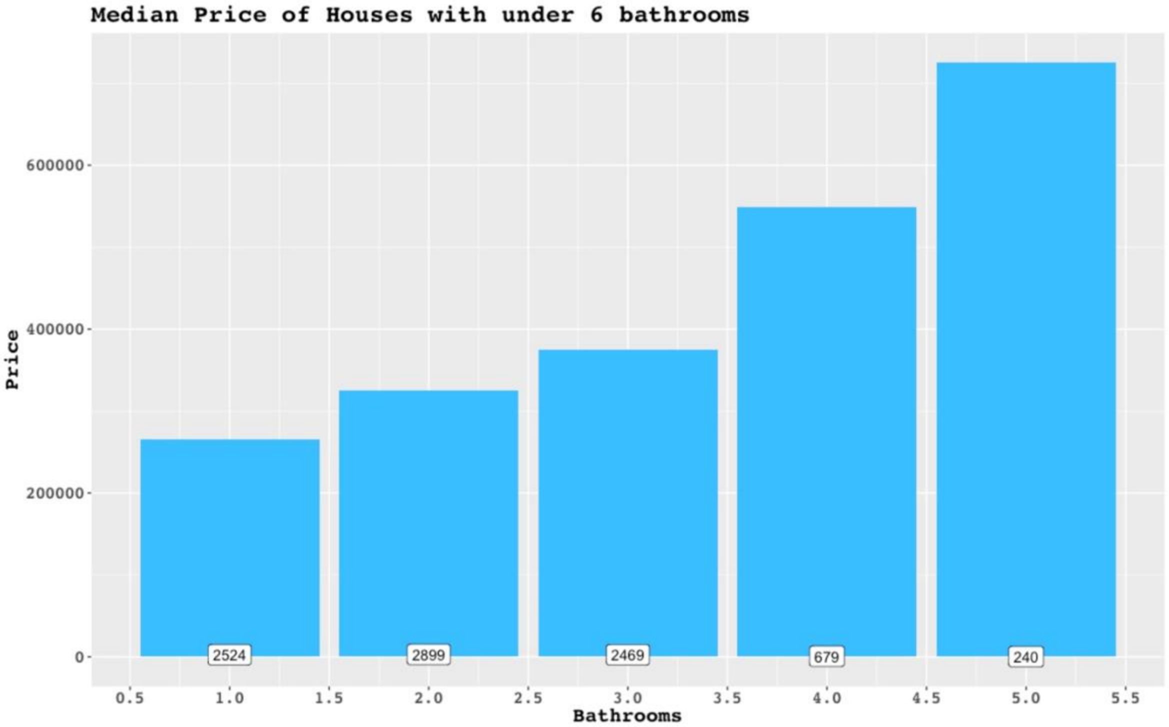


Source: Original data from the research, it can be found in the repository linked in the Appendix

A clear outlier in the first graph is the house with 6000 square meters (Size axis) and price under €500,000 (Price axis). In the second graph, it is much clearer the trend that the bigger the property, the more expensive it gets.

The number of bathrooms and bedrooms were also studied in conjunction with the price variable and what could be seen is that the more bedrooms or bathrooms a house has, the more expensive it gets. Figure 4 shows that about 7900 out of 9000 observations have up to 3 bathrooms and about 97% of the properties have up to 5 bathrooms. Thus, it is possible to figure that the prediction model will have much more information on houses with up to 3 or 5 bathrooms than houses with the number of bathrooms over that threshold.

Figure 4: Median Price of Houses with under 6 bathrooms



Source: Original data from the research, it can be found in the repository linked in the Appendix

Analysis of variables of type *factor*

In this part of the analysis, the relationship between price and the factor variables: *property type, county, townOrNeighbourhood* and *berRating* will be studied. The statistical method used was the **ANOVA (Analysis of Variance)** (Fisher, R.A - 1925) so that the significance of those variables with relation to the price can be understood. The aov function from the stats package was used in R to compute the ANOVA.

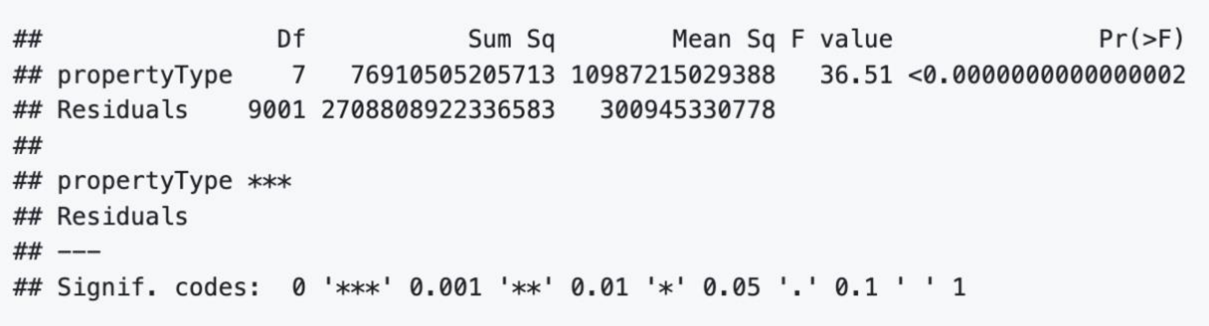
The hypothesis test used across those factors is the following:

- **H0** *(p-value > 0.05 & small f-value)*: It means that the variables do not have a strong relationship and therefore the categorical variable does not contribute to the prediction of the target variable.
- **H1** *(p-value < 0.05 & large f-value)*: It means that the variables have a strong relationship, which indicates that the categorical variable is significant when trying to understand changes in the target variable.

Property Type

For the *propertyType* variable, the ANOVA and f-critical values were calculated. The Figure 5 shows the output of the aov function when computing the linear model between *price* and *propertyType.*

Figure 5: Output of the aov function showing the p-value between *price* and *propertyType.*



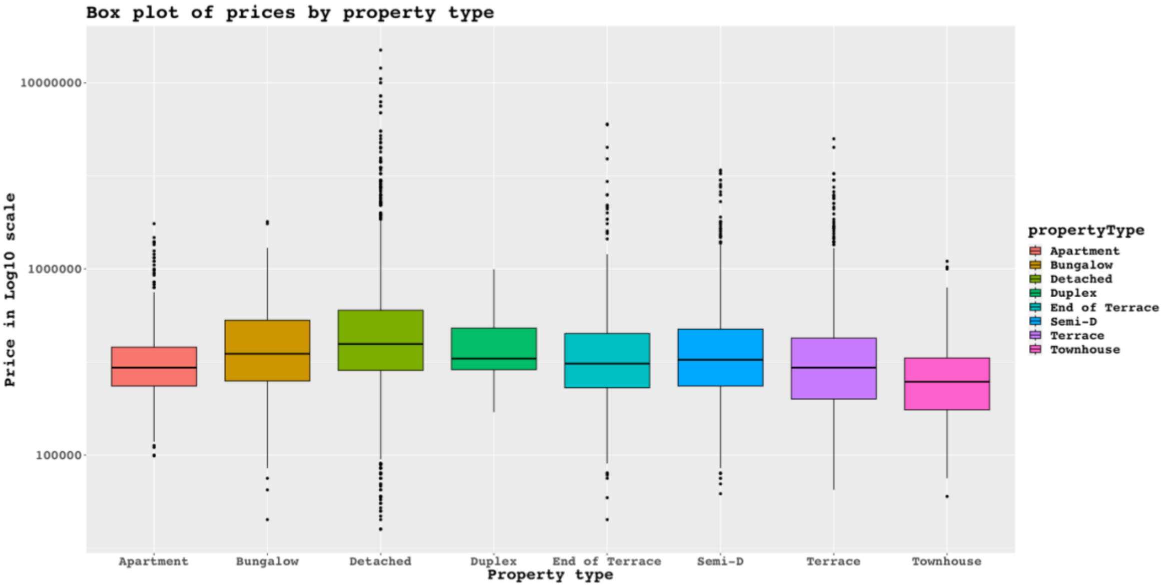
```
##                  Df             Sum Sq         Mean Sq F value               Pr(>F)
## propertyType    7   76910505205713 10987215029388   36.51 <0.0000000000000002
## Residuals    9001 2708808922336583   300945330778
##
## propertyType ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source: Original data from the research, it can be found in the repository linked in the Appendix

The f-critical was also calculated and its value is *0.9659243.* Given those results above, based on the **p-value** that is smaller than *0.05* and the f-critical value being significantly smaller than the f-value calculated from the ANOVA step, the null hypothesis (**H0**), can therefore be **rejected**. By assuming the Alternative Hypothesis (**H1**), it was possible to say that the *propertyType* variable helps explain variations in the house price.

For completeness' sake, the box plot of *prices* by *propertyPrice* was also computed to help visualise the relationship between those variables, for each property type it is possible to see the distribution of the price values and their median. Figure 6 depicts the box plot mentioned.

Figure 6: Box plot between price and property type*.*



Source: Original data from the research, it can be found in the repository linked in the Appendix

Analysis of *county*, *berRating* and *townOrNeighbourhood*

Similar to what was seen for *propertyType*, the alternative hypothesis proved to be true for *county*, *berRating* and *townOrNeighbourhood.* In Table 4, it is possible to see the p-value, f-value and f-critical for each of those variables – propertyType was added to the table for completeness.

Table 4: Summary of the ANOVA analysis across factor variables

Variable	p-value	f-value	f-critical
propertyType	<0.0000000000000002	36.51	0.9659243
county	<0.0000000000000002	24.52	0.9659071
berRating	<0.0000000000000002	34.84	0.9659157
townOrNeighbourhood	<0.0000000000000002	11.31	0.9658226

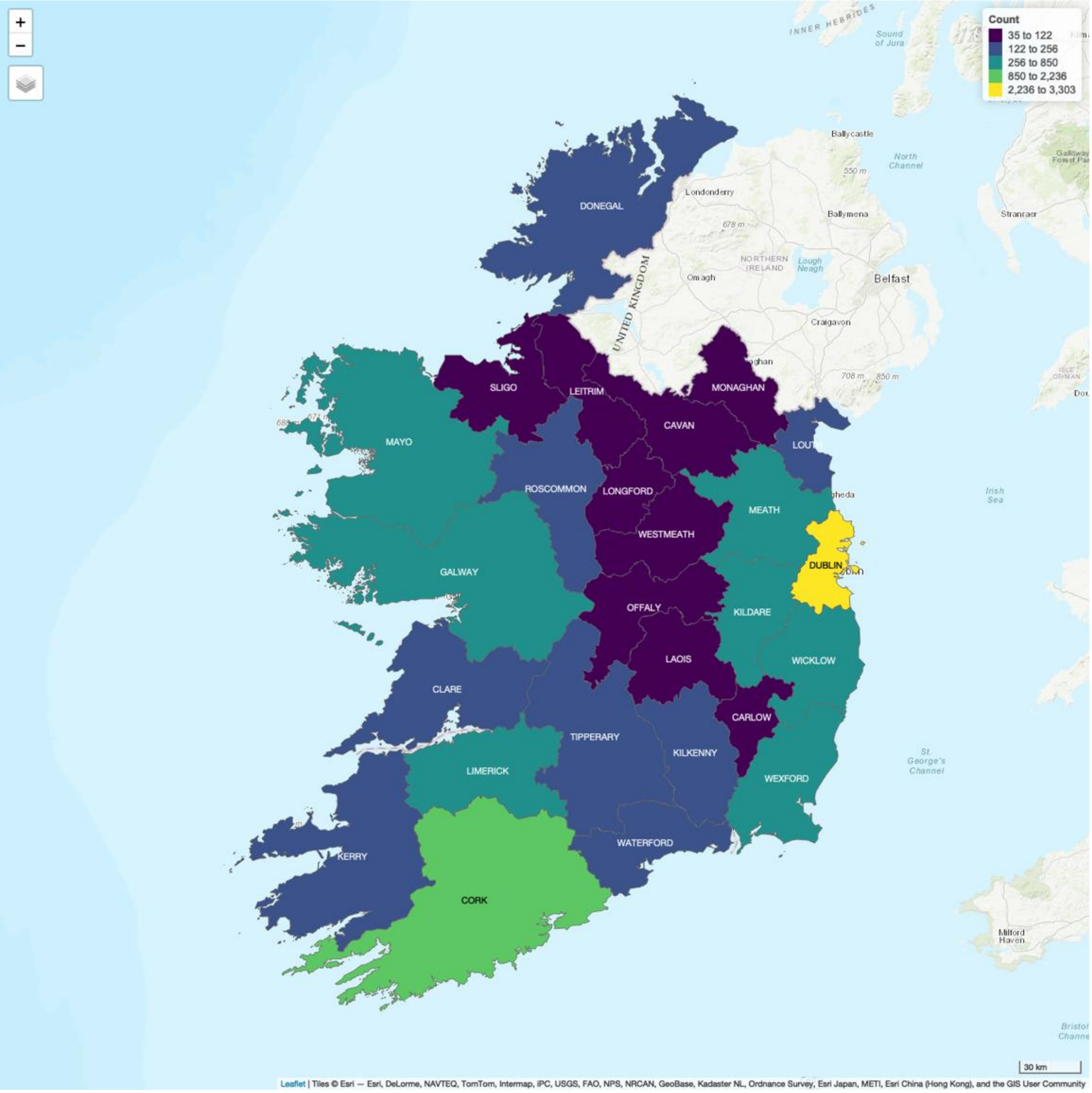
Source: Original data from the research, it can be found in the repository linked in the Appendix

Further evidence of the significance and relevance, such as box, frequency a density plots of those variables can be found in the Appendix.

Spatial visualisation of properties across the country

The housing dataset was also examined from a spatial perspective. This is important because in this study, the aim is also to try to understand how the spatial location of a house can influence its price. Therefore, visualising the distribution of the houses per county is important because it gives an idea about the supply of houses in that area. Figure 7 illustrates how the counties with the main cities – Dublin, Cork and Galway hold most of the house supply, this fact is expected as those areas have the highest population density.

Figure 7: Density spatial graph of the number of houses per county.

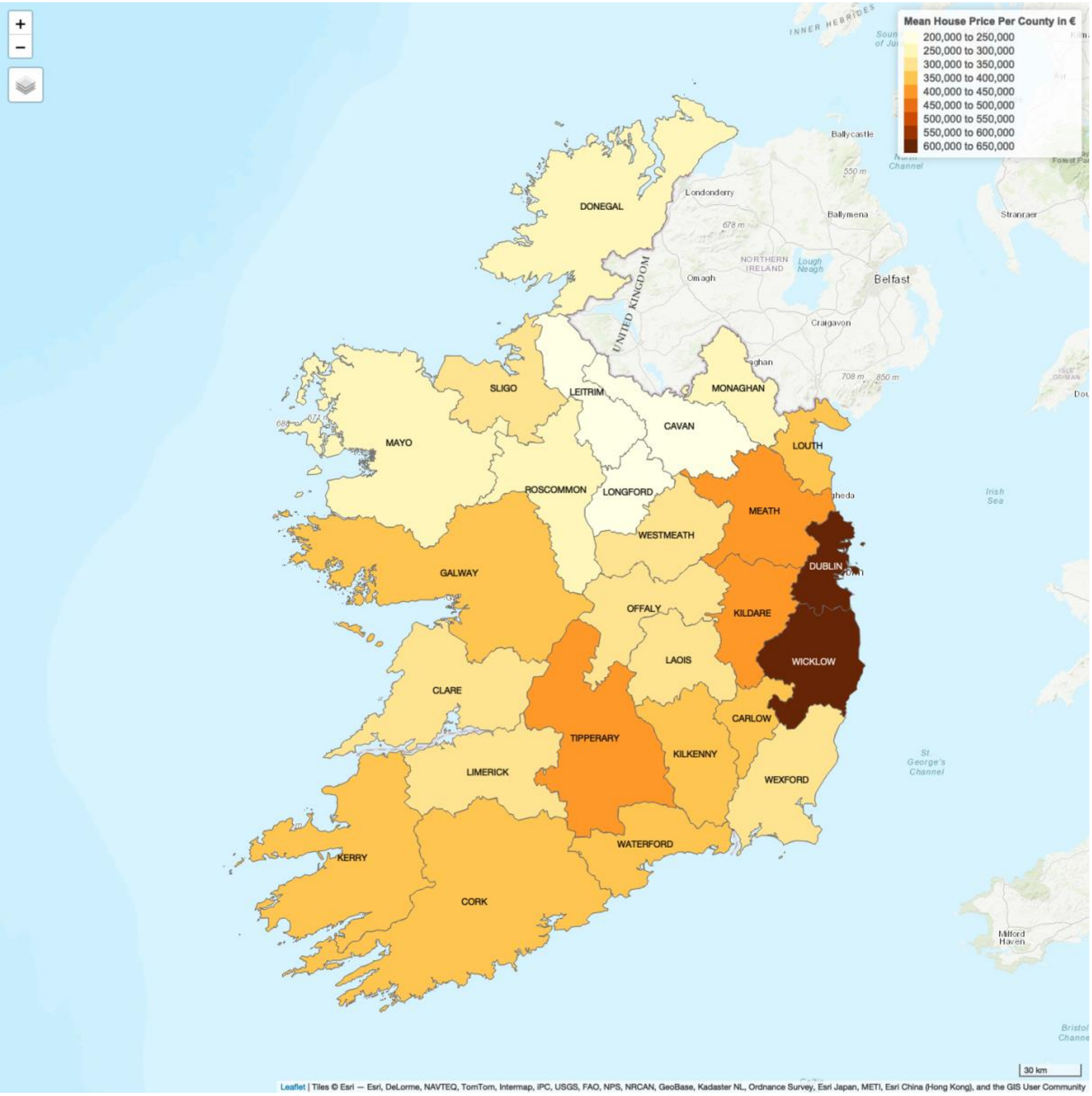


Source: Original data from the research, it can be found in the repository linked in the Appendix

In Figure 7, another point that was possible to infer is around the "Neighbouring effect" where the supply of houses in one county impact the supply in adjacent counties, take Dublin and Cork as an example, the closer a county is to them, the more properties those have on sale. It is also quite clear that the counties located in the middle of the country have a much lesser number of properties on sale, i.e., the two biggest clusters of counties, 35-122 and 122-256 count density, with the shortest supply of houses are predominantly located in the midlands, this fact further contributes to the neighbouring effect assumption.

Moreover, the analysis on the mean house price per county also shows that the neighbouring effect is once more present, the mean price of a county impacts the counties adjacent to it. Figure 8 shows clusters of counties that share same mean price per house, especially the counties in the south of the island – Kerry, Cork, Waterford, Kilkenny, and Carlow. Dublin and Wicklow counties have a quite strong relationship and that seemingly affect counties Kildare and Meath.

Figure 8: Choropleth Map of Mean House Price per County

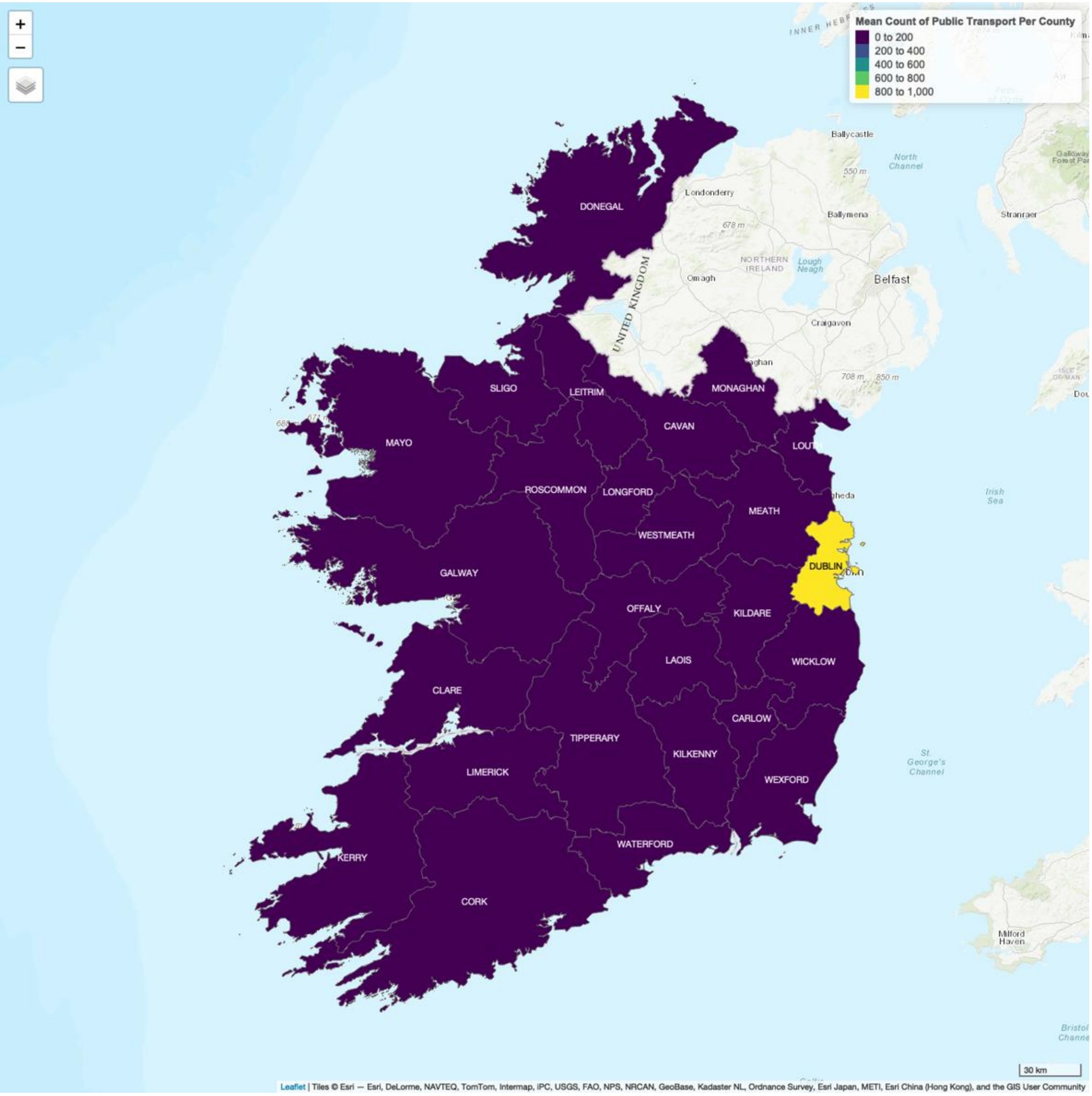


Source: Original data from the research, it can be found in the repository linked in the Appendix

Spatial visualisation of socio-economic characteristics

One of the main goals in this study is to find out whether socio-economic factors contribute to the valuation of a house. Thus, spatial visualisation is a useful tool to understand how the different areas in the country may be more privileged due to their access to a wider variety of public transport, hospital, education, and security infrastructure.

Figure 9: Choropleth Map of Mean Count of Public Transport per County

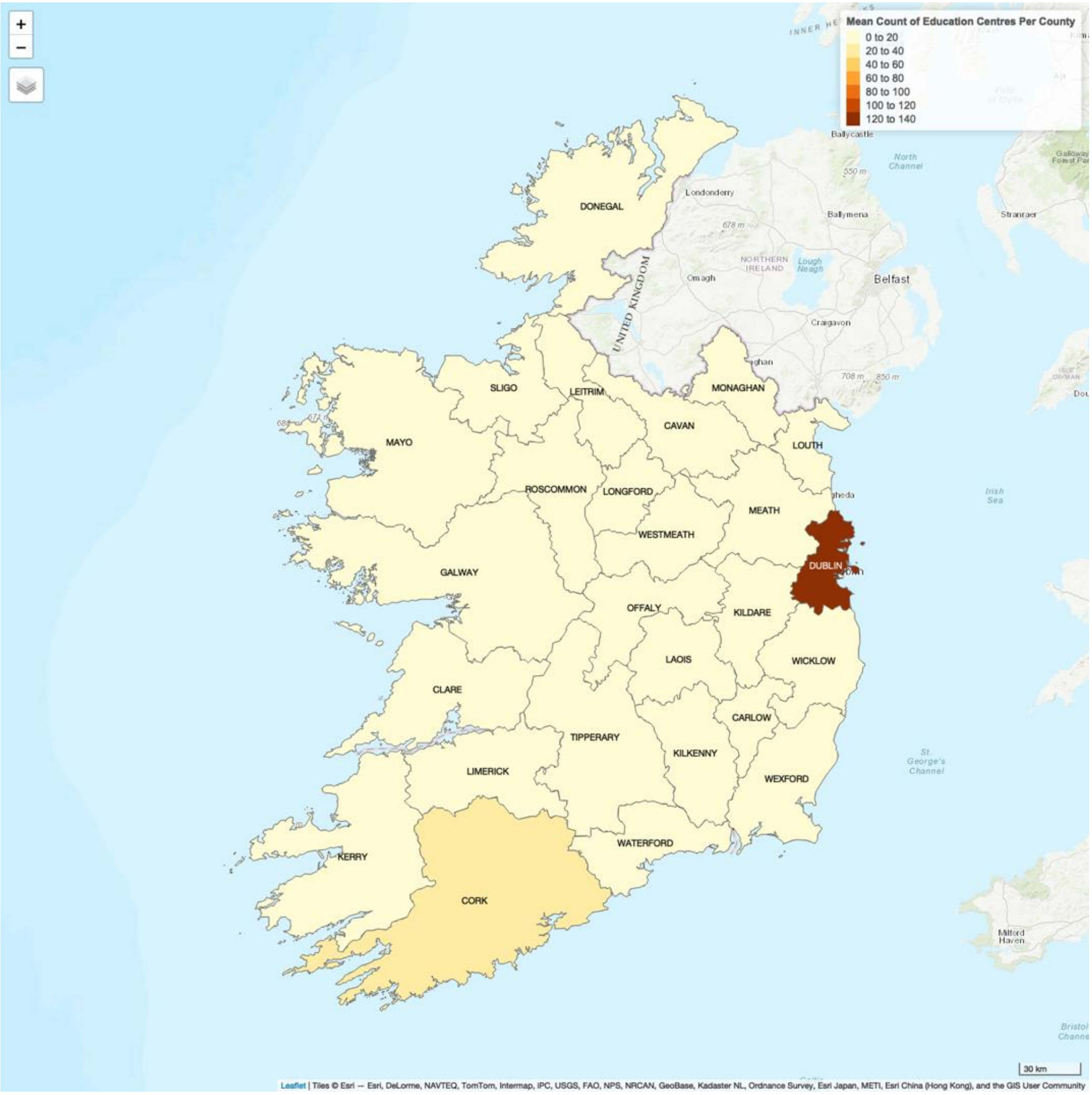


Source: Original data from the research, it can be found in the repository linked in the Appendix

Figure 9 depicts that those properties in county Dublin benefit from a much better access to public transport. The overwhelming difference between Dublin and other counties may indicate that the houses in the countryside are lacking on access to public transport. It might also clarify the question as to why people would be willing to pay more for a house near county Dublin.

A similar effect can be seen in Figure 10 that shows the average number of education centres – schools and universities, within a 5 km radius from the property on a given county. Dublin and Cork counties have the highest number of education centres average per house. However, Dublin County is still the clearest outlier on the map, having over 6 times more schools and universities closer to properties than most of the country.

Figure 10: Choropleth Map of Mean Count of Public Transport per County



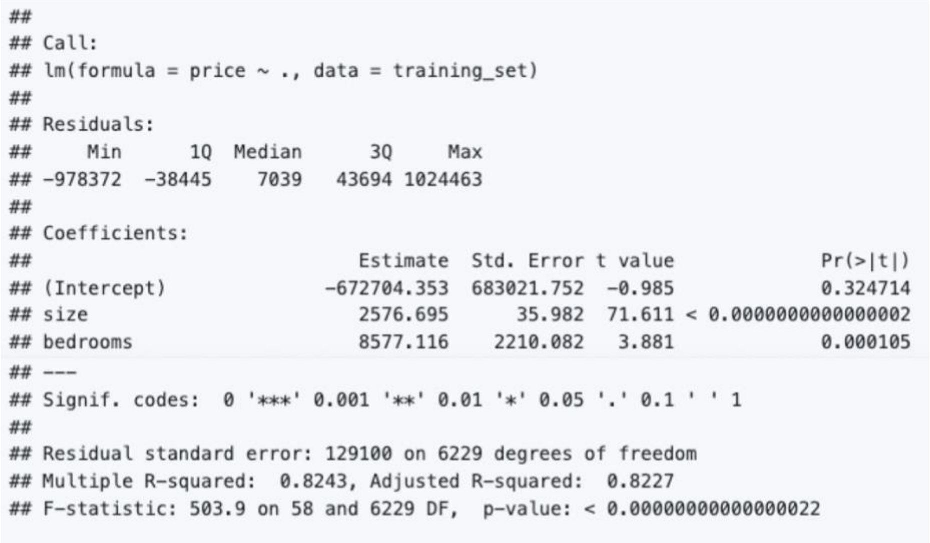
Source: Original data from the research, it can be found in the repository linked in the Appendix

Model Training

GLM – Linear Regression Model

In this study, the first model considered was the GLM Linear Regression, this model is commonly used as the baseline in many machine learning projects due to its simplicity, easy training and the good results it yields. The *lm* function from the stats package in R was used to compute the first model, this model used the price variable as its target variable and all the other variables as its predictors, the model had an R-squared of 0.8243 and an Adjusted R-squared of 0.8227. In Figure 11, it is possible to see the partial output of the lm function – some of the output was omitted for brevity’s sake.

Figure 11: Partial output of lm function



```
##
## Call:
## lm(formula = price ~ ., data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -978372  -38445    7039   43694 1024463
##
## Coefficients:
##                             Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)              -672704.353  683021.752  -0.985             0.324714
## size                        2576.695      35.982  71.611 < 0.0000000000000002
## bedrooms                    8577.116    2210.082   3.881             0.000105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 129100 on 6229 degrees of freedom
## Multiple R-squared:  0.8243, Adjusted R-squared:  0.8227
## F-statistic: 503.9 on 58 and 6229 DF,  p-value: < 0.00000000000000022
```

Source: Original data from the research, it can be found in the repository linked in the Appendix

While the model presented a reasonable R-squared, most of the dummy variables as well as latitude and longitude had a very high p-value and proved not to be statistically significant when taking into consideration the other predictors. Thus, the *step* function, which comes from the stats package in R, was used to perform the stepwise procedure so it could find the best predictors for the model to use. The stepwise procedure is an iterative process whereby different predictors with the highest statistical significance are added or removed from the model to find a combination of predictors that minimises some error function such as the Residual Sum of Errors (RSS). The stepwise procedure ensures that all the predictors left in the model are statistically significant and it also helps removing predictors with multicollinearity issues.

The results from the linear regression model after stepwise did not differ much from the first model, nevertheless with the stepwise model, only statistically significant predictors were kept. After that, the next step was to apply a normality test on the residual errors to figure out whether those followed a normal distribution and therefore the linear model was indeed suitable for the problem. Many error normality tests can be used: Shapiro-Francia [33], Anderson-Darling [34] and many others. The test of choice was the Anderson-Darling via the function ad.test from the nortest package in R. Anderson-Darling was used as its implementation in R could handle data frames with over 5000 rows. Furthermore, the Anderson-Darling test makes the same hypothesis test as Shapiro-Francia, where:

- **H0**: The null hypothesis is such that when *p-value* >= 0.05, the residual errors follow a normal distribution.
- **H1**: The alternative hypothesis is such that when *p-value* < 0.05, the residual errors do not follow a normal distribution.

The result from the Anderson-Darling test showed a p-value of 2.2e-16, which accepts the alternative hypothesis. However, the Box-Cox [35] transformation on the target variable was applied to bring the target variable into a normal distribution. After applying Box-Cox and putting the model under the Anderson-Darling test once more, it was proven that the residual errors did not follow a normal distribution and therefore the GLM Linear model could not be used for this problem.

The results of the linear model with and without stepwise as well as with and without Box-Cox can be seen on Table 5. Please note that when comparing models, the Adjusted R-squared should be used.

Table 5: Results of GLM Linear Models

Model	R-squared	Adjusted R-squared
Linear model	0.8243	0.8227
Linear model Stepwise	0.8225	0.822
Linear model box-cox	0.8722	0.871
Linear model box-cox stepwise	0.872	0.871

Source: Original data from the research, it can be found in the repository linked in the Appendix

Regression Trees

Regression trees were chosen as the next approach as they can work well with complicated data. Regression Trees were used in the following manner:

1. Initial naive tree model was trained using the function rpart from the rpart package in R was used with its default parameters and by providing the price as tree’s target and the remaining of the dataset as its predictors. By default, rpart already performs a 10-fold cross validation when training the model.
2. Then, a grid search on the *minimum splits,* which is the minimum number of observations in a node so a split can happen, and *max tree dep parameter, which*

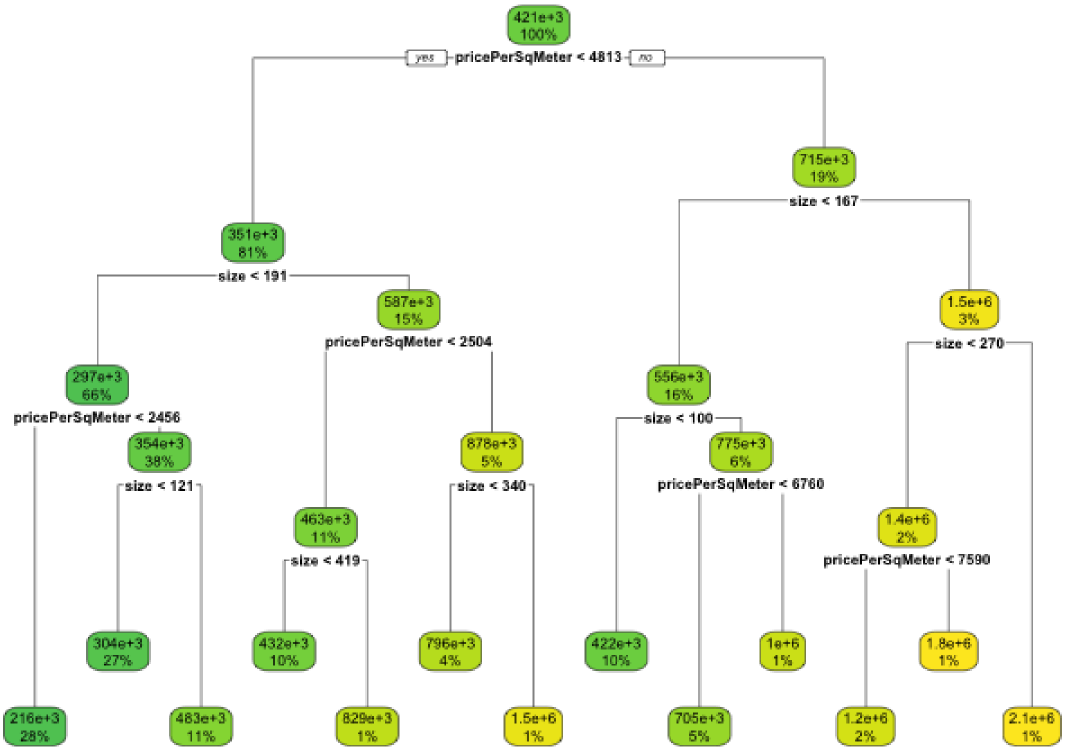
represents the maximum number of internal between the root node and a leaf node, was done. The grid search was constructed so that we had the minimum split varying from 3 to 30 variables (28 values) and for the maximum depth parameter we had a range from 15 to 50 nodes (36 values).

3. The grid search used 1008 (28 x 36) combinations of parameters during tuning of the model on the training dataset.
4. The final parameters used were minimum splits set to 3 and maximum depth set to 15 with and a cost complexity penalty (applied on the number of leaf nodes), *cp*, of 0.01.

The tuned regression tree presented a Residual Mean Square of Errors (RMSE) of 122957.30 in the validation dataset. Which meant that its prediction on a given house could be off by €122,957.30. In the testing dataset, the tree had a RMSE of 131472.00, which means that the prediction of house price could be off by €131,472.00.

Figure 12 shows the final tuned tree and the variables it used to split. The only predictors used were pricePerSqMeter and size.

Figure 12: Tuned Regression tree



Source: Original data from the research, it can be found in the repository linked in the Appendix

Random Forests

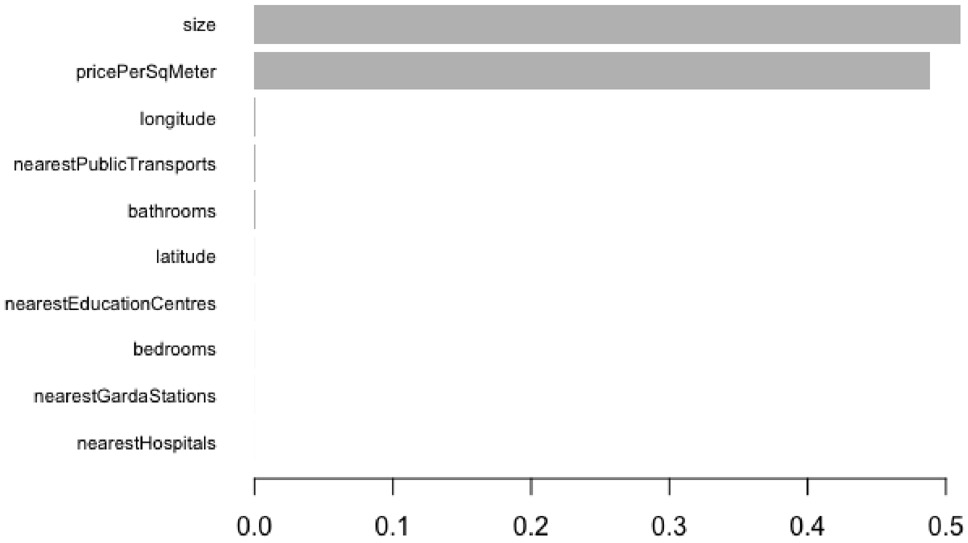
As the regression tree did not have a satisfactory result, the next approach was to try an ensemble model, like Random Forests [36] which can make use of many different trees that combine different parts of the training dataset, in a process called *bootstrapping*, and make use of different combination of predictor variables to find an aggregation of trees, *bagging*.

The random forest implementation used to train the model was the ranger implementation from the ranger package in R. The choice for this implementation is due to its computational performance, ranger has a better performance against other common implementations and that proves very useful when tuning the model. The steps used to train the random forest are similar to the ones used for the regression tree, what differs is the choice of hyperparameters. The following approach was used:

1. A grid search using the parameters: *mtry* – number of predictors to randomly use in the model was from 5 to 57 predictors with a step of 2; *node_size* – minimum number of observations within a leaf node was from 3 to 30 with a step of 2 , and *sample_size* – percentage of samples used to train on – the following sizes were used: 55%, 63.2%, 70% and 80%, were used to tune the model. A total of 972 combinations of those values was then used in the tuning phase.
2. The random forest ranger model was tuned, and the optimal values found were: *mtry* of 57 predictors, *node_size* of 3 and *sample_size* of 0.8 (80%).

The tuned random forest had a RMSE of 11061.66 on the validation dataset, which meant that its estimates on a real scenario would be off by €11,061.66. This result is impressive, and it is over 10 times better than the one given by a regression tree. Furthermore, in the testing dataset the Random Forest had a RMSE of 19489.01, which in contrast is still far better than a regression tree, but it was twice the RMSE found in the validation set. Lastly, Figure 13 shows the most important variables used by random forest to make its decision.

Figure 13: Most important variables used by tuned random forest.



Source: Original data from the research, it can be found in the repository linked in the Appendix

Extreme Gradient Boosting

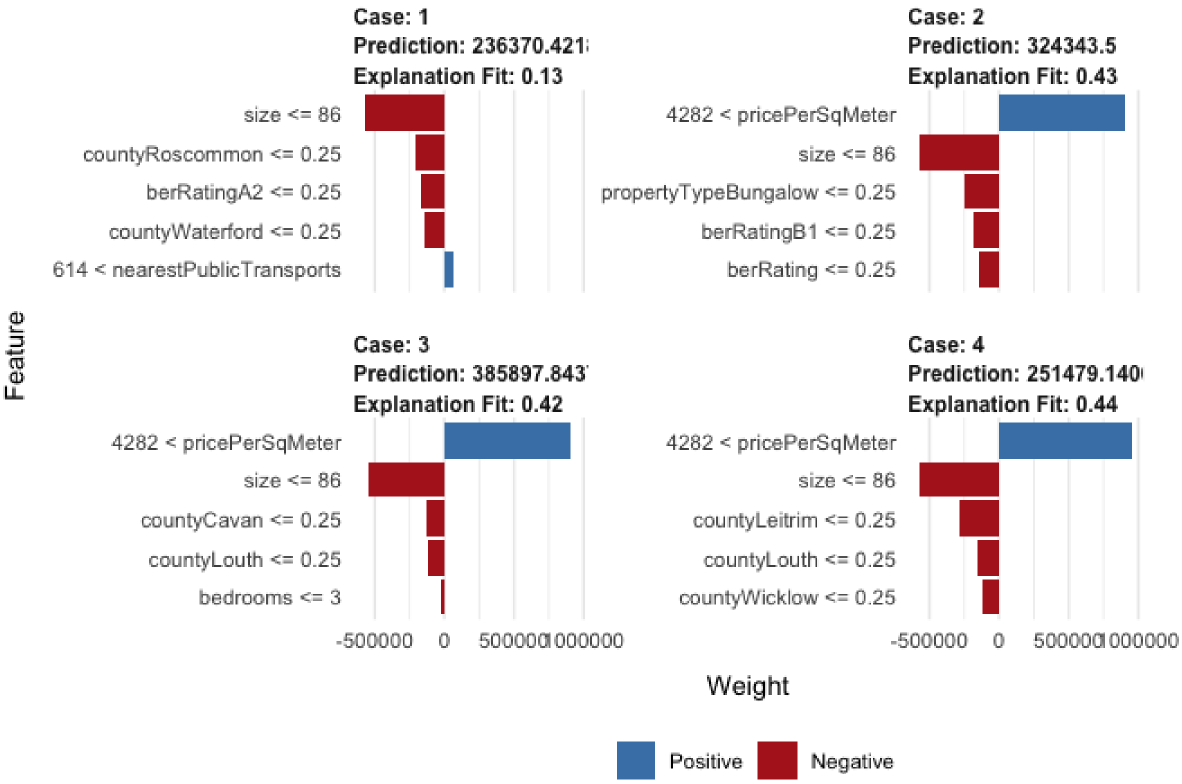
Even though Random Forest delivered a satisfactory result, Extreme Gradient Boosting (XGB) was used to try and find an optimal result. XGBoost is similar to random forest, however, instead of trying to build deep trees, XGBoost builds many shallow trees, sometimes with just one node to create a collection of weak trees that learn from each other.

The steps used to train and tune the XGBoost model are similar to the previous models, the following hyperparameters were tuned:

1. A grid search using the parameters: *eta* – which is the learning rate of the algorithm, *max_depth – which is the maximum number of internal between the root node and a leaf node*, *min_child_weight* – which represents the minimum sum of observations in each node, *subsample* – this is the ratio of training dataset that will be used, *colsample_bytree* – this parameter sets the ratio of predictors used to train each tree, were used. In the end, 162 combinations were created for this grid search.
2. The XGBoost model was then trained using the grid search and its optimal hyperparameters values were *eta* of 0.05, *max_depth* of 3, *min_child_weight* of 3, *subsample* of 0.65 and *colsample_bytree* of 1.0.

The XGB tuned model obtained a RMSE of 12727.64 on the validation dataset, which meant that its predictions might be wrong by €12,727.64. Nevertheless, that is slightly higher than what was obtained by Random Forests. When using the testing dataset, the model obtained a RMSE of 14717.2 or €14,717.2. Finally, Figure 14 presents the most important features used by XGBoost to make the predictions on given sample set.

Figure 14: Features used by XGBoost on sample from training set.



Source: Original data from the research, it can be found in the repository linked in the Appendix

A summary of the results across the three ensemble models can be seen on Table 6. In the table, it can be seen that Regression Trees had the worst performance, whereas Random Forest had the best performance on the validation dataset, nevertheless its performance on the testing set was way worse. Extreme Gradient Boosting was overall the best performing model as it performed quite well on the testing set and proved that it did not overfit on the validation set.

Table 6: Results of GLM Linear Models

Model	RMSE – Validation set	RMSE – Testing set

Regression Tree	€122,957.30	€131,472.00
Random Forest	€11,061.66	€19,489.01
Extreme Gradient Boosting	€12,727.64	€14,717.2

Source: Original data from the research, it can be found in the repository linked in the Appendix

Conclusion

Being able to effectively value properties in an environment where their prices have been on a steady rise for the last 10 years has become an important task in to avoid overpaying for houses that do not necessarily base their valuation on its characteristics, spatial location, and access to public amenities. Furthermore, understanding if socioeconomic factors, such as the number of hospitals and schools or access to public transport, play a role on the valuation of a house is also an important consideration for buyers and policymakers.

This study aimed at building a machine learning model that can make use of variables on the house's characteristics, spatial location, and socioeconomic factors to effectively predict the price of a house.

House characteristics, mostly price per square meter and size proven to be the most used across the 3 machine learning models tested. Nevertheless, those characteristics alone did not yield satisfactory results as it was shown by the regression tree RMSE and the reason why better performing algorithms, like Random Forest and XGBoost used other variables as well.

Access to public transport, such as train stations and bus stops, was among the top 4 most important features used by Random Forest to make its predictions as it can be seen on Figure 13, although its importance was relatively small when compared to size and price per square meter, it proved to be an important feature. Similarly, having garda stations, hospitals and schools were part of the 10 most important variables used by Random Forest, however to a far less extent than public transport.

Spatial location was used by XGBoost on its decision process to mostly accurate make predictions as Figure 14 shows. However, more importantly, spatial data proved to be very useful to understand the spatial relationship amongst counties in Ireland and how prices and number of houses on sale influence its neighbours.

References

[1] – Trading Economics. Ireland Residential Property Prices. Available at: <https://tradingeconomics.com/ireland/housing-index>. Accessed on: 30th of March 2022.

[2] – Parliamentary budget Office. Snapshot of the Housing Market in 2021. Available at <https://data.oireachtas.ie/ie/oireachtas/parliamentaryBudgetOffice/2021/2021-12-06_snapshot-of-the-housing-market-in-2021-part-1_en.pdf>. Accessed on: 29th of March 2022.

[3] – Financial Times. Ireland unveils record spending to tackle housing crisis. Available at: <https://www.ft.com/content/ea4cf916-c5c1-4f5c-8fce-59792aed8162>. Accessed on: 1st of August 2022.

[4] – McQuinn, Kieran. 2017 Irish House Prices: Déjà Vu all over again? - Quarterly Economic Commentary. Available at <https://www.esri.ie/system/files/media/file-uploads/2017-12/QEC2017WIN.pdf#page=96>

[5] – Cost of living crisis: New figures reveal who is hit the hardest by inflation. Available at: <https://www.independent.ie/business/personal-finance/latest-news/cost-of-living-crisis-new-figures-reveal-who-is-hit-hardest-by-inflation-41871591.html>. Accessed on: 1st of August 2022.

[6] Allen-Coghlan, Matthew et al. (2020) - Estimating the cost of Irish housing for the CPI: A rental equivalence approach, ESRI Working Paper, No. 676, The Economic and Social Research Institute (ESRI), Dublin.

Available at <https://www.econstor.eu/bitstream/10419/237947/1/WP676.pdf>

[7] – Residential Property Price Index June 2022. Available at: <https://www.cso.ie/en/releasesandpublications/ep/p%2Drppi/residentialpropertypriceindexjune2022/>. Accessed on: 12th of August 2022.

[8] – Kok, S.H., Ismail, N.W. and Lee, C. (2018), "The sources of house price changes in Malaysia", *International Journal of Housing Markets and Analysis*, Vol. 11 No. 2, pp. 335-355. https://doi.org/10.1108/IJHMA-04-2017-0039.

[9] - Bork, L 2017, What drives metropolitan house prices in California? in *Reserve Bank of New Zealand conference on Housing, household debt and policy.* pp. 1, Reserve Bank of New Zealand: Housing, household debt and policy, Wellington, New Zealand, 11/12/2017.

[10] – Torset, W 2018, An empirical analysis of the Norwegian housing market – What drives the house price? , Norwegian School of Economics. https://openaccess.nhh.no/nhh-xmlui/bitstream/handle/11250/2561254/masterthesis.PDF.

[11] – Tedin, D. J. M. and Faubert V. (2020), Housing Affordability in Ireland, Economic Brief 061. December 2020. Brussels. PDF. 20pp.

[12] - Hurley, Aoife & Sweeney, James. (2022). Irish Property Price Estimation Using A Flexible Geo-spatial Smoothing Approach: What is the Impact of an Address?. The Journal of Real Estate Finance and Economics. 10.1007/s11146-022-09888-y.

[13] - Jason Deegan. A spatial regression of Irish residential property prices. Available at: < https://jasondeegan.com/spatial-regression/>. Accessed on 9th of June 2022.

[14] – Selectra. The House Crisis Ireland 2022: fact or fiction. Available at <https://selectra.ie/moving/guides/renting/housing-crisis-ireland>. Accessed on: 29th of March 2022.

[15] – Irish Times Newspaper. Housing is "number one" crisis facing young people – Taoiseach. Available at: <https://www.irishtimes.com/news/politics/housing-is-number-one-crisis-facing-young-people-taoiseach-1.4559764>. Accessed on: 29th of March 2022.

[16] - RTE News. "Collapse" in home ownership among young adults – report. Available at <https://www.rte.ie/news/business/2022/0113/1273467-parliamentary-budget-office-on-housing/>. Accessed on: 27th of February 2022.

[17] Property Price Register Ireland Dataset. Available at <https://www.propertypriceregister.ie/>. Accessed on: 9th of June 2022.

[18] Daft.ie API Version 1. Available at <https://api.daft.ie/doc/v1/ >. Accessed on: 1st of August 2022.

[20] Garda stations directory. Available at < https://www.garda.ie/en/contact-us/station-directory/ >. Accessed on 10th of February 2023.

[21] Health Service Executive page on gov.data.ie. Available at< https://data.gov.ie/dataset/list-of-hospitals-in-ireland/resource/f6727f58-a6bc-45f9-9657-84c9eecfd5b7 >. Accessed on 10th of February 2023.

[22] National Transport Ireland, National Public Transport Access Nodes (NaPTAN). Available at < https://data.gov.ie/dataset/national-public-transport-access-nodes-naptan >. Accessed on 10th of February 2023.

[23] Irish Rail Realtime API – getAllStationsXML service. Available at < http://api.irishrail.ie/realtime/realtime.asmx?op=getAllStationsXML >. Accessed on 10th of February 2023.

[24] gov.ie – schools directory. Available at < https://www.gov.ie/en/directory/category/495b8a-schools/?page=1>. Accessed on 10th of February 2023.

[25] Education in Ireland – where can I study. Available at <https://www.educationinireland.com/en/Where-can-I-study-/>. Accessed on 10th of February 2023.

[26] Google Geocoding API. Available at < https://developers.google.com/maps/documentation/geocoding/overview >. Accessed on 10th of February 2023.

[27] Fisher, R.A.: Statistical Methods for Research Workers. Oliver & Boyd, Edinburgh (1925)

[28] Law of Supply and demand. Available at < https://www.investopedia.com/terms/l/law-of-supply-demand.asp >. Accessed on the 10th of February 2023.

[29] Web Scrapping. Available at < https://en.wikipedia.org/wiki/Web_scraping >. Accessed on the 20th of April 2023.

[30] distHaversine function from the geosphere package in R. Available at < https://www.rdocumentation.org/packages/geosphere/versions/1.5-18/topics/distHaversine >. Accessed on 20th of April 2023.

[31] ggplot2 package in R. Available at < https://ggplot2.tidyverse.org/ >. Accessed on 20th of April 2023.

[32] tmap – thematic maps in R. Available at < https://cran.r-project.org/web/packages/tmap/ >. Accessed on 20th of April 2023.

[33] SHAPIRO, S. S., & WILK, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

[34] Anderson, T.W., Darling, D.A.: A test of goodness of fit. J. Am. Stat. Assoc. **49**, 765–769 (1954)

[35] Box GEP, Cox DR (1964) An analysis of transformations. J Roy Stat Soc, Ser B 26:211–252

[36] Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324

Appendix

The source code for the R notebook and Python scripts used in this analysis are available at < https://github.com/marcosmcb/housing-dataset-geospatial-analysis >.Three R notebooks were created for doing the data cleaning, exploratory analysis and modelling.