WHAT SOCIO-ECONOMIC AND GEOGRAPHICAL FACTORS INFLUENCE HOUSE PRICES IN IRELAND

Marcos Cavalcante Barboza

Ana Julia Righetto

Agenda

- Introduction
- Material and Methods
- Results and Discussion
- Conclusion

Introduction
Motivation

House prices **increased** by **167%** from 2012 to 2022.

Between 1995 and 2007 home prices in Ireland **increased** by 474% the sharpest increase across countries in the **OECD**.

Government will allocate **€4bn annually** to build **300,000** homes by end of decade.

*“**Housing** is **number one** crisis facing young people”* – says **Irish Prime Minister**

Introduction

Motivation

Macro-economic shocks: interest rates, GDP growth, access to credit and exchange rates.

Regional Factors: supply and demand, scarcity of labour, rent legislation and household income.

Geospatial variables: address, proximity to other cities.

Introduction

Objectives

Does 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**.

HOSPITALS

SCHOOLS AND UNIVERSITIES

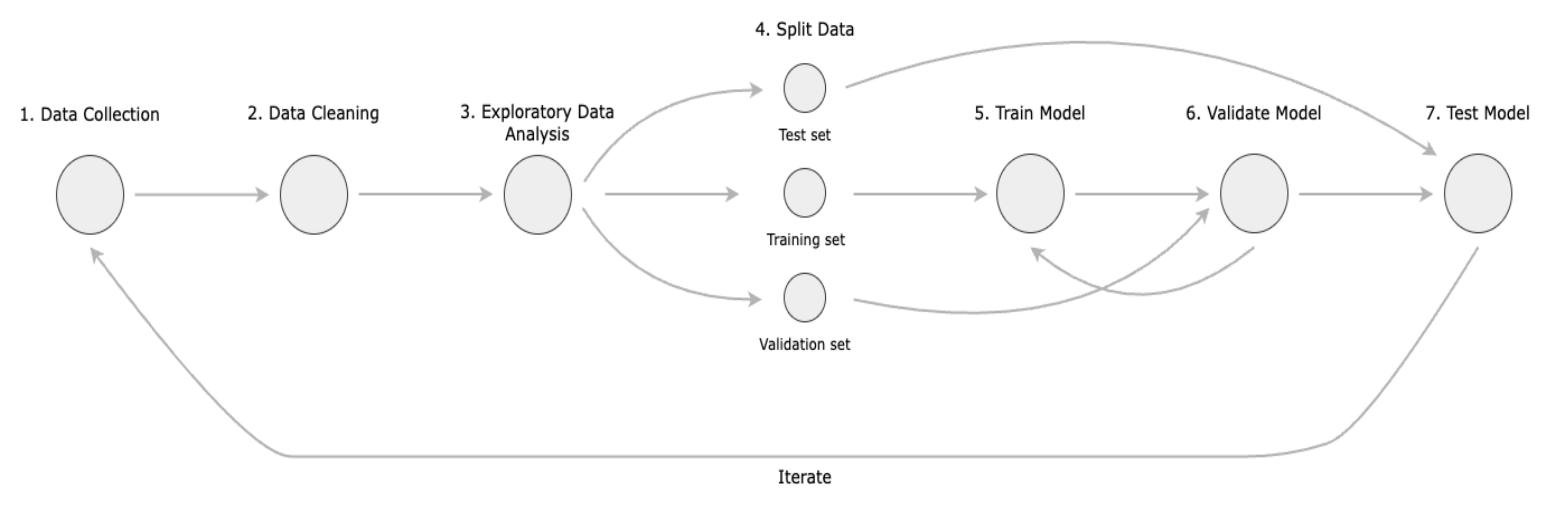
GARDA STATIONS

BUS AND TRAIN STOPS

HOUSE CHARACTERISTICS

Material and Methods

Overview

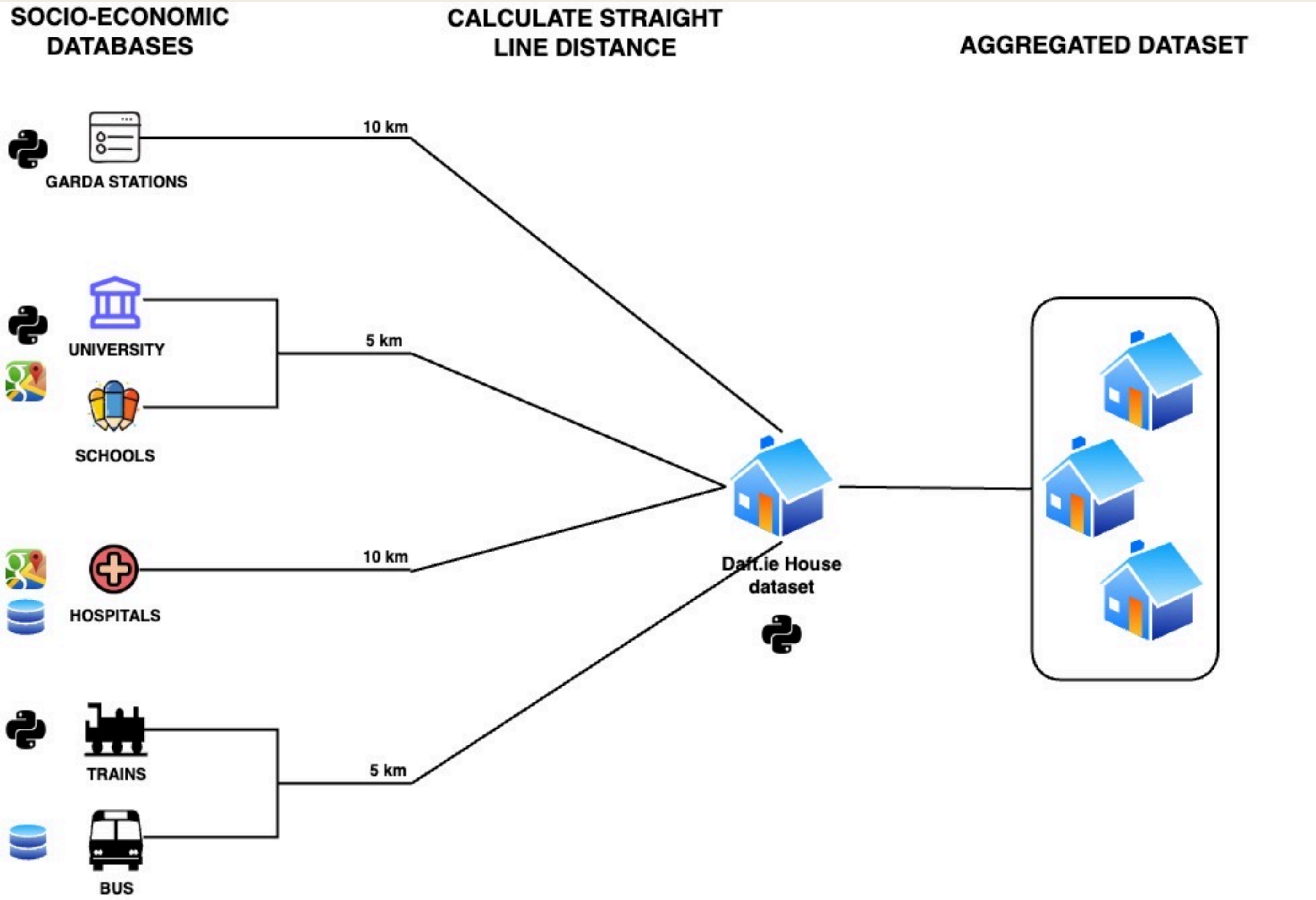


Material and Methods

Data Collection

- **House dataset** obtained from **Daft.ie** property website.
- Socio-economic datasets collected via ***webscrapping*** with ***python** scripts* and from ***data.gov.ie*** data portal.
- **Google Maps API** was used for enriching datasets.

Dataset	Total Observations
Housing Prices	100327
Bus stops - transport	16225
Train stations – transport	159
Schools – education	3531
Universities – education	22
Hospitals - health	38
Garda stations - security	566



Material and Methods

Data Cleaning and Preparation

Data Cleaning

Removal of duplicates: id and URL.

Removal of missing values: price, size (~ 26%) and bathrooms (~ 2%).

Feature Engineering

pricePerSqMeter created from the division between house price and size.

Location: created **county** and **townOrNeighborhood.**

Count of **hospitals**, **education centres**, **transport** and **garda stations** within a given radius.

Material and Methods

Final Dataset

- 9009 observations
- 1 target variable – *price*
- *15 features*
- ***Geospatial variables:*** *longitude, latitude, county, townOrNeighbourhood.*
- ***Socio-economic variables:*** *nearestHospitals, nearestGardaStations, nearestEducationCentres and nearestPublicTransports.*

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

Material and Methods

Exploratory Data Analysis (EDA)

Summary statistics

Bivariate plots with *ggplot2*

Spatial visualisations with ***tmap***

Pearson Correlation

Analysis of Variance - ***ANOVA***

Box-plot

Material and Methods

Model Selection and Training

Dataset randomly split into: ***training* – *70%*, *validation* - *15%*** and ***testing sets* – *15%***

4 machine learning algorithms: **GLM – linear regression, regression tree**, **random forests** and **extreme gradient boosting** (XGB)

Hyperparameter tuning via **grid search and**

Results and Discussion

Exploratory Data Analysis - EDA

- **75%** of the houses are under **€500,000**.
- Very few houses above **€4 million**.
- Dataset seems to be **right skewed**.

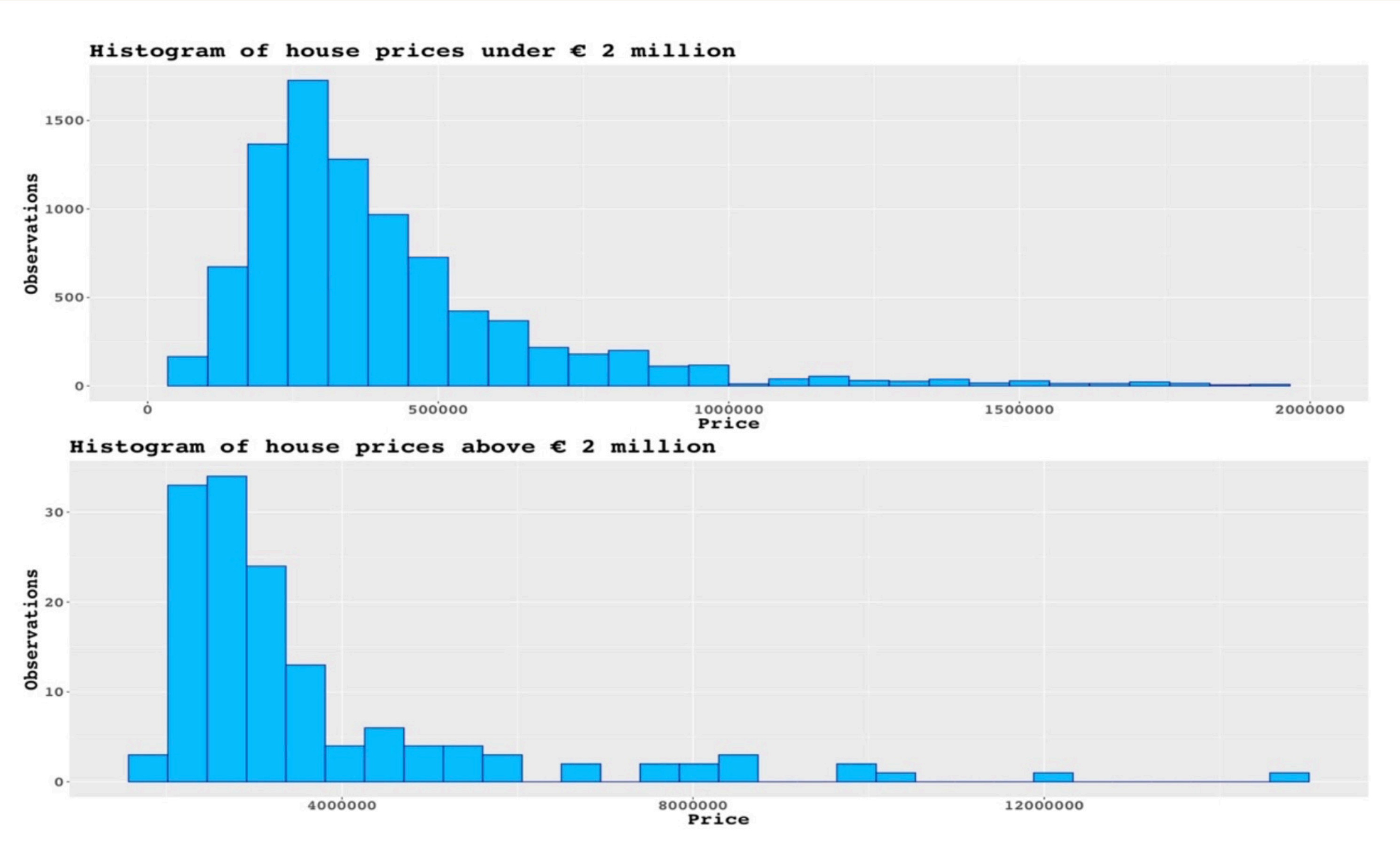


Figure 1: Histograms of the price variable.

Results and Discussion

Exploratory Data Analysis - EDA

- **Moderate positive correlation** between **price** and **size** as well as **price** and **pricePerSqMeter**.

- **Weak positive correlation** between price and the variables: **bathroom**, **bedroom**, **nearestHospitals**, **nearestGardaStations**, **nearestEducationCentres** and **nearestPublicTransports.**

- **Strong positive correlation** between **bathrooms** and **bedrooms**; and amongst the socio-economic variables.

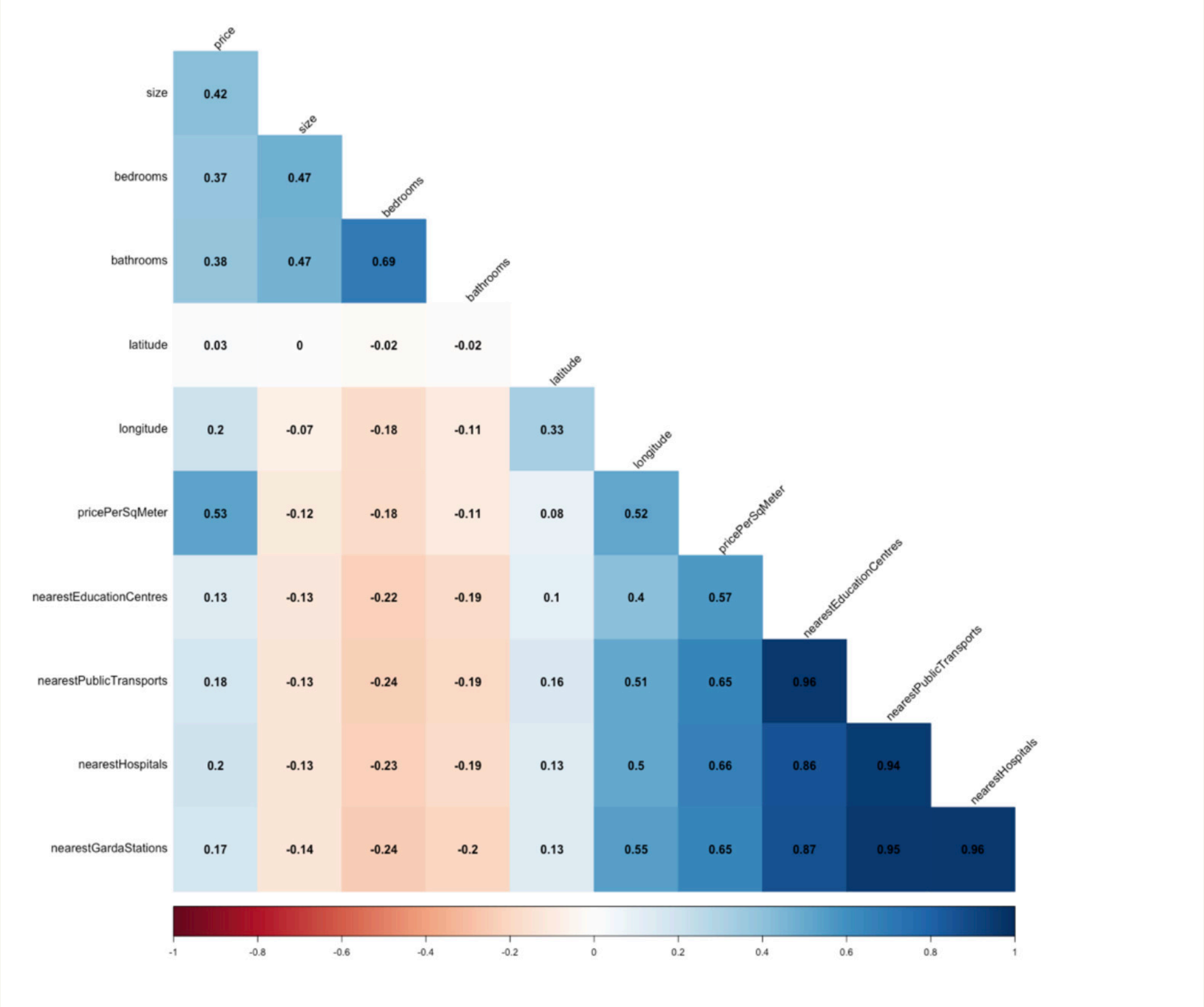


Figure 2: Correlation matrix of the numeric variables in the dataset.

Results and Discussion

Exploratory Data Analysis - EDA

- The **bigger** the house, the **more expensive** it gets.
- A clear **outlier** in the first graph is the house with **6000 square meters** (Size axis) and price under **€500,000** (Price axis).

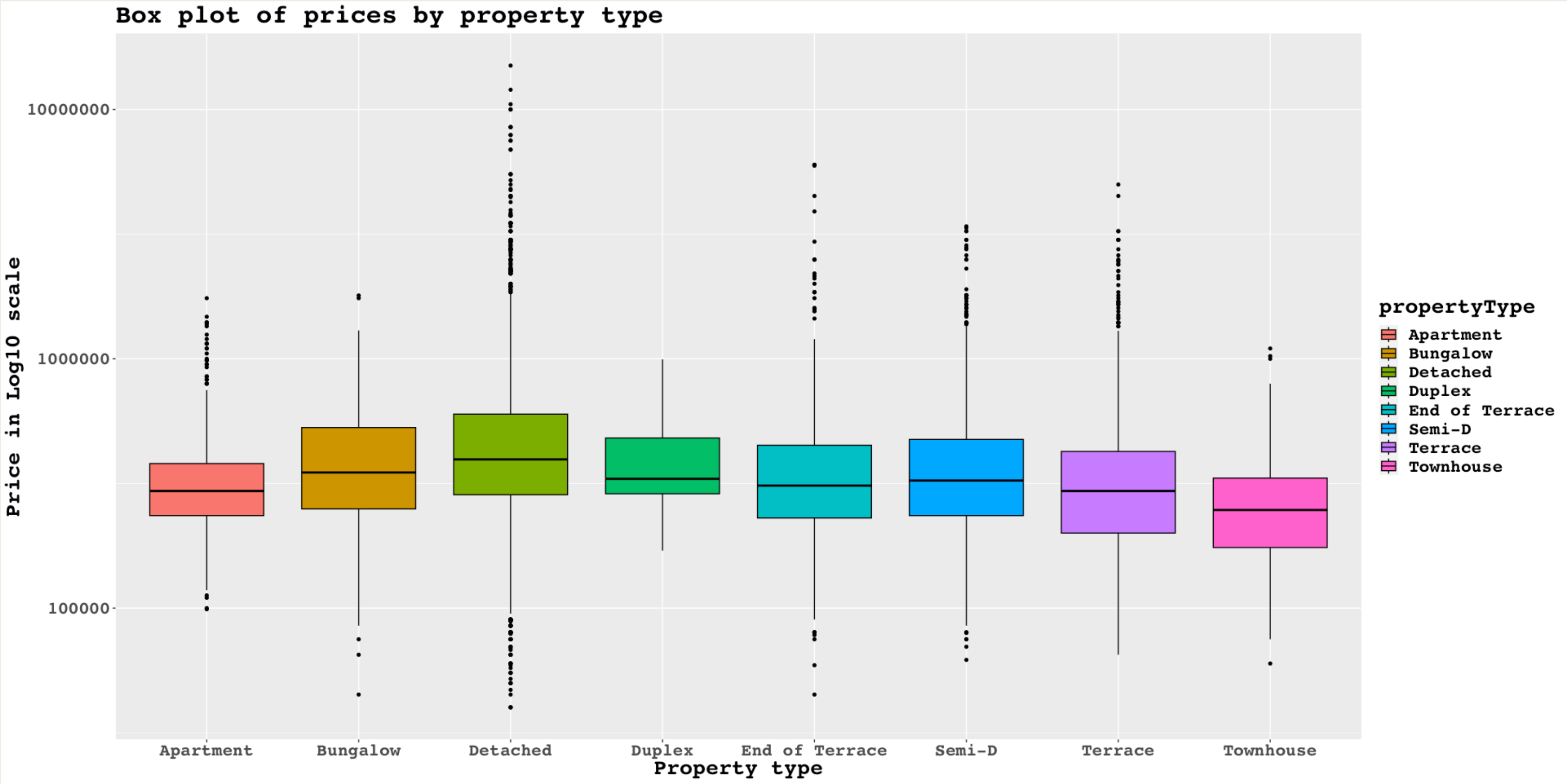
Figure 3: Bivariate graphs of houses under and over €2,000,000 by their size.

Results and Discussion

Exploratory Data Analysis - EDA

- **Apartments**, **terrace** and **townhouse** property types have the lowest prices.

- **Detached**, **bungalow** and **duplex** property types are usually more expensive and have more outliers.



Appendix: Box plot of prices by property type.

Results and Discussion

Exploratory Data Analysis - EDA

- Counties with the largest cities – **Dublin**, **Cork** and **Galway** hold most of the house supply.

- **"Neighbouring effect"** where the supply of houses in one county impact the supply in adjacent counties.

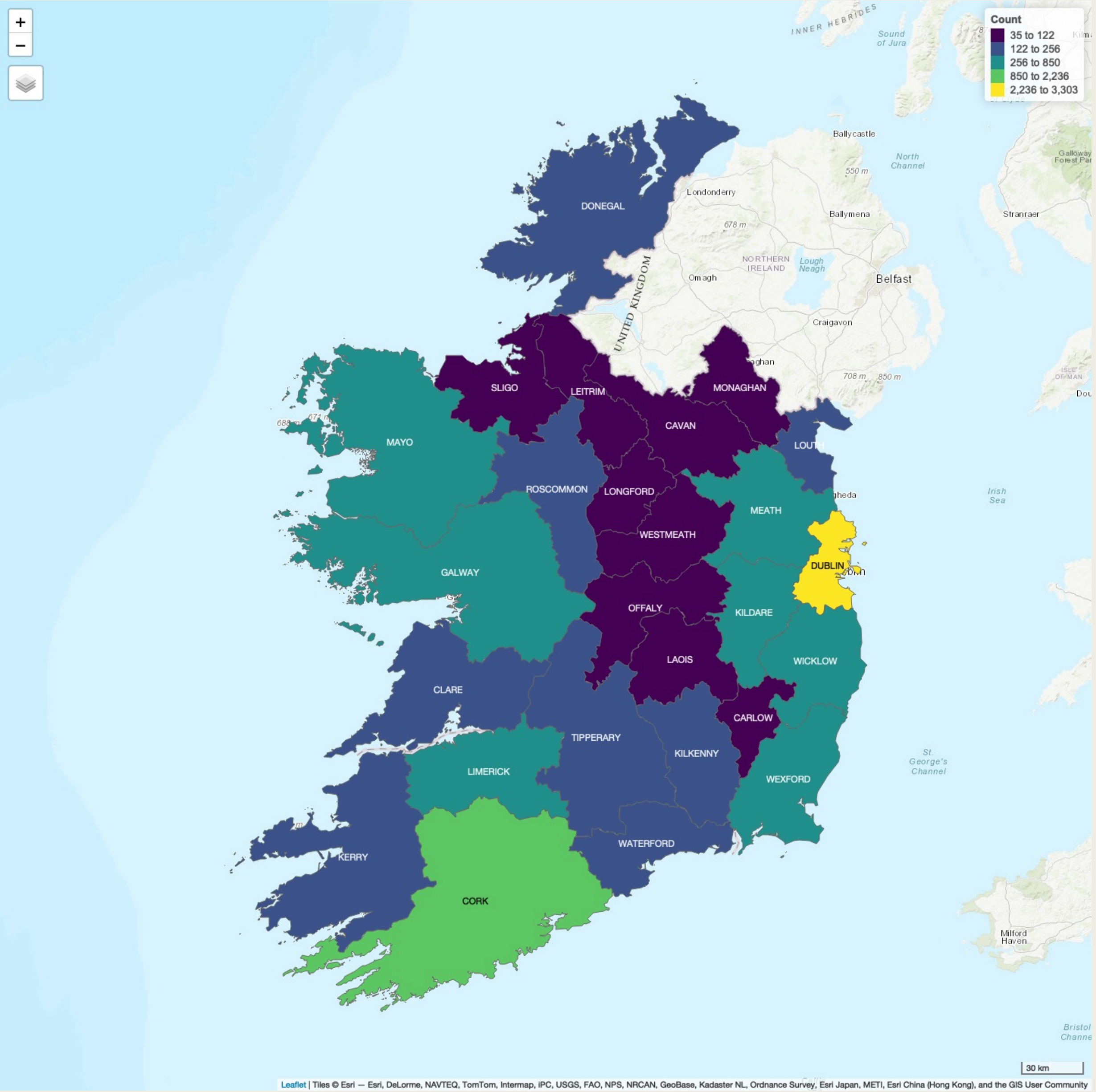


Figure 7: Density spatial graph of the number of houses per county.

Results and Discussion

Exploratory Data Analysis - EDA

- 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.**

- Counties in the **Midlands** and **North West** affect one another and thus are clustering together.

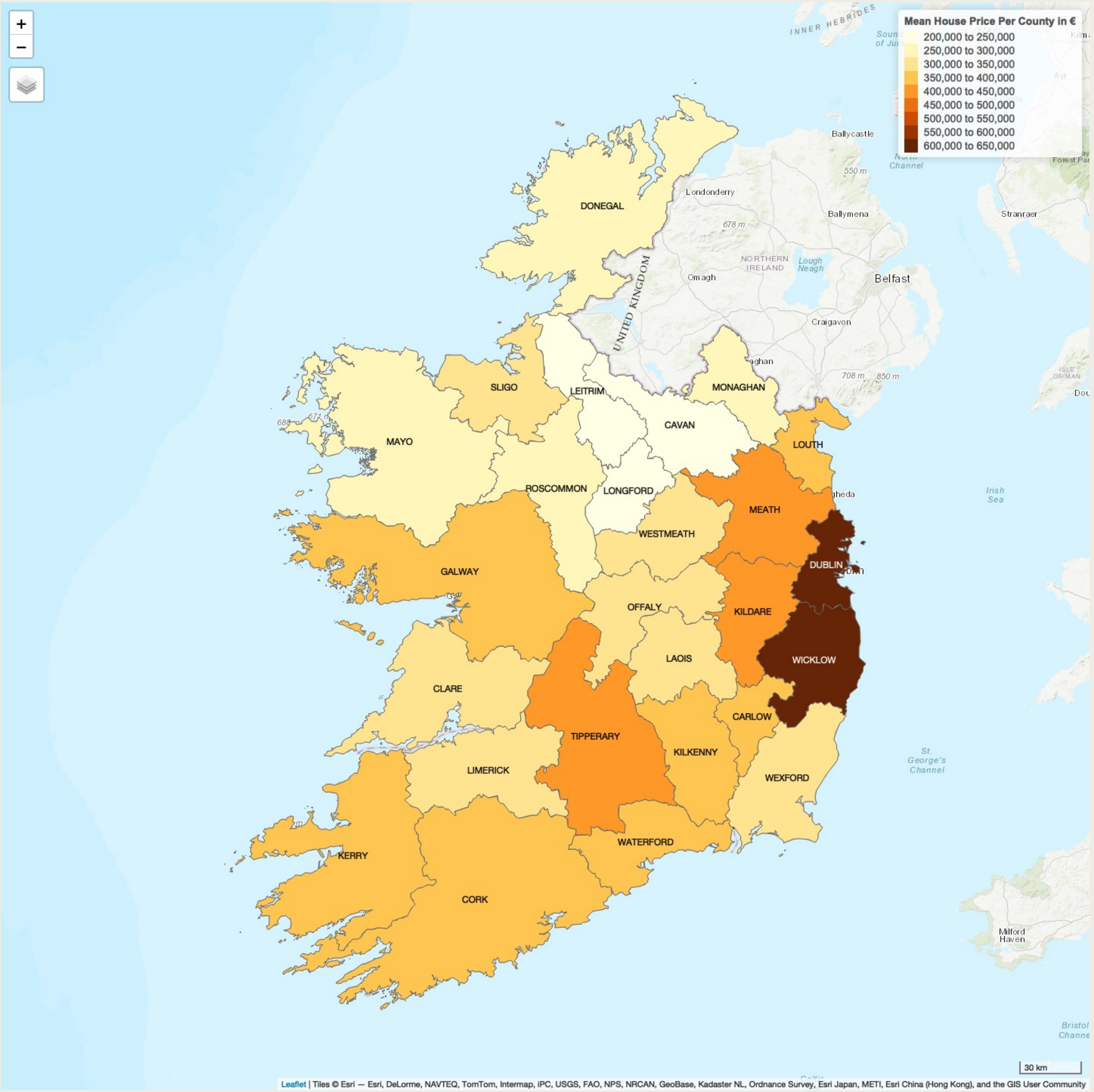


Figure 8: Choropleth Map of Mean House Price per County.

Results and Discussion

Exploratory Data Analysis - EDA

- Properties in county **Dublin** benefit from a much **better** access to **public transport.**

- **Early** indication that the houses in the **countryside** are **lacking** on **access** to **public transport.**

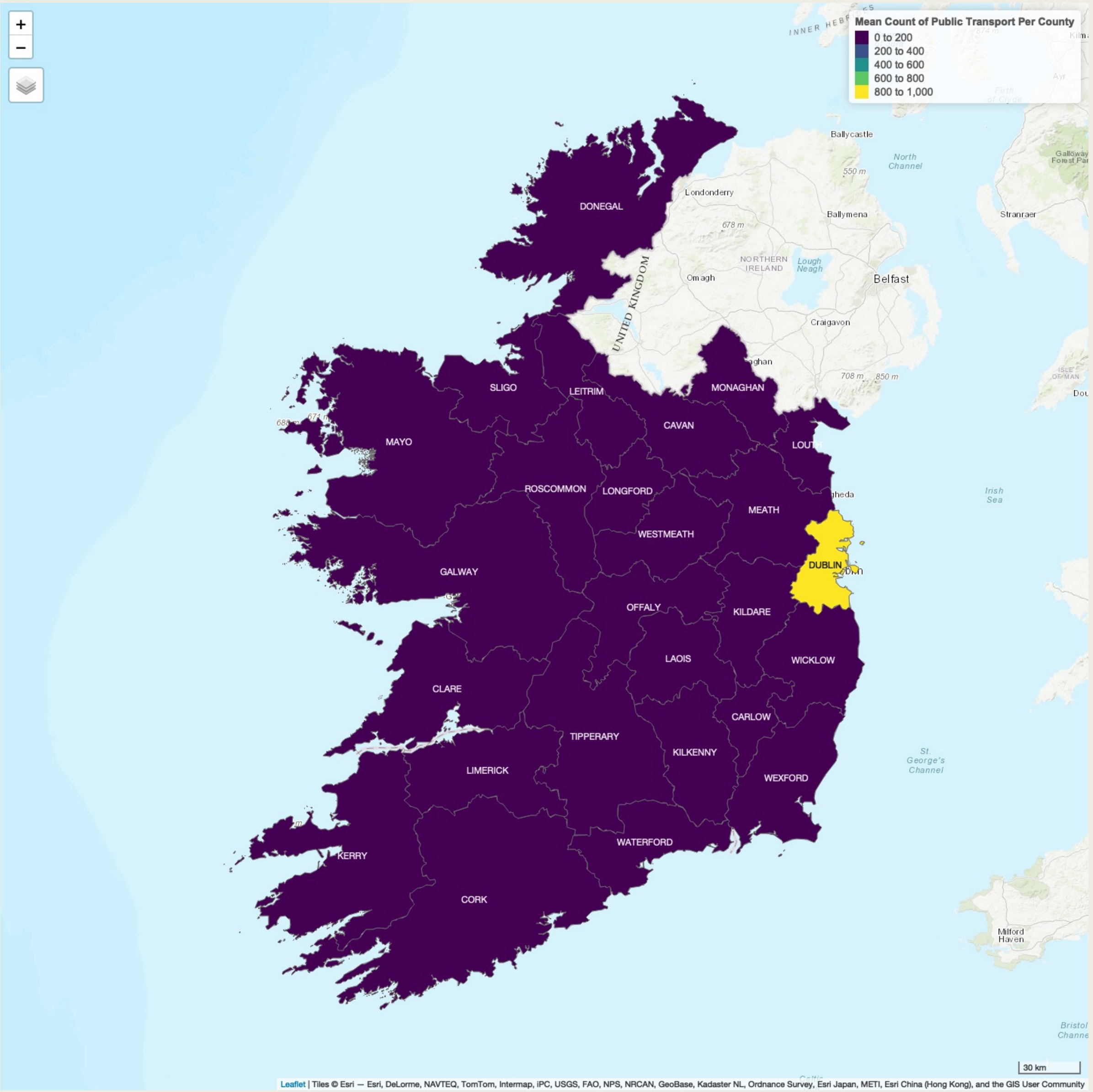


Figure 9: Choropleth Map of Mean Count of Public Transport per County.

Results and Discussion

Model Selection – GLM – Linear Regression

- **Baseline** model.
- **Stepwise** procedure was used to find statistically significant predictors.
- **Box-Cox** transformation found lambda value of -0.1280778.
- **Shapiro-Francia** *(Anderson-Darling)* applied to validate errors distribution - *p-value* of 2.2e-16 **reject** null hypothesis.
- Linear Regression model **rejected** due to errors **not** following **normal distribution**.

$$\boldsymbol{y} = \boldsymbol{\beta_0} + \beta_1 * \boldsymbol{x_1} + \beta_2 * \boldsymbol{x_2} + \beta_3 * \boldsymbol{x_3} + \cdots + \beta_p * \boldsymbol{x_p}$$

GLM – Linear Regression Formula

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

Table 5: Results of GLM Linear Models

Results and Discussion

Model Selection – Regression Tree

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Sum of the Squared Errors formula

- Quite **Interpretable** and easy to understand **variable importance**.
- **rpart** implementation.
- Minimise ***Sum Squared Errors (SSE).***
- Tuning of ***minsplit*** and ***maxdepth*** parameters.
- Grid searched across 1008 combinations.
- **RMSE** (Root Mean Squared Error) of 131472.00, which suggests that the prediction of house price could be off by €131,472.00.

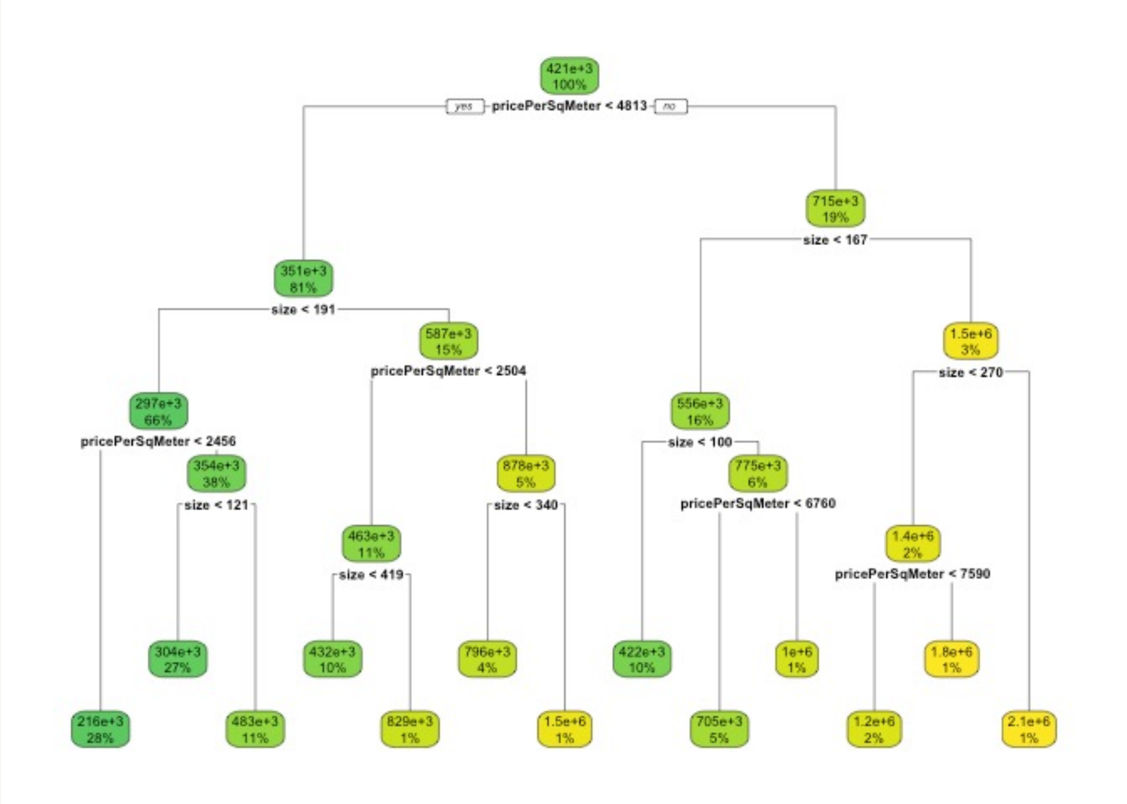


Figure 12: Tuned Regression tree.

Results and Discussion

Model Selection – Random Forests

- Similar to regression trees.
- ***Bootstrapping*** subsets of the dataset.
- ***Bagging*** training multiple decision trees on bootstrapped samples.
- ranger implementation
- Grid search of ***mtry***, **node_*size*** and ***sample_size*** – total of **972 combinations.**
- ntrees (1000), mtry (57) node_size(3)
- **RMSE** of €19,489.01 - far better than a regression tree.

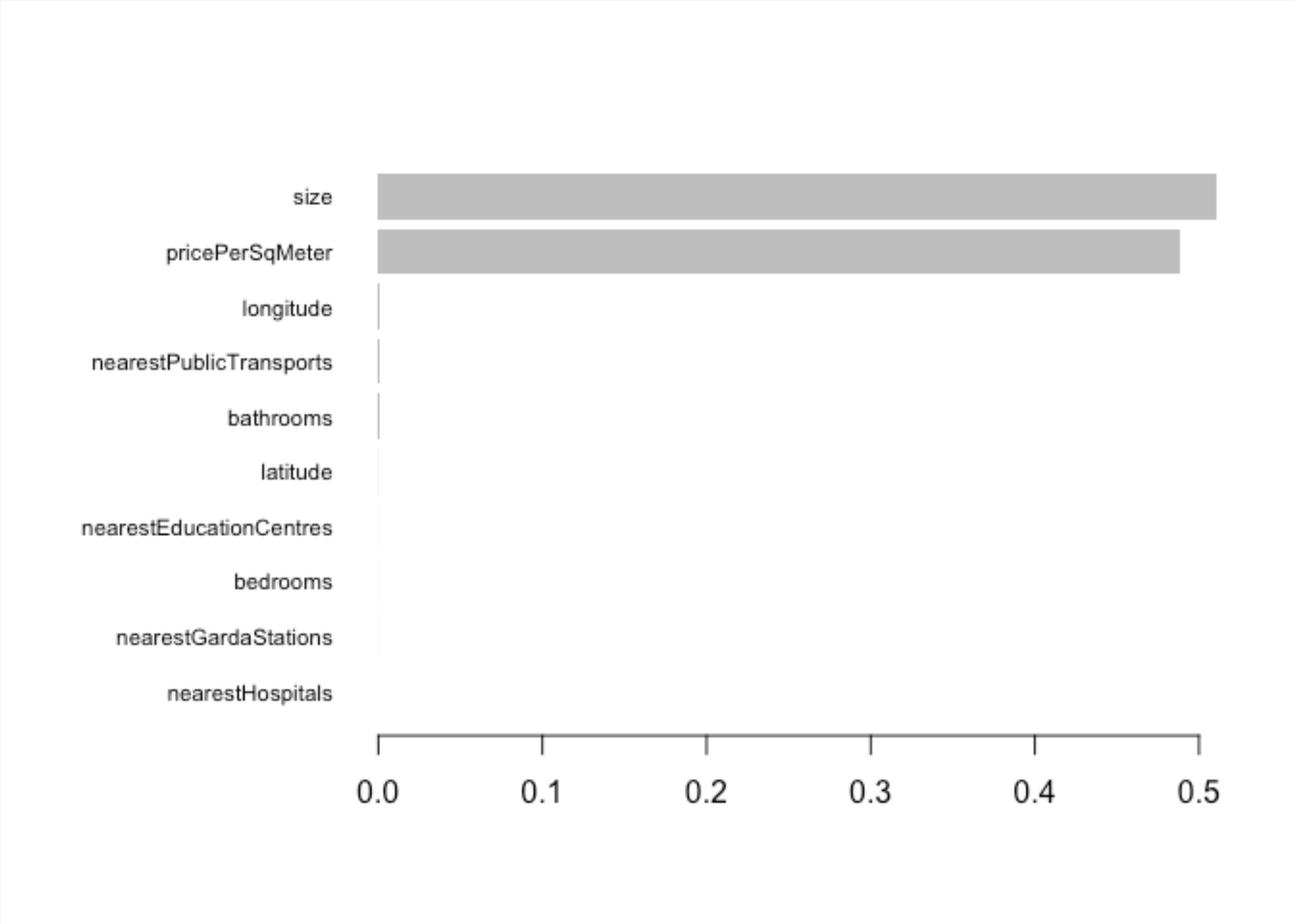


Figure 13: Most important variables used by tuned random forest

Results and Discussion

Model Selection – XGB - eXtreme Gradient Boosting

- Collection shallow / weak trees.
- Grid search on parameters: number of trees, depth of trees (3), learning rate (0.05) and subsampling (0.65) – 2800 rounds.
- 162 combinations.
- Model obtained a **RMSE** of 14717.2 or €14717.2

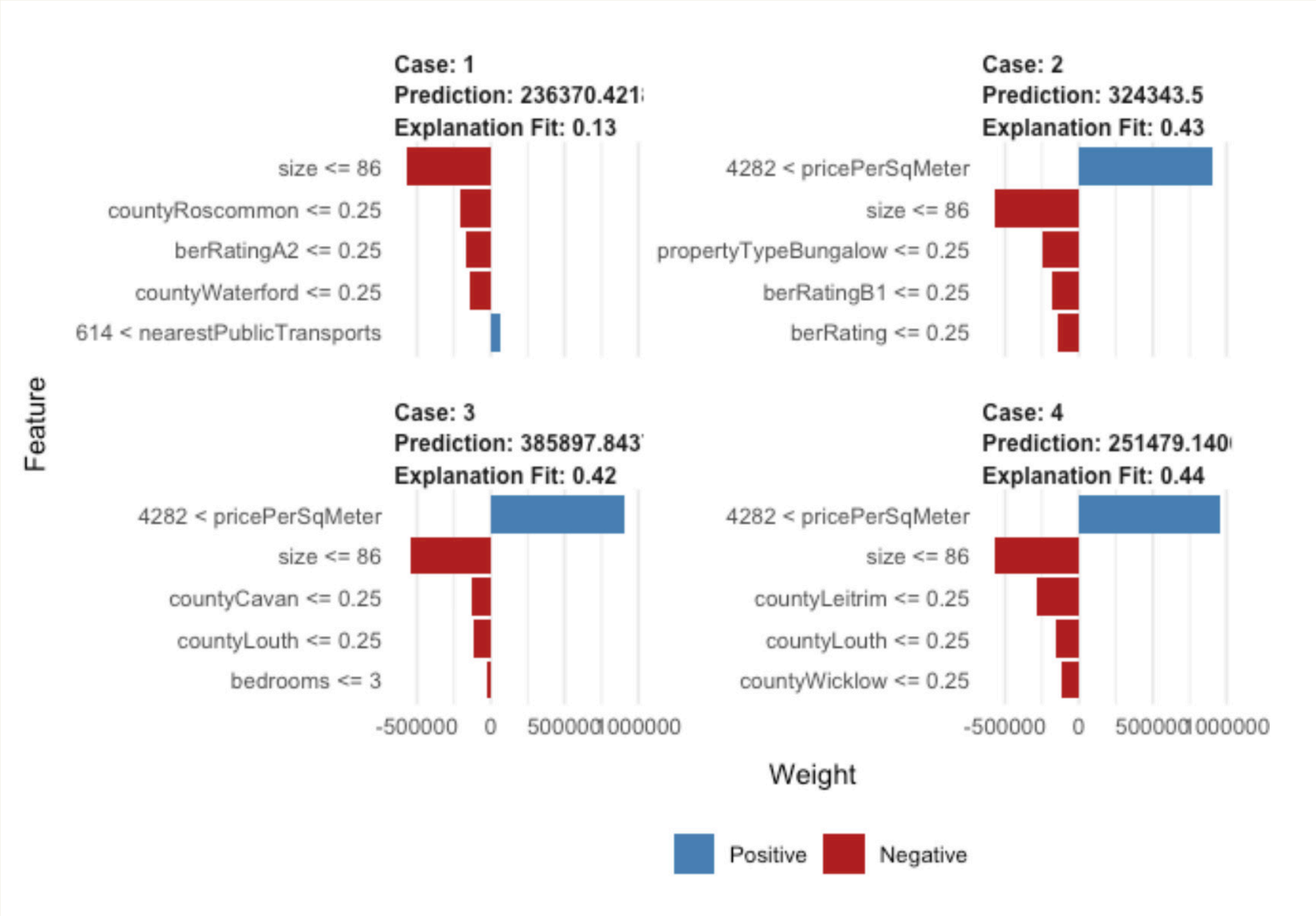


Figure 14: Features used by XGBoost on sample from training set.

Discussion

- **Removal of outliers:** to get the methods to have a satisfactory performance, some outliers were removed using the z-score method and about 200 observations were removed where their ***size*** and *price* was above 4 deviation standards.

Conclusion

- Access to public transport, such as train stations and bus stops, was among the top 4 most important features used by Random Forest.

- 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 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.

Future work

- Investigate Gaussian GLMs
- Create Google Chrome extension to plug on property websites
- Spatial Regression