Analysis of Ireland Property Prices

Ginu Varghese
Department of Computing Science and Mathematics
Dundalk Institute of Technology
Dundalk, Co.Louth, Ireland
d00251842@student.dkit.ie

Abstract: **The house prices are increasing every year and it is demanding the analysis and prediction of property prices. The main objective of this study is to perform an initial analysis on the property prices of Ireland from 2010 to 2022 collected from the PSRA website. Factors influencing the house prices county, year, province, location of property, property size and type are considered for analysis. Statistical test and visualizations were generated for the analyzing the effect of different attributes on property prices and the results implied that the province, year, property size, and location are statistically significant and are affecting the house prices significantly.**

Keywords: property prices, statistical analysis, visualization, province.

I. Introduction

An accurate house price prediction is important for prospective homeowners, real estate developers, investors, banks, governments, tax assessors, insurers, and mortgage lenders (Frew and Jud 2020; Ihre 2019). House price in Ireland has been highly volatile due to several factors such as availability, material cost, population density, and rising rents (Reddan 2018). Several prospective homebuyers are struggling to secure enough money for their first home (Coughlan 2022). Having an accurate prediction regarding the house price will help these buyers to plan for their first home. Different machine learning models such as Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT) can predict house prices considering several factors that are not considered while using traditional methods for house price predictions.

While traditional methods use factors such as type, size, quality of finish, location, number of floors, area of the property, availability of facilities such as schools, hospitals, and grocery stores; modern machine learning algorithms can factor in several more such as inflation, salary, environmental factors, and geography (Hurley et al. 2022). Furthermore, machine learning will help to identify house price determinants that are selectively applicable for the stakeholder who will be using these predictions. However, individual contributors that influence the house price will be different for each state and country. So, it is always a benefit for the stakeholders to know the changes that are predicted to happen soon to the property market to act accordingly.

The house prices in Ireland have been increasing every year and the prospective homeowners are finding it difficult to find a home within their budget. Ireland faced a construction boom with wage growth, bank credit and rapid increase in property prices in the early 2000s (Jose Doval Tedin et al. 2020). However, from 2007 to 2013, during the crisis time the house prices decreased sharply by almost 50%. As the economy recovered, the prices increased from 2013 onwards and by 2020 the rents have reached 32% higher than the previous years (Jose Doval

Tedin et al. 2020). Even though the house prices have hit a new record, beyond the records during the Celtic tiger years (FitzGerald 2007), several investors are considering properties as a long-term investment, which is further driving the house prices. Most of the model that predicts housing prices are not defining the investment aspect of owning a house as one of the variables that can affect the housing prices. Traditional methods have totally ignored the trends that are observed in major cities like Dublin, Manchester, Barcelona where vulture investors are investing in properties instead of companies, which is one of the highly influential factors in driving the house prices (Petrov 2009). COVID-19 pandemic has shown a few factors that can drive the house prices by 50% (Sullivan 2021). Disruptions in supply chain and unavailability of the work force has also been a factor that has to be considered while using different methods for predicting house prices. COVID-19 pandemic has provided the opportunity to work from home that caused less expenses in daily commute, coffee, and other expenditures that in turn added to the savings for buying houses. When people started working from home, proper workspace became a necessity. This has led to owning houses farther from cities but are with better facilities. Different countries have reported this to be a factor to drive house prices in rural areas of the countries. Also at least a few people have considered the opportunity to work from home as a factor to consider buying houses that are far from cities but have more facilities (Sullivan 2021). These factors have affected an indirect increase in housing prices, especially in Ireland. During the COVID-19 pandemic people were unable to spend money on holidays, dining out, outdoor activities, entertainments like movies and concerts, and instore shopping which accumulated into the savings and thereby increasing the buying power. Thus, those savings were converted to long term investments and a major portion of that investment was in real estate which drove the property prices even further. Furthermore, a country like Ireland which highly depends on external sources for building materials that are being brought to Ireland by freight will be affected and house prices will be highly influenced by any changes that will be seen in supply chain (Gazette Desk 2021). Along with these conflicts between countries can drive the material cost and fuel cost that will affect the house prices and these events are highly unpredictable (MacFarlane 2022). In this scenario, it is important to know the trends in the property market and to understand the factors influencing the property prices.

Machine learning algorithms can assist us in predicting the housing prices while considering all these factors. These models can be trained using data that has all the mentioned attributes. Generally, scholars have used hedonic model, Linear Regression (LR), Random Forest (RF), and Support Vector Machine (SVM) algorithms for predicting house prices (Ja’afar et al. 2021). The data used for training machine learning models will be a key factor in determining the accuracy of the model. Historically, Ireland has large database in relation to price and traditional factors that affect the price of property. Having a model that can be trained with this data will result in better accurate predictions for the future. Along with this the same model can be expanded to be used in different countries that are having similar structures.

This research will be based on both statistical and machine learning methods to analyse and predict the property prices of Ireland from 2010- 2022. For statistical analysis I will be using hedonic regression, linear regression, and ANOVA methods. Machine learning methods such as SVM, DT, XG-Boost, and Linear regression (LR) algorithms will be employed to predict the property prices. Hedonic model is one of the most popular methods used for house price prediction and it considers the house as a combination of many attributes (Limsombunc et al. 2004). The main goal of hedonic model is to estimate the contribution of different attributes to the price of property (Montero and Fernández-Avilés 2014). SVM has been considered by many researchers since it works well with high dimensional data, unstructured and semi-structured data. Also, the outliers have less influence on the SVM, and larger amount of data can be modelled using SVM (Advantages of Support Vector Machines (SVM) n.d.). In addition to this SVM has been found extremely popular in commercial field for predicting the sales of the company (Ho et al. 2020). When it comes to RF, it is been used for prediction in many applications due its ability to reduce the over fitting (Ho et al. 2020). RF also provides higher accuracy compared to other models and it can also deal with the larger datasets (Mbaabu 2020). Compared to other algorithms, DT requires less time for data preparation while pre-processing and data normalization and scaling is not required for it. Both classification and regression problems can be solved using the DT (K 2019). In XG-Boost algorithm, it has the in-built ability to deal with the missing values and is faster than the gradient boosting machine (GBM). It is also referred as the regularized form of GBM because it has in-built Lasso regression (L1) and Ridge regression (L2) which reduces the overfitting (Kumar 2019). Linear regression is considered as one of the simplest algorithms and the over fitting problem in modelling can be avoided by using the regularization and cross validation techniques. It also works well with systems that has less computational power and has a noticeably lower time complexity (Waseem 2022). The research will be based on the property sales data of Ireland from 2010 – 2022 and it is collected from the Residential Property Price Register page of the PSRA website. The research scope and goals of the research are listed in the Table 1 and the core technologies used in the research are mentioned in Table 2.

Research questions	Project Goals
1. Has the property prices increased over the years?	• Find out the property prices in each year. • Find out the counties in which most of the houses are sold. • Find out the counties in which the least number of houses sold. • Find out the house prices in each county/province. • Find out the year in which most houses were sold. • Find out the trend in the property pricing over the years. • Find out either the New or secondhand dwelling has the highest prices and is it inclusive of VAT. • To analyze how the house prices depends on the type of house and its size.
2. Which county has the highest prices for properties?	
3. Does the property price depend on which county the house is in? Is there any relationship between the counties and the prices?	
4. Is there any relationship between the size of the properties, type of properties (like new or second hand) and the property prices?	

Table. 1. Research questions

Technology
Programming language: Python (Jupyter Notebook, VS, Spyder)
Python libraries: • Pandas • Numpy • Matplotlib • Seaborn • scipy.stats • statsmodels
For visualization: Seaborn, Matplotlib, Plotly
For version control system: GitHub.

Table 2. Technology used for the research

This paper is structured as follows; section II gives the related works, the life cycle of the project is described in section III, data exploration is presented in section IV, methods using for the later research are mentioned in section V, the section VI the results give the outcome of the initial analysis performed, ethical considerations are mentioned in section VII, conclusion in section VIII and finally references are given in section IX.

II. Literature Review

Several studies have been performed in the past to analyze and predict the real estate and residential property prices to aid the customers as well as the developers. Hurley & Sweeney (2022) studied the impact of post codes and address in Irish property prices based on the data from January 2018 to November 2018. They focused on analyzing the property prices in Dublin with the dataset of 5028 properties for the development of geospatial statistical models where the post codes and addresses are being mislabeled. They also performed text mining to create additional variables that describes the features of the properties and its surroundings. A spatial hedonic regression model was used to separate the spatial and non-spatial contributions of property features to the resale value. Generalized additive models (GAM), regression kriging and Geographically Weighted Regression (GWR) have also been used since these methods provide greater interpretability with smaller data requirements. Three different Machine learning (ML) methods like Decision Tree, Random Forest and K-nearest neighbor algorithms were also applied on the data to evaluate how these ML models perform with smaller datasets.

Their models shown a reduction in median absolute percentage error with an increasing model complexity i.e., 12% for the hedonic model and 9.6% for linear model with spatial surface. And the authors also stated that although ML models are widely used for property prices prediction, they did not have the probability-based uncertainty intervals and the interpretability of the statistical spatial models. They also added that the random forest models may not fit well with the areas outside Dublin since the housing turnover is low outside Dublin and in that case statistical spatial modelling can work more efficiently. In Ireland, where property valuations are currently based on comparison to recently sold neighboring properties, the authors claim their model has higher applicability and will help to improve property tax computations and site value estimates because the model is not only based on the property value but uses the spatial location scaling (Hurley et al. 2022).

Machine learning models were used to predict house prices in Godavari district of Andhra Pradesh, India by Thamarai and Malarvizhi (2020). The models were built to help the people buy suitable houses for their needs. Decision tree regression, decision tree classification and linear regression models were performed on the data based on the attributes of the property like number of bedrooms, age of the house, availability of school near the house and shopping malls available nearby the house location. Attribute selection algorithms were used to remove the redundant features to reduce the impurity in the process before splitting the data for modelling.

To predict the availability of houses according to the requirement of the user, they used the decision tree classifier which gives responses like yes or no to show whether a house is available or not. Along with this, regression methods like decision tree regression and multiple linear regression were used to predict the prices of the houses. The dataset used for the modelling was a real-time data acquired from the Godavari district with all the attributes of the house and modelled using the scikit learn, a machine learning tool.

The main dataset was divided into train and test data and the decision tree classifier is performed using the training dataset. The accuracy of the model is then checked using the test data. Mean Squared Error (MSE), Mean Absolute Error (MAE) and root mean squared error (RMSE) were used to evaluate the performance of both the classification and regression models. The house price prediction with decision tree algorithm produced an output with some data record prices predicted with lesser deviations. From the multiple linear regression, it is found that the number of bedrooms is the feature having high influence on house price and age of the house is the feature having less influence on house price. From the performance metrics it is found that the prediction of house price using multiple linear regression has higher performance than the prediction using the decision tree regression. The authors also stated that the developed model can be used to predict the availability and prices of houses for any new attributes according to the users and the overall accuracy can be increased in the future with a large dataset and by identifying the best features.

Quang et al. (2020) conducted a study to compare housing prices prediction using traditional and advance machine learning methods. They used three different machine learning models: Random Forest, XGBoost and LightGBM. Further, two machine learning techniques are used, Hybrid Regression and Stacked Generalization Regression for prediction. For the analysis, the Beijing house price data from 2009 and 2018 was used and feature engineering was performed to select the appropriate features. In addition to this, exploratory data analysis was done to discover the patterns in the data. The machine learning models were applied on the training data which is split from the actual data and Root Mean Square Logarithmic Error (RMSLE) was used for evaluating the performance of the models. Results from the study shows that the Random Forest method is prone to overfitting even though it has lowest errors. The Hybrid regression method is better performing among all three methods. They also stated that although the Stacked regression method has the worst time complexity, it is the best choice when accuracy is considered(Quang et al. 2020).

To understand the factors influencing the property prices, Decision Tree (DT), Random Forest (RF) and linear regression (LR) methods were used by (Yee et al. 2021) using the Malaysian dataset from April 2017 to December 2019, to help the buyers and sellers who need to finance in the property market. The accuracy of the models is evaluated based on the R squared, RMSE and MAE values. Natural Language Processing was used to transform the data so that it is readable by the machine. The outcome of the study explained that the RF is the better model with higher accuracy compared to the DT and LR.

Alen Ihre (2019) discussed the machine learning algorithms K-Nearest Neighbour (K-NN) and Random Forest (RF) regression to predict the house prices using the Ames dataset of 3000 observations. Five-fold cross validation was performed on the dataset to minimize the bias and grid search algorithm was utilized to select the best number of hyperparameters for the prediction. Finally, the RF was found to be the best performing model based on the MAE values. However, there is small differences in the actual price and the predicted price, and the author suggested that the results could be improved by using a larger and less biased dataset (Ihre 2019).

Artificial Neural Network (ANN) is used for analysing the real estate price by (Shi and Li 2009) to evaluate the house price determinants. An improved Genetic Algorithm (GA) was used to optimize the weights of the neural network. The results of the study shown that the GA-ANN is more capable of determining the house price determinants more time efficiently and the errors of the HGA-ANN model was found to be lower than the back propagation (BP) and the genetic algorithm (GA) models.

Support Vector Machine (SVM) has been employed for predicting and forecasting house prices by Mu et al. (2014), Phan (2018), Ho et al. (2020), Chen et al. (2017), Gu et al. (2011) and Wang et al. (2014).

House value forecasting based machine learning methods by Mu et al. (2014) aimed at helping the developers and government to take decisions regarding developing real estate in the Boston area. The authors collected the data from the UCI data sets and support vector machine (SVM), least squares support vector machine (LSSVM), and partial least squares (PLS) algorithms were applied on the training data to predict the housing value. From the prediction results it is found that the SVM and LSSVM has better efficiency with the nonlinear data. PLS algorithm is better for linear data due to the simplicity of the algorithm. They also added that to achieve best forecasting effect and an optimal solution, SVM can be used (Mu et al. 2014).

Phan (2018) utilized the SVM technique to predict the house prices in the Melbourne city of Australia to help the house buyers and sellers. Neural network (NN), Polynomial regression, Linear regression and Regression Tree models were also developed along with SVM to identify the best fit. The data used in the study was downloaded from the Kaggle website and it has the house sold houses transaction from 2016 to 2018. Data imputation and descriptive analysis techniques were performed on the data prior to modelling. Along with this, principal component analysis (PCA) was also performed to select the desired features and stepwise method was utilized for subset selection. The results shown that the SVM with the subset selection method gives the best efficiency with lower errors. When regression tree and linear regression delivered almost equal prediction result, the polynomial regression gave better accuracy with lower errors. The neural network seemed to be not working well with the available dataset.

The authors also stated that regression tree and neural network worked more faster than the SVM, where PCA with SVM took more time than SVM with stepwise (Phan 2018).

SVM methods were employed by Ho et al. (2020) for predicting the property prices in Hong Kong. Random forest (RF) and Gradient Boosting Machine (GBM) were also utilized along with SVM to compare the performance of algorithms. The dataset used was a sample data with over 40000 housing transactions in a time of 18 years. Correlation matrix was used to determine the features which should be included in the models. The results from the performance metrices revealed that the RF and GBM were able to estimate the house prices better than the SVM with smaller errors. They also found that the ML algorithms need more computation time than the traditional Hedonic pricing model and among the three models used, SVM is the better choice for forecasting when speed is the priority and RF and GBM should be used if the accuracy is considered (Ho et al. 2020).

To predict the housing prices in the Taipei city of Hong Kong, SVM models were implemented by Chen et al. (2017). By using stepwise multi regression, the support vectors were found, and a SVM hedonic price model was built using the support vectors, the structural and the spatial variables to predict the house prices in the Taipei city. The SVM model is then developed based on the identified support vectors to forecast the future housing prices for the data from 2007 to 2010. One of the advantages in using SVM is that it is not depending on the probability distribution assumption and hence it could plot the input variables into a high dimensional feature space. To compensate the bias variance trade-off, five-fold cross validation has been used for testing and training in the analysis. The outcome from the study points out that the SVM can be considered as a superior approach which legitimise the issues in the multiple regression analysis and combining the hedonic approach with SVM is feasible for non-linear modelling (Chen et al. 2017).

Gu et al. (2011) used the SVM methods along with hybrid genetic algorithm methods to forecast the house prices in China. SVM has proven to be one of the best algorithms in both classification and regression in lots of applications. In the study, genetic algorithm (GA) has been used instead of grid algorithm to optimize the parameters of the SVM since, GA is more time efficient and the G-SVM is developed. The results of the study revealed that the G-SVM method giving more accuracy than the Grey Algorithm (GM) which is used in the past to predict the house prices (Gu et al. 2011).

Machine learning algorithms has used to forecast the real estate prices by many researchers and Wang et al. (2014) used SVM to forecast the real estate price with particle swarm optimization (PSO) in the Chongqing city in China. One of the reasons to choose SVM is its ability to conquer the ‘Curse of dimensionality’. To identify the parameters of SVM, PSO method is used instead of GA and grid algorithm since it is easy to enforce. The actual data was divided into train and test data for the modelling. The study shown that the PSO-SVM model is performing better than the BP neural network methods used by other researchers.

PSO has also used by Alfiyatin et al. (2017) for predicting the house prices to help the builders to decide the selling price of the house and to help the buyers to set the right time to buy the house. PSO and regression analysis was implemented on the houses data in the Malang city of Indonesia within 2014-2017. Hedonic regression was chosen as the regression prediction model and PSO to select the appropriate features. The error prediction values are found to be higher for the regression model compared to the PSO-regression model. Hence the study proved that the combination of PSO with regression can give the minimum prediction error.

Support Vector Machine Regression (SVR) has used by several researchers to predict the real estate and housing prices due to its efficiency and application. A SVR model was used by Li et al. (2009) to understand the possibility of predicting real estate prices in China from 1998-2008. The results of the model are compared with the Back Propagation Neural Network (BPNN) model to analyze the performance of the model. Several indicators like CPI, loan interest rate, real estate price, real estate investment, income etc. were used to forecast the real estate prices. The analysis demonstrated that the SVR model works better than the BPNN model for real estate forecasting based on the MAE, MAPE and RMSE. The study also proved that the SVR method is an efficient approach to forecast the real estate price (Li et al. 2009).

SVM regression with Gaussian kernel was utilized by Miao et al. (2021) to predict the property prices in Boston area. The aim of the study was to help the people to estimate the price of the house based on the properties of the house. They selected the most important features using the decision tree with ID3 algorithm, divided the data to train and test data and SVR is employed on the data with the gaussian kernel to predict the prices and compared the model with different regression methods to analyze the performance. The study proved that the SVR with gaussian kernel has more efficiency than the KNN, decision tree and SVR with linear kernel. But the model still has some drawbacks which are mentioned by the authors and the important one is that some of the factors that impact the house prices cannot be measured such as the cultural tolerance of the neighbors and so on.

Jiao Yang Wu (2017) also performed a study to analyse the house price prediction using SVR based on the housing sales data of Kings County, USA with an aim to help the buyers and sellers. Feature selection methods like Random Forest selector, Lasso ridge and Recursive Feature Extraction and the feature extraction method PCA is done prior to building the model. Further to this, parameter tuning, and transformation techniques have been used to improve the accuracy of the model. But the results from the experiments shown that there is not much difference in the performance of the model with feature extraction and feature selection. The Radial Basis Function (RBF) kernel with SVR was found to be the best one among the performed combinations. The author also pointed out that in future other machine learning models like XGBoost and other feature engineering methods can be applied on the data (Wu 2017).

Hedonic Regression is one of the estimation and prediction method preferred by the researchers when it comes to prices. It is used in most of the scenarios when a price variable is to be considered. Property prices in Croatia is studied by Kunovac and Zagreb (2019) using Hedonic regression based on the data collected from different sources. One of the goals of the research was to propose how the hedonic models can be used for the evaluation of residential property. The hedonic model built allows the evaluation of some attributes of the property such as age, location and so on. The results of the analysis shown that the micro location of the property should also be considered to the hedonic models and to commonly used other models to improve the prediction of residential house prices.

Abdulai and Owusu-Ansah (2011) used hedonic regression to determine the house price determinants in Liverpool, The United Kingdom over a period of 18 years from 1990-2008. They also analysed how the past and present buyers valued the property features. The regression model was able to explain almost 75% of variation in the housing prices and all the variables in the dataset was found to be statistically significant. Another thing they found was that the price of the new properties is almost double than the old properties and the detached houses are more expensive than the flats. When the past buyers before 1999 focused more on the number of bedrooms, bathrooms and detached houses, the buyers after 2000 more value the number of floors, gardens, and showers (Abdulai and Owusu-Ansah 2011).

Selim (2008) identified the house price determinants in Turkey using hedonic regression using the 2004 household survey data. Environmental factors were not included in the data for analysis and natural logarithm of the house price is treated as the dependent variable. To estimate the hedonic model, ordinary least square method is utilized. The heteroscedasticity present in the model was eliminated using the White's heteroscedasticity consistent coefficient covariance matrix. And the results of the analysis shown most of the variables to be significant and the house prices seems to be higher for houses with more rooms. The variables such as water system, pool, type of the house, number of rooms, size of house, locational attributes and building structure are the found to be the most significant variables that influence the house prices.

Limsombunc et al. (2004) compared the hedonic regression with Artificial Neural Network (ANN) model on house price prediction based on a sample dataset of 200 New Zealand houses. The age of the house, size of house, bedrooms, bathrooms, and other features were considered for the experiment. The heteroscedasticity consistent coefficient covariance matrix and weighted least squares method were used instead of the ordinary least square method to eliminate the heteroscedasticity issues. However, the heteroscedasticity problem was not completely removed. The results from the study revealed that even though the hedonic model was able to explain almost 70% of variance in the model, it did not outperform the neural network model. The authors mentioned that the small dataset and the lack of environmental features may be some reasons for the poor performance of the hedonic model.

Following researchers have used supervised and semi-supervised ML models for predicting property and real estate prices (Ihre 2019; Phan 2019; Quang et al. 2020; Mu et al. 2014; Ho et al. 2020; Chen et al. 2017; Gu et al. 2011; Wang et al. 2014; Pow and Janulewicz 1995; Alfiyatin et al. 2017; Li et al. 2009) and have employed the RMSE, MAE, MSE, MAPE, and R Squared performance metrices for evaluating the models. Either regression or classification supervised learning techniques are utilized by most scholars for prediction. There are certain algorithms which are preferred by several scholars such as the RF, DT, LR, K-NN, XG-Boost, SVM and ANN algorithm (Ja'afar et al. 2021).

From the previous studies the most preferred algorithm is the RF. It is one of the ensemble methods used in machine learning and it is generally utilized for predicting the real estate prices due to its less error compared to other algorithms (Shinde and Gawande 2018). Another most chosen algorithm is SVM, and it is applied in both regression and classification problems. Since SVM can better deal with bias and can use high dimensional data, it is used as recognition in classification problems (Ja'afar et al. 2021). DT, XG-Boost and LR are also preferred by the researchers because, compared to other algorithms, DT requires less time for data preparation while pre-processing and data normalization and scaling is not required for it. Both classification and regression problems can be solved using the DT (K 2019). In XG-Boost algorithm, it has the in-built ability to deal with the missing values and is faster than the gradient boosting machine (GBM). It is also referred as the regularized form of GBM because it has in-built Lasso regression (L1) and Ridge regression (L2) which reduces the overfitting (Kumar 2019). Linear regression is considered as one of the simplest algorithms and the over fitting problem in modelling can be avoided by using the regularization and cross validation techniques. It also works well with systems that has less computational power and has a noticeably lower time complexity (Waseem 2022).

III. Life Cycle Of Project

The project is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to indicate Machine Learning aspects and ensure a successful execution of the project. The whole life cycle of the project consists of six stages: data collection, data preparation, data exploration, data modelling, evaluation, and deployment.

- Data Collection: The data was collected from the Residential Property Price Register website, which is available to the public for download. The data has the property selling details of Ireland from 2010 - 2022 with the address, post codes, county, prices, and several other attributes.
- Data Preparation: Data preparation involves the cleaning of the data including removing the data recorded in the Irish language, dealing with the missing values, and adding extra data such as province and latitude/longitude details to the available data.
- Data Exploration: Exploratory data analysis and summary statistics were performed to explore the patterns in the data. Data visualizations were performed on the data using different plots to understand the trend of the house pricing.
- Modelling: Statistical and machine learning techniques were employed on the data to predict the house prices. Dependent variable was selected from the data prior to modelling.
- Evaluation: To evaluate the models, performance metrices like MSE, RMSE and MAE were used and along with this, graphs were used to identify the patterns and outliers in the data.
- Deployment: In deployment stage, the review of the project is performed, and future improvements are suggested.

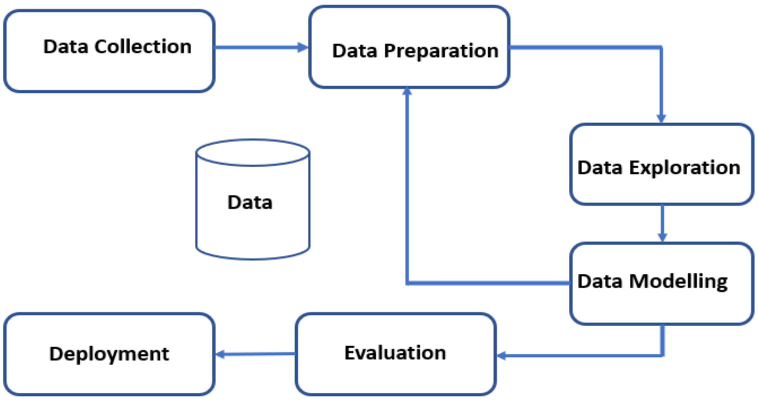


Figure. 1. Life Cycle of Project based on CRISP-DM

IV. Data Exploration

A. *Data Collection*

The data is collected from the Residential Property Price Register page of PSRA website which provides the residential property sales data of Ireland since January 2010. The data is collected as csv files for the ease of the analysis. Prior to this, an attempt was made to collect data from daft.ie, which is a property sales application of Ireland. Unfortunately, it failed since they were not willing to provide their data to a third person. The data from Kaggle and CSO was found to be insufficient due its size and incomplete information. Finally, the data is collected from the https://propertypriceregister.ie. website which has sufficient size and information for the analysis.

B. *Data Preparation*

Data wrangling is performed on the collected data and outliers and missing values are treated for further analysis. To expand the actual data, province, latitude & longitude, and location columns were added. The province details were collected from the 'IrelandTownList' website as a csv file, and it is then merged to the actual dataset. To access the latitude and longitude, an API is needed, and it was created through the https://opencagedata.com/ which is a free and secured website for creating free and paid API keys. The first option for creating API key was Google API but it is paid and not affordable so, I couldn't use it. Python Geopy was also tried for getting the latitude and longitudes however, the Jupiter Notebook was not able to process all the data and hence that attempt was also in vain. Finally, API is created through 'opencagedata' website, and the latitude and longitude of each county are accessed and merged to the data. This is because, only 2000 requests can be processed in a single a day using the free API and the dataset has more than 500000 rows which makes it impossible to process the entire data. Another column 'location' is added to the data which contains the details of the place of sale either as 'Dublin' or 'Outside Dublin' for the analysis. Unemployment, income, and homelessness data were supposed to be added to the actual data for further analysis unfortunately, these data were not available for all the years since 2010 and as per county and thus those details were not considered for analysis.

The combined data was then cleaned using Python pandas library and missing values are treated. There are many missing values in the data, but they are kept same since removing them will affect the entire analysis and modelling. The duplicates were checked, datatypes were corrected and Irish text in the data were converted to English using Python. The address values were changed to string title as some of the values were in uppercase. Additional columns were created for date variable with month and year for further analysis and visualization.

C. *Data Description*

The dataset contains the details of the residential property sales of Ireland from 2010 - 2022. It was collected from the Residential Property Prices Register website and has 516586 rows × 9 columns. The data has the date of the sale, the price of the property, county, and other related information about the sale. After removing duplicates and adding the additional columns there were 515792 rows × 15 columns. Table. 3 describe the variables in the dataset. The variables month, year, province, location, latitude, and longitude were additionally created for the analysis.

Type of Variables			
Variable Name	**Category**	**Type**	**Description**
date_of_sale	Date	Datetime	The date of the property sale
address	Nominal Categorical	String	The address of the sold property
postal_code	Nominal Categorical	String	The postal code of the sold property
county	Nominal Categorical	String	The county name of the sold property
price	Continuous Numerical	Float	The price of the sold property
FMP	Nominal Categorical	String	The information about whether the sold property price is full market price or not.
VAT_exclusive	Nominal Categorical	String	The information about whether the sold property price is VAT exclusive or not.
property_description	Nominal Categorical	String	
property_size_description	Nominal Categorical	String	
province	Nominal Categorical	String	The province name of the sold property
lat	Continuous Numerical	Float	

lon	Continuous Numerical	Float	
location	Nominal Categorical	String	The information about whether the sold property is in Dublin or outside Dublin.
year	Discrete Numerical	Integer	The year in which property is sold.
month	Discrete Numerical	Integer	The month in which property is sold.

Table. 3. Variables in the dataset

V. Methods

To analyze and predict the property prices different machine learning as well as statistical techniques will be utilized in the research and this section will give an overview of those methods. The proposed work will be implemented using scikit learn, a machine learning and statistical modeling tool in Python.

A. Scikit Learn

Scikit learn is a useful Python library for machine learning and it contains different tools for machine learning and statistical modelling including classification, regression, and clustering. It is developed upon the Scientific Python (SciPy) which must be installed before using scikit learn (Brownlee 2014).

The stepwise execution of a model using scikit learn is as follows (Andrade 2021).

1. Import the required libraries.
2. Load the required dataset.
3. Split the data into train and test data.
4. Fit the model into the data.
5. Predict the dependent variable for the test data.

B. Linear Regression

Linear regression is a supervised machine learning algorithm which identifies the best fit linear line between the dependent and independent variable. In other words, it identifies the linear relationship between the dependent and the independent variable. There are two types of linear regression, simple linear regression, where there is only one independent variable and multiple linear regression, where there are multiple independent variables (Deepanshi 2021).

Simple linear regression equation is given by, $y = b_0 + b_1x$, where b_0 is the intercept, b_1 is slope, x is the independent variable and y is the dependent variable.

Multiple linear regression equation is given by, $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \dots . + b_nx_n$, where b_0 is the intercept, b_1,b_2,b_3…,b_n are slopes of the independent variables x_1,x_2,x_3,x_4…,x_n and y is the dependent variable.

The basic assumptions for linear regression are as follows,

1. There should be a linear relationship between the independent and dependent variable. This can be verified using scatter plot between the variables.
2. The independent and dependent variables should be normally distributed. This can be identified using the KDE plots and histograms.
3. The spread of the residuals should be constant for all values of independent variable, which can be verified using the residual plot.
4. There should be no correlation between the independent variables and the error terms should be normally distributed.
5. There should be no autocorrelation.

When the assumptions are violated, it leads to the decrease in accuracy which in turn affects the predictions and makes the high errors (Deepanshi 2021).

C. *Support Vector Machine*

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression problems. The main goal of the SVM is to find a hyperplane in an n-dimensional space where n is the number of features that classifies the data points with maximum distance. SVM is built using the support vectors, which are the data points that are close to the hyperplane and are used to maximize the distance (Gandhi 2018; Ray 2017).

D. *Decision Tree*

Decision Tree (DT) is a supervised machine learning algorithm which can be utilized for both classification and regression problems. It is a tree like structure with root node and then branches into solutions as like in a tree. There are root nodes, decision nodes and leaf nodes in a DT where, root nodes are the beginning node of the DT and the decision nodes are the nodes which we get after splitting the root nodes and the leaf nodes, where further splitting is not possible. In DT, the overfitting can be reduced by pruning, the cutting down of some nodes (Saini 2021; Amrutha 2022).

E. *Random Forest*

It is a supervised learning algorithm used for both classification and regression problems. It has number of decision trees on subsets and takes the average of it to predict the accuracy. The higher the number of trees in the model, the lower the chance of overfitting (Yiu 2019).

F. XG-Boost

Extreme Gradient Boosting or XG-Boost is an ensemble learning method. It is an extension of gradient boosted decision trees and are designed to improve the performance of the model. XG-Boost has the regularized learning feature, which helps to reduce the overfitting (analyticsvidhya 2018).

The boosting ensemble technique has mainly 3 steps,

1. A primary model F0 is defines for predicting the target dependent y. This will be related to the residual (y-F0)
2. New model h1 is fit to the residuals from the step1.
3. The boosted version (F1) of F0 is obtained by combining F0 and h1. Similarly, new models are created after each residual to improve the performance. The mean squared error of each new model will be lower than previous model. The residuals can be minimized through m iterations (analyticsvidhya 2018).

$$F_1(x) < -F_0(x) + h_1(x)$$

$$F_m(x) < -F_{m-1}(x) + h_m(x)$$

G. Hedonic Regression

Hedonic regression is a regression method which analyses the impact of different attributes on the price of a good. In hedonic regression, the dependent variable will be the price of the good and the independent variables will be the attributes of the good that influence the price. It uses the ordinary least squares and other techniques to estimate how the attributes affect the price of real estate like house.

Hedonic regression function is defined as, pi = j (ci), where p is the price of good i, and ci is the vector of attributes related to the good. The attributes can be location, structure of the property, environmental properties, and accessibility to the property. The hedonic regression is mostly used to estimate the property prices in real estate industry (investopedia 2021; CFI 2022).

H. ANOVA Test

Analysis of Variance (ANOVA) is a statistical test which helps to find out whether the difference between different groups of data is statistically significant. It also helps us to understand the relationship between the dependent and independent variables. One-way ANOVA compares the effects of independent variable on dependent variables and two-way ANOVA does the same with multiple independent variables(Kenton 2022).

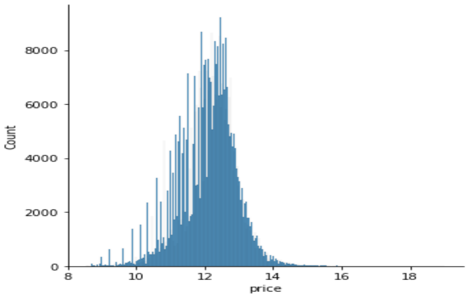
VI. RESULTS

Initial analyses and visualizations were made to understand the patterns in the data and to analyze the attributes contributing to the property prices. This section describes the outcome of the initial analyses performed.

*A. **Visualizations***

a. Univariate Visualizations

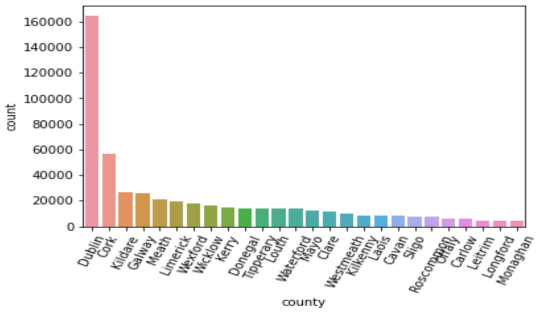
price



- There are outliers in the data.
- Some of the properties were sold at higher prices and some at very lower prices.

Table. 4. Univariate for price

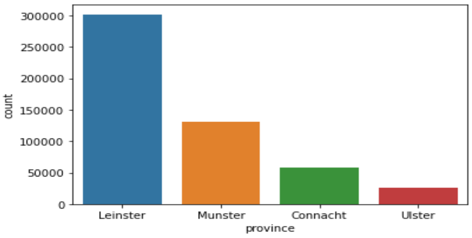
county



- Most of the houses were sold in Dublin and least in Monaghan.

Table. 5. Univariate for county

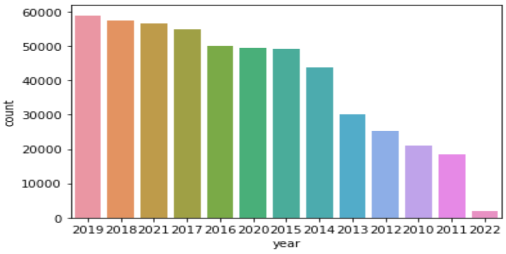
province



- Most of the sold properties are at Leinster province and least number of properties are sold at Ulster.

Table. 6. Univariate for province

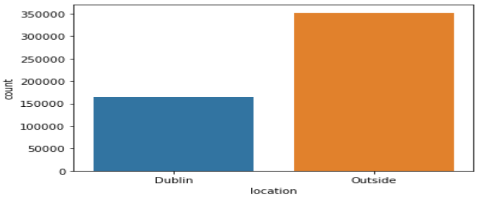
year



- Highest number of property selling was in 2019 and least selling were there in 2022.

Table. 7. Univariate for year

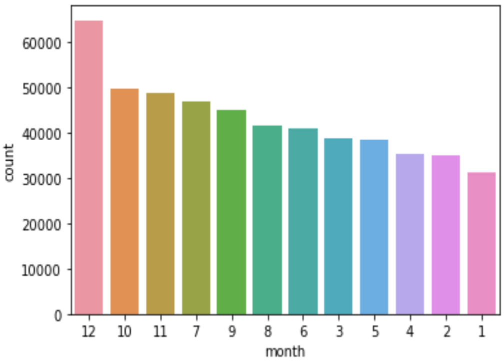
location



- Almost half of the observations account for Dublin. So, most of the houses were sold in Dublin.
- Rest half of the houses were sold in counties other than Dublin.

Table. 8. Univariate for location

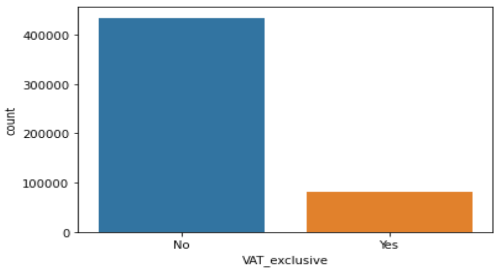
month



- Most of the property selling were in the month of December and least in January.

Table. 9. Univariate for month

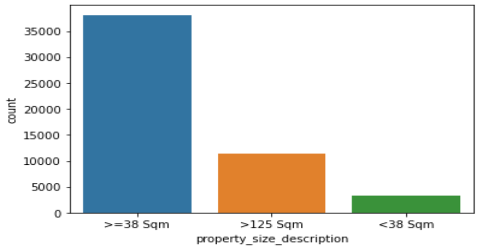
VAT_exclusive



- Most of the sold property prices were VAT inclusive.

Table.10. Univariate for VAT_exclusive

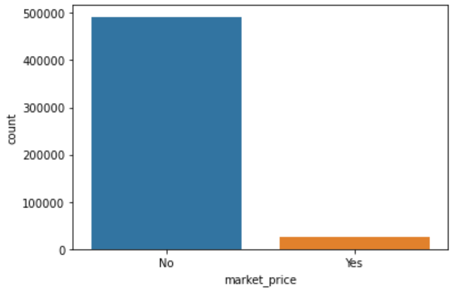
property_size_description



- Most of the sold houses were with greater than or equal to 38 Sqm and least houses with less than 38 Sqm.

Table. 11. Univariate for property size

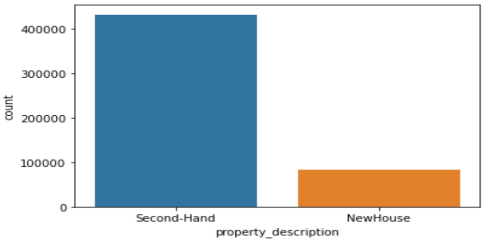
FMP



- Most sold prices were not full market price.

Table. 12. Univariate for FMP

property_description

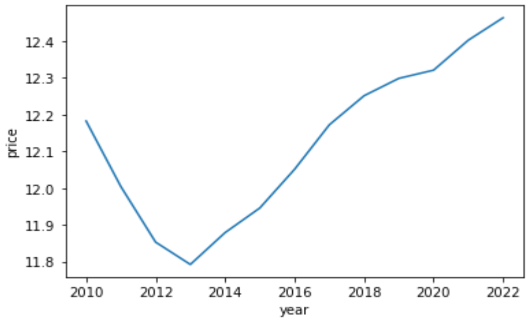


- Most of the sold properties were secondhand houses than new houses.

Table. 13. Univariate for property description

b. Bivariate Visualizations

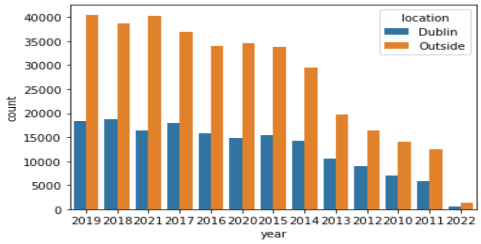
year and price



- The house price shows a decrease in the year 2013 and then shows significant increase.

Table. 14. Year-price bivariate

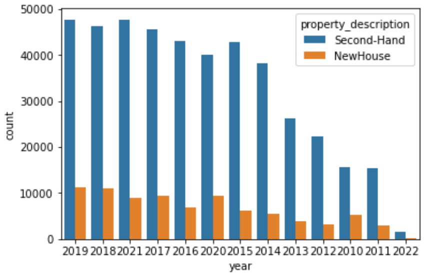
year and location



- In each year most of the sold houses are in Dublin i.e., almost half of the observations are in Dublin.
- Number of sold houses are decreasing each year.

Table. 15. Year-location bivariate

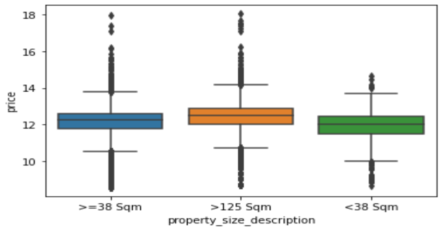
year and property_description



- Secondhand houses are sold in higher numbers than the new houses every year.

Table. 16. Year-property description bivariate

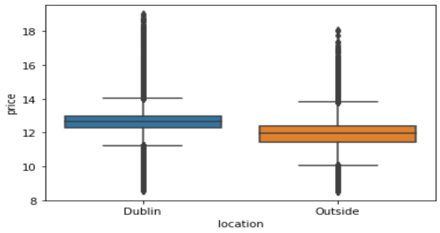
price and property_size_description



- There are outliers present.
- IQRs does not overlap, and whiskers are almost equal.

Table. 17. Year-property size bivariate

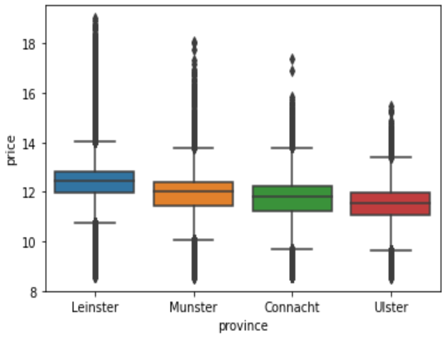
location and price



- There are outliers in the data. IQRs are not overlapping, and whiskers are of different length.
- Seems like there very high and low-price houses sold in Dublin and Outside Dublin.

Table. 18. location-price bivariate

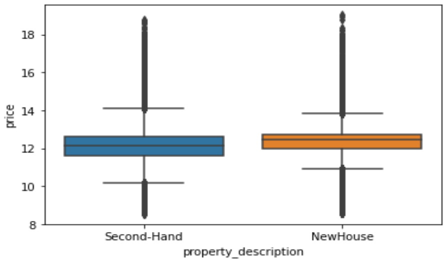
price and province



- There are outliers present.
- IQRs does not overlap, and whiskers are almost equal.

Table. 19. Year-price bivariate

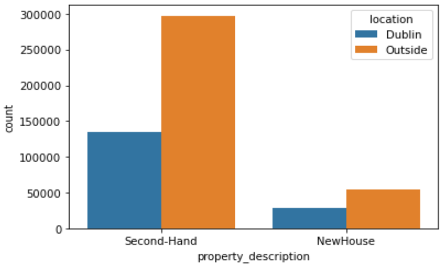
price and property_description



- There are outliers present.
- IQRs are almost overlapped, and whiskers are oof different.

Table. 20. Property description-price bivariate

location and property_size_description

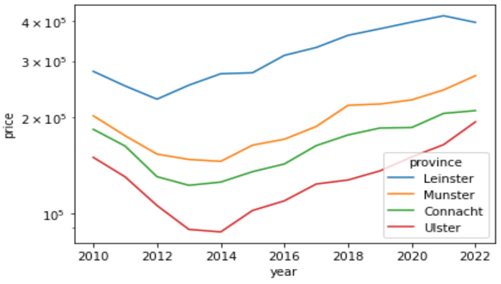


- Almost half of the sold houses were in Dublin in the case of both New and Secondhand houses.

Table. 21. location-property size bivariate

c. Multivariate Visualizations

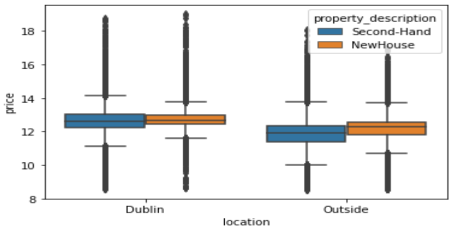
price, year, and province



- The price of the houses in each province has shown a decrease in the year 2013 and then rapidly increases in each year.

Table. 22. Multivariate1

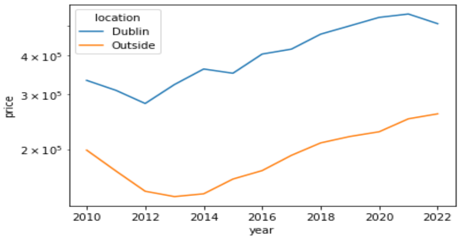
location, price, and property_description



- There are outliers present and IQRs does not overlap.
- Indicates that very high and low prices were sold in Dublin and outside.

Table. 23. Multivariate2

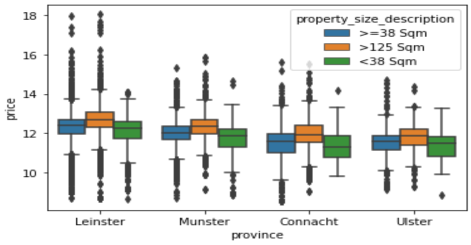
price, year, and location



- The price of the houses in Dublin and other counties has shown a decrease in the year 2013 and then rapidly increases in each year.

Table. 24. Multivariate4

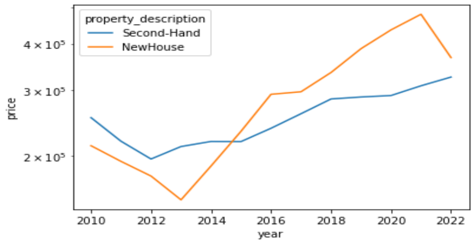
province,price and property_size_description



- There are outliers which indicates that there are houses sold in each category with very high and low prices.

Table. 25. Multivariate5

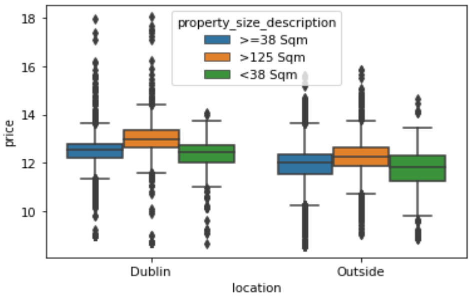
year, price, and property_description



- The price of Second-hand and new houses has decreased in the year 2013 and then increased in next years, except for new houses in the year 2021.

Table. 26. Multivariate6

location,price and property_size_description



- There are outliers which indicates that there are houses sold in and outside Dublin with very high and low prices.

Table. 27. Multivariate7

Further plots were made to understand the data for later analysis. The maximum price of sold property in each county were plotted. It is revealed that, Dublin has the sold properties with highest prices around 182 million Euro and the least sold price is in county Offaly with 1.4 million Euro.

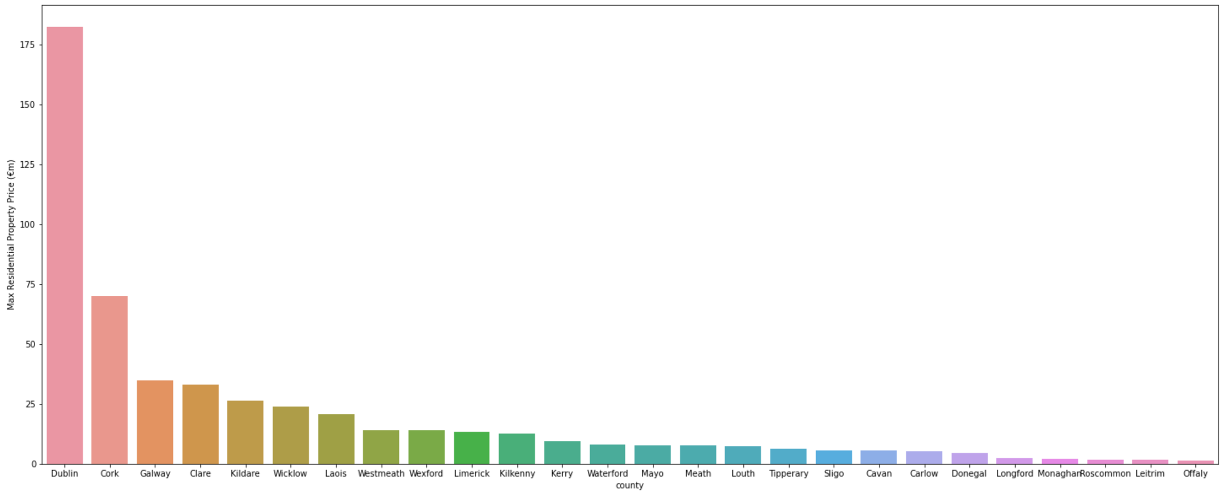


Figure. 2. Maximum property prices in counties

Similarly, the minimum sold prices in each county is plotted and it shows that, county Cork has the minimum property price of 5000 Euro and county Carlow has the minimum property price of 7000 Euro.

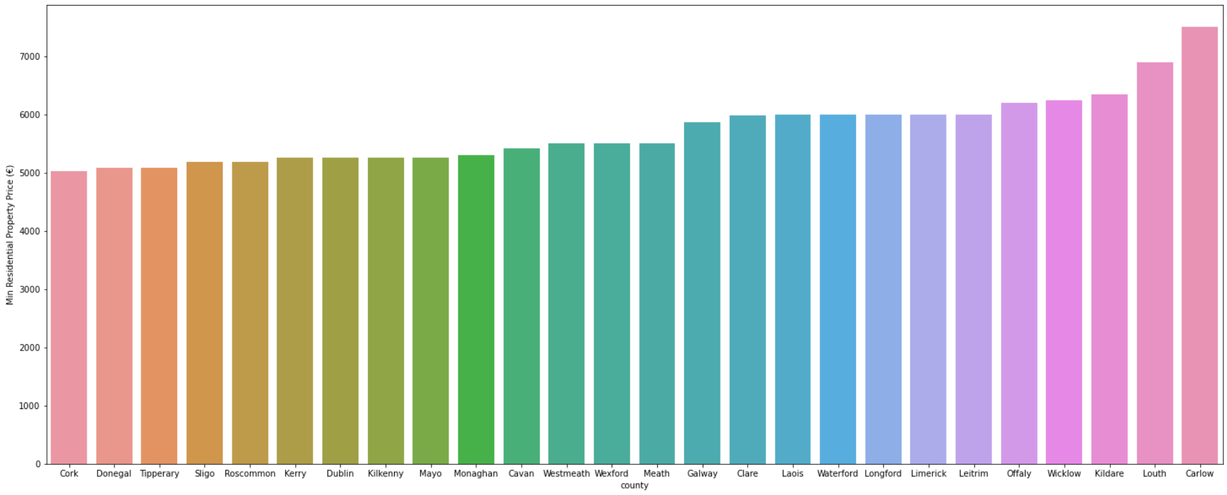


Figure. 3. Minimum property prices in counties

To understand the relation between the size and type of the property and price, a bar chart is plotted between the ‘property_size_description’ variable and the median of the prices and ‘property_description’ variable and median of the prices. From that, it is understood that the houses with greater than 125 square meters are having higher prices than the others and new houses are having higher prices than the second-hand houses.

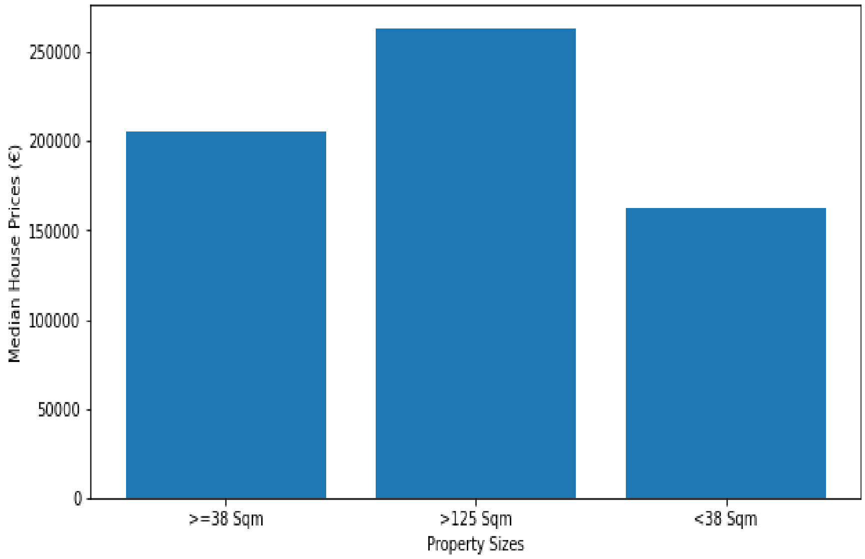


Figure. 4. Property sizes and median property prices

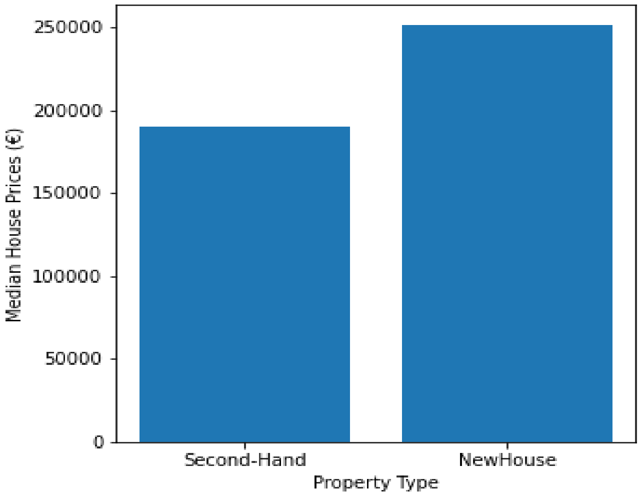


Figure. 5. Property types and median property prices

From the univariate and bivariate plots, it is evident that property prices in Dublin are far higher than other counties. So, to analyze the price of properties inside Dublin a bar chart is plot between the Dublin post codes and price. It shown that, within Dublin, the post code area Dublin 14 has the highest property prices and Dublin 10 has the lowest property prices. This implies that the post codes are also having an impact in property prices.

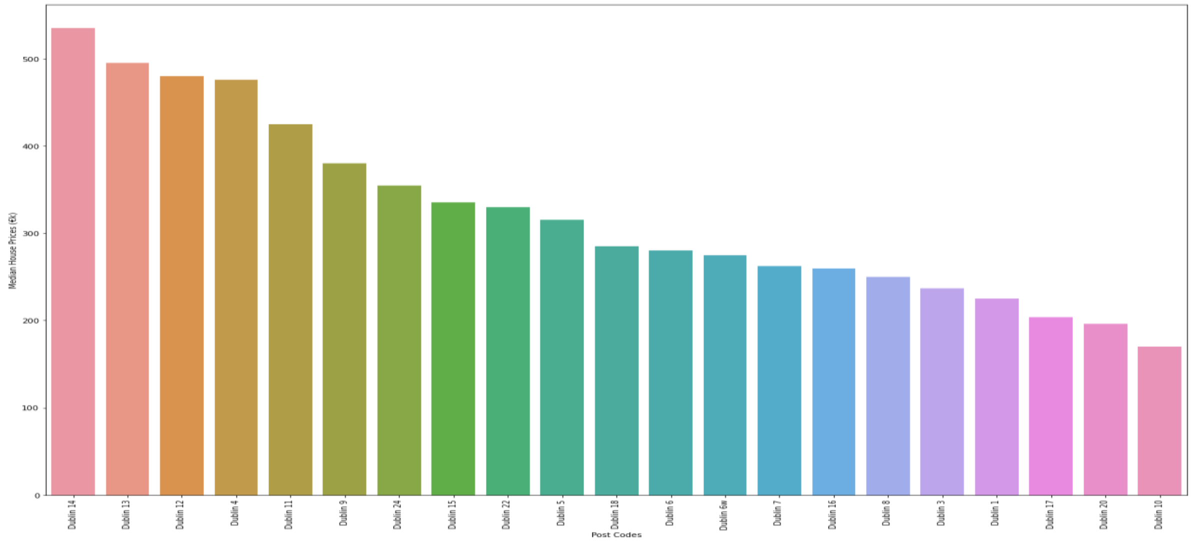


Figure. 6. Postal codes and median property prices

B. ***Initial Analysis***

a. Statistical Analysis

For statistical analyses, a random sample of 500 observations are selected from the total population to get an accurate plots and results for ANOVA. The number of samples chosen was based on an online calculation that considered 95% confidence level and 5% margin of error (Qualitrics 2022). For MLR, the entire dataset is used and even though the assumptions are not fully satisfied, I have generated the MLR model to interpret the coefficients and to estimate the significant variables. ANOVA tests were performed as a part of initial analysis to understand the relationship between the dependent and independent variables. Price variable is chosen as the dependent variable and log of the price variable was applied to meet the assumptions of the tests.

- **ANOVA Test**

1. *Full model*

The assumptions to be met for two-way ANOVA are as follows:

1. Observations should be sampled independently.
2. The variance of data in the different groups should be the same.
3. Each sample should be normally distributed.

A two-way model ANOVA is generated excluding the property size variable since it has many missing values.

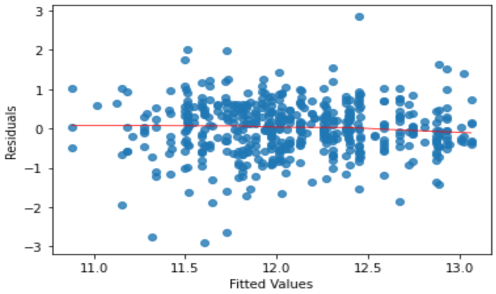


Figure. 7.

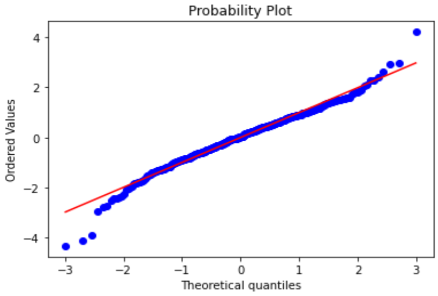
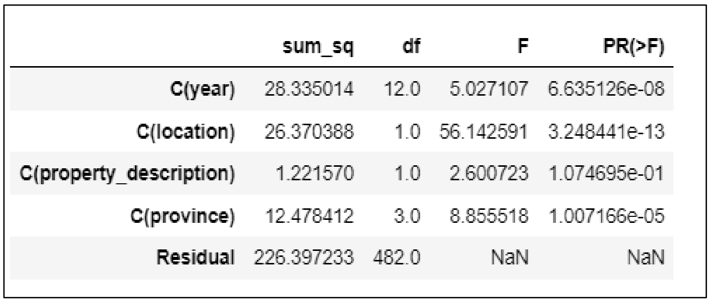


Figure. 8.



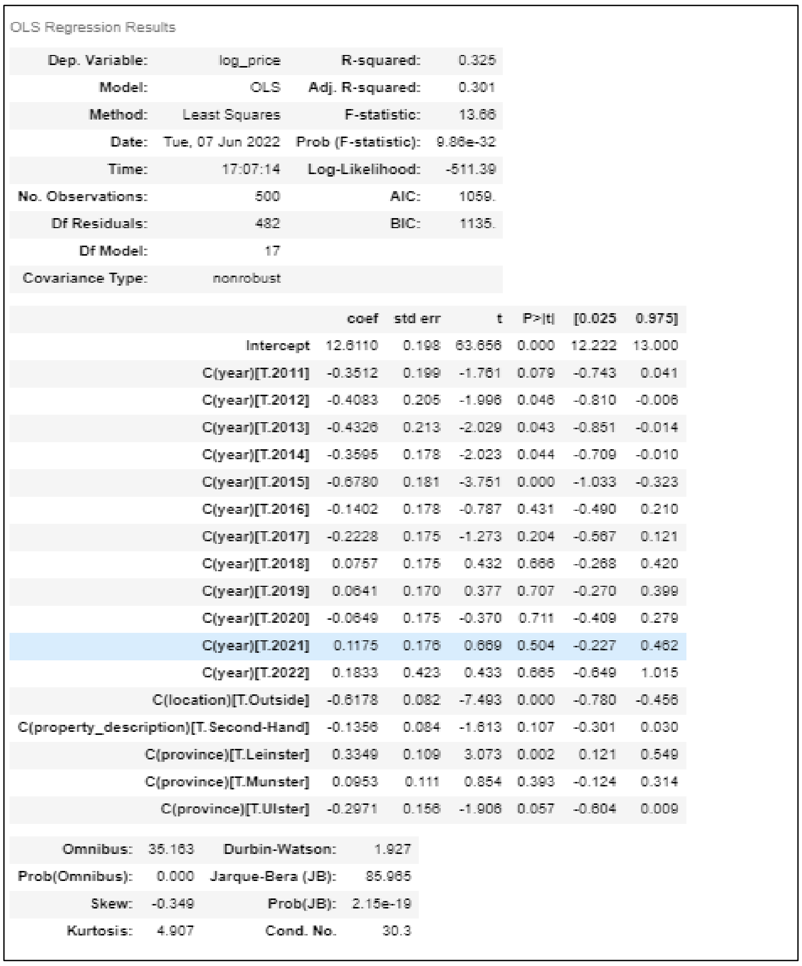
	sum_sq	df	F	PR(>F)
C(year)	28.335014	12.0	5.027107	6.635126e-08
C(location)	26.370388	1.0	56.142591	3.248441e-13
C(property_description)	1.221570	1.0	2.600723	1.074695e-01
C(province)	12.478412	3.0	8.855518	1.007166e-05
Residual	226.397233	482.0	NaN	NaN

Figure. 9. ANOVA results for full model

Interpretation of the model as per *Figure. 9.* and *Figure. 10.*

First Hypothesis: H_0: μ2010 = μ2011 = μ2012 = μ2013 = μ2014= μ2015 = μ2016 = μ2017 = μ2018 = μ2019= μ2020 = μ2021 = μ2022 (All means are equal) Where μi is the sample mean for year i. H1: Not all the means of the sample population is same. (at least on mean is different so, there is an effect with year on price) P-value = 6.6x10^{-8}< 0.05, hence, reject H0. Inference: Year is statistically significant; it is indicated that a change in year will affect the price or log price significantly.	Second Hypothesis: H_0: $\mu_D = \mu_{OD}$ H_1: $\mu_D \neq \mu_{OD}$ Where: μ_D is the sample population mean for the log price when the property location is Dublin, μ_{OD} is the sample population mean for the log price when the property location is outside Dublin. P-value = 3.2x10^{-13}< 0.05, hence, reject H0. Inference: Location is statistically significant. So, changing the location will impact the price or log price significantly.
Third Hypothesis: H_0: $\mu_{new} = \mu_{used}$ H_1: $\mu_{new} \neq \mu_{used}$ Where: μ_{new} is the sample population mean for the log price when the property is new, μ_{used} is the sample population mean for the log price when the property second-hand. P-value = 1.07x 10^{-1}< 0.05, hence, reject H0. Inference: Property type is statistically significant; it shows that changing the type will impact the price or log price significantly.	Fourth Hypothesis: H_0: $\mu_{leinster}$ = $\mu_{munster}$ = μ_{ulster} = $\mu_{connacht}$ H1: Not all the means of the sample population is same. (at least on mean is different so, there is an effect with province on price) P-value = = 1.0x 10^{-6}<0.05. hence reject H0. Inference: Province is statistically significant; it shows that changing the province will impact the price or log price significantly.

Table.28. Two-way ANOVA interpretation



OLS Regression Results

Dep. Variable:	log_price	R-squared:	0.325
Model:	OLS	Adj. R-squared:	0.301
Method:	Least Squares	F-statistic:	13.66
Date:	Tue, 07 Jun 2022	Prob (F-statistic):	9.86e-32
Time:	17:07:14	Log-Likelihood:	-511.39
No. Observations:	500	AIC:	1059.
Df Residuals:	482	BIC:	1135.
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.6110	0.198	63.656	0.000	12.222	13.000
C(year)[T.2011]	-0.3512	0.199	-1.761	0.079	-0.743	0.041
C(year)[T.2012]	-0.4083	0.205	-1.996	0.046	-0.810	-0.006
C(year)[T.2013]	-0.4326	0.213	-2.029	0.043	-0.851	-0.014
C(year)[T.2014]	-0.3595	0.178	-2.023	0.044	-0.709	-0.010
C(year)[T.2015]	-0.6780	0.181	-3.751	0.000	-1.033	-0.323
C(year)[T.2016]	-0.1402	0.178	-0.787	0.431	-0.490	0.210
C(year)[T.2017]	-0.2228	0.175	-1.273	0.204	-0.567	0.121
C(year)[T.2018]	0.0757	0.175	0.432	0.666	-0.268	0.420
C(year)[T.2019]	0.0641	0.170	0.377	0.707	-0.270	0.399
C(year)[T.2020]	-0.0649	0.175	-0.370	0.711	-0.409	0.279
C(year)[T.2021]	0.1175	0.176	0.669	0.504	-0.227	0.462
C(year)[T.2022]	0.1833	0.423	0.433	0.665	-0.649	1.015
C(location)[T.Outside]	-0.6178	0.082	-7.493	0.000	-0.780	-0.456
C(property_description)[T.Second-Hand]	-0.1356	0.084	-1.613	0.107	-0.301	0.030
C(province)[T.Leinster]	0.3349	0.109	3.073	0.002	0.121	0.549
C(province)[T.Munster]	0.0953	0.111	0.854	0.393	-0.124	0.314
C(province)[T.Ulster]	-0.2971	0.156	-1.906	0.057	-0.604	0.009

Omnibus:	35.163	Durbin-Watson:	1.927
Prob(Omnibus):	0.000	Jarque-Bera (JB):	85.965
Skew:	-0.349	Prob(JB):	2.15e-19
Kurtosis:	4.907	Cond. No.	30.3

Figure. 10. Model summary for two-way ANOVA

- **MLR**

The assumptions to be met for MLR are as follows:

1. There should be linear relationship between the dependent and independent variables.
2. The residuals should be normally distributed.
3. There should be no collinearity.

A MLR model is generated excluding the property size variable since it has many missing values. Square root transform is applied on the log value of price variable to deal with the outliers and make it symmetrical.

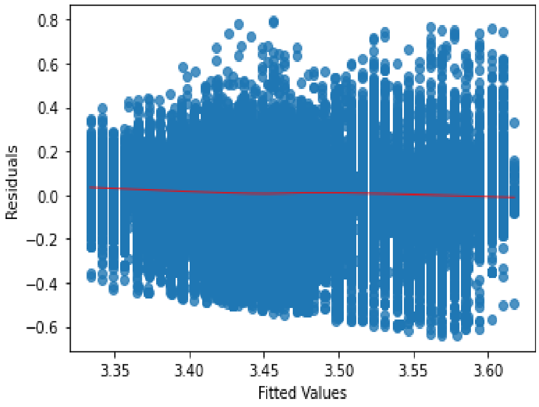


Figure. 11

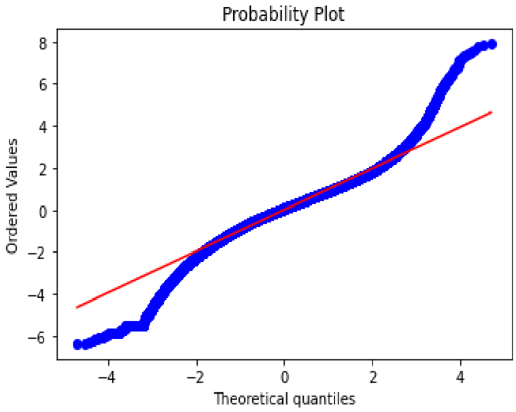


Figure. 12

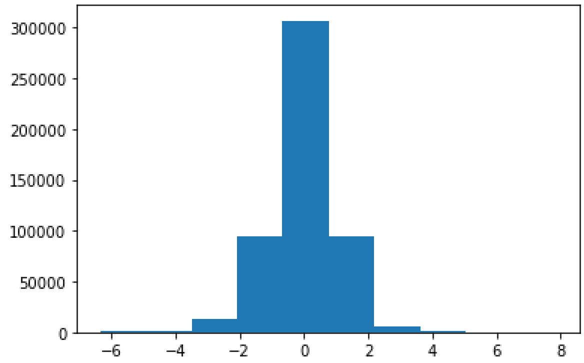
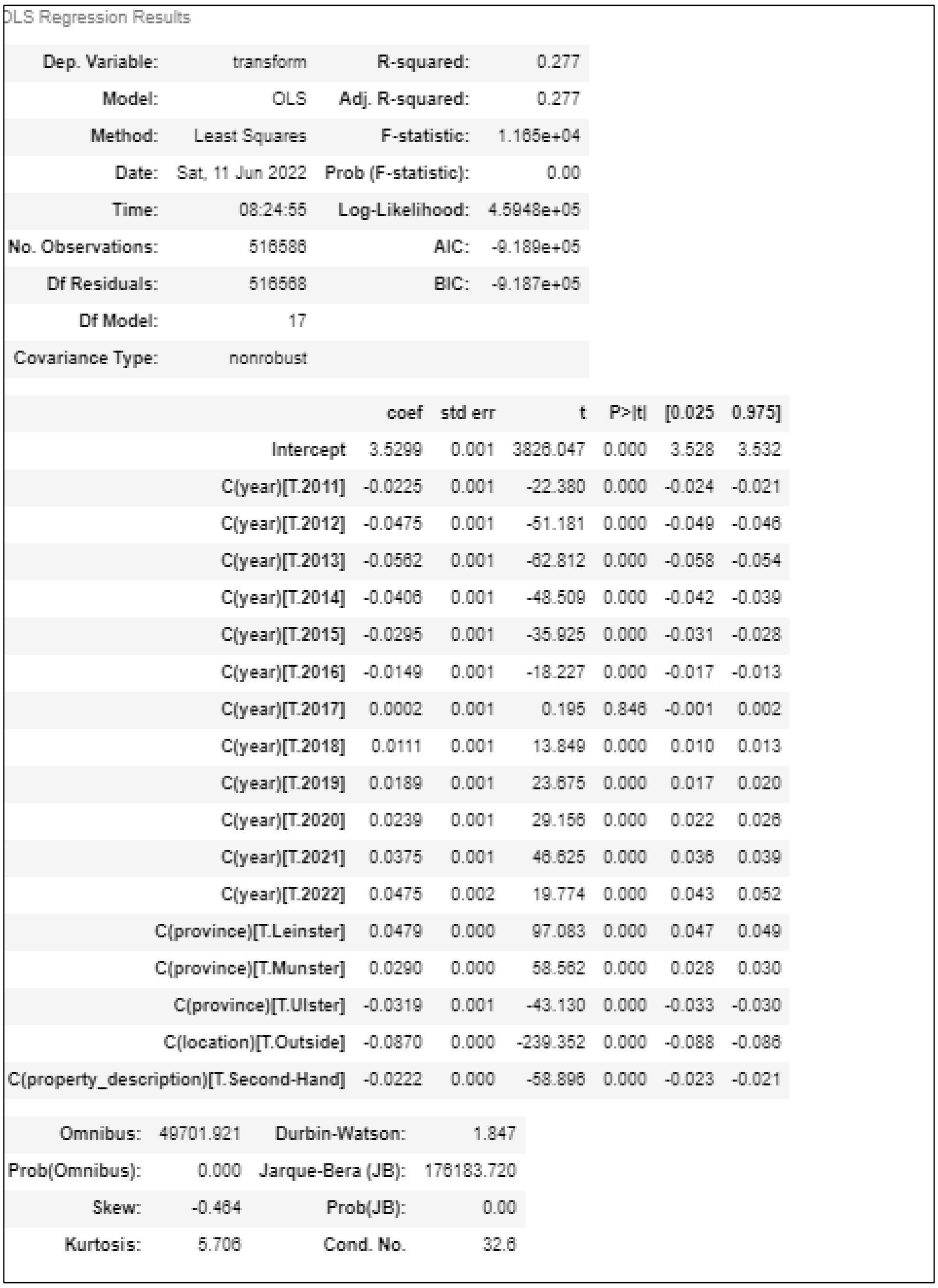


Figure. 13



OLS Regression Results

Dep. Variable:	transform	R-squared:	0.277
Model:	OLS	Adj. R-squared:	0.277
Method:	Least Squares	F-statistic:	1.165e+04
Date:	Sat, 11 Jun 2022	Prob (F-statistic):	0.00
Time:	08:24:55	Log-Likelihood:	4.5948e+05
No. Observations:	516586	AIC:	-9.189e+05
Df Residuals:	516568	BIC:	-9.187e+05
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.5299	0.001	3826.047	0.000	3.528	3.532
C(year)[T.2011]	-0.0225	0.001	-22.380	0.000	-0.024	-0.021
C(year)[T.2012]	-0.0475	0.001	-51.181	0.000	-0.049	-0.046
C(year)[T.2013]	-0.0562	0.001	-62.812	0.000	-0.058	-0.054
C(year)[T.2014]	-0.0406	0.001	-48.509	0.000	-0.042	-0.039
C(year)[T.2015]	-0.0295	0.001	-35.925	0.000	-0.031	-0.028
C(year)[T.2016]	-0.0149	0.001	-18.227	0.000	-0.017	-0.013
C(year)[T.2017]	0.0002	0.001	0.195	0.846	-0.001	0.002
C(year)[T.2018]	0.0111	0.001	13.849	0.000	0.010	0.013
C(year)[T.2019]	0.0189	0.001	23.675	0.000	0.017	0.020
C(year)[T.2020]	0.0239	0.001	29.156	0.000	0.022	0.026
C(year)[T.2021]	0.0375	0.001	46.625	0.000	0.036	0.039
C(year)[T.2022]	0.0475	0.002	19.774	0.000	0.043	0.052
C(province)[T.Leinster]	0.0479	0.000	97.083	0.000	0.047	0.049
C(province)[T.Munster]	0.0290	0.000	58.562	0.000	0.028	0.030
C(province)[T.Ulster]	-0.0319	0.001	-43.130	0.000	-0.033	-0.030
C(location)[T.Outside]	-0.0870	0.000	-239.352	0.000	-0.088	-0.086
C(property_description)[T.Second-Hand]	-0.0222	0.000	-58.896	0.000	-0.023	-0.021

Omnibus:	49701.921	Durbin-Watson:	1.847
Prob(Omnibus):	0.000	Jarque-Bera (JB):	176183.720
Skew:	-0.464	Prob(JB):	0.00
Kurtosis:	5.706	Cond. No.	32.6

Figure. 14

Interpretation of the model coefficients as per *Figure. 14*.

y = 3.52 - 0.02 year_2011 – 0.05 year_2012 - 0.06 year_2013 - 0.04 year_2014 - 0.03 year_2015 - 0.01 year_2016 + 0.0002 year_2017 + 0.01 year_2018 +0.02 year_2019 + 0.02 year_2020 + 0.04 year_2021 + 0.05 year_2022 + 0.05 province_leinster +0.03 province_munster - 0.03 province_ulster - 0.09 outside_dublin - 0.02 second_hand.

3.52 is the estimated intercept i.e., average y when all x are zero.

For 1 year increase in year_2011, the average increase in property price will be -0.02, keeping all other variables constant.

For 1 year increase in year_2012, the average increase in property price will be -0.05, keeping all other variables constant.

For 1 year increase in year_2013, the average increase in property price will be -0.06, keeping all other variables constant.

For 1 year increase in year_2014, the average increase in property price will be -0.04, keeping all other variables constant.

For 1 year increase in year_2015, the average increase in property price will be -0.03, keeping all other variables constant.

For 1 year increase in year_2016, the average increase in property price will be -0.01, keeping all other variables constant.

For 1 year increase in year_2017, the average increase in property price will be 0.0002, keeping all other variables constant.

For 1 year increase in year_2018, the average increase in property price will be 0.01, keeping all other variables constant.

For 1 year increase in year_2019, the average increase in property price will be 0.02, keeping all other variables constant.

For 1 year increase in year_2020, the average increase in property price will be 0.02, keeping all other variables constant.

For 1 year increase in year_2021, the average increase in property price will be 0.04, keeping all other variables constant.

For 1 year increase in year_2022, the average increase in property price will be 0.05, keeping all other variables constant.

For 1 unit increase in province_leinster, the average increase in property price will be 0.05, keeping all other variables constant.

For 1 unit increase in province_munster, the average increase in property price will be 0.03, keeping all other variables constant.

For 1 unit increase in province_ulster, the average increase in property price will be -0.03, keeping all other variables constant.

For 1 unit increase in outside_dublin, the average increase in property price will be -0.09, keeping all other variables constant.

For 1 unit increase in second_hand, the average increase in property price will be -0.02, keeping all other variables constant.

The property price is having positive increase after year 2017, it shows that the price has been increased significantly after 2017.

All p-values are < 0.05, except for year_2017 which 0.08 this implies that the variables year, location, province, and property_description are statistically significant.

Table. 29. MLR Results interpretation

VII. Ethical Considerations

This section describes the ethical challenges faced in the different stages of the research as a data scientist and the ethical concerns with the data. To avoid the ethical issues, the data was collected from a website which is accessible to the public for download and use and no personal information were used for analysis. The ethical challenges faced are as follows.

A. *Data storage, security, and responsible data stewardship*(Vallor et al.)

Even though the data does not contain any personal information, to restrict the loss of the data from the system and to avoid the misuse of it especially the API key, a folder is created in Google drive and one drive to save the details related to the dissertation. The folder is secured with passcode to avoid the misuse of it. Anti-virus packages are installed in machines to reduce the risk of hacking and virus.

B. *Data hygiene and data relevance*(Vallor et al.)

To add additional data to the dataset, details are obtained from genuine websites with proper referencing. The API key to get the latitude is taken from a website which follows GDPR and sends requests as HTTPS.

C. *Identifying and addressing ethically harmful data bias*(Vallor et al.)

The dataset has a column with postal code, and it contains the postal code of Dublin only. So, if I use that column in my modelling there may be a chance of bias towards Dublin than other places. Even though it doesn't affect any individuals and doesn't create any harms, but for the better modelling and to reduce bias, I may drop that column.

The ethical issues related to data are also considered in the research and they are,

A. *Harms to Privacy and Security*

To avoid the issues with privacy and security, no personal information is using in the research. The data is collected from an open website, and the data is free to public for downloading and using. Reuse and download of the data are permitted with acknowledging the website and the data cannot be used in a misleading way. The data does not contain any personal information and hence there is no ethical harms related to privacy and security (Vallor et al.).

B. *Harms to Fairness and Justice*

Since no personal data is being used, there is no issue related to discrimination and bias related to gender or age and hence there is no issue related to sexism, racism, and ableism. Since the data is considering the whole Ireland and the properties sold and not about people, there is no other ethical issues related to fairness and justice also (Vallor et al.).

C. *Harms to Transperancy and Autonomy*

To avoid the transparency issues, the data is collected from the free website, and anyone can visit the data without any issues. The research is not using any humans for the analysis or any personal data and hence there is no issues related to transparency and autonomy and no human rights have been violated (Vallor et al.).

VIII. Conclusion And Futurework

Property prices in Ireland were analyzed using statistical analysis and visualizations as a part of the initial analyses for the research. Different plots were created using seaborn and matplotlib for the entire dataset which is downloaded from the PSRA website and ANOVA test was generated for the random sample size of 500 samples from the entire data. The results from the ANOVA test implies that the attributes ‘year’, ‘province’, ‘property_description’ and ‘location’ has significant effect on the property prices. From the visualizations it is revealed that the property prices are higher in Dublin compared to other counties and more houses are sold in Dublin also. The post codes of the county Dublin are also having significant effect on the property prices where Dublin 14 is having higher property prices and Dublin 10 is having lower prices.

For future analysis, hedonic regression and machine learning algorithms including linear regression, SVM and DT will be used to predict the property prices. Machine learning models will be compared based on the performance metrices and best model will be identified.

Timeline for proposed work:

The timeline for the proposed work is almost three months from June 10 to September 10. The proposed plan for the research is to spend 8 hours daily with 4 hours each for report and coding.

By June end, the statistical analyses will be completed. Prediction using machine learning is planning to complete by July. Work will be reported along with the modelling and Iam planning to complete the report by the mid of August and then review it.

IX. References

Abdulai, R. and Owusu-Ansah, A. (2011). *(PDF) Hedonic regression analysis of house price determinants in Liverpool, England* [online]., pp.6–31. Available from: https://www.researchgate.net/publication/286675668_Hedonic_regression_analysis_of_house_price_determinants_in_Liverpool_England [accessed 1 June 2022].

Advantages of Support Vector Machines (SVM). Available from: https://iq.opengenus.org/advantages-of-svm/ [accessed 2 June 2022].

Alfiyatin, A.N., Febrita, R.E., Taufiq, H. and Mahmudy, W.F. (2017). Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study : Malang, East Java, Indonesia. *International Journal of Advanced Computer Science and Applications* [online], 8(10). Available from: www.ijacsa.thesai.org [accessed 1 June 2022].

Amrutha. (2022). *Decision Tree Machine Learning Algorithm - Analytics Vidhya* [online]. *https://www.analyticsvidhya.com/* [online]. Available from: https://www.analyticsvidhya.com/blog/2022/01/decision-tree-machine-learning-algorithm/ [accessed 5 June 2022].

analyticsvidhya. (2018). *XGBoost Algorithm | XGBoost In Machine Learning* [online]. *analyticsvidhya.com* [online]. Available from: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/ [accessed 5 June 2022].

Andrade, F. (2021). *A Simple Guide to Scikit-Learn — Building a Machine Learning Model in Python | by Frank Andrade | Towards Data Science* [online]. *https://towardsdatascience.com/* [online]. Available from: https://towardsdatascience.com/a-beginners-guide-to-text-classification-with-scikit-learn-632357e16f3a [accessed 5 June 2022].

Brownlee, J. (2014). *A Gentle Introduction to Scikit-Learn* [online]. *https://machinelearningmastery.com/* [online]. Available from: https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn-a-python-machine-learning-library/ [accessed 5 June 2022].

CFI. (2022). *Hedonic Regression Method - Overview, Application, Function* [online]. *https://corporatefinanceinstitute.com/* [online]. Available from: https://corporatefinanceinstitute.com/resources/knowledge/other/hedonic-regression-method/ [accessed 5 June 2022].

Chen, J.H., Ong, C.F., Zheng, L. and Hsu, S.C. (2017). Forecasting spatial dynamics of the housing market using Support Vector Machine. *International Journal of Strategic Property Management* [online], 21(3), pp.273–283.

Coughlan, M. (2022). *How everything is different for today's first-time buyers* [online]. Available from: https://www.rte.ie/news/primetime/2022/0130/1276779-how-everything-is-different-for-todays-first-time-buyers/ [accessed 2 June 2022].

Deepanshi. (2021). *Linear Regression | Introduction to Linear Regression for Data Science* [online]. *https://www.analyticsvidhya.com/* [online]. Available from: https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/ [accessed 5 June 2022].

FitzGerald, G. (2007). *What caused the Celtic Tiger phenomenon? – The Irish Times* [online]. *The Irish Times* [online]. Available from: https://www.irishtimes.com/opinion/what-caused-the-celtic-tiger-phenomenon-1.950806 [accessed 8 June 2022].

Frew, J. and Jud, G.D. (2020). Estimating the Value of Apartment Buildings. *https://doi.org/10.1080/10835547.2003.12091101* [online], 25(1), pp.77–86. Available from: https://www.tandfonline.com/doi/abs/10.1080/10835547.2003.12091101 [accessed 2 June 2022].

Gandhi, R. (2018). *Support Vector Machine — Introduction to Machine Learning Algorithms | by Rohith Gandhi | Towards Data Science* [online]. *https://towardsdatascience.com/* [online]. Available from: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 [accessed 5 June 2022].

Gazette Desk. (2021). *Supply chain constraints will push up house prices, say builders* [online]. *Law Society of Ireland* [online]. Available from: https://www.lawsociety.ie/gazette/top-stories/2021/05-may/supply-chain-constraints-will-push-up-house-prices-say-builders [accessed 2 June 2022].

Gu, J., Zhu, M. and Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine. *Expert Systems with Applications* [online], 38(4), pp.3383–3386.

Ho, W.K.O., Tang, B.S. and Wong, S.W. (2020). Predicting property prices with machine learning algorithms. *https://doi.org/10.1080/09599916.2020.1832558* [online], 38(1), pp.48–70. Available from: https://www.tandfonline.com/doi/abs/10.1080/09599916.2020.1832558 [accessed 1 June 2022].

Hurley, A.K., Sweeney, James, Hurley AoifeHurley, A.K., Hurley, A. and Sweeney, J. (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 2022* [online], 5 May 2022, pp.1–39. Available from: https://link.springer.com/article/10.1007/s11146-022-09888-y [accessed 30 May 2022].

Ihre, A. (2019). Predicting house prices with machine learning methods., 2019.

investopedia. (2021). *Hedonic Regression Definition* [online]. *investopedia.com* [online]. Available from: https://www.investopedia.com/terms/h/hedonic-regression.asp [accessed 5 June 2022].

Ja’afar, N.S., Mohamad, J. and Ismail, S. (2021). Machine learning for property price prediction and price valuation: A systematic literature review. *Planning Malaysia* [online], 19(3), pp.411–422.

Jose Doval Tedin, M., Faubert, V. and Economy, E. (2020). Housing Affordability in Ireland., December 2020. Available from: https://ec.europa.eu/info/publications/economic-and-financial-affairs-publications_en. [accessed 2 June 2022].

K, D. (2019). *Top 5 advantages and disadvantages of Decision Tree Algorithm | by Dhiraj K | Medium* [online]. *Medium* [online]. Available from: https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a [accessed 2 June 2022].

Kenton, W. (2022). *Analysis of Variance (ANOVA) Definition & Formula* [online]. *investopedia.com* [online]. Available from: https://www.investopedia.com/terms/a/anova.asp [accessed 7 June 2022].

Kumar, N. (2019). *The Professionals Point: Advantages of XGBoost Algorithm in Machine Learning* [online]. *theprofessionalspoint* [online]. Available from: http://theprofessionalspoint.blogspot.com/2019/03/advantages-of-xgboost-algorithm-in.html [accessed 2 June 2022].

Kunovac, D. and Zagreb, K.K. (2019). Residential Property Prices in Croatia., 2019.

Li, D.Y., Xu, W., Zhao, H. and Chen, R.Q. (2009). A SVR based forecasting approach for real estate price prediction. *Proceedings of the 2009 International Conference on Machine Learning and Cybernetics* [online], 2, pp.970–974.

Limsombunc, V., Gan, C. and Lee, M. (2004). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. *American Journal of Applied Sciences* [online], 1(3), pp.193–201.

MacFarlane, I. (2022). *How the war in Ukraine could affect UK property – Show House* [online]. *https://www.showhouse.co.uk/* [online]. Available from: https://www.showhouse.co.uk/news/how-the-war-in-ukraine-could-affect-uk-property/ [accessed 2 June 2022].

Mbaabu, O. (2020). *Introduction to Random Forest in Machine Learning | Engineering Education (EngEd) Program | Section* [online]. *Section* [online]. Available from: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/ [accessed 2 June 2022].

Miao, D., Tang, H. and Wang, B. (2021). Support Vector Regression with Gaussian kernel for Housing Prices Prediction. *Journal of Physics: Conference Series* [online], 2021, pp.1–9. Available from: https://iopscience.iop.org/article/10.1088/1742-6596/1994/1/012023/pdf [accessed 1 June 2022].

Montero, J.-M. and Fernández-Avilés, G. (2014). Hedonic Price Model. *Encyclopedia of Quality of Life and Well-Being Research* [online], 2014, pp.2834–2837. Available from: https://link.springer.com/referenceworkentry/10.1007/978-94-007-0753-5_1279 [accessed 2 June 2022].

Mu, J., Wu, F. and Zhang, A. (2014). Housing Value Forecasting Based on Machine Learning Methods. , 2014. Available from: http://dx.doi.org/10.1155/2014/648047 [accessed 31 May 2022].

Petrov, M. (2009). *The vulture has landed | EurobuildCEE* [online]. Available from: https://eurobuildcee.com/en/magazine/866-the-vulture-has-landed [accessed 2 June 2022].

Phan, D. (2018). Housing Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia. *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)* [online], 2018, pp.1–8. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8614000 [accessed 1 June 2022].

Phan, T.D. (2019). Housing price prediction using machine learning algorithms: the case of Melbourne city, Australia. *Proceedings - International Conference on Machine Learning and Data Engineering, iCMLDE 2018* [online], 15 January 2019, pp.35–42. Available from: https://researchers.mq.edu.au/en/publications/housing-price-prediction-using-machine-learning-algorithms-the-ca [accessed 25 May 2022].

Pow, N. and Janulewicz, E. (1995). Applied Machine Learning Project 4Prediction of real estate property prices in. *Machine Learning* [online], 20(3), pp.273–297.

Qualitrics. (2022). *How to Determine Sample Size in Research | Qualtrics* [online]. *Qualitrics* [online]. Available from: https://www.qualtrics.com/uk/experience-management/research/determine-sample-size/ [accessed 11 June 2022].

Quang, T., Minh, N., Hy, D. and Bo, M. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science* [online], 174, pp.433–442.

Ray, S. (2017). *SVM | Support Vector Machine Algorithm in Machine Learning* [online]. *https://www.analyticsvidhya.com/* [online]. Available from: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ [accessed 5 June 2022].

Reddan, F. (2018). How high can they go? Six factors affecting Irish house prices – The Irish Times. *The Irish Times* [online], 18 September 2018. Available from: https://www.irishtimes.com/business/personal-finance/how-high-can-they-go-six-factors-affecting-irish-house-prices-1.3628349 [accessed 2 June 2022].

Saini, A. (2021). *Decision Tree Algorithm - A Complete Guide - Analytics Vidhya* [online]. *https://www.analyticsvidhya.com/* [online]. Available from: https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/ [accessed 5 June 2022].

Selim, S. (2008). DETERMINANTS OF HOUSE PRICES IN TURKEY: A HEDONIC REGRESSION MODEL Sibel SELİM. , 9(1), pp.65–76.

Shi, H. and Li, W. (2009). Fusing Neural Networks, Genetic Algorithms and Fuzzy Logic for Analysis of Real Estate Price. , 2009, pp.1–4. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5362675 [accessed 2 June 2022].

Shinde, N. and Gawande, K. (2018). VALUATION OF HOUSE PRICES USING PREDICTIVE TECHNIQUES. *International Journal of Advances in Electronics and Computer Science* [online], 2018, pp.2393–2835. Available from: http://iraj.in [accessed 3 June 2022].

Sullivan, A. (2021). *House prices: 'Wall of money' hits European real estate | Business | Economy and finance news from a German perspective | DW | 03.06.2021* [online]. Available from: https://www.dw.com/en/house-prices-wall-of-money-hits-european-real-estate/a-57765308 [accessed 2 June 2022].

Thamarai, M. and Malarvizhi, S.P. (2020). Information Engineering and Electronic Business. *Information Engineering and Electronic Business* [online], 2, pp.15–20. Available from: http://www.mecs-press.org/ [accessed 30 May 2022].

Vallor, S., William, J. and Rewak, S.J. An Introduction to Data Ethics MODULE AUTHOR: 1. Available from: https://techethics.ieee.org [accessed 9 April 2022 a].

Vallor, S., William, J. and Rewak, S.J. *An Introduction to Data Ethics MODULE AUTHOR: 1* [online]. Available from: https://techethics.ieee.org [accessed 5 June 2022 b].

Wang, X., Wen, J., Zhang, Y. and Wang, Y. (2014). Real estate price forecasting based on SVM optimized by PSO. *Optik* [online], 125(3), pp.1439–1443.

Waseem, M. (2022). *Linear Regression for Machine Learning | Intro to ML Algorithms | Edureka* [online]. *edureka* [online]. Available from: https://www.edureka.co/blog/linear-regression-for-machine-learning/ [accessed 2 June 2022].

Wu, J.Y. (2017). Housing Price prediction Using Support Vector Regression. , 2017.

Yee, L.W., Abu Bakar, N.A., Mohd Zainuddin, N.M. and Mohd Yusoff, R.C. (2021). Using Machine Learning to Forecast Residential Property Prices in Overcoming the Property Overhang Issue., 2021, pp.1–6. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9573830&casa_token=_yLGMqWPaLAAAAAA:bQg91cipVoaA5gHUGRqe1MsvooGwR_WjCbHx-xgvWqsLXaEbPpiN6WdyOnt45_GP7BmPgNvL0KkBJA&tag=1 [accessed 2 June 2022].

Yiu, T. (2019). *Understanding Random Forest. How the Algorithm Works and Why it Is… | by Tony Yiu | Towards Data Science* [online]. *Towards Data Science* [online]. Available from: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 [accessed 11 June 2022].

APPENDIX