[**Using Machine Learning to Analyze and Predict Ireland's Housing Crisis 3**](#_c555letwbjl5)

[Introduction 3](#_q983y024uqi5)

[Problem Definition 4](#_ro5qx7usj2mi)

[Project Scope 5](#_o5b9ti5oj7w5)

[Key Phases of the Project: 5](#_1qw6rzhpr53m)

[**Methodology 5**](#_uk8plmkhdb1s)

[Success Criteria and Indicators 5](#_93dwnjxg5f7w)

[Ethical Considerations 6](#_2ezehsb2spn6)

[Project Plan: Execution 7](#_1fcffilbrs6d)

[**Exploratory Data Analysis (EDA) 7**](#_kqgulox194vn)

[Data Description 7](#_n9jee9pv65za)

[Data Visualization 8](#_vbb48cdskzmy)

[**Pre-processing 9**](#_ti4ni0lz6as2)

[Data Understanding 9](#_qefucqs2olxv)

[Data Preparation and Cleaning 10](#_d6l3r4m5534o)

[Encoding and Scaling 11](#_6overpoai3nj)

[Feature analysis selection 11](#_1yxz0kh0kwaa)

[**Modeling and Evaluation 12**](#_2kmr6ddj64o5)

[Data Splitting 12](#_oy48kwcnuf1d)

[X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42) 12](#_dsbrw1ij7456)

[**Linear Regression 13**](#_rptcpbpki7j5)

[Support Vector Regression (SVR) 14](#_d8i17tyqkcc)

[) 15](#_igb3tiebggqg)

[Performance Comparison 15](#_r1jyx854gcxa)

[**Visualisation Results 15**](#_xr0l5181y5zt)

[**Challenge and Strategies 16**](#_3tb4jozfu6lg)

[**Conclusion and Recommendations 16**](#_a5bbqw37mw1m)

[Reference list 16](#_86pewvlojpd1)

# **Using Machine Learning to Analyze and Predict Ireland's Housing Crisis**

## **Introduction**

People dream of having their own home. Everyone deserves a suitable place to live, and some would even say that it's a basic right to have a house to call your own. We are currently witnessing a housing crisis in Ireland that has only worsened over the years, despite the government introducing several programs to assist the population, such as the Help-to-Buy scheme and the First Home Scheme.

The pressing issues of housing availability and affordability have created significant barriers for potential homeowners. Property prices continue to rise at rates that outpace wage growth, and the rental market offers little relief, with monthly payments often exceeding potential mortgage payments. There aren't enough houses for everyone, and it is a big challenge to save enough money when faced with exorbitant prices, leaving many, especially younger people, without hope of purchasing a home.

The impacts of the housing crisis are numerous: Irish people leaving their own country, people living in poor conditions, and homelessness, along with the social and psychological effects of being unable to own a home. The crisis has led to increased mental health concerns, delayed family formation, and reduced economic mobility. Additionally, younger generations face an uncertain future, where owning a home feels more like a distant dream than a realistic goal.

The ripple effects extend beyond housing itself, affecting Ireland's economic competitiveness, social fabric, and demographic patterns. Many skilled professionals are choosing to relocate abroad, creating a brain drain that could have long-term implications for the country's development. The crisis disproportionately affects certain groups, including young professionals, single-parent households, and essential workers who often find themselves priced out of the communities they serve.

Addressing this issue requires innovative solutions and different approaches to ensure that the basic human need for housing is met. This includes exploring new construction technologies, alternative financing models, and data-driven policy making. This project aims to predict future house prices and utilize machine learning to help people prepare for the future of something so essential: having a roof over their heads and a place to call home.

Objective

*The primary objective of this project is to develop a machine learning model that accurately predicts house prices in Ireland, helping potential homeowners assess viability and providing insights into current and short-term housing market trends. There’s also other objectives as such:*

* *Analyze historical housing price trends across Dublin, Cork, Galway, Limerick, Waterford, and other regions in Ireland to identify patterns and influencing factors.*
* *Develop and compare machine learning models to predict housing prices for 2025 and 2026 for both new and second-hand houses across major Irish regions.*
* *Identify regions with the highest and lowest projected price increases to provide actionable insights for homebuyers, investors, and policymakers.*
* *Evaluate how key factors impact housing prices, including market supply of new and second-hand homes, government programs, economic indicators, and interest rates.*
* *Create an accessible report translating complex predictive analytics into practical insights for diverse stakeholders including first-time buyers, renters, developers, and politicians.*
* *Establish a sustainable framework for ongoing housing price prediction that guides future homeowners through purchasing challenges.*

## **Problem Definition**

Ireland has been experiencing a housing crisis for at least a decade, if not longer. The root cause is complex, linked to:

* High demand and limited supply
* Wages and the current economy making it difficult for the general population to afford current house prices
* Immigration, which accounts for 15.5% of Ireland's population
* Platforms like Airbnb, where homeowners can earn more money by renting through the platform rather than selling or leasing their homes

This has a direct impact on homelessness, as many people cannot afford to buy or rent due to high prices. The crisis also negatively affects mental health and well-being, with many feeling hopeless and considering leaving the country as the situation shows no signs of resolving soon.

For some context and background, we can consider that the Irish house market has faced consistent pressure from the demand of new houses and high prices for buying and renting.

This issue also impacts several types of people, including first-time buyers, renters, developers and politicians.

## **Project Scope**

This capstone project spans two semesters and will analyze historical data from data.gov.ie, which has a dataset with information about the average house price on the main areas in Ireland: Dublin, Cork, Galway, Limerick, Waterford and the rest of Ireland, and it has a total of 589 rows.

While other EU nations and individual Irish counties may face similar housing challenges, this project will remain concentrated on these main areas to maintain clarity and relevance. By keeping the scope, we aim to produce more targeted insights that can be directly applied locally, also ensuring the findings are both practical and impactful for Ireland's housing market.

## **Key Phases of the Project:**

1. Data Understanding and Preparation:  
   * Analyze the dataset for trends and anomalies.
   * Ensure data quality by addressing missing values and inconsistencies.
2. Development of Predictive Models:  
   * Use machine learning techniques to predict housing prices for 2025 and 2026.
   * Evaluate multiple algorithms to ensure accuracy and interpretability.
3. Insights and Reporting:  
   * Identify counties with the highest and lowest predicted price increases.
   * Develop a user-friendly report summarizing findings for prospective homeowners and politicians.
4. Ethical Considerations:  
   * Address biases in the dataset to ensure equitable insights.
   * Evaluate the impact of predictions on vulnerable populations.

# Methodology

## **Success Criteria and Indicators**

The success of this project will be measured by checking if it achieves its main goals and creates useful results for different users. Below are the key criteria for success:

1. Model Accuracy  
   * The predictions for house prices in 2025 and 2026 should be close to real data when it becomes available.
   * The model should work well for all the main regions included in the project.
2. Clear Insights  
   * The project should identify which areas (according to the dataset) in Ireland will see the highest and lowest increases in house prices.
   * The final report should be easy to understand and helpful for people planning to buy a house and for policymakers.
3. Positive Impact  
   * The report should help people make better decisions about buying houses by providing useful information.
   * Politicians should be able to use the results to improve housing programs and strategies.
4. Ethical and Social Awareness  
   * The project should not create unfair results or ignore certain groups of people or regions.
   * It should explain clearly how the model works and why certain predictions are made.
5. On-Time Delivery  
   * The work should follow the planned timeline and finish all parts of the project on time.
   * Each part of the project, such as data preparation and model building, should meet the quality expected by supervisors.
6. Long-Term Value  
   * The model and results should be easy to update with new data so they can stay useful in the future.
   * Policymakers or other organizations should find the results helpful enough to consider using them.

## **Ethical Considerations**

Housing has long been recognized as a determinant of health, as highlighted in a 1934 report by Britten (R. Britten, The Relation Between Housing and Health). This is particularly relevant for vulnerable populations such as the homeless, refugees, and immigrants, who often face historical and systemic challenges related to housing. I will ensure that this project does not introduce any bias and that it is sensitive to the challenges faced by those striving to purchase a home, by ensuring the data analyzed doesn't contain any specific information such as nationality, gender and social situation. This approach ensures that we're maintaining objectivity by focusing only on variables related to housing, and so the insights will remain impartial.

## 

## 

## 

## **Project Plan: Execution**

The implementation of this project will follow a structured approach that incorporates regional analysis and socioeconomic factors to ensure comprehensive predictions.

Our analysis will include comparing price trends across Dublin, Cork, Galway, Limerick, and Waterford, while also considering employment rates, population changes, and economic indicators that influence housing markets.

We'll evaluate how different stakeholders are affected by the housing crisis and assess the effectiveness of existing housing policies like the Help-to-Buy scheme and planning regulations.

From a technical perspective, our machine learning approach will use feature engineering, time series analysis, and ensemble methods to improve prediction accuracy, while ensuring our models are interpretable for decision-making.

## 

# **Exploratory Data Analysis (EDA)**

## **Data Description**

During Ireland's ongoing housing crisis, where many people cannot buy homes, data analysis helps us understand the market and plan for the future. This analysis looks at two housing price datasets:

* HousingPrices.csv: Old housing price data from 2000 onwards. This data
* HousingPrices2021.csv: New housing sale data

These datasets show the housing market in different parts of Ireland, including Dublin, Cork, Galway, Limerick, Waterford, and other areas.

They show prices for new houses and second-hand houses from 2000 to 2021. The prices are in Euros.

We cleaned the data in these ways:

* We divided counties into standard areas
* We divided houses into two types: "New House Prices" and "Second Hand House Prices"
* We joined the two datasets to make one dataset with more information, as the first dataset only had data until 2016
* We checked data from 2010-2016 (where both datasets had information) to make sure they matched

Summary Statistics

As people in Ireland struggle to find homes, our data shows:

1. Geographic Price Variation: Dublin has the highest house prices. This is because many jobs are in Dublin. There are big price differences between cities and other areas, making it harder for some people to find homes.

Property Type Differences: New houses and second-hand houses have different prices. This affects who can buy them. First-time buyers often find it harder than people who already own homes.

Time Trends: Our data shows many housing market ups and downs. Dublin's market changes a lot, making it hard for people to know when to buy.

Data Checking: We checked the differences between our datasets in years where they overlap to make sure our final dataset is correct.

## **Data Visualization**

To better understand the housing crisis that makes many Irish people delay having families, move to other countries, or live in bad conditions, we made these pictures:

**Box Plots by Area:** These show how house prices are different in different areas. Dublin houses cost much more, which is why many workers cannot live where they work. The plots show how much prices change within each area.

**Time Series Analysis:** These track how prices changed over time in Irish cities. They show how house prices went up faster than wages. The charts show market cycles and how different cities have different price trends.

**Price Distributions:** Histograms show how house prices are spread out, separated by house type. These show how some types of houses are easier to buy than others, which affects social classes.

**Comparative Box Plots:** These show both area and house type together. They help us see how these things work together to make houses easy for some people to buy but too hard for others.

**Predictive Visualizations:** Looking at the future for people who want to buy homes, these charts show predicted house prices until 2026. The bar chart shows predicted 2025 prices by county, and the line chart shows 5-year price changes across different counties. These help people and government make plans.

These pictures show Ireland's housing market, its past patterns, and future trends. They are both tools for analysis and clear pictures of the crisis that is changing Irish society, making people leave Ireland, delay having families, and causing problems for the country's social and economic health.

# **Pre-processing**

## **Data Understanding**

In this project, I worked with two complementary datasets:

* The first dataset was selected from data.gov.ie, providing valuable historical house price information across major regions ( add link)
* The second dataset was obtained from Kaggle (https://www.kaggle.com/datasets/erinkhoo/property-price-register-ireland), offering more recent property transaction data

After integration, my combined dataset contained 634 rows and 4 columns, spanning from 2000 to the most recent year available in the Kaggle dataset.

**Key Variables:**

* **Statistic Label:** Categorizes houses as:
  + New House Prices
  + Second Hand House Prices
* **Years Coverage:** When houses were purchased
* **Area Coverage:**
  + National (country-wide statistics)
  + Dublin
  + Cork
  + Galway
  + Limerick
  + Waterford
  + Other areas
* **Price Information:** All values in Euros

My initial assessment showed not many empty values within the dataset. The data is also consistent with all values being in Euros.

## **Data Preparation and Cleaning**

My first steps for data preparation involved checking for missing values through reviewing the dataset structure and performing basic statistical analysis.

I filtered the data to focus specifically on entries from 2000 onwards, which provides more relevant information for the current housing crisis. For the first dataset (from data.gov.ie), I only kept records from years 2000-2009, and for the second dataset (from Kaggle), I transformed the SALE\_DATE column to extract only the year information, and included all available years up to 2021

For property categorization, I used string matching to identify new versus second-hand properties in the second dataset. I also mapped counties to standardized areas and kept the 5 main counties in Ireland, and with smaller counties, I grouped into "Other areas" for consistency.

I handled missing values in the Area column by filling them with "Other areas" to ensure data completeness.

I simplified both datasets by selecting only the essential columns (Year, Area, Property Type, VALUE) to create consistent structures before merging them.

To ensure data quality, I compared both datasets in overlapping years (2010-2016) by calculating percentage differences. This validation step confirmed reasonable consistency between the datasets.

I then integrated the data by combining pre-2010 records from the first dataset with the second dataset. After integration, I dropped any remaining rows with NA values in the VALUE column to ensure clean analysis.

The processed dataset was exported to a CSV file, allowing for easy access in future analysis steps.

## **Encoding and Scaling**

I created numerical codes for categorical variables using the factorize method, converting Area and Property Type into Area\_Code and Property\_Type\_Code. This encoding step was necessary for the machine learning models.

I introduced feature scaling for my modeling approach. For the SVR model specifically, I implemented RobustScaler which handles outliers better, which is important for the Dublin housing data that showed significant price outliers.

## **Feature analysis selection**

After preparing and cleaning my housing price datasets, I conducted a thorough feature analysis to identify the most relevant variables for my predictive models. My goal was to select features that would effectively capture the dynamics of Ireland's housing market while maintaining model parsimony.

From my exploratory data analysis, I identified three key features that demonstrated to be the best options to showcase the information necessary for housing prices:

**Year**: Yearly information is crucial for capturing market trends, economic cycles, and the overall evolution of housing prices. The data showed distinct patterns across different time periods, with significant price fluctuations corresponding to economic events and policy changes.

**Area**: We need to be aware of the locations to predict housing prices, especially in a country like Ireland that has 5 major cities. My analysis confirmed substantial regional variations, with Dublin consistently showing premium pricing compared to other regions.

**Property Type**: The distinction between new houses and second-hand properties revealed meaningful price differentials. This variable reflects important aspects such as building condition, modern amenities, and government programs like the help-to-buy, that provides support for first time buyers buying a new house.

For my models, I needed to change Area and Property Type into numbers since machine learning models can't work with text. I used the factorize method to create Area\_Code and Property\_Type\_Code.

My final features for predicting house prices were:

* Year (numerical)
* Area\_Code (converted from Area)
* Property\_Type\_Code (converted from Property Type)

With house price (VALUE) as what I wanted to predict, these features gave me a good balance of accuracy and simplicity. I didn't include other variables that showed similar information to these three or didn't help much in prediction.

The charts I created, like scatter plots and boxplots, helped confirm these were good choices. And with the boxplots, we can see how house prices change based on area and property type.

## 

# Modeling and Evaluation

## Data Splitting

To ensure my models could generalize well to unseen data, I split my dataset into training and testing portions. I used the train\_test\_split function from scikit-learn with a test size of 20% and a random state of 42 for reproducibility:

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

This split provided me with:

* A training set to teach the models the patterns in the housing data
* A test set to evaluate how well the models would perform on new, unseen data
* Training and Test Data

My training data consisted of:

* X\_train: Feature values for Year, Area\_Code, and Property\_Type\_Code
* y\_train: The corresponding house prices (VALUE)

The test data followed the same structure:

* X\_test: Feature values held out for evaluation
* y\_test: The actual house prices for testing prediction accurac

# Linear Regression

I first implemented a Linear Regression model as my baseline approach. This model assumes a linear relationship between the features and the target variable:

model = LinearRegression()

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

After training, I evaluated the model using two key metrics:

* R² Score: Measures how well the model explains the variance in the data
* Root Mean Squared Error (RMSE): Measures the average prediction error in euros

The Linear Regression model's performance was captured by these metrics, which served as my baseline for comparing more complex models.

Random Forest

To potentially capture more complex patterns in the housing data, I implemented a Random Forest Regressor. This ensemble method creates multiple decision trees and averages their predictions:

rf\_params = {

'n\_estimators': [100, 200],

'max\_depth': [10, 20, None],

'min\_samples\_split': [2, 5]

}

I used GridSearchCV to find the optimal hyperparameters for the Random Forest model:

rf\_grid = GridSearchCV(RandomForestRegressor(random\_state=42), rf\_params, cv=5, scoring='r2')

rf\_grid.fit(X\_train, y\_train)

rf\_model = rf\_grid.best\_estimator\_

The best parameters were selected from the grid search, optimizing the model's performance. I also calculated cross-validation scores to ensure the model was performing consistently across different subsets of the data.

The Random Forest model's performance on the test set was measured using its R² score, allowing for direct comparison with the Linear Regression model

## **Support Vector Regression (SVR)**

I also implemented a Support Vector Regression model as an alternative approach. SVR works differently from both Linear Regression and Random Forest by trying to find a boundary that contains most of the data points within a certain margin.

For the SVR model, I used a more advanced scaling technique:

robust\_scaler = RobustScaler()

X\_train\_robust = robust\_scaler.fit\_transform(X\_train)

X\_test\_robust = robust\_scaler.transform(X\_test)

y\_scaler = RobustScaler()

y\_train\_scaled = y\_scaler.fit\_transform(y\_train.values.reshape(-1, 1)).ravel()

The RobustScaler was particularly important for this model because it handles outliers better than standard scaling. This was valuable for the Dublin housing data, which had significant price outliers compared to other regions.

I configured the SVR model with these parameters:

svr\_model = SVR(

kernel='rbf',

C=1000,

gamma='auto',

epsilon=0.01

## **)**

These parameters include:

* kernel='rbf': Using a Radial Basis Function kernel to capture non-linear relationships
* C=1000: A high penalty for error, making the model try harder to fit the training data
* gamma='auto': Letting the algorithm determine the right influence of each training example
* epsilon=0.01: Setting a small margin of error that's acceptable without penalty

After training on the scaled data, I had to transform the predictions back to the original scale to evaluate the model properly:

svr\_pred\_scaled = svr\_model.predict(X\_test\_robust)

svr\_pred = y\_scaler.inverse\_transform(svr\_pred\_scaled.reshape(-1, 1)).ravel()

## **Performance Comparison**

I compared the performance of all three models:

* Linear Regression: R² = 0.6767
* Random Forest: R² = 0.9900
* SVR: R² = 0.8297

The Random Forest model performed significantly better than both the Linear Regression and SVR models. With an R² score of 0.9900, it was able to explain most of the variance in housing prices. This suggests that the relationship between my selected features and housing prices has complex patterns that the Random Forest algorithm captured well.

The SVR model with RobustScaler showed moderate performance, while the Linear Regression achieved the lowest R² score among the three models.

Based on these results, I used the Random Forest model to make predictions about future housing prices across different counties in Ireland. These predictions can help people understand potential housing price trends in the context of Ireland's ongoing housing crisis.

## 

# Visualisation Results

# Challenge and Strategies

# Conclusion and Recommendations

## **Reference list**

www.linkedin.com. (n.d.). Understanding the Housing Crisis in Ireland: Causes, Consequences, and Solutions. [online] Available at: https://www.linkedin.com/pulse/understanding-housing-crisis-ireland-causes-solutions-tafura-khatun-lltme/.

Britten, R.H. (1934). The Relation between Housing and Health. Public Health Reports (1896-1970), [online] 49(44), pp.1301–1313. doi:https://doi.org/10.2307/4581354.

of the Taoiseach, D. (2024). Migration - the Facts. [online] Www.gov.ie. Available at: https://www.gov.ie/en/collection/aeea0-migration-the-facts/.

GeeksforGeeks. (2019). ML | Handle Missing Data with Simple Imputer. [online] Available at: <https://www.geeksforgeeks.org/ml-handle-missing-data-with-simple-imputer/>.

<https://realpython.com/python-matplotlib-guide/>

<https://realpython.com/linear-regression-in-python/>

Introduction to Machine Learning with Python: A Guide for Data Scientists Paperback –

by [Andreas C Müller](https://www.amazon.co.uk/s/ref=dp_byline_sr_book_1?ie=UTF8&field-author=Andreas+C+M%C3%BCller&text=Andreas+C+M%C3%BCller&sort=relevancerank&search-alias=books-uk)

<https://www.w3schools.com/python/python_try_except.asp>

<https://www.evalacademy.com/articles/how-to-combine-data-from-multiple-sources-for-cleaning-and-analysis>

<https://saturncloud.io/blog/pandas-how-to-concatenate-dataframes-with-different-columns/#:~:text=To%20concatenate%20dataframes%20with%20different%20columns%2C%20we%20use%20the%20concat,new%20dataframe%20that%20combines%20them>.

<https://www.statology.org/pandas-factorize/>

<https://medium.com/@nandiniverma78988/neural-network-regression-implementation-and-visualization-in-python-d5893713ed79>

<https://visualstudiomagazine.com/Articles/2023/05/01/regression-scikit.aspx>

https://proclusacademy.com/blog/robust-scaler-outliers/

**Github Repo**

Overall Notes:  
  
When combining the two datasets for data from 2010 to 2016  
Average absolute difference: 19.36%

The differences aren't too bad. We can probably combine them.

Include that I added a new dataset