Housing Price Analysis in Ireland

The real estate sector has always been an important part of the economy, with property prices being a strong indicator of a country's economic condition. In Ireland, as in many other countries, the real estate market has been influenced by political, economic, and social factors that have caused property prices to fluctuate over time.

# Importance of Analysis

The real estate market is complex and vast, containing a huge amount of data. Often, traditional methods of accurately analyzing the various market patterns are flawed and insufficient. Machine learning can bridge this gap, as it has tools capable of processing large volumes of data and making accurate predictions and classifications.

This is crucial for helping investors and buyers understand future property price trends, whether for buying and selling or for making investments in the sector. Machine learning methods can also assist politicians and urban planners in creating better housing policies through detailed analysis of regional market conditions.

# Problems to be Addressed

## Future Price Determination

To adequately plan for investors and consumers, effective property price forecasting is crucial. Forecasting algorithms can identify trends of rising or falling prices based on historical data, reducing market uncertainties and developing better strategies in the real estate market.

# Justification for Using Forecasting and Classification Algorithms

Machine learning algorithms can identify complex patterns in data that are often not visible through more traditional analysis methods. Regarding housing prices:

1. Forecasting algorithms can model relationships between multiple variables and effectively predict future prices.
2. Classification algorithms help categorize data into groups or classes, such as price ranges or market categories (market price vs. non-market, for example).

# Project Objectives

Predict future property prices in Ireland, using historical data to train predictive models. This objective aims to provide valuable insights that can benefit not only investors and planners but also ordinary individuals in their real estate decision-making processes.

# Characterization

## About Dataset

### Context

All residential properties sold in Ireland from 2010 to May 28th, 2021.

Dataset retrieved from Kaggle at the address: https://www.kaggle.com/datasets/erinkhoo/property-price-register-ireland/data

Collaborator: Erin Khoo (Owner)

### Content

Rows: 476,745

Columns: 9

Memory usage: 32.7MB

Variables:

* SALE\_DATE => Date of sale (dd/MM/yyyy) | datetime64[ns]
* ADDRESS => Address | string
* POSTAL\_CODE => Postal Code | string
* COUNTY => County | string
* SALE\_PRICE => Price (€) | float32
* IF\_MARKET\_PRICE => Not Full Market Price | int8
* IF\_VAT\_EXCLUDED => VAT Exclusive | int8
* PROPERTY\_DESC => Description of Property | string
* PROPERTY\_SIZE\_DESC => Property Size Description | string

### Acknowledgements

Data sourced from the publicly available site: https://propertypriceregister.ie

Source: National Property Price Registry

License: EU ODP Legal Notice

## Data Preprocessing

### Data Cleaning

To prepare the dataset for analysis, the following actions were taken:

1. Columns POSTAL\_CODE and PROPERTY\_SIZE\_DESC were removed due to containing a very high percentage of null data (81.17% and 88.92%, respectively).
2. Data with unrecognized characters in the PROPERTY\_DESC column were changed to null and subsequently deleted.
3. 770 records were deleted due to being considered duplicates.

These actions were sufficient for data cleaning, removal of null data, and maintaining them in good condition.

### Data Preparation

The SALE\_DATE variable was split into SALE\_MONTH and SALE\_YEAR to facilitate and simplify temporal analysis, allowing for the identification of seasonal trends over time or grouping data by specific time intervals.

### Conversion of Categorized Values to Numeric

Converting categorical variables into numerical values allows for a wider range of algorithms to be used in the Machine Learning model, and in some models, it is required that all input variables be numeric. Additionally, it helps to standardize the data and make it easier to interpret and analyze.

1. **One-Hot Encoding (Dummy Variables) -** In the COUNTY variable, there are several categories, and the technique of 'dummy variables' is used to ensure that the numerical values assigned to the categories are not interpreted as ordinal values (i.e., values that have a specific order).
2. **Value Transformation** - The variable PROPERTY\_DESC has only two values ('Second-Hand Dwelling house /Apartment' and 'New Dwelling house /Apartment'), and we can replace them with 0 and 1, respectively.