Caroline Alves Ferreira

2022188@student.cct.ie

A Roof Over Our Heads: Homeownership Challenges in Ireland

CA 2 – Capstone Report

Link GitHub: <https://github.com/CCT-Dublin/ca1-capstone-project-proposal-C2022188.git>

A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Caroline Alves Ferreira |
| *Student Number* | 2022188 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA 1 – Capstone Report |
| *Assessment Due Date* | Sunday, 08th December 2024 |
| *Date of Submission* |  |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Figures ii](#_Toc184143105)

[Introduction 1](#_Toc184143106)

[Business Understanding 2](#_Toc184143107)

[Data Understanding 3](#_Toc184143108)

[Data Preparation 5](#_Toc184143109)

# Figures

[Figure 1 - Info of dataset "Rent as Percentage of Disposable Income" 3](#_Toc184143120)

[Figure 2 - Info of dataset "RTB Average Monthly Rent Report" 4](#_Toc184143121)

[Figure 3 - Info of dataset "House Prices" 4](#_Toc184143122)

[Figure 4 - Handling Missing Values in the 'House Prices' Dataset 5](#_Toc184143123)

[Figure 5 - Dropping Columns in the 'House Prices' Dataset 6](#_Toc184143124)

[Figure 6 - Dropping Columns in the 'RTB Average Monthly Rent Report' Dataset 6](#_Toc184143125)

[Figure 7 - Missing Values in the 'RTB Average Monthly Rent Report' Dataset 6](#_Toc184143126)

[Figure 8 - Missing values by location in the 'RTB Average Monthly Rent Report' Dataset 7](#_Toc184143127)

[Figure 9 - Missing Values by Quarter in the 'RTB Average Monthly Rent Report' Dataset 8](#_Toc184143128)

# Introduction

The housing crisis is one of the most talked-about problems in Ireland today. Property prices continue to rise, affecting both rent and the dream of owning a home. Families paying rent, especially in Dublin, one of the most expensive cities in Ireland, find it very difficult to save enough money to buy their own homes. For single people, the challenge is even greater, as reaching the income level needed for mortgage approval can feel impossible without additional support.

This project will explore the main barriers to affordable housing by analysing rent costs across regions and comparing them with income levels. It will also look at house prices in different areas, highlighting how they have consistently increased year after year. To structure the work, the report will follow the CRISP-DM framework, covering Social Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.

# Business Understanding

The first stage of the CRISP-DM framework focuses on understanding the problem and the needs of the project. Since housing prices are a social issue, this stage is called Social Understanding. The goal is to identify the main objectives of the project and understand the challenges related to the crisis.

Ireland is experiencing a serious housing crisis caused by high demand and a lack of available homes. This imbalance has led to rising house prices and rents, making housing less affordable, especially for low-income families. Many people cannot save enough money to buy a home because of the high cost of living and strict mortgage rules. According to Hearne (2017), this crisis is linked to economic inequality. Housing policies often treat homes as investments rather than a basic human need, making homeownership even harder to achieve for many people.

The crisis is not just about the shortage of homes. Potts (2020) explains that the real issue is affordability. While everyone needs a place to live, people can only live where they can afford it. Gillespie (2018) points out that private housing developments tend to focus on higher-income groups, leaving lower-income families with very few options. This housing crisis affects not only individuals and families but also society as a whole. Hansen (2013) states that affordable and secure housing is essential for physical and mental health, children’s education, and family stability. However, Hearne (2020) argues that weak tenant protections and treating housing as a financial asset have made it even harder for many people to find safe and affordable homes.

The housing crisis is a complex issue that affects not only individuals and families but also the overall stability of society. Understanding the financial barriers and regional disparities is essential to address this problem effectively. This project will focus on identifying these challenges, analyzing the social and economic impacts of the crisis, and providing data-driven insights to propose potential solutions for making housing more affordable and accessible in Ireland.

# Data Understanding

The data understanding phase of CRISP-DM involves taking a closer look at the data available for mining. This step is critical in avoiding unexpected problems during the next phase--data preparation--which is typically the longest part of a project (IBM, 2021.). The data for this study was collected from reliable sources, mainly the Central Statistics Office (CSO) in Ireland. The datasets cover important aspects of the housing crisis, such as rent, house prices, disposable income, and inflation. These datasets are key to understanding the current situation and finding solutions.

The tools and technologies that will be used in the project are:

o   Pandas

o   Matplotlib

o   Seaborn

One dataset, *"Rent as Percentage of Disposable Income"*, shows how much of their income tenants spend on rent. It highlights the financial pressure of rent in different regions. By using the .info() function, it is possible to observe that the dataset contains 124 entries and 5 columns: Statistic Label, Rental Year, Local Authority, Unit, and Value, and there are no missing values.

A screenshot of a computer

Description automatically generated

Figure - Info of dataset "Rent as Percentage of Disposable Income"

Another key dataset is the "RTB Average Monthly Rent Report," which provides detailed information on average monthly rents across various regions. The dataset consists of 393,372 entries with 7 columns. During the data cleaning process, it will be necessary to address missing values by imputing them with the average for each region. Overall, the dataset is reliable, offering strong coverage across both time periods and locations.

A screenshot of a computer

Description automatically generated

Figure - Info of dataset "RTB Average Monthly Rent Report"

The *"House Prices"*  dataset contains information about average house prices over time. It helps identify trends and changes in the housing market. Using the .info() it is possible to see that the dataset contains 334,656 entries and 8 columns:

A screenshot of a computer

Description automatically generated

Figure - Info of dataset "House Prices"

It is observed that in the column “VALUE” are 826 missing values, which will be dropped in the next stage of data cleaning process. Overall, this dataset is well-organized by year, regions and value, with no significant missing data relative to the total number of entries.

All the data has been carefully prepared to ensure consistency across sources and suitability for analysis. With this strong foundation, the next step will be to explore these patterns further and connect them to the housing challenges faced in Ireland.

# Data Preparation

According to Conroy (2023), data preparation involves selecting, cleaning, merging, formatting, and transforming data to make it suitable for use in a data product. This phase focuses on resolving issues and errors in the dataset to ensure it is clean and ready for machine learning model implementation.

The first step in cleaning this dataset “House\_Prices” is to remove rows that have missing values, as they do not contribute to the analysis:

A screenshot of a computer program

Description automatically generated

Figure 4 - Handling Missing Values in the 'House Prices' Dataset

The next step of the data cleaning, for this dataset, was dropping the columns that are not useful:

A screenshot of a computer

Description automatically generated

Figure 5 - Dropping Columns in the 'House Prices' Dataset

Another dataset where it was necessary to drop columns was the *"RTB Average Monthly Rent Report,"* as those columns were deemed irrelevant:

A screenshot of a computer

Description automatically generated

Figure 6 - Dropping Columns in the 'RTB Average Monthly Rent Report' Dataset

During the Data Understanding process, it was observed that this dataset has some missing values. By using the command .isnull().sum(), we can see an overview of how much data is missing in the *VALUE* column:

A white rectangular object with a white border

Description automatically generated

Figure 7 - Missing Values in the 'RTB Average Monthly Rent Report' Dataset

The next step involves identifying the locations where the *VALUE* column has missing data and counting the number of missing entries for each location. This helps pinpoint areas with significant data gaps.

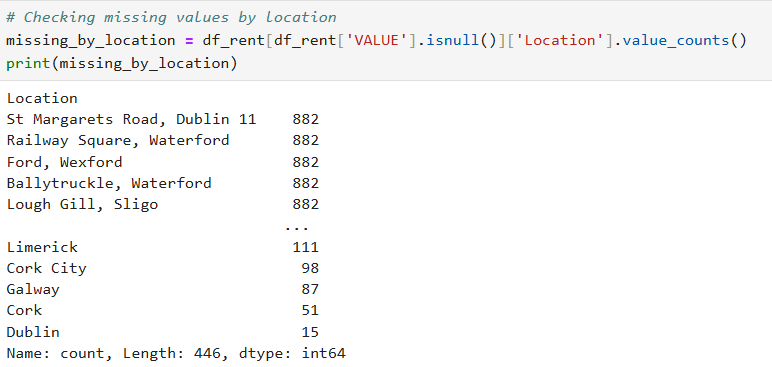


Figure 8 - Missing values by location in the 'RTB Average Monthly Rent Report' Dataset

Similarly, the data is analysed to count the number of missing entries in the *VALUE* column grouped by the *Quarter*. This helps to check if the missing data is concentrated in specific time periods:

A screenshot of a computer

Description automatically generated

Figure 9 - Missing Values by Quarter in the 'RTB Average Monthly Rent Report' Dataset

After these analyses, the next step was to fill the missing values in the *VALUE* column by replacing them with the median *VALUE* of the corresponding *Location* group. The transform function is used to ensure the replacements are applied at the group level while keeping the DataFrame structure intact:

A screen shot of a computer code

Description automatically generated

Figure 10 - Handling Missing Value in the 'RTB Average Monthly Rent Report' Dataset

Finally, after filling the missing values using the location-specific medians, any remaining missing values were replaced with the overall median of the *VALUE* column. This ensures that the missing data is handled effectively with minimal impact on the overall dataset quality.

In the other dataset, *"RTB Tenants"* , since there were no missing values, the only modification made was dropping the columns that were irrelevant.

A screenshot of a computer

Description automatically generated

Figure 11 - Info of dataset 'RTB Tenants'