Project Report

GitHub URL

https://github.com/MartyRath/UCDPA_MartyRath

Abstract

The price of houses in Ireland seems to constantly be on an upward trend. I am in the market to purchase my first home. I chose my project subject based on this, looking at the Irish property register from 2010 to May 2021. With a strict budget of 110K, and a preference for living in Leinster, I am hoping my deep dive into this dataset will be invaluable to making an informed and rounded decision on purchasing the right property. Plots made with Matplotlib.

Introduction

(Explain why you chose this project use case)

I chose this project as I am planning on buying my first home in the next year somewhere in Leinster. This is the most practical form of project I could do. The insights made in this project will guide and influence my decision when buying a house. This seems to be the appropriate application of the given project. My current house budget is €110K, which will limit options significantly. This necessitates my project for real world application. I will deem this project a success if I find the insights useful to purchasing the best possible first home I can afford. Top priorities of this project are; when and where to buy in Leinster?

Dataset

I am planning to buy a house in Ireland, and so chose a dataset from Ireland. The dataset is a CSV file from Kaggle.com, which was sourced from the Irish Property Price Register (source below). The Kaggle dataset was cleaner than downloading directly from the Price Register site, and suits the project's needs perfectly.

The dataset of the price register covers all houses sold in Ireland from 2010 to May 28th 2021.

The vast amount of data and time frame can be used to discover long-term trends, while also being up-to-date enough to gain relevant, contemporary insights.

I also added inflation data from 'Inflation rate, average consumer prices (Annual percent change)'. I sourced this information from the International Monetary Fund. (source below)

Implementation Process

Firstly, I imported necessary libraries. There included:pandas

- pandas as pd
- NumPy as np
- Matplotlib.pyplot as plt
- Seaborn as sns
- xlrd

Then, I imported my main dataset, 'Property_Price_Register_Ireland-28-05-2021.csv', as a Dataframe using pandas 'read_csv'.

I printed the dataset head, shape, and its columns.

Its columns cover:

- 'SALE_DATE' - dates properties were sold
- 'ADDRESS' - property addresses
- 'POSTAL_CODE' - postcode or eircode
- 'COUNTY' - county
- 'SALE_PRICE' - price sold
- 'IF_MARKET_PRICE' - if sold at market price at time of sale
- 'IF_VAT_EXCLUDED' - Generally, all sales of old properties are exempt from VAT in Ireland
- 'PROPERTY_DESC' - if the property is old or new
- 'PROPERTY_SIZE_DESC' - the size in square metres of the property

I used iloc to print the first and last rows, which showed the date range to be 2010-01-01 - 2021-05-18.

<u>Cleaning:</u>
'SALE_DATE',
'ADDRESS',
'POSTAL_CODE',
'COUNTY',
'SALE_PRICE',
'IF_MARKET_PRICE',
'IF_VAT_EXCLUDED',
'PROPERTY_DESC',
'PROPERTY_SIZE_DESC'

Drop duplicates:
I started by ensuring there were no duplicate entries. I used drop duplicates, with the subset columns ‘DATE’ and ‘ADDRESS’. These seemed specific enough as together there could be multiple entries of the same address.

Missing values:
I checked for missing values, using df.isna() and sum(). I made a plot using Matplotlib:

The only missing values appeared in ‘POSTAL_CODE’ and ‘PROPERTY_SIZE_DESC’, with over 80% missing in each.

I considered using the ‘fillna()’ method on ‘PROPERTY_SIZE_DESC’, but with my emphasis being more on price and general location, I found I could drop both these columns entirely.

Dropping columns:
I dropped any columns which would not be directly relevant to my needs using .drop.

‘POSTAL_CODE’ and ‘ADDRESS’
I have a preference to live in Leinster, but that is as far as location specifics would be required. I found the ‘COUNTY’ column to be sufficient for my research.
‘IF_MARKET_PRICE’
If market price was offered in 1 or 0, without specifying if above or below. This

didn’t seem relevant enough than insights which could be gained from other columns present.
‘IF_VAT_EXCLUDED’
This is indicative of the selling of a residential home, and would not offer any guidance on how to buy.

Remaining columns:
'SALE_DATE',
'COUNTY',
'SALE_PRICE',
'PROPERTY_DESC',

Then, I began exploring the remaining, relevant data.

‘PROPERTY_DESC’ seemed only to have two values, 'New Dwelling house /Apartment' and 'Second-Hand Dwelling house /Apartment'.
I checked this by printing unique values for this column using a for loop coupled with .unique.
I found that many of the values were Irish versions of the New/Second-hand inputs. So, I used .replace.
Now, there remain only two values in ‘PROPERTY_DESC’, denoting ‘New’ and ‘Second-Hand’ properties.

I formatted the ‘SALE_DATE’ column using pandas datetime.

I then created two new columns from ‘SALE_DATE’, ‘YEAR’ and ‘MONTH’, using pandas dt.year and dt.month respectively.
This will be convenient for later reference.

As I am focusing on Leinster properties, I modified my data to reflect this. I used the isin() method to create a new Dataframe, ‘Leinster’. I input the twelve Leinster counties, thus dropping counties outside Leinster.

I decided to limit my data from 2010-2020. I used the year column and less than or equal to to achieve this. This will lead to more rounded results for calculations such as cheapest time of year to buy, as not including half of 2021.

I created a new CSV file as an extra backup of my work so far, using to_csv.

Merging Dataframes:

I thought I would add an additional column with annual percentage change in inflation.

I imported the excel file using pandas ‘read_excel’, used ‘.drop’ for unnecessary rows and columns, and ‘reset_index()’.
The inflation Dataframe had years as columns, and the percentage change in inflation as rows.

I decided to make a new Dataframe, extracting years and inflation change to lists.
For the years column, I used ‘.columns’ and ‘tolist()’.
For the percentages row, is used ‘.values’, then used zero to index this nested list.
I used my two lists, years and percentages, to make a dictionary
Then, I converting this to a new Dataframe, ‘inflation’, with columns ‘YEAR’ and ‘INFLATION’:

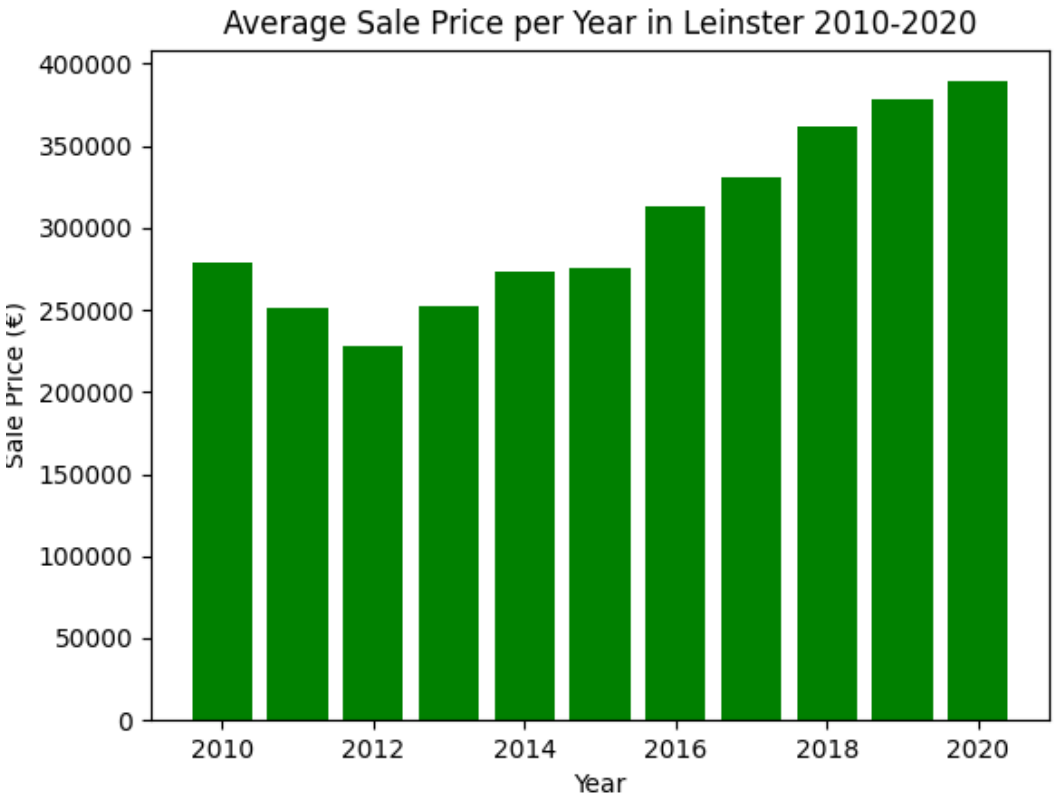
Finally, I merged this with my main ‘Leinster’ Dataframe using ‘.merge’ on ‘YEAR’.

Exploratory Analysis
I used ‘.groupby’, ‘.agg’, NumPy, comparison operators, and a custom function for exploratory analysis.
Some of my most interesting highlights for this are:

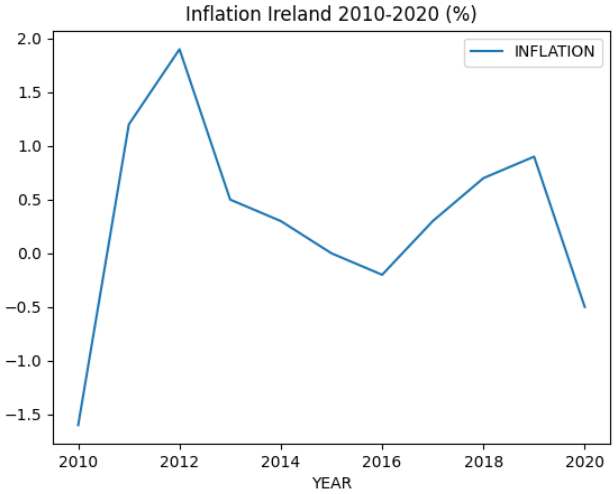
- In 2020 some houses sold for under 7K, with the cheapest for just €6500 in Dublin.
- The cheapest house sold in Wexford in 2020 was just €7000.
- There were over 1800 houses sold in Leinster in 2020 under 110K, with 258 of those sold in Wexford under 110K.
- I used ‘.groupby’ and ‘.agg’ to get the following results for 2020:
- min: €6,500
- max: €134,261,040
- mean: €389,144
- median: €290,000
- Custom function: As there are obvious outliers in this data, I defined a custom function to find the InterQuartile Range in Sale Prices, over standard deviation. IQR shows where the bulk of the data lies.
  IQR in 2020 was €196,500.

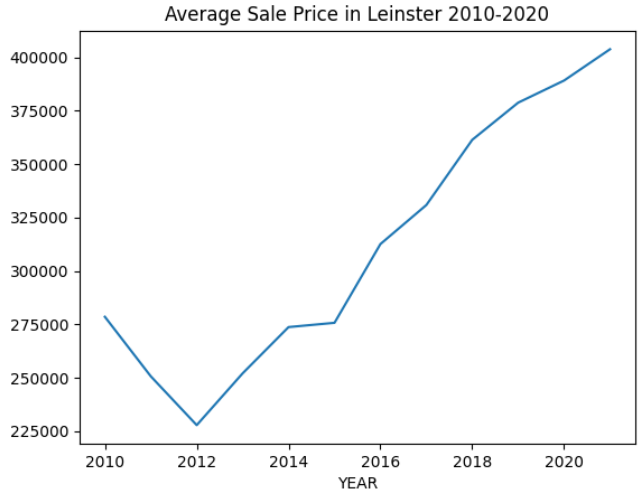
Results

I started by viewing the average sale price per year in Leinster, shown below. As suspected, house prices are steadily increasing, with over 10k increase from 2019-2020. (379K-389K~)



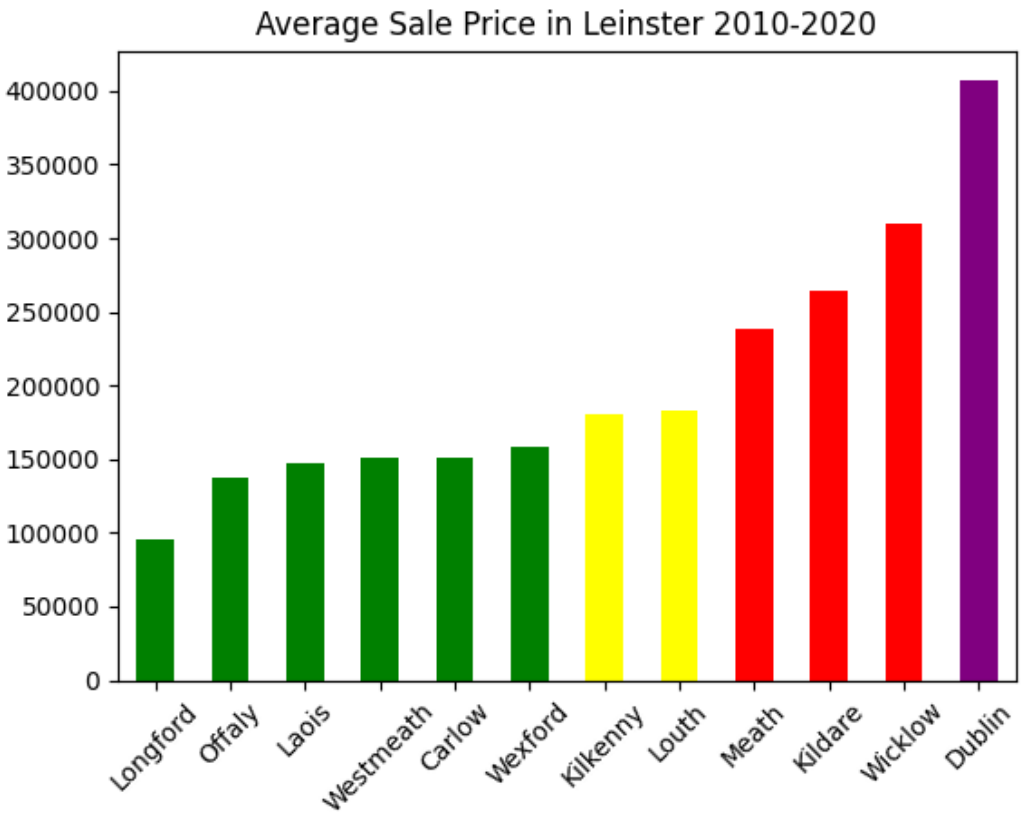
This made me wonder if there was a correlation between house price increases and inflation?





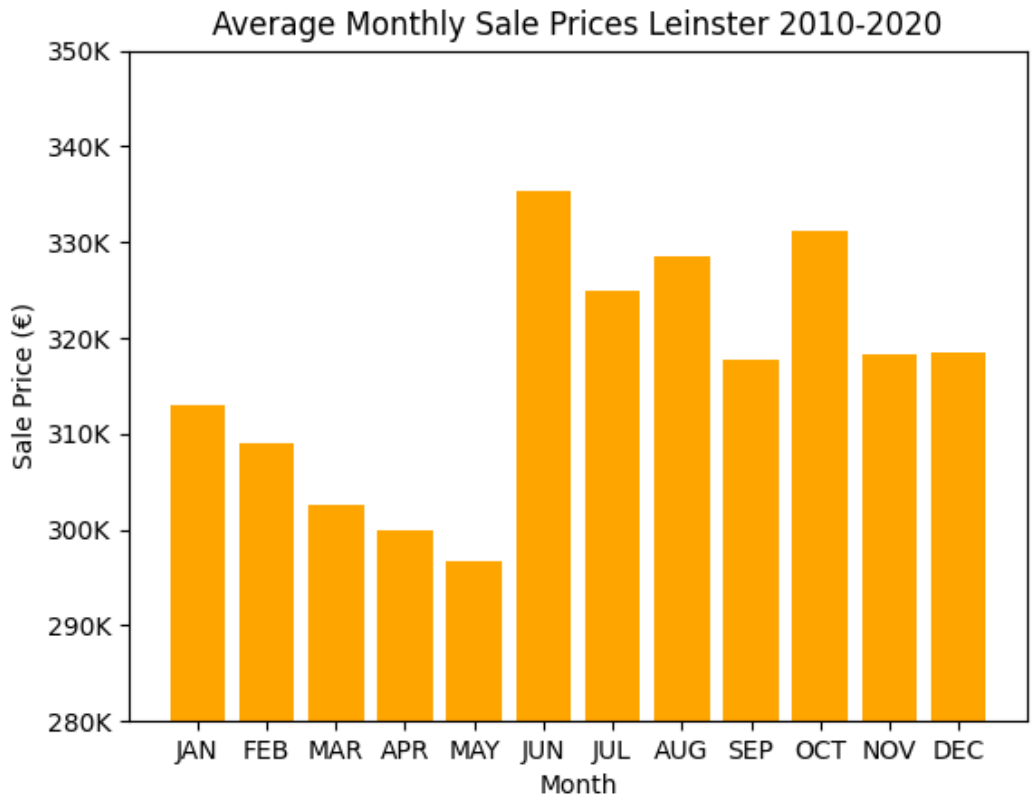
There does not seem to be any clear correlation between inflation and property in Ireland. This means inflation cannot predict house prices.

Next, as my budget is 110K, and the mean price in Leinster 2020 being €389,144 I looked at the cheapest counties.



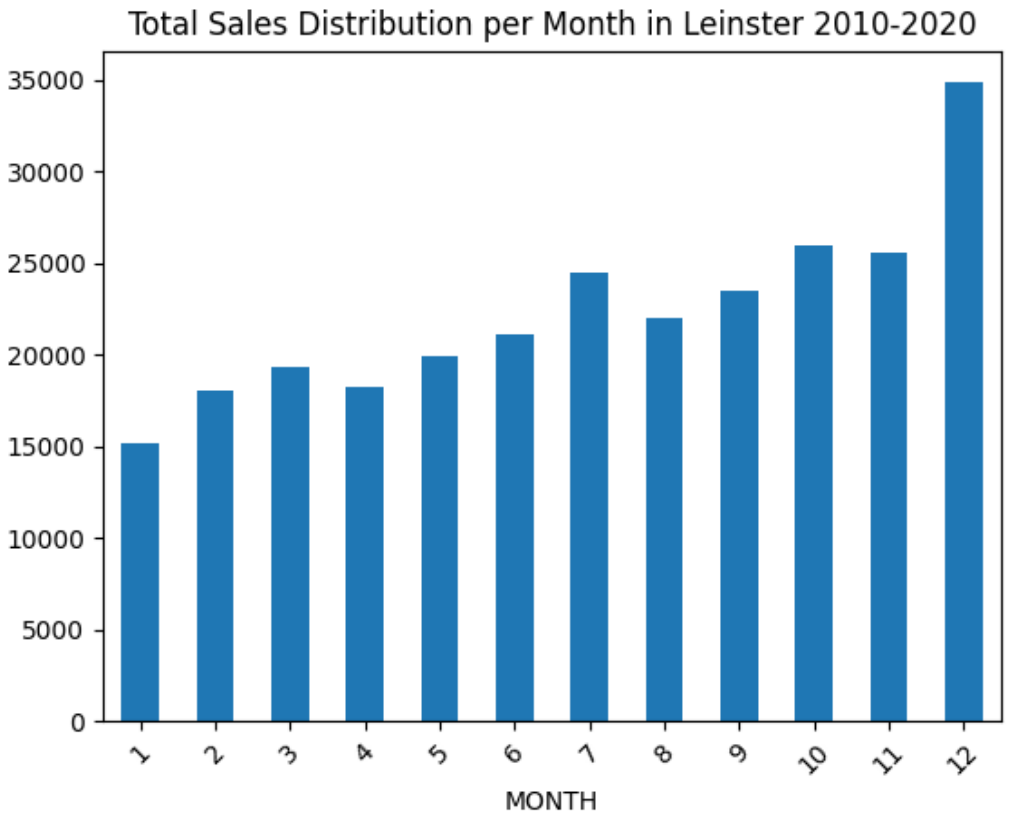
The graph shows Dublin to be, by far, the most expensive county. The counties in green seem much more accommodating to the 110K budget.

Next, I looked to see if there is any particular time of year which is cheapest to buy.



To my surprise, May is the lowest by a substantial amount. There is a steady decline from Janurary to May, then a massive increase from June on. The mean price in May here is €296,684, while the mean price is June is €335,428. That is a difference of €38,744.

This insight proved more valuable than I would have previously anticipated. I was baffled, and wanted to find a reason. I thought perhaps the amount of houses sold, low or high, in a given month might be the reason.



Comparing how many sales were made to mean sale price shows no correlation. May is the cheapest month to buy, but is average when it comes to house sales.

This information has proven invaluable to my house hunt, and will guide the buying of my first home.

I think the next step in my learning and research will be data scraping for counties on Daft.ie to set up an alert of ads that fit the criteria outlined in this project.

Insights

- House prices are steadily increasing year by year.
- There does not seem to be a correlation between inflation and house prices.
- The cheapest month to buy is by far May.
- The cheapest average counties to buy in are Longford, Offaly, Laois, Westmeath, Carlow and Wexford. The others may be out of reach with this budget.
- The cheapest time to buy a house next seems to be in May 2022.

References

- Dataset: https://www.kaggle.com/erinkhoo/property-price-register-ireland
- Kaggle data source: https://propertypriceregister.ie
- Irish Inflation/CPI source: https://www.imf.org/external/datamapper/PCPIPCH@WEO/IRL?zoom=IRL&highlight=IRL