**MSc in Data Analytics**

**Continuous Assessment 2**

Author: Ramon Joulia

e-mail: [sbs23007@student.cct.ie](mailto:sbs23007@student.cct.ie)

Student ID: sbs23007

GitHub link: <https://github.com/RamonJoulia/MSC_DA_InterGr_CA2>

Word count: 3056

**Table of contents**

[1. Introduction 2](#_Toc135736278)

[2. Materials and Methods 2](#_Toc135736279)

[2.1. Python libraries and packages used: 2](#_Toc135736280)

[2.2. Datasets and sources (direct links in appendix): 4](#_Toc135736281)

[2.3. Data cleaning and preparation: 4](#_Toc135736282)

[3. Result and Discussion 5](#_Toc135736283)

[3.1. Statistics: 5](#_Toc135736284)

[3.1.1. New dwellings completion by year in Ireland: 5](#_Toc135736285)

[3.1.2. Actual Irish housing market: 6](#_Toc135736286)

[3.1.3. Residential price index comparison between some European countries: 8](#_Toc135736287)

[3.1.4. Hypothesis tests: 10](#_Toc135736288)

[3.2. Machine Learning: 11](#_Toc135736289)

[3.2.1 Clustering and Classification: 11](#_Toc135736290)

[3.2.2 Time Series Analysis and prediction: 14](#_Toc135736291)

[3.2.3 Sentiment Analysis: 15](#_Toc135736292)

[4. Conclusions 17](#_Toc135736293)

[5. Appendix 17](#_Toc135736294)

[5.1. Links for datasets: 17](#_Toc135736295)

[5.2. Web scraped data: 17](#_Toc135736296)

[5.3. Other data: 18](#_Toc135736297)

[6. References 18](#_Toc135736298)

# 1. Introduction

The construction sector plays a vital role in most countries economy and Ireland is not the exception. This project aims to explore various aspects related to the Irish housing market, new dwellings completion and residential price indices.

First, knowing that the housing supply is one of the key factors on the price, the analysis focused on Ireland’s construction of new dwellings by year, reviewing aspects like proportion of each property type and floor area size evolution over time.

Then, from a real sample of house listings across Ireland at the time of this analysis, confidence intervals were calculated for the entire Irish housing market.

After, the analysis focused on the study of the residential price index and contrasting it with other European countries to find differences or similarities over time, also prediction and forecasting machine learning techniques were applied. The outcome of this task combined with the previous, help to predict future housing prices for different property types.

And finally, topics and sentiment analysis machine learning techniques were applied to compare Irish and worldwide news titles about construction and house prices.

# 2. Materials and Methods

The report was entirely performed using Python (version 3.9.13) as programming language and Jupyter Notebook (Anaconda3, version 6.4.12) as the integrated development environment (IDE).

## 2.1. Python libraries and packages used:

* General use and statistics:
  + pandas,
  + seaborn,
  + numpy,
  + re.
  + requests
  + bs4.BeautifulSoup.
  + json.
  + matplotlib,
  + statistics,
  + scipy.stats (st, Shapiro, mannwhitneyu).
  + statsmodels.api.
  + pingouin.kruskal.
* Machine Learning:
  + sklearn.model\_selection (train\_test\_split, cross\_val\_score, GridSearchCV).
  + sklearn.ensemble (RandomForestClassifier, RandomForestRegressor).
  + sklearn.linear\_model (LogisticRegression).
  + sklearn.cluster (KMeans).
  + sklearn.naive\_bayes (MultinomialNB)
  + sklearn.preprocessing (StandardScaler).
  + sklearn (metrics).
  + keras.models (Sequential).
  + keras.layers (Dense).
  + keras\_tuner.tuners (RandomSearch).
  + pmdarima (auto\_arima)
  + statsmodels.tsa.arima.model (ARIMA)
  + nltk.corpus (stopwords)
  + nltk.stem (WordNetLemmatizer)
  + fsklearn.feature\_extraction.text (CountVectorizer)
  + sklearn.decomposition (LatentDirichletAllocation)
  + nltk.sentiment.vader (SentimentIntensityAnalyzer)
  + nltk.tokenize (word\_tokenize, RegexpTokenizer)
  + nltk.probability (FreqDist)
  + wordcloud (WordCloud)

For dataset reading and manipulation was explored ‘Polars’ as a possibility instead of Pandas and it was found that while Polars performance in terms of time of execution for grouping, sorting and aggregations on large datasets is considerably better than Pandas, the last is more syntactically appealing and easier to learn. The decision whether to use Pandas or Polars will depends on the size of the dataset and how important performance is for the project (Chaudhary, 2023).

## 2.2. Datasets and sources (direct links in appendix):

* Ireland New dwellings by property type by year from the Central Statistics Office public database.
* Residential monthly price index for European countries from Eurostat public database, table name “Construction producer prices or costs, new residential buildings - monthly data” and filtering for the indicator “Output price index in construction”.
* House market in Ireland, for real sample data about house prices, a web scraping tool was designed and applied on Daft.ie website capturing almost seven thousand house listings with actual asking price, floor area and property type.
* For topics and sentiment analysis, three different sources were reviewed and scraped. “The Journal” housing news section, “World Construction Today” which is a worldwide news website specialized in the construction topic and “The Irish Times” searching news with house prices as keywords.

## 2.3. Data cleaning and preparation:

For the Residential monthly price index, the full dataset downloaded from eurostat website consists of 17905 rows and 11 columns of which 6 were not relevant for this report for being generic labels for eurostat way of collecting data. The remaining 5 columns were rearranged in a way that the resulting dataset is easier to interpret, summarize and also to make it useful for training machine learning algorithms for prediction and classification.

For Ireland new dwellings construction by property type and year, three different datasets from the “Central Statistics Office” on “New Dwelling Completions” were merged, one for the total number by type of house, other for the average size in square meters for all types and the third for the average size by property type and its weight in the total mix of all types. The resulting dataset containing all that information together were rearranged in a way that is easier to interpret and summarize.

Regarding the housing market in Ireland dataset resulting from the web scraping tool developed for this project, a first raw result of about seven thousand two hundred house listings were filtered by property types and removed the ones where the floor area was expressed in different units such as acres which was most of the times referring to the land an not the actual house leaving a total of almost seven thousand real house listings with asking price, floor area and property type across all Ireland.

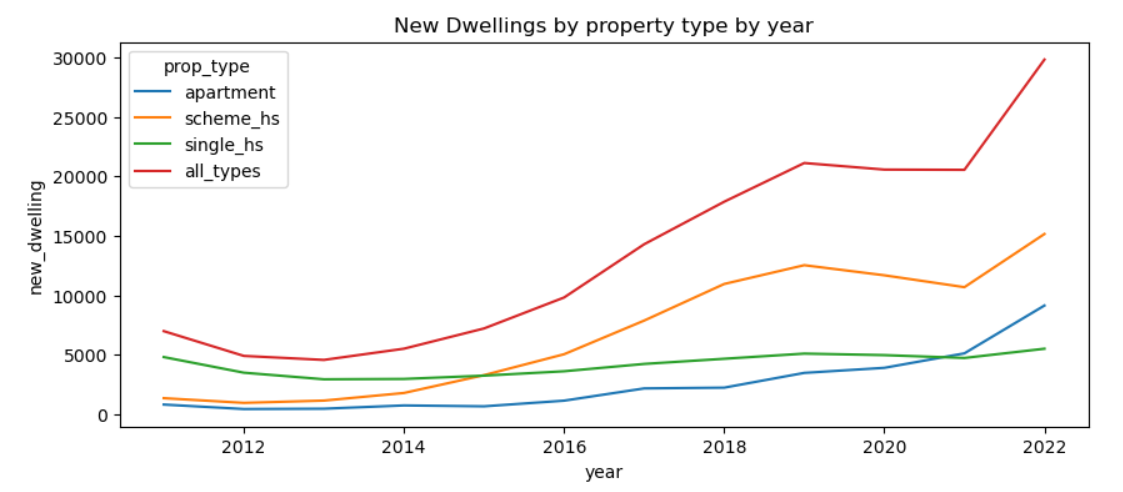
On the matter of topics and sentiment analysis three different datasets were generated from web scraping tools designed for each source with the aim of collecting news titles on the topic of construction and house prices. Also, a fourth dataset with public labeled sentiment information was created in order to train the model, the only cleaning technique before the preprocessing for machine learning models was to remove end of line symbols and unnecessary spaces.

# 3. Result and Discussion

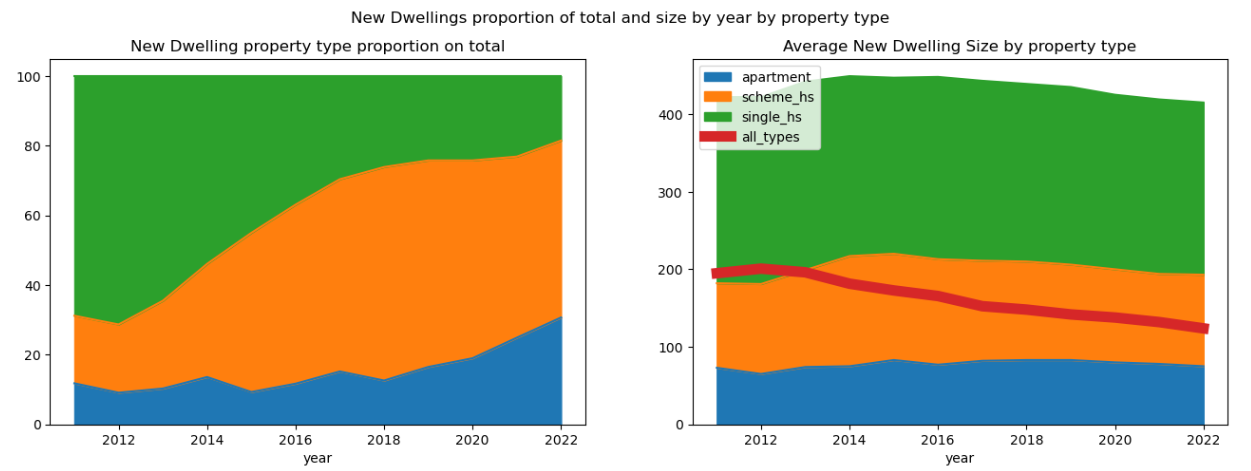
## 3.1. Statistics:

### 3.1.1. New dwellings completion by year in Ireland:

During the period of ten years between 2012 and 2022, according to the data provided by the Central Statistics Office on ‘New Dwelling Completions” the total amount of new dwellings for all property types went from almost five thousand to almost thirty thousand averaging a 20% year on year increase on the national housing supply, mostly due to the categories of apartment and scheme house with 35% and 31% annual increase respectively leaving only 5% for single houses.



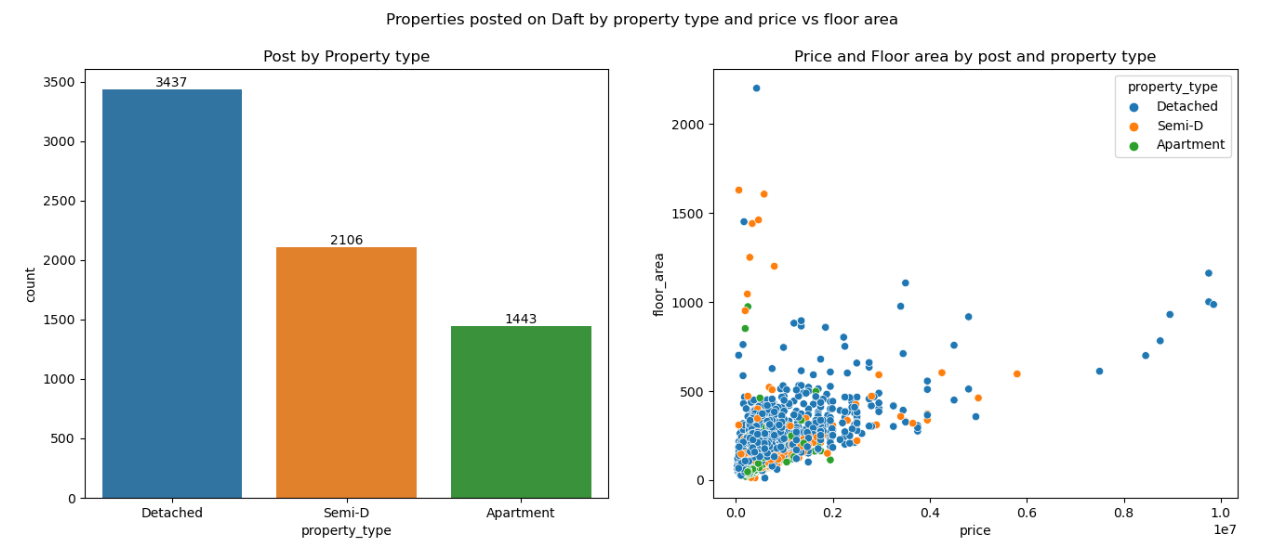
Moreover, from the initial 70% share on all new dwellings completed in 2012 for single houses it went to less than 20% of all new dwellings in 2022 while scheme house and apartment went from 20% and 10% in 2012 to 50% and 30% in 2022 respectively of all new dwelling completions.



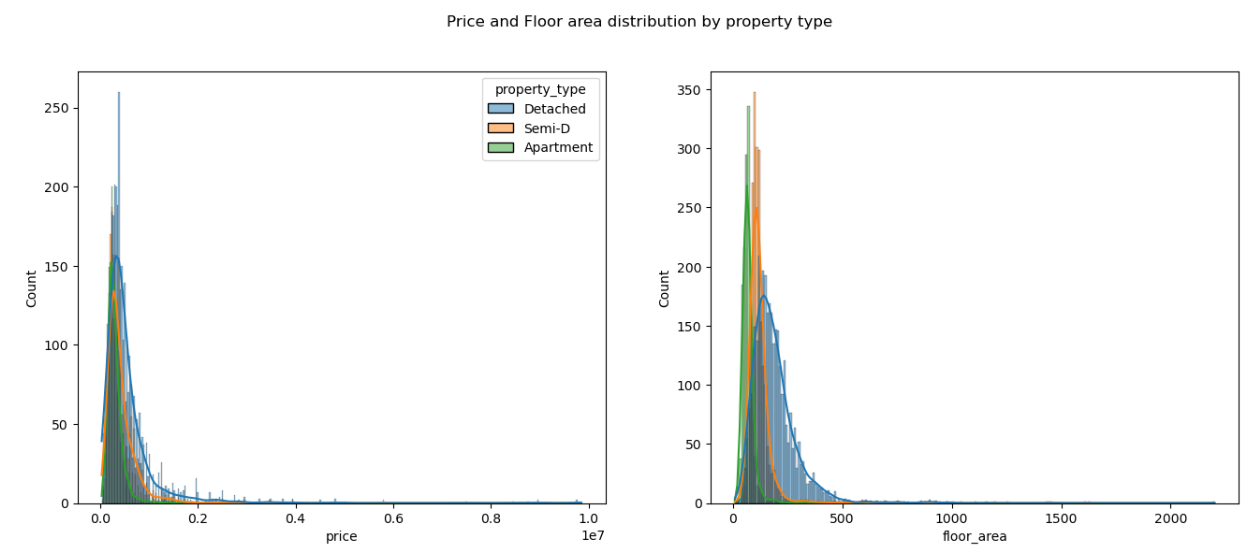
Regarding the average size of new dwellings by year, it kept constant during the same period of time but the mentioned change in property type proportions over the years has decreased the average new house size from around 200 sqm to 124 sqm in the same period of time.

### 3.1.2. Actual Irish housing market:

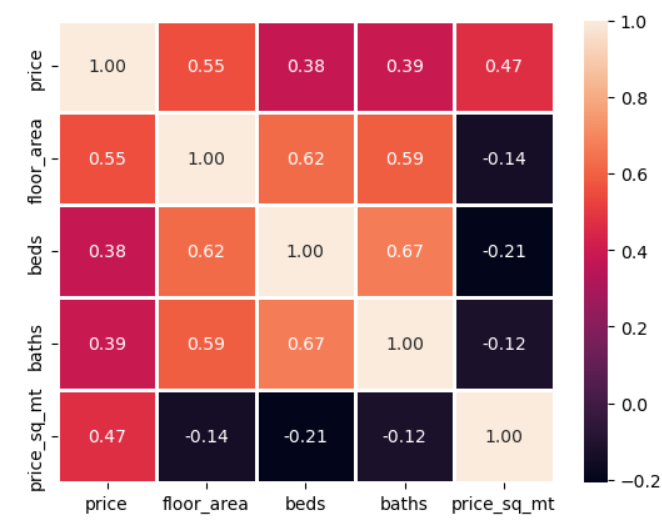
The analysis included almost seven thousand properties listed on “Daft.ie” at the time of this project, of which 49% (3.437) were detached houses, 30% (2.106) were semi-detached houses and 21% (1443) were apartments, with a median of €345.000 and 120 sqm for price and floor area respectively.



Price and floor area distribution for all property types showed to be right skewed thus the median would be more representative.



In addition, on the correlation between variables analysis was observed that floor area and the number of beds and baths are the highest correlated (0.69 and .59 respectively) followed by the price (0.55), but not as high as it was expected to be. This could be due to variables outside of the scope of this project that could be also affecting the price and it would be worth investigating such as location or general condition of the house.



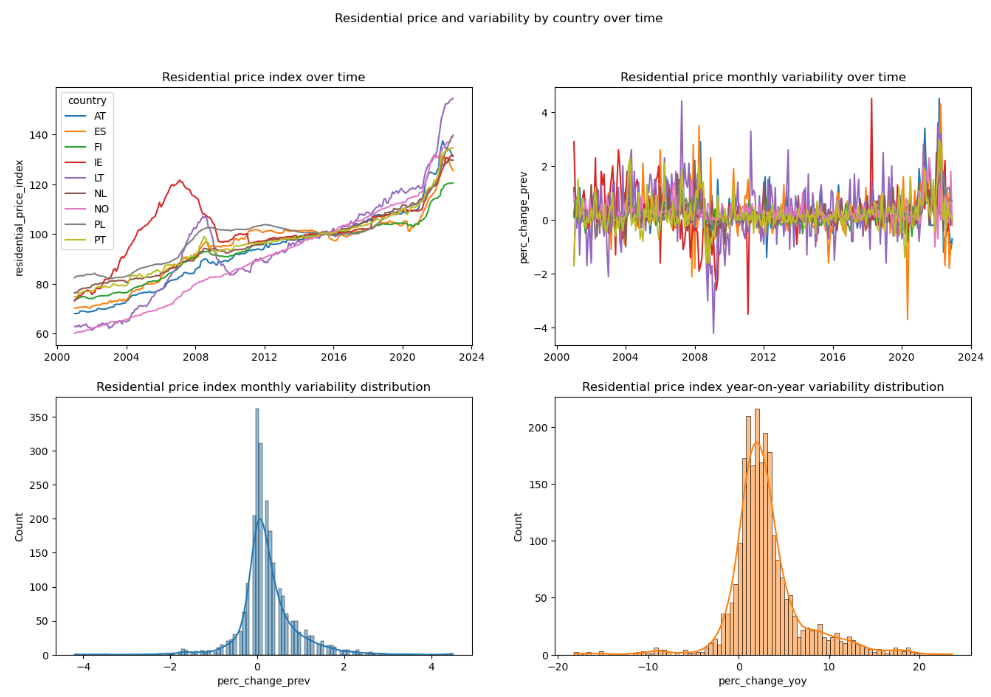
From the sample of all houses listed on “Daft.ie”, with a 95% of confidence level, the average price and floor area size by property type for the entire population resulted as follows:

* Detached house:
  + Average price: between €529.192 and €571.494.
  + Average floor area: between 187 and 194 sqmt.
* Semi Detached house:
  + Average price: between €415.276 and €449.348.
  + Average floor area: between 120 and 128 sqmt.
* Apartment:
  + Average price: between €292.528 and €310.637.
  + Average floor area: between 70 and 75 sqmt.

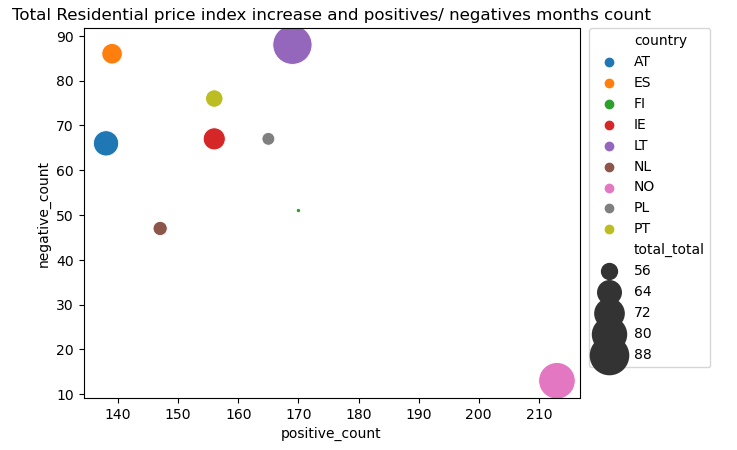
### 3.1.3. Residential price index comparison between some European countries:

The dataset includes the monthly residential price index (base 2015=100) and the percentage change with previous month as well as year-on-year, from January 2001 to December 2022 for Ireland (IE), Spain (ES), Lithuania (LT), Netherlands (NL), Austria (AT), Poland (PL), Portugal (PT), Finland (FI) and Norway (NO).

The residential price index over time has a clear overall increasing trend with some periods of negative months for most of the countries included in the data, while the monthly variability does not show any clear pattern other than higher variability from the year 2000 to 2009 and from 2020 onwards. The percentage change month-on-month and year-on-year distribution showed to be shifted to the positive side as expected due to the increasing trend in the price index.



A more in dept analysis on the amount of positives and negatives months compared with the total growth over the years by country showed some interesting things like the country with the highest growth over the years (Lithuania) is also the country with the highest count of negative months. In the following visualization the size of the bubble is the total percentage growth in the residential price index.



### 3.1.4. Hypothesis tests:

First, Shapiro-Wilk test was used to understand if the percentage change over moth variable was normally distributed as the distribution chart had suggested. The result for this test was a p-value lower than 0.05 therefore the null hypothesis of being normally distributed was rejected and therefore non-parametric tests were applied to compare between populations.

Kruskas-Wallis test was applied with the following null and alternative hypothesis:

* H(0): there is no difference in the residential price monthly variability between countries.
* H(1): at least one of the countries is different than the others.

The result of this test including all countries (9) resulted in a p-value < 0.05 therefore H(0) was rejected because the difference in the median for at least one of the countries is statistically significant.

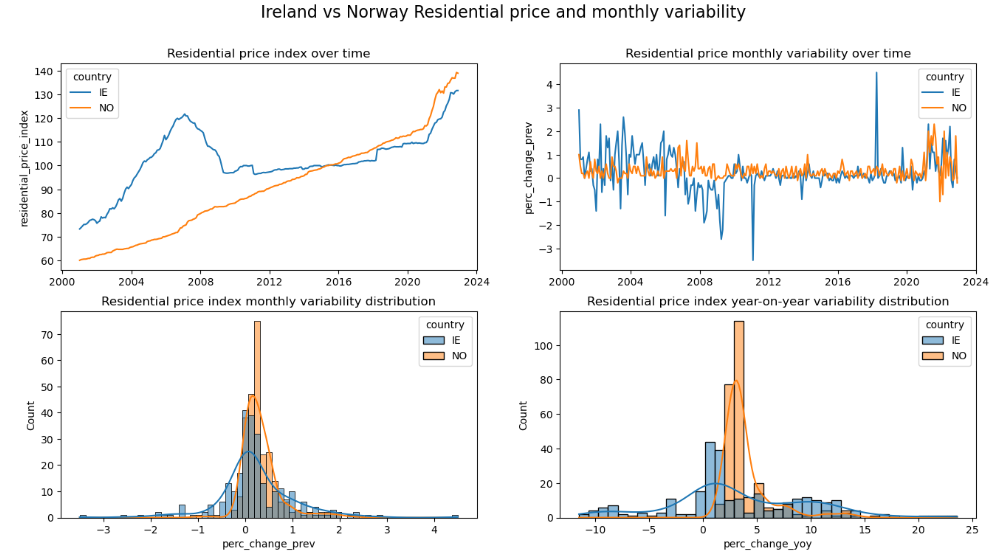
In order to understand if only one of the countries was different than the rest, the same test was repeated nine times (the total number of countries) removing one country each time and testing the remaining eight. The only test which resulting p-value is > 0.05 was when Norway was removed from the pool of countries, thus we can confirm with 95% confidence that Norway is the only country which its residential price monthly variability median is different than the rest.

U-Mann Whitman test was applied to compare Ireland’s residential price monthly variability against each country with the following null and alternative hypothesis:

* H(0): there is no difference and both countries has similar monthly residential price variability.
* H(1): there is difference in the monthly residential price variability.

Same as Kruskas-Wallis test, the only country comparison which resulting p-value rejected H(0) was when testing Ireland against Norway.

As a result of these tests, a more in-depth analysis was required to understand why Norway’s residential price index over time since 2001 is different than the rest of the countries.



Monthly and year-on-year variability distribution for Norway is more concentrated on the positive side compared to Ireland's distribution. Norway monthly price variability between 0% and 1% explains 90% of all variability while for Ireland is only 63%. Also, negative Ireland monthly price variability explains 25% of all variability while for Norway is only 0.05%.

The reason for this stability on the residential price index growth during this period of time and that Norway's economy was relatively less affected by the financial crisis compared to other countries, can be attributed to factors such as effective regulation and surveillance of Norwegian banks, low interest rates, the removal of prior taxation on houses, high job security, and a robust national banking system (Lidtveit, 2018)

## 3.2. Machine Learning:

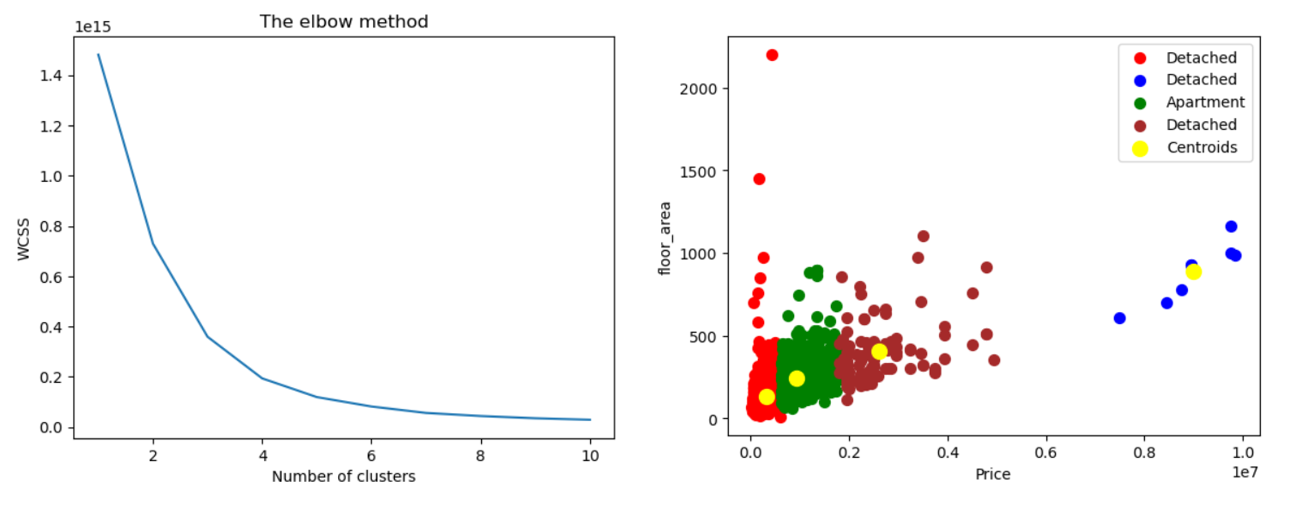
This section is divided into three subsections. The first includes unsupervised learning for clustering and two supervised learning classification algorithms. The second subsection consists on a time series analysis applying some regression models for prediction and forecasting into the future. And the third subsection combines different machine learning models for topic and sentiment analysis on labeled and unlabeled data.

### 3.2.1 Clustering and Classification:

This unsupervised and supervised learning techniques were applied to the actual Irish housing market dataset.

The first task was to fit an unsupervised clustering algorithm, in this case ‘Kmeans”, with unlabeled data using the house price and floor area size for apartments and detached houses, to test if there was a relationship between these two features and its respective property type based on the distance of each data point with their cluster’s centroid. The optimum number of clusters for this test was found as four by applying the elbow method. After comparing the resulted classification with the original data, was observed that the four different classes could be labeled as;

* Class 1: cheap detached houses, probably due to location or general condition.
* Class 2: apartments.
* Class 3: expensive detached houses.
* Class 4: more expensive detached houses.



Regarding the supervised learning classification models, these were applied to the full dataset with the property type as a target variable to test two different models, Random Forest Classifier and Keras Sequential Artificial Neural Network. Both models were first applied with defaulted hyperparameters or minimum modification and a second time with more hyperparameter tuning applying GridSearchCV and cross validation for Random Forest and Keras-Tuner for Sequential ANN.

Random Forest Classifier:



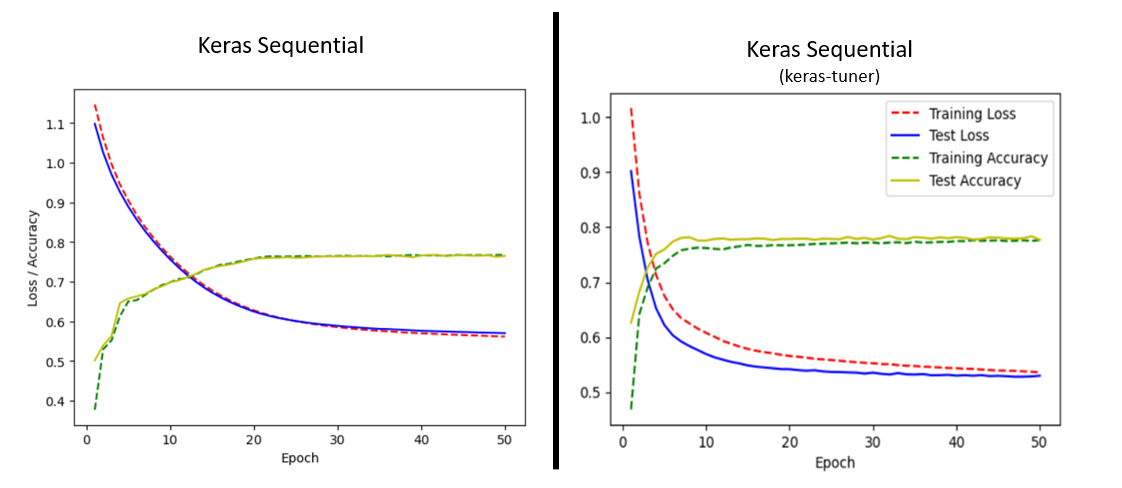
Random Forest Classification summary table:

|  |  |  |
| --- | --- | --- |
| Metric | Random Forest Classifier (default) | Random Forest Classifier (GridSearchCV) |
| Accuracy (training) | 0.99 | 0.87 |
| Accuracy (testing) | 0.74 | 0.76 |
| Cross Validation\* mean | 0.74 | 0.77 |
| Cross Validation\* std | 0.03 | 0.03 |

\*Cross validation cv=20

From this comparison was observed that the hyperparameter tuning did not notably improve the performance of the model but it performed 2-3% better.

Keras Sequential Artificial Neural Network:



Artificial Neural Network Classification summary table:

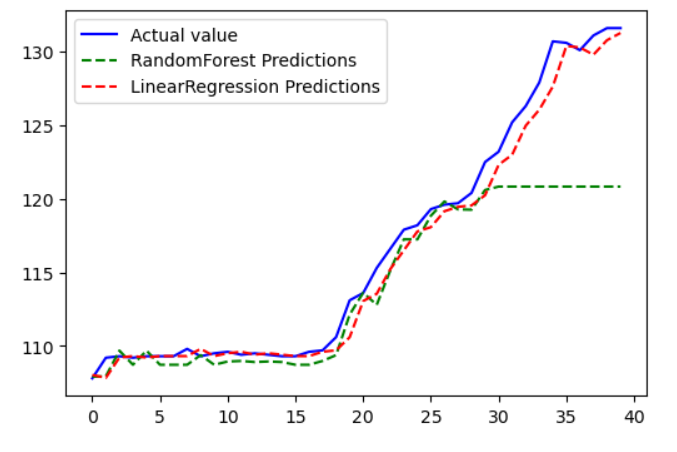
|  |  |  |
| --- | --- | --- |
| Metric | Keras Sequential | Keras Sequential  (with keras-tuner) |
| Test accuracy | 0.76 | 0.78 |
| Test loss | 0.57 | 0.53 |
| Hidden Layers | 1 | 1 |
| Neurons | 32 | 512 |

From this comparison was also observed that the improvement after applying hyperparameter tuning was around 2% on testing accuracy and that it reached to best accuracy and lower loss faster than the previous (less epochs needed). Overall, the best performing classification model applied on the dataset was keras sequential with keras-tuner “best\_model” which was with one hidden layer and 512 neurons.

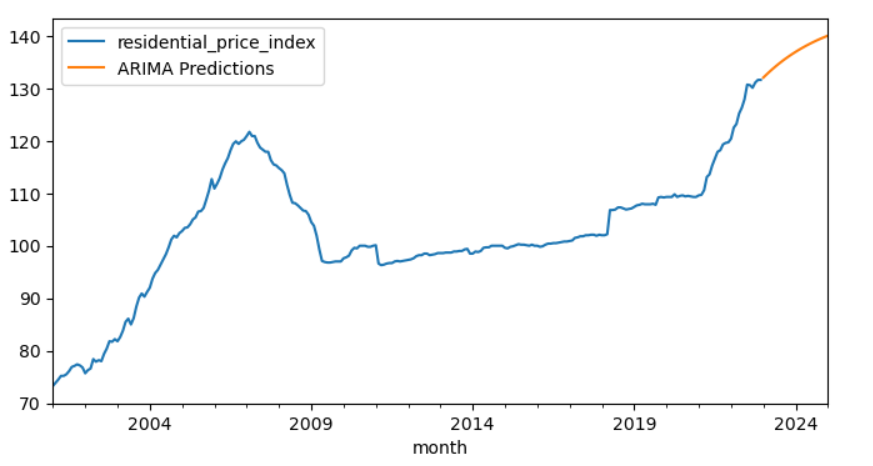
### 3.2.2 Time Series Analysis and prediction:

The time series analysis was performed on the residential price index for Ireland from January 2001 to December 2022, three different regression models were applied for prediction and forecasting into the future.

For the first comparison between sklearn Linear Regression and Random Forest Regressor, a new feature was added to the dataset with the record of the previous month index as input variable (X) for prediction and the model performance was measured with the root mean square error. Both models were trained and tested with the same sets and the liner regression resulted as a better model for prediction in this case with a root mean square error of 1.09 against 4.32 for Random regressor.



For forecasting into the future was applied Autoregressive Integrated Moving Average (ARIMA) model which is similar to the previous linear regression applied using lag values to predict the target variable but ARIMA allows to do the same without manually creation of features. This model has three parameters, ‘p’ which is the number of lag observations, ‘d’ which is the number of times the raw observations are differenced and ‘q’ which is the size of moving average window. Moreover, “PyPI” package has a function called auto\_arima which identifies the most optimal parameters for an ARIMA model, in this case p=2, d=1 and q=1. The resulting forecast showed that the residential price index is expected to continue on its growing trend with an increase of 5.9% from January 2023 to January 2025.

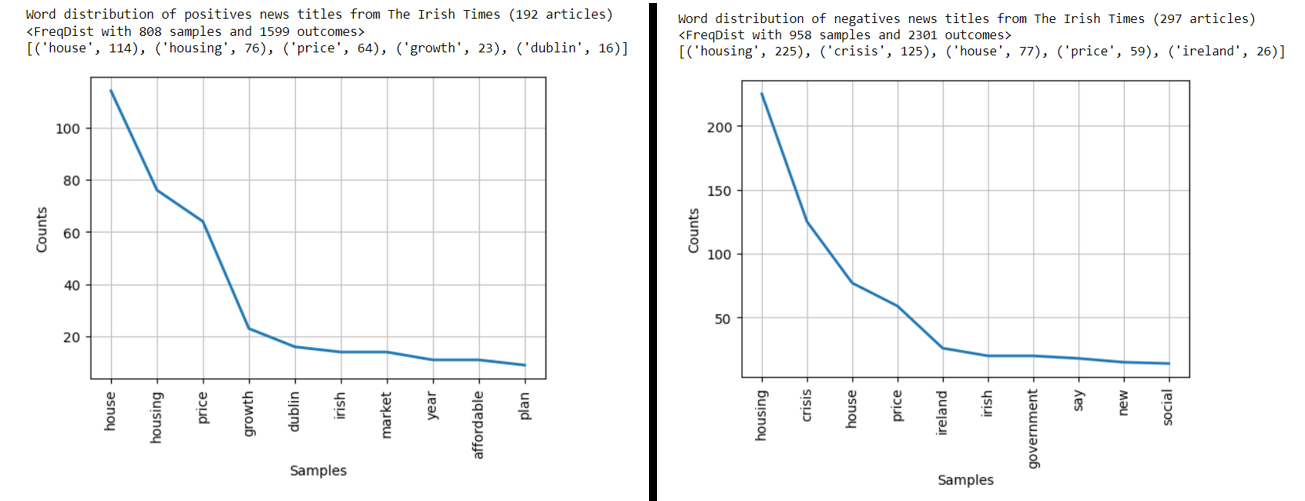


### 3.2.3 Sentiment Analysis:

For topic analysis on general construction news, “The Journal” was compared with “World Construction Today” with the aim of understanding what are the main topics in Ireland news against worldwide news. In order to be able to apply Latent Dirichlet Allocation (LDA) model, the data needed to be preprocessed by keeping only letters, removing stop words which are common words that does not have real meaning for the analysis, also words were lemmatized which is a process of converting words to its base form and finally converted to a matrix of token counts by applying sklearn CountVectorizer method. LDA model was applied to get the main three topics and the top six words per topic to understand its meaning. The result was as follows:

* Irish ‘The Journal” three main topics on housing:
  + Topic 1: New home for homeless.
  + Topic 2: Rental crisis/ evictions.
  + Topic 3: New property price in Dublin.
* Worldwide “World Construction Today” main three topics:
  + Topic 1: New equipment for construction industry.
  + Topic 2: New projects contracts.
  + Topic 3: New building technology.

Sentiment analysis was performed on the “Irish Times” dataset, which is a collection of news titles on the topic of house prices, to understand which were the most frequent words on positives and negatives news titles by applying Valence Aware Dictionary for Sentiment Reasoning (VADER) NLTK module which basically provides a sentiment score based on the words present on each news title, if the resulting compound score for the title is higher or equal to 0.05 or lower or equal to -0.05 it would be labeled as positive or negative respectively. The resulting most frequent words for positives and negatives titles are very similar with some words like ‘growth’, ‘affordable’ and ‘plan’ for positives and ‘crisis’, ‘government’ and ‘social’ for negatives.





Since VADER predictions accuracy can’t be directly calculated because its input was unlabeled data, another model can be trained with labeled data and applied on the joined dataset (labeled and VADER results) to compare accuracy in both section of the dataset. The second model applied was a Multinomial Naïve Bayes classifier trained on a public labeled dataset (link in appendix). The model performed very good on the labeled section of the dataset with 94,5% accuracy and poorly on the vader predicted section of the dataset with 54,6% which could mean that VADER prediction was not accurate or the text does not have a clear sentiment.

# 4. Conclusions

Despite the fact that the housing offer in Ireland has increased several times since the 2000s, the residential price has moved in the same way probably due to an increasing demand and it is expected to continue on this path for the following months.

From the real market sample taken at the moment of this project and considering the forecast results, by January 2025 the average detached house in Ireland would cost between €553.295 and €597.523, the average semi-detached house between €434.190 and 469.814 and the average apartment between €305.852 and 324.785.

From the topic and sentiment analysis was also interesting to find that Irish news topics were more pessimistic compared to worldwide news on general housing. Also, there were more negative news titles than positives in Irish news on the topic of house prices.

# 5. Appendix

## 5.1. Links for datasets:

* New Dwelling Annual:
  + NDA02.csv (<https://data.cso.ie/table/NDA02>).
  + NDA07.csv (<https://data.cso.ie/table/NDA07>).
  + NDA08.csv (<https://data.cso.ie/table/NDA08>).
* Residential monthly price index: <https://ec.europa.eu/eurostat/databrowser/view/STS_COPI_M/default/table?lang=en&category=sts.sts_cons.sts_cons_pri>

## 5.2. Web scraped data:

* House market Ireland: <https://www.daft.ie/property-for-sale/ireland>
* “The Journal” news titles: <https://www.thejournal.ie/housing/news/>
* “World Construction Today” news: <https://www.worldconstructiontoday.com/news/>
* “The Irish Times” news titles: <https://www.irishtimes.com/search/?query=house%20prices>

## 5.3. Other data:

* Labelled data for sentiment analysis (amazon, imbd and yelp reviews txt): <https://www.kaggle.com/datasets/marklvl/sentiment-labelled-sentences-data-set>

# 6. References

Chudhary, P. (2023) “Pandas 2.0 vs Polars: The Ultimate Battle” Available at: https://medium.com/cuenex/pandas-2-0-vs-polars-the-ultimate-battle-a378eb75d6d1 (Accessed: May 2023).

Lidtveit, M, 2018, ‘*An Analysis of the Norwegian Housing Cycle*’, MSc thesis, Copenhagen Business School, Copenhagen. <https://research-api.cbs.dk/ws/portalfiles/portal/59028513/Martin_Lidtveit_Kristin_Albrigtsen.pdf>