# Table of content

# Table of figures

# Abstract

*100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary, 100 words summary.*

# Introduction

# 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).

### 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)

### Datasets and source (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 used 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.

### Data cleaning and preparation: Add why the dataset is useful that way

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 other 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, 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 were 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 the only cleaning technique before the preprocessing for machine learning models was to remove end of line symbols and unnecessary spaces.

# Result and Discussion

1. Statistics;

### 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.

(Add line and staked charts)

### 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.

(Add st\_daft\_1)

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

(Add st\_daft\_2)

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.

(Add st\_daft\_3)

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 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.

### 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.

(Add st\_rpi\_1)

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.

(Add st\_rpi\_2)

### 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 rest.

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, 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.

(Add st\_rpi\_3)

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)

## 2. Machine Learning:

This section is divided into three subsections. The first includes unsupervised learning for clustering with Kmeans 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.

### 2.1 Clustering and Classification:

### 2.2 Time Series Analysis and prediction:

### 2.3 Sentiment Analysis:

For topic analysis on general construction news, “The Journal” was compared with “World Construction Today” in order to understand what are the main topics in Ireland news against worldwide an Irish source were General housing construction news titles from

# Conclusions

Actual price (confidence interval), Arima prediction, future price

# Appendix

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>

## 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>

## Other data:

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

# References

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>