# Housing in Ireland

# Comparative approach through sentiment analysis and machine learning tools for understanding the Ireland housing sector.

# MSc in Data Analytics

Author: Alex Santini

e-mail: [sbs23065@student.cct.ie](mailto:sbs23065@student.cct.ie)

Student ID: 23065

GitHub repository: <https://github.com/alexCCTcollege/CA-2>

Word count (not abstract, titles, references, figure description or citations): 3297

## Abstract

This data analysis project aims to explore various aspects of the housing market in Ireland.

The primary objectives include conducting time series analysis and future prediction on the number of houses sold in Ireland; the findings will be presented through a comprehensive dashboard specifically designed for construction professionals, and stakeholders providing them with valuable information for decision-making and market analysis.

Additionally, a classification algorithm will be developed to differentiate between newly built houses and second dwellings using the same dataset employed for time series analysis aiming to categorize houses accurately, supporting market segmentation and targeted marketing strategies.

Lastly, sentiment analysis will be conducted on Google News headlines, comparing news related to Ireland's housing market with news from foreign countries: this analysis will shed light on public sentiment, perceptions, and overall media coverage, enabling a comprehensive understanding of Ireland's housing market in a global context.

Hopefully. The results of this analysis project will provide valuable insights for construction professionals, policymakers, and stakeholders in the housing sector.

## Introduction

Scope of this project is to give a comprehensive view over the housing sector of Ireland and its challenges: Sentiment analysis will be performed with the intent of gathering opinions around the issue and then compare the results with state-of-the-art data understanding techniques; is Ireland different from other countries when it comes to housing? What factors are the most important?

Once defined the main problems and surrounding debate around the issue two output in the form of classification and prediction tools will be released for stakeholders and policymakers.

## Project planning note

The entire project comprehensive of planning, research and implementation took form in 30 days with an approximate average workload of 3/4 hours per day.

Due to the relatively short timeframe available; the overall strategy has been to perform research first and throughout most of the timeline while performing the first steps in data collection and cleaning before going through the implementation/testing and reporting phase.

As expected, data cleaning and EDA employed more than half of the overall resources; for full breakdown of each phase please refer to Tab 1 at the end of the paper.

## Main Libraries, testing and optimisation

#### Pandas over PySpark

In order to conduct our analysis a range of programming libraries has been critically analysed and researched; in particular paradigms and trade-offs have been analysed for two specific tools: Pandas and PySpark (Singh 2021).

When it comes to data science: both libraries are among the most popular choices by researchers and professionals being both open source and well known in the field.

Pandas offers functions and structures for working with structured datasets, it can manipulate data using a wide range of functions similar to a structured query language such as join, merge, filtering and so much more; due to its overall simplicity pandas tends to be the first choice for data analysis.

PySpark on the other hand offers a wide range of solutions when it comes to big data and scalability of the task which pandas could not handle; it provides an interface into Apache Spark: main advantage of Spark relies in its ability to distribute computing tasks across multiple machines.

In terms of Performances PySpark works better than pandas when dealing with larger datasets by using distributed systems while pandas can handle datasets only up to few Gigabytes.

Speed is also in favour of PySpark when working with huge amount of data due to its parallel computation feature as pandas stores all data in memory while PySpark is able to retrieve only the requested data from the disk: this particular feature is called “lazy processing”.

However, the intent of this paper is to present key insights gained from relatively small datasets: the data and models performed are measured and optimized for a quick reading and quick computation; often in this research smaller samples are preferred over memory intensive datasets as the final research output wants to be similar to a prototype than to a full-scale production algorithm or product.

For the reasons listed above Pandas is to be preferred for its ease of use over PySpark: it does not require setting up a distributed computing cluster for running the code, meaning a fastest approach when it comes to EDA and overall data understanding; Pandas has great integration capabilities with other libraries such as matplotlib, scikit learn and other statistical libraries; Due to its huge community Pandas also offers huge support through forums and documentation as millions of users are able to document their code online on a scale not comparable to PySpark.

In summary due to the light nature of this project Ease of utilization has been chosen over complex environments as performances and memory consumption did not pose a threat in the realization of this project.

#### Testing and optimisation

The overall strategy for testing the validity of the code has been formulated by employ two main points:

1. A frequent use of samples during data pre-processing and transformation: either by using functions like “DataFrame.head()” or manually selecting samples, this step proven to be essential for understanding and testing the right approach after major transformation and important code blocks.
2. A frequent use of printing statements in for loops in order to trace back any possible error within the for-loop logic, even though especially useful during the testing phase.

Optimisation has been achieved through multiple means:

1. In scenarios where the size of datasets created limitations, a common approach has been to select smaller samples to reduce computational costs for algorithms. The overall objective was to ensure that the more time-intensive algorithms, such as creating sentiment analysis datasets or running classification algorithms, complete within approximately 10 minutes.
2. For bigger libraries only particular parts of modules have been imported to optimize computation resources.
3. The methods employed in the code have been chosen with an eye towards potential scalability issues in the future. For instance, under-sampling techniques have been selected to address the class imbalance problem in classification. This choice was made considering the large amount of data available in the PPR datasets.
4. Where dataset has been created by web scraping using functions (sentiment analysis) to optimize efficiency, the data has been stored in databases, using MySQL, after the initial iteration. Subsequently, instead of re-running the code, the data has been read from the database, saving computational resources and time.

## Data understanding

As mentioned, this project is aiming to deliver three main pieces of work: a sentiment analysis on housing related headlines; a time series analysis on number of houses built with future predictions and a classification algorithm for classifying new and second-hand dwellings.

#### Sentiment analysis dataset

Sentiment analysis has been performed by scraping article headlines from Google news for three English speaking countries: Ireland, US and UK.

The license for the chosen library for scraping data can be found here: [Google news license](https://github.com/Iceloof/GoogleNews/blob/master/LICENSE).

Scope of the analysis is to compare the sentiment from different countries; the analysis is considering only English-speaking countries to avoid issues and biases due to translation.

The library chosen proven to be the right choice mainly for two reasons: instead of scraping webpages for specific newspapers; random scraping on Google news have drawn a more comprehensive datasets consisting of various data sources.

Google News also provided the ability of extract headlines by dates, giving the ability to drawn insights from different countries and time periods.

Main limitation of this approach would be that data extraction is restricted to the first 100 articles for each inserted period, with periods set by me as 6 months long. This means that articles are not solely extracted based on their published date but, as per any google search, Search engine optimizations might have played a role in which articles have been picked per each sample of N=100; although the sample size is big enough to be representative.

The terms chosen for the google search have been “housing Ireland”, “housing UK” and “housing USA” respectively; terms have been chosen as wider and high level as possible to avoid biases.

#### Time series analysis and classification tool dataset

Data for this two parts comes from the residential property price register, license here: [Property price register license](https://www.propertypriceregister.ie/website/npsra/pprweb.nsf/page/ppr-home-en).

It consists of roughly half million registered properties data; as stated: “It includes Date of Sale, Price and Address of all residential properties purchased in Ireland since the 1st of January 2010, as declared to the Revenue Commissioner”.

For the classification part, due to its massive size a random sample of N=50000 have been used for the analysis as the scope of the project was to create a working demonstration of both modelling and dashboard for construction professionals which in a hypothetical scenario might have been properly scaled further.

Dataset size has not being an issue for the time series purpose as data has been summarized by counting instances by sales date de facto shrinking it to a more manageable size.

#### Other datasets

Ireland and Scotland most recent Census data have also been used throughout the project for comparing housing across countries, unfortunately as of the time of this project, the latest Census data for Ireland has not been published. Therefore, the available data for Ireland must be based on the 2016 Census. licenses here: [Scotland Census 2021](https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/households/household-estimates/2021), [Ireland 2016 Census (both Glossary and Data)](https://www.cso.ie/en/aboutus/whoweare/copyrightpolicy/).

## Sentiment analysis

As introduced, first part of this project aims to better understand the overall debate around housing in Ireland and to compare its sentiment with other English-speaking countries.

As discussed, the approach taken has been to utilize news headlines as main source of data, why headlines and not for example subreddit or tweets?

Housing can be a quite sensitive subject; relying solely on people comments might introduce biases; headlines on the other hand tend to depict a more neutral perspective returning more consistent data.

however , choosing news as data source entails several implications: the data gathering process utilizes samples therefore relying on size to be representative; using samples means also having slightly different results if choosing different periods or running the code in a different time.

As shown in Tab.2 we managed to retrieve more than 100 headlines per year per country with UK having the highest number of articles followed by Ireland and then US; as expected, housing tends to be a local issues, it’s understandable that a search for “housing US” did not retrieve as many news as per “Housing Ireland”; another possible approach would be to compare cities instead of countries.

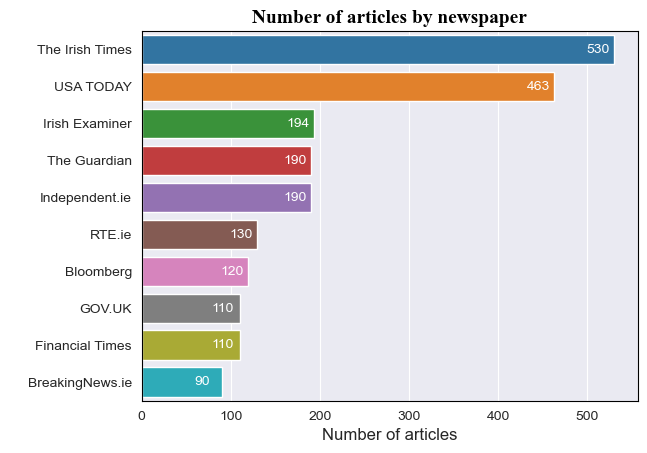


Figure 1 Number of articles retrieve by newspaper

We assigned a polarity score (from -1 to 1) to each data point using the Text Blob library.

The VADER library, although an option, was not suitable due to being specific tuning for social media sentiment, which did not align well with the assumed neutrality of the data.

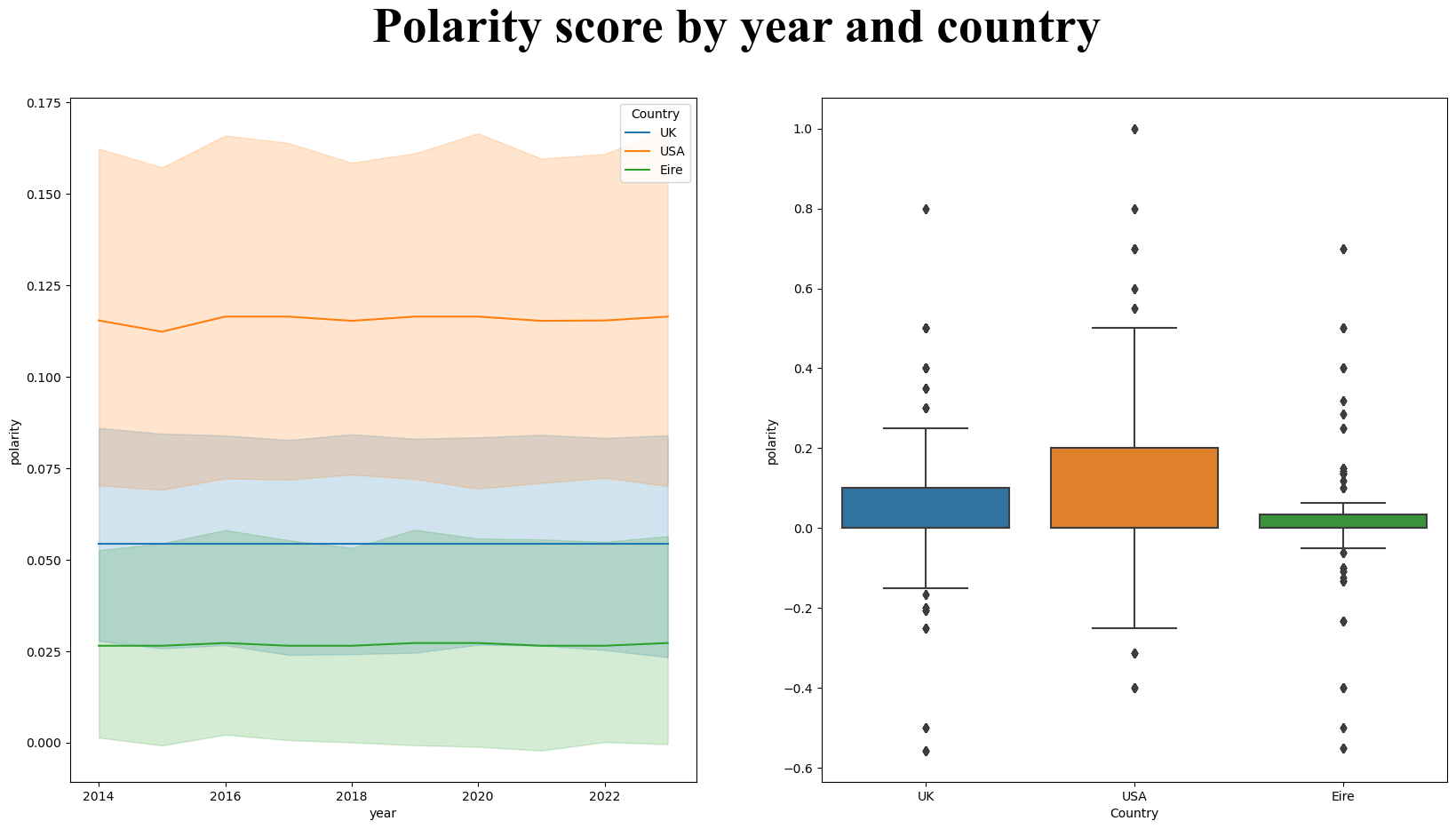


Figure 2Polarity scores by year and country

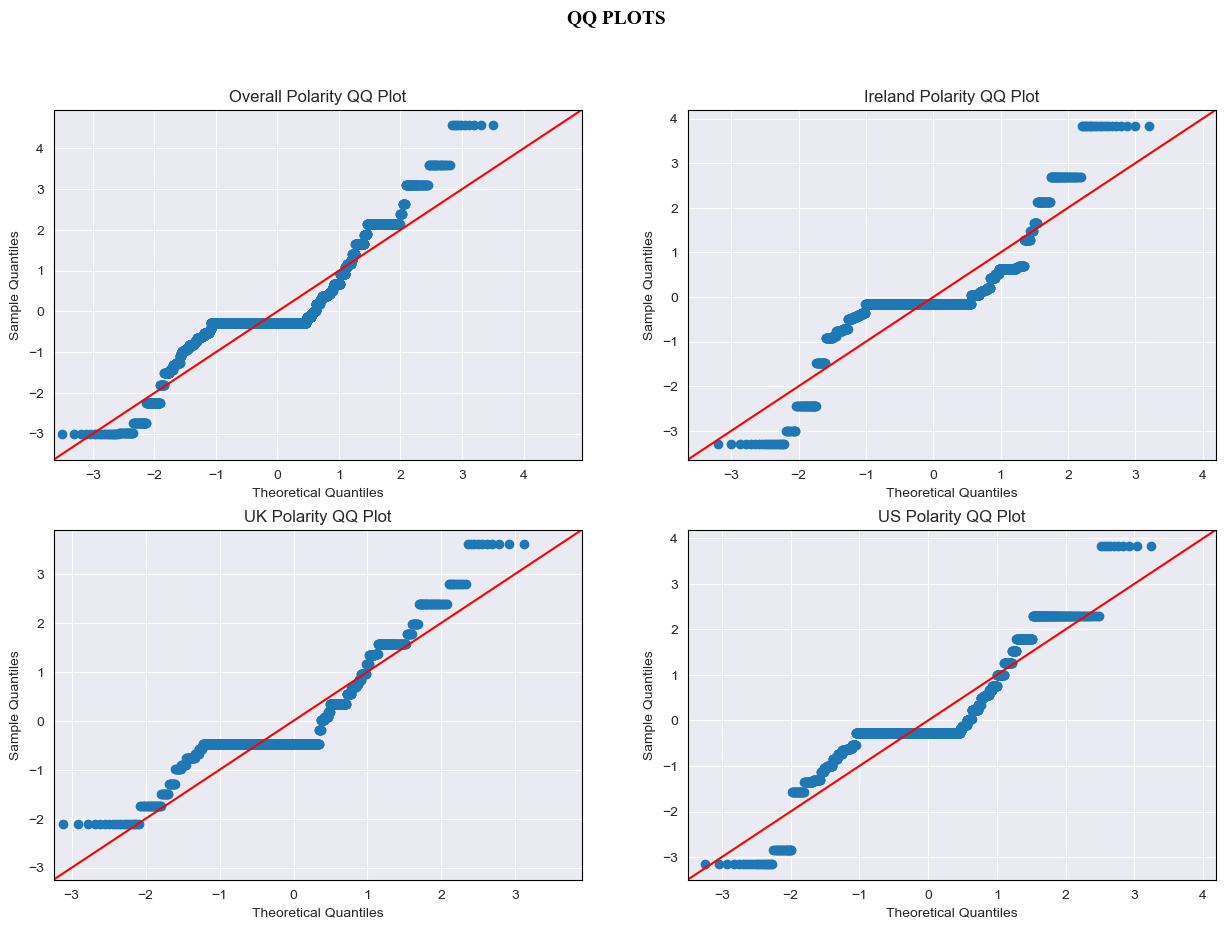
As expected, Tab 3 and Figure 2, polarity scores averaged around 0 for all three countries due to the neutral nature of the data, however the distributions seem to imply a substantial difference in scores: especially between Ireland and US.

Figure 3QQ plot for countries

In order to deepen the understanding of the polarity distributions: multiple Statistical test have been run in order to asses if the differences in scores were statistically significant.

As shown in fig 3 the data did not appear to be normally distributed; to assess if the difference across the three countries were significant a Kruskal Wallis nonparametric test has been performed with the following null and alternative hypothesis:

A picture containing text, font, receipt, white

Description automatically generated

Having a very small p-value, Kruskal Wallis test did not found evidence of the three groups having the same median values.

We can safely say that the three countries have significantly different scores, what about Ireland and UK? Do UK and Ireland share the same polarity scores?

Another nonparametric, the Mann Whitney U test has been conducted for comparing UK and Ireland with the following hypothesis:

A black text on a white background

Description automatically generated with low confidence

Once again, the test has accepted the alternative hypothesis stating that Ireland and UK have statistically different polarity scores on the news data.

Another way of comparing the data is to have a look at the most frequent word which appeared on the headlines by country, as shown in Tab 4 while general terms associated with the housing topic like “prices” “market” or “mortgage” were prominent in UK and USA there seems to be an Irish Specificity with the term “Crisis” being by far the most appeared word in the data.

To better understand the Ireland sentiment specifically a topic modelling analysis has been run using Latent Dirichlet allocation, even thought criticised for its inability to scale efficiently (not a problem in our case) LDA presents a great alternative for gaining better understanding of a text corpus.

Developing such models implied multiple data transformation techniques such as modifying the strings (capitalization), applying specific stop words, vectorizing, lemmatizing the text and much more.

Specifically, Lemmatization has been selected over stemming as recognized more suitable for the job due to the limited vocabulary available as all data is related to the housing topic: stemming the word might lead to loosing valuable information compared to saving the specific lemma of a word.

The first 5 Results for the LDA modelling are shown in fig 4.

A black screen with white text

Description automatically generated with low confidence

Figure 4 LDA output, first 5 topics

Lastly a Word cloud has been drawn to visually represents the most prominent words of the analysis fig 5

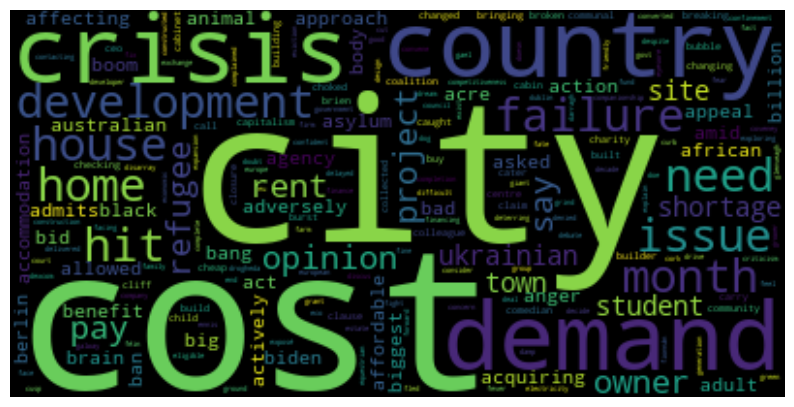


Figure 5 Word Cloud for Ireland corpus

## Classification tool

Second output of this project is to provide possible stakeholders with a classification tool which could correctly classify if a dwelling is new or not by using the price property register and provide some analysis on prices and residencies within Ireland.

Many times, in data analytics, the lack of data can be a true challenge, this is why this section does not simply try to create a plausible classification algorithm; the true aim is to classify instances based on little data as possible: Price and address.

The real value added will come from the ability to being able to extract the data firstly and then tuning the model accordingly.

The strategy is to train a model on a hybrid dataset comprehensive of the price numeric feature and features derived by apply Term Frequency - Inverse Document Frequency on the address data (Onan 2016).

The main concern for this approach has been scalability: the Datasets consisted of half million observations and applying TFIDF on such massive datasets would create relevant performance issues; this is why the project has been threated as a demonstration by using only a sample of the total data.

Another approach would have been to apply bag of words instead of TFIDF; although due to the format of the address, using TFIDF I wanted to give more weight to uncommon concurrency such as specific street names rather than common words which are going to be present in multiple lines such as “street” or “road” (Li-Ping 2002).

In terms of feature selection, two approaches have been tested PCA and Feature Importance using Random Forest Classifier: the latter has been selected as definitive due to being easier to understand in front of stakeholders when presenting findings although PCA maintain the advantage in terms of possible scalability.

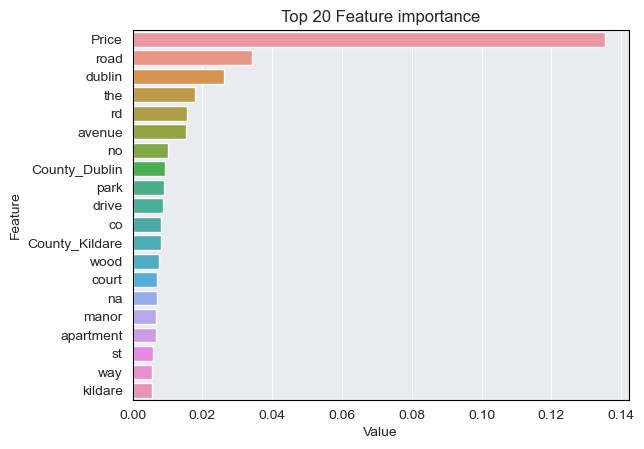


Figure 6 Top 20 Feature importance

Price attribute will have to be scaled as other features (words) will have lower magnitude: as the analysis of prices will show, a great number of outliers have been found this attribute which is why Robust scaler has been chosen as scaling method as using a standard scaler would have result in a distorted view of the original distribution.

As shown in Tab 5 and Tab 6 the average price appear to be quiet different between the new and second hand dwellings, the data also appears to be quiet skewed; statistical tests have been run and the data appears to be not Normal both by employing a Shapiro or a Kolmogorov tests (Kolmogorov tests is better suited for larger samples as in this case).

A nonparametric Mann Whitney U test has been used to asses if the median price of the new and second-hand appartment are equal and the result inequivocably reject the null hypothesis: the two medians are NOT equal.

Class imbalance fig.7 was particularly pronounced, undersampling has been chosen instead of oversampling because of the aformentioned possible issues with scalability.

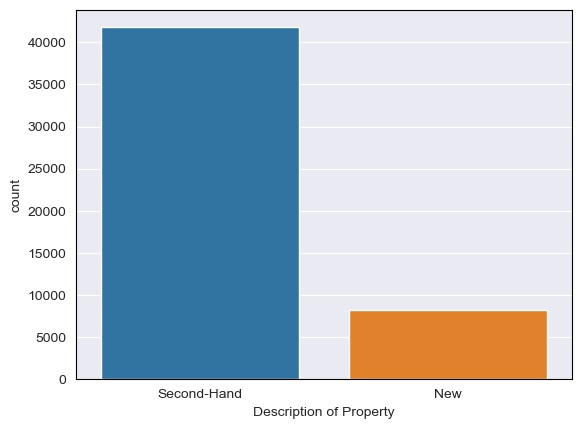


Figure 7 Class imbalance

For the choice of model: Support Vector Classification has been chosen as model and tuned by GridsearchCV over the two main parameters “Kernel” ans “C”, unfortunately “gamma” could have not been tuned due to performance issues; a 5-fold cross validation has been applied.

Main reason for choosing SVM is that this models perform well with high number of features as in this case with thousands of features extracted by using TFIDF; however SVM does not perform well with very large datasets as in this case, which is why the approach since the beginning was to use a sample of the Price Property Register.

Results are reported in fig 8 and confusion matrix in fig 9.

A picture containing text, font, screenshot, design

Description automatically generated

Figure 8 SVM Scores

The accuracy and F1 score are quiet promising if we consider the starting point of the analysis as only two attributes (price and address) were considered.

For future references other possible solution to be taken in consideration might be Naïve Bayes classifiers for the easy scalability and Ensemble methods for trying in retrieve a better precision score.

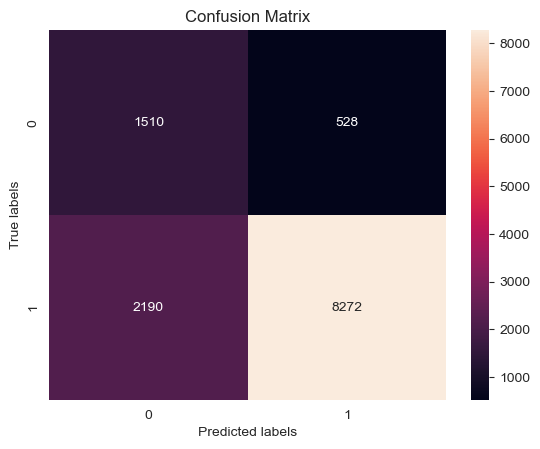


Figure 9 SVM confusion matrix

## Time series

Evidence seems to suggest a lack of offer in terms of housing for Ireland, a good estimator for this might be the average household size by county.

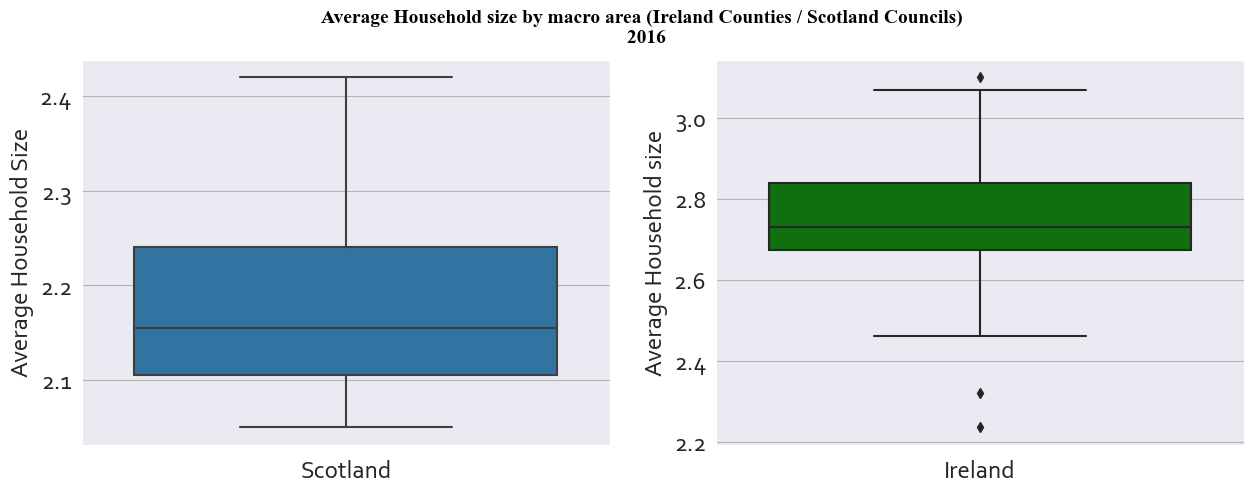
Data from the 2016 has been studied and compared to the Scottish Census data to explore further these statistics, why Scotland? Due to being close in terms of size and population and with a vibrant capital compared to the rest of the country (as Dublin with Ireland) Scotland looks like the perfect natural experiment for a comparison.

Figure 10 Boxplot Ireland Vs Scotland household size

As shown in figure 10 Ireland appears to have a slightly higher average household size by area.

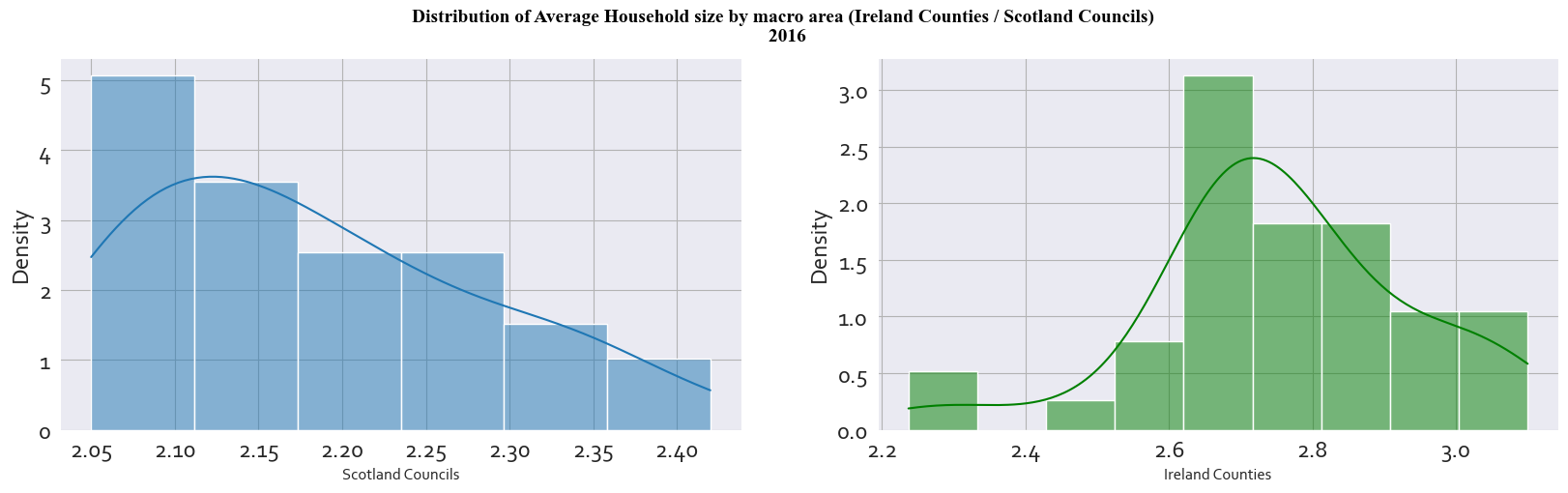


Figure 11 Ireland vs Scotland distribution of household size

Statistical tests have been run for comparing the two distributions: the distributions appear Normal after a Shapiro test, however Levene test found that the two variances are not equal, as so nonparametric Mann Whitney U test has been completed and the two distributions appear to have significantly different medians.

As shown Ireland seems to have quiet a problem when it comes to housing offer which is why the last output of this project aims to be an interactive tool for predicting the number of houses in the near future (up to December 2024) by county.

The dataset of choice is the Property Register, data has been grouped and transformed by date in order to have the number of property sold by date.

As shown in figure using a simple rolling average, data consisted of the last 12 years of sales; it appears that peak are regularly showing around the end of each year.

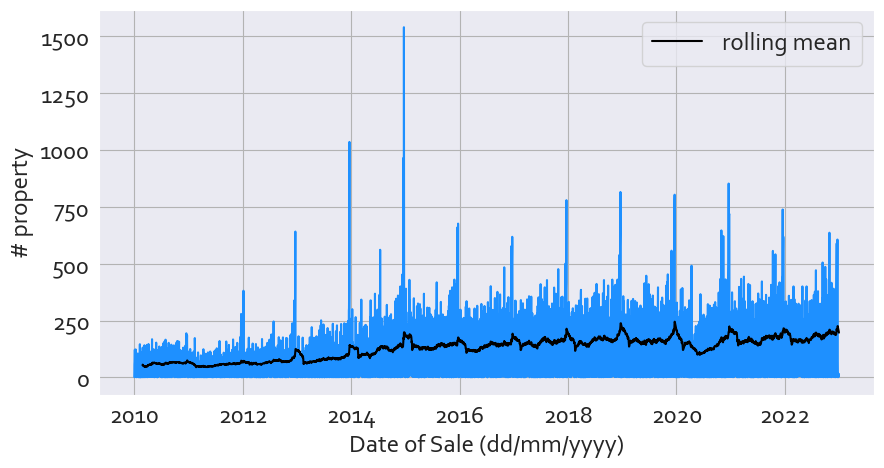


Figure 12 Rolling average for all county data

Before jumping into the proper time series analysis other statistical tests have been run, first of all: can we consider time an independent variable or other factors can shape the amount of property sold by year?

For example, we can group the number of properties sold by two attributes: years and area (for this example defined as Dublin or outside Dublin) can we say that “year” appears independent from the “Dublin” category?

To answer this question a Chi Square test of independence has been run with the following hypothesis:

A picture containing text, font, screenshot, white

Description automatically generated

The retrieved p-value is lower than 0.05, we reject the null hypothesis: the observations between the "Dublin" and "Not Dublin" category is NOT independent of the year.

With this information we decided that, for the dashboards, we will run multiple models based on the county of the properties instead of a generalized model trained on all the data for Ireland.

Last prerequisite for a time series analysis is to check is the distribution is stationary, to achieve that Augmented Dickey Fuller tests have been run with the following hypothesis:

A picture containing text, font, white, screenshot

Description automatically generated

This test has been put within a function and run for each selected county; the modelling will activate only if the H1 condition is met.

Recent years have seen a resurgence of deep learning modelling for time series; however, these types of models tend to be overly complex (Elsayed 2021), especially in settings with high number of features and observations like in this case; on the other hand, gradient boosting models can perform just as well but resulting in more easy to understand outputs (Luo 2021).

As result, the overall approach for modelling has been to split the date attributes into its primary attributes (day of the week, week of the year, month and so on) and then to apply a XGBoost model with Gridsearch to tune 8 total parameters, alternatively sine and cosine might have been applied to the extracted attributes in order to emphasize the circular nature of the feature.

As shown in fig 13 (not county specific model) the model is able to detect seasonality and trends, in order to complement the analysis a confidence interval has been drawn for the results obtained for the years 2023 and 2024: with an alpha value of 0.05; the expected intervals for the predictions can be set between 165.9 and 172.9.

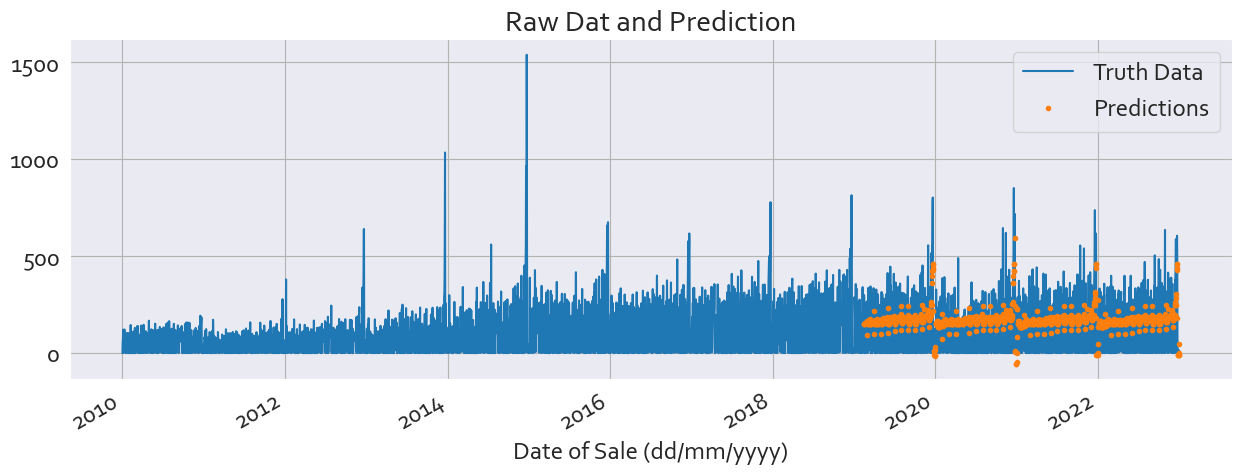


Figure 13 Test results for all county data

The final output for 5 counties has been provided in two different versions in the form of interactive dashboards: version 1 as a birds eye view from 2010 to 2024 for stakeholders interested in the overall trend were they can zoom in and out by clicking with their mouse; Version 2 is specific for capacity building as a slider for all the years have been provided and will benefit stakeholders who are more interested in seen the future workload within a specific timeframe.

The mean square error for each specific county model can be seen after calling the function.

## References

Elsayed, S., Thyssen, T., Rashed, A. G. A., & Schmidt-Thieme, L. (2021). Do we really need Deep Learning for time Series Forecasting?

Joshi, S. (2019). Time Series Analysis and Forecasting of the US Housing Starts using Econometric and Machine Learning Model.

Li-Ping Jing, Hou-Kuan Huang and Hong-Bo Shi (2002). Improved feature selection approach TFIDF in text mining. Proceedings. International Conference on Machine Learning and Cybernetics. doi:https://doi.org/10.1109/icmlc.2002.1174522.

Luo, J., Zhang, Z., Fu, Y. and Rao, F. (2021). Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. Results in Physics, 27, p.104462. doi:https://doi.org/10.1016/j.rinp.2021.104462.

MIT Press. (n.d.). Advances in Knowledge Discovery and Data Mining. [online] Available at: <https://mitpress.mit.edu/9780262560979/advances-in-knowledge-discovery-and-data-mining/>.

Onan, A., Korukoğlu, S. and Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications, 57, pp.232–247. doi:https://doi.org/10.1016/j.eswa.2016.03.045.

Singh, P.K. (2021). Manage Data with PySpark. pp.15–37. doi:https://doi.org/10.1007/978-1-4842-7777-5\_2.

www.iza.org. (n.d.). Housing Satisfaction, Homeownership and Housing Mobility: A Panel Data Analysis for Twelve EU Countries. [online] Available at: <https://www.iza.org/publications/dp/2318/housing-satisfaction-homeownership-and-housing-mobility-a-panel-data-analysis-for-twelve-eu-countries>

‌

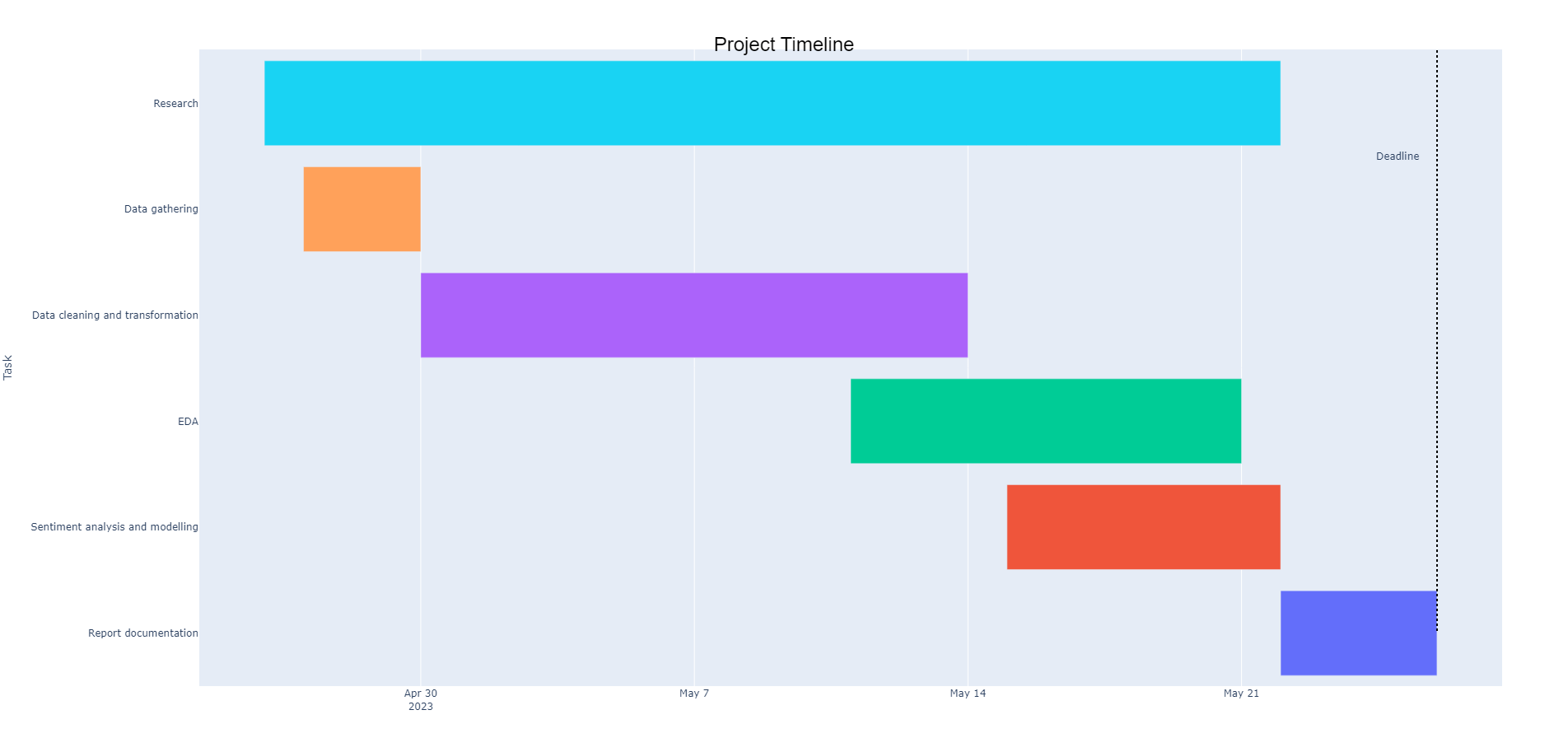


Table 1 Project Timeline

Tab 1. Project timeline

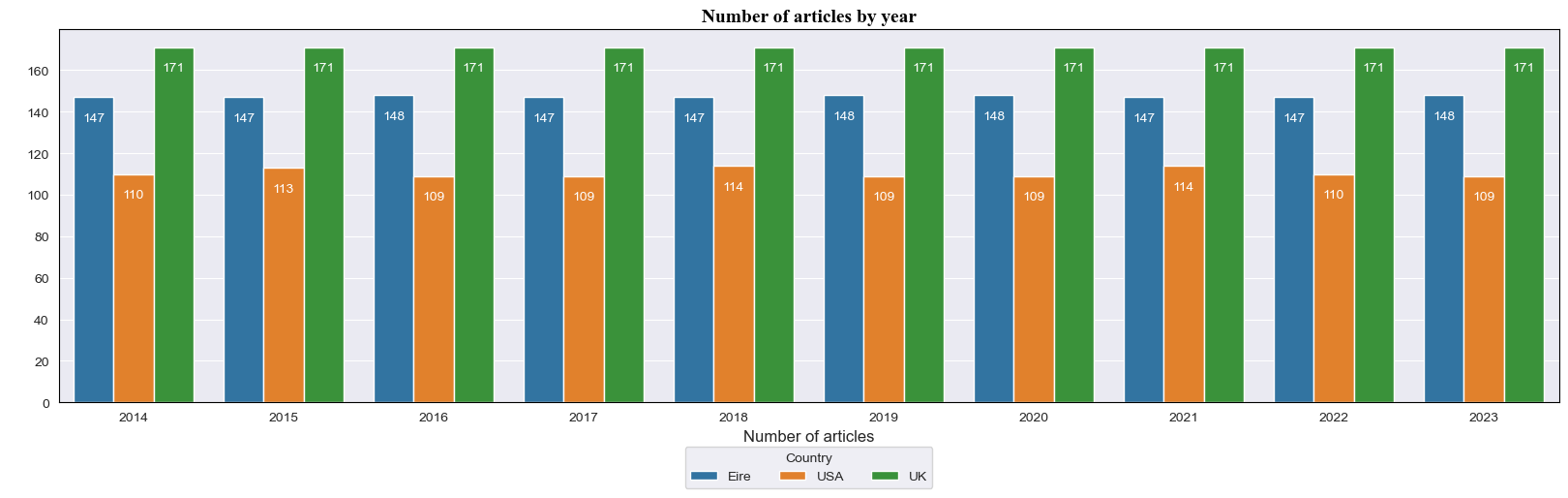


Table 2 Number of articles by year and Country

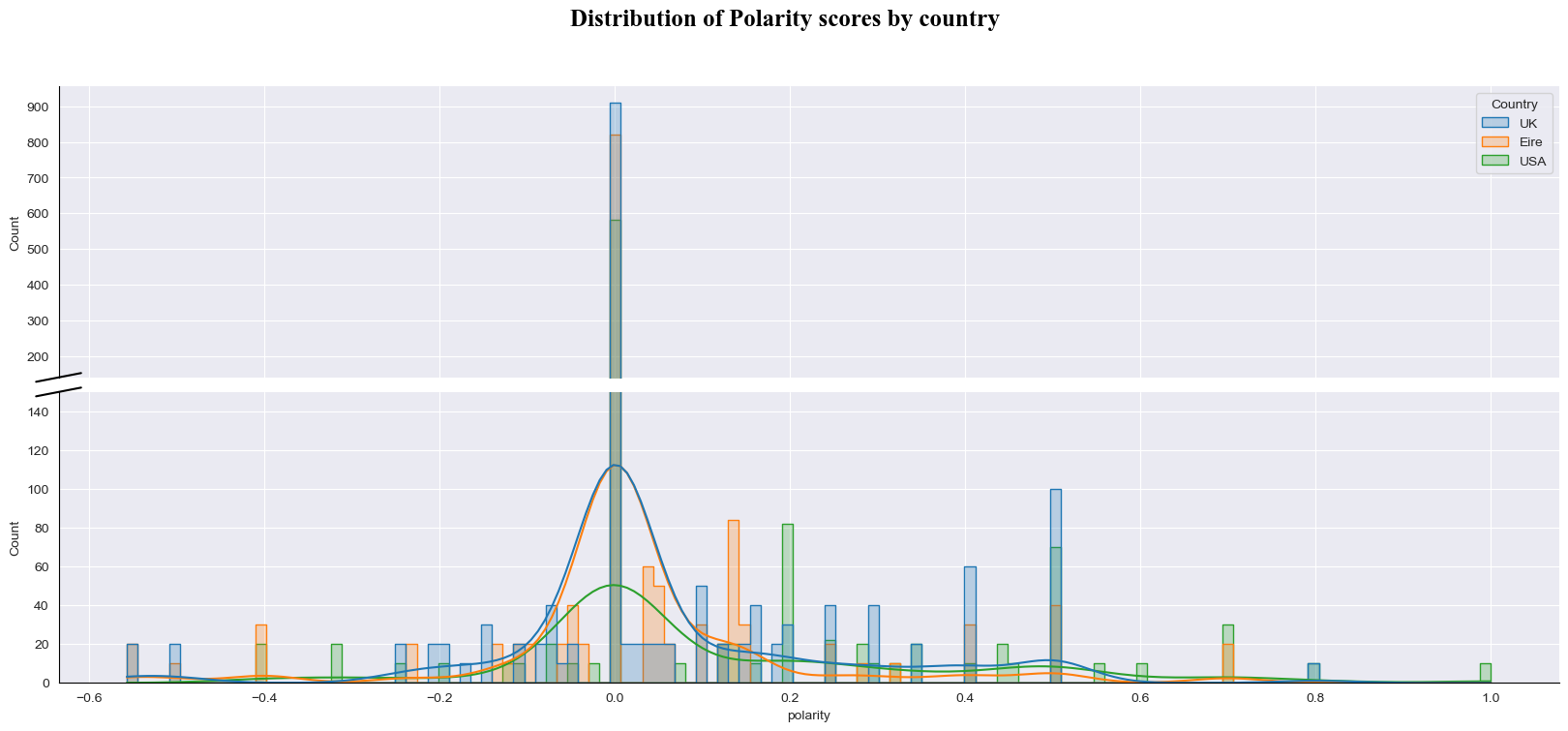


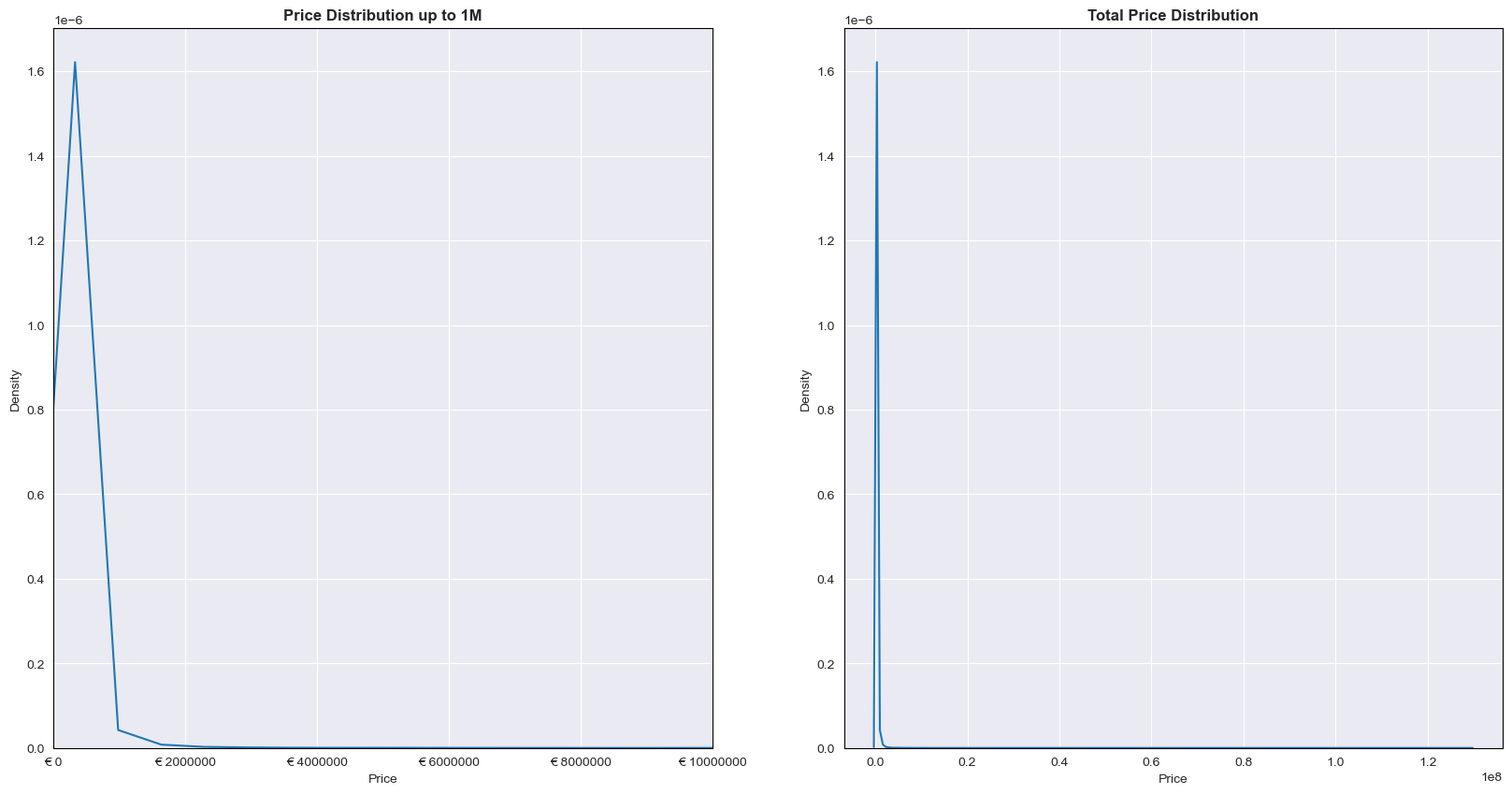
Table 3 Distribution of polarity scores by year \*please notice that the chart is interrupted and the scale changes after value 140 to better shown values.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Table 4 Top 5 most frequent words by Country

Table 5 Boxplot for new and second-hand dwellings prices

Table 6 Price distributions for all prices in PPR (on the left up to 1M) (on the right all prices)