Addressing Ireland's Housing Crisis: ***Forecasting Real Estate Prices Using Regression Analysis and Supervised Machine Learning***

Magdalena Sliwa
National College of Ireland
Data Analytics student
Dublin, Ireland

Abstract— **The housing crisis in Ireland, compounded by factors such as COVID-19, high inflation, surging housing demand, war in Ukraine, and supply deficiencies, necessitates innovative strategies for forecasting real estate prices. This paper employs a Supervised Machine Learning regression model to estimate real estate prices from 2022 to 2030. Visualizations highlight the evolution of real estate prices from 2012 to 2021, while forecasted prices for the next decade in Counties Cork and Dublin are presented using line charts and tables. Comparative insights between both counties shed light on regional variations. Furthermore, a comparison of historical and projected housing price increase rates provides a nuanced perspective on the housing crisis. The results and insights gleaned from this study hold significant implications for the banking sector and can serve as valuable assets for the government in formulating long-term strategies to address Ireland's housing challenges. Data for this study has been sourced from the Property Services Regulatory Authority in Ireland, covering real estate prices from January 1, 2012, to December 31, 2021.**

Keywords — **Real estate, Supervised Machine Learning, Data Analysis, Forecasting, Housing crisis**

I. INTRODUCTION

According to the KBC Bank Ireland [1] and Central Statistics Office (CSO) real estate prices have been increasing since 2014 with a stable growth rate until 2019. Handopo and Rahadi [2] explored in their paper why Generation Z faces major financial obstacles around purchasing a property due to demand increase and limited supply.

Therefore, the housing sector in Ireland has experienced significant influences from various factors since 2000. One notable factor is the recent surge in annual inflation, reaching its highest level in May 2022 since 1984. This inflationary pressure has had a notable impact on the housing market, contributing to the ongoing increase in real estate prices. The annual inflation has been the highest in May 2022 since 1984. [3] Moreover, the COVID-19 pandemic has introduced unprecedented challenges, particularly in the employment sector. The shift towards remote work arrangements has led to an increase in the number of employees working remotely. This shift has, in turn, influenced the demand for properties, as individuals seek homes that accommodate their remote work needs. [4]

Furthermore, the arrival of Ukrainian refugees to Ireland from the beginning of the Russian invasion of Ukraine on 24th February 2022 requires more accommodating facilities and it has increased the Irish population to approximately 35,000 people. In fact, population augmentation is the significant factor determining house prices in Ireland. Consequently, these combined factors have contributed to the continued growth in real estate prices in Ireland. As individuals adapt to new work environments and economic conditions evolve, the housing market responds accordingly, shaping the trajectory of real estate prices.

This paper aims to utilize machine learning algorithms, specifically Linear Regression and forecast algorithms in Tableau Software, to predict house prices for the next decade. By identifying regular patterns in historical data, both algorithms are expected to accurately forecast future values. Furthermore, the study will leverage these predictions to compare housing prices in 2030 between Counties Cork and Dublin.

This comparative analysis will provide insights into regional variations and trends in real estate markets. Additionally, the third objective of the study is to establish the housing price increase rate for two distinct timeframes: from 2012 to 2021 and from 2022 to 2030. By quantifying these increase rates, the study aims to shed light on the pace of growth in real estate prices over these periods, offering valuable insights into the dynamics of the housing market in Ireland.

As a result, the insights derived from this study could serve as a valuable asset to support the Irish Government's long-term plan to address the housing system in Ireland. Despite the launch of the "Housing for All" initiative, which aims to construct 33,000 homes annually until 2030, the outcomes of this project can complement existing efforts and potentially inform the development of new initiatives. Moreover, these insights can assist in optimizing the allocation of funds and resources within the "Housing for All" initiative, ensuring its effectiveness in meeting the evolving needs of Ireland's population. Additionally, the findings may prompt the exploration of innovative approaches and interventions to address specific challenges within the housing sector [5].

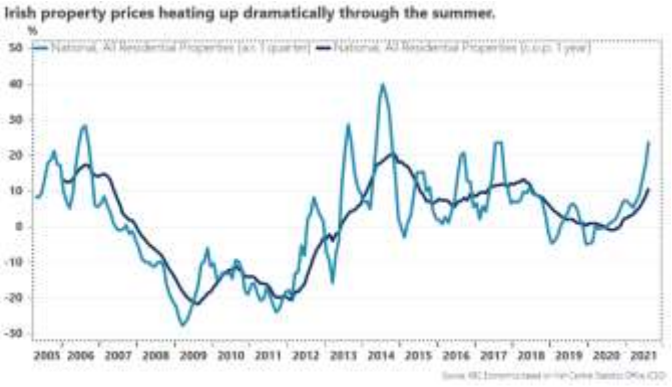


Fig.1. Real estate prices from 2005 to 2021 in Ireland. Source: KBC https://www.kbc.ie/w/irish-property-prices-surge-while-broader-inflation-pressures-build

In addition, the implementation of Machine Learning models enables the prediction of future real estate prices for both Counties Cork and Dublin, leveraging trained models. While the primary objective is to forecast housing prices for the next 10 years, comparing future property prices between Counties Cork and Dublin can aid individuals in making informed decisions about their choice of residence.

This comparison can be particularly valuable for homebuyers seeking to optimize their mortgage eligibility. By assessing projected property prices in both counties, individuals can align their financial resources with the anticipated costs of homeownership, thus ensuring a more informed and strategic approach to mortgage applications.

Additionally, property investment relies heavily on thorough research and informed decision-making. Therefore, the predicted values derived from objective research papers, such as those utilizing machine learning algorithms for real estate price forecasting, hold significant value for investors.

II. BACKGROUND

The literature review section of this paper delves into the insights gathered from various academic papers sourced online, particularly focusing on machine learning and housing price prediction studies. The objective is to select an appropriate machine learning algorithm for the research. By reviewing a range of articles, valuable insights into the methodologies employed in this technological field are gained. The real estate market has been significantly influenced by several factors in recent years, including high inflation, increased demand for property purchases, and rising energy costs, among others. Information gathered from the KBC website has further enriched the rationale behind the chosen topic for this research paper.

Machine learning, a subset of artificial intelligence, is defined by Jijo and Abdulazeez as the study of computational algorithms resulting from the integration of statistics and computer science. Additionally, Grybauskas, Pilinkiene, and Stundziene [8] conducted a study testing 15 different machine learning models in predictive analysis, showcasing the variety available for data scientists and the potential for more precise testing. The study by Grybauskas, Pilinkiene, and Stundziene [8] specifically explores machine learning algorithms for housing price prediction. Their research highlights the precision and accuracy of the Support Vector Regression algorithm, while also comparing it with other algorithms such as Random Forest and Gradient Boosting Machine. Their findings reveal that Support Vector Regression generates fewer prediction errors compared to the other algorithms tested.

Both authors Thamarai and Malarvizhi [9] recommend the utilization of the Scikit-Learn toolbox for machine learning, highlighting its foundation on both scientific and numerical SciPy and NumPy libraries. This toolbox, employed through programming in the Python language, aids in achieving research objectives and training models. Additionally, SciPy and NumPy libraries are utilized for visualizations and scientific calculations, enhancing the analytical capabilities of the research.

In their study, Vineeth, Ayyappa, and Bharathi [10] outlined and elucidated three models for house price prediction: Simple Linear Regression, Multiple Linear Regression, and Neural Networks. Their findings suggest that the Simple Linear Regression model presents a favourable approach to forecasting real estate prices for counties Dublin and Cork.

Furthermore, the real estate crisis in Ireland has broader societal implications, impacting the country's overall poverty rate. According to Social Justice Ireland, the crisis has contributed to a notable increase of 7.5% in the poverty rate, underscoring the urgency of addressing housing challenges in Ireland from a holistic perspective. [6]

III. METHODOLOGY

A. *Dataset background and acquisition*

Datasets chosen for the analysis and housing price prediction in Ireland are publicly available CSV files from the data repository of the Residential Property Price Register [4]. The accuracy of the real estate price data across all counties in Ireland is expected to be high, as the datasets are provided by a statutory body responsible for licensing and regulating the property services sector n Ireland.

Moreover, the datasets include variables describing the type and size of the property, date of sale, providing comprehensive information for analysis. The time frame of the data spans from 1st January 2012 to 31st December 2021., encompassing the previous 10 years, worth of information.

B. *Data pre-processing and analysis*

Data pre-processing as a crucial step, enables data analysts to prepare the dataset appropriately for further analysis. The first initial task of this process is identifying missing values, which influenced the decision of data reduction, resulting in the removal of 5 variables from the dataset using the '.drop' method. Additionally, all missing values were replaced with 'NaN' values to standardize the dataset for analysis. This ensures consistency and facilitates subsequent data manipulation and analysis processes. This method enables to ensure the integrity and reliability of the dataset for accurate analysis.

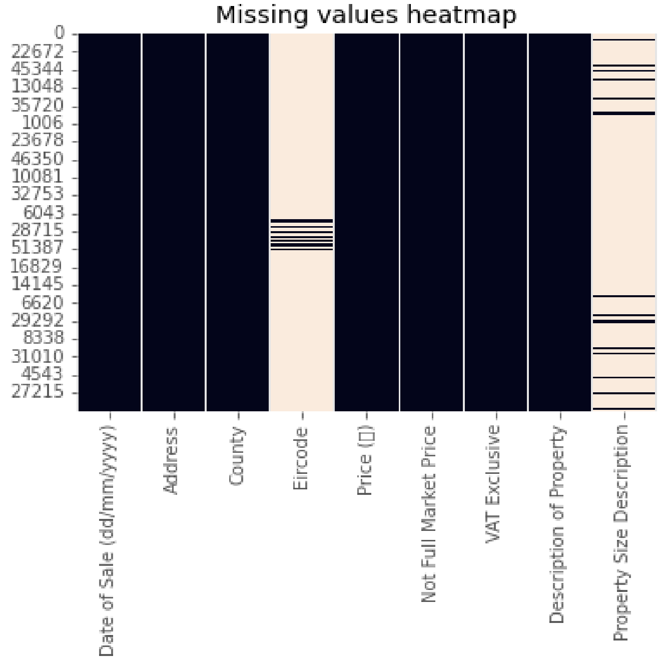
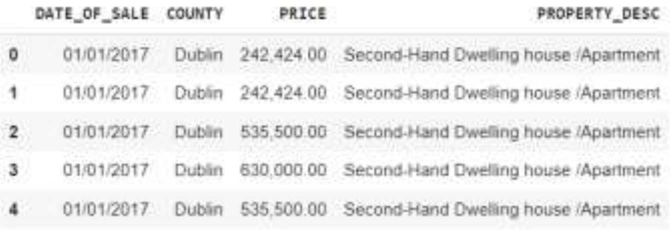


Fig.2. Missing values heatmap, **Python**

Renaming fields is indeed a strategic approach to enhancing the understandability of the dataset, it can also facilitate further programming by using shorter variable names. The '.rename' method is a convenient tool that allows renaming all the variables within the dataset. By its implementation, data analysts can modify the names of columns to make them more descriptive and intuitive. This can improve clarity and ease of interpretation during data analysis. Additionally, using shorter variable names can streamline programming and make code more concise and readable.

Overall, renaming fields using the '.rename' method is an effective strategy for improving the usability and accessibility of the dataset, both for analysis and subsequent programming tasks.



	DATE_OF_SALE	COUNTY	PRICE	PROPERTY_DESC
0	01/01/2017	Dublin	242,424.00	Second-Hand Dwelling house /Apartment
1	01/01/2017	Dublin	242,424.00	Second-Hand Dwelling house /Apartment
2	01/01/2017	Dublin	535,500.00	Second-Hand Dwelling house /Apartment
3	01/01/2017	Dublin	630,000.00	Second-Hand Dwelling house /Apartment
4	01/01/2017	Dublin	535,500.00	Second-Hand Dwelling house /Apartment

Fig.3. Dataset view after pre-processing processes implementation

For conducting analysis, the conversion of the PRICE datatype from object to integer within the 'astype()' method in Pandas was required accordingly. This conversion ensures that the PRICE variable is in a numerical format, which is essential for performing mathematical operations and statistical analysis. Furthermore, as the outliers impact the statistical analysis and distort the accuracy of results, identifying outliers eliminates noisy data. Descriptive statistical methods and Boxplots were utilized for this purpose, particularly focusing on the 'PRICE' numerical value.

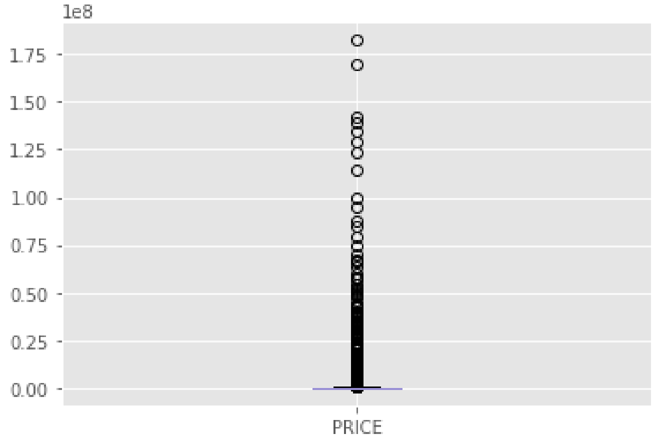


Fig.4. PRICE variable's outliers presented by boxplot, **IBM SPSS**

The removal of outliers has taken place accordingly within the use of NumPy.percentile() method and NumPy library. The outliers were removed with the use of the IQR approach, which is an acronym for Interquartile Range. Three quartiles (Q1, Q2, Q3 for 25%, 50%, and 75% accordingly.) Those were replaced with NumPy.nan as the Null values. The 'dropna()' method allowed the removal of the outliers from the dataset completely.

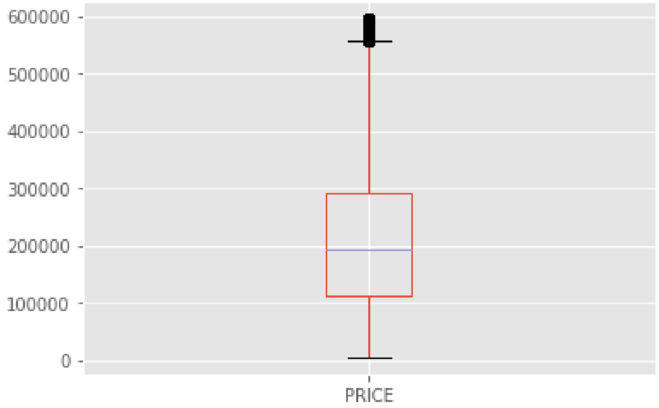


Fig.5. PRICE variable after removing outliers presented by boxplot, **IBM SPSS**

After the pre-processing method, the dataset was converted into an XLSX file format for analysis and visualisation creation in IBM SPSS and Tableau Software. Additionally, a column named 'Date of Sale (dd/mm/yyyy)' was transformed into three columns using Pandas library. By default, this transformation split the date into Day, Month, and Year as distinct variables, facilitating further analysis in IBM SPSS.

Determining causal relationships between the independent and dependent variables requires estimating the dependent 'y' variable value based on the values of independent variables 'x'. This objective is achieved through the regression model within the Machine Learning algorithm of Linear Regression.

The final shape of the real estate prices dataset includes **450282 records** and **4 variables.**

C. *Project Management*

Time management is indeed crucial for the successful execution of any project, and utilizing tools like the Gantt chart can help in planning and organizing tasks effectively. By allocating tasks across the timeline of the project, starting from 24th May 2022 to the deadline of 7th August 2022, it becomes possible to avoid stress and heavy workloads at the last stages of the project. The project timeline can be divided into several key stages, with the first stage focusing on conducting secondary research. This involves reviewing various scholarly papers, articles, and datasets, as well as drawing from personal and professional experience to choose a topic for the project.

The risks associated with this paper include technological problems such as SPSS License expiry, difficulties with finding appropriate scholarly papers, and performance issues with tools like Spyder. There is a risk that the chosen datasets may not be as relevant or comprehensive as expected, impacting the accuracy and validity of the analysis results. Changes in market conditions or data availability could also affect the relevance of the chosen datasets. The RAID log was used to log any issues that arose over time.

IV. RESULTS

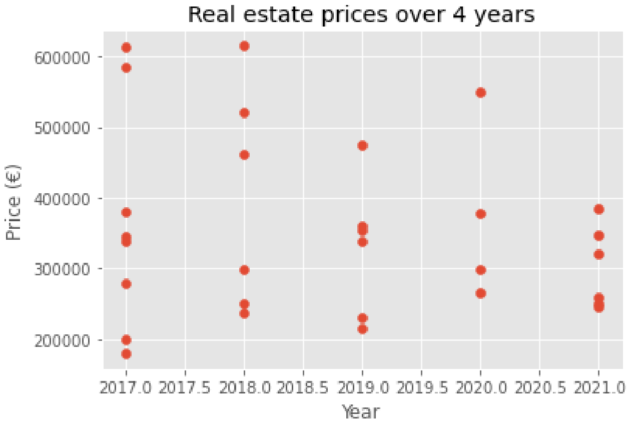


Fig.6. Real estate prices over 4 years scatter plot, Python

Figure 6 likely illustrates a non-linear distribution of data, indicating that a Linear Regression algorithm may not be suitable for analyzing the entire dataset. In such cases, it is common to preprocess the data by filtering rows based on specific conditions using the Pandas library. This approach allows for the extraction of relevant data and the creation of a sample dataset that focuses on specific regions or criteria. In this scenario, filtering rows by conditions using the Pandas library enabled the extraction of data pertaining to housing prices for County Dublin and County Cork specifically. By creating a sample dataset that focuses on these regions, analysts can potentially achieve a more accurate analysis and better understand the relationship between independent and dependent variables within these specific areas. Using this filtered sample dataset, alternative machine learning algorithms or regression techniques more suitable for non-linear data distributions can be applied to perform further analysis and make predictions. This approach ensures that the analysis is tailored to the characteristics of the data and increases the likelihood of obtaining meaningful insights.

A. County Dublin analysis, visualisations

Various data pre-processing strategies such as grouping average price per each of 10 years, converting lists to Pandas DataFrames, and filtering rows by conditions allowed the creation of a new dataset for the Linear Regression analysis and implementation of Machine Learning models accordingly. The implementation of Linear Regression requires two arrays, the input (x) and the output (y).

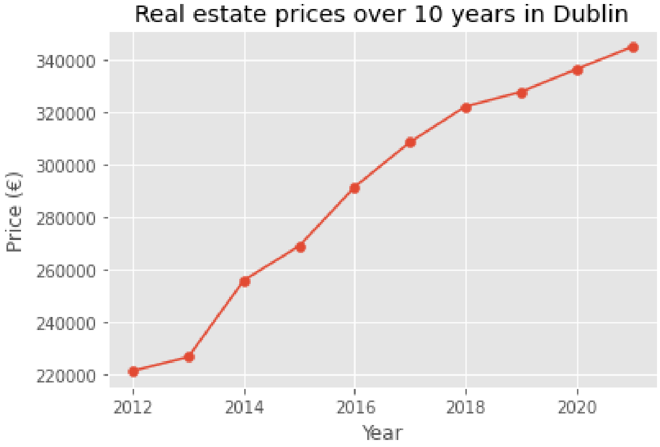


Fig.7. Linear Regression chart displaying real estate prices over 10 years in County Dublin, Python

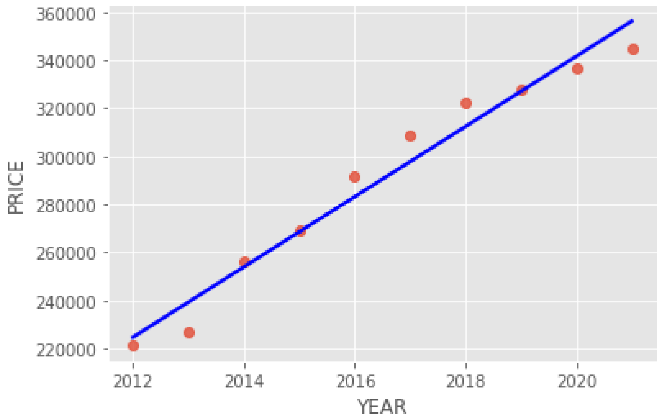


Fig.8. Linear Regression Scatter Plot, Python

B. *County Cork Analysis, visualisations*

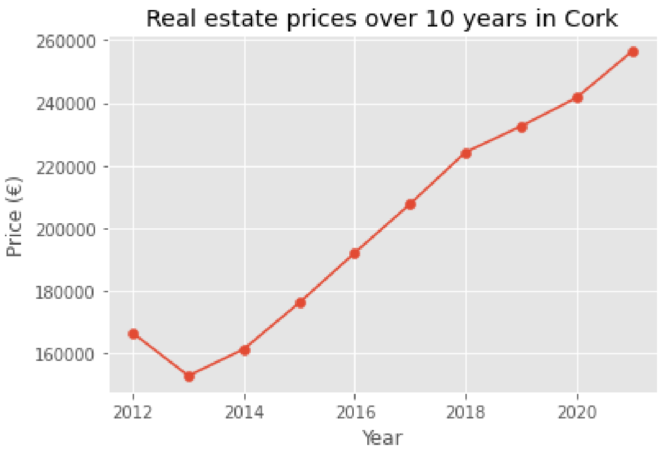


Fig.9. Linear Regression chart displaying Real estate prices over 10 years in County Cork, **Python**

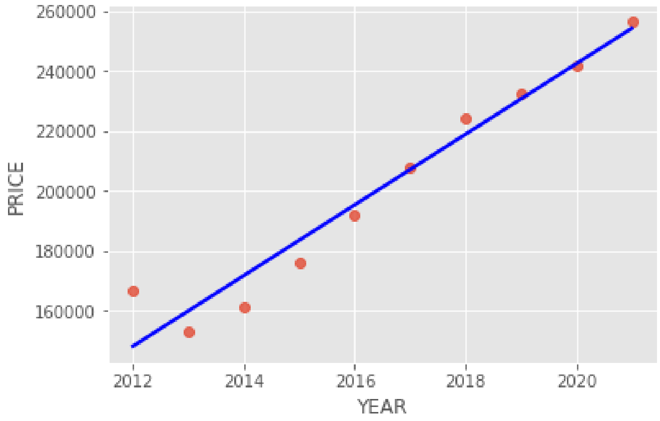


Fig.10. Linear Regression Scatter Plot, **Python**

TABLE I
HOUSING PRICES PREDICTION FOR COUNTY CORK

Prices prediction 2022 - 2030	
County Cork	
2022	*€266,016*
2023	*€277,819*
2024	*€289,622*
2025	*€301,424*
2026	*€313,227*
2027	*€325,030*
2028	*€336,832*
2029	*€348,635*
2030	*€360,048*

TABLE I displays the predicted average real estate prices for County Cork from 2022 to 2030, derived from the Machine Learning model. The prices are represented in Euros (EUR).

C. *Machine Learning Regression model*

Before creating the Regression model, reshaping data from one-dimensional to two-dimensional is essential for compatibility with the model. This reshaping process is typically accomplished using the '.reshape()' method in Python, ensuring that the input data is in the appropriate format for the model. Furthermore, creating Linear Regression Model is required to fit the existing data of the average real estate prices in County Dublin to the model.

After fitting the model to the data, the final step is to present the results. This includes interpreting the coefficients of the regression model, assessing the goodness-of-fit metrics (such as R-squared), and visualizing the relationship between the independent and dependent variables. The results presentation provides valuable insights into the predictive capabilities of the model and its effectiveness in explaining the variation in real estate prices over time in County Dublin.

D. *County Dublin vs County Cork results explanation*

TABLE II
LINEAR REGRESSION MODEL RESULTS

	Linear Regression results			
	R^2	**Adj. R^2**	**R-value**	**P-value**
County Dublin	96.6%	96.2%	98.3%	3.52E-07
County Cork	95.1%	94.5%	97.5%	1.65E-06

The TABLE II results explain the Linear Regression Model performance. **The lower value of the Adjusted R-squared than the R-squared confirms the model accuracy and correct fit measure for linear models.** There is a 96.6% fit of the linear regression for County Dublin analysis, whereas the fit of the linear regression for County Cork is equal to 95.1%.

The coefficient of determination uses other metrics to establish how much of the total variation of the targeted variable is explained by the variation in the regression line. **Therefore, the Correlation Coefficient is equal to 97.5% for County Cork and 98.3% for County Dublin meaning the relationship between the two chosen variables is strong for both counties.**

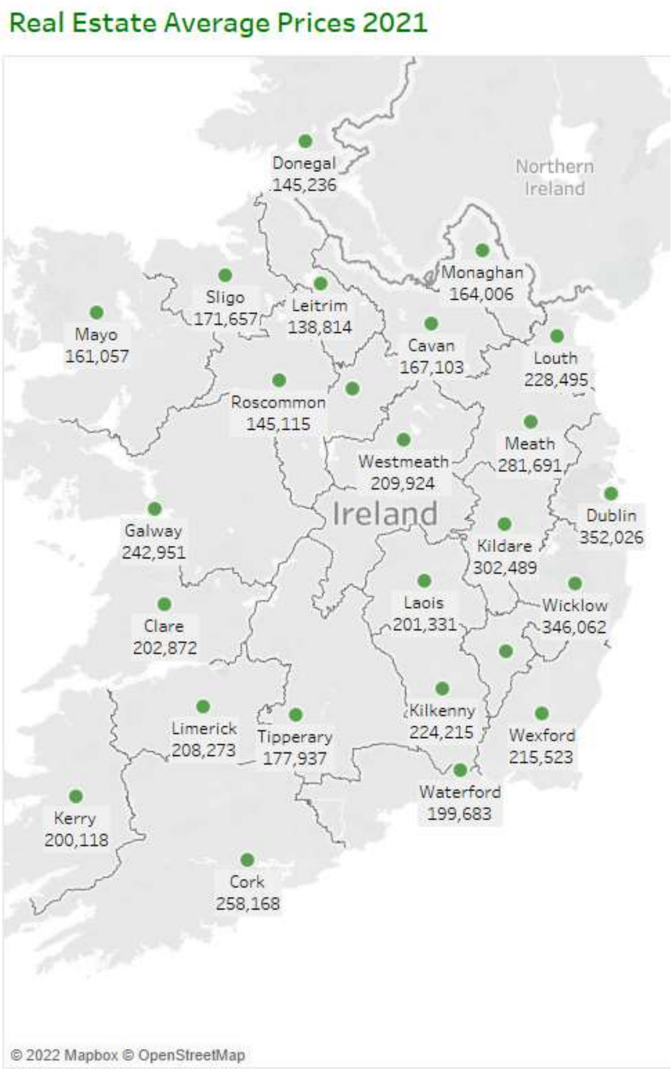
In theory, the p-value determines the probability of extreme results in the statistical hypothesis test, assuming the Null Hypothesis to be correct.

TABLE III
REAL ESTATE PRICES INCREASE FROM 2012 TO 2021 IN IRELAND

Housing Price increase from 2012 to 2021

	2012	*2021*	*Increase rate*
County Cork	€166,401	€258,168	55%
County Dublin	€221,565	€352,026	59%

Table III represents the real estate price increase in Ireland over 10 years from 2012 to 2021 has been significant for both counties Cork and Dublin and the increase rate is equal to approximately 55% for County Cork and 59% for County Dublin.



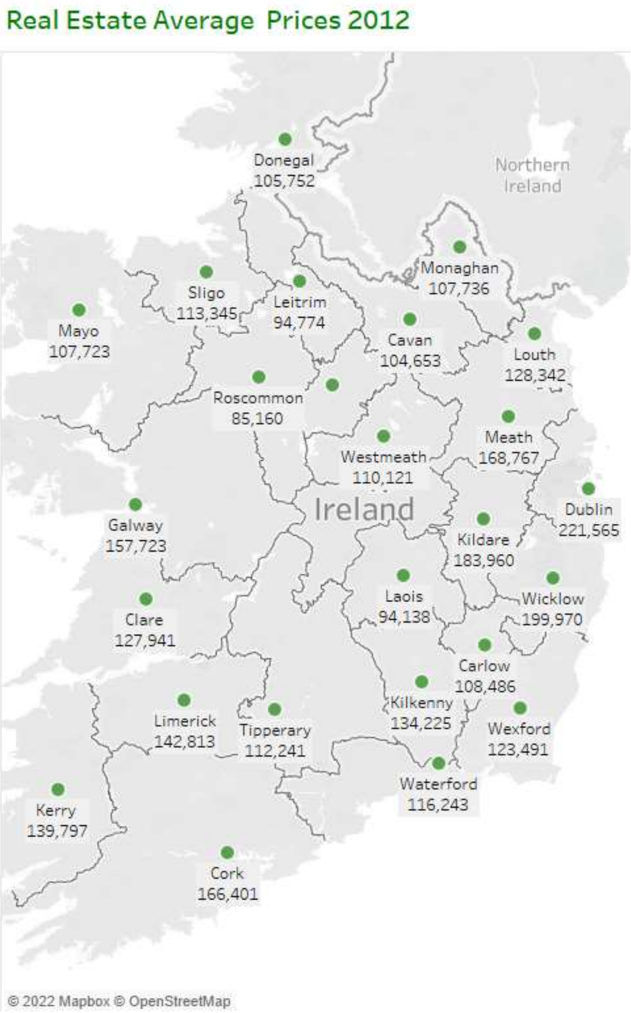


Fig.*11*. Average real estate prices dashboard in 2012 across Ireland, **Tableau Software**

Indeed, County Dublin consistently maintains the highest average prices of real estate compared to other regions in Ireland. The comparison of average prices from 2012 to 2021 reveals a significant increase in real estate values, particularly in County Dublin. This trend is evident in the geographical dashboards presented in Figures 11 and 12. Overall, these geographical dashboards offer valuable insights into the spatial and temporal dynamics of real estate prices in Ireland, highlighting County Dublin's position as a key driver of growth in the housing market.

Fig.12. Dashboard displaying real estate prices in 2021 across Ireland, **Tableau Software**

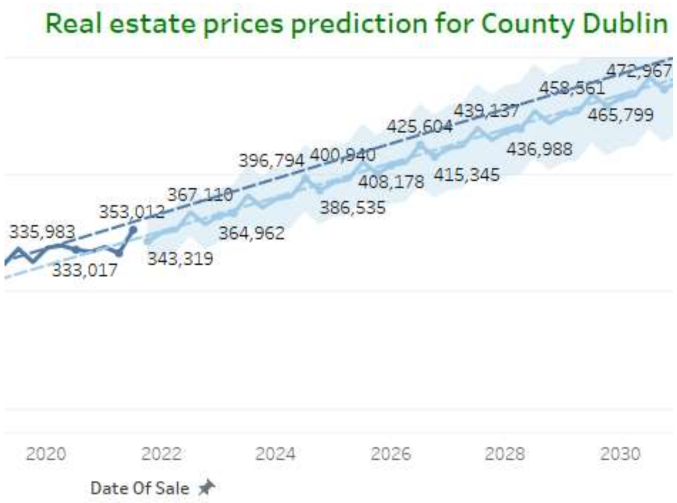


Fig.13. Prediction of housing prices until 2030 in County Dublin, **Tableau Software**

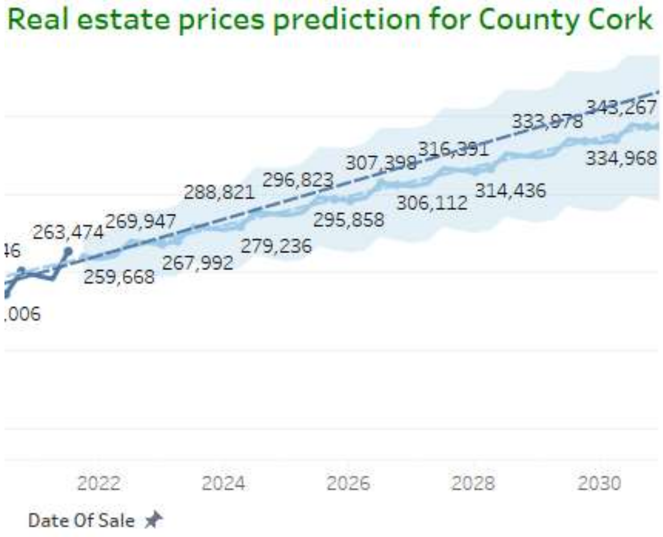


Fig.16. Prediction of housing prices until 2030 in County Cork, **Tableau Software**

Figures 13 and 14 likely present visualizations of housing price predictions for the real estate sectors in County Cork and County Dublin, respectively, spanning the period from 2021 to 2030. These visualizations were created using forecasting algorithms in Tableau Software and provide insights into the estimated price growth over the next decade.

V. CONCLUSION

The research findings reveal that housing prices have increased at a similar rate for both County Cork and County Dublin from 2012 to 2021, with approximately a 55% increase for County Cork and a 59% increase for County Dublin. This indicates a consistent upward trend in real estate prices in both regions over the past decade. However, TABLE IV shows that the projected average prices from 2021 to 2030 will experience a slower rate of increase, with around a 33% increase for County Cork and a 34% increase for County Dublin. This suggests a moderation in the pace of price growth compared to the previous decade. Despite the slowdown in the rate of increase, the findings confirm that housing prices will continue to rise in both counties. This indicates sustained demand in the housing market, driven by factors such as population growth, economic expansion, and limited housing supply.

It's noteworthy that despite government campaigns and solutions aimed at addressing housing affordability and supply issues, the research findings suggest that housing prices will continue to increase. This underscores the complexity of the housing market and the challenges associated with achieving long-term affordability and stability.

TABLE IV
INCREASE RATES FOR THE FORECASTED REAL ESTATE PRICES

	Real estate prices forecasting		
	2021	*2030*	*Increase rate*
County Cork	*€258,168*	*€343,267*	*33%*
County Dublin	*€352,026*	*€472,967*	*34%*

The prediction of housing prices based on a single numerical variable raises concerns about the accuracy of the forecasting process. As the predicted timeframe increases, the estimation brackets widen, leading to potential inaccuracies in the predicted values. Additionally, the real estate sector is influenced by various factors beyond numerical variables, which should be thoroughly examined to improve the accuracy of predictions. Furthermore, it is recommended to use a sufficient number of samples for machine learning processes, typically more than 8 samples, to ensure robust model training. Despite the sample size limitation, merging new datasets into the existing dataset was undertaken to enhance the performance of the machine learning model.

It's noted that the predicted real estate prices for Counties Cork and Dublin obtained from training the Linear Regression Model differ from the forecasted values generated using Tableau Software. Jack Dwyer [11] suggests that Tableau forecasts may have low accuracy, and obtaining predictions through Python programming is recommended for more accurate output.

Machine learning techniques and running suitable algorithms on data to forecast outcomes are also advised. Future research should focus on advanced techniques such as Artificial Intelligence, particularly Machine Learning, to enhance the accuracy of predictions and protect sectors from potential crises. Additionally, investors would benefit from real estate predictions conducted by independent researchers, as they may offer unbiased insights compared to those provided by financial institutions.

VI. REFERENCES

[1] KBC, “Irish property prices surge while broader inflation pressures build.”, 2022. [Online]. Available: https://www.kbc.ie/w/irish-property-prices-surge-while-broader-inflation-pressures-build [Accessed on: 15th June 2022].

[2] J.J.Handopo and R.A.Rahadi, “Analysis of Generation Z’s Housing Affordability and Preferences Based on Price to Income Ratio in Jakarta Region”, 2021. [Online]. Available: https://www.researchgate.net/publication/357865837_Analysis_of_Generation_Z%27s_Housing_Affordability_and_Preferences_Based_on_Price_to_Income_Ratio_in_Jakarta_Region [Accessed on: August 6, 2022].

[3] Trading Economics, “Ireland Inflation Rate”, 2022. [Online]. Available: https://tradingeconomics.com/ireland/inflation-cpi#:~:text=The%20annual%20inflation%20in%20Ireland,rise%20in%20the%20previous%20month [Accessed on: June 10, 2022].

[4] Residential Property Price Register, “Residential Property Price Register”, 2022. [Online]. Available: https://www.propertypriceregister.ie/ [Accessed on: June 1, 2022].

[5] gov.ie., “Housing for All-a New Housing Plan for Ireland.”, 2022. [Online]. Available: https://www.gov.ie/en/publication/ef5ec-housing-for-all-a-new-housing-plan-for-ireland/ [Accessed on: June 11, 2022].

[6] Social Justice Ireland, “Housing for all strategy won’t solve housing crisis.”, 2021. [Online]. Available: https://www.socialjustice.ie/article/housing-all-strategy-wont-solve-housing-crisis [Accessed on: July 8, 2022].

[7] B. T. Jijo and A. M. Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning” , 2021. [Online]. Available: https://www.researchgate.net/publication/350386944_Classification_Based_on_Decision_Tree_Algorithm_for_Machine_Learning [Accessed on: August 5, 2022].

[8] A. Grybauskas, V. Pilinienе, and A. Stundziene, “Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic”, *Journal of Big Data*, vol.8, pp.105. 2021. [Online]. Available: https://doi.org/10.1186/s40537-021-00476-0 [Accessed on: July 20, 2022].

[9] M. Thamarai and S. P. Malarvizhi, " House Price Prediction Modeling Using Machine Learning", *International Journal of Information Engineering and Electronic Business*, vol.12, no.2, pp. 15-20, 2020. Accessed: 1.08.2022 [Online]. Available: https://www.mecs-press.org/ijieeb/ijieeb-v12-n2/IJIEEB-V12-N2-3.pdf

[10] N. Vineeth, M. Ayyappa, and B. Bharathi, “House Price Prediction Using Machine Learning Algorithms.”, *Communications in Computer and Information Science*, vol.837, pp. ,2018. Accessed: 1.08.2022 [Online]. Available: https://doi.org/10.1007/978-981-13-1936-5_45

[11] D. Maulud and A. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning.”, *Journal of Applied Science and Technology Trends*, vol.01, pp.140 – 147.

[12] Askpython, (2022). Detection and Removal of Outliers in Python – An Easy to Understand Guide. Available at: https://www.askpython.com/python/examples/detection-removal-outliers-in-python [Accessed on: 20th June 2022].

[13] Jack Dwyer (2018) “Tableau Python Forecasting: Increase Your Accuracy!”, 2018. [Online]. Available: https://www.blastanalytics.com/blog/tableau-python-forecasting-improve-your-accuracy [Accessed on: August 6, 2022].