Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

MONETARY & FINANCIAL ECONOMICS

MASTER´S FINAL WORK

DISSERTATION

SANAM SAMADANI

OCTOBER - 2021

MASTER

MONETARY & FINANCIAL ECONOMICS

MASTER´S FINAL WORK

DISSERTATION

FORECASTING REAL ESTATE PRICES IN PORTUGAL

BASED ON A DATA SCIENCE APPROACH

SANAM SAMADANI

SUPERVISION: Prof. CARLOS J. COSTA

OCTOBER - 2021

To my beloved uncle who is
no longer of this world, and
to all people whose lives got
affected by COVID

GLOSSARY

AIC- Akaike's Information Criterion

ANN- Artificial Neural Network

ARIMA- Auto-Regressive Integrated Moving Average

BIC- Bayesian Information Criterion

GDP- Gross Domestic Product.

HQIC- Hannan-Quinn Information Criterion

OLS- Ordinary Least Squares.

SARIMA- Seasonal Auto-Regressive Integrated Moving Average

ABSTRACT, KEYWORDS AND JEL CODES

The evolution of the European residential market is notorious over the last ten years. House prices in the E.U. rose by 30.9 per cent between 2010 and the first quarter of 2021. The prices of homes in Portugal has risen almost 50 per cent during 11 years. Considering this previous argument, I propose the following research question: How to predict real estate prices. In this context, my research aims to analyze the prices' evolution and understand the main components impacting the price of real estate. First, using the time series analysis, I use ARIMA to analyze the prices of real estate and the number of buildings sold since the first quarter of 2009, almost one year after the great recession in Portugal, until the third quarter of 2020 which was during the COVID-19 pandemic. The model was fitted and the prediction line was accurized with an upward trend. The second approach consists of analyzing the impact of five independent variables on real estate prices. To understand the most relevant components, regression analysis has been performed. I used OLS to analyze the impact of independent variables (crime rate, selected waste rate, tax rate, purchasing power and tourism rate) on real estate prices. Crime rate and tourism are negatively correlated while purchasing power, selected waste rate and tax rate are positively correlated with real estate prices. Then, I compared the accuracy of the result with neural networks and other types of regression analysis. Results were not much better than with linear regression. It is also essential to consider that this approach has some limitations, especially regarding the analysis's granularity. The data has been collected from INE and PORDATA, the databases of contemporary Portugal, to construct the models and forecast house prices in Portugal.

KEYWORDS: real estate; prices; prediction; ARIMA; Regression; ANN; OLS

JEL CODES: C1; C45; C53; E24; R3; R31.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENT

First and foremost, this project is especially dedicated to Professor Carlos. J. Costa assisted me throughout this path to complete my work. I am also grateful to my colleagues and friends for numerous discussions and unlimited help.

Also, I would like to dedicate this project to my dear father, Amir Hossein, who has been a wonderful supporter until my research was complete, and to my beloved mother, Anahita, who has been encouraging me for months.

Finally, I would like to be thankful for my dear sister, Asa, who was emotionally beside me, and my lovely maternal grandparents, Pedram & Manouchehr, whose warm words gave me courage, hope, and energy to stay strong & focused.

1. INTRODUCTION

The real estate industry is critical to the Portuguese economy. Housing is a household's most valuable asset. In 2017, Portuguese real estate accounted for 48% of total family wealth. (Pinheiro,2019).

As indicated by Pinheiro, 2019, on the demand side of the real estate sector, one factor that has prompted a lift has been normal among most nations in the euro region. The environment of low-loan costs with which the ECB has settled and invigorated the European economy lately. The lower yield of monetary resources that led to this environment increased the demand for housing for investment purposes. Likewise, the strong growth in tourism activity has intensified demand for real estate among investors. Finally, the dynamics witnessed in the real estate market in recent years have also been influenced by demand for real estate among non-inhabitants, particularly, under the Golden Visa conspire took on in 2012, which stimulated the demand for housing by granting tax benefits non-residents who purchase a property. A balance in demand is expected amid a slowdown in traveler movement.

Residential real estate provides refugee and wealth preservation. Central banks, financial supervisory bodies, investors, and homeowners all benefit from having access to accurate house price forecasts. For a variety of reasons, forecasting national house price indexes is difficult. The major difficulty is the limited duration of time-series data available. House price indices are generally generated monthly or quarterly, limiting the length of computed indices and complicating model development and testing. Residential real estate provides safety and wealth preservation. Central banks, financial supervisory bodies, investors, and homeowners all benefit from having access to accurate house price forecasts. For a variety of reasons, forecasting national house price indexes is difficult. The major difficulty is the limited duration of time-series data available. (Milunovich, 2020)

COVID-19, an ongoing pandemic of coronavirus disease 2019, has been affecting Portugal since March 10th of 2020. The pandemic has caused severe global economic disruption, including the largest global recession since the U.S. Great Depression (International Monetary Fund 2020). 'Based on Idealista, 2020 and Idealista, 2021 recent records, due to COVID-19, it is vital to analyze the fluctuations of the house prices in the following years. Many foreign buyers hesitate to invest in real estate business due to the result of Brexit and the COVID-19 pandemic. On the other hand, Portugal's housing market hardly fluctuated, considering other prices rising in 2020. The upward trend of house prices in Portugal during last 10 years is something that continued in 2020 despite the COVID-19 pandemic, which reduced the number of operations in the market. Property prices in Portugal soared by an average of 5.9 per cent in 2020, and in the fourth quarter of the year. This gain in prices was 2.7 per cent. The increase in house

prices and rents are much higher in Portugal than the average for E.U. and Eurozone countries. In the third quarter of 2020, compared with the same period in 2019, house prices increased by 4.9 per cent in the Euro area and 5.2 per cent in the E.U., with Portugal rising above the average with property price increases of 7.1 per cent.' (idealista.pt, 2021).

COVID-19 has created a lot of uncertainty in the real estate market. Individually, social distancing precautions have limited house views, which are essential in the selling process, forcing buyers and sellers to reassess their plans. Sellers are increasingly seeking certainty. To reduce the danger of virus spread, several brokers provide property visits through Skype and FaceTime. Brokers are also asking potential buyers to pre-register for viewings around. It allows the brokers evaluate client's degree of interest and likelihood of purchasing. (Gujral et. al, 2021). Additionally, thousands of workers worldwide have been made redundant or placed on a temporary, unpaid leave of absence. Inevitably, this significantly impacts individuals' abilities to pay rent, mortgages, and various household expenditures (Pickford, 2021).

In the United Kingdom, the government has advised buyers and lenders to postpone talks during the shutdown, thus halting deals. Many people are concerned that we may experience another financial crisis, such as 2008, which would influence real estate confidence. Banks in the United Kingdom have begun to take measures; high-street lenders now want up to 40% deposits before approving a new mortgage. However, it is too early to predict COVID-19's impact; the virus has yet to influence the property market. (Read, 2020).

Households' indebtedness measured as housing loans in terms of disposable income increased from 25 per cent in the mid-1990s to almost 90 per cent by the end of 2007, in a context of rising disposable income and low-interest rates. Nevertheless, over this period, house prices barely changed. After the sub-prime crisis, housing loans have been contracting since 2011, reflecting banks' deleveraging. Interest rates exhibited much volatility in the first two years of the financial crisis (Lourenço & Rodrigues, 2017)

According to Eurostat, for five consecutive years, the value of houses on the Portuguese market has annual changes of more than 6 percent, a value from which the European Commission considers that a market is at risk of a price bubble. In the national territory, the peak was recorded between 2018 and 2019 with annual increases of 6-8 per cent in this index (house price index) - and even in a year marked by the pandemic- as was 2020- the house prices index increased by 7.4 per cent, the fourth largest increase recorded in Portugal in the last ten years.

The evolution of the European residential market is notorious over the last ten years. House prices in the E.U. rose by 30.9 per cent between 2010 and the first quarter of 2021. The prices of homes in Portugal has risen almost 50 per cent during 11 years. (Sousa, 2021).

Considering this previous argument, I propose the following research question: How to predict real estate prices? In this context, this research has the main purpose to analyze the prices' evolution and understand the main components impacting the price of real estate. First, using the

time series analysis, I use ARIMA to analyze the prices of real estate and the number of buildings sold since the second quarter of 2009, almost one year after the great recession in Portugal, until the third quarter of 2020 which was during the COVID-19 pandemic. We use ARIMA because of its use as one of the methods frequently used in social science to estimate housing sales (Temür et al., 2019). One of the most relevant and broadly used time series models is the autoregressive integrated moving average (ARIMA) model. The popularity of the ARIMA model is due to its statistical properties and the well-known Box-Jenkins methodology in the model building process (Wilson, 2016).

The evolution of the European residential market is notorious over the last ten years. House prices in the E.U. rose by 30.9 per cent between 2010 and the first quarter of 2021. The prices of homes in Portugal has risen almost 50 per cent during 11 years (Sousa, 2021).

The second approach used consists of analyzing the impact of five independent variables on real estate prices have been analyzed. In other studies, similar to what Mach, 2017 has presented in a paper, he features of the buildings have been studied while the main independent variables of this analysis are as follows: purchasing power, tax rate, waste rate, tourism and crime rate. Third, by using various types of regressions, such as OLS, Bayesian regressions, polynomial regression, ridge regression, lasso regression as well as applying Neural Network in the data, the impact of aforementioned independent variables on real estate prices will be studied to compare the models and find the best fit. Finally, in the last part, the conclusion and will be presented.

Using the previous approach, I developed a causal model and an ARIMA model. The causal model showed the importance of purchase power as an element for predicting prices. Time series allowed to confirm the rise of prices, even in the pandemic period.

This research was presented at the 16th Iberian Conference on Information Systems and Technologies and published in IEEE (Samadani & Costa, 2021)

The dissertation has the following structure: introduction, literature review, methodology data and results and conclusions. The introduction is this chapter, then, the literature review presents the main background concepts and related works. The methodology, data and results' chapter present the methodological approach followed, data collected and main results. The last chapter presents the main conclusions of the research.

2. LITERATURE REVIEW

In Portugal, the rate of homeownership has risen from 57 per cent in 1981 to 73 per cent in 2011. The number of second homes also increased significantly, from 7 percent in 1981 to 20 per cent in 2011 of total family dwellings' (INE,2012). Financial markets also directly impacted the housing sector, contributing to the predominance of policies focusing on demand that have stimulated homeownership using credit. The expansion of the European integration contributed by successive lowering of interest rates. The result was an oversupply of housing already noticeable at the beginning of the 2000s. The progressive withdrawal of public support for housing loans and the decrease in public investment in major public works led to a prolonged crisis in the construction sector, which further dragged the Portuguese economy into stagnation (idealista.pt,2021).

According to Temur and colleagues, 2019, the surplus in the housing supply may result in unintended price decreases, which results in firms operating in this sector facing various problems, such as not being able to sell what they produce or having to sell at a cheaper price. The failure of the housing supply to meet demand affects the welfare of people who desire to buy a home as a safe haven for their money.

Lourenço et al. (2017) examined the relationship between major economic fundamentals and house price changes in Portugal during and after the financial crisis. They consider (1966Q1 to 2017Q2) to test the relationship between a set of selected independent variables and real house price growth, enabling them to identify the directions and extent of the relationship. The results suggest that most indicators, including interest rate and GDP growth, behaved analogously during and after the financial crisis. Nevertheless, since the significance and magnitude of parameter estimates may change when the market is in crisis, they also consider a framework that allows for breaks. House prices in Portugal have increased lately but are still below pre-crisis levels in real terms. Allowing for breaks (i.e. different regimes) makes it possible to have a fresher look at fundamentals. During the 2007 to 2011, the fact that residential gross fixed capital formation dropped may have prevented house prices from declining even more during that period attenuating the contraction of housing demand.

Moreover, foreigners' growth in housing investment may have prevented house prices from falling further in 2011-2017. Low (or even negative) interest rates may affect house prices through alternative saving options. Following the years of the sovereign debt crisis, non-residents' housing investment slowed. Finally, the results of the Probit model show that the likelihood of future positive home price increase in Portugal is still strong.

Zillow research conducted data on the economic effect of the global pandemic. Typically, the economic activities fall for 6-18 months and then recover more slowly. During SARS Hong Kong house prices did not contract significantly. However, transaction volumes fell by 33-72 per cent as customers avoided human contact (like avoiding travel restaurants and public gatherings). After the epidemic was over, transactions quickly returned to a previous condition considering the average volumes. During standard recessions, home prices and transaction volumes may go down, but this is not always the case (e.g. the 2001 recession). It is hard to precisely forecast the probability of an epidemic-related downturn and how such a downturn could provoke a standard recession because this depends on how COVID-19 progresses and how this progress interacts with pre-existing recession risks and policy responses (ranging from doing nothing to shutting down entire cities for months at a time). Because consumers wish to avoid nonessential human contact, the 2003 SARS pandemic led to a temporary fall in monthly real estate transactions from 33 per cent to 72 per cent vs baseline for the epidemic duration, while real estate prices held steady.

Meanwhile, during the current episode in China, news reports and early data provided by Goldman Sachs (2020) indicate a near shutdown in the volume of Chinese real estate transactions. Wong (2008) concludes based on transaction data covering 44 housing estates that the onset of SARS coincides with a 1.6 per cent decrease in house prices versus a pre-SARS trend. In addition, she finds that the onset of the SARS epidemic coincides with a 72 per cent reduction in transaction volumes for these estates. She explains this pattern (small price reductions coincided with a considerable reduction in volume) as customers adopt a *wait and see* approach. This pattern happened for Chinese house prices during January and February of 2020 as it remained stable from December to January (+0.27 per cent). However, the volume of transactions has fallen by 90 per cent to 98 per cent from regular (Gudell, 2021).

Del Giudice et al., 2020. studied the effect of COVID-19 in the Italian real estate markets. According to them, the housing prices will reduce 4.16 per cent in the short-run and 6.49 per cent in the mid-run, “predatory” housing may occur in the short-run, leading to change in the national and local economic geography (idealista.pt, 2021). The greatest danger for the national and local economy is the income impoverishment that will arise as an effect induced by forced inaction. There are two factors to be reckoned. On the one hand, the impoverishment will result in an effect induced by involuntary inactivity for many productive and commercial sectors. On the other hand, a new future propensity of families who will prioritize saving to protect themselves from other future difficulties. In the short term, the effect of this situation reflects quickly on housing sales and real estate price, which for the residential market, could drop between 1.3 per cent and 4 per cent in the two-year period 2020-2021, and then slightly ticked up in 2022,

according to what the Nomisma Institute affirms (*Società di consulenza Strategica e Aziendale*, 2021).

Fragoso (2017) and Bełej & Cellmer (2014) showed macroeconomic variables' effect on real estate prices. Grum, & Govekar,. (2016) suggest that there are cultural differences that contribute to different effects of macroeconomic variables. Berlemann, & Freese (2013) suggested the impact of commercial property prices as stock markets do not react on interest rate variations, questioning the impact of monetary policy on house prices. GDP (gross domestic product), the NAHI (net average household income), and imports and exports play an important role in the market, and interest rates negatively impact the increase of bank valuation (Fragoso 2017). By applying artificial neural networks and linear multiple regression (LMR), ANN has shown better results for larger datasets, while LMR has shown fair results when applied to smaller datasets. Also, the study demonstrated that Regression models are well fitted for an explanatory analysis while performing poorly when used as predictive models. Temür et al. (2019) chose ARIMA as a linear model because it is one of the most used methods in social science to estimate housing sales in Turkey. The financial crisis that began in August 2007 with the mortgage crisis caused by the housing and real estate markets in the United States evolved into a global economic crisis beginning in September 2008. After the real estate bubble burst, housing market analysis has become even more critical. It precipitated the rest of the world's economic crisis, particularly in the United States. As a result, accurate estimation of real estate sales is critical in order to achieve a balance of supply and demand. Furthermore, as primary housing markets have become more integrated with secondary markets, the computation of housing prices has become increasingly important to investors who must choose between portfolios composed of other investment assets and portfolios composed of housing securities. (Temür et al., 2019).

The effects of property taxes on local land, housing, and labour markets in Germany in research found a negative effect of property tax and housing investment. In rural areas, real estate taxes decline building permit and capital investment. Also, the rents and house prices decrease if the tax increases in the short-run (Löffler & Siegloch, 2017).

A significant influence on home prices is purchasing power, as home prices are likely to follow when buyer purchasing power falls. (*The Source of Home Price Movement*, 2021). Another strand of literature specializes that house prices depend positively on income per capita, wealth, and population growth and negatively on the mortgage interest rate. On the other hand, studies suggest that the house price and income relation may be negative at times. (Özmen et al., 2019, Case & Shiller, 1988)

Pour et al. (2013) show that the earnings-house price relationship was negative in Iran amid the huge oversupply of property and a large volume of construction activity during the periods of real economic growth. Frischtak et al. (2012) investigated one example of this mechanism as it pertains to the connection between crime and house by using a recent policy experiment in Rio de Janeiro. Where government installed permanent police stations in low-income communities (or favelas). Their empirical work show that decreasing crime does benefit lower-valued properties disproportionately, reducing the disparity among properties. The inclusion of previous crime rates as a driver of present property prices is the mechanism in the model that causes decreasing returns. This work adds to a number of ongoing research topics, ranging from studies of conflict economics to wealth distribution analyses. The fact that these estimates considered the crime rate as exogenous may have skewed the elasticity estimates if, for example, crime occurs disproportionately in impoverished neighborhoods with low property values or if criminals seek regions with higher-priced properties. The recent studies that do instrument crime (Gibbons, 2004 and Tita et al., 2006) again find a significant negative relationship, an effect that is particularly pronounced for violent crimes.

Many studies emphasize that income often plays a key role in forecasting house price growth rates (Case & Shiller, 1990). While homeownership is often viewed to enable households to build wealth, threats to the value of that investment may limit its appeal. One such threat is a crime, which may reduce the desirability of ownership in affected neighborhoods (Tita et al., 2006).

The effect of crime on property prices in an urban county in the United States, Pierce County Washington, has been evaluated in a study (Angelov. 2020). They developed their model and tested it on their data (random forests, decision trees, and artificial neural networks). Buying a house can be one of the most important decisions in many people's lives. Estimating a fair market value is critical for both clients and retailers. Property buyers should be aware of the crime rate in the neighborhood in which the property is located. The crime rates are not always displayed by real estate companies or websites. Most customers consider physical factors such as size, number of bedrooms, bathrooms, and so on. However, one of the most important factors to consider when purchasing a property is the crime rate. Also, if sellers or neighbors want to get a better price for their properties, they should spread the word about crime reduction. Local police and other government agencies play an important role. Is that when purchasing a home, buyers should be aware of the crime rates. Also, if sellers or neighbors want to get a better price for their properties, they should spread the word about crime reduction. Local police and other government agencies play an important role. (Angelov, 2020)

Olijade and Lizamn (2016) discussed how property values in residential areas were affected by crimes. They also talked about how different types of crimes (burglary, street crimes, and vandalism) influenced housing prices. To anticipate the influence of various types of crimes on residential property values, the authors employed logistic regression analysis. The data for the model was taken from several surveys conducted in Southwestern Nigerian areas. Street crime, vandalism, robbery, and violent crime were split into numerous independent variables. All of the predictors had a direct impact on property prices; however, violent crimes were the most important factor negatively impacting residential property values. The authors argued that violent crimes have the highest impact on property values because such crimes produce fear in residential communities. They concluded that the findings of their logit model were consistent with other published papers; that is, the model supported the hypothesis that neighborhood crimes negatively influence residential property values. This negative relationship can decline property investments, which in turn deteriorates the neighborhood and declines property tax that can be used to fight residential crimes.

Ofori (2021) using regressions and ANOVA, assessed the extent to which waste disposal sites impact residential rental values in Ghana. In low-income countries such as Ghana, waste dumping is one of the most severe environmental problems. He collected data from both secondary and primary sources using a mixed-method cross-section linear hedonic price model was adopted to support the research objectives. The research showed that all the residential properties that were far from the waste disposal sites had higher rental values than those that were close, as almost all the p-values were less than alpha, thus ($0.000 < 0.05$)

Also, Iman & Gan (2013) surveyed, the locational effect of noise and water pollution on adjacent property values and their analysis revealed that noise and water contamination were seen as a discomfort and therefore all properties located close to noise pollutants were sold at lower prices. Rocha et al. (2008) analyzed the noise influence on real estate values because the Portuguese Noise Code imposes building limitations on municipal urban areas exceeding established noise limits. In these locations, and until mitigating methods enable noise reduction, private contractors refuse their building permits for excessive noise reduction. In 2003, the Portuguese government issued a law concerning Real Estate Taxation. As municipalities have to communicate the inadequacy of vacant land for building construction, the following steps is the reduction of nominal real estate value and the taxes income.

According to Mc Donald (1993), using data from the 1980s for the six counties in metropolitan Chicago, an econometric model of changes in the property tax rate was presented. The elasticity of demand for commercial or industrial real estate in a county that is a part of a

larger metropolitan area is likely to be significant. Therefore, the long-run effects on the commercial and industrial tax base of increasing the property tax rate are likely to be significant and negative.

3. METHODOLOGY, DATA & RESULTS

In what concerns the methodology, I followed a data science approach (Costa & Aparicio, 2020, 2021, Aparicio, et al. 2019).). I started to understand the building market, especially using the literature review. Then I collected Data mainly from INE and PORDATA.

First, I have applied this model to study the patterns of real estate prices in Portugal from the first quarter of 2009 until the third quarter of 2020. The model helps to determine the direction of changes in the real-estate prices. To understand the evolution and predict prices, I perform a time series analysis, including ARIMA (Tomas et al., 2018.). SARIMA (Seasonal Autoregressive Integrated Moving Average) or Seasonal ARIMA is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. Then, I would explain ARIMA and apply it to the data obtained from INE and get the analysis results.

Next, I used regression analyzed. An explanatory model will be created and estimated. OLS analysis and artificial neural network will be applied to the data obtained from PORDATA to observe the results of analysis. Then, the accuracy of training and test subsets of other types of regression analysis and artificial neural networks will be compared to see which one fits better.

3.1. TIME SERIES

It adds three new hyperparameters to specify the autoregression (A.R.), differencing (I), and moving average (M.A.) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality. Four seasonal elements are not part of ARIMA that must be configured: P is seasonal autoregressive order, D is seasonal difference order, Q is seasonal moving average order, and m is the number of time steps for a single seasonal period.

“Similar to other studies it is obvious that forecasting the real estate market is neither an easy task to accomplish nor naive, mechanistic approaches. Nevertheless, any forecasting approach that consistently provides better odds than those from tossing a coin in making the correct investment decision should merit careful examination. In fact, technical analysis has been developed and applied to finance for many years. In time series, it is important to identify the data series in the following processes: (a) is the data random? (b) does the data have a trend? (c) model identification and (d) testing for model adequacy.

If a series is random, the correlation between successive values in a time series is close to zero. However, if the observations of the time series are statistically dependent on or related to one another, then the Box-Jenkins (ARIMA) methodology is appropriate.

By looking at autocorrelation coefficients for time lags of more than one period, one can determine additional information on how values of a given time series are related. This method produces forecasts that are likely to be more accurate than the forecasts produced by other approaches. The ARIMA models have also proved to be excellent short-term forecasting models for a wide variety of time series because short-term factors are expected to change slowly. The simpler autoregressive and moving average models are actually special cases of the ARIMA classes. Moving averages are popular for determining turning points, which is when a market trend changes direction. The basic concept behind the application of moving averages is that, when a price series crosses the correct moving average of itself, the price series will continue in the direction of the crossing. Moving averages are also useful for filtering the effects of cycles of the known periods in data. The simple models can contain either autoregressive or moving average components but not both. A mixed autoregressive and moving average model with both components is known as the ARIMA model" (Tse, 1997). Similar to studies of McKenzie,1984 and Temur et al, mentioned in literature review, ARIMA has been applied for time series analysis on this set of data.

3.1.1. DATA AND TIME SERIES (ARIMA)

Here, I have applied this model to study the patterns of real estate prices in Portugal from the first quarter of 2009 until the third quarter of 2020. The model helps to determine the direction of changes in the real estate prices.

I started by analyzing the evolution of volume, quantity and prices. This analysis was performed considering new buildings' values and total buildings' values.

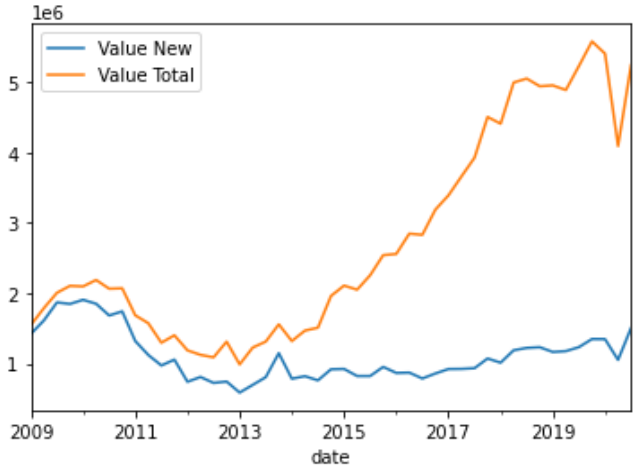

Figure 1: Evolution of total and new real estate value, source: Data from INE.pt

The evolution of the quality of building sold is similar to the evolution of the total value. The figure shows an increase in the values since the first quarter of 2009. After 2010, there was a significant decline in the value amid recession in Portugal. Prices started to grow significantly as the recession period ended. As COVID-19 lockdown measures were implemented in March 2020, the value of buildings fell, while it grew again in the second quarter of 2020.

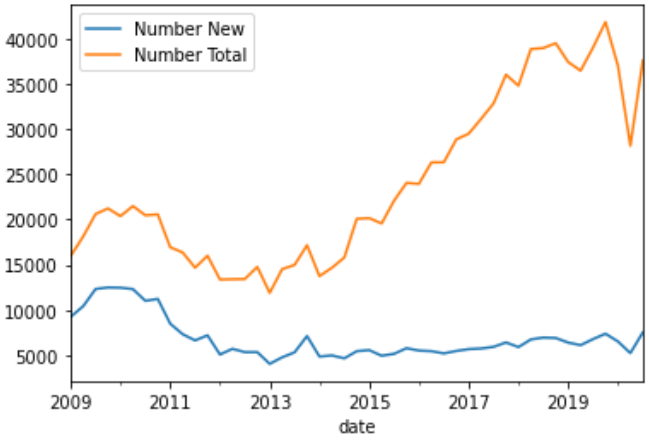

Figure 2: Evolution of the number of buildings sold source: Data from INE.pt

[1]

Figure 2 shows that the total number of houses sold during this period has almost the same upward trend as figure 1.

[1] All graphs have been created by python

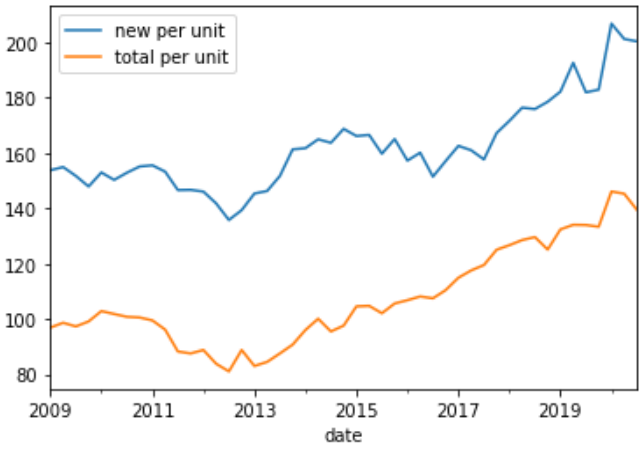

Figure 3: Evolution of prices of buildings sold per unit, source: Data from INE.pt

Figure 3 also shows an upward trend of the total value of buildings sold per unit at a softer pace compared to Figures 1 and 2.

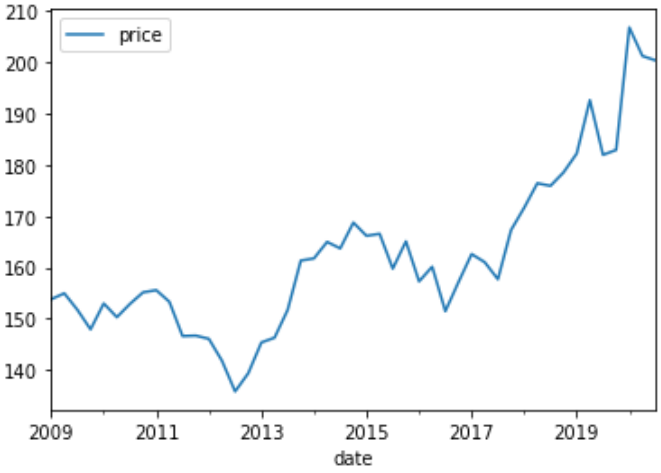

Figure 4: Evolution of real estate prices, source: Data from INE.pt

Figure 4 is representative of the evolution of houses in Portugal with a similar upward trend, ranging from an average of EUR 618 thousand until more than EUR 3.1 million.

3.1.2. RESULTS AND TIME SERIES

The ARIMA model was created and fitted. The dependent variable is the price of real estate with 47 observations, indicating 47 quarters since the first quarter of 2009. The best result was for AR (4), with one first difference. A low P-value, less than 0.05, means we can reject the null hypothesis. Here only AR-L4 seems to be less than 0.05 and statistically significant. Also, the independent variable the constant β the error term is sigma2, or ξ is 0.000 and statistically significant. It calculated the t-statistic: t = estimated coeff. /std.error of coeff to 1.96 ACF of the

residuals, if it is a good model, all autocorrelations for the residual series should be nonsignificant. Box-Pierce (Ljung) tests for possible residual autocorrelation at various lags. The Ljung Box test is pronounced “Young” and is sometimes called the modified Box-Pierce test that the errors are white noise. The Ljung-Box (L1) (Q) is the LBQ test statistic at lag 1 is, and the Prob (Q) is 0.02, and P-value is 0.89. Since the probability is above 0.05 we cannot reject the null that the errors are white noise.

Heteroscedasticity tests that the error residuals are homoscedastic or have the same variance. The summary statistics show a statistic of 4.23 and a p-value of 0.01, which rejects the null hypothesis and the residuals show variance.

Jarqua-Bera tests for the normality of errors tests the null that the data is normally distributed against an alternative of another distribution. There is a test statistic of 8.08 with a probability 0.02 which means we reject the null hypothesis, and the Data is not normally distributed.

The Log-Likelihood, AIC, BIC, and HQIC help compare one model with another.

The log-likelihood function identifies a distribution that fits best with the sampled data. While it is useful, AIC and BIC punish the model for complexity, which helps make the ARIMA model parsimonious. Additionally, as part of the Jarque-Bera test, we see the distribution has a smaller skew and a larger Kurtosis.

Considering the prices, in the following analysis, the main results of an ARIMA analysis has been provided:

Table 1:SARIMAX results

Dep. Variable	Price	No. Observations	47
Model	ARIMA	Log-Likelihood	-143.568
AIC	297.136	HQIC	300.561
BIC	306.279		

Table 2: Covariance results

	Coef	Std err	z	P> [z]	[0.025	0.9759]
ar.L1	-0.2008	0.177	-1.135	0.256	-0.548	0.146
ar.L2	-0.0573	0.158	-0.362	0.718	-0.368	0.253
ar.L3	0.2427	0.197	1.232	0.218	-0.143	0.629
ar.L4	0.4115	0.204	2.016	0.044	0.011	0.812
Sigma2	29.5050	7.159	4.121	0.000	15.473	43.537

Table 3: Additional statistics SARIMAX

Ljung-Box (L1) (Q)	0.02	Jarque-Bera (JB)	8.08
Prob (Q)	0.89	Prob (JB)	0.02
Heteroskedasticity (H)	4.23	Skew	0.72
Prob (H)	0.01	Kurtosis	4.47

This ARIMA model was used to forecast the housing prices. One way to measure the accuracy of the forecast is using the RMSE. For this specific model, the RMSE number is 8.463. In the chart, the red line represents the forecast, and the blue is the price series.

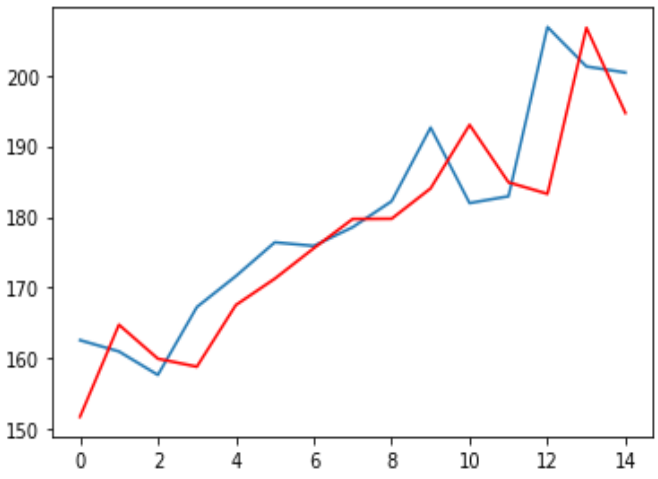

Figure 5: Prediction line, source : Data from INE.pt

3.2. REGRESSION ANALISIS

Linear regression is a strong tool for investigating the relationships between multiple variables by relating one variable to a set of variables.

Ordinary least squares (OLS) is a method for estimating a linear regression model that minimizes the discrepancies between the observed sample values and the model's fitted values. The regression coefficients that minimize the sum of squared differences between the outcome variable and the linear combination of explanatory factors are used to get the OLS estimations. However, causal relationships can be one of the drawbacks of using this tool. Also, it is not always a straightforward rule to choose a correct specification, and some complexities might exist. Finally, a regression model needs to be tested statistically, especially if estimates of interest appeared very sensitive to the specification used or to the set of explanatory variables included. (Verbeek, 2017)

Artificial Neural networks are software or hardware-implemented systems built on the principle of organization and functioning of their biological analogue- the human nervous system. The artificial neural networks imitate the properties of a biological neuron. The input receives a certain number of signals, each of them is the output of another neuron. The neural network's power comes from two sources: first, the diversity of information processing, and second, the capacity to self-learn, or the ability to generalize. The capacity to acquire an educated answer based on material that was not encountered during the learning process is referred to as generalization. These characteristics enable neural networks to handle difficult issues. Some advantages of neural networks are: (Safronov, 2017)

1. Solving problems with unknown regularities,
2. Resistance to noise in the input data, 3. Adaptation to environmental changes,
4. Potential ultra-high performance.

On the other hand, neural networks ignore many properties of their biological counterparts.

For example, it ignores the time delay, which has an impact on the system's dynamics. It does not consider the impacts of the frequency modulation function or the synchronization function of the biological neuron, and it generates an output signal instantly. (Safronov, 2017)

In this part, the data set comprises yearly independent variables in all municipalities in Portugal. Data on real estate prices, waste, crime, purchasing power, and tax rate were collected from PORDATA and 327 observations.

As an approach to our analysis, I consider a standard multiple linear regression framework to examine the relationship between prices of real estate and a set of covariates.

$$Price_t = \alpha_0 + \alpha_1 PurchasingP_{t-1} + \alpha_2 Crime_{t-1} + \alpha_3 Tax_{t-1} + \alpha_4 Tourism_{t-1} + \alpha 5 Wastet - 1 + et$$

Where $Price_t$ corresponds to the average of building sold prices in each municipality in the year 2018, $Crime_t$ is the crime rate per 1000 residents, tax_t is the total tax rate in each municipality, $Tourism_t$ is the transaction rate of tourists per 100 residents, and $Waste_t$ is the selected waste rate.

Table (1) describes the independent variables in the analysis. Table (II), table (III), and table (IV) are the results of ordinary least square (OLS) estimates of and robust standard errors.

Table 4: Features description

FEATURE NAME	DESCRIPTION
PurchasingPower	Purchase power per capita of the population from the municipality
CrimeRate	Crime rate per capita of the population from the municipality
TaxRate	The Portugal property transfer tax (transfer of ownership of real estate)
Tourism	Transaction rate of tourists per 100 residents
WasteRate	Percentage of selected waste

The model has as observations average data from municipalities. So, the granularity does not give enough information to analyze this reality with precision.

I created a regression model where the impact on prices of 2018 houses are explained by purchasing power, crime, taxes, and tourism rate and waste rate. We used the OLS (Ordinary least square) and fitted the model. The main statistics may be seen in the following table. The main statistics may be seen in the following table. As R-squared is 77% and F-statistics is higher than 4, we conclude that the model is a good fit.

Table 5: Main statistics

Dep. Variable:	price2018	R-squared:	0.771
Model:	OLS	Adj. R-squared:	0.767
Method:	Least Squares	F-statistic:	215.8
Date:	Tue, January 26 2021	Prob (F-statistic):	2.26e-100
Time:	20:44:26	Log-Likelihood:	-3907.1
No. Observations:	327	AIC:	7561.
Df Residuals:	321	BIC:	7583.
Df Model:	5		
Covariance Type:	Non-robust		

After fitting the model, the relative importance and significance of each variable are presented in Table 6.

Table 6: Results

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Const	-7.593e+04	7109.203	-10.681	0.000	-8.99e+04	-6.19e+04
Purchasing Power	1241.9925	93.044	13.348	0.000	1058.940	1425.045
CrimeRate	-610.6191	195.320	-3.126	0.002	-994.889	-226.350
Tax	227.1894	14.862	15.287	0.000	197.951	256.428
Toursim	-3.7013	1.177	-3.145	0.002	-6.017	-1.386
Selected Waste	7.775e+04	1.97e+04	3.947	0.000	3.9e+04	1.17e+05

A brief analysis allows verifying that all the features are significant for the P value less than 0.005.

The results show that purchasing power and tax rate and waste are positively correlated to building sold prices while crime rate and tourism and are negatively correlated.

‘The reason that tourism has a negative effect on the prices of houses might be that the regions with more tourists can increase the crime rate, waste disposal, traffic, noise, and

pollution, which can decline the demand for buyers and reduce the prices.' (aware impact, 2019)

Table 7: Additional statistics,

Omnibus:	60.56	**Durbin-Watson:**	1.24
Prob (Omnibus):	0.000	**Jarque- Bera (JB):**	142.319
Skew:	0.908	**Prob (JB):**	1.25e-31
Kurtosis:	5.674	**Cond. No.**	2.44e+04

To improve the analysis, we used other models. I split the sample into 70% to train and 30% to test. We compared linear regression with ridge regression, lasso regression, Bayesian regression, polynomial regression, and neural networks. Our network's architecture has two layers with 9 and 7 neurons, ReLU activation, max_iter of 5000, and as solver the 'lbfgs' (Limited-memory Broyden–Fletcher–Goldfarb–Shanno).

Table 8: Comparison of models

Model	Accuracy on the training subset	Accuracy on the test subset
Linear Regression	0.763	0.642
Ridge Regression	0.762	0.644
Lasso Regression	0.763	0.642
Bayesian Regression	0.757	0.642
Polynomial Regression	0.818	0.426
Neural Network	0.755	0.619

As it shows, all the models have similar results in both training and test subsets with an acceptable level of accuracy. Polynomial regression improves the training accuracy

4. CONCLUSION

Considering the total values of building sales, there was an increase until 2010. Then, the total value reduced but increased since 2013. Before the mid of 2012, house prices fell around 3% on average per year from the beginning of the financial crisis (Del Giudice et al., 2020). Since 2013, Portugal's housing sector has experienced significant changes, with a strong upturn in real estate transactions and a rise in housing prices. This rise in housing prices has resulted from the conjunction of a low supply of real estate property and significant growth in demand. (Nicola et al., 2020)

As we expected, linear regression allows an acceptable accuracy on the training subset. Neural networks and polynomial regression improve the training subset's accuracy but with lower accuracy on the test subset. This is a consequence of the possible overfit limitation of those approaches, but more sophisticated approaches may be used, like in other studies. (Custódio et al., 2020)

This work aims to analyze the prices' evolution and understand the primary components impacting the price. To understand the evolution and predict prices, I perform a time series analysis, including ARIMA. I concluded an essential variability of quantity sold from this analysis, but prices have a crescent trend. It is possible to predict prices using ARIMA with some errors, even with the last year's COVID impact.

To understand the most relevant components, a regression analysis has been performed. I conclude that purchase power, crime, taxes, and tourism may substantially affect the prices. The results showed the higher purchasing power and higher taxes positively affect the prices of houses. On the other hand, the municipalities with a higher crime rate and waste will have lower prices in buildings. Additionally, the tourism rate will have a negative impact on the prices of real estate, as the demand might decline in municipalities with more tourists. This can be for a higher rate of crime rate, waste rate, pollution, noise and traffic. In this analysis, I used several approaches, like neural networks. Results were not much better than with linear regression. It is also essential to consider that this approach has some limitations, especially regarding the analysis's granularity.

In future, I recommended to use other approaches and algorithms such as LSTM (Long short-term memory), random forest or the Prophet library. Also, this study can be extended to further territory divisions (such as *freguesias*) as long as data becomes available to focus policies. Moreover, other variables can be used to affect real estate

prices for future studies since, in this research, causality effects have been found and limited the analysis. Another interesting dimension will be the analysis of costs. In fact, real estate prices depend also significantly on building and maintenance prices. It is expected that Covid will contribute to an increase in inflation, as result of supply chain problems. On the other hand, building construction costa are increasing as consequence of increase of costs of permits and technical audits and the increase in complexity of the components involved. Other important approach would be the possibility analyzing the impact of using technologies in the context of the buying process. In fact, according to several researchers, the use of theologies could contribute to the reduction of transaction costs (Costa, 1996). In fact, information technologies, specially Internet allows a easier comparability between vendors, but also may contribute to the reduction of some costs, like advertisement. .

5. REFERENCES

Aparicio, S., Aparicio, J. T., & Costa, C. J. (2019). Data Science and AI: Trends Analysis *2019 14th Iberian Conference on Information Systems and Technologies* (CISTI), 2019, pp. 1-6, https://doi.org/ 10.23919/CISTI.2019.8760820.

Aware impact. (2019, November 8). 10 Negative Effects of Tourism You Should Know About. Aware Impact. https://awareimpact.com/negative-effects-of-tourism/

Angelov, P. (2020). Using Machine Learning Algorithms To Analyze The Impact Of Crime On Property Values. Issues In Information Systems. https://doi.org/10.48009/1_iis_2020_55-61

Bełej, M., & Cellmer, R. (2014). The effect of macroeconomic factors on changes in real estate prices-response and interaction. *Acta Scientiarum Polonorum. Oeconomia*, 13(2), 5-16.

Berlemann, M., & Freese, J. (2013). Monetary policy and real estate prices: a disaggregated analysis for Switzerland. International Economics and Economic Policy, 10(4), 469-490.

Case, K. E., & Shiller, R. J. (1990). Forecasting Prices and Excess Returns in the Housing Market. Real Estate Economics, 18(3), 253–273.

Case, K., & Shiller, R. (1988). The Efficiency of the Market for Single-Family Homes (No. w2506; p. w2506). National Bureau of Economic Research. https://doi.org/10.3386/w2506

Costa, C. (1996). Internet e estratégia empresarial. Journal Revista Portuguesa de Marketing, 1, 3.

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A Methodology to Boost Data Science. 2020 *15th Iberian Conference on Information Systems and Technologies (CISTI*), 1–6. https://doi.org/10.23919/CISTI49556.2020.9140932

Costa, C. J., & Aparicio, J. T. (2021). A Methodology to Boost Data Science in the Context of COVID-19. *Advances in Parallel & Distributed Processing, and Applications*: Proceedings from PDPTA'20, CSC'20, MSV'20, and GCC'20, 65. https://doi.org/10.1007/978-3-030-69984-0_7

Custódio, J., Costa, C., & Carvalho, J. (2020). Success Prediction of Leads – A Machine Learning Approach. *2020 15th Iberian Conference on Information Systems and Technologies (CISTI*), 1–6. https://doi.org/10.23919/CISTI49556.2020.9141002

Del Giudice, V., De Paola, P., & Del Giudice, F. P. (2020). COVID-19 Infects Real Estate Markets: Short and Mid-Run Effects on Housing Prices in Campania Region (Italy). Social Sciences, 9(7), 114. https://doi.org/10.3390/socsci9070114

Frischtak, C., & Mandel, B. R. (2012). Crime, House Prices, and Inequality: The Effect of UPPs in Rio. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1995795

Gibbons, S. (2004). The Costs of Urban Property Crime. *Economic Journal*, 114(499), 441–463.

Gudell, S. (2021, September). Information From Past Pandemics, And What We Can Learn: A Literature Review. https://finance.yahoo.com/news/information-pastpandemics-learn-literature-222750122.html

Gujral, V, Palter, R. Sanghvi, A & Vickery, B. (2021). COVID Promo. McKinsey & Company. Retrieved September 19, 2021, from http://ceros.mckinsey.com/coronavirus-promo

Grum, B., & Govekar, D. K. (2016). Influence of macroeconomic factors on prices of real estate in various cultural environments: Case of Slovenia, Greece, France, Poland and Norway. Procedia Economics and Finance, 39, 597-604.

idealista.pt. (2021). Property prices in Portugal have soared in the last decade.

Idealista. Retrieved Sep 19, 2021, from https://www.idealista.pt/en/news/property-for-sale-in-portugal/2021/01/21/845property-prices-in-portugal-have-soared-in-the-last-decade idealista.pt. (2020). The effects of COVID-19 on house prices in Portugal in 2021.

Idealista. Retrieved Sep 19, 2021, from https://www.idealista.pt/en/news/property-for-sale-in-portugal/2021/02/22/877-theeffects-of-covid-19-on-house-prices-in-portugal-in-2021

Iman, A. H. M., & Gan, C. (2013). Community Loss Of Residential Value From Water And Noise Pollution. Journal of Techno-Social, 5(2), Article 2. https://publisher.uthm.edu.my/ojs/index.php/JTS/article/view/1420

International Monetary Fund- Portugal and the IMF

Januário, J. F. (2017). The influence of macroeconomic factors on Portuguese Real Estate: A statistical approach. IST.

Löffler, M., & Siegloch, S. (2017). "Property Taxation, Housing, and Local Labor Markets: Evidence from German Municipalities." Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association Vol. 110 , pp. 1-37.

Lourenço, R. F., & Rodrigues, P. M. M. (2017). "House prices in Portugal—What happened since the crisis?" Economic Bulletin and Financial Stability Report Articles and Banco de Portugal Economic Studies, 41-57.

Mc Donald, J. F. (1993). Local property tax differences and business real estate values. The Journal of Real Estate Finance and Economics, 6(3), 277–287. https://doi.org/10.1007/BF01096962

McKenzie, Ed. (1984). General exponential smoothing and the equivalent arma process. Journal of Forecasting, 3(3), 333–344. https://doi.org/10.1002/for.3980030312

Milunovich, G. (2020). Forecasting Australia's real house price index: A comparison of time series and machine learning methods. Journal of Forecasting, 39(7), 1098-1118.

Mach, Ł. (2017). The Application of Classical and Neural Regression Models for the Valuation of Residential Real Estate. Folia Oeconomica Stetinensia, 17(1), 44–56. https://doi.org/10.1515/foli-2017-0004

Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., & Agha, R. (2020). The socio-economic implications of the coronavirus pandemic (COVID-19): A review. International Journal of Surgery (London, England), 78, 185– 193. https://doi.org/10.1016/j.ijsu.2020.04.018

Ofori, P. (2021). Waste Disposal Sites and Residential Rental Values Nexus: An Appraisal of Agogo Asante Akyem Dumps. African Journal of Science, Technology, Innovation and Development, 13(2), 223–233. https://doi.org/10.1080/20421338.2020.1830542

Olajide, S. E., & Lizam, M. (2016). Determining the Impact of Residential Neighbourhood Crime on Housing Investment Using Logistic Regression. Traektoriâ Nauki = Path of Science, 2(12), 6.8-6.17.

Özmen, M. U., Kalafatcılar, M. K., & Yılmaz, E. (2019). The impact of income distribution on house prices. Central Bank Review, 19(2), 45–58. https://doi.org/10.1016/j.cbrev.2019.05.001

Pickford, J. (2021). Subscribe to read | Financial Times. Financial Times. Retrieved September 19, 2021, from https://www.ft.com/content/e30ccb84-6799-11ea-800dda70cff6e4d3%20(2020)

Pinheiro, G, Belo, D (2019) Portugal and the future of housing, Portugal and the future of housing (caixabankresearch.com)

Pour, M. S., Khani, P. N., Zamanian, G., & Barghandan, K. (2013). Specifying The Effective Determinants Of House Price Volatilities In Iran. Journal of Economics and Business, 2, 6.

Read, S. (2020, March 31). Coronavirus: U.K. mortgage market goes into partial lockdown. BBC News. https://www.bbc.com/news/business-52106119

Rocha, C. & Carvalho, A. (2008). Portuguese Real Estate Taxation, Land Use and Noise. INTER NOISE 2008..

Safronov, O. (2017). A neural network based model for mass non-residential real estate price evaluation of Lisbon, Portugal. Msc Dissertation, ISEGI, UNL.

Samadani, S. & Costa, C (2021) "Forecasting real estate prices in Portugal : A data science approach," 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), 2021, pp. 1-6, https://doi.org/10.23919/CISTI52073.2021.9476447.

Società di consulenza Strategica e Aziendale. (2021). Nomisma. Retrieved September 26, 2021, from https://www.nomisma.it/

Sousa, V. (2021). Bolha imobiliária em Portugal e na Europa? O que dizem os dados do Eurostat. idealista.pt/news . https://www.idealista.pt/news/imobiliario/internacional/2021/08/10/48433-bolhaimobiliaria-em-portugal-e-na-europa-o-que-dizem-os-dados-do-eurostat

Temür, A. S., Akgün, M., & Temür, G. (2019). Predicting Housing Sales In Turkey Using Arima, Lstm, And Hybrid Models. Journal of Business Economics and Management, 20(5), 920–938. https://doi.org/10.3846/jbem.2019.10190

The source of home price movement: Buyer purchasing power | firsttuesday Journal. (2021). https://journal.firsttuesday.us/the-source-of-home-price-movement-buyerpurchasing-power/27418/

Tita, G. E., Petras, T. L., & Greenbaum, R. T. (2006). Crime and Residential Choice: A Neighborhood Level Analysis of the Impact of Crime on Housing Prices. Journal of Quantitative Criminology, 22(4), 299–317. https://doi.org/10.1007/s10940-006-9013-z

Tomás, D., Costa, C, Gaivão, J. & Carvalho, J. (2018), "Time series for incidences, orders and invoicing forecast" Proceedings of CAPSI. 36. https://aisel.aisnet.org/capsi2018/36.

Tse, R. Y. C. (1997). An application of the ARIMA model to real-estate prices in Hong Kong. Journal of Property Finance, 8(2), 152–163. https://doi.org/10.1108/09588689710167843

Verbeek, M. (2017). Using linear regression to establish empirical relationships. IZA World of Labor. https://doi.org/10.15185/izawol.336

Wilson, G. (2016). Time Series Analysis: Forecasting and Control,5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. Journal of Time Series Analysis, 37, n/a-n/a. https://doi.org/10.1111/jtsa.12194

6.APPENDIX

```
import pandas as pd from
matplotlib import pyplot
from pandas.plotting import autocorrelation_plot
from statsmodels.tsa.arima.model import ARIMA from
math import sqrt
from sklearn.metrics import mean_squared_error
df=pd.read_excel('https://github.com/masterfloss/dataRealEstate/blob/m
ain/Portugal.xlsx?raw=true')
df['price']=df["Value New"]/df["Number New"]
serie=df[['price','date']]
serie1=serie.set_index('date') serie1.plot()
pyplot.show()
autocorrelation_plot(serie1) pyplot.show()
serie1.index = pd.DatetimeIndex(serie1.index.values, freq=serie1.index
.inferred_freq) # fit model p =
4 # number of lags d =1  #
degree of differencing.
q =0 # size of the moving average window

model = ARIMA(serie1, order=(p,d,q)) model_fit
= model.fit()


# summary of fit model
print(model_fit.summary()) # line plot of
residuals residuals =
pd.DataFrame(model_fit.resid)
residuals.plot() pyplot.show()
# density plot of residuals
residuals.plot(kind='kde') pyplot.show()
# summary stats of residuals
print(residuals.describe()) # split
into train and test sets X =
serie1.values size = int(len(X) * 0.70)
```

```
train, test = X[0:size], X[size:len(X)]
history = [x for x in train]
predictions = list()

# walk-forward validation for t in
range(len(test)):     model =
ARIMA(history, order=(p,d,q))
model_fit = model.fit()     output =
model_fit.forecast()     yhat = output[0]
predictions.append(yhat)     obs =
test[t]     history.append(obs)
    print('predicted=%f, expected=%f' % (yhat, obs))

# evaluate forecasts
rmse = sqrt(mean_squared_error(test, predictions)) print('Test
RMSE: %.3f' % rmse)

# plot forecasts against actual outcomes
pyplot.plot(test) pyplot.plot(predictions,
color='red') pyplot.show()
```

```
predicted=151.781667, expected=162.598906 predicted=164.777341,
expected=161.002441 predicted=159.979375, expected=157.689137
predicted=158.852862, expected=167.271819 predicted=167.550702,
expected=171.637598 predicted=171.277259, expected=176.407034
predicted=175.569707, expected=175.903843 predicted=179.702648,
expected=178.543702 predicted=179.738638, expected=182.210485
predicted=184.025319, expected=192.610120 predicted=192.993593,
expected=181.948200 predicted=184.873914, expected=182.908425
predicted=183.254122, expected=206.800860 predicted=206.699004,
expected=201.230004 predicted=194.676632, expected=200.372706
Test RMSE: 8.463
```

OLS and ANN

```
import pandas as pd import
statsmodels.api as sm from
statsmodels.stats.outliers_influence import
variance_inflation_factor

Out[2]:
Unnamed: 0              object
Price                  float64
```

```
purchacingPower      float64 crime
float64 wage
float64 waste
float64 wasteSel
float64 tourism2018
object grad
float64 tourism
float64 wage
float64 waste
float64 tax
float64 dtype: object
In [2]: df=pd.read_excel('realEstate1.xlsx') df1=df
df1["waste"]=pd.to_numeric(df1.waste2018, errors='coerce')
df1["tourism"]=pd.to_numeric(df1.tourism, errors='coerce')
df1["wasteSel2018"]=pd.to_numeric(df1.wasteSel,
errors='coerce') df1["wage"]=pd.to_numeric(df1.wage,
errors='coerce') df1['waste']=df1["wasteSel"]/df1["waste"]
df1=df.dropna() df1.dtypes
In [14]:
y=df1['price2018']
X= df1[['purchacingPower','crime','waste','tourism','tax']]
X = sm.add_constant(X) results = sm.OLS(y, X).fit()
results.summary() Out[14]:
```

OLS Regression Results			
Dep. Variable:	price	R-squared:	0.771
Model:	OLS	Adj. R-squared:	0.767
Method:	Least Squares	F-statistic:	215.8
Date:	Sat, September 25 2021	Prob (F-statistic):	2.26e-100
Time:	19:30:50	Log-Likelihood:	-3774.3
No. Observations:	327	AIC:	7561.
Df Residuals:	321	BIC:	7583.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.593e+04	7109.203	-10.681	0.000	-8.99e+04	-6.19e+04
purchacingPower	1241.9925	93.044	13.348	0.000	1058.940	1425.045
crime	-610.6191	195.320	-3.126	0.002	-994.889	-226.350
waste	7.775e+04	1.97e+04	3.947	0.000	3.9e+04	1.17e+05
tourism	-3.7013	1.177	-3.145	0.002	-6.017	-1.386

tax	227.1894	14.862	15.287	0.000	197.951	256.428

Omnibus:	60.560	Durbin-Watson:	1.240
Prob(Omnibus):	0.000	Jarque-Bera (JB):	142.319
Skew:	0.908	Prob(JB):	1.25e-31
Kurtosis:	5.674	Cond. No.	2.44e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.44e+04. This might indicate that there are strong multicollinearity or other numerical problems.
In [4]:

```
# VIF dataframe vif_data =
pd.DataFrame()
vif_data["feature"] = X.columns

# calculating VIF for each feature vif_data["VIF"] =
[variance_inflation_factor(X.values, i)
for i in range(len(X.columns))]
   print(vif_data)
feature        VIF 0
const  26.125925 1
purchacingPower   1.475335
2                  crime   1.887041
3                  waste   1.332964
4                  tourism   1.792031
5                  tax   2.547069
In [5]: from sklearn.neural_network import
MLPRegressor from sklearn.model_selection import
train_test_split y=df1['price2018']-df1['price2000']
X=
df1[['purchacingPower','crime','IMTpercapita','IMIpercapita','waste']]
X_train, X_test, y_train, y_test = train_test_split(X, y,
random_state=1)
 regr = MLPRegressor(random_state=1,
max_iter=5000).fit(X_train, y_train)  regr.predict(X_test[:2])

regr.score(X_test, y_test)
Out[5]:
0.6016827668819571
In [6]:
# Linear Regression from sklearn import linear_model reg =
linear_model.LinearRegression() reg.fit(X_train, y_train)
print('Accuracy on the training subset:
```

```
{:.3f}'.format(reg.score(X_train, y_train))) print('Accuracy on the
test subset: {:.3f}'.format(reg.score(X_test, y_test)))
LR_ATrain=reg.score(X_train, y_train)
LR_ATest=reg.score(X_test, y_test)
Accuracy on the training subset: 0.763 Accuracy
on the test subset: 0.642
In [7]:
# Ridge Regression from sklearn
import linear_model reg =
linear_model.Ridge (alpha = .5)
reg.fit(X_train, y_train)
print('Accuracy on the training
subset:
{:.3f}'.format(reg.score(X_train,
y_train))) print('Accuracy on the
test subset:
{:.3f}'.format(reg.score(X_test,
y_test)))

Ridge_ATrain=reg.score(X_train, y_train)
Ridge_ATest=reg.score(X_test, y_test)
Accuracy on the training subset: 0.762 Accuracy
on the test subset: 0.644
In [8]: #Lasso Regression from sklearn import linear_model reg =
linear_model.Lasso(alpha = .5) reg.fit(X_train, y_train)
print('Accuracy on the training subset:
{:.3f}'.format(reg.score(X_train, y_train))) print('Accuracy on the
test subset: {:.3f}'.format(reg.score(X_test, y_test)))

Lasso_ATrain=reg.score(X_train, y_train)
Lasso_ATest=reg.score(X_test, y_test)
Accuracy on the training subset: 0.763 Accuracy
on the test subset: 0.642
In [9]:
#Bayesian Regression from sklearn import linear_model reg =
linear_model.BayesianRidge(compute_score=True) reg.fit(X_train,
y_train) print('Accuracy on the training subset:
{:.3f}'.format(reg.score(X_train, y_train))) print('Accuracy on the
test subset: {:.3f}'.format(reg.score(X_test, y_test)))

Bayesian_ATrain=reg.score(X_train, y_train)
Bayesian_ATest=reg.score(X_test, y_test)
Accuracy on the training subset: 0.757 Accuracy
on the test subset: 0.642
In [10]:
#Polynomial Regression from sklearn.pipeline import
Pipeline from sklearn.preprocessing import
PolynomialFeatures from sklearn.linear_model import
LinearRegression
reg = Pipeline([('poly', PolynomialFeatures(degree=2)),('linear',
LinearRegression(fit_intercept=False))]) reg.fit(X_train, y_train)
```

```
print('Accuracy on the training subset:
{:.3f}'.format(reg.score(X_train, y_train))) print('Accuracy on the
test subset: {:.3f}'.format(reg.score(X_test, y_test)))


Polinomial_ATrain=reg.score(X_train, y_train)
Polinomial_ATest=reg.score(X_test, y_test)
Accuracy on the training subset: 0.818 Accuracy
on the test subset: 0.426
In [11]: #Neural Network from sklearn.neural_network import
MLPRegressor reg = MLPRegressor(random_state=1,hidden_layer_sizes =
(9,7), activation='relu', max_iter=5000, solver='lbfgs')
reg.fit(X_train, y_train) print('Accuracy on the training subset:
{:.3f}'.format(reg.score(X_train, y_train))) print('Accuracy on the
test subset: {:.3f}'.format(reg.score(X_test, y_test)))

NN_ATrain=reg.score(X_train, y_train)
NN_ATest=reg.score(X_test, y_test)
Accuracy on the training subset: 0.755 Accuracy
on the test subset: 0.619
In [12]: print("Model                    | Accuracy on the training
subset | Accuracy on the test subset") print("Linear Regression
|             {:.3f}              ".format(LR_ATrain),'|
{:.3f}'.format(LR_ATest)) print("Ridge Regression         |
{:.3f}
".format(Ridge_ATrain),'|             {:.3f}'.format(Ridge_ATest))
       print("Lasso Regression          |             {:.3f}
".format(Lasso_ATrain), '|            {:.3f}'.format(Lasso_ATest))
print("Bayesian Regression      |             {:.3f}
".format(Bayesian_ATrain), '|
{:.3f}'.format(Bayesian_ATest)) print("Polinomial
Regression    |             {:.3f}
".format(Polinomial_ATrain), '|
{:.3f}'.format(Polinomial_ATest))
print("Neural Network Regression|             {:.3f}
".format(NN_ATrain), '|             {:.3f}'.format(NN_ATest)) Model
| Accuracy on the training subset | Accuracy on the test subset
Linear Regression        |             0.763               |
0.642
Ridge Regression         |             0.762               |
0.644
Lasso Regression         |             0.763               |
0.642
Bayesian Regression      |             0.757               |
0.642
Polinomial Regression    |             0.818               |
0.426
Neural Network Regression|             0.755               |
0.619
```