CT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

Number of words in the document is 3684

the link to github : https://github.com/Sama4723/MSC_DA_CA2

Module Title:	ML, Dataprep, Stat,p rpgraming
Assessment Title:	CA1
Lecturer Name:	
Student Full Name:	Samah Ahmed Mohamed Ahmed
Student Number:	sbs23090
Assessment Due Date:	14/04/2023
Date of Submission:	14/04/2023

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

May 2023

Contents

Abstract

The datasets for prices of houses in Ireland, France and UK were studied to get insights about how the prices vary form one country to another. A machine learning modeling was applies to predict the prices of houses in Ireland. The model performance on the training sets was acceptable with accuracy of 75% on average. However the accuracy of the testing set was extremely low. Statistic models conclude that the prices in Uk and Ireland vary.

Chapter 1

Data preparation and visualization

1.1 Dataset Insights

In this project will be following the Grisp DM mangement framework because it is most commonly used approach for data science projects.Further the nature of the dataset is suitable to follow the steps of this particular framework as will be shown in this report.

The dataset is about Dublin Residential Property Price Register for 2017. It was downloaded as a csv file form data.gov.ie, License by cc-by and last updated on August 31, 2019. The dataset contains number of 9 Variables, 17952 rows, 20295 missing Cells and 22 dublicate rows. The 9 features are:
Postal Code, Property Size Description,[1] Address, Price, County, Date of Sale (dd/mm/yyyy), Not Full Market Price and Description of Property.

The motivation to choose this dataset is the number of samples is generous and the features are clear and understandable. Beside this, the number of missing samples is not big.

Problem statment of this report is to study indepth the prices in Dublin for properties and to get insight about how the proces varies within different areas in Dublin and according to the descroption of the property itself.

Column	Non-Null Count
Date of Sale	17952 non-null
Address	17952 non-null
County	17952 non-null
Price	17952 non-null
Not Full Market Price	17952 non-null
Description of Property	17952 non-null
Postal Code	11543 non-null
Property Size Description	4066 non-null
VAT Exclusive	17952 non-null

Table 1.1: Table listing all columns and number of observations available

1.2 EDA

The number of missing values is not considered large with a percentage of only 12% of the dataset as indicated in the bar chart number 1.1. Two variables have missing values, Property Size Description variable missing 77% of the values while, Postal Code missing 35% as indicated in table 1.2. There is 1 numerical variable ,Price, and 8 catogorical variables.

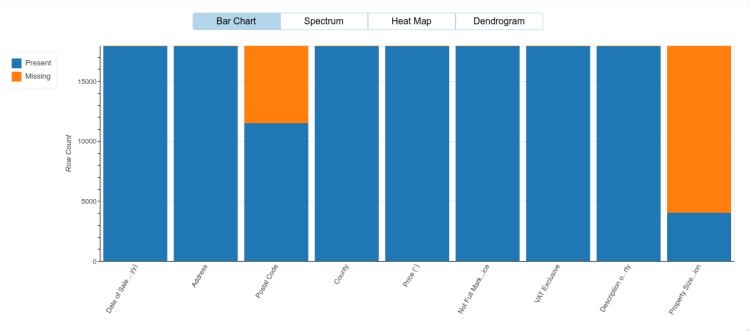

Figure 1.1: Bar chart illustrate missing values

Column	Variable Types	Missing	Unique values
Date of Sale	categorical	0	303
Address	categorical	0	17796
County	categorical	0	1
Price	numerical	0	3405
Not Full Market Price	categorical	0	2
Description of Property	categorical	0	3
Postal Code	categorical	6409	23
Property Size Description	categorical	13886	3
VAT Exclusive	categorical	0	2

Table 1.2: A table listing all variables types, number of missing values, and unique values

All the variables are plotted to get insight about the distribution of the values and the highest and the lowest number of counts. All the graphs were generated using the EDA dataprep.eda function. In 1.2 shown the first 10 addresses.
In 1.3 the highest number of counts is for the Dublin 15 value then comes the Dublin24 for the Postal Code variable.

Graph 1.4 states that 250,000 the most common price for properties in Dublin in 2017.
The graph 1.5 for Not Full Mareket Price indicates that most of the prices are not the full price in the market. Graph 1.6 agrees with graph 1.5 that most the prices not including the vat.
The Description of property stating that 77% of the counts are for the second hand houses and 22% for the new houses 1.2
The Property Size Description variable graph 1.2 states that the property size is greater than or equal to 38 square meters and less than 125 square meters with counts of 17%.

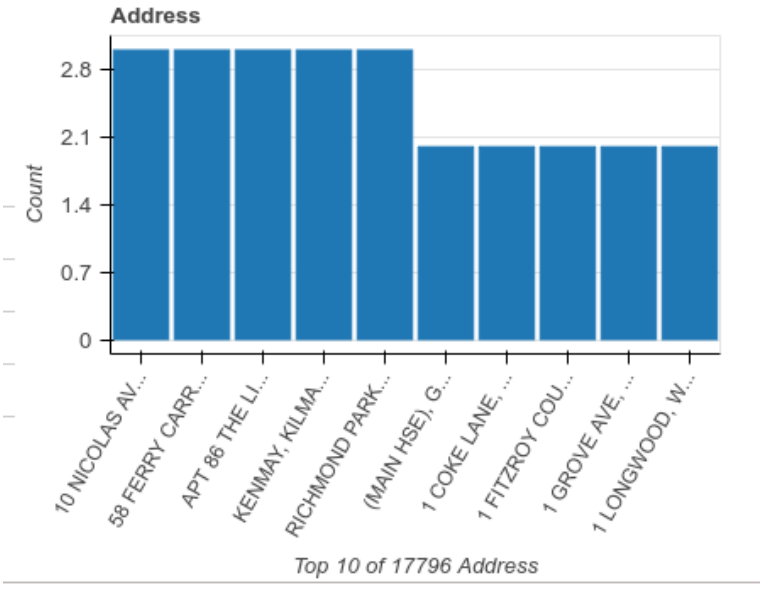

Figure 1.2: Number of counts top 10 values for Address variable

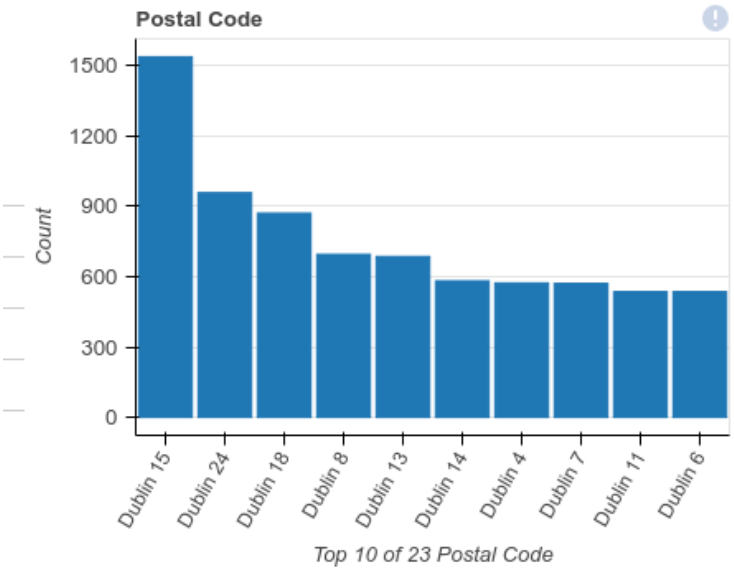

Figure 1.3: Number of counts top 10 values for Postal Code variable

1.3 Data Preparation

1.3.1 cleaning the data

The Column Property Size Description was not concidered in this analysis due to the large number of missing cells ,therefore, the variable was deleted. The column Price needed to be cleaned and processed to be ready for plotting and and machine learning modeling. The values were stored as string with comma in the middle. The comma was deleted beside the Euro symbol to convert the type of the variable form object to float. All the steps are clearly written in the jupyter notebook attached. All the nan values were filled with zeros. Further, imputaion for the missing values in the Postal Code column were generated using the column Address to extract the corresponding Postcode with the help of Goole maps API. All the details are listed in the Jupyter nootebook and further justification and declaration stated in the next section.

Figure 1.4: Number of counts top 10 values for price

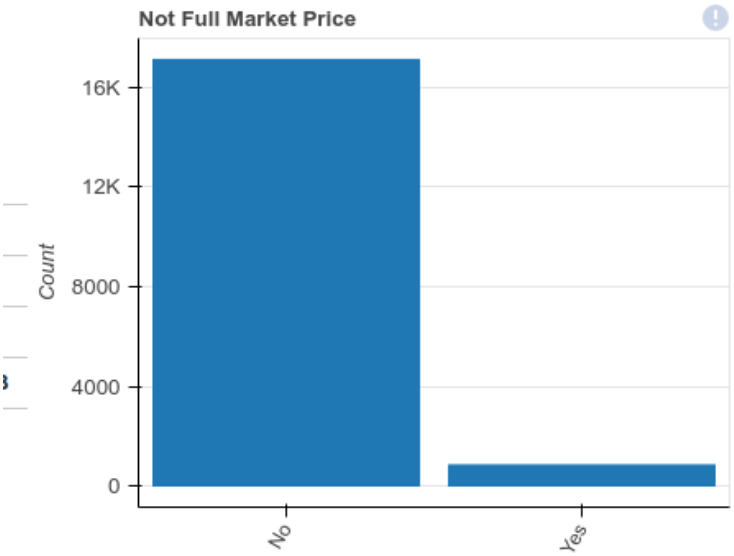

Figure 1.5: Number of counts top 10 values for Not Full Mareket Price variable

1.3.2 Imputation of missing values

The variable Postal Code missing a 1065 values. Since there is a column with the full address Google maps API were used to geocode the addresses to get the corresponding lattitude and longitude and the presice postcode. A formatted address was gathered in a jason format file then converted to csv format. The addresses were parsed to get all the 3000 postcodes. The motivation behind choosing the Postal Code variable to analysis the prices of the houses is that it is essential to examin if the prices varies according to the locaion, for example, is the property price in Dublin 15 is the same as Dublin 3?.

The plan of getting postcodes form google maps was not working as expected because the generated postcodes were writted in a very chaotic order; some addresses have 4 entrries, others have 5. Some has the postcode as Dublin 3, Dublin7 others, have the postcode as aircode. However, the longitude and the latitude of each location were generated successfully and plotted on a map

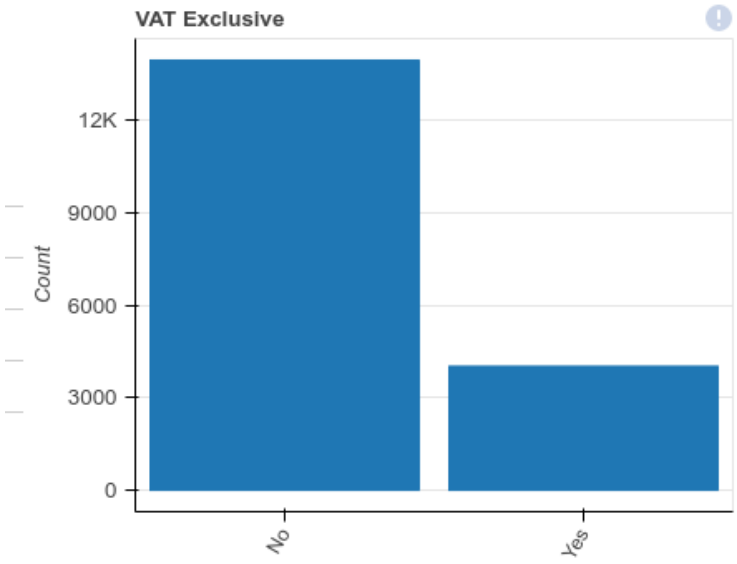

Figure 1.6: Number of counts top 10 values for VAT Exclusive variable

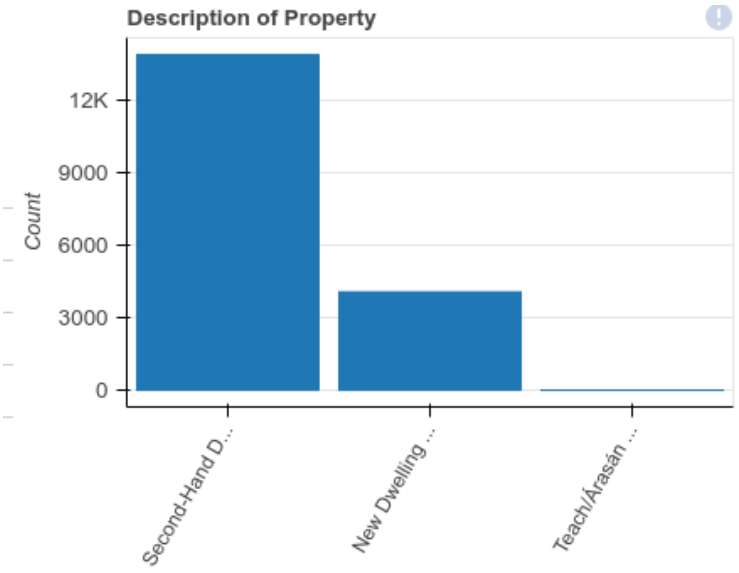

Figure 1.7: a histogram showing the counts of the top 10 values of the Description of Property variable.

illustrated in the figure 1.3.2. After all, the postcodes were added to the dataframe and encoded using a label encoder to make it understandable by the machine learning model. However the resulted accuracy by the trained model were extremly low, this is justified by the unaccurate parsing of the postcode due to the very low quality of the initialy gathered data. Another reason, is due to using a lable encoder instead of using a one hot encoder. The label encoder label the postcodes as numbers 170, 132,168,44,... these labels not suitable for the models as the model interpert the numbers as ranked values not labels. The one hot encoder was not chosen to encode the addresses because this will produce a tremendous amount of columns.

1.4 Descriptive statistics

The large number of samples in the dataset is challenging to work on with the limited available computing resources. Therefore, the dataset was shuffeled to take a representative sample [2]

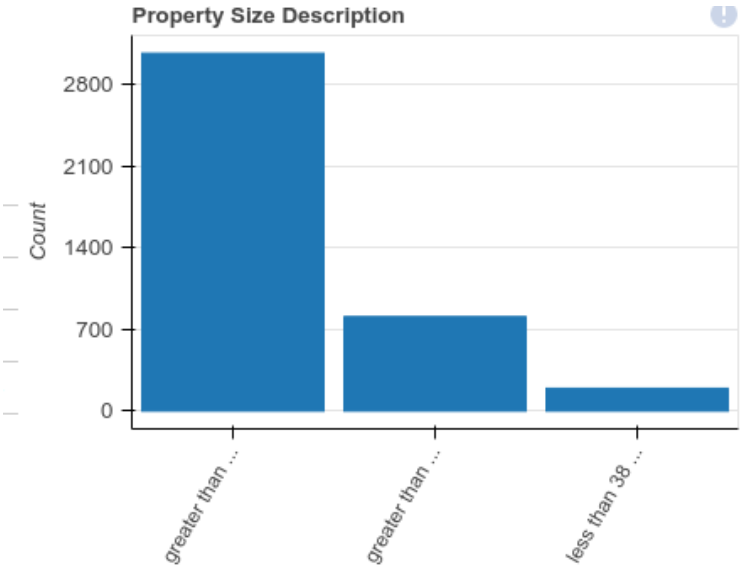

Figure 1.8: A histogram showing the to 10 values of the roperty Size Description variable .

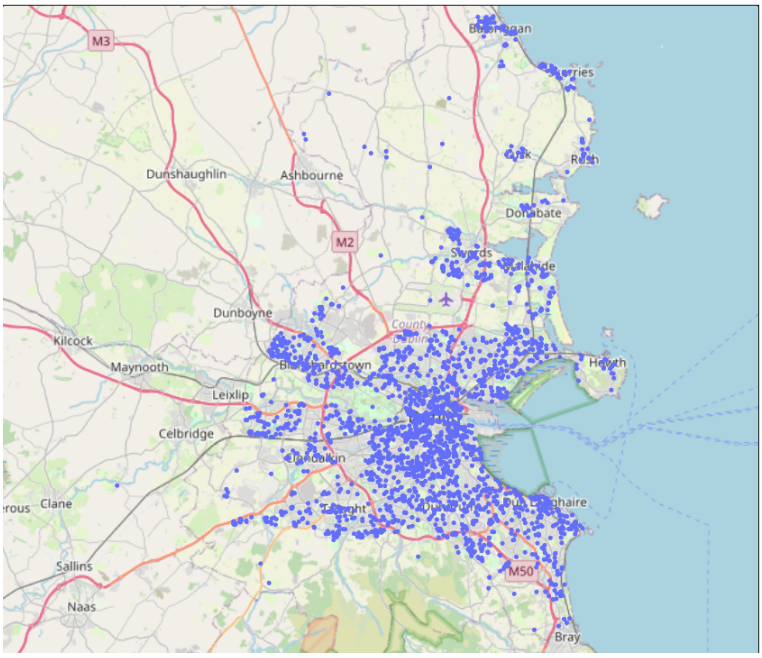

Figure 1.9: Google map with all the address scattered on it.

of 3000 observations to be 3000. The variables are all categorical except the price column, the table 1.4 list the counts,number of unique values, most frequent value, the frequnecy of the most frequent value for the full dataset. For Address variable there is an address has frequent 3 this was concidered as duplicate and all the dublicate rows were deleted in the data processing step.

	Date of Sale (dd/mm/yyyy)	Address	Postal Code	County	Price (')	Not Full Market Price	VAT Exclusive	Description of Property	Property Size Description
count	17952	17952	11543	17952	17952	17952	17952	17952	4066
unique	303	17796	23	1	3405	2	2	3	3
top	14/12/2017	58 FERRY CARRIG RD, COOLOCK, DUBLIN 17	Dublin 15	Dublin	250,000.00	No	No	Second-Hand Dwelling house /Apartment	greater than or equal to 38 sq metres and less...
freq	256	3	1538	17952	198	17119	13937	13885	3064

Figure 1.10: Table showing all the descriptive statistics for the categorical variables.

	Price
count	3.000000e+03
std	3.819725e+05
min	5.675000e+03
25%	2.422465e+05
50%	3.210000e+05
75%	4.700000e+05
max	7.600000e+06

Table 1.3: A table listing decrriptive statstics for price variable

1.4 and 1.4

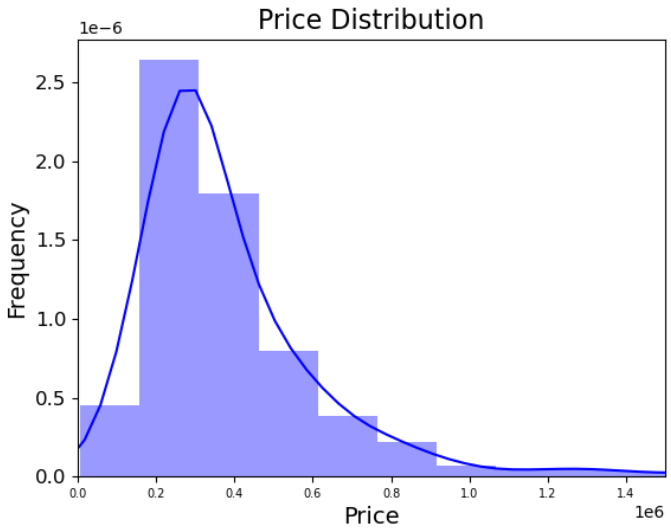

Figure 1.11: The histogram shows the distribution of prices.

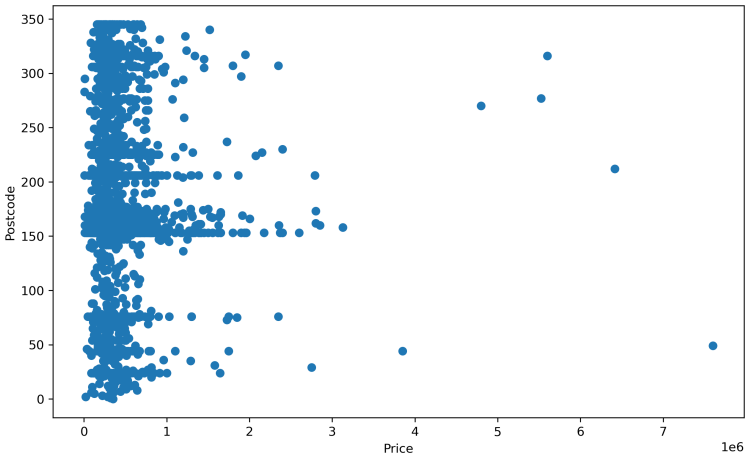

Figure 1.12: TScatter plot for prices versus postcode.

The Price column has a standard deviation of 4.416,> mean, which indicates that the data are well spread from the mean. Quarter, half, and 75% of the dataset are listed in the table 1.3. In the graph 1.4 the distribution of the prices is skewed to the left which indicates the larger number of prices are

between 200.000 and 40.000.The scatter plot1.4 shows the prices versus the postcode.It is clear that there is no correlation bewteen the two variables i.e it not clear that houses in Dublin 4 for example has more prices than houses in Dublin 15.

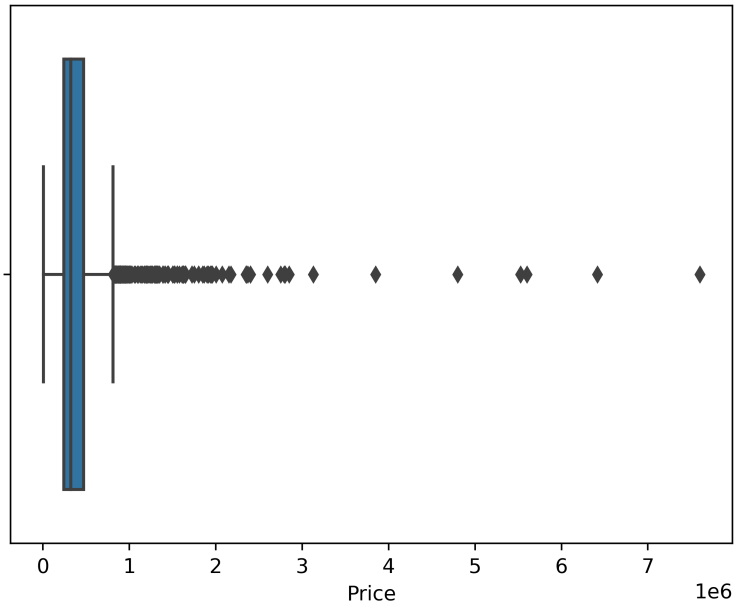

Figure 1.13: Prices box plot .

Two other datasets were studied in this analysis, one for houses prices in France and the other for houses prices in UK. The price variable in dataset from France has count = 5.450000e+02 the mean of the total prices in the sample is 4.766729e+06, the standard deviation is 1.870440e+06 which indicates that the data is well spread from the mean the lowest prices are around 1.750000e+06 Euro while the highest ones are around 1.330000e+07 Euro. 25$\%$ of the prices are around 3.430000e+06 Euro and half of the prices are around 4.340000e+06 Euro.

For the house prices in UK there are, 100.000000 counts, the mean of the total prices in the sample is 72859.350000, the standard deviation is 55767.459497 which indicates that the data is well spread from the mean the lowest prices are around 4000.000000 Pound while the highest ones are around 312500.000000 Pound. 25$\%$ of the prices are around 36837.500000 Pound and half of the prices are around 53225.000000 Pound.
Confidence interval The population mean for prices of houses in Ireland at confidence interval of 95$\%$ is $(397819.89203099883, 425167.8451823335)$, for UK is $(61793.87615240995, 83924.82384759007)$ and for France is $(4609345.1502212, 4924113.345191644)$. It is clear that France has the highest mean for prices then Ireland and then UK. **T test** Performing a t-test to compare houses prices in Ireland and France:

-for dataset of prices in Ireland we pick a sample of n=10, std =723300.21 , mean=421632.36 significance level 95$\%$ -for dataset of prices in France we pick a sample of n=12, std =1870439.61 , mean=4766729 significance level 95$\%$
- Substituting the critical values: the number of degrees of freedom = 20, and the alpha = 0.025 in the app to ge x= 2.423.

-The hypothesis are H0: the mean values for the tow populations are equal mu1 = mu2 H1: the mean vlaues for the tow populations are different mu1 =! mu2.

We accept H0 if t < -2.423 or t > 2.423 and We reject H0 if -2.423 <= t <= 2.423 We found t < -2.423 hence, we reject H0. The interpretaion there is enough evidence at 5% significance level to say that there are significant differences between prices of houses in France and Ireland. Note all the calculations are stated clearly in the Jupyter notebook.

Performing a t-test to compare houses prices in Ireland and UK:

for dataset for prices in $\mathbf{Ireland}$ we pick a sample of n=10, std =723300.21 , mean=421632.36 significance level 95%
-for dataset for prices in $\mathbf{UK}$ we pick a sample of n=12, std =55767.4594, mean=72859.35 significance level 95%
- Substituting the critical values: the number of degrees of freedom = 20, and the alpha = 0.025 in the app to ge x= 2.423 -The hypothsis are H0: the mean values for the tow populations are equal mu1 = mu2 H1: the mean vlaues for the tow populations are different mu1 =! mu2

We accept H0 if t < -2.423 or t > 2.423 We reject H0 if -2.423 <= t <= 2.423
We found $t > 2.423$ hence, we accept H0. The interpretaion there is enough evidence at 5% significance level to say that there are no significant differences between prices of houses in UK and Ireland.

A **Shapiro test** was done to check the normality of the data, the data is showing a normality behaviour in agree with the QQ plots in the Jupyter notebook.

levene test was done to check the homogeinity of variance between prices in Ireland and UK, the results shows the variances are equal.

ANOVA test Since the datasets are normal, homogeneous and independant we applied the ANOVA to check if the prices of houses in Ireland are different form prices in France. The result is there is no different between France and Ireland.
non parametric test U-mann test and ANOVA were applied to investigate the results of the parametric tests done. The non parametric tests agrees with the parametric tests. We can concluse there is a difference between prices in Ierland and France.

1.5 Machine learing modeling

The supervised machine learning technique is used to model the dataset in this analysis. A set of regression models applied to model the "Price" as the target variable, the independat variables are "Description of Property" and "Property Size Description". The nature of the "Price" variable is a continous numerical variable. Therefore, the chosen model is regression not classification. Regression models are used to forcast and predict how the prices of properties in dublin change according to the independant variables "Description of Property" and "Property Size". Further, regression mod-

els are alligned with the problem statment of this analysis, and to identify patterns and relationships within a dataset.

Natural language processing techniques were used to convert the text data in the independatnt variables to numbers. Two different vactorizers were used, the first is CountvectorizerT and the second is F-IDF both of them gave very similar accuracy after running the models. The reason that the tow vactorizers doesn't perform differently is the nature of the text in hand is very short and contains less unique words.
Then splitting the data to training data with size 80$\%$ and testing data with size 20% of the sample.

The first, model used to train the data is the linear regression model with an accuracy of 0.73 for training training set and accuracy of -12713 for testing set. The R^2 score of training is acceptable while the one for testing is extremly worse. The very low score for the testing set can be interperted as the model failed to generlize the results obtained during training and there is a high chance of overfitting the model to the data.

The second, model used to train the data is ridge regression with and average accuracy of 0.75 for training set and average accuracy of -0.3 for testing set. Ridge regression performed better than linear regression in the training and testing set. This implies that ridge model is better in generalization and less likely to overfitting the data.

The third, model used to train the data is lasso regression model with an average accuracy of 0.75 for training set and -0.34 for tesing set the performance of the lasso model is similar to the ridge model.
The over all performance of the three models is quite similar due to the quality of the data and small number of features available for the modeling. In the next paragraph the hyperparameter tunning is discussed in detail.
Choosing between the three models is dependant on the aim of he analysis and the nature of data. In general , ridge is more suitable for dataset with high number of feature as it is good in choosing the important features over the less important ones, while lasso is suitable to get an easier and more understandable model. For this analysis particularly, the two models performance are very similar.
Hyperparameter tunning
GridSearchCV method used to tune the value of α to get the best results for modeling the data.

For the ridge regression model changing the value of α doesn't have effect on the accuracy for the training set on average the accuracy is 0.75 while, for testing set accuracy a higher value for α increased the accuracy. That is explained by the higher α is the the more restricted the mode, hence; the coeffiecients of the models are samller in magnitude. All the details of the values used for α are listed in the Jupyter notebook.

Lasso regression model shows a similar predictive performance as the ridge model when $\alpha = 0.01$. The class ElasticNet combines ridge and lasso models also applied to model the data, the resulting Mean Squared Error on test set is 180624879269.62692 which is a very high value and very low

accuracy of 0.04345.

K-Fold Cross Validation

To test authenticity of the modeling out comes the cross validation function is used for all the three models used in this analysis. For linear regression the accuracy usign cross validation function is on average 0.7 for the training set and -0.2 for the testing set when the data is splitted to 5 parts i.e kfold =5. While, when kKfolds = 3 is slightly increased and the test score is the same.

For ridge an lasso models the number of kfolds doen't affect the accuracy. And the resulted accuracy is apporximately colse the one produced my the modle initialy.

model	training score	testing score
Lnear regression	0.79	1.00
Ridge regression	0.72	-0.08
Lasso regression	0.79	-0.00

Table 1.4: Table listing scores for each model

Table 1.4 list the testing and training scores for each model. The training and testing score increased for linear regression model after the $-ypredstest-$ value taken into account to calculate the score instead of the Y test value. The same value was used to calculate the scores for the other two models. The models are not performing well on predicting the prices of properties in Dublin.

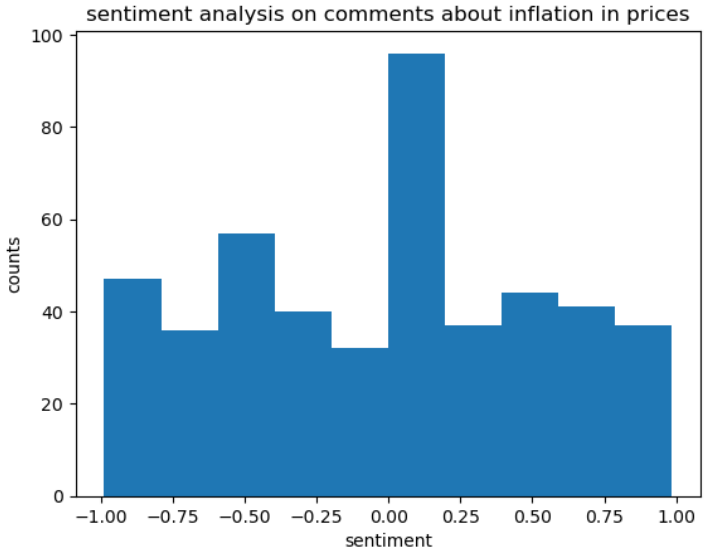

Figure 1.14: Distribution of comment sentiment from 1 to -1.

1.5.1 Sentiment Analysis

To perform a sentiment analysis a collection of comments made by people about the inflation of houses prices in Ireland form the www.reddit.com website. The vaderSentiment library imported to use the SentimentIntensityAnalyzer function to analyse the comments. Figure 1.5 illustrate the distribution of the sentiment for each comment, the highest counts are neutral comments. However, the graph shows more counts on the negative side. This indicates the dissatisfaction of the people

about houses prices in Ireland. This result is expected as the inflation of prices is affecting the prices in the whole world.

1.6 Conclusion

The problem stated in this analysis is to study the datasets of prices of properties in Ireland, UK and France. The quality of the dataset and nature of features contains in the dataset affect the insights gained form this analysis and wether it reflects the reality of the growing prices in the whole world.

The quality of the data collected for houses in Ireland is extremely challanging to process and visiualize as it doesn't contain a valuable features that affect the prices directly such as the area, the number of rooms and dettahced or semi-dettached. Therefore, the machine learning modeling was not able to predict the prices accurately for the testing sets. Since the prices are continous variables, a regression model was chosen to perform the modeling. The regression models perform better when there is enough number of features this is another reason of the very low attained accuracy.

A number of statistical tested applied to check how the prices varies form one contry to another. To conclude prices in ireland vary from prices in France and more similar to prices in UK. This results contradict the reality, the prices in Ireland are the highst in Europe recently according to the indepandant jornal article about this topic.

A better prediction and forcasting of prices could be a cheived in the future by appling some feature engineering techniques.

Bibliography

[1] AureŽlien GeŽron. *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, 2017. ISBN: 978-1491962299.

[2] N.A. Weiss. *Introductory Statistics*. Pearson Education, 2012. ISBN: 9780321691224. URL: https://books.google.ie/books?id=%5C_r5ucgAACAAJ.