SOLENT UNIVERSITY

FACULTY OF BUSINESS LAW AND DIGITAL TECHNOLOGIES

MSc Applied AI and Data Science

Academic Year 2021-2022

RUTH BABATUNDE

House Price Predictive System

Supervisor: Dr Drishty Sobnath

September 2022

This report is submitted in partial fulfilment of the requirements of Solent University for the degree of MSc Artificial Intelligence and Data Science

ACKOWLEDGEMENT

I would begin by appreciating the Almighty who has saw me through these years. Thank

you, Lord, for your Grace.

I am grateful to my supervisor Dr Drishty Sobnath for her great contribution towards the

quality of my project. Thank you for your time and advice.

My gratitude also goes to all my tutors in Applied AI and Data science department as they have all impacted my life one way or the other.

In a special way, I am grateful to my husband Sunday Osaele, babe I see your concerns and sacrifices and I am grateful for the words of hope and encouragement you lavished me with. I am deeply grateful.

To my baby Gloria Osaele, thank you for keeping me up at night with your disturbances. It made me learn how to manage my time.

To my parents Mr. and Mrs. Babatunde, thank you for the constant reminder that you care

and expect the best from me. I have not and will not let you down. My singular sister

Oluwafunmito, you know I love you just like that right? We will always be relevant.

Furthermore, I want appreciate my colleagues that helped me in one way or the other during child birth and this research in like of Simisola, Michael, Peter, Idris and the whole class of Data science 2022.

ABSTRACT

Nowadays, people view homes as an investment strategy rather than just a place to live. To increase house equity, accurate housing price forecasting is crucial. House price fluctuations in the real estate market are caused by the effects of other features that are correlated to housing prices.

This research examined and investigated house payment data obtained from the UK land registry in order to more effectively and properly assess house pricing. In this study, an effort has been made to develop a predictive model for evaluating the price based on the price-influencing variables. Modeling explorations use boosting algorithms like Extreme Gradient Boost Regression (XG Boost) as well as regression techniques like Linear Regression, Decision Tree, and Random Forest models. By performing a comparative study on the prediction errors obtained between various models, the models were utilized to construct a predictive model and select the best-performing model. Here, an effort was made to build a predictive model for evaluating the price based on those characteristics.

The outcomes of the experiment demonstrate that XGboost has the highest prediction accuracy, with a prediction accuracy score of 0.90. The XGboost algorithm performs better at generalization and resilience in data prediction than linear regression and decision tree models, and it also guards against overfitting.

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

1.1 Background

Housing has gone beyond just providing a place to live, it has become an investment that affects the economy vice versa. One of the most significant and expensive purchases a person will ever make is a home. A home reflects the significance that is embodied, home is simply not only bricks and mortar or a shelter from wind and weather but comes with deeper meaning and attachment (Blunt et all, 2006).

House price is now a marketable, lasting, irreplaceable commodity and investment (Mansi Jain et al, 2020). People invest in real estate, and in recent years, the price of homes has been rising yearly. As a result, investors have been able to profit from it as people started to take notice. The housing market currently has a significant impact on the financial sector and the overall economy due to the increase in home investments. (Xinyu Yang et al, 2017).

House prices are a significant impression of the economy, and its value ranges are of great concerns for the clients and property dealers.

Also house price prediction refers to a concept of evaluating property prices by using various techniques. It serves as a first-hand assistant for people in purchase or sale of properties. (Maida Ahtesham,2020)

Accurate and timely statistics and the publication of home price trends are crucial for stabilizing the real estate market and promoting economic growth because housing prices are strongly tied to everyone's lives. (Yuchen Ni,2022). However, house prices are affected by many factors including physical conditions, locations, number of bedrooms and others. For example, economic factors are part of house price influencer: Recently in august 2022, Rightmove properties announced the fall in house price as the first ever in 10 years (Guardian news, UK) due to fewer people

looking to move because of inflation on living. This is just to proof the relationship between house price and economy. Also, in the (Telegraph news), there was an announcement and I quote" London price will fall by 12pc over the next 2 years as buyers in the country's most expensive market are hammered by the cost-of-living crisis". Housing price escalate every year which eventually reinforced the need of strategy or technique that could predict house prices.

People are concerned about house price trends as the housing industry is starting to boom, so house price prediction has grown into an essential field of study that aids in strategic decisions regarding buying and selling homes.

In recent years, purchasers have been able to plan their investments ahead of time and save time and money by projecting the future trajectory of house prices. Real estate firms can also plan their strategic marketing initiatives in advance to optimize rewards. (Yihao Chen et al, 2021).

Several Machine Learning models and methodologies have been used to estimate house sale price in recent years, as studies have shown that the housing market is strongly related to monetary policy, social networks, the stock market, etc. Thus, forecasting home prices can assist the government in stabilizing the real estate market and preventing price fluctuations. Additionally, this might assist individuals and investors in deciding when and where to invest in real estate (Choujun Zhan et al, 2022).

Every day, houses are sold, and buyers are anxious about whether they are receiving a fair price for the home. They are also concerned about knowing when to invest and the finest locations.

This study looks into and compare different models to pick the best fit for the system that points the pattern and trends in the chosen borough.

1.2 Problem Statement

There has been a rise and fall in house price in recent time which has made investors not to know where and when to invest in England. Past few research has predicted price with a level of accuracy, this study proposes a prediction model that predict price with a better accuracy than previous researches.

1.3 Research Hypothesis

There are a number of hypothesis that has been confirmed influences house prices over the years which are:

- Old properties will have less price than new one
- Terrace and semi-detached property will have more price compared to flat.
- City property prices will be more
- County and District also have impact on property price

This research wants to confirm that and also investigate the influence of a property's age (whether an old or new property) on price.

1.4 Aim

To create a predictive model for house price.

1.5 Objectives

To identify the features that influences house price

To develop a predictive model for house price

To develop a graphic user interface (GUI) through website that demonstrates the model

1.6 Justification

There has been a lot of study on house price prediction using different datasets from different nations, but there hasn't been much done with the dataset from England.

Using data from Melbourne, (Guangliang Gao et al., 2022) conducted extensive research elucidating several aspects that affect housing price. He advised further investigation be done to determine if the year of construction (which is the property age) influences price. This research's goal is also to help buyers, homeowners, and investors make educated judgments on the direction of the housing market.

1.7 Scope

The project is within the limit of four cities which are Manchester, Birmingham, London and Southampton in England with the history data of five years which is 2017-2021.

1.8 Chapter Breakdown

This research is further broken down into:

Chapter Two: A deep literature review of related works surrounding house price generally

Chapter Three: Describes the methodology used in achieving the goals

Chapter Four: Discusses the implementation process, development and deployment.

Chapter Five: Discusses the results and benchmark

Chapter Six: contains the Conclusion, limitation and recommendation

2. LITERATURE REVIEW

2.1 Overview

The price of housing is a significant indicator of the status of the economy and can aid in the planning of future investments by real estate developers. It's crucial that both buyers and sellers are well-informed when making judgments. (Feng Wang et al, 2019). No matter where they are from or how much money they have, the majority of individuals aspire to own a home. The application of regression algorithms can help establish a reasonable pricing.

The classic method of predicting home prices was first employed more than 40 years ago, and it has since been refined with the use of various models and data from various places with an emphasis on the factors that influence home prices. The effects of several factors, such as location, house size, and proximity to a school, have been studied in the past. A few factors, including the type of house, the number of rooms offered, parking space, etc., as well as government legislation, might influence the price of a home. In this assignment, we'll examine how the type of housing affects the cost. whether it is a freshly constructed or existing home. (Yong Piao et al,2019)

Real estate stakeholders can utilize house price prediction as a useful tool to help them make better decisions, such as looking for potential homes that fit their budgets or continuously watching the market for the optimum time to sell a home. (Jie Cao et al, 2022)

A model for predicting home prices aims to comprehend the key elements influencing price changes in a certain area. Numerous models based on conventional statistical methods have been put forth, however due to the complexity of the component, they are insufficient.

The complexity and nonlinearity of the influencing factors make house price prediction a difficult subject. A deep learning-based model for predicting home prices is put out and put to the test using actual housing data. (Feng Wang et al,2019)

Studies have revealed that the location of a home and its surrounding neighborhoods have a substantial impact on the price of a home. However, when the size of the housing data grows, the prediction performance of such systems becomes unsatisfactory. (Jie cao et al, 2022)

For many real estate stakeholders, including home owners, buyers, and investors, accurate house forecast is of utmost importance.

2.2 The Housing Market in United Kingdom

The term "financialization" refers to "the rising importance of financial markets, financial motives, financial players, and financial institutions in the operation of the domestic and international economies" and describes the process that the UK economy has undergone over the past forty years. (Epstein, 2005). The housing markets in the UK have been severely affected by this phenomenon. Financial institutions and investors are now more able to bet on housing and other financial products that use home as collateral thanks to financial deregulation and the privatization of the stock of social housing. Real estate has become "simply another asset class" as a result of this. (Van Loon and Aalbers, 2017).

Due to the resulting rise in housing values, landlords and homeowners have realized large capital gains, while private renters have seen an increase in rent. When consecutive Conservative governments significantly altered the ownership and regulation of both the housing and financial systems in the 1980s, the financialization of the UK housing market really got going. A privatized system has been established as a result of the surge in consumer credit spurred by the Thatcher

administration's policies. This indicates that a greater proportion of aggregate demand the desire for all products and services in the economy now depends on people having access to cheap credit, notably through mortgage lending (Crouch, 2009). Real estate has developed into a new asset class, as can be shown. This is clear throughout the world, but especially in the most financially integrated nations. Most multinational banks' business models now include mortgage lending and securitization, which has strengthened the connections between housing and finance (Aalbers, 2016). More people than ever before are able to invest in real estate without actually buying a specific piece of property through the housing market by purchasing funds that buy a portfolio of properties on their behalf. Instead of being a basic commodity like it once was, real estate—and particularly residential property—has evolved into an asset over which bankers can speculate. The ability of investors to speculate on property values has now supplanted the human right to dwelling, which is protected by international law (Rolnik, 2013).

2.3 Factors Affecting House Price

It can be difficult and confusing to predict the price of a home. Numerous elements, such as location, room orientation, retailers, decoration, neighborhoods, schools, traffic patterns, and security concerns, among others, could affect a home's price. It is impossible to consider all the variables involved in price prediction. Scientists find it difficult to collect precise data and forecast price changes because this circumstance is always changing. Therefore, it is challenging to enhance price prediction without a better algorithm model that focuses on the important variables that affect the real price.

In small areas, house prices are spatially auto-correlated, but they are also spatially heterogeneous in different parts of the world. According to research, there are numerous factors that affect property prices in England at various geographic sizes.

House prices are influenced by macro-structural political, economic, and demographic issues, such as regional economic development, infrastructure provision, and migration policies, at the largest national or regional dimensions. House prices are influenced by urban shape, local economic conditions and amenities, and the availability of various transportation options at a scale that may encompass cities, local authorities (LAs), or travel-to-work zones. (Hamnett et al, 2019). The character of neighboring homes, neighborhood amenities (like parks and schools), local public goods (like open space), and the accessibility of public transportation all have an impact on housing prices at a larger scale. While house prices on the same street or in the same neighborhood tend to be comparable, they might differ due to physical characteristics such home size, age, structural design, and historic value. (chi et al,2021)

2.4 Machine Learning and Artificial Intelligence (AI)

In many academic disciplines, technology is rapidly advancing. This frequently results in a change in the way firms and corporations manage their operations. As a result, artificial intelligence (AI) is now used as a problem-solving approach in practically every industry, including but not limited to the following domains: business intelligence and insight, augmented reality, and supply chain, to name a few.

Nowadays, AI is employed to create inventions that meet human demands. (Christain Muhlroth et al, 2020). However, obtaining references from data cannot be used to enable the employment of AI in providing such support. Machine learning is a component of AI that helps with obtaining these insights (ML). (Winky K et al, 2021)

By definition, machine learning is the study of how computer systems are used and developed to learn and adapt without explicit instructions by examining and deriving conclusions from data patterns using algorithms and statistical models. It is so important in the twenty-first century that it is used practically everywhere, from

commonplace things like a search engine and an email filter to more challenging issues like predicting consumer behavior or our topic, predicting property prices. ML is a branch of AI that develops mathematical models based on prior (historical) data to predict future outcomes of any event. As a result, machine learning has become a significant factor in recent years.

Around the world, ML is currently being used to enhance company skills and discover insights through data analysis. (Gan srirutchataboon et al, 2021)

All businesses are going towards the expanding AI and ML stock market, and the financial industry isn't left out. This is true in many facets of our lives, such as medical diagnosis and weather prediction. (Nivitha Shree et al, 2022). With the current technological revolution, massive amounts of data are extracted daily from practically all organizations and people. On our computers, phones, networks, etc., there is a ton of data. For gathering and storing data, many organizations and companies have various resources. Data science is a profession that involves gathering data and analyzing it to gain insightful information (DS).

Data is the driving force behind technological advancements, and employing predictive models, any outcome is now feasible (Ayush Varma et al,2018). The field of data science (DS) focuses on understanding and gaining knowledge from any kind of data by using methodologies, procedures, algorithms, and models. Any computer that wants to learn from the past must first be given data, after which it must use the knowledge, it has obtained to make decisions in many fields. The discipline of data science has greatly developed and grown, drawing scholars from both scientific and non-scientific fields. For instance, a coffee business would like to know how many customers they receive year-round. Analyzing the historical data they have on hand will reveal this. 2019 (Coleman SY)

Data is produced every day in significant quantities, dimensions, and values, as has been said. Using the information provided to determine your next decision is crucial. This choice could be applied to future financial planning, including mortgage planning. The choosing of a home is one future for which one should plan. 2019 (Vijay Kotu)

2.4.1 History of Machine Learning

Although the concept of regression, or the act of building a function to describe a set of data points, was not developed until around 1800, machine learning algorithms did not appear until 1952. In 1805, Legendre created and published "the method of least squares" to evaluate how well a function suited a huge set of data. The first effective cost function with a mathematical foundation is developed. This concept was developed by mathematicians and scientists like Gauss and Morkov, who used it to

create formulas for the upcoming a century. The regression was an extremely challenging process, though, as there were no computers (or even calculators) accessible at the time. Everything began to alter in the 1950s with the introduction of the machine learning idea. To execute linear regression, a unique type of calculator was developed. as suggested by the name. Utilizing a linear function to make predictions based on supplied data points is known as linear regression. By minimizing the cost function of linear regression, a best fit linear function may be obtained for nearly any dataset (squared error). However, when it originally debuted, it didn't seem to be that helpful.

The new era of data science began with the first use of the term "Big data" in 2005. Many ideas can now be fulfilled, including decision tree regression. Additionally, numerous novel regression techniques built on time-tested concepts, like random forest regression and gradient boost regression, have been created, enabling machine learning accuracy to rise year after year. Although the complexity and

precision of each approach differs, the housing price prediction problem can be solved using gradient boosting regression, random forest regression, linear regression, and polynomial regression. (Songyi Bai,2022)

2.5 Related work

A house is typically viewed as a heterogeneous good with a variety of utility-relevant characteristics. Numerous studies have looked at the relationship between home price and home attributes. A house's price can be thought of as a quantitative representation of a group of these features. (Jie cao et al, 2022)

For example:

(Guangliang Gao et al., 2022) tested three MTL-based algorithms for predicting house prices under each task description and explored two categories of task definitions, single-profile one and multiple-profile one. They analysed experimental data and found that the task definitions chosen for the house price forecast can greatly influence forecast performance, as well as the selection of MTL-based approaches. They also noted that the performance of MTL-L1 and MTL-L21 is closer to MTL-Graph, that the overall performance is better based on numerous profiles, and that combining two profiles to construct tasks ensures lower RMSE and MAE values.

They confirm, based on their analysis, that task definitions and method selections have a stronger influence on prediction performance than method selections alone. Additionally, we see that different characteristics have varying effects on home prices. In order to increase the applicability of the proposed strategy, they advised exploring profiles that are closely tied to time and have an impact on home prices. They also recommended using multi-task feature learning models.

With the help of four fundamental kernels and two distinct sets of input variables, (Danh Phan, 2019) constructed an SVM model. The best parameters are obtained by performing tuned functions. The greatest prediction outcomes come from regression trees and polynomial regression, but neural networks appear to struggle with this dataset. Despite having an over-fitting problem, PCA and tailored SVM give a comparatively greater level of accuracy. According to their investigation, SVM was slower than regression trees and neural networks, although they required more training time. Their study looked for practical models for predicting home prices. For further deployment, they advised combining the Step-wise and SVM models.

In their study using the TensorFlow framework, (Feng Wang et al, 2019) proposed a deep learning method. The model was trained using the Adam optimizer, and the trend in anticipated housing prices largely matches the reality.

In their study, they used the ARIMA model, a well-known time series forecasting technique, to forecast the trend of home price. A support vector regression (SVR) model was used to carry out the comparison experiments in order to assess the proposed approach. The experimental findings demonstrate a performance contrast between the two models. Both the suggested model and the SVR model accurately predicted the training data, but the SVR model has a significant difference between some data's predicted value and actual value. This research proposes a deep learning-based ARIMA model for predicting home prices..

Using photos of houses as the dataset, (Gan srirutchataboon et al, 2021) employed CNN to extract features from the images. To increase the accuracy of price in Thailand, ensemble machine learning models containing XGBoost, Random forest, and AdaBoost were utilised. Their analysis revealed that XGBoost performed better than other distinct methods.

To fit and predict the prices, (Zhen pen et al.2020) used XGboost, decision trees, and multiple linear regression models. According to the study, XGboost performed better than other models with a 0.9251 score.

The Artificial Neural Network (ANN), XGBoost, Random Forest, Linear Regression, and Support Vector Machine models were compared by (Yong Piao et al. in 2019). SVM did worse with 59% accuracy, whereas ANN did better with 87% accuracy.

Additionally, (Yegi Feng et al., 2022) employed random forests for prediction and contrasted the outcomes with models like decision trees, linear regression, and support vector regressor and discovered that random forests outperformed them all by 81%.

(Zeng Peng et al., 2019) used the XGboost algorithm, linear regression, and decision tree to predict the sale of second-hand homes in China. They showed that xgboost had the highest prediction accuracy of 0.85 and can strengthen a model and prevent overfitting in comparison to decision tree and linear regression. The model is only applicable to pre-owned homes.

For the purpose of projecting housing prices, (Madhuri et al., 2019) employed multiple regression approaches such gradient boosting and adaptive boosting. Their suggested approach might be used by sellers to calculate cost of sales and by purchasers to get information. Gradient boosting was found to have performed better on the data in the comparison study, and it was suggested that more research be done on resale property price prediction using a different method.

To create an ensemble model for housing pricing in Turkey, (Temur et al., 2019) merged the autoregressive integrated moving average (ARIMA) model and the long short term memory (LSTM) model. Their hybrid model was shown to be superior to the other individual models utilised in comparison. They sought to forecast the volume of sales that Turkey would experience the next year. They suggested that

local cities or regions be studied in addition to projecting sales throughout the entire nation.

As CNN is typically used for image processing, (Choujun Zhan et al, 2022) utilised CNN to predict house prices in Taiwan. They did this to see if CNN could also be used to predict time series data. They recommended using LSTM, GRU, and RNN for additional research.

Python libraries were used by all of these studies to analyse and examine their individual data sets.

3. METHODOLOGY

This research follows a qualitative approach because the output of this research deals with numbers that is house price and also how much the age of building affects or influences house price.

3.1 Methods Used

In order to achieve the objectives of the research, various methods were used at every phase of the implementation cycle.

Below are the various methods used to carry out activities at various stages in this research.

3.1.1 Literature Review Approach

A thorough review of literature was carried out to understand the past research others have done. A systematic review was done in which PRISMA methods was used. These enabled me to gather literatures of past research to ascertain the hypothesis that factors that influences house price varies with the main one being location, type of house and structure of the property.

However other literatures have studied other influencing factors that affect house price like economy, location etc this research also studied if the age of the property affect price. Prisma method was used because it assists in keeping track of what needs to be done to acquire relevant materials through the prisma checklist.

3.1.2 Data Collection

Data was collected from UK land registry through web scrapping. The data contains property description with market value on every record. The data contains 10 features and It contains 524022 instances spanning through 4 regions which are Birmingham, London, Manchester and Southampton with history data of 5 years.

3.1.3 Data Preparation

A bit of pre-processing was done to prepare the data for analysis. Python libraries like pandas and datetime were used for the following

Property type features category: property type feature was reduced by removing the 'O' which stands for others. It does not significantly influence the outcome of the prediction. It was also dropped because others as a property type can't really be defined.

Date of Transfer: year and month were extracted from 'Date_of_Transfer' to aid further analysis.

3.1.4 Data Analysis

In order to see the details of what the data entails analysis was carried out. Visualization using python libraries such as seaborn and matplotlib was used to see the data via histogram, bar chart, pie chart, line graph and box plot. Python libraries were used because of the variety of code that could be tweaked to get different angle to visualize.

3.1.5 Data Pre-processing

All categorical variable were encoded using on hot encoding method. one hot encoding was chosen because it allows rescaling to be easy.

Data scaling

Data pre processing includes a stage called scaling, which applies independent variables or data properties. In essence, it aids in normalizing the data within a specific range. It occasionally aids in accelerating an algorithm's calculations. There are different ways of scaling which are standardization, normalization and MinMaxScaler. For this research, MinmaxScaler was used.

MinMaxScaler keeps the original distribution's shape. The information present in the original data is not materially altered. Keep in mind that the MinMaxScaler does not lessen the significance of outliers. The MinMaxScaler's feature has a default range of 0 to 1.

3.1.6 Modelling

Various models were used and compared with each other to pick the best that best predict house price. The below models were used:

XGboost

Extreme Gradient Boosting is the full name of the technique known as XGboost, which is used by numerous regression and classification trees. One could think of the XGboost algorithm as a variation on the GBDT method. The inclusion of regularization terms lowers the model's risk of overfitting and increases model accuracy based on the GBDT (Gradient Boosting Decision Tree) algorithm. The algorithm's goal is to develop trees continuously by adding new trees and changing existing features. Additionally, each time a tree is added, a new function is really learned to suit the residual of the previous prediction. The score of a sample is projected when training is complete. The sample's attributes will determine which leaf node in each tree it will land on. Each leaf node correlates to a score, and the projected value of the sample is simply required to equal the matching score for each tree (Manasa J et al, 2020)

Decision Tree

DT is a popular non-parametric supervised learning technique for regression and classification. Its objective is to develop a model that uses straightforward decision rules to infer the value of target variables from data attributes. It is possible to think of the decision tree as a set of if-then rules. The decision tree is converted into if-then rules using the following procedure: A guideline is created along each leaf node's path from the decision tree's root node; The characteristics of the

internal nodes on the path match the c whereas the classes of the leaf nodes correspond to the rule's conclusions, its variations. One crucial characteristic of the decision tree's route is that it is both complete and mutually exclusive. Every situation is covered by a guideline or a path (Zhen Peng et al, 2019).

Linear Regression

A linear combination of independent variables is used to estimate a continuous dependent variable in the simplest regression model, known as linear regression. Although the model is too straightforward to adequately reflect the intricacy of the London housing market, linear regression contains many core ideas that many other regression models are built upon.

Ridge Regression

The analysis of [MR]multiple regression on multicollinear data is done using the Ridge Regression tool. Multicollinearity (mcl) is the presence of nearly linear connections among variables that are independent of one another. Ridge regression places a unique kind of condition on the parameters.

Random Forest (RF)

RF designates a classifier that trains and predicts data using multiple trees. A RF is, to put it simply, made up of many decision trees. The training set that each tree uses is sampled from the entire training set, which means that some samples from the complete training set may appear in a tree's training set more than once or never at all. The characteristics that are utilized to train the nodes of each tree are chosen at random from all features according to a predetermined ratio without replacement.

3.2 Hyperparameter Tunning

Hyparameter is used in the models to aid optimal result

For example; hyperparameters in random forest are the variables and thresholds used to split each node during training.

Scikit-Learn uses a set of default hyperparameters for all models but these are not guaranteed to be the optimal for a problem.

A form of a trial-and-error approach is used to get the best parameters using grid search.

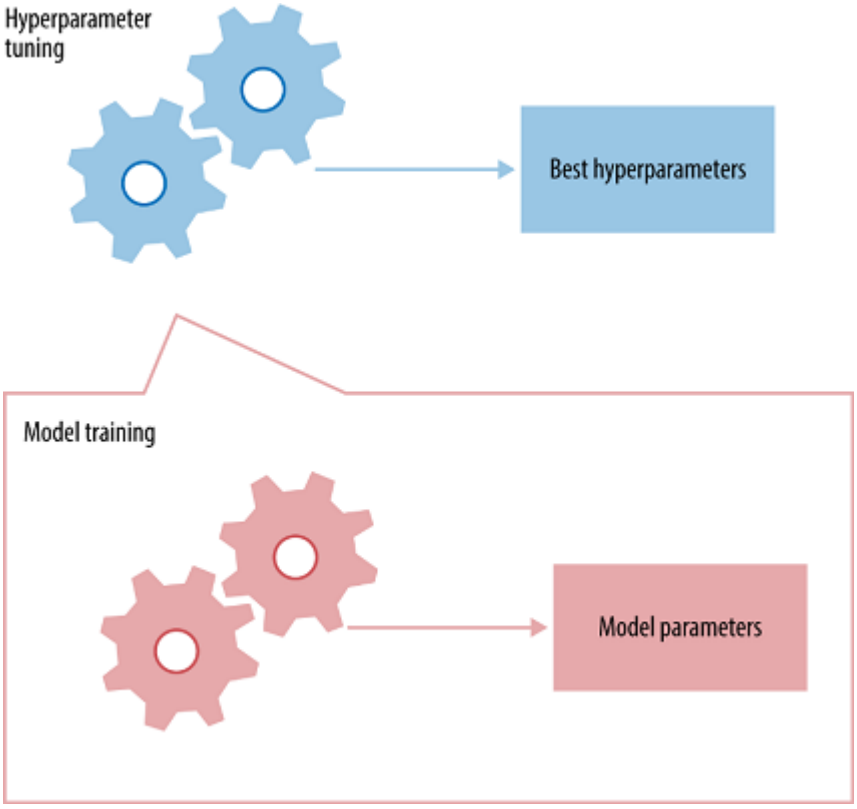


Figure 1: Hyperparameters and Parameters (Adapted from'Towards Data Science')

The best hyperparameter is known from evaluating each model on the testing data. Grid search was used to determine the best parameters to use for tunning.

3.3 Programming used

Python

Python is a high-level programming language for programming that has a wide range of applications. In 1991, Guido Van Rossum produced it, and it became available. It makes it possible to programme clearly at both local and big scales. Python supports a number of programming standards, including as object-arrayed, practical, and procedural. Python is a language that is simple to read. While other programming languages employ punctuation, it uses English keywords. Python uses whitespace rather than wavy sections to separate clauses. Python was primarily created to make reading programmes simple. Python supports a wide range of libraries, including pandas, numpy, scipy, matplotlib, etc., as well as a wide range of packages, including Xlsx Writer. Python is a very useful language for web development and software innovation. It is frequently used to create web apps. It might very well be used to browse and edit documents. It may very easily be utilised to conduct sophisticated scientific procedures. Python has become a highly well-known language since it may make progress at different points. Python code can be composed and then executed. Python is a very important language since it allows for programme updates without requiring additional time and effort. Python supports numerous functional frameworks.(Mansi Jain et al,2020)

Evaluation metrics

In this research, some evaluation methods were utilized including R^2, MAE, MSE, and RMSE, detailed explained as following:

R^2: It is a measurement that is defined as the percentage of the dependent variable's volatility that can be predicted from the independent variable.

MAE: In this case, the actual and expected values of y are two different continuous variables.

MSE: Instead of utilizing the absolute value, the mean square error squares each difference before adding them all up.

RMSE: The mean square error is the arithmetic square root of MSE.

4. IMPLEMENTATION

To implement the objectives, Jupyter notebook IDE is used in this project. It is an open-source application that enables us to exchange and create documents with python code, visuals and narrative text. It comprises tools for data preparation, data transformation, numerical value simulation, model creation using statistics, data visualization, and machine learning. The following were the processes used to achieve the objectives.

.

4.1 Data Collection and Understanding

The data was collected through web scrapping. An actual official data from HM land registry of the data.gov.uk would be used to make the predictions which is also a real life data. Figure 2 shows what the data looks like.

	ID	Price	Date_of_Transfer	Postcode	Property_	Old/new	Duration	PAON	SAON	Street	Locality	Town/Cit	District	County	Category_	Record_Status
2	{8A78B2B0-5	50000	30/04/2019 00:00	M14 5PU	F	N	L	10-Dec	FLAT 6	CURZON A	VENUE	MANCHES	MANCHES	GREATER I	B	A
3	{8A78B2B0-5	140000	16/05/2019 00:00	M24 1DF	O	N	F	111		MILL FOLD	MIDDLETC	MANCHES	ROCHDALI	GREATER I	B	A
4	{8A78B2B0-5	90000	14/05/2019 00:00	M22 9PX	S	N	F	19		CHESHAM	AVENUE	MANCHES	MANCHES	GREATER I	B	A
5	{8A78B2B0-5	83000	16/04/2019 00:00	M38 9RA	S	N	F	45		BARON FC	LITTLE HU	MANCHES	SALFORD	GREATER I	B	A
6	{8A78B2B0-6	90000	12/04/2019 00:00	M8 0RG	T	N	F	8		NASMYTH	STREET	MANCHES	MANCHES	GREATER I	B	A
7	{8A78B2B0-6	112500	26/04/2019 00:00	M43 6DZ	T	N	L	18		GORSEYFI	DROYLSDE	MANCHES	TAMESIDE	GREATER I	B	A
8	{8A78B2B0-6	370000	09/04/2019 00:00	M14 6PA	T	N	F	118		MOSELEY	FALLOWFI	MANCHES	MANCHES	GREATER I	B	A
9	{8A78B2B0-6	69000	22/03/2019 00:00	M27 6AD	T	N	F	8		WESLEY ST	SWINTON	MANCHES	SALFORD	GREATER I	B	A
10	{8A78B2B0-6	100000	24/05/2019 00:00	M11 1AQ	T	N	F	59		CHEADLE	STREET	MANCHES	MANCHES	GREATER I	B	A
11	{8A78B2B0-6	850000	08/05/2019 00:00	M13 0DX	O	N	F	LINCOLN H	FLAT 4	LINCOLN	GROVE	MANCHES	MANCHES	GREATER I	B	A
12	{8A78B2B0-6	90000	08/02/2019 00:00	M40 2JN	T	N	L	49		OLD CHURCH	STREET	MANCHES	MANCHES	GREATER I	B	A
13	{8A78B2B0-6	304000	18/04/2019 00:00	M22 4BU	D	N	F	4		SHAWDENE	ROAD	MANCHES	MANCHES	GREATER I	B	A
14	{8A78B2B0-6	1170000	16/05/2019 00:00	NW6 1EB	S	N	F	5		AGAMEMNON	ROAD	LONDON	CAMDEN	GREATER I	B	A
15	{8A78B2B0-6	55620	25/04/2019 00:00	SW5 0JJ	O	N	F	13		BRAMHAM	GARDENS	LONDON	KENSINGT	GREATER I	B	A
16	{8A78B2B0-6	310000	29/03/2019 00:00	SE18 2EL	T	N	F	113		KIRKHAM	STREET	LONDON	GREENWI	GREATER I	B	A
17	{8A78B2B0-6	1028770	14/03/2019 00:00	NW1 0QT	O	N	F	42		ST PANCRAS	WAY	LONDON	CAMDEN	GREATER I	B	A
18	{8CAC1318-A	1100000	07/06/2019 00:00	SW19 8BN	T	N	F	30		ASHEN	GROVE	LONDON	MERTON	GREATER I	A	A
19	{8CAC1318-A	905000	20/06/2019 00:00	SW19 8BH	T	N	F	35		DURNSFORD	AVENU	LONDON	MERTON	GREATER I	A	A

Figure 2: Initial Dataset

The data contains property description with market value on every record. The data contains 10 columns/variables/features. There are no missing or null values in the

data. It contains 524022 instances spanning through 4 regions which are Birmingham, London, Manchester and Southampton with history data of 5 years. The available features are:

Unique identifier: a unique automatic generated number at every recorded sale

Price: price sales

Property type: D Detached, S= Semi-Detached, T=Terraced, F- Flats/Maisonettes, O= Other.

Old/New: Y= newly built property, N = an established residential building

Duration: F= freehold, L=leasehold

Town/City

District

County

Category Type: Indicates the type of price paid. A = Standard Price paid (full market value), B = additional price paid (repossessions, buy-to-lets).

Record Status: A- Addition, C = Change, D= delete.

4.1.1 Variable data types

The data types of the features are found in the below table:

Table 1: Data Types

Features	Data type
Price	Int
Property type Old/New	Categorical

Duration Town/city District County Category Type	

Independent and Non-independent variable:

Price – Dependent/target variable

While features: Property type, old/New, duration, Town/city, District, County and Category type are dependent variable.

4.2 Data Preparation

A bit of pre-processing was done to prepare the data for analysis. The data was prepared to showcase the analysis deeper. The features prepared are:

Property type features category: property type feature was reduced by removing the 'O' which stands for others. This was dropped because it made 0.4% of the total records and it does not significantly influence the outcome of the prediction. It was also dropped because others as a property type can't really be defined. Flats, terraced, detached and semi-detached was used.

Date of Transfer: Year and month were extracted from 'Date_of_Transfer' and kept on a separate column to do time series and show the time trends on distribution of prices

Also there was no missing values. Figure 3 confirms that

```
empty_data = df[df.isna().any(axis=1)]
print(empty_data)
```

```
Empty DataFrame
Columns: [Price, Date_of_transfer, Property_type, old_new, duration, Town_city, District, County, category_type]
Index: []
```

```
#object_feat = lon_data.dtypes[lon_data.dtypes =='O'].index.values
```

Figure 3: Empty variables

4.3 Data Analysis

In order to understand the data better and know the distribution of price over the different locations including the various relationship that exist among the variables, various univariate and multivariate visualization was carried out. Python libraries such as seaborn and matplotlib was used in carrying out the visualizations and it was carried out using histogram, bar chart, pie chart, line graph, box plot.

These libraries were used because they consist of a lot of attribute that makes it easier to tweak in order to get desired result. These libraries were also used for statistical analysis because the libraries automatically generate the statistics without manually calculating. Starting with visualizing the four cities as a whole.

Displaying the effect of price over the four cities

Figure 4 below shows the distribution of price over the four cities.

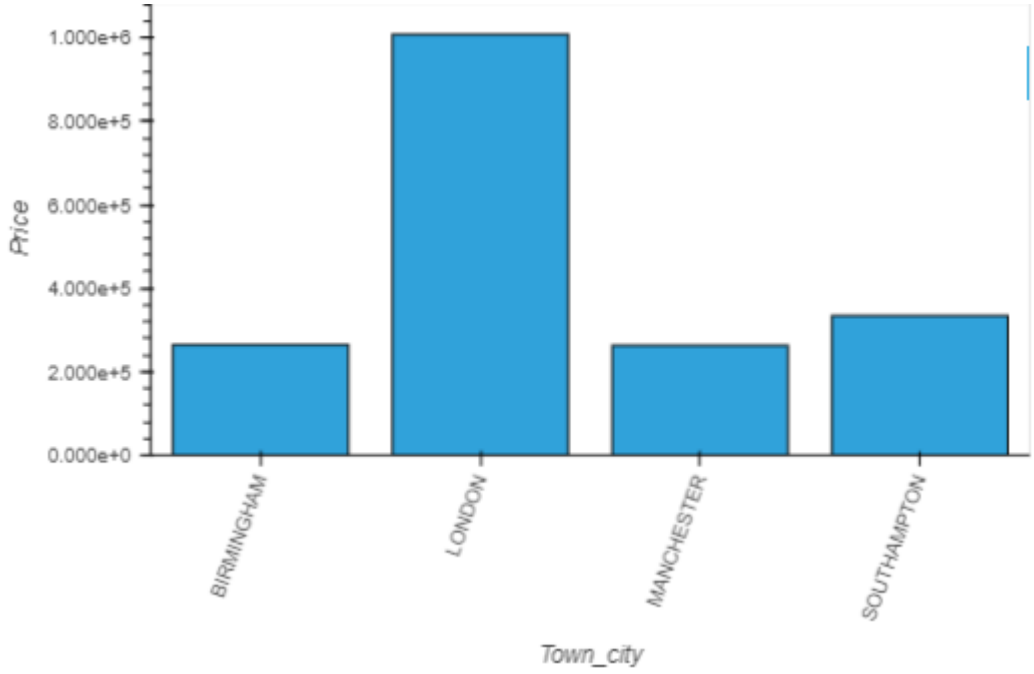


Figure 4: Price distribution over the cities

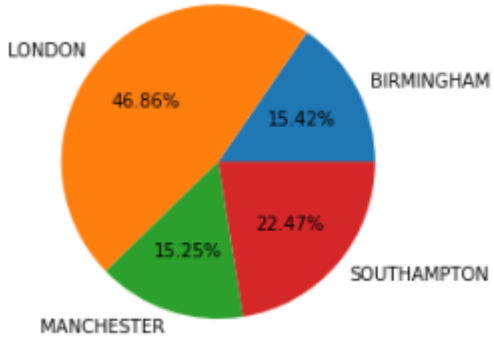


Figure 5: Price pie chart of the cities

Both figure 4 and 5 confirms London has the highest cities that has the highest house price and highest investment. This is also confirmed on (the telegraph of 24th august,2022) saying London is the most expensive market. The below line graph in figure 6 shows the distribution of price over the years from 2017, it could be seen

that London has been known to have the highest price over the years which rose in 2020 and fell in 2021 followed by Southampton with the second highest and has been a constant all through the years. Birmingham and Manchester are on the same level with the distribution of price from the graph

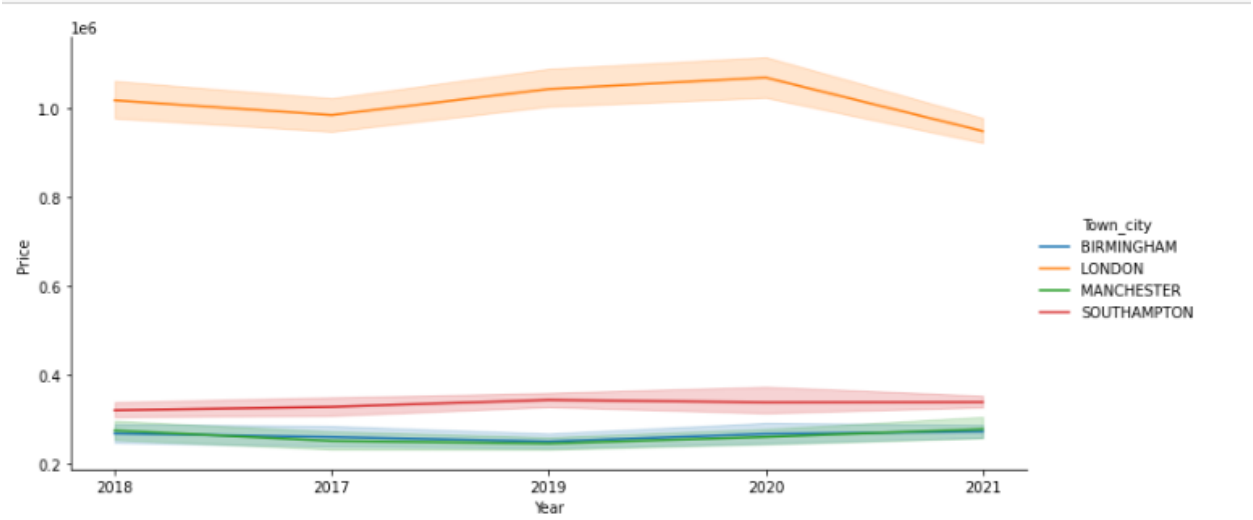


Figure 6: Line graph showing price over the cities

```
Town_city: 0 nulls, 4 unique vals, most common: {'LONDON': 327899, 'MANCHESTER': 85394}
District: 0 nulls, 69 unique vals, most common: {'BIRMINGHAM': 63665, 'MANCHESTER': 38536}
County: 0 nulls, 12 unique vals, most common: {'GREATER LONDON': 327802, 'GREATER MANCHESTER': 85393}
```

Figure 7: Showing the cities with highest sales

In figure 6, even though Southampton is the second city with a higher house price, Manchester is the second city that has many sales across the years according the house count in figure 7. It shows Manchester a second city in District, County and City.

The figure 8 below shows the price distribution over the entire city, the graph shows the data is right skewed meaning it does not follow a normal distribution. There would be need to normalize the data.

Figure 8: Price distribution of sales

Figure 9 below shows property type price popularity across the cities. Flats is the least expensive

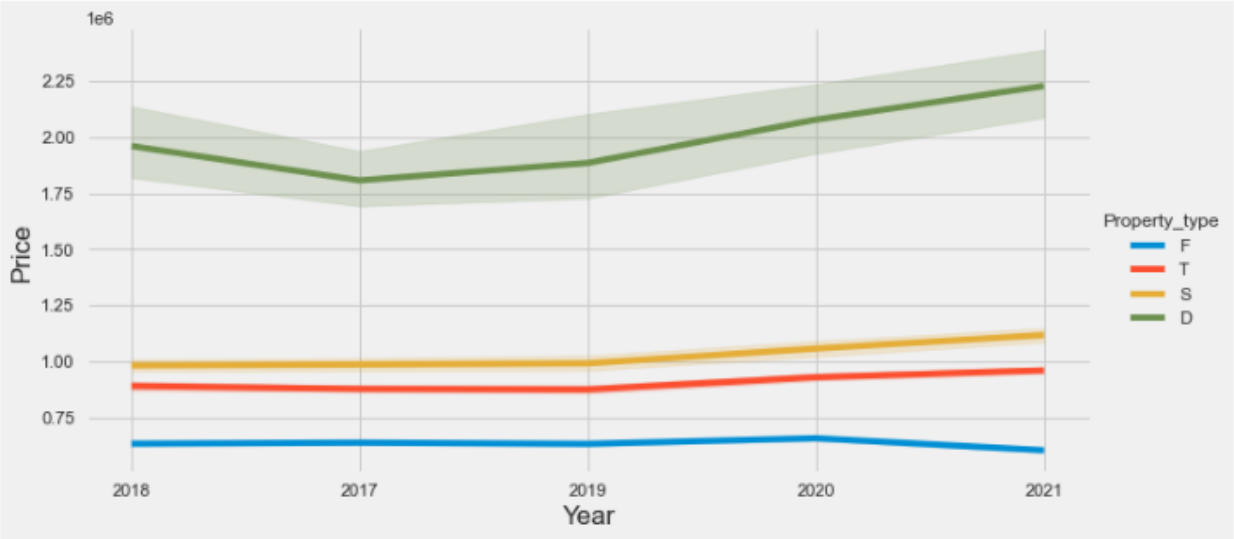


Figure 9: Price distribution over the Property Type

Figure 10 shows the relationship between price and the age of property either N(Established property) or Y(Newly built). From the graph established property is in demand more than newly built due to the obvious reason that newly built house are more expensive which is also seen below

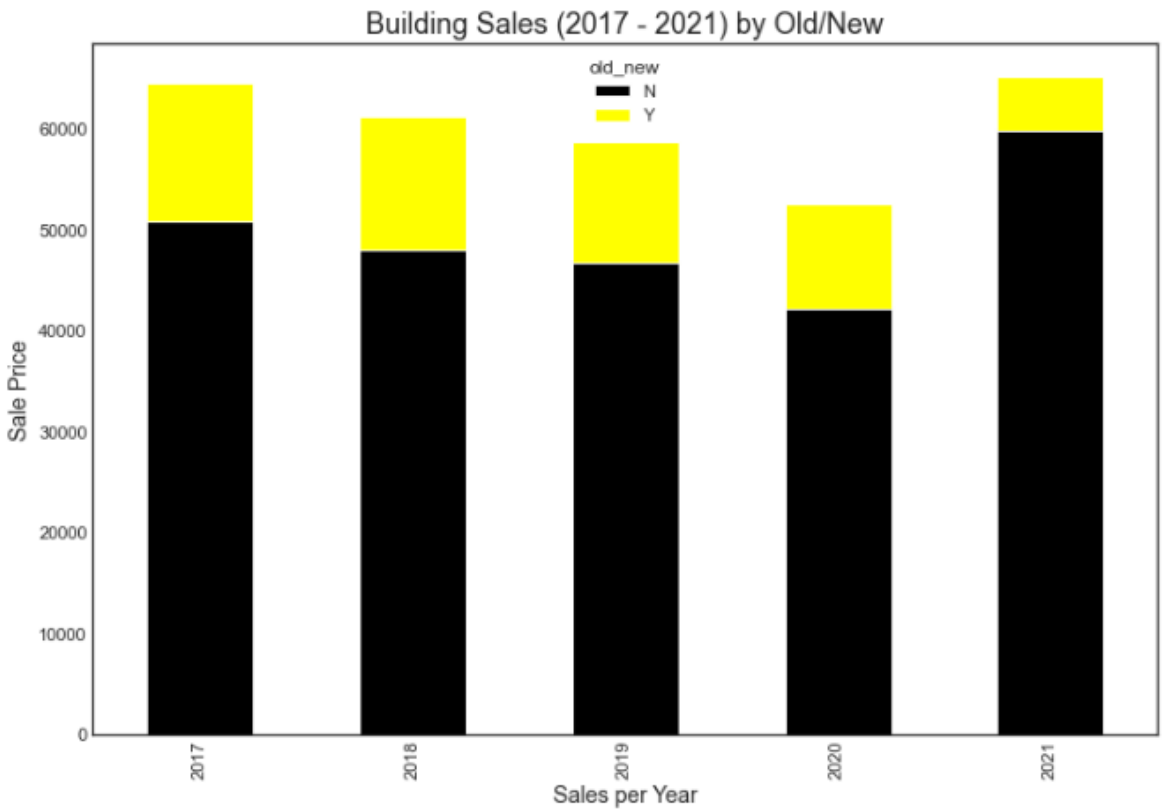


Figure 10: Sales per year on property type

The below also shows average sale per year across the cities. The price was more expensive in 2019 as seen below

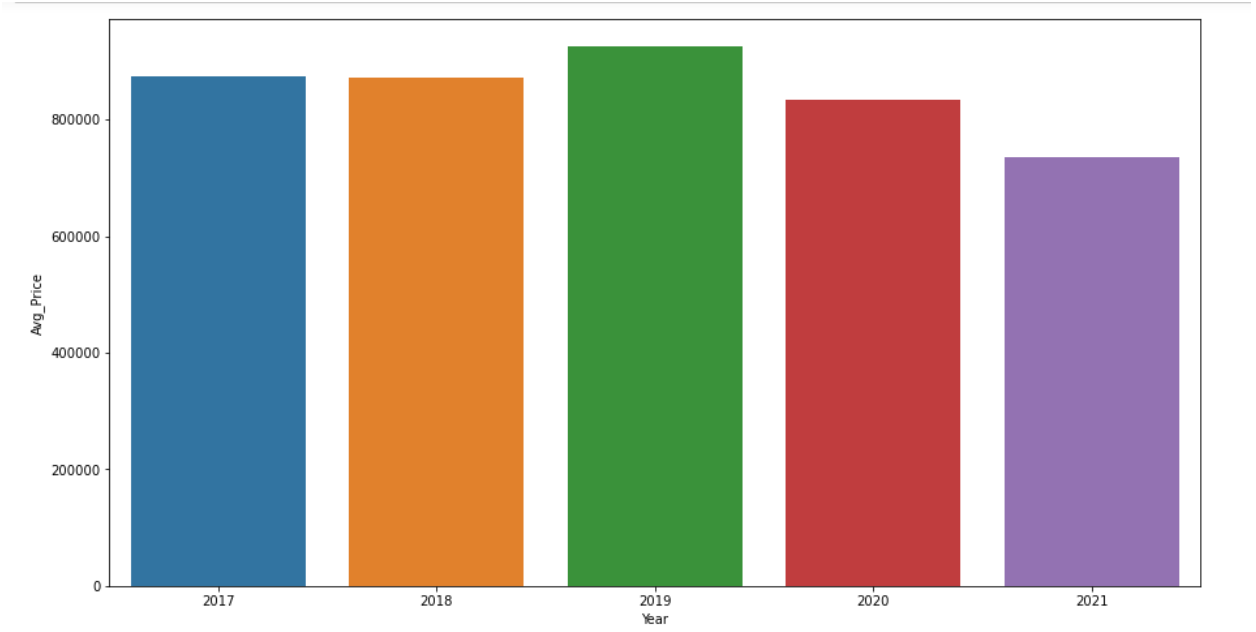


Figure 11: Average house price over the years

Old/New

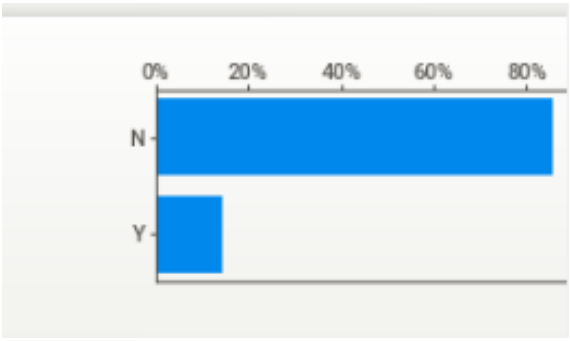


Figure 12: Number of house sold according to the house's age

Generally speaking, figure 12 above shows that established buildings are more in demand in the four cities than newly built house. As the analysis is done deeper the reason will be known.

Duration:

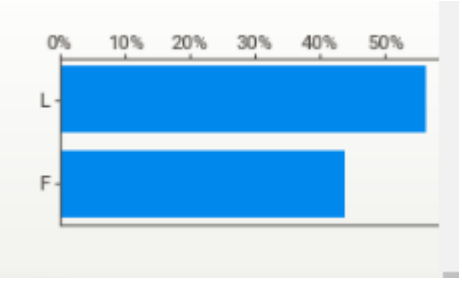


Figure 13: Number of house sale according to duration

In the duration graph above in figure 13, the demand for leasehold property are more than freehold. Leasehold property are properties that only the build is bought while freehold property are properties that both the lad and the building is bought.

County

Prices across the four city counties

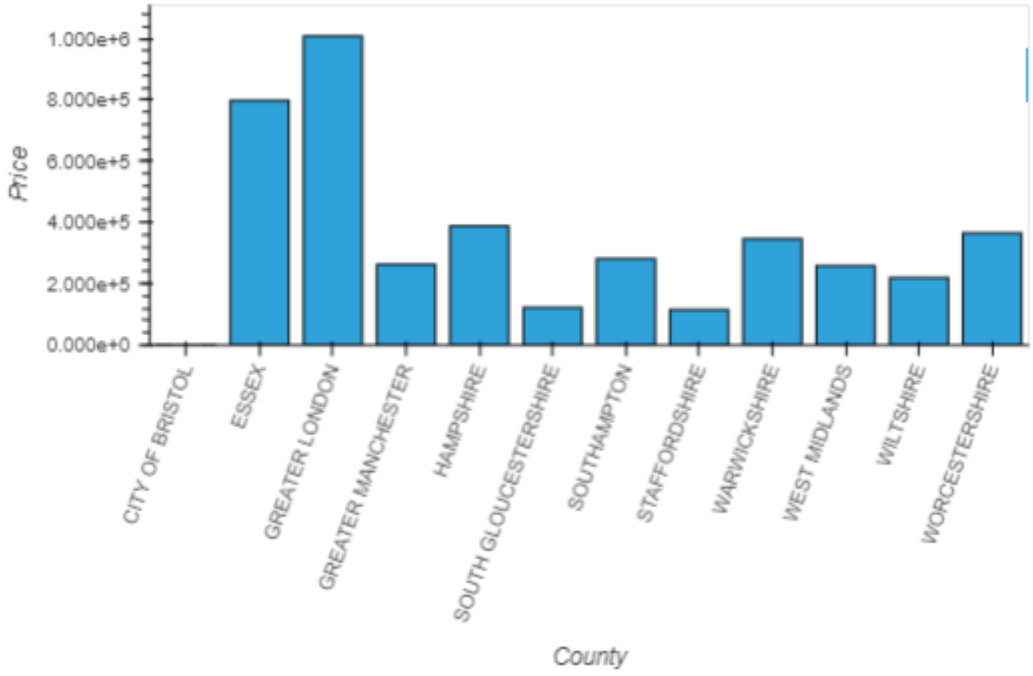


Figure 14: Price distribution across the counties

Figure 14 shows counties with high price across the four cities. The first 2 are from London followed by Hampshire and Southampton which is in Southampton.

In order to further see what was happening in each city, the cities were further analysed individually to better understand the regions.

Starting with London, as it has been ascertained from previous visualization that London has the highest number of house sales. Diving further into London data, figure 7 below shows the that through the number of years that was researched, flat property type was sold more, followed by terraced, followed by semi-detached and then detached.

The sale of terraced, semi-detached and detached was constant all through the years except 2021 with a slight rise in the sales of terraced. The sale of flat went down in 2020 probably due to covid 19 pandemic.

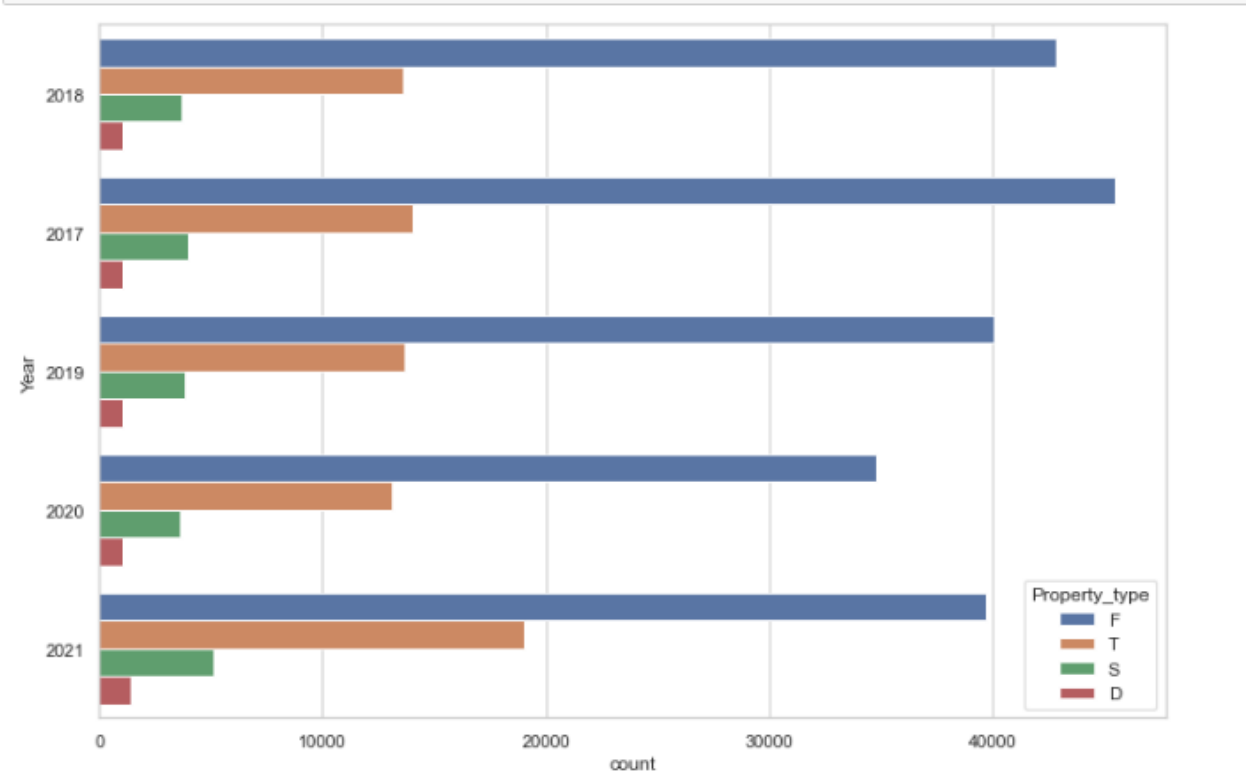


Figure 15: Number of property type sold through the years

From the figure above, one thing certain is that there has been demand for flat in London than any other type of property. Reason being that there are mor workers moving into London demanding smaller houses. Even though there was a fall in demand for flat during the pandemic, the demand has risen again. (Metro property news).

The below shows the prices of old or new over property types across the cities

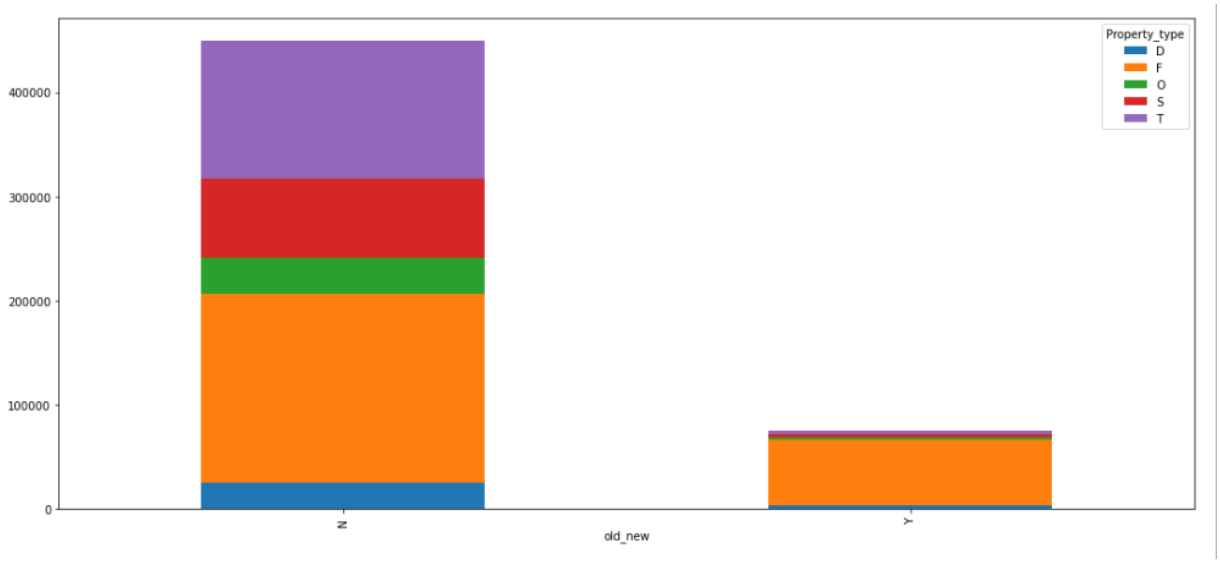


Figure 16: Price of old/new across the property type

The diagram shows more of old building and flat in high demand across the cities.

Visualizing the number of sales

The below analysis is to know and see the number of sales made in each of the features across the cities

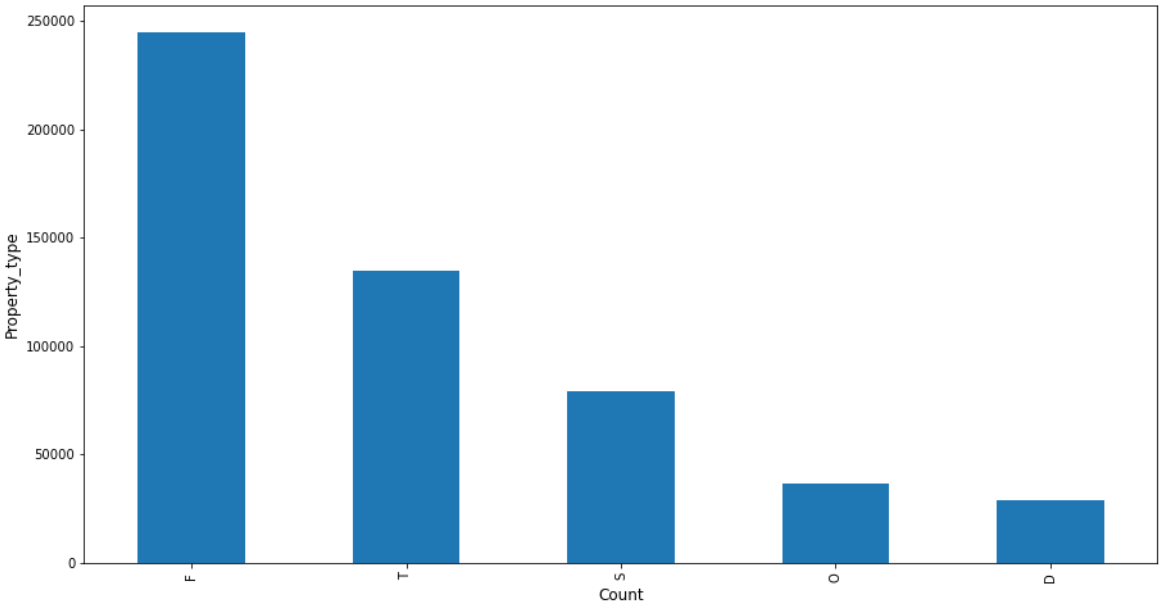


Figure 17: Sale of property type across the cities

In the whole four cities, more flats were sold compared to other property type due to the fact that they come in handy.

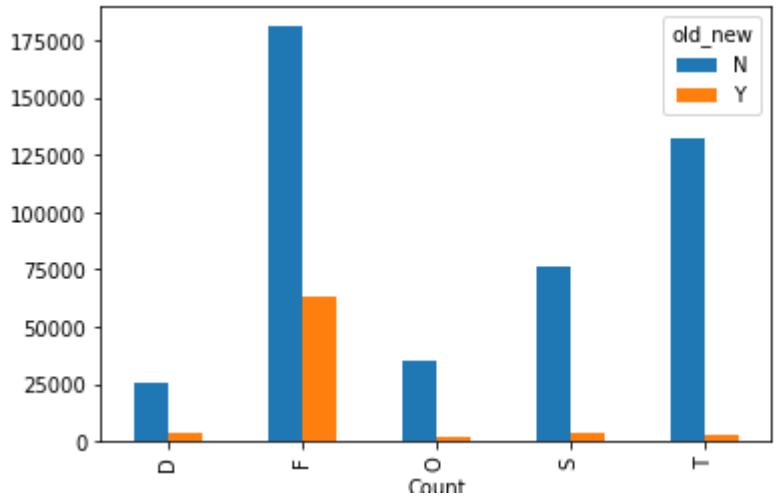


Figure 18: Sale of property type that are either old/New

This is showing the distribution of property type that are either newly built or old. From the diagram, established buildings across the types of property were In demand. This can be linked to figure 12 that shows that newly built home are more expensive than old houses.

Also going further, figure 19 below is showing the number of property type that are either old or new ad also either freehold or lease hold. The diagram below shows that they were more of established buildings (old houses) that are flat and also leasehold across the four cities that made sales throughout the five years followed by established houses that are terraced and leasehold.

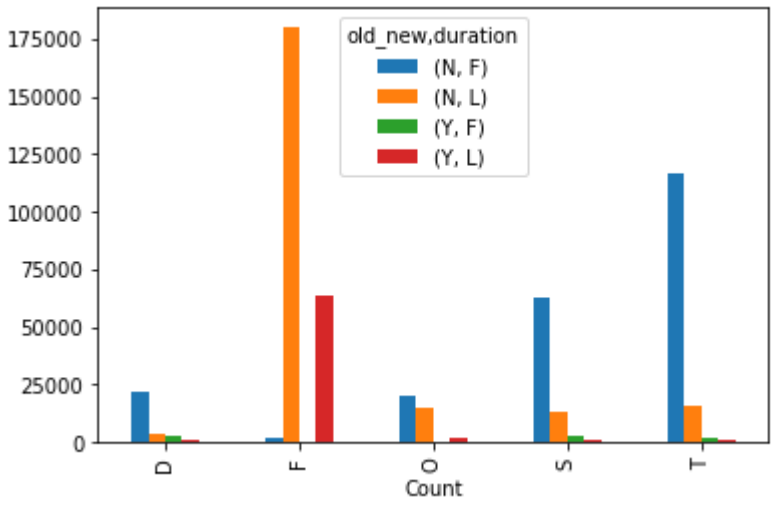


Figure 19: sales count of house count in demand.

The figure 20 and figure 21 below shows the number of sales in each county and city respectively. Though Southampton show the second most expensive city from figure 4, Manchester is the second city to have the highest sales

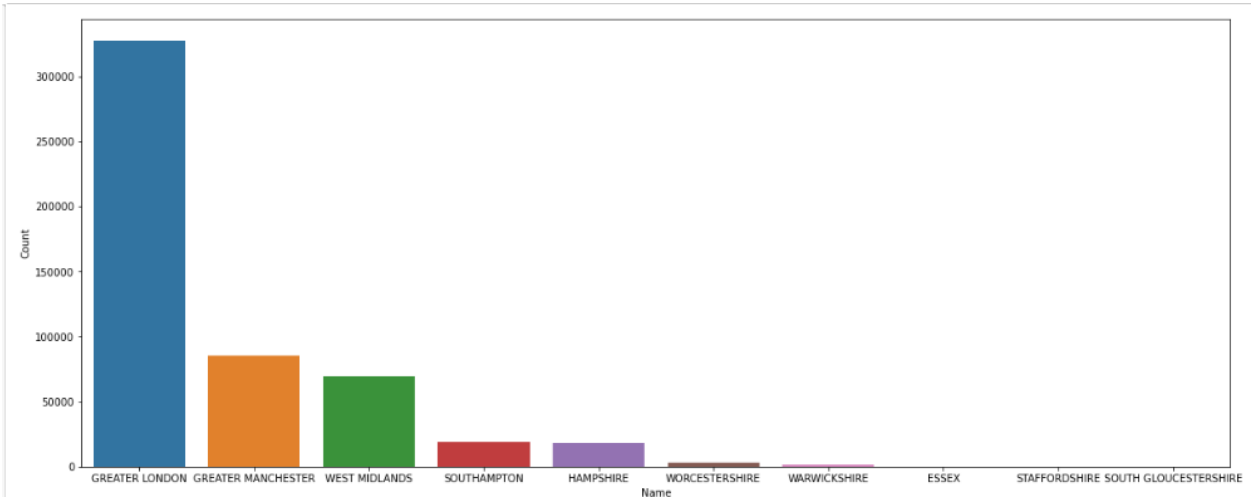


Figure 20: county count of house sale

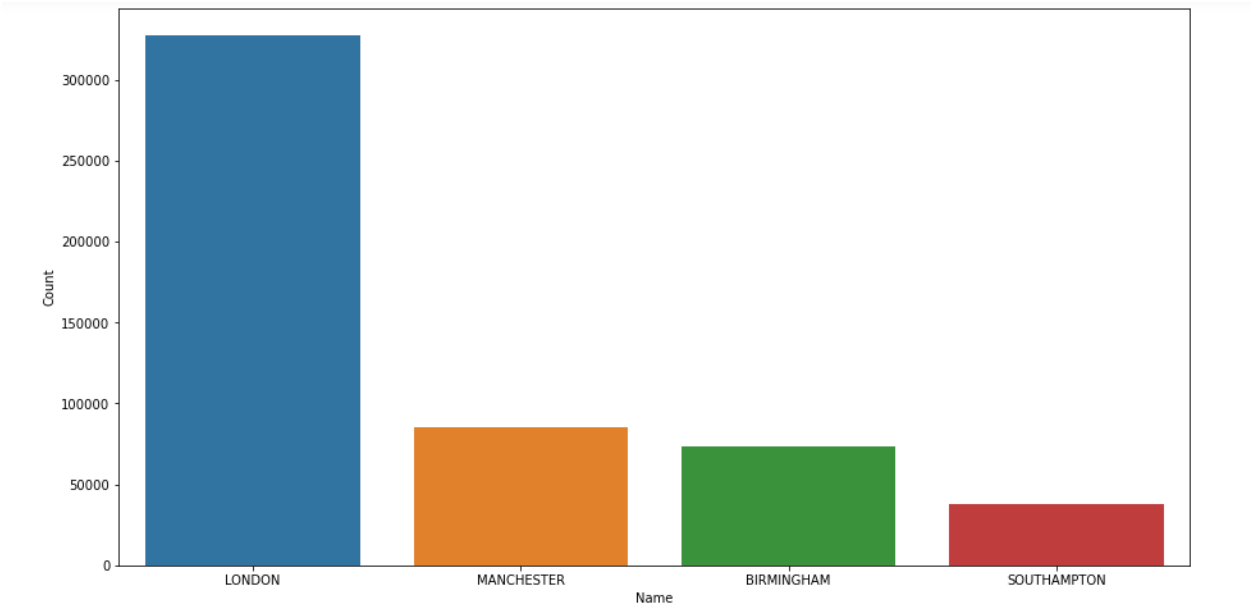


Figure 21: city count of house sale

Taking Southampton as a case study for analysis:

The below figure shows the number of old or new buildings that are either freehold or leasehold that was bought over the years.

It shows high demand of old buildings that are freehold. This is different from figure 13 where the general analysis was done on the four cities that showed leasehold in higher demand. For Southampton freehold buildings were highly demanded, this can be the reason why Southampton was the second city with the highest house price since more of the buildings sold were freehold.

Freehold means the occupant owns the building and the land while leasehold mean the occupants owns only the building.

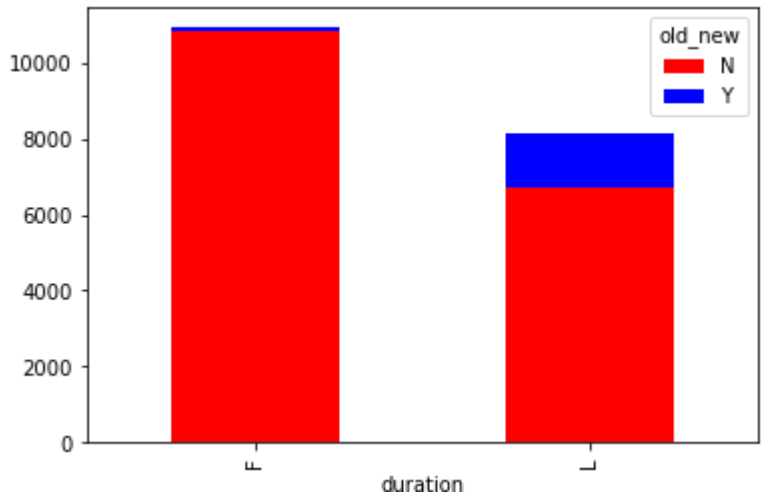


Figure 22: Sale count for Duration.

The below figure shows the sales count for property type. Analysis shows that they have been averagely a good demand for the types of property across the years except for 2020 which might likely be due to the pandemic.

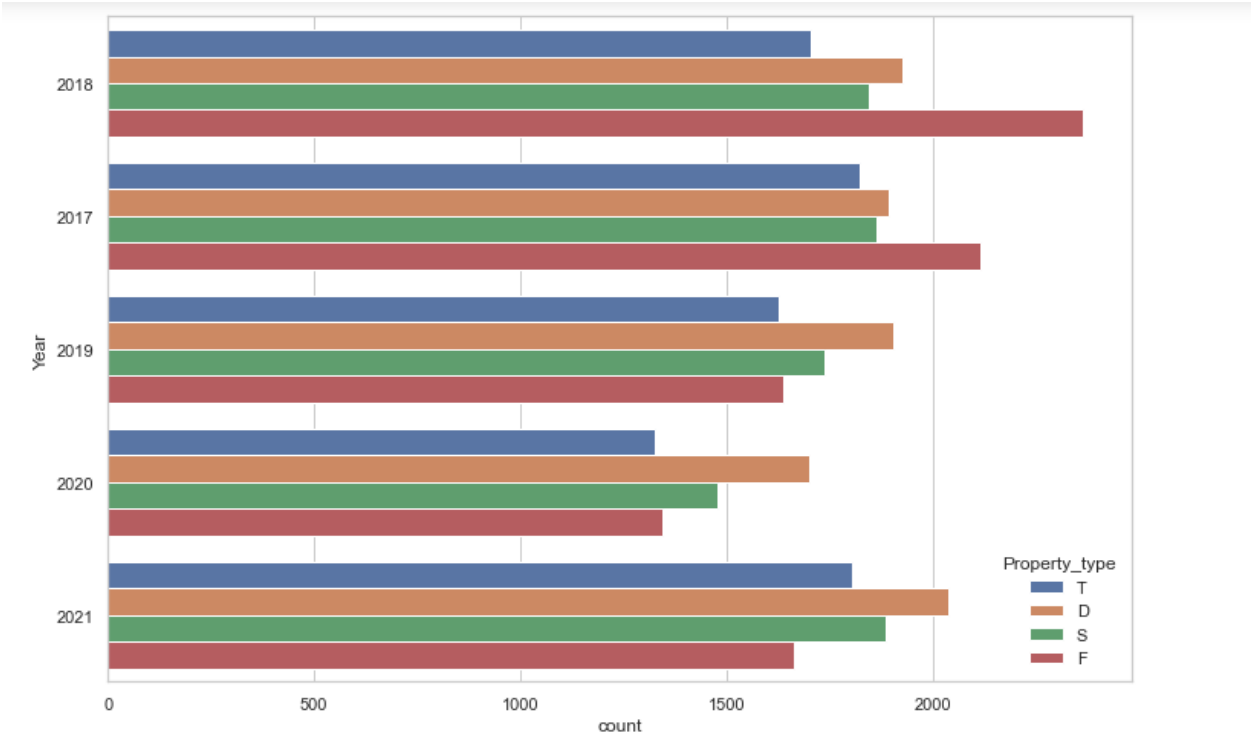


Figure 23: sales count for property type across the years in Southampton

Southampton average price sale has been rising from 2018 in the figure below

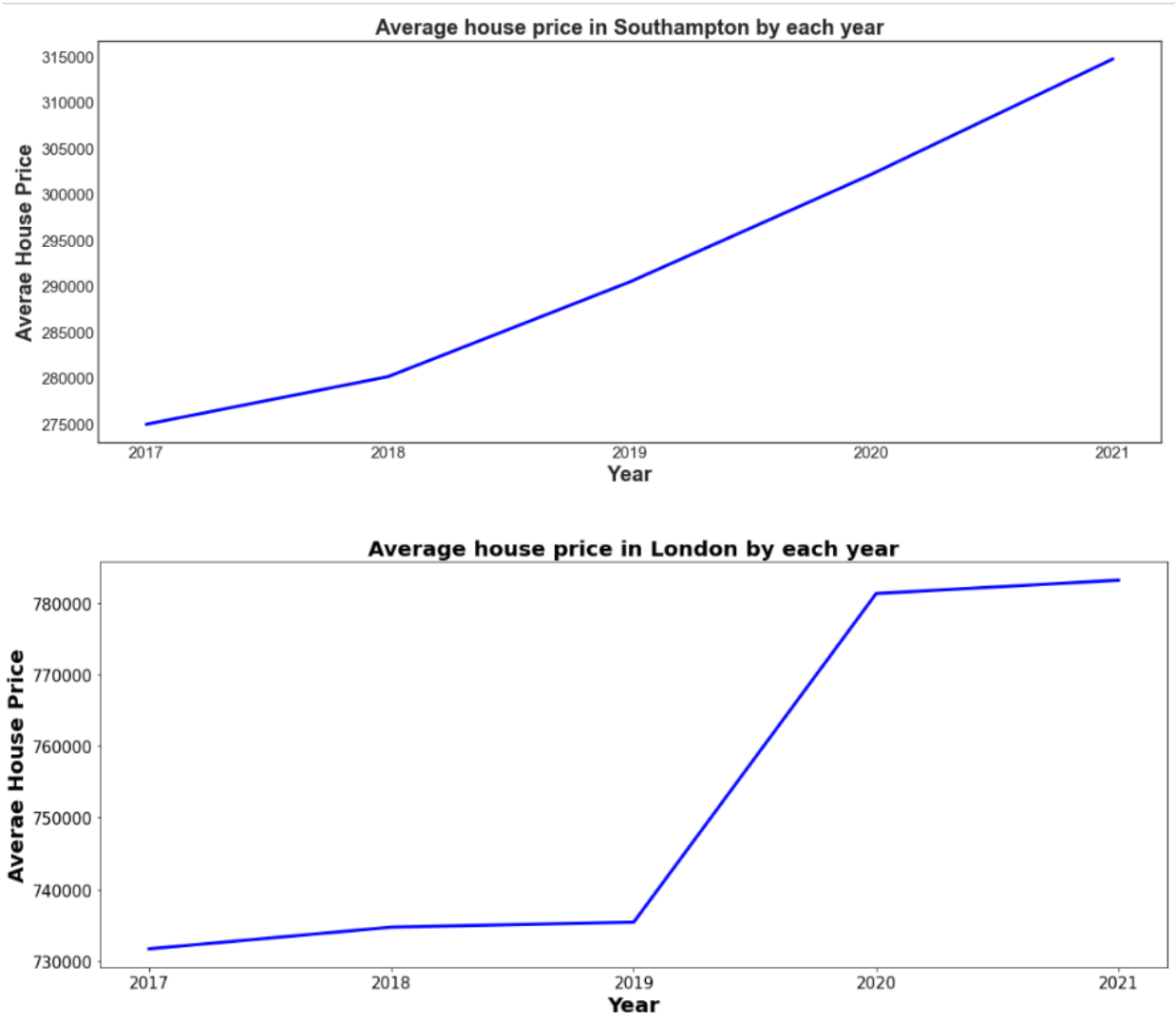


Figure 24: Average price in London versus Southampton

Looking at the graph on figure, it is seen that house price was strong from 2017 t0 2019 but between 2019 and 2020 the price skyrocketed. This is confirmed from (Yopa property UK) that the growth rate was 2% each year until 2020, there was an increase of 7.4% as pent up buyer demand from covid induced lockdowns was released into the market ad it increased further in 2021 by 10.8%.

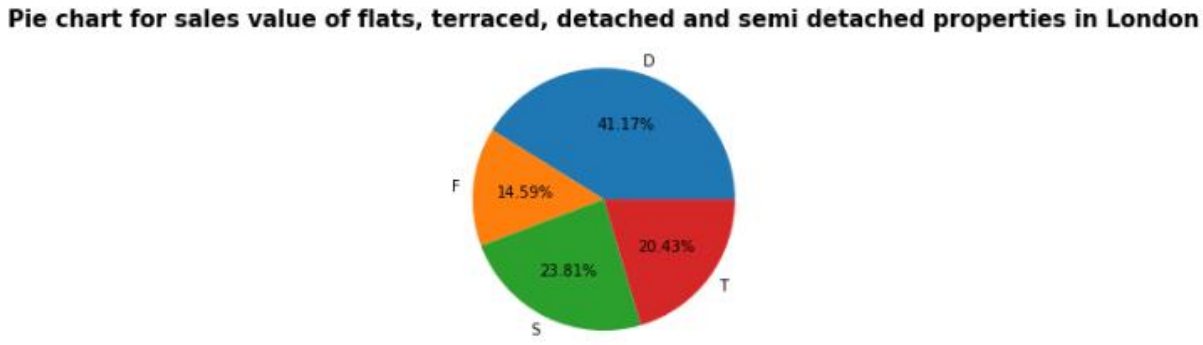


Figure 25: Property type average price ranking London

Figure shows detached building are more expensive than other type of property. This is due to the fact that detached building are spacious and not attached to any other property. (The Mirror news uk), confirmed that even though Flats are more in demand, detached building are more expensive due to the space they have. The line graph below also confirms the prices of the property types.

Like its observed in London's property type, detached apartment are also the most expensive in Southampton as seen in figure 26 below

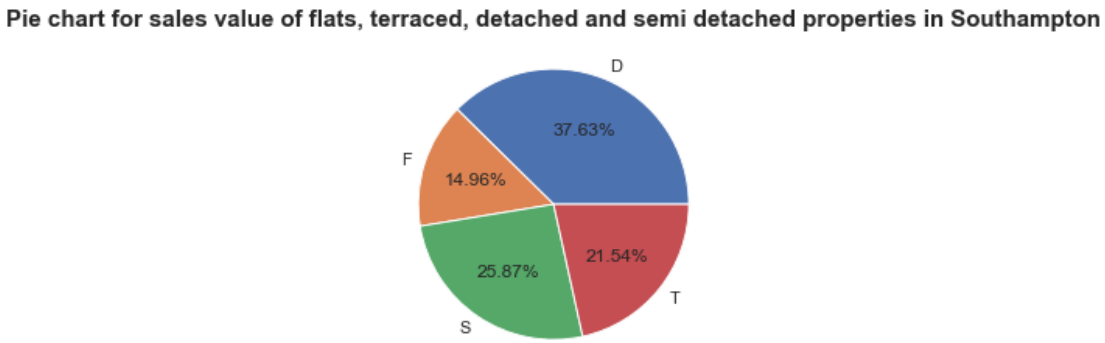


Figure 26: Property type average price ranking Southampton

This confirms the price for flat has always being the lowest price of the four.

4.3.1 Financial Analysis

The below figure shows the statistical analysis of the 'Price' column.

	Price
count	5.244190e+05
mean	7.345539e+05
std	4.216840e+06
min	1.000000e+00
25%	2.050000e+05
50%	3.850000e+05
75%	6.250000e+05
max	5.943000e+08

Figure 27: Statistical Summary on Price

Skewness

The data skewness needs to be corrected next. The skewness of the data can affect the outcome of Linear regression. In order to test for the normal distribution of the price, skewness was done and figure 28 below shows the data is positively skewed.

```
hp["Price"].skew()
```

```
33.05297394123192
```

Figure 28: Skewness

In this case, data skewness is 33.05. The positive skew means that the median is closer to the bottom of the box. It is far more effective to employ the log

transformation because positive skew is a concern. Additionally, log transformation will improve linear relationships.

4.4 Data Pre-processing

Preparing the data would be carried out to make the data correct and suitable for the model to work with.

Because of how huge the data is and five cities has been chosen to explore which are Birmingham, London, Manchester and Southampton over the period of five years i.e. from 2017-2021. Southampton was extracted out to make the major city to be explored. The data is cleaned to extract these cities over the period from the whole data file. After which few columns will to be dropped because they are irrelevant to the analysis which are: unique identifier, duration and record status (they are all 'A').

4.4.1 Columns were dropped

Below are the columns that were dropped because of their less impact on the data

Unique identifier

Post code

House number

Date of transfer

Record status: this was dropped because the values all 'A'.

```
df.drop(columns=['Unique_identifier', 'Postcode', 'Paon', 'Saon', 'Street', 'Locality', 'record_status'], axis=1, inplace=True)
```

Figure 29: Dropped columns

4.4.2 Removing Outliers

The outliers are in those point far away from others in figure 31 and also the outliers are all values outside the box boundaries in figure 30, they are not in the range 10.2 to 13.5. The outliers can be either replaced with the median for numeric values or mode for categorical values or dropped. Since the 'Price' feature is numerical, the outliers are replaced with the median value.

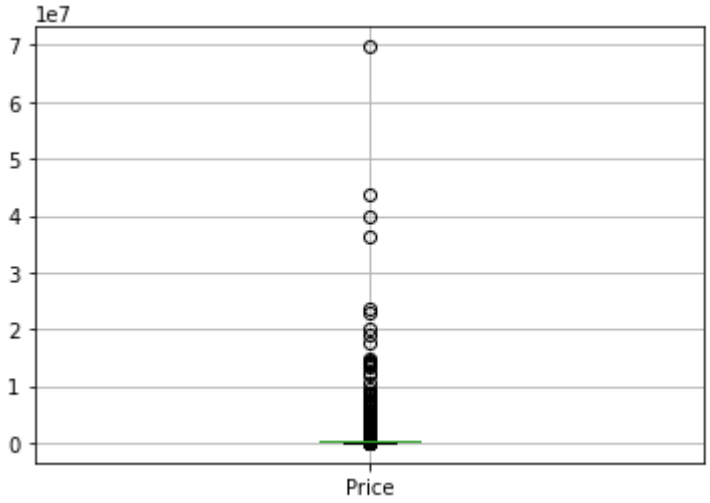


Figure 30: Boxplot for price showing the outliers

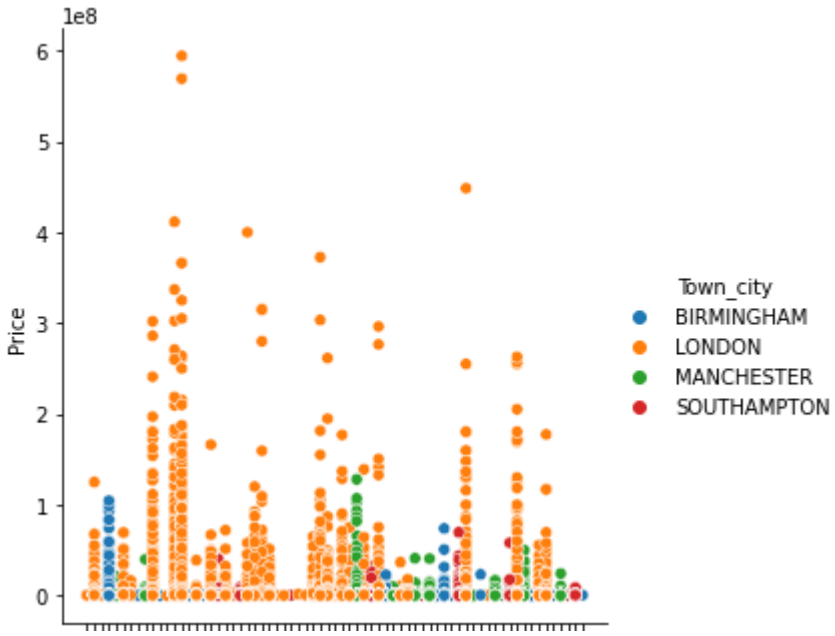


Figure 31: Scatter plot showing the price outliers across the cities

4.4.3 Feature Encoding

Another process of preprocesing done is to encode the categorical variables in the data. Encoding categorical variable is necessary because machine learning model only accept numerical variables. The variables are converted to numbers such that the model can understand and extract valuable information.

One hot encoding was done on the categorical variables except Town, County and district. One hot encoding was used because it has been known to make training data more useful and expressive. It is also known to allow rescaling easy.

Label encoding was used for Town, District and County was label encoding is used for data.one hot encoding was not used for these three variables so as to reduce high dimensionality.

The below figure shows the data after encoding

	Property_type	District	County	dur_F	dur_L	new_N	new_Y	cat_A	cat_B
38	2	3	1	1	0	1	0	1	0
39	4	3	1	1	0	1	0	1	0
382	1	0	0	0	1	1	0	1	0
564	2	3	1	0	1	1	0	1	0
565	3	3	1	1	0	1	0	1	0
...	...	...	...	...	...	...	...	...	...
104732	2	3	1	1	0	1	0	1	0
104733	3	2	0	1	0	1	0	1	0
104734	3	4	0	1	0	1	0	1	0
104735	3	0	0	1	0	1	0	1	0
104736	1	2	0	0	1	1	0	1	0

Figure 32: Encoded Data

4.4.4 Scaling

Because data comes in different scale, the data is normalized to move them into the same scale, standardizing and scaling techniques in the python library.

The figure below shows the data after scaling

```
Scaled data:
 [[0.6 1.  1.  0.  1.  0.  1.  0.  0.  0.  0.  1. ]
 [0.6 1.  1.  0.  1.  0.  1.  0.  1.  0.  0.  0. ]
 [0.  0.  0.  1.  1.  0.  1.  0.  0.  1.  0.  0. ]
 [0.6 1.  0.  1.  1.  0.  1.  0.  0.  0.  0.  1. ]
 [0.6 1.  1.  0.  1.  0.  1.  0.  0.  0.  1.  0. ]
 [0.6 1.  1.  0.  1.  0.  1.  0.  0.  0.  0.  1. ]
 [0.4 0.  1.  0.  1.  0.  1.  0.  1.  0.  0.  0. ]
 [0.4 0.  1.  0.  1.  0.  1.  0.  0.  0.  0.  1. ]
 [0.6 1.  1.  0.  1.  0.  1.  0.  1.  0.  0.  0. ]
 [0.6 1.  1.  0.  1.  0.  1.  0.  0.  0.  0.  1. ]]
```

Figure 33: Scaled Data

4.4.5 Feature Importance

After data pre-processing, factors that affect property prices in the UK data used can be revealed from the four cities analysed. The relations between variables are represented in figure via heatmap visualization and Random Forest classifier library.

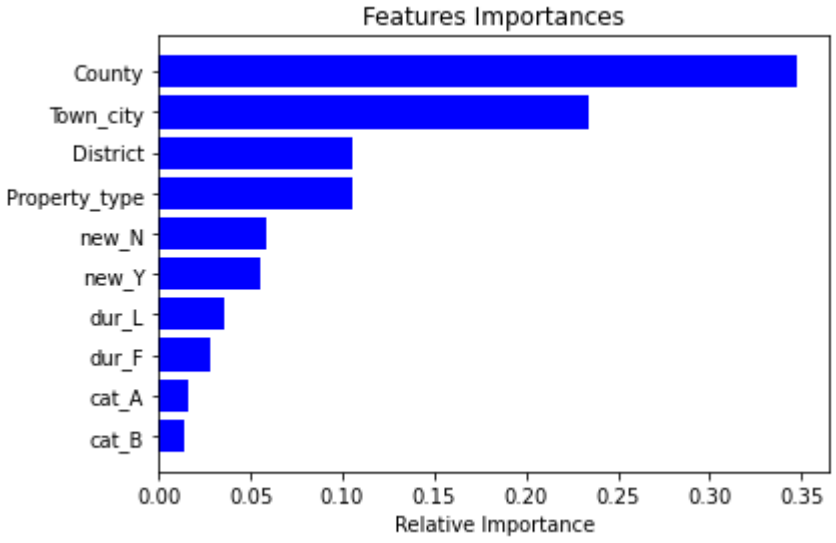


Figure 34: Feature Importance

From figure 34 above, County has about 35% of the variable make up that influences the price in UK house data followed by the city then district. This highly confirms the hypothesis from past research that location is an influencer of house price. Property type makes up 13% of price influencer followed by the age of the building. Building age is one of what this research looked into to determine how much does the age of the building contributes to house price. From the analysis of the UK data, figure 34 shows the age of a building represent about 7% of the price. It can be said that the age of a house is only one of the criteria used in estimating price. It actually would not be one of the most important.

4.5 Modelling

The models used was chosen based on the fact that the excepted output of this project is quantitative, the problem to be solved is a regression problem. Price prediction cannot be labelled as a classification problem since the problem isn't classified into two or more classes.

The models used are linear regression, random forest, Decision tree and XGboost

The data was splitted

```
x_train, x_test, y_train, y_test = train_test_split(X, y_data, test_size=0.3, random_state=1)
print(x_train)
```

Figure 35: Data splitted into Train and Test

After which the data was fitted into the models using '. fit' function as seen below

```
#from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestRegressor
model = RandomForestRegressor(n_estimators=100, random_state=2)
#model = RandomForestRegressor()
#grid_search = GridSearchCV(estimator = model, param_grid = param_grid,
 #                          cv = 3, n_jobs = -1, verbose = 2)
model.fit(x_train, y_train)
predi = model.predict(x_test)

print(predi)
```

Figure 36: Model fitting

4.6 Evaluation

For this study, we evaluated model's performance using metrics: the coefficient of determination, R2, RMSE (Root Means Square Error). The result from the evaluation is found in chapter 5

4.7 Software Artefact

Now that our model is built and evaluated, the model is deployed on the website. Streamlit library in python was used. It is an open-source **python** framework for building web apps for Machine Learning and Data Science. Streamlit is easy to use and light weighted. The figure 37 below shows the dashboard on the site. It has to modules on the site.

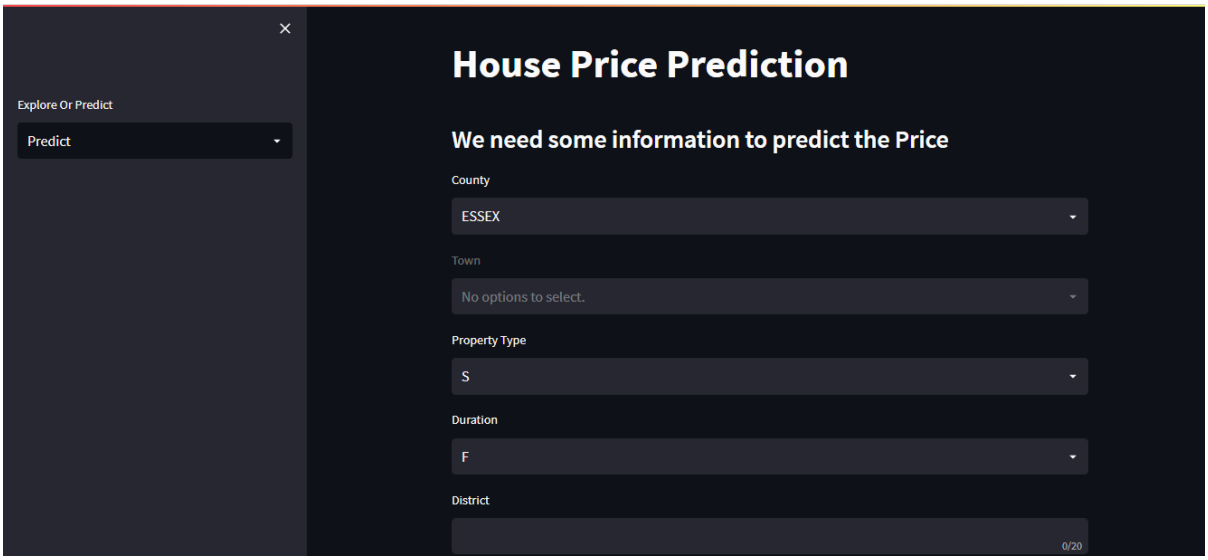


Figure 37: Interface to fill the form

While under the explore tab, there are different graphs to explore showing the trends of house price. Figure 38 below shows a sample of the view

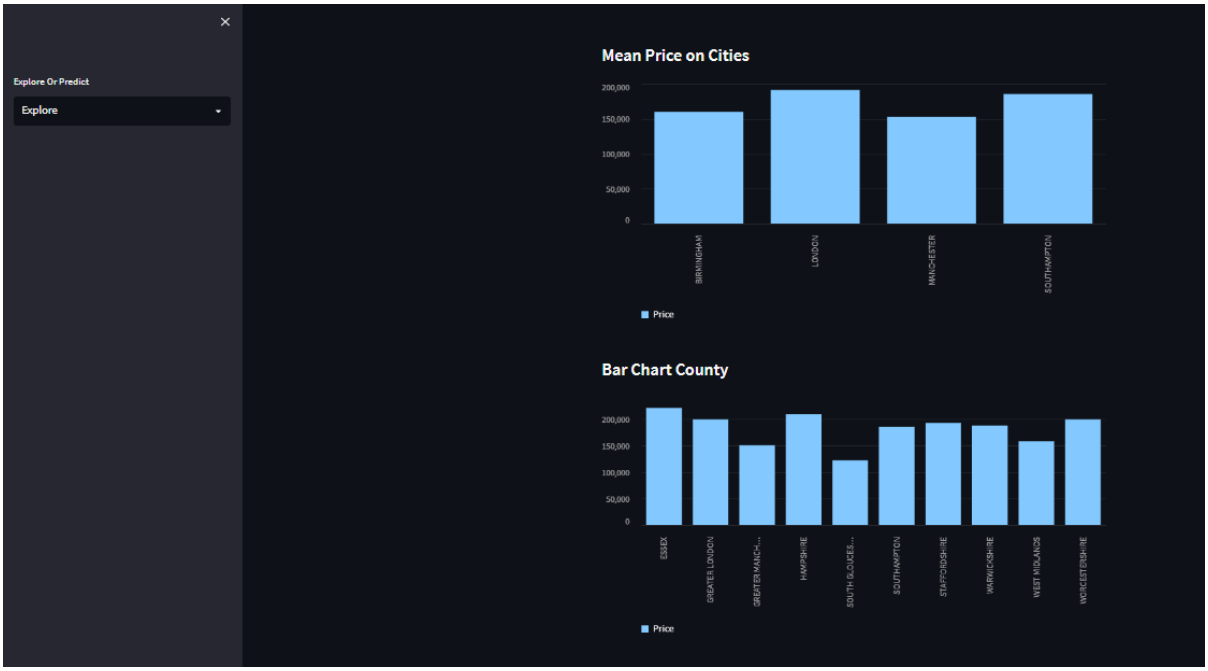


Figure 38: Showing the explore tab

4.8 Technology Choice

The hardware used in implementing this project are:

An 8gb RAM

500GB Hard disk

Windows 10 operating system

Software specifications

Program language: Python language

Website: python streamlit

Browser: Google chrome

5. RESULTS AND DISCUSSION

5.1 Results

After going through the processes in chapter 4, the result gotten from evaluating is stated below in table 2. The results display the mean square error and R square result gotten from the models used.

Table 2: Results Table

Models	Mean Square Error	R^2 Score
Decision Tree	205922.62	0.750
Linear Regression	214278.04	0.704
XGBoost	0.446	0.905
Random Forest	0.435	0.903

5.2 Discussion

This research investigates different models for housing price prediction. Four different types of Machine Learning models used are Random Forest, XGBoost, Decision Tree and Linear regression. these models were compared and analyzed for optimal solutions.

Even though all of the model used achieved some levels of results, different models have their own pros and cons. The Random Forest and XGBoost models had same R2 score but Random Forest has the lowest error but is prone to be overfitting. Its time complexity is high since the dataset has to be fit multiple times.

From table 2 above, it is clear that the XGboost model has the best prediction effect. The core idea of the XGboost algorithm is based on the gradient to promote tree,

which can better adapt to the unbalanced data sets, and training algorithm performance is also extremely high. Additionally, XGboost is less prone to fitting, performs well, and has higher generalization capabilities for many non-linear regression issues. The XGboost algorithm model explains the impact of attributes on the home price better than the conventional linear regression model and decision tree model. Since not all factors affecting home prices are linear, XGboost offers significant advantages when analyzing such unbalanced data.

5.2.1 Benchmarking

The result output from this research was compared with models used by Yeng fang et al. (2022) and Manasa J et al. (2020) in predicting house price.

From Yeng Fang et al.2022, he compared Random Forest, Decision Tree and Linear Regression. He got 0.81, 0.699, 0.66 respectively on R2. This research results out performs all giving a better result at prediction.

The XGBoost models used in this research was also compared with (Manasa J et al. 2020) who used XGBoost model to predict price and also had 0.75 R2 score even after optimizing the result with cross validation.

This is to say the four models used in this research outperforms the accuracy gotten from previous research. The objective has been satisfied by

6. CONCLUSION

Few works have been done on predicting house price in UK, maybe due to the fact that there are few data that capture the happenings in UK. Comparing this research result with previous work of (Yeng Fang et al,2022), there was 10% increase on the accuracy with minimal error rate. The research was able to compare about 4 different models to select the best optimal result.

From the aim, objectives, research question ad problem statement at the beginning of this research, it can be said that the following occurred successfully:

The aim which is to develop a predictive model was done by picking the model with a better R^2 score of the four models compared.

To improve on the accuracy of the predictive models that was used in previous research. The improvement gotten was iterated in Chapter 5.2

This research also agrees with the below hypothesis that state:

- Old properties will have less price than new one
- Terrace and semi-detached property will have more price compared to flat.
- City property prices will be more
- County and District also have impact on property price

All of these hypotheses were confirmed in chapter four which shows London happens to be the capital city having more house sales and higher price. Also, the feature importance which is figure 21 confirms that location and property type have a lot of influence on house price.

Also, from the analysis done in chapter 4, it is confirmed that a lot of features affect price depending on the features that comes with a dataset but there are some basic features that highly influence the price which make up the hypothesis.

Another hypothesis to be added from the analysis done is that the price in a city does not determine the number of sales made. This can be confirmed from figure 4 where Southampton was the second highest price after London but and also the highest number of sales while Manchester had third highest price but second number of sales. It is safe to say aside the general influencer of house price like location and property type, the metropolis of the city matters.

A system that aims to provide an accurate prediction of housing prices has been developed. The system makes optimal use of Linear Regression, Random Forest Regression, Decision Tree Regression and XGBoost. The efficiency of the algorithm has been further increased with use of hyperparameters. The system will satisfy customers by providing accurate output and preventing the risk of investing in the wrong house.

6.1 Limitation

This analysis of this research is limited to four cities which are London, Southampton, Birmingham and Manchester.

The project took a long time to find relevant research dataset. Though the dataset used is an actual dataset gotten from UK land registry, the dataset set lacks some house features that can be analysis in different ways to see their impact on price. The aim was to find a dataset that has a variety of variables and can be analyzed in several ways.

6.2 Future Work

Addition of larger cities to the model could be added, which will allow our users to explore more houses in different region, get more accuracy and thus come to a proper decision.

Also, the dataset can be expanded by adding the features it lacks. A full dataset can be gotten from Rightmove Agency but it is a paid platform because they are a profit-making company. So, if there is time, the researcher can access the features of a property through the postcode and match it with the sales in the land registry.

The accuracy of the system can be improved. Several more cites in England or UK in general can be included in the system. if the size and computational power increases of the system.

Also, a learning system can be created which will gather users feedback and history so that the system can display the most suitable results to the user according to his preferences.

REFERENCES

AALBERS, M.(., 2016. The Financialization of Housing: A Political Economy Approach. *Routledge, London, UK.,*

AYŞE SOY TEMÜR1, MELEK AKGÜN, GÜNAY TEMÜR, 2019. PREDICTING HOUSING SALES IN TURKEY USING ARIMA, LSTM AND HYBRID MODELS. *Journal of Business Economics and Management,* , 920-937

AYUSH VARMA *et al.,* 2020. House Price Prediction Using Machine Learning And Neural Networks.

BLUNT, A AND DOWNING, R, 2006. *Home.* Abingdon: Routledge

CH.RAGA MADHURI, ANURADHA G, M.VANI PUJITHA, 2019. House Price Prediction Using Regression Techniques: A Comparative Study . *International Conference on smart structures and systems ICSSS 2019,*

CHI, B. *et al.,* 2021. Shedding new light on residential property price variation in England: A multi-scale exploration. *Environment and Planning B: Urban Analytics and City Science,* 48(7), 1895-1911

CHOUJUN ZHAN *et al.,* 2020. Housing prices prediction with deep learning: anapplication for the real estate market in Taiwan. , 719-724

CHRISTIAN MÜHLROTH, 2020. Artificial Intelligence as Innovation Accelerator.

COLEMAN, S., 2019. Data science in industry . *20th European Conference on Mathematics for Industry,*

CROUCH, C., 2009. Privatized Keynesianism: an unacknowledged policy regime. *The British Journal of Politics and International Relations,* , 382–399

DANH PHAN, 2018. Housing Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia. *2018 International Conference on Machine Learning and Data Engineering,*

ANON., 2022. Demand for London flats soars as more seek to buy first property. Feb 11,

EPSTEIN, G., 2005. Introduction: financialization and the world economy, *Financialization and the World Economy, Edward Elgar, Cheltenham, UK.,*

EVENING STANDARD, 2022. Demand for London flats soars as more seek to buy first property. "Feb 22,"

FENG WANG, YANG ZOU, HAOYU ZHANG AND HAODONG SHI, 2019. House Price Prediction Approach based on Deep
Learning and ARIMA Model. *2019 IEEE 7th International Conference on Computer Science and Network Technology,*

GAN SRIRUTCHATABOON *et al.,* 2021. Stacking Ensemble Learning for Housing Price Prediction: a Case Study in Thailand. *International Journal of Engineering & Technology,*

GUANGLIANG GAO, ZHIFENG BAO, JIE CAO, A. K. QIN ,TIMOS SELLIS, 2022. Location-Centered House Price Prediction: A Multi-Task
Learning Approach. *ACM Transactions on Intelligent Systems and Technology,*

HAMNETT C AND READES J, 2019. Mind the gap: Implications of overseas investment for regional house price divergence in Britain. *Housing Studies,* , 388-406

J. MANASA, R. GUPTA and N. S. NARAHARI, 2020. Machine Learning based Predicting House Prices using Regression Techniques. *- 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA).* pp.624-630

MADHURI, C.R., G. ANURADHA and M.V. PUJITHA, 2019a. House price prediction using regression techniques: a comparative study. *2019 International conference on smart structures and systems (ICSSS).* IEEE, pp.1-5

MANSI JAIN, HIMANI RAJPUT, NEHA GARG, PRONIKA CHAWLA, 2018. Prediction of House Pricing Using Machine Learning with Python. *International Conference on Electronics and Sustainable Communication Systems (ICESC 2020),* , 570-574

METRO PROPERTY, 2022. Demand for house rise again after pandemic.

MIRROR NEWS, 2022. House prices for this type of property are soaring as demand rises. May 18,

NIVITHA SHREE, R.H. *et al.,* 2022. Price Prediction of House using KNN based Lasso and Ridge Model . *International Conference on Sustainable Computing and Data Communication Systems (ICSCDS-2022),* , 1521-1527

OMAR ALJOHANI, 2021. Developing a stable house price estimator using regression analysis. *ICFNDS 2021,*

S. BAI, 2022. Boston house price prediction: machine learning. *- 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP).* pp.1678-1684

SAIYAM ANAND, PRINCE YADAV, ADARSH GAUR, INDU KASHYAP, 2021. Real Estate Price Prediction Model. IEEE, pp.541-543

THE TELEGRAPH, 2022. London the most expensive market falls. Aug 12,

VAN LOON, J. AND AALBERS, M. B., 2017. How real estate became ‘just another asset class’:the financialization of the investment strategies of Dutch institutional investors,European Planning Studies. , 221–240

VIJAY KOTU, B.D., 2019. *Data Science: Concepts and Practice*. Jonathan Simpson

WINKY K.O. HOA, BO-SIN TANGB AND SIU WAI WONG, 2021. Predicting property prices with machine learning algorithms. *JOURNAL OF PROPERTY RESEARCH,* , 48-70

XINYU YANG, ZESHENG YIN, JIAYI LI, 2021. Housing Price Mathematical Prediction Method through Big Data Analysis and Improved Linear Regression Model. pp.751-754

Y. CHEN, R. XUE and Y. ZHANG, 2021a. House price prediction based on machine learning and deep learning methods. - *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)*. pp.699-702

Y. CHEN, R. XUE and Y. ZHANG, 2021b. House price prediction based on machine learning and deep learning methods. - *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)*. pp.699-702

Y. FANG, T. LI and H. ZHAO, 2022. Random Forest Model for the House Price Forecasting. - *2022 14th International Conference on Computer Research and Development (ICCRD)*. pp.140-143

Y. NI, 2022. A housing price prediction method based on neural network. - *2022 International Conference on Big Data, Information and Computer Network (BDICN)*. pp.592-595

YONG PIAO, ANSHENG CHEN, ZHENDONG SHANG, 2019. House Price Prediction Based on CNN. *9th International Conference on Information Science and Technology (ICIST),*

YONGQIONG ZHU, 2020. Stock price prediction using the RNN model. *Journal of Physics Conference Series,*

YOPA PROPERTY, 2022. Increase rate of house price.

ZHENPENG, Q., Yincheng Han, 2019. Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboostAlgorithm. *International Conference on Advanced Infocomm Technology,* , 168-172

APPENDIX A-ETHICS

Ethics form

Project status

Status

Approved

Actions

Date	Who	Action	Comments
10:20:00 29 June 2022	Drishty Sobnath	Supervisor approved	
02:06:00 29 June 2022	Ruth Babatunde	Principal investigator submitted	

Ethics release checklist (ERC)

Project details

Project name: House Price Predictive System

Principal investigator: Ruth Babatunde

Faculty: Faculty of Business, Law and Digital Technologies

Level: Postgraduate

Course: Applied AI and Data science

Unit code: COM726

Supervisor name: Drishty Sobnath

Supervisor search:

Other investigators: None

APPENDIX B- CODE SNIPPEX

Importing Libraries

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.graph_objs as go

from datetime import datetime
#from pandarallel import pandarallel
#from imblearn.over_sampling import SMOTE
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import PolynomialFeatures
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import KFold
from sklearn.pipeline import Pipeline
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import OneHotEncoder
#from sklearn.ensemble import IsolationForest
from sklearn.metrics import mean_absolute_error
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
#from sklearn.ensemble import IsolationForest
from sklearn.ensemble import RandomForestClassifier
from plotly.offline import iplot
```

APPENDIX C-Data Loading

```
df = pd.read_csv('house_price.csv')
df
```

	Unique_identifier	Price	Date_of_transfer	Postcode	Property_type	old_new	duration	Paon	Saon	Street	Locality	Town_city
0	{80E1AA98-09F7-7BF8-E053-6C04A8C00BF2}	160000	2018-12-14	M19 2WD	S	N	L	18	NaN	WILSTHORPE CLOSE	NaN	MANCHESTER
1	{7C2D0701-0B0C-4963-E053-6B04A8C07B97}	100000	2018-10-19	M30 0PX	T	N	L	24	NaN	LEWIS STREET	ECCLES	MANCHESTER
2	{7C2D0701-0B0E-4963-E053-6B04A8C07B97}	206000	2018-10-29	M9 4HR	O	N	L	163	NaN	MOSTON LANE	NaN	MANCHESTER
3	{7C2D0701-0B10-4963-E053-6B04A8C07B97}	700000	2018-09-18	M34 3RU	O	N	F	FLAIR FLOORING SUPPLIES LTD	NaN	GREY STREET	DENTON	MANCHESTER
4	{7C2D0701-0B14-4963-E053-6B04A8C07B97}	73000	2018-11-21	M18 8NR	S	N	F	63	NaN	PINNINGTON ROAD	NaN	MANCHESTER
...	...	...	...	...	...	...	...	...	...	...	...	...
524414	{D707E536-0CD1-0AD9-E053-6B04A8C067CC}	679000	2021-08-02	E16 2PY	T	N	L	COMMODORE HOUSE, 2	38	ADMIRALTY AVENUE	NaN	LONDON
524415	{D707E536-0CD2-0AD9-E053-6B04A8C067CC}	480000	2021-09-14	SW15 2QJ	F	N	L	AYDARTH COURT 39-41	FLAT G	OAKHILL ROAD	NaN	LONDON
524416	{D707E536-0CD3-0AD9-	210000	2021-11-29	SE5 0BJ	F	N	L	GATEKEEPER	FLAT	SCENA WAY	NaN	LONDON

APPENDIX D- Data cleaning

```
df.drop(columns=['Unique_identifier', 'Postcode', 'Paon', 'Saon', 'Street', 'Locality', 'record_status'], axis=1, inplace=True)
```

```
df = df[df['Property_type'] != 'O'].copy()
```

```
#out_file = "C:/Users/Admin/Documents/pp/"
#df = df[df['Town_city'] =="SOUTHAMPTON"].copy()

# display the data set to show that it has been created correctly
#df
#lon_data.to_csv(out_file+"south.csv", index =False)
```

```
empty_data = df[df.isna().any(axis=1)]
print(empty_data)
```

```
Empty DataFrame
Columns: [Price, Date_of_transfer, Property_type, old_new, duration, Town_city, District, County, category_type]
Index: []
```

```
hp['Price'].describe()
```

```
count    5.244190e+05
mean     7.345539e+05
std      4.216840e+06
min      1.000000e+00
25%      2.050000e+05
50%      3.850000e+05
75%      6.250000e+05
max      5.943000e+08
Name: Price, dtype: float64
```

```
hp.drop(columns = 'Unique_identifier', axis = 1, inplace = True)
hp.drop(columns = 'duration', axis = 1, inplace = True)
hp.drop(columns = 'category_type', axis = 1, inplace = True)
hp.drop(columns = 'record_status', axis = 1, inplace = True)

hp.head()
```

	Price	Date_of_transfer	Postcode	Property_type	old_new	Paon	Saon	Street	Locality	Town_city	District	County
0	160000	2018-12-14	M19 2WD	S	N	18	NaN	WILSTHORPE CLOSE	NaN	MANCHESTER	MANCHESTER	GREATER MANCHESTER
1	100000	2018-10-19	M30 0PX	T	N	24	NaN	LEWIS STREET	ECCLES	MANCHESTER	SALFORD	GREATER MANCHESTER
2	206000	2018-10-29	M9 4HR	O	N	163	NaN	MOSTON LANE	NaN	MANCHESTER	MANCHESTER	GREATER MANCHESTER
3	700000	2018-09-18	M34 3PU	O	N	FLAIR FLOORING	NaN	GREY STREET	DENTON	MANCHESTER	TAMESIDE	GREATER

APPENDIX E

```
tow = df.groupby(by='Town_city').median().reset_index()
```

```
tow
```

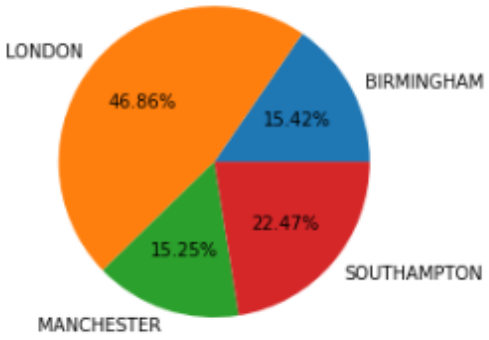
	Town_city	Price
0	BIRMINGHAM	175000.0
1	LONDON	531665.0
2	MANCHESTER	173000.0
3	SOUTHAMPTON	255000.0

```
plt.pie(tow['Price'],labels=tow['Town_city'], autopct='%1.2f%%')
```

```
([<matplotlib.patches.Wedge at 0x18b56708430>,
  <matplotlib.patches.Wedge at 0x18b56708b50>,
  <matplotlib.patches.Wedge at 0x18b567112b0>,
  <matplotlib.patches.Wedge at 0x18b567119d0>],
 [Text(0.9733835903477818, 0.5123713360851307, 'BIRMINGHAM'),
  Text(-0.8409781919587085, 0.7090526642287313, 'LONDON'),
  Text(-0.3462876127314299, -1.0440713046860197, 'MANCHESTER'),
  Text(0.8370389335720555, -0.7136986925058475, 'SOUTHAMPTON')],
 [Text(0.5309365038260627, 0.27947527422825313, '15.42%'),
  Text(-0.45871537743202273, 0.386755998670217, '46.86%'),
  Text(-0.18888415239896172, -0.5694934389196471, '15.25%'),
  Text(0.4565666910393029, -0.3892901959122804, '22.47%')])
```

```
plt.pie(tow['Price'],labels=tow['Town_city'], autopct='%1.2f%%')
```

```
([<matplotlib.patches.Wedge at 0x18b56708430>,
  <matplotlib.patches.Wedge at 0x18b56708b50>,
  <matplotlib.patches.Wedge at 0x18b567112b0>,
  <matplotlib.patches.Wedge at 0x18b567119d0>],
 [Text(0.9733835903477818, 0.5123713360851307, 'BIRMINGHAM'),
  Text(-0.8409781919587085, 0.7090526642287313, 'LONDON'),
  Text(-0.3462876127314299, -1.0440713046860197, 'MANCHESTER'),
  Text(0.8370389335720555, -0.7136986925058475, 'SOUTHAMPTON')],
 [Text(0.5309365038260627, 0.27947527422825313, '15.42%'),
  Text(-0.45871537743202273, 0.386755998670217, '46.86%'),
  Text(-0.18888415239896172, -0.5694934389196471, '15.25%'),
  Text(0.4565666910393029, -0.3892901959122804, '22.47%')])
```



APPENDIX F

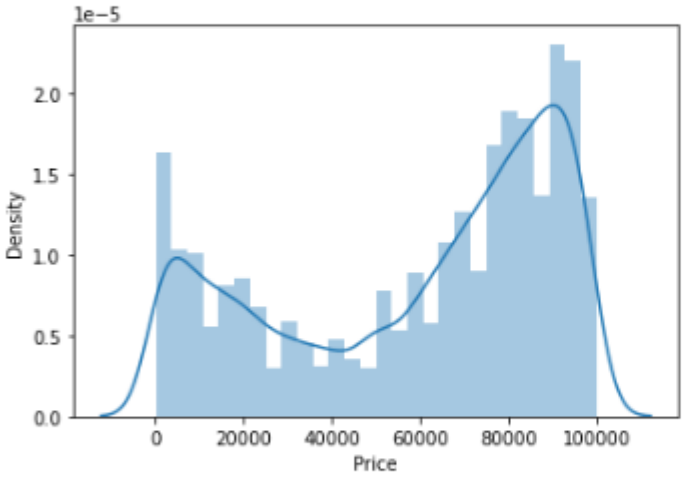
```
plt.figure(figsize=(12,8))
sns.distplot(df2['Price'], color='r')
plt.title('Distribution of Sales Price', fontsize=18)

plt.show()
```

```
C:\Users\Admin\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distp
d will be removed in a future version. Please adapt your code to use either `displot` (a figur
xibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

```
# Distribution of Price under 100 thousand
sns.distplot(df_s[df_s["Price"] < 100000]["Price"]);
```

```
C:\Users\Admin\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureW
d will be removed in a future version. Please adapt your code to use either `disp
xibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

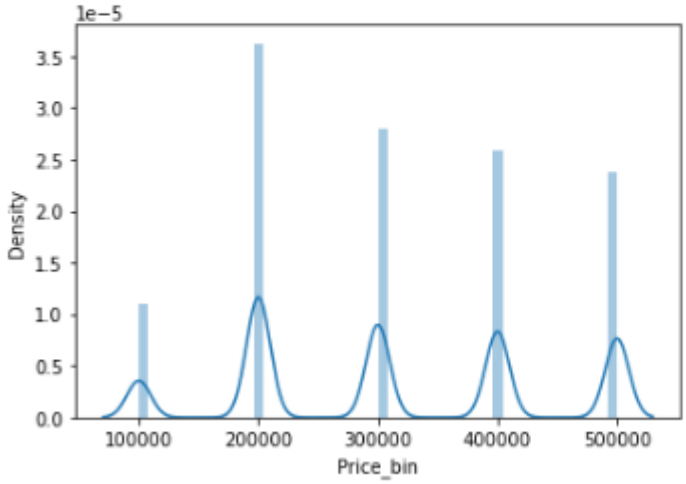



```
# Creating new column as Price_bin
bins = [0, 100000, 200000, 300000, 400000, 500000]
labels = [100000, 200000, 300000, 400000, 500000]
df_s["Price_bin"] = pd.cut(df_s["Price"], bins = bins, labels = labels)
```

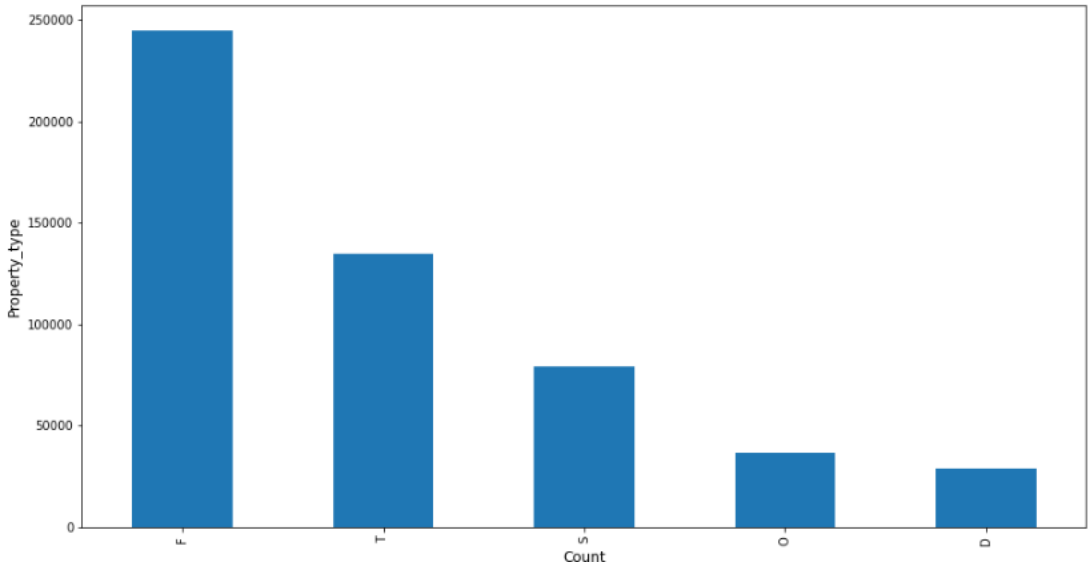
APPENDIX G

```
# Distribution with binning of Price column
sns.distplot(df_s["Price_bin"]);
```

```
C:\Users\Admin\anaconda3\lib\site-packages\seaborn\distributions.py:2619: Futur
d will be removed in a future version. Please adapt your code to use either `di
xibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

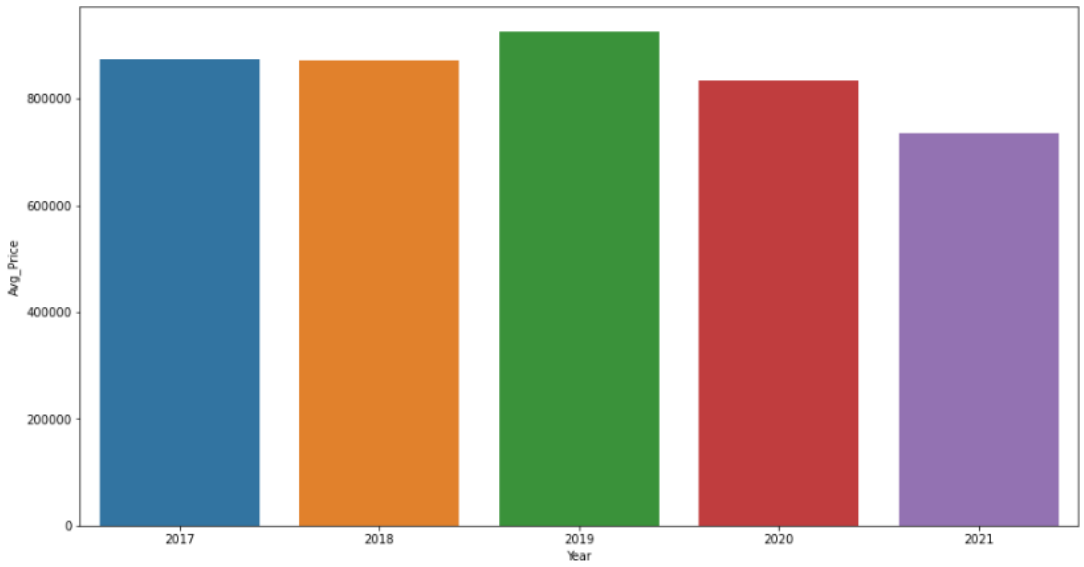


```
# Visualizing Property Type
plt.figure(figsize=(15,8))
ax = df_s["Property_type"].value_counts().plot.bar()
ax.set_xlabel("Count", fontsize = 12)
ax.set_ylabel("Property_type", fontsize = 12)
plt.show()
```



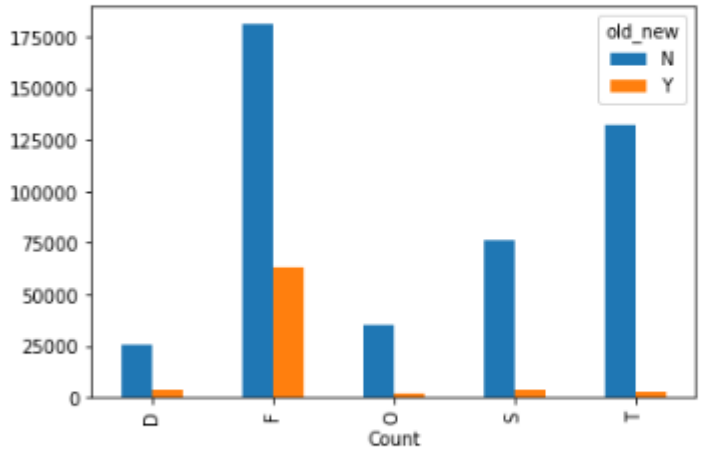
APPENDIX H

```
# Yearly Average Price
plt.figure(figsize=(15,8))
ax = sns.barplot(data = df_yearly, x = "Year", y = "Avg_Price")
```



```
# Visualizing Property Type Vs Old/New
plt.figure(figsize=(15,8))
ax = pd.crosstab(df_s["Property_type"], df_s["old_new"]).plot.bar()
ax.set_xlabel("Count")
plt.show;
```

```
<Figure size 1080x576 with 0 Axes>
```




```
# Visualizing Duration
plt.figure(figsize=(15,8))
ax = df_s["duration"].value_counts().plot.bar()
ax.set_xlabel("Count", fontsize = 12)
```

APPENDIX I

```
hp["Price"].skew()
```

```
33.05297394123192
```

```
log_transform_price = np.log(hp["Price"])

hp['Price'] = log_transform_price

print('New skewness:', log_transform_price.skew())
```

```
New skewness: 0.3627905811536933
```

```
hp_mean = hp["Price"].mean()
hp_median = hp["Price"].median()

sns.displot(data = hp, x = 'Price')
plt.axvline(x = hp_mean, color = 'blue')
plt.axvline(x = hp_median, color = "red", linestyle = '--')
```

```
<matplotlib.lines.Line2D at 0x22b97d6f910>
```

```
lower = 10.2
higher = 13.5
```

```
price_outliers_below = hp.loc[hp['Price'] < lower]
price_outliers_abow = hp.loc[hp['Price'] > higher]

print(price_outliers_below['Price'].count(), "entries having 'Price' value lower than ", lower)
print(price_outliers_abow['Price'].count(), "entries having 'Price' value greater than", higher)
```

```
61 entries having 'Price' value lower than  10.2
89430 entries having 'Price' value greater than 13.5
```

```
price_mean = hp["Price"].mean()
price_median = hp["Price"].median()

print('Mean:', price_mean)
print("Median:", price_median)
```

```
Mean: 12.845863803375098
Median: 12.86099861326992
```

```
lower = 10.2
higher = 13.5
hp["Price"] = hp["Price"].parallel_apply(lambda x: 10.2 if x < 10.2 else x)
hp["Price"] = hp["Price"].fillna(price_mean)

hp["Price"] = hp["Price"].parallel_apply(lambda x: 13.5 if x > 13.5 else x)
hp["Price"] = hp["Price"].fillna(price_mean)
```

APPENDIX J

```
lin_df['Town_city'] = lin_df['Town_city'].factorize()[0].astype('float32')
lin_df['District'] = lin_df['District'].factorize()[0].astype('float32')
lin_df['County'] = lin_df['County'].factorize()[0].astype('float32')

lin_df
```

	Price	Postcode	old_new	Paon	Saon	Street	Locality	Town_city	District	County	Year	Month	Property_Type_is__D	Pr
0	11.982929	M19 2WD	0	18	NaN	WILSTHORPE CLOSE	NaN	0.0	0.0	0.0	2018	12	0	
1	11.512925	M30 0PX	0	24	NaN	LEWIS STREET	ECCLES	0.0	1.0	0.0	2018	10	0	
4	11.198215	M18 8NR	0	63	NaN	PINNINGTON ROAD	NaN	0.0	0.0	0.0	2018	11	0	
5	11.580584	M35 0ES	0	90	NaN	ASHTON ROAD WEST	FAILSWORTH	0.0	2.0	0.0	2018	10	0	
6	11.156251	M24 5LS	0	30	NaN	TALKIN DRIVE	MIDDLETON	0.0	3.0	0.0	2018	10	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
524414	13.428376	E16 2PY	0	COMMODORE HOUSE, 2	38	ADMIRALTY AVENUE	NaN	1.0	36.0	1.0	2021	8	0	
524415	13.081541	SW15 2QJ	0	AYDARTH COURT 39-41	FLAT G	OAKHILL ROAD	NaN	1.0	25.0	1.0	2021	9	0	
524416	12.254863	SE5 0BJ	0	GATEKEEPER BUILDINGS, 5	FLAT 22	SCENA WAY	NaN	1.0	30.0	1.0	2021	11	0	
524417	12.765688	SW18 4GY	0	11	FLAT 53	MAPLETON CRESCENT	NaN	1.0	25.0	1.0	2021	4	0	
524418	12.867471	SE3 9FZ	0	COTTAM HOUSE, 305	FLAT 804	KIDBROOKE PARK ROAD	NaN	1.0	24.0	1.0	2021	8	0	

487981 rows × 16 columns

```
feat_cols = ['District', 'County', 'old_new', 'Year', 'Month', "Town_city","Property_Type_is__D", "Property_Type_is__F",
           'Property_Type_is__S', 'Property_Type_is__T']
```

```
X = lin_df[feat_cols]
y = lin_df['Price']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)
```

```
scaler = StandardScaler()

normalized_array = scaler.fit_transform(X_train[feat_cols])
X_train = pd.DataFrame(normalized_array, columns=feat_cols)

X_train
```

	District	County	old_new	Year	Month	Town_city	Property_Type_is__D	Property_Type_is__F	Property_Type_is__S	Property_Type_is__T
0	1.032916	-0.225299	-0.420972	-1.358846	-0.165553	-0.164029	-0.250508	0.996065	-0.440158	-0.618482
1	0.432007	-0.225299	2.375458	0.727919	0.422730	-0.164029	-0.250508	0.996065	-0.440158	-0.618482
2	0.689540	-0.225299	-0.420972	-1.358846	-0.459695	-0.164029	-0.250508	-1.003951	-0.440158	1.616861
3	-1.714095	-1.125300	2.375458	0.727919	-0.753836	-1.472208	-0.250508	0.996065	-0.440158	-0.618482
4	0.260319	-0.225299	-0.420972	1.423508	-0.165553	-0.164029	-0.250508	0.996065	-0.440158	-0.618482
...	...	...	...	...	...	...	...	...	...	...
365080	1.370718	1.125300	-0.420972	1.423508	-0.459695	1.472208	-0.250508	-1.003951	2.271099	-0.618482