**HOUSE PRICE PROGNOSIS IN U.K. USING MACHINE LEARNING AND VISUALIZATION METHODOLOGIES**

**Student Name: Ronak Bhavik Barot Student ID: 22036660**

**Email:** [**xrb0002@my.londonmet.ac.uk**](mailto:xrb0002@my.londonmet.ac.uk)

**Abstract:**

Things that come to a home buyer’s mind while purchasing a property generally include metrics like the total area of the property, the number of bedrooms, terraced garden etc. However, this paper proves that factors influencing house prices are much more than just those. The key objective of this paper is to provide useful insights regarding housing policies to the U.K. government and also to the housing business personnel who are responsible for making critical business decisions. This paper methodically describes the findings of the machine learning and visualization techniques used to predict house prices in U.K. This paper successfully achieves this primary objective using supervised machine learning model named Linear Regression and the most powerful visualization tool called Power BI. Power BI tool has helped in creating some very insightful visuals which aid crucial problem solving related to the housing industry. Linear Regression model performed better than expected and showed tremendous accuracy. This paper can prove more than a handful for personnel who are associated with the housing industry in U.K. Whether one is buying/selling or analysing the housing market, this paper may prove a game changer as it accurately describes the factors contributing to the house prices and predicts house prices with pin point accuracy. This paper draws some interesting conclusions such as people nowadays prefer remote and silent properties rather than the hustle and bustle of metropolitans. Other machine learning methodologies can also be used to analyse house prices. Clustering can be a good alternative and can give fruitful results.

**Keywords: U.K.- United Kingdom, Linear Regression, Power BI**

1. **Introduction**

According to a report by The Greater London Authority [1],” London’s house prices are considerably higher, and have been rising at a faster rate than the country as a whole. This is particularly true in desirable central London boroughs with median house prices in 2014 as high as £860,000 in Westminster (up 11.4 per cent annually in the five years since 2009) and £1.2 million in Kensington and Chelsea (up 12.2 per cent annually in the five years since 2009) based on Land Registry data. This compares to a London borough low median house price of £215,000 in Barking and Dagenham (up 6.1 per cent annually in the five years since 2009), which is still higher than the national average for England and Wales of £192,000 (up 2.6 per cent annually in the five years since 2009).” Statistics like these prompted the motivation for this paper’s study.

This paper explores the trend in house prices in U.K in the past 18 years and successfully uses it to predict these prices for the forthcoming 13 years. This objective is achieved using machine learning and Power BI methodologies. Data is collected from the country’s national data store and house prices all along the country are studied, explored and predicted.

1. **Literature Review**

One study by Hamad Ahli [2] on house price classification selected clustering algorithms for this purpose. The author used one of the popular housing data sets of Dubai and classified the into clusters and used the same for their classification. The study helped understand the pain points and opportunity areas of the property market in Dubai which is also one of the most booming industries.

Another study by Zufikley et.al [3] verified the use of Artificial Neural Network (ANN), Support Vector Regression and XGBoost as the most efficient models compared to others for successfully predicting house price. These models were developed based on several input attributes and they work significantly positive with house price. The study also concluded that locational attributes are the main attribute in predicting house price.

Finally, a study by R Manjula et.al [4] predicted the real estate value in India using multivariate regression models. The study realized a linear model gives a high bias (under fit) whereas a complexity-based model gives a high variance (overfit). Data Scientist tends to overfit their models which can be reduced by ridge regression and LASSO. The study thus, concluded that LASSO was the most appropriate model for the purpose.

These papers were included in the research as they all include data from relevant and reliable sources, and they successfully achieve the primary motive to predict the house prices in their vicinities. They were also chosen as they worked on most recent data (all worked on data available from the past 15 years). Their modeling style and documentation were also up to the mark and thus they achieved fruitful results and provided valuable conclusions which can assist the housing industry in a great manner.

These papers successfully analyse and predict house prices in various countries. However, they include input parameters which focus only on the property itself and not on factors encouraging the buyer to purchase a property in a particular vicinity. For example, the AQI i.e., the Air Quality Index, the noise pollution in that area, nearest superstore, school, hospital and rail/bus station etc. are more than important in people considering to buy/sell a property nowadays. These areas require further research with the help of technologies at hand these days as people are well informed and alert with the help of internet. These, factors are very significant in making a housing decision nowadays and were not given due importance in any of the research conducted over the internet. This led to the study included in this paper as it almost felt like the need of the hour to include them.

There were many other papers related to house price estimations which were not included owing to their usage of inappropriate and unreliable data sources and some were rejected as they were not properly documented as using them in the research could have led to faulty results and biased conclusions and ultimately had adverse effects on the findings of this paper. The data used in this paper was collected from the Land Registry available on the internet and is reliable as it is published by the U.K. government itself.

Various regression methods were used by different authors predicting house prices and after careful consideration and in-depth research of all these, Linear Regression was chosen as the model to be applied in this paper. The sole reason behind this was that here, continuous variables are used to predict another continuous variable i.e., house price. This model gave excellent results with splendid accuracy and thus, justified its selection in a more than satisfying manner.

1. **Methodology**

The data set used in this paper contains information about housing prices in the U.K. and is accessible easily on the country’s national data store. This data is published by the Land Registry and is thus more than reliable. The data set used in this paper is a huge one as it contains 54 attributes with 134829 data points. The target variable is “AveragePrice” that is predicted using all other attributes. The data analytics life cycle is described below:

1. **Discover**

The number of bedrooms or the total square feet area are not the only factors people are considering while purchasing properties but factors such as local AQI (Air Quality Index), closest rail/bus station are equally important. The amount of influence such neglected factors have, on a housing decision, is the business problem this paper attempts to solve.

People are nowadays well informed and so, are more health conscious than ever and hence, parameters that keep them healthy physically and mentally are given more importance these days while purchasing properties. There is enough data available on the internet to support this theory. Hence, an attempt is made in this paper to find the most influential factors contributing to house prices in U.K.

1. **Data Preparation**

This is the most essential and time-consuming step in data analytics as the data needs to be well structured and organised so that the model can give accurate results at a later stage. The data set used here is huge as it contains 54 attributes and 134829 data points. All of the data was checked and it was found that many attributes had null values. This was done using the data filtering feature in Power BI where one can view the values one each attribute contains. The Power BI tool uses “Blank” expression to indicate null values. In this case, the number of null values each attribute had were huge so dropping the rows having null values or replacing them with zero was not a good option. Hence, adequate research was done on replacing the null values with median. **The null values were replaced by median and not mean as the values here are continuous and replacing them with median gives more accurate results.**

There are many methods in Power BI with how this can be achieved and the simplest of those was followed here. Initially, clicking on “Transform data” tab leads to the power query environment. Then clicking on “Replace values” tab on the top, the null values are replaced with zero.

This can also be achieved using the power query. The M code to do this is as follows [5]:

**= Table.ReplaceValue(#"Changed Type", null, 0, Replacer.ReplaceValue, {"Sample"})**

After replacing null values with zero, the median of the whole column was taken into account and all the null values of that column were replaced by the calculated median. The M code to achieve this is as follows:

= Table.ReplaceValue(#"Changed Type",null,**List.Median(#"Changed Type"[Sample]),**Replacer.ReplaceValue,{"Sample"})

Finally, all attributes were checked for having null values using data filtering and it was found that all attributes had non-null values. The categorical attributes had no null values so they did not need data cleaning. Hence, it was safely concluded that the data set had no anomalies and was ready for further processing.

1. **Plan Model**

The machine learning models researched for analysing the data include Artificial Neural Network (ANN), Support Vector Regression, XGBoost, Clustering and Linear Regression. Out of all these Linear Regression was chosen for data modeling as continuous quantitative parameters were used to predict continuous value of house prices. Another reason for choosing this model was that it provided the best results for types of data sets used in this paper as compared to other models. The data used in this paper mostly contains numerical variables which are linearly distributed amongst themselves. Thus, Linear Regression would be of great help to predict the house price in this project. The choice of Linear regression is justified by the amount of accuracy it provided which will be discussed in detail at a later stage.

1. **Build Model**

The applied model of Linear Regression showed outstanding results with noteworthy accuracy. Linear Regression was applied to the target variable versus the four most correlated attributes and it gave correlation coefficients in the range of 0.85-1.00 and R-Squared values in the range 0.8-0.9. The plots of linear regression showed positive linear correlation. These metrics are proof that the applied model is very accurate and can be successfully used for predicting house prices.

Thus, the linear regression model was used to predict house estimate house prices for the next 13 years. This was done using the forecasting feature of Power BI. These predicted values were then compared with the actual values and they showed remarkable similarities. All these plots will be presented at a later stage.

Identifying the most correlated variables by using correlation coefficients against the target variables is the best strategy for applying linear regression which is also known as feature selection. The threshold value chosen here was 0.8.

1. **Communicate Results**

The linear regression model was used to predict house prices. The primary objective of this paper was to analyse the given data and accurately predict house prices for the next 13 years which was achieved successfully. The Power BI dashboard created is more than useful for taking crucial business decisions as it depicts the performance metrics affecting the house prices in U.K. These valuable insights can be more than useful for the government to design policies aiding planning and development of the society. It can also be utilised by housing business personnel to do critical analysis and form profitable business policies.

The four most correlated attributes to the target variable were “DetachedPrice”,

”SemiDetachedPrice”,”OldPrice” and “TerracedPrice”. These are crucial parameters and were on expected lines but the surprising omission from this list was the “PriceIndex” as it one can easily correlate it to the target variable. However, it was found to be less correlated to the target variable than the four variables mentioned earlier.

1. **Apply Live**

The ethical considerations for using the data set and data analysis methodologies include using reliable and trustworthy data sources, using appropriate modeling methods for the data collected by researching proper papers and avoiding misuse of the data or any other tool for malicious practices.

There are many commercial risks to this paper as people dealing with real estates would not wish that their customers have such transparencies related to house pricing as it would reveal their own commissions and the amount of money they make from broking properties in and around U.K.

The findings of this paper may not be favourable for real estate dealers, but for the customers it provides great insights as to where, when and what sort of property one should prefer buying in U.K without getting cheated and at a reasonable and close to market price.

Thus, the six steps of data analytics lifecycle were successfully implemented and gave more than useful insights to aid the housing market in U.K. Data preparation was the most crucial and time-consuming part but it proved why it is worth investing time in it as it gave splendid results at the data modeling stage. Overall, each step had its own significance and was helpful in achieving the objectives of this paper.

1. **Data Modeling**

Sir Francis Galton was the pioneer of Linear Regression as he was the first person to propose it in 1894[6]. Linear regression can be described as a mathematical testing method used to quantify the relationship of variables of a dataset. Corelation analysis can only numerically depict this dependency. But Linear regression can visualise it mathematically. **Y= mx + c** is one of the most popular equations used in academics also known as the equation of a straight line. Linear regression uses this simple equation to calculate the dependency of the variable y on the variable x mathematically. The R2 score, also called as the regression co-efficient, used in Linear regression is nothing but the measure of this particular dependency.

The four key variables that significantly contribute to the model prediction are “DetachedPrice”,”SemiDetachedPrice”,”OldPrice” and “TerracedPrice”. These are the variables that had a regression coefficient of more than the threshold value of 0.8.

Power BI is a unified, scalable platform for self-service and enterprise business intelligence (BI). Connect to and visualise any data, and seamlessly infuse the visuals into the apps you use every day [7].

Power BI enables one to import raw data from any file virtually or from the local device. In this case, raw data file was downloaded in the csv format on the local device and then imported into Power BI using the “Get data” feature of Power BI. This data is then loaded into Power BI using the “Load Data” feature. One can transform the raw data before loading it into Power BI but transformation of large data sets is usually done after loading them and thoroughly exploring the data and the same procedure was followed here.

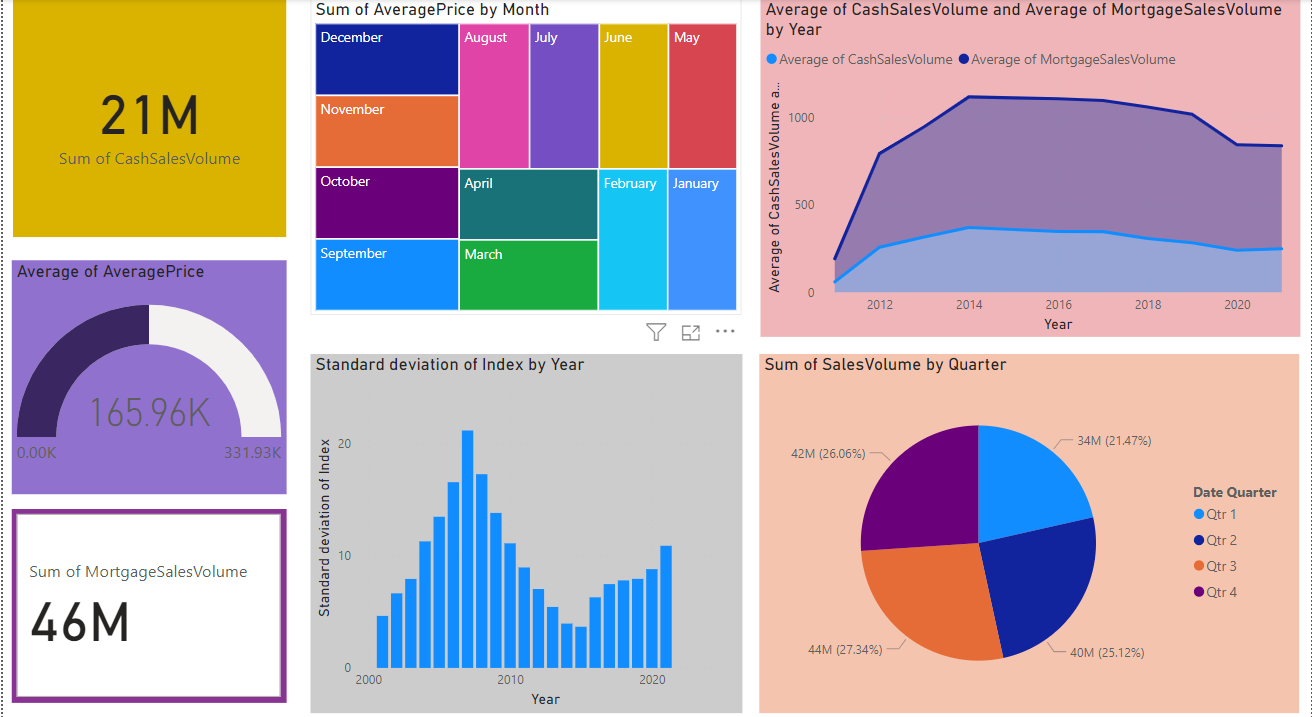
There were many missing values in the data set at hand and the decision was taken to replace them by median of the respective column as dropping those rows was not a good option as the number of rows having null values were large. After replacing null values with zero, the median of the whole column was taken into account and all the null values of that column were replaced by the calculated median. The outliers were ignored as they were small in number.

Power BI has many modeling features. Initially, the regression coefficient and R2 score were calculated using the “New quick measure”. Then the scatter plot visualization was chosen and the four most correlated variables were plotted against the target variable. The “Trend line” feature was activated to get the line of best fit. Finally, the “Forecasting” feature was used to estimate house prices for the next 13 years. The Linear Regression perfectly integrated with the visuals generated as all the four plots showed positive linear correlation with the target variable. The visuals clearly proved that the chosen model was the most appropriate one for the data set used here.

1. **Results And Discussions**

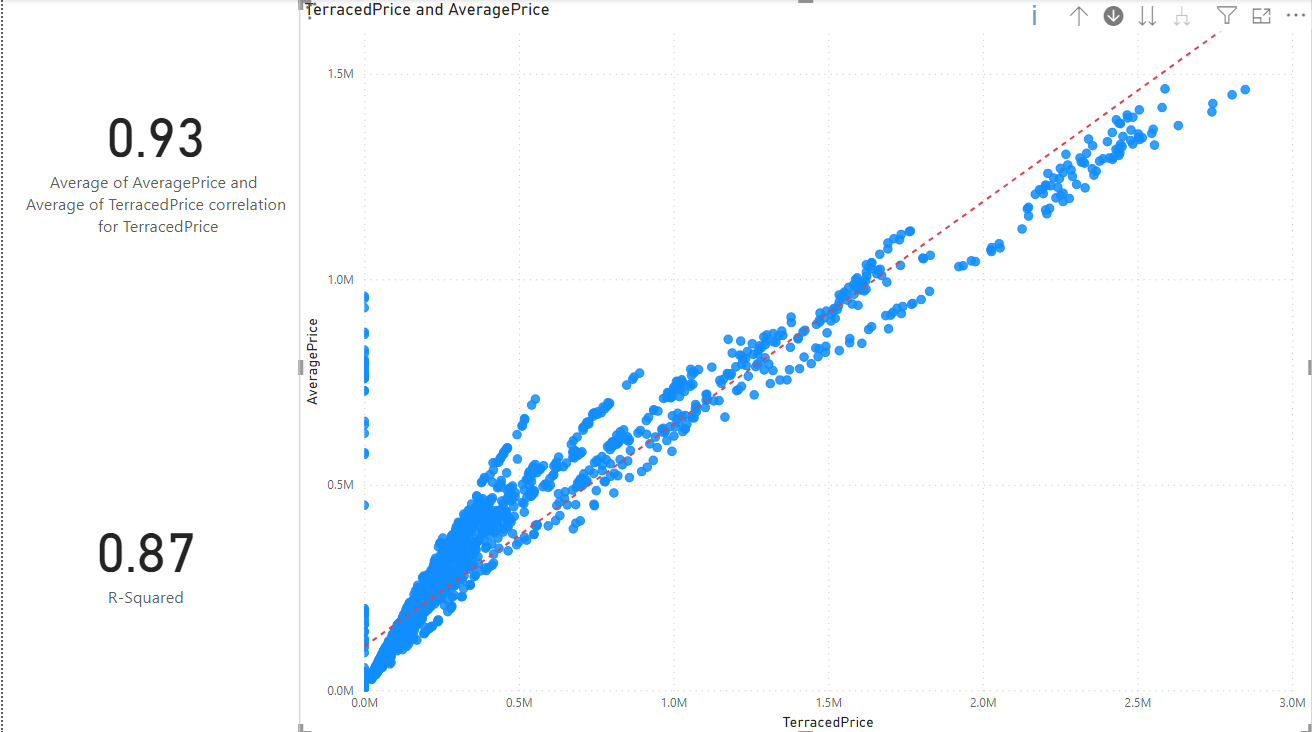
The Power BI dashboard created in this paper includes plotting the performance metrics against the target variable using stacked area chart, pie chart, tree map, stacked column chart and some key metrics using cards. The next visual uses cards to show the regression coefficients and R2 score along with scatter plots showing linear regression with trend lines. The final visual uses line chart to forecast house prices for the next 13 years using forecasting feature.

The main findings of the visuals state that the house prices are evenly distributed along the period of time. The house prices show positive linear correlation, meaning they are bound to increase steadily over the period of time. The house prices are overly dependent on only a few parameters which is evident from the visuals that are described below.



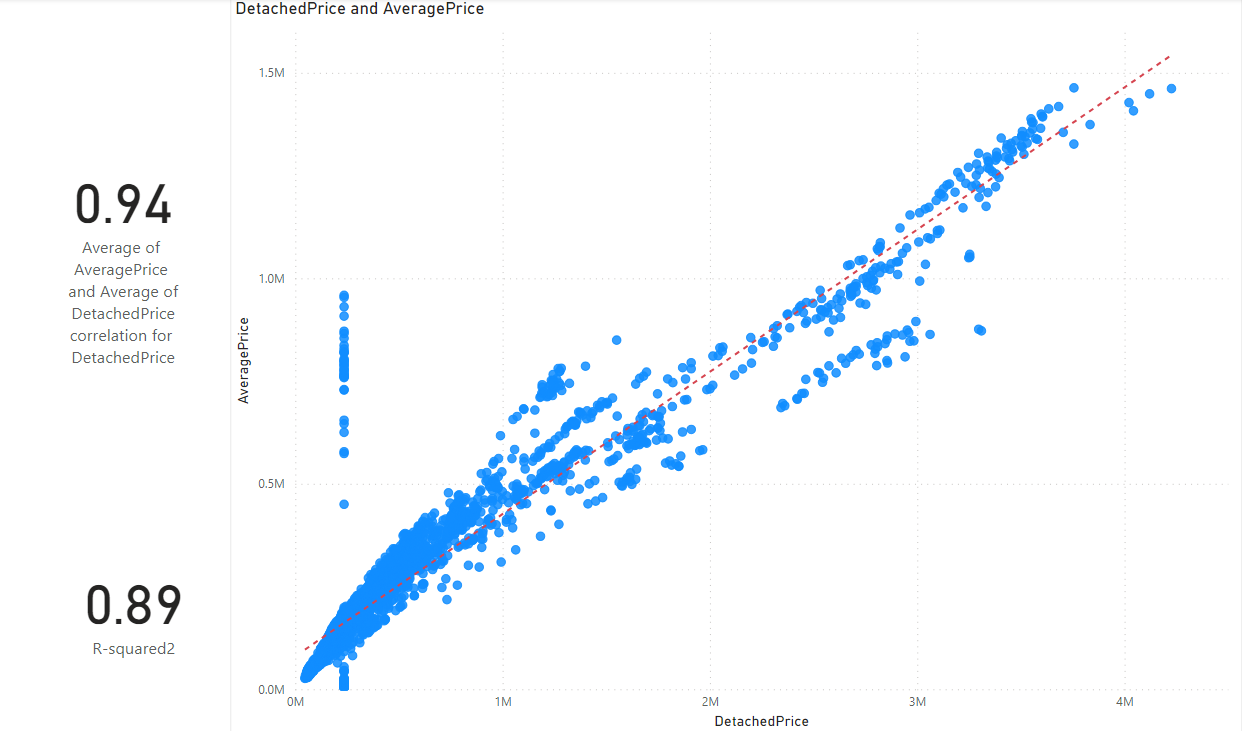
**Fig 1. Power BI Dashboard**

The above dashboard was created in Power BI from the available data. The main essence of creating this dashboard was to assist business stakeholders and responsible personnel so that they can take crucial decisions based on the data that ultimately contribute to the growth of the housing industry in U.K. This dashboard provides information about the sum of average prices per month which can help in analysing price related information month-wise. It also illustrates the standard deviation in index price of each property by year which is again crucial in evaluating the yearly change in index prices of the properties. Similarly, the sum of total sales per quarter is also shown which again is vital in making various analysis quarterly. It also visualizes the comparison between the cash sales and mortgage sales per year which provides information to compare which payment type the customer has preferred. Finally, the sum of cash sales and mortgage sales along with the average price of the properties are displayed as they can be helpful in predicting future values of properties.



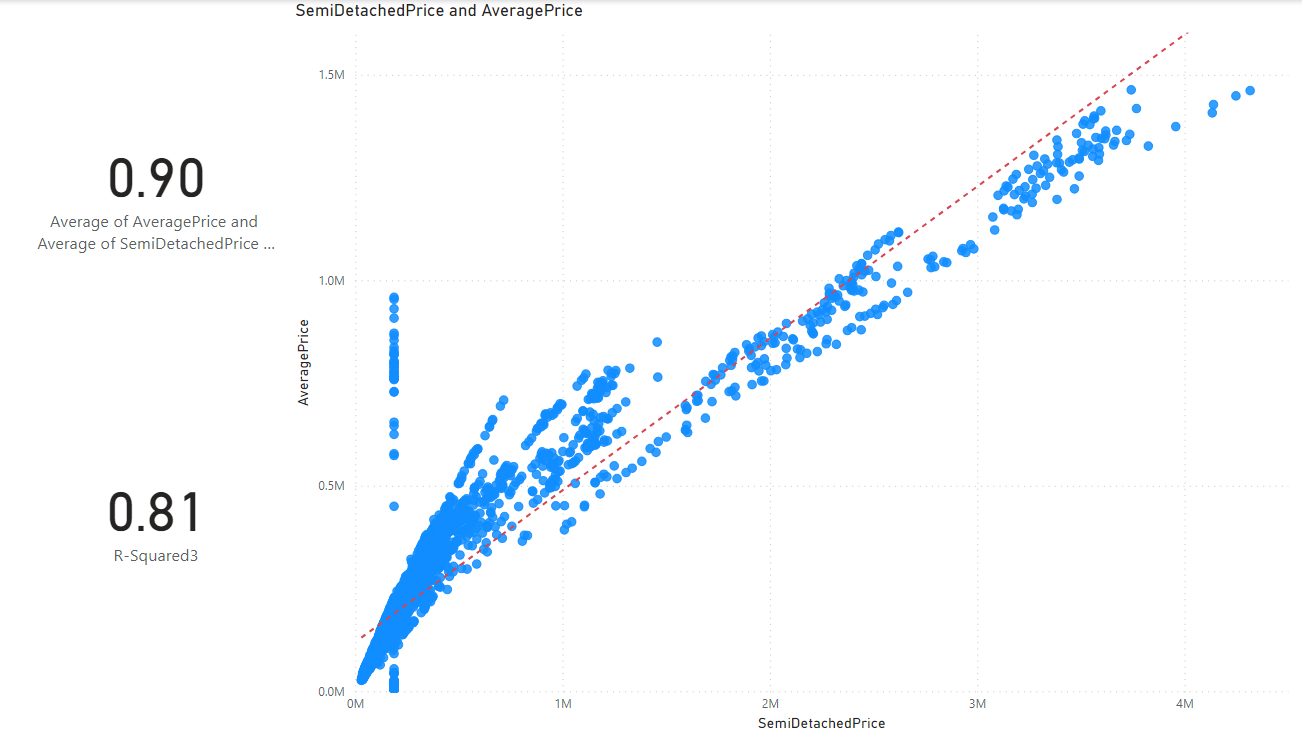
**Fig 2. Linear Regression of “AveragePrice” v/s “Terracedprice”**

The high correlation coefficient, R-Squared value and the scatter plot give enough evidence to conclude that **“AveragePrice” and “TerracedPrice” show positive linear relationship**.



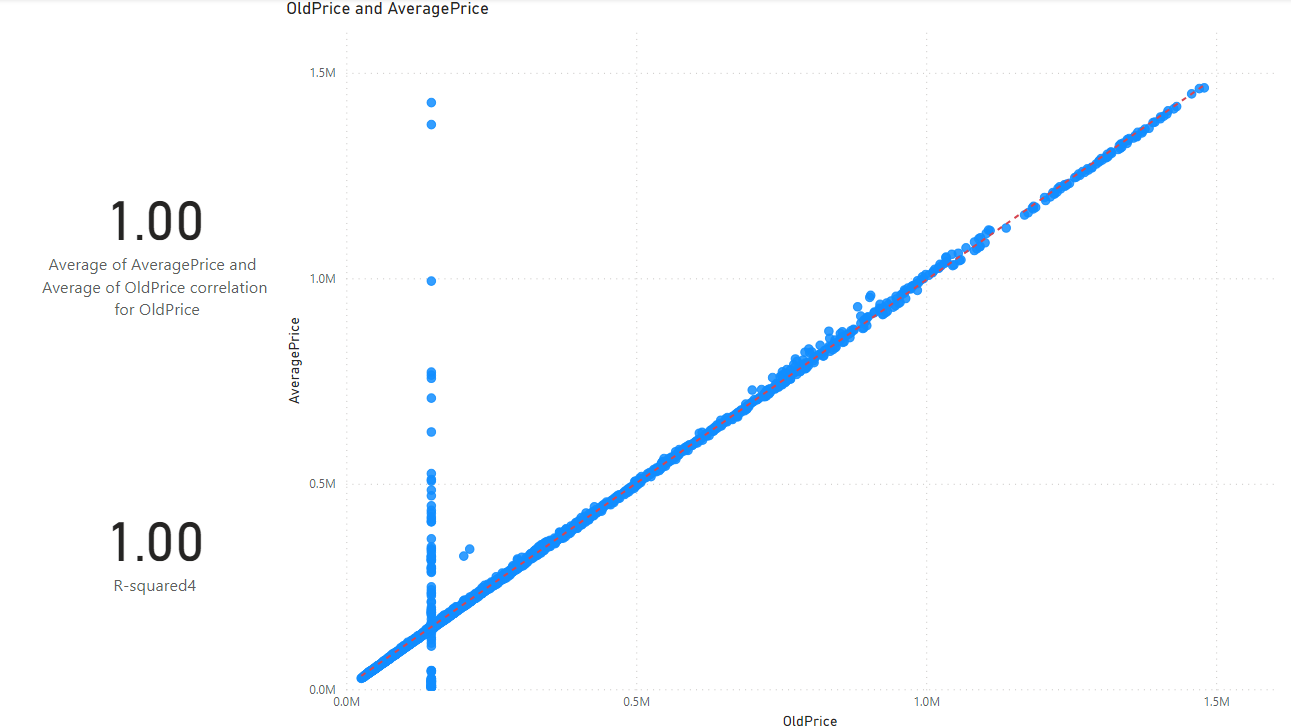
**Fig 3. Linear Regression of “AveragePrice” v/s “Detachedprice”**

The high correlation coefficient, R-Squared value and the scatter plot give enough evidence to conclude that **“AveragePrice” and “DetachedPrice” show positive linear relationship**.



**Fig 4. Linear Regression of “AveragePrice” v/s “SemiDetachedprice”**

The high correlation coefficient, R-Squared value and the scatter plot give enough evidence to conclude that **“AveragePrice” and “SemiDetachedPrice” show positive linear relationship**.

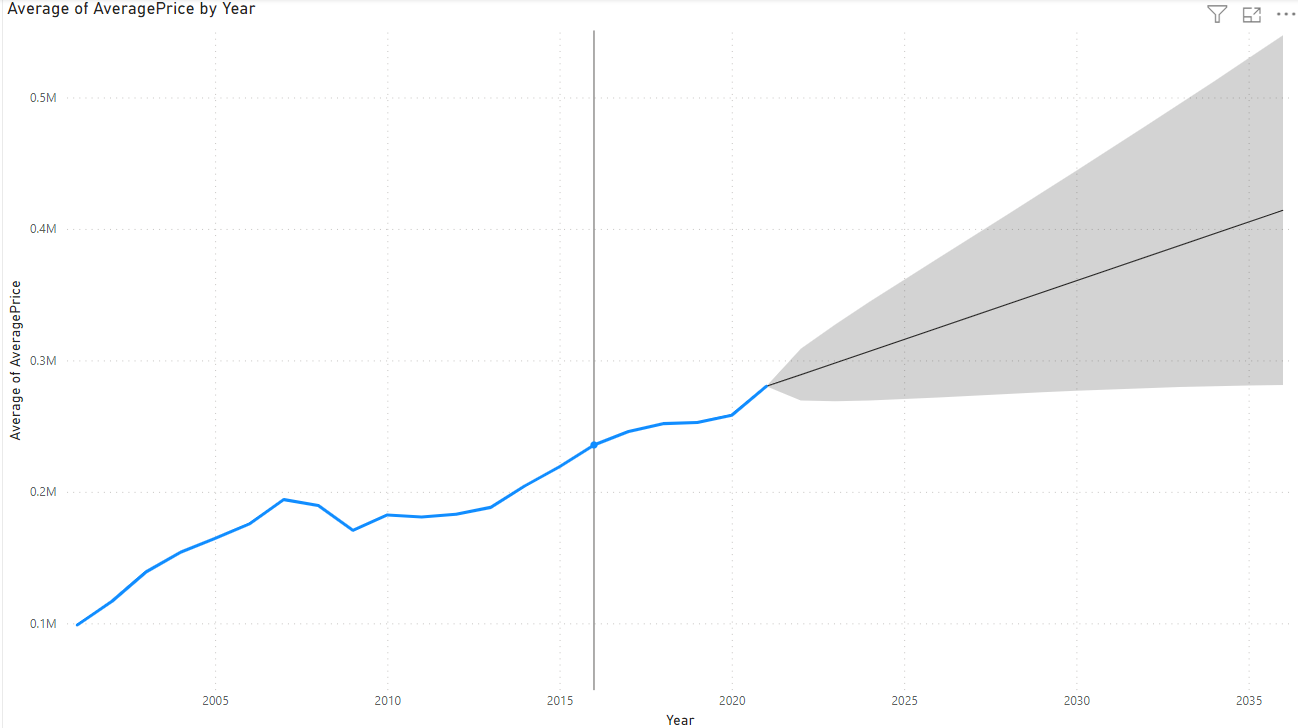


**Fig 5. Linear Regression of “AveragePrice” v/s “Oldprice”**

The high correlation coefficient, R-Squared value and the scatter plot give enough evidence to conclude that **“AveragePrice” and “OldPrice” show positive linear relationship**. But here, it is a case of **overfitting** as both regression coefficient and R-square value are 1.

These plots have certain outliers but they can be ignored as they are less in number. These can be certain exceptions where house prices can deviate under unforeseen circumstances for example, as they did during Covid-19 times.

These plots strongly reinforce the fact that Linear Regression model has given outstanding results on the given data set and they do complement each other brilliantly.



**Fig 6. Forecasting of House Prices till year 2036**

The applied model was then used for predicting the house prices in future which is also known as forecasting. The average house prices were predicted until year 2036 as seen in the figure above.

The Power BI visuals presented above clearly indicate that the house prices in U.K. are directly proportional to their earlier prices which is quite sensible as well. These visuals assert the fact that Linear Regression is generally the go-to model in predicting these types of values.

1. **Conclusion and Recommendation**

The findings of this paper are like gold dust for anyone related to the property market in U.K as it provides exactly the parameters and methodologies needed to predict housing prices in U.K. It provides great insights to buyers/sellers who wish to buy/sell properties in U.K. It also provides pricing transparency related to housing in U.K which in turn would benefit the development of the real estate industry in U.K in a great manner.

It is always true that there are two sides of a coin and similar is true in this paper’s case as well. There are many commercial risks to this paper as people dealing with real estates would not wish that their customers have such transparencies related to house pricing as it would reveal their own commissions and the amount of money they make from broking properties in and around U.K. The real estate industry is one of the most recently booming industries as more and more people are choosing U.K as their destination for civilisation. Hence, along with positives, the findings of this paper may not be favourable for real estate dealers, but for the customers it provides great insights as to where, when and what sort of property one should prefer buying in U.K without getting cheated and at a reasonable and close to market price.

The conception that only real estates in metropolitan cities are in demand is a myth. People nowadays, prefer remote and silent locations to stay peacefully and away from all the hustle and bustle of large cities. Pollution too, has played a major role in people’s selection of location to live nowadays. Hence, remote and smaller houses too are equally in good demand in the real estate market.

One astonishing fact that was clearly visible from the model’s results’ accuracy was that despite working on such a large dataset, the models showed stunning results. Generally, working on a small dataset gives very good accuracy results, which also may result in overfitting the models, but in this case, the dataset was very large and still the models performed brilliantly. Thus, it can be concluded that the dataset has very reliable variables containing actual numbers which are of great assistance in predicting the house prices in U.K. Also, it has the right parameters needed to predict housing prices of a region.

Further research is recommended on the lines of finding the influence of factors such as AQI, noise pollution, proximity to rail/bus stations etc. in determining the house prices in U.K.

**References**

[1]  *Joel Marsden “House prices in London – an economic analysis of London’s housing market” [2015], Accessed from* [*https://www.coursehero.com/file/63872183/house-prices-in-londonpdf/*](https://www.coursehero.com/file/63872183/house-prices-in-londonpdf/)

[2] *Ahli, Hamad, "House Price Classification using Clustering Algorithms" (2022). Thesis. Rochester Institute of Technology. Accessed from* [*https://scholarworks.rit.edu/theses/11381*](https://scholarworks.rit.edu/theses/11381)

[3] *Zulkifley, Nor & Rahman, Shuzlina & Nor Hasbiah, Ubaidullah & Ibrahim, Ismail. House Price Prediction using a Machine Learning Model: A Survey of Literature. International Journal of Modern Education and Computer Science. 12. 46-54. [2020]*

[4] *R Manjula*et al*2017*IOP Conf. Ser.: Mater. Sci. Eng.***263****042098 [2017]*

[5] Accessed from <https://smantindata.com/replace-null-values-in-power-query/>

[6] Accessed from <https://www.researchgate.net/publication/349477129_House_Price_Prediction>

[7] Accessed from <https://powerbi.microsoft.com/en-gb/what-is-power-bi/>.