PRESENTATION SCRIPT

# Introduction

# Data Description

The real estate data has 43 variables in all which have been classified into 4 different categories:

* transactional (mainly involving price and dates)
  + price\_1 and price\_2 for the 1st and 2nd transaction prices respectively
  + date\_1 and date\_2 corresponding to those sales.
  + days between sales is their difference
  + perc\_change\_p\_1\_p2 is the percentage change in prices
* socio-economic
  + consisting of different social and economic factors affecting the price of real estates
  + mainly it assigns scores on the basis of these factors like employment, income
* sustainability (denoting the energy efficiency) and
* geographical (denoting the various regions of England and Wales).

The last two variables will be discussed in the following slides.

Next, we come to the repeated sales prices. The plots of the logarithms of the percentage change and

the time period between two transactions are shown and it is clear that both of them are not normal.

The summary numbers out the various aspects like mean, median, normality, etc. of several variables including the two prices, their logarithms, percentage change and time period. The value of the normality also proves the fact that both the plots in the previous slide was not normal.

Energy efficiency has denoted by the EPC (Energy Performance Certificate) which ranges from a scale of 1 to 100 and has been divided in 7 categories: 1-20, 21-38.... 92-100 naming them A, B, C .... G. So, A is the best and G is the worst.  
The EPC summary chart shows that none are of type A also very few are of type G with the highest being of type G.

These are the geographical distributions of the real estates among England and Wales.

So, we have labelled the states of England in different colors. In our data the variables are in the form of one dot vector, i.e., if the house is in London, then London has value 1 and rest are 0. Let’s check the fraction of houses in these regions.

North-west 19.99%

North East 3.66%

Yorkshire and the Humber 17.66%

West midlands 14.25%

East midlands 10.35%

East of England 6.64%

London 8.59%

South-west 6.83%

South- east 9.68%

Now, we will look at various Index of Multiple Deprivation (IMD) Variables which includes a discrete and a continuous plot (i.e., rank and decile) of every variable which includes Income Deprivation rank, Crime Rank, Employment Deprivation etc. Here 1 is the best and 10 is the worst.

# Hedonic Regression

Let us look at some preliminary plots before moving to regression.

Here is the plot of mean of logarithm of prices plotted against the various EPC variables. As you can see here there are no EPC A because our dataset does not contain those variables.

Moving on to the boxplot of ln prices again plotted against the different EPCs. This shows that there is a slight but some trend. In the x-axis 2 refers to EPC B, 3 to EPC C and so on

Here comes the third plot, where we again plot the mean of ln prices but now against the different regional variables.

The 4th is a boxplot plot of the ln prices plotted against the regional variables, i.e., 1 refers to north east, 2 to north-west and so on.

Now let me explain the key idea behind hedonic models and construct a hedonic index.

The hedonic regression framework is widely believed to have been developed by Court in 1939. The main idea of hedonic models is to decompose the characteristics of similar heterogeneous assets, and give them separate values. A dummy variable can take one of two values, either 0 or 1. If we consider the EPC\_D for example, its dummy variable would be 1 when the house is of EPC rating D, and 0 otherwise.

So, we move to formulate the model. We take P\_it to be the Price at time t and i denotes the variable number. We model the ln price with multiple

regression. Let's run the regression with R. We get the estimated coefficients.

Moving on to Partial Changes, if we change x\_jt by delta x\_jt then the Price changes from P\_t to P'\_t. Now performing some jugglery here, we can see that the ratio of the two prices is exponent of the coefficient times the change. Note one thing that in hedonic regression the change is discrete. And it can be 1 or -1 because we are considering the dummy variables. Without any loss of generality, we may assume that it is 1. Now the exponents of this coefficients give us a measure for the change in prices as the dummy variable is changed. These values are known as hedonic index.

Also here are some model choices we have made. Note that while categorizing into k categories we require k-1 variables. However, we have k dummy variables. So, one must be held out. We have held out epc D because it has the highest fraction of data. For similar reasons we have again referenced on north west.

Let's construct the hedonic index. We just plot the exponents of the coefficients. Here is the graph.

We can see that there are peaks in London. This shows that if the house shifts to London (hypothetically speaking) then there is huge change in price. Also, we can notice

a dip in change of price in epc g. Moreover, this dip makes it lesser than one. Which shows that if non epc d house is made g then there is a decrease in price. Now why D?

because we have referenced on d [as explained in previous slide].

Let’s move on to the diagnostic plots of this regression

1. Residual vs Fitted

This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model doesn’t capture the non-linear relationship.

1. Normal Q-Q

This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It’s good if residuals are lined well on the straight dashed line.

1. Scale-Location

It’s also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It’s good if you see a horizontal line with equally (randomly) spread points.

1. Cook’s Distance

We plot Cook’s Distance against row number, we can see if highly influential points exhibit any relationship to their position in the dataset. We can see there are too less points with high Cook’s Distance. This shows that although the data had many outliers pointed out by the boxplot but they mostly have low influence.

1. Residuals vs Leverage

This plot helps us to find influential cases (i.e., subjects) if any. Not all outliers are influential in linear regression analysis (whatever outliers mean). We look for cases outside of a dashed line, Cook’s distance. When cases are outside of the Cook’s distance (meaning they have high Cook’s distance scores), the cases are influential to the regression results. We can barely see Cook’s distance lines (a red dashed line) because all cases are well inside of the Cook’s distance lines.

1. Cook’s Distance vs Leverage

Cook's distance and leverage are used to detect highly influential data points, i.e., data points that can have a large effect on the outcome and accuracy of the regression. High leverage observations are ones which have predictor values very far from their averages, which can greatly influence the fitted model.

The contours in the scatterplot are standardized residuals labelled with their magnitudes.

# Price vs Socio-Economic

Here we are going to model the price variables i.e., price\_1 and price\_2 with respect to the socio-economic variables, i.e., we will concentrate on the scores which are continuous values.

In linear regression, we assume that the errors are normally distributed. This assumption allows us to construct confidence intervals and conduct hypothesis tests. We also know that the predicted value has a normal distribution under a fixed value of explanatory variables. Hence, the original variables better be normal. However, the histograms clearly suggest that they are not. They are left skewed.

By transforming our target variable, we can (hopefully) normalize our errors (if they are not already normal). We can use the Box-Cox transformation to transform the Y into as close to a normal distribution as the Box-Cox transformation permits. Now, in Box-Cox, we try out these transformations of Y. Then we choose the of lambda that provides the best approximation for the normal distribution of our response variable.

Basically, we start with y and then transform it to f(y) and then we perform the regression.

Now, we come to the part of selecting lambda. This log-likelihood function is an estimator just like method of moments. Here, it estimates the lambda. So, what Box-Cox does basically is reduces the standard deviation. In order to do that the we choose the lambda for which log-likelihood is maximized. R does the whole process for us by checking values of lambda from -2 to 2. This is the log-likelihood plot for price\_1. We can see lambda is quite close to 0.

This is the log-likelihood plot for price\_2. However, the lambda is not 0 here. Here the lambda comes out to be -0.3 approximately.

Now this is the price\_1 variable as seen in one of the previous slides. Since lambda was zero, we apply log to price\_1.

Here is the outcome. We can see that it is much closer to normal.

Next this is the price\_2 variable as seen in one of the previous slides. Here lambda is -0.30303, we apply y^lambda-1/lambda to price\_2.

Hence, we get the following output, which also looks like normal.

Then we run the regression to obtain the coefficients. Here is a fancy plot of the coefficients. We can see that crime and barrier score hardly has any effect on the price\_1

Similarly, in the case of price\_2 too. Most of the coefficients are small but after taking the inverse transform of the function. It may become quite relevant.

Then we move to the partial residue plots. So here are the partial residue plots corresponding to each variable. Again, we can see that crime and barrier has hardly any effect. Moreover, these plots reveal possibility of non-linearity in data. R fits a mean line through each of these plots. It is in dotted red hence very faint. Those two are quite close. Therefore, our model is quite accurate. Now, you may ask if the residual plot revealed the possibility of non-linearity, then why we check for partial residues. It’s because they show us the non-linearity of each variable explicitly hence its more useful often.

# Repeat Sales Index

In 1963, Baily, Muth and Nourse developed a methodology for constructing a real estate index. The idea was simple, the coefficients of the index at each designated period can be estimated by running an ordinary least square regression. The beauty of the repeat sales index methodology lies in its simplicity. Our aim is to assign index to each year. We want to make sure that these indices are close to the change in prices of houses.

Let us assume that the house was first sold in year\_i and then in year\_j. Now we want the ratio of price\_1 and price\_2 to be close to the ratio of the indices of these two years. We take logs on both side in order to make it linear. We set the index of the base year to be 1. In our case the index of 1995 is 1. Now set beta\_i to be the log of index of year\_i. We take y to be the difference of ln price\_1 and price\_2.

Now we model it as multiple linear regression. Now the question comes what are x\_ij. It is the coefficient of this log indices. As we can notice here that it is -1 during the first sale, 1 during the second and 0 otherwise. Now we estimate with OLS to get the estimated values of these indices.

Before moving to regression, we need to preprocess the data to get the required values. Let’s take this example. So, this is our data and now we need to get the log change price which is the difference of the column 2 and 4. Also we need to get the X matrix.

After getting the values, it will look somewhat like this. As u can see 2006 has 0 since it was not sold during that year. Similarly, we also have the log change. Now, notice that we don’t need this column of 2006. Why? Because we already have set that index to be 1, i.e., beta to be 0. Hence, our required part is the shaded area.

Finally, we perform the regression. The exponents of the coefficients are the index levels at each year. Then we plot those indices. We can clearly see the major retractions in 1998 and 2011. We can also notice that the index level for this repeat sales index had reached its highest level in 2012. Hence, we can infer that there is an inflation of house prices over these years in UK.

# Conclusion

Landlords can make informed investments when there is a clear link between price and energy efficiency. In addition, we have taken regional aspects into account. We linked the pricing to socio-economic parameters in the second model. The empirical findings confirm that energy efficiency characteristics have a minor but considerable impact on both transaction prices.

This has opened the gateways of many future extensions. Like, the current study was unable to control the correlation between factors. The graph clearly shows that there is high correlation among some variables which we assumed to be independent. Better transformations or deletion of correlated columns may help in those cases.

Furthermore, our model was solely devoted to the prices and we failed to generate any link between the EPCs and Regions. This plot also shows that such a link may be evident. The data can be extended to much larger database too, in order to generate a robust idea on this split incentive problem.

We move to the limitations.

* For hedonic regression, a substantial amount of data must be collected and processed with. However, there were only 4201 values in the data used.
* The Hedonic Price Index calculates how much people are prepared to pay for alleged differences in environmental quality and its implications. However, if individuals are unaware of the link between environmental attributes and property worth, the value will not be reflected in the price.
* Because it only uses data on units that have sold multiple times throughout the sample period, the RSI approach is inefficient. As a result, a sample selection bias problem arises.
* The target variable is harder to grasp if lambda is non-zero for price 2 in Box Cox Transformation than if we just applied a log transform.

Every variable is subjected to a theoretical analysis. To eliminate assumptions, better statistical approaches might be utilized. Many assumptions may not hold true for the wider populace. The entire investigation was carried out for educational grounds.

Here are our references.

We'd like to express our gratitude to Dr. Rituparna Sen for her significant assistance as a mentor and for giving us the opportunity to learn more about the subject. We'd also want to show our thanks to the writers for their thorough research. Finally, we'd want to express our appreciation to several of our classmates, with whom we discussed our worries and found it easier to navigate when we encountered difficulties.