

Analysis of UK Cities House Prices based on City Economic Fundamentals

Capstone Project 1 | Final Report

1 Background

The project examines the movement of house prices for 62 major cities in the UK in the period from 2003/4 to 2015/16 (the start and end years vary depending on availability of data) and which factors influence house price dynamics.

The data set includes information on indicators for each city, including average house prices, housing stock, population, employment, average wages and other data. Each of these factors was correlated to house prices across the 62 cities, in order to identify the factors where there is a relationship of statistical significance.

2 Problem and client

Understanding the behavior of house prices is of huge interest to banks, real estate investors, property developers, and the buying public. A typical client for this analysis would be a UK-based mortgage lender with a portfolio of loans across the country. The results of this exercise will help answer the following questions:

- What key factors influence the movement of house prices in major cities the UK? What makes house prices certain cities more expensive than in others?
- Can the cities be grouped based on their economic fundamentals and risk profile?
- Can we use any of these factors to predict the movements of house prices by city e.g. one year from now?
- How is the overall level of house prices in the UK related to the FTSE100 stock market index?

3 Data source and steps to clean data

The data set used is published by the Centre for Cities (www.centreforcities.org). The site has a data tool where data can be downloaded in the form of a .csv file.

In order to arrive at a clean data set, the following actions were performed:

- Importing: reading the .csv file into a pandas data frame and exploring content of 63 rows (one row for each city) and 400+ columns (indicators)

- Cleaning: removing data from the footer, removing columns that contain no data and one row with missing price data, converting text rows into floating point numbers
- Manipulation: extracting slices of the data set to obtain specific annual data of interest (house prices, population, wages, employment, housing stock, etc.)

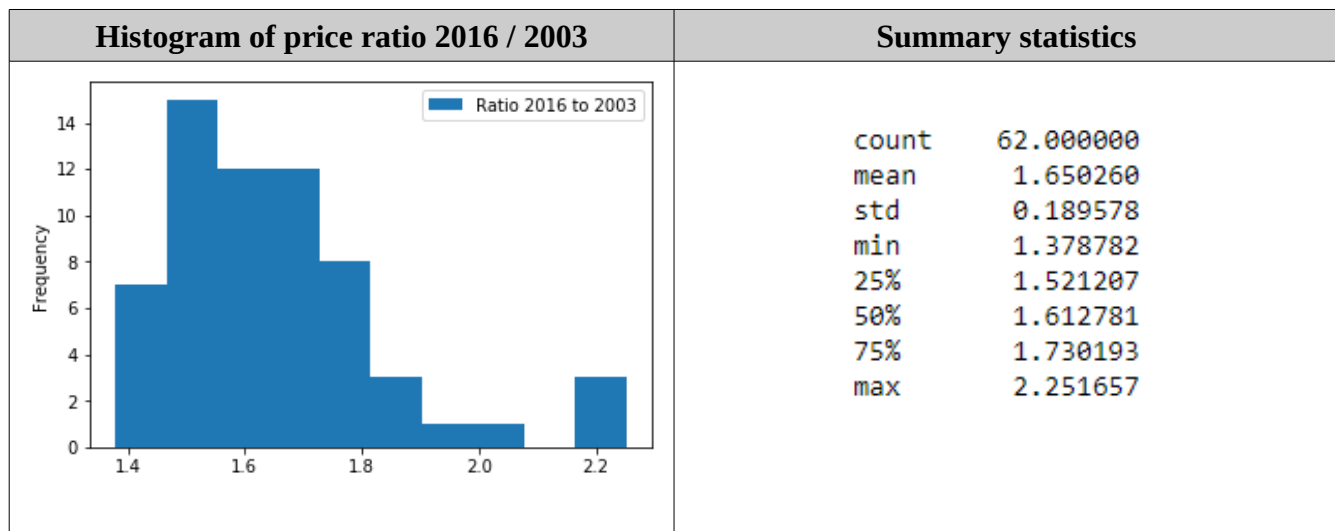
The final data set contains 62 rows and 400 columns, of which one column represents the city name and the remaining 399 columns are indicators of data type float.

For the correlation of FTSE100 data with average UK house prices, the House Price Index data set was used from www.gov.uk and merged with monthly data on FTSE100. The two data sets were merged together using dates as the joining parameter.

4 Exploratory data analysis

4.1 Prices shift from 2003 to 2016

The preliminary analysis involved examining the movement of house prices for the cities and understanding their progression over the period 2003-2016.

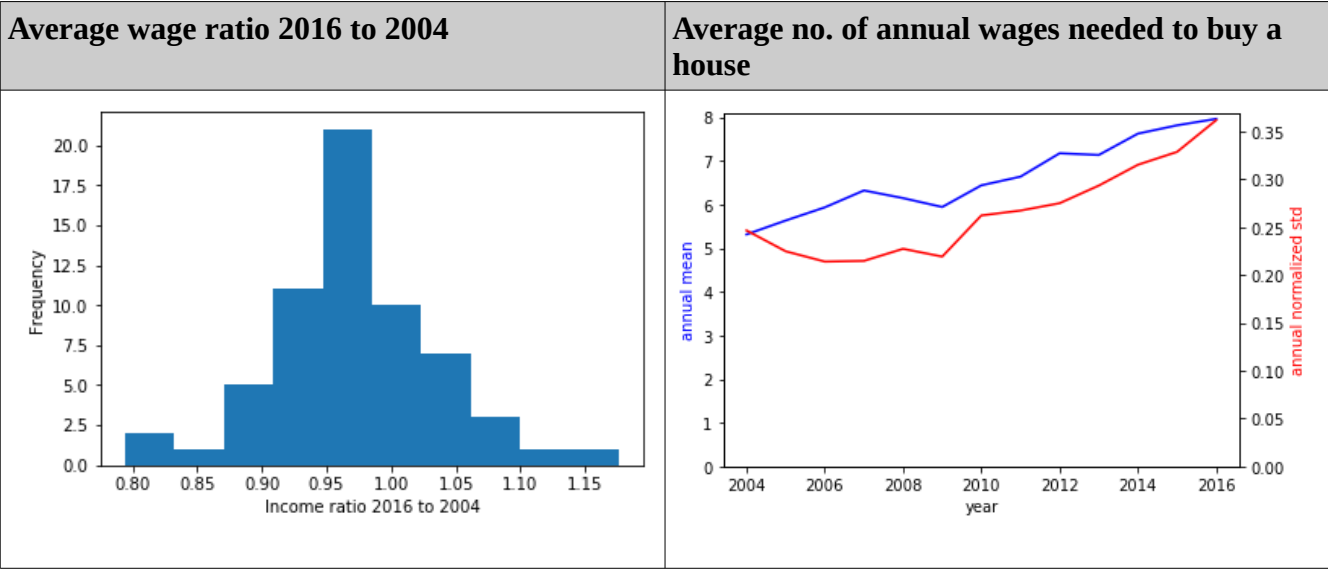


The analysis shows that prices have increased in all of the cities from 2003 to 2016. The highest increase was a factor of 2.25x, the lowest 1.38x and the mean 1.65x.

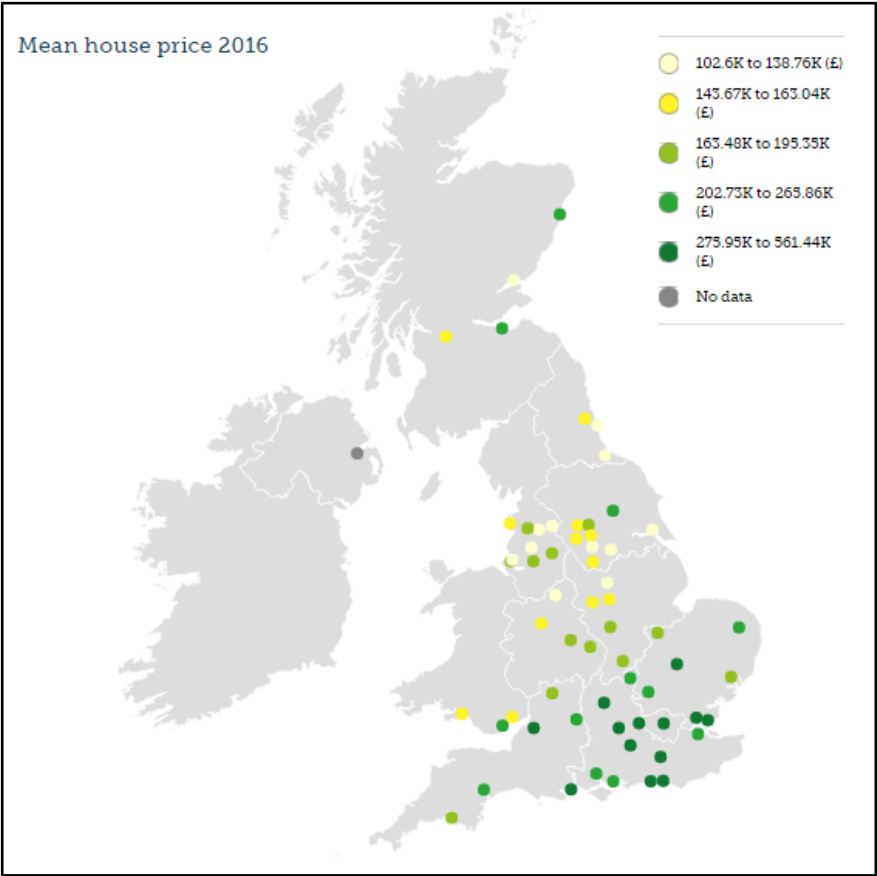
4.2 Affordability

Although house prices have risen by an average of 60% from 2004 to 2016, wages have on average remained nearly constant. As a result, it has become increasingly harder for an average worker to afford a house. Considering the 62 cities, on average 5.3 annual salaries were needed to buy a house in 2004, whereas by 2016 this figure has increased to 8.0 years. The standard deviation of prices

(normalized by the mean price in each year) has also increased, meaning that the variation between cities has become more extreme.

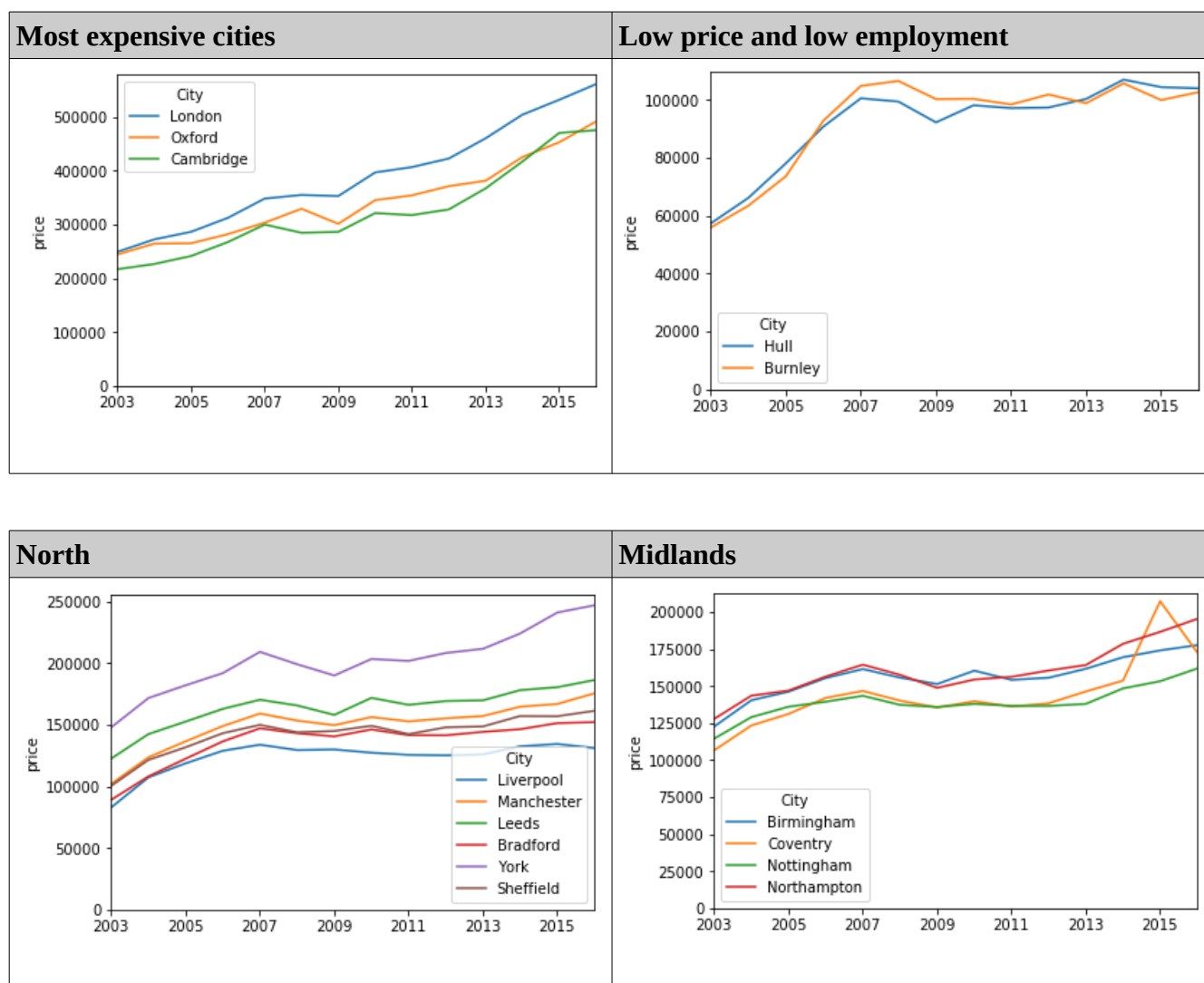


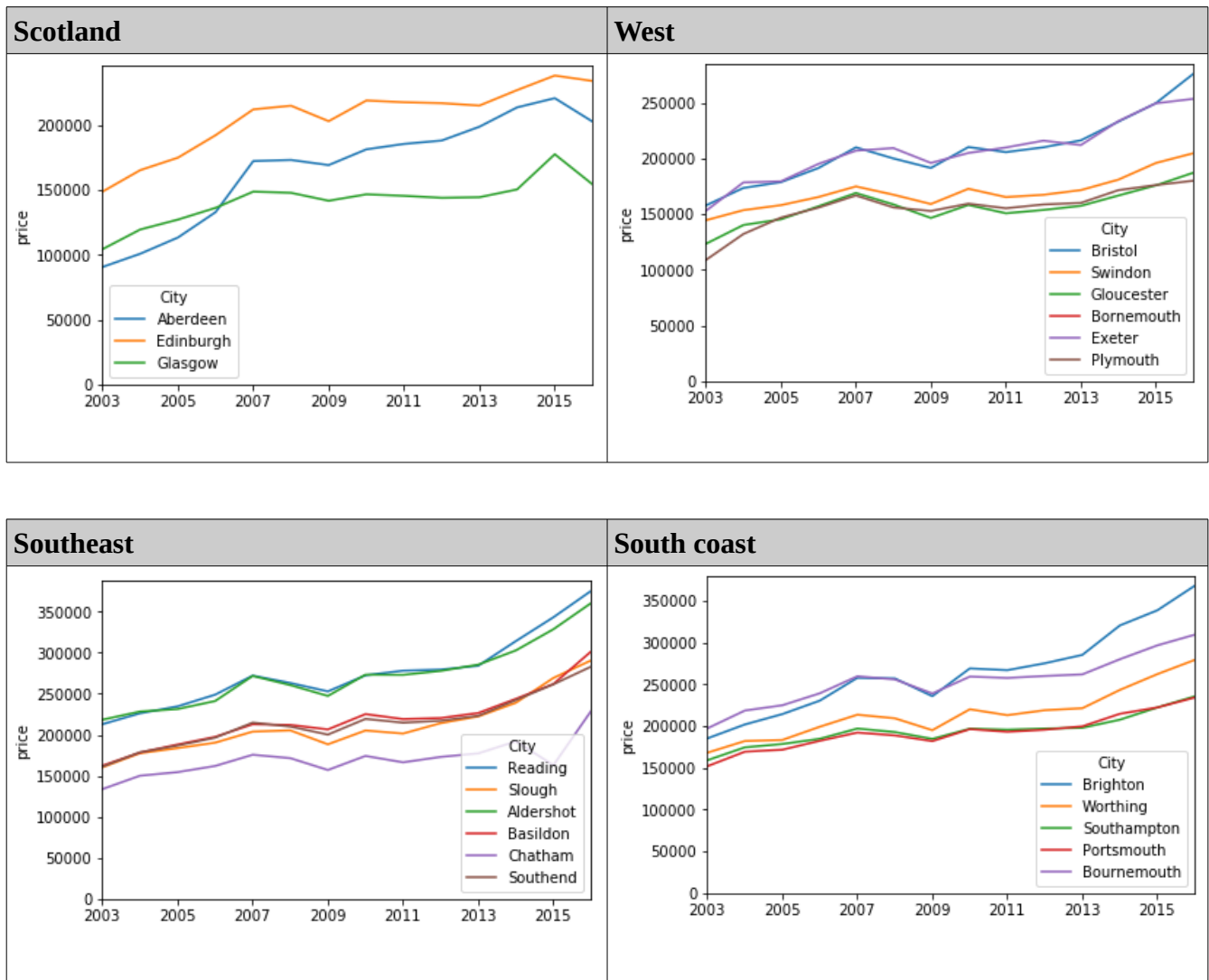
4.3 Location-based analysis



Location-based prices for the year 2016 from the Centre for Cities website are shown on the map above. This clearly shows the geographical distribution of prices, noting that the Southeast is the priciest (green dots), followed by a few cities on the South coast and the West. Cities get progressively cheaper the further north they are in the Midlands, while the northern industrial belt has considerable variation. A small number of towns in Scotland and the North have prices that match those in the Southeast.

The price movement of some specific cities are shown below:





The purpose of this project is look beyond location parameters and explore city data in order to identify factors that have influenced observed house prices in the given period.

5 Influence of economic / demographic factors

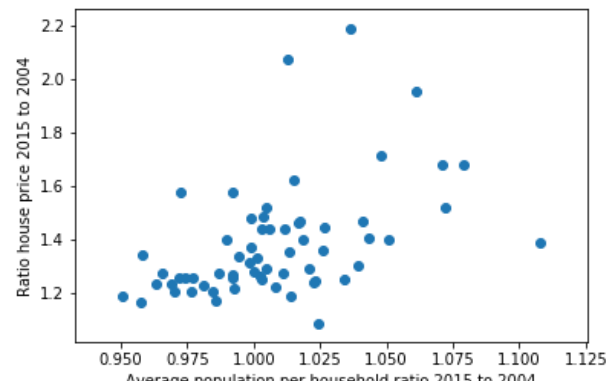
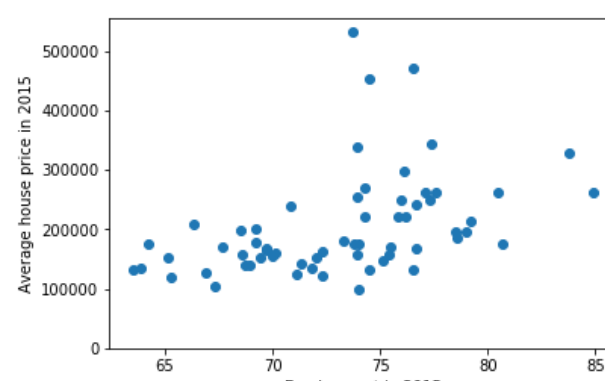
The influence of economic and demographic factors on house prices yields the following results:

- Wages do not have a statistically significant correlation to house prices
- The employment rate of a city correlates positively with house prices
- The ratio of population per unit of housing stock (

As seen from the chart below, there is no statistically significant relationship between wages and house prices, with a p-value of 0.92. Hence we can accept the null hypothesis and conclude that wages also do not have an effect on prices.

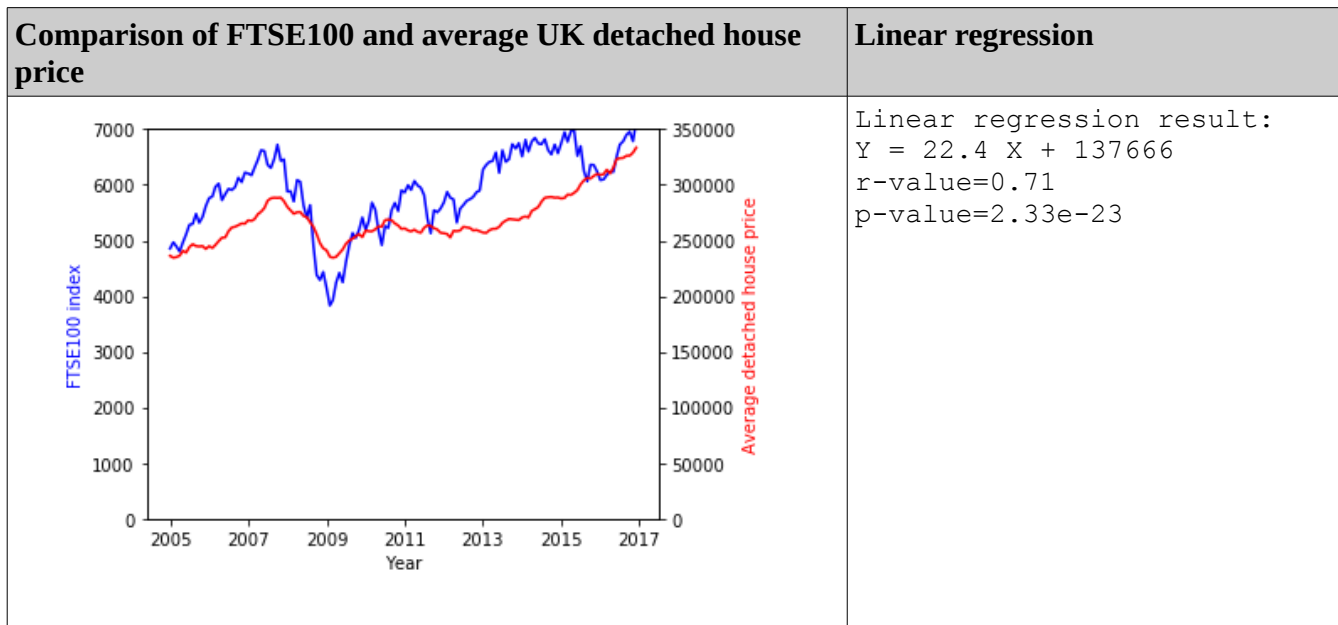
The results show a positive correlation, with a p-value of 0.0001, hence statistically significant to lead us to reject the null hypothesis. However the causality can be argued both ways, i.e. that a shortage of housing stock has the effect of pushing up prices, or alternatively that an increase in prices has forced a higher number of people to share a house on average.

The results for 2015 show that there is a positive correlation, with a p-value of 0.0005, hence statistically significant and we can reject the null hypothesis. The slope of the regression line corresponds to an average price increase of GBP 7,733 for each percentage point of higher employment. Apart from the three outliers that have average prices far above the average (London, Cambridge and Oxford), for the rest of the cities the employment rate is a significant factor in determining house prices.

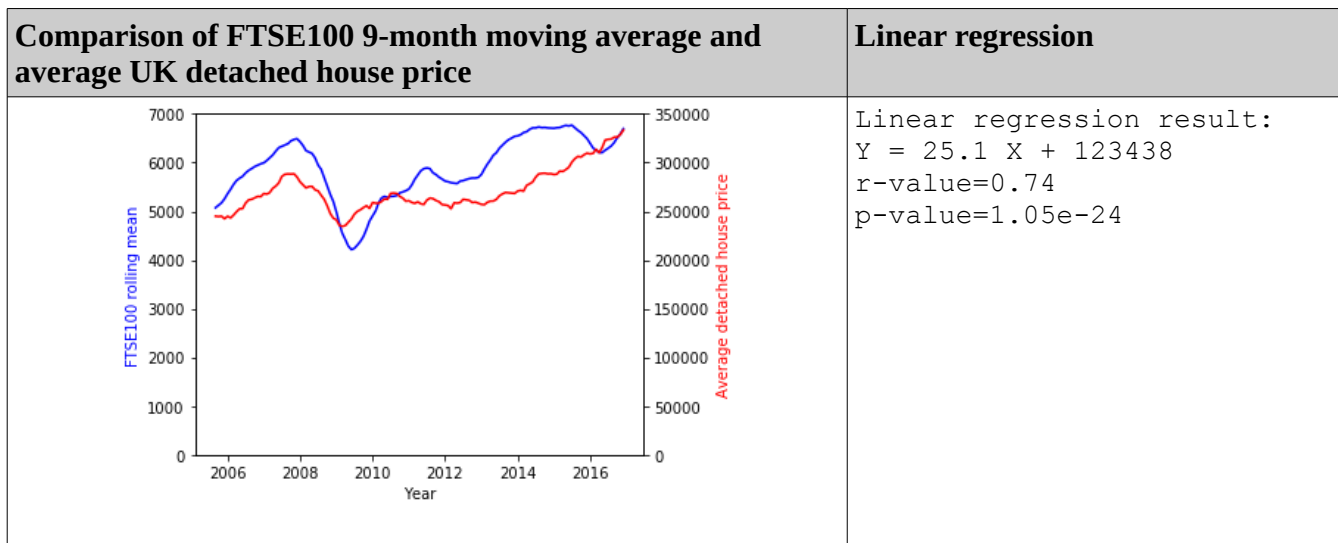
Ratio of population to housing stock	Employment
 <p>Linear regression result: $Y = 3.10 X - 1.75$ $r\text{-value}=0.467$ $p\text{-value}=0.000128$</p>	 <p>Linear regression result: $Y = 7734 X - 362057$ $r\text{-value}=0.430$ $p\text{-value}=0.000494$</p>

6 Correlation with the stock market

Further information on UK house prices is available from www.gov.uk, including the House Price Index (HPI) which tracks prices for different regions and different types of property (detached house, semi-detached house, terraced house and flat). This data provides an overall UK average, for each property type and the detached house UK average was used to compare prices with the FTSE100 index (noting that the four property types have 98-99% correlation between them so any of the types can be used).



Next, a moving average of the FTSE100 was compared with house prices as house prices seem to react slower than the stock market index.



Indeed, the FTSE100 9-month moving average provides a slightly closer correlation with the house prices, with a p-value of 0.74 vs. 0.71 before the moving average was taken into account.

7 Clustering Analysis

The conclusions of the statistical analysis lead us to explore if the cities exhibit a strong clustering behavior based on their economic fundamentals. As demonstrated in the analysis, the cities have varied

widely in terms of employment and housing supply/demand, while a few cities have been identified as outliers. A clustering exercise has been performed to answer these questions.

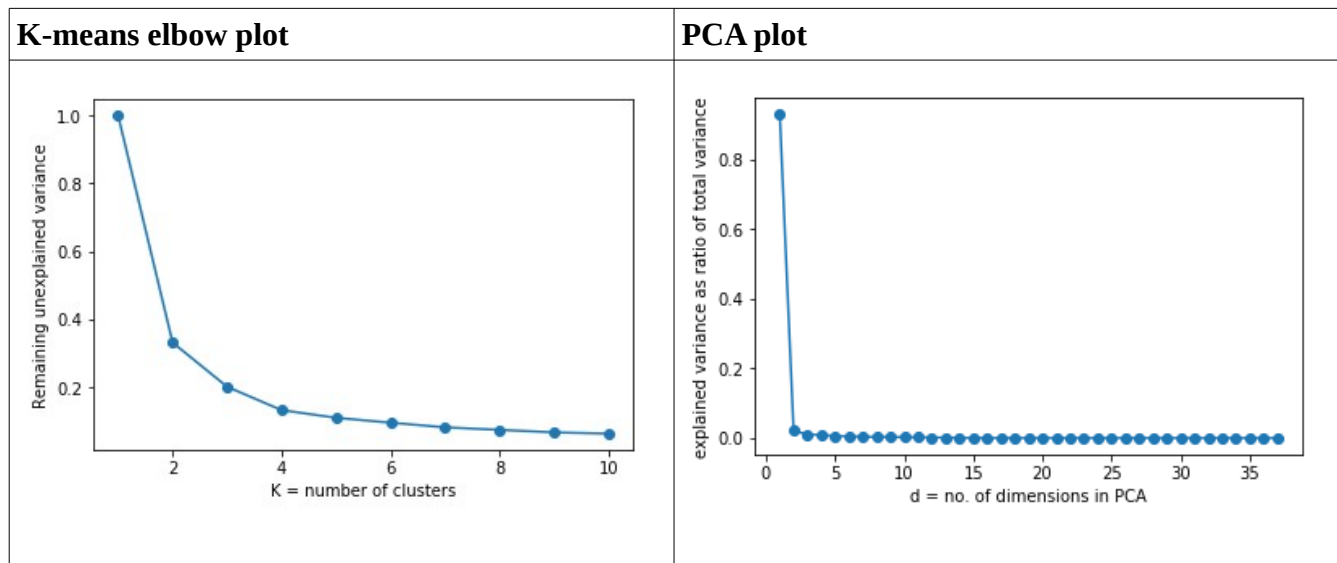
The parameters examined in the clustering exercise are:

- employment
- wages
- ratio of population to housing stock

For each of the parameters the data was sliced from 2004 to 2015. This data formed a features matrix of 36 columns (3 features x 12 years for each) and 62 rows observations (i.e. cities). The clustering analysis is therefore unsupervised, i.e. performed without feeding in any information on the prices.

7.1 K-Means

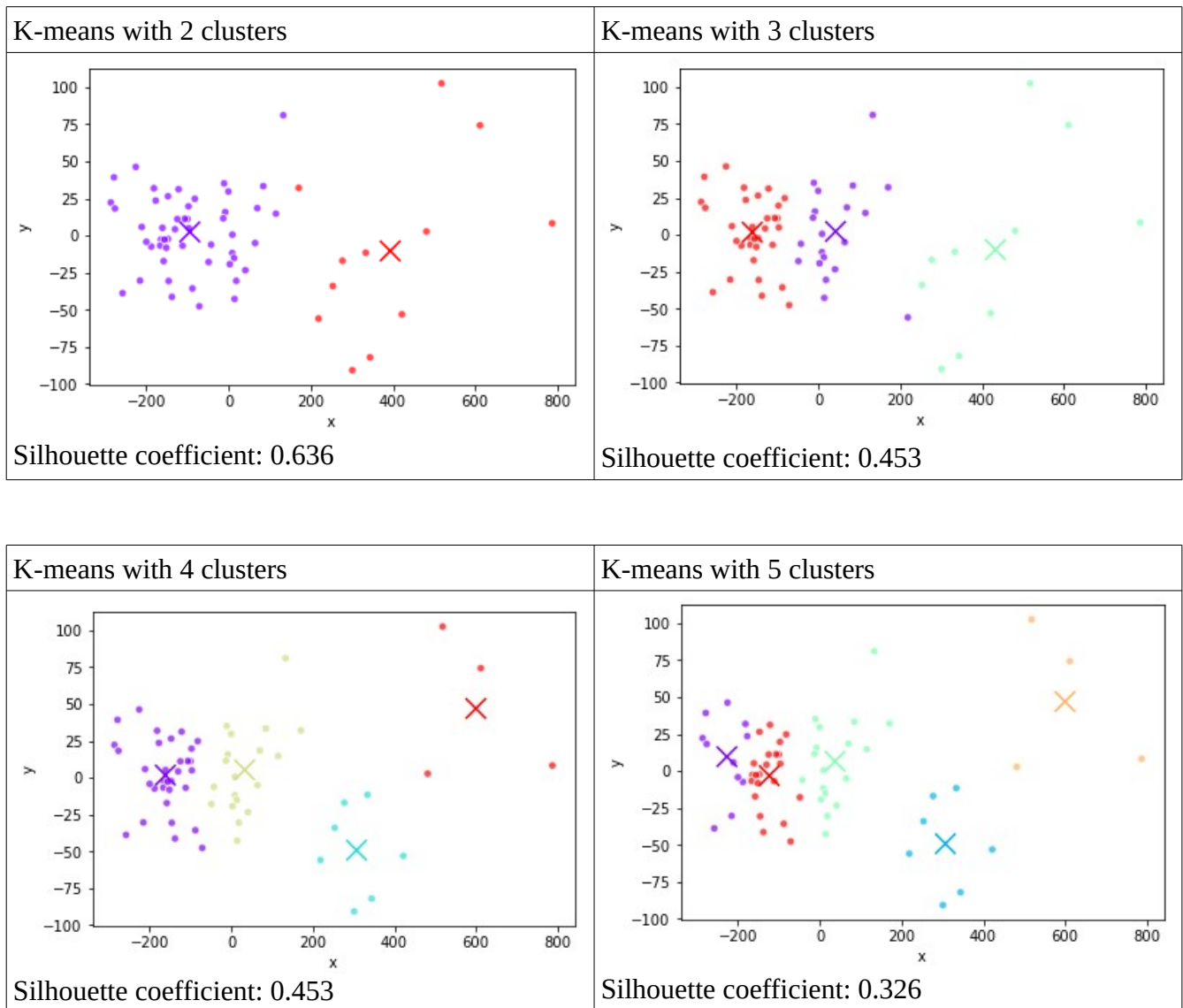
First, a K-means clustering was performed. In order to determine a reasonable value for the parameter K, an elbow plot was made on the data. In addition, a PCA plot was made to determine how well a 2-D plot explains the data.



The K-means elbow plot shows a well-formed elbow, with 2-5 clusters being a reasonable number of clusters to examine.

The PCA analysis shows that two dimensions explain almost all of the variance in the data. However this also leads us to question if there is sufficient variability in the data set or if the variance might be explained by a single feature.

The results of the K-means clustering are shown below:



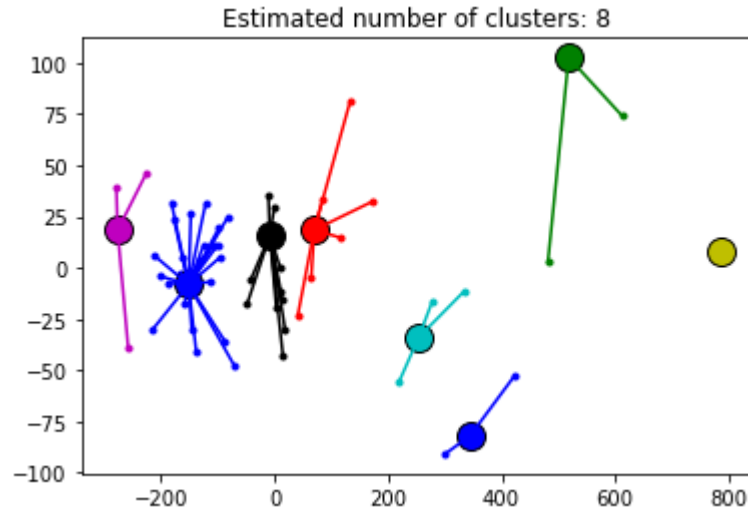
Based on this analysis, 2 or 3 clusters describe the data well and there is no benefit in increasing beyond 4 clusters.

7.2 Affinity Propagation

Further, an affinity propagation clustering algorithm was performed on the data set. This method determines the optimum number of clusters and is a suitable method for identifying groups that exhibit a similar behaviour pattern. Indeed, the scikit-learn documentation provides an example for grouping stocks using this method.

The affinity propagation method requires the tuning of some parameters, key of which is the ‘damping’ coefficient, which can result in the method yielding a different number of clusters depending on the value of the damping coefficient. In this case however the algorithm always produces 8 clusters, as demonstrated below.

Affinity propagation clustering

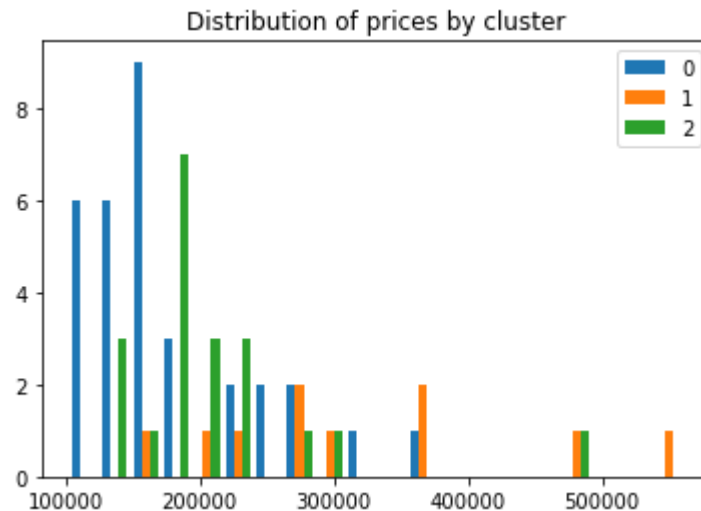


Silhouette coefficient: 0.278

Comparing the results of the two methods, for the given data set, the K-Means with 2 and 3 clusters are the most intuitive grouping outcomes.

7.3 Results of the Clustering Analysis

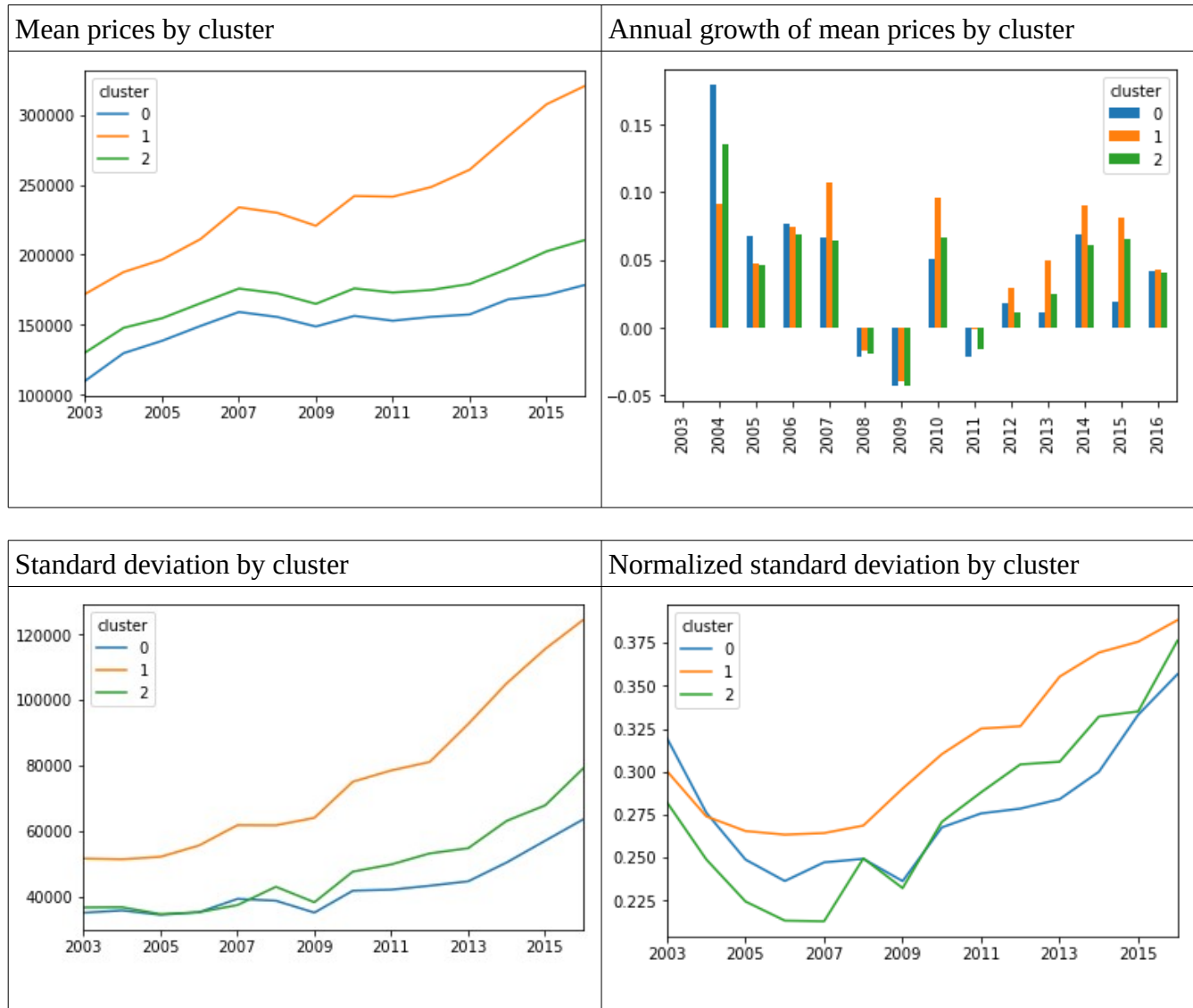
Histogram of 2016 prices by cluster for the 3-cluster K-Means method



The clustering was then applied to 2016 house prices to determine if the clustering based purely on economic fundamentals (from the features matrix) would translate into a notable difference in the house prices. The resulting bar chart shows that the clusters show some distinction in the pricing patterns, but there is also significant overlap between the clusters. Hence the clustering does not provide a conclusive result in terms of classifying cities by house price for one year alone.

However, looking at the mean and standard deviation of prices by cluster over the entire period provides more useful insight into the clustering. The difference between clusters may be summarized as follows:

- Cluster 0: low price and low variance cities that has a strong period of growth in 2004-2005 but struggled to recover at the same pace as the other two clusters post-2009
- Cluster 1: the high mean and high variance cluster containing the most expensive cities (London and its peripherals plus Cambridge).
- Cluster 2: priced above cluster 0 and with a stronger growth rate post-2009 (in particular stronger growth in 2015)



For the client, this exercise provides a mechanism to group cities into a small number of cluster and view its portfolio in terms of clusters rather than in terms of individual cities. Mortgage lenders are likely to view their portfolios as groups split by regions – this clustering method that is based on economic fundamentals. Hence the clustering may be useful for the following purposes:

- Assessing the portfolio exposure, risk and expected return in terms of clusters by analyzing the specifics of each cluster
- Fine-tune their strategy by using the clustering analysis for portfolio optimization purposes
- Define distinct policies for different cities based on the clusters they belong to

The clustering method and number of clusters can be selected to best suit the client's needs.

8 Linear Regression Modeling

For the client, the most useful tool arising from this exercise would be to determine a robust and reliable pricing model for each city based on its historic house prices and changes to economic fundamentals.

The ordinary linear regression model was used to test the predictive power of the data set under a variety of scenarios and determine the best model for predicting house prices in a given city based on known data.

For the linear regression model, the features matrix was augmented with FTSE 100 data, noting that this figure is the same for all cities. The average value of the FTSE 100 index was taken for each year.

8.1 Linear Regression on Economic Indicators Only

The first attempt to fit an ordinary linear regression (OLS) model involved using only the economic indicators for the features matrix, without feeding any price data:

In the observed year t , the OLS model was fitted using the formula:

$$Y_t = aX_t + b$$

where:

- Y_t is the vector of prices for each city in year t
- X_t is the features matrix for year t
- a and b are the fitted parameters in the OLS model

For the future year $t+1$, the prices were predicted using the following formula:

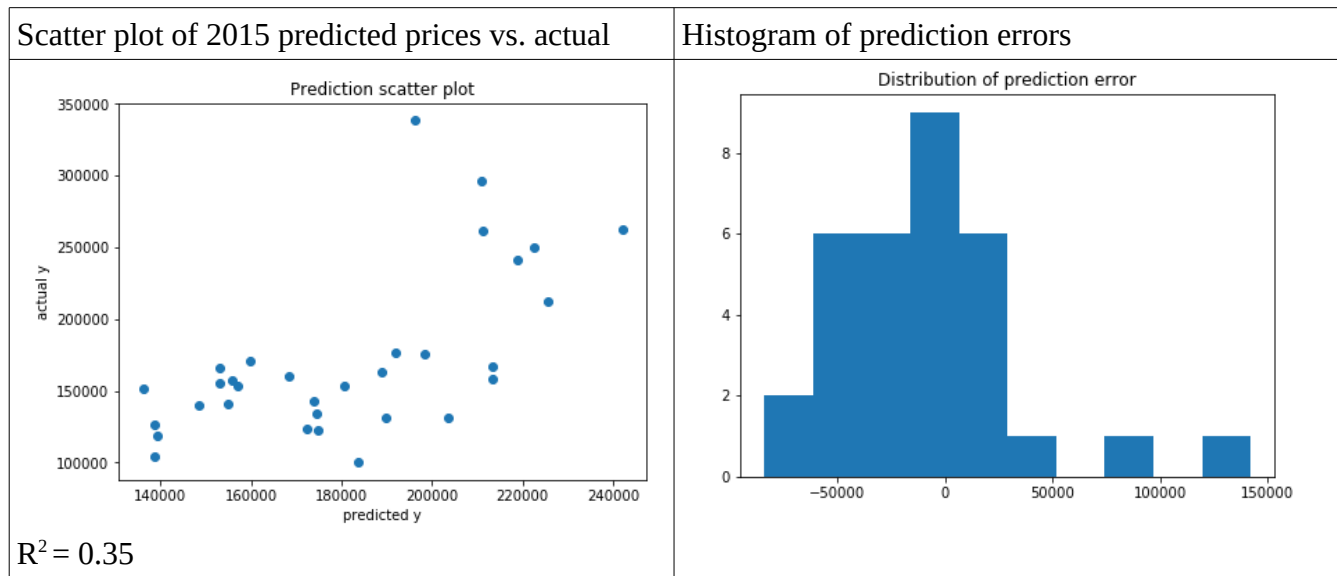
$$\hat{Y}_{t+1} = aX_{t+1} + b$$

where:

- \hat{Y}_{t+1} is the predicted vector of prices for each city in year $t+1$
- X_{t+1} is the features matrix for year $t+1$
- a and b are the fitted parameters in the OLS model from the previous year t

The model takes the actual observed X matrix from the year $t+1$ in the model, i.e. it looks into the future values of economic indicators. The reason for this is that we are trying to determine how well the economic indicators predict the actual house prices Y_{t+1} .

This analysis was performed for each available year and fitted to the next year. In each case, the scatter plot of \hat{Y}_{t+1} vs. Y_{t+1} was made, as well as a histogram of prediction errors. A typical result is presented below for Cluster 0:



The R^2 value ranges between 0.26 and 0.54 depending on the year. However, in some cases the prediction error is too large to be used as a predictive model.

Next, the log transform of features and prices was taken and an OLS model fitted. This gave only marginally improved results, with R^2 values between 0.30 and 0.58. For the 10 predicted years, the log-transformed model was better than the standard model in 9 out of 10 years.

9 Linear Regression on Economic Indicators and Past Prices

In order to improve the error margin, the features matrix was augmented to include prior year prices. Hence the trial matrix X_t contains the actual house prices Y_{t-1} .

A number of tests were performed in order to arrive at the best fitting model for house prices in year t :

Test 1: Using economic data for year t and prices for year $t-1$ to predict prices

Test 2: Using economic data for years t and $t-1$, plus prices for year $t-1$. The Idea here is to capture the change in conditions from year $t-1$ to year t

Test 3: Removing all economic fundamentals (wages, employment and population to housing stock) and using only the FTSE value for year t and house prices for year $t-1$ to fit prices for year t

Test 4: Removing the FTSE value from the features matrix and using only house price data for year $t-1$ as a predictor for prices in year t

In Tests 1-3 the results were taken as the score of the linear regression fit function (which gives the R^2 value), whereas for Test 4 the R^2 was calculated directly using the correlation function.

The test was performed on different clusters individually in order to achieve the closest fit.

9.1 Results of the Linear Regression Modeling

The results of the analysis are presented below:

Cluster 0	test1	test2	test3	test4
predicted year				
2007	0.9651	0.9506	0.9558	0.9858
2008	0.7067	0.8619	0.7339	0.9937
2009	0.9734	0.9327	0.9786	0.9937
2010	0.8142	0.8489	0.7955	0.9982
2011	0.8618	0.8875	0.8666	0.9985
2012	0.9750	0.9544	0.9870	0.9983
2013	0.9756	0.9687	0.9782	0.9952
2014	0.9858	0.9495	0.9868	0.9980
2015	0.9908	0.9883	0.9901	0.9962
average	0.9165	0.9270	0.9192	0.9953

Cluster 1	test1	test2	test3	test4
predicted year				
2007	0.9815	0.9808	0.9840	0.9965
2008	0.8226	0.8223	0.8268	0.9953
2009	0.9615	0.9629	0.9676	0.9921
2010	0.8118	0.8103	0.8178	0.9918
2011	0.8684	0.8866	0.8730	0.9969
2012	0.9731	0.9725	0.9741	0.9984
2013	0.9937	0.9928	0.9964	0.9985
2014	0.9534	0.9522	0.9547	0.9975
2015	0.9573	0.9604	0.9577	0.9909
average	0.9248	0.9267	0.9280	0.9953

Cluster 2	test1	test2	test3	test4
predicted year				
2007	0.9803	0.9741	0.9795	0.9904
2008	0.8462	0.7995	0.8586	0.9921
2009	0.8883	0.9107	0.9000	0.9953
2010	0.7589	0.7256	0.7694	0.9981
2011	0.8744	0.8619	0.8417	0.9975
2012	0.9893	0.9843	0.9882	0.9988
2013	0.9963	0.9964	0.9951	0.9992
2014	0.9751	0.9777	0.9746	0.9985
2015	0.9618	0.9602	0.9588	0.9831
average	0.9189	0.9100	0.9184	0.9948

It turns out that the best predictor of house prices in each case is Test 4, i.e. simply the vector of prices in the previous year. Although the other 3 tests also give good results, the dominant variable is the previous year's price and the other features introduce noise in the fitting process that amplify errors in the prediction.

10 Conclusions

The statistical analysis leads to the following key take-away remarks:

1. House prices across the 62 major cities in the UK have risen by an average of 65% from 2003 to 2016.
2. As wages have on average remained at the same level, owning a house has become increasingly difficult to afford for the average worker, needing 8.0 annual wages to buy a house compared to 5.3 annual wages in 2003.
3. Population on its own does not correlate to house prices. However, population taken in combination with the housing stock in a city does have a statistically significant correlation with house prices.
4. Employment is an important factor in determining house prices in a city.
5. The 3 most expensive cities (London, Cambridge and Oxford) can be considered outliers as their prices are not explained by the factors covered in this analysis.
6. In the UK as a whole, house prices are strongly correlated with the FTSE100 stock market index and follow the same pattern, but with less intense reactions and with a lag behind the

FTSE100 index. A 9-month rolling average of the FTSE100 index provides the closest correlation with the average house prices.

From the clustering and linear regression modeling analysis, the following can be concluded:

1. The selected set of features can be clustered in various ways using either K-means clustering or the affinity propagation method. There are no clear natural divisions between the clusters so that one clustering solution could be chosen as ideal.
2. In the selected 3-cluster division, the house price dynamics for each cluster do reflect some distinct features, but there is still significant overlap between clusters to make a useful pricing classification. The clustering can be useful to the client in terms of an alternative method for assessing their portfolio as cities within a cluster do show some similarity in behaviour. This may be a better method of grouping cities compared to e.g. grouping by region (although this was not tested here).
3. In terms of predicting future house prices, the most accurate predictor is the previous year's prices.
4. Although there is a degree of correlation between house prices and indicators such as the FTSE 100 index, employment figures for a city and the ratio of population to housing stock in a city, adding these features to the linear regression model does not improve the results, rather it introduces noise in the predictions that increases the prediction error.
5. As a possible future improvements to the model, the following can be tested:
 - 5.1. A more sophisticated algorithm, such as the random forest method, may be more successful in identifying the effect of other features besides the prior year price
 - 5.2. Testing over smaller / larger cluster sizes and several years of test data