A Visual Analytics Approach to London Housing Prices

Abin Abraham

Abstract: We analyse London Housing Price data looking for distinguishing patterns in housing price using a visual analytics approach. Focus is on analysing housing prices and its relationship with different demographic, social and economic variables across space and time, while highlighting the implications for potential investors in real estate. Subsequently we identify the drivers behind the housing prices using a regression analysis. Throughout this paper, we explain the interplay between visual analysis techniques and computational methods to gain knowledge.

Index Terms : Visual Analytics , London Housing Prices, Multi variate analysis, Regression Analysis

1 Problem Statement

London's house prices are one of the highest in the world. This paper aims to consider the level of London's house prices from an visual analytics perspective. It looks at the fundamental factors that drive the housing market and tries to analyse it using socio-economic associations.

The dataset is combined from 3 sources- ward profiles, wellbeing scores and land registry house prices of London. Though this analysis is focused on multivariate analysis of housing prices, it is fundamentally a human behaviour problem as housing prices are driven by the market based on demand and supply. The sourced data is suitable for our analysis as it encompasses data regarding housing prices and ward profile indicators – social, economic, safety, environment and transport which are key to understanding human needs.

This paper aims to perform analysis and arrive at interpretations to help an investor in housing in London through research questions such as:

1. Are there places in London which are more expensive in terms of housing prices than others?
2. Can we cluster housing prices based on the data and thus device strategies for investment?
3. To what extent has there been a rise in housing prices in London over time?
4. Can we explain the housing prices in terms of the characteristics of the population?
5. To what extent can a single socio-economic estimator for housing price be determined?

We will perform our analysis using a visual analytics approach and iteratively improve our understanding based on visualization and models to answer the research questions.

2 State of the Art

With the rapid growth of information, the data has become ubiquitous, complex and multi-dimensional, resulting in the need to introduce visual analytics to tackle these complexities[1]. A phenomenon with more than 2 explanatory variables is termed as multi-variate and they presents challenges to visualization, due to increase in dimensions.

The most common techniques in Multivariate analysis are: Multiple Regression Analysis, Parallel coordinates, scatter plots, Discriminant Analysis, Factor Analysis, Cluster Analysis, Principal Component Analysis. We will explore the state of art in visual analytics for housing prices, followed by novel techniques proposed to tackle problems with the existing techniques in multivariate analysis.

There has been an exclusive study on the visualization for real estate market in Australia by Mingzhao et al where an interactive visual analytics system is proposed [2]. They have identified that geographical related information is seldom captured in existing systems, and systems are not 'location aware' with different level of detail. Their system has 5 level details - basic, transportation, facilities, education and census profiles for a location displayed over in space using choropleth maps, dot maps, stream graphs, spider chart and glyphs on maps which are at multiple levels of granularity-regional, suburb and property. Shengwen et al [3] has performed a spatio-temporal analysis of housing prices in China and identified an underlying clustering in pattern of changes in house prices. Nearest neighbor analysis and geographic weighted regression are used to emphasize that multivariate relationships vary over space. Data from social media has been used to build an association with housing prices. Most, if not all the visualization methods, are applicable for our project as well. We will be exploring the association of housing prices with similar details using spatial and regression analysis tools.

To combat the limitations of existing tools for multi-variate analysis which mostly assume normality of distributions and their inability to understand the inherent influencing dimensions for the phenomenon, Turkay et al[4] has proposed a dual visualization model for simultaneous visual analysis of multivariate datasets using dimensions and actual data in what they term as 'items space' and 'dimensions space. The data set they have used is pertaining to DNA microarray analysis, but a sample analysis has been done using the Boston neighborhood housing dataset which has data regarding median housing price value, crime rate, economic status. Analysis is performed iteratively by selecting items and dimensions, from selecting specific set of data in focus using 'brushing operations' and later more interpretation are built using 'focus+context' visualizations. A selection in either of the spaces updates the other helping better understanding of high dimensional data.

A common technique used in multi-variate analysis is pair wise scatter plots. Their simplicity compared to other multivariate analysis techniques, familiarity among users, and clarity of visualization makes them quite popular. A key disadvantage is that a single scatter plot can only reliably visualize few dimensions which may not be practical for many realistic datasets. Elmqvist et al has proposed a technique employing 3D animated transition 'graphics as well as point color, shape, and size as graphical properties, for visual exploration of multi-dimensional data in scatterplot matrices' [5].

3 Properties of the Data

There are three datasets which we will use for our analysis and they have majorly ratio and interval datatypes with ward codes in the nominal datatype.

The first dataset consists of ward wellbeing scores sourced from the government website for London data based on data between 2009-2013[7]. These ward level well- being scores represent an aggregated calculation of well-being indicators of the resident population based on 12 indicators across health, economic insecurity, safety, education, transport and environment. Each score is relative to the average score for England and Wales which is considered as 0. A positive score implies a greater than average score while a negative score implies a below average score.

The second dataset consists of ward profiles sourced from the government website for London data based on a study conducted in 2014[8]. The ward profiles provide information on demographic and other related data for each ward in Greater London. They describe the population in terms of parameters like diversity, household income, life expectancy, housing, crime, benefits, land use, deprivation, and employment. The third dataset has the land registry prices in London from 1995-2014 from London Datastore. All three datasets are linked using the 'New ward code' and has the lowest granularity at that level.

Firstly, we combine the different data sources ward profiles and wellbeing scores to consolidate data for our analysis, where each row represents an observation for the ward and column its attribute. After merging the two data files for using new ward codes, the main dataset of 659 rows corresponding to each ward in London and 128 columns are obtained. It has features which are only of ratio and interval data-types. Next, land registry house prices which has separate columns for each year were transformed separately by reshaping from wide to long. All the transformations and data cleaning were performed using Python.

There are no missing values in the dataset. Columns not relevant for our analysis were dropped and null values were filled with median values in relevant columns. We also standardised row header names. Data was later normalized using Standardization on a [0, 1] Scaling.

Viewing the data distribution of housing prices in London.

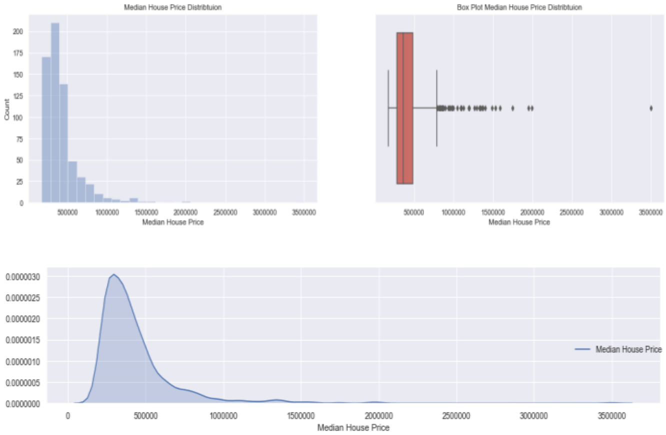


Figure 1 Histogram, Box-Cox plot and Kernel Density plot for housing price

Based on the histogram we can see that the median value for all wards in London is around $365000 with data being almost normal in distribution. Minimum being $173000 and maximum $3.5 million We have quite a large number of outliers. We will handle this right before our modelling so that we can observe the impact visually. We will remove 44 records with housing price > $792500 which was the obtained after using the interquartile method where outlier critical value is defined as Q3 + 1.5 * IQR where Q3 refers to the 3rd quartile and IQR refers to the difference between the 1st Quartile and 3rd Quartile.

4 Analysis

4.1 Approach

Our objective is to answer the research questions using a visual analytics approach. We will employ tools like Tableau and Python which will be used for iterative computational analysis with an interactive interface , backed by analytical reasoning, to derive visualizations and interpretations through multiple feedback loop process as described by Keim et al. [6].

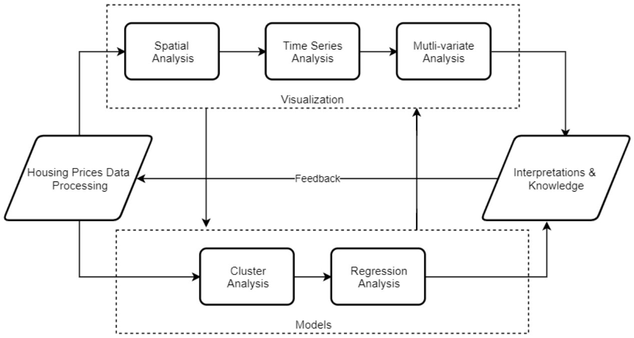


Figure 2 Visual analytics approach

We will list down the identified tasks in the subsequent sections. Just as the workflow suggests, we switch between visualization and models to better understand the phenomenon as required and thus many of the steps overlap with the next.

4.1.1 Data Processing

3 datasets are merged using the ward code as key followed by outlier detection and normalization. All categorical variables have been considered as dimensions and numerical variables as measures for exploration in Tableau. Shape file with ward details as of 2011 is loaded to build the map of London. Though 128 columns are available our focus will only be on- 'Median House Price (£) - 2014','Population - 2015','Median Age - 2013','Population Density- 2013','Rate of All Ambulance Incidents per 1,000 population - 2014','Number of jobs in area - 2013','Median Household income estimate (2012/13)','Crime rate - 2014/15','Average Public Transport Accessibility score - 2014','Unemployment rate 2013','% with Level 4 qualifications and above - 2011','Homes with access to open space & nature, and % greenspace - 2013' and yearly prices from 1995-2014.

4.1.2 Visualization of patterns

Spatial analysis

Choropleth maps are used to understand spatial distribution of housing price using diverging multiple hues at borough and ward level and thus identify region with high/low prices.

Time series analysis

Time series data analysis will be used to understand the trends in pricing growth and drill down to identify specific outlier in patterns over the years.

Multivariate analysis

The pairwise scatter plots and correlation matrix help to visually determine the extent of the relationship of the socio-economic variables with the housing price and features to retain i.e. not highly correlated. The correlation coefficients will quantify it and contribute to the regression analysis.

4.1.3 Computational Models

Cluster analysis

We will explore building clusters using a partition-based clustering method like K-Means to group wards based on similarity in housing prices. To determine the optimum number of clusters we employ techniques which are both computational and visual – silhouette analysis with scores to determine the level of consistency within a cluster and elbow method using distortion scores. The results will be plotted on a choropleth map.

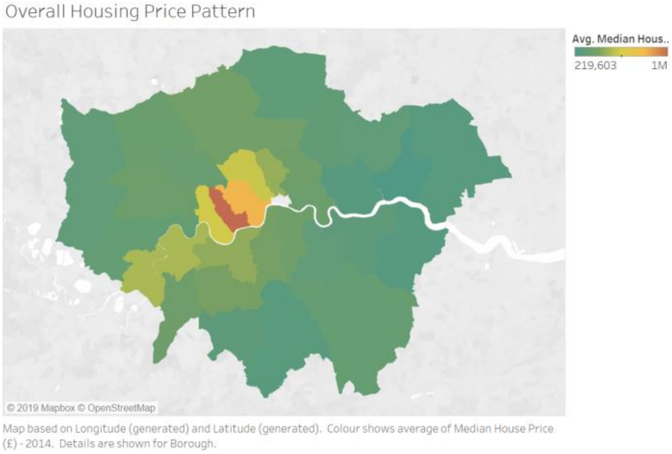
Regression Analysis

Regression Analysis will be used here for inferential statistics to better understand the phenomenon of housing prices. We will try multiple linear regression as a computation model using Ordinary Least Square regression (OLS) to refine our understanding of the housing prices and their association with socio-economic variables. Finally we attempt to estimate the housing prices using a single predictor from the previous results which can serve as an alternative to an investor.

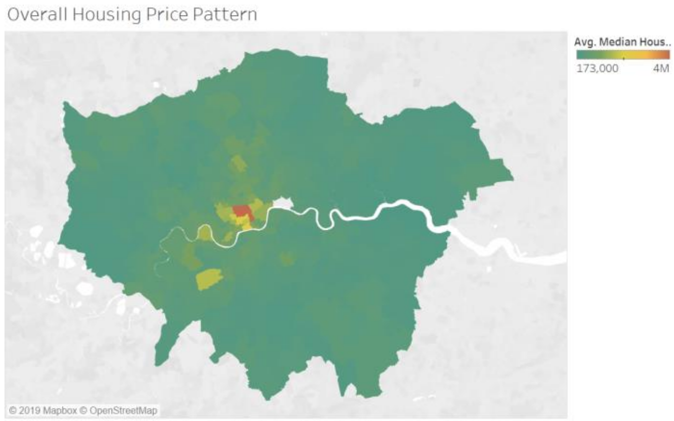
4.2 Process

4.2.1 Spatial Analysis

We start by visualizing the spatial distribution of median housing prices at a borough level using a choropleth map. Choropleth map is the most commonly used and effective method for summarized spatial information with geo-shapes.


Overall Housing Price Pattern

Map based on Longitude (generated) and Latitude (generated). Colour shows average of Median House Price (£) - 2014. Details are shown for Borough.



Map based on Longitude (generated) and Latitude (generated). Colour shows average of Median House Price (£) - 2014. Details are shown for Borough and Ward name.

Figure 3 Choropleth map for median housing price in Greater London i. Borough level ii. Ward level

We can observe that the prices range from $0.2M at Barking and Dagenham to $1.2M at R B of Kensington and Chelsea. The central boroughs have a higher median price as compared to the outer boroughs and the price seems to follow a pattern of gradual increase as we move from the periphery to Central London which can be attributed to the proximity to the city. But is this even the case at ward level? Let us dig one more level deeper to observe the median housing price at ward level and we can see that there is a spike in housing prices in certain ward pockets alone instead of the whole central borough. Westminster- Knightsbridge and Belgravia is the highest median housing price ward, but this is not in the highest median priced borough which is RB of Kensington and Chelsea. Thus, our initial understanding from borough level housing price pattern has been updated with the new knowledge at ward level. If we had not drilled down to ward level, we may have made the incorrect interpretation that all the central boroughs were expensive.

4.2.2 Clustering

Let us try clustering using K-means and assessing it using the computational model - silhouette coefficient measures how similar an object is to the other objects in its own cluster versus those in the neighbor cluster[11]. The silhouette score (SS) values range from 1 to – 1. A value of SS close to 1 indicates that the dataset is well clustered and there is clear separation. We can infer that the object is similar to the other objects in the same cluster. A value of SS close to -1 indicates that the dataset is poorly clustered and there doesn't seem to be a clear similarity between objects in the same cluster and that shift to another cluster may improve the overall score.

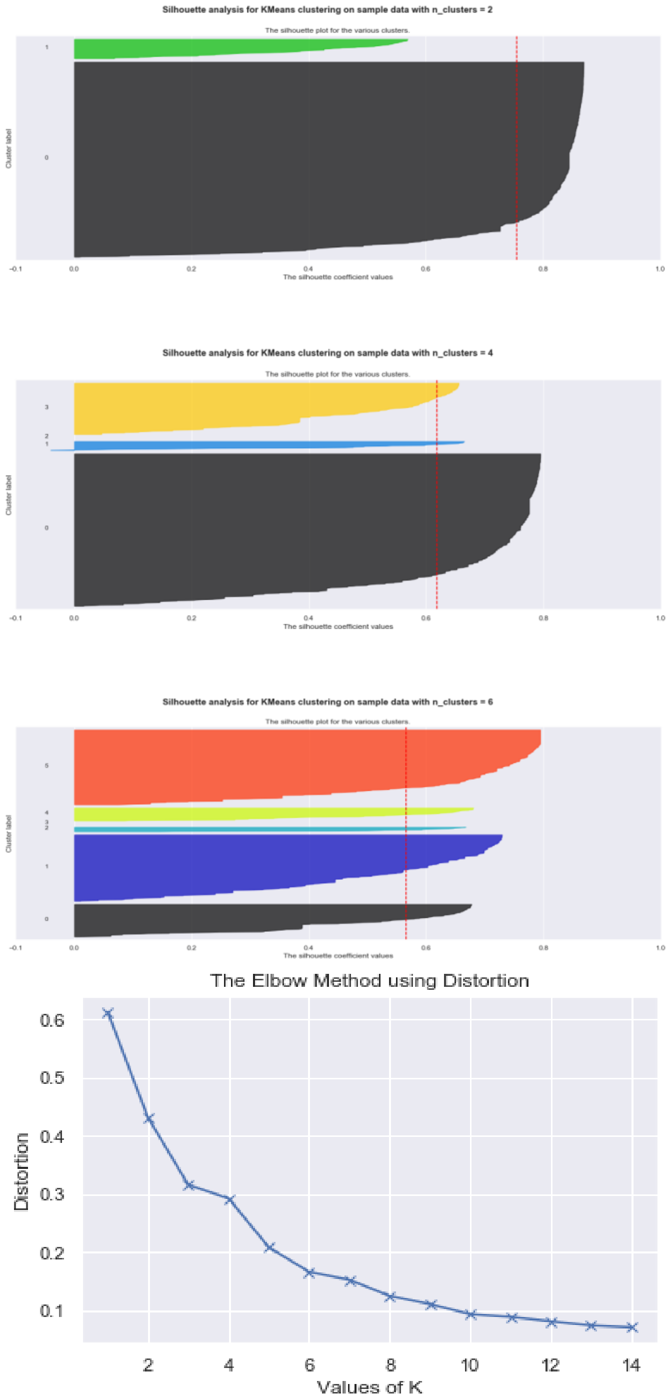


Figure 4 Cluster analysis – i,ii,iii Silhouette plot for 2,4,6 clusters iv. Elbow method for determining the K value

Based on the silhouette plots we may be tempted to opt for the highest value for the SS which is for 2 clusters.

```
For n_clusters = 2
The average silhouette_score is : 0.7548883906729306
For n_clusters = 3
The average silhouette_score is : 0.632061742000422
For n_clusters = 4
The average silhouette_score is : 0.6185841491895102
For n_clusters = 5
The average silhouette_score is : 0.5784380455269241
```

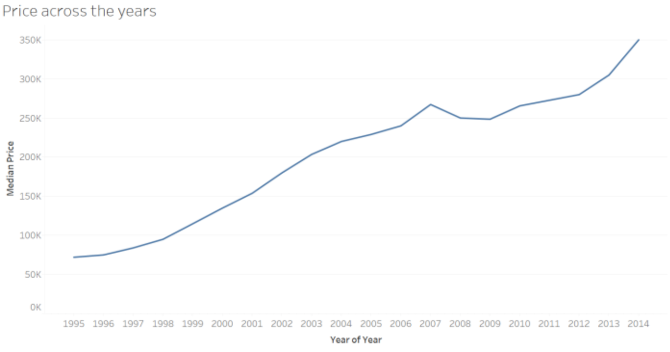
```
For n_clusters = 6
The average silhouette_score is : 0.5658653597736717
For n_clusters = 7
The average silhouette_score is : 0.5687235863346455
For n_clusters = 8
The average silhouette_score is : 0.5761307635046684
For n_clusters = 9
The average silhouette_score is : 0.578439656974032
For n_clusters = 10
The average silhouette_score is : 0.5664222203018969
```

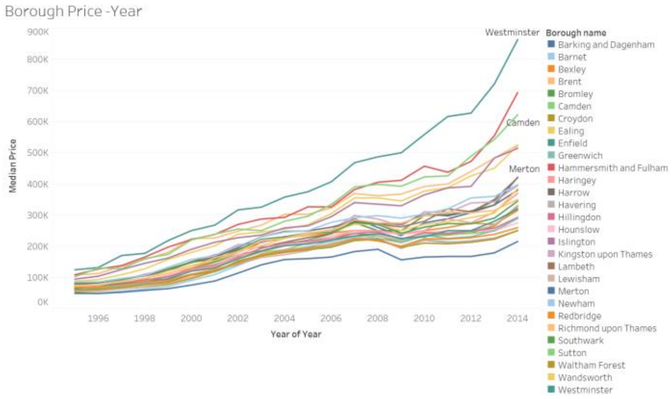
But this does not provide a reasonable cluster which can be beneficial for recommendations while visualizing spatially. In addition, for 4 clusters there is a negative value for one of the clusters. Employing the elbow method we can see that the optimum cluster value is 6 post which the distortion is continuously diminishing. Thus we have employed both computational model and visualisation to arrive at a decision on the number of *clusters. The Silhouette score of 0.5658 though may not* be the highest but there are no negative scored clusters in the silhouette plot unlike for 4,7, clusters.

Hence, we decide to opt for 6 clusters and observe the choropleth map (depicted in the section for 'Results'.) We have a clear ward level clusters to understand how the housing prices vary now and can group wards based on clusters. Thus an investor can shortlist multiple wards based on their budget which will falls into a cluster.

4.2.3 Time Series Analysis

Next, we analyze the housing prices across time. The general trend across London has risen over the years with a blip in 2008 and 2009 which can be attributed to the recession. We subsequently drill down the median house prices to borough and ward levels





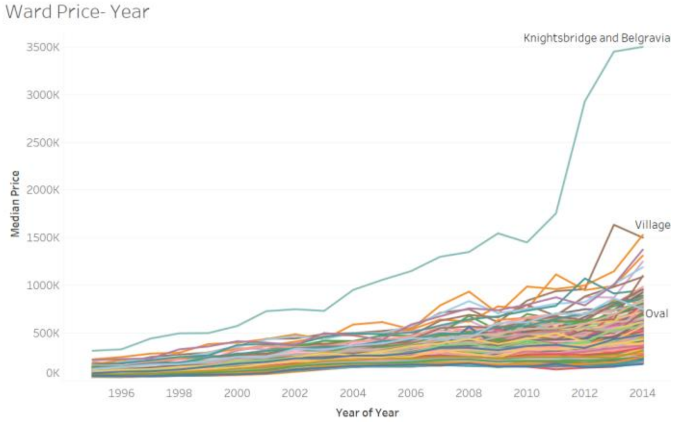


Figure 5 House prices trend over time 1996-2014 i. Overall London ii. Borough Level iii. Ward Level

We can see that though there has been a rise in prices across the years, some boroughs have more steeply risen – Westminster, Camden, Hammersmith and Fulham and some remain stable- Barking and Dagenham. Drilling down one more level, we can see that we have a total outlier in price rise in the ward of Knightsbridge and Belgravia. Has there always been a price rise – how did the boroughs fare in 2009 recession?

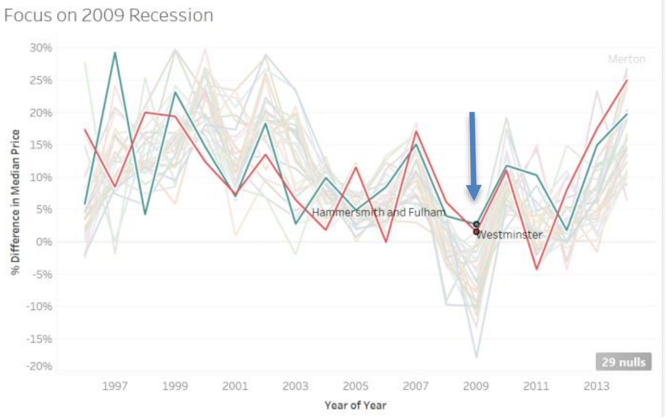


Figure 6 Percentage difference in housing prices 1996-2014

Only the boroughs of Westminster and Hammersmith and Fulham did not have a depression in prices during the recession of 2009, while all the rest did. In addition, Hammersmith and Fulham has never had a depression in prices while Westminster did in 2011.

4.2.4 Multi-variate Analysis

Next let us try to understand the factors that influence housing price by taking a scatter plot between various factors – median income, crime rate, public transport, unemployment, education level, open space access, population, median age, population density, ambulance incident rate and number of jobs.

Pair-plots to understand the relationship between the features

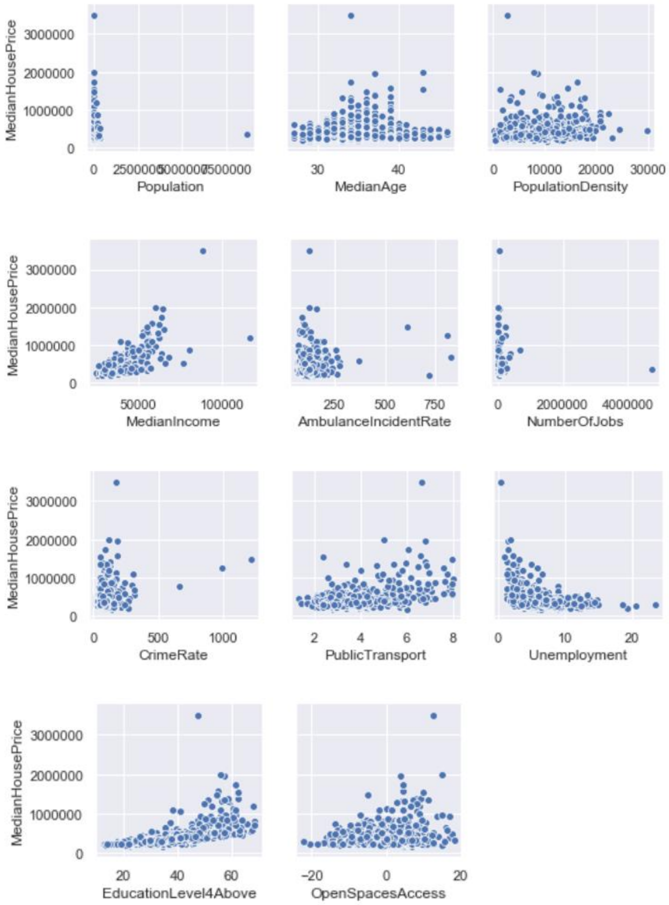


Figure 7 Pair-plot of housing price with socio-economic variables

The relationships here are interesting. Relative levels of median income, public transport, degree-level educational attainment is most correlated with housing prices. Median age, ambulance incident rates and crime rates appear quite random with no distinct pattern to housing prices. We can see how the outliers may distort our analysis. So, we remove it based on details found in our initial data analysis.

We can try to build a multivariate regression model to take account of the effects of these variables on the housing price. Multivariate linear regression assumes that there is a linear relationship between the feature and target, features and target understudy follow a normal distribution, lack of collinearity between features and homoscedasticity.

Based on the scatter plot is it is established that there is relationship between majority of the features with the housing price. Through the regression model we attempt to understand to what degree of explainability can each of the socio-economic variables contribute to the housing prices.

Collinearity between each of the features and dependant variables is determined.Let us quantify the extend of the relationship with a Pearson's correlation matrix.

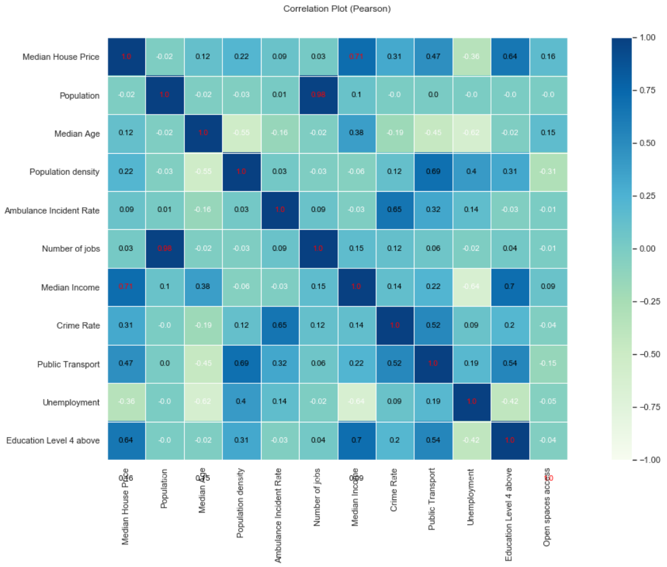
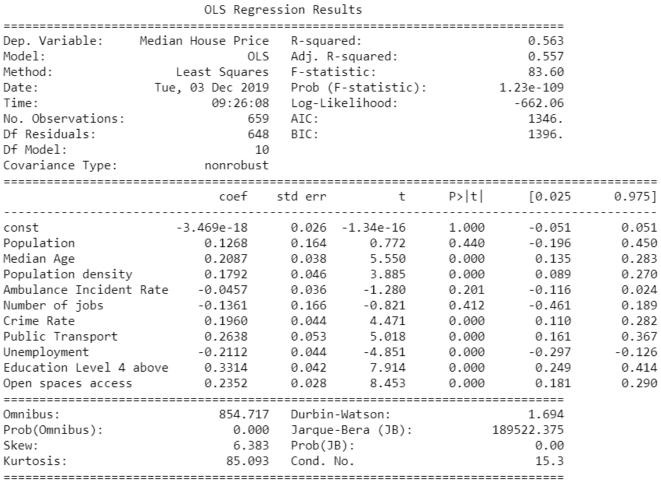


Figure 8 Correlation matrix for housing price and socio-economic indicators

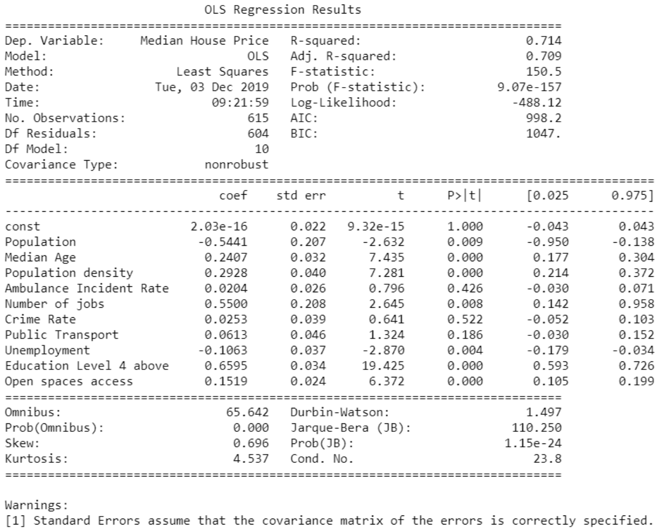
We see that median income is highly correlated with the housing price with person coefficient of 0.71. We will have to eliminate this before we proceed with building computation models.

Next, we build a computational model to better understand the relationship using ordinary least square method (OLS) of multiple regression. The focus will be on model building to identify the impact of features listed. The next two tables compare our results with the outliers and without outliers on the regression.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Median House Price   R-squared:                       0.563
Model:                            OLS   Adj. R-squared:                  0.557
Method:                 Least Squares   F-statistic:                     83.60
Date:                Tue, 03 Dec 2019   Prob (F-statistic):          1.23e-109
Time:                        09:26:08   Log-Likelihood:                -662.06
No. Observations:                 659   AIC:                             1346.
Df Residuals:                     648   BIC:                             1396.
Df Model:                          10
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                   -3.469e-18      0.026  -1.34e-16      1.000      -0.051       0.051
Population                  0.1268      0.164      0.772      0.440      -0.196       0.450
Median Age                  0.2087      0.038      5.550      0.000       0.135       0.283
Population density          0.1792      0.046      3.885      0.000       0.089       0.270
Ambulance Incident Rate    -0.0457      0.036     -1.280      0.201      -0.116       0.024
Number of jobs             -0.1361      0.166     -0.821      0.412      -0.461       0.189
Crime Rate                  0.1960      0.044      4.471      0.000       0.110       0.282
Public Transport            0.2638      0.053      5.018      0.000       0.161       0.367
Unemployment               -0.2112      0.044     -4.851      0.000      -0.297      -0.126
Education Level 4 above     0.3314      0.042      7.914      0.000       0.249       0.414
Open spaces access          0.2352      0.028      8.453      0.000       0.181       0.290
==============================================================================
Omnibus:                      854.717   Durbin-Watson:                   1.694
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           189522.375
Skew:                           6.383   Prob(JB):                         0.00
Kurtosis:                      85.093   Cond. No.                         15.3
==============================================================================
```

Table 1: Regression with outliers included



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Median House Price   R-squared:                       0.714
Model:                            OLS   Adj. R-squared:                  0.709
Method:                 Least Squares   F-statistic:                     150.5
Date:                Tue, 03 Dec 2019   Prob (F-statistic):          9.07e-157
Time:                        09:21:59   Log-Likelihood:                -488.12
No. Observations:                 615   AIC:                             998.2
Df Residuals:                     604   BIC:                             1047.
Df Model:                          10
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                     2.03e-16      0.022   9.32e-15      1.000      -0.043       0.043
Population                 -0.5441      0.207     -2.632      0.009      -0.950      -0.138
Median Age                  0.2407      0.032      7.435      0.000       0.177       0.304
Population density          0.2928      0.040      7.281      0.000       0.214       0.372
Ambulance Incident Rate     0.0204      0.026      0.796      0.426      -0.030       0.071
Number of jobs              0.5500      0.208      2.645      0.008       0.142       0.958
Crime Rate                  0.0253      0.039      0.641      0.522      -0.052       0.103
Public Transport            0.0613      0.046      1.324      0.186      -0.030       0.152
Unemployment               -0.1063      0.037     -2.870      0.004      -0.179      -0.034
Education Level 4 above     0.6595      0.034     19.425      0.000       0.593       0.726
Open spaces access          0.1519      0.024      6.372      0.000       0.105       0.199
==============================================================================
Omnibus:                       65.642   Durbin-Watson:                   1.497
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              110.250
Skew:                           0.696   Prob(JB):                     1.15e-24
Kurtosis:                       4.537   Cond. No.                         23.8
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Table 2: Regression with outliers removed

Comparing the results, we are overwhelmed with how the goodness of fit -R squared which determines the proportion of variance in the dependent variable which can be explained by the independent variables, is improved by just removing the outliers. Also, we can see that the most important features totally change. With outliers we may interpreted that among our analysis factors, education level, public transport, open space access are the top 3 factors that influence housing price. But on removing them, we get the actual influencers which are education above 4 levels, number of jobs in the area and population.

To confirm our findings visually, lets us try to plot the choropleth map for education above 4 levels to view how it stands with respect to the earlier spatial observations for the housing price in Figure 3.

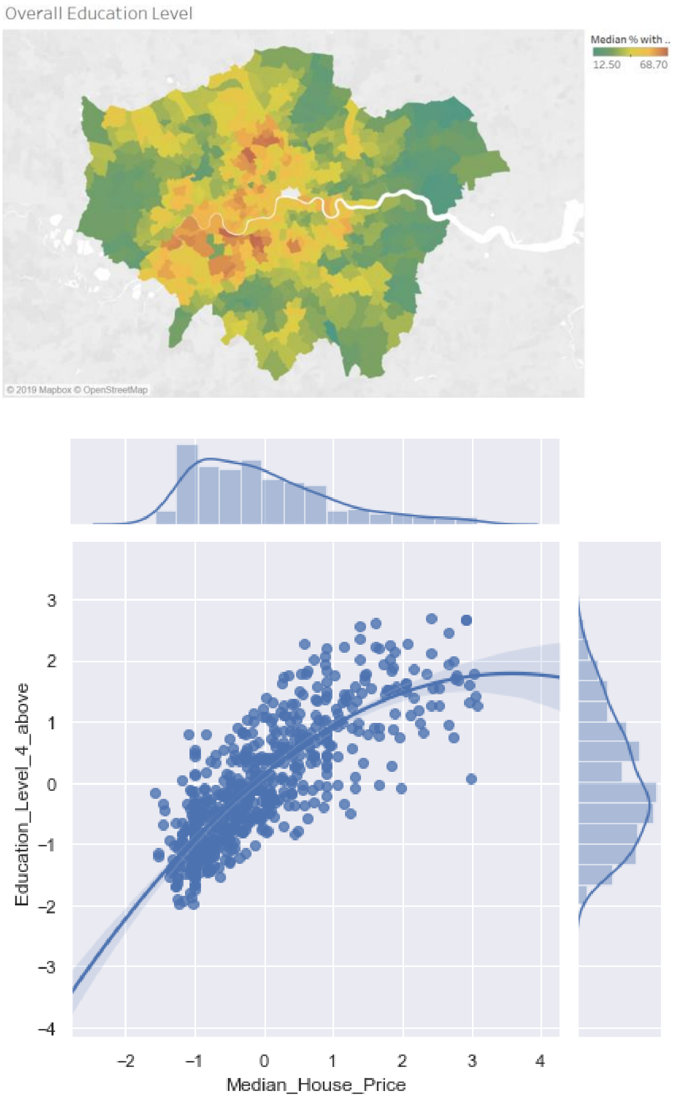


Figure 9 Education Level 4 as predictor i. Choropleth map ii. Scatter plot with regression line order 2

We can observe that there is a close similarity with the patterns of the housing price at ward level, though the intensity differs. Let us now try a univariate linear regression to assess the prediction and analyze the residuals. The second order regression gives a better result than the linear regression based on our analysis visually.

This confirms our stoppage point for our analysis as we have found the best estimator for housing prices from our dataset and answered all the proposed research questions.

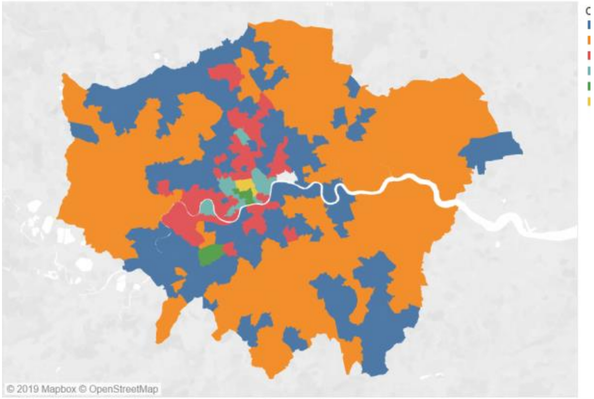
4.3 Results

1. Are there places in London which are more expensive in terms of housing prices than others?

There is a huge variation in housing prices as we move from the periphery to the central wards.

2. Can we cluster housing prices based on the data and thus device strategies for investment?

   Optimally we can group wards into 6 clusters.



Clusters of Housing Prices

© 2019 Mapbox © OpenStreetMap

Map based on Longitude (generated) and Latitude (generated). Colour shows details about Clusters (1). Details are shown for Borough and Ward name.

Figure 10 Six Clusters for Housing Prices

3. To what extent has there been a rise in housing prices in London over time?

   London housing prices have gradually risen from 1995 to 2014 with depreciation across almost all boroughs except two during the recession in 2009.

4. Can we explain the housing prices in terms of the characteristics of the population?

   The most important socio-economic indicators that are associated with housing prices are the actual influencers which are education above 4 levels, number of jobs in the area and population. Using the 10 features we were able to build a computational model with goodness of fit 0.7

5. To what extent can a single socio-economic estimator for housing price be determined?

   Education above 4 levels is the best estimator for housing price when considered as a second order regression function.

5 Critical Reflection

Our findings suggest that housing prices in Greater London are not randomly distributed but there is a pattern and analyzing spatial, temporal patterns and the associated social, economic, population parameters can help any potential investor or home buyer to make an informed decision on an investment opportunity in housing in London.

Spatial analysis by successive iteration of increase in granularity from borough to ward on the choropleth maps helped us to identify more specific patterns which are not visible when viewed at an aggregated level. Consequently, the data patterns are assumed to be uniform at the chosen granularity level which may be a strong assumption to make. Using K Means clustering we successfully clustered Greater London into 6 clusters. Silhouette plots helps to visualize the scores for each cluster, observe any negative values and decide the optimum number based on the average along with the elbow method. An investor can thus potentially look at investing in one or more of the clusters based on their budget.

Pairwise scatter plots of different socio-economic factors with housing price helped to distinguish which factors are associated to housing prices. Bivariate analysis using Pearson correlation assumes normal distribution of data, linearity and homoscedasity. Kendall’s rank correlation and spearman rank correlation are other alternatives which are non-parametric and does not make assumptions on the distribution.

Multinomial regression analysis was used to determine the association of various socio-economic factors and housing prices. Based on the OLS regression we determined that the most important socio-economic indicators that are associated with housing prices are education above 4 levels, number of jobs in the area and population. Linear regression assumes that there is a linear relationship which may not always be true.

Time series analysis helped to understand the pattern of housing prices over the years including percentage growth rate over the years. An investor is generally interested in an asset which gives the highest gain in the shortest time or yields. Based on the time series analysis, we can shortlist boroughs/ wards with maximum growth. In addition, the growth in number of enterprises setup (a proxy for number of jobs), population and number of schools (a proxy for number of educated people) can help the investor foresee the future value of their investment. An extension to the work on analysis of time series price data would be to build a model to predict the housing prices in the future based on historical data.

Housing prices are influenced by the external socio-economic factors as well as the internal – floor area, number of bedrooms, bathrooms etc. Our analysis is limited to only the external factors. It is worth noting that the data suffers from a selection bias to users who participated in the survey on the wellness indicators and is therefore not a totally precise representative of the population. For future research and extension of this project for people dealing with similar data, the next-step can be to carry out a study on the inter-ward associations and influence on pricing and also include the internal factors determining housing prices.

REFERENCES

The list below provides examples of formatting references.

[1] Liu, Shusen, Dan Maljovec, Bei Wang, Peer-Timo Bremer, and Valerio Pascucci. 2017. 'Visualizing High-Dimensional Data: Advances in the Past Decade'. IEEE Transactions on Visualization and Computer Graphics 23 (3): 1249–68.

[2] Li, Mingzhao & Bao, Zhifeng & Sellis, Timos & Yan, Shi & Zhang, Rui. 2018, 'HomeSeeker: A Visual Analytics System of Real Estate Data', Journal of Visual Languages & Computing. 45. 10.1016/j.jvlc.2018.02.001.

[3] Li, S, Ye, X, Lee, J, Gong, J & Qin, C 2017, 'Spatiotemporal Analysis of Housing Prices in China: A Big Data Perspective', Applied Spatial Analysis and Policy, vol. 10, no. 3, pp. 421-433. https://doi.org/10.1007/s12061-016-9185-3

[4] Turkay, C., Filzmoser, P. and Hauser, H. (2011). 'Brushing dimensions--a dual visual analysis model for high-dimensional data', IEEE Transactions on Visualization and Computer Graphics, 17(12), pp. 2591-2599. doi: 10.1109/TVCG.2011.178

[5] N. Elmqvist, P. Dragicevic and J. Fekete, 'Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation', in IEEE Transactions on Visualization and Computer Graphics, vol. 14, no. 6, pp. 1539-1148, Nov.-Dec. 2008

[6] Keim D., Andrienko G., Fekete JD., Görg C., Kohlhammer J., Melançon G. 2008, 'Visual Analytics: Definition, Process, and Challenges' In: Kerren A., Stasko J.T., Fekete JD., North C. (eds) Information Visualization. Lecture Notes in Computer Science, vol 4950. Springer, Berlin, Heidelberg

[7] London Well Being Scores London https://data.london.gov.uk/dataset/london-ward-well-being-scores

[8] London Ward Profiles https://data.london.gov.uk/dataset/ward-profiles-and-atlas

[9] House prices in London – an economic analysis of London's housing market https://www.london.gov.uk/sites/default/files/house-prices-in-london.pdf

[10] INM433 Visual Analytics (PRD1 A 2019/20) Week 7 practical http://www.staff.city.ac.uk/~sbrm048/w7_practical.html

[11] Silhouette score coefficient https://en.wikipedia.org/wiki/Silhouette_(clustering)

Table of word counts

Problem statement	253
State of the art	517
Properties of the data	480
Analysis: Approach	501
Analysis: Process	1408
Analysis: Results	199
Critical reflection	499