Housing Prices in London and the entire United Kingdom

Group 3, DS522 Data Acquisition and Analytics, City University of Seattle

Ean Pfeiffer ([pfeifferean@cityuniversity.edu](mailto:pfeifferean@cityuniversity.edu)), David Rodriguez-Jenkins ([rodriguezjenkinsdav@cityuniversity.edu](mailto:rodriguezjenkinsdav@cityuniversity.edu)), Brock Saxen ([brock@cityuniversity.edu](mailto:brock@cityuniversity.edu))

**Abstract**

Our project is a comparison of home pricing between London counties and other regions of the United Kingdom. We are applying principles and techniques learned throughout this course to analyze datasets on the topic. We plan to explore the dataset and analyze the following: correlations between housing prices and location, correlations between property type and county, and old versus new construction based on location. The datasets are both from 2018, while they are a few years old, they are from the same time and will work well as a comparison. One dataset covers housing prices and information specific to London and the other dataset covers housing prices throughout the entire UK. Additional UK Housing information may be included later. Techniques that we will be using to analyze the data will be updated later as they’re from modules 5-10 and we don’t have access to those at this time. Overall, this project will show a good representation of housing information and how that differs from large cities to more rural areas. ---------- We need to add a brief summary of the results -----

**Keywords:** property type, regression, model, price

**1. INTRODUCTION**

Property prices have been a contentious subject ever since the housing collapse in 2008 which was then followed by an unprecedented rise in housing prices in 2020 which left a large swath of the population unable to afford a home. Real estate has historically been considered an easy investment vehicle which increases in value and keeps pace with inflation. Although it is not a liquid asset it can be used to rent out, inherited later, and be sold later for more than it was purchased for. Our team has taken a keen interest in researching housing prices and identifying insights derived from housing data. Our team was able to find large data sets on Kaggle containing UK housing data for London and various other UK areas. Our goal is to analyze and find insights on the different properties being sold and how their value has increased/decreased over time. We can also analyze how much value properties lose as they move away from major cities like London as well as other major cities near them.

**Problem Statement**

Using the UK datasets acquired on Kaggle we would like to determine how much the recent spikes have affected the UK. It would also be of interest to see how much value the properties have increased as well as how they are valued relative to the distance between London and various other major UK cities. The dataset will include the years from 1995 to 2018, additionally we will perform deep dive analysis on the 2018 information that is available.

**Motivation**

Utilizing the techniques taught in this class by using large dataset of UK housing transaction we will be able to visualize key figures, create projections, and find insights. Ideally, findings found from the analysis of the data set will allow us to present key findings.

**Approach**

Using technique learned throughout the class we will be taking the large datasets and converting them to pandas data frames which can then be manipulated to visualize key insights and explore possible findings. Ideally, we develop an approach to cleaning the data, identifying the key pieces of information, optimize our methods by developing functions that can be used efficiently, and visualizing our data that can be easily understood by a wide audience.

**2. BACKGROUND**

We are using the knowledge gained and methods learned in this class as well as from other classes for this project. To preprocess, find insights and visualize the information we deem valuable. Keys to success are the following: collecting the data and identifying useful data, cleaning, and preparing the data, exploring the data, and finally visualizing and accurately detailing the results of our labor.

For collecting the data, we plan on utilizing the 2 UK housing data sets that can be found on Kaggle along with possibly utilizing other datasets that can be provided by the UK government website. For the data preparation process our team will need to determine the quality of the data that is being used and the completeness of the dataset.

For example, our team will need to determine how much of the data is missing and decide on the overall consistency of the datasets being used. For the exploration portion find and visualize, averages, modes, and other statistics we might find valuable. Having a good understanding of our data will help us in determining the best methods to utilize to find key insights.

**3.** **RELATED WORK**

Related work for this can be found on Kaggle.com, where we gathered the datasets; and includes, but is not limited to:

<https://www.kaggle.com/datasets/hm-land-registry/uk-housing-prices-paid>

https://www.kaggle.com/datasets/arnavkulkarni/housing-prices-in-london

**4.** **APPROACH**

**Implementation**

This process will include a massive set of data produced from active research prepared on the housing prices of the UK. The implementation of this data will be as follows: Initial analysis, pre-processing, trimming the data, then manipulating the columns and rows to produce a data frame more appropriate for the size and scope of our implementation. Once this is achieved, we can then proceed with examining the relationships between the multiple points of data to exercise our code.

**Design**

<additional comments needed>

**Technologies Used**

The project was developed using Jupyter notebooks. While additions to the file were updated in teams. This allowed the team to collaborate using a few files that were updated in real time.

At the current moment our team is using the following libraries to support our project:

We will use the following libraries to support our data exploration, mining and cleaning.

A screen shot of a computer screen

Description automatically generated with low confidence

Figure 1

Pandas is a Python programming library for data manipulation and analysis.

Numpy is Numerical Python, the core library for numeric and scientific computing.

Matplotlib is a comprehensive library for creating static, animated and interactive visualization in Python.

Expand on folium, mpu geopy, and heatmap plugin

**5. DATA COLLECTION**

The datasets chosen for the assignment were found on Kaggle.com and are comparable as is, but we were able to create Data Frames to make them more directly comparable and easier to analyze. One of our datasets covers home sale information specifically in London in 2018 and the other covers all the UK from 1995 through 2017. In addition to pricing information the datasets include other things that we were able to compare such as property type, square footage, location, number of bedrooms and bathrooms, and how long the homes were for sale. Both datasets are comma-separated value files(csv), and we used a variety of methods learned in this class to analyze the data.

**6.** **DATA CLEANING**

We ran the following commands to remove duplicates and missing data. Additionally, we modified Data Frames to make the data more comparable. One dataset was home prices for just 2018, and the other was from 1995 through 2017. Dropping the data that was from 1995 until 2016 and focusing on the data from 2017 made the data more comparable to the other dataset.

A screenshot of a computer program

Description automatically generated with medium confidence

Figure 2

**7.** **DATA EXPLORATION**

Figure 3

A screenshot of a computer

Description automatically generated with medium confidence

Figure 3

Preparing the data was the most time-consuming part of the process. There were multiple angles to approach the imported information, that it became imperative to isolate the truly necessary points of information. After clearing out the data to perform a deep dive on the desired range, we began to isolate the relationships between county, region, city, and sale date, including the size of the home regarding a single family or generational family unit. Once isolating the data was achieved it became much easier to filter through data to observe those relationships.  
  
A step farther that was one path to determine the relationships between present and future housing prices came from the need to isolate housing by county, and then town. Creating a copy of the data, setting it aside, and then using that to create a plotted graph isolates the data into a specific view to be more discernable to try and show any relationship between geographical location and the difference in housing prices.

A screenshot of a computer

Description automatically generated

Figure 4

The above indicates the processing and evaluation of the total count of data on isolating the data.

Figure 5

A screen shot of a computer

Description automatically generated with medium confidence

Figure 5

Isolation and separation of 2017 Data

A screenshot of a computer

Description automatically generated with medium confidence

Figure 6

Graphical representation of Home Price and County. It became clear that the West Midlands had the highest home price in the evaluated years. It would be interesting to see the depth of population in an area and its potential correlation to home prices. Do more heavily populated areas have higher home prices? If so, could this be due to competitive pricing systems in the county?

One angle of evaluating the data was specifically looking at two districts and comparing the mean pricing of the homes between the two. By isolating the county data of both Hampshire and Greater London, and then comparing prices by mean values between districts within that county we were able to see which districts were the most expensive based on geographic location. While looking at the charts associated with Greater London County, it was clearly observed that the City of London contains the highest mean price of homes by nearly a third evaluated against the districts of City of Westminster and Kensington and Chelsea. The same was done with Hampshire, which showed that the highest mean of home prices existed in Winchester. (Figure 7)

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 7

**7. EVALUATION**

Much of the found data and graphical representations can be used to thoroughly evaluate the different reasons, causes, and isolated specific instances of inflation over time. By using this method of examination, we can be expected to note the most expensive areas under the most expensive living conditions through either inflation or naturally occurring cost of living.

**8. FINDINGS**

Findings were what we expected with this. The closer to larger cities and more populated areas, housing tends to be more expensive; and this is the case in the United Kingdom as well. The closer you get to London the more expensive homes tend to be. There’s also an unexpected correlation between increasing square footage of a home and price. This is supported by plots labeled as Flat Prices in London by Square Footage (Figure 7), Home Prices in London by Square Footage(Figure 8), and DataUK.

A picture containing screenshot, graphics

Description automatically generated

Figure 8

Figure shows flat/apartment prices in relation to square footage in London.

A picture containing screenshot, map, graphics

Description automatically generated

Figure 9

Figure shows single family home prices in relation to square footage in London.

**9. CONCLUSIONS**

Homes in metropolitan areas are much more expensive than the country as a whole; and finding. The average home price in London in 2017 was between £1.864.172 and £1.877.659 for a home between 1712 and 1766 square feet depending on the chart. The average home price in all the UK is £328.828. So, the United Kingdom has a wide range of prices.

**10. FUTURE WORK**

Future work that we would like to try to do would be to plot out the prices of homes and their location. Some sort of geographical map with prices in relation to their proximity to large metropolitan areas. Currently our project explores the prices in some specific areas and the whole country. What we have doesn’t visually show how close a given listing is to London or Wimbledon. So that’s something that we would like to pursue if we had the time and expertise to do so.

**11. ACKNOWLEDGEMENTS**

Place before the references.

**12.** **REFERENCES**

<https://www.kaggle.com/datasets/hm-land-registry/uk-housing-prices-paid>

https://www.kaggle.com/datasets/arnavkulkarni/housing-prices-in-london

<https://pandas.pydata.org/docs/user_guide/index.html>

https://matplotlib.org/3.5.3/api/\_as\_gen/matplotlib.pyplot.html