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 applied principles and techniques learned throughout this course to analyze datasets on the topic. We explored the datasets and analyzed 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, and 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 used to analyze the data included data cleaning, creating data frames, and creating charts and graphs. Our team's data exploration of the UK datasets allowed us to compare housing prices of areas outside the metropolitan areas as well as comparing the various counties and areas in London.

**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 counties and various other UK areas. Our goal was to analyze and find insights on the different properties being sold and what key metrics cause their variation in price. We looked for correlations based on square footage, median housing prices, and location, and utilized data visualization techniques to better understand the datasets to come to our conclusions.

**Problem Statement**

Using the UK datasets acquired on Kaggle we utilized some of the techniques learned throughout this class to acquire unique insight on the UK housing market based on information that is freely available. It was of interest to see how various attributes like square footage, average prices for various other areas, number of bedrooms and bathrooms, and property type; before finding ways to visualize the information. The datasets used included pricing information from 2017 and 2018. More specifically the 2018 dataset is mainly centered on London and its surrounding areas and counties.

**2. BACKGROUND**

We used some of the knowledge gained and methods learned in this class as well as from other classes for this project. The techniques learned will help with preprocessing, finding insights and visualizing the information we deem valuable. Our group felt that some of the keys to success with this project included 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 used the following datasets with UK housing data: Housing Prices in London (Kulkarni, 2020) and UK Housing Prices Paid (HM Land Registry, 2020). These datasets were found on Kaggle. We considered the possibility of utilizing other datasets that can be found on the UK government website. However, we decided to stick with the two that we chose because of the quality and completeness of the datasets we gathered, we didn’t feel like it was necessary to go with a government source.

**3.** **APPROACH**

**Implementation**

This process included a massive set of data produced from active research prepared on the housing prices of the UK. The implementation of this data was 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 was achieved, we then proceeded with examining the relationships between the multiple points of data to create our code.

**Technologies Used**

The project was developed using Jupyter notebooks, while additions to the file were updated in teams and code was updated in our Git repo. This allowed the team to collaborate using a few files that were updated in real time. The libraries we used throughout this project utilized the following: Pandas, NumPy, Folium, Mpu, Matplotlib.pyplot, and Geopy. The libraries were used to support our data exploration, mining and cleaning.

Pandas is a data manipulation library that allowed us to create data frames which could easily manipulate and adjust. The library is incredibly useful in allowing data preparation and analysis. Throughout the project there are several instances where columns and values are utilized to create new data to be examined. Matplotlib was extensively used to generate various graphs scatter plots to visualize our data. Mpu provides various math operations but was mainly used to calculate the distance between two locations using their latitude and longitude. The folium, Geopy, and heatmap plugin with Geopy was utilized in constructing heatmaps over the London area.

**4. DATA COLLECTION**

The datasets chosen for the assignment were found on Kaggle.com. Utilizing the Kaggle data sets 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.

**5.** **DATA CLEANING**

We ran the following commands to remove duplicates and missing data in figure 1. For dropna, dropped rows where an element is missing; while in place=True returns a new data frame. Drop\_duplicates removes duplicate rows and isnull().sum() will give us the number of missing values, which is also shown in the figure 1.

A screenshot of a computer program

Description automatically generated with medium confidence

*Figure 1*

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. In the Housing Prices in London dataset there were various areas that only had one or two homes sold in the area. It was necessary to remove these homes as there were not a significant number of homes sold to represent that area effectively. The Housing Prices in London dataset also had whitespaces present in the data and it made it difficult to iterate through and apply functions correctly. These white spaces were removed utilizing strip() function which removes spaces at the end and beginning. The filter\_df function allowed us to create a data frame column that removes the white spaces in the location column, filter by housing type (in this case “houses”), and finally remove locations that appear less than 10 times in the data.

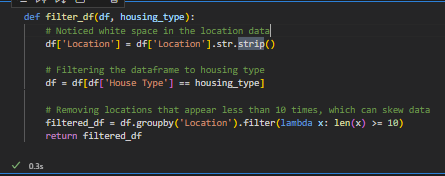


Figure 2

**6.** **DATA EXPLORATION**

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. We needed to approach the information from multiple angles so that we were able to isolate only the information that we needed. 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 in the UK Housing Prices Paid dataset. Once isolating the data was achieved it became much easier to filter through data to observe those relationships.  
  
In order to explore the relationships between present and future housing prices we needed 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 to isolate 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 figure shows the processing and evaluation of the total count, during the process of isolating the data.

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

While looking at our 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?

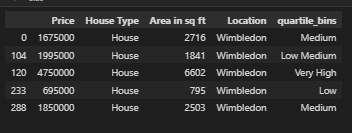
One angle of evaluating the data was specifically looking at two districts and comparing the mean pricing of the homes between the two. By isolating the county data of both Hampshire and Greater London, and then comparing prices by mean values between districts within that county we were able to see 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.

A picture containing text, screenshot, diagram, plot

Description automatically generated

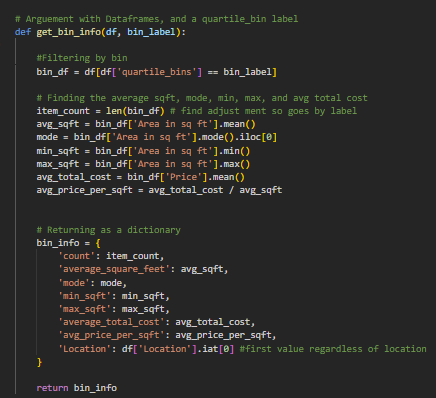
*Figure 7*

Then came the data exploration for Housing Prices in London (Kulkarni, 2020) after preprocessing was done in various ways. For this data we focused on houses sold in the London area in order to create a process that could be repeated for the other house types. The first step was creating a method to break the houses down into bins based on their square footage. For example, if a house sold in Wimbledon has an area sq ft larger of 6602sqft it would get a “Very High” label and a house that sold in Wimbledon with 795ft would be labeled “Low”. The binning can help us understand the relationship with house size, possibly provide better visualization as well as possibly finding outliers in the data.



*Figure 8*

With the tagged houses we generated a function to provide information of the tagged properties giving the following: avg\_sqft, mode, min and max sqft, average total cost, and average price per square feet.



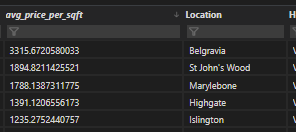
*Figure 9*

With the get\_bin\_info function, in conjunction with a binning function, and a location list we were able to iteratively go through and create a new data frame with all London location and bin information. In addition coordinates were added to the location using gelocator.geocode as well as adding in the percent cost difference in price per square foot. For example, Wimbledon had an average price/sq ft of £741.32 for houses tagged in the “Very High” bin, while “Low” housing had an average price of £937.06 price/sq ft. Essentially, buying a house in the “Very High” category in Wimbledon netted you a 20.8% cheaper price on you sq ft. The figure below shows the 5 locations that experienced the largest discounts as the houses increased in size:



*Figure 10*

Using the mpu library it was also possible to calculate the distance between the most expensive part of London and other areas in London. Looking at the data frames we noticed that Belgravia had the highest price/sq ft and initially we wanted to use it, but it was an incredible outlier (being it is one of the wealthiest areas in London), so we went with the second highest, St.John’s Wood, which wasn’t as much of an outlier when compared to Marylebone.



*Figure 11*

With that being done, we utilized the coordinates to find the distance between St.John’s Woods and various locations in London so that we could see how much cheaper price/sq ft gets as you move away from one of the priciest areas in the world in 1 kilometer increments. There were 24 locations, figure 12 shows the 5 farthest locations in the dataframe. Interestingly, for every 1km you move away from St.John’s Woods on average it get £84.69/ sq ft cheaper.

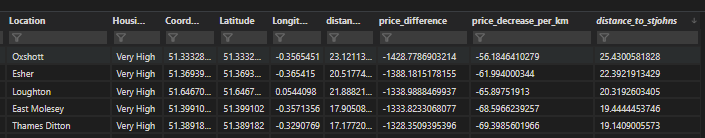


Figure 12

**7. EVALUATION**

Using the data we gathered and graphical representations we can evaluate and hypothesize some of the different reasons, causes, and isolated specific instances of inflation over time. By using this method of examination, we were able to find 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 its price decreasing per sq ft. This can be seen in the plots labeled as Flat Prices in London by Square Footage (Figure 13), Home Prices in London by Square Footage(Figure 14), and DataUK.

A picture containing screenshot, graphics

Description automatically generated

Figure 13

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

A picture containing screenshot, map, graphics

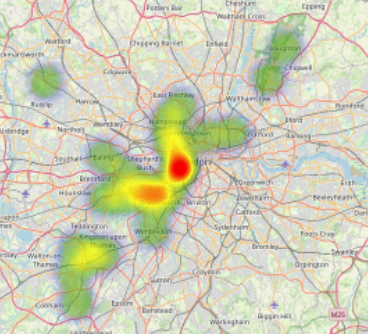
Description automatically generated

Figure 14

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

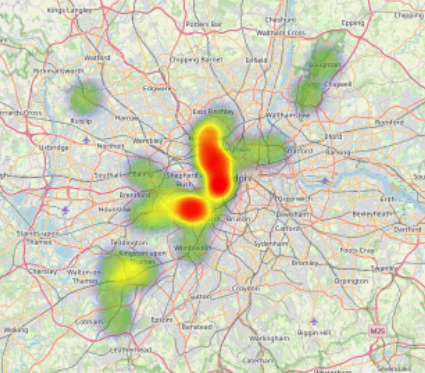
The following are the heat maps that were generated to show “Low”, “Medium, and “Very High” houses. The heat map shows the areas with the highest price/sq ft. Based off these maps the most expensive per square footage homes are in the regions of Westminster and Regent Park. While a majority of the housing data resides on the west part of the Thames river. This could be an indication that a lot less housing and more industry is in the southeastern portion of London. It should also be noted that housing tagged in the “Medium” range had pricing that were more similar based on the heat indication in the center.

Low:



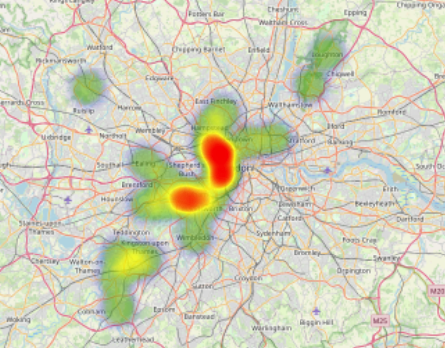
*Figure 15*

Medium:



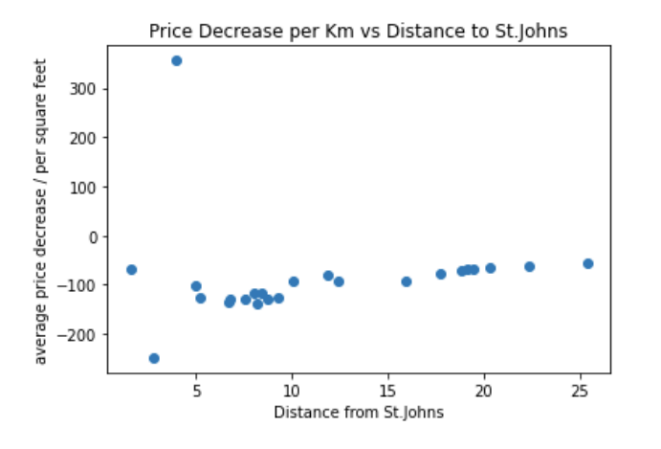
*Figure 16*

Very High:



*Figure 17*

The following is a scatter plot of the average price decrease/ per square feet:



*Figure 18*

The decrease stays relatively the same. The outlier being Belgravia, London we removed from the data.

**9. CONCLUSIONS**

Homes in metropolitan areas are much more expensive than the country. 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. With the 2018 dataset we were able to determine that, based on housing prices, Westminster and Regents Park area cost exceedingly more than the other areas of London. There was also surprising data that showed that in some areas, as you purchased more square feet the average price per sq ft could go to as much as 15% when compared with housing from the “low” bin. Lastly, as you move 1km away from the most expensive areas of London you stand to save as much as £87 per square foot compared to houses in the more expensive areas. Overall, there is still a lot of information that could be further extracted, but the datasets allowed us to fully utilize a lot of the techniques learned throughout the course.

**10. FUTURE WORK**

Future work would consist of finding out if the trends found act similar on a year-to-year basis. With year-to-year data we could see if the price per sq ft trends we found, based on distance from expensive areas continues or if there’s a change. We could also see if the discounted price per sq ft with an increase in square footage continues unchanged, increases, or decreases in more recent years. We could also add major nearby cities and compare prices of cities of larger sizes like Cambridge, Oxford, Manchester, and possibly other large cities in other countries. Population density could also be utilized in tandem with the UK housing data.

**11.** **REFERENCES**

GeoPy Contributors. (n.d.). Welcome to GeoPy’s documentation! Welcome to GeoPy’s documentation! - GeoPy 2.3.0 documentation. <https://geopy.readthedocs.io/en/stable/>

Pandas. (n.d.). User guide#. User Guide - pandas 2.0.2 documentation. <https://pandas.pydata.org/docs/user_guide/index.html>

Hunter, J., Dale, D., Firing, E., Droettboom, M., & Matplotlib development team. (n.d.). Matplotlib 3.7.1 documentation#. Matplotlib documentation - Matplotlib 3.7.1 documentation. <https://matplotlib.org/stable/>

KULKARNI, A. (2020). Housing Prices in London. Retrieved June 08, 2023 from <https://www.kaggle.com/datasets/arnavkulkarni/housing-prices-in-london>.

HM LAND REGISTRY. (2020). UK Housing Prices Paid. Retrieved June 08, 2023 from <https://www.kaggle.com/datasets/hm-land-registry/uk-housing-prices-paid>

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). <https://doi.org/10.1037/0000165-000>

Martin Thoma, mpu, (2022), GitHub repository, Retrieved June 08, 2023 from https://github.com/MartinThoma/m