Adam Lang

3/20/2024

Springboard Data Science Fellow

London Housing Case Study - Conclusions

- What did you find?
    - The first thing I noticed when preprocessing the data was that the ‘City of London’ is not officially considered 1 of the 32 boroughs of London. However, I decided it might be interesting to compare a dataframe with and without the ‘City of London’ to see how it may change the maximum or most expensive boroughs and where the ‘City of London’ fits into the ranking of most expensive boroughs.
    - However, in the end, it did not matter whether ‘City of London’ was included or not for modeling the ratio of house prices for the years 1998 to 2018. The ‘City of London’ was 27th overall and did impact the final top 15.
    - The main finding of this case study was that from the years 1998 to 2018, the borough of ‘Hounslow’ had the highest price ratio of 0.251483 with ‘Richard upon Thames’ a close second at 0.249678.
    - The borough with the lowest price ratio for the years 1998 to 2018 was ‘Hackney’ at 0.161335.
- Which borough is the most expensive?
    - If we look at the ‘raw data’ before we compute the price ratios for the years 1998 to 2018, then the most expensive borough is ‘Kensington & Chelsea’ at $1,517,127. The next 2 most expensive are ‘Westminster’ at $1,117,407.52 and then the ‘City of London’ at $1,005,695.
    - As I mentioned above in the overall findings of the case study though, the greatest price ratio increase was seen in the borough of ‘Hounslow’.
- Any other interesting trends?
    - The most expensive borough as seen in the ‘raw data’ was ‘Kensington & Chelsea’ but it was in the bottom 15 in terms of the price ratio from 1998 to 2018 which means the prices did not change much over that period of time which is interesting.
    - I was able to groupby and aggregate on the 'max' metric, and found that the borough 'Kensington & Chelsea' holds the top 11 most expensive average prices with 2022 being the most expensive year. So not only was this borough the most expensive but it maintained this.

    - What is interesting though is that if you look at the ‘max’ metric the top 20 most expensive prices regardless of the borough occurred in years 2012 to 2022.
    - If we look at the ‘bottom 20’ most expensive prices, we see that ‘Barking & Dagenham’ takes the top 2 for the years 1995 at $53,700.35 and 1996 at $53,853.51. What is interesting is that if we look at the bottom 20 more closely, these prices all occurred between 1995 and 1997.
    - Aggregating on the standard deviation we can see that ‘Kensington & Chelsea’ has the greatest std at $99,814.79 for the year 2021 and the ‘City of London’ a close second at $88,264.04. What is most interesting about this is that both of these boroughs make up most of the top 10 standard deviations so while ‘Kensington & Chelsea’ had the highest price of the boroughs it also had the highest standard deviation.
    - The borough of Croydon had the lowest standard deviation at $263.52 and this occurred in the year 1995 with the borough of Bexley a close second at a standard deviation of $388.05 also in 1995. The top 10 lowest standard deviation values were all in the year 1995 which makes sense as home prices did not change that much back then.
    - Lastly, I decided to quickly examine some seasonality in the prices by adding a column to the dataframe that included the ‘City of London’ (df_with_london). I did this by extracting the month into a separate column called ‘Month_extracted’.
        - First I aggregated on the ‘max’ average price and found that of course the borough of ‘Kensington & Chelsea’ was still in the top 20 prices. However, the most interesting thing was that the top 10 prices occurred between March and September in the year 2021. For the most part the rest of the top 20 prices were all in spring or summer months with a few outlier months in the fall and winter (October, November, December, January). This goes along with my hypothesis that home prices tend to increase in warmer months.
        - I also used this aggregation for the lower price boroughs and of course ‘Barking & Dagenham’ had the top 20 bottom prices in this aggregation. In terms of seasonality by month, these seemed to be mostly in the late fall to early winter (October, November, December) and winter (January, February, March).
- How did you arrive at your conclusion?
    - To get to these conclusions I used the groupby and aggregate functions in pandas by using the ‘max’, ‘mean’, ‘std’ functions. I also ranked the data

using the sort_values by changing the ascending to False or True to get descending and ascending respectively.
    - I also calculated the price ratio from 1998 to 2018 by calculating the average price for 1998 divided by the average price for 2018.
    - I was also able to evaluate the month and year variables in my analysis by sorting the values on average price.
- What were the main challenges you encountered? How did you overcome them? What could you not overcome?
    - The biggest challenge of working with this data was the preprocessing. The process of transposing the data frame, resetting the index, moving the first row to the column headers, then renaming the columns, then using the pd.melt function and finally renaming the columns again as well as changing the 'Average_price' column datatype took awhile. Lastly I had to deal with missing values which are always a problem in 'real world data'.
    - To overcome the data 'cleaning' issues, I followed a step by step process performing each data transformation then checking the outcome. I felt taking a step by step approach to data cleaning is paramount so you don't miss anything that could skew your final results.
    - The other issue I encountered was related to the missing or null values. I had to determine which 'boroughs' were not actually borough's and needed to be dropped. I had to incorporate some NLP techniques here because I had to find the difference in the lists of boroughs - some were all uppercase and some had a number in them - these were much different than the other borough names and I had to 'bucket them' to be able to compare and contrast them with the other variables. I had to figure out what the actual 32 boroughs should be. In the end I was able to determine that the 'City of London' is not actually considered a borough, but I did create 2 data frames for analysis, 1 with and 1 without the 'City of London'. One caveat to this is that to verify what the true list of 32 boroughs are I went to the source data and uploaded a list of these boroughs and then used the python set.difference() function to help determine what variables did not match this list.
    - There weren't any issues that I was not able to over come.
- Is there anything you'd like to investigate deeper?
    - If I was just working with this dataset, I would like to look further into the seasonality of the data and the average price and how it changed for certain months. For the most part based on my analysis it seems that the prices were higher between April and September and lower between October and March, but this may not always be true.

- I would also like to introduce other datasets that can be found on the London Datastore and consider other factors that may contribute to the borough home prices including but not limited to:
    - Land Area & Population Density of the boroughs
    - Income and Earnings of home buyers and owners
    - Occupation and Education and how this relates to who is buying homes and in which borough.
    - Culture and Construction - were there any specific construction projects such as businesses, entertainment (i.e. restaurants, bars, sports arenas) that may contribute to the change in the average price of a home?