Analysis of London Real Estate Market

Recommending Similar Properties on Sale

May 4th, 2020

---

1. Background And Problem Definition

Real estate industry is based on identifying customer's taste and presenting properties that are in alignment with customer needs and preferences. Customers typically contact the agent once they are interested in a property listed with the company. Even if customer may not necessarily purchase the property they were first interested in, if they are provided relevant recommendations, they can view and purchase other similar properties listed with the same company. Hence, given customer's interest in a specific property, being able to identify and recommend similar properties is a crucial skill for realtors to close a sale. Rather than leaving these recommendations to realtor's experience and subjective evaluation, a data-based recommendation would provide and objective, more accurate evaluation of the similarity across different properties and enhance the trust between customers and their agents as well.

In addition to being an essential challenge for real estate agents, recommending similar properties is also crucial to the success of real estate recommender websites such as https://www.compass.com/. If you click on a specific property in Compass.com or similar websites and go back to the home page, other similar properties are listed on the screen based on the ones recently viewed.

In this study, my goal is to develop a data-based tool that can be used by real estate companies to identify similar property listings. This way, their agents can provide relevant and similar recommendations to a customer who is interested in a specific property and maximize their chances of making a sale.

2. Data

I am interested in implementing an unsupervised method to avoid labeling input by the agents, which would introduce subjectivity to the evaluations. Given the nature of the problem requires identifying similarity across different data points, I am planning on using K-means clustering to identify properties similar to each other.

For this project, I will particularly focus on London real estate, which is one of the prime locations throughout the world. Given the cost of housing and extensive availability of properties throughout the year, making relevant and accurate recommendations to potential buyers has a lot of value for real estate businesses in London. I also choose to focus on properties on sale as opposed to rental units.

Clustering similar properties on sale in London real estate market requires combining several datasets:

- List of properties on sale
  In order to get up-to-date data on property listings in London, I will use a web scraping app to download the data from RightMove.co.uk. The app used is called rightmove-webscraper and is available on GitHub by toby-p. This app is embedded in my code.
  While downloading the data from RightMove.co.uk, I particularly focused on properties on sale, the listings from the past 14 days and those with listing price less than or equal to £2 million. Among other information, for each property, this dataset includes:
    - Address
    - Price
    - Number of bedrooms

- List of postcodes and district names in London

The name of the district that each property is located in is not available in the RightMove dataset. Therefore, I will use web scraping to download the table that includes the list of postcodes and district names in London from https://en.wikipedia.org/wiki/London_postal_district

- Coordinates of each property using OpenCage API
  RightMove dataset provides the address, but not the exact coordinates of the listed properties. In order to obtain this information, I used OpenCage API to convert address information to coordinates. This way, for each listing, I obtained
    - Latitude, and
    - Longitude,

  which will be inputs to the queries using Foursquare API.

- List of venues nearby for each property using Foursquare API
  The last component of the data is the list of venues in 500 meters proximity of each listing obtained through Foursquare API using the latitude and longitude information. In order to simplify the venue categories and reduce the computational effort, I will particularly focus on the top 20 venue categories around each property.

  This last component adds 20 more columns to the dataset that includes the number of pubs, coffee shops, cafés, hotels, grocery stores, Italian restaurants, pizza places, parks, gyms / fitness centers, sandwich places, bakeries, bars, Indian restaurants, burger joints, supermarkets, bus stops, cocktail bars, plazas and theaters.

Once the aforementioned datasets are merged, we will have both the data describing the property such as price, number of bedrooms, district, and data describing the neighborhood such as the number of top 20 most common venues in 500m. vicinity.

Using this dataset, my goals are two-fold:

- First, analyze this data to understand the cost and availability of properties on sale in London and the relative expense of different districts.

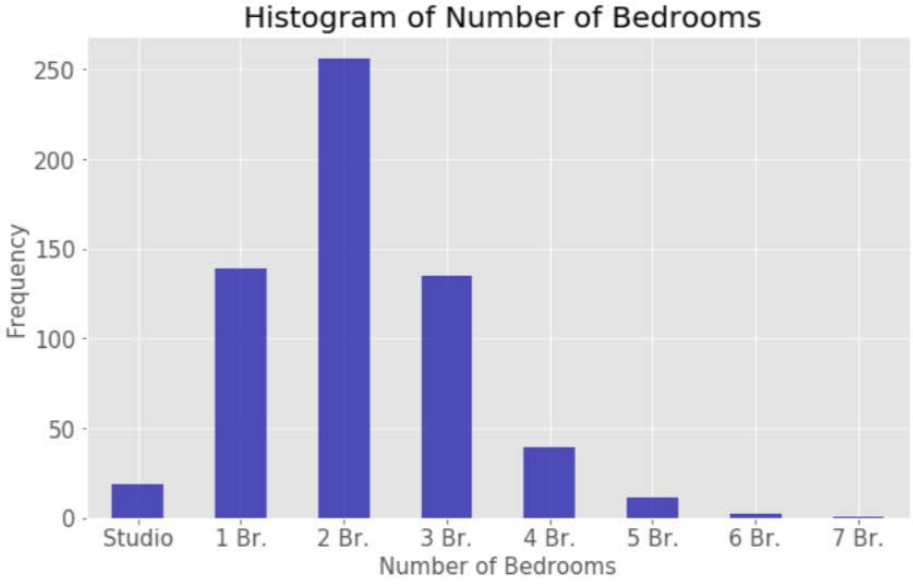
- Second, apply K-means clustering on this dataset to identify similar clusters of properties. In practice, the output of this analysis would be the basis for a realtor's recommendation to their customers. For example, if a customer got in touch with the realtor regarding a specific property in Cluster 1, even if the customer is not interested in buying this property, the realtor can recommend other properties in Cluster 1 as they are similar across multiple dimensions.

3. Methodology and Results

In this section, I present my analysis on the dataset that is merged as explained in the previous section (called listing_data_venues).

3.1 Exploratory Analysis

I first start by examining the data to understand its basic structure. I start by analyzing the distribution of number of bedrooms of the properties on sale.

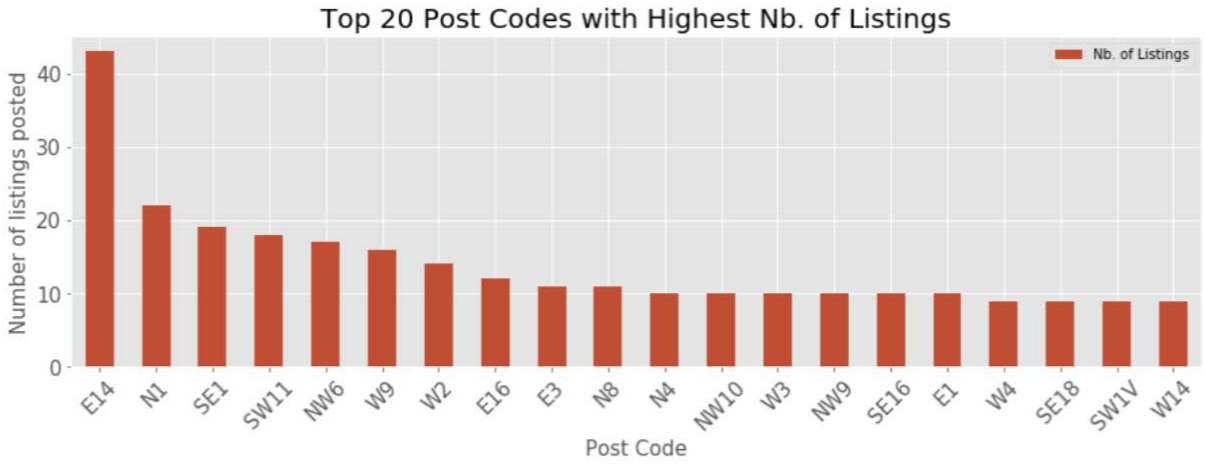


As seen in the histogram, vast majority of the properties on sale have 1, 2 or 3 bedrooms, while 2 bedroom apartments are the majority.

Next, I examine the range of prices of the properties on sale. Recall that my search on RightMove.co.uk was filtered for properties of price below £2 million.

Majority of the listings are between £500K - £1M range. A more meaningful data to look at here could have been price per square foot. However, this information was not available on RightMove.

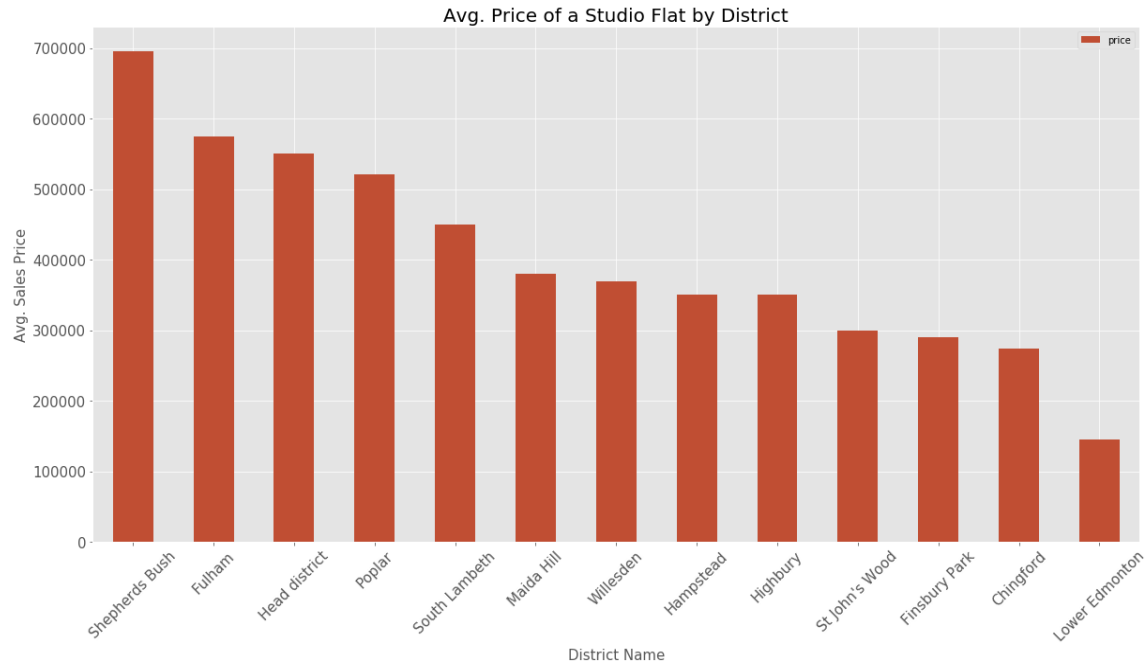
Next, I sort and plot the postcodes with the highest number of property listings.



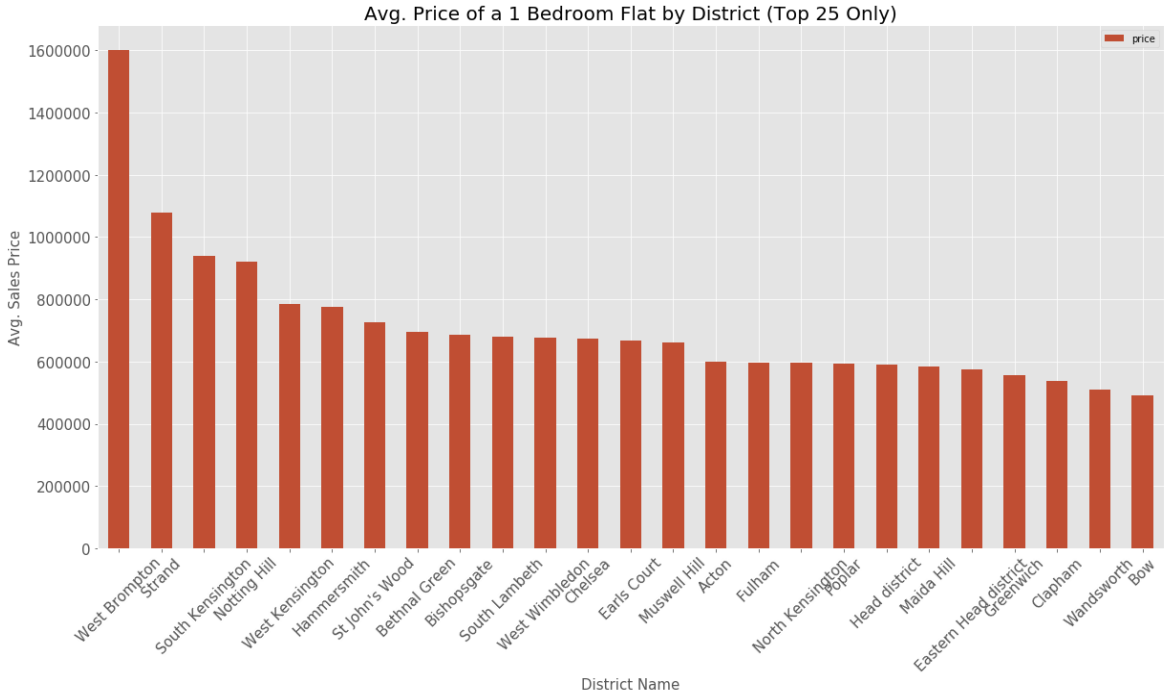
As the figure shows, postcode E14 has significantly more property listing than other areas. Further analysis shows this is the area with predominantly higher new developments and high-rise residential buildings.

Another goal in the exploratory analysis is to understand the relative cost of real estate across districts in London. Since I do not have the square foot information, I will instead look at the breakdown of each number of bedrooms separately.

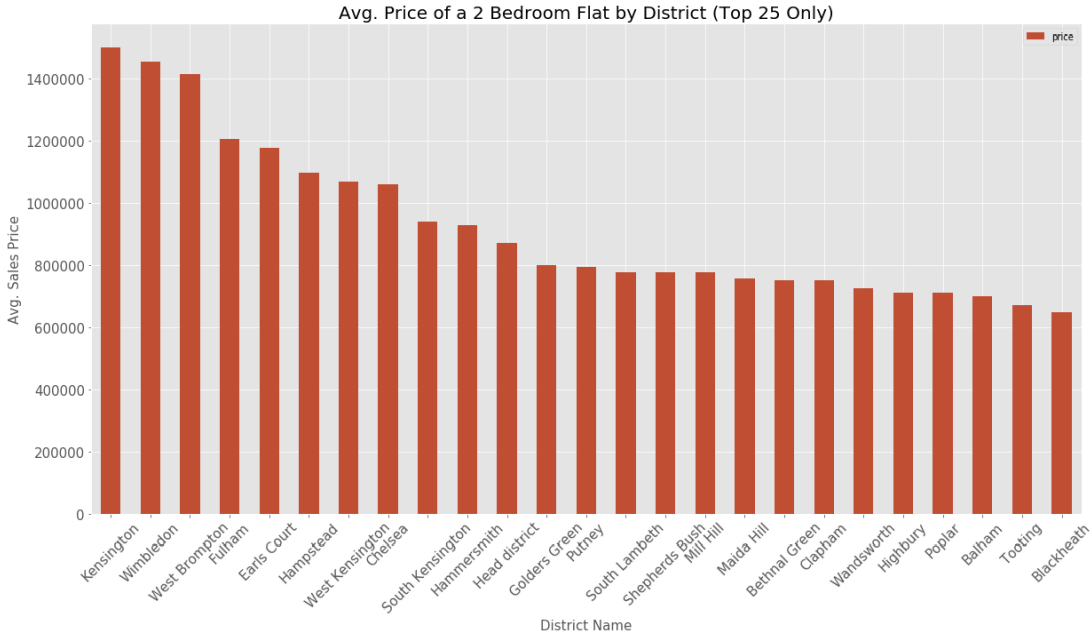
First, I plot the <u>average price of studio apartments</u> across districts.



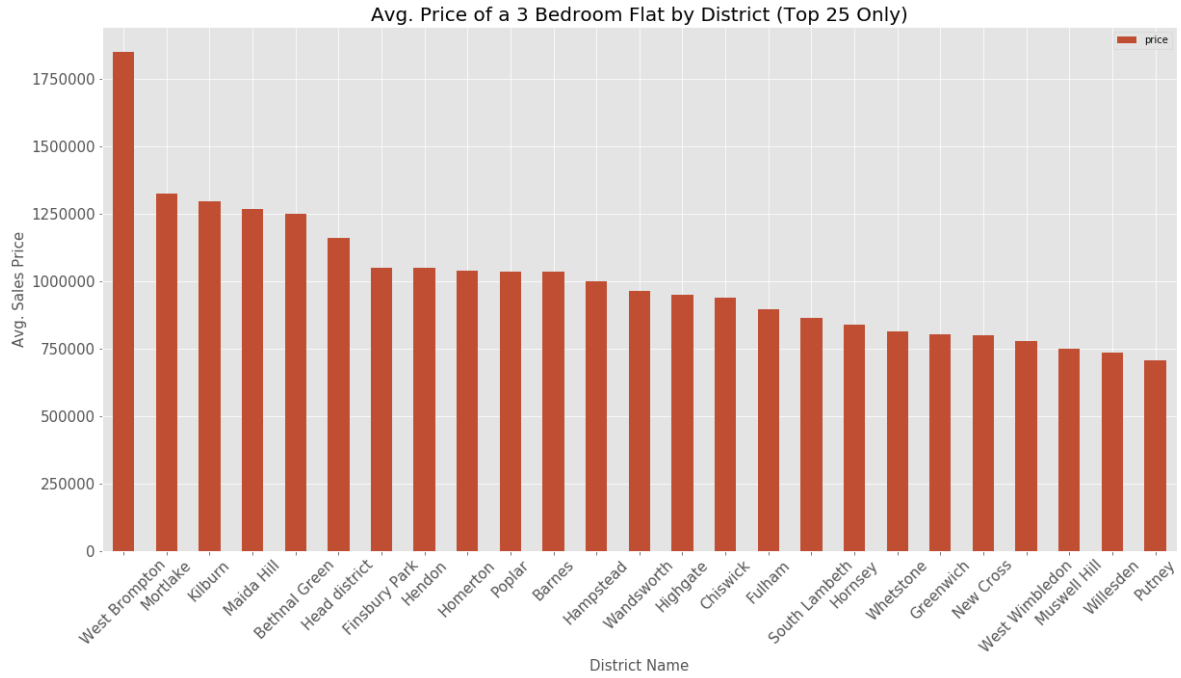
Next I'll plot the average price of 1 bedroom apartments across most expensive 25 districts.



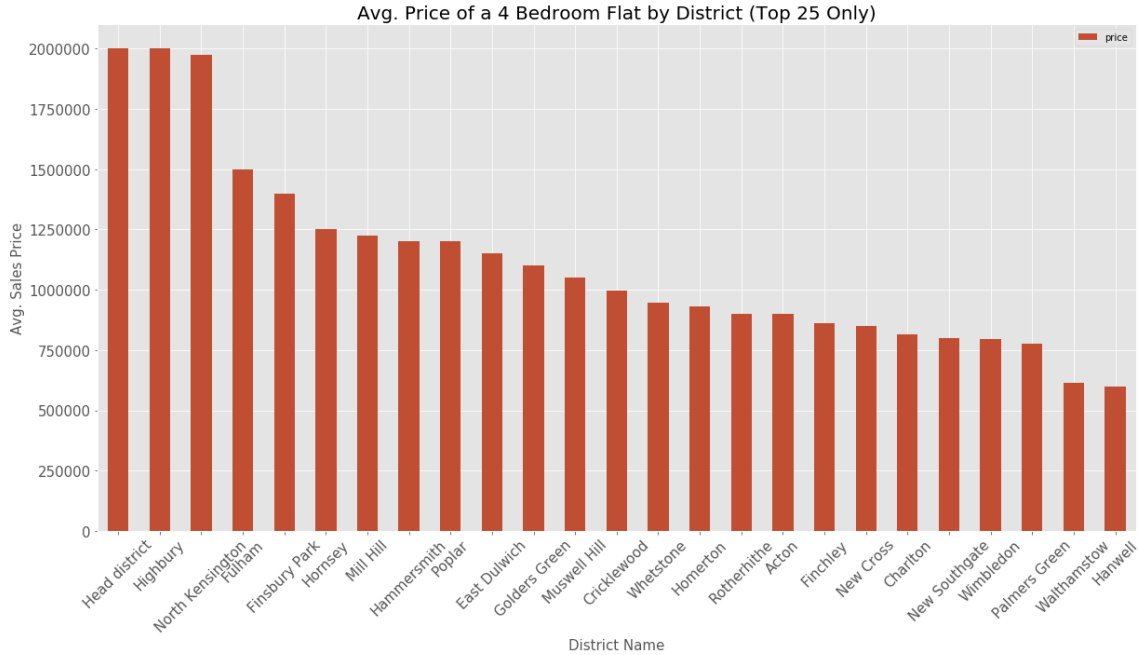
Next I'll plot the average price of 2 bedroom apartments across most expensive 25 districts.



Next I'll plot the average price of 3 bedroom apartments across most expensive 25 districts.



Next I'll plot the average price of 4 bedroom apartments across most expensive 25 districts.



The analysis above shows us that the relative cost of different districtschanges by the flat size. For smaller flats such as Studio or 1 or 2 bedrooms, most expensive neighborhoods are South Kensington, Notting Hill, West Kensington, Fulham and St. John's Wood. For larger flats that are better suited for families, most expensive districts are West Brompton, Highbury, North Kensington and Mortlake. I should also note that the deduction for larger flats could be biased as the number of listings in each district is very few to make a general statement.

3.2 Clustering Property Listings

In this section I will focus on identifying clusters of similar listings that real estate recommendations would be based on. I define the input that clustering will be based on as follows:

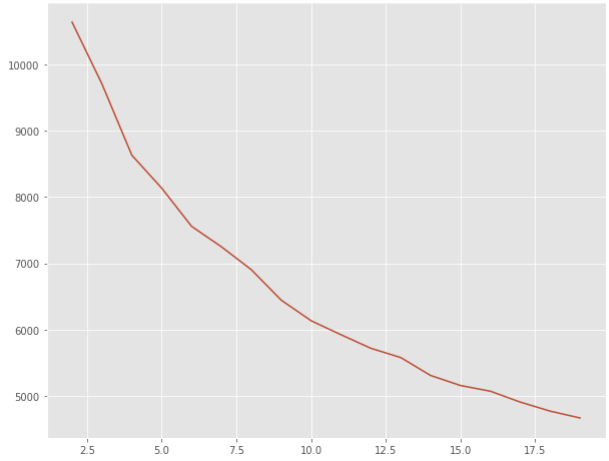
```
from sklearn.preprocessing import StandardScaler
X = listing_data_venues[['price','number_bedrooms', 'Pub', 'Coffee Shop', 'Café', 'Hotel', 'Grocery Store',
       'Italian Restaurant', 'Pizza Place', 'Park', 'Gym / Fitness Center',
      'Sandwich Place', 'Bakery', 'Bar', 'Indian Restaurant', 'Restaurant',
       'Burger Joint', 'Supermarket', 'Bus Stop', 'Cocktail Bar', 'Plaza',
       'Theater' ]] .values

X = np.nan_to_num(X)
Clus_dataSet = StandardScaler().fit_transform(X)
Clus_dataSet
```

The inputs include the price, number of bedrooms as well as the number of top 20 venue categories in the 500m. vicinity of each property obtained from Foursquare API.

Next, I will find the appropriate number of clusters to be used. To do this, I use K-means clustering by varying number of clusters and plot the resulting score as follows:

```
scores=[]

for clusterNum in range(2,20,1):
    k_means = KMeans(init = "k-means++", n_clusters = clusterNum, n_init = 12)
    k_means.fit(Clus_dataSet)
    labels = k_means.labels_
    scores.append(-1*k_means.score(Clus_dataSet))


print(scores)

plt.figure(figsize=(10,8))
plt.plot(range(2,20,1),scores)
```

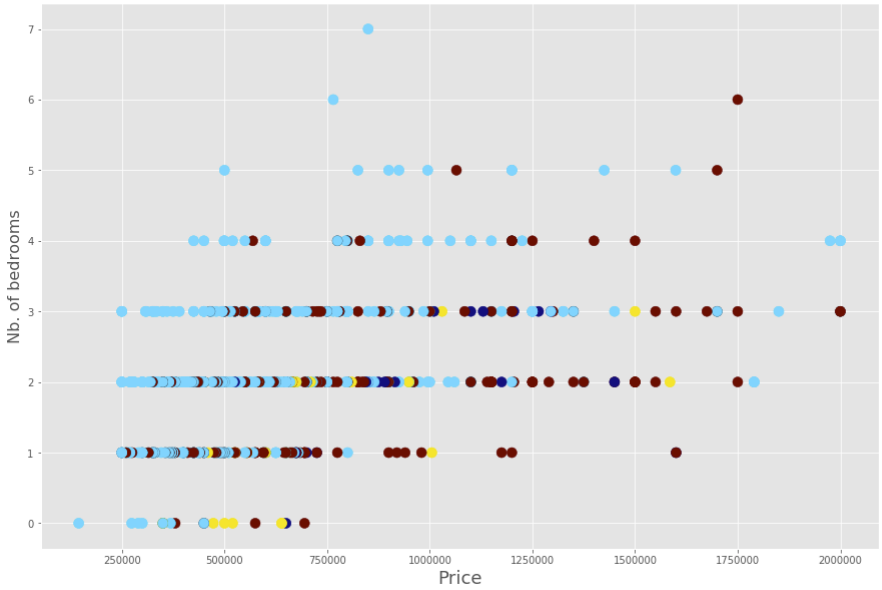
```
    [10635.377159326468, 9704.410622660705, 8629.586518516307, 8133.364856094965, 7557.3991573
    208609, 4771.243254248324, 4668.08036404795]

2]: [<matplotlib.lines.Line2D at 0x7f1abe329828>]
```



Using elbow method, 4 clusters seem to be a reasonable selection. So in the rest of the analysis, I use clustering with k=4.

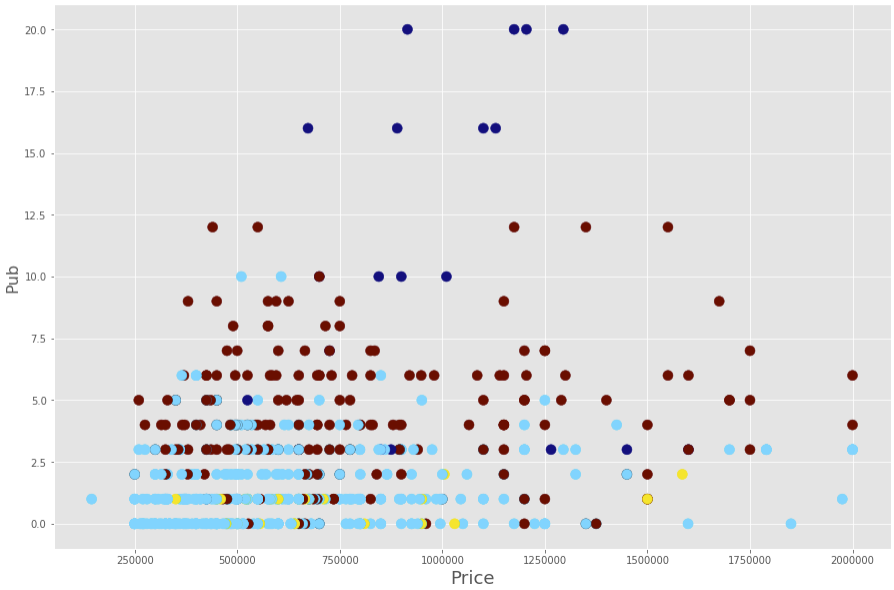
Once the listed properties are clustered into four, I next examine whether some basic trends could be identified across clusters. First, I examine the clusters on the price – number of bedrooms plane as follows.

```
plt.figure(figsize=(15,10))
plt.scatter(listing_data_post['price'], listing_data_post['number_bedrooms'], s=100, c=listing_data_post['Clus_km'], cmap='jet')
plt.xlabel('Price', fontsize=18)
plt.ylabel('Nb. of bedrooms', fontsize=16)

plt.show()
```



Next, I examine the clusters on the price – number of pubs in 500m. vicinity plane (note that pubs seem to be the most common nearby venues in London).

```
plt.figure(figsize=(15,10))
plt.scatter(listing_data_post['price'], listing_data_post['Pub'], s=100, c=listing_data_post['Clus_km'], cmap='jet')
plt.xlabel('Price', fontsize=18)
plt.ylabel('Pub', fontsize=16)

plt.show()
```

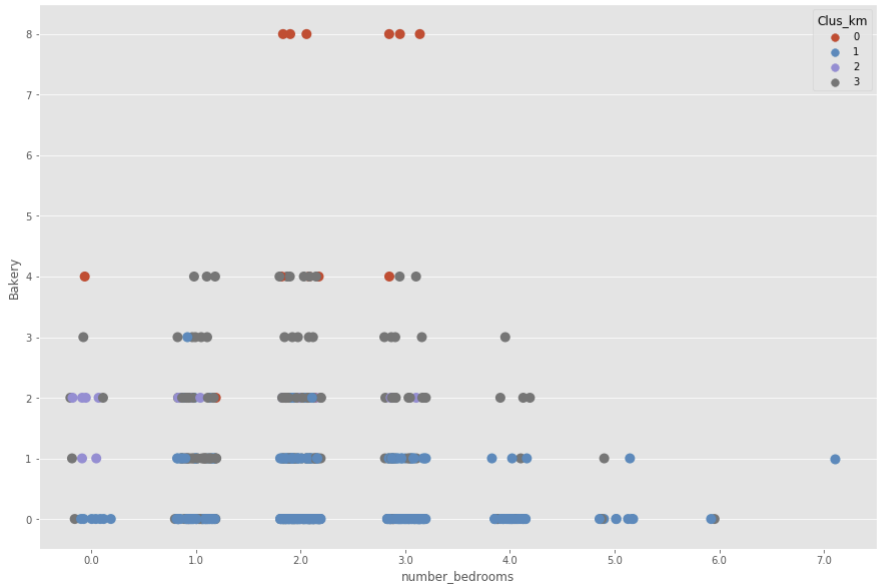


Finally, I plot the clusters on the number of bedrooms – number of bakeries in 500m. vicinity plane.

```
import seaborn as sns
```

```
plt.figure(figsize=(15,10))
sns.stripplot('number_bedrooms', 'Bakery', data= listing_data_post, jitter=0.2, hue='Clus_km', size=10)
```

```
78]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1abd761ba8>
```



4. **Discussion**

This analysis results in a couple of observations. 1. Price by itself is far from being a sufficient similarity measure to base recommendations on. 2. Number of nearby pubs seem to be a good indicator of the cluster the property belongs to. Roughly, the clusters are distributed based on 0-3, 3-10 and more than 10 pubs. The number of pubs is likely a proxy for how central and lively the neighborhood is. A similar relationship holds for the number of nearby bakeries as well. 3. The number of bedrooms by itself is insufficient to characterize similar properties that belong to the same cluster.

I also check whether clusters are mostly based on postcodes as that could explain the clusters' dependency on the number of venues around. The data frame below illustrates the postcode and the distinct cluster numbers the properties in the postcode belong to.

```
listing_data_post.groupby('postcode')['Clus_km'].unique()
```

```
postcode
E1        [0, 1]
E10          [1]
E11          [1]
E12          [1]
E13       [1, 0]
E14    [1, 2, 0]
E15          [1]
E16          [1]
E17       [0, 1]
E18          [1]
E2     [3, 1, 0]
E20       [0, 1]
E3           [1]
E4           [1]
E6           [1]
E7           [1]
E8           [1]
E9           [1]
EC2          [0]
N1     [1, 0, 3]
N10       [1, 0]
N11          [1]
N12          [1]
N13          [1]
N14          [1]
N15          [1]
N16       [1, 0]
N17          [1]
N19          [0]
N2        [1, 0]
          ...
SW12         [1]
SW13         [1]
SW14         [1]
SW15      [1, 0]
SW16         [1]
SW17      [0, 1]
SW18      [1, 0]
SW19         [1]
SW20         [1]
SW3          [0]
SW4          [0]
SW5          [0]
SW6       [0, 1]
SW7          [0]
SW8          [1]
SW9          [1]
W10          [1]
W11       [0, 1]
W12       [0, 1]
W13       [0, 1]
W14       [0, 1]
```

As this data frame shows, each postcode can include properties belonging to different clusters. This shows that real estate recommendations made to customers based on the postcode would likely fall short and a cluster analysis would provide more accurate and comprehensive recommendations of similar properties.

5. Conclusion

In this project, I tried to come up with a data-based approach to providing recommendations for customer that are on the market to buy a property. In real estate, customers typically contact the agent once they are interested in a property listed with the company. Even if customer may not necessarily purchase

the property they were first interested in, if they are provided relevant recommendations, they can view and purchase other similar properties listed with the same company. Hence, given customer's interest in a specific property, being able to identify and recommend similar properties is a crucial skill for realtors to close a sale.

The k-means clustering approach used in this project is one approach that can easily be used by real estate companies to provide accurate and up-to-date recommendations to their customers. For example, a real estate company can simply use the RealMove.co.uk data I used in this study and obtain the clusters of each listed property. This way, when a customer calls in who is interested in a specific property, even if the customer does not purchase this property, the agent can recommend other properties that belong to the same cluster, and hence similar in nature. In practice, clusters should be updated at least every day to make sure all available listings posted online are included in the analysis.