

# Analysis of London Real Estate Market

Recommending Similar Properties on Sale

May 4<sup>th</sup>, 2020

# Background

Real estate industry is based on identifying customer's taste and presenting properties that are in alignment with customer needs and preferences.

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.

# Background

- Recommending similar properties is also crucial to the success of real estate recommender websites such as <https://www.compass.com/>.

## Compass Exclusives

Be the first to browse exclusive listings before they hit the market.

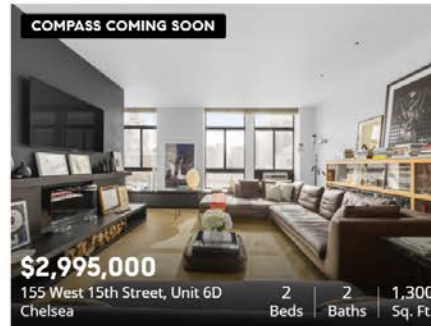
**COMPASS COMING SOON**



**\$1,049,500**  
367 Rea Street  
North Andover, MA 01845

4	3.5	4,186
Beds	Baths	Sq. Ft.

**COMPASS COMING SOON**



**\$2,995,000**  
155 West 15th Street, Unit 6D  
Chelsea

2	2	1,300
Beds	Baths	Sq. Ft.

**COMPASS PRIVATE EXCLUSIVES**

Work with a Compass agent to see private exclusives.

[Learn More](#)

**COMPASS COMING SOON**



**Contact Agent**  
40 Hampton Road  
Scarsdale, NY 10583

6	8	9,911
Beds	Baths	Sq. Ft.

**COMPASS COMING SOON**

**Coming Soon**

**\$1,359,000**  
9049 Hargis Street  
Los Angeles, CA 90034

3	1.75	1,366
Beds	Baths	Sq. Ft.

**COMPASS COMING SOON**



**\$4,495,000**  
6851 Cahuenga Park Trail  
Los Angeles, CA 90068

8	7	5,287
Beds	Baths	Sq. Ft.

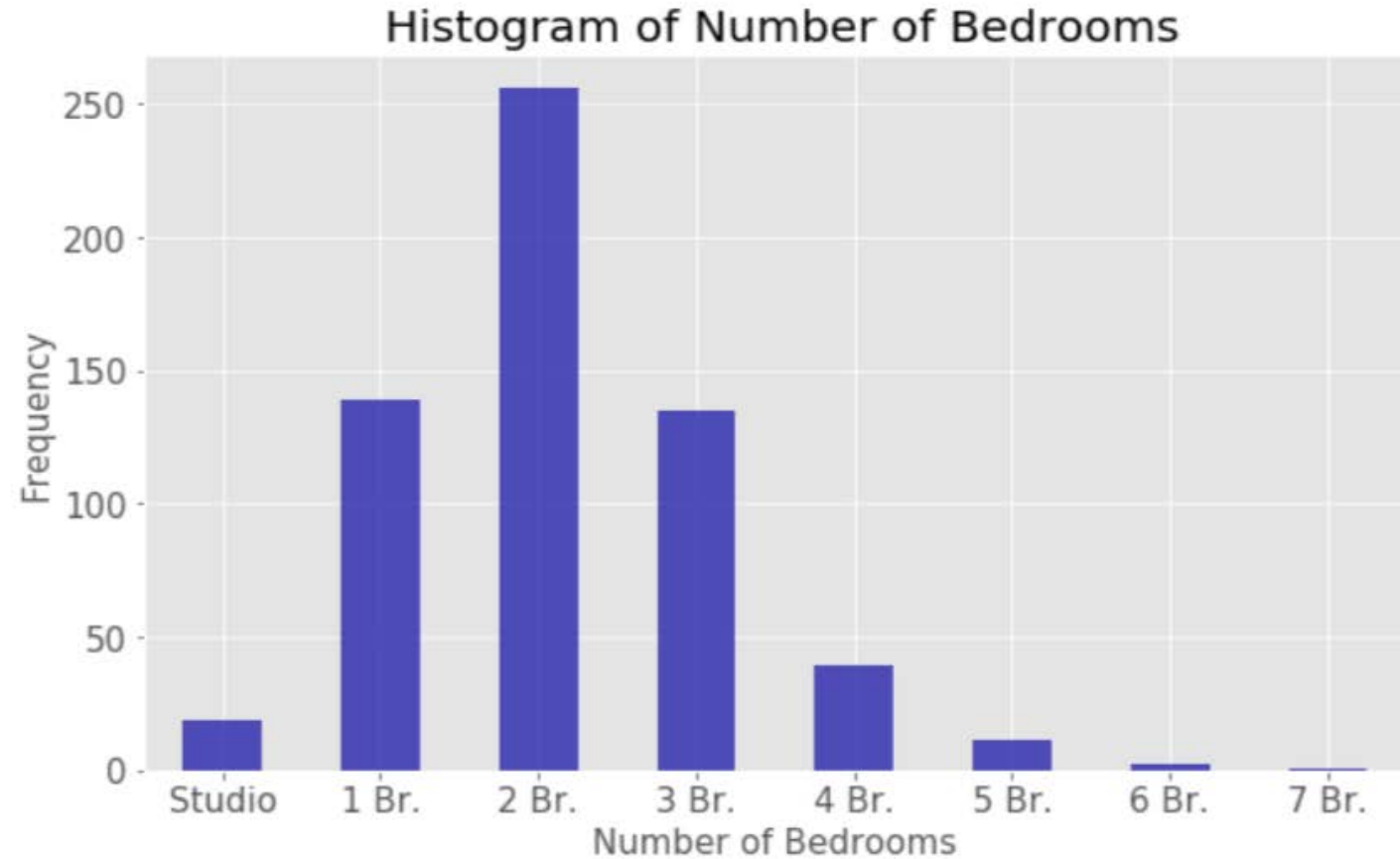
# Goal of the Project

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.

# Data

- Clustering similar properties on sale in London real estate market requires combining several datasets:
- List of properties on sale  
RightMove.co.uk data obtained from rightmove-webscraper.
- List of postcodes and district names in London  
Scraped from [https://en.wikipedia.org/wiki/London\\_postal\\_district](https://en.wikipedia.org/wiki/London_postal_district)
- Latitude and longitude of each property using OpenCage API
- List of venues nearby for each property using Foursquare API

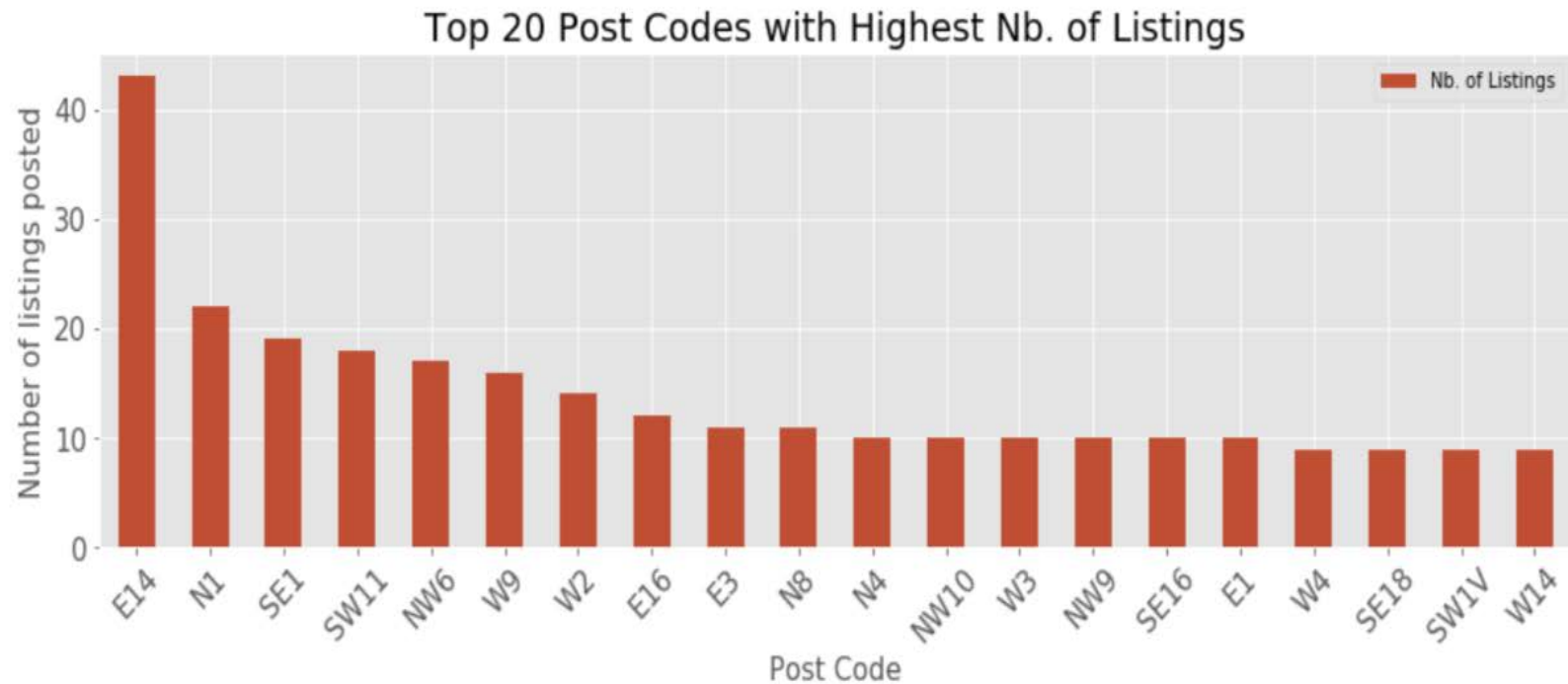
# Methodology and Results – Exploratory Analysis



# Methodology and Results – Exploratory Analysis

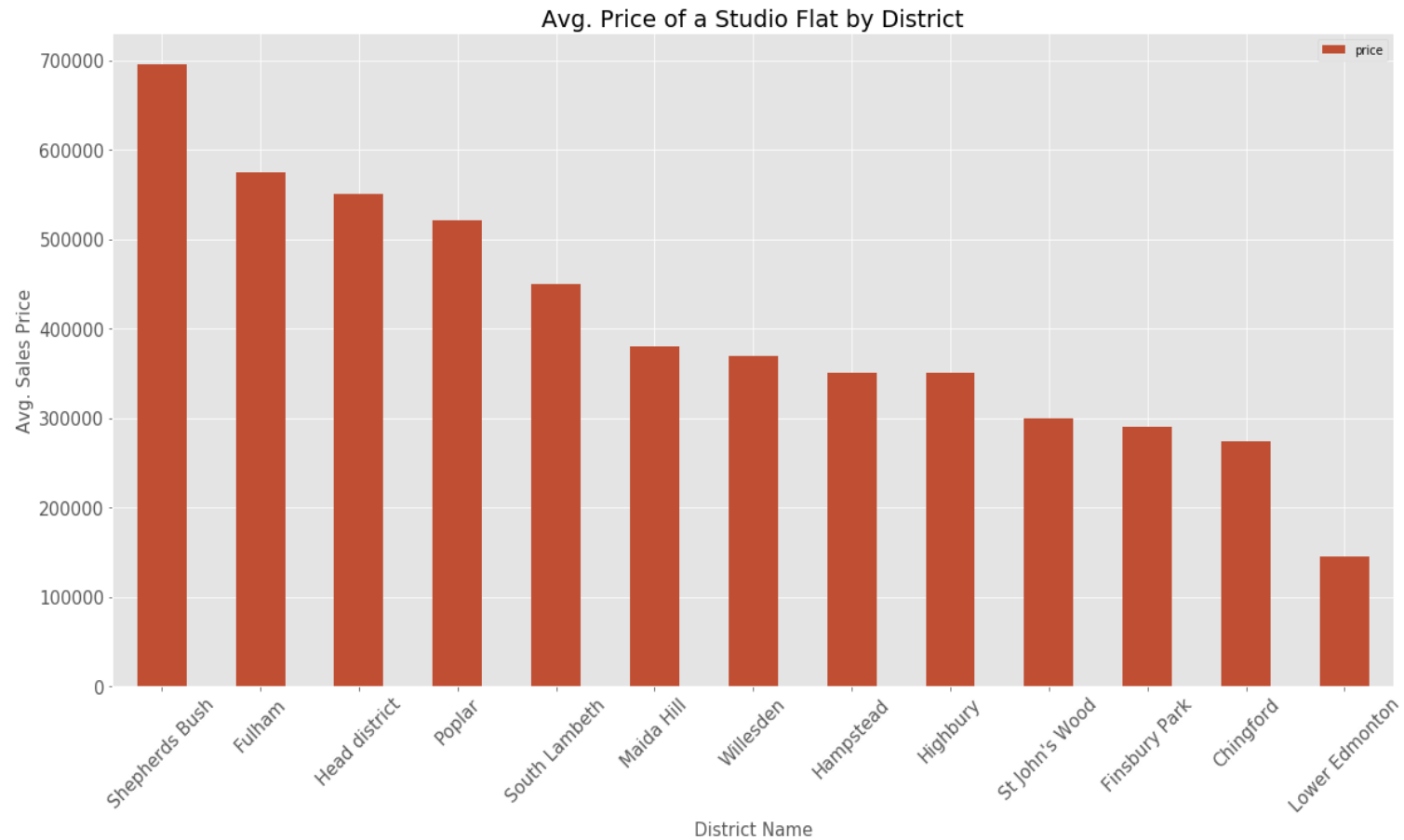


# Methodology and Results – Exploratory Analysis

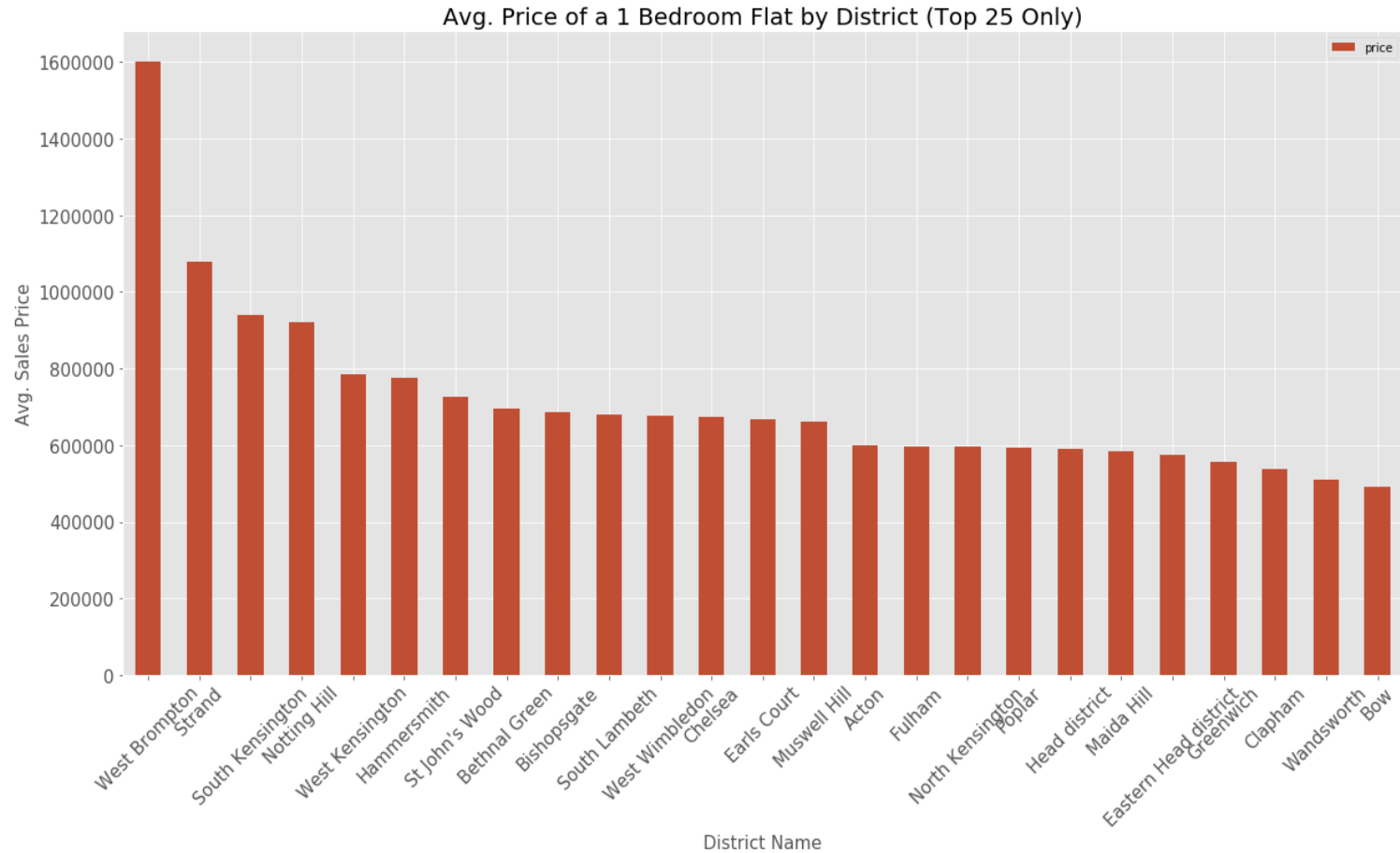




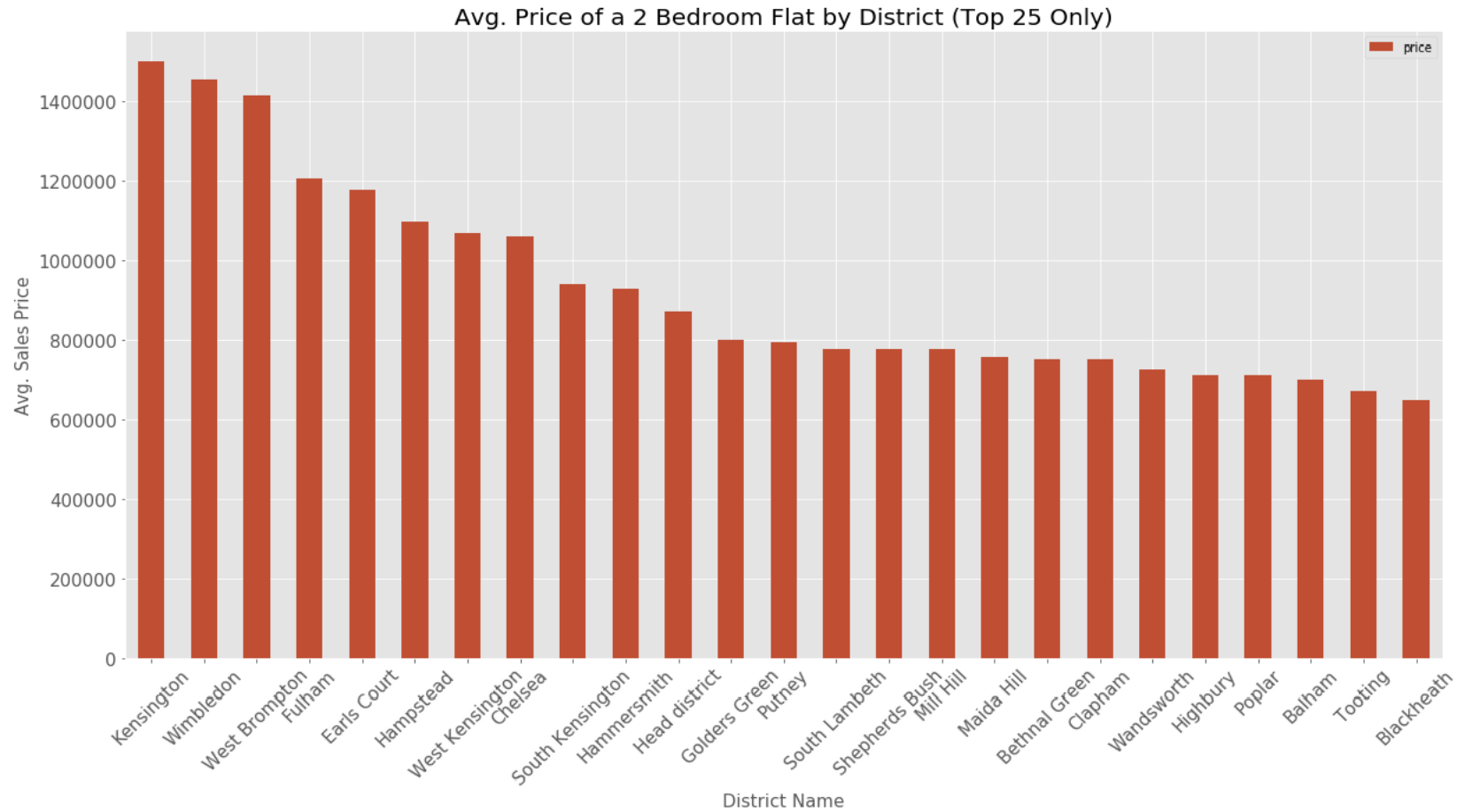
# Methodology and Results – Exploratory Analysis



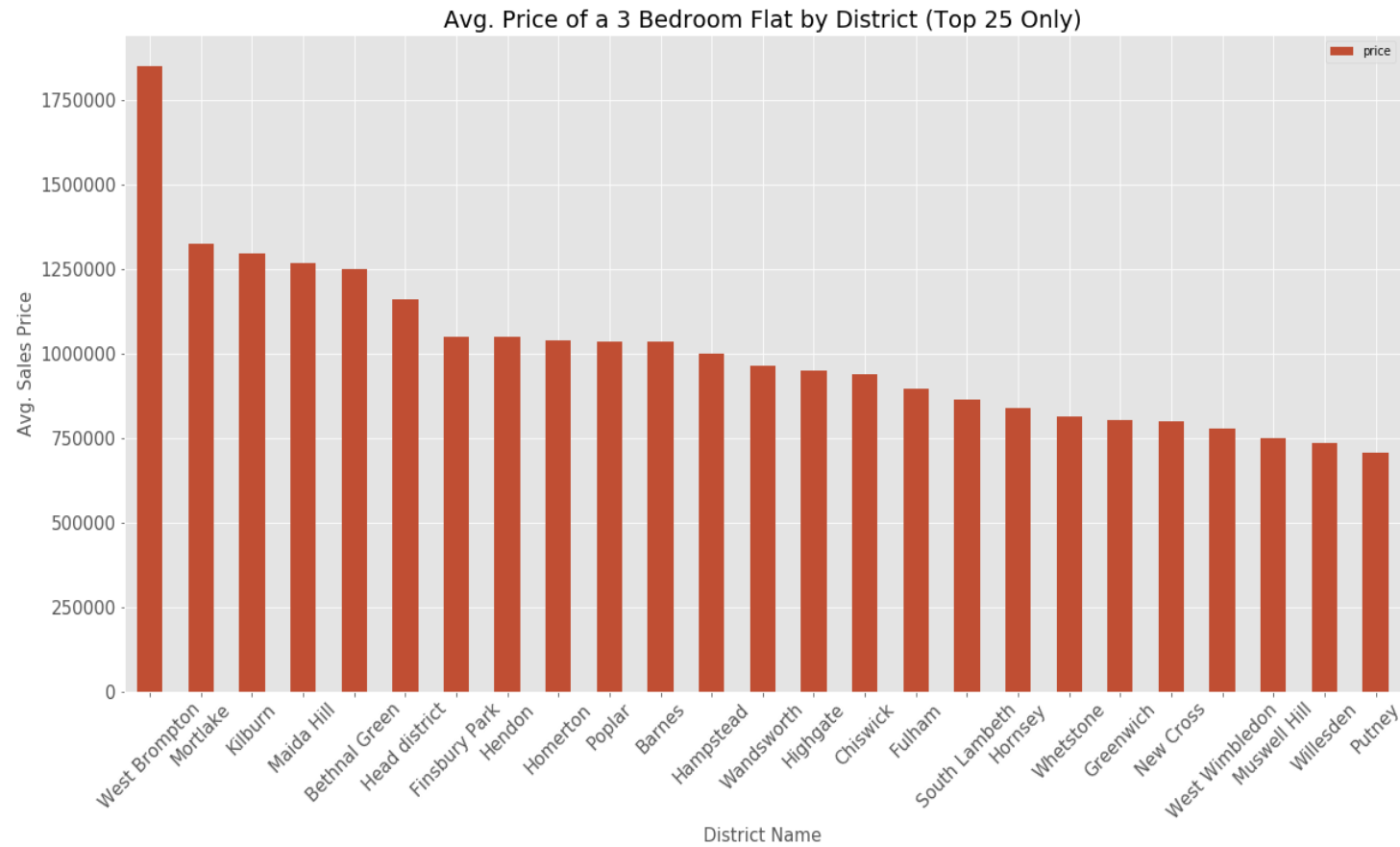
# Methodology and Results – Exploratory Analysis



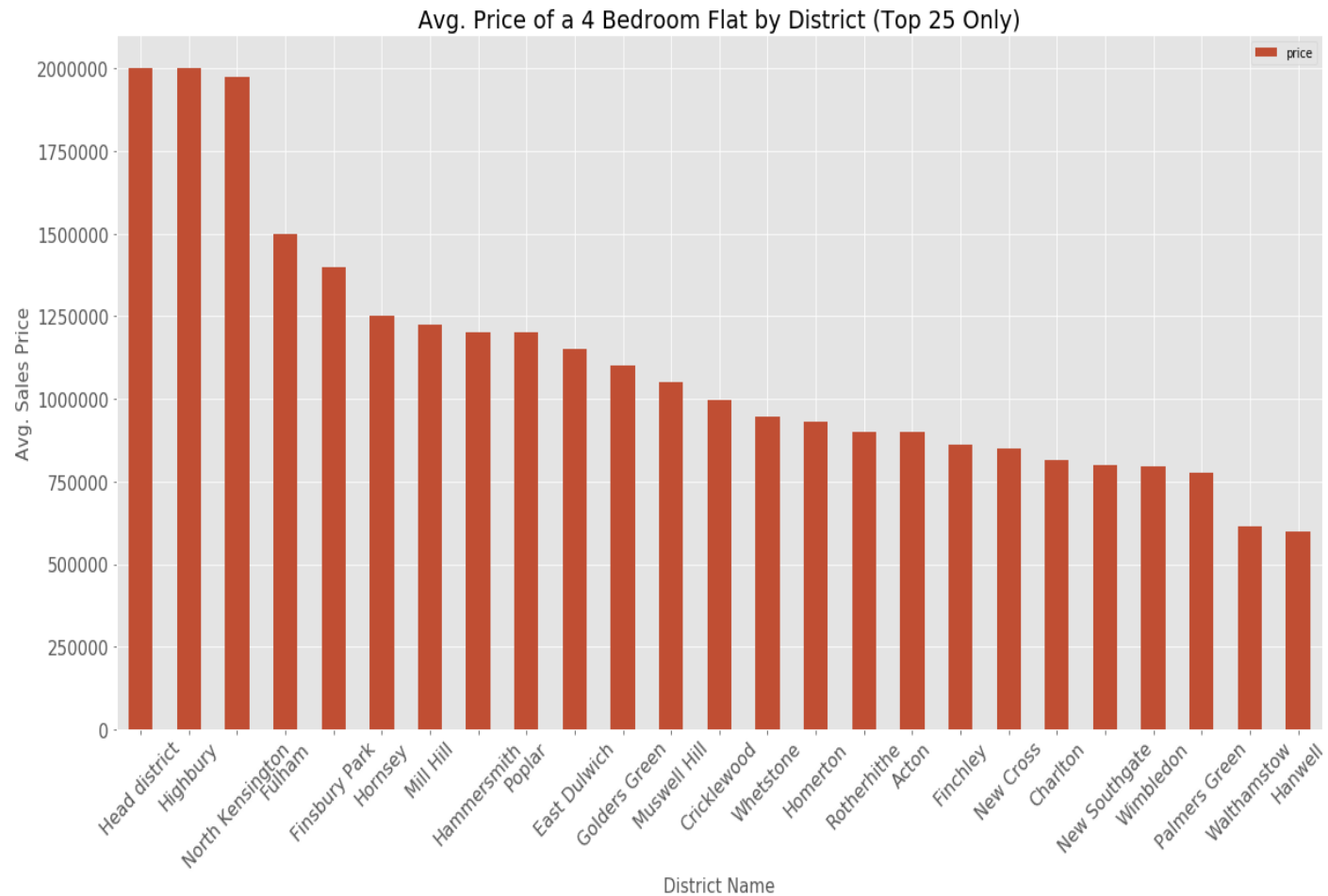
# Methodology and Results – Exploratory Analysis



# Methodology and Results – Exploratory Analysis



# Methodology and Results – Exploratory Analysis



# Methodology and Results – K-means Clustering

```
) 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
```

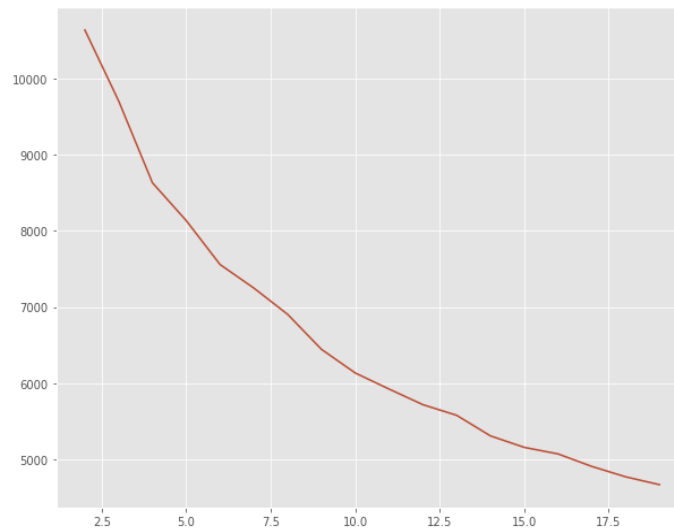
The inputs include the price, number of bedrooms as well as the number of top 20 venue categories in the 500m. vicinity of each property obtained from Foursquare API.

# Methodology and Results – K-means Clustering

- Next, I will find the appropriate number of clusters to be used.

```
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)
```

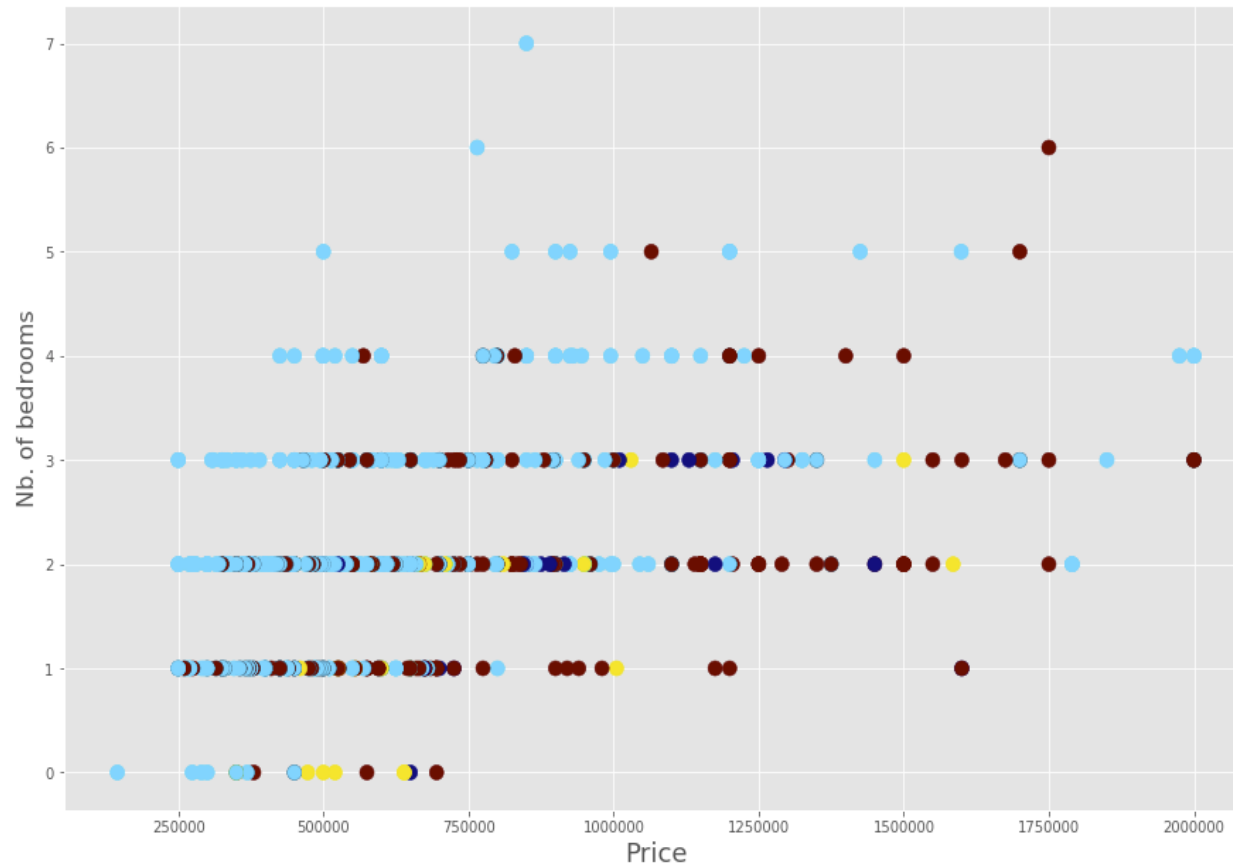
[10635.377159326468, 9704.410622660705, 8629.586518516307, 8133.364856094965, 7557.39915738,  
208609, 4771.243254248324, 4668.08036404795]  
2]: [matplotlib.lines.Line2D at 0x7f1abe329828]



**Proceed with k=4.**

# Methodology and Results – K-means Clustering

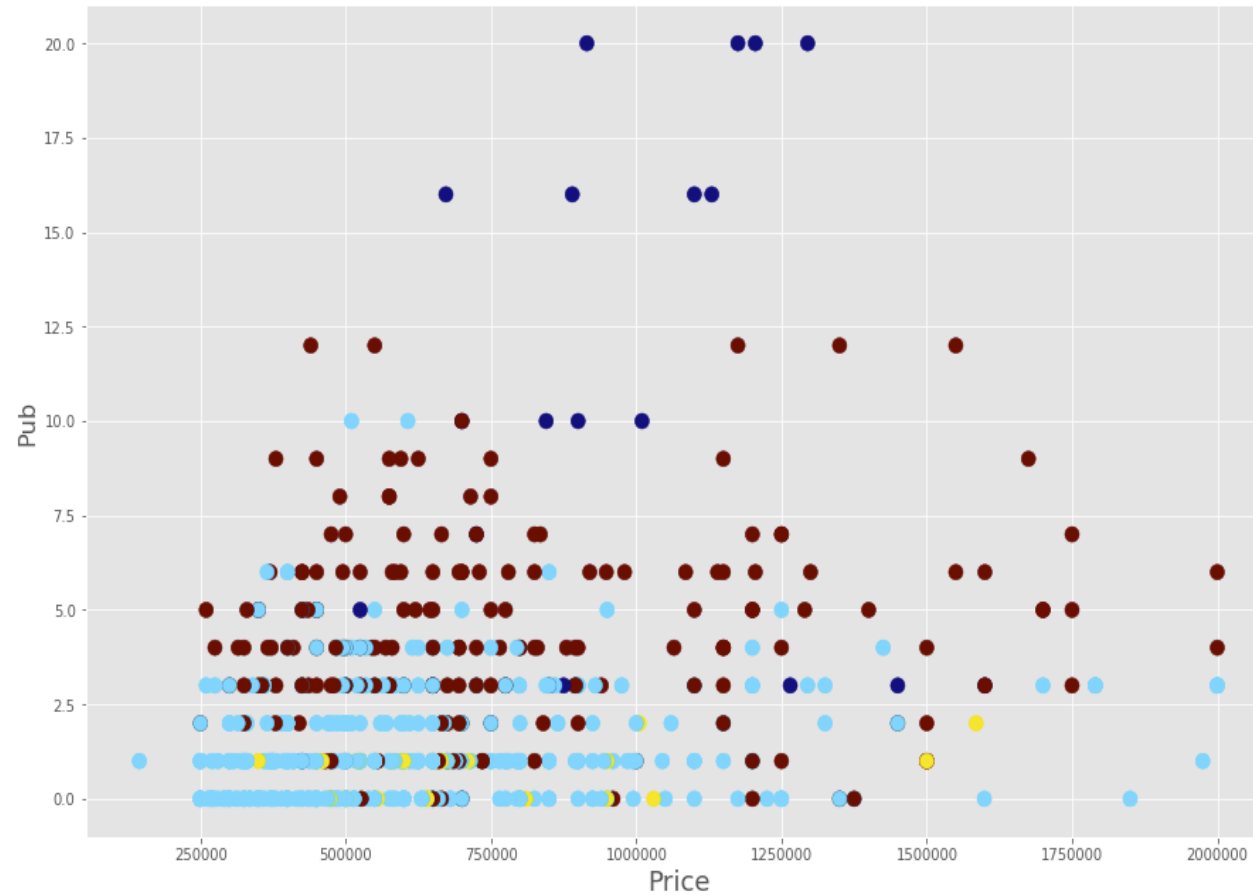
```
plt.figure(figsize=(15,10))
plt.scatter(listing_data_post['price'], listing_data_post['number_bedrooms'], s=100, c=listing_data_post['Clus_km'], cmap='jet')
plt.xlabel('Price', fontsize=18)
plt.ylabel('Nb. of bedrooms', fontsize=16)
plt.show()
```





# Methodology and Results – K-means Clustering

```
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()
```

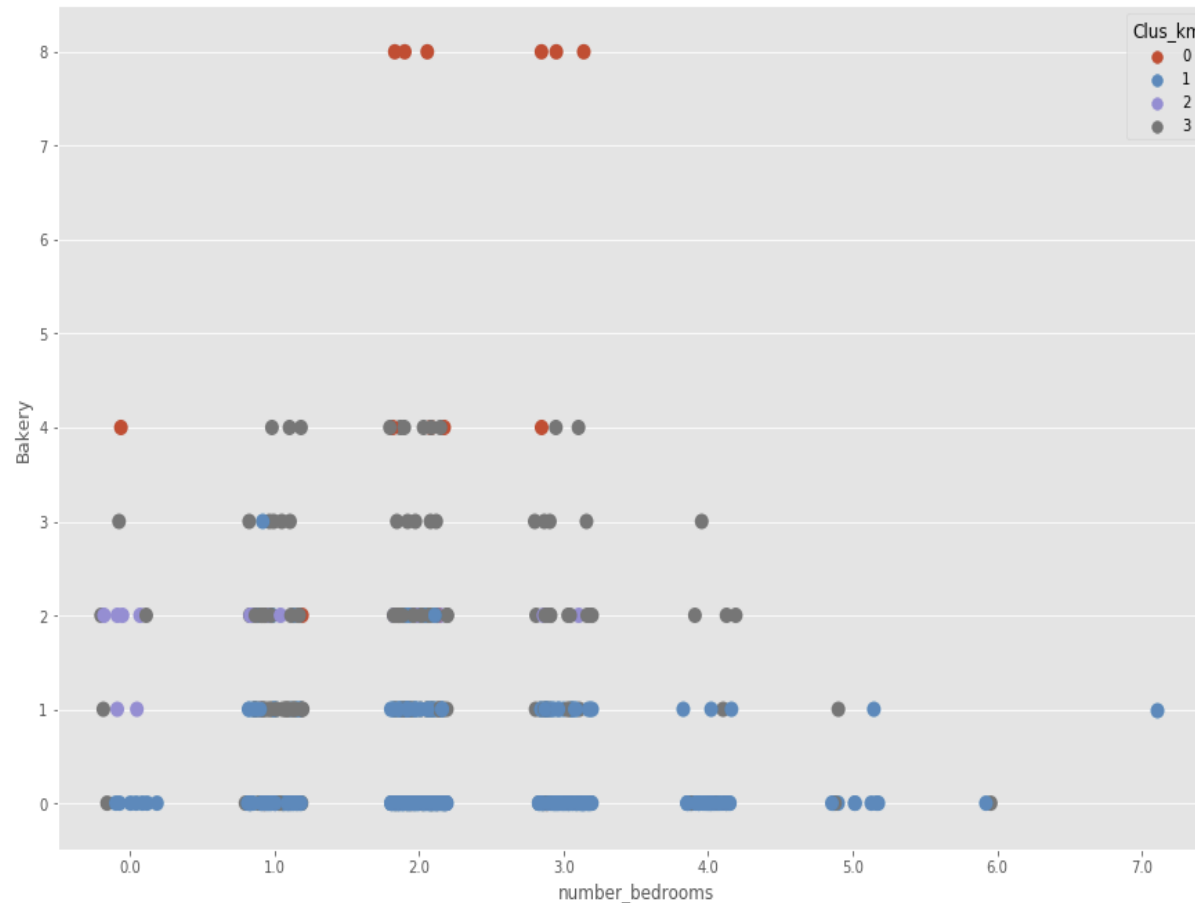


# Methodology and Results – K-means Clustering

```
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>



# Discussion – Key Observations

- Price by itself is far from being a sufficient similarity measure to base recommendations on.
- 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.
- The number of bedrooms by itself is insufficient to characterize similar properties that belong to the same cluster.
- 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.

# Conclusion

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.