Add credit and caption

Housing Rent Prices and Venues Data Analysis of London - A geographical clustering approach for house seekers.

Understanding the environment - London Population 2019

9,176,530

The latest official estimate of the population of London comes from the Office for National Statistics. According to their data, the estimated population of Greater London in 2016 was **8,787,892**. The metro **population in 2019 is estimated** to be as much as **9.18 million.**

The Census in the United Kingdom takes place every ten years, with the most recent one completed in 2011 this means that we are close to the next demographic data collection.

London's population makes it by far the largest city in the United Kingdom. The second-largest city in the UK - **Birmingham** - has a population of 1.1 million which is **11,98% only the population of the Capital City. London** is also the largest city in the European Union, **twice the size of Dublin** and **three times the size of Rome.**

It is the third-largest city in Europe, behind Istanbul (14.8 million) and Moscow (10.3 million), and the 27th most populous metro area in the world.

Ethnicity

London as a city is **considerably more diverse** than the rest of the United Kingdom. Across England and Wales, 86% of the population is white based on the 2011 Census, but in London that number falls to 59,79%. The white proportion of London's population increases when travelling away from the city centre.

Is important to understand **how the diversity of ethnicity** of the population **change the way we analyze the housing market.** People have **different needs and priority** to satisfy and, when it's time to rent a flat, they are more likely available to rent a flat in an area with the presence of specific venues at walk distance.

White: Gypsy or Irish Traveller[Note 2]					8,196	0.10%	n/a
White: Other[Note 3]			594,854	8.29%	1,033,981	12.65%	▲ 73.82%
White: Total	**5,333,580**	**79.80%**	**5,103,203**	**71.15%**	**4,887,435**	**59.79%**	**▼ 4.23%**
Asian or Asian British: Indian	347,091	5.19%	436,993	6.09%	542,857	6.64%	▲ 24.24%
Asian or Asian British: Pakistani	87,816	1.31%	142,749	1.99%	223,797	2.74%	▲ 56.78%
Asian or Asian British: Bangladeshi	85,738	1.28%	153,893	2.15%	222,127	2.72%	▲ 44.34%
Asian or Asian British: Chinese[Note 4]	56,579	0.84%	80,201	1.12%	124,250	1.52%	▲ 54.92%
Asian or Asian British: Other Asian	112,807	1.68%	133,058	1.86%	398,515	4.88%	▲ 199.51%
Asian or Asian British: Total	**690,031**	**10.33%**	**946,894**	**13.20%**	**1,511,546**	**18.49%**	**▲ 59.63%**
Black or Black British: African	163,635	2.44%	378,933	5.28%	573,931	7.02%	▲ 51.46%
Black or Black British: Caribbean	290,968	4.35%	343,567	4.79%	344,597	4.22%	▲ 0.3%
Black or Black British: Other Black	80,613	1.20%	60,349	0.84%	170,112	2.08%	▲ 181.88%
Black or Black British: Total	**535,216**	**8.01%**	**782,849**	**10.92%**	**1,088,640**	**13.32%**	**▲ 39.06%**
Mixed: White and Black Caribbean			70,928	0.99%	119,425	1.46%	▲ 68.38%
Mixed: White and Black African			34,182	0.48%	65,479	0.80%	▲ 91.56%
Mixed: White and Asian			59,944	0.84%	101,500	1.24%	▲ 69.33%
Mixed: Other Mixed			61,057	0.85%	118,875	1.45%	▲ 94.70%
Mixed: Total[Note 5]			**226,111**	**3.15%**	**405,279**	**4.96%**	**▲ 79.24%**
Other: Arab[Note 6]					106,020	1.30%	n/a
Other: Any other ethnic group			113,034	1.58%	175,021	2.14%	▲ 54.84%
Other: Total	**120,872**	**1.81%**	**113,034**	**1.58%**	**281,041**	**3.44%**	**▲ 148.63%**
Total	6,679,699	100.00%	7,172,091	100.00%	8,173,941	100.00%	▲ 13.97%

Introduction and Business problem presentation

We could identify **3 main reasons** why a flat doesn't fit the customer needs:

- *The flat looks old and stale*
- *The neighbour hasn't the expected commodities nearby*
- *The price is too high for that particular flat or out of budget*

Our goal with this Notebook is to have a systematic way to analyze the offers posted by ***RightMove.co.uk*** to produce a map of the best opportunities in the city. If you are looking for a new flat and you like your actual neighbour, we can provide you with a list of all the opportunity on the market.

For this project, I'm going to create simple software that scrapes the website **RightMove** to collect an updated list of flat for rent, **collect and analyze the main venues near each flat** offered on the market using **Foursquare** and **cluster them** in order to divide the housing market **into 20 groups by venues similarity in a radius of 500 meters.**

Methodology

For this particular analysis, we are going to *collect updated data from RightMove.co.uk.*

To do so, I decided to spend time developing a web scraping application using **Beautiful Soup 4**, but then I discovered a repository on GitHub offered by *toby-p* and available on his profile, that provides an easy way to **scrape Rightmove!**

- type
- address
- URL
- agent_url
- postcode
- number_bedrooms
- search_date

A record will look like the following:

	price	type	address	url	agent_url	postcode	number_bedrooms	search_date
0	5265	3 bedroom maisonette	Abingdon Road, High Street Kensington, London W8	http://www.rightmove.co.uk/property-to-rent/pr...	http://www.rightmove.co.uk/estate-agents/agent...	W8		Add description
1	1500	1 bedroom apartment	Askew Road, London, W12	http://www.rightmove.co.uk/property-to-rent/pr...	http://www.rightmove.co.uk/estate-agents/agent...	W12	1.0	2019-12-04 19:53:51.373561
2	1278	Studio flat	Montagu Row, Marylebone, London W1U	http://www.rightmove.co.uk/property-to-rent/pr...	http://www.rightmove.co.uk/estate-agents/agent...	W1U	0.0	2019-12-04 19:53:51.373561

The address is in the format *"Street, City, Postcode"* and is an unstructured field but for our purpose, we can leave as it is. Instead, the PostCode present a *"limited"* format because we have the first two/three digits only. **This is not accurate enough to collect meaningful data about the venues around the flat.**

To solve this problem, we are going to use **OpenCage Geocoder API** to look up coordinates from a postal address. **This is a case when an unstructured field becomes helpful.**

To associate each rent offer to a District, we are going to join the data table with a second dataset presenting two columns:

- District Name
- PostCode

This dataset had been created **scraping a Wikipedia Table** (available here) with the data we need for the analysis.

When the data are collected and merged into a single data frame, we are going to **cluster** them **using the K-Means algorithm**. To visualize **geographic details** and the distribution of the offers in London, I plotted **2 meaningful maps using folium** Python library:

- **Clusters map**: this map shows the distribution of the clusters using colours to identify each cluster.

the average price for a **studio flat, 1 bedroom flat, 2 bedrooms flat, 3 bedrooms flat and 4 bedrooms flat grouped by the District Name**.

One of the questions to answer is to **quantify the magnitude of the impact of the location (District) on the average price for each category** of apartments and identify the best number of bedrooms we should look for to **minimize the geographic influence on the monthly fee**.

Finally, I concluded the project by asking **the user to input the following data**:

- *Your address*: this input is used to analyze the neighbourhood you are living in and to use this information to find the cluster you belong to.
- *The number of bedrooms you are looking for*: this input is used to filter the results of the cluster you belong to.
- *The Maximum monthly fee (budget)*

This part of the analysis has as an output, a data frame with a list of filtered results based on your preferences.

Results

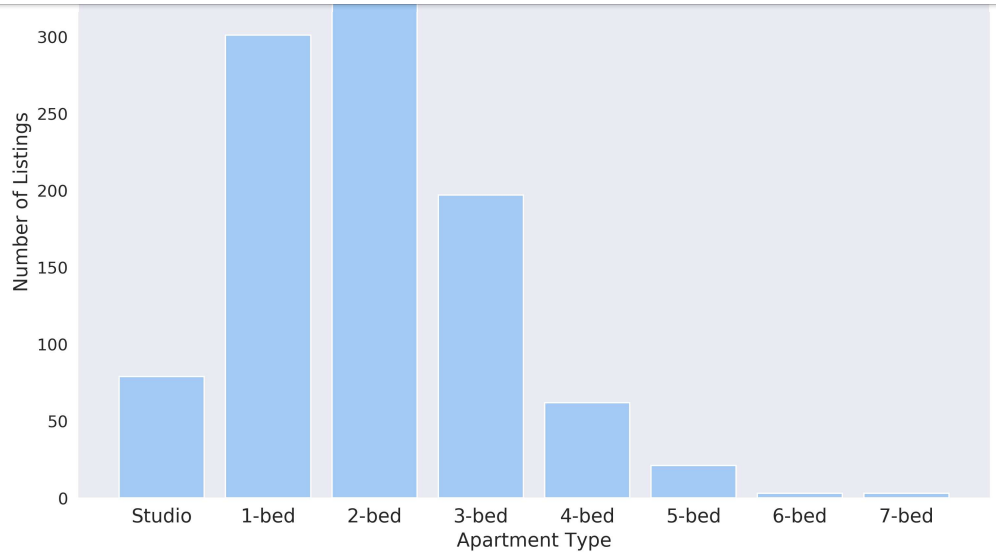
As expected, the price of a flat can't be forecasted by the venues around it only but **the market could be filtered effectively to help the house-seeker to find the best opportunities available**. Nevertheless, *is possible to develop a pricing model based on the characteristics of the flat* (number of bedrooms, number of bathrooms etc)*, the District the flat belongs to, and the presence of some key venues nearby the flat*. An example of key factors could be the presence of supermarkets with high reputation, a public transport stop, schools or Universities, Hospital. **The correlation between price and these categories is low but still important based on the preferences of the final users.**

The goal of this project is **to provide everyone with a way to scrape the housing market and identify the best offers that fit the user's personal needs.**

Data Exploration

We can take a look at the raw data and visualize different data distributions to start to understand how the market is organized.

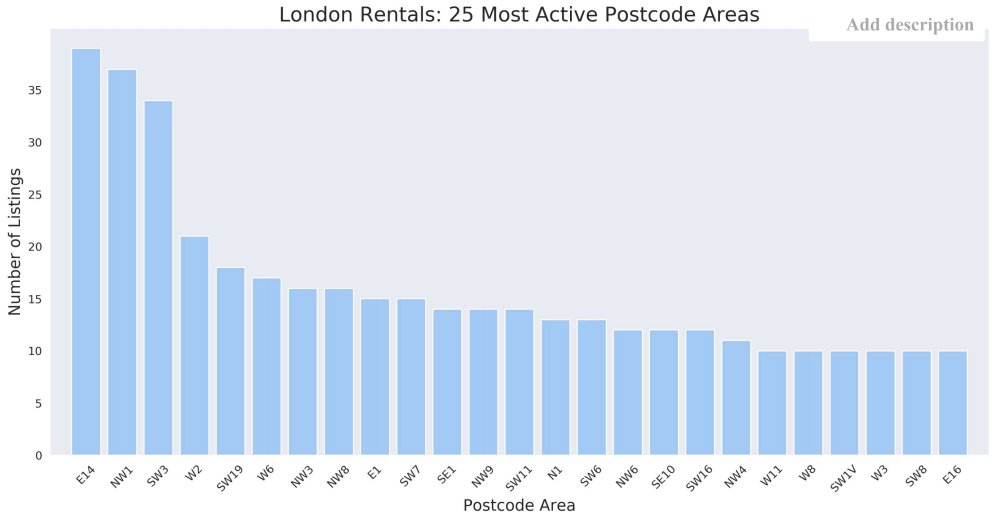
We are going to **plot a bar chart** that shows the **number of offers listed by the number of bedrooms.**



As we can see, the **most frequent** offers on the market are about **2 bedrooms apartment category** followed by the **1 bedroom apartment category.**

Which district has the most offers in London?

We can plot a new **bar chart to explore the distribution of offers by postcode.**



Data Manipulation

In order to make the results easier to read and interpret, I decided to **join each postcode to the District Name**. This step is of **primary importance to make the results enjoyable to the final user.**

To define the location of each District, we are going to identify the **latitude and longitude** of the centre **of each District using OpenCageData.**

This API permits to search and grab geographical information based on latitude and longitude or address. We are going to use the postcode of each District to expand our data

0	E1	Eastern Head district	51.489334	-0.144055
1	E2	Bethnal Green	51.489334	-0.144055
2	E3	Bow	51.514947	-0.093046
3	E4	Chingford	51.507322	-0.127647
4	E5	Clapton	51.514947	-0.093046

Having the latitude and longitude of each district could be useful for future and more in-depth analysis such as the distance from the centre of the District as an independent variable that could impact the price of the flat.

However, for this project, it is even more important to expand our database by **associating each apartment with its geographical coordinates.**

Using OpenCage to expand the record of each flat

We can follow the same process to **expand the geographic information related to each flat**. **OpenCage Data** permits to collect geographical information **using the address as a query.**

In order to have enough data to work with, is better to collect the following details:

- latitude of the flat
- longitude of the flat
- county
- complete postcode
- state district
- suburb

The dataset looks like the following:

	i	address	Latitude_a	Longitude_a	county	Postcode_complet		
0	0	Abingdon Road, High Street Kensington, London W8	51.501600	-0.213300	NaN	W8	NaN	NaN
1	1	Askew Road, London, W12	51.505232	-0.244442	London Borough of Hammersmith and Fulham	W12 9HD	Greater London	Brook Green
2	2	Montagu Row, Marylebone, London W1U	51.508500	-0.125700	NaN	W1U	NaN	NaN

Add description

To be sure that each address has been merged with the correct series of new information, I used the 'i' column as a flag.

This solution permitted me to double-check that the previous index matched the new index, this means that **I dropped the records where key data are missing.**

- **District Name**
- **Latitude** *and* **Longitude**: lat and lng of the "District Name"
- **Address_y**: a copy of the original address that we'll drop. I used it to be sure that the DataFrames have been merged correctly
- **Latitude_a** *and* **Longitude_a**: lat and lng of the flat.
- **County**
- **Postcode_complete**: an extension of the original PostCode
- **State_district**
- **suburb**

Foursquare API - Find the most common venues near the flat

Collect meaningful information about the most common venues around each flat.

This step is crucial to **clustering the market training a K-Means model.**

We are going to use the **Foursquare API** to collect the **first 100 venues** in a **radius of 500 meters around each flat posted on Rightmove.**

Address and Venues

The process of collecting meaningful information about the venues around each flat is the key to perform the clustering analysis for this project.

The following table shows the number of venues collected for the first three records from the dataset:

address_x	i	price	type	address Latitude	address Longitude	Venue	Venue Latitude	Ver... Lo...	Ver... Ca...
1 Warminster Road, South norwood, SE15	298	925	1 bedroom flat	1	1	1	1	1	1
Hargood Road, Blackheath, London, SE3	1039	2250	4 bedroom terraced house	5	5	5	5	5	5
Heathway Court	540	2492	3 bedroom apartment	1	1	1	1	1	1

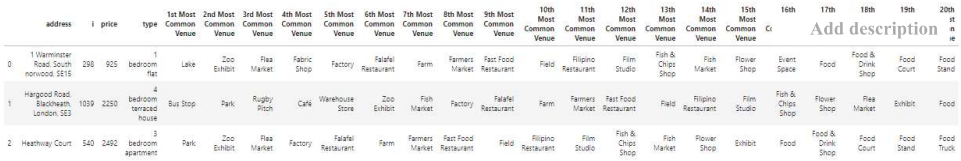
Add description

In this case, the first 3 records are not particularly significative of the process but, as we will see later on, some addresses present a larger amount of venues nearby.

We are working with categorical variables. **A categorical variable** is a variable that can take on one of a **limited**, and usually **fixed number**, **possible values**, assigning each individual or other units of observation to a particular group or nominal category on the basis of some **qualitative property.**

To make the analysis faster and tinier, we are going to find the **top 20 most common venues** that will be used to develop the **clustering analysis**.

Finally, we have a dataset ready to be used for our purpose!



	address	i	price	type	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	16th	17th	18th	19th	20th
																					Add description			
0	1 Warminster Road, South norwood, SE15	298	925	1 bedroom flat	Lake	Zoo Exhibit	Flea Market	Fabric Shop	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Film Studio	Fish & Chips Shop	Fish Market	Flower Shop	Event Space	Food	Food & Drink Shop	Food Court	Food Stand
1	Hargood Road, Blackheath, London, SE3	1039	2250	4 bedroom terraced house	Bus Stop	Park	Rugby Pitch	Café	Warehouse Store	Zoo Exhibit	Fish Market	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Film Studio	Fish & Chips Shop	Flower Shop	Flea Market	Exhibit	Food
2	Heathway Court	540	2492	3 bedroom apartment	Park	Zoo Exhibit	Flea Market	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Film Studio	Fish & Chips Shop	Fish Market	Flower Shop	Exhibit	Food	Food & Drink Shop	Food Court	Food Stand	Food Truck

Clustering using K-means

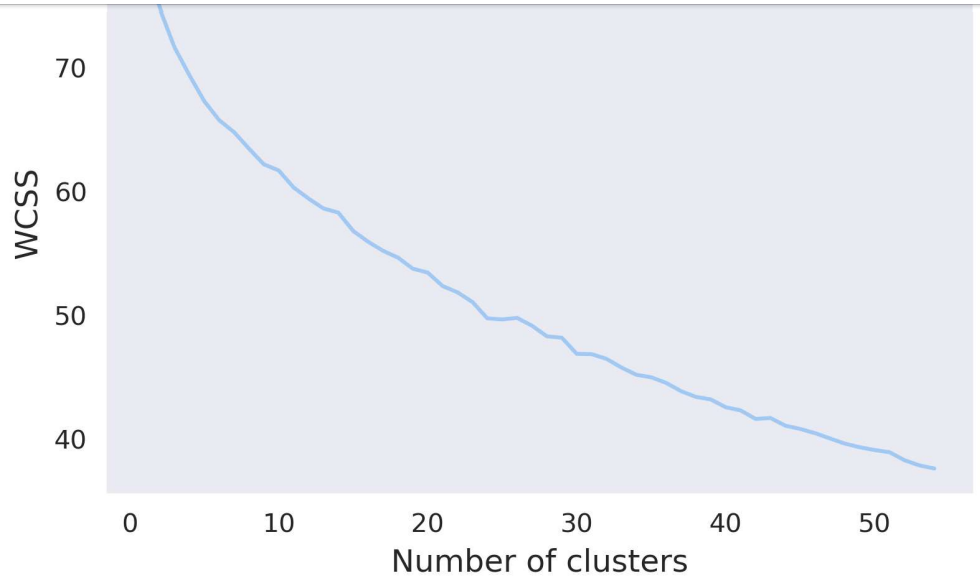
Now that we have all the information we need to divide the market by geographic similarity, it's time to split it into clusters.

We are going to use the **K-Means algorithm** for this purpose.

K-Means algorithm is one of the most popular **unsupervised machine learning algorithms**. Normally, the unsupervised algorithms make inferences using unlabelled dataset. The goal of this unsupervised machine learning technique is to find similarities in the data point and group similar data points together. In case of the **K-Means** clustering algorithm, it aims to **partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean**, serving as a prototype of the cluster.

One method to decide the optimal k is known as **"the elbow method"**.

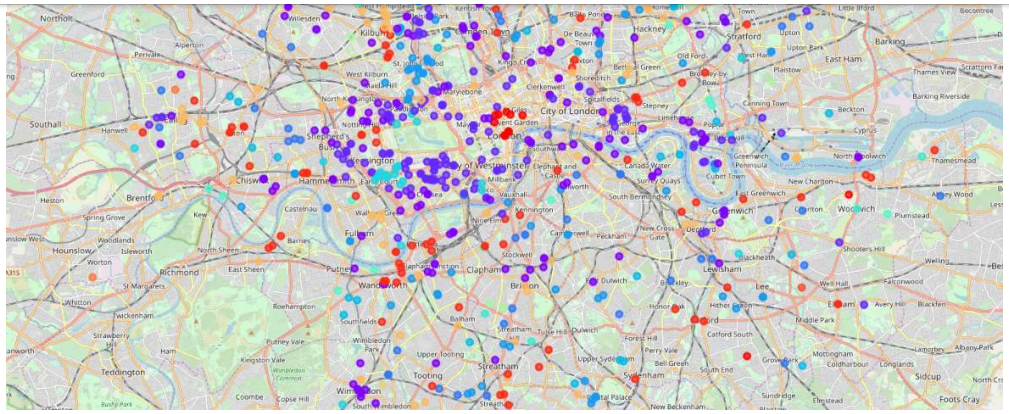
The "elbow" method helps data scientists to select the optimal number of clusters by fitting the model with a range of values for K. If the line chart resembles an arm, then the "elbow" (the point of inflection on the curve) is a good indication that the underlying model fits best at that point.



In this case, **the elbow is not definitely clear.** We will use a **k = 20** because, after several tests, the number of offers in each cluster looks better distributed into the model.

London Clusters Distribution Map

To show that the clusters are not selected by density, we can plot the map relative to the dataset. **Each colour identifies a specific cluster** and the dots represent an offer and its geographic location.

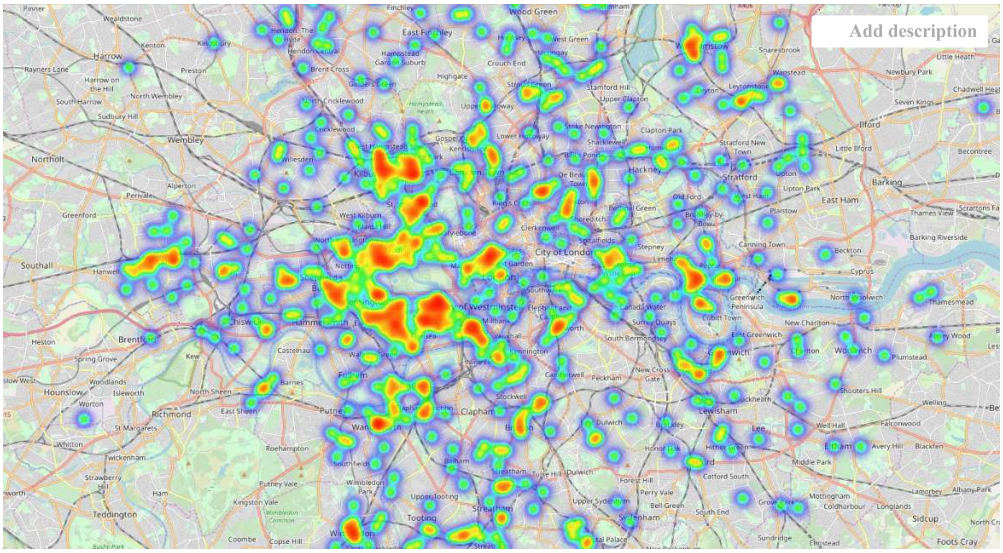


Housing market - Heating Map

Is meaningful to plot a **heating map to identify** the areas that present a **high volume of offers**. This particular view changes daily because of the high activity of the market.

A dataset with historical series records would be able to identify if there is or not any phenomenon of seasonality or peak of activity related to the expiring of the tenancy agreements.

Another interesting analysis would be interesting to conduct, would be to indagate the relation between the housing market and Brexit announcements. **Did the Brexit accelerate the cycles of the housing market promoting a major number of short term tenancy agreements to face the higher uncertainty on the future in the Country for foreign citizens?**

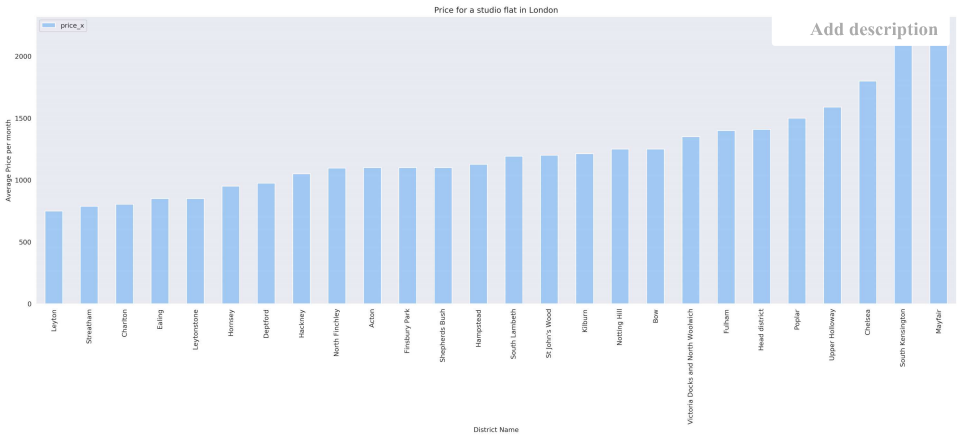


Looking at the heating map, there are **some areas** of the East-Side that **are too cold**, in my opinion. **This is probably a consequence of the algorithm Rightmove uses to show the offers to the user.** In fact, querying the website to look for the offers in London without any other specifications, it shows there are 30,805 results but in reality, **there are 25 offers per page divided into 42 pages readable** by the user, that means **1050 offers available** in total (the same number of the offers analyzed in this article).

landscape view of the average price of the 5 most common type of flat in relation with the district. The **standard deviation** of the price of a **studio flat** in London is **378,1329** and the **average price** is **£1.233,09**. We can define the coefficient of variation to be able to compare the type of flats and the price.

The coefficient of variation (CV), also known as relative standard deviation (RSD), is a **standardized measure of the dispersion** of a probability distribution or frequency distribution. It is often **expressed as a percentage and is defined as the ratio of the standard deviation to the mean (or its absolute value).**

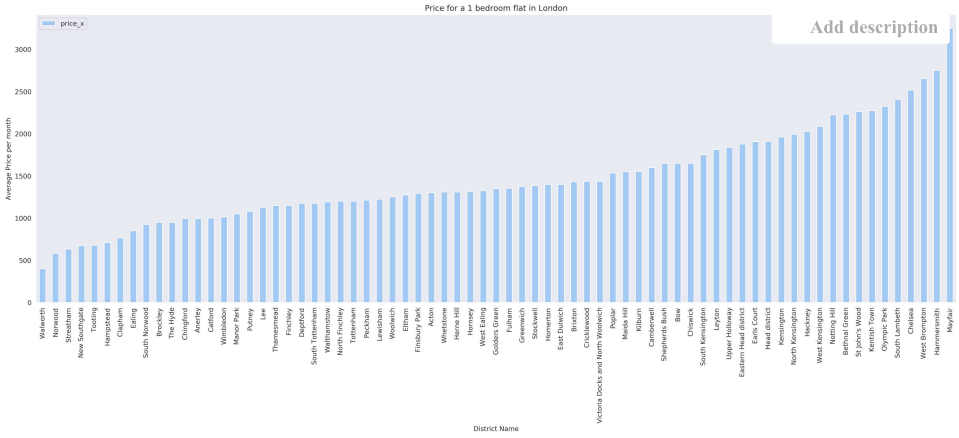
In this case, the **coefficient of variation** is equal to **0,306653**.



The average price for a 1 bedroom flat by district

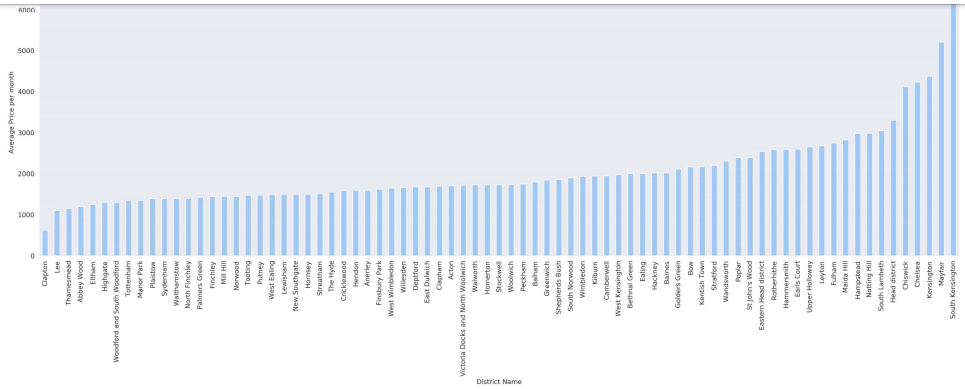
The **standard deviation** of the price of a **1 bedroom flat** in London is **561,9852** and the **average price** is **£1.468,28**.

The **coefficient of variation** is equal to **0,382750**



The average price for a 2 bedroom flat by district

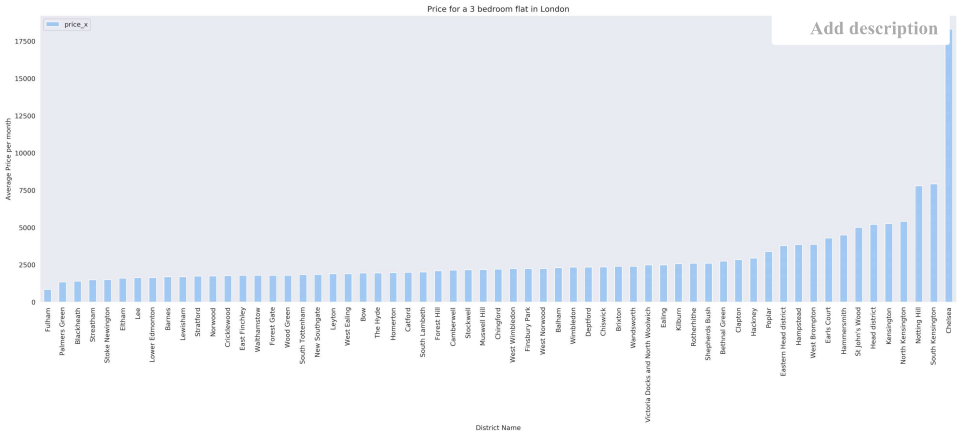
The **standard deviation** of the price of a 2 bedroom flat in London is **949,9799** and the **average price** is **£2.050,92.**



The average price for a 3 bedroom flat by district

The **standard deviation** of the price of a **3 bedroom flat** in London is **2.449,1497** and the **average price** is **£2.892,65**.

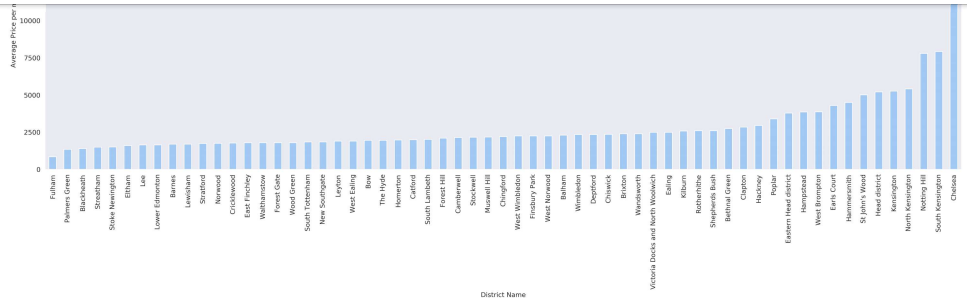
The **coefficient of variation** is equal to **0,846678**.



The average price for a 4 bedroom flat by district

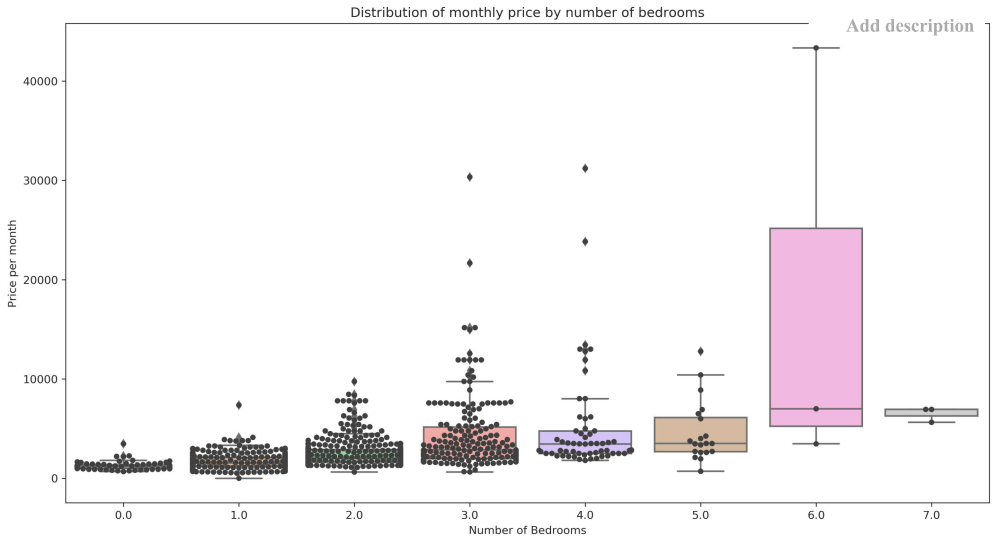
The **standard deviation** of the price of a **4 bedroom flat** in London is **5.176,5294** and the **average price** is **£4.891,59**.

The **coefficient of variation** is equal to **1,058249**.



The tendency of the price by the number of bedrooms

In the charts above, we can notice that **the variation of the price increases more than proportionally to the number of bedrooms**. This means that the impact of the district on the price is weaker on small apartments. In the boxplots relative to the two and three bedrooms flats, we can easily see **the presence of outliers**, typically located in South-Kensington, Chelsea or Mayfair.



There may be several reasons for such market behaviour:

- the price of a studio flat or a one-bedroom flat is too high compared to the average salary and people find hard to afford it.
- the flats are old and poorly maintained
- renters prefer other channels to post this specific category of flat

And moreover.

The Final Input

The goal of this analysis is not only **to inform people about the rentals in London** but also **to provide a way to filter the market based on their needs**.

budget.

After the inputs have been collected, the algorithm analyzes the address and provide the user with a list of offers with a filtered list of offers by venues similarity.

Conclusion

I developed this project to test, once again, my ability to face a business problem implementing a data-based solution. The entire process has been developed using Python and all the documentation, the screenshots and some data in CSV format are available on GitHub.

Any comment and suggestions to implement and improve the analysis are welcome.

Thank you!