London Real Estate - Rent Analysis

Find the flat that better fit your needs

Location based Clustering

30th November 2019

Federico Sciuca

[Contact me on LinkedIn (https://www.linkedin.com/in/federico-sciuca/)](https://www.linkedin.com/in/federico-sciuca/)

Table of Contents

1. Executive Summary 2. Introduction and Business problem presentation 3. Methodology 4. Results 5. Let's start the analysis! 1. Libraries installation and data collection 2. Data Exploration 3. Data Manipulation 5. Foursquare API - Find the most common venues near the flat 6. Clustering using K-Means Algorithm 7. London Clusters Distribution Map 8. Housing market - heating map 6. The Final Input: find your best offer! 7. Conclusion

Executive Summary

This is not the end of the study path but just the start!
As an analyst, I'm interested in finding new insight to understand a certain phenomenon but as a marketer, I also have a client-centric vision of the businesses and one of the powerful ways to make business is implementing solutions that permit the user to make his/her decision fast and easy.

Everything that saves time and efforts to the user has a great value on the market.

Empowering people to make an informed decision is a great way to improve the world we're living in.

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 a simple software that scrape the website RightMove to collect an updated list of flat for rent, analyze each offer using Foursquare and cluster them to divide the housing market in 20 groups with similar neighbours.

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 here (https://github.com/toby-p/rightmove_webscraper.py), that's provide a easy way to scrape RightMove!

This script collect the following data:

- price
- type
- address
- url
- agent_url
- postcode
- number_bedrooms
- search_date

A record will look like the following:

id	price	type	address	url	agent_url	postcode	nu
0	2210.0	2 bedroom terraced house	Lifford Street, SW15	http://www.rightmove.co.uk/property-to-rent/pr (http://www.rightmove.co.uk/property-to-rent/pr)...	http://www.rightmove.co.uk/estate-agents/agent (http://www.rightmove.co.uk/estate-agents/agent)...	SW15	

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**.

In order **to solve this problem**, we are going to use OpenCage Geocoder API (https://opencagedata.com/) 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 (https://en.wikipedia.org/wiki/London_postal_district)) with the data we need for the exercise.

When the data are collected and merged into a single data frame, we are going to **cluster** them using the **K-Means algorithm** to divide the market into *20 different clusters***.** To have an idea of the distribution on the territory of the offers, I plotted **2 meaningful maps**:

- Cluster map: this map shows the distribution of the clusters using colours to identify each cluster.
- Heating map: this map shows the areas with a higher number of offers**.**

To better understand the market, **I** decided to develop some **bar plots** to easily identify the **average price for a studio flat, 1 bedroom flat and 2 bedroom flat grouped by the District Name**.
I've also plotted a **correlation matrix** to identify if there are strong unexpected correlations between features in the dataset**.**

Finally, **I decided to conclude the project by asking the user to insert some 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.

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

Results

As expected, the **price of a flat can't be forecast based on the venues around it only** and there is, of course, a strong correlation between the number of bedrooms and price. *Nevertheless, is possible to develop a prediction model to take in consideration not only the characteristics of the flat but also the District the flat belongs to and the presence of some key venues near 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 to the final users**.

The **goal of this notebook** is to provide to everyone a way to scrape the housing market and **identify the best offers that fit the user's personal needs**. Providing an ideal address and the number of bedrooms the user is looking for, he/she can easily understand which area of London is the best for his/her next step.

Let's Start!

Libraries Installation

First of all we need to install all the libraries we need for the analysis.