Final Capstone Project

How to decide where to rent a flat in Prague

How to spend five minutes to choose an ideal location instead of five days

1. A description of the problem

Introduction

As a student, I have spent 2 years in Prague, Czech Republic. This city is famous for its beauty, but in fact it is one of the most attractive cities for workers and students all around the Europe, because it combines three elements: it's cheap, it's safe, and it's interesting. A met a lot of people from France, Russia, UK, even Brazil who came there to work and were enjoying their life, because it's way more relaxing than in crowded capitals. And all those people have something in common: they needed to rent a flat.

Why it's a problem?

Prague is small. There are not a lot of options: Czech people prefer to buy their own places to live, there are almost no professional landlords who owns the whole building to rent out. Since that, looking for a good apartment is a real challenge you have to invest time and money to.

Of course, professional realtors may help you, and they usually do; but it's really sad for both of you to spend you time to go to look at a 'perfect option' that you eventually decline because you just didn't like the neighborhood? And yes, like most European cities, Prague is really unhomogeneous: sometimes you would refuse an apartment because the surrounding looks abandoned, but in fact there's everything you need round the corner.

You can ask you realtor, but would even a professional read you mind and understand what you want exactly? Or maybe, it's better to spend five minutes for exploring the infrastructure using Data Science, instead of spending five days looking at options which don't worth your time? You answer.

Implementation

First of all, this system can be used by realtors themselves. It saves time, and it's not rude like just sending clients to explore locations on Google Maps. They can show it in the office to reduce the list of options. Also, extra data may be added: for example, criminality rate. It's not so important for Prague, because literally all its locations are safe, but it can be meaningful if we decide to implement this scheme in bigger cities.

In the end, this analysis will save our time and raise the quality of our service: the less time clients lose, the more satisfied they are. And they may be amazed how accurate our offers are. It means they would likely recommend us to their colleagues and acquaintances.

Final Capstone Project

How to decide where to rent a flat in Prague

How to spend five minutes to choose an ideal location instead of five days

2. A description of the data

Which data we need

Although the most popular transport in Prague is tram, foreigners prefer subway: tram lines and routes can puzzle a person who doesn't speak Czech. It means that a foreigner would rather choose an apartment depending on a subway station.

We can easily access this data on English Wikipedia: https://en.wikipedia.org/wiki/List_of_Prague_Metro_stations. Sometimes Wikipedia information can be a little outdated. To check if it is, we can use official website of Czech Transport: https://czech-transport.com/. Its data are full, but not really structured; so it's better to use final data from Wikipedia.

3. Methodology section

During my project, I am going to use the following instruments:

- **Pandas** and **Numpy** library for preparing and processing dataframes.
- **Beautiful Soup** for extracting data from HTML.
- **Json**, **json_normalize and Geopy Nominatim** to work with geospatial data.
- **Folium** for creating maps.
- **Matplotlib** for plotting the results on maps.
- **Sklearn** for KMeans-clustering.
- **Foursquare API** to obtain actual information about infrastructure.

I will go by following steps:

1. To get required data from open reliable sources.
2. To observe what I should do to prepare the data.
3. To transform data in the way it will be convenient to explore.
4. To visualize prepared data to be sure that it has no errors.
5. To explore data using KMeans-clustering in order to get relevant categories of locations.
6. To explore each cluster to find out its main characteristics.
7. To enunciate the results of the exploration.

4. Results section

The data exploration shows that all locations centered on subway stations in Prague can be divided in five clusters, each having its own characteristics.

1. Locations suitable for clients who prefer to rent an apartment just to sleep in it. Far from the center, lack of pubs, bars and coffee places around, no cultural venues or parks. This locations is perfectly fit for solo tenants, who works a lot and prefer spend their free time in the downtown, and don't want to / just can't pay extra for 'good infrastructure' around.
2. Most common locations with a lot of public places. It's suitable for solo tenants or couples who prefer to have everything nearby. At the same time, these locations might be the most crowded and loud, so they're no suitable for old people and for families with little kids.
3. These locations are far from the center and don't really have excellent infrastructure, however there are way more places for entertaining that in cluster 1, so it makes them ideal for student groups or solo tenants who prefer to spend free time somewhere around the own place.
4. These locations combine some characteristics from cluster 2 and cluster 3. In fact, they have everything: entertainments, emergency things like pharmacies or household goods. This makes these locations convenient for young families, even with children, because there are a lot of parks and playgrounds around.
5. These locations seem convenient for old couples and families, who would prefer a quiet, nice neighborhood without loud bars around. The infrastructure there is really good; however, there is lack of places like bars, ect. (Except a casino on Luka, but it seems to be an outlier).

5. Discussion section

Of course, infrastructure is just a one measurement of how good an apartment is. However, just saying 'good infrastructure' is not enough. For one it can be an entertaining place with a lot of pubs, for another it is a quiet suburb with gym and park around the corner. But this system present a quick and handy way to show, *why* the infrastructure is really good or bad, and really can motivate the client to chooser decline the option.

However, it's not perfect. First of all, we need way more data to make a well-considered decision. But we also can add more data in our scheme: for example, how good the schools and hospitals nearby, or, maybe, are there any industrial enterprises nearby? No one would like to live next to the factory.

Second question: are these venues really good? For example, Chinese Restaurant can be a small café with plastic chairs, or can be a fancy place with Michelin-level cousin. Maybe if we are interested, we really need extra information: for example, rating of these venues.

But it we have already saves five days just by doing this simple job.

6. Conclusion

I believe that some agencies are already using such systems, especially in big cities where time means a lot of money. But it should be presented worldwide, because it's really the most time-consuming and exhausting part of looking for an apartment: to physically present there and to have a look.

The real implementation of the system would allow clients and realtors spend their time only on what is important: to observe the condition of the flat and to meet the owner. Agencies would benefit from that, because the quality of their services would be really high, with fewer resources spent. And it's actually the real mission of Data Science: better results with less time expenses.

Thank you for your attention!