Capstone Project - The Battle of the Neighborhoods

Coursera/ Applied Data Science Capstone by IBM

Orestis Papantoniou

Project title:

Avoiding roadworks in Berlin: optimal location for a new coffee shop.

Introduction: Business Problem

1.1. Background of the Problem:

Berlin is one of the most famous destinations for tourism, culture and entertainment worldwide. A modern metropolis that in the last 10 years has been expanding rapidly. Because of its well-known club, music and artistic scene, the German capital is a pole of attraction for large crowds all over the year.

However, apart from being an exciting city, a lot of problems come together as a side-effect. For instance, the housing problem has become an acute social and political matter. Furthermore, the growing population agglomeration has provoked a constant deficiency of the urban infrastructure facilities. Consequently, public and private utility companies for water, wastewater, household waste removal and energy efficiency face an extended demand for construction of new facilities, as they not only need to replace and upgrade aging infrastructure, but also to meet the new requirements. As a result, large parts of the city seem to be under permanent reconstruction. Berlin since many years looks like a huge construction site with persistent problems in traffic and public transport.

Beside the "housing" problem, the "roadworks" problem has become one of the favorite topics of political debate. In 2014 the Senate for City Development decided to impose fixed-term restrictions to new excavation works as one of several measures with which construction should become more bearable for the citizens (§ 12 BerlStrG '*Berliner Straßengesetz*'). This means: newly built road lanes and side lanes may not be dug up again for five years; a period of three years applies to newly created walkways and bike paths. This restriction aimed to compromise the necessity for construction activity and a the same time ensure that traffic, cyclists and pedestrians are disturbed as little as possible. Supply companies need to tune their construction measures at an early stage with each other, but also with the road building authorities.

The coordination of the construction measures should achieve transparency and planning security for everyone involved and affected. The Geoportal of the City-State of Berlin displays the current roadworks prohibitions with approximately monthly updates.

1.2. The Problem.

Supposedly we are assigned with the task of suggesting some optimal spots for a new cafe in the most hippy and central (out of total 12) boroughs of Berlin, let´s say Mitte and Neukölln. By inquiring the optimal location for a new coffee shop, we have to take into consideration not only the standard competitor analysis but also to predict how the situation of the roadworks is going to look like in the first years after the business start. We will be facing not only the problem of a city ***already crowded with coffee places and bars***; We have to particularly spot those locations ***with the lowest probability of facing a construction measure in the first few years***. A coffee place at a road under construction is a less attractive option for the target group, in terms of noise, accessibility and quality/attractiveness of the surroundings. Furthermore, especially for such a small business like a coffee shop, the first 2-3 years are extremely crucial for building clientele and becoming established in their market field. A construction measure at the vicinity directly after the cafe opening could be catastropfical and existence- threatening.

1.3 Interest

The target group is wide enough to define, as coffee shops belong to a service sector on huge demand in Berlin. Specifically we have a certain focus on ***groups facing mobility challenges***, like parents with baby carriages and disabled persons on wheelchairs. This group is not to be underestimated, as Berlin is a proud baby-friendly city and the spectacle of young parents dragging huge baby/kid carriages with their bike is quite regular. Besides, accessibility for differently -abled people and people with reduced mobility is always highly ranked within the social agenda of the authorities.

Data
According to the above described problem, factors that will influence our decision are:

- current roadworks or roadworks that finished recently at various neighborhoods of the city that could come intro consideration
- number of other coffee shops in the neighborhood
- distance of neighborhood from city center

Following datasets will be used:

candidate areas will be generated algorithmically and approximate addresses will be obtained from the Geoportal of the Senate Department for Environment, traffic and climate protection of Berlin.

After excluding vicinities which with certainty will be undergoing construction works in the next 3-5 years, we will explore the competitive situation and the number of other coffee shops in every neighborhood using the RestAPI of Foursquare.

Data acquisition and cleaning

To begin with, I downloaded the .xls file with all current excavation bans in Berlin. Then I load the data with Python Pandas in order to read them as a DataFrame. I ll have a look at the first 5 rows. (source of the dataset: "Geoportal Berlin / Aufgrabeverbote")

Our dataset displays information about the current excavation bans all over the city, ordered by borough, neighborhood, and street. Columns 4 and 5 give information about the part of a street (or nearby streets) that are covered from the ban. The next 2 columns display the area in meters that is covered from the ban. Then we get information about the type of ban, if it affects the road or the pavement or the bike path and finally the exact period of time that the ban refers to.

Methodology

Feature Selection

I started with a statistical summary of the columns including only object-typed attributes.

I could see how many unique values, which is the top value and the frequency of top value in the object-typed columns.

From the above we get informed that the current excavation bans affect 10 different boroughs and 70 different neighborhoods of Berlin. The borough with highest frequency actually is not relevant for us, as it is located quite at the verge of the city. From the dataset we will keep only the data pertaining to the boroughs of Mitte and Neukölln.

I spotted all unique Boroughs for which there is a current roadworks ban: Now for reasons of simplicity in the context of this project, I merged the expiration date columns into one column that indicates the end of excavations restriction for each street, regardless of whether it concerns the road part or the pavement part. In case of streets affected from both road and pavement works, I did the merging of the exp.date of the last towards the first type, as the restrictions for road excavation last longer (5 years) than those on the pavement (3 years). In the end I kept the value that lasts longer. Last step, I dropped the remaining NaN values

Now, since for the purpose of this project the most important factor for the selection of optimal locations was how safe our choice is gonna be in terms of probability of local construction works, I gave a quick overview of which parts of the city look like the safest right now through a choropleth visualization. From the expiration date of each re-constuction ban, I calculated how much time of 'immunity' remains for each street, in days. First, I found how many days remain until the streetworks-ban for all rows expires. I took as time measurement the current date. Then I added it as a new column.

Choropleth map:

I found online by github a GeoJson file with the Boroughs of Berlin as polygon shapes.

The boroughs with the darkest shades of purple have a big concentration of streets freshly excavated, subsequently longterm safe. The boroughs the interest us, Mitte and Neukölln, are somewhere in the middle. The data frame seems lacking information about a few districts, as we see a small part of the map with no color at all (white). Unfortunately the part of the city which has the darkest shade is quite far away from the center and didn’t come into closer consideration.

In the next step I worked only with the 2 boroughs of interest, Mitte and Neukoelln. I extracted their data to 2 seperate datadrames.

Afterthat, I used the Geopy geolocator Nominatim to add the coordinates of each affected street. As the geolocator looks for the coordinates in the column with the streetnames, I had to add for each street the prefix "Neukoelln, Germany" so that I did get the right ones and not form some other namesake street all over Germany.

Nnext step was be to spot all similar coffee shops located at the boroughs of Mitte and Neukoelln, and group them in two categories: first one being situated in a street (partially or fully) affected by the ban law, second one being situated in a restrictions-free street. As "similar coffee shop" I defined a shop that has a declared wheelchair accessibility. For this part of our survey, I used the Fourthsquare API in order to request relevant information for these coffee shops

F**oursquare API calls**

I performed a venue search for the 2 boroughs seperately.

We're interested in venues in 'coffee' category, but only those that are proper coffe shops. We did not want to include restaurants or other places that might serve coffee as well. From the documentation of Foursquare API, are 5 categories IDs relevant to a coffee shop:

venue categories

1) coffee shop

1) 4bf58dd8d48988d1e0931735

2)Turkish Coffeehouse 56aa371be4b08b9a8d5734c1

3) Coffee Roaster 5e18993feee47d000759b256

4) Café 4bf58dd8d48988d16d941735

5) Corporate Coffee Shop 5665c7b9498e7d8a4f2c0f06

After collecting and analyzing the json files, I visualized the concentration of the cafes in the 2 boroughs in relation to their geographical proximity to the streets in question. On the map we can see with the blue pins the streets that are currently under new constructions ban. With the lily circle we can see the cafes, scattered around the streets of borough Neukoelln.

Further Analysis of the collected data: Binnig

Follwing I created a merged dataframe with all streets under construction ban in both Mitte and Neukölln.

Binning the dates into 3 categories

From the above dataframe, I binned the category 'Ban days left' into 3 categories, say: 'ban ends soon', 'ban ends medium-term' and 'ban ends longterm'

Plotting with Histogram

To begin with, l plotted the histogram of ban days left, to see what their frequency distribution looks like. We would like 3 bins of equal size bandwidth so we used numpy's linspace(start_value, end_value, numbers_generated function. I built a bin array, with a minimum value to a maximum value, with bandwidth calculated above. The bins will be values used to determine when one bin ends and another begins.

Visualization of the first results with heat map

With the creation of a heat map I visualized which parts of Mitte and Neukoelln are "hot", meaning that the streetworks-free time is running out soon

We can see that the best parts are not in the city center,but rather at the southest parts of Neukölln.

Conclusion and Discussion

It is obvious from the above that this is just a preliminary analysis of a much complexer project.

A lot more steps are to be done: the relatively still safe part of the two boroughs can be usued as centroid in order to measure their distances to the cafes. The results with the longest distances will be the locations that could be the most optimal, always within the framework of this simplistic project, which had to leave multiple other factors aside.

A more complex and extended version of this project would also take into consideration int he final decision, the proximity to locations that are less probable getting affected of future construction works, like parks or squares.

At the final stage the end candidate locations could be clustered in order to create major zones of interest (containing greatest number of potential locations) and addresses that would be the initial point for further exploration.

The aim of this project was to show that there are a lot of diffirent and unpredictable factors in every problem seeking solution. There are varius and nowadays almost endless data resources, that can be used to extract useful information and dare prodictions that otherwise couldn't have been made.