The Battle of Neighborhoods: Where to eat for cheap in Curitiba

Cesar Henrique Sanfelice

6/4/21

The Coursera Applied Data Science Capstone Project

A. Introduction

Curitiba is one of the most important cities in the southern part of Brazil. As capital of the state of Paraná, the city has more than 1,800,000 habitants and covers an area of 430.9 km². It is the 7th largest Brazilian city and 4th largest in the Southern Cone (the south part of South America). The city has the largest population and the largest economy in Southern Brazil.[5] The urban area of Curitiba is looked after by 26 local governments and has 3,335,588 people living there.

Curitiba is a city in south Brazil, capital of Paraná. It has 75 neighborhoods, divided into 9 administrative regions. The city itself has a population of 1.9 million people. It's continuously built-up urban area, is the eighth largest city in Brazil.

Let’s see how the neighborhoods are distributed:

Curitiba

Divisão administrativa

Curitiba is considered a leading regional city, with a strong market, culture, art, finance and education areas. Its industrial district hosts larges national and international companies, over last decades the district is continuously growing because of the city's important role in government and commercial business.

Therefore, in this project I would like to help tourists or even residents with an analysis of the rated restaurants distribution and finally with a quick guide of the best places and region to eat for cheap in Curitiba.

Some questions we are looking to answer:

- Which areas/neighborhoods have more restaurants in Curitiba?
- What are the best-rated restaurants to eat for cheap in Curitiba?
- Is it possible to define the best location to eat for cheap in Curitiba?

Target audience:

People interested on eating well for less, tourists or residents who looks for a guide of great food for cheap. Business Analyst, who wish to analyze the neighborhoods of Curitiba using python, Jupiter notebook and some machine learning techniques.

B. Data Description

1. Data Requirements and Collection

In this section, I will list and describe the data that will be used to analyze Curitiba neighboorhoods.

Based on criteria listed above the following data will be utilized in our analysis:

- **1st Data:** https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Curitiba
  The list of Curitiba neighborhoods, with area, population and average net income;
- **2nd Data**: https://nominatim.openstreetmap.org/
  Coordinates from each neighborhood;
- **3rd Data**: https://developer.foursquare.com/
  Restaurants lists database, containing Category, ID and coordinates;
- **4rd Data**: https://developer.foursquare.com/
  Restaurants ratings, likes, tips and price;

2. Data Cleaning and Feature Extraction

- The first data is a Wikipedia page about Curitiba neighborhoods. Neighborhoods are spitted into nine boroughs, each one in one table, that contains information of population, households, average income and area. We will scrape the page and create a data frame consisting of seven columns: **Neighborhood, Area, Men, Women, Total, Households, Avg. Income.** With the Area and Total (total population) columns we can calculate **Population Density** and create another column;

- Then, we will be using the **Geocoder Nominatim** python package to retrieve the **Neighborhoods coordinates**. This is our second database and we stored as csv file (CWB Coordinates.csv) and joined into our main curitiba_df database. The result was a database with 76 neighborhoods and 10 columns that looks like the following:

	Neighborhood	Area	Men	Women	Total	Households	Avg. Income	Population Density	Latitude	Longitude
0	Ganchinho	11.20	3667	3658	7325	1921	767.35	654.017857	-25.572076	-49.263667
1	Sitio Cercado	11.12	50631	51779	102410	27914	934.95	9209.532374	-25.542701	-49.269106
2	Umbará	22.47	7280	7315	14595	17064	908.70	649.532710	-25.568169	-49.285699
3	Abranches	4.32	5463	5702	11165	3154	1009.67	2584.490741	-25.361474	-49.272054
4	Atuba	4.27	6156	6476	12632	3627	1211.60	2958.313817	-25.387500	-49.206606
5	Bacacheri	6.98	10762	12344	23106	7107	3029.00	3310.315186	-25.396850	-49.234456
6	Bairro Alto	7.02	20244	21789	42033	12071	1211.60	5987.606838	-25.405822	-49.207660
7	Barreirinha	3.73	8079	8942	17021	5024	1272.18	4563.270777	-25.368564	-49.260455
8	Boa Vista	5.14	13677	15714	29391	9212	1817.40	5718.093385	-25.388479	-49.243713
9	Cachoeira	3.07	3811	3927	7738	2091	908.70	2520.521173	-25.353982	-49.257271
10	Pilarzinho	7.13	13358	14549	27907	7883	1211.60	3914.025245	-25.396348	-49.287557

- The third data will be obtained from **Foursquare**, and we will use **Foursquare API** to access it. We utilize the neighborhoods coordinates to retrieve as many venues as possible around a specific radius. We considered a radius os 1000m from each neighborhood coordinate and that resulted on overlapping of areas between close neighborhoods. Due to this overlap, some venues were duplicate on our database and the solution for that was to calculate the distance from its neighborhood center coordinate and keep only the closest one, which resulted on 2344 restaurants from an initial list of 2993 restaurants;

- The fourth database is a complement to the third, but relies on **Foursquare API** Premium Calls, which are limited to 500 per day. As we have 2344 restaurants to pull ratings, likes, tips and price, we had to split our request in five different days. Considering the goal of our project, only

evaluated restaurants must be in our final database and after dropping rows with NaN in the columns of interest (ratings, likes, tips and price), 1440 restaurants remained for our analysis.