Cutting the Gordian Knot – with Data!

Evidence-based decision-making is frequently held up as a solution to bias, shortsightedness, and the status quo. The value of a data analyst lies in her ability to objectively validate a course of action or to uncover insights that were previously unrecognized or considered counterintuitive.

When issued the challenge for my Coursera capstone project of using geographic data to solve a problem, I looked forward to building my portfolio. The course presented the final assignment as an opportunity to demonstrate my valuable analytical skills to my peers and prospective future employers. Instead, because of the particular problem I selected as the focus for my project, my work ultimately shed more light on the limits and inefficiencies of building your own rigourous data models. As I'll explain in this blog post, my approach to solving the problem was no faster nor more effective than what basic research using existing internet tools would have yielded.

Required to use data from Foursquare to solve a problem, I opted to use it to identify hotel venues in a rural region of Western Canada. I compared the results to the locations of existing networks of mountain bike trails using information from TrailForks. My objective was to find potential venues for a mountain bike race in a region where no major events currently exist. I was looking for trails that were long and varied enough to host a mountain bike race and which had hotels nearby to host participants who traveled to attend the event.

Methodology

I started my search by importing the necessary libraries into Python and establishing a connection to Foursquare. I then asked Foursquare to pull a list of all the hotels within a given 200km radius of Steinbach, a small town of approximately 15,000 people in southeastern Manitoba. Foursquare returned just seven results. This presented the first and most significant shortcoming in my data analysis: I received an incomplete dataset. To illustrate this point, please see figure 1 below, which plots all of the hotels Foursquare returned on a map. Contrast that with the 32 hotel venues that Google found to the south and east of Winnipeg, to say nothing of the dozens of results that Google or a travel website would find within Winnipeg itself (a city of 750,000).

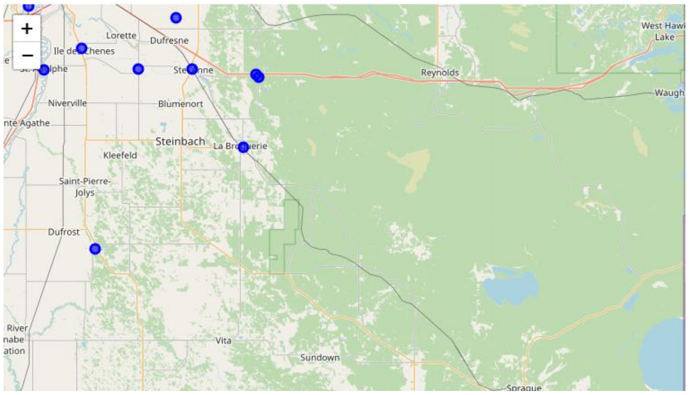

Figure 1: A map generated in Python, showing hotels identified by Foursquare.

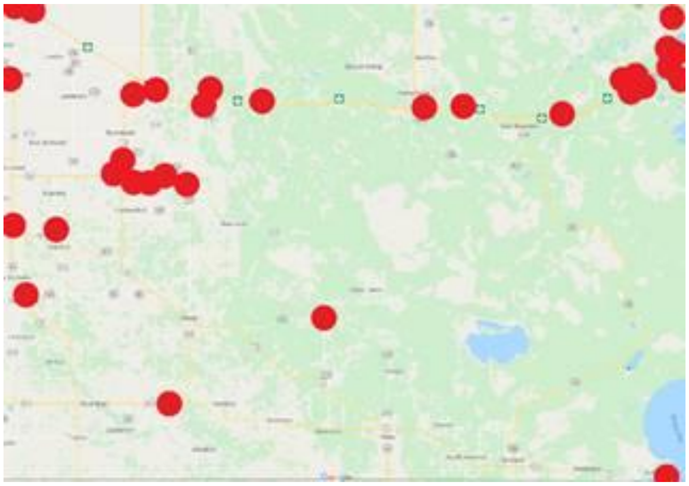

Figure 2: A screenshot from Google Maps, with highlights of hotel locations.

Perhaps it’s because I focused on a rural region. After all, Coursera used two major metropolitan cities to showcase the potential of Foursquare in its course material. I wondered if part of the issue was that Foursquare had become less relevant than other online platforms since the Coursera course was developed (its earliest reviews date back to September 2018 and according to LinkedIn, the course instructor’s biography is at least one year out of date). However, a 2019 New York Magazine article stated Foursquare’s database of 105 million locations rivals maps generated by Google and Facebook. Regardless, in this case the ability to use an API to query Foursquare provided me with information that was worse, not better, than a layperson would yield from a basic internet search.

I then overlaid the coordinates of the hotels Foursquare returned for me with the coordinates of the trail networks on TrailForks to present both simultaneously on a map of the region. Finally, I used colour-coding to separate the trail networks into two categories: those which were longer than 15km, and those which were not.
I should note that the reason I relied on TrailForks to provide me with the locations of trail networks is because Foursquare yielded limited results on this front.

Results

In the end, my map did not reveal any locations that met my specified criteria for venues for a mountain bike race. Although there were four trail networks of 15km or longer, none had hotels nearby according to Foursquare. As a result, my analysis was actually inferior to what a Google Maps search could accomplish because Google Maps found multiple hotel options near some of the longer trail networks.

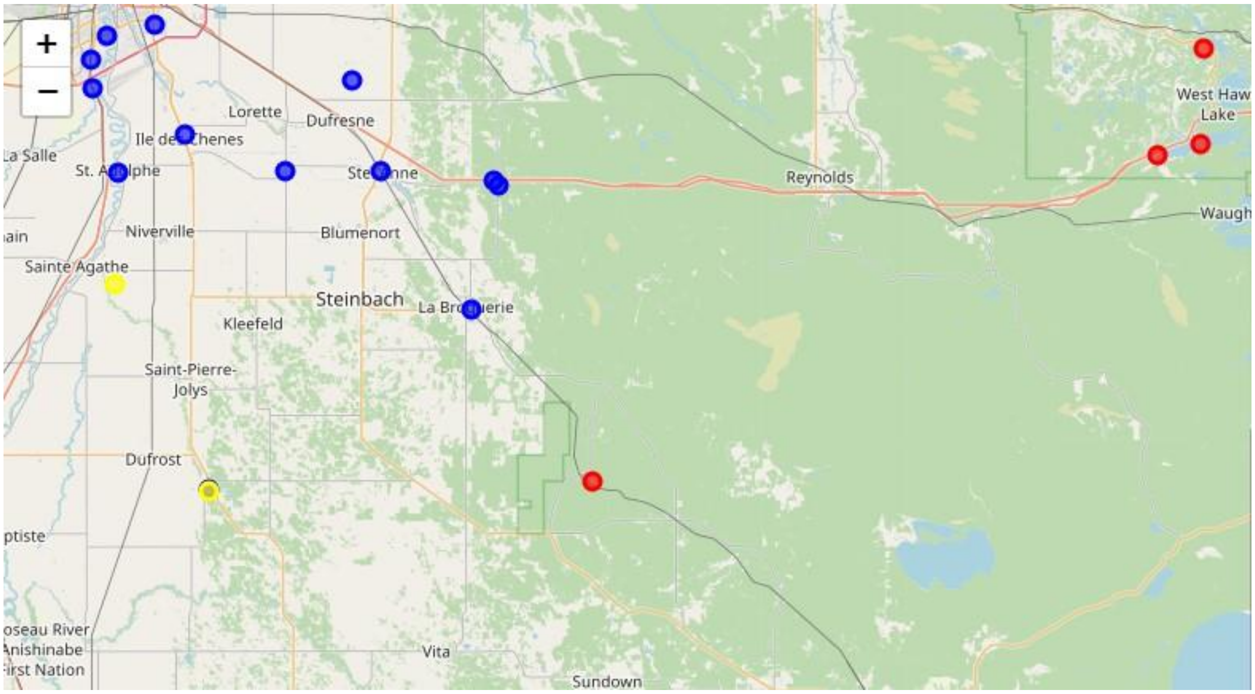

Figure 3: A map of southeastern Manitoba showing mountain bike trail networks of 15km or longer as red dots, and shorter trail networks as yellow dots. The blue dots represent hotel venues.

My use of the Folium library and Python allowed me to generate an interactive map showing all of the venues that I considered. In the end, however, the quality of the insight generated by my analysis didn’t justify the time and effort or showcase the full potential of a data analyst’s skillset.

Although market research was suggested as one possible direction for the capstone project and selecting the location for a new restaurant was provided as an example, my selected problem was too narrow to take full advantage of powerful data analysis tools. There were too few trail and hotel options in the rural area I studied for Foursquare and Python to generate insights that couldn’t have been possible otherwise. Instead, basic website research and consultation with people who mountain bike in the region could lead to recommendations at least as good as the two candidates my research landed on.

One other limitation that my work identified was the potential unreliability of Foursquare. I queried Foursquare for hotel data for the same region on four separate occasions while working on this project. On one of those four occasions, Foursquare found dozens more results, including many within the city of Winnipeg and others within the rural region that was my focus. When I re-ran the query on subsequent attempts, despite not making any changes to my query, I received significantly fewer results. While I’m admittedly a novice with respect to using Foursquare API, I would not rely on it as a data analysis tool if its results cannot be reliably reproduced.

Ultimately, the insight that I uncovered in the data I studied wasn't about mountain bike races. Rather, it was the sometimes-necessary reminder that a good data analyst needs to be able to query data in a structured and rigourous way, but she also needs to be able to select the right set of tools for the job. The preliminary phase of market research for this project – identifying potential venues – could have been accomplished as quickly and more easily by using existing web tools, or consulting people with direct knowledge of the area and mountain biking.

Author's Note: I don't have an actual blog for data science where I could post this, so I have saved my "blog post" as a PDF file for your reference. I did, however, try to use the sort of tone and style that might suit a blog post. I hope that's acceptable.