Coursera Capstone

IBM Applied Data Science Capstone

Where to Move in Indianapolis, Indiana

By: Allison Fernandez

March 2020

A view of a city

Description automatically generated

Introduction

Moving is one of life’s most stressful events. It places a financial, emotional and physical burden on those who experience it. For someone who is relocating to a brand new area, finding a neighborhood that fits his or her lifestyle and personality is especially important. As a result, they fall down a rabbit hole of blogs, reviews and skewed information on the internet to make informed decisions on where to relocate if they have no prior contacts in the area. As a result, I have decided to take a city that I am quite unfamiliar with to undergo this scenario with data-driven decisions.

Business Problem

The objective of this capstone project is to analyze and select the best locations within Indianapolis to recommend to a family who enjoys fitness, being outdoors, exploring cafes, and going out to eat. Through web scraping, clustering and visualizing data, a fictitious family will be able to decide which neighborhood among the top five safest zip codes in Indianapolis to move to. This neighborhood will contain the most venues that appeal to the family’s interests.

Target Audience

This project is useful to new college grads or 20-somethings that are moving for new jobs and new beginnings.

Data

To solve the problem, we will need the following data:

The safest neighborhoods in Indianapolis.

The Zip Codes of the safest neighborhoods in Indianapolis.

The latitudes and longitudes of these zip codes.

Venues that are within these zip codes.

Sources of Data and Methods to Extract Them

I will be scraping crime data from neighborhoodscout.com, a website that lists real estate, demographics, crime, and general overviews for various zip codes and neighborhoods across the United States. In order to web scrape this site, I will be using Python requests and beautifulsoup packages.

After that, I will use the FourSquare API to get venue data for the top five safest neighborhoods. Foursquare is the most trusted, independent location data and technology platform for marketers, developers and explorers in our digital age. It will provide a wealth of information on nearby venues to appeal to the personalities of our fictional family for the smoothest transition possible.

Methodology

I began by installing BeautifulSoup to find the top five safest zip codes/neighborhoods in Indianapolis. BeautifulSoup is a great package that allows you to extract the text from html code that builds web pages. Using neighborhoodscout.com/in/Indianapolis/crime as my URL, I “prettified”, or extracted, the information from this webpage with a BeautifulSoup function. By right-clicking on the table I wanted to extract from this web page, I was able to “inspect” the element and find out its class and attribute. This led to me grabbing the list of Indianapolis’ safest neighborhoods in plain text, which I cleaned up and converted into a pandas dataframe.

Since there were only five neighborhoods to find the zip code for, I quickly looked them up and created a list to append to the existing dataframe. By this point, I had the safest neighborhoods and their corresponding Zipcodes into a table, like so: A screenshot of a cell phone

Description automatically generated

Next came the process of finding each area’s latitude and longitude. Public.opendatasoft.com/explore is an open source web page that hosts this type of information in tables that you can export as csv files. I downloaded the csv to my desktop and then re-uploaded it into my Jupyter notebook to start aggregating the data into one place.

A screenshot of a cell phone

Description automatically generated

After adding Neighborhoods:

A screenshot of a cell phone

Description automatically generated

To proceed, I installed geopy and geopandas to be able to start visualizing and mapping various venues within these zipcodes. Using the FourSquare API, I was able to pull all venues within the five zip codes. Using the “unique” function, I discovered that there were a total of 74 unique categories with 143 total venues in my five areas of interest.

I scanned the 74 unique categories and realized there were only a select few (like gyms, restaurants, coffee shops and parks) that I was interested in examining for this fictional family. I filtered out all extraneous locations to have the final list of relevant venues.

A screenshot of a cell phone

Description automatically generated

Results

Finally, I used Folium to add markers for each venue onto a map to visualize the data. Each marker contained a “popup” listing its corresponding ZipCode. A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

Upon final inspection, I noticed that the largest cluster of relevant venues resided in zip code 46237, or the Arlington Ave/Shelbyville Road neighborhood. From this observation, we can assume that Arlington is one of the safest, most palatable neighborhoods for our fictitious family.

Discussion

From our observations in the results section, we can see that the Arlington neighborhood has the highest number of relevant venues for our fictional family. While there are other safe zip codes that contain similar venues, none of them are as concentrated as zip code 46352. Another observation shows that southeast Indianapolis may have a higher population and more things to do than these other areas, implying that it is not as big of a farming area as the other areas observed. Although it ranks fourth on the list of safest neighborhoods, this is out of more than [113](https://en.wikipedia.org/wiki/List_of_Indianapolis_neighborhoods) neighborhoods within Indianapolis. It’s still a great choice.

Limitations and Suggestions for Future Research

In this project, I examined data based off of five neighborhoods. Further research could color-code each zip code in terms of safety and examine the distance each one is from the airport, downtown Indianapolis, and nearby states (you know, in case of a zombie apocalypse and the need to flee from the state to a safe area.)

Conclusion

In this project, we have identified the socioeconomic and logistical struggles of moving to a brand new area. We web scraped, extracted, cleaned, transformed and visualized data to find the most appropriate neighborhoods for a new family. To answer the initial question – which was which safe neighborhood would be best suited for our family – Arlington neighborhood in zip code 46352 comes out as the clear winner. The findings of this project can help college graduates who need to move to a new city for their first job. I would have liked to use this after I graduated college. When I started with IBM, I was told I needed to move to southern California and report to their Costa Mesa office in two weeks. If I had been able to look at neighborhoods that would be a good fit for me, it would have eased my nerves in starting the next chapter of life.