Optimal Locations to Build Additional Hospitals in New Jersey

IBM Data Science Capstone Project Report

By: Michael Manian

1. Introduction

1.1 Background

Hospitals are an integral part to society. Hospitals, also referred to as medical centers, are health care institution that provide patient treatments with specialized medical, nursing staff and medical equipment. They serve to provide treatment and help cure illnesses all around the world. Over the years, technology and science have discovered over 10,000 diseases that could affect humans and most of those diseases do not have a cure. In order to survive global pandemics, there must be enough medical supplies and hospitals to take care of those affected. I have decided to do my project on this topic as it is an important real-life scenario and has an impact considering all that is going on across the world with the current COVID-19 pandemic.

1.2 Problem

COVID-19 has proved to be one of the worst pandemics the world has faced. The United States especially got hit the hardest with this virus compared to other countries. Back in March, when the virus first hit the US, New Jersey was one of the many states that were affected the most. It got to the point where there were not enough hospital beds for those infected across the tristate. Hospitals were jam packed and nurses and doctors were overwhelmed and running out of resources. Taking into account the number of municipalities, population, income, and number of hospitals, this project aims to show where an addition of new hospitals would be most beneficial in the state of New Jersey. This not only will help in future pandemics, but also help on a regular basis in terms of distributing patients across different hospitals closest to them to reduce hospital wait time etc.

1.3 Interest

Adding new hospitals in New Jersey would be of interest to the Department of Health and Human Services, considering they are the ones in charge of enhancing the health and well-being of all Americans through providing effective health and human services. It would also be of interest to any fresh doctors and nurses looking for a job.

2. Data

2.1 Data sources

I used the following data sources:

- Basic data about New Jersey, including counties and municipalities was scrapped from **Wikipedia**.
- Population data per county as of early 2020 and median income data per county as of 2018 was also scrapped from **new jersey demographics**, and **Indexmundi**, respectively.
- **Python's Geopy library** was used to locate the geographical coordinates of each county. Geopy was not able to find two of the coordinates so I used **Google Maps** to manually get those.
- **Foursquare API** was used to gather all the hospitals located near each county.

2.2 Data usage/cleaning

Data that was extracted from all three websites was scrapped and combined into one table for readability. This includes counties, municipalities, population, and income. The Wikipedia website had lots of excess information on its table such as municipal type, form of government, community established, and year incorporated. I dropped all these columns since they provided no value to solve the problem. The website also contained population data from 2010 and 2017, which I dropped and replaced with the 2020 data scrapped from the new jersey demographics website. The geographical coordinates for each New Jersey county was gathered from the Geopy library and appended to the final table. Two locations were not found so I manually had to append the latitude and longitude through searching on Google Maps. Next, hospital data was gathered from Foursquare. This included hospital latitude and longitude as well as the hospital name. Foursquare ended up gathering hospitals from New York, Pennsylvania, and Delaware since they are relatively close to New Jersey. Since I am only focused on New Jersey, I dropped the hospitals that were from a different state. Once all the New Jersey hospitals were gathered, I was able to get the number of hospitals per county to see which counties have more hospitals and which have less. Out of all this, only the number of counties, population, income, and the number of hospitals per county were used to solve the problem. These selected data columns were then normalized in order to test and determine correlation with each other. The goal of this is to locate counties that did not have many hospitals and correlate that to the population and income in that area, and the overall density of how many towns reside in the county. Finally, this data will be used to inference a solution to the problem stated and, in the end, figure out the ideal locations for building additional hospitals.