IBM Data Science Capstone Project

Opening a New Shopping Center in Nashville, Tennessee

By: Jarrett Smith

Introduction:

For many residents, shopping centers are a great way to relax and enjoy themselves with friends and family. Shopping centers provide customers with a wide variety of options including dining, grocery shopping, movies, and fashion/clothes outlets. So instead of residents traveling to multiple places to get their necessary things, they can just go to a mall where they can accomplish everything on their to do list without traveling from store to store. For retailers, the shopping center location is a crucial aspect to a mall being successful. Successful shopping centers tend to be located in a high population area so more customers can shop there. With the population of Nashville increasing yearly, more and more shopping centers are being built across the entire city. This also affects property developers because building in close proximity to shopping centers increases their value. Of course, many factors should be considered before blindly building a shopping center within a densely populated area. However, the location of the shopping center has historically been a significant influencer of whether it will be a successful business or not.

Business Problem:

The objective of this IBM Data Science capstone project is to analyze a data set and select the most ideal locations in Nashville to open a new shopping center. By using data science and machine learning techniques learned in this specialization, this project aims to provide solutions to answer the business question: Within the city of Nashville, if a property developer is looking to build and open a shopping center, where would you recommend they open it?

The target audience of this project would focus on property developers and investor looking to open or invest in new shopping centers in the city of Nashville, Pennsylvania.

Data:

To solve our established business problem, we need the following data:

1) List of neighborhoods in Nashville. This establishes our boundary for the entire data set.

2) Latitude and longitude coordinates of the neighborhoods within Nashville. This data is required in order to plot the map with the neighborhoods and also get the venue data.

3) Venue data that is related to the shopping center. This data will be used to perform the clustering technique on the neighborhoods.

Sources of the Data & Methods for Extraction:

The Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Nashville,_Tennessee) contains a list of the 90 neighborhoods located in Nashville, Tennessee. The web scraping technique will be utilized to extract the neighborhood information from the Wikipedia webpage. The beautifulsoup packages and requests from Python will help with the extraction process. In order to obtain the latitude and longitude coordinates of the neighborhoods, the Python Geocoder package will be used. Once the spatial coordinates are obtained, we will use the Foursquare API to get the venue data for the Nashville neighborhoods. With this particular dataset, we will be focused on the shopping center/mall category of venue data to solve our business problem. In the next section, I will discuss in more detail the specific methods that will be used in this capstone project, such as machine learning and data visualization.

Methodology:

First, the list of Nashville, Tennessee neighborhoods was collected from the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Nashville,_Tennessee). Web scraping was utilized using Python requests and the beautifulsoup package to extract the neighborhood list. In order to obtain the geographical coordinates (latitude and longitude) of the neighborhoods extracted from Wikipedia, I used the geocoder package. The geocoder allowed us to convert addresses of the neighborhoods into geographical coordinates. After gathering the data, they were transformed into a pandas data frame for easier data manipulation. Then, the data (neighborhoods) was visualized in map view using the folium package. This allowed us to perform a check to make sure the latitude and longitude values obtained from the geocoder were accurately placed in Nashville, Tennessee.

Next, I used the Foursquare API to get the top 100 venues that are within a defined radius of 2000 meters. My credentials (Foursquare ID and Foursquare secret key) were generated from creating a Foursquare Developer Account. I then made API calls to Foursquare passing in the geographical coordinates of the neighborhoods within a Python loop. Foursquare returned the data in a JSON format. Next, I extracted the venue name, venue category, and venue geographical coordinates. With this extracted data, I was able to tell how many venues were returned for each neighborhood. I then analyzed each of the neighborhoods by grouping the rows by neighborhood and taking the mean frequency of venue occurrence. This step is also helpful because this helped us to partially prepare the data for clustering. Lastly, I filtered out every venue with the category of "shopping malls" since we are trying to answer the business question of : 'Where is the best place in Nashville, Tennessee to build a shopping mall'.

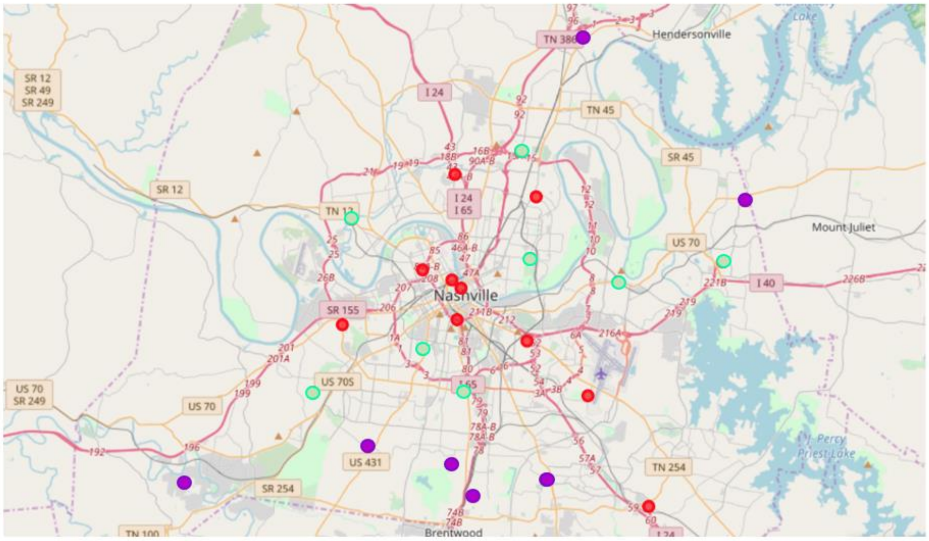
Finally, I performed clustering on the neighborhood data by using the k-means clustering algorithm. K-means clustering identifies the k number of centroids, and then allocates every data point to the nearest cluster. This is done while keeping the centroids as small as possible. I then clustered the neighborhoods into three different clusters based on the frequency of occurrence of shopping malls. The frequency of shopping malls was already calculated in the previous step. The results of the neighborhood clustering allowed us to identify which neighborhoods have the fewest number, or frequency, of shopping malls. This information will then be used to determine the best location to build a shopping mall.

Results:

The results obtained from the k-means clustering showed that we can categorize the Nashville neighborhoods into three distinct clusters based on shopping mall frequency.

- Cluster 0: Neighborhoods with a moderate frequency of shopping malls
- Cluster 1: Neighborhoods with no existence of shopping malls
- Cluster 2: Neighborhoods with a high frequency of shopping malls

The results of the clustering are shown in the image below. Cluster 0 is green, Cluster 1 is purple, Cluster 2 is red.



Discussion:

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the central area of Nashville, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still have very few shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of shopping malls and suffering from intense competition.

Future Work:

In this project, we only consider one factor, the frequency of occurrence of shopping malls. There are other factors such as population and income of the Nashville residents that could influence the location decision of a new shopping mall. However, income data was not available to analyze in this study. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall.

Conclusion:

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.