**IBM Data Science Capstone Project**

***Opening a New Shopping Center in Pittsburgh, Pennsylvania***

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 Pittsburgh 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 Pittsburgh 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 Pittsburgh, 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 Pittsburgh, Pennsylvania.

**Data:**

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

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

2 ) Latitude and longitude coordinates of the neighborhoods within Pittsburgh. 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/List_of_Pittsburgh_neighborhoods>) contains a list of the 90 neighborhoods located in Pittsburgh, Pennsylvania. 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 Pittsburgh 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.