**Selling the Outdoors in Lima, Peru**

**Introduction**

I am a business consultant working for a venture capital firm in New York City. One of the owners has ties to Peru and has procured a contract to open a retail outdoors store in Lima. It is my job to recommend a location.

Choosing a location for a retail store involves factoring in many socio-economic variables. First off, the location must include enough people to buy enough goods to pay for the costs of the operation. Moreover, the nearby population must be large enough to make the investment of capital worthwhile, so the projected revenue must provide a large enough margin to fend off future competition. A second key factor in choosing a location is the presence, or lack thereof, of competition. If so, how many competitors are there, and is there enough local purchasing power to support all of us? Finally, a third key factor is the presence of similar stores. If there is already a group of stores selling products to our target population, then that will tell me the marketing of our operation will be easier and less costly because consumers are already present in the proposed location.

**Data**

To properly address the business problems, I will be leveraging data science techniques, such as clustering, to segment the city into smaller units, using the longitude and latitude of each district in Lima as the starting center points. I will then be mapping those geocoordinates to a foursquare API, which will provide a list of nearby places and the distance to the center of each segmented cluster. Next, I will analyze the group of places nearby each cluster in order to gauge whether similar businesses already exist (and if so, how many) and/or if there is enough of a market to support the new venture.

**Methodology**

I chose to code in the python language because it is the one of which I am most familiar, and it’s capable of running the machine learning algorithms that I need to arrive at an optimal location. To begin, I used the requests library to connect to a Wikipedia page that contained list of districts in Lima. Once connected, I used the beautiful soup library to extract the district information and the pandas library to frame the data. Then I dropped all the irrelevant information, leaving me with district name and zip code. From here, I used the geopy and geocoder libraries to grab the latitude and longitude for each district name, which I then appended to my data frame. Next, I used the folium library to map each district. I then connected to the foursquare API to locate venues within each district. I converted the venue categories into binary integers using the “get dummies” function. This allowed me to group and sort each district by the category types to find which places were most frequent in which location. After this, I used k-means clustering to classify each district based on the portfolio of venues in the area.

**Results**

After running the k-means clustering algorithm on the district-level data for Lima, I was left with five clusters, each which a set of features distinct enough to warrant a unique classification. The first cluster included the two districts that already have “Sporting Goods” venues. It also includes the two districts with the most “Athletics & Sports” venues. This confirms that those two latter districts would be great locations to set up an Outdoors Store.

**Discussion**

Lima is a large city with a lot of districts, but it lacks a deep market for outdoors equipment. There is a lot of opportunity to open a new market. It would be wise to invest in Lima. I suggest raising capital and renting out a store in either the San Borja or Magdalena del Mar districts.