IBM BY COURSERA

Applied Data Science Capstome Project

OPENING A NEW SPORT UNIT IN BERLIN

By Vera Rykalina

January, 2020

Contents

1 Introduction

1.1 Berlin City

Berlin is largest city in Germany with almost 4 million inhabitants. Not only is Berlin the capital of Germany but also it is the second most populous city of the European Union. The city has its special vibrant atmosphere and full of multicultural communities which appeals foreigners from around the world. Due to its high quality of life, Berlin is truly considered to be a scientific hub attracting specialists from different industrial and academic sectors. However, the most valuable characteristics for living in Berlin is its ecological situation. Berlin is a green city. What doest it mean? First of all, there are dozens of parks, forest areas and recreation spots. Besides, people separate rubish, recycle plastic, ride bycicles and are obessed about bioshops in a good way. In addition to it, and what is especially inportant, people in Berlin strive to lead a healthy life style. The inhabitants prefer clean eating and actively do sports. In this report I would like to consider a possibility to open a new sport unit in Berlin preferably in the localities which are lacking sport centers. The bussiness desion should be made based on several parameters, mainly: a location, a population of the district and a number of the neigbouring sport units.

1.2 Business Challenge

The objective of the given capstone project is to analyse and select the best locations in Berlin city to open a new sport unit (gym or fitness studio). Using data science methodology and machine learning algorithm such as clustering, the capstone project aims for providing solutions to answer the fowllowing business question: What are the best locaties for a new sport unit in Berlin?

1.3 Target Audience

The given project can be particularly useful for a potential property developer or an investor who is planning to launch or invest in a new fitness unit in Berlin. The project goal is very challenging as the city is currently having a great number of sport sports but on the other hand, very fast growing dynamic and active population of Berlin is demanding in terms of variety of sport activities and vacant memberships in already existing sport

centers. Besides, there are localities, which have a limited number of the sport schools and centers due to historical or economical reasons but still with increasing demand for such units.

2 Data

2.1 Type of Data

To answer the business question, one should collect the following data:

- A list of the localities in Berlin.
- Demographic characteristics of each locality.
- Latitude and longitude coordinates of the localities. This type of data is required for locality mapping and venue data establishment.
- A venue data preferentially anchored in the data related to a gym or a fitness center for clustering.

2.2 Data Mining

Berlin city consists of 95 localities. A list of the localities of the German capital can be found using a Wikipedia page https://en.wikipedia.org/wiki/Category:Localities_of_Berlin. The Web scraping for the data is performed using two Python packages: Requests and Beautifulsoup. The geographical coordinates of the localities are obtained by Python Geocoder package. Foursquare API is leveraged for the venue data acquisition. The Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API can provide a developer with many categories of the venue data. The focus of this projects is a category linked with fitness centers.

To sum up, the given project should make use of many data science skills from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and geovisualization (Folium).