Introduction to the IBM's applied data science capstone project

This notebook introduces a business problem which is meant to be solved using data science and machine learning and will also walk through the steps on how the data is collected, cleaned and prepared to be fed in the machine learning model.

Step 1: Introduction / Business understanding

I have chosen this data science project as an opportunity to work on a business idea that I formed in my head since college. My dream is to set up a chain of artificial turfs for various sports and activities like football, cricket, fitness workshops etc. This has huge upside potential in India, especially in Kolkata which is my hometown, because:

- The parks and fields available to the public are not well maintained and well suited for a seemless playing experience.
- The parks and playing fields are not always available (to suit the needs of working class) due to lack of infrastructure (proper lighting at night, nets, designated playing areas, level playing fields, etc.)
- Interrputions caused due to religious ceremonies, fairs, political gatherings etc.
- Artificial turfs equipped with sheds will be a bonus as turfs can also be utilized in bad weather (like rainy days or in hot summer days etc.)

If a business model is followed where we offer our target audience a designated playing time with decent infrastructure, people will be more likely to pay fees as people are becoming more and more health conscious and trying hard to be fit and active in their fast paced lives.

One of the most import decisions to make before starting this business is to determine the location where the turf has to be set up to reach maximum potential. In this data science project I have tried to utilize which I have learnt from the 9 courses and designed a machine learning model using the K-means algorithm to find the optitum location to set up the chain of artificial turfs.

Step 2: Business problem

To determine the location or neighbourhood where setting up turfs would be the most beneficial we will use an unsupervised learning algorithm which is k-means clustering. Using foursquare data we will gather data of nearby venues of each neighbourhood or wards. By exploratory data anaylsis we can determine the ratio of occurence of most common venues of each ward.

Then we will feed the dataset into the k-means clustering algorithm to cluster the wards to their likeness to each other. Our goal is to find out which clusters will be beneficial for us to set up artificial turfs.

- First we will discard the cluster where there are the most parks and playgrounds. People living near playgrounds will thick twice paying for artificial turfs when they can play for free.
- Secondly, we will look out for wards which has ease of connectivity. Wards with more number of bus stops, train stations will be beneficial for us.
- We will also look for wards with greater number of sports shops in range.