Data Science Capstone Project

The Battle of Neighbourhoods – Mumbai vs. New York

By: Akash Kadam

August 2020

Introduction

Mumbai and New York, two cities that never sleep. Both are the financial hub of their respective nations. Each of them attracts immigrants (who come to follow their dreams).

Mumbai, formerly known as **Bombay**, capital of **Maharashtra** state, **India**. It is the country’s financial and commercial centre and its principal port on the Arabian Sea. Located on Maharashtra’s coast, Mumbai is India’s most-populous city and it is one of the largest and most densely populated urban areas in the world. Mumbai is home for 1.84 Cr people. The total area of Mumbai is 603.4 km^2. Average temperature of Mumbai is 26.9°C.

New York City, officially **the City of New York**, city and port located at the mouth of Hudson River, north-eastern U.S. It is the largest and most influential American metropolis. New York is the most populous and most international city in the United States of America. New York City is in reality a collection of many neighbourhoods.

City’s population is 1.9 Cr. Area is 790 km^2. Average temperature of New York is 17.1°C.

Business Problem

The Objective of this project is to make an attempt to analyse the neighbourhoods in each of these two cities and try to understand what is popular in them and what they have to offer to someone who is contemplating to make a choice on seeking a life in either of the cities. The decision to choose one over another would depend on popular venues in the neighbourhoods in each of these cities.

Target Audience

People who would be interested in this study are those who would like to create a projection of potential life and activities in these metropolitan city neighbourhoods if the subject moves to live in one of them.

Data

For any "Data Science Project" data is of paramount importance. To solve the problem, we will need the following data:

- List of neighbourhoods in Mumbai and New York.
- Latitude and Longitude coordinates of those neighbourhoods. This is required in order to plot the data on map and also to get the venue data.

- Venue data, particularly the top venues in 500 meter radius of the neighbourhoods. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them:

- **Mumbai**

  This Wikipedia page contains a table named “Mumbai neighbourhood coordinates”. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages.

- **New York**

  We will download the New York data using “wget” command in Python
  “wget -q -O 'newyork_data.json' https://cocl.us/new_york_dataset”. Note that this is a .json file and we need to import json library in Python to handle this file.

- **Venues Data**

  Now, we will use Foursquare API to get the venue data for the neighbourhoods. Foursquare API will provide many categories of the venue data.