COURSERA CAPSTONE PROJECT - IBM DATA SCIENCE SPECIALIZATION

HOUSING DATA ANALYSIS

For the Neighborhoods of Mumbai

By: Devesh Poojari

Email: deveshpoojari7@gmail.com

2020

About the Report

This report is for the final course of the Data Science Specialization. A 9 course series created by IBM and hosted on Coursera Learning platform, it covers the entire process in Data Science from collecting data to making predictive conclusions based on the data.

The problem and analysis approach are left for the learner to decide and explore on his own, with a requirement to leverage the Foursquare location data to explore or compare neighborhoods or cities of our choice or to come up with a problem that we can use the Foursquare location data to solve.

Introduction

Mumbai, formerly Bombay, city, capital of Maharashtra state, southwestern 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. It was built on a site of ancient settlement, and it took its name from the local goddess Mumba—a form of Parvati, the consort of Shiva, one of the principal deities of Hinduism—whose temple once stood in what is now the southeastern section of the city.

In addition, the city's commercial and financial institutions are strong and vigorous, and Mumbai serves as the country's financial hub. It suffers, however, from some of the perennial problems of many large expanding industrial cities: air and water pollution, widespread areas of substandard housing, and overcrowding. The last problem is exacerbated by the physical limits of the city's island location. Area about 239 square miles (619 square km). Pop. (2001) 11,978,450; urban agglom., 16,434,386; (2011) 12,478,447; urban agglom., 18,414,288.

Housing is largely privately owned, though there is some public housing built by the government through publicly funded corporations or by private cooperatives with public funds. Mumbai is extremely crowded, and housing is scarce for anyone who is not wealthy.

(For that reason, commercial and industrial enterprises have found it increasingly difficult to attract mid-level professional, technical, or managerial staff.) In an attempt to stem the ongoing immigration of unskilled labour that has increased the city's indigent and homeless population, city planners have encouraged enterprises to locate across Mumbai Harbour—notably in Navi ("New") Mumbai—and have banned the development and expansion of industrial units inside the city; their efforts, however, have been largely unsuccessful.

Because of the limited physical expanse of the city, the growth in Mumbai's population has been accompanied by an astounding increase in population density. By the early 21st century the city had reached an average of some 77,000 persons per square mile (29,500 per square km). Settlement is especially dense in much of the city's older section; the wealthy areas near Back Bay are less heavily populated.

With the above mentioned problem in the city of Mumbai it becomes increasingly important to analyse the housing data.

Objective:

The objective of this Capstone Project is to build a system with help of data to:

1) Analyze property prices and show the property prices in forms of a heatmap to understand the distribution of property prices across the city

2) Cluster similar Neighbourhoods and analyze the types of Neighbourhoods in the city

3) To check whether the surrounding venues affect the price of a house. If so, what types of venues have the most affect.

Target Audience:

The target audience for this report are:

1) Potential Buyers who can roughly estimate the value of a house based on the surrounding venues and average price.

2) Real Estate makers and planners who can decide what kind of venues to put around their products to maximize selling price.

3) House sellers who can optimize their advertisement.

Data Description

The data for this project has been retrieved and processed through multiple sources, giving careful consideration to the accuracy of the methods used to solve the problem, we will need the following data:

- Data containing Neighborhoods of Mumbai and average housing prices. This defines the scope of this project which is confined to Mumbai.
- Latitude and Longitude data of all the above Neighborhoods. This is required in order to plot the map and also get the venue data.
- Venue data, particularly data related to venues in the vicinity of neighborhoods. We will use this data to perform clustering on the neighborhoods.

Sources of Data and Method to extract them:

1. The required data of neighborhood and their housing prices was collected from 99 acres.com. 99Acres is an Indian real estate database website founded in 2005. At the time of collecting the data there were 906 neighborhoods spread out in 9 distinct sub cities. We will be using Web Scraping techniques to extract the data from the website, with the help of Python requests and BeautifulSoup packages. The collected data is stored as CSV for further use in the project.

2. The Latitude and Longitude Data of the 906 neighborhoods were obtained by normal Google Search with automated web scraping. Here we will use BeautifulSoup and Python requests to parse data from HTML to required format and automate the process of searching on Google and scraping the data with the help of Selenium in python and store the data in CSV format for further use in the project.
3. After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data and help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (BeautifulSoup), automation(Selenium), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).