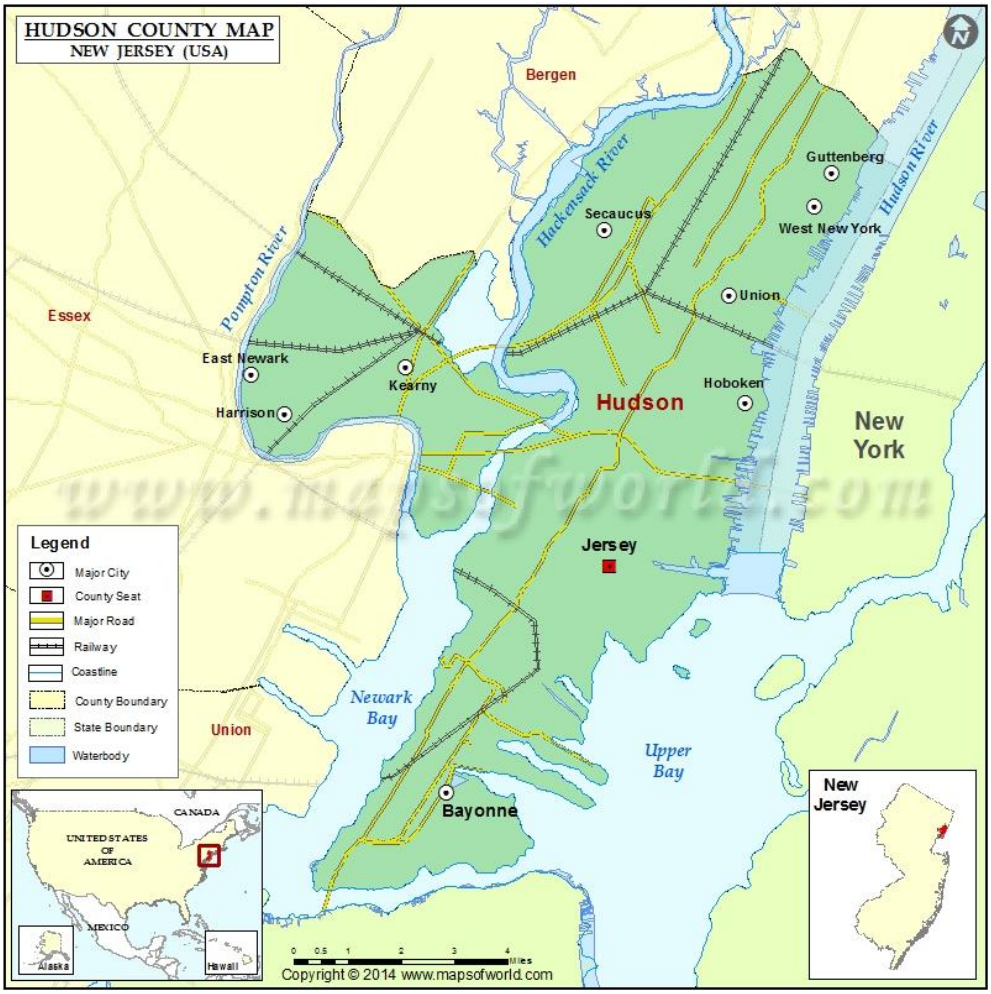
Coursera Capstone Week 4

IBM Applied Data Science Capstone

Recommendation for best location to build new housing complex in Hudson County, New Jersey

Saheli Ghosh

June, 2020



1) Discussion of the background and description of the problem

a) Problem background

Hudson County is one of the fastest growing county in New Jersey according to census data (https://www.nj.com/data/2018/03/the_fastest_growing_counties_this_past_decade.html) and belong to the largest metropolitan area in USA (https://en.wikipedia.org/wiki/Metropolitan_Statistical_Areas_of_New_Jersey).
Many people working in New York City prefer to live in Hudson County due to its proximity and well connected transportation system to the city. After NYC has been hit hard by Covid-19 there has been an exodus of city dwellers to suburbs (https://www.nytimes.com/2020/05/08/realestate/coronavirus-escape-city-to-suburbs.html).

b) Problem description:

For many home construction companies and investors a major problem is to decide in which neighborhood in a certain county homes will be most marketable. For example the proximity of train station or bus terminals might be useful for a certain group of buyers whereas parks, schools, grocery store or restaurant might be attractive to another group. This project will analyze the data for all the municipalities in Hudson County and provide a recommendation for optimal marketable location for construction of a housing complex.

2) Target audience:

Target audiences for this project are home construction companies, real estate investors, real estate agents as well as future home buyers.

3) A description of the data and how it will be used to solve the problem

a) Data:

To solve the problem, we will need the following data:

i) List of municipalities in Hudson County.
ii) Latitude and longitude coordinates of the municipalities which will be used to plot the map and also to get the venue data.
iii) Venue data for each of the municipalities which will be used to perform clustering on the neighborhoods

b) Sources of data and methods to extract them

i) First we will use web scraping techniques using Python libraries requests and Beautiful Soup to extract the name of the municipalities from The Wikipedia page about Hudson County (https://en.wikipedia.org/wiki/Hudson_County,_New_Jersey) which has the name of all 12 municipalities.

ii) Secondly we will use python geocoder library to get the geographical coordinates i.e. latitude and longitude of each of the municipalities of the county.

iii) After that we will use Foursquare API to get the venue data for each of the municipalities. Foursquare is a social location service with a database of about 60M+ point of interest and 941 Venue Categories. We will be comparing the statistics of the

categories like bus stops, train stations, school, parks, grocery stores and restaurants for the municipalities in the Hudson County.

iv) Next we will use K-means clustering algorithm to group similar data points together and draw inferences on which municipalities will be best suited for construction of a new housing complex.

v) Finally we will use folium to visualize the resulting clusters.

4) Conclusion

Thus in this project we will use various technical, statistical and data science skills like web scraping, working with API interface, data cleaning, data wrangling, machine learning algorithm, statistical inferences and visualization technique.