Coursera Capstone Project
Applied Data Science

IBM – Coursera

Data Science Specialization

By

MANDADI.KARTHIK

TABLE OF CONTENT

01-Import Libraries
02- Define Foursquare Credentials
03- Define the city and get its latitude & longitude
04- Search for Hotels & clean dataframe
05- Search for Parks & clean dataframe
06- Search for Restaurants & clean dataframe
07- Search for Cafeteria & clean dataframe
08- Search for Shopping Stores & clean dataframe
09- Generate map to visualize hotels, shopping stores and Cafeteria and how they cluster together.
10- Generate map to visualize Park, Restaurant and Cafeteria and how they cluster together

INTRODUCTION

Hyderabad, India, differentiates itself with **low living costs**. According to our city rankings, this is a good place to live with high ratings in **housing**, **startups** and **safety**.

Hyderabad is the capital and largest city of the Indian state of Telangana and *de jure* capital of Andhra Pradesh.Hyderabad City has a population of about 6.9 million, with about 9.7 million in Hyderabad Metropolitan Region, making it the fourth-most populous city and sixth-most populous urban agglomeration in India. With an output of US$74 billion, Hyderabad is the fifth-largest contributor to India's overall gross domestic product.

This report is for the final course of the Data Science Specialization. A 9- courses series created by IBM, hosted on Coursera platform. The problem and the analysis approach are left for the learner to decide, with a requirement of leveraging the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.

The main goal will be exploring the neighborhoods of HYDERABAD city in order to convey the information regarding surrounding venues,hotels,parks and etc.

DATA DESCRIPTION

The availability of geo data which can be used to visualize the dataset onto a map.

The dataset will be composed from the following two main sources:

1. The dataset will be composed from the following two main sources:

2.https://commons.wikimedia.org/wiki/Category:Suburbs_of_Hyderabad,_India

	name	categories	address	lat	lng	postalCode	state
0	Hotel Imperial	Residential Building (Apartment / Condo)	Station road	17.391178	78.470230	NaN	Telangana
1	Woodbridge Hotel	Hotel	11-4-649/C, AC Guards Lakdi-ka-Pu	17.402915	78.460450	NaN	Telangana
2	Taj Mahal Hotel	Hotel	4-1-999, Abidi Road Chirag Ali Road	17.391942	78.476915	NaN	Telangana
3	The Golkonda Hotel	Hotel Bar	10-1-124 Masab Tank	17.404457	78.453950	500028	Telangana
4	Hotel Savera	Hotel	Chaderghat	17.377101	78.491940	500036	Telangana
5	Hampshire Plaza Hotel	Hotel	6-1-79 & 80, Lakdikapul	17.405397	78.464407	NaN	Telangana
6	Hotel Mandakini Jaya International	Hotel	Hanuman tekdi road	17.385698	78.476934	NaN	Andhra Pradesh
7	Hotel Nayaab	Food Court	Charminar	17.368504	78.477480	NaN	NaN
8	Hotel Rajdhani	Hotel	Siddiamber Bazaar	17.378295	78.474986	NaN	Telangana
9	Alliance Woodbridge Hotel	Hotel	11-4-649/C AC GUARDS LAKDI KA POOL	17.402977	78.459991	NaN	Telangana

Fig:Hotel data

	name	categories	address	lat	lng	postalCode	state
0	KBR Park	Park	KBR National Park	17.425369	78.425063	NaN	Telangana
1	Park Hyatt Hyderabad	Hotel	Road No. 2 Banjara Hills	17.424764	78.429600	500034	TG
2	The Park Hotel	Hotel	22 Raj Bhavan Rd	17.423605	78.462401	500082	Telangana
3	Park Continental Hotel	Hotel	Mehdipatnam	17.399034	78.448108	500057	TG
4	The interior park	Furniture / Home Store	NaN	17.389930	78.470202	NaN	NaN
5	Park view	Residential Building (Apartment / Condo)	NaN	17.466397	78.485148	NaN	NaN
6	The Park View	Smoke Shop	NaN	17.424646	78.423025	NaN	NaN
7	Jalagam Vengalrao Park	Park	Road No 1	17.422526	78.448826	NaN	Telangana
8	Raheja Mindspace IT Park, Building 9	Office	NaN	17.394800	78.470600	NaN	NaN
9	Hotel Green Park	Hotel	Ameerpet Rd, Greenlands, Begumpet	17.436276	78.453892	560016	Telangana
10	Chacha Nehru Park	Park	Masab Tank	17.401969	78.449486	500028	Telangana
11	oxy rich park	Park	NaN	17.400322	78.454988	NaN	NaN
12	OYO The City Park	Hotel	The city park, 5-8-457-489, Chirag AlicLane, A...	17.391635	78.474416	500001	TG

Fig: Parks

The dataset has 50 samples and more than 300 features. The number of features may vary for different runs due to FourSquare API may returns different recommended venues at different points in time. The number of features is much bigger than the number of samples. This will cause problem for the analysis process. Detail and countermeasurement will be discussed further in the next section.

METHODOLOGY

The assumption is that real estate price is dependent on the surrounding venue. Thus, regression techniques will be used to analyze the dataset. The regressors will be the occurrences of venue types. And the dependent variable will be standardized average prices. At the end, a regression model will be obtained. Along with a coefficients list which describes how each venue type may be related to the increase or decrease of a neighborhood's real estate average price around the mean. Python data science tools will be used to help analyze the data.

First insight using visualization:

In order to have a first insight of HYDERABAD city real estate average price between neighborhoods, there is no better way than visualization. The medium chosen is Choropleth map, which uses differences in shading or coloring to indicate a property's values or quantity within predefined areas. It is ideal for showing how differently real estate priced between neighborhoods across the New York city map. The map (Figure shows high price in neighborhoods that located around HYDERABAD AND SECUNDRABAD. It's where most businesses, tourist attractions and entertainments located. So, the venue types that can attract many people are expected to have the most positive coefficients in the regression model.

Linear Regression:

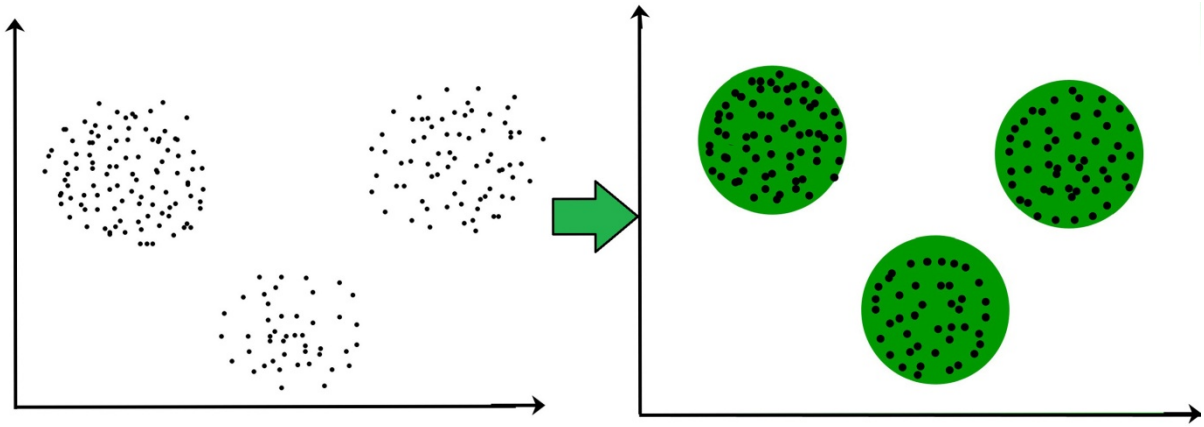
Linear Regression was chosen because it is a simple technique. And by using Sklearn library, implementing the model is quick and easy. Which is perfect to start the analyzing process. The model will contain a list of coefficients corresponding to venue types. R2 score (or Coefficient of determination) and Mean Squared Error (MSE) will be used to see how well the model fit the data.

CLUSTERING IN MACHINE LEARNING

Introduction to Clustering
It is basically a type of *unsupervised learning method* . An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

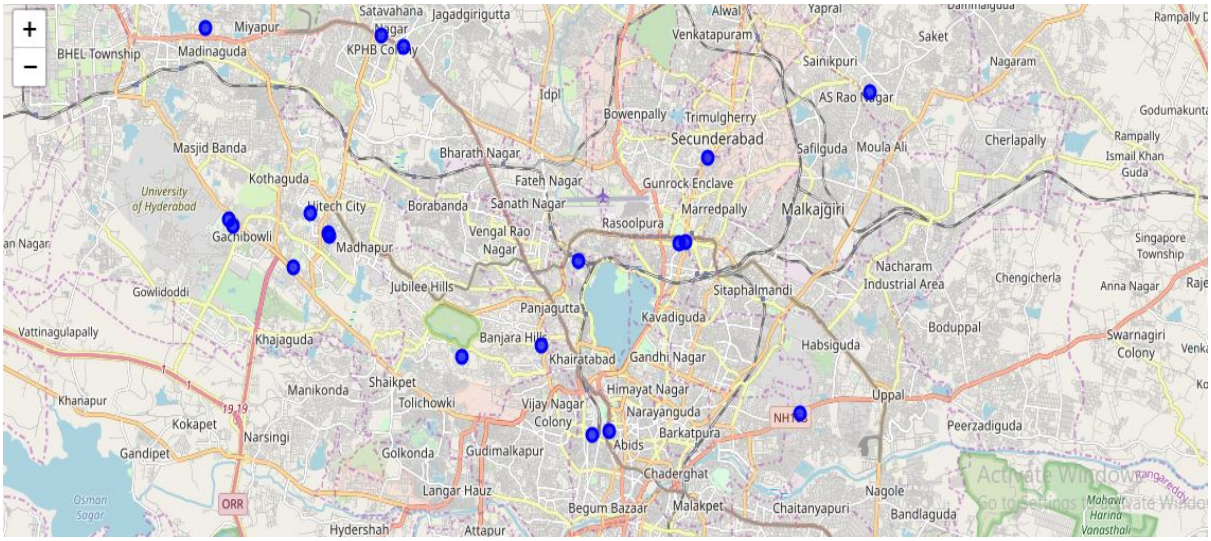
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

For ex– The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.

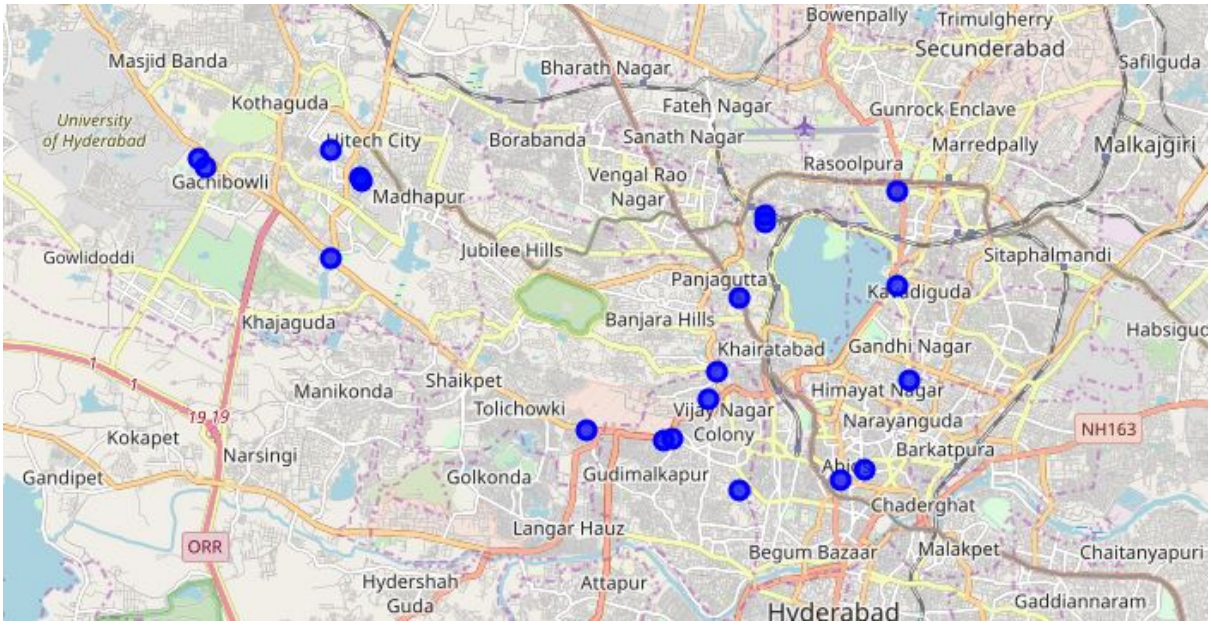


RESULTS

The company has to put a good program, including a hotel of residence, a hall for meetings, places of landscape to visit, stores for shopping, restaurants and cafes. So the company's purpose is to make a list of places of landscape in Hyderabad, including the nearest restaurants, cafes, and shopping stores for each place.



map to visualize hotel neighbourhood including shopping stores and Cafeteria



map to visualize park neighbourhood including Restaurant and Cafeteria

DISCUSSION:

The real challenge is constructing the dataset: -

1.Usually the needed data isn't publicly available.
2.When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.
3.For data obtained through API calls, different results are returned with different set of parameters and different point of time. Multiple trial and error runs are required to get the optimal result.
4. Even after the dataset has been constructed, lots of research and analysis are required to decide if the data should be kept as is or be transform by normalization or standardization.

It can be considered the most important process in the whole data science pipeline. Which can affect the most on the result.

CONCLUSION:

It's unfortunately that the analysis couldn't produce a precise model or showing any strong coefficient correlation for any venue type. But we can still get some meaningful and logical insights from the result.

Doing this project helps practicing every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into.

Some notes on the analysis result:

1.This project is done by a web developer who only started selfstudying Data Science for 4 months. So please take it with a grain of salt.

2.The coefficients only show correlation, not causation. So, if your neighborhood average price is low, please don’t go destroying the surrounding bars and food trucks. There might be another reason.

Toward the person that went through this project, many thanks for the time and patient.