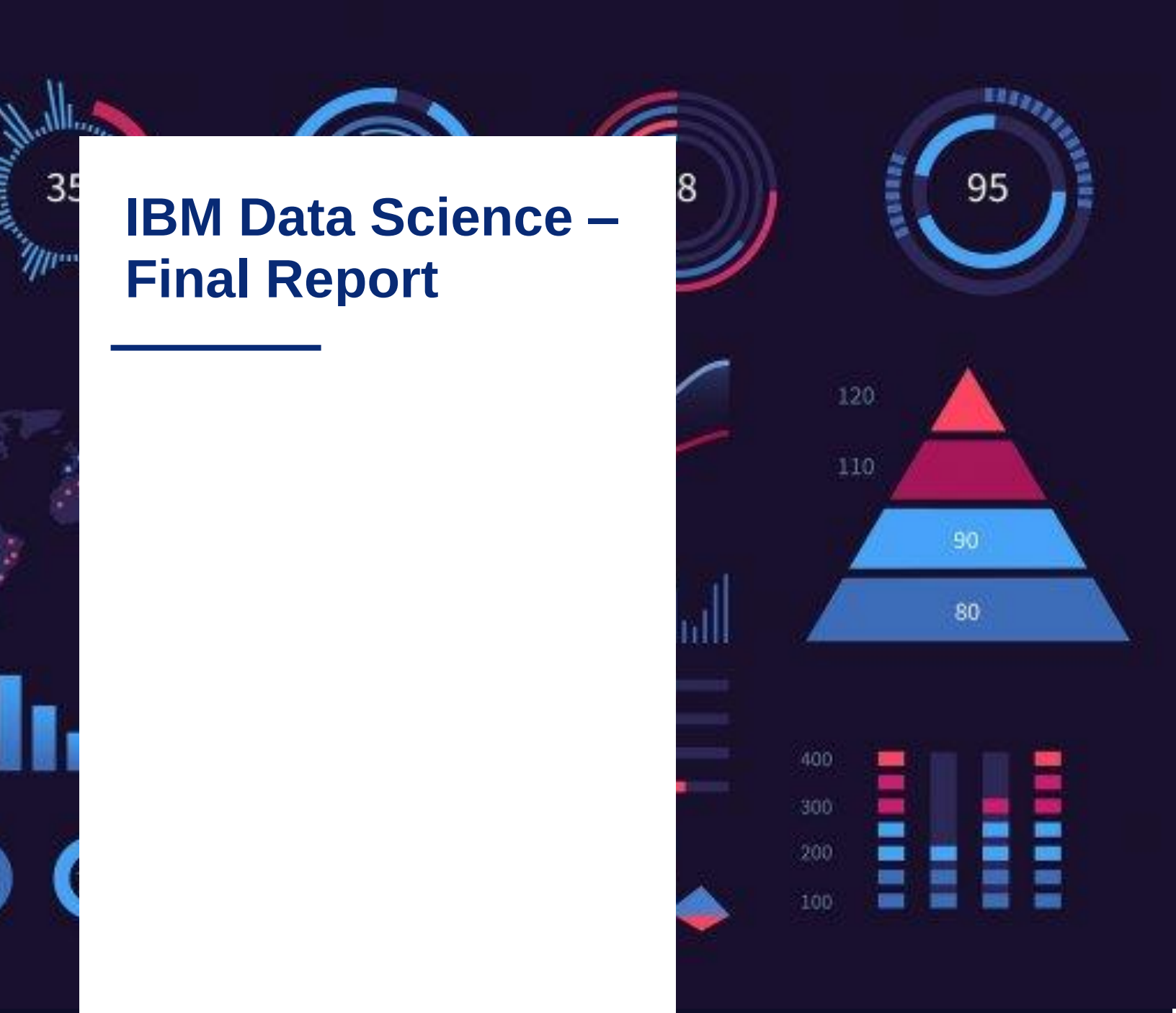

IBM Data Science – Final Report

APRIL 26

Authored by: Eswar Sunder

The Battle of the Neighborhoods

This report is a part of the final IBM capstone project.

Problem statement:

A Multinational Company X wants to setup its first branch in India. The choice of the place where the company is going to setup will depend on the venues in that area. In order to identify the ideal place, the company is interested to do analysis on the Central Business Districts of India.

The Company has many businesses and wants to keep the identity confidential. The analyst doesn't know for which business the company are picking the place to remove any bias from the analysis. So, the company profile or type of business or any other information is unknown.

Central Business districts are the commercial and business center of the city. In other words, this is the "happening" places in the city. In India, there are around 12 top tier cities and each of them have their own one or many districts.

The outline of the objectives:

- Identify the CBDs in India
- Collect the venue data of the CBDs.
- Perform a clustering analysis among the CBD to identify the similar groups
- Profile the clusters based on the analysis

We cannot omit any place as we don't know the nature of the company/business.

Data Sources and Description:

In order to identify the CBDs, we can scrape the Wikipedia website https://en.wikipedia.org/wiki/List_of_central_business_districts. A snippet of the table that should be scrapped is shown below.

Ahmedabad and Gandhinagar	India	Gujarat International Finance Tec-City
Bangalore	India	MG Road Shivajinagar, Bangalore Central Business District, Electronic City,Whitefield
Chandigarh	India	Sector 17
Chennai	India	Anna Salai, T Nagar, Parry's Corner, Nungambakkam
Coimbatore	India	Avinashi Road, Gandhipuram, RS Puram
Hyderabad	India	Nampally, HITEC City, Nanakramguda, Manikonda, Gachibowli, Koti and Himayatnagar

After scraping and cleaning the dataset we will have the CBDs across the 12 cities of India. Using Foursquare API, we can get the information of the specific Area.

Now we have a data with venues and the dataset is rich for any clustering exercise.

There are going to be 2 levels of data. One is a city level data, and another is CBD level data. Any information that is common for the city for example crime rate, etc. will not be used in the clustering exercise so that the final cluster won't have inherent bias based on the city information. If we use this information, then CBDs in the same city is more likely to be clubbed together.

In the end the company will be given the clusters and it can pick choose any place based on its business needs. To make the decision making easier, company will also be given a profile of cities as well.