# Applied Data Science Specialisation - IBM Course

# Capstone Project Documentation

**CLUSTERING SMART CITIES AROUND THE WORLD BASED ON THEIR VENUES**

Bhaskar Rao

# Contents

# 1. Introduction

A smart city integrates technology, knowledge, infrastructure and governance to become more responsive to its residence and thereby improving their quality of life. They do so by improving city services and develop strategies for safety, transport, communication and health. As more than 50% of all the people in the world live in an urban environment, it becomes more and more important for city planners, government official and stakeholders to apply modern solutions and development strategies, often utilizing technology, to become a 'smart city'. To facilitate and tract their smart city initiatives, city officials generally refers to the benchmarks provided by independent bodies. One such organisation is Eden Strategy Institute who rate top 50 smart cities based on vision, leadership, policies, initiatives etc. City leadership should also have information about how similar is their city to one of the smart city. This can be done by either looking at transport spend, population, health facilities or education facilities. Apart from these macro measures, a city's characteristic can also define by it's social life. Social life can be gauged by looking at the popular social hotspots around the city, which can be determined by the popular places/venues around it. This can be a new approach of finding similarities between cities and grouping them together based on their social dynamics, which will give the leadership a new perspective to compare their city.

# 2. Problem Statement

Top 50 smart cities around the world are selected and analysed. A novel way to cluster them together, based on their social signature (social venues), is explored to generate insights. City officials can use the clusters generated to compare how similar or different is their city to a smart city in terms of social life. This will help them evaluate and plan strategies more suitable for their residence, keeping in mind their social signature.

# 3. Data acquisition

## 3.1 Data Sources

The data about the smart cities is present on the smartcitygovt.com website. The data is for 2018/2019. This contains ranking of 50 global cites around 10 themed operation:

1. Funding Smart City Initiatives
2. Developing a Smart City Strategy
3. Smart Clusters & Innovation Districts
4. Digital Inclusion in Smart Cities
5. The Promise of Open Data
6. Co-creating the smart city
7. Smart City Leadership Models
8. Sharing Knowledge Across Cities
9. Preparing a Smart Workforce
10. Beyond Affordability and Efficiency

This list of 50 top global smart cities is extracted by web-scrapping https://www.smartcitygovt.com/.

The latitude and longitude of these city were generated by using the ***geopy*** python library. These city latitude and longitudes are queried against the foursquare data to get the most popular venues around their city centre.

Foursquare database provided top 10 popular venue around city centre (using ***geopy's*** latitude and longitude) are collected for each of these 50 cities. All these are combined to create the analytical data set, a snapshot of which is provided below

| City | Total Score | Vision | Leadership | Budget | Financial Incentives | Support Programmes | Talent-Readiness | People Centricity | Innovation Ecosystem | Smart Policies | Track Record | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| London | 33.5 | 3.1 | 4.0 | 3.0 | 4.0 | 3.0 | 3.1 | 3.0 | 4.1 | 3.1 | 3.1 | 51.507322 | -0.127647 |
| Singapore | 32.3 | 3.0 | 4.0 | 3.0 | 4.1 | 3.0 | 3.1 | 2.0 | 3.1 | 4.0 | 3.0 | 1.357107 | 103.819499 |
| Seoul | 31.4 | 3.1 | 3.0 | 3.0 | 2.2 | 3.0 | 3.0 | 4.1 | 3.0 | 3.0 | 4.0 | 37.566679 | 126.978291 |
| New York | 31.3 | 3.0 | 3.0 | 3.0 | 3.1 | 3.0 | 3.1 | 3.0 | 4.0 | 2.0 | 4.1 | 40.712728 | -74.006015 |
| Helsinki | 31.2 | 3.0 | 2.0 | 4.0 | 3.1 | 3.0 | 4.0 | 3.0 | 3.1 | 2.0 | 4.0 | 60.167410 | 24.942577 |

# 4. Methodology

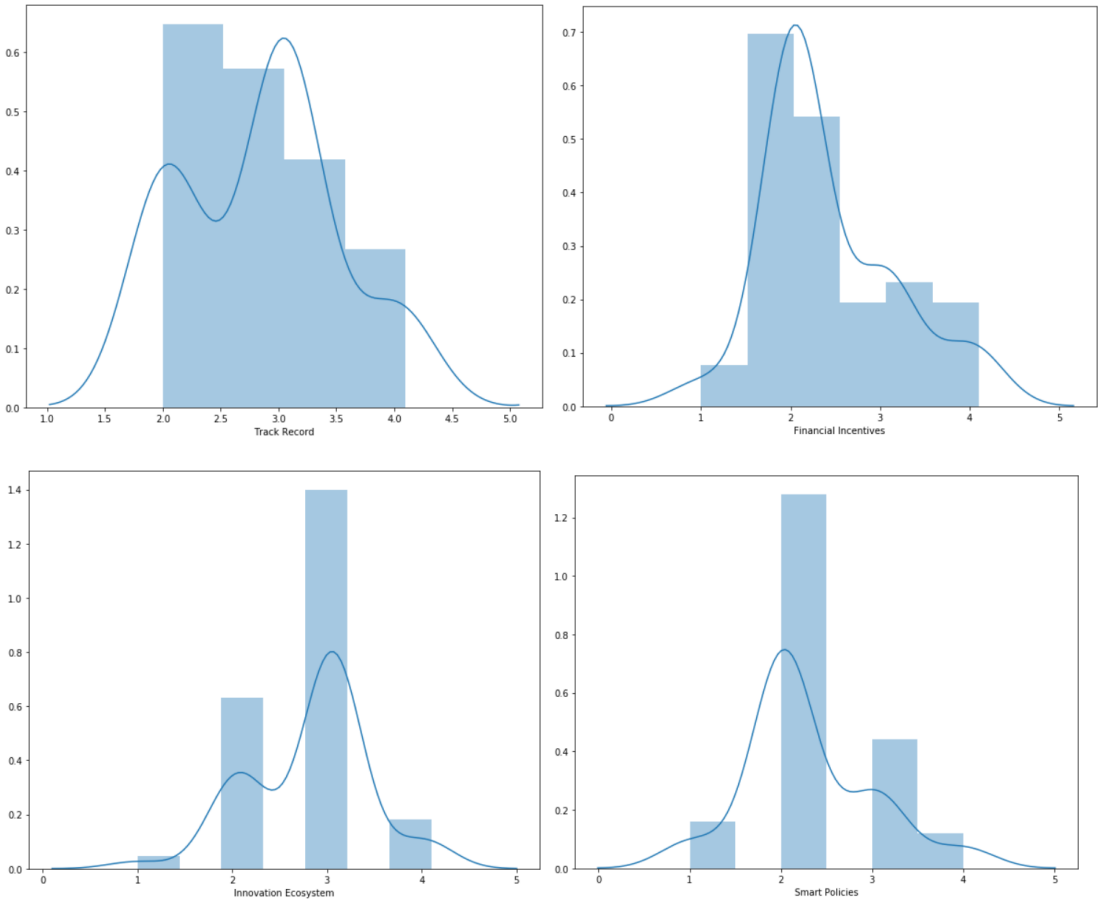
## 4.1 Exploratory Data Analysis

### 4.1.1 Geographical Analysis of cities

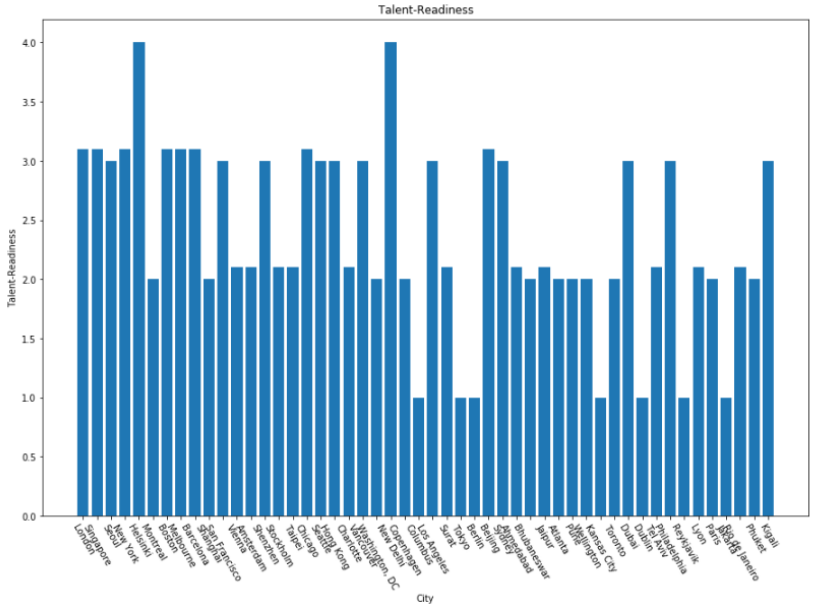
Since this exercise is about unsupervised clustering, there will not be a target variable. To see the geographic location of these cities, they are plotted on a world map. We can see that most of the cities are spread around Europe, North America, South East Asia. It is also very interesting to note that most of the cities at present near the ocean.


### 4.1.2 Analysing the distribution of scores across different parameters

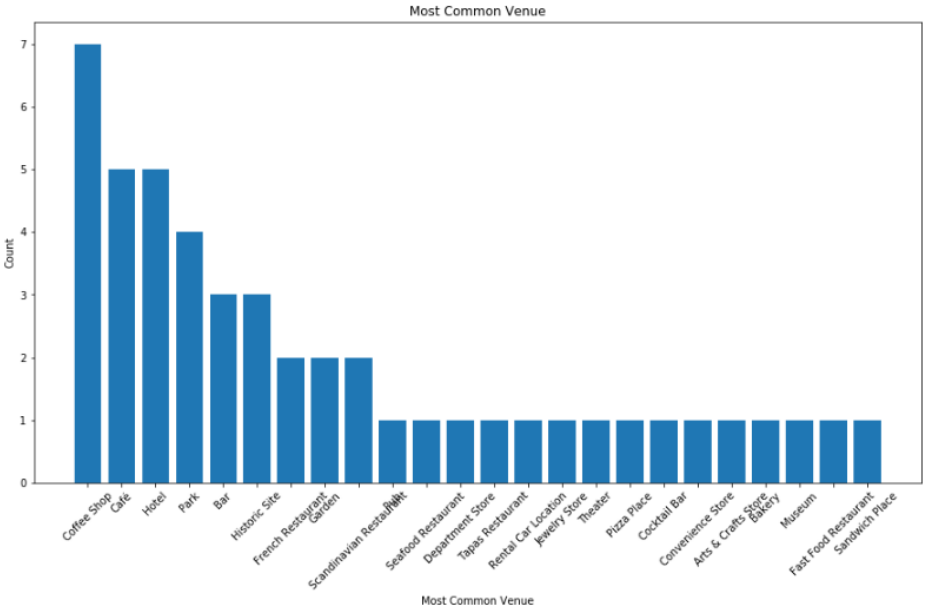
Each of the parameters which have been used to score the cities have been analysed by looking at their distribution using histograms. This tells us if there is a parameter which is skewed or if they follow a normal distribution. The parameters which showed interesting distributions are plotted below




Most of these cities were scored below average for track record and financial incentives whereas for innovative ecosystem, they received above average score. They score just about average for smart policies.  Also, talent readiness has a lot of variability around scores for different cities, which can be seen in the bar plot below.




### 4.1.3 Most popular venue type across all 50 cities




Coffee shops, cafés and hotels are the topmost popular spots around these smart cities.

## 4.2 Clustering smart cities

### 4.2.1 Data set creation

Since the aim of this project was to cluster these cities based on their popular venues, one-hot-encoding is performed on these venues for all 50 cities. A snapshot of which is provided below.

| Zoo | African Restaurant | Alsatian Restaurant | American Restaurant | Antique Shop | Arcade | Arepa Restaurant | Argentinian Restaurant | Art Gallery | Art Museum | Arts & Crafts Store | Asian Restaurant | Astrologer | Au Re |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |

### 4.2.1 Clustering algorithm

K-means clustering algorithm is selected for this unsupervised learning project, as the aim is to discover relations/insights from these cities (in terms of venues). From towardsdatascience.com article,

"**Kmeans** algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way kmeans algorithm works is as follows:

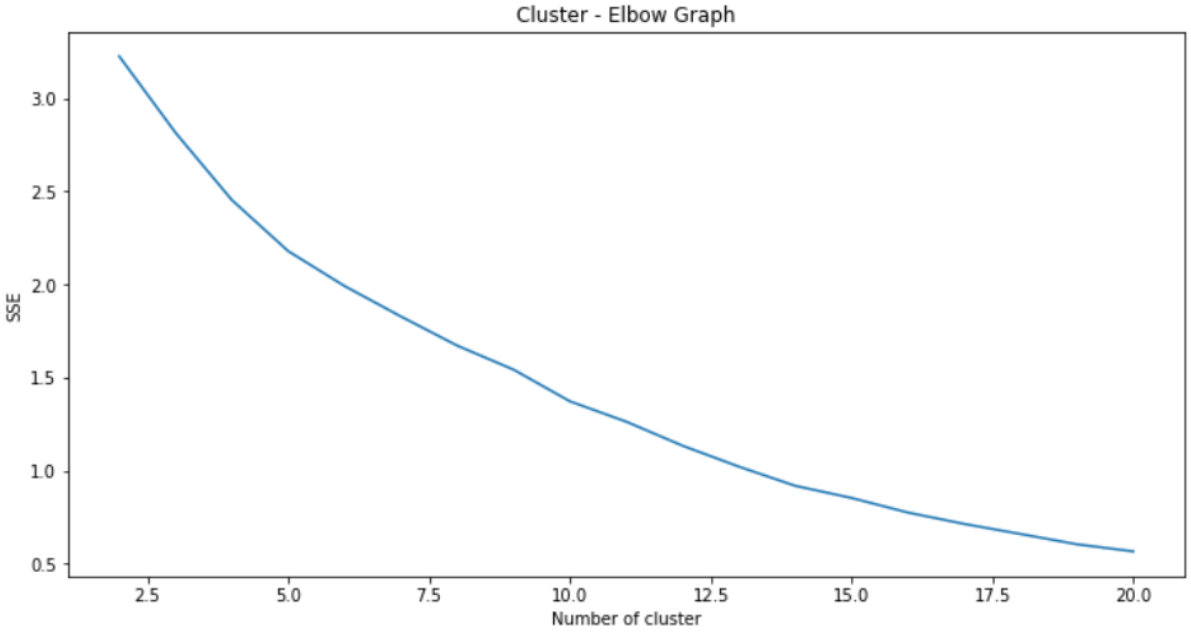
1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.

- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster. "

# 5. Results

K-means clustering is done for the 50 smart cities, using their venues as features. An elbow plot is generated to determine the optimal number of clusters, based on their *SSE*. The plot has been printed below and using which, cities are clustered into 7 clusters.




The 50 cities were grouped into 7 clusters, they are provided below, based on the similarities of their popular venues

**Clusters**

| 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Copenhagen | Ahmedabad | Pune | Kigali | Barcelona | Atlanta | Bhubaneswar |
| Dubai | Amsterdam | Shenzhen | Montreal | Dublin | Beijing | New Delhi |
| Jakarta | Berlin | Sydney | | Los Angeles | Rio de Janeiro | Shanghai |
| Melbourne | Boston | | | Lyon | Singapore | |
| Stockholm | Charlotte | | | Seoul | | |
| Tel Aviv | Chicago | | | Taipei | | |
| Wellington | Columbus | | | Tokyo | | |
| | Helsinki | | | Toronto | | |
| | Hong Kong | | | Vancouver | | |
| | Jaipur | | | Vienna | | |
| | Kansas City | | | | | |
| | London | | | | | |
| | New York | | | | | |
| | Paris | | | | | |
| | Philadelphia | | | | | |
| | Reykjavik | | | | | |
| | San Francisco | | | | | |
| | Seattle | | | | | |
| | Washington, DC | | | | | |


*Plotting the clusters on the world map*

## 6. Discussion

From the clustering results, it can be seen that majority of cities fall under cluster 1. The top 2 most popular venues around these cities are coffee shops and hotels. For the smallest cluster, i.e 3, most popular venues are coffee shops and convenience stores. A city's leadership can analyse their popular venues and can compare their city with one of the clusters. This will inform them how similar or different is their social signature is with one of these clusters. This comparison can be used to take inspirations about how smart cities similar to their cities are tackling problems and take learnings from them to develop their own cities.

## 7. Conclusion

Smart cities were clustered based on their 'social signature', which is defined by their popular venues. Tools and techniques like K-means clustering, foursquare API, geopy and leaflet are used for this project. A new way of using the social factor of a city to compare and contrast with other cities is explored and implemented.