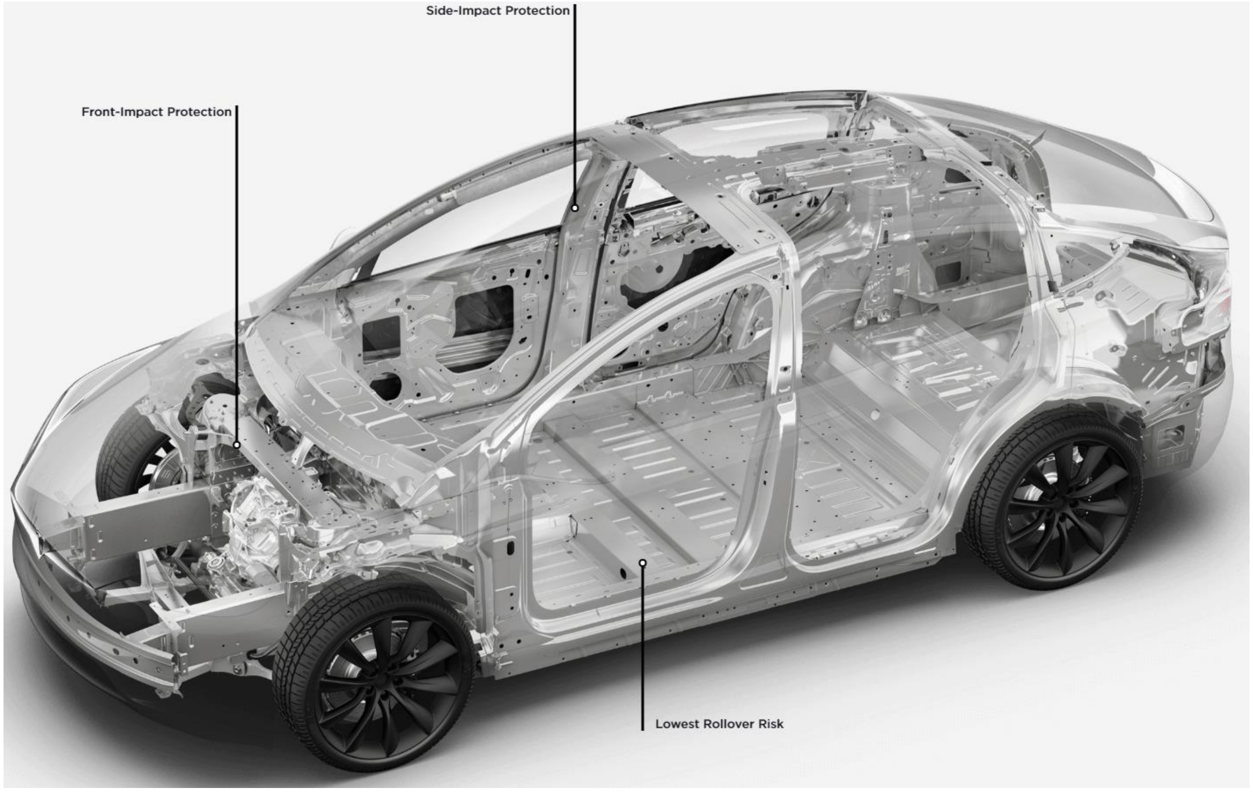
Visualizing the Future Prospect of Electric Vehicle: A K-means Clustering-based Spatial Analysis of Electric Vehicle Charging Stations in Boston, MA

Capstone Final Project—the Battle of Neighborhoods

Wang Xi

Contents


Side-Impact Protection
Front-Impact Protection
Lowest Rollover Risk

1. Introduction

Electric vehicles (EVs) are both economic and ecological vehicles which get their power from rechargeable batteries inside the car. Since they have a lot of advantages as producing nearly no carbon emissions or pollution, being cost effective and less noisy; the main disadvantage of these vehicles are recharge related problems.

One approach to deal with this problem is to construct electric vehicle charging stations (EVCS). A proper EVCS also should be located very carefully to maximize EV usage. Thus in this project, a K-means clustering-based spatial analysis is applied to demonstrate the surrounding built-environment clustering situation of each EVCS site, and explore EV drivers' possible consumption propensity to the surrounding build-environment when charging their vehicles.

2. Data Acquisition

In this part, let me go through a top-down introduction of my database. My dataset comprises of 11 features that are the following in Table 1:

Table 1: Dataset Explanation

Data Name	**Name Interpretation**
elec_car_station_data	the geolocation of the electric car charging stations json data (the whole United States)
evcs	electric vehicle charging station pandas dataframe (the whole United States)
evcs_boston	the electric vehicle charging station location in Boston, MA.
evcs_boston_map	the visualization map of the electric vehicle charging station location in Boston, MA.
evcs_venues	the nearby venues characteristics of EVCS in Boston, MA.
evcs_onehot	the dummies dataframe of EVCS in Boston, MA.
evcs_grouped	EVCS grouped by neighborhood and taken the mean of the frequency of occurrence of each category in Boston, MA.
num_top_venues	the number of top visiting venues near EVCS in Boston, MA
evcs_venues_sorted	EVCS nearby venues sorted by from the 1st most common venue to the 10th most common venue in Boston, MA.

evcs_merged	EVCS dataframe merged with EVCS nearby venues sorted by from the 1st most common venue to the 10th most common venue in Boston, MA.
map_clusters	the visualization map of the electric vehicle charging station clusters by using K-means clustering methodology in Boston, MA.

Based on definition of our problem, factors that will influence our decision are: the number of existing leisure facilities in the neighborhood (categorized by venue type), and the clustering situation of the existing places in the neighborhood.

Following data sources will be needed to extract/generate the required information:

The data was obtained from Alternative Fuels Data Center, the office of Energy Efficiency & Renewable Energy, U.S. Department of Energy official website. On the other hand, the number of restaurants and their type and location in every neighborhood will be obtained using Foursquare API.

3. Methodology

In this project I will direct my efforts on detecting areas of Boston that have high electric vehicle charging station density, particularly defining the category of the surrounding leisure facilities. I will limit my analysis to area radius ~500m and limit 100 venues around each electric vehicle charging station.

In first step I will visualize the electric vehicle charging station geolocation (Latitude and Longitude) focusing on City of Boston in MA, by creating folium map.

Second step I will focus my attention on collecting the required data: location and type (category) of every leisure place within 500m (10-min walking distance) from each EVCS (according to Foursquare categorization). I have also limited 100 venues around each EVCS. Therefore, in my analysis it will not only be calculation and exploration of leisure facility density across different areas of Boston by using folium map package, but also I will add pop-up text that would get displayed when you hover over each marker, which can display the name of each electric vehicle charging station when hovered over.

In third and final step I will create clusters (using k-means clustering) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for each electric vehicle charging station exploration and search for the nearby venues.

(1) Customer segmentation is the practice of partitioning a customer base into groups of individuals that have similar characteristics. It is a significant strategy, as it allows the business to target specific groups of customers, so as to more effectively allocate marketing resources.

(2) Clustering can group data only unsupervised, based on the similarity of customers to each other. It will partition customers into mutually exclusive groups. For example, in this project, into five clusters. The customers in each cluster are similar to each other demographically. And then we can create a profile for each group, considering the common characteristics of each cluster. Clustering means finding clusters in a dataset, unsupervised. A cluster is a group of data points or objects in a dataset that are similar to other objects in the group, and dissimilar to data points in other clusters.

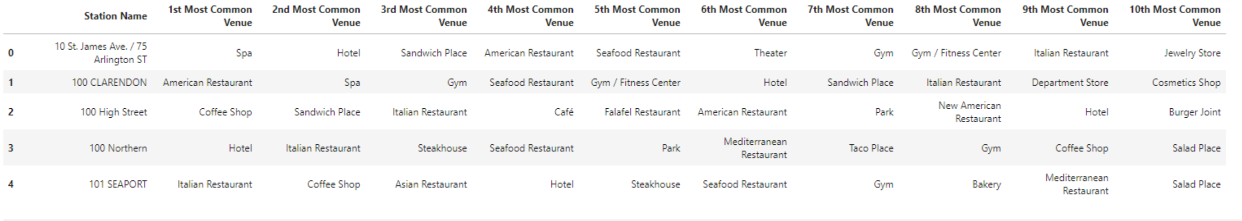
(3) In clustering however, the data is unlabeled and the process is unsupervised. In my analysis, I will use a clustering algorithm-k-means to group similar customers as mentioned, and assign them to a cluster, based on whether they share similar attributes, such as geographic position, surrounding neighborhood status and so on. K-Means can group data only unsupervised based on the similarity of customers to each other. K-Means is a type of partitioning clustering, that is, it divides the data into K non-overlapping subsets or clusters without any cluster internal structure or labels. This means, it's an unsupervised algorithm. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar. K-Means tries to minimize the intra-cluster distances and maximize the inter-cluster distances.

4. Exploratory Data Analysis

(1) Analyze Each Neighborhood

I created one hot encoding, which is dummy variable, for each EVCS based on different neighborhood category. And then normalize this newly-created data to calculate the relative visiting frequency. Figure 1 showed that the top 10 venues for each neighborhood.

Figure 1: the Top 10 Venues for Each Neighborhood



	Station Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	10 St. James Ave. / 75 Arlington ST	Spa	Hotel	Sandwich Place	American Restaurant	Seafood Restaurant	Theater	Gym	Gym / Fitness Center	Italian Restaurant	Jewelry Store
1	100 CLARENDON	American Restaurant	Spa	Gym	Seafood Restaurant	Gym / Fitness Center	Hotel	Sandwich Place	Italian Restaurant	Department Store	Cosmetics Shop
2	100 High Street	Coffee Shop	Sandwich Place	Italian Restaurant	Café	Falafel Restaurant	American Restaurant	Park	New American Restaurant	Hotel	Burger Joint
3	100 Northern	Hotel	Italian Restaurant	Steakhouse	Seafood Restaurant	Park	Mediterranean Restaurant	Taco Place	Gym	Coffee Shop	Salad Place
4	101 SEAPORT	Italian Restaurant	Coffee Shop	Asian Restaurant	Hotel	Steakhouse	Seafood Restaurant	Gym	Bakery	Mediterranean Restaurant	Salad Place

(2) Cluster Neighborhoods

I set K as 5, and visualized the resulting clusters as shown in Figure 2 below.

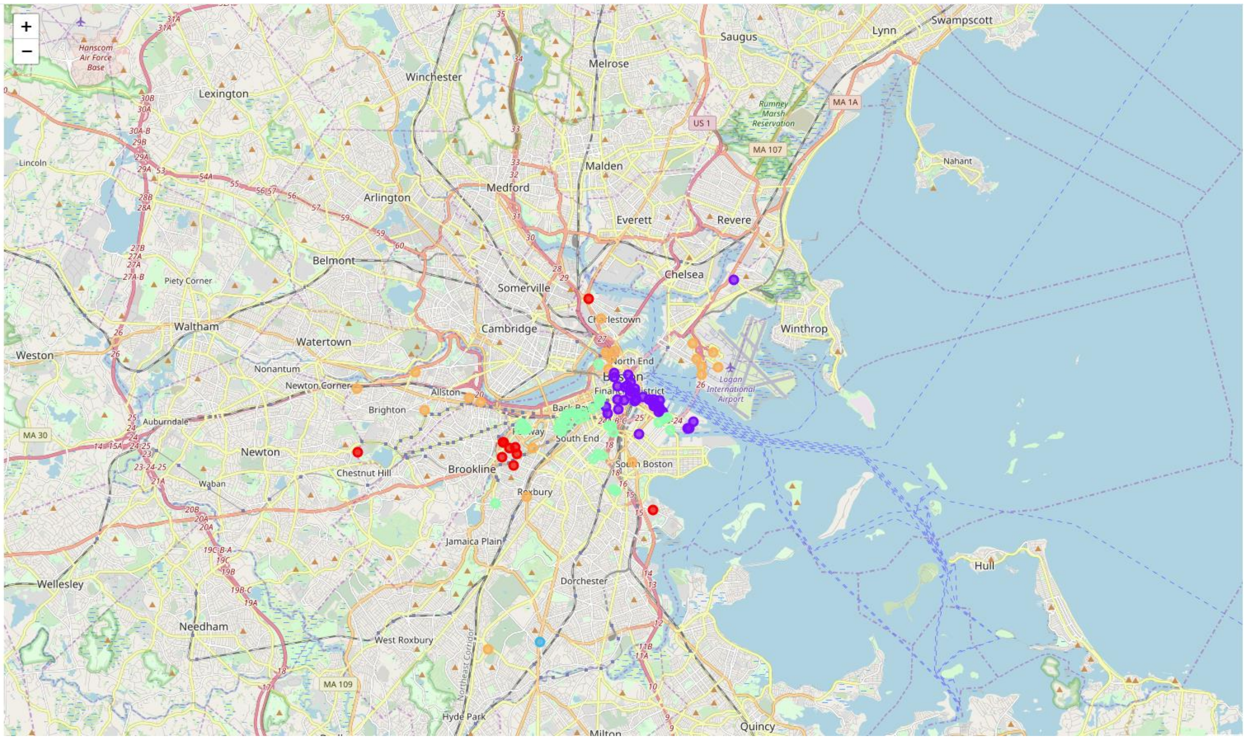


Figure 2: the cluster final map

(3) Examine Clusters

I exchanged the “cluster labels” from 0 to 4 respectively, to display five different clusters. According to the jupyter notebook, cluster 1 is red point on the map, and there are 10 venues in total. Cluster 2 is purple point on the map, and there are 35 venues dropped into this cluster. Cluster 3 is blue point on the map, and it has only one venue. Cluster 4 is brilliant green point on the map, and it has 41 venues in this range. Finally, cluster 5 is brown point on the map with 20 venues in the end.

5. Discussions

Essentially, determining the number of clusters in a data set, or k as in the k-Means algorithm, is a frequent problem in data clustering. The correct choice of K is often ambiguous because it's very dependent on the shape and scale of the distribution of points in a dataset. There are some approaches to address this problem, but one of the techniques that is commonly used is to run the clustering across the different values of K and looking at a metric of accuracy for clustering. This metric can be mean, distance between data points and their cluster's centroid. Which indicate how dense our clusters are or, to what extent we minimize the error of clustering. Then, looking at the change of this metric, we can find the best value for K. But the problem is that with increasing the number of clusters, the distance of centroids to data points will always reduce. This means increasing K will always decrease the error. So, the elbow point is determined where the rate of decrease sharply shifts. It is the right K for clustering. (This method is called the elbow method) However, pre-specifying the number of clusters is not an easy task.