**Data Science Capstone Project**

**Battle of Neighborhood**

**Exploring Venues in Most Popular US Cities for Tourists using Foursquare API**

1. **Introduction**

In this project, we will explore similarities and differences among top five popular US cities for travelers based on the ranking data provided by *thrillist.com*. Cities ranking from 1st to 5th are New Orleans, Portland, Miami, Nashville, Charleston.

Today, most of tourists tend to make decisions for travel destinations based on venues’ ratings provided by applications such as Foursquare, which is also frequently used for travel planning including exploring, dining and lodging. For both now and future, assisting users to make personalized traveling plans would be a direction of great potential for the development of rating apps. To realize the personalization of traveling plans, an application is required to be able to perform classification and evaluation of neighborhoods in cities. For this particular study, we will take an initial step by comparing neighborhoods based on distribution of venues and their ratings, which will contribute to future development of learning algorithm for travel plan design.

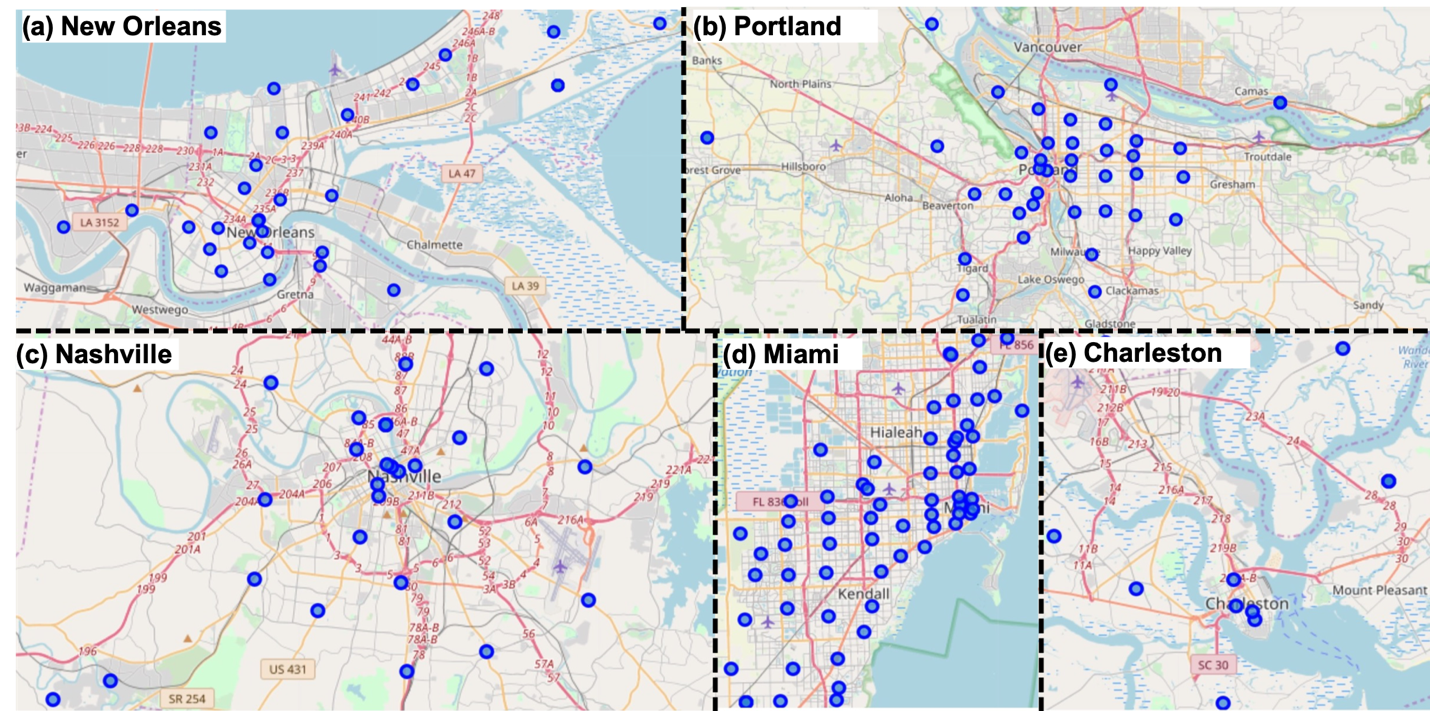
The main objectives of this project include:

1. Identify venues in neighborhoods contained in five cities and comment on feature for each city;
2. Measure similarities between five cities;
3. Make suggestions for travelers on planning design.
4. **Data Discussion**

In this project, neighborhoods are represented by their postal codes. Postal codes and related latitude, longitude data are downloaded from [*https://public.opendatasoft.com*](https://public.opendatasoft.com). Numbers of neighborhoods studied for each city are summarized in Tab.1. Distribution of neighborhoods are presented in Google Map as shown in Fig.1(a-e). 100 venues are collected from Foursquare API, within 500m from center of each neighborhood.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| City | New Orleans | Portland | Nashville | Miami | Charleston |
| Number of Neighborhoods | 66 | 63 | 45 | 96 | 19 |

**Tab.1.** Number of neighborhoods studied for each city.

****

**Fig.1.** Distribution of neighborhood to be explored in cities presented on Google Map.

1. **Methodology**

At first step, coordinates of postal codes for each city are obtained from *https://public.opendatasoft.com*; coordinate data are then used to retrieve venue information from Foursquare API.

Based on venue category hierarchy provided by Foursquare, we classify venues into a set of primary categories that considered important to tourists, i.e., 'Arts & Entertainment', 'Event', 'Food', 'Nightlife Spot' 'Outdoors & Recreation', 'Shop & Service', 'Travel & Transport'.

To quantify the similarity among cities, we need to define a vector to represent each city. Assume that an arbitrary region can be represented by a collection vector which measures the distribution of different kinds of venues:

Where Ci is the number of venues of the same kind I and r is the center coordinate of the region. Notes that C(r) is a normalized vector. The collection vector for the city is then defined as average over all neighborhoods, i.e.,

In this study, the collection vector will be an array with size of 7, representing weights of each primary venue category.

Next, the city-to-city similarity is evaluated by two methods, i.e., the Kendall rank correlation coefficient and a simple Euclidean distance between collection vectors.

Finally, we create a function to assist user to choose a city to explore. The user input will be a set of integer number ranging from 0 to 5, representing how much the user cares about merits of the city, i.e., nature, art, entertainment, dining, shopping, transport. The function will then generate a score for each city through weighted summation referring to the collection vector.