Similarities between Popular US Cities for Tourists: an Exploratory Study Using Foursquare API

Weixuan Li

1.Introduction

In this project, I will explore the top five popular United States cities for travelers based on the ranking provided by *thrillist.com*. Cities ranking from 1st to 5th are New Orleans (LA), Portland (OR), Miami (FL), Nashville (TN) and Charleston (SC).

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 assist designing personalized traveling plans, an application is required to be able to perform classification and evaluation of neighborhoods or cities. For this particular study, I will take an initial step by clustering neighborhoods based on recommended venues, and evaluate cities by their diversities in culture, food, shopping, etc. This study provides some insights into future development of algorithm for travel plan design.

The main objectives of this project include:

1. Clustering neighborhoods in the five cities;
2. Quantify similarities between five cities;
3. Make suggestions for travelers on planning design.

Impression of Five Cities

2.Data Acquisition and Cleaning

2.1.Data Source

In this project, neighborhoods are represented by their postal codes. Postal codes and related latitude, longitude data are downloaded from public website https://public.opendatasoft.com. Recommended venues near each neighborhood are requested from *Foursquare API*. The full category hierarchy file provided by *Foursquare* is employed to validate and correct venue categories.

2.2.Data Cleaning

Available postal codes for New Orleans, Portland, Nashville, Miami and Charleston are extracted from the original table. I assume that each postal code represents one neighborhood. Multiple neighborhoods sharing a same postal code are not counted. 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.

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

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

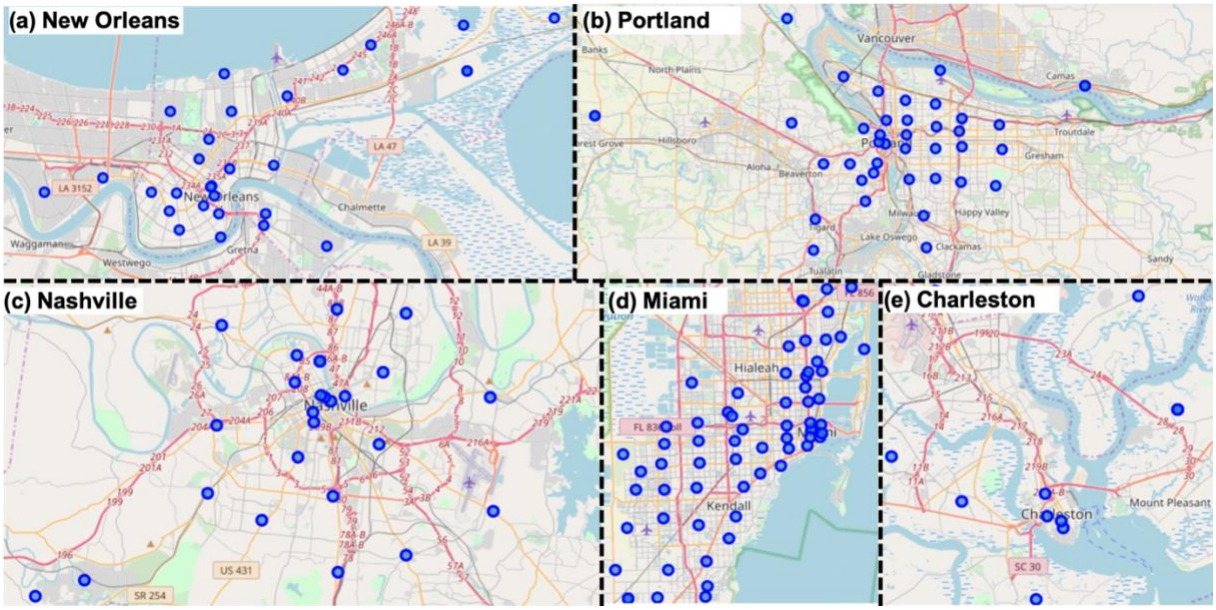


Fig.1. Distribution of neighborhoods presented on Google Map; neighborhoods are marked with blue filled circles.

To explore each neighborhood, 200 recommended venues are collected from *Foursquare API*, within 1000m from the neighborhood center. Venue information include: zip code, coordinates, venue name, venue categories. The Full category hierarchy is then used to identify the primary venue category for classification. An example of cleaned table of venues in New Orleans is shown below in Tab.2.

	Neighborhood Zip	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Primary
0	70112	29.956804	-90.077570	Handsome Willy's Bar and Lounge	29.957193	-90.077989	Dive Bar	Nightlife Spot
1	70112	29.956804	-90.077570	Pho Tau Bay	29.956489	-90.078376	Vietnamese Restaurant	Food
2	70112	29.956804	-90.077570	Rollin' Fatties	29.955108	-90.075870	Food Truck	Food
3	70112	29.956804	-90.077570	Saenger Theatre	29.956432	-90.072603	Performing Arts Venue	Arts & Entertainment
4	70112	29.956804	-90.077570	The Roosevelt New Orleans	29.954343	-90.072298	Hotel	Travel & Transport
...	...	...	...	...	...	...	...	...
1265	70199	29.987528	-90.079501	St. Louis Cemetery No. 2	29.982629	-90.087530	Cemetery	Outdoors & Recreation
1266	70199	29.987528	-90.079501	Loa VIP Viewing Area	29.989060	-90.089231	Music Venue	Arts & Entertainment
1267	70199	29.987528	-90.079501	American Seafood	29.988933	-90.069574	Seafood Restaurant	Food
1268	70199	29.987528	-90.079501	On the Bayou	29.983742	-90.088615	American Restaurant	Food
1269	70199	29.987528	-90.079501	McDonald's	29.983632	-90.070860	Fast Food Restaurant	Food

Tab.2. Example of cleaned table of 1270 venues in New Orleans.

2.3.Feature Selection

After data cleaning, tables of venues for New Orleans, Portland, Nashville, Miami, Charleston contain 1270, 2102, 1575, 2462, 634 rows and 8 columns respectively. Noted that only certain features need to be considered by tourists. For example, venues with primary features as " Colleges & universities", " Professional and other places" are relatively unimportant for a tourist compared to other features.

Therefore, I select seven primary categories for analysis, i.e., art & entertainment, event, food, nightlife spot, outdoor & recreation, shop & service and travel & transport. A summary of venue categories is shown in Tab.3.

	New Orleans	Portland	Nashville	Miami	Charleston
Art & Entertainment	98	43	161	84	16
Event	0	0	0	0	0
Food	580	873	633	1066	341
Nightlife Spot	141	180	193	95	42
Outdoors & Recreation	41	177	96	189	28
Shop & Service	262	671	349	809	147
Travel & Transport	131	129	95	109	33

Tab.3. Number of venues with different categories in New Orleans, Portland, Nashville, Miami and Charleston.

3.Methodology

At first step, coordinates related to available postal codes are extracted from https://public.opendatasoft.com. Latitude and longitude data are then used as parameters in the request for venue information from *Foursquare API*.

A collection vector of summation of seven selected primary features is defined to describe each neighborhood. A k-means clustering is performed to classify neighborhoods of each city based on the collection vector. Distribution of neighborhoods with different classes on the map is then analyzed to provide a basic understanding of city planning.

Similarity among cities are evaluated in terms of functionality and diversity. For functionality, we assume that the normalized collection vector is able to represents distribution of primary categories of venues in each neighborhood:

$$C(r) = \frac{[C_1\ C_2\ \ldots\ C_N]}{\sum C_i}$$

Where C_i 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 normalized collection vector for the city is then obtained by taking average over all neighborhoods, i.e.,

$$C_{city} = \frac{1}{N}\sum_{i=1}^{N} C(r_i)$$

In this study, the normalized collection vector will be an array with size of 7, representing weights of each primary venue category. The functionality similarity is evaluated by two methods, i.e., the Kendall rank correlation coefficient and the simple Euclidean distance between collection vectors.

For the diversity, I considered multi-culture diversity, varieties of entertainment, shopping and outdoor. The diversity is quantified by Shannon-Wiener index, a measurement of entropy, i.e.,

$$H = -\sum_{i=1}^{N} p_i ln p_i$$

where p_i is the portion of i-th secondary venue category in all venues.

Finally, features for each city will be discussed based on analysis results. According to specialty of each city, suggestions will be made for tourists according to their preference.

4.Exploratory data analysis

4.1. Clustering of neighborhoods

A collection vector containing number of primary venue categories is calculated for each neighborhood. Based on collection vectors, neighborhoods are then divided into three types using k-mean clustering method. Distribution of neighborhood classes are presented on maps shown in Fig.2.

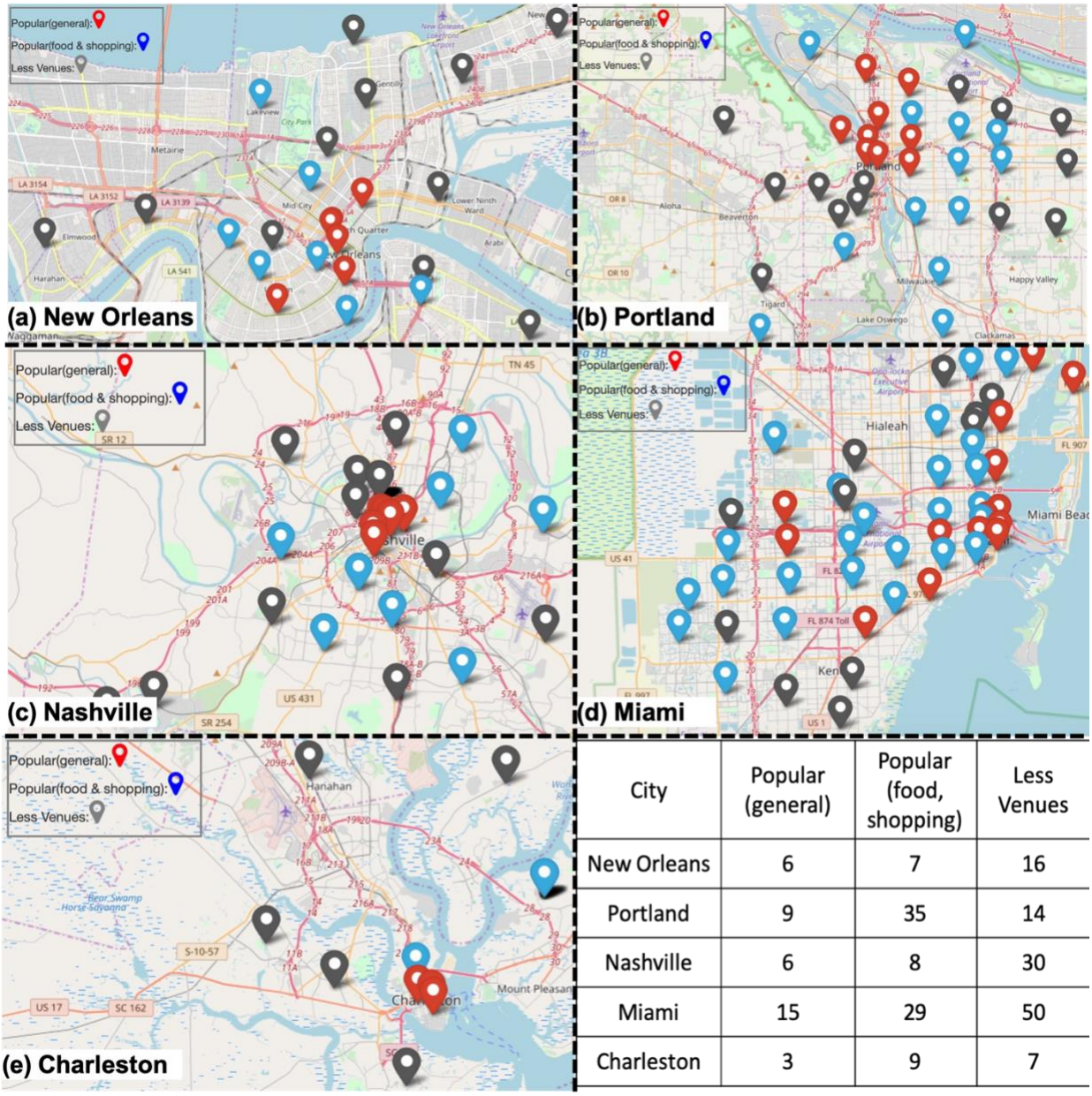
Based on clustering results, three types of neighborhoods are identified with distinctive features:

Type I: this type of neighborhoods has not only the most venues, but also the best variety among all neighborhoods. These features could make type I neighborhoods most popular in the city;

Type II: this type of neighborhoods has fewer venues compared with type I neighborhoods. The majority of venues are found with category of food or shopping. Common venues in these neighborhoods could be shopping plazas;

Type III: this type of neighborhoods has relatively fewer venues in all categories. These neighborhoods could be residential areas or natural, historical sites.

In Portland, Nashville and Charleston, Type I neighborhoods are found to be highly focused in downtown areas near river or sea, as shown in Fig.2(b, c, e). In New Orleans and Miami, Type I neighborhoods are majorly distributed along the river or coastline. Similarly in all five cities, distribution of Type II and Type III neighborhoods is relatively sparse all over the city area. As an overview of neighborhoods, I found that neighborhoods Nashville and Charleston exhibit a central-divergent distribution, while neighborhoods in Portland and Miami exhibit a grid-like distribution. For New Orleans, neighborhoods distribute relatively randomly.



City	Popular (general)	Popular (food, shopping)	Less Venues
New Orleans	6	7	16
Portland	9	35	14
Nashville	6	8	30
Miami	15	29	50
Charleston	3	9	7

Fig.2. Clustering of neighborhoods based on their collection vector for five cities (a-e). Red, blue, gray colors represent popular neighborhood with most various venues, popular neighborhood with most venues as food and shopping, neighborhood with less venues, respectively.

4.2. Analysis of city similarity

4.2.1. Functionality

Using the method discussed in Sec.3, normalized collection vectors for each city with features listed in Tab.3 are calculated. Results are shown in Tab.4.

City	Normalized Collection Vector
New Orleans	[0.078, 0.000, 0.463, 0.113, 0.033, 0.209, 0.105]
Portland	[0.021 0.000, 0.421, 0.087, 0.085, 0.324, 0.062]
Nashville	[0.105, 0.000 0.414, 0.126, 0.063, 0.229, 0.062]
Miami	[0.036, 0.000, 0.453, 0.040, 0.080, 0.344, 0.046]
Charleston	[0.026, 0.000, 0.562, 0.069, 0.046, 0.242, 0.054]

Tab.4. Normalized collection vector for New Orleans, Portland, Nashville, Miami and Charleston, with features of art & entertainment, event, food, nightlife spot, outdoor & recreation, shop & service, travel & transport.

The city similarity in terms of functionality is then estimated with Kendall correlation coefficient and Euclidean distance. City similarities between each pair of cities are summarized in Tab.5 and Tab.6.

	New Orleans	Portland	Nashville	Miami	Charleston
New Orleans	1.000	0.809	0.809	0.619	0.904
Portland	0.809	1.000	0.809	0.809	0.904
Nashville	0.809	0.809	1.000	0..619	0.714
Miami	0.619	0.809	0.619	1.000	0.714
Charleston	0.904	0.904	0.714	0.714	1.000

Tab.5. Kendall rank correlation coefficient matrix for New Orleans, Portland, Nashville, Miami and Charleston.

	New Orleans	Portland	Nashville	Miami	Charleston
New Orleans	0.000	0.153	0.080	0.176	0.135
Portland	0.153	0.000	0.135	0.064	0.168
Nashville	0.080	0.135	0.000	0.166	0.178
Miami	0.176	0.064	0.166	0.000	0.156
Charleston	0.135	0.168	0.178	0.156	0.000

Tab.6. Euclidean space matrix collection vectors of New Orleans, Portland, Nashville, Miami and Charleston.

Considering calculation results from two methods, strong similarities are identified for pairs of (Portland, Miami) and (New Orleans, Nashville). Charleston is a relatively special city compared with other four.

4.2.2. Diversity

4.2.2.1. Restaurant

In *Foursquare*, restaurant category is often contains key word of nation/region. The diversity of restaurant types is therefore considered a direct reflection of multi-culture in a city. Venues with sub-category as "restaurant" are extracted and analyzed. Noted that restaurant without national/regional feature are not considered here. Statistical data are shown in Fig.3 and Shannon-Wiener Indices are presented in Fig.4.

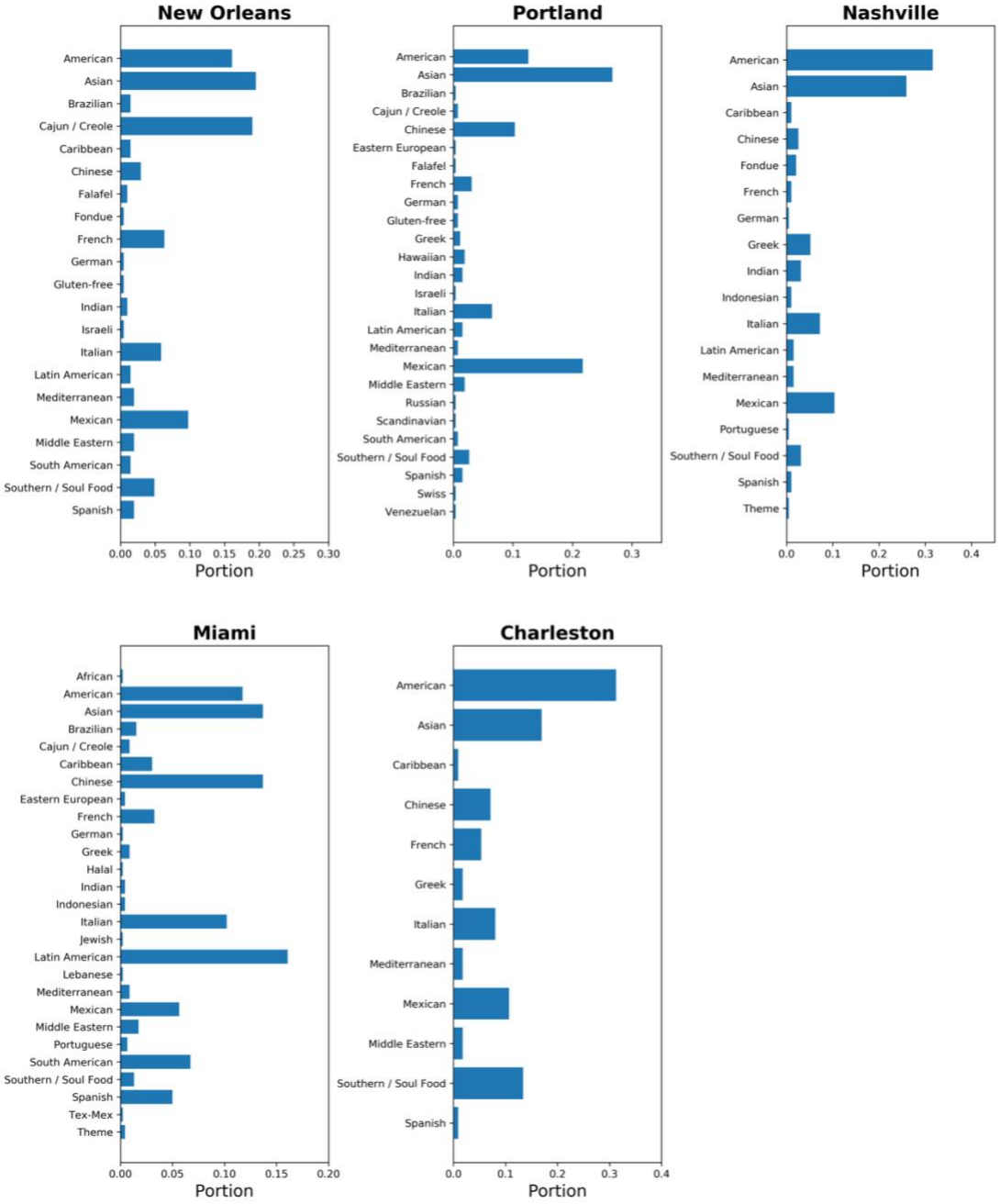


Fig.3. Portion of restaurants with different national/regional types for New Orleans, Portland, Nashville, Miami and Charleston.

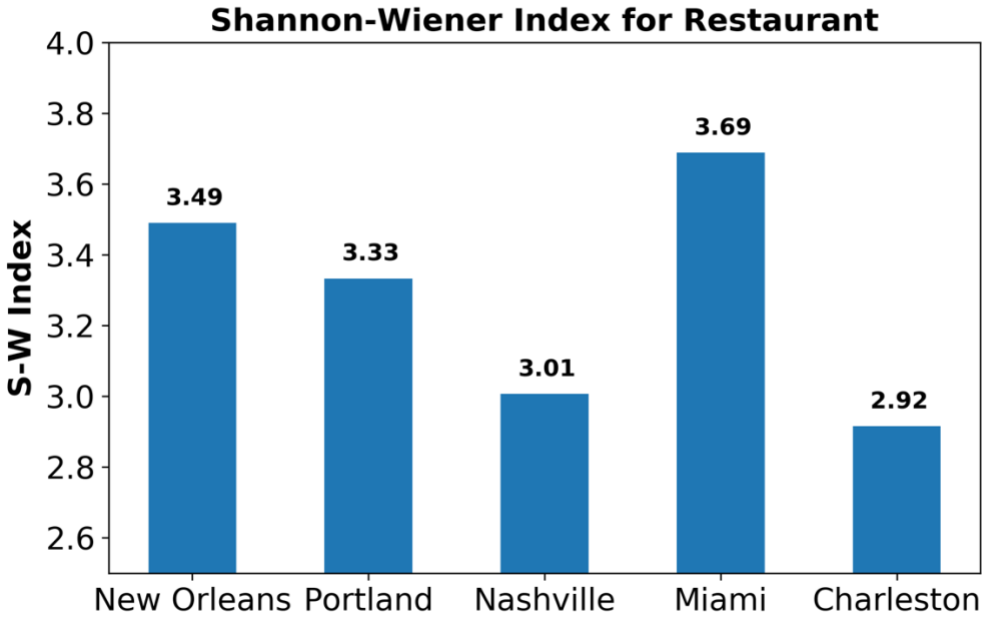


Fig.4. Shannon-Wiener Index for restaurants in New Orleans, Portland, Nashville, Miami and Charleston.

Among five cities, Miami has a great variety of restaurants and exhibit the highest Shannon-Wiener index of 3.69. New Orleans and Portland exhibit similar index but have different dominant types of Cajon and Mexican. Relative lower indices are found for Nashville and Charleston. American food dominates in both cities.

Regarding restaurant diversity, cities can be divided into three categories: Miami, (New Orleans, Portland), (Nashville, Charleston) with high to low diversity.

4.2.2.2. Entertainment

Entertainment, including event, art and nightlife, is another factor that worth considering for tourists. Venues with primary categories of “Art & Entertainment”, “Event” and “Nightlife Spot” are extracted and analyzed. Statistical data and Shannon-Wiener indices are shown in Fig.5 and Fig.6.

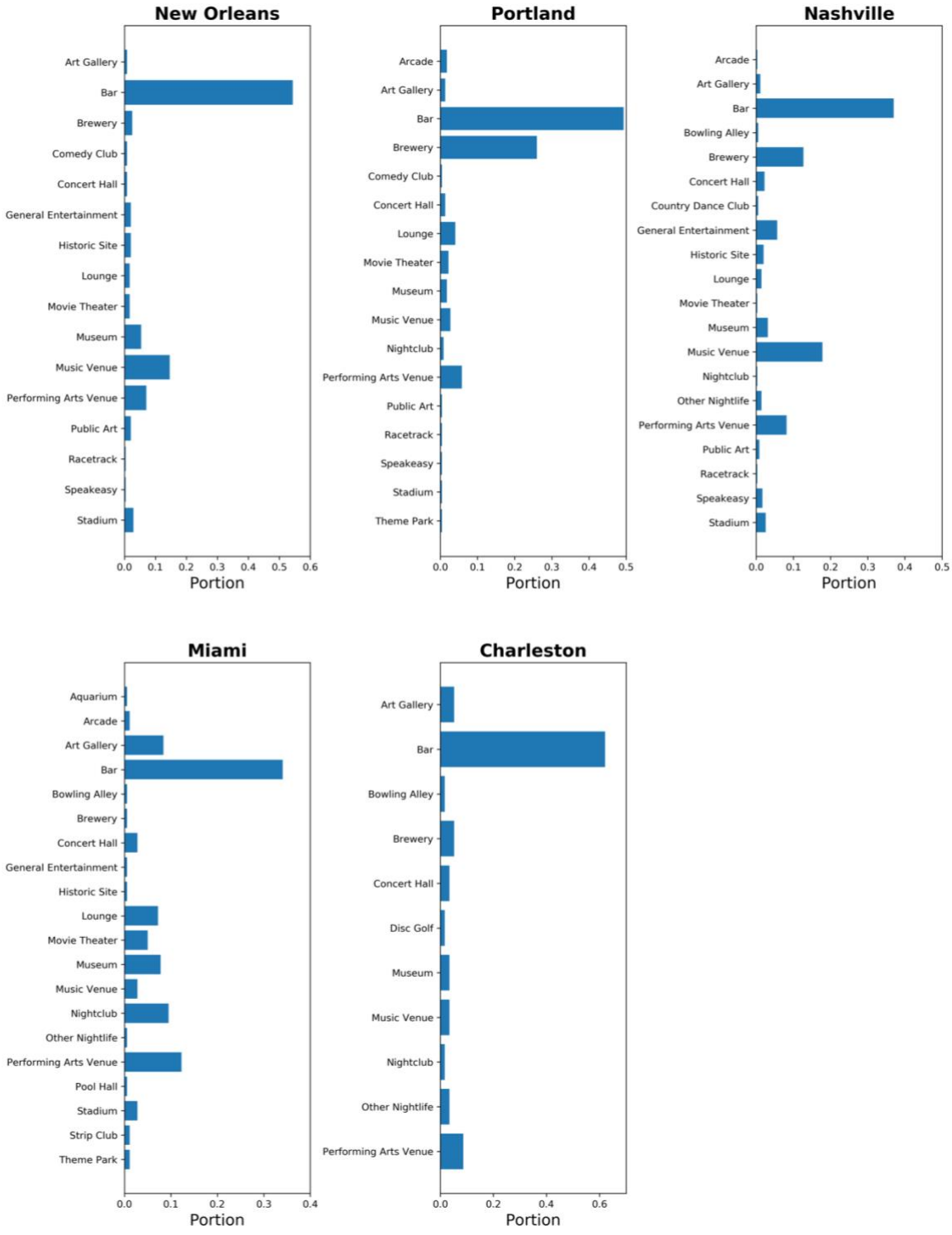


Fig.5. Portion of different types of entertainment for New Orleans, Portland, Nashville, Miami and Charleston.

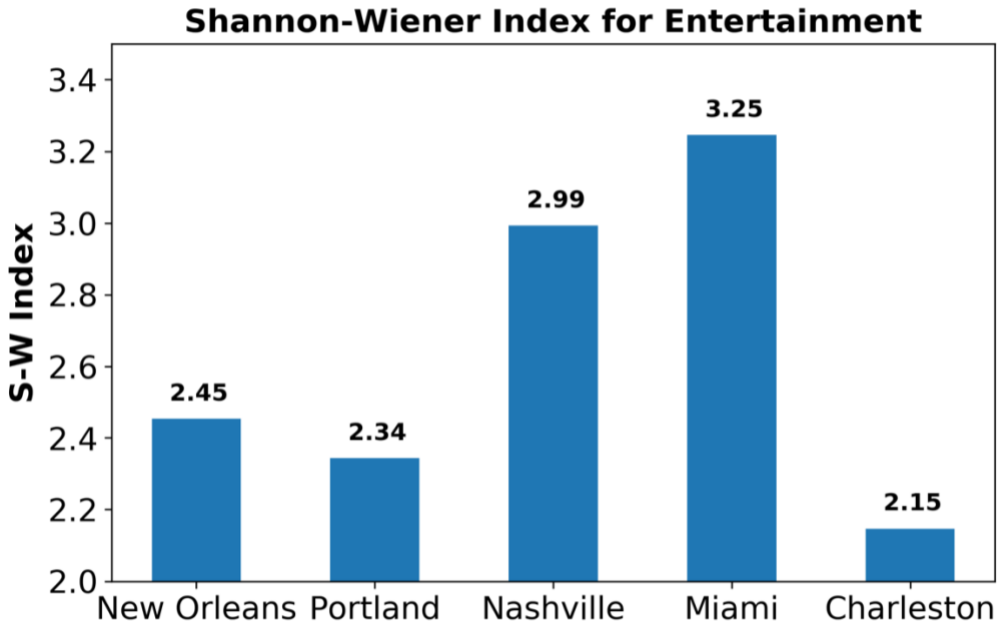


Fig.6. Shannon-Wiener Index for entertainments in New Orleans, Portland, Nashville, Miami and Charleston.

Miami and Nashville exhibit highest SW-indices among five cities. Other three cities have relatively lower SW-indices. It is worth mentioning that venues with type of bar dominate in all cities. Apart from this common feature, each city has its own specialty. New Orleans, Nashville have a great number of music venues and Portland is filled with brewery venues. Venues in Miami shows great variety of several types and Charleston exhibit a focus on art.

In terms of entertainment diversity, cities are categorized into three types: (Miami, Nashville), (New Orleans, Portland), Charleston.

4.2.2.3. Outdoor

Venues including park, lake, spring, trail, etc. are important for tourists who are fan of outdoor activities. In this section, venues with type of “outdoor & recreation” are extracted. Statistical data and Shannon-Wiener indices are shown in Fig.7 and Fig.8.

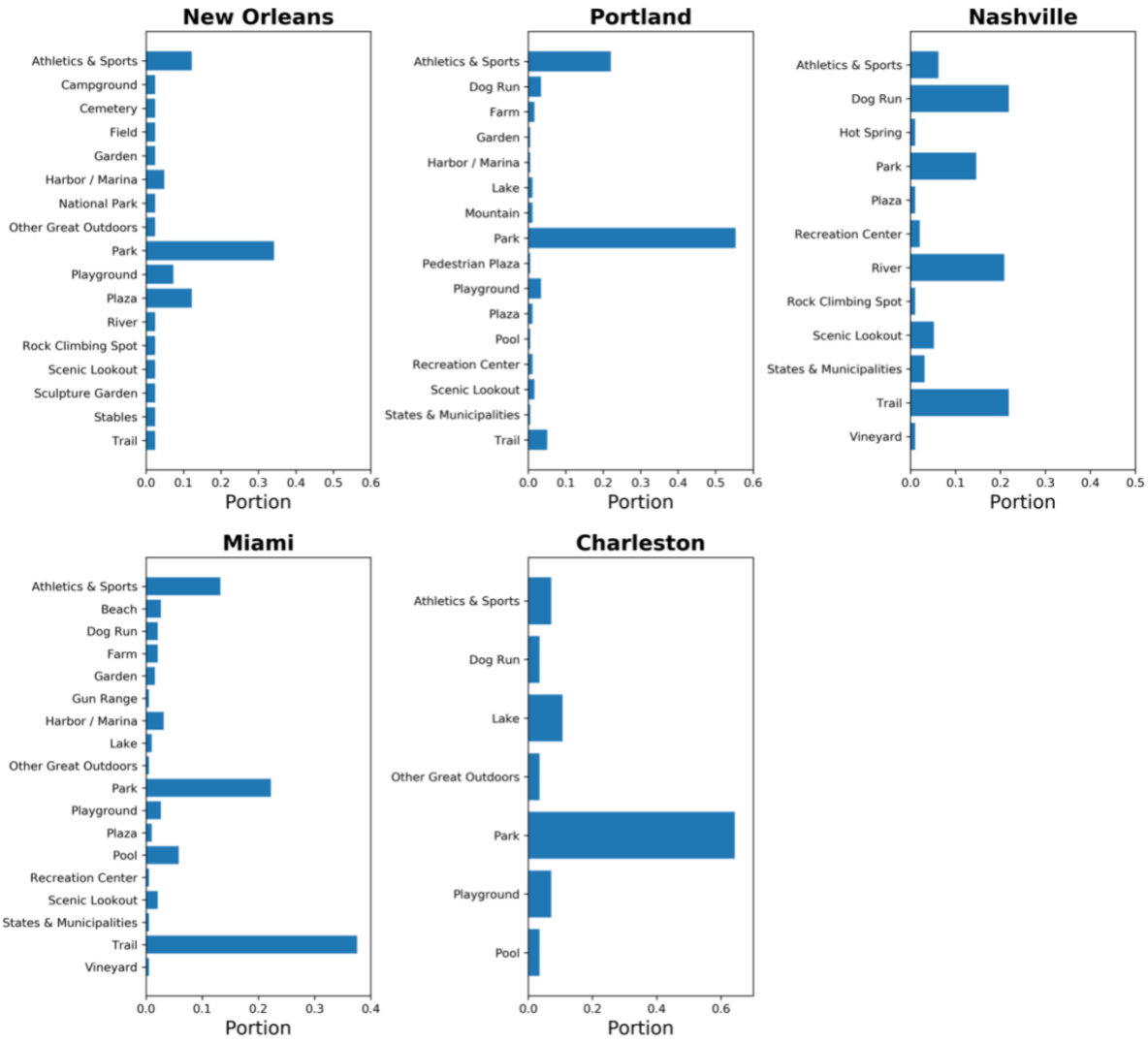


Fig.7. Portion of different types of outdoor venues for New Orleans, Portland, Nashville, Miami and Charleston.

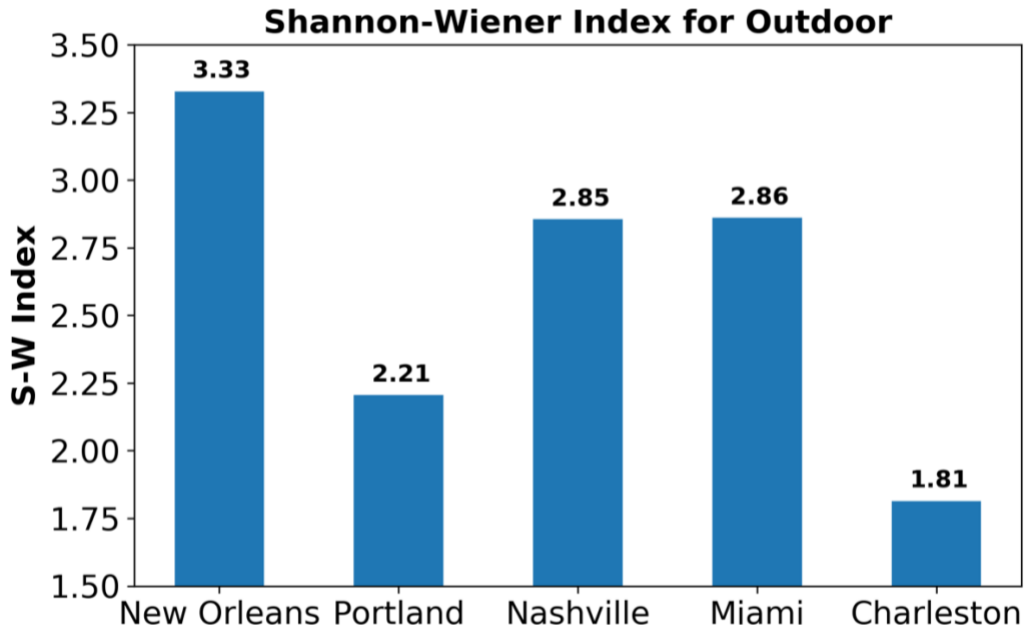


Fig.8. Shannon-Wiener Index for outdoor venues in New Orleans, Portland, Nashville, Miami and Charleston.

In this case, cities can be categorized into three types based on SW-index: New Orleans, (Nashville, Miami), (Portland, Charleston). In addition, New Orleans, Portland and Charleston show emphasis on park. Trail venues dominates in Miami. In Nashville, there is a balance among dog run, park, river and trail.

4.2.2.3. Shopping

One of purposes for a tourist to visit a city would be shopping. Various stores can attract numerous visitors. In this section, venues with category of “shop & service” are extracted. Statistical data and Shannon-Wiener indices are shown in Fig.9 and Fig.10.

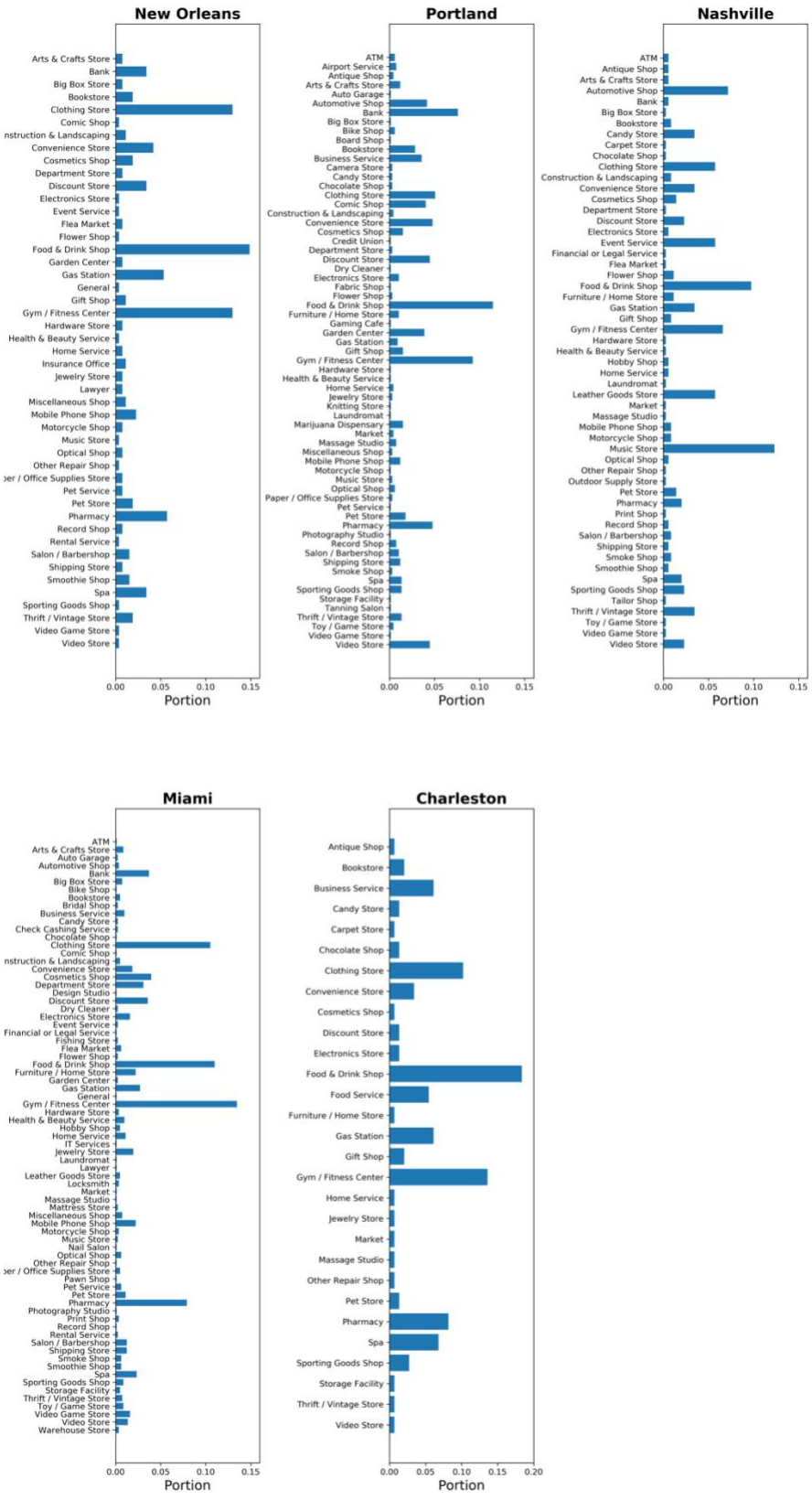


Fig.9. Portion of different types of shopping venues for New Orleans, Portland, Nashville, Miami and Charleston.

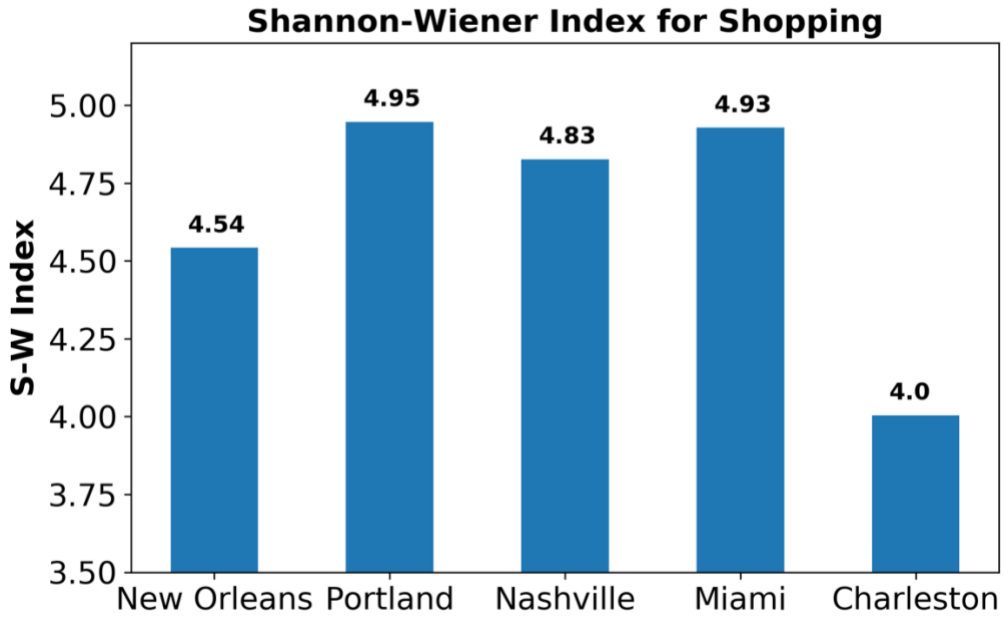


Fig.10. Shannon-Wiener Index for shopping venues in New Orleans, Portland, Nashville, Miami and Charleston.

It is shown in Fig.9 that all cities exhibit great diversity in shopping. In terms of SW-index, Portland, Nashville and Miami are fairly similar. New Orleans and Charleston shows relatively lower index. Among different venues, large number of food & drink shop, gym center and clothing store is a common feature for all cities. Apart from that, it worth mentioning that Nashville exhibit a unique feature, i.e., the large portion of music store.

5. Discussion

Based on data analysis results, features for the five cities are summarized below.

New Orleans: Multi-culture

As shown by the neighborhood clustering results, different types of neighborhoods are relatively randomly distributed over the city area. Most popular neighborhoods, containing most various venues, are located near the Mississippi River. In terms of diversity, New Orleans exhibit high SW indices in restaurant and outdoor activities. For restaurants, American, Cajun and Asian foods dominate. Mexican food also takes an appreciable part. For outdoor activities, New Orleans exhibit a great variety with a dominant category of park. Relative lower SW indices are found for shopping and entertainment, which are dominated by clothing store and bar. Overall, New Orleans is a multi-cultural city with various outdoor activity choices. Also, bar and clothing stores are highly recommended for tourists.

Portland: Shopping Heaven

Neighborhoods in Portland show a grid-like distribution. Popular neighborhoods are focused in downtown near Willamette river. In Portland, restaurants have dominant types of Asian, Mexican, American, Chinese and Italian. The SW index for entertainment is relatively low, with a great emphasis on alcohol (Bar and Brewery). Park and sports are dominating outdoor activities in Portland. Most importantly, Portland has the highest SW index in shopping. Such feature is strongly recommended for visitors.

Nashville: City of Music

In Nashville, neighborhoods distribution is central-divergence like. Popular neighborhoods are focused in downtown area. Based on the SW index of restaurant, Nashville is not a multi-culture city and it is dominated by American culture. Nashville has best music related venues, which dominate in both entertainment and shopping. For outdoor activities, Nashville shows emphasis on river, trail and dog run, which are unique among five cities. Overall, despite that Nashville does not have a great diversity, its unique features for river and music are highly recommended for romantic tourists.

Miami: Multi-culture

Neighborhoods in Miami are distributed grid-like. There is no obvious concentration of popular neighborhoods, which are distributed along the coastline. Miami is a multi-cultural and multi-functional city. It has the highest SW index in restaurant, entertainment and the second highest SW index in shopping. Besides, the great emphasis on art-related venues is an unique feature for Miami. Overall, Miami is highly recommended for visitors who enjoy multiple cultures.

Charleston: Small, Lovely

The area of Charleston is smaller compared to other four cities. Popular neighborhoods are focused in downtown area in south of Charleston. SW indices for Charleston are relatively lowest. However, evaluation solely based on diversity is not fair for this city. Charleston shows great emphasis on American culture, bar and park. It is recommended for a relative short period travel and joy.

For quantification, I defined a simple function to generate score for each city considering two kinds of user preferences as inputs. For the first input type, the user provides a set of number between 0 to 5, showing how much they care about primary features including art & entertainment, event, food, nightlife spot, outdoor & recreation, shop & service and travel & transport. Based on normalized collection vectors, the function then calculate the score for each city by weighted average. For the second input type, the user indicates specific venues to explore. The function will search venues with input key words and count the number of venues. The number is regarded as the score of the city. Fig.11 shows an input-output example for the defined function.

```
choose method: 1 or 2
1
entertainment:5
event:3
food:4
nightlife:5
outdoor:2
shop:2
travel:3
Scores:
New Orleans= 3.60
Portland= 3.23
Nashville= 3.59
Miami= 3.18
Charleston= 3.46
```

```
choose method: 1 or 2
2
How many features you want:1
please indicate (lower case):art
['art']
New Orleans=25.0
Portland=21.0
Nashville=37.0
Miami=61.0
Charleston=5.0
```

Fig.11. An input-output example for defined function for providing suggestions for users.

6. Conclusion and Future Suggestions

In this project, I analyzed neighborhoods and venues in five popular tourists' cities in United States based on data downloaded from the rating application *Foursquare*. Features important to travelers, i.e., art & entertainment, event, food, nightlife spot, outdoor & recreation, shop & service and travel & transport, are selected for data analysis. Neighborhoods are clustered by the k-means algorithm. Distinctive features in neighborhood distribution are identified in each city. In addition, city similarities are analyzed in aspects of functionality and diversity. For functionality, Miami & Portland, New Orleans & Nashville show good similarity. In terms of diversity, each city shows unique features compared to others. Finally, I made suggestions for travelers based on data analysis results.

Future study can be conducted on personalized travel plan design. Based on user required features, and analysis of venues' locations, spatial distributions and related lodge, transport, an optimized itinerary can be designed.

Data Source

City ranking data:
https://www.thrillist.com/travel/nation/best-us-cities-to-spend-a-weekend-nashville-austin-charleston-providence

Postal codes and Coordinates:
https://public.opendatasoft.com

Venue information and Category hierarchy:
Foursquare API

References

1. Daniel Preoţiuc-Pietro, Justin Cranshaw, and Tae Yano. 2013. Exploring venue-based city-to-city similarity measures. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. ACM, New York, NY, USA, Article 16, 4 pages. DOI: https://doi.org/10.1145/2505821.2505832
2. López Baeza, Jesús & Cerrone, Damiano & Männigo, Kristjan. (2017). Comparing two methods for Urban Complexity calculation using Shannon-Wiener index. WIT Transactions on Ecology and Environment. 226. 369-378. 10.2495/SDP170321.