HOW TO IMPROVE THE LIVABILITY OF SOME OKLAHOMAN CITIES

A study of Livability.com's Top-100 places to live in 2020
and lessons for Oklahoma City, Tulsa and Stillwater.

Hernan Fernandez-Barriales Lopez

November 2020

FINAL ASSIGNMENT – CAPSTONE PROJECT

How to improve the livability of some Oklahoman cities

A study of Livability.com's Top-100 places to live in 2020
and lessons for Oklahoma City, Tulsa and Stillwater

1. Introduction

City officials have always been interested in improving the living conditions of their fellow citizens, while making their cities more appealing for national and international tourism. While the reasons behind this could lie within the selfless interest in making others' lives better, there are usually more "selfish" reasons for this including the interest in getting re-elected for the position, the generation of more tax revenue, and the recognition attained when doing an outstanding job by media and other officials nationwide.

The availability of more and more data to quantify the impact of several factors in the living standards of the population, assess the efficiency of implemented measures and ordinances, and to compare the overall performance to other cities across the nation have proven a very valuable asset that continues to be mined by the most modern administrations.

This report will study data from the 100 cities identified in www.livability.com as 2020's Top-100 Best Places to Live in the US, and try to extract information that could provide guidance for city officials to replicate the success formulas of these "top places to live". In particular, the lessons learned through the analysis of readily available data will be applied to provide recommendations to three major cities in the State of Oklahoma (Oklahoma City, Tulsa and Stillwater) in order to become the first Oklahoman city to make it into the list of Top-100 places to live in the US.

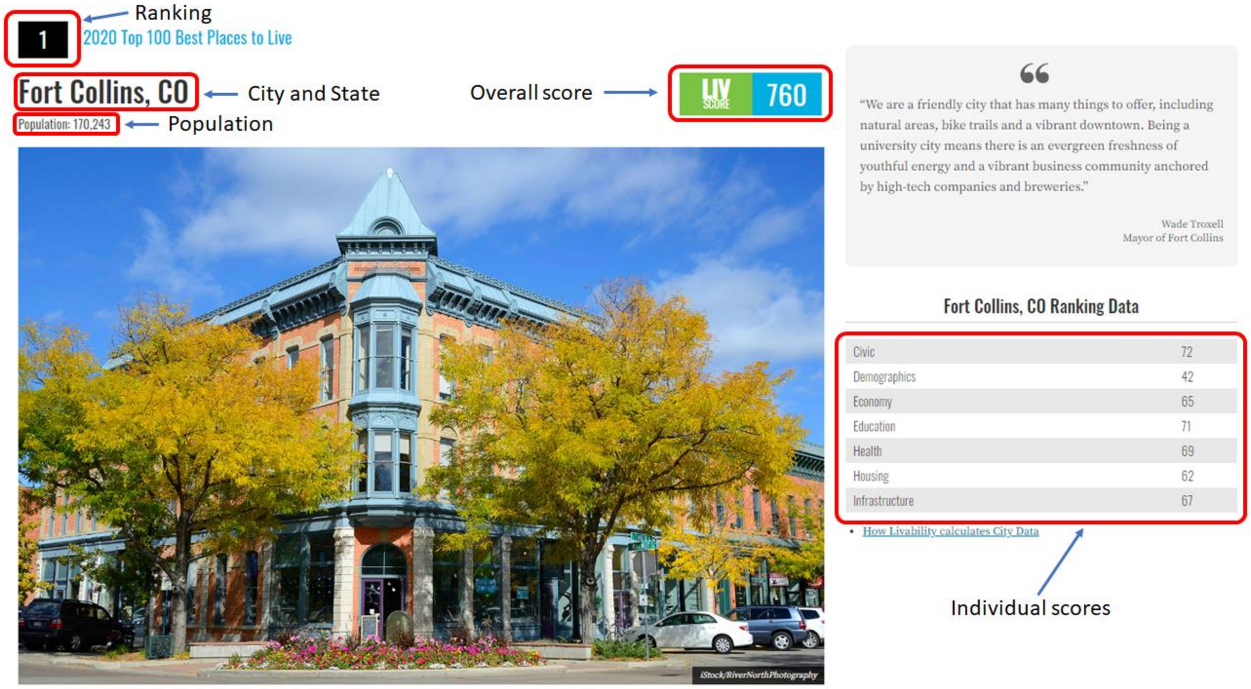
2. Data

2.1. Initial approach at data collection

The initial data frame was collected manually from a report put together by the website livability.com (https://bit.ly/2K4df3x). The report is a result of the statistical analysis of polling data from 1,000 US cities, a study performed in collaboration with Ipsos. This report provides a series of numerical indicators of different aspects of city live (civics, demographics, economy, education, health, housing and infrastructure), calculated using raw data from a wide array of sources, such as the US Census Bureau, US Department of Housing and Urban Affairs, USDA, and many more.

Using these 7 individual indicators, an overall "Livability Score" (LivScore) is calculated and used to rank the top 100 places to live. There is not a proper description explanation of either how the individual indicators are calculated, or how these are used to come up with the final score.

A sample of one of the 100 pages is shown next:



“We are a friendly city that has many things to offer, including natural areas, bike trails and a vibrant downtown. Being a university city means there is an evergreen freshness of youthful energy and a vibrant business community anchored by high-tech companies and breweries.”

Wade Troxell
Mayor of Fort Collins

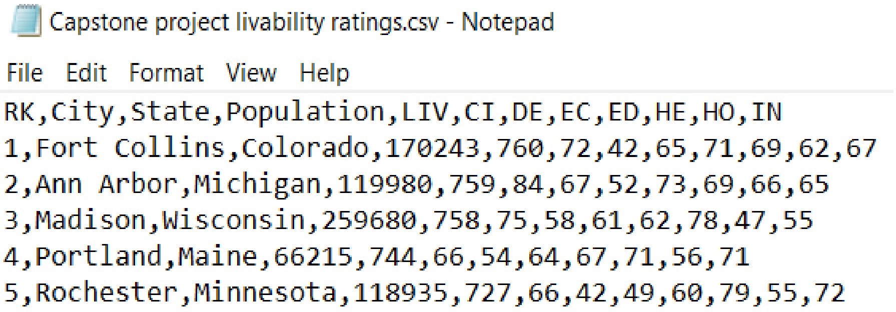
Fort Collins, CO Ranking Data

Civic	72
Demographics	42
Economy	65
Education	71
Health	69
Housing	62
Infrastructure	67

- How Livability calculates City Data

Figure 1. Page of the 1st-ranked city according to Livability.com's TOP 100 Best Places to Live: Fort Collins, Colorado

Having the information spread through 100 individual pages, it was considered faster to browse through them manually collecting the few bits of information required from each page than to develop a code to automatize the data collection. The information was collected in a comma-separated-values (csv) file titled “Capstone project livability ratings.csv”, which will be available in the Jupyter Notebook. The first five rows of data along with the column names can be seen in the following image:


Capstone project livability ratings.csv - Notepad
File Edit Format View Help

```
RK,City,State,Population,LIV,CI,DE,EC,ED,HE,HO,IN
1,Fort Collins,Colorado,170243,760,72,42,65,71,69,62,67
2,Ann Arbor,Michigan,119980,759,84,67,52,73,69,66,65
3,Madison,Wisconsin,259680,758,75,58,61,62,78,47,55
4,Portland,Maine,66215,744,66,54,64,67,71,56,71
5,Rochester,Minnesota,118935,727,66,42,49,60,79,55,72
```

Figure 2. 5 first rows of the csv file containing the raw data

While a few insights can be extracted from this raw data, the core of this project will deal with analyzing these cities using data from the different categories of venues in their city centers using the FourSquare API.

The Livability report has given us the names of 100 cities throughout the United States that are worth investigating further. We can feed these names to a Python program that will first extract relevant data about the types of venues present in these cities, and then will help us interpret and extract conclusions from this data.

How do we intend to do this? The general categories of the venues present in each of the cities will be extracted using the popular FourSquare API, and their relative frequency used to identify various “types of cities” (i.e. a nightlife hotspot, an outdoors paradise, an artistic center, etc.). The different general categories, as described in the FourSquare website, are the following:

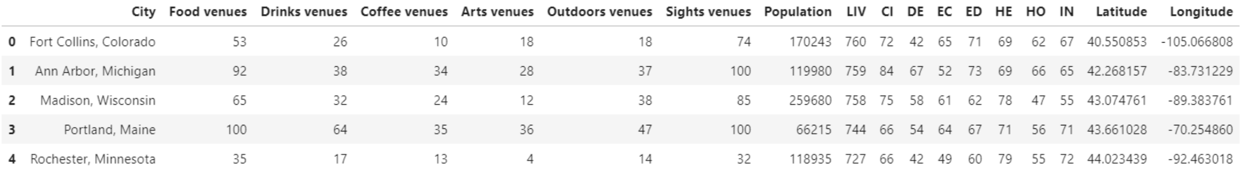
a. Food: Restaurants, food trucks, delis, etc.
b. Drinks: Places to meet and have a drink, such as breweries, clubs, pubs, etc.
c. Coffee: Places to sit and have a warm beverage, such as a café.
d. Shops: Places to buy different products, such as a shopping mall, an antique shop, etc.
e. Arts: Music venues, museums, galleries, etc.
f. Outdoors: Parks, trails, gardens, etc.
g. Sights: Landmarks such as statues, monuments, etc.

A program will be built that extracts and stores the total number of venues of a particular type for each of the cities using the FourSquare API. This data set will then be evaluated to determine the different shares of each type of venue on each of the cities, and how they compare to each other (using a clustering technique such as K-Means).

Then, the case study entities of Oklahoma City, Tulsa and Stillwater will be evaluated to determine which type of city (i.e. which cluster) they belong to using another machine learning technique (K-Nearest Neighbors). Finally, we will determine which of the Top 100 cities are more similar to the three Oklahoman cities using Euclidean distance.

With the insights accumulated throughout this process, a series of recommendations will be drafted that could help the three Oklahoman cities to climb up in the ranking, maybe reaching one year the Top 100 cities ranking.

The following image shows a sample of the type of data that will be available after collection and cleanup:

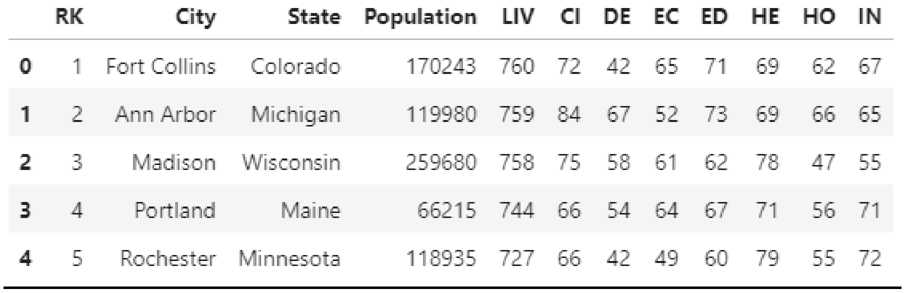


	City	Food venues	Drinks venues	Coffee venues	Arts venues	Outdoors venues	Sights venues	Population	LIV	CI	DE	EC	ED	HE	HO	IN	Latitude	Longitude
0	Fort Collins, Colorado	53	26	10	18	18	74	170243	760	72	42	65	71	69	62	67	40.550853	-105.066808
1	Ann Arbor, Michigan	92	38	34	28	37	100	119980	759	84	67	52	73	69	66	65	42.268157	-83.731229
2	Madison, Wisconsin	65	32	24	12	38	85	259680	758	75	58	61	62	78	47	55	43.074761	-89.383761
3	Portland, Maine	100	64	35	36	47	100	66215	744	66	54	64	67	71	56	71	43.661028	-70.254860
4	Rochester, Minnesota	35	17	13	4	14	32	118935	727	66	42	49	60	79	55	72	44.023439	-92.463018

Figure 3. Top 5 rows of the cleaned Python data frame used in this study.

2.2. Data acquisition

As it was mentioned earlier, the initial data from livability.com was acquired manually and recorded in a csv file that is imported into our python code at the beginning of the process.



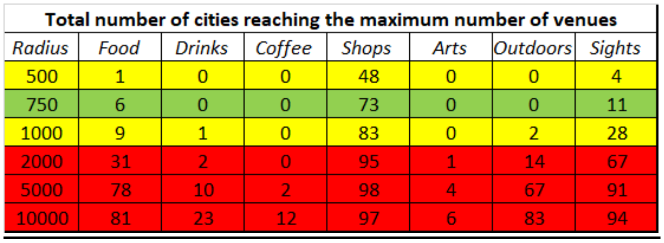
	RK	City	State	Population	LIV	CI	DE	EC	ED	HE	HO	IN
0	1	Fort Collins	Colorado	170243	760	72	42	65	71	69	62	67
1	2	Ann Arbor	Michigan	119980	759	84	67	52	73	69	66	65
2	3	Madison	Wisconsin	259680	758	75	58	61	62	78	47	55
3	4	Portland	Maine	66215	744	66	54	64	67	71	56	71
4	5	Rochester	Minnesota	118935	727	66	42	49	60	79	55	72

Figure 4. First 5 rows of data from the imported csv file.

The latitude and longitude of each of the cities is then obtained using a loop and the geocode function. After appending the coordinates to the data frame, another loop will be used to generate a list of the FourSquare urls that will be used to collect the data. These urls will use the following inputs:

a. Location: the City and State names will be used instead of the coordinates, as geocode doesn't do a great job assigning the coordinates to the actual city center.
b. Radius: the maximum distance from these coordinates that the API will return results from. This is set manually.
c. Limit: the maximum number of venues that will be returned. Also set manually.
d. Section: the venue type. This is set as one of the major categories in each of the urls generated for each city.

Here we encounter our first problem. The FourSquare API restricts the number of results to 100 (i.e. setting the limit above to any number above 100 will automatically truncate it to 100). This introduces a restriction in the radius we can set in our analysis if we want to avoid all cities to return the number of venues as 100. The code was ran setting different values for radius, and it was determined that the best results would be obtained for a 750-meter radius. By seeing the results of the different iterations, we see we will have to drop the shop venues column, but at 750 meters we can conduct our analysis of the rest of the venue types dropping just a few cities.

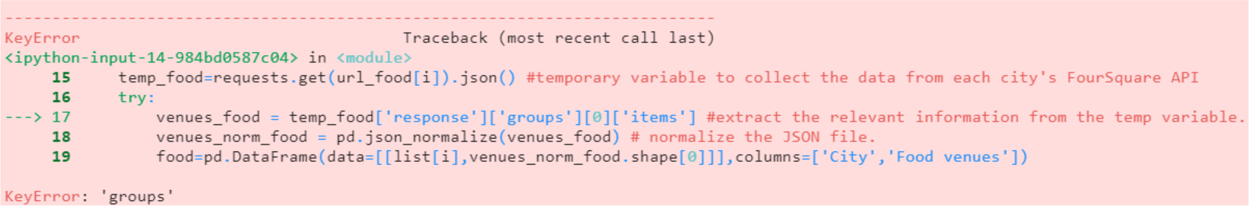


Total number of cities reaching the maximum number of venues							
Radius	*Food*	*Drinks*	*Coffee*	*Shops*	*Arts*	*Outdoors*	*Sights*
500	1	0	0	48	0	0	4
750	6	0	0	73	0	0	11
1000	9	1	0	83	0	2	28
2000	31	2	0	95	1	14	67
5000	78	10	2	98	4	67	91
10000	81	23	12	97	6	83	94

Figure 5. Selection matrix for radius.

We then move into the actual FourSquare data collection step, where we use another while loop to feed the 7 urls for each venue type (Food, Drinks, Coffee, Shops, Arts, Outdoors and Sights) for each of the 100 cities and store the number of venues of each type within each city into a variable.

The API is not very reliable, and can return the following error message somewhat randomly in a particular step of the loop, and the next time you run the loop it won't happen.



```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
<ipython-input-14-984bd0587c04> in <module>
     15     temp_food=requests.get(url_food[i]).json() #temporary variable to collect the data from each city's FourSquare API
     16     try:
---> 17         venues_food = temp_food['response']['groups'][0]['items'] #extract the relevant information from the temp variable.
     18         venues_norm_food = pd.json_normalize(venues_food) # normalize the JSON file.
     19         food=pd.DataFrame(data=[[list[i],venues_norm_food.shape[0]]],columns=['City','Food venues'])

KeyError: 'groups'
```

Figure 6. FourSquare API error that occurs somewhat randomly.

To avoid stopping the loop every time this happens, a try/except code is introduced, so every time we get this error, the particular value will show as blank and the code will continue to run. The loop is then ran a couple of times and the results combined to avoid losing valuable data.

The code was ran a few times, and all of the times the same values where obtained. This allowed to "fill in the blanks" left by the error showed in Figure 6. The values where then stored in another csv file that will reduce the time of running the code considerably.

Finally, the same loop is used to acquire the same data for 3 Oklahoman cities that we want to evaluate against the Top 100 Best Cities to Live in 2020, that is, Oklahoma City, Tulsa and Stillwater. This concluded the data acquisition portion of this project.

2.3. Data cleaning

As it was briefly mentioned earlier, the limit imposed by the FourSquare API of just 100 results per query demands that we perform some cleaning of the data.

The first thing that needs to be done is to remove altogether the row for Shop venues, as almost 3 in 4 cities (73%) return 100 venues. The reason why we shouldn't accept these values is because they are not representative of the true number of shops found in the 750 given radius (i.e. 100 could mean the city indeed has exactly 100 shops in that area or, more likely, that it has any number from 100 to 1,000 stores).

The same problem was encountered using this radius with 6 cities that provided 100 results for Food venues and 11 that provided 100 results for Sights venues. After giving it thorough consideration, it was decided to drop those individual cities rather than dropping the entire venue categories, resulting in a 87-city data frame containing values of Food, Drinks, Coffee, Arts, Outdoors and Sights venues.

We are now ready for the proper data analysis portion of the project.

3. Methodology and Results

After the cleaning process, we have two sets of valuable information to group the different successful cities into clusters; the 7 livability ratios described in the Top 100 Places to Live in the US report, and the number of venues of each of the 6 types collected through our python program using the FourSquare API.

3.1. General information

The distribution of the 100 cities included in the Livability.com study are shown in Figure 7:

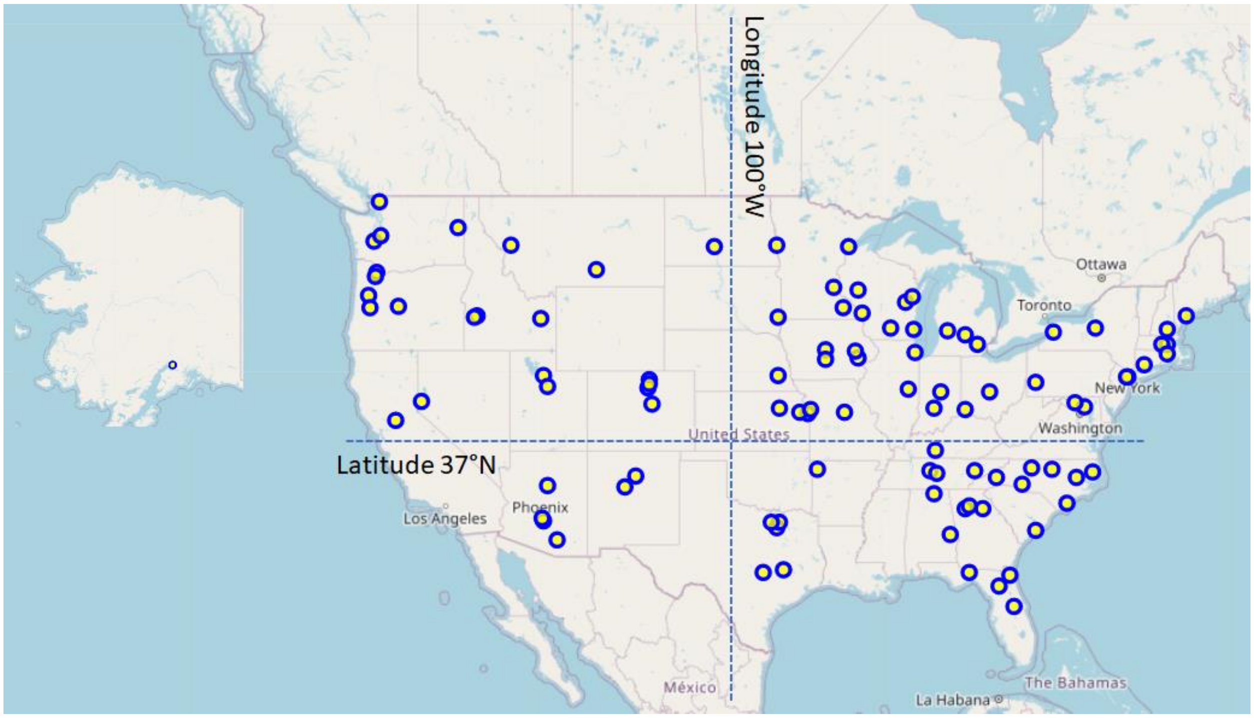


Figure 7. Distribution of the Top 100 Best Cities to live in the US in 2020 as per Livability.com

At a first glance, we can see the number of cities east of the 100°W meridian is way more densely populated, with a pretty straight line connecting Fargo, ND, in the north and Pfluggerville, TX, in the south acting as a sort of boundary. Using our python code, we determined that 70% of the cities lie to the east of this boundary. When it comes to North-South Distribution, we can identify another sort of boundary at the 37°N parallel, having 67 cities north of that boundary and 33 south of it.

Running a simple line of code, we can determine which State has the most representation in the Top 100 Cities (Wisconsin, with 6 cities). On the other hand, Oklahoma is one of the 9 states with no representation in the Top 100. The following table summarizes the results:

Table 1. Number of cities representing each US State in the Top 100.

Top 100 cities	States
6	Wisconsin
5	North Carolina, Texas, Washington
4	Arizona, Colorado, Florida, Iowa, Oregon
3	Georgia, Idaho, Kansas, Michigan, Minnesota, New York, South Carolina, Tennessee
2	Alabama, Illinois, Indiana, Maryland, Massachusetts, Missouri, Montana, New Mexico, North Dakota, Ohio, Utah
1	Alaska, Arkansas, California, Connecticut, Kentucky, Maine, Nebraska, Nevada, New Hampshire, New Jersey, Pennsylvania, Rhode Island, South Dakota
0	Delaware, Hawaii, Louisiana, Mississippi, **Oklahoma**, Vermont, Virginia, West Virginia, Wyoming

A histogram will show us the general distribution of the LIV scores (see figure 8). It resembles a normal distribution centered at 650.

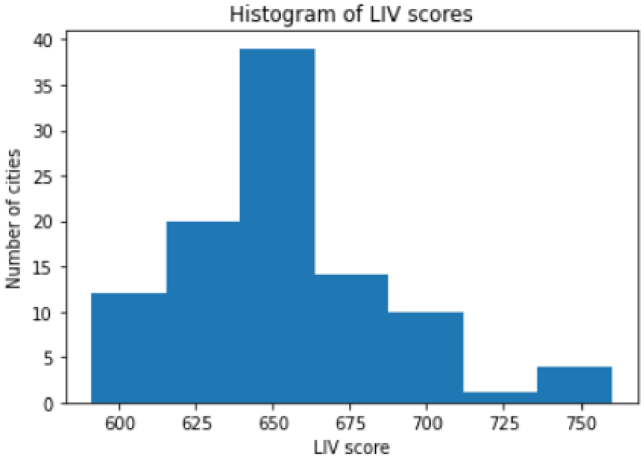


Figure 8. Distribution of LIV scores in Top 100 cities.

Finally, lets focus on the individual scores on each of the 7 categories studied in the Livability.com report; civics, demographics, economy, education, health, housing and infrastructure. The following table summarizes the top 5 cities in each of those categories:

Table 2. Top-5 cities in each category of the Livability.com report.

	#1	#2	#3	#4	#5
LIV	Fort Collins, CO	Ann Arbor, MI	Madison, WI	Portland, ME	Rochester, MN
Civics	Carmel, IN	Oak Park, IL	Franklin, TN	Ann Arbor, MI	Columbia, MD
Demographics	Jersey City, NJ	Providence, RI	New Haven, CT	St. Paul, MN	Sacramento, CA
Economy	Lansing, MI	Loveland, CO	Tucson, AZ	Vancouver, WA	Spokane, WA
Education	Marietta, GA	Overland Parks, KN	Gainesville, FL	Ann Arbor, MI	Bloomington, IN
Health	Iowa City, IA	Rochester, MN	Madison, WI	Eau Claire, WI	Manhattan, KN
Housing	Winston-Salem, NC	Billings, MT	Asheville, NC	Auburn, AL	Champaign, IL
Infrastructure	Anchorage, AK	Rochester, MN	Flagstaff, AZ	Manhattan, KN	Portland, ME

It is surprising that Fort Collins, ranked #1 in the Livability Score, doesn't appear in any other Top-5, which must mean it ranks relatively high in almost every category. In fact, a few lines of code indicate us that they rank 8th in Infrastructure, 10th in Education, 13th in Economy, 15th in Civics, 16th in both Housing and Health, and 74th in Demographics.

3.2. Venue analysis

The core idea of this project is to evaluate the composition of the city centers of each of the Top 100 cities identified by Livability.com. As it was previously described, the FourSquare API divides the different venue types into 7 general categories; food, drinks, coffee, shops, arts, outdoors and sights. Through our code, we studied the total number of venues of each category in a 750-meter radius around what the FourSquare API considers the city center of each of the 100 cities.

Similarly to we did earlier with the Livability.com scores, the next table shows the top 5 cities for each of the venue categories returned by our program.

	#1	#2	#3	#4	#5
Food	Sacramento, CA (76)	Charleston, SC (71)	Columbia, SC (71)	Vancouver, WA (70)	La Crosse, WI (70)
Drinks	Wilmington, NC (58)	Des Moines, IA (57)	Orlando, FL (54)	Providence, RI (51)	Boise, ID (46)
Coffee	Champaign, IL (48)	Sacramento, CA (40)	Bellingham, WA (37)	Providence, RI (34)	Tempe, AZ (32)
Arts	Santa Fe, NM (56)	Charleston, SC (42)	Grand Rapids, MI (41)	Asheville, NC (40)	Boise, ID (37)
Outdoors	Columbia, MD (67)	Columbus, OH (66)	Providence, RI (60)	St. Paul, MN (58)	Sioux Falls, SD (56)
Sights	Asheville, NC (99)	Boise, ID (97)	Santa Fe, NM (92)	Kansas City, MO (91)	Anchorage, AK (91)

At a first glance, it can be inferred that the most common venues encountered are those categorized as "Sights", while coffee places seem to be the least common venues category. Lets explore this through a bar chart showing the average number of each of these categories in our data frame (Figure 9):

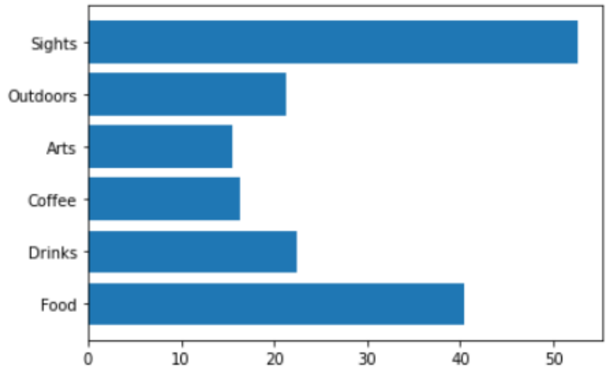


Figure 9. Average number of venues per category in Top 100 cities

As it turns out, the less common venue category is Arts (average of 15), followed by Coffee (16), Outdoors (21), Drinks (23), Food (40) and Sights (53).

We can now compare these average values to the city considered the best to live in 2020 (Fort Collins, Colorado), and those of our case study of Oklahoman cities (Oklahoma City, Tulsa and Stillwater).

Table 3. Number of venues of the average Top 100 city, the Top1 city, and the 3 studied Oklahoman cities

Venue categories	Average Top100	Fort Collins, CO	Oklahoma City, OK	Tulsa, OK	Stillwater, OK
Food	40	53	59	53	13
Drinks	23	26	33	35	5
Coffee	16	10	14	11	16
Arts	15	18	26	15	16
Outdoors	21	18	38	32	9
Sights	53	74	74	94	25
Total	168	199	244	240	84

Table 3 provides some interesting results. Oklahoma City and Tulsa have considerably more venues on almost every category (with the exception of Coffee venues, which lies somewhere around the average value, and Art venues, where Tulsa exactly matches the average). Lets dive into the results a bit more in detail.

Fort Collins, as its top ranked position in the Top 100 may have suggested, has 31 more venues than the average, with a bigger proportion of food, drinks and sights venues. On the other hand, it has a somewhat average amount of coffee, arts and outdoors venues.

Oklahoma City, with 45 more venues than Fort Collins, improves on the total number of venues of each category except for tying the number of sights venues.

Tulsa compensates having slightly fewer food venues than the capital city of Oklahoma with a few more drinks venues, and 94 sights venues that would put it on the Top 3 of the Top 100 cities. Its arts venues match those of the average Top 100 city, while the outdoors activities are better than average.

Finally, Stillwater clearly underperforms the food, drinks, outdoors and sights categories, while tying the number of coffee and arts venues of the average Top 100 city

3.3. City clustering

We now have two different methods to cluster the set of cities that remain from the Top 100 after data cleaning (87). These two methods consist of using the categories provided by Livability.com (civics, demographics, economy, education, housing, health and infrastructure), and using the number of venues within each category as obtained from FourSquare.

We will use the K-means method for both cases, and once the Top 100 cities data has been assigned to a cluster, the K-Nearest Neighbors method will be used to assign each of the 3 studied cities (Oklahoma City, Tulsa and Stillwater) to one of the clusters. It must be noted that, since we lack information about the Livability.com categories for the Oklahoman cities, we will only be able to assign them to a Venues cluster.

The code was ran multiple times to determine the most appropriate number of K-means. After evaluating K ranging from 2 to 10, it was decided to use 4 clusters. The reason behind this is that with 4 clusters, the subsequent K-Nearest Neighbor method provided the best accuracy of estimation of the Test Set for both methods. Once we have set the K-means value of K at 4, then the proper value of k for the K-NN method was evaluated in a similar manner, reaching the conclusion that the best number of neighbors to use in our analysis is for k=3. These two optimization graphs are shown in Figure 10.

It should be noted that due to the random nature of the K-Means method (the initial cluster center-points are assigned randomly), it is possible that subsequent runs of the code will result in slightly different accuracies and results.

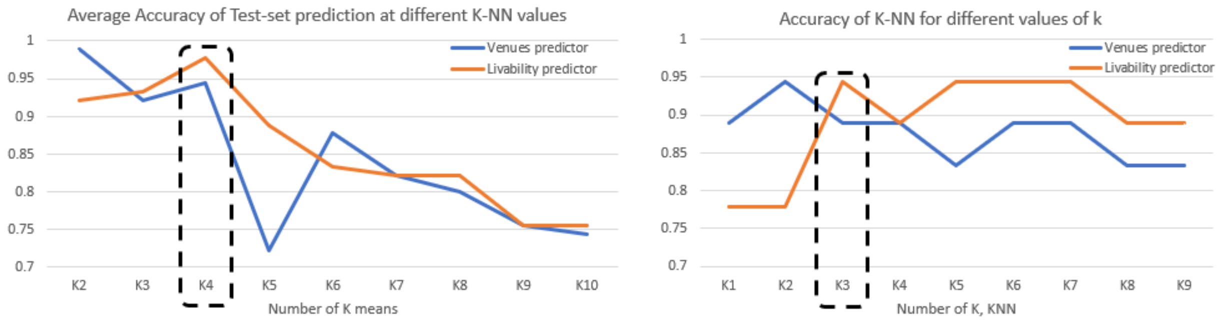


Figure 10. Optimization of k values for K-means and K-NN methods

Having set these k constants to 4 for K-Means, and 3 for K-NN, we are ready to proceed with the clustering of the dataset of cities.

Clustering by venue category:

Running the code developed to extract the number of venues of each category, we obtain 4 groups with the characteristics described in Table 4.

Table 4. Clustering of cities by number of venues of each category

Cluster #	Number	Food	Drinks	Coffee	Arts	Outdoors	Sights
0	5	28	26	36	15	13	26
1	28	49	30	15	16	22	69
2	12	65	39	24	36	45	86
3	42	29	13	12	10	15	35
Mean	-	*40*	*23*	*16*	*15*	*21*	*53*

Cluster #0 is formed by just 5 cities (Tempe, AZ, Champaign, IL, Bellingham, WA, Vancouver, WA and Des Moines, IA), the highest ranked one being Tempe #26. It contains below-average number of venues in most categories, except for drinks (+3) and coffee (+20). We will define this cluster as “Coffee cities”.

Cluster #1 contains 28 cities, 4 of which conform the Top 10 according to Livability.com (#1 Fort Collins, CO, #3 Madison, WI, #7 Fargo, ND, and #8 Durham, NC). Cities belonging to this cluster have more venues of the food (+9), drinks (+7), and sights (+16) categories, while keeping an average number of coffee, arts and outdoors venues. We will call this cluster “Sightseeing cities”.

12 cities conform cluster #2, including #6 and #10 of the Top 100 (Asheville, NC, and Columbus, OH, respectively). This category clearly overperforms the average city in every single venue category; food (+25), drinks (+16), coffee (+8), arts (+21), and sights (+33), which leads us to calling this cluster “Vibrant cities”.

Finally, the most common cluster (#3) contains 42 cities, including #5 (Rochester, MN) and #9 (Sioux Falls, SD). These cities have fewer number of venues of every single category; food (-11), drinks (-10), coffee (-4), arts (-5), outdoors (-6), sights (-18). This cluster will be defined as “Peaceful cities”.

It should be noted that some of the Top 100 cities (including #2 Ann Arbor, MI, and #4 Portland, ME) have been removed since the number of sights and food venues surpassed the 100 limit set by the FourSquare API, so the final number of cities studied using venue numbers is 87. Figure 11 displays these cities colored based on the cluster they belong to (Coffee cities in light green, Sightseeing cities in blue, Vibrant cities in red, and Peaceful cities in black). Note that Anchorage, AK has been left out of the map, but it belongs to the Sightseeing cities.

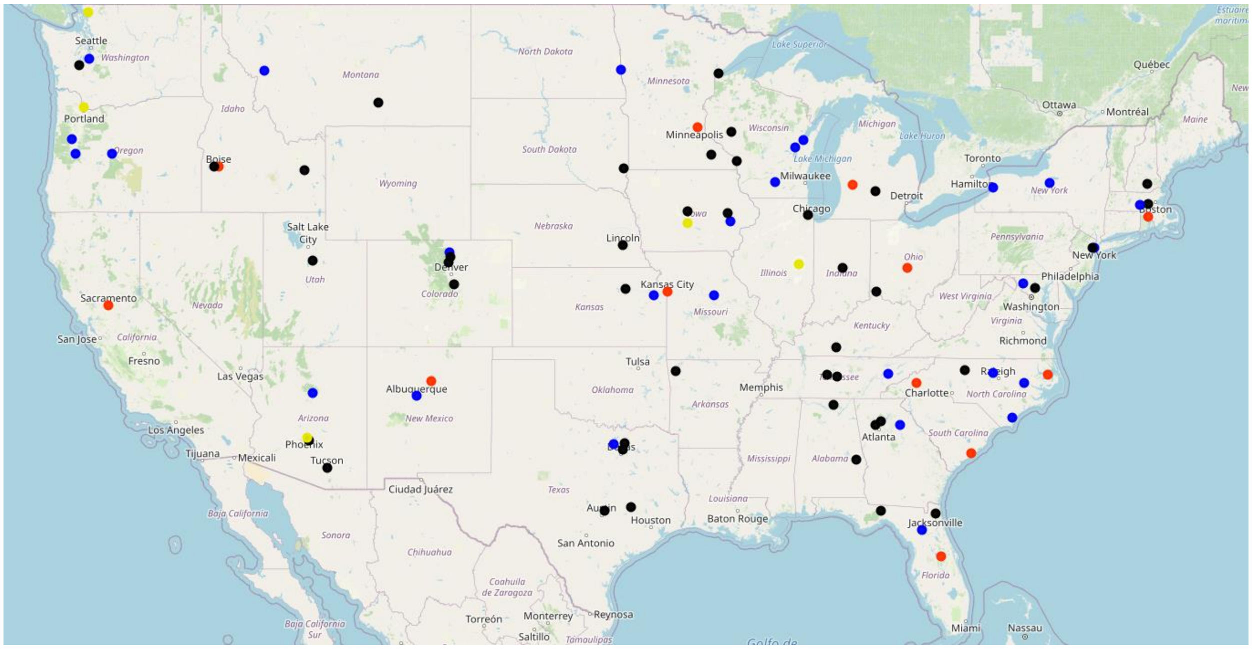


Figure 11. Map of the Top 100 cities grouped into Coffee, Sightseeing, Vibrant and Peaceful cities.

Clustering by livability score:

The other method we can use to cluster the studied cities is by taking into account the different scores provided in Livability.com. The only drawback to this is the fact that we will not be able to include our studied Oklahoman cities to these clusters, as we lack the livability score information.

Nevertheless, we could obtain some interesting information to shed some light on the algorithm used by Livability.com, so lets go ahead and carry out a second clustering using these 7 livability scores. The results are shown in Table 5.

Table 5. Clustering of cities by livability score

Clust #	Number	CI	DE	EC	ED	HE	HO	IN
0	24	52	57	56	66	45	58	48
1	10	76	56	46	52	59	45	29
2	35	60	40	56	63	65	52	56
3	18	36	66	62	56	55	44	54
Mean	-	*55*	*52*	*56*	*61*	*56*	*51*	*50*

Cluster #0 contains 24 of the 87 cities in our data frame, including #8 (Durham, NC) of the Top 100. It can be characterized by a lower than average Health score (-11), higher than average Education (+5), Demographics (+5) and Housing (+7). We will refer to cluster #0 as “High Education, Low Health”.

Cluster #1 contains just 10 cities, none of them positioned in the Top 10. These cities have an outstanding level of Civics (+21), lower than average Economy (-10) and Education (-9), and very poor Infrastructure (-21). From now on, we will call this cluster “High Civics, Low Infrastructure”.

Cluster #2 is the most common, with 35 cities belonging to it. Six of the Top 10 cities belong to this cluster (#1 Fort Collins, CO, #3 Madison, WI, #5 Rochester, MN, #6 Asheville, NC, #7 Fargo, ND, #9 Sioux Falls, SD), which could indicate that this combination of factors is favored by Livability.com’s algorithm when ranking the cities. These cities have a considerably higher Health score (+9), slightly higher Infrastructure (+6) and Civics (+5), and very low score on demographics (-12) when compared to the average. We will call this cluster “High Health, Low Demographics”.

Cluster #3 contains 18 cities, including Columbus, Ohio (ranked #10 overall). These cities score very low on Civics (-19), relatively low in Housing (-7) and Education (-5), but compensate this with a high score in Demographics (+14) and Economy (+6) when compared to the average city. We will call this cluster “High Demographics, Low civics”.

These clusters have been plotted on a US map (see Figure 12), with the “High Education Low Health” cities being plotted as light green, the “High Civics Low Infrastructure” cities plotted as blue, the “High Health Low Demographics” cities plotted in red, and the “High Demographics Low Civics” cities plotted in black. Anchorage, AK, belongs to High Health Low Demographics’ group.

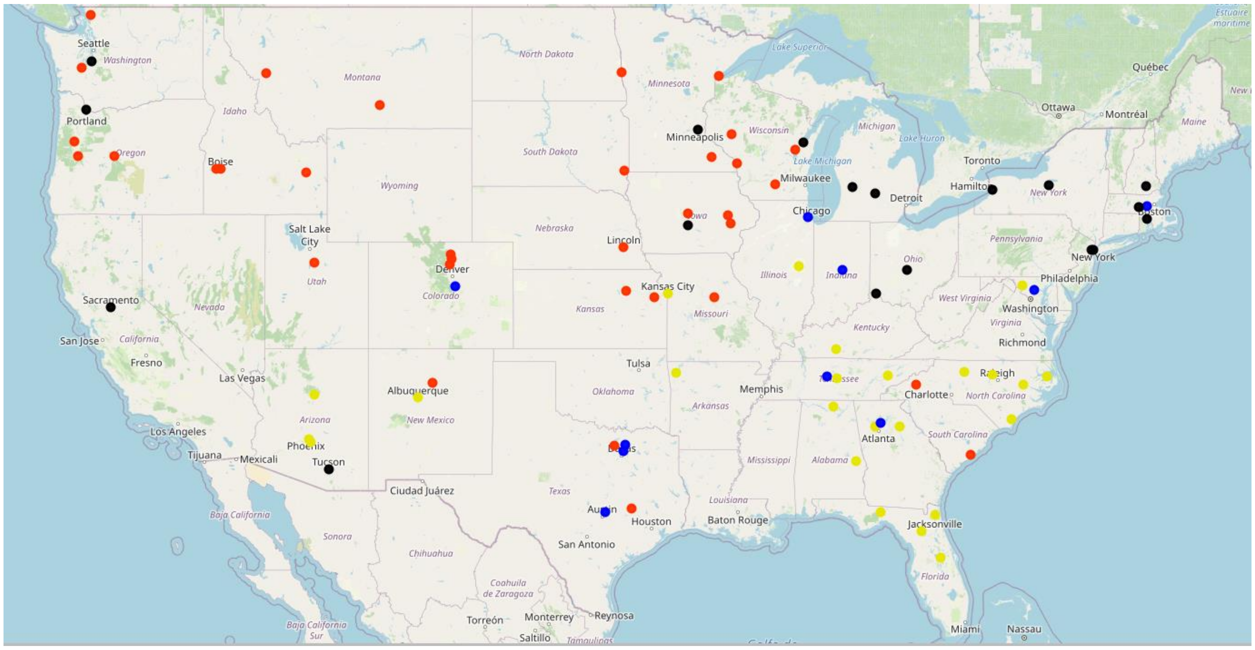


Figure 12. Map of the Top 100 cities grouped into 4 categories of Livability.com's ratings

3.4. Case Study: Oklahoman cities evaluation

As it has been made abundantly clear, no Oklahoman city has been selected in the Top 100 Cities, and our goal is to try to provide some guidance to improve the situation for three major cities of the State of Oklahoma; Oklahoma City (capital), Tulsa (second largest city) and Stillwater (home of the Oklahoma State University).

It was previously discussed in section "3.2 Venue Analysis" that both Oklahoma City and Tulsa have a larger number of venues of almost every category studied than the average city in the Top 100. But we should go one step further and see which of the other cities included in the Livability.com report are more similar to these three Oklahoman cities.

The machine learning method that will be at play here is the K-Nearest Neighbor. As we discussed earlier on, running several times the code varying the k constant, we reached the conclusion that the best value for this constant is k=3. The accuracy of the method, was tested using 20% of the already studied cities (18 cities), having used the remaining 69 for training the model. The venues model returned a train test accuracy of 0.957, and a test accuracy of 0.889. On the other hand, the livability model returned a train test accuracy of 0.928 and a test accuracy of 0.944, although this method won't be used as we don't have access to the data of livability scores of our Oklahoman cities.

Both Oklahoma City and Tulsa were assigned to Cluster #1, or Sightseeing Cities. This is a promising result, as this cluster includes 4 of the top 10 cities including the ones ranked as #1 (Fort Collins, CO) and #3 (Madison, WI).

Table 6. Comparison between venues in Oklahoma City, Tulsa and Sightseeing Cities.

	Food	Drinks	Coffee	Arts	Outdoors	Sights	Total
Oklahoma City	59	33	14	26	38	74	244
Sightseeing Cities	49	30	15	16	22	69	201
Tulsa	53	35	11	15	32	94	240

As pointed out on Table 6, the similarity between the two biggest cities of Oklahoma and the Sightseeing Cities to which they were assigned to is surprising. Both Oklahoma City and Tulsa surpass the number of Sights venues and outdoor venues considerably, and Oklahoma City has an important edge on Arts venues.

Stillwater was assigned to Cluster #3, or Peaceful Cities (alongside Top #5 Rochester, MN, and #9 Sioux Falls, SD). Table 7 shows a higher dissimilarity in this case, with Stillwater considerably underperforming in food (-16), drinks (-8), outdoors (-6) and sights (-10) venues. Their modest advantages in terms of coffee (+4) and arts (+6) venues aren't enough to cut the 30 venues deficit when compared to the average Peaceful City.

Table 7. Comparison between venues in Stillwater and Peaceful Cities.

	Food	Drinks	Coffee	Arts	Outdoors	Sights	Total
Stillwater	13	5	16	16	9	25	84
Peaceful Cities	29	13	12	10	15	35	114

This analysis concludes with determining the Oklahoman cities' most similar Top 100 cities using cosine similarity. The approach here is to compare the venues vector of the studied Oklahoman cities to the list of Top 100 cities for which data of venues is available. The results are displayed in Tables 8 through 10.

Table 8. Cosine Similarity matrix for Top 5 matches of Oklahoma City.

	Knoxville, TN	Buffalo, NY	Tacoma, WA	Charleston, SC	Boise, ID	Average	Oklahoma City
Food	44	49	53	71	69	57	59
Drinks	23	34	33	30	46	33	33
Coffee	11	16	19	21	27	19	14
Arts	16	19	20	42	37	27	26
Outdoors	30	30	37	46	37	36	38
Sights	57	68	80	88	97	78	74
Totals	181	216	242	298	313	250	244
Cluster	2	2	2	1	1	1.6	1
Ranking	96	89	75	48	35	69	NA
LIV	599	615	634	650	657	631	NA
Similarity	99.86	99.59	99.44	99.34	99.27	99.50	NA

Oklahoma City has been paired with cities ranging on the lower end of the Top 100, but the similarity between the average number of venues in these 5 cities and that of Oklahoma is staggering. The highest difference between the number of venues is just 5 points in Coffee venues. Two out of these 5 cities belong to the assigned cluster of Sightseeing Cities, but it's worth noting that the three most similar cities belong to the Vibrant Cities. If the city officials of Oklahoma City managed to make the list of Top 100 Cities in 2021, the average LIV score of 631 suggest it would do so in position #78.

Table 9. Cosine Similarity matrix for Top 5 matches of Tulsa.

	Greenville, SC	Missoula, MT	Appleton, WI	Fargo, ND	Gainesville, FL	Average	Tulsa
Food	51	41	36	45	37	42	53
Drinks	31	31	29	34	30	31	35
Coffee	12	16	9	11	9	11	11
Arts	11	13	13	15	13	13	15
Outdoors	26	24	17	19	25	22	32
Sights	77	74	69	86	61	73	94
Totals	208	199	173	210	175	192	240
Cluster	2	2	2	2	2	2	1
Ranking	49	20	35	7	67	36	NA
LIV	647	671	655	701	637	662	NA
Similarity	99.71	99.65	99.55	99.52	99.48	99.58	NA

Tulsa differs more from the average of the 5 cities it's been paired to. The main differences come in the Food (+11 in favor of Tulsa), Outdoors (+10) and especially Sights (+21) venues. These cities have an overall better ranking than those of Oklahoma City (average of #36), with Fargo being in the Top 10, and an average livability score of 662. This score would put Tulsa (in the event that it was included in 2021s Top 100 Cities report) at a #32 ranking. Interestingly, the most similar cities to Tulsa according to the cosine similarity all belong to the Vibrant Cities, but Tulsa has been assigned to the Sightseeing Cities.

Table 10. Cosine Similarity matrix for Top 5 matches of Stillwater.

	Bowling Green, KY	Duluth, MN	Santa Fe, NM	Pflugerville, TX	Huntsville, AL	Average	Stillwater
Food	20	24	64	7	18	27	13
Drinks	11	11	27	5	19	15	5
Coffee	13	11	27	5	19	15	16
Arts	21	20	56	7	15	24	16
Outdoors	7	14	32	6	18	15	9
Sights	27	41	92	16	47	45	25
Totals	99	121	298	46	136	141	84
Cluster	0	0	1	0	0	0	3
Ranking	83	75	88	65	95	81	NA
LIV	616	636	606	638	591	616	NA
Similarity	97.30	96.30	95.96	95.66	94.89	96.02	NA

Stillwater most similar cities are even more divergent than those obtained for Tulsa. Saving the similitude in Coffee venues (+1 for Stillwater) and Outdoor venues (-6), the rest of the categories differ by -8 (Arts), -10 (Drinks), -14 (Food) and -20 points (Sights) from the average most similar city. Based on the average Livability Ranking, Stillwater would have 616 points (#88).

4. Discussion

4.1. Results summary

The combination of available data in Livability's Top 100 Cities to Live in the US, along with data of number of venues in city centers extracted using the FourSquare API has allowed us to extract valuable information that can be of help to city officials of Oklahoman cities to improve the livelihood of their citizens while making their cities more appealing to other fellow Americans.

While the results of this study have been disclosed along with the methodology in section 3 of this report, I would like to summarize the most interesting findings here.

Probably the first takeaway resulting from this study is that both Oklahoma City and Tulsa have a lot of potential based on the number of venues of different categories available to the public. Oklahoma City in particular offers a great number of arts venues, while Tulsa's number of Sights venues puts it in the Top 3 of the Best Cities using that same metric. Stillwater, on the other hand, has a lack of venues of all types that makes it harder for them to aspire to entering the group of Top Cities.

Table 11. Ranking of Oklahoma City, Tulsa and Stillwater in each of the venues' categories.

	Food	Drinks	Coffee	Arts	Outdoors	Sights	Total	Avg
Oklahoma City	17	17	47	13	11	24	**15**	**22**
Tulsa	23	10	61	33	19	3	**17**	**25**
Stillwater	81	82	35	27	68	75	**83**	**61**

When assigning the cities to a specific cluster, both Oklahoma City and Tulsa joined the group of Sightseeing Cities, which has an overwhelming representation of Top 10 cities, such as Fort Collins, Colorado, Madison, Wisconsin, Fargo, North Dakota, and Durham, North Carolina. The city officials may well consider study the specifics of these cities to try to come up with similar strategies to achieve excellence.

Stillwater joined the group of Peaceful Cities. Rochester, Minnesota, and Sioux Falls, South Dakota, are the top ranked members of this group, which indicates that having fewer venues in their city centers can be used to their advantage.

When studying the specific makeup of venues of the Top 100 cities, the 5 most similar cities to each of our 3 Oklahoman cities were determined using cosine similarity. In this particular analysis, Tulsa seems to be the city with the highest promise to one day enter the Top 100 Cities in the US. This is so because these 5 most similar cities have an average livability score of 662, which would represent a ranking of #32. It would be of special interest to conduct further studies on the city of Fargo, North Dakota, as it holds a ranking of #7 and has a cosine similarity of 99.52% with Tulsa.

Oklahoma City performed slightly poorer in this case, with Boise, Idaho, being the best qualified city out of the 5 most similar (#35). The average livability score of 631 would result in an approximate #78 position in the ranking.

Stillwater has been matched with cities on the end of the ranking, with Pflugerville, Texas, being the best qualified at #65. The average livability score of these 5 cities, 616, would place a hypothetical Top-100 Stillwater at the ranking position of #88.

4.2. Future research

This study had an important limitation imposed by the little number of venues that FourSquare API allowed us to collect in each query. Therefore, we had to limit the scope of the study to the immediate 1.8 km^2 around the city center (or a 750m radius). This can result in an underestimation of venues of less densely developed cities, or the overestimation of those cities that consist of a very active city center englobed in more peaceful suburbs. Future research should be conducted to avoid this limitation, via running multiple queries per city, or using a less stringent API.

The relationship between our venue data and the some or all of the scores assigned by Livability.com could be studied, as that could help determine whether the actions taken by the Oklahoman cities are going in the right direction in order to make the Top 100 list next year.

5. Conclusion

The two most populous cities of Oklahoma, Oklahoma City and Tulsa, are at a great position to become the first representative of the State to join the Top 100 Cities ranking put together by Livability.com. Based on the large number of different categories of the venues that make up their city centers, they show similitude with some of the best cities in the US. This study suggests that the city officials study more in depth the leading cities of what we have denominated Sightseeing Cities (i.e. Fort Collins, Madison, both belonging to the Top 3) and how can they adopt some of their success strategies to improve the lives of their citizens.

On the other hand, Stillwater may benefit from following the example set by cities with a smaller number of venues (like Rochester, Minnesota, and Sioux Falls, South Dakota) but that still are ranked very highly in the Top 100 cities study from Livability.com.