**Clustering US Cities Based on Crime Statistics**

**Cemile Sahin**

**for**

**IBM Data Science Professional Certificate on Coursera**

**1. Introduction**

One of the important considerations for most people to make a city home is the safety record of the city. Everything else being the same, people would prefer cities with lower criminal profiles to those with higher criminal profiles. My internet search for "important factors when choosing a city to live" features crime rate in all the entries on the first page of results.

* Where Should I Live? 14 Important Factors When Deciding the Best Place to Live: <https://www.moneycrashers.com/where-should-i-live-decide-best-places/>
* The Most Important Factors in Choosing a Place to Live: <https://medium.com/@jinlinh/the-most-important-factors-in-choosing-a-place-to-live-39f39f2fc7ed>
* 11 Essential Questions to Answer “Where Should I Live?”: <https://www.openlistings.com/blog/11-factors-to-help-you-decide-where-to-live/>
* Where Should I Live? 6 Factors to Consider When Choosing a Neighborhood: <https://www.tchabitat.org/blog/how-to-choose-a-neighborhood>
* How to Choose Where to Live: <https://www.thespruce.com/find-best-place-to-live-2435883>
* Factors to Consider When Deciding Where to Live: <https://www.homes.com/blog/2017/07/factors-consider-deciding-live/>
* 7 Things to Consider When Choosing a Neighborhood: https://realtybiznews.com/7-things-to-consider-when-choosing-a-neighborhood/98753418/

In this report, I will look into the crime statistics for the largest 100 US cities, look at the correlation between different types of crimes, and cluster them based on their crime statistics. This will help someone who is considering moving from one city to another to compare the cities in terms of their safety record.

Crime is, of course, not the only criterion people consider in choosing which city to move. Amenities available in the city is also an important consideration. In the second part of the report, I will use the Foursquare data to cluster cities based on the venues available in these cities. I will also investigate if there is any relation between the clusters based on crime statistics and clusters based the venues.

**2. Data**

In this report, I use two main sources of data.

1. Crime statistics for the largest 100 US cities

2. Foursquare location data for the same 100 US cities

I obtained the crime statistics from a Wikipedia page which is based on FBI Uniform Crime Reports statistics from 2017. <https://en.wikipedia.org/wiki/List_of_United_States_cities_by_crime_rate>

Number of crimes committed per 100,000 people are listed in three major categories: Violent Crime, Property Crime, and Arson. Under Violent Crime, subcategories are Murder, Rape, Robbery, and Aggravated Assault. Under Property Crime, the listed crimes are Theft, Burglary, and Motor vehicle Theft. Arson is the only crime listed under Arson. The table also includes the population data.

I will discuss the exploratory data analysis I did on this data set in the Methodology section.

I also used the Foursquare location data for the 100 cities listed in the Wikipedia page. I extracted information about the venue name and the venue category.

The purpose of using Foursquare venue data is twofold: i) to see if there is any connection between crime statistic clusters and venue clusters, and ii) to look at the top venues available in the lowest crime cities if we can choose among them based on the venues available.

**3. Methodology**

**3.1 Crime Data Exploration**

I extract my main source of data from the Wikipedia page that provides the crime statistics table for the 100 largest US cities.

Arguably, different types of crimes are correlated with each other. If a city has a low number of murders, one can assume that it also has a low number of thefts, for instance. Using the corr().mean() method, I check if that is indeed the case.

**Average Correlation with Other Variables**

|  |
| --- |
| Population 0.049337 |
| Violent crime - Total 0.673517 |
| Murder/nonnegligent manslaughter 0.553335 |
| Rape 0.489713 |
| Robbery 0.611428 |
| Aggravated assault 0.627215 |
| Property crime - Total 0.596165 |
| Burglary 0.595968 |
| Larceny-theft 0.475138 |
| Motor vehicle theft 0.554206 |
| Arson Arson 0.502066 |

Looking at the correlation between different types of crimes, I notice that Rape has a low correlation with other types of crimes. One reason could be the fact that most rape is not related to the safety of the city as 80% of rape is committed by someone the victim knows.

Perpetrators of Sexual Violence Often Know the Victim 8 out of 10 rapes are committed by someone known to the victim. 'https://www.rainn.org/statistics/perpetrators-sexual-violence'

As such, it is reasonable to drop rape statistics in evaluating safety of a city.

Another question that comes to mind is the difference between burglary, robbery, and larceny (theft). A web search provides the following explanations as to the difference between burglary and larceny and the difference between robbery and larceny. In short, robbery involves a threat of violence, and burglary involves entering a building and violating the privacy of one's home.

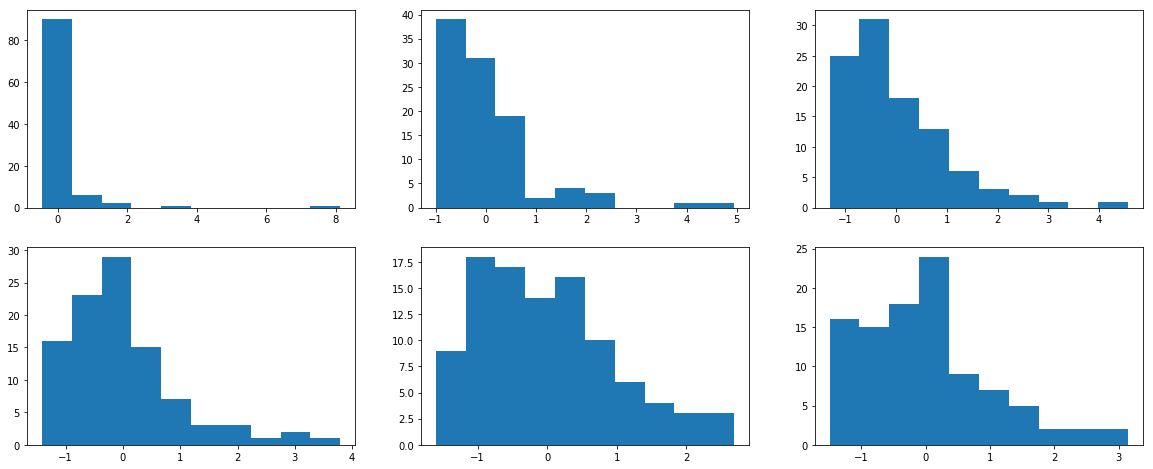
* 'Burglary (entering a building with the intent to commit a crime inside) and larceny (theft) are two different crimes, although burglaries are often committed for the purpose of theft. Burglary laws are intended to protect the sanctity and privacy of people’s homes and other structures. Laws against larceny protect personal property.' <https://www.criminaldefenselawyer.com/resources/whats-difference-between-burglary-and-larceny.htm>
* "Robbery and larceny are both categorized as theft crimes. However, they are very different from one another. Robbery is generally defined as: the unauthorized taking of another’s personal property from their presence, with the use of force or threats of serious bodily injury, and with intent to permanently deprive them of the object. Larceny is defined as: the unauthorized taking of another person’s property with the intent to permanently deprive them of the use of the property. So, the basic distinction between robbery and larceny is that robbery involves the use of force, whereas larceny doesn’t." <https://www.legalmatch.com/law-library/article/robbery-vs-larceny-charges.html>

Based on this information, I decide to drop Larceny/Theft as well as Rape and Arson. The simpler data frame contains the following variables: Murder, Robbery, Assault, Burglary, and MV Theft.

When I compare the mean values for different variables, I notice that there is large variation. The mean number of murders is less than 12 while the mean number of burglaries is 647. This will give much more weight to Burglary values using any statistical methods. Therefore, I decide to standardize all the numerical values in the data frame.

It is worth to point out that Population is negatively correlated with Murder, Burglary, and MV Theft. Although it is positively correlated with Robbery and Assault, the sizes are smaller. This suggests that larger cities being more dangerous is a myth.

A plot of sorted Population and crime statistics show that safest places are about one standard deviation below the mean in all crime categories. Most crimes, and especially murder, are a major problem in a handful of cities. Based on these we may expect to get one small cluster with a huge crime problem.



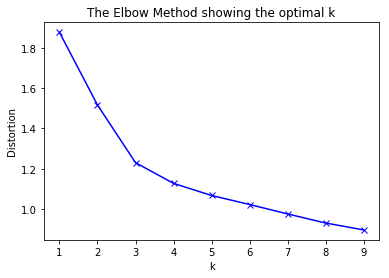
**3.2 Crime Clusters**

I cluster 100 largest US cities based on crime data using the kmeans clustering library.

First, I create a features vector, X, by dropping State, City, and Population. I drop Population because I do not want the clusters to be based on population, only on crime statistics.

Before running the kmeans clustering algorithm, I first use the elbow method to determine the appropriate number of clusters.

The elbow method plots distortion which depends on the Euclidian distance between each observation (here US cities) and the cluster center where the distance is a measure of the dissimilarity.



The elbow method indicates that k=3 is a reasonable choice for the number of clusters. The decline in distortion as we go from 3 clusters to 4 clusters is smaller than that as we go from 2 clusters to 3 clusters. An alternative is to choose k=7, but it will be difficult to interpret results with such fine clustering. Therefore, I will use k=3.

Then, I carry out clustering using 3 clusters. I discuss the results in Section 4.

**3.3 Foursquare Location Data Exploration**

In this section, following the Week 3 lab example, I will obtain the location data for the largest 100 US cities. I will then cluster them based on the venues available choosing the number of clusters using the elbow method. Finally, I will check if there is any relation between clusters based on crime statistics and clusters based on venue categories.

Now that the dataframe has the coordinates, I can use it to extract information from Foursquare.

I start with defining a function to extract the nearby venues. I limit the number of venues to 100 and set the radius for the search to 5000 meters.

Then, I call on the function with my dataframe, and call the resulting dataframe 'venues.'

There are 9992 venues in the venues dataframe only 8 fewer than the full 10,000 if all cities had 100, the limit set by the search.

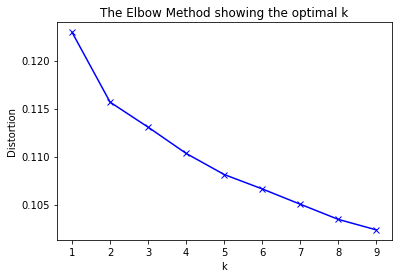
There are 397 unique venue categories.

For the features matrix that will be used for clustering, I create dummies for each of these categories.

Next, I take the mean of each dummy for each city using the groupby() method. Next, I define a function to obtain the most common venues in each city.

**3.4 Venue Category Clusters**[**¶**](http://localhost:8888/notebooks/Documents/LearnPython/IBM/Course9/Crime%20in%20US%20Cities.ipynb#3.4-Venue-Category-Clusters)

I use the elbow method for some visual cues to the appropriate number of clusters.



The Elbow Method suggests that k=2 is a reasonable choice for the number of clusters since the decrease in distortion is much lower as we go from 2 clusters to 3 clusters as compared to going from no clustering to 2 clusters. Furthermore, the distortion is quite low to begin with.

**4. Results**

In this section, I will first discuss the results from the crime statistics analysis. Then, I will turn to the results from the Foursquare location data.

**4.1 Clusters Based on Crime Statistics**

This is a summary of the three clusters:

|  |  |  |  |
| --- | --- | --- | --- |
| Safety Cluster | 0 | 1 | 2 |
| Number of cities | 38 | 11 | 51 |
| Population | -0.023131 | -0.196266 | 0.059567 |
| Murder | -0.623741 | 1.836600 | 0.068619 |
| Robbery | -0.782984 | 1.991968 | 0.153760 |
| Assault | -0.784848 | 1.897106 | 0.175609 |
| Burglary | -0.868993 | 1.561164 | 0.310763 |
| MV Theft | -0.776015 | 1.654719 | 0.221307 |

The first cluster consists of 38 cities. The average city size is about average. All crime statistics are below the mean values, from .62 standard deviations below the mean for Murder to .86 standard deviations below the mean for Burglary. I will call this group "Safe Cities."

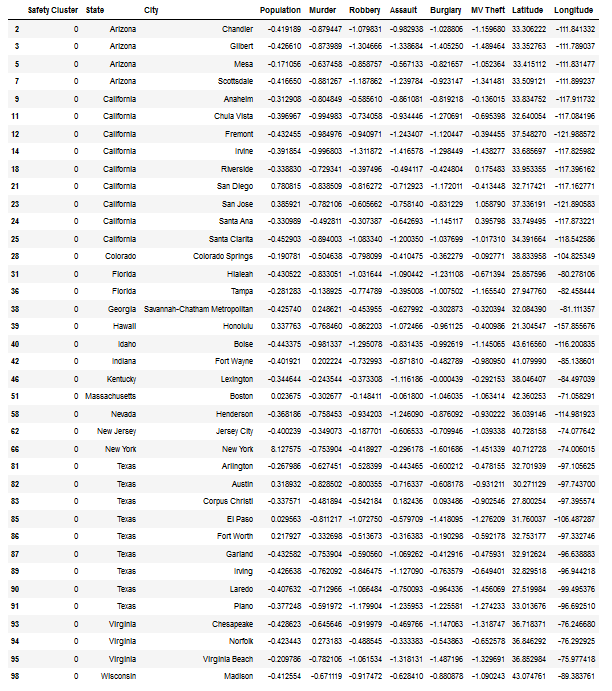
The second cluster consists of 11 cities. Their population is slightly lower than the average, and their crime statistics are significantly higher: from 1.56 standard deviations above the mean for Burglary to 1.99 standard deviations above the mean for Robbery. I will call this group "Dangerous Cities."

The third cluster consists of 51 cities, more than half of the cities. Their average is slightly higher than the average population, and their average crime rate is slightly higher than the average. I will call this group "Larger Cities."

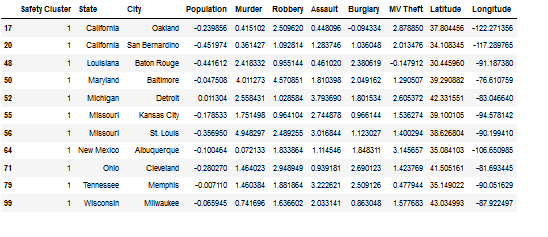
Based on this information, I would recommend someone concerned about average city safety to move to a city within the first group and avoid cities in the second group.

Which are the cities in each group?

Cluster 1:



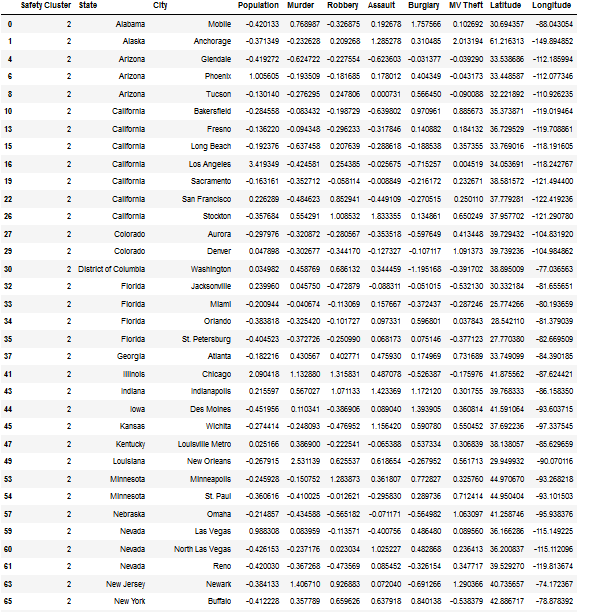
Cluster 2

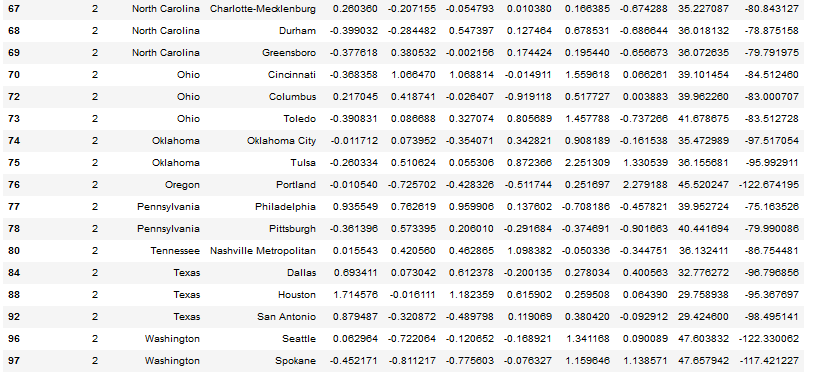


When we look at the poverty rates, we notice that most of the cities in the second group has some of the highest poverty rates in the country. For instance, the poverty rate in Detroit, Michigan is 39.3%, in Cleveland, Ohio, it is 35.4%, and in San Bernardino, CA it is 32.4%. Compare this to the poverty rate in Fremont, CA is 6%, in Gilbert, Arizona it is 6.7%, and in Plano, Texas, it is 7.7%. Not surprisingly, all these cities are in the Safe Cities group.

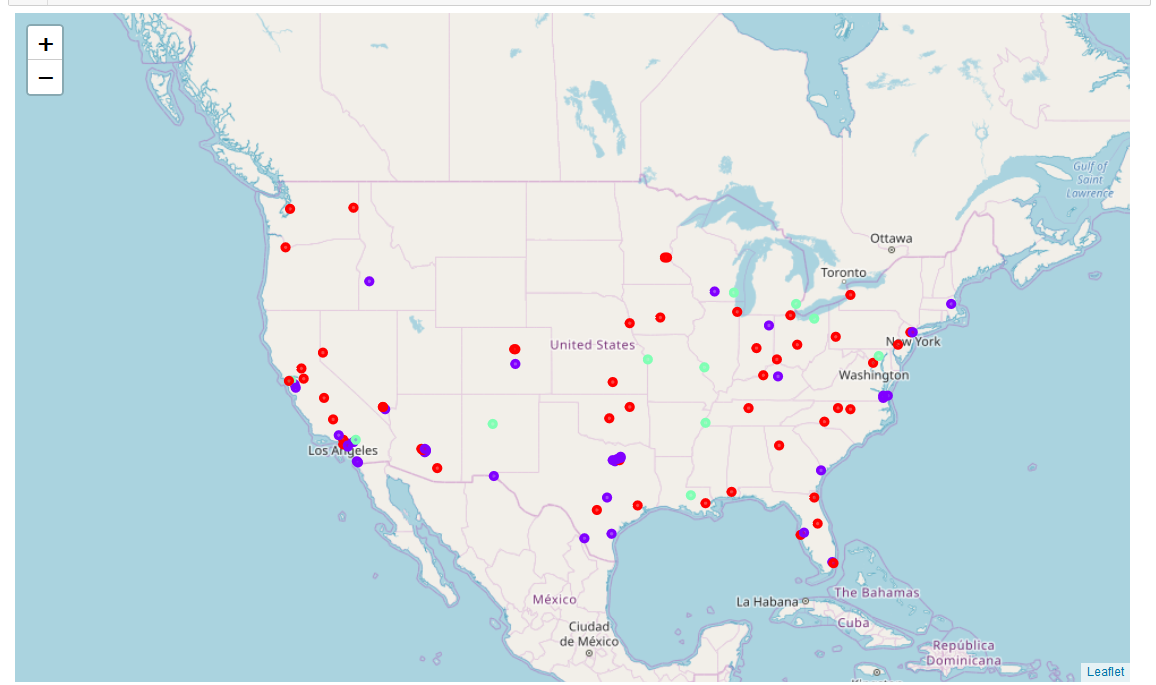
Source: 'https://www.indexmundi.com/facts/united-states/quick-facts/cities/rank/percent-of-people-of-all-ages-in-poverty'

Cluster 3





Now, we can see which cities are included in each group on a map.



The purple points are the Safe Cities, green points are the Dangerous Cities, and the red points are the Larger Cities.

**4.3 Clusters Based on Venue Categories**

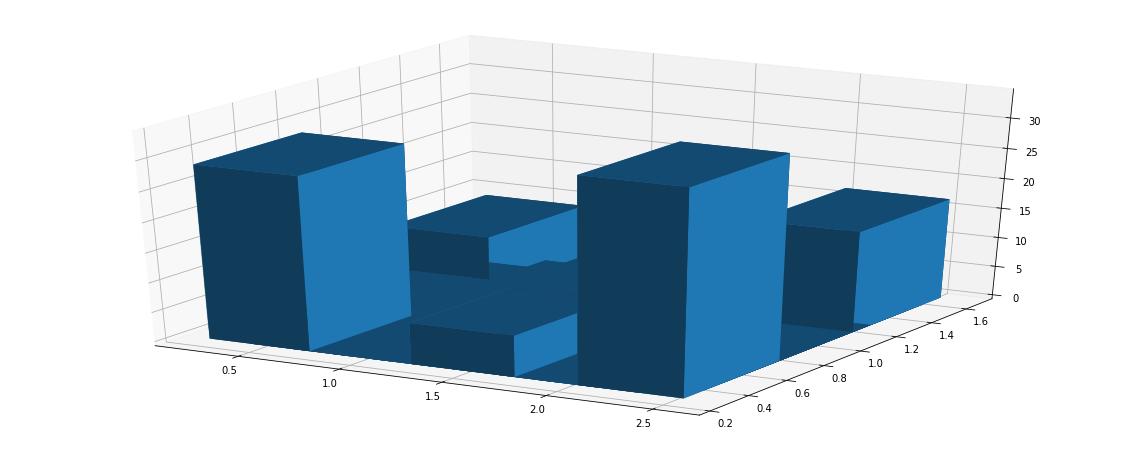
As with the clusters based on crime statistics, I start with looking into summary information about each cluster. The first cluster consists of 70 cities and the second one consists of 30 cities.

The first cluster has a slightly higher population than the average and slightly lower crime rates across all crime categories.

The second cluster has a slightly smaller population than the average and its crime rates are slightly higher than the average.

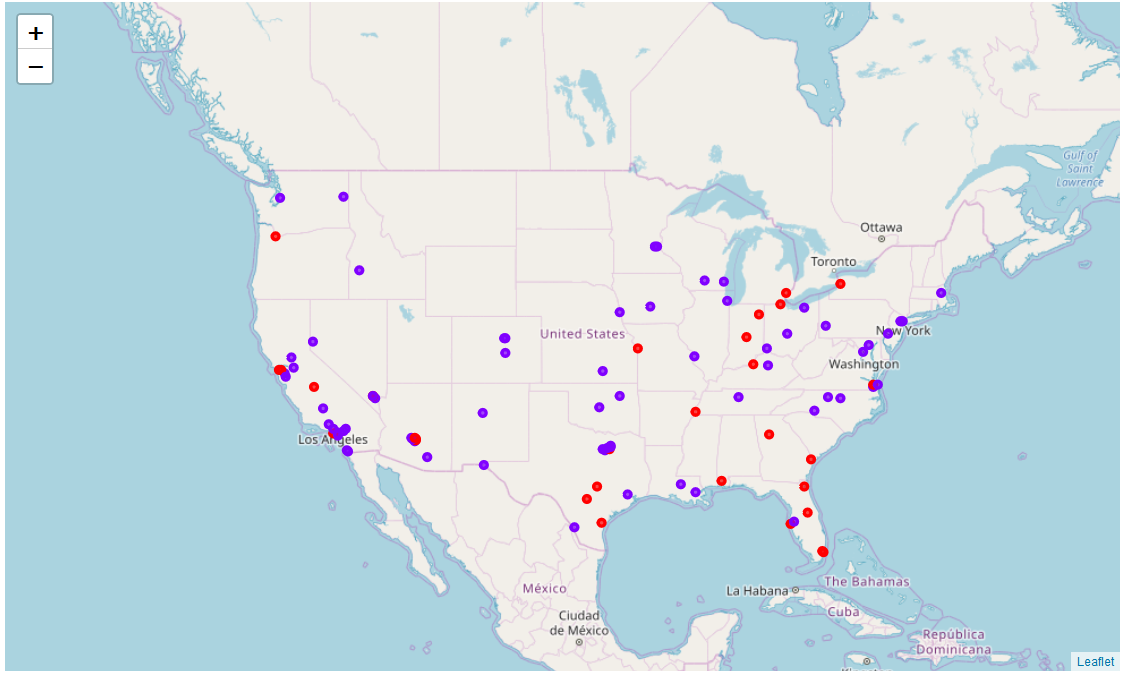
Scanning through the common venue categories does not offer much difference.

We can try to visualize the relation between the two types of venues using a 3-D plot.



There does not seem to be any pattern.

**4.4 Mapping Clusters Based on Venues**



There is not much of a geographic pattern either.

**5. Discussion**

Results show that although it is possible to have a meaningful clustering based on crime statistics, it is not the case for the venue categories. These are the largest cities in the US, and they offer a variety of venues to their large number of residents. The venues are not related to the crime statistics in cities, and these cities are not that different from each other.

It is possible to see that from the Elbow Method plots as well. Even with a single cluster, the distortion is 0.1225 while the distortion at a single cluster with crime statistic is around 1.9.

**6. Conclusions**

This report shows that some US cities have significantly higher crime statistics than others. In most cities, crime is local, and it is possible to live in low-crime areas even in high crime cities. But for someone who prefers to live in a low crime city, there are 11 cities to be avoided.

On the other hand, largest 100 US cities do not differ much from each other in terms of the venues they offer. Furthermore, there is not much of a relation between the crime statistics and the type of venues common in a city. Hence, there is no need to focus on venues as one decides where to live in the US assuming that small towns are not desired/not an option.