IBM Data Science Capstone Project
Does post-secondary education affect the rate of crime in cities?

An investigation of the relation between the increase in the number of academic centers and libraries and a decrease in the rate of crime in American cities.

Iman Soltani

1. Introduction (Business Problem)

Everyday many different types of crimes are committed in the cities. Among the developed countries the United States suffers from one of the highest crime rates. In fact, in 2020 the crime index of the United States has been determined as 47.7. This is although the rate of crime has decreased significantly in the US in the past few years. In 2012 United states was considered as one of the creepiest countries in the world, with a crime index of 64.93 (1).

This high rate of crime can be considered as a major issue threatening safety in American cities. The crimes in the cities are divided into two main groups of violent crime and property crime. A violent crime according to the definition is considered as where the offender uses or intimidates to use force on the victim. Such violence can be objective like in the case of rape or murder, or as a form of coercion. Violent crime range extends from harassment to homicide, and not all of them involve using weapons. Some of the more well-known violent crime types include robbery, hijacking, carjacking, rape, kidnapping, shooting, torture, … (2).

On the other hand, Property crime as its name points to usually deals with private property, and it is committed to obtaining money or any other similar benefit. The more known types of property crimes include theft, shoplifting, vandalism, larceny, and burglary (3).

Many problems could be argued as the possible reasons behind the high rate of crime in American cities. The economic issues that the people under the poverty line would struggle with can be considered as the major motivation. Also, gun freedom is considered as one of the reasons behind the high rate of violent crime in these cities. However, it is commonly believed that an increase in the level of education of people may play a major effect on decreasing the cities' crime rates. So, in this study, it is endeavored to scrutinize the existence of such a correlation. It should be added that even a possible correlation may not point to the existence of causation.

To measure the level of education in the cities, one may consider the number of academic and educational institutions per capita as a possible measure of the education level in a city. To make

it more specific we can account for the number of pos-secondary academic institutions as a measure of the education level of people in the vicinity. Although many people may leave that city after graduation intendedly or unwantedly to pursue their careers still it can be assumed that most of the people would remain in the same city upon graduation.

2. Data

2.1. Crime rate in American cities

To determine the rate of crime in American cities, a Wikipedia article on the list of a selected United States cities by crime rates, was used. In this article, there is a table containing about 90 most populous American cities along with their crime rates per 100,000 people based on Federal Bureau of Investigation (FBI) Uniform Crime Reports (UCR) statistics from 2017. Both violent crime data including murder and nonnegligent manslaughter, rape, robbery, aggravated assault, as well as property crime including burglary, larceny-theft, motor vehicle theft, and arson were included. The FBI has used the 2009 population estimate in this table (4).

2.2. Geographic coordinates of the selected most populous American cities.

In order to determine the geographic coordinates of the selected most populous American cities, we used the geopy.geocoder library to convert the name of the cities into their corresponding coordinates.

2.3. Number of post-secondary schools in American cities

To get detailed information about the cities post-secondary schools we used Foursquare, which is a location data provider website. So, we constructed some URL to make Foursquare API calls to search the cities for universities and colleges. Then, we received the json file of the corresponding detailed information upon calling get request.

3. Methodology

3.1. Information scrapping from the crime data table in American cities.

We imported Pandas and NumPy libraries, and we used read_html function from Pandas library to read the webpage table data into a data frame. We performed data cleaning steps on the table like dropping the rows where the total crime values were missing using the dropna method.

```
Cdf1.dropna(subset=['Total Violent Crime','Total Property crime'], axis=0)
```

The numbers in the cities were removed using str.replace method as the following.

```
Cdf2['City'] = Cdf2['City'].str.replace('\d+', '')
```

3.2. Determining the geographic coordinate of the cities.

First, we added a column with the name of both city and state to avoid any mistake in case of the cities with similar names.

```
Cdf3['City, State']=Cdf3[['City','State']].apply(lambda x: ', '.join(x[x.notnull()]), axis = 1)
```

Then, we used geopy.geocoder and imported the Nominatim function to do the geocoding of the cities' names into geographic coordinates. To define an instance of the geocoder, we need to define a user_agent. We will name our agent us_explorer, as shown below.

```
geolocator = Nominatim(user_agent="us_explorer")
for i in range(Cdf3.shape[0]):
    location = geolocator.geocode(Cdf3.loc[i, 'City, State'])
    Cdf3.loc[i, 'Latitude'],  Cdf3.loc[i, 'Longitude'] = location.latitude, location.longitude
```

3.3. Creating a map of American cities using Folium library.

To plot the map we used matplotlib library with cm and colors modules. Also, it was needed to install and import the folium graphical map rendering library. Again geopy geocoder was used to determine the geographic coordinate of the United States.
Then, a map of the United States with the selected cities superimposed on top was created.

```
# create a map of United States using latitude and longitude values
map_USA = folium.Map(location=[latitude, longitude], zoom_start=3)

# add markers to map
for lat, lng, city, state in zip(Cdf3['Latitude'], Cdf3['Longitude'], Cdf3['City'], Cdf3['State']):
    label = '{}, {}'.format(City, State)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_USA)

map_USA
```

3.4. Exploring the American cities using Foursquare API

The Foursquare API calls was used to explore the selected American cities and cluster them. Initially, we registered with Foursquare and received the client ID and Secret needed to call Foursquare API.

```
CLIENT_ID =     # your Foursquare ID
CLIENT_SECRET =   # your Foursquare Secret
VERSION = '20201227' # Foursquare API version
LIMIT = 1000 # A default Foursquare API limit value is 100
```

Then, we constructed the appropriate URL to make a Foursquare API call, upon which we could explore the universities and colleges in the selected American cities. Then a get request was called to receive a json file containing detailed information about the selected cities academic institutions. We began this process for only one sample city Mobile, Alabama to practice its viability and how to interpret the received json file data, which is in the form of a dictionary, with different keys and their corresponding values. So, we picked the appropriate data in the json file and pursued data cleaning to make an appropriate data-frame containing the required data. Accordingly, we picked the relevant data columns. Since the relevant information, which is the type (Category) of the venues, was in the name key of the categories column. So, we defined a function to extract the value of the name from the column. Finally, we chose the venue categories to exclude the venues that are not in fact universities. After that we felt satisfied with the acquired and cleaned data, we do the same procedure for all of the selected cities, using a for loop to automatize it, like the following.

```
def getSchools(names, latitudes, longitudes, radius=5000):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url =
'https://api.foursquare.com/v2/venues/search?&query=University+College&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID, CLIENT_SECRET, VERSION, lat, lng, radius, LIMIT)

        # make the GET request
        results = requests.get(url).json()['response']['venues']

        # return only relevant information for each nearby venue
```

```
    for v in results:
        try:
            venues_list.append([(
                name, lat, lng,
                v['name'],
                v['location']['lat'],
                v['location']['lng'],
                v['categories'][0]['name'])])
        except:
            pass

    city_colleges = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    city_colleges.columns = ['City',
                'City Latitude',
                'City Longitude',
                'College',
                'College Latitude',
                'College Longitude',
                'College Category']

    return(city_colleges)
```

Then we wrote a code to run the above function on each city and create a new dataframe called colleges.

```
colleges = getSchools(names=Cdf3['City'],
                      latitudes=Cdf3['Latitude'],
                      longitudes=Cdf3['Longitude']
                     )
```

In the category college category column there were still many irrelevant names, considering our goal. They included 1.Elementary Schools, 2.High School, 3.Medical School (which is a school of an already listed university), College Administrative Building, College Lab, College Gym. Therefore, we picked the venues with the name ‘University’ included in their category.

```
colleges3=colleges2[colleges2['College Category'].str.contains('University')]
```

The returned academic centers unique categories were determined using the following code.

```
print('There are {} uniques categories.'.format(len(colleges4['College Category'].unique())))
```

So, 3 categories are returned as College & University, General College & University, and University. In order to separate the selected universities and colleges into those 3 categories, we created the college_onehot, using the get_dummies function from the Pandas library.

```
colleges_onehot = pd.get_dummies(colleges4[['College Category']], prefix="", prefix_sep="")
```

We grouped the academic centers for each city.

```
city_grouped1= colleges_onehot.groupby('City').sum().reset_index()
```

Then, to have the number of academic centers per capita, we divided their numbers by the population for each city, using the following code.

```
master_df['University per capita']=master_df['University']/master_df['Population']

master_df['College & University per capita']=master_df['College &
University']/master_df['Population']

master_df['General College & University per capita']=master_df['General College &
University']/master_df['Population']
```

Finally, we cleaned the table one more before clustering to drop the unpopulated data rows.

3.5. Clustering cities based on their post-secondary schools

We performed clustering, which is an unsupervised machine learning method, to group the cities based on their post-secondary schools according to our defined categories. Therefore, we used KMeans algorithm to do the clustering of the cities.
So, initially, we imported the KMeans module from sklearn.cluster library.

```
from sklearn.cluster import KMeans
```

Since, we wanted to minimize the number of possible clusters, to group the cities academic institutions in only 3 levels, we chose the cluster number of 3.

```
 # set number of clusters
kclusters = 3
```

```
cities_edu_clusters = master_df2[['University per capita', 'College & University per
capita','General College & University per capita']]

# run k-means clustering
#creating the KMeans clustering object and fitting it to the real data in one line
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(cities_edu_clusters)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

3.6. Clustering cities based on their total crime rates

Similar with the procedure used to cluster the cities based on their universities, we used KMeans algorithm to cluster the cities based on their total violent crime rates and total property crime rates into 3 groups.

```
# set number of clusters
kclusters = 3

cities_crime_clusters = master_df2[['Total Violent Crime', 'Total Property crime']]

# run k-means clustering
#creating the KMeans clustering object and fitting it to the real data in one line
kmeans1 = KMeans(n_clusters=kclusters, random_state=0).fit(cities_crime_clusters)

# check cluster labels generated for each row in the dataframe
kmeans1.labels_[0:10]
```

4. Results

Our initial map shows the United States and selected most populous cities superimposed on top of it that were plotted using folium library.

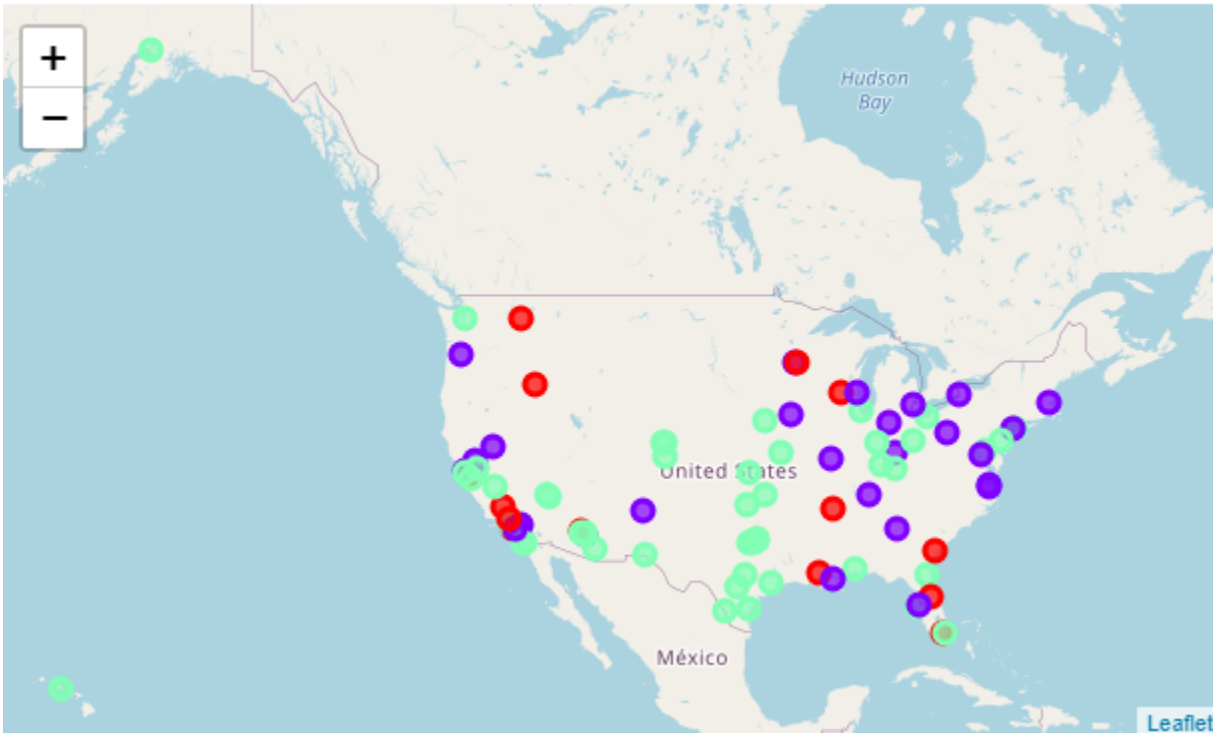
4.1. Clustering cities based on their post-secondary schools

Upon making Foursquare API calls and data cleaning we found 558 postsecondary academic centers in 91 most populated American cities. The following table shows the first 5 rows of the resulting data.

	Edu Cluster Labels	City	University per capita	College & University per capita	General College & University per capita	Latitude	Longitude
0	1	Albuquerque	0.000011	0.0	0.000004	35.084103	-106.650985
1	1	Anaheim	0.000011	0.0	0.000006	33.834752	-117.911732
2	2	Anchorage	0.000003	0.0	0.000003	61.216313	-149.894852
3	2	Arlington	0.000008	0.0	0.000005	32.701939	-97.105624
4	1	Atlanta	0.000010	0.0	0.000004	33.748992	-84.390264

A quick review of the resulting table shows that the cluster 1 shows the cities with the highest number of universities per capita. The cluster 0 shows a relatively high number of General College and University per capita, whereas, the cluster 2 shows the lowest number of General College and University per capita. The number of College and University per capita is low for all 3 clusters.

The result of visualizing the determined clusters over the map using the Folium library is displayed in the following, where the red color assigns the cities grouped in cluster 0. The colors violet and light green show the clusters 1 and 2, respectively.



4.2. Clustering cities based on their crime rates

Clustering of the cities based on their crime rates using the KMeans algorithm resulted in 3 groups. The following image shows the upper part of the resulting table.

Crime Cluster Labels	Edu Cluster Labels	City	University per capita	College & University per capita	General College & University per capita	Total Violent Crime	Total Property crime	Latitude	Longitude
0	1	Irvine	1.810840e-05	0.000000e+00	7.243359e-06	61.21	1316.48	33.685697	-117.825982
0	2	Gilbert	8.261390e-06	0.000000e+00	0.000000e+00	85.51	1385.85	33.352763	-111.789037
0	2	Plano	3.443372e-06	0.000000e+00	0.000000e+00	149.79	1733.74	33.013676	-96.692510
0	0	Santa Clarita	4.622140e-06	9.244280e-06	1.386642e-05	162.70	1424.08	34.391664	-118.542586
0	0	Fremont	0.000000e+00	4.230691e-06	1.269207e-05	182.34	2150.46	37.548270	-121.988572
0	2	Henderson	3.341297e-06	3.341297e-06	3.341297e-06	185.11	1833.04	36.030113	-114.982619
0	0	Hialeah	8.394191e-06	0.000000e+00	2.937967e-05	198.52	2213.55	25.857596	-80.278106
0	2	Irving	8.262346e-06	0.000000e+00	0.000000e+00	226.80	2539.43	32.829518	-96.944218
0	2	Honolulu	2.019419e-06	2.019419e-06	3.029128e-06	246.37	2774.38	21.304547	-157.855676
0	1	Chandler	1.604139e-05	0.000000e+00	0.000000e+00	259.47	2329.61	33.306160	-111.841250
0	0	Boise	4.431112e-06	8.862223e-06	8.862223e-06	279.16	2444.64	43.616616	-116.200886
0	2	Chula Vista	7.377107e-06	0.000000e+00	3.688553e-06	298.04	1432.27	32.640054	-117.084196
2	2	Garland	4.232930e-06	0.000000e+00	0.000000e+00	316.62	3032.47	32.912624	-96.638883
0	2	Laredo	0.000000e+00	3.836283e-06	3.836283e-06	321.86	2483.61	27.519984	-99.495376
2	2	Lexington	6.204783e-06	0.000000e+00	6.204783e-06	350.88	3782.13	38.046407	-84.497039
0	1	Anaheim	1.131862e-05	0.000000e+00	5.659310e-06	354.56	2630.45	33.834752	-117.911732

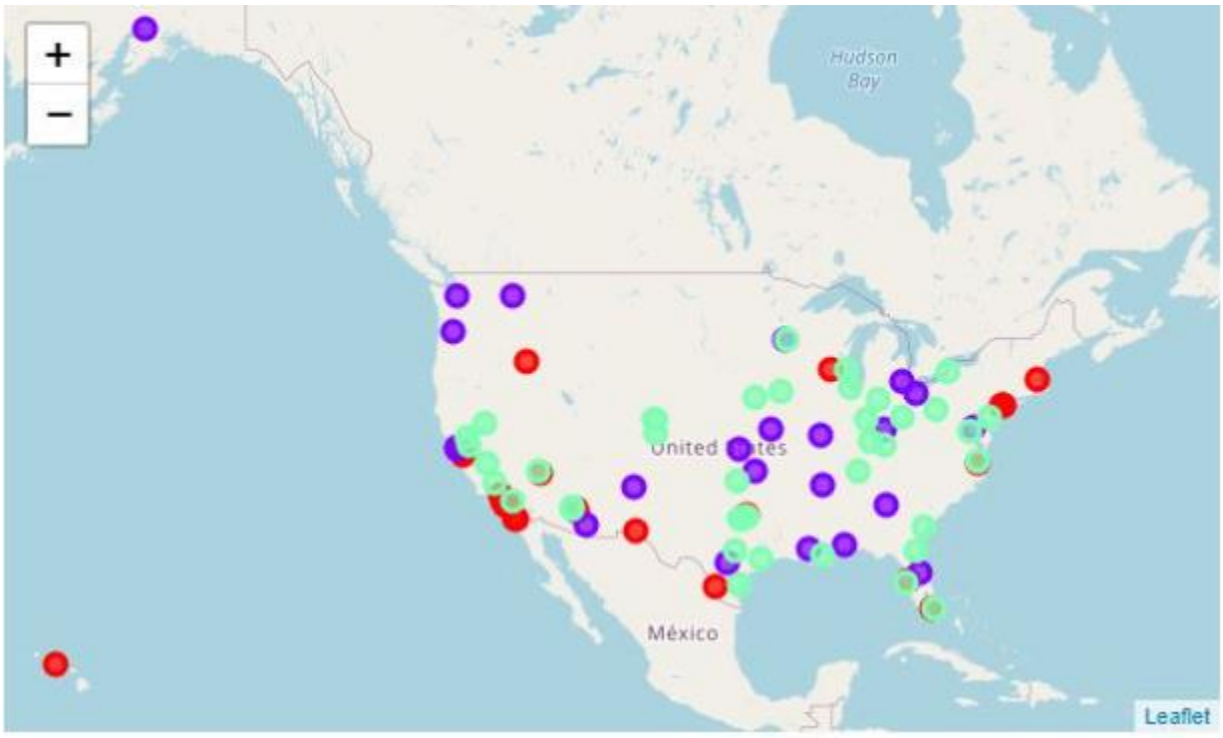
Cluster 0 showed the cities with the lowest Total Property Crime rate. Cluster 2 showed the cities with the mid-level of the Total Property Crime rate, while cluster 1 demonstrated the highest rate of this type of crime.
When it came to the Total Violent Crime rate, again the cluster 0 showed the lowest crime rate. Despite existing some contradicting data, the general trend demonstrated a mid-level of Total Violent Crime for cluster 2 and the highest level for cluster 1, which were almost along with the Total Property Crime trend.

In the next step, the cities clusters were visualized using the Folium map and the following code.

```
# create map (of Toronto)
map_clusters1 = folium.Map(location=[latitude, longitude], zoom_start=4)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map

markers_colors = []
for lat, lon, poi, cluster in zip(crime_df1['Latitude'], crime_df1['Longitude'], crime_df1['City'],
crime_df1['Crime Cluster Labels']):
    label = folium.Popup(str(poi) + ' Crime Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
```

```
    [lat, lon],
    radius=5,
    popup=label,
    color=rainbow[cluster-1],
    fill=True,
    fill_color=rainbow[cluster-1],
    fill_opacity=0.7).add_to(map_clusters1)

map_clusters1
```

The following map shows the cities clusters based on their rates of crime, where the color red stands for cluster 0, the violet color signifies cluster 1, and cluster 2 is signified with the color green.



Making a comparison between the result of the two different clustering experiences showed that while there were cities to show a correlation between their low crime rate and a high number of universities, the exceptions were observed as well. For instance, cities like Albuquerque, Oakland, St Louis, Memphis, Baton Rouge, and Detroit with good levels of academic centers still proved very high crime rates.

5. Discussion

The clustering results for the cities post-secondary educational centers showed a trend in data, where cluster 1 showed the cities with the highest number of University per capita, and the highest number of General College and University per capita was related to cluster 0, while, cluster 2 demonstrated the lowest number of General College and University per capita. However, in some cases, those trends were not supported by the results of clustering done for the rate of crime of the cities.

In the case of the rate of crimes, the trends of clustering results looked clearer, where the clusters 0, 2, and 1 showed the lowest to the highest rates of crime, particularly for property crimes and, with some exceptions, for violent crimes rates.

Comparing both clustering trends showed a level of correlation between them. However, there were exceptional cases as well. The particular exceptions were the cities like Albuquerque, Oakland, St Louis, Memphis, Baton Rouge, and Detroit, where despite good academic levels, the rates of crimes displayed showed very high levels.

It may be argued that other factors are affecting the rate of crime as well. For example, the economic condition of a city can affect the rate of crime as we see in the cities like Baton Rouge, Memphis, and St Louis. Another exception is the city of Oakland, CA. Oakland is a short distance from Berkeley CA, which is the home to many good academic centers. Since we set for the radius of 5000m to count for the post-secondary schools near each city, Berkely's academic centers would increase the academic grade around Oakland. However, Oakland has been historically considered as a poor region in the San Francisco Bay Area that affects the rate of crime in this city.

Another possibility might be that the university graduate number count for a small minority of the city’s population, and this intellectually elite minority is not affecting the rate of crime. Maybe an increase in a lower level of education, for instance, secondary education, would be more of a determining factor, when it comes to the rate of crime.

6. Conclusion

In this study it was attempted to investigate the rate of crime in most populated American cities, as well as their level of post-secondary education. For the rate of crime, both violent crime rates and property crime rates data were scrapped from the relevant websites. After data cleaning, the K-Means clustering as an unsupervised machine learning algorithm was performed on them.

Finally, the clusters were visualized into a map using the Folium library. Accordingly, the cities were divided into 3 clusters based on their crime rates.

Similarly, to investigate the level of post-secondary education in those most populated American cities, the numbers of universities were determined by making Foursquare API calls. Then data cleaning was performed on the query results, which was a json file containing the relevant detailed information about those cities. Then, K-Means clustering was conducted on the prepared data to categories the cities based on the numbers of their post-secondary educational centers into 3 groups. In the same way, visualization was done for education-based clustering of the cities, using the Folium library.

Despite observing some contradicting exceptions, many cases of correlations were observed between two different clustering results, pointing to a decrease in the rate of crime upon an increase in the number of universities in the cities. The results confirmed the assumption that an increase in the post-secondary education level of people may decrease their intentions towards committing the crime. On the other hand, for the cities with a good level of academic education and still a high rate of crimes, economic problems may be considered as the crime controlling factors.

References

1. https://www.numbeo.com/crime/rankings_by_country.jsp?title=2016
https://worldpopulationreview.com/country-rankings/crime-rate-by-country

2. https://en.wikipedia.org/wiki/Violent_crime

3. https://en.wikipedia.org/wiki/Property_crime

4. List of United States cities by crime rate - Wikipedia:
https://en.wikipedia.org/wiki/List_of_United_States_cities_by_crime_rate