# Establishing a relationship between social indicators and the occurrence of violent events

## A. Introduction

Biologically humans are programmed to be aggressive since this is part of their survival instinct, instead violence is a learned behavior, it can be assumed that there are certain elements in the environment that trigger this type of behavior.
The objective of this study is to determine whether social indicators such as health, education, infrastructure, commerce and public places, have an impact on the manifestation of criminal acts or not (we will take crime as an indicator of violence in society). In the case of finding a relationship, these indicators could be used as input variables in the construction of a predictive model of violent events, which could be used by users who are interested in knowing the level of risk of a certain location or the model could be used by government entities to define preventive measures in those locations with high risk.

As an initial hypothesis, it is established that if there are social indicators that could influence acts of violence, these variables should have a correlation with the target variable (number of crimes)
To demonstrate this relationship first, we establish the expected result in the event that the original hypothesis was correct, we will call it Cluster Hypothesis graphic, this will be our reference value. additionally we will build two additional graphs the Cluster_Predictors and the Cluster_Venues which serve to validate the hypothesis, the predictive variables of the first graph will be the social indicators and the second will have as predictive variables the most frequent public places of interest in each locality (restaurants, local commercial, pharmacies, museums, public institutions, etc…)

## B. Data

In the constructed Data Set, each row represents a region of my country of origin Chile and each column is a social indicator and a numerical variable, the Data set was constructed from various sources which are detailed below:

### b.1 General Data

General variables such as name of the region, regional capital, demographic, area and density. They were extracted from Wikipedia, we use the regional capital variable to obtain the coordinates using the Python Geopy library

Source: https://es.wikipedia.org/wiki/Chile_Regions

### b.2 Social Indicators

The education, health, GDP, housing, public investment, labor and commercial production indicators were obtained from the public data sets created by the National Statistics Institute of Chile (INE)
Source: https://datosabiertos.ine.cl/home

### b.3 Variable Objective (crimes)

The data corresponding to violent events was obtained from the annual reports delivered by Carabineros de Chile and the Investigation Police.
http://datos.gob.cl/dataset

### b.4 Geolocation variables

For this data, the Geocoder Api was used to obtain geographical points such as latitude and longitude, to obtain the places of interest (restaurant, beaches, bar, etc.) I use the Foursquare Api

## C. Methodology

### c.1 Construction of the Data Set and choice of predictor variables,

After cleaning, reducing and grouping the different Data Set, the columns of interest are copied to the Data-set-end. The objective of this step is to generate a Data set that contains all those characteristic elements of the specific region, formatting each column, new names are given to columns and new index is created, leaving the final data set as follows

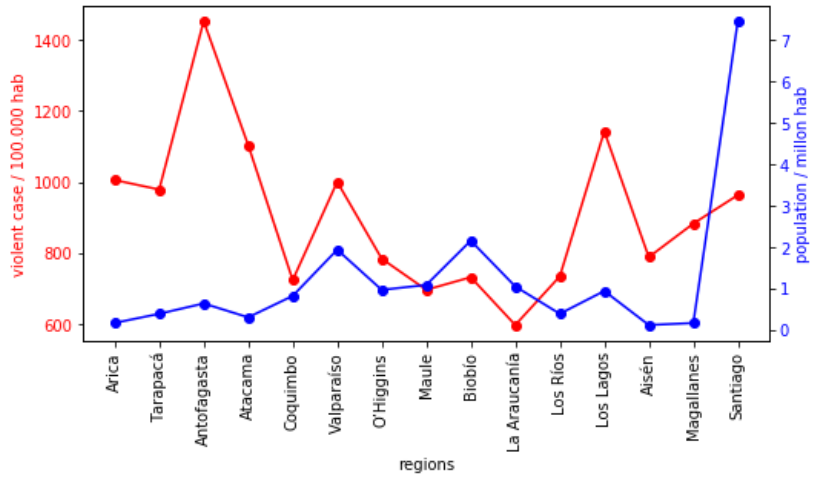
Initial Data Set:

```
df.columns

Index([
        'demographic',
        'average remuneration','crimes per 100,000',
        'violations per 100,000','allegation',
        'medium overcrowding','critical overcrowding',
        'deficient sanitation', 'precarious housing',
        'Potable Water', 'Sewer', 'Sewage Treatment',
        'public investment','number of students',
        'school retention bonus','mental disability subcide',
        'family subcide', 'drinking water subcide',
        'solidarity bonus', 'subcide cedula', 'family ethical income bonus',
        '0 scholar years', '1 scholar years', '2 scholar years',
        '3 scholar years', '4 scholar years', '5 scholar years',
        '6 scholar years', '7 scholar years', '8 scholar years',
        '9 scholar years', '10 scholar years', '11 scholar years',
        '12 scholar years', '13 scholar years', '14 scholar years',
        '15 scholar years', '16 scholar years', '17scholar years',
        '18 scholar years or more'],

dtype='object')
```

Graphic: cases of violence and population by region




In the graph it can be clearly seen that the acts of violence do not have a linear relationship with the number of people in each locality, more people do not mean more crimes

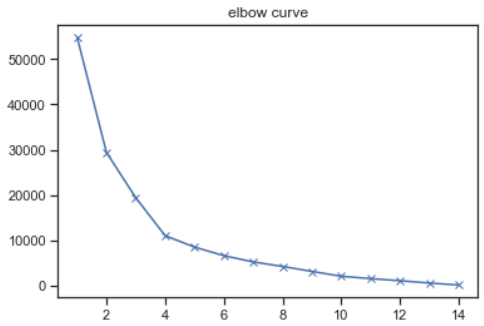
The pingouin library is used to obtain a table of dispersions that relates the objective-taget variable ('crime per 100,000') to all other variables, the results are used to reduce the number of predictors of the Data Set leaving only

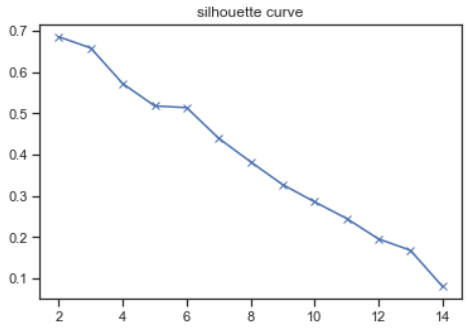
those strongly related variables with the target variable, the reduction of variables in the Deta set will allow avoiding an eventual overfitting when adjusting the variables in some model

| | X | Y | method | tail | n | r | CI95% | r2 | adj_r2 | z | p-unc | BF10 | power |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | crimes per 100,000 | deficient sanitation * | pearson | two-sided | 15 | -0.274 | [-0.69, 0.28] | 0.075 | -0.079 | -0.281 | 0.323822 | 0.498 | 0.170 |
| 18 | crimes per 100,000 | subcide cedula | pearson | two-sided | 15 | -0.264 | [-0.68, 0.29] | 0.070 | -0.085 | -0.270 | 0.342006 | 0.482 | 0.160 |
| 19 | crimes per 100,000 | family ethical income bonus | pearson | two-sided | 15 | -0.241 | [-0.67, 0.31] | 0.058 | -0.099 | -0.246 | 0.387498 | 0.449 | 0.141 |
| 13 | crimes per 100,000 | school retention bonus | pearson | two-sided | 15 | -0.234 | [-0.67, 0.32] | 0.055 | -0.103 | -0.238 | 0.400417 | 0.441 | 0.136 |
| 17 | crimes per 100,000 | solidarity bonus | pearson | two-sided | 15 | -0.227 | [-0.66, 0.32] | 0.052 | -0.107 | -0.231 | 0.415981 | 0.432 | 0.130 |

| | X | Y | method | tail | n | r | CI95% | r2 | adj_r2 | z | p-unc | BF10 | power |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | crimes per 100,000 | allegation | pearson | two-sided | 15 | 0.500 | [-0.02, 0.81] | 0.250 | 0.125 | 0.549 | 0.057603 | 1.663 | 0.499 |
| 22 | crimes per 100,000 | avg_edu | pearson | two-sided | 15 | 0.570 | [0.08, 0.84] | 0.325 | 0.212 | 0.648 | 0.026510 | 3.03 | 0.635 |
| 5 | crimes per 100,000 | critical overcrowding | pearson | two-sided | 15 | 0.621 | [0.16, 0.86] | 0.386 | 0.284 | 0.727 | 0.013446 | 5.196 | 0.735 |
| 21 | crimes per 100,000 | PIB/usd | pearson | two-sided | 15 | 0.714 | [0.32, 0.9] | 0.510 | 0.428 | 0.895 | 0.002789 | 18.691 | 0.889 |
| 1 | crimes per 100,000 | average remuneration | pearson | two-sided | 15 | 0.786 | [0.46, 0.93] | 0.618 | 0.555 | 1.061 | 0.000507 | 76.935 | 0.965 |

The educational data (years of education) were reduced to a single variable that averages the years of schooling by each region, the data with low correlation with the target variable were taken from the dataset since together with those variables that have a certain linear dependence with another variable within dataset, they do not contribute anything when evaluating the data in some model.
Due to the small number of rows in the dataset I decide to use the Kmeans model which allows grouping the nodes (row) of the dataset using a criterion of similarity between the variables, to use Kmeans we must calculate the number of ideal nodes that the model will use ( number K), to calculate this value I use the elbow method and use a second method (silhouette) to confirm the result




With the elbow method the ideal K is not clearly visible, but with the silhouette method it is possible to see clearly that for this data set the ideal amount of clusters is 2, we use Kmeans using this value.

```
#Cluster the data using the optimal n_cluster
kmeans = KMeans(n_clusters=2, random_state=0).fit(df_kmeans)
labels = kmeans.labels_
centroids = kmeans.cluster_centers_

#Glue back to originaal data
df['clusters'] = labels
print(labels)
```

```
[0 1 1 1 0 0 0 0 0 0 0 0 0 0 1]
```

The array (labels) shown above, stores the classification for each of the regions of the Data Set, the table below shows the characteristics of the centroids found by Kmeans, this table allows you to establish the weight of each variable and how each predictor affects in the final classification.

```
# characteristic elements of centroids
# red is clouster 0
# green is cluster 1
array2=pd.DataFrame(name_kmeans['Y'])
array2['cluster_red']=kmeans.cluster_centers_[0]
array2['cluster_green']=kmeans.cluster_centers_[1]
array2
```

| | Y | cluster_red | cluster_green |
|---|---|---|---|
| 1 | average remuneration | 355222.363636 | 492593.000000 |
| 21 | PIB/usd | 13.794727 | 28.414000 |
| 5 | critical overcrowding | 2.527273 | 3.375000 |
| 22 | avg_edu | 8.304681 | 8.738194 |
| 3 | allegation | 17.409091 | 30.275000 |
| 4 | medium overcrowding | 16.790909 | 20.400000 |
| 9 | Sewer | 93.863636 | 97.850000 |
| 7 | precarious housing | 11.390909 | 18.925000 |
| 10 | Sewage Treatment | 90.736364 | 94.925000 |
| 6 | deficient sanitation * | 19.881818 | 18.000000 |

Dataframe predictor variables

```
df_predictor=pd.DataFrame()
df_predictor['region']=df['region']
df_predictor['clouster_predictors']=df['clusters']
df_predictor
```

| | region | clouster_predictors |
|---|---|---|
| 0 | Arica | 0 |
| 1 | Tarapacá | 1 |
| 2 | Antofagasta | 1 |
| 3 | Atacama | 1 |
| 4 | Coquimbo | 0 |
| 5 | Valparaíso | 0 |
| 6 | O'Higgins | 0 |
| 7 | Maule | 0 |
| 8 | Biobío | 0 |
| 9 | La Araucanía | 0 |
| 10 | Los Ríos | 0 |
| 11 | Los Lagos | 0 |
| 12 | Aisén | 0 |
| 13 | Magallanes | 0 |
| 14 | Santiago | 1 |

### c.2 Obtaining geolocation variables

A new temporary dataframe is created to store the location data, the 'capitals' column of the main dataframe is copied to be used as an entry in the Api Geocode, with the Geocode results the following dataframe is created:

```
df_coord
```

| | capital | geocode | point | longitude | latitude |
|---|---|---|---|---|---|
| 0 | arica | (Arica, Provincia de Arica, Región de Arica y ... | (-18.478518, -70.3210596, 0.0) | -18.478518 | -70.321060 |
| 1 | iquique | (Iquique, Provincia de Iquique, Región de Tara... | (-20.2140657, -70.1524646, 0.0) | -20.214066 | -70.152465 |
| 2 | antofagasta | (Antofagasta, Provincia de Antofagasta, Región... | (-23.6463741, -70.3980033, 0.0) | -23.646374 | -70.398003 |
| 3 | copiapo | (Copiapó, Provincia de Copiapó, Región de Atac... | (-27.3664897, -70.3322733, 0.0) | -27.366490 | -70.332273 |
| 4 | la serena | (La Serena, Provincia de Elqui, Región de Coqu... | (-29.9026615, -71.2520136, 0.0) | -29.902662 | -71.252014 |
| 5 | valparaiso | (Valparaíso, Provincia de Valparaíso, Región d... | (-33.0458456, -71.6196749, 0.0) | -33.045846 | -71.619675 |
| 6 | rancagua | (Rancagua, Provincia de Cachapoal, Región del ... | (-34.170249, -70.7407427, 0.0) | -34.170249 | -70.740743 |
| 7 | talca | (Talca, Provincia de Talca, Región del Maule, ... | (-35.4266305, -71.6661153, 0.0) | -35.426631 | -71.666115 |
| 8 | concepcion | (Concepción, Provincia de Concepción, Región d... | (-36.8270776, -73.0502683, 0.0) | -36.827078 | -73.050268 |
| 9 | temuco | (Temuco, Provincia de Cautín, Región de la Ara... | (-38.7362442, -72.5905979, 0.0) | -38.736244 | -72.590598 |
| 10 | valdivia | (Valdivia, Provincia de Valdivia, Región de Lo... | (-39.8141965, -73.2458525, 0.0) | -39.814197 | -73.245852 |
| 11 | puerto montt | (Puerto Montt, Provincia de Llanquihue, Región... | (-41.4718121, -72.939621, 0.0) | -41.471812 | -72.939621 |
| 12 | coyhaique | (Coyhaique, Provincia de Coyhaique, Región Ays... | (-45.5711804, -72.0684863, 0.0) | -45.571180 | -72.068486 |
| 13 | punta arenas | (Punta Arenas, Provincia de Magallanes, Región... | (-53.1625446, -70.907785, 0.0) | -53.162545 | -70.907785 |
| 14 | santiago | (Santiago, Provincia de Santiago, Región Metro... | (-33.4377968, -70.6504451, 0.0) | -33.437797 | -70.650445 |

The latitude and longitude column is used as entries in the Api de Foursqueare which will allow us to obtain the public places (venues) that characterize each locality (public places refers to restaurants, tourist places, general commerce, museums, typical areas, institutions public)
With the information obtained from Foursqueare a dataframe is created that contains all the places of interest belonging to each locality, assigning it geographical position, name of the place, and category to which it belongs, the resulting dataframe contine 1143 places of interest.

```
chile_venues.head()
```

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | arica | -18.478518 | -70.32106 | Rayú | -18.462694 | -70.304099 | South American Restaurant |
| 1 | arica | -18.478518 | -70.32106 | Valle de Azapa | -18.492309 | -70.280021 | Field |
| 2 | arica | -18.478518 | -70.32106 | La Fontana | -18.484163 | -70.303503 | Ice Cream Shop |
| 3 | arica | -18.478518 | -70.32106 | Playa El Laucho | -18.487818 | -70.326610 | Beach |
| 4 | arica | -18.478518 | -70.32106 | Playa Chinchorro | -18.454318 | -70.301581 | Beach |

```
chile_venues.shape
```

```
(1143, 7)
```

The places (Venues) are categorical variables must be converted into numerical variables for that uses the ONE-HOT-DUMMY method that converts each categorical variable into a new column and establishes the presence or not in a particular row using binaries (0 does not exist in that row, 1 exists in that row) this new information is stored in the Data Set “chile_onehot”
Using this method will allow you to group the results in their respective regions and establish which places (venue) have the greatest presence in each region, the results are grouped by region (rows) and the columns are sorted in descending order according to their average presence within the data set, finally a new dataset is obtained that shows the first 10 places that have more presence for each region this data is stored in the data set “df_venues_chile”

```
df_venues_chile
```

| | Region | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | antofagasta | Beach | Hotel | Restaurant | Pizza Place | Peruvian Restaurant | Chinese Restaurant | Sandwich Place | Bistro | Soccer Field | Scenic Lookout |
| 1 | arica | Beach | Restaurant | Hotel | Ice Cream Shop | Italian Restaurant | Latin American Restaurant | Surf Spot | Chinese Restaurant | Plaza | Gym |
| 2 | concepcion | Pizza Place | Beach | Restaurant | Burger Joint | Café | Peruvian Restaurant | Nightclub | Hotel | Theater | Coffee Shop |
| 3 | copiapo | Beach | Hotel | Pub | Seafood Restaurant | Historic Site | History Museum | Pizza Place | Nightclub | Diner | Restaurant |
| 4 | coyhaique | Café | Restaurant | Bed & Breakfast | Hotel | Scenic Lookout | Pizza Place | Bar | Burger Joint | Boat or Ferry | Sushi Restaurant |
| 5 | iquique | Beach | Restaurant | Hotel | Latin American Restaurant | Plaza | Sushi Restaurant | Park | Museum | Theme Park | Pizza Place |
| 6 | la serena | Beach | Seafood Restaurant | Café | Restaurant | Ice Cream Shop | Coffee Shop | Cupcake Shop | Bakery | Soccer Field | Resort |
| 7 | puerto montt | Hotel | Scenic Lookout | Beach | Seafood Restaurant | German Restaurant | Restaurant | BBQ Joint | Tea Room | Coffee Shop | Dessert Shop |
| 8 | punta arenas | Restaurant | History Museum | Café | Scenic Lookout | Other Great Outdoors | Tea Room | Bed & Breakfast | Coffee Shop | Pizza Place | Gastropub |
| 9 | rancagua | Vineyard | Mountain | Restaurant | Scenic Lookout | Park | Wine Bar | Hotel | Winery | Café | Pizza Place |
| 10 | santiago | Park | Scenic Lookout | Hotel | Pizza Place | Peruvian Restaurant | Museum | Sandwich Place | French Restaurant | Theater | Café |
| 11 | talca | Park | Vineyard | Plaza | Restaurant | Farmers Market | Hotel | Beach | Coffee Shop | Ice Cream Shop | Mediterranean Restaurant |
| 12 | temuco | Hotel | Café | BBQ Joint | Seafood Restaurant | Beach | Plaza | Italian Restaurant | Burger Joint | Restaurant | Ski Area |
| 13 | valdivia | Café | Restaurant | Hotel | Bar | Scenic Lookout | Bed & Breakfast | Beach | Seafood Restaurant | Park | Brewery |
| 14 | valparaiso | Scenic Lookout | Pizza Place | Restaurant | Beach | Café | Ice Cream Shop | Historic Site | Italian Restaurant | Peruvian Restaurant | Coffee Shop |

I apply Kmeans to the Dataset “chile_onehot” to create a grouping of the data and in this way to be able to classify each region based on its public places, this data is stored in the dataframe “df_kmeans_venues” with the column name ‘clousters_venues’

```
# set number of clusters, using the same n_cluster of befor
kclusters = 2

chile_grouped_clustering = chile_onehot.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(chile_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_

#  resultado de arriba [0 1 1 1 0 0 0 0 0 0 0 0 0 0 1]
```

```
array([0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1])
```

```
neighborhoods_venues_sorted['clousters_venues']=kmeans.labels_
df_kmeans_venues=neighborhoods_venues_sorted[['Region','clousters_venues']]
df_kmeans_venues
```

| | Region | clousters_venues |
|---|---|---|
| 0 | antofagasta | 0 |
| 1 | arica | 0 |
| 2 | concepcion | 0 |
| 3 | copiapo | 0 |
| 4 | coyhaique | 1 |
| 5 | iquique | 0 |
| 6 | la serena | 0 |
| 7 | puerto montt | 1 |
| 8 | punta arenas | 1 |
| 9 | rancagua | 1 |
| 10 | santiago | 1 |
| 11 | talca | 0 |
| 12 | temuco | 1 |
| 13 | valdivia | 1 |
| 14 | valparaiso | 1 |

## D. Results

The partial results obtained from the previous section are used to construct the final Data Set, the 'region' column identifies the location that is being classified, it will also serve as an index of the dataframe, the column 'target_value' in the value of the rate of crimes produced in each location this variable is the objective variable used in the analysis, the variable 'clouster_hypothesis' will be the reference clouster (ideal data grouping) the other partial results will be compared with this clouster which will measure the degree of success of the study , the 'clouster_predictors' is the clouster result created based on the predictor variables, and finally the 'clouster_venues' column is the resulting classification after the analysis performed based on the public places of each locality

```
df_final
```

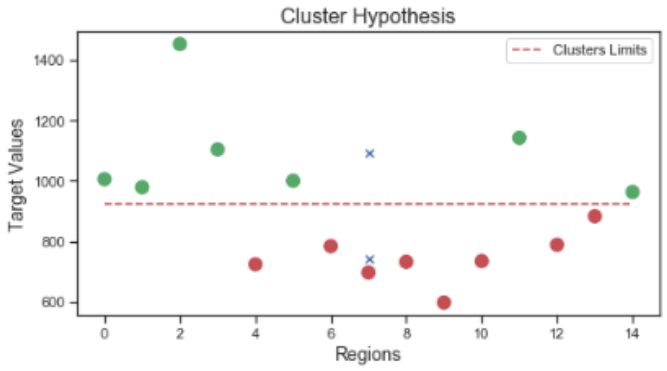
| | region | target_value | cluster_hypothesis | cluster_predictors | clouster_venues |
|---|---|---|---|---|---|
| 0 | Arica | 1005.639149 | 1 | 0 | 0 |
| 1 | Tarapacá | 978.908480 | 1 | 1 | 0 |
| 2 | Antofagasta | 1452.540381 | 1 | 1 | 0 |
| 3 | Atacama | 1103.930592 | 1 | 1 | 0 |
| 4 | Coquimbo | 724.207164 | 0 | 0 | 0 |
| 5 | Valparaíso | 1000.416674 | 1 | 0 | 1 |
| 6 | O'Higgins | 784.271108 | 0 | 0 | 1 |
| 7 | Maule | 696.743799 | 0 | 0 | 0 |
| 8 | Biobío | 732.158965 | 0 | 0 | 0 |
| 9 | La Araucanía | 597.679971 | 0 | 0 | 1 |
| 10 | Los Ríos | 734.509848 | 0 | 0 | 1 |
| 11 | Los Lagos | 1142.473118 | 1 | 0 | 1 |
| 12 | Aisén | 788.798489 | 0 | 0 | 1 |
| 13 | Magallanes | 883.064821 | 0 | 0 | 1 |
| 14 | Santiago | 963.294736 | 1 | 1 | 1 |

Cluster_hypothesis graphic:
Red color represents cluster of value 0
Green color represents cluster of value 1

```
categories = df_final['cluster_hypothesis']
colormap = np.array(['r', 'g', 'b'])
plt.figure(figsize=(8,4))
plt.scatter(df.index, df['crimes per 100,000'], s=100, c=colormap[categories])
plt.scatter(7,centroids_only_target[0,0], marker="x", color='b')
plt.scatter(7,centroids_only_target[1,0], marker="x", color='b')
plt.title('Cluster Hypothesis',fontsize=16)
plt.plot([0,14],[923,923], 'r--',label='cluster divition')
plt.legend(['Clusters Limits'])

plt.xlabel('Regions',fontsize=14)
plt.ylabel('Target Values',fontsize=14)


plt.savefig('ScatterClassPlot.png')
plt.show()
```

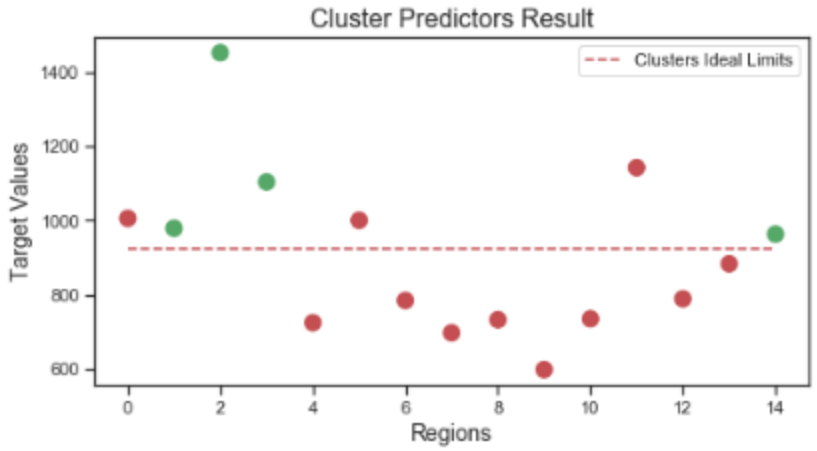



Cluster_predictors graphic:
Red color represents cluster of value 0
Green color represents cluster of value 1

```
categories = df_final['cluster_predictors']
colormap = np.array(['r', 'g', 'b'])
plt.figure(figsize=(8,4))
plt.scatter(df.index, df['crimes per 100,000'], s=100, c=colormap[categories])
plt.title('Cluster Predictors Result',fontsize=16)
plt.plot([0,14],[923,923], 'r--',label='cluster divition')
plt.legend(['Clusters Ideal Limits'])

plt.xlabel('Regions',fontsize=14)
plt.ylabel('Target Values',fontsize=14)


plt.savefig('ScatterClassPlot.png')
plt.show()
```

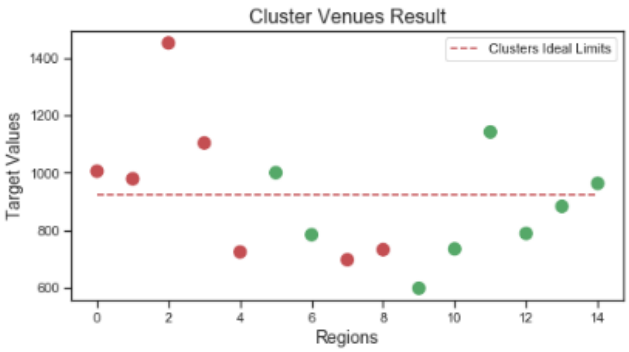



Cluster_venues chart:
Red color represents cluster of value 0
Green color represents cluster of value 1

```
categories = df_final['clouster_venues']
colormap = np.array(['r', 'g', 'b'])
plt.figure(figsize=(8,4))
plt.scatter(df.index, df['crimes per 100,000'], s=100, c=colormap[categories])
plt.title('Cluster Venues Result',fontsize=16)
plt.plot([0,14],[923,923], 'r--',label='cluster divition')
plt.legend(['Clusters Ideal Limits'])

plt.xlabel('Regions',fontsize=14)
plt.ylabel('Target Values',fontsize=14)


plt.savefig('ScatterClassPlot.png')
plt.show()
```



## E. Discussion

With a larger set of data, better predictive results could be obtained because this would allow us to use other approaches in the analysis, unfortunately the raw data needed to do this is not easily obtainable and not public. In more developed countries, public institutions such as the police and ministries make public access Apis available to citizens for data consultation. It would be very useful for Chile to have such initiatives.

## F. Conclusion

The initial hypothesis of the project establishes the relationship that should exist between acts of violence (number of crimes) and the set of predictive variables.

For validity, the hypothesis proceeds to the construction of the Cluster_Predictors and the Cluster_venues graphs to establish the relationship between them and the Cluster_Hypothesis. When comparing the results between the Cluster_Hypothesis and the Clouster_Predictors, it can be seen that there is 57% success in the classification of the cases of higher crime rate and 78% of success in the total classification of crimes, so it could be said that the set of predictors is correlated with the target variable, in the same way by predicting with certainty less than 60% in the cases of interest, I find that the result is not entirely conclusive and that a second review of the data and the model is needed to find better results.

In the comparison of Cluster_venues and Cluster_Hypothesis, the results indicate 57% of successes in cases of cases of high violence, but they only have 57% of successes in relation to the total of the classified cases, the amount of predictive variables used in this Kmeans model (192 predictor variables), makes the model classify less accurately.

none of the comparisons made achieved conclusive results that validated the initial hypothesis (more than 80% of correctness in the classification), without doubt the study carried out delivered unexpected results that merit a second review, such as analyzing the behavior of the submitted Kmeans models to a smaller number of predictors or to analyze how the model behaves when delivering predictors of a different nature to those presented in this study. At the moment and being this the first version of the study that seeks the relationships between social indicators and the occurrence of violent events, I have determined that no social factors have been found that influence the violent behavior of people.