Ciavarro Cristina 253188
Di Natale Marco 255660
Nadhomi Timoty 255089
Okeke Chisom Prince 255450

Analysis of crimes in the USA

Project work

Introduction: why this data?

The data we have chosen derive from our curiosity. American culture fascinates us and the differences that emerge in contrast with ours have led us to ask some questions. In particular, there is a famous aspect of American justice, which is the center of numerous debates even within America itself, which has often led us to ask ourselves questions: the death penalty. The subject is delicate, but we thought that if we had studied crime in the United States, particularly in the different states, perhaps we could have better understood the dynamics of American justice that seem so different from ours. So the data set we have chosen concerns criminality in America. Data were collected by the FBI and our study focused from 2010 to 2016. We are satisfied with the choice made because the study has surprised us: we left with an idea that at the end of the work was completely overturned.

Description of data set

From the official website of the FBI it is possible to download the data collected by the federal police from 1993 to 2016. The data are available per year in csv files. However every year has different types of data sets that contain different information. To do our studies we needed to download different types of data sets and then rebuild the data set useful for our purposes. In particular we downloaded 9 data sets:

~ 7 data sets show the number of different crimes compared to American cities grouped according to the states they belong to (which for convenience we will call "Data set of crimes in the USA"),

~ 1 data set containing the coordinates of the places we have to show in our studio,

~ 1 data set containing all the crimes that took place in Los Angeles in 2016 (which for convenience we will call "Data set of crimes in Los Angeles").

This is the composition of a part of the initial Data set of crimes in the USA:

	State	City	Year	Population	Violent crime	Murder	Rape	Robbery	Aggravated assault	Property crime	Burglary	Larceny theft	Motor vehicle theft	Arson
0	ALABAMA	HUNTSVILLE	2010	180.105	596.000	4	31.0	231.000	330.000	4.758	1.262	3.113	383.000	10.0
1	NaN	NaN	2011	NaN	518.000	4	41.0	158.000	315.000	3.427	989.000	2.073	365.000	15.0
2	NaN	MOBILE	2010	251.106	808.000	10	27.0	291.000	480.000	7.099	1.761	4.759	579.000	46.0
3	NaN	NaN	2011	NaN	695.000	18	24.0	280.000	373.000	6.650	1.897	4.366	387.000	28.0
4	ALASKA	ANCHORAGE	2010	291.826	1.191	9	129.0	213.000	840.000	4.816	603.000	3.806	407.000	36.0
5	NaN	NaN	2011	NaN	1.172	4	135.0	226.000	807.000	4.406	506.000	3.616	284.000	59.0
6	ARIZONA	CHANDLER	2010	236.123	330.000	2	33.0	107.000	188.000	3.716	694.000	2.868	154.000	23.0
7	NaN	NaN	2011	NaN	335.000	2	28.0	79.000	226.000	3.580	624.000	2.820	136.000	24.0
8	NaN	GILBERT	2010	208.453	92.000	1	12.0	28.000	51.000	2.129	425.000	1.629	75.000	13.0
9	NaN	NaN	2011	NaN	79.000	1	8.0	21.000	49.000	1.851	371.000	1.426	54.000	14.0
10	NaN	GLENDALE	2010	226.721	513.000	4	25.0	165.000	319.000	6.088	1.120	4.300	668.000	21.0
11	NaN	NaN	2011	NaN	535.000	6	20.0	197.000	312.000	7.249	1.093	5.393	763.000	40.0
12	NaN	MESA	2010	439.041	855.000	8	60.0	241.000	546.000	7.319	1.313	5.516	490.000	42.0
13	NaN	NaN	2011	NaN	877.000	5	77.0	237.000	558.000	7.254	1.303	5.446	505.000	44.0
14	NaN	PEORIA	2010	154.065	143.000	5	19.0	32.000	87.000	2.287	415.000	1.693	179.000	6.0
15	NaN	NaN	2011	NaN	137.000	1	16.0	25.000	95.000	2.391	542.000	1.678	171.000	3.0
16	NaN	PHOENIX	2010	1.445.632	3.818	52	254.0	1.497	2.015	30.457	7.367	19.220	3.870	164.0
17	NaN	NaN	2011	NaN	3.888	63	279.0	1.512	2.034	30.746	8.425	18.573	3.748	154.0
18	NaN	SCOTTSDALE	2010	217.385	176.000	2	18.0	50.000	106.000	3.251	620.000	2.513	118.000	7.0

Figure 1

This is the first of the 7 data sets that we downloaded from the FBI website and reports the data collected for the year 2010 and the prediction that the federal police did for 2011. Each of these data sets is structured as follows:

COLUMNS

State: state name,

City: city where the crime occurred,

Year: Year to which the crime refers,

Population: Population of the city in question,

Violent crime: number of occurrences for this type of crime,

Murder: number of occurrences for this type of crime,

Rape: number of occurrences for this type of crime,

Robbery: number of occurrences for this type of crime,

Aggravated assault: number of occurrences for this type of crime,

Property crime: number of occurrences for this type of crime,

Burglary: number of occurrences for this type of crime,

Larceny theft: number of occurrences for this type of crime,

Motor vehicle theft: number of occurrences for this type of crime,

Arson: number of occurrences for this type of crime.

The final date set, not yet cleaned, containing all 7 years studied by us, has 3582 rows and 14 columns.

This is the composition of the initial Data set of crimes in Los Angeles:

	DR_Number	Date_Reported	Date_Occurred	Time_Occurred	Area_ID	Area_Name	Reporting_District	Crime_Code	Crime_Code_Description	MO_Codes	...	S
0	1208575	03/14/2013	03/11/2013	1800	12	77th Street	1241	626	INTIMATE PARTNER - SIMPLE ASSAULT	0416 0446 1243 2000	...	
1	102005556	01/25/2010	01/22/2010	2300	20	Olympic	2071	510	VEHICLE - STOLEN	NaN	...	
2	418	03/19/2013	03/18/2013	2030	18	Southeast	1823	510	VEHICLE - STOLEN	NaN	...	
3	101822289	11/11/2010	11/10/2010	1800	18	Southeast	1803	510	VEHICLE - STOLEN	NaN	...	
4	42104479	01/11/2014	01/04/2014	2300	21	Topanga	2133	745	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	0329	...	

Figure 2

This is the initial data set we used to analyze the city of Los Angeles in more detail, and is structured as follows:

COLUMNS

DR Number: district number,

Date Reported: report date,

Date Occurred: crime date,

Time Occurred: exact crime time,

Area ID: Area ID,

Area Name: Area name,

Reporting District: LAPD Reporting Districts,

Crime Code: Crime Code,

Crime Code Desc: description of the corresponding crime code,

MO Codes: codes corresponding to an additional crime description,

Victim Age: age of the victim,

Victim Sex: sex of the victim,

Victim Descent: Ethnicity/Descent code:

A - Other Asian,

B - Black ,

C - Chinese ,

D - Cambodian ,

F - Filipino ,

G - Guamanian ,

H - Hispanic/Latin/Mexican ,

I - American Indian/Alaskan Native ,

J - Japanese ,

K - Korean ,

L - Laotian ,

O - Other ,

P - Pacific Islander ,

S - Samoan ,

U - Hawaiian ,

V - Vietnamese ,

W - White ,

X - Unknown ,

Z - Asian Indian,

Premise Code: building/place code where the crime took place,

Premise Desc: description of the building/place where the crime took place,

Weapon Used Cd: code corresponding to the weapon used for the crime,

Weapon Desc: description of the weapon used for the crime,

Status Code: code corresponding to the status of the crime (case settled with a man arrested, the investigation continues etc..),

Status Desc: description of the corresponding status code,

Crm Cd 1: crime code 1 that correspond to the field Crime Code,

Crm Cd 2: this field is not empty if another crime code can be assigned to the crime,

Crm Cd 3: this field is not empty if a third crime code can be assigned to the crime,

Crm Cd 4: this field is not empty if a fourth crime code can be assigned to the crime,

Address: address where the crime took place,

Cross Street: the name of the street crossing,

Location: coordinates of the place where the crime took place.

The data set, not yet cleaned, contain data related to different years, infact it is big and contain 1584316 rows and 28 columns. However we have used only the crimes occurred during 2016 for our study and in this way the rows are reduced to 20441.

Data cleaning

As for data cleaning, we first created a single data set by combining the 7 data sets that reported crimes in the United States each year. As each data series shows crimes in US cities during a given year (for example, 2010, as we saw in Figure 1 in the previous paragraph) with the prediction for the following year (for example, 2011), we have eliminated the row containing the prediction made by the FBI. For this reason we have used only the rows of the current year (in this case 2010), dropping the rows of prediction, since we had the real data of the crimes committed (in our example we had the real information concerning the crimes that happened in 2011 because they were on one of the other 6 data sets describing 2011). We did this for each of the seven data sets that show crimes in the United States, but for the 2016 data set we used FBI predictions for 2017 to compare them with those made by us at the end of our study on crime in Los Angeles. So we had to clean up, format and prepare the

dataset so that it could solve our problems but still be able to respond to our queries as simply as possible. Then we had to add the geographic coordinates data set for each state.

As for the NaNs in the crimes columns, we used the following logic to replace them: since each city could have a row with information about crimes for each year, if in a given year a city presents a NaN value for a crime, we check if that same city has non-NaN values for the same crime both for the year before and for the year after. If both of these values are present then we have replaced the NaN value with the average between these two values. If only the previous value is different from NaN then we have taken this value. If only the next value is different from NaN we have taken the next value. If the city appears only once in the 7 data sets and therefore does not have any of these two values, then we put the NaN value equal to 0. This could happen because unfortunately the seven data sets are not all the same: in fact it happened that one status of which information was present in the 2011 data set did not appear in the 2013 and 2015 data sets, or that, for example, an Ohio city was present only in the 2016 data set and in none of the other 6.

Always referring to figure 1, you can see that the name of each state was reported only once a year. This is because the different data were shown according to the cities of each state. But we had to show the data according to the status, so we had to fill the empty row in the State column with the name of the state to which the respective city in the next column belonged. Then we added the data of all the cities of each state, thus reporting information on crimes according to the states and no longer according to the cities.

For the second part it was necessary to download a new data set because the data available in the old one were few and did not allow us to make predictions. Furthermore, the old data set did not report where the crime had occurred, while the new one gives us the time, the road, the district and the coordinates of where the crime took place.

Because each line of this new data set reports a crime with related information, we needed to take the different types of crimes located in each row making them the columns of this new data set. This was necessary to get the old man's look at the new data set, thus being able to do the same kind of predictions on crimes.

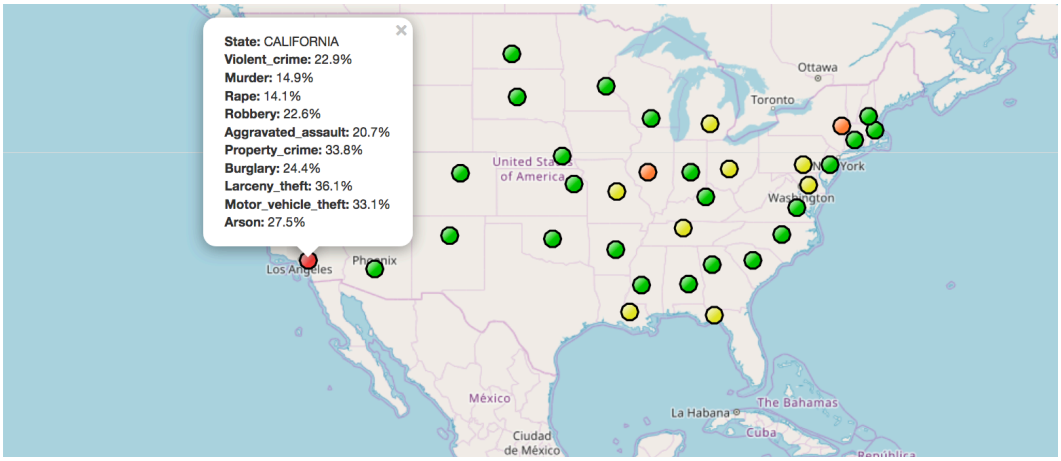
As for the NaNs on this data set, they were present in a different way: a crime could bring back a NaN, for example, in the columns related to information about the victim (gender , age, race, and so on ...), but this because the crime occurred did not involve victims (for example, fortunately, vehicle theft does not usually have victims). So we replaced the NaNs with a 0 which indicated that the information was not present.

Exploratory analysis

Our study has moved according to a zoom: we started from the general situation of the United States and then concentrate on the most affected State in the crime rate (California), finally studying its most populous city (Los Angeles). The thing that we found interesting initially was to see how the different crimes behaved between 2010 and 2016, state wise. Then we saw what the behavior of the states was for each year, seeing precisely, that the state with the highest crime rate was in all seven years California with a high gap from all states, even for Texas which, despite winning the second place, is located far from California. Then we focused our studies on California, particularly Los Angeles, seeing thanks a cluster how crimes were focused on the map. So we have studied what were the most dangerous roads, the most widespread crimes, the most used weapons, the gender of the victims killed according to each street and the average age of the victims according to gender. Finally we predicted the crimes for 2017 according to different linear regression algorithms and compared them with the actual data of 2017 to verify the reliability of our prediction.

The statically study we have done are as follow:

1) To calculate the mean values which shows the crime with the highest impact in USA:



2) To calculate the standard deviation which shows the spread of the crime over the years interval 2010-2016, with the following results:

The **most frequent crime from 2010-2016 in Los Angeles** is ARSON because it has the highest mean value and maximum frequency. In the considered year interval, our

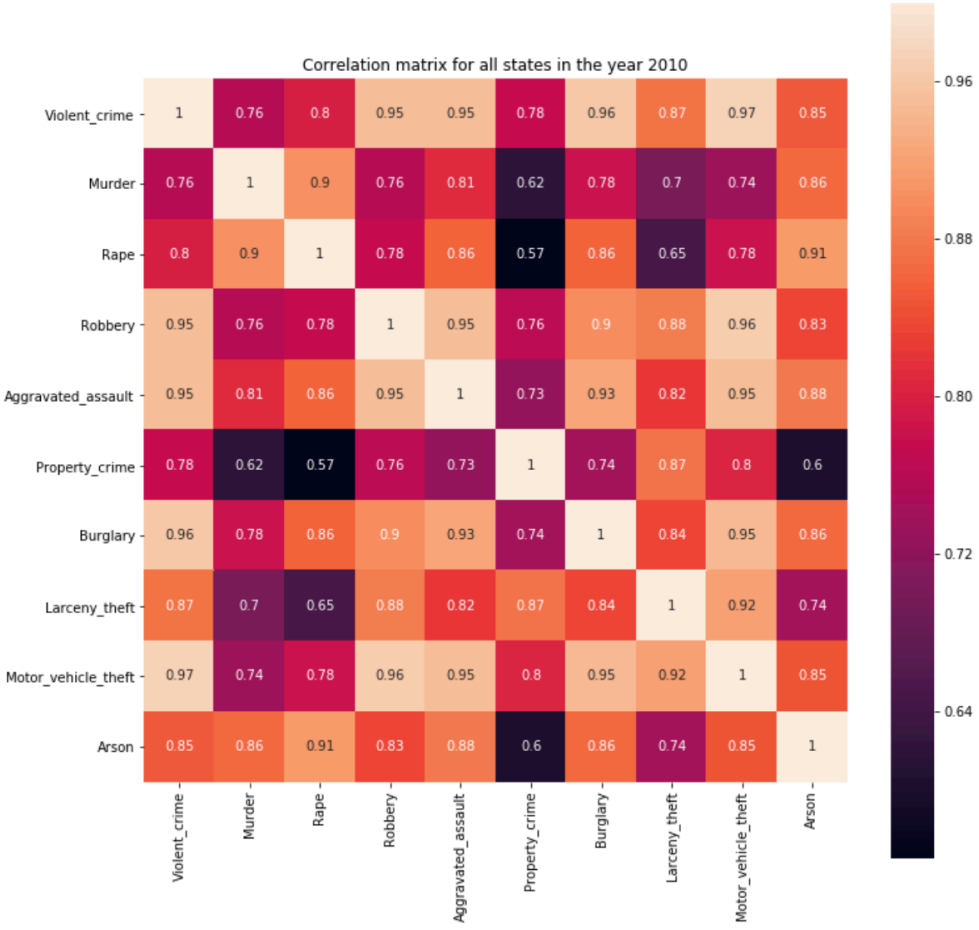
analysis reviewed that ARSON attained it's maximum occurrence in 2013 and minimum in 2015 in Los Angeles.

The crime BURGLARY has **the lowest standard deviation** this shows that the crime rate spread is close to the mean, that is, from 2010-2016, the rate of occurrence is close to each other.

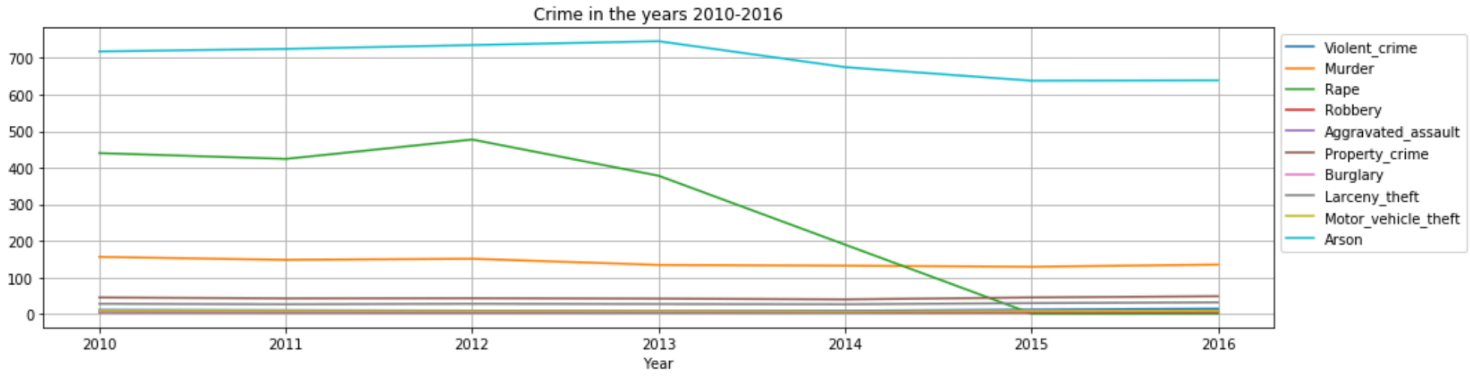
The **less frequent crime from 2010-2016 in Los Angeles** is ROBBERY because it has the lowest mean value. In the considered year interval, our analysis reviewed that ROBBERY attained it's maximum occurrence in 2010 and minimum in 2014 in Los Angeles.

The crime, RAPE has the **highest standard deviation** this shows a high fluctuation of the crime between 2010 - 2016. According our analysis, from 2013 to 2016 Los Angeles experience a sharp decrease in RAPE, this may have resulted from FBI modification of the definition for RAPE and other RAPE related crimes have been classified in another category of crimes not contained in our dataset.

3) The correlation matrix which shows how various crimes are related to each other. The crimes with a correlation value of more than 0.75 shows that the two crimes are strongly correlated that is the increase to one of the crime leads to the increase of the other crime and vice versa. This example shows the correlation matrix of crimes in America in the year 2010:



4) We calculated the Line Plot graph which clearly shows the increase and decrease of crimes in the given year interval 2010-2016. In the following example we can see crimes in Los Angeles in the given year interval 2010-2016



Unsupervised learning

For unsupervised learning we used the K-Nearest Neighbors clustering algorithm. It seemed to us wise to use this type of algorithm precisely because its purpose is to use a data set in which data points are separated into different classes to predict the classification of a new sampling point.

Moreover, since it is not parametric, it makes no assumptions about the distribution of the underlying data, and we have found this requirement to see how the crimes were distributed in the city, seeing in particular what kind of crimes took place in every street. In fact this type of algorithm is quite useful in the analysis of data in the "real world" or, more properly, in the analysis of data that do not obey the typical theoretical assumptions made (as in linear regression models, for example, that we have used precisely to predict according to a specific rule the number of crimes for a year of which we had no data).

Furthermore, K-Nearest Neighbors is also a lazy algorithm because it is an unsupervised learning algorithm, so there is no explicit training phase or it is very minimal (therefore quite fast).

The lack of generalization leads the algorithm to retain all training data, so all (or most) of training data is needed during the test phase and this unfortunately slows down processing and takes up a lot of memory. For this reason we have clustered on "only" 1000 elements, and also to make the map more legible for the purposes of the examination.

The results of this clustering could be used, in a more specific study, to understand how to arrange the police patrols in the city, to organize the streets more subject to certain crimes in order to control those who attend them or even preventing crime or limiting the damage that entails. Surely having a result with such a great visual impact makes the study of the phenomenon easily understandable and readable for anyone interested in consulting or studying a "map of the crimes of Los Angeles" (Figure 3).

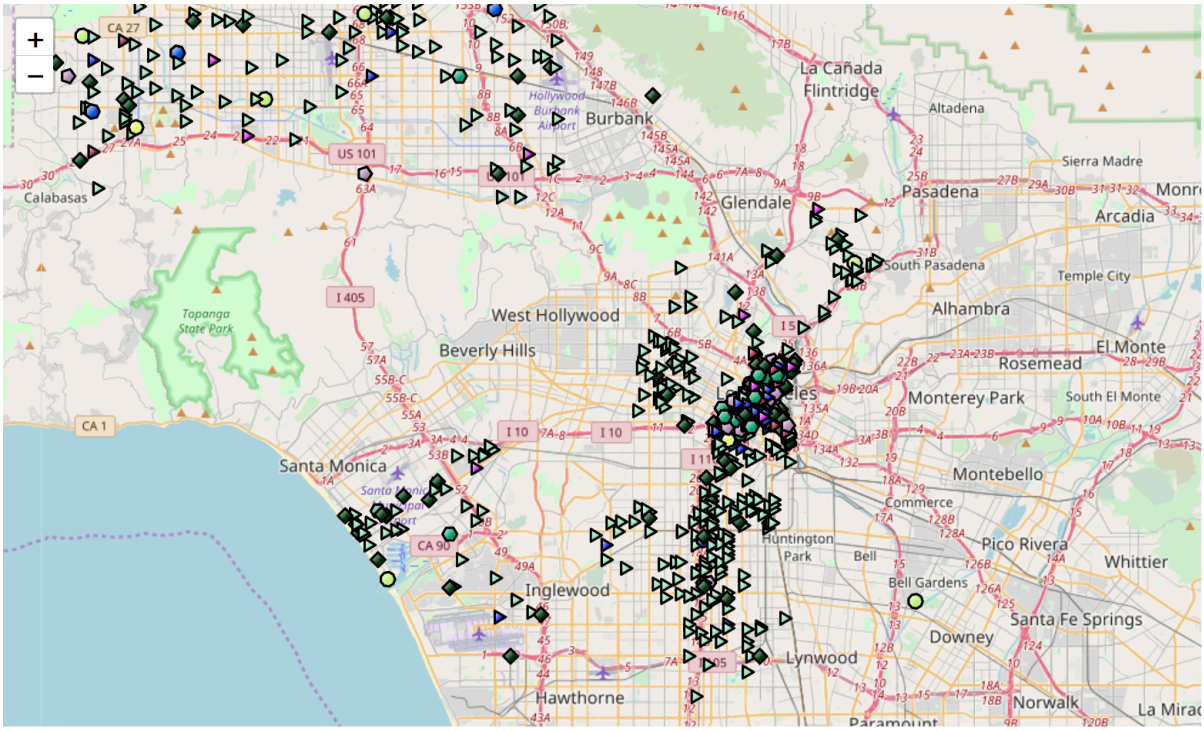


Figure 3: result of our clustering

Supervised learning

The goal of supervised learning is to learn a function that, given a sample of data and desired output, best approximates the relationship between input and output observable in the data. In other words supervised learning is typically done when we want to map input to a continuous output.

In our study on Supervised Learning we applied various Regression algorithms such as Regression Tree, Big Random Forest, Random Forest and Bagging to predict crime in 2017 we found out that the Regression is the best for prediction and has the greatest number among the different algorithms of regression which lead to the best fitting regression curve for

prediction. In conclusion from our study we can say that the Linear Regression is the best method for Supervised Learning .

As for the prediction made for 2017 in Los Angeles these are printed at the end of the notebook. As we have previously mentioned, we compared the results of our prediction with the prediction made by the FBI and we have had good results. Here is the result:

FBI prediction vs Our prediction

```
Data = Data[Data['City'] == 'LOS ANGELES']
Data[Data['Year'] == 2017]
```

	State	City	Year	Population	Violent crime	Murder	Rape	Robbery	Aggravated assault	Property crime	Burglary	Larceny theft	Motor vehicle theft	Arson
85	CALIFORNIA	LOS ANGELES	2017	4.007.905	14.44	138	1.212	5.165	7.925	50.122	8.252	32.363	9.507	651.0

ARSON Prediction: **408**
BURGLARY Prediction: **14423**
ROBBERY Prediction: **9040**
AGGRAVATED ASSAULT Prediction: **7516**

The comparison of the two forecasts naturally only shows the crimes present in both data sets (Los Angeles data set vs the FBI's prediction of the crimes done in Los Angeles in 2017 that was present in the first data set having the crimes of all United States).

Instead this is the result of the comparison of the real data concerning the crimes in 2017 in Los Angeles vs the prediction made by us for 2017. To not exaggerate with the images we put only a forecast, you can see the result of the forecasts in detail on the notebook.

Real Data for 2017 vs Our prediction

```
LA_2017 = DataSetLosAngeles[DataSetLosAngeles['Date_Occurred'].str.contains("2017")]
LA_2017['Date_Occurred'] = pd.to_datetime(LA_2017['Date_Occurred'])

LA_2017 = LA_2017[['Date_Occurred','Crime_Code','Crime_Code_Description','N']].groupby(by=['Date_Occurred','Crime_Code'
```

Prediction: **17872**

```
LA_2017[LA_2017['Crime_Code'] == 624].groupby(by=['Crime_Code','Crime_Code_Description'],as_index=False).sum()
```

	Crime_Code	Crime_Code_Description	N
0	624	BATTERY - SIMPLE ASSAULT	12701

Conclusions

As mentioned initially, our study surprised us because we never thought that the state with the highest crime rate was California, but that above all it had a crime rate so high compared to all the other states. The same surprise was given by the study itself: seeing the graphical results on the map, the tendency and the result of the prediction algorithms have made us the customers of our project a little bit, because we were able to experiment and study what we were interested in and it intrigued us, assigning us tasks that we had not initially foreseen, because we were driven by the curiosity and the questions that the study had aroused.

One of the most interesting results we were able to draw was that from our study, in reference to our Data Deceleration, in 2013 the FB1 UCR Program initiated the collection of rape data under a revised definition within the summary based Reporting System. The term Forcible was removed from the offense name, and the definition was changed to penetration this clearly explains why in our Old Dataset in general crimes rate reduced from 2013 -2016 and also there is a strong positive correlation between our new Data set and old Dataset which explains more why crime rates reduced from 2013-2016 , the new Dataset has many attributes which shows that many crimes has been redefined and this leads to a decrease in crime rate.

As we said in the introduction we chose this data set because we wondered if in the US the presence of the death penalty affected or not the crime rate. We found that California and Texas, respectively the states with the highest crime rate between 2010 and 2016 in the US, were still active, but in particular in California the statute of the death penalty was contested constitutionally, while Texas is the state with the highest number of executions in the US. In particular, in 2015, 86% of executions focused on Texas, Georgia and Missouri. On 6 November 2012, California voted on the possible abolition of the death penalty, maintaining it with 52% of the vote and since 2013 many courts have suspended the application in California. To date, Texas is considering adapting to the temporary suspension of the death penalty.