Università degli Studi di Pisa

Distributed Data Analysis and Mining

Crime Analysis

Emanuele Sciancalepore (678521)
Riccardo Pisano (677463)
Cristian Leone (673375)

Academic year 2024/2025

Contents

1 Introduction

In this project, we analyze crime statistics from the Los Angeles Police Department (LAPD), focusing on a comprehensive dataset that reflects incidents of crime in the City of Los Angeles dating back to 2020. This dataset[1], which includes 982,000 records and 28 features, is transcribed from original crime reports that were originally typed on paper, so there may be some inaccuracies within the data. Address fields are only provided to the nearest hundred block to maintain privacy. The data offers valuable insights into a wide range of crimes, from violent offenses like homicides and assaults to property crimes such as theft and burglary. We aim to explore crime trends, geographical variations, and temporal patterns, helping us understand the dynamics of criminal activity in Los Angeles during this period.

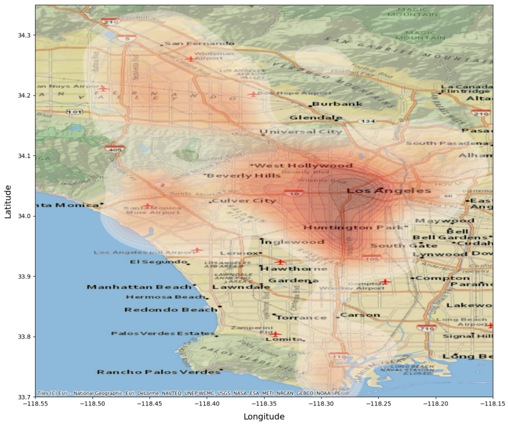
2 Exploratory Data Analysis

As can be seen from Table 1, the variables within the dataset are almost all categorical or string-based. For this reason, we have adopted specific techniques for the research questions that will be discussed in Section 4. It is important to emphasize that each row in the dataset represents a crime committed.

The first step we took was to check for missing values. Table 1 shows the number of missing values present in each column. Our goal was to fill them correctly and have all the data at our disposal, so we could work with a complete dataset that takes full advantage of Spark's large-scale computational capabilities. As we will see in Section 3, each variable was analyzed and managed with specific strategies.

From Figure 1, it can be observed how crimes are geographically distributed, being denser in the city center compared to the outskirts.

Regarding duplicates, we did not find any, with the help of the identification number DR_NO for our search. However, we noticed some inconsistencies in the data at this stage. For example, the variable "Vict Age" (although it has no missing values) contains about 150,000 records with improbable ages, such as 0 years old and negative ages. We assume these values were errors during the recording process. Another example of inconsistency is the similarity between "DATE_OCC" and "TIME_OCC." While "DATE_OCC" correctly reports the date and time of the crime, "TIME_OCC" only reports the time, but in an incorrect format (e.g., "DATE_OCC" = 07/23/2019 21:30 and "TIME_OCC" = 2130). Another inconsistency involves the variables within "Crm Cd," which are redundant compared to "Crm Cd 1." For all these inaccuracies (and many others), several changes have been made, as detailed in the next section.



Tiles (C) Esri -- National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC

Figure 1. LA heatmap

[1] Source:`https://www.kaggle.com/datasets/arpitsinghaiml/u-s-crime-dataset?resource=download&select=Crime_Data_from_2020_to_Present.csv`

Table 1. Description of variables in the dataset.

Name	**Description**	**Data Type**	**NaN**
DR_NO	Division of Record Number	String	0
Date Rptd	Date the crime was reported (MM/DD/YYYY)	Date	0
DATE OCC	Date the crime occurred (MM/DD/YYYY)	Date	0
TIME OCC	Time the crime occurred (in 24-hour format)	String	0
AREA	Numerical code of the geographic area (from 1 to 21)	String	0
AREA NAME	Name of the geographic area or patrol division	String	0
Rpt Dist No	Four-digit code representing a sub-area	String	0
Part 1-2 No	Indicates if the crime is a Part 1 (serious) or a Part 2 (less serious) offense	Number	0
Crm Cd	Crime code or classification number	String	0
Crm Cd Desc	Description of the crime code	String	0
Mocodes	Motivations or circumstances related to the crime	String	145262
Vict Age	Age of the victim (two-digit numerical value)	Number	0
Vict Sex	F - Female, M - Male, X - Unknown	String	138445
Vict Descent	Victim's descent code (e.g., H - Hispanic, W - White, B - Black, etc.)	String	138456
Premis Cd	Code indicating the type of premise, vehicle, or location where the crime occurred	Number	14
Premis Desc	Description of the provided premise code	String	585
Weapon Used Cd	Code indicating the type of weapon used in the crime	String	656471
Weapon Desc	Description of the weapon	String	656471
Status	Current status of the case	String	1
Status Desc	Description of the case status	String	0
Crm Cd 1	Additional Crime codes if applicable	String	11
Crm Cd 2	Additional Crime codes if applicable	String	913763
Crm Cd 3	Code for an additional crime, less severe than Crime Code 1	String	980327
Crm Cd 4	Code for a further crime, less severe than Crime Code 1	String	982574
LOCATION	Address of the crime incident rounded to the nearest hundred block	String	0
Cross Street	Cross street of the rounded address	String	830789
LAT	Latitude	Number	0
LON	Longitude	Number	0

3 Preprocessing

In this section, we will discuss all the operations performed on the data. Obviously, we are aware that with the decisions made, we might risk altering the original structure of the data too much, introducing a lot of synthetic data. Therefore, the following will be a list explaining all the modifications made:

- **Vict Age**: as we have already mentioned, it contains negative ages and many ages set to 0. In this situation, we decided not to manipulate the column's distribution too much by replacing these anomalous values with the appropriate proportions. A similar approach was applied to **Vict Sex** and **Vict Descent**, with particular attention to mapping the latter column to have more understandable values;

- **Weapon Used Cd and Weapon Desc**: both columns present the same number of missing values. Specifically, the field is empty when no weapon is used in the crime, so both columns will contain the strings "0" and "NO WEAP", respectively;

- **Premis Cd and Premis Desc**: we tried to find a correspondence between these columns, but unfortunately, there are codes in Premis Cd that are always associated with missing values, making it impossible to fill Premis Desc. In the absence of alternatives, we decided to replace them with the strings "000" and "UNKNOWN CODE", respectively;

- **Cross Street**:this column has about 90% missing instances. We tried filling it using the API, but even when the data is present, it is poorly formatted. Additionally, since we have many other geographic pieces of information like **LOCATION**, **LAT**, **LON**, and **AREA**, this column was deemed unnecessary for our analysis. Therefore, it was dropped;

- **Crm Cd**: this column is redundant compared to Crm Cd 1, so after making the necessary changes (see 'DataCleaning.ipynb'), we decided to remove it. Speaking of **Crm Cd 1, Crm Cd 2, Crm Cd 3, and Crm Cd 4**, when we encountered missing values, we filled them with 0 (since some individuals may only have one charge);

In addition to these changes, we added two columns that might be useful in the future: "**season**" and "**time_of_day**". We created them using the information contained in **DATE OCC** (from which the time was removed, as it is already reported) and **TIME OCC** (which was converted into a date format).

A separate discussion is necessary for the column **Mocodes**. For this variable, we decided to experiment with a Data-Driven approach. We believe that, by leveraging the information available in our dataset, it is possible to reconstruct the modus operandi of a crime. Therefore, we chose to use Pattern Mining (FP-Growth).

We made changes to some variables used as input, which can be observed in the code ('PM_Mocodes.ipynb'), such as binning and encoding on multiple columns. Among the generated association rules, we retained only those with a 'lift' greater than a threshold (lift ¿ 1) and with **Mocodes** values in the 'consequent' but not in the 'antecedent.' An example follows:

$$\textbf{Spousal abuse-Simple Assault , Hispanic/Latin/Mexican} \longrightarrow \textbf{0913\_mocodes}$$

This association rule demonstrates how the variables 'Spousal abuse-Simple Assault' and 'Hispanic/Latin/Mexican' can be correlated with the variable 0913 in **Mocodes** (which means 'Victim knew suspect'). This shows that, in this example, a logical pattern can exist within the association rules.

The resulting new table was then converted into a dictionary, which we used to fill the missing **Mocodes** values in the original dataset.

4 Learning

4.1 Q1: Which ethnicities are most affected?

For this Research Question, we tried to implement several mini-classification tasks with the target column being the **Vict Descent** variable, which contains strings representing the victim's ethnicity. The chosen (base) model was a *Sequential Multilayer Neural Network*, starting with a dense layer of 256 units with ReLU activation and L2 regularization. This is followed by a series of three dense layers with 128, 64, and 32 units, respectively, each with ReLU activation and L2 regularization. There is a 50% dropout rate to reduce overfitting. The final layer is dense, with a number of units equal to the number of classes and a softmax activation function. However, this model was simply a baseline and was slightly modified based on the classification task and the changes in the data. The columns used for this section are the following: [**'AREA NAME', 'Mocodes', 'Vict Descent', 'Crm Cd Desc', 'Vict Sex', 'Premis Desc', 'Weapon Desc', 'Status Desc', 'time_of_day', 'season'**]. Most of these are categorical columns, which we transformed into numeric values using *StringIndexer*. The results of the various tasks can be seen in Table 2.

Data	Neurons	Accuracy	Activation
Unbalanced	256/128/64/32	0.47	relu/softmax
Weighted	256/128/64/32	0.002	relu/softmax
4 Clusters	256/128/64/32	0.46	relu/softmax
SouthAmericaVsAll	256/128/64/32	0.62	relu/sigmoid

Table 2. Classification Table

As shown in the table, the results are far from satisfactory. The performance of the *neural network* is excessively tied to the data distribution. In the first task, the "Unbalanced" case in Table 2, only a few values exceed an accuracy/F1-Score level higher than 0.00. In fact, values such as Hispanic, Black, and White (being the most common and popular) reach F1-scores of 0.60, 0.38, and 0.07, respectively, while the rest is set to 0.00 or slightly above.

The situation does not seem to improve with the Weighted dataset, where the *neural network* seems to perform poorly, achieving an overall accuracy of 0.002.

In the third task, where the dataset is divided into four clusters (as observed in 'ResearchQ1.ipynb'), the performance seems to improve slightly. In our opinion, the accuracy is similar to the first task due to the distribution of the neural network's training variables, even though the target column is different (e.g., the first cluster, 'Sud American', has the same F1-score as Hispanic in the first task, as in the cluster division, in fact all the element of Sud American are Hispanic).

The best results are obtained in the fourth and final task, where the classification is binary, labeling everything that is not "Sud American" as "Rest." In this case, the "Rest" class, being the least represented, manages to achieve an F1-score of 0.4.

4.2 Q2: Is there a seasonal pattern in crime types?

In this analysis, we aimed to determine whether there is a relationship between crime types and seasons. To answer this question, we performed an unsupervised analysis using clustering algorithms, specifically *K-Means* and Bisecting *K-Means.*

The first step involved a targeted selection of a few relevant variables, which were converted into numerical values where necessary. These variables were then vectorized, combining them into a single vector, and finally normalized to ensure uniformity in scale and proper interpretation by the clustering algorithms.

With the data prepared for clustering, we conducted several experiments to identify the optimal number of clusters (k), evaluating the results using the Silhouette Score and the Elbow Method. Both algorithms were tested, but *Bisecting K-Means* showed the best performance with $k = 17$ and a Silhouette Score of 0.3746. Based on these parameters, we divided the data into 17 clusters and subsequently analyzed the distribution to identify potential seasonal trends.

The analysis reveals that the distribution of crimes across seasons is relatively homogeneous, with a consistent predominance of certain clusters, particularly clusters 0, 9, 10, and 15, in all seasons. Cluster 0 was associated with assaults and battery, cluster 9 with burglary and serious crimes, cluster 10 with vehicle theft and related offenses, and cluster 15 with petty theft and animal cruelty. This homogeneity suggests that there is no evident seasonal trend and that crimes tend to be distributed similarly regardless of the season.

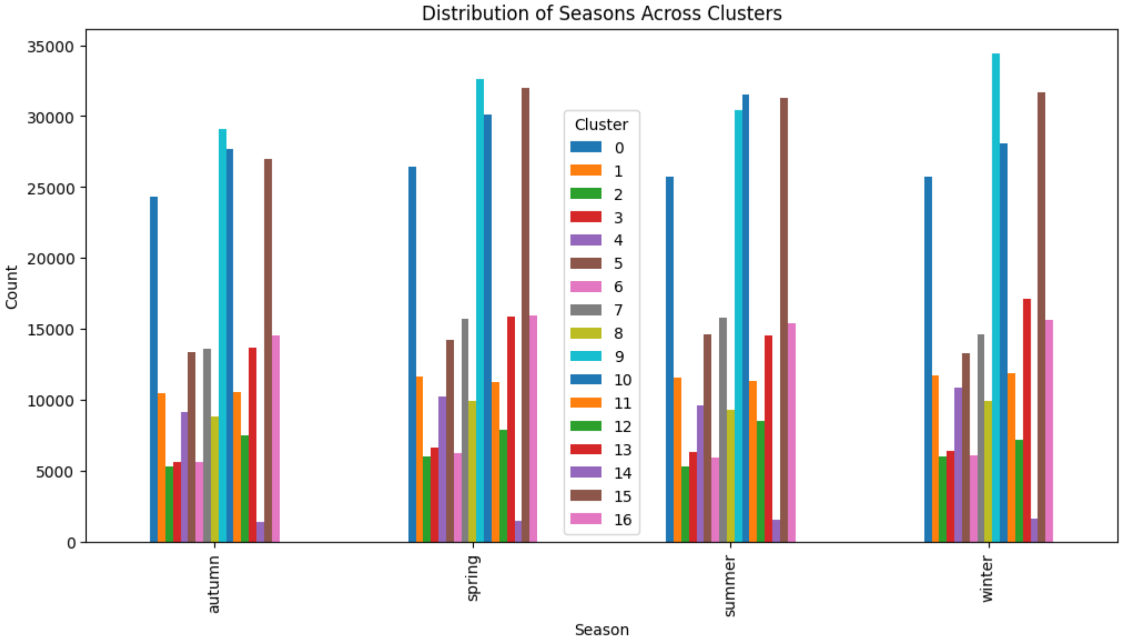


Figure 2. Distribution of crimes across seasons and clusters.

4.3 Q3: Can cluster be identified based on crime types?

In this subsection, we focused on the variable **Crm Cd Desc**, which includes many types of crimes (about 140 categories). In Image 4, we show the top 15 most frequent crimes in Los Angeles during the period we analyzed. Car theft is the most common, followed by simple assault and burglary from vehicles. Lower in the list, we noticed some crimes that are very similar but differ only in the extent of the damage. Our goal here was to reduce the variability of this variable to prepare for the next research question in subsection 4.4. To do this, we used two algorithms, *Latent Dirichlet Allocation (LDA)* and *K-Means*, both applied to textual data that had undergone the same preprocessing.

First, we used a PDF from the Los Angeles Police Department portal to map the **Mocodes** variable into textual descriptions. This gave us a better understanding of how crimes were classified. Then, we worked on the textual data about the crime description, the modus operandi, the premise description, and the weapon used. We cleaned the data by removing unnecessary details, like parts in parentheses, numbers (which described damage levels in theft but were not useful), and words like articles, diminutives, and symbols that didn't add meaning. This step was quite important to giving the algorithms cleaner and more meaningful information, making it easier to find useful patterns.

After preparing the data, we ran the algorithms with different values of k (number of clusters or topics to identify). For each value, we calculated various metrics: perplexity and coherence score for *LDA*, and SSE (Sum of Squared Errors) and silhouette score for K-Means. Both graphs agreed on the optimal number of clusters, which is 20. The performance results were: coherence score (cv_score) = 0.56, perplexity = 4.4, and silhouette score = 0.09.

The metrics for *LDA* suggest a decent thematic coherence, with clusters that

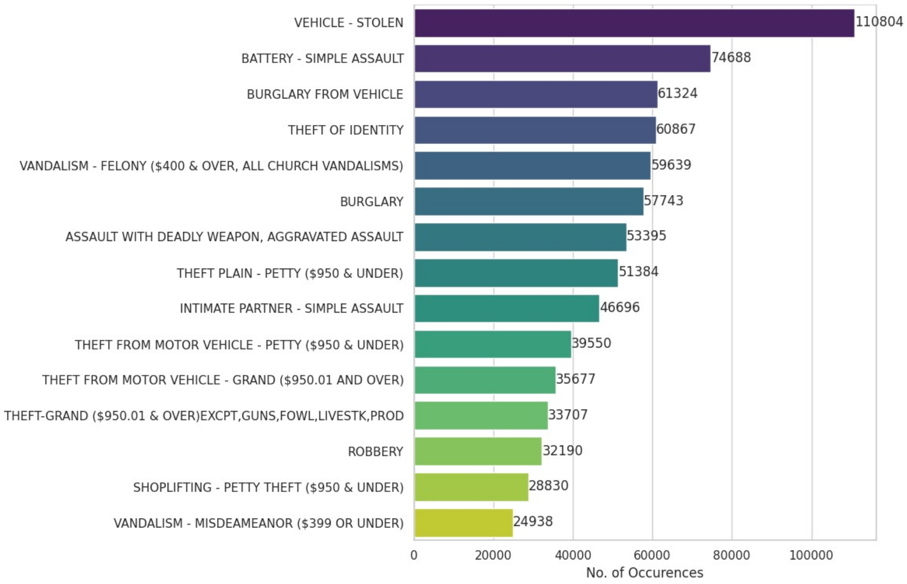


Figure 3. Top 15 crimes

capture meaningful patterns. A quick semantic analysis of the clusters confirmed the algorithm's ability to identify consistent groups. For example, analyzing the topic with the highest number of records, characterized by terms such as vehicle, property, removes, street, stolen, theft, motor, suspect, grand, order, it became clear that this topic refers to vehicle and property thefts. On the other hand, *K-Means* showed a very low silhouette score (0.09), indicating weak separation between clusters and significant overlap among categories. This issue is also reflected in the data distribution: the largest cluster in *K-Means* contains about 470,000 records (about 50% of the total), compared to LDA's largest cluster with 190,000 records. While *K-Means*' larger clusters suggest a lack of clear separation, the smaller ones might represent more specific groups. In contrast, *LDA*'s more balanced and homogeneous clusters seem better for identifying distinct themes.

To evaluate which algorithm works better with our data, we plan to train a classification model using **Crm Cd Desc** as the target variable, clustered first with *LDA* and then with *K-Means*, and compare their performance. This approach will help us decide which segmentation method is more useful for practical purposes.

4.4 Q4: How can the type of crime be identified in cases of incomplete reports?

This section is related to the previous one, as we used the string created in the previous section (removing all the crime-related information to avoid redundancy and overfitting). The task is divided into two sub-tasks: *LDA Cluster Classification* and *K-means Classification*. The names assigned to the labels were determined based on our domain knowledge.

For this task, we decided to use a **Convolutional Neural Network** (CNN), given its suitability for working with textual data. The texts were preprocessed in the same way as in Q3, similarly for both mini-tasks.

The data, performances, and characteristics of the CNN are as follows:

- **LDA Cluster Classification**: In this task, we used a CNN consisting of an initial embedding layer followed by a convolutional layer and a pooling layer. Then, we created two dense layers with 64 and 32 neurons, respectively, each regularized using L2 regularization. Dropout was applied after each layer, set at 30% and 20%. In this case, the classification yielded very satisfactory results: the overall accuracy is 0.85. The model is not only satisfactory in terms of performance but also for correctly classifying less-represented labels (e.g., 'Sexual Offenses', with only 3,000 instances, achieved an f1-score of 0.83).

- **K-Means Classification**: In this task, we used a slightly different CNN model: here, the dense layers consist of only two layers with 128 (embedding) and 64 neurons, respectively, using two dropout layers set at 0.5. Once again, the performance is very satisfactory, with the model achieving an overall accuracy of 0.84. However, the accuracy and f1-scores seem more extreme compared to LDA, as the smaller clusters show much higher accuracy than the larger ones. Therefore, we believe this is due to the similarity of instances within the same cluster—larger clusters tend to mix records that are not as similar to each other.

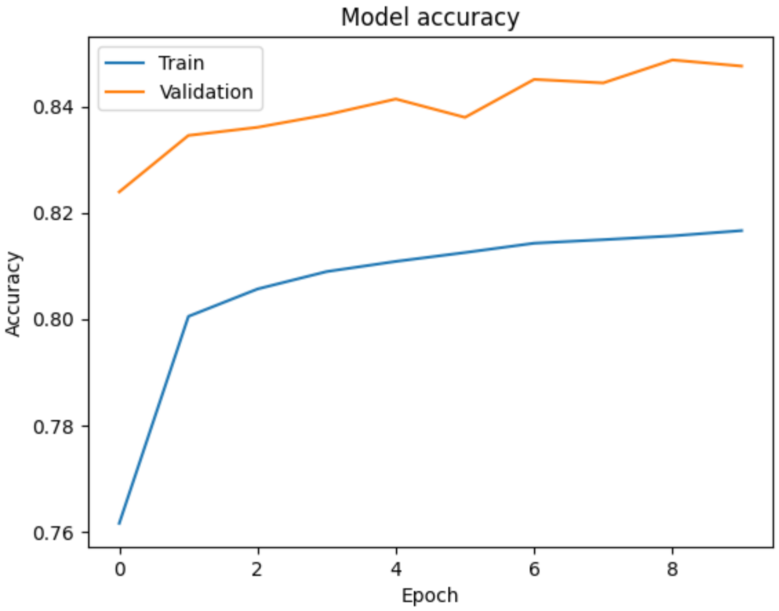


Figure 4. LDA Learning curve

5 Conclusion

This project has allowed us to explore various clustering and classification techniques applied to a large-scale dataset of crime incidents from Los Angeles. Leveraging Spark, we managed to process a dataset with over one million records efficiently.

However, we encountered certain limitations, primarily due to the absence of specific libraries that restricted the implementation of more advanced techniques.

Among all the analyses, the one that convinced us the most was **Q4**, particularly the clustering performed using *Latent Dirichlet Allocation (LDA)*. This approach showed better performance overall, primarily because the resulting classes were more balanced compared to *K-Means*. With *K-Means*, one label dominated the dataset, covering a significant portion of the records. In contrast, *LDA* provided a more equitable segmentation, making it more effective in identifying distinct and meaningful patterns in the data.