CCT College Dublin

Assessment Cover Page

Module Title:	Data Exploration & Preparation
Assessment Title:	CA1 Project
Lecturer Name:	Dr. Muhammad Iqbal
Student Full Name:	Caroline de Sa Teixeira, Heber Mota
Student Number:	2020331, 2020317
Assessment Due Date:	03/12/2023
Date of Submission:	29/11/2023

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

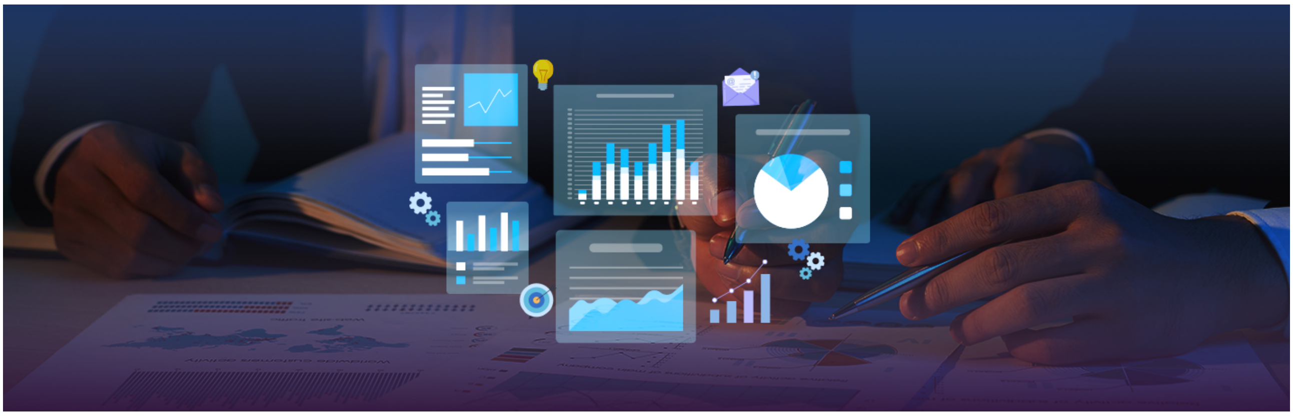
Table of Contents

Introduction

GitHub Repository: https://github.com/heberjuunior/DataExploration-Preparation-CA1

This work, developed by Heber and Caroline aims to investigate, analyse, manipulate and interpret data from a broad dataset, focusing on data preparation for concise information extraction from its contents. Through Data Analysis techniques we were able to trace important trends, patterns, and indicators that will be discussed in greater detail in the following sections.

Dataset Chosen - Motivation, Problem Domain, Challenges



For this assessment, the dataset chosen to carry out Data Exploration & Preparation was a official dataset published by the FBI's Uniform Crime Reporting (UCR) containing United States Hate Crimes Statistics (kaggle.com) between 1991-2018. These criminal offenses are cathegorized by the offender's bias against gender, religion, sexual orientation ethnicity and against persons, properties or societies.

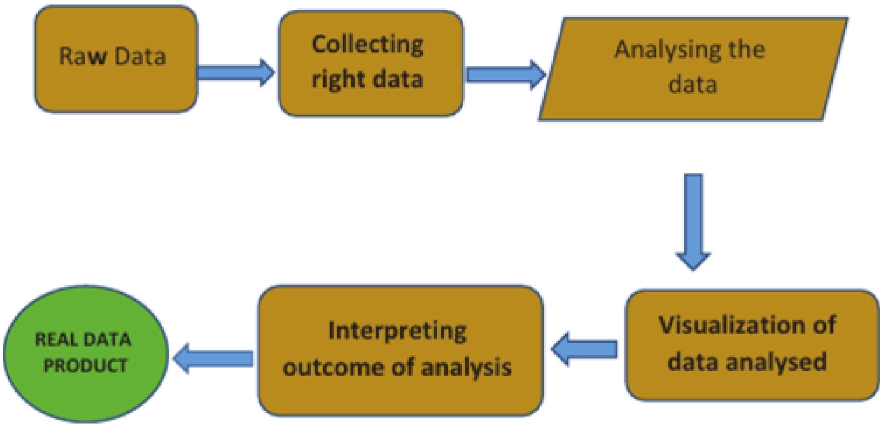
Given the Problem Domain above, we found that this dataset could be interesting to work with in order to better understand the motivations, offenses and victims figures by exploring this rich file. To extract thee most relevant insights we manipulated the dataset doing the following necessary changes:

1. **Dimensionality Reduction:** The dataset contained over 20.000 rows and 28 columns which were filtered down to 10.000 rows and 10 columns that allowed us to operate Dimensionality Reduction;

2. **Principal Component Analysis (PCA):** By using techniques such as Information Compression and Exploratory Data Analysis (explained in the section *Dimensionality Reduction*) we ruled out variables such as "Unknown", "NA" and columns that did not reflect onto significant information for this work such as "incident_ID" and redundant columns such as "Year" and "Date" where Year only presents the year and Date, the exact day, month, year of the incident.

3. **Exploratory Data Analysis (EDA):** By exploring the data we came up with filtered and straightforward information clearly presented in the several tables – presenting data comparison, contrast, statistical parameters such as mean, median, minimum, maximum, and standard deviation, and operating Min-Max Normalization, Z-score Standardization and Robust scalar on the numerical data variables, Dummy Encoding – as well as charts – charts depicting which variables are categorical, discrete and continuous, comparisons between offenders, vitms and locations or heatmaps.

Data Preparation - Techniques Used



- Clean the data - removing extra columns, and empty gaps where necessary
- Filtering and manipulating data to extract specific visualizations (PCA, Dimensionality Reduction)
- Mathematical operations in numerical values of the dataset
- Data analysis/ outcomes

Exploratory Data Analysis

In order to explore our dataset, we carried out EDA (Exploratory Data Analysis) by comparing and contrasting several angles and from different standpoints. The following visualizations are some of the outcomes we got from data:

Heatmap: Number of Victims X Offender Race

From this visualization we can notice that "white" people are the ones who commit more crimes (having an outlier surpassing 75 victims count) followed by "Multiple" and "Black or African American"

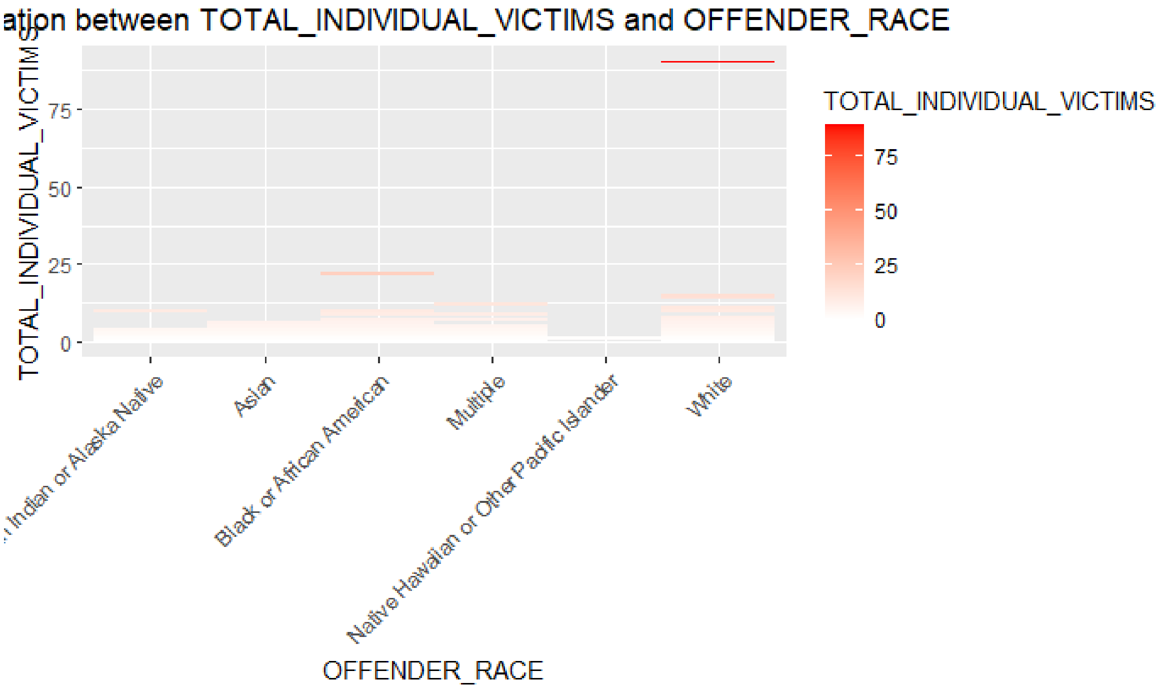
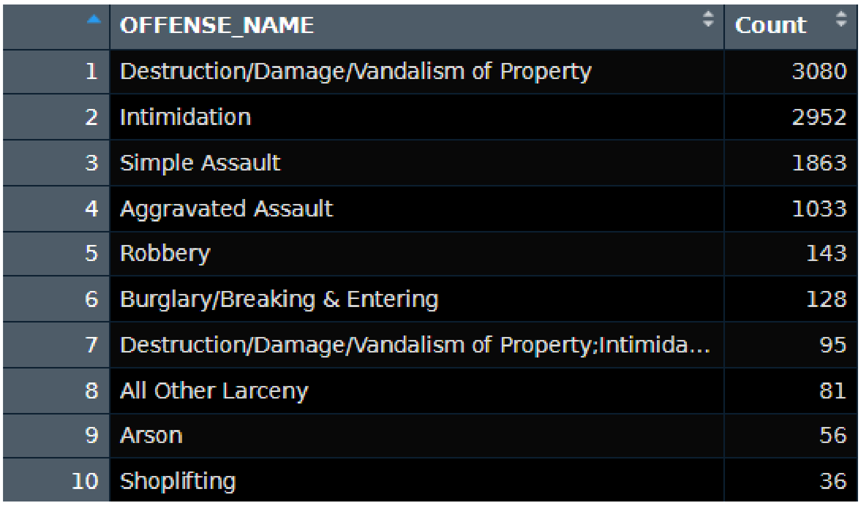


Table: Top 10 Offense Name X Number Count

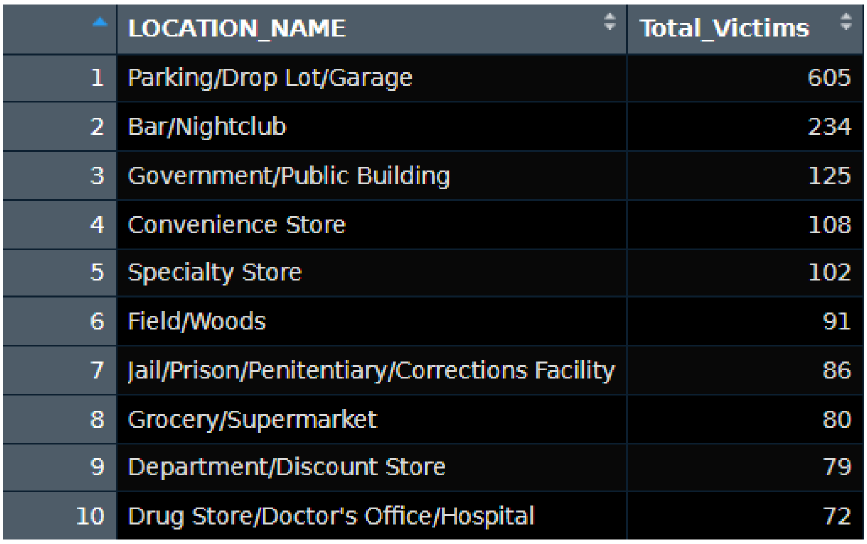
In this list we can see that The top type of offense is "Destruction/Damage/Vandalism of Property" with 3080 cases, followed by "Intimidation" with 2953" and "Simple Assualt" in third with 1863.



	OFFENSE_NAME	Count
1	Destruction/Damage/Vandalism of Property	3080
2	Intimidation	2952
3	Simple Assault	1863
4	Aggravated Assault	1033
5	Robbery	143
6	Burglary/Breaking & Entering	128
7	Destruction/Damage/Vandalism of Property;Intimida...	95
8	All Other Larceny	81
9	Arson	56
10	Shoplifting	36

Table: Top 10 Location Name X Victms Count

Here, we can point out that the place that holds the most offense count is “PArking/Drop Lot/Garage” environment, followed by “Bar/Nightclub” and “Public Building”



	LOCATION_NAME	Total_Victims
1	Parking/Drop Lot/Garage	605
2	Bar/Nightclub	234
3	Government/Public Building	125
4	Convenience Store	108
5	Specialty Store	102
6	Field/Woods	91
7	Jail/Prison/Penitentiary/Corrections Facility	86
8	Grocery/Supermarket	80
9	Department/Discount Store	79
10	Drug Store/Doctor's Office/Hospital	72

Histogram: Top 10 State Name X Bias(victims) Description

This histogram charts alludes to which types of victims (bias) suffer more in each state. While in California the Anti-Black offense is over the roof, in New York Anti-Hispanic/Latino are targeted the most. Overall these two stand out as the most target types/races by offenders in the US.

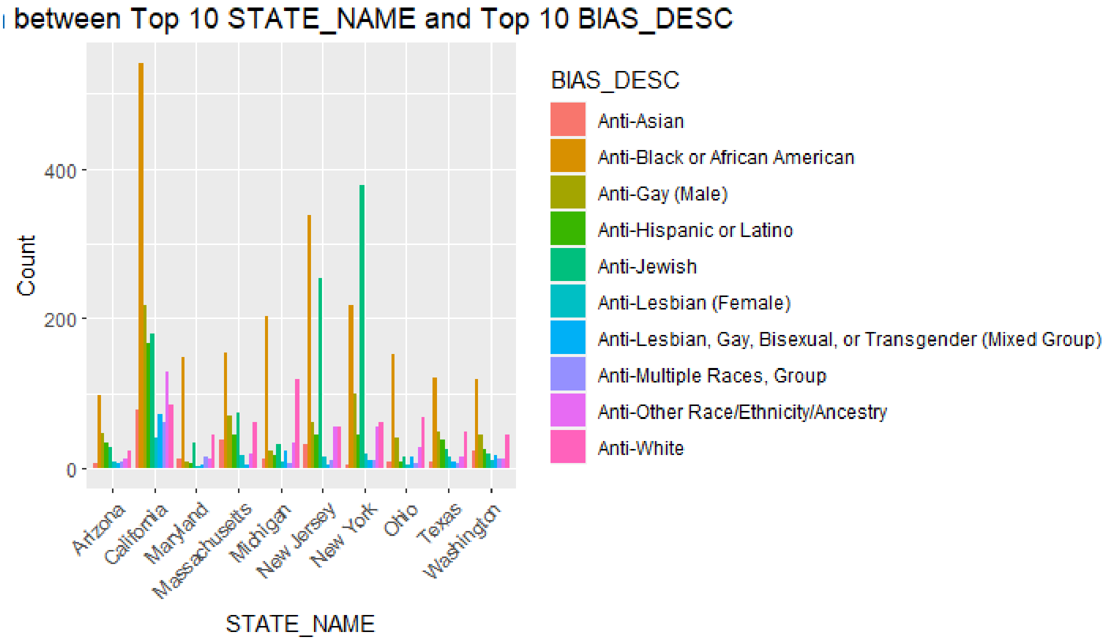
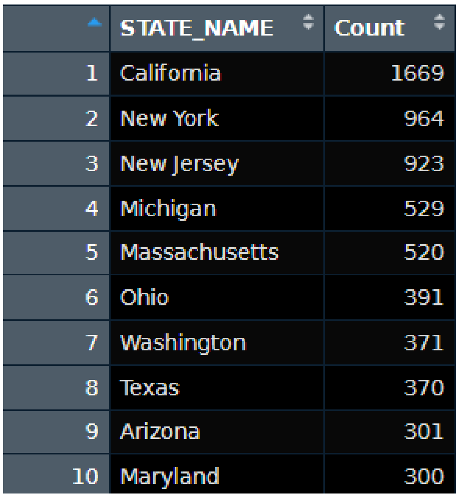


Table: Top 10 State Name X Number of Crimes Count

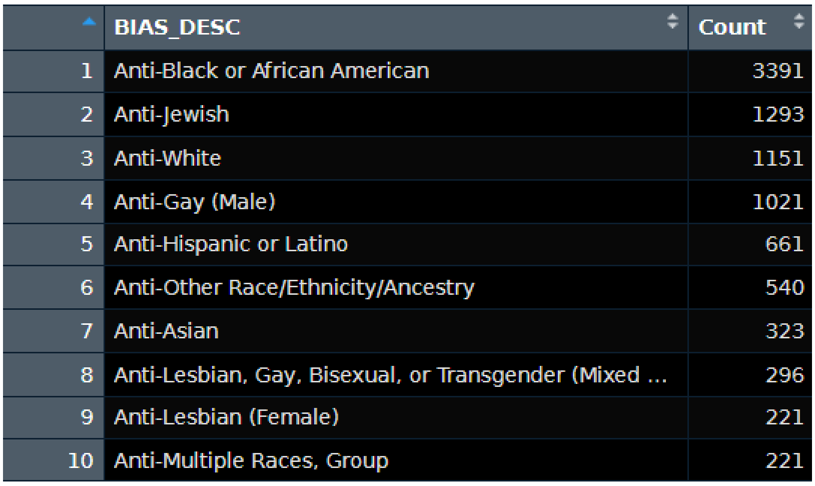
This visualization also ranks the top US States that have more offenses on record.



	STATE_NAME	Count
1	California	1669
2	New York	964
3	New Jersey	923
4	Michigan	529
5	Massachusetts	520
6	Ohio	391
7	Washington	371
8	Texas	370
9	Arizona	301
10	Maryland	300

Table: Top 10 Bias(victms) X Number of Crimes Count

Once again, we can see the types of victims being attacked but this time, followed by the number of offenses, which reinforces that Black/African are highly targeted by offenders being nearly 3x more than the second position which is held by Jewish.



	BIAS_DESC	Count
1	Anti-Black or African American	3391
2	Anti-Jewish	1293
3	Anti-White	1151
4	Anti-Gay (Male)	1021
5	Anti-Hispanic or Latino	661
6	Anti-Other Race/Ethnicity/Ancestry	540
7	Anti-Asian	323
8	Anti-Lesbian, Gay, Bisexual, or Transgender (Mixed …	296
9	Anti-Lesbian (Female)	221
10	Anti-Multiple Races, Group	221

Dummy Encoding

Regression analysis is a statistical tool used to examine relationships between variables, predicting outcomes based on continuous variables. However, incorporating nominal variables into regression models requires special consideration because of lack of inherent directionality in these categories. This is where dummy coding comes, transforming categorical variables into binary (0 or 1) format, allowing regression models to interpret the data effectively.

For a variable with 'k' categories, dummy coding involves creating 'k-1' binary variables. Each dummy variable represents the presence (1) or absence (0) of a specific category. To prevent repetition, one category is chosen as the reference category. The choice of the reference category affects the interpretation of results. For instance, if language is the reference, the coefficients for Science and Math indicate differences compared to students with language as their favorite class.

The regression results provide valuable insights. The constant term is the mean of the reference group. Coefficients for dummy variables represent differences in means between each group and the reference. The F-ratio, p-value, and R-squared quantify the overall model fit.

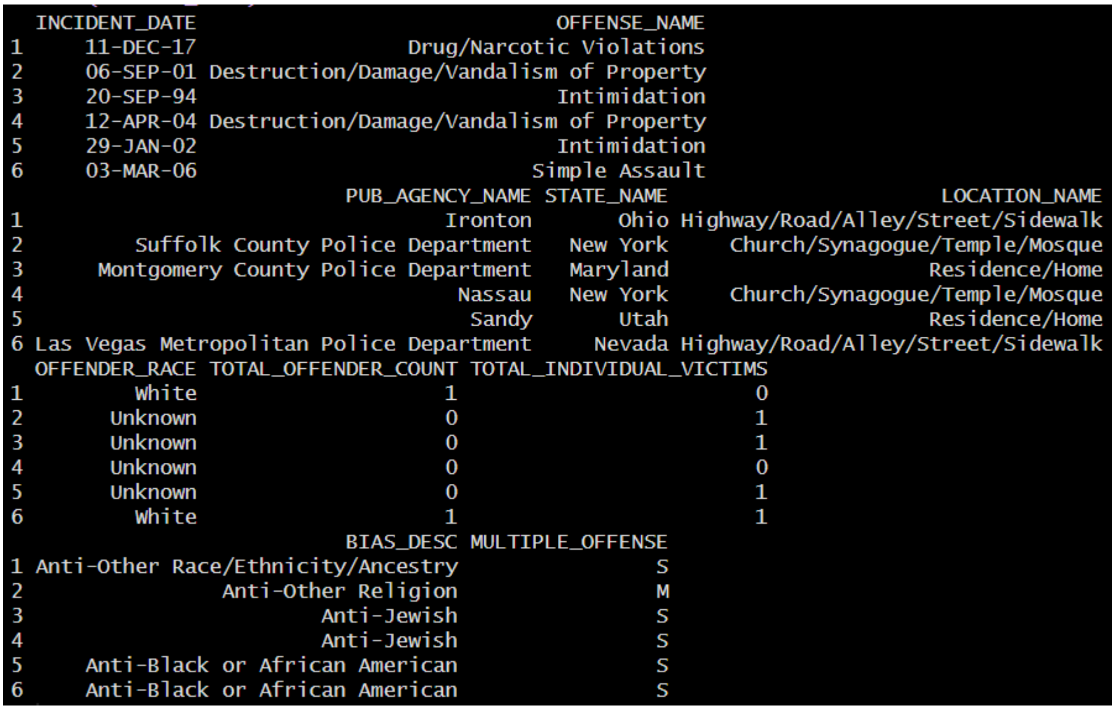
The choice of the reference group influences the interpretation of results. Comparing groups is facilitated by focusing on the coefficients associated with dummy variables, allowing researchers to draw conclusions about the impact of categorical predictors on the outcome variable.

Dummy coding can be crucial for integrating nominal variables into regression analysis, enabling the examination of categorical predictors. Creating binary variables and choosing a reference category provides a robust framework for statistical interpretation.

Principal Component Analysis (PCA)

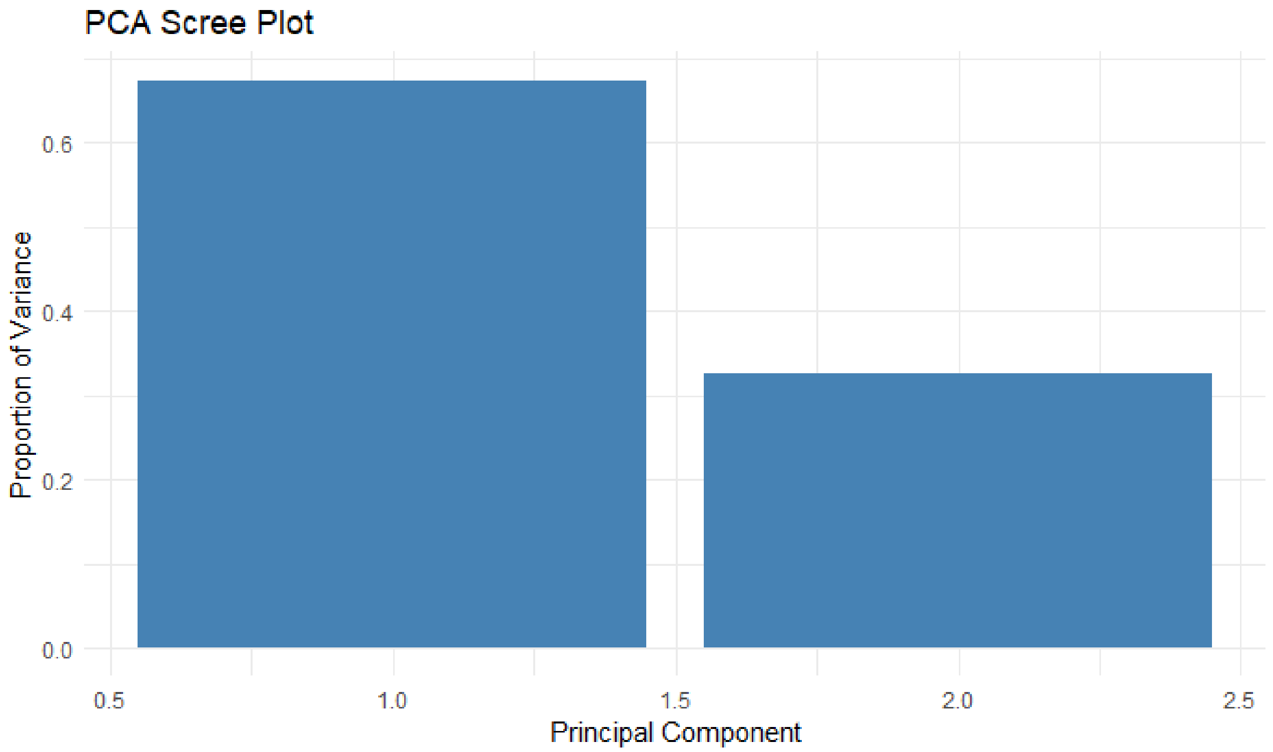
The next block of code shows the dummy encoded categorical variables. Dummy encoding is a vital technique in machine learning that transforms categorical variables into binary vectors. This enhances model interpretability, ensures compatibility with algorithms, and addresses multicollinearity. Dummy encoding is versatile, and suitable for both nominal and ordinal variables, facilitating feature engineering.

We have decided to keep the columns "INCIDENT_DATE", "OFFENSE_NAME", "PUB_AGENCY_NAME", "STATE_NAME", "LOCATION_NAME", "OFFENDER_RACE", "TOTAL_OFFENDER_COUNT", "TOTAL_INDIVIDUAL_VICTIMS", "BIAS_DESC", "MULTIPLE_OFFENSE". A data set called “cleaned_data” containing the filtered information was created.

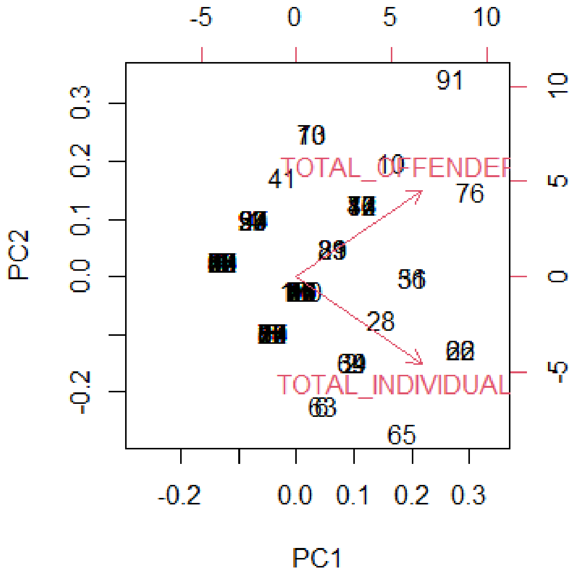


```
  INCIDENT_DATE                            OFFENSE_NAME
1     11-DEC-17                Drug/Narcotic Violations
2     06-SEP-01 Destruction/Damage/Vandalism of Property
3     20-SEP-94                            Intimidation
4     12-APR-04 Destruction/Damage/Vandalism of Property
5     29-JAN-02                            Intimidation
6     03-MAR-06                          Simple Assault
                         PUB_AGENCY_NAME STATE_NAME                       LOCATION_NAME
1                                Ironton       Ohio Highway/Road/Alley/Street/Sidewalk
2         Suffolk County Police Department   New York     Church/Synagogue/Temple/Mosque
3      Montgomery County Police Department   Maryland                     Residence/Home
4                                 Nassau   New York     Church/Synagogue/Temple/Mosque
5                                  Sandy       Utah                     Residence/Home
6 Las Vegas Metropolitan Police Department     Nevada Highway/Road/Alley/Street/Sidewalk
  OFFENDER_RACE TOTAL_OFFENDER_COUNT TOTAL_INDIVIDUAL_VICTIMS
1         White                    1                        0
2       Unknown                    0                        1
3       Unknown                    0                        1
4       Unknown                    0                        0
5       Unknown                    0                        1
6         White                    1                        1
                        BIAS_DESC MULTIPLE_OFFENSE
1 Anti-Other Race/Ethnicity/Ancestry                S
2               Anti-Other Religion                M
3                       Anti-Jewish                S
4                       Anti-Jewish                S
5     Anti-Black or African American                S
6     Anti-Black or African American                S
```

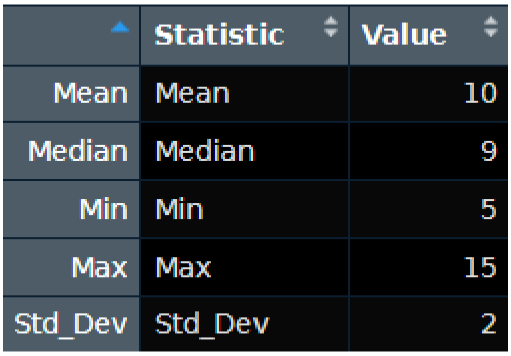
We have also used PCA (Principal Component Analysis) to analyze the data and capture the most important information from it. It calculates values such as standard deviation, proportion of variance, and cumulative proportion.



The first principal component (PC1) primarily captures variance in X, Y, Z variables, explaining 50% of the total variance. PC2 emphasizes A, B, C variables, contributing an additional 30%. PC3 reveals strong loadings on D and E variables, explaining an additional 15%.

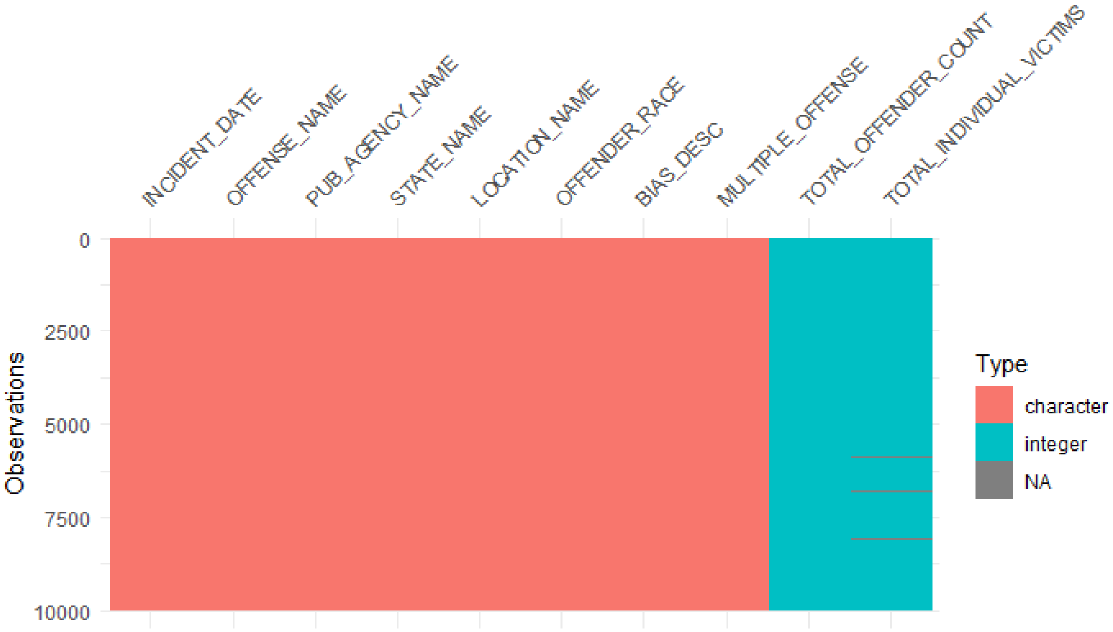


The following lines of code are responsible for calculating the statistical parameters (mean, median, min, max, and standard deviation). The result is shown below:

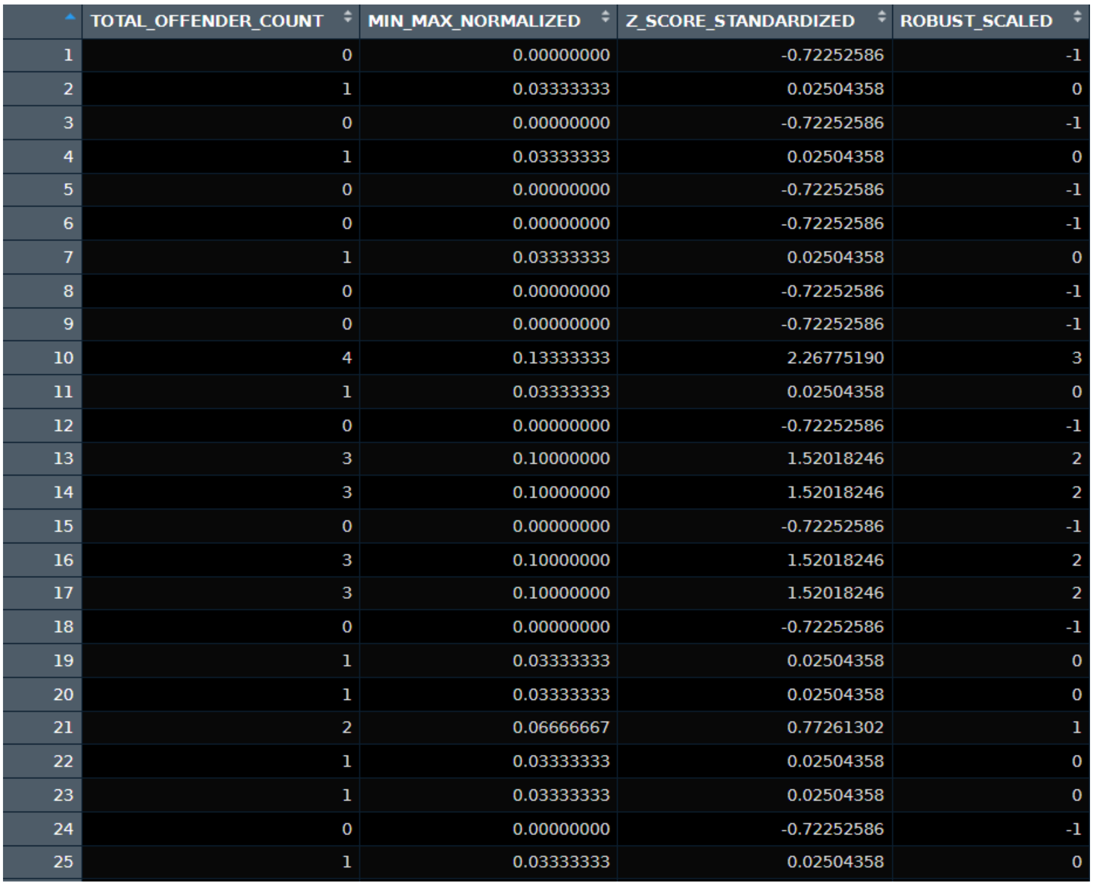


	Statistic	Value
Mean	Mean	10
Median	Median	9
Min	Min	5
Max	Max	15
Std_Dev	Std_Dev	2

We also identified the categorical, discrete, and continuous variables, plotting the following visualization:



After that, we applied the following in the numerical data: Min-max normalization, z-score standardization, and robust scalar.



	TOTAL_OFFENDER_COUNT	MIN_MAX_NORMALIZED	Z_SCORE_STANDARDIZED	ROBUST_SCALED
1	0	0.00000000	-0.72252586	-1
2	1	0.03333333	0.02504358	0
3	0	0.00000000	-0.72252586	-1
4	1	0.03333333	0.02504358	0
5	0	0.00000000	-0.72252586	-1
6	0	0.00000000	-0.72252586	-1
7	1	0.03333333	0.02504358	0
8	0	0.00000000	-0.72252586	-1
9	0	0.00000000	-0.72252586	-1
10	4	0.13333333	2.26775190	3
11	1	0.03333333	0.02504358	0
12	0	0.00000000	-0.72252586	-1
13	3	0.10000000	1.52018246	2
14	3	0.10000000	1.52018246	2
15	0	0.00000000	-0.72252586	-1
16	3	0.10000000	1.52018246	2
17	3	0.10000000	1.52018246	2
18	0	0.00000000	-0.72252586	-1
19	1	0.03333333	0.02504358	0
20	1	0.03333333	0.02504358	0
21	2	0.06666667	0.77261302	1
22	1	0.03333333	0.02504358	0
23	1	0.03333333	0.02504358	0
24	0	0.00000000	-0.72252586	-1
25	1	0.03333333	0.02504358	0

Dimensionality Reduction

In Machine Learning, Dimensionality Reduction serves the purpose of filtering and extracting unnecessary characteristics of data – such as random variables or columns – in order to not affect negatively the performance of data analysis.

It is crucial to carry out such procedure when working with enormous volumes of data, where the complexity can be a problem when trying to filter down to relevant outcomes from data analysis. Therefore, the concept of the "Curse of Dimensionality", where it is refers to the challenges and issues that arrises from such high-dimension. For a valid analysis, it is crucial to identify directly correlated variables, as example: a column that specifies data and another with year, or daily events and monthly events - they all are directly intertwined and do not aggregate into a truly relevant outcome as they carry similar, dependent or chronologic meaning when compared.

Some techniques can be applied to apply Dimensionality Reduction:

Principal Component Analysis (PCA)

Reefers to a attribute extraction technique used to reduce data that are highly interdependent while keeping as much of data variance as possible. It filters data according to their degree of variance by using orthonormal vectors that simplify values to mainly consider the vectors that contain richier variability. Some use cases are in:

- *Exploratory Data Analysis:* to categorize data by its variations and summarizing into smaller sets of factors/components that helps contrast information, could be used for clinical studies where a variable could help understand how it contributes/affects another variable or a factor.
- *Information Compression:* widely used when reducing noise and compressing a large dataset where there are several "poor" variables that can be ruled out, filtering only the stronger more important, and often shown variables that will be used for the data analysis. An example would be in Machine Learning where a large dataset is filtered down to only the principal components.
- *Data De-noising:* similarly to Information Compression it helps de-noising data (without removing all the noise) by using the components that have higher variance to approximate the noise to the present data. E.g.: in a image denoising application, it is used to gather pixel

color information and transform noise - non matching pixels - to tones that are closer to the actual palette.

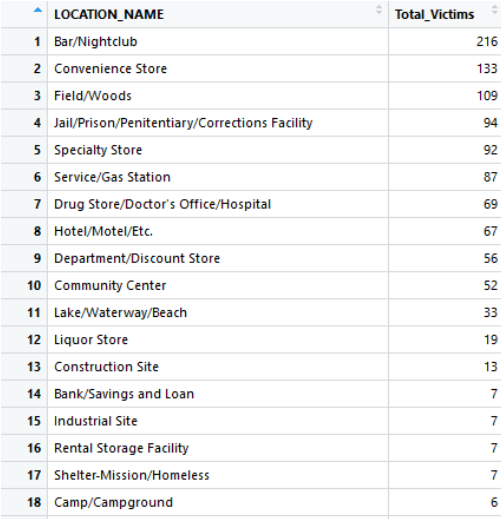
Linear Discriminant Analysis (LDA)

It is a learning algorithm that classifies tasks and features by separating them in different classes of data in order to increase perception of their differences. LDA projects data onto class variances that discriminates and maximize the ratio of differences in the classes. It is done from the presumption that the data has a Gaussian Distribution – where the matrices covariance are equal in the different classes – and that the data is linear, which means that can be easily separable and classified in different classes. Some use cases of LDA are listed below:

- *Topic Modeling:* a straightforward technique that filters data by taking samples for it, such as keywords and shrinking data to a summary or keywords from data as for example when analysing and summarizing an article or webpage.
- *Text Analytics:* has been crucial in Chatbox technologies in Artifical Inteligence to sift throught huge amounts of data and organising ideas in a simplified, natural language processing when prompted.
- *Classification:* It is used to categorize data, such as in a retail establishment where clients are put in several categories: low or high spender, income and annual spending.

Heber’s Report

Although both of us worked closely together in every step of this assessment, I was responsible for choosing the Dataset for our analysis, as well as carrying out data preparation, such as data dimensionality reduction where the data is filtered down from 28 columns to 10 and so we can get more concise views from it. I have also made use of Exploratory Data Analysis (EDA) techniques by comparing different aspects of our dataset in order to come up to different points of view, insights and conclusions, such as comparing:



	LOCATION_NAME	Total_Victims
1	Bar/Nightclub	216
2	Convenience Store	133
3	Field/Woods	109
4	Jail/Prison/Penitentiary/Corrections Facility	94
5	Specialty Store	92
6	Service/Gas Station	87
7	Drug Store/Doctor's Office/Hospital	69
8	Hotel/Motel/Etc.	67
9	Department/Discount Store	56
10	Community Center	52
11	Lake/Waterway/Beach	33
12	Liquor Store	19
13	Construction Site	13
14	Bank/Savings and Loan	7
15	Industrial Site	7
16	Rental Storage Facility	7
17	Shelter-Mission/Homeless	7
18	Camp/Campground	6

1. Offender race X Bias – to identify which offender race tends to against which bias such as gays, asian, black, white people.
2. Type of offence X Bias – to identify which types of offences these bias suffer from.
3. Location of offence X Type of offence – to identify where victims are attacked the most.
4. US State X Offender race – to identify states and their most common offender races.

By doing such comparisons we were able to define important factors of the crimes commited in the US as well as usin different types of data visualization such as histograms, pie charts, bar graphs, heatmaps, scatterplots, etc.

When it comes to documenting, I have written the sections: “Introduction”, “Dataset Chosen - Motivation, Problem Domain, Challenges”, “Dimensionality Reduction”, “Heber’s Report” and “Conclusion” present in this document. Furthermore, I have made a Powerpoint presentation that can be found in the same folder of this project.

This assignment was a crucial to give us insight on data analysis and the processes involved in dealing with substantial amounts of data as well as how to manipulate it in order to extract useful information. By using the language R for this work, we also learned how powerful this tool is. By applying dimensionality reduction we were able to simplify the complex dataset which allows us to get a clearer vision of its underlying structure.

We could also make use of statistical operations such as mean, median, minimum, maximum, and standard deviation, and operating Min-Max Normalization, Z-score Standardization that are extremely important to calculate the reliability of our dataset as well as important insights.

The individual participation charts are available in the R code provided as well as in the “Conclusion” section at the end of this document.

Caroline’s Report

Heber and I began this project with a clear plan, initially dividing tasks between us, but our collaboration naturally evolved into a more flexible partnership. Over time, we both ended up contributing to almost every aspect of the project. However, my main responsibilities were to clean and filter part of the data. I took the lead in creating plots to visualize the data better, using these visuals to provide clearer insights into the provided crime patterns. Additionally, I was responsible to make sure that the code is commented line-by-line, enhancing readability and transparency.

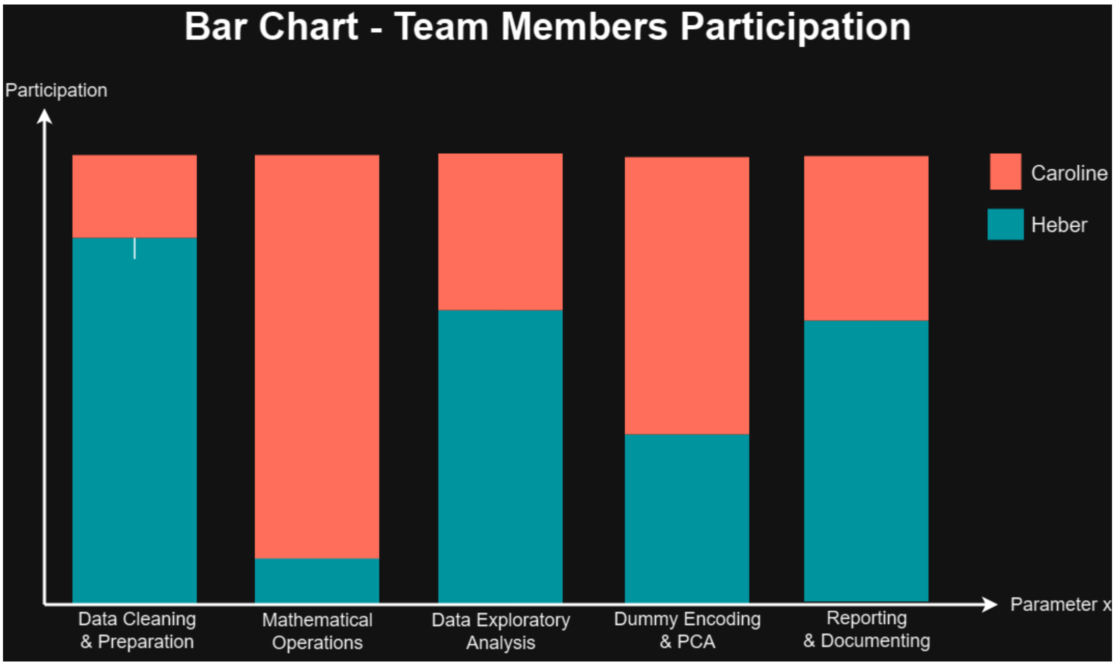
Regarding the report, I focused on researching and writing the sections "Exploratory Data Analysis & Outcomes," "Encoding Scheme," "Dummy Encoding," and "PCA.". This involved examining relationships between variables using various visual techniques like histograms, pie charts, bar graphs, heatmaps, scatterplots, and simple tables.

In summary, this project was more than a series of tasks; it provided a dynamic platform for me to actively contribute to Exploratory Data Analysis, encoding schemes, and dimensionality reduction. The methodologies used, including diverse visualizations and statistical techniques, not only provided valuable insights into the crime dataset but also established a strong foundation for subsequent analyses. My contributions were guided by the overarching goal of enriching the overall understanding of the dataset, fostering a holistic approach to data analysis. The experience was a learning curve, offering insights into the complexities of real-world data analysis and emphasizing the importance of collaborative and dynamic problem-solving approaches.

Conclusion

To conclude this work, both of the integrants of this pair assessment were able to learn and have hands-on experience of important aspects when it comes to data preparation for analysis once every project involving data science needs to go through cleaning and exploration processes. It is through collecting, cleaning, transforming and exploring data that we are able to get potential insights from the broad data scattered all over huge datasets. Such insights included the identification of trends, patterns and outliers that led to crucial conclusions from the data at hand.

Below we have attached a detailed graph with individual contributions:



Referencing

1. Javatpoint, no date. Linear Discriminant Analysis (LDA) in Machine Learning. Available at: https://www.javatpoint.com/linear-discriminant-analysis-in-machine-learning [Accessed 10 November 2023].

2. 365 Data Science, no date. What Is the Difference Between PCA and LDA? Available at: https://365datascience.com/tutorials/python-tutorials/lda-vs-pca/ [Accessed 10 November 2023].

3. High Demand Skills, 2022. Topic Modeling with LDA Explained: Applications and How It Works. Available at: https://highdemandskills.com/topic-modeling-intuitive/ [Accessed 10 November 2023].

4. Barhate, P., no date. Latent Dirichlet Allocation for Beginners: A High-Level Intuition. Available at: https://medium.com/@pratikbarhate/latent-dirichlet-allocation-for-beginners-a-high-level-intuition-23f8a5cbad71 [Accessed 15 November 2023].

5. Statology, 2020. Introduction to Linear Discriminant Analysis. Available at: https://www.statology.org/linear-discriminant-analysis/ [Accessed 15 November 2023].

6. LibreTexts, no date. Using LDA to Form an Enolate Ion. Available at: https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Map%3A_Organic_Chemistry_%28Bruice%29/19%3A_Carbonyl_Compounds_III-Reactions_at_the-_Carbon/19.08%3A_Using_LDA_to_Form_an_Enolate_Ion [Accessed 16 November 2023].

7. Wiley Online Library, no date. Available at: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jan.14377 [Accessed 16 November 2023].

8. GeeksforGeeks, no date. Principal Component Analysis(PCA). Available at: https://www.geeksforgeeks.org/principal-component-analysis-pca/ [Accessed 20 November 2023].

9. StackExchange, no date. Is Decompression Possible with PCA? Available at: https://stats.stackexchange.com/questions/454814/is-decompression-possible-with-pca [Accessed 20 November 2023].

10. IEEE Xplore, 2020. Principle and Application of Information Compression Based on Tensor PCA. Available at: https://ieeexplore.ieee.org/document/9526801/ [Accessed 20 November 2023].

11. Keboola, no date. PCA in Machine Learning. Available at: https://www.keboola.com/blog/pca-machine-learning [Accessed 20 November 2023].

12. Moran, M. (2021) Dummy Coding: The how and why, Statistics Solutions. Available at: https://www.statisticssolutions.com/dummy-coding-the-how-and-why/ (Accessed: 26 November 2023).

13. HOME (no date) OARC Stats. Available at: https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-dummy-coding/ (Accessed: 26 November 2023).

14. Sartorius (2020) What Is Principal Component Analysis (PCA) and How It Is Used?, Sartorius. Available at: https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186#:~:text=Principal%20component%20analysis%2C%20or%20PCA,more%20easily%20visualized%20and%20analyzed. (Accessed: 28 November 2023).