|  |  |  |  |
| --- | --- | --- | --- |
|  | Exploratory Data Analysis | |  |
|  | |
| 23/02/2023 | |
|  | Ministry of Social Development | |  |
| Mihir Bachkaniwala |  |  |
|  | |  |

# Exploratory Data Analysis (EDA)

* Understanding the data: This involves getting to know the data you are working with by reviewing the data's source, structure, and contents. It is essential to understand the data's features, such as its type, range, and distribution, to make informed decisions about which EDA techniques to use.
* Cleaning the data: This step involves detecting and correcting errors and inconsistencies in the data to ensure that the results of the EDA are accurate and reliable. Data cleaning may involve imputing missing values, identifying and removing outliers, and standardizing the data.
* Exploring the data: This step involves visualizing the data to identify patterns, trends, and relationships between variables. The goal is to identify interesting features of the data that may provide insights into the underlying phenomenon being studied.
* Summarizing the data: This step involves summarizing the data using descriptive statistics such as measures of central tendency, dispersion, and correlation. These statistics can provide a quick overview of the data and help identify potential issues or interesting patterns.
* Drawing conclusions: Based on the results of the EDA, you can draw conclusions about the data and the phenomenon being studied. This step involves synthesizing the insights gained from the data analysis and identifying potential follow-up analyses that could provide more information or answer specific research questions.

In summary, a good EDA involves thorough understanding and cleaning of the data, effective exploration and visualization techniques, and careful summarization and conclusion drawing.

## Understanding the data

|  |  |  |  |
| --- | --- | --- | --- |
|  | Demographic.csv | Event.csv | Injury.csv |
| Number of records and columns | Rows: 1621  Columns:7 | Rows:3076  Columns:11 | Rows:3224  Columns:8 |
| Data types | Int64, varchar | Int64, varchar | Int64, varchar |
| Missing values | No missing values | Site 3 Name: 21.39%  Site 2 Name: 2.40% | Injury: 19.6 %  Body Part: 4.74% |
| Duplicates | No Duplicates | No Duplicates | 460 Duplicates found |

* The absence of values in the Site 2 and Site 3 Name columns indicates that the events occurred outside of the site. As such, no further action is required.

## Cleaning the data

* The process of removing duplicate records in the 'Injury.csv' file resulted in the elimination of 460 rows. After this deduplication step, the number of rows in the file decreased to 2764.
* The rows with 'N/A' or 'Unassigned' values were removed because most of the cases were either closed or waiting for closure and required simple or no investigation.

## Exploring the data

## Summarizing the data

* Approximately 80% of the cases that did not specify any particular body part were found to be associated with psychological distress.

## Drawing conclusions