|  |  |  |  |
| --- | --- | --- | --- |
|  | 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.

The objective of this report is to conduct an exploratory data analysis on an incident reporting dataset from the MSD (Ministry of Social Development). This dataset contains row-level information for incident events, including details about injuries sustained, hazards involved, demographic information of employees, and details about the incident events such as the location, severity, and actions taken after the incident.

The purpose of this analysis is to gain insights into the data and uncover patterns and trends that may inform the MSD's decision-making process regarding incident prevention and management. This report will focus on analyzing the data using various data visualization techniques and statistical methods to answer questions such as which regions have the highest incident and lowest incident rates, the summary of the top 5 types of incidents, and the top 5 common body parts injured.

The report will start with a data cleaning process to ensure that the data is accurate and complete. Then, we will conduct exploratory data analysis, which involves the use of visualizations and descriptive statistics to understand the data better. Finally, we will summarize our findings and provide recommendations to the MSD based on the insights obtained from the data analysis.

## Understanding the data

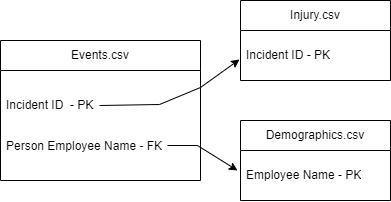
Based on the initial analysis done with Excel and Jupyter Notebook, the dataset consists of three tables: Demographic.csv, Event.csv, and Injury.csv. The Demographic table has 1621 rows and 7 columns, with data types of Int64 and varchar. There are no missing values or duplicates in this table.

The Event table has 3076 rows and 11 columns, with data types of Int64, varchar, and datetime64. There are missing values in the Site 3 Name and Site 2 Name columns, which indicates that some events occurred outside of the site. Additionally, there are duplicates in this table, with 460 duplicate records found.

The Injury table has 3224 rows and 8 columns, with data types of Int64 and varchar. There are missing values in the Injury and Body Part columns, with 19.6% and 4.74% of the data missing, respectively.

The earliest record in the dataset is dated 10 February 2018 and the latest record is on 19th January 2021.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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, datetime64  Note: Data type for date changed to datetime64. | 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 |



## Cleaning the data

Based on the initial analysis, the data cleaning process involved the removal of duplicate records and rows with missing or unassigned values. The 'Injury.csv' file had 460 duplicate rows, which were removed to decrease the total number of rows to 2764. Additionally, rows with 'N/A' or 'Unassigned' values were eliminated as they did not require any intervention or medical care.

The 'Incident Date' column was modified to have a more precise data type of datetime64. This modification will facilitate the analysis of the incident data over time.

To find relationships between columns across the data frames, we merged the data frames. This process will allow us to investigate relationships between injury types, demographics, and incident events.

The data visualization phase will involve the use of graphs, charts, and other visual aids to explore the relationships between different variables in the dataset. The goal of this phase is to identify any patterns, trends, or anomalies in the data that may be useful for understanding the incident reporting system and improving workplace safety.

## Exploring the data

* The incidents were analyzed based on age group, gender, ethnicity and salary range to determine their demographic distribution.
* In order to examine the regions with the highest and lowest incident rates, we combined the data from multiple sources into a single dataframe.
* Upon reviewing the column names in the Events and Demographics dataframes, it was noticed that the column name required for merging was not consistent. Therefore, it was necessary to change the column name 'Person Employee Name' to 'Employee Name' in the Events dataframe to match the column name in the Demographics dataframe.
* incident rates by region were plotted
* total number of incidents per region was plotted
* Summary on the top 5 type of incidents was plotted
* Top 5 common body parts injured
* Top 5 Position Titles with most injuries
* calculate injury count by salary range
* Timeseries plot for weekly and monthly frequency of incidents
* Types of events captured at sites plotted
* Plotted incident status of this data set
* Plotted the potential consequences on the incidents
* Plotted the Treatments of Injuries
* Visualising the Nature of Injuries

Demographic Distribution:

To analyze the demographic distribution, incidents were categorized based on age group, gender, ethnicity, and salary range. The following bar graphs were plotted to represent the distribution of incidents based on each of these attributes:

Bar graph showing the number of incidents by age group - Fig 1.3

Bar graph showing the number of incidents by gender - Fig 1.1

Bar graph showing the number of incidents by ethnicity group - Fig 1.2

Bar graph showing the number of incidents by salary range - Fig 1.4

Regional Incident Rates:

To examine the regions with the highest and lowest incident rates, data from multiple sources was combined into a single dataframe. The following plots were created:

Bar graph showing the total number of incidents per region - Fig 1.6

Bar graph showing the highest and lowest incident rates - Fig 1.5

Incidents Report:

A summary of the top 5 types of incidents was created using a bar graph – Fig 1.7.

A bar graph was plotted to show the top 5 common body parts injured - Fig 1.8.

Position Titles with Most Injuries:

A bar graph was plotted to show the top 5 position titles with the most injuries – Fig 1.9.

Injury Count by Salary Range:

A bar graph was created to show the injury count by salary range – Fig 2.0.

Time Series Plot:

A time series plot was created to show the weekly and monthly frequency of incidents - Fig 2.1, Fig 2.2.

Incident Status:

A bar graph was plotted to show the incident status of this data set - Fig 2.5.

Potential Consequences:

A bar graph was plotted to show the potential consequences of the incidents - Fig 2.6.

Treatments of Injuries:

A horizontal bar graph was plotted to show the treatments given for the injuries - Fig 2.7.

Nature of Injuries:

A bar graph was plotted to show the distribution of nature of injuries - Fig 2.8.

Overall, the above visualizations provide insights into the demographic distribution of incidents, the regions with the highest and lowest incident rates, the types of incidents, and the nature of injuries.

During the Exploratory Data Analysis (EDA) phase, we analyzed the files to gain insights into the demographic distribution of incidents, the regions with the highest and lowest incident rates, the types of incidents, and the nature of injuries. We categorized incidents based on age group, gender, ethnicity, and salary range to examine the demographic distribution. We created several bar graphs to visualize the distribution of incidents based on each of these attributes. The bar graph representing the number of incidents by gender showed that female employees had a higher number of incidents compared to male employees. The bar graph representing the number of incidents by ethnicity showed that incidents were reported across various ethnicities, with the highest number of incidents reported by group 'B' employees. The bar graph representing the number of incidents by age group showed that employees in the age band 1 reported the highest number of incidents.

We also examined the regions with the highest and lowest incident rates by combining data from multiple sources into a single dataframe. We plotted a bar graph showing the total number of incidents per region and a bar graph showing the highest and lowest incident rates. The bar graph representing the highest and lowest incident rates showed that Region K had the highest incident rate while Region M had the lowest incident rate. The bar graph representing the total number of incidents per region showed that Region G had the highest number of incidents while Region J had the lowest number of incidents.

Furthermore, we analyzed the types of incidents and nature of injuries by creating several bar graphs. A bar graph representing the top 5 types of incidents showed that "Sprain/ Strain" was the most common type of incident reported, followed by "Bruising/Crushing." The bar graph representing the top 5 common body parts injured showed that the "Back" was the most commonly injured body part. The bar graph representing the nature of injuries showed that "Ergonomic" were the most common type of injury reported. Overall, the visualizations created during the EDA phase provide valuable insights into the distribution of incidents and the nature of injuries reported.

## Summarizing the data

* Approximately 80% of the cases that did not specify any particular body part were found to be associated with psychological distress.
* Most of the incidents in the data set have a closed status
* Common Body Parts Injured: Back, Multiple Locations, Wrist, Shoulders and Neck.
* Case Managers had the highest number of injuries.
* M1 Salary band recorded the highest number of injuries.
* Customer service delivery site had the most injuries while National office and offsite injuries were relatively not many.
* There were more cases with moderate consequences followed by minor.
* A major proportion of incidents required no treatment.

After analyzing the Injury dataset, it was observed that a significant percentage of cases did not specify any particular body part but were associated with psychological distress. This highlights the importance of addressing mental health concerns in the workplace. Additionally, the majority of incidents in the dataset were closed, indicating that appropriate measures were taken to resolve the issues. The common body parts injured were found to be the back, multiple locations, wrist, shoulders, and neck. This information can be used to prioritize workplace safety measures to prevent injuries in these areas.

Further analysis of the dataset revealed that Case Managers had the highest number of injuries, indicating that their job roles may have a higher risk of incidents. The M1 salary band was associated with the highest number of injuries, which highlights the need for targeted safety measures for employees in this salary range. The Customer service delivery site had the most injuries compared to other sites, indicating the need for site-specific safety measures. However, the National office and offsite injuries were relatively low in number, suggesting that the current safety measures may be effective in preventing injuries at these locations.

The incidents in the dataset were classified based on their potential consequences, and it was observed that a major proportion of incidents had no or minor consequences. However, a significant number of cases had moderate consequences, indicating the need for appropriate measures to prevent incidents from escalating to more severe outcomes. Interestingly, a majority of incidents required no treatment, which suggests that many incidents were minor in nature. Overall, these findings provide valuable insights into the nature of incidents and highlight the need for targeted safety measures to prevent injuries in the workplace.

## Drawing conclusions

Future enhancement for advance tracking of unassigned cases, we can add another timestamp with the updated information on event captured. Using the concept of slow changing dimensions, we might be able to do a more in-depth analysis of the data.

References for bar charts

|  |  |
| --- | --- |
| Fig 1.1 - Distribution of demographics by gender |  |
| Fig 1.2 - Distribution of demographics by ethnicity |  |
| Fig 1.3 - Distribution of demographics by Age Band | Fig 1.4 - Distribution of demographics by Salary Range |
| Fig 1.5 - Regions with the highest incident and lowest incident rates | Fig 1.6 - total incidents by region |
| Fig 1.7 - Summary on the top 5 type of incidents | Fig 1.8 - Top 5 common body parts injured |
| Fig 1.9 - Top 5 Position Titles with most injuries | Fig 2.0 - Plot of injury count by salary range |
| Fig 2.1 - Weekly frequency of incidents | Fig 2.2 - Monthly frequency of incidents |
| Fig 2.3 - Plotting the locations of injuries | Fig 2.4 - Events captured at Client Service Delivery |
| Fig 2.5 - incident status | Fig 2.6 - Potential consequences of the incidents |
| Fig 2.7 - Treatments of Injuries | Fig 2.8 - Nature of Injuries |