

# Exploratory Data Analysis

JOB ASSESSMENT EXERCISE

MIHIR BACHKANIWALA



**MINISTRY OF SOCIAL  
DEVELOPMENT**

TE MANATŪ WHAKAHIATO ORA

## TABLE OF CONTENTS

<b>EXPLORATORY DATA ANALYSIS (EDA)</b>	2
<b>UNDERSTANDING THE DATA</b>	3
<b>CLEANING THE DATA</b>	4
<b>EXPLORING THE DATA</b>	4
<b>SUMMARIZING THE DATA</b>	5
<b>REFERENCES FOR VISUALIZATIONS</b>	6
DISTRIBUTION OF DEMOGRAPHICS BY GENDER	6
DISTRIBUTION OF DEMOGRAPHICS BY ETHNICITY	6
DISTRIBUTION OF DEMOGRAPHICS BY AGE BAND	7
DISTRIBUTION OF DEMOGRAPHICS BY SALARY RANGE	7
REGIONS WITH THE HIGHEST INCIDENT AND LOWEST INCIDENT RATES	8
TOTAL INCIDENTS BY REGION	8
SUMMARY ON THE TOP 5 TYPE OF INCIDENTS	9
TOP 5 COMMON BODY PARTS INJURED	9
TOP 5 POSITION TITLES WITH MOST INJURIES	10
PLOT OF INJURY COUNT BY SALARY RANGE	10
WEEKLY FREQUENCY OF INCIDENTS	11
MONTHLY FREQUENCY OF INCIDENTS	11
PLOTING THE LOCATIONS OF INJURIES	12
EVENTS CAPTURED AT CLIENT SERVICE DELIVERY	12
INCIDENT STATUS BREAKDOWN FOR THIS DATASET	13
POTENTIAL CONSEQUENCES OF THE INCIDENTS	13
TREATMENTS OF INJURIES	14
NATURE OF INJURIES	14
DISTRIBUTION OF NATURE OF INJURY FOR ERGONOMIC TYPE	15
<b>DRAWING CONCLUSIONS</b>	16

# EXPLORATORY DATA ANALYSIS (EDA)

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.

This report has been created for the exploratory data analysis conducted, which summarizes the findings and insights gained from the analysis. The report has been documented in a Word file and is also available in PDF format. Additionally, the code used to perform the analysis has been implemented in Jupyter Notebook, which is also accessible on my Github repository.

Link for Github Repository - <https://github.com/MihirB-Byte/MSD-Assessment>

The exploratory data analysis involved various tasks such as data cleaning, data merging, data visualization, and data interpretation. Quick insights were generated in Jupyter Notebook to identify patterns and trends in the data. The findings were summarized and presented in the report, which also includes visual representations of the analyzed data. Overall, the exploratory data analysis aimed to provide a better understanding of the data and generate insights that can be used to inform decision-making processes.

## UNDERSTANDING THE DATA

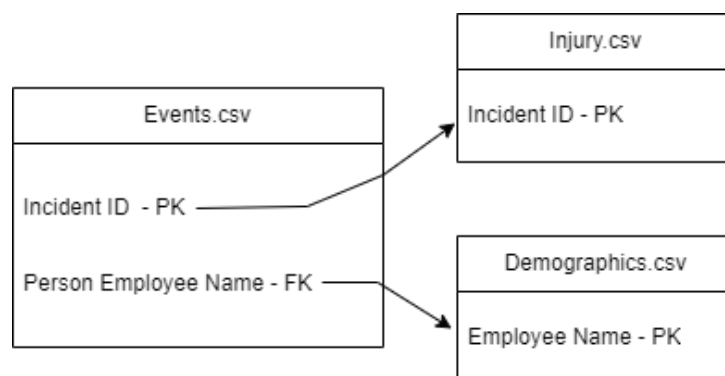
Based on the initial analysis done with Excel and Jupyter Notebook, the dataset consists of three tables: Demographic.csv, Event.csv, and Injury.csv. The Demographic table has 1621 rows and 7 columns, with data types of Int64 and varchar. There are no missing values or duplicates in this table.

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

- Duplicate rows were removed from the 'Injury.csv' file, resulting in a decrease in the total number of rows to 2764.
- Rows with 'N/A' or 'Unassigned' values were eliminated as they did not require any intervention or medical care and most instances had a status as closed.
- The 'Incident Date' column was modified to have a more precise data type of datetime64 to facilitate analysis of the incident data over time.
- The data frames were merged to investigate relationships between injury types, demographics, and incident events.
- Data visualization techniques, such as graphs, charts, and other visual aids, were used to explore relationships between different variables in the dataset.

## EXPLORING THE DATA

- During the Exploratory Data Analysis (EDA) phase, we examined demographic distribution, incident rates, types of incidents, and nature of injuries.
- Incidents were categorized based on age group, gender, ethnicity, and salary range.
- Several bar graphs were created 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.
- Regions with the highest and lowest incident rates were examined by combining data from multiple sources into a single dataframe.
- Bar graphs were plotted to show the total number of incidents per region and 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.
- Types of incidents and nature of injuries were analyzed by creating several bar graphs.
- The most common type of incident reported was "Sprain/ Strain", followed by "Bruising/Crushing."
- The most commonly injured body part was the "Back."
- The most common type of injury reported was "Ergonomic."
- A deeper analysis of Ergonomic injuries showed a high number of incidents related to workstation setup.
- Ergonomic training and equipment may be necessary to improve workstation ergonomics, thereby reducing ergonomic injuries.

## SUMMARIZING THE DATA

The Injury dataset analysis reveals that mental health concerns in the workplace should not be overlooked, as a significant percentage of cases associated with psychological distress did not specify any particular body part. This finding underlines the importance of prioritizing mental health support for employees. Additionally, the most common body parts injured were identified as the back, multiple locations, wrist, shoulders, and neck, providing a clear focus for safety measures to prevent injuries in these areas. Targeted safety measures should also be considered for employees in the Case Manager role, as this group had the highest number of injuries.

The analysis also highlighted the need for site-specific safety measures, as the Customer service delivery site had the highest number of injuries compared to other sites. The M1 salary band was associated with the highest number of injuries, indicating the need for targeted safety measures for employees in this salary range. It is noteworthy that the National office and offsite injuries were relatively low in number, which suggests that the current safety measures may be effective in preventing injuries at these locations.

The potential consequences of incidents were classified, and it was observed that 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, suggesting that many incidents were minor in nature. This highlights the importance of implementing effective safety measures to prevent even minor incidents from occurring.

Furthermore, a deeper analysis of Ergonomic injuries showed a high number of incidents related to workstation setup. This indicates that ergonomic training and equipment may be necessary to improve workstation ergonomics, thereby reducing ergonomic injuries. In conclusion, this exploratory data analysis provides valuable insights into the nature of incidents, and the findings should be used to inform targeted safety measures to prevent injuries in the workplace.

REFERENCES FOR VISUALIZATIONS

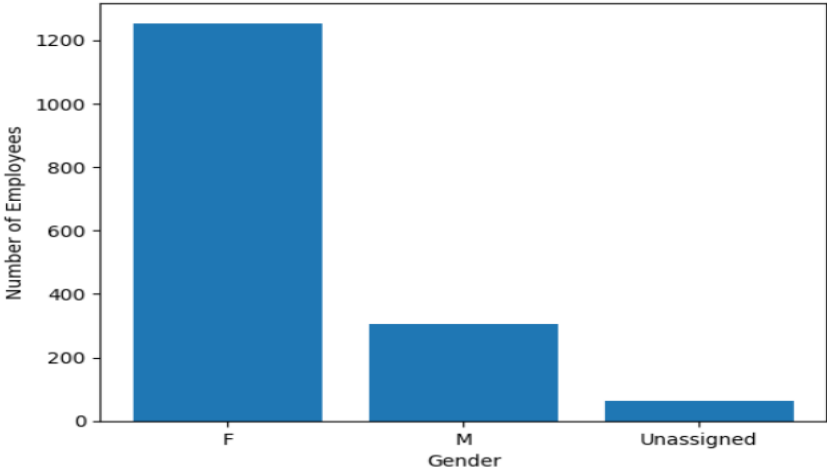
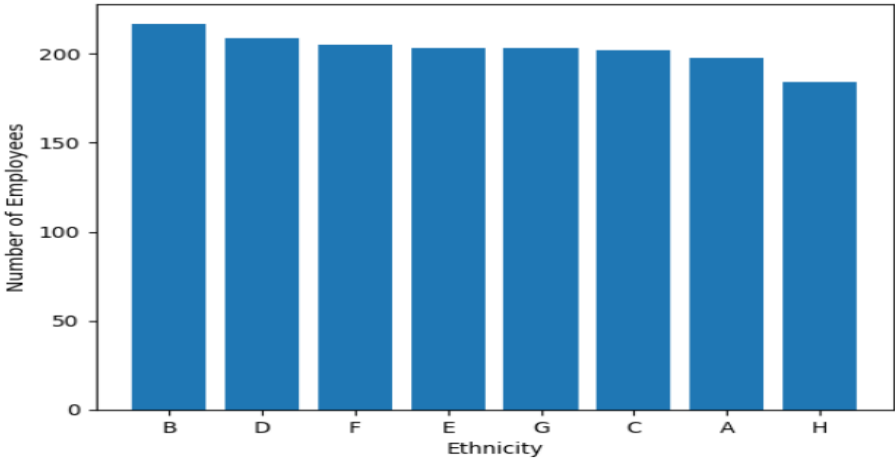
FIG 1.1	<div>DISTRIBUTION OF DEMOGRAPHICS BY GENDER</div> <div><p>Distribution of demographics by Gender</p><table border="1"><thead><tr><th>Gender</th><th>Number of Employees</th></tr></thead><tbody><tr><td>F</td><td>1250</td></tr><tr><td>M</td><td>300</td></tr><tr><td>Unassigned</td><td>50</td></tr></tbody></table></div>	Gender	Number of Employees	F	1250	M	300	Unassigned	50										
Gender	Number of Employees																		
F	1250																		
M	300																		
Unassigned	50																		
FIG 1.2	<div>DISTRIBUTION OF DEMOGRAPHICS BY ETHNICITY</div> <div><p>Distribution of demographics by Ethnicity</p><table border="1"><thead><tr><th>Ethnicity</th><th>Number of Employees</th></tr></thead><tbody><tr><td>B</td><td>210</td></tr><tr><td>D</td><td>205</td></tr><tr><td>F</td><td>202</td></tr><tr><td>E</td><td>200</td></tr><tr><td>G</td><td>200</td></tr><tr><td>C</td><td>200</td></tr><tr><td>A</td><td>195</td></tr><tr><td>H</td><td>180</td></tr></tbody></table></div>	Ethnicity	Number of Employees	B	210	D	205	F	202	E	200	G	200	C	200	A	195	H	180
Ethnicity	Number of Employees																		
B	210																		
D	205																		
F	202																		
E	200																		
G	200																		
C	200																		
A	195																		
H	180																		

FIG 1.3

## DISTRIBUTION OF DEMOGRAPHICS BY AGE BAND

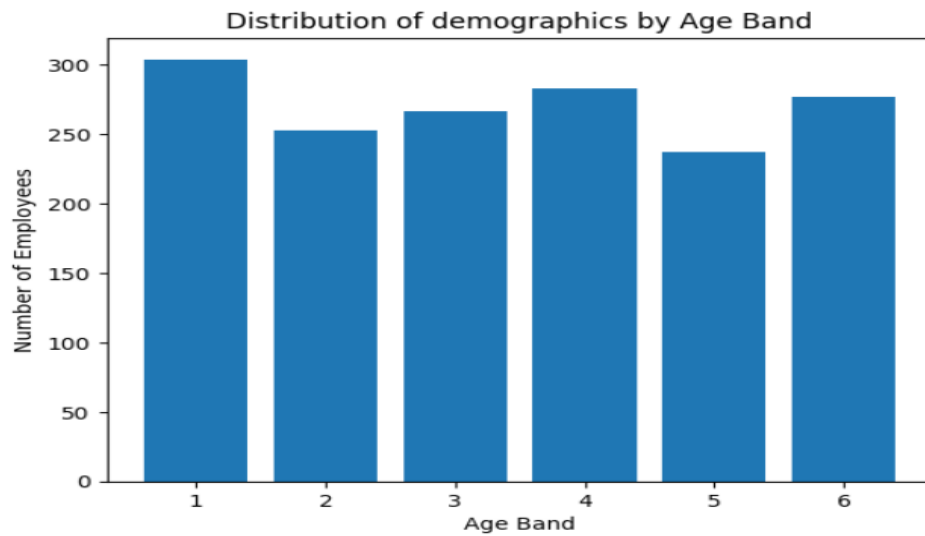


FIG 1.4

## DISTRIBUTION OF DEMOGRAPHICS BY SALARY RANGE

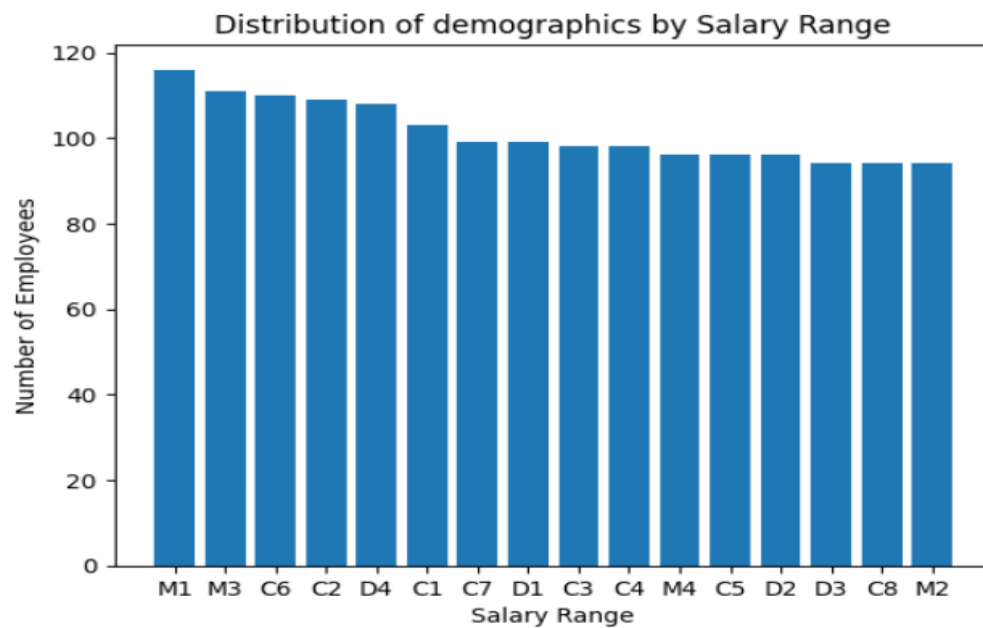




FIG 1.5

## REGIONS WITH THE HIGHEST INCIDENT AND LOWEST INCIDENT RATES

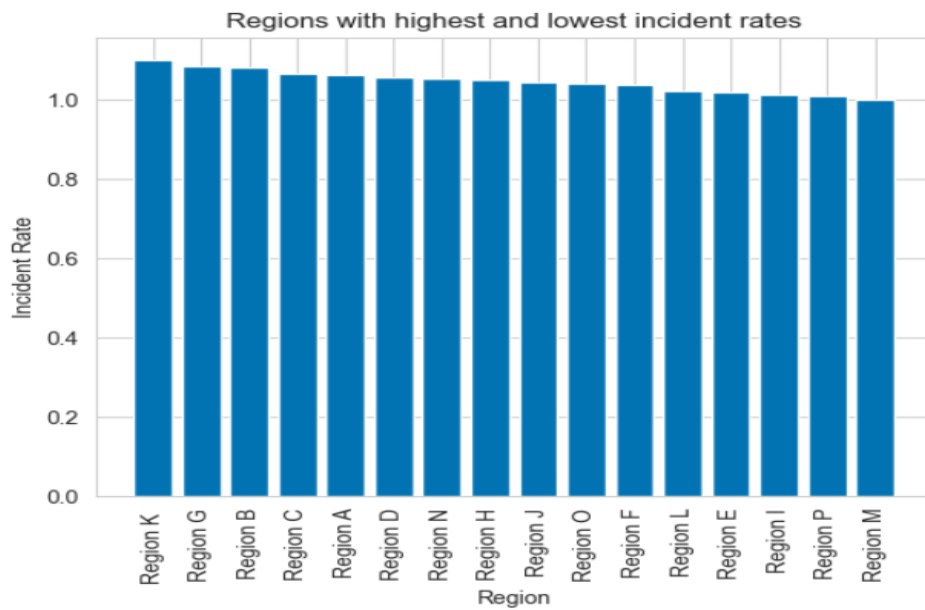


FIG 1.6

## TOTAL INCIDENTS BY REGION

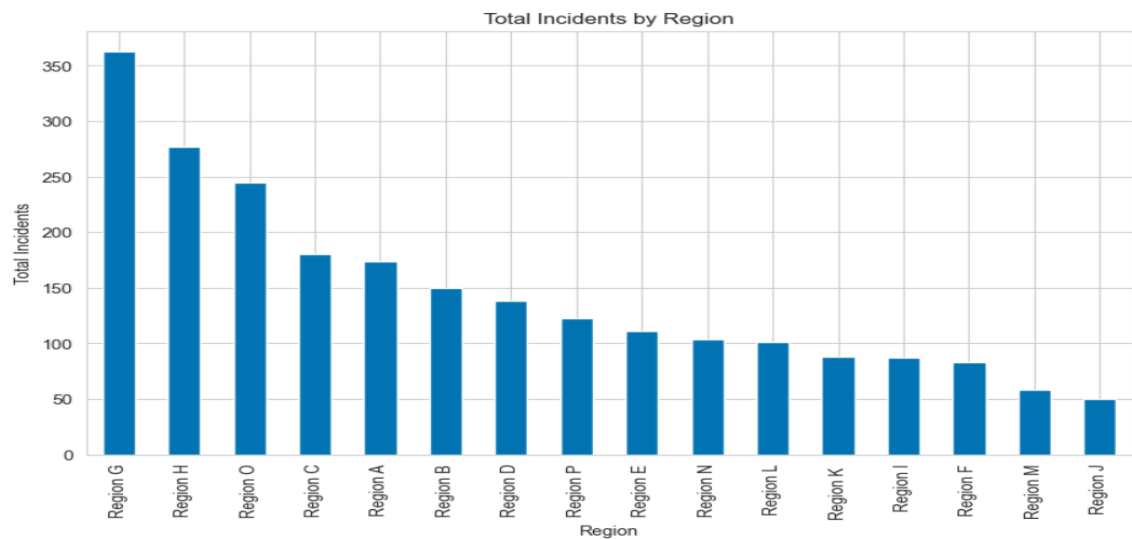


FIG 1.7

## SUMMARY ON THE TOP 5 TYPE OF INCIDENTS

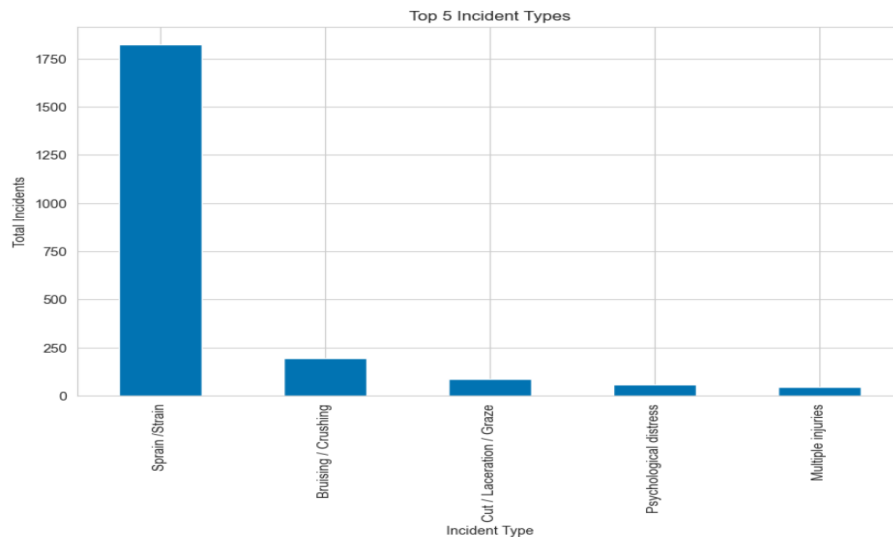


FIG 1.8

## TOP 5 COMMON BODY PARTS INJURED

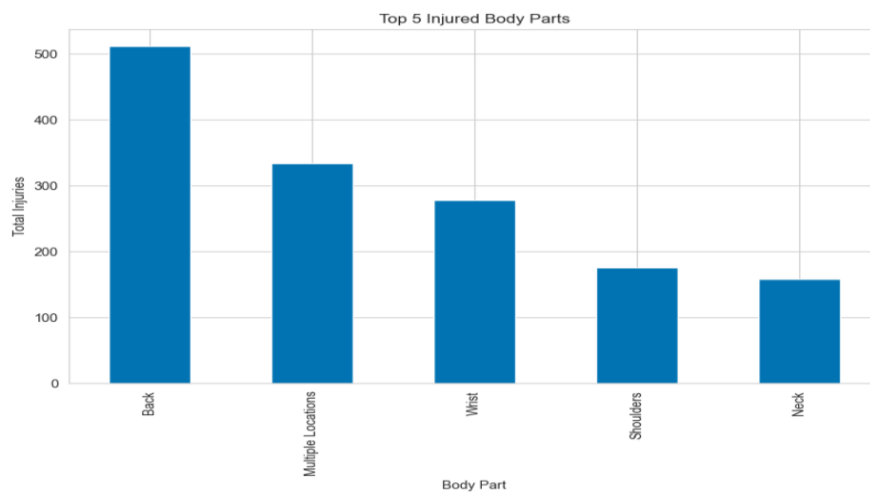


FIG 1.9

TOP 5 POSITION TITLES WITH MOST INJURIES

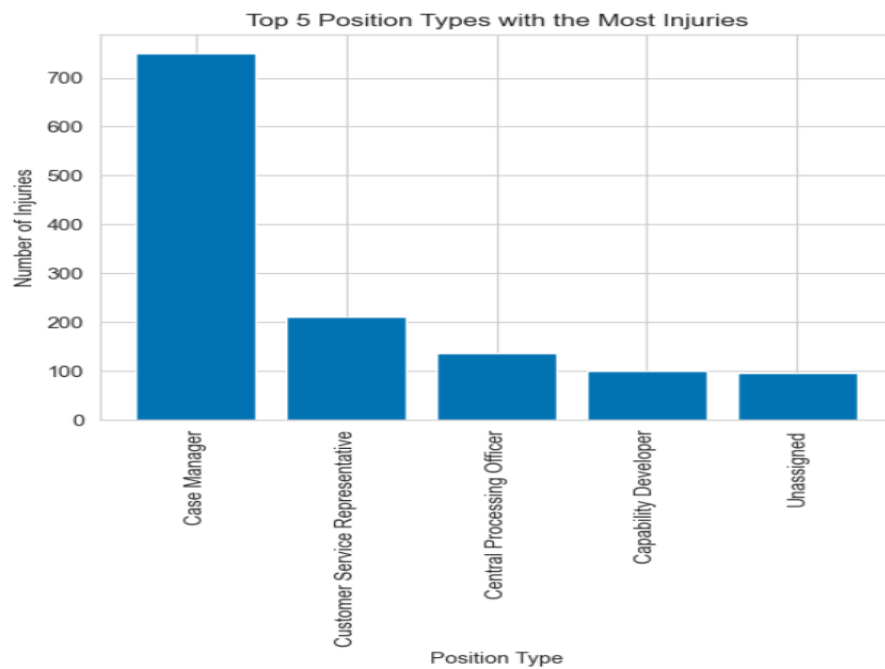


FIG 2.0

PLOT OF INJURY COUNT BY SALARY RANGE

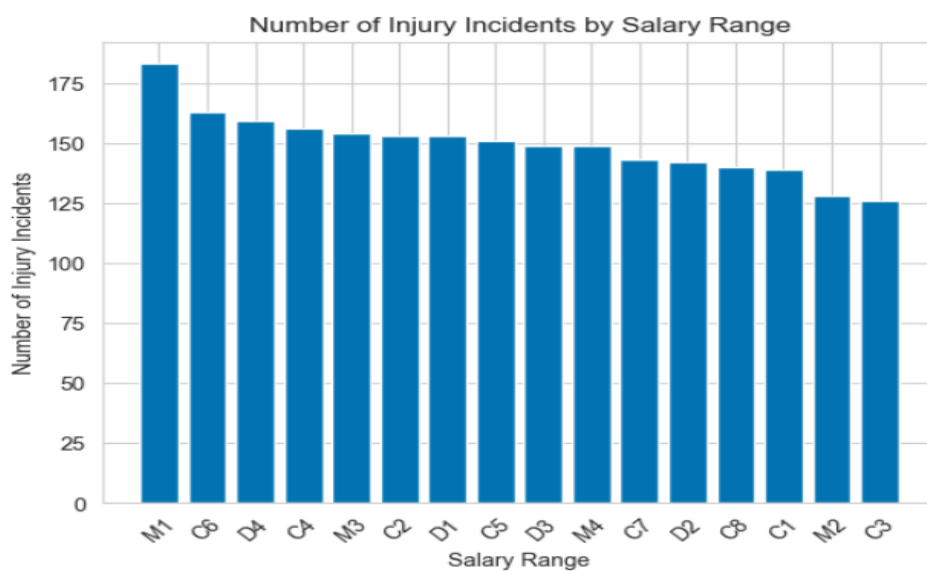


FIG 2.1 WEEKLY FREQUENCY OF INCIDENTS

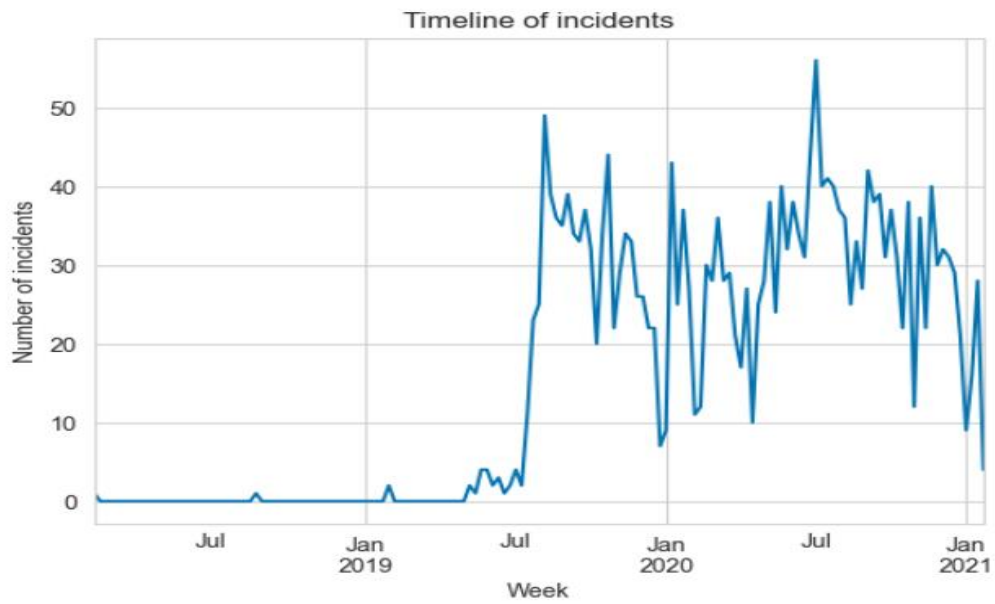


FIG 2.2 MONTHLY FREQUENCY OF INCIDENTS

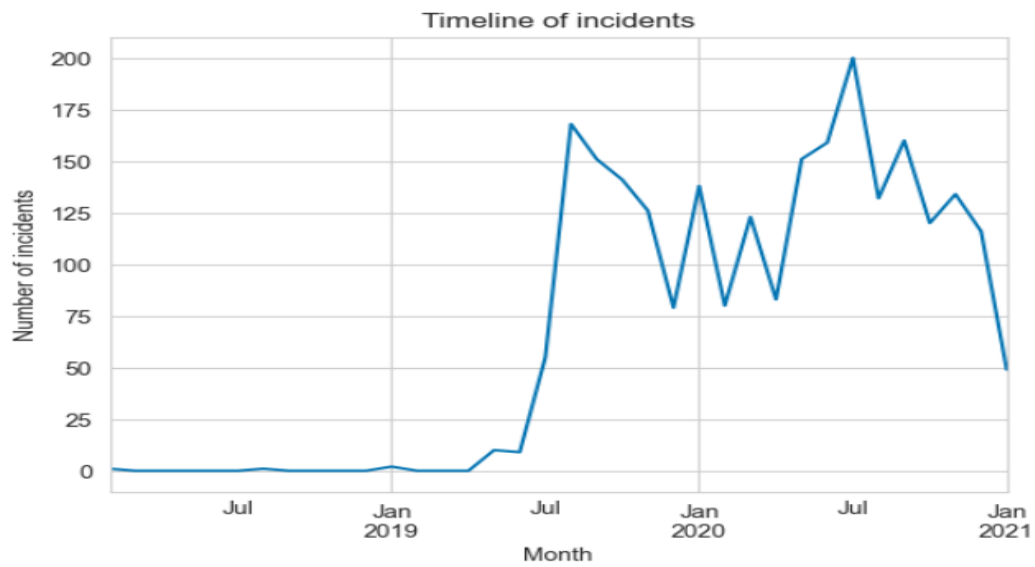


FIG 2.3

## PLOTTING THE LOCATIONS OF INJURIES

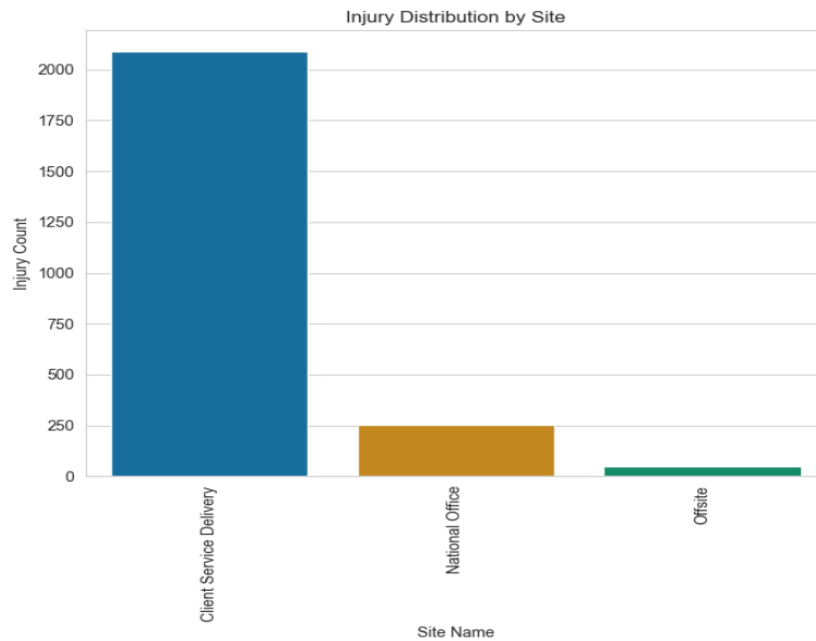


FIG 2.4

## EVENTS CAPTURED AT CLIENT SERVICE DELIVERY

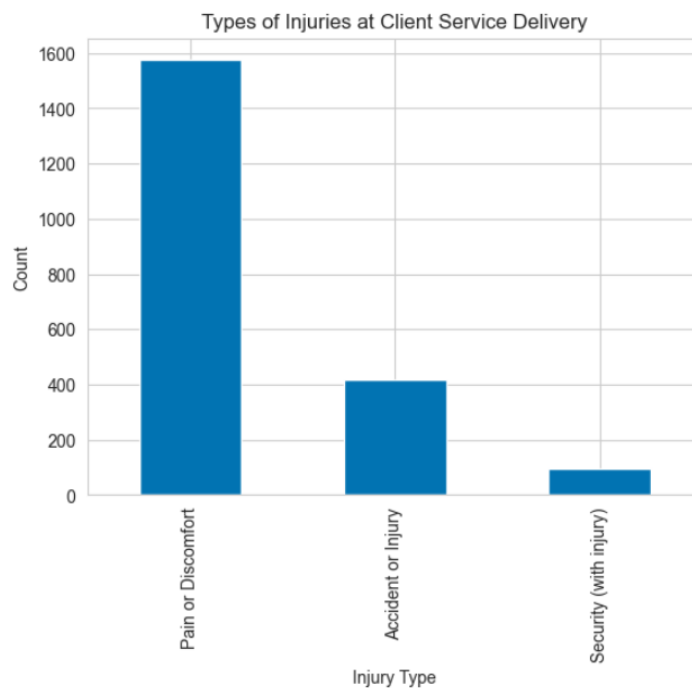


FIG 2.5

## INCIDENT STATUS BREAKDOWN FOR THIS DATASET

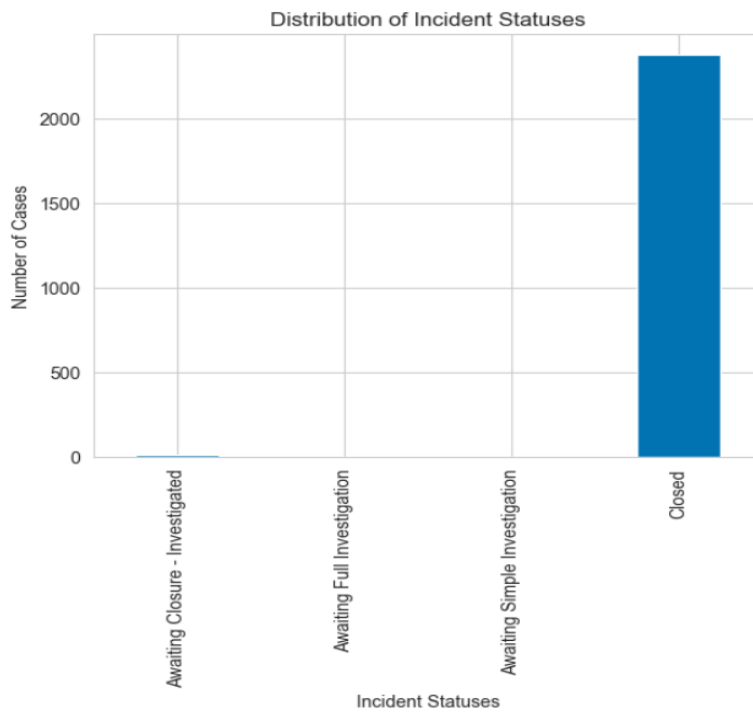


FIG 2.6

## POTENTIAL CONSEQUENCES OF THE INCIDENTS

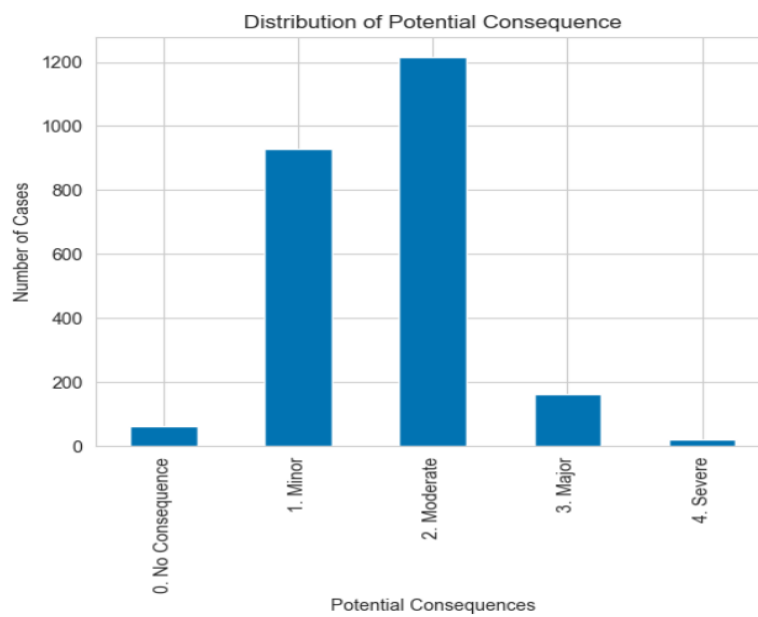


FIG 2.7

TREATMENTS OF INJURIES

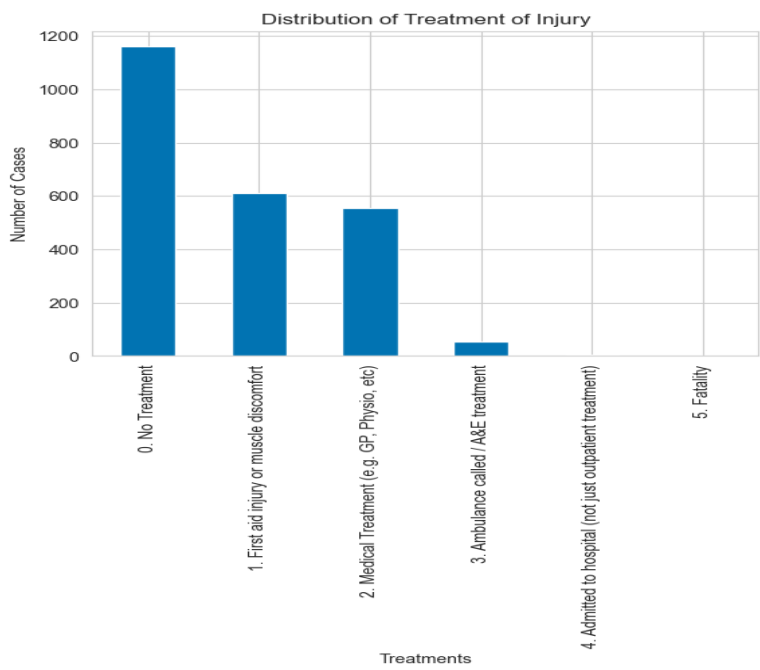


FIG 2.8

NATURE OF INJURIES

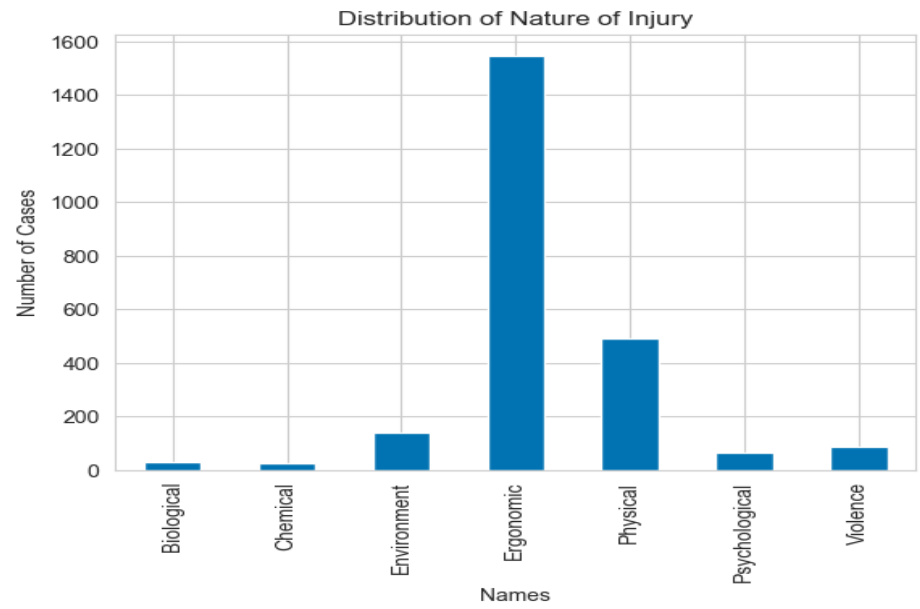
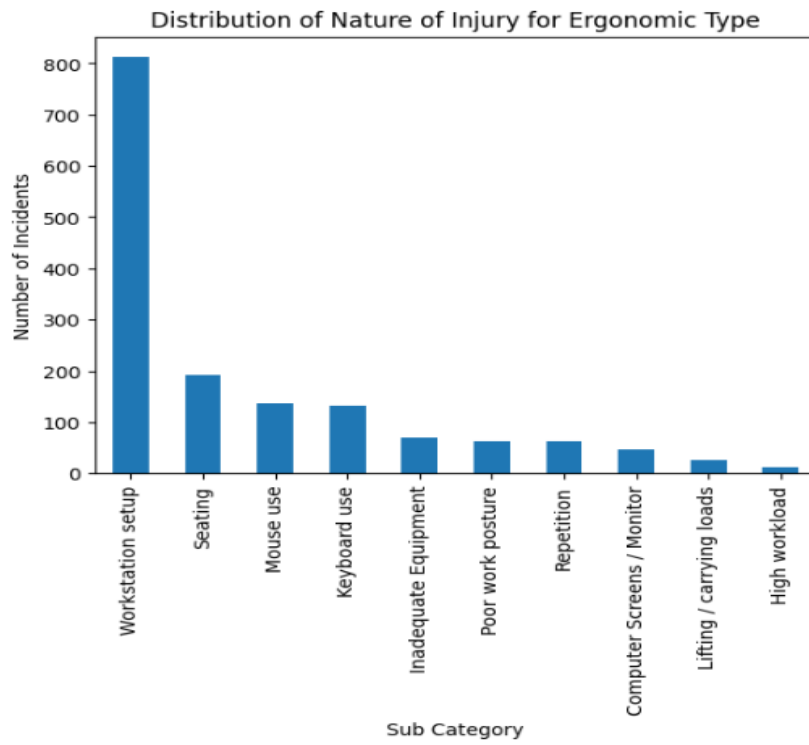


FIG 2.9

## DISTRIBUTION OF NATURE OF INJURY FOR ERGONOMIC TYPE





## DRAWING CONCLUSIONS

Firstly, the data analysis provides 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. The analysis helps to identify the common body parts injured, position titles with the most injuries, and the injury count by salary range.

Secondly, the data shows that there is a high incidence of psychological distress associated with incidents that did not specify any particular body part. Additionally, the closed status was the most common incident status, and Case Managers had the highest number of injuries.

Finally, the data analysis also reveals that future enhancements can be made for advanced tracking of unassigned cases by adding another timestamp with updated information on event captured. This could help in doing a more in-depth analysis of the data using the concept of slow changing dimensions.

Overall, the EDA provides valuable insights into the incidents and injuries, which can help organizations take steps to prevent them from occurring in the future.