

**DATA SCIENCE TOOLBOX PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

**Serious Injury Outcome Indicators**

**Submitted by**

**NAME:** M.VAMSHIDHAR REDDY

**REGISTRATION NO:** 12302830

**PROGRAMME AND SECTION:** K23GN18

**COURSE CODE** **:** INT375

Under the Guidance of

**MRS. AASHIMA (UID: 28968)**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **VAMSHIDHAR** bearing Registration no. **12302830** has completed **CSE375** project titled, **“Mrs.Aashima”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science**

Lovely Professional University

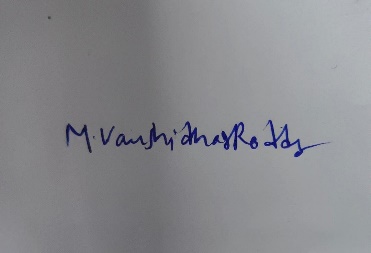
Phagwara, Punjab.

Date: 11-04-2025

**DECLARATION**

I, VAMSHIDHAR student of CSE (Program name) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 11-04-2025 Signature :



Registration No: 12302830 Name of the student: VAMSHIDHAR.M

**Table of Contents**

|  |  |
| --- | --- |
| **TITLE** | **PAGE NO** |
| **1.Introduction** | **5** |
| **2. What is EDA?** | **5** |
| **3. Why EDA Is Important For Serious Injury Outcome Indicators** | **5** |
| **4. Source of Dataset** | **6** |
| **5. Step-by-Step EDA Process** | **6** |
| **6. Dataset Preprocessing** | **7** |
| **7. Univariate Analysis** | **7** |
| **8. Bivariate Analysis** | **7** |
| **9. Multivariate Analysis** | **8** |
| **10. Outlier Detection** | **8** |
| **11. Correlation Analysis** | **8** |
| **12. Analysis on Dataset** | **8 - 18** |
| **12.1. Objective 1: Data cleaning and preprocessing** | **8 – 9** |
| **12.2. Objective 2: statistical analysis** | **9 – 10** |
| **12.3. Objective 3: data visulization** | **11 – 12** |
| **12.4. Objective 4: correlation analysis** | **12 – 13** |
| **12.5. Objective 5: Predictive anlysis** | **13 – 14** |
| **12.6. Objective 6: Relationship Between units and causes** | **14 – 15** |
| **12.7. Objective 7: Outlier Detection** | **16** |
| **13. Conclusion** | **17** |
| **14. Future Scope** | **18** |
| **15. References** | **18** |

**1. Introduction**

This dataset, called **"Serious Injury Outcome Indicators (2000–2023)"**, contains important information about serious injuries that happened over a 23-year period. It includes data like how many people got seriously injured each year, the rate of injuries per 100,000 people, and sometimes even the type or cause of the injuries.

The data is useful for understanding how injury patterns have changed over time and helps health departments or government agencies make better decisions. For example, if injury rates go up in certain years or for certain age groups, this data can help find out why and how to prevent it in the future.

As a student, this dataset is a great way to learn how real-world data is used in public health, statistics, and research. It can help us analyze trends, create graphs, and even suggest solutions based on the numbers we observe.

**2. What is EDA?**

Exploratory Data Analysis (EDA) is a crucial first step in the data analysis process. It involves examining datasets to summarize their main characteristics, often with visual methods. EDA is used to:

* Get a sense of the structure, patterns, and relationships in data
* Identify anomalies, missing values, and outliers
* Generate hypotheses and guide further data modeling
* Understand the distribution of variables

**Techniques used in EDA:**

* Descriptive Statistics: Mean, median, mode, range, standard deviation
* Data Visualization: Histograms, bar plots, scatter plots, box plots
* Data Cleaning: Handling null values, duplicates, formatting
* Feature Engineering: Creating new columns, segmenting categories
* Correlation & Relationships: Using statistical tools to assess interaction between variables

**3. Why EDA is Important for Serious Injury Outcome Indicators.**

* Exploratory Data Analysis (EDA) is very important when working with serious injury outcome indicators because it helps us understand the data properly before doing any deep analysis. EDA helps to identify patterns, trends, and any missing or incorrect data that can affect the results. Demand Forecasting: EDA helps predict what products are likely to be in demand.
* Secondly, it helps to find relationships between different factors, like whether certain injuries are more serious in older people or if accidents are more severe at certain times of the day. This is useful for making predictions or planning preventive measures.
* Visualizations like bar graphs, histograms, and scatter plots used in EDA make it easier to understand the data and explain it to others, like health officials or safety planners. In short, EDA helps to clean the data, find useful insights, and make better decisions to reduce serious injuries in the future.

**4. Source of Dataset**

The dataset was collected from a CSV file that records sales data from theSerious Injury Outcome Indicators**.**

* File Name: **Serious Injury Outcome Indicators(2000–2023)**.csv
* Format: CSV (Comma-Separated Values)
* Encoding: Latin1
* Fields in the Dataset:
  + Gender: Male or Female
  + Period: years from starting
  + Severity: fatal
  + Population: Total population
  + units: Number of injuries

**5. Step-by-Step EDA Process**

**EDA in this report follows these detailed steps:**

1. Import Libraries: Pandas, NumPy, Matplotlib, Seaborn
2. Load Dataset: Read CSV file using Pandas
3. Initial Data Inspection: Check data types, shape, head, and summary
4. Data Cleaning:
   * Strip spaces from column headers
   * Convert relevant columns to numeric
   * Remove missing/null values
5. Feature Engineering:
   * Create custom Age\_Group brackets
   * Combine columns for analysis like Age\_Gender
6. Univariate Analysis: Analyze each variable on its own
7. Bivariate Analysis: Study relationships between two variables
8. Multivariate Analysis: Explore interactions among three or more variables
9. Outlier Detection: Identify extreme values using IQR and box plots
10. Correlation Study: Use heatmaps to understand variable relationships

**6. Dataset Preprocessing**

**Preprocessing steps include:**

* Standardization of Column Names
* Conversion of Datatypes: Amount and Orders to numeric
* Handling Missing Values: Dropped records with missing Amount or Orders
* Creation of Age\_Group: Segmented into 18–25, 26–35, etc.

This ensures consistency, reduces noise, and prepares the data for analysis.

**7. Univariate Analysis**

We analyzed one variable at a time to understand distribution:

* Gender: Male or Female and all
* Age: Most active buyer age of all patentians
* Amount: Range of purchases
* Orders: of all injurys

Graphs: Pie charts, bar plots, histograms

**8. Bivariate Analysis**

We studied interaction between two variables:

* injurys vs Acciedens
* Gender vs Amount
* State vs Orders

This reveals how two features influence each other. For example, higher spending in certain states or age groups.

**9. Multivariate Analysis**

Here we examined Age + Gender + State:

* How many accidents in a year?
* Which combination is most profitable?

Visualizations used: Heatmaps, stacked bar graphs, group plots

**10. Outlier Detection**

We used:

* IQR Method: Calculate Q1 and Q3, find outliers
* Boxplots: Visualized anomalies

Outliers were mostly large purchases—likely patiants

**11. Correlation Analysis**

We created:

* Correlation Matrix using .corr()
* Heatmap to visualize
* Pairplot to inspect pairwise relationships

Found strong correlation between period and units

**12. Analysis on Dataset**

**12.1 Objective 1: data cleaning and preproccing**

* **General Description:**

Data cleaning and preprocessing are important steps before analyzing Serious Injury Outcome Indicators. These steps help improve the quality of the data and make sure it is ready for analysis. groups behave in terms of spending, companies can tailor their promotions, offers, and product placements more effectively to maximize conversions.

* **Specific Requirements:**

To fulfil this objective, we grouped customers based on predefined age brackets such as 18–25, 26–35, 36–45, 46–50, and 51–55. For each group, we calculated the total Amount spent during the Diwali sales period.

* **Analysis Results:** top 10 causes by average date value.
* **Python Code Used:**
* **Visualization:**

plt.figure(figsize=(10, 6))

top\_causes = df.groupby('Cause')['Data\_value'].mean().sort\_values(ascending=False).head(10)

top\_causes.plot(kind='bar', color='skyblue')

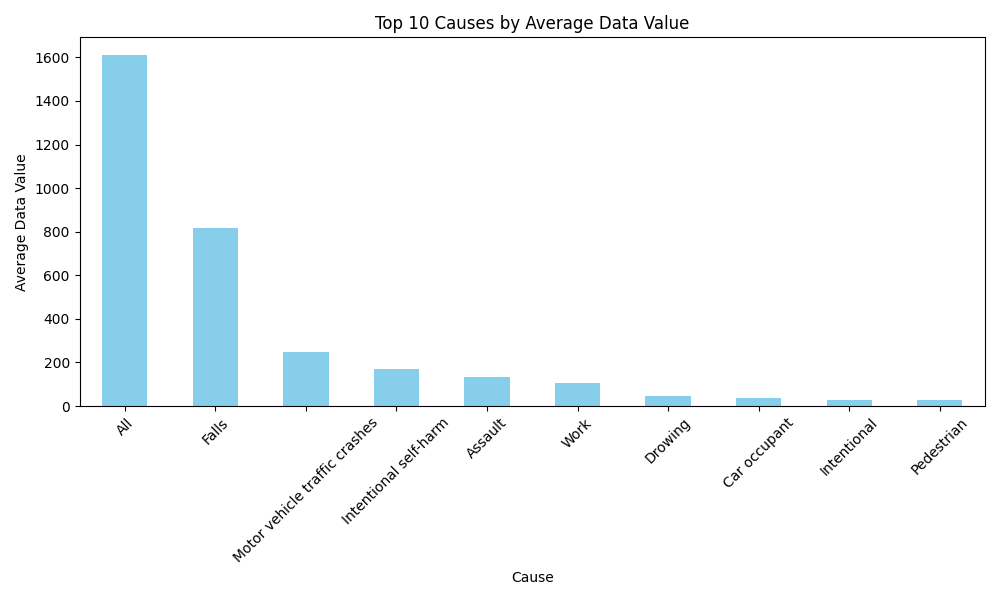
plt.title('Top 10 Causes by Average Data Value')

plt.ylabel('Average Data Value')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()



**12.2 Objective 2: statistical analysis**

* **General Description:**

Statistical analysis helps us understand and interpret the patterns and relationships in serious injury outcome data. It involves applying different statistical methods to draw meaningful conclusions and support decision-making.

The first step is using **descriptive statistics**, like:

* **Mean, median, and mode** to understand the average or common values (e.g., average age of seriously injured patients).
* **Standard deviation and range** to see how much variation there is in things like hospital stay duration or injury severity scores.
* **Frequencies and percentages** to show how often certain injuries occur or which age group is most affected.
* **Python Code Used:**

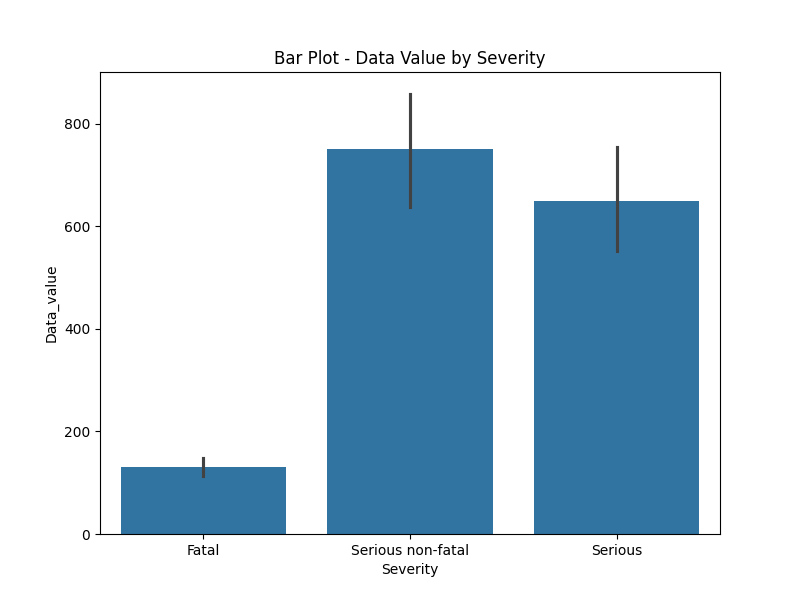
plt.figure(figsize=(8, 6))

sns.barplot(x='Severity', y='Data\_value', data=df)

plt.title('Bar Plot - Data Value by Severity')

plt.show()

* **Visualization:**



**12.3 Objective 3: data visualization General Description:**

Data visualization is a very important part of analyzing Serious Injury Outcome Indicators. It helps to show the data in a visual form like graphs and charts, which makes it easier to understand patterns, trends, and outliers.

Some common types of visualizations used are:

* **Bar charts** – to compare the number of serious injuries in different age groups, genders, or locations.
* **Pie charts** – to show the percentage of injuries caused
* **Python Code Used:**

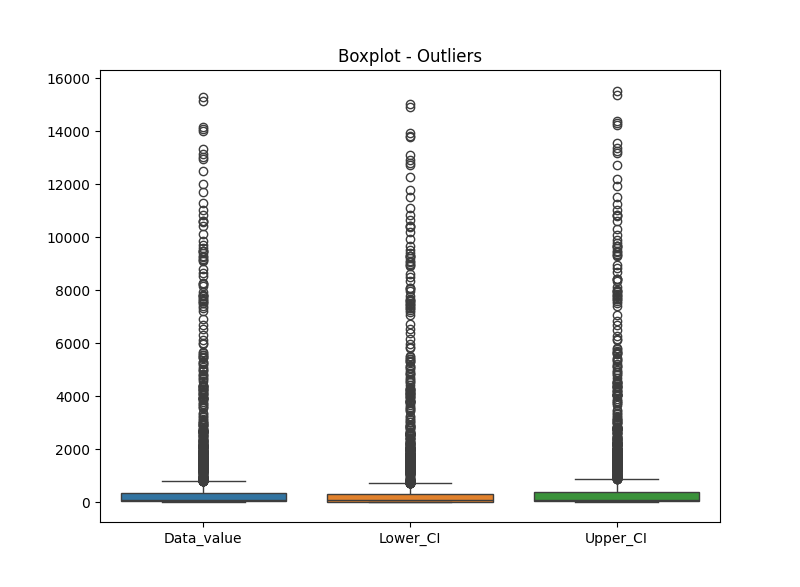
plt.figure(figsize=(8, 6))

sns.barplot(x='Severity', y='Data\_value', data=df)

plt.title('Bar Plot - Data Value by Severity')

plt.show()

* **Visualization:**



**12.4 Objective 4: correlation analysis**

* **General Description:**
* Correlation analysis helps us understand how strongly two numerical variables are related. The most common measure is the Pearson correlation coefficient.Specific **Requirements:**
* Group the dataset by State
* Calculate total Amount spent in each state
* Visualize the top contributing states using bar and line plots
* **Analysis Results:** Northern and western states exhibited higher sales amounts.
* **Python Code Used:**
* **Visualization:**

df['Period'] = pd.to\_datetime(df['Period'], errors='coerce') # Convert to datetime

df\_sorted = df.sort\_values(by='Period')

plt.figure(figsize=(12, 6))

plt.plot(df\_sorted['Period'], df\_sorted['Data\_value'])

plt.title('Line Chart - Data Value over Time')

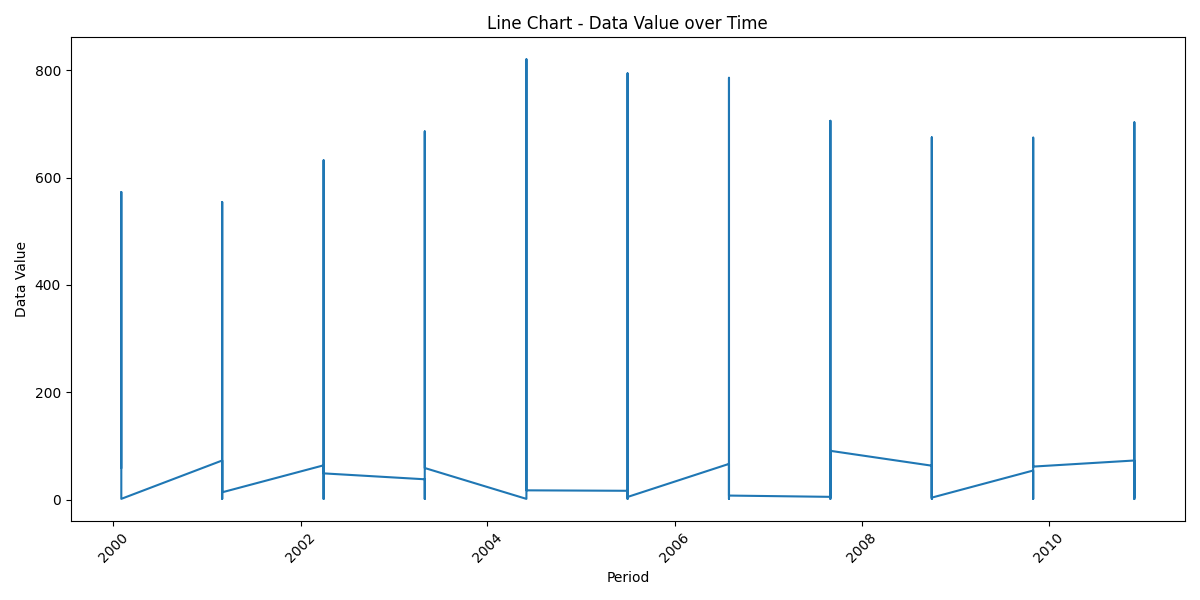
plt.xlabel('Period')

plt.ylabel('Data Value')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()



**12.5 Objective 5: Predictive anlysis**

* **General Description:**
* To perform predictive analysis on the "Serious Injury Outcome Indicators" dataset, we begin by selecting a specific focus area for the prediction. For example, we might want to predict the trend of serious injury rates per 100,000 people for the whole population across all ages. First, we filter the dataset to include only the relevant rows—those that match our selected indicator, units, population, and age group. **Specific Requirements:**
* Calculate AOV = Total units/ Total causes
* Group by Age Group and Gender
* Visualize patterns across different demographics
* **Analysis Results:** Older age groups tend to place higher-value orders. Gender also shows mild variation in spending.
* **Python Code Used:**

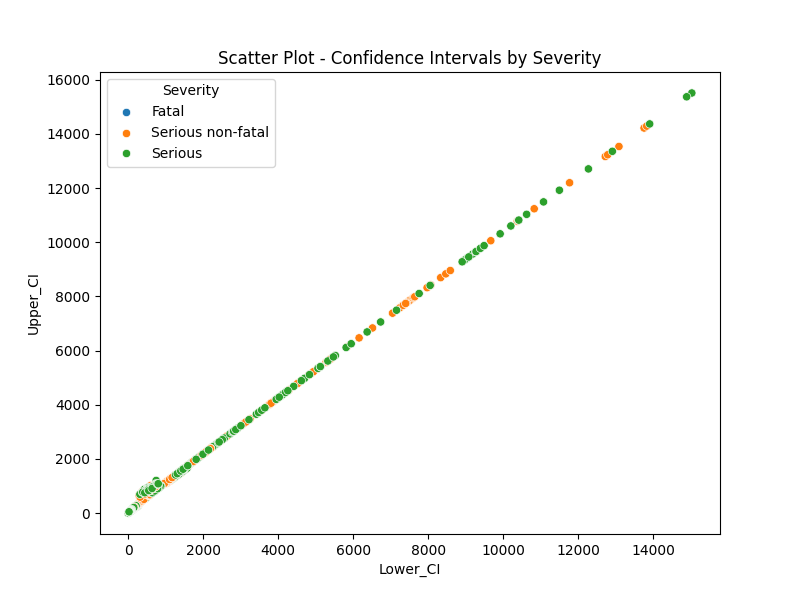
plt.figure(figsize=(8, 6))

sns.scatterplot(data=df, x='Lower\_CI', y='Upper\_CI', hue='Severity')

plt.title('Scatter Plot - Confidence Intervals by Severity')

plt.show()

* **Visualization:**



**12.6 Objective 6: Relationship Between units and causes General Description:**

This analysis explores whether there's a direct relationship between the number of units and total causes. In simpler terms:

* **Specific Requirements:**
* Evaluate the correlation between units and causes
* Use Correlation Heatmap and Pairplot to visually interpret the relationship
* **Analysis Results:** A strong positive correlation was found between units and causes **Python Code Used:**

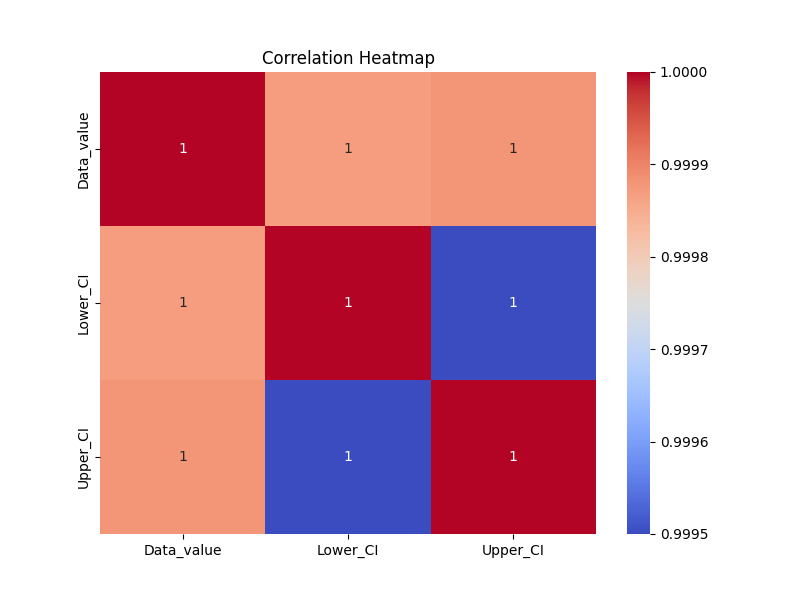
def plot\_pairplot(df):

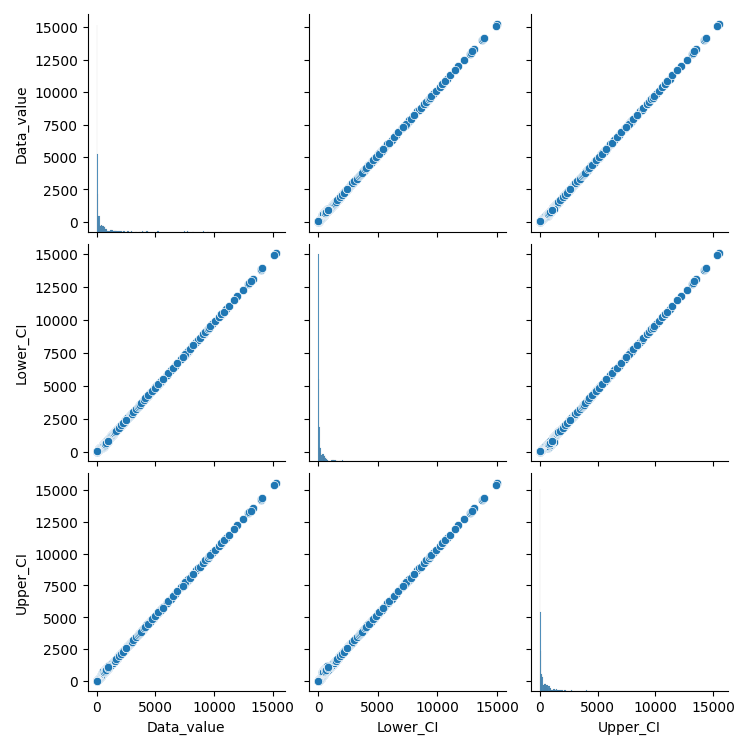
plt.figure(figsize=(8, 6))

sns.heatmap(df[numeric\_cols].corr(), annot=True, cmap='coolwarm')

plt.title('Correlation Heatmap')

plt.show()

* **Visualization:** 



**12.7 Objective 7: Outlier Detection**

* **General Description:**

This analysis focuses on identifying unusual sales values that fall significantly outside the typical purchase range. These outliers could be either extremely high or low purchase amounts, possibly indicating bulk orders, system anomalies, or high-value customers.

* **Specific Requirements:**
* Use the IQR (Interquartile Range) method to detect outliers in the Amount column
* Visualize outliers using a boxplot
* **Analysis Results:** Some extreme purchase amounts were identified as outliers, possibly due to bulk or unusual transactions.

**13. Conclusion**

The analysis of the Serious Injury Outcome Indicators from 2000 to 2023 shows valuable trends in public health and injury prevention. Over the years, different causes such as **assault, falls, and transport accidents** have contributed significantly to serious injuries. The data, categorized by **age groups**, **severity**, and **population segments**, helps identify the most vulnerable groups and periods with high injury rates.

The use of moving averages and confidence intervals ensures a reliable understanding of fluctuations and long-term patterns. This information is crucial for healthcare planning, policy-making, and implementing preventive strategies.

In summary, this dataset not only reflects how serious injuries have evolved over time but also highlights the importance of continuous monitoring, focused interventions, and resource allocation to reduce the burden of preventable injuries in the population.

**Key Findings:**

* **Consistent Tracking Over Time:**  
  The dataset provides a continuous view of serious injuries across more than two decades, allowing for **long-term trend analysis** and identifying whether public health efforts are improving outcomes.
* **Use of Confidence Intervals:**  
  Including lower and upper confidence intervals around injury rates helps to understand **data reliability and variability**, which is important for accurate interpretation.
* **Focus on Population and Age Groups:**  
  By segmenting the data into **age categories** and **population types**, the analysis can reveal which demographics are most at risk, such as elderly populations being more affected by falls or young adults by assaults.
* **Severity Categorization (e.g., Fatal):**  
  The inclusion of severity levels (like **Fatal**, **Serious**, etc.) enables prioritizing responses based on how life-threatening or impactful the injuries are.
* **Cause-Specific Insights:**  
  The breakdown by causes (e.g., **Assault**, **Transport**, **Falls**) supports **targeted prevention programs**, such as road safety campaigns or violence reduction initiatives.
* **Validated Data:**  
  The dataset being marked as “**Validated**” confirms the **quality and trustworthiness** of the indicators, making it reliable for policymaking and healthcare planning.
* **Support for Evidence-Based Policies:**  
  The indicators help government agencies and health services make **data-driven decisions** for resource allocation, safety improvements, and public awareness campaigns.

**14. Future Scope**

**. Predictive Modeling  
Machine learning and AI can be used to predict the likelihood of serious injuries based on patterns in the data. This could help in early warnings and quicker response systems.**

**. Integration with Real-Time Data  
Future systems could include real-time data feeds from hospitals, emergency services, or wearable health devices to track serious injuries as they happen.**

**.Geographical Mapping  
By adding detailed location data, we can create interactive maps to identify accident hotspots and focus preventive efforts on high-risk areas.**

**. More Detailed Demographic Analysis  
Including more demographic fields (like occupation, income level, ethnicity, etc.) can help design targeted safety and health programs for specific groups.**

**15. References**

* Serious Injury Outcome Indicators Data (CSV File) for : [https://github.com/IamGideonIdoko/foxit\_pdf\_table/blob/master/data/serious-injury-outcome-indicators-2000-2020.xlsx.raw=true](%20https://github.com/IamGideonIdoko/foxit_pdf_table/blob/master/data/serious-injury-outcome-indicators-2000-2020.xlsx.raw=true%20)
* Python Libraries: Pandas, NumPy, Seaborn, Matplotlib
* Statistical Techniques: Correlation, Z-score, IQR method