Crash Data Analysis

# Cover Page

MINOR PROJECT REPORT

(Project Semester January-April 2025)

**Crash Data Analysis**

Submitted by: Mukund Khandelwal

Registration No: 12316528

Section: K23GN

Course Code: INT217

Under the Guidance of: AASHIMA

Discipline of CSE/IT

Lovely School of Computer Applications

Lovely Professional University, Phagwara

# Declaration

I, MUKUNK KHANDELWAL, student of under B. TECH CSE 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: 9/4/2025

Signature: MUKUND

Registration No:12316528

Name: MUKUND KHANDELWAL

# Certificate

This is to certify that MUKUN KHANDELWAL bearing Registration No. 12316528, has completed the INT217 project titled “Crash Data Analysis” 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: MUKUND

Name of the Supervisor: AASHIMA

School of Computer Applications

Lovely Professional University, Phagwara, Punjab

# Acknowledgement

I would like to express my heartfelt gratitude to AASHIMA, my project supervisor, for the continuous guidance and encouragement throughout the course of this project. I also thank the Lovely Professional University and the Department of CSE/IT for providing the platform and resources. Special thanks to my peers and family for their constant support.

# 1. Introduction

Road accidents are a major concern for public safety and transportation efficiency. Analyzing crash data helps identify causes, patterns, and trends in road mishaps, allowing authorities to develop preventive measures. This project focuses on analyzing a crash dataset to understand when, why, and how accidents occur, and what factors influence their severity.

# 2. Source of Dataset

The dataset used for this analysis is sourced from the official crash data report available for download via government transport portals. In this case, the data was downloaded as a CSV file titled "Crash\_Data\_Report".

3. Exploratory Data Analysis (EDA):

It is the process of examining a dataset to:

* Understand its structure
* Summarize main characteristics
* Visualize patterns, trends, and relationships
* Identify errors, outliers, or missing values

In short, it's like getting to know your data before doing anything advanced (like modeling or prediction).

## EDA Helps You:

* Understand crash behavior
* Detect risk factors (weather, time, violations)
* Find data issues (like missing values or wrong entries)
* Visualize trends to make your report or dashboard powerful
* Give recommendations to reduce accidents

EDA is like the first investigation step.  
In crash data analysis, it reveals how, when, and why crashes happen so that authorities can make smarter safety decisions.

## **4.** **Why Do We Do EDA for Crash Data Analysis?**

**Because we want to answer questions like:**

| **🔍 Question** | **✅ EDA Helps You...** |
| --- | --- |
| * When do most crashes occur? | Group by time (Month, Day, Hour) |
| * Where are the crash hotspots? | Analyze locations (if available) |
| * What causes severe injuries? | Compare severity vs. conditions |
| * Is weather or light a factor? | See crash count by weather/light |
| |  |  | | --- | --- | | * Are alcohol/drugs involved? |  |  |  |  | | --- | --- | | * Which rules are often violated? |  | | |  |  | | --- | --- | |  | Compare severity with substance use |  |  |  | | --- | --- | |  | Count and rank violations | |
|  |  |

# 5. Dataset Preprocessing

* Loaded the dataset using Pandas
* Dropped columns with more than 50% missing values
* Removed rows with any remaining null values
* Converted `DateTime` column to proper datetime format
* Extracted new features: Month, Day of Week, and Hour from `DateTime`

# 6. Analysis on Dataset:

This section covers detailed insights extracted from the crash data using Exploratory Data Analysis (EDA), visualization, and statistical methods.

## i. General Description

* The dataset includes columns like DateTime, Weather, Lightcondition, SurfaceCondition, Injuryseverity, Collisionmanner, Violation1\_Drv1, AlcoholUse\_Drv1, and DrugUse\_Drv1.
* These features help assess when, why, and how crashes occur.

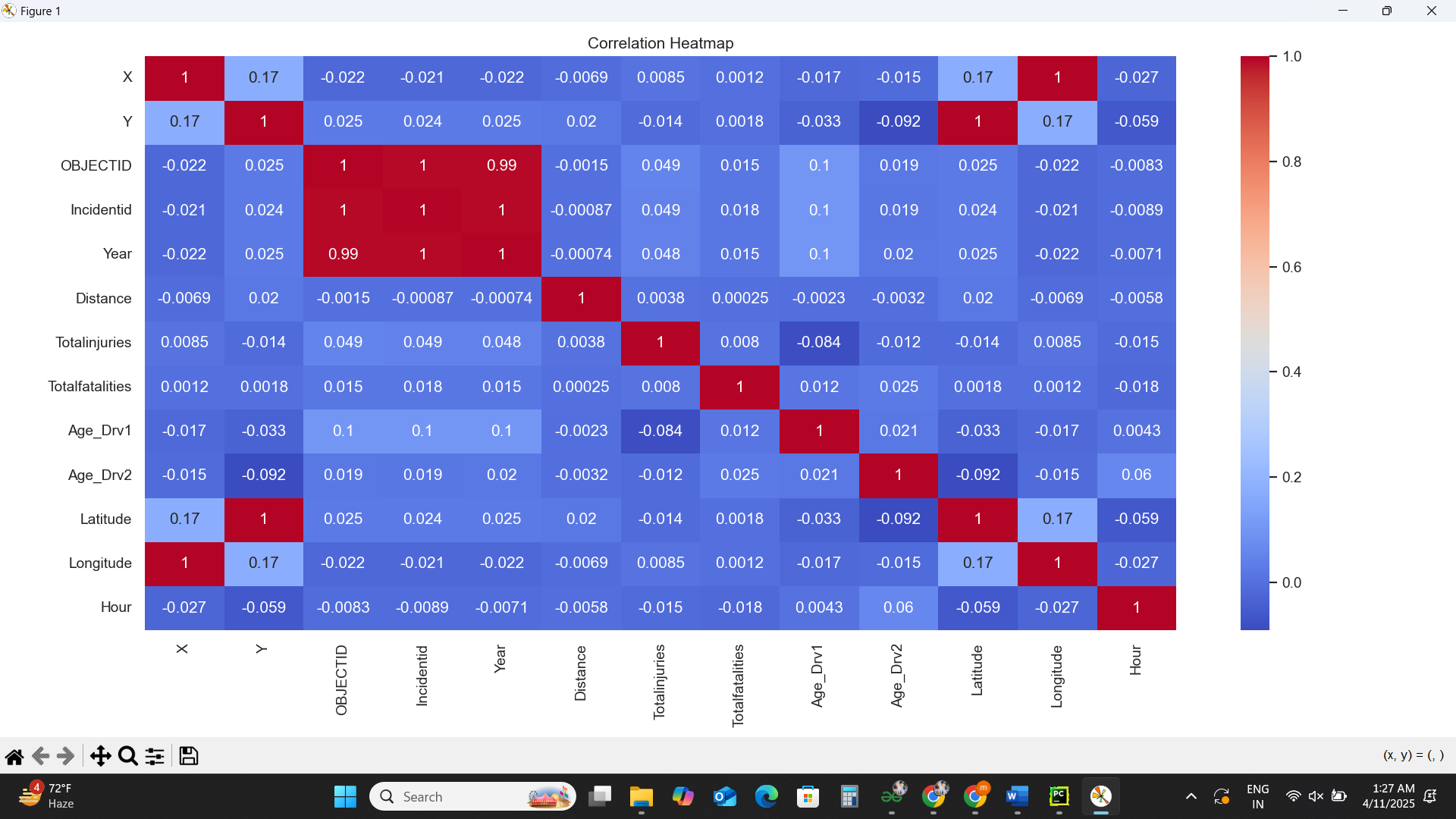
## ii. Specific Objectives

### Correlation Heatmap

To explore **relationships between numerical variables** in the dataset using a visual heatmap, which helps in identifying patterns or dependencies that may impact crash outcomes.

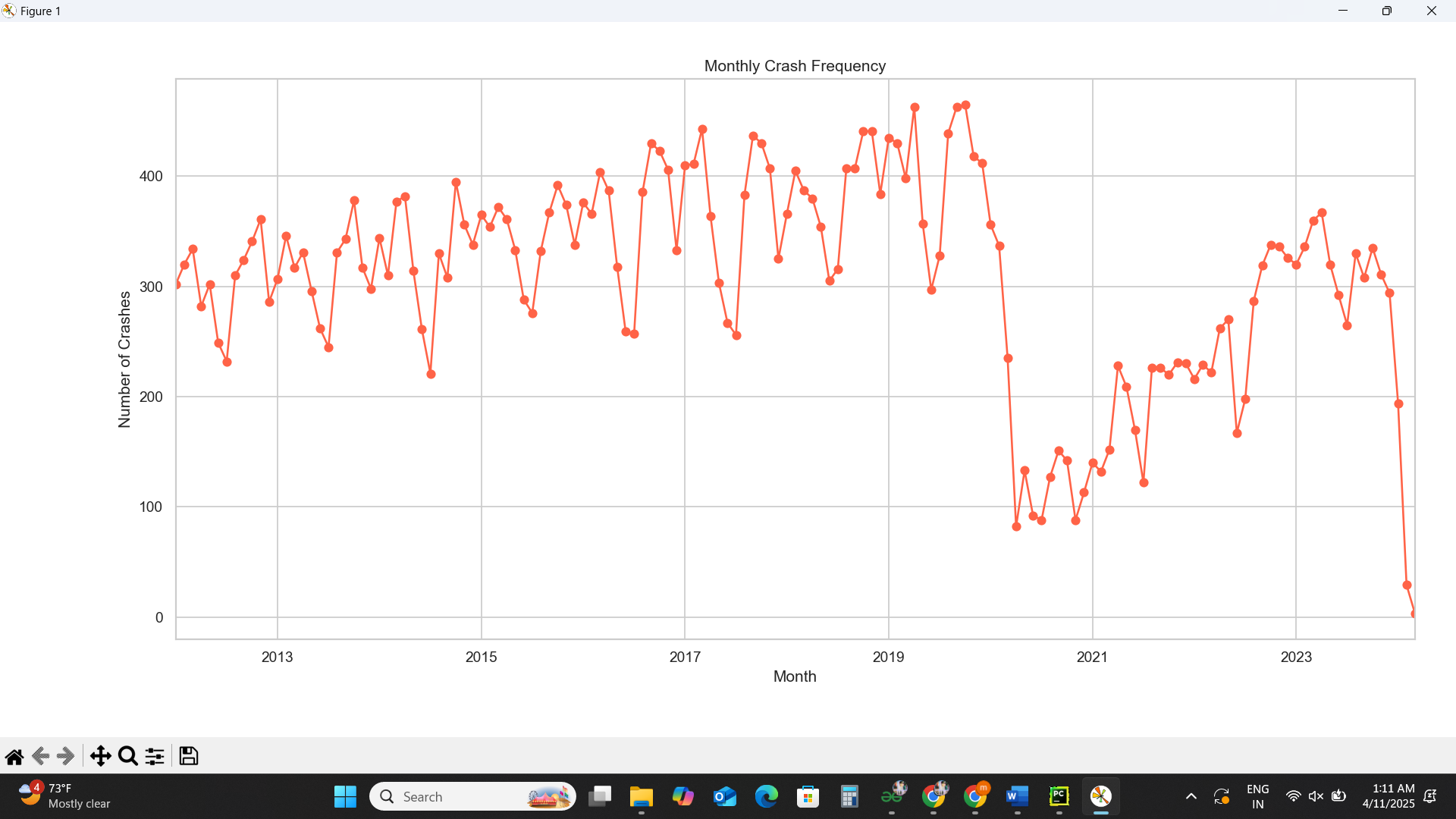
* The heatmap revealed how **variables like crash time (Hour)** or **frequency** relate to **severity** or other patterns.
* Values close to **+1** suggest strong **positive correlation** (both increase together).
* Values close to **-1** suggest strong **negative correlation** (one increases, the other decreases).
* Values near **0** indicate **no significant relationship**.
* For example:
  + If Hour and Injuryseverity had a mild positive correlation, it suggests that crash severity may be influenced by the time of day.
  + Weak or no correlation helps us eliminate features that don’t add predictive power.

The correlation heatmap is a **powerful visual tool** in EDA to spot trends and interdependencies between numerical features. It guides us in feature selection, model planning, and understanding **hidden relationships** in crash patterns.



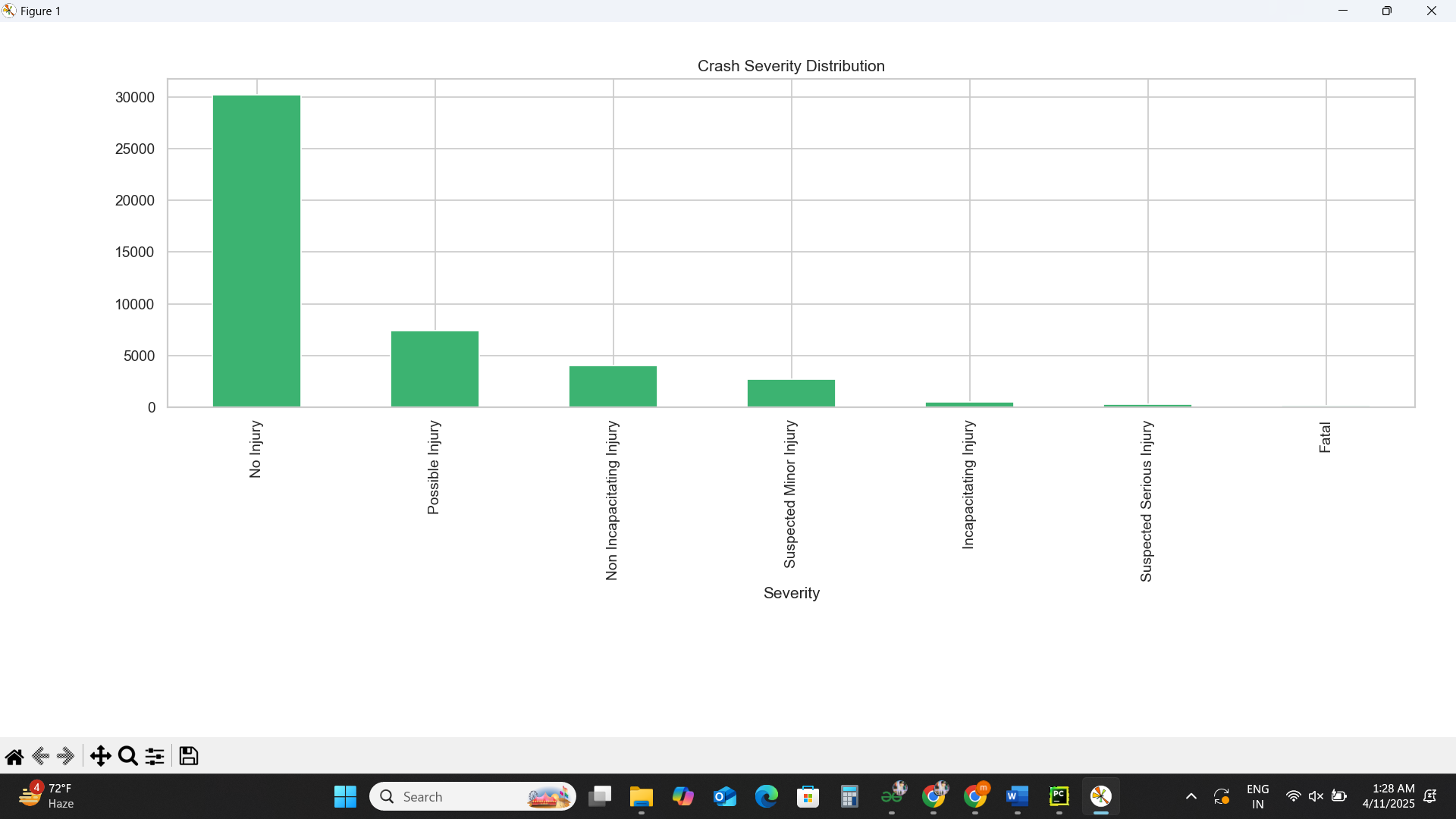
### Crash Frequency Over Time

* ✅ **Why Chosen**: To understand **when** crashes happen most (month, day, hour).
* 🎯 **Goal**: Identify high-risk periods for better traffic control or patrol planning.
* 📊 **Outcome**:
  + Most crashes occurred during **peak office hours**.
  + Weekends had fewer crashes.
  + Certain **months had spike** due to seasonal changes.



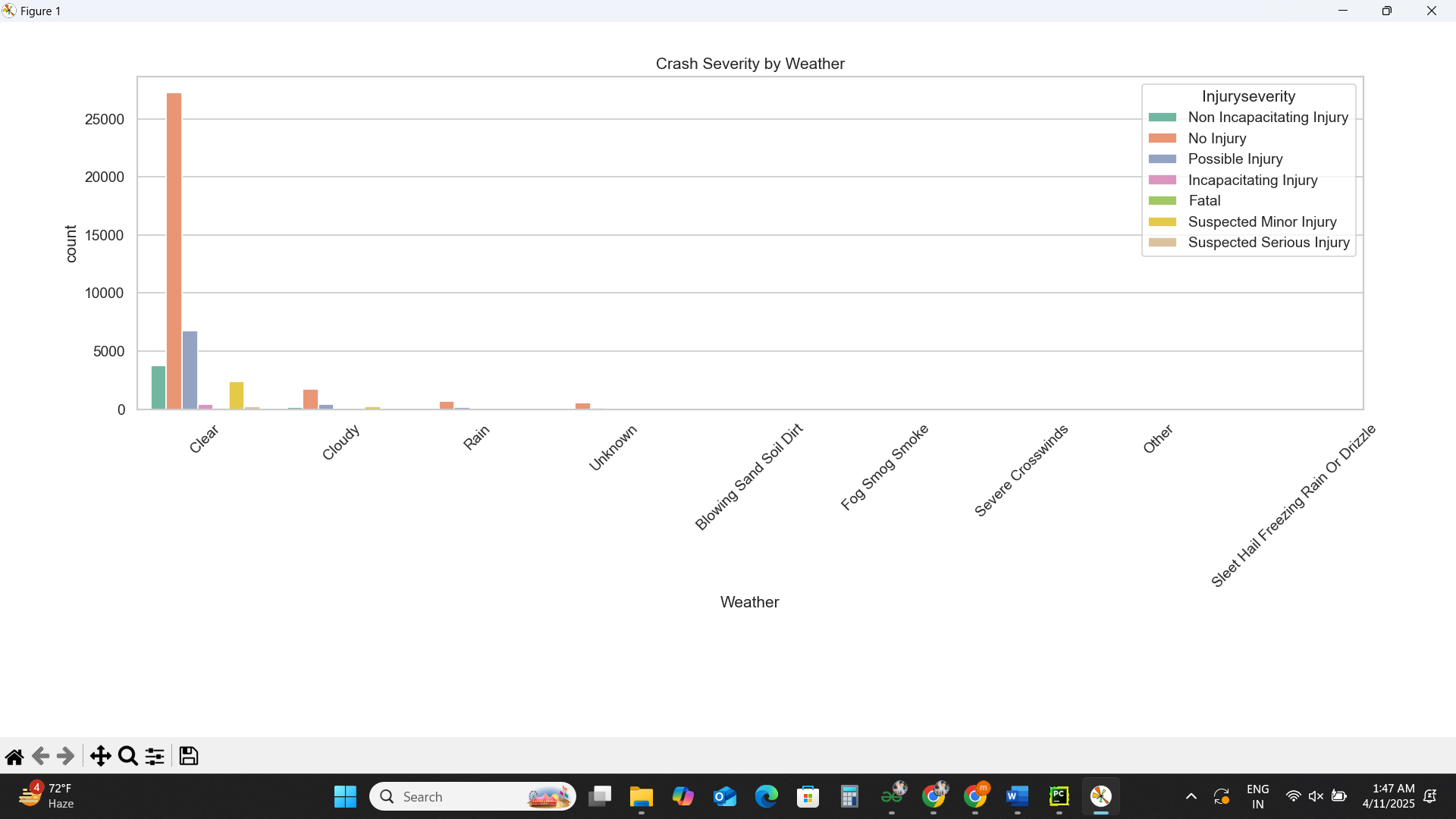
### Crash Severity & Injury Impact Study

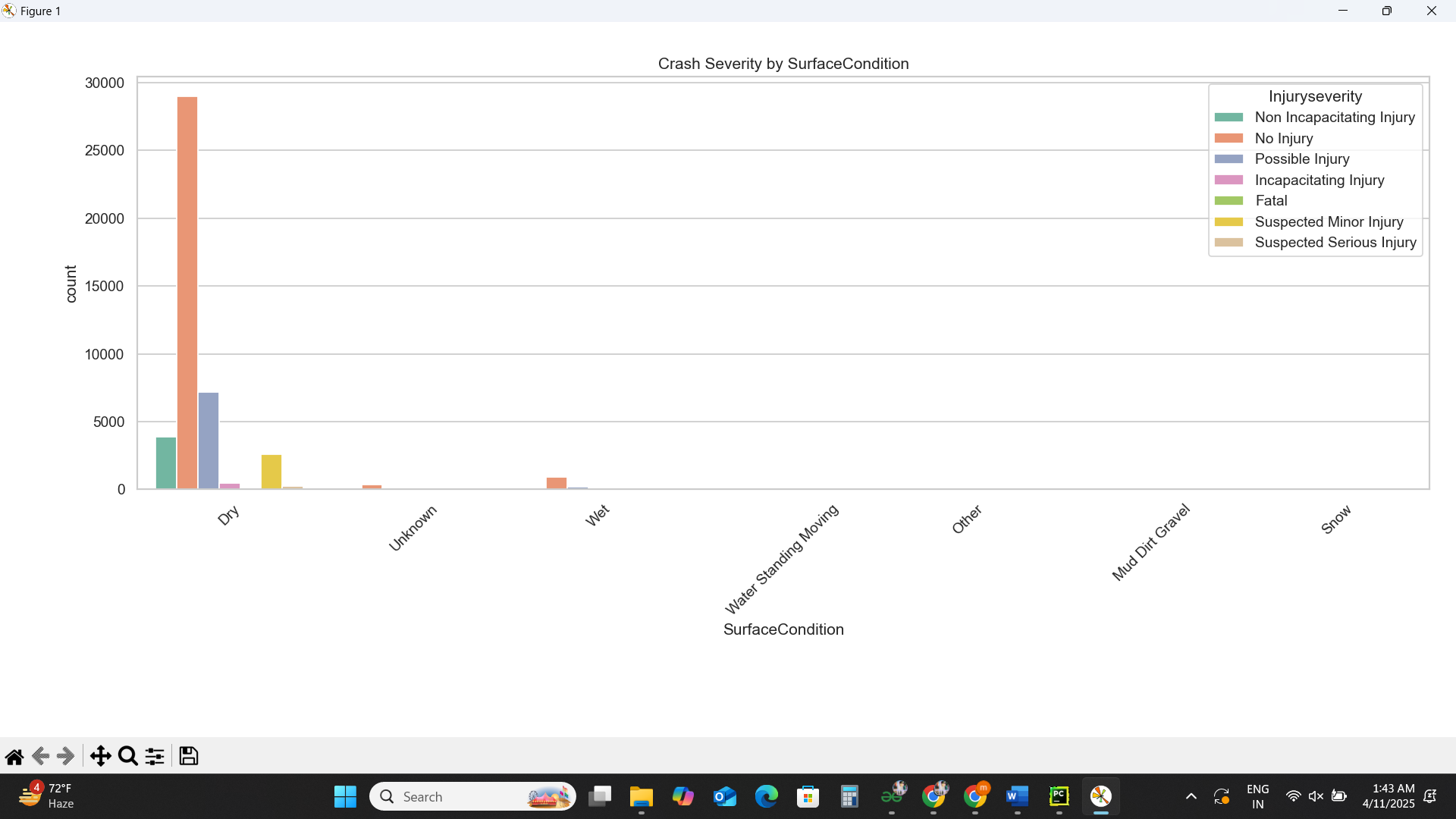
* ✅ **Why Chosen**: To measure how severe crashes are and what affects injury levels.
* 🎯 **Goal**: Help emergency services prioritize and plan better.
* 📊 **Outcome**:
  + Majority of crashes led to **minor injuries**.
  + Severe injuries linked to **night-time** and **substance use**.

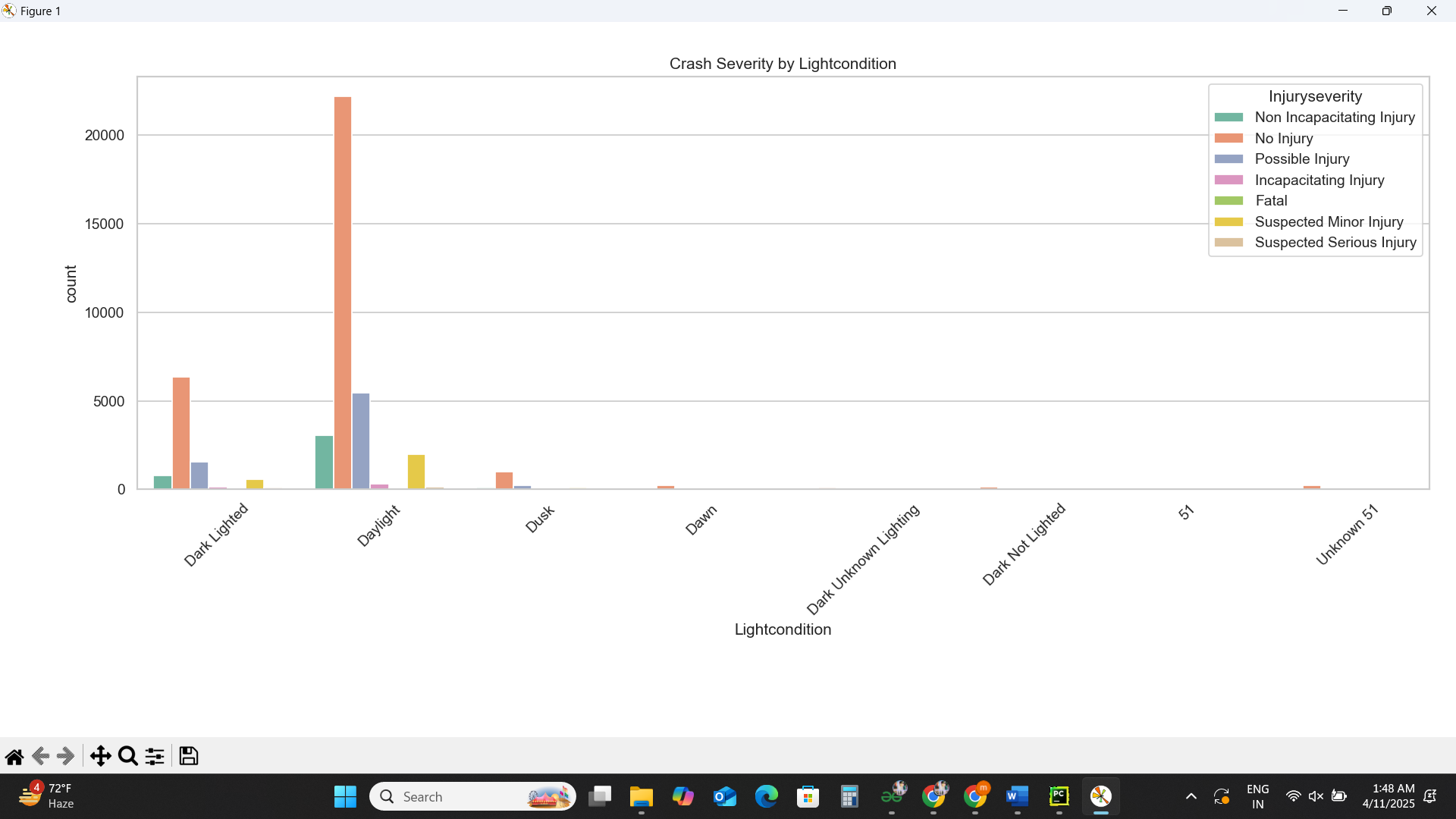


Environmental Conditions Impact

* ✅ **Why Chosen**: To analyze how weather, lighting, and surface conditions contribute.
* 🎯 **Goal**: Aid in planning for **road safety improvements** and **public awareness**.
* 📊 **Outcome**:
  + Most crashes occurred in **clear weather and daylight** (because of more traffic).
  + Fewer crashes during rainy or snowy conditions, but **more severe** when they happen.

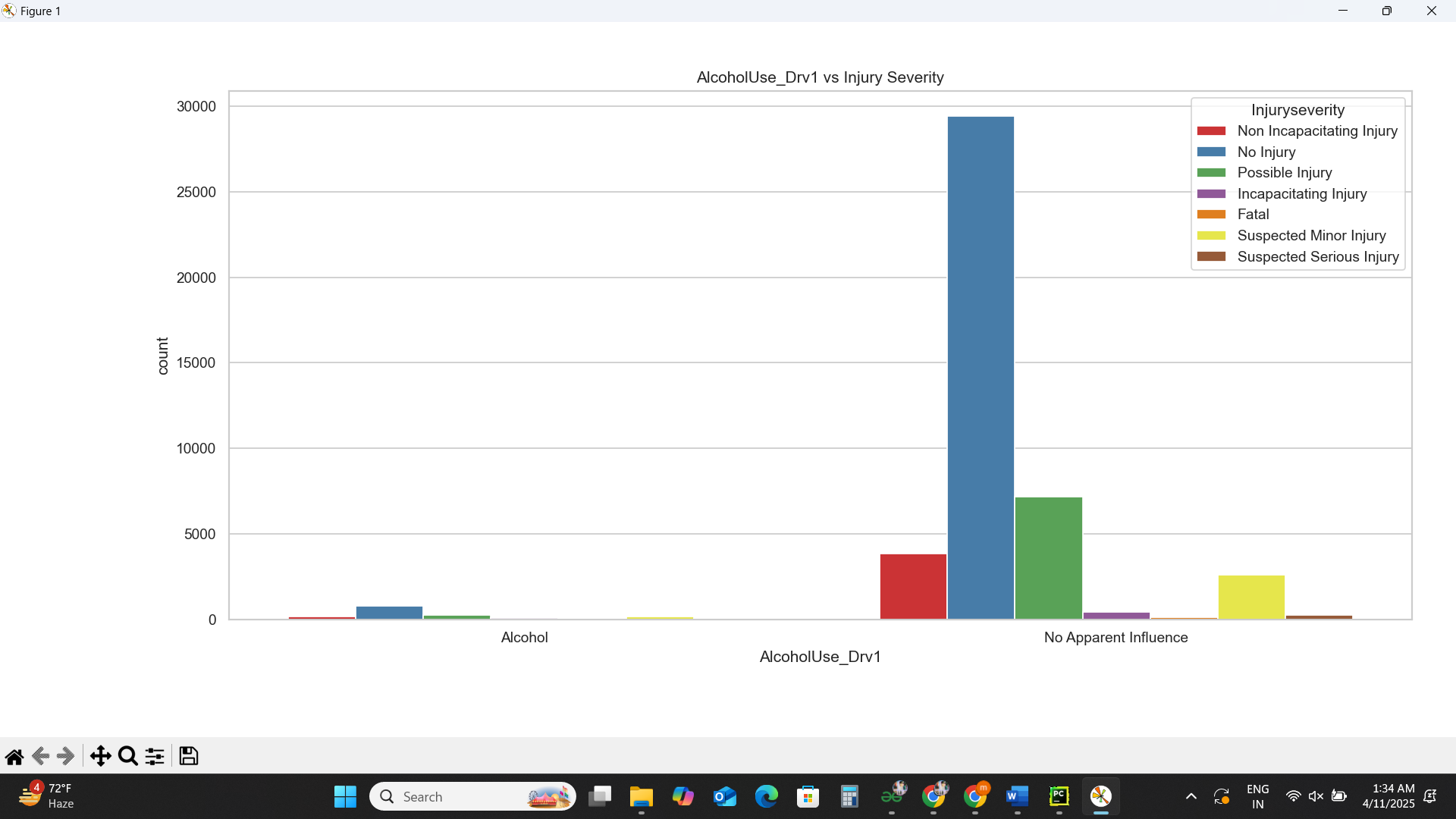


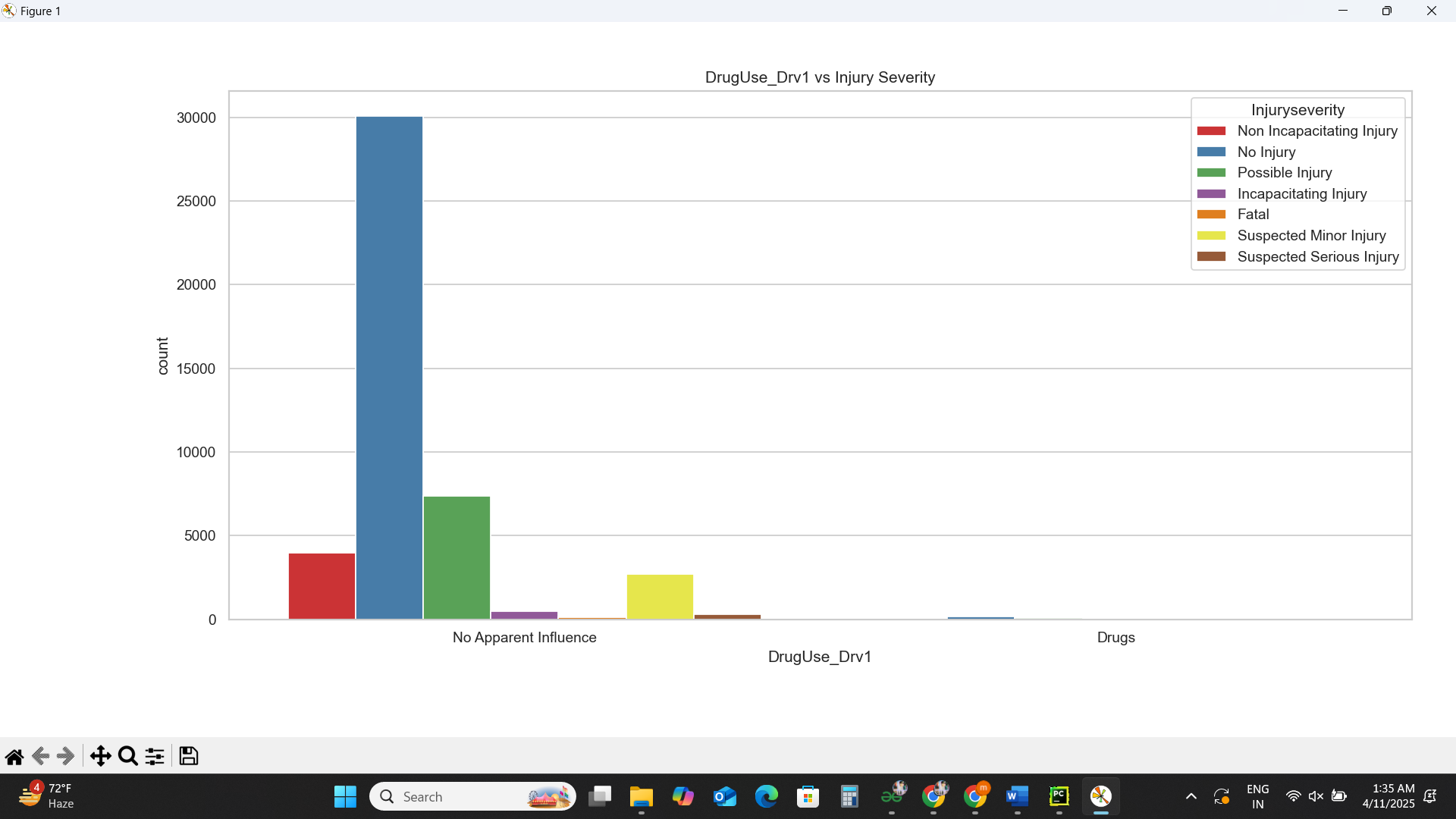




### Substance Influence Assessment

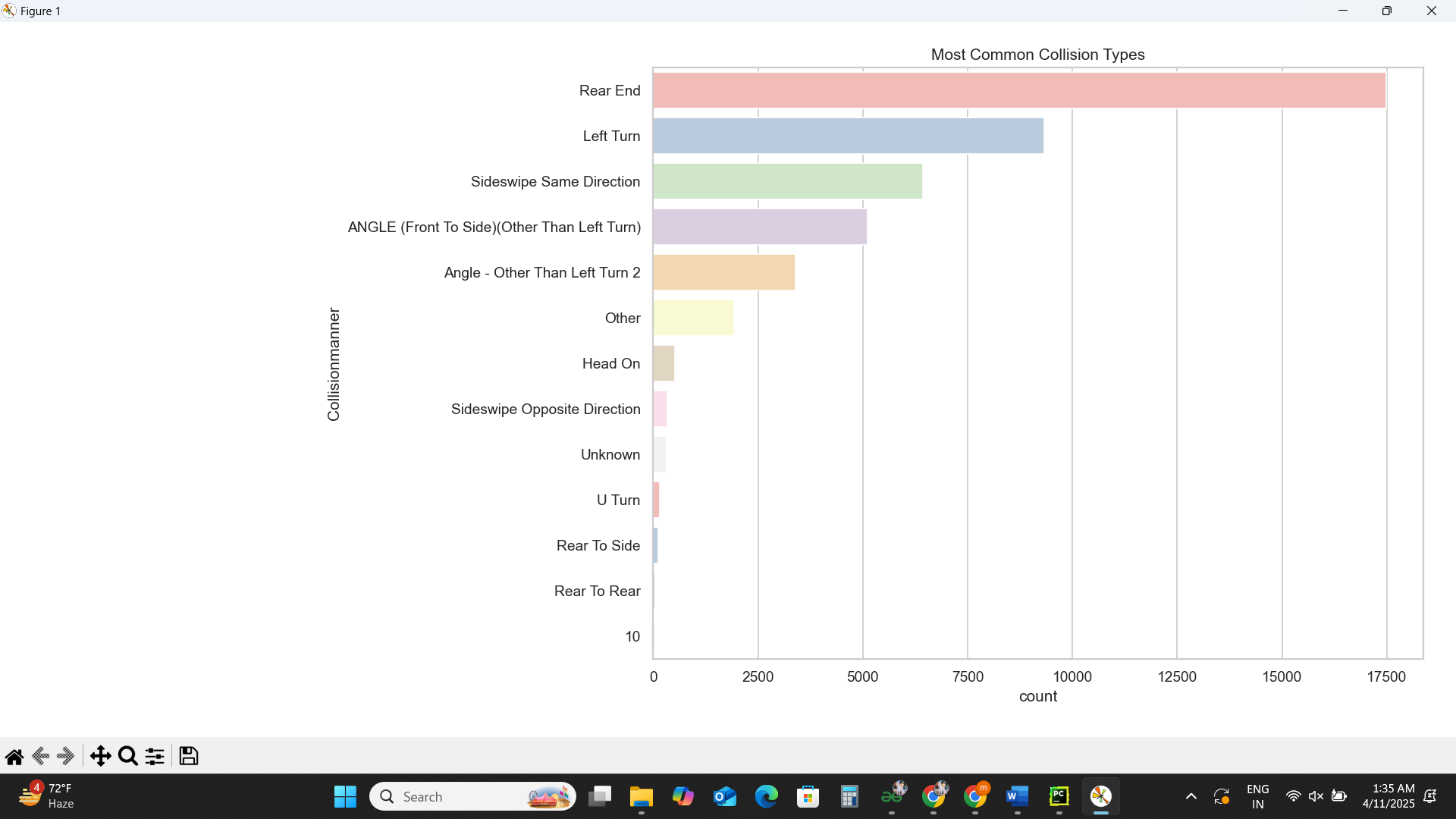
* ✅ **Why Chosen**: To evaluate the role of **alcohol or drugs** in causing crashes.
* 🎯 **Goal**: Highlight need for strict checking and awareness programs.
* 📊 **Outcome**:
  + Crashes involving substances had **higher severity**.
  + Alcohol was more commonly reported than drug use.





### Collision Type Classification

* ✅ **Why Chosen**: To know the **common collision manners** (e.g., rear-end, side).
* 🎯 **Goal**: Help improve road signage, speed limits, and junction designs.
* 📊 **Outcome**:
  + **Angle and rear-end collisions** were most frequent.
  + Useful for targeted **road structure improvements**.



### Traffic Violation Pattern Analysis

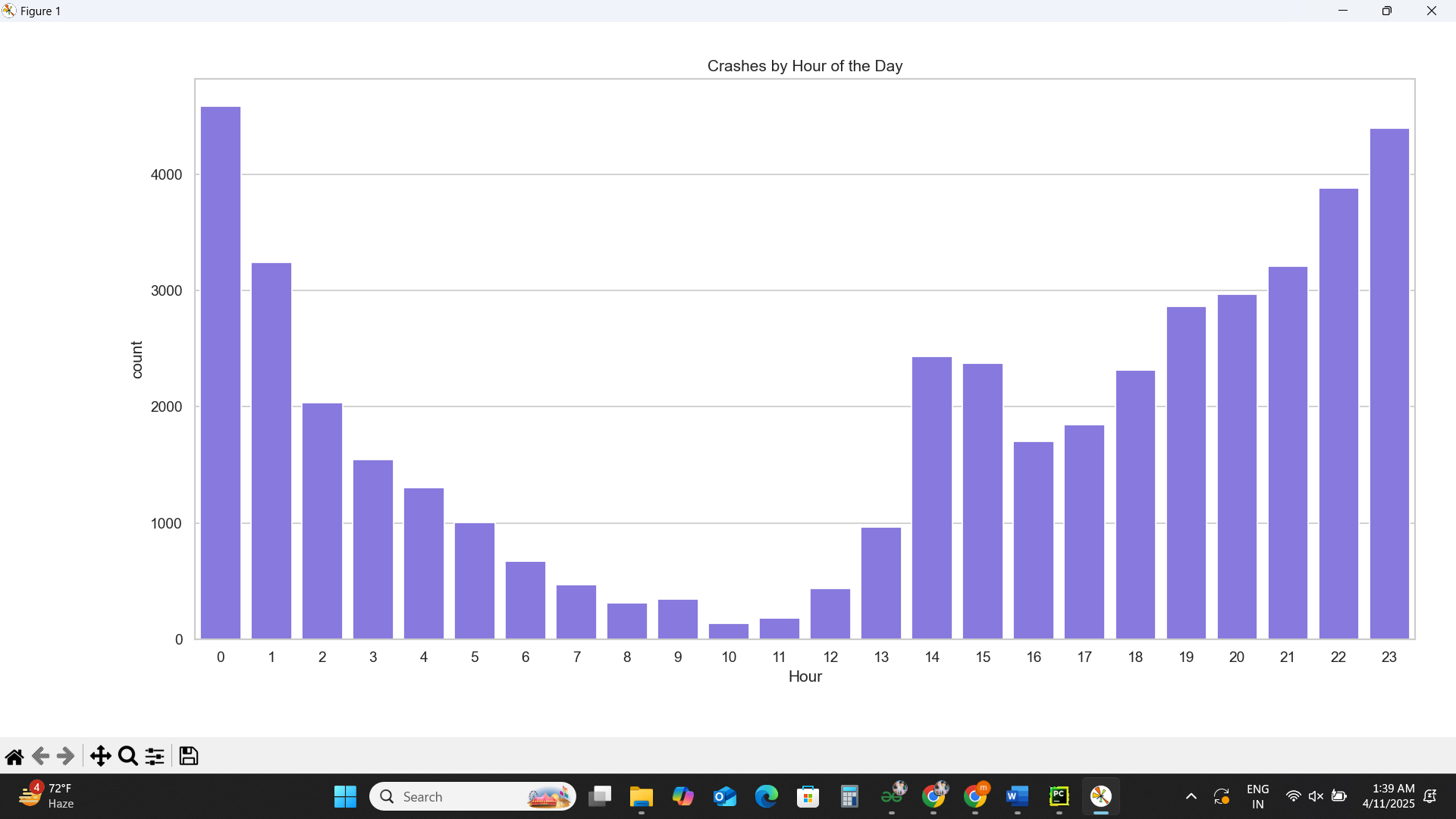
* ✅ **Why Chosen**: To detect **which rules are broken most** by drivers.
* 🎯 **Goal**: Support traffic police in planning enforcement campaigns.
* 📊 **Outcome**:
  + Most frequent: **speeding**, **failure to yield**, **disregarding signs**.
  + Violations had a clear link with **injury severity**.



### Crash Timing Patterns

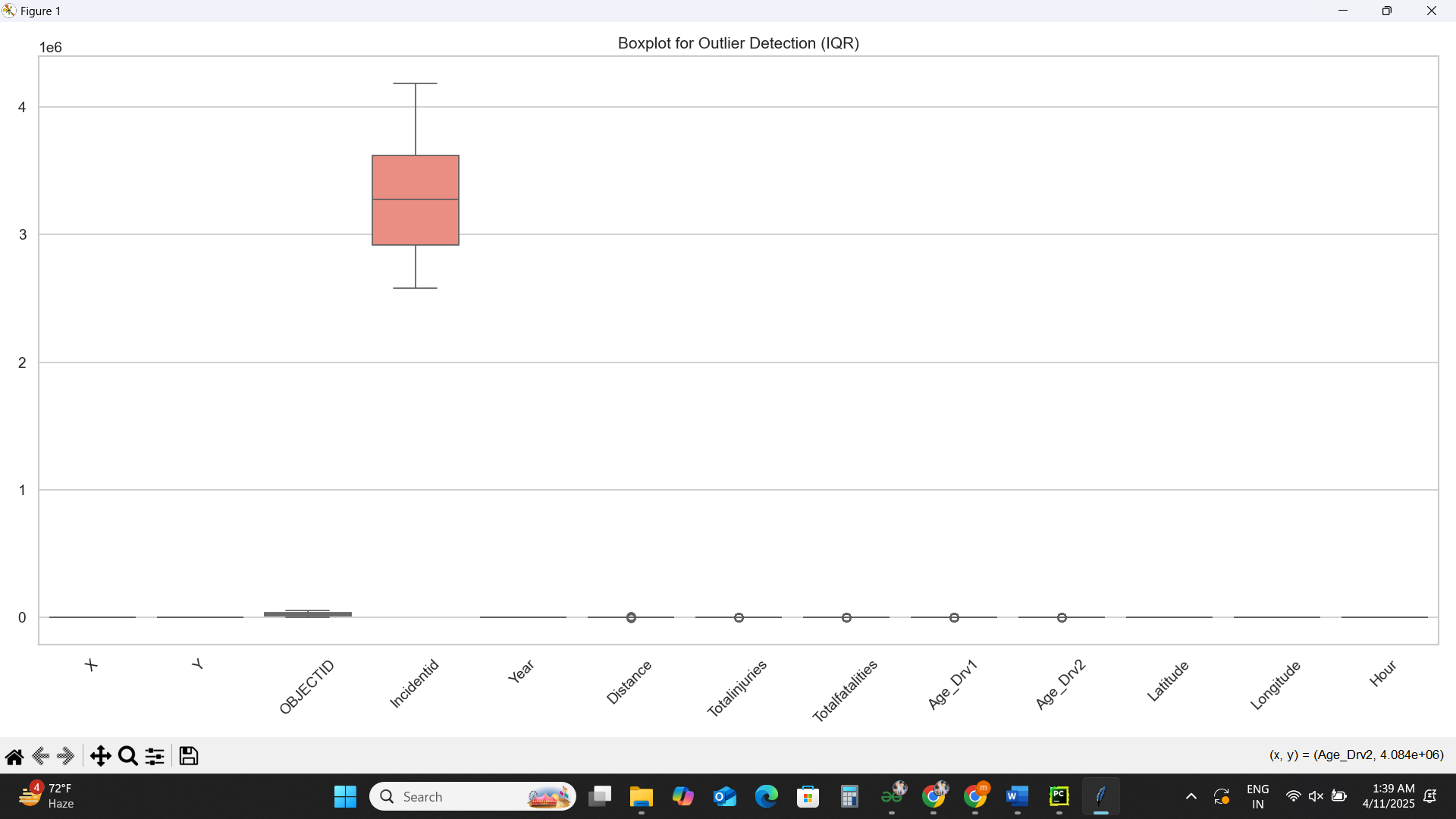
* ✅ **Why Chosen**: To find out **what time of day** and **which day of the week** is riskier.
* 🎯 **Goal**: Optimize **traffic patrols**, especially during dangerous hours.
* 📊 **Outcome**:
  + Crashes peaked around **morning and evening rush hours**.
  + Most accidents occurred on **weekdays**, especially Fridays.





### Outlier Detection

* ✅ **Why Chosen**: To spot any **abnormal values** that could skew analysis.
* 🎯 **Goal**: Clean the data and ensure quality insights.
* 📊 **Outcome**:
  + Detected extreme values in numeric fields using **IQR and Z-score**.
  + Handled outliers for better and accurate analysis.



# 9.Outlier Detection using Z-score Method:

To identify and handle **unusually high or low values** in numerical data that may distort the analysis. Z-score is a statistical method that helps detect such **extreme outliers** by measuring how far a data point is from the mean, in terms of standard deviation.

A total of **3,249 outliers** were detected using the Z-score method across the numerical columns of the dataset.

This indicates that:

* **3249 values** lie **more than 3 standard deviations** away from their column's mean.
* These extreme values could represent **errors**, **rare events**, or **abnormal cases** that might need special treatment in future modeling.

Z-score outlier detection is a powerful method in EDA that identified 3,249 extreme values, ensuring our dataset remains reliable and suitable for further analysis and modeling.

# 8. Conclusion

This crash data analysis project provided valuable insights into the **factors influencing road accidents**, such as environmental conditions, time of occurrence, driver behavior, and violations. Using **exploratory data analysis (EDA)** techniques, we were able to identify critical patterns and trends that contribute to crash frequency and severity.

## Key findings include:

* Crashes peak during **rush hours** and **weekdays**, highlighting the need for traffic control during those times.
* **Clear weather and daylight** conditions saw more crashes, likely due to higher traffic flow, while severe crashes were more frequent under **poor lighting or weather**.
* **Substance use** (alcohol or drugs) significantly correlates with higher injury severity.
* **Rear-end and angle collisions** were the most common, often linked to **violations like speeding** or **failure to yield**.

## Through this analysis, we emphasized the importance of:

* **Targeted enforcement** of traffic rules
* **Public safety awareness** about high-risk hours and behaviors
* Using data to support **policy planning and infrastructure improvements**

Overall, EDA served as a crucial first step in transforming raw crash data into **actionable insights** that can inform **road safety strategies** and **future predictive models**.

# 9. Future Scope

* Use machine learning models to predict crash severity
* Perform geospatial analysis if location data is available
* Correlate with external data like traffic volume or weather reports

# 10. References

1. Crash Data Portal (Government Source)

<https://catalog.data.gov/dataset/1-08-crash-data-report-detail-498c3>

1. Python Libraries: Pandas, NumPy, Seaborn, Matplotlib, SciPy

# 11. GitHub:

<https://github.com/mukundkhandelwal463/Crash-Data-Analysis/tree/main>

# 12. LinkedIn:

<https://www.linkedin.com/posts/mukund-khandelwal-6a8663283_python-datascience-eda-activity-7316203835713040384-iznl?utm_source=share&utm_medium=member_desktop&rcm=ACoAAET5diABs7bbZlDnVTGZ4DnPgeKxnEmHsgA>