**Final Report - Analysing the Relationship Between Weather Events and Traffic Accidents in NYC (2019–2021)**

**Project Overview:**

This project explores the impact of weather conditions especially precipitation and temperature on traffic accidents in New York City. Using real-world datasets from 2019 to 2021, we investigated correlations and built machine learning models to predict daily crash counts based on weather features.

**The primary goals** of this study were:

* to analyse trends and relationships between weather and accidents
* to determine weather features significantly influence crash frequency
* to build and evaluate models to predict crash counts using environmental data

**Data Collection and Preparation:**

We used three key datasets:

* **NYC Vehicle Crashes (2012–2023)** – filtered for 2019–2021
* **NYC Daily Weather Data (2019)** – temperature and precipitation
* **NYC Hourly Weather Data (2020–2021)** – aggregated into daily averages

The final dataset was constructed by merging these sources and engineering features like:

* bad\_weather (precip > 3mm or temp < 0°C)
* temp\_category (Very Cold, Cold, Mild, Warm)
* is\_weekend (binary)

**Data Cleaning & Enrichment**

* Converted dates into a unified format
* Aggregated hourly data to daily averages (2020–2021)
* Enriched dataset with bad\_weather, weekend flags, and temperature categories
* Removed rows with missing or extreme outliers

**Exploratory Data Analysis (EDA):**

We performed multiple visualizations:

* **Scatter plot:** Precipitation vs crash counts
* **Line plot:** Monthly average crashes
* **Heatmap:** Weekday & month crash intensity
* **Boxplot:** Crash distribution across temperature categories

**Hypothesis Testing:**

We conducted a Pearson correlation test:

* **H₀:** There is **no correlation** between precipitation and crash frequency
* **H₁:** There is **a correlation** between precipitation and crash frequency
* **p-value = 0.505** → Since p > 0.05, **we failed to reject the null hypothesis**

**Machine Learning Phase:**

We trained and evaluated three models to predict daily crash counts:

**Models Used**

* **K-Nearest Neighbours (KNN)**
* **Decision Tree Regressor**
* **Random Forest Regressor**

**Evaluation Metrics**

Each model was evaluated using:

* **RMSE** – Root Mean Squared Error
* **MAE** – Mean Absolute Error
* **R² Score** – Goodness of fit

Decision Tree Regressor gives the best performance

**Final Remarks**

This study demonstrated how machine learning methods especially tree-based algorithms can uncover complex relationships in real-world datasets. Findings from this work can inform public transportation safety planning and weather-related traffic interventions in urban areas like NYC.