# Motor Vehicle Collision Analysis: Documentation

**1. Project Overview**

This project analyse motor vehicle collision data to identify patterns, trends, and significant insights. The analysis focuses on time-based trends, factors contributing to crashes, and statistical tests such as ANOVA to explore relationships between variables.

**2. Data Source**

* **Dataset Name**: Motor\_Vehicle\_Collisions\_Crashes.csv
* **Rows**: 1 Million+
* **Columns**: 29
* **Key Features**:
  + **CRASH DATE**: Date of the crash.
  + **NUMBER OF PERSONS INJURED** / **NUMBER OF PERSONS KILLED**.
  + **LATITUDE, LONGITUDE**: Geographical location of crashes.
  + **BOROUGH**: NYC borough where the crash occurred.
  + **CONTRIBUTING FACTOR VEHICLE 1/2/3**: Contributing crash factors.

## ****3. Data Cleaning and Preprocessing****

* **Missing Values**:
  + Columns like LATITUDE, LONGITUDE, and crash-related factors were filled using mode or excluded where appropriate.
* **Outliers**:
  + Latitude and Longitude values were assessed using boxplots.
* **Date Conversion**:
  + Converted CRASH DATE to datetime format.
  + Extracted **Year** and **Month** for time-based analysis.

**Example Code**:

df\_copy['CRASH DATE'] = pd.to\_datetime(df\_copy['CRASH DATE'], errors='coerce')

df\_copy['YEAR'] = df\_copy['CRASH DATE'].dt.year

df\_copy['MONTH'] = df\_copy['CRASH DATE'].dt.month

## ****4. Exploratory Data Analysis (EDA)****

### **Trends Over Time**

* **Objective**: Analyze yearly crash trends.
* **Visualization**:
  + **Histogram** for yearly distribution.
  + **Line Plot** for crash trends over years.

Python

# Plotting histogram for CRASH DATE distribution

plt.figure(figsize=(5, 3))

df\_copy['CRASH DATE'] = pd.to\_datetime(df\_copy['CRASH DATE'], errors='coerce')

df\_copy['YEAR'] = df\_copy['CRASH DATE'].dt.year # Extracting the year for histogram

plt.hist(df\_copy['YEAR'].dropna(), bins=range(int(df\_copy['YEAR'].min()), int(df\_copy['YEAR'].max()) + 1),

color='c', alpha=0.7, edgecolor='black')

plt.title('Histogram of Crashes by Year')

plt.xlabel('Year')

plt.ylabel('Frequency')

plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.show()

### **Heatmap of Crashes by Month and Year**

* **Objective**: Understand monthly and yearly crash intensity.
* **Visualization**: Heatmap

Python

heatmap\_data = df\_copy.groupby(['YEAR', 'MONTH']).size().unstack()

plt.figure(figsize=(12, 8))

sns.heatmap(heatmap\_data, cmap='coolwarm', annot=True, fmt='d', linewidths=0.5)

plt.title('Heatmap of Crashes by Year and Month')

plt.xlabel('Month')

plt.ylabel('Year')

plt.show()

## ****5. Statistical Analysis****

### **Hypothesis Testing: ANOVA**

* **Objective**: Determine if the mean number of injuries significantly varies across boroughs.
* **Test**: One-Way ANOVA.
* **Null Hypothesis (H₀)**: The mean number of injuries is the same across all boroughs.
* **Alternative Hypothesis (H₁)**: At least one borough has a different mean number of injuries.

**Steps**:

1. Remove rows with missing values in the columns BOROUGH and NUMBER OF PERSONS INJURED.
2. Group the data by BOROUGH to create arrays of injury counts for each borough.
3. Perform the ANOVA test using scipy.stats.f\_oneway.
4. Evaluate the F-statistic and p-value to interpret the results.

Python

from scipy.stats import f\_oneway

# Remove rows with missing values in BOROUGH and NUMBER OF PERSONS INJURED

df\_anova = df\_copy.dropna(subset=['BOROUGH', 'NUMBER OF PERSONS INJURED'])

# Group data by Borough

borough\_groups = [group['NUMBER OF PERSONS INJURED'].values

for name, group in df\_anova.groupby('BOROUGH')]

# Perform ANOVA Test

f\_stat, p\_value = f\_oneway(\*borough\_groups)

# Output results

print("ANOVA Results:")

print(f"F-statistic: {f\_stat:.4f}")

print(f"P-value: {p\_value:.4f}")

if p\_value < 0.05:

print("Result: Reject the Null Hypothesis (Significant differences exist).")

else:

print("Result: Fail to Reject the Null Hypothesis (No significant differences).")

### **Insights**:

* **If P-value < 0.05**:  
  Reject H₀ → At least one borough has a significantly different mean number of injuries.
* **If P-value ≥ 0.05**:  
  Fail to reject H₀ → No significant difference in the mean number of injuries across boroughs.

## ****6. Visualizations Summary****

1. **Histogram**: Crash frequency distribution by year.
2. **Line Plot**: Trends in crashes over time.
3. **Heatmap**: Crash intensity by month and year.
4. **Boxplot**: Outlier analysis for LATITUDE and LONGITUDE.

## ****7. Key Insights****

* **Trends**: Significant peaks in crash frequencies during specific years.
* **Time-based Patterns**: Higher crashes in certain months, possibly due to weather or holidays.
* **Statistical Significance**: ANOVA results reveal borough-specific crash frequency variations.

## ****8. Conclusion****

This project provided critical insights into motor vehicle collisions using data visualization and statistical methods. The analysis can inform future urban safety policies and interventions.

## ****9. Future Scope****

* Incorporating weather data for crash correlations.
* Exploring machine learning models to predict crash-prone locations.
* Analyzing time-of-day patterns for targeted safety improvements.