**Data preprocessing - NYC Motor Vehicle Crashes**

**Mission objectives**

Be able to use pandas

Be able to clean a data set

Be able to do prepare a data set for a machine learning model

**Mission**

Information about all the traffic accidents that happened in New York City.

Clean the data set (missing values, etc) and prepare it for a machine learning model.

* **We want to predict which streets are the most dangerous in NYC**

**Must-have features**

The dataset contains no missing values ("" or null)

No duplicates.

Values are consolidated

Data format is correct

No blank spaces (ex: " I love python " => "I love python")

1). To select the right columns and remove the irrelevant columns

We want to predict the most dangerous streets in NYC. Which means the streets where there are the most accidents.

print(original\_dataset.columns.tolist())

""" to remove the columns that are not relevant """

dataset1 = original\_dataset.drop(columns=['off\_street\_name', 'cross\_street\_name', 'borough', 'location'], inplace=True)

2). To remove the missing values, replace ‘ ‘ by Unknown

# drop all rows with any NaN and NaT values df1 = df.dropna() print(df1)

tree\_row.health = tree\_row['health'].fillna('Unknown')

tree\_row.spc\_common = tree\_row['spc\_common'].fillna('Unknown')

tree\_row.guards = tree\_row['guards'].fillna('Unknown')

tree\_row.sidewalk = tree\_row['sidewalk'].fillna('Unknown')

tree\_row.problems = tree\_row['problems'].fillna('Unknown')

2).To figure out if there are any missing value in the dataframe,

and sum up the total for each column

df.isnull().sum()