Cairo university  
Faculty of engineering  
Computer engineering department  
Big Data [**CMP4011**]   
Project proposal

**US Accidents Analysis**

**Team 14**

|  |  |  |
| --- | --- | --- |
| Name | section | BN |
| احمد أسعد درويش محمد درويش | 1 | 1 |
| عمر فريد عبد العاطى لملوم | 2 | 4 |
| محمد نبيل عبد الفتاح فهمى | 2 | 19 |
| ممدوح احمد محمد محمد عطيه | 2 | 26 |

Presented to:

**Eng. Omar Samir**

# *Project Proposal*

# Problem Description

Our project will use a big dataset of US accidents from Kaggle to understand why accidents happen. We'll use different data analysis methods to find patterns in the data, like what makes accidents more likely in certain areas or weather conditions, and predict how severe can it become relative to many factors featured in the dataset.  
The motivation behind this idea is all about making roads safer by giving valuable information to people who can make a difference, like *government officials*, *transportation departments*, and even *everyday drivers*.

# Dataset Link : US Accidents

(49 columns – 7.7 Million records – 3 GB) 🡺 [Dataset link](https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)

# Planned approach

* Wewill use **PySpark** as a framework.
* EDA phase steps: Understanding variables - Data cleansing (e.g.: check nulls – Default values) – dropping unwanted columns – visualization) and more as we explore the data.
* *Gain insights* : for ex: Analyzing the relationship between weather conditions and accident severity – rate of accidents in cities – Map plot of severity of accidents in the US , and so on…
* Predictive analytics include:
  1. [classification - Logistic regression]  
     🡺 Predicting the **severity** of an accident based on the factors involved (e.g. : weather conditions – surrounding POIs – place – time …)
  2. [Regression - Linear Regression]  
     🡺 Predicting **accident Duration** as indicator of impact on *traffic flow.*
* Descriptive analytics include:
  1. [Association Rules] 🡺 find obvious correlation between different circumstances (e.g.: time & place of accidents)
  2. [Clustering - K-Means clustering]  
      🡺 Accident location clusters based on latitude information.
* Algorithms to be implemented using Map reduce: Linear regression.
* Cloud to be used 🡺 GCP.

# Project Pipeline

# Problem analysis

We performed an EDA over the data , with the following goals and objectives:

* ***Understanding Accident Trends***: EDA can reveal trends in accident frequency over time, seasonal variations, and any notable spikes or decreases.
* ***Identifying High-Risk Areas***: Analysis of accident locations can help identify regions with a high frequency of accidents, assisting authorities in implementing targeted safety measures.
* ***Examining Contributing Factors***: By exploring variables such as weather conditions, road conditions, and time of day, insights can be gained into factors contributing to accidents.

## Data preprocessing steps:

* 1. Get null counts for each row
     1. Count null values for each column
     2. Calculate total row count
     3. Subtract non-null counts from total row count to get null counts for each column
     4. Display null value counts for each column
     5. # Drop nulls
  2. Calculate missing value percentages for each column
     1. Filter out columns with non-zero null value percentages

## Data visualization:

* 1. Explore city column:
     1. Show the first few rows of the 'City' column
     2. Count the number of unique cities in the 'City' column
        1. Number of unique cities: 11729
     3. Group the data by 'City' and count the number of occurrences
     4. Sort the result in descending order based on the accident counts
     5. Get the city with the highest number of accidents

**City with the highest number of accidents: Miami**

**Number of accidents in Miami: 158736**

* + 1. Plot the first 10 cities with the highest number of accidents

صورة تحتوي على نص, لقطة شاشة, الخط, رقم

تم إنشاء الوصف تلقائياً

* + 1. Plot a histogram of number of accidents :

صورة تحتوي على نص, لقطة شاشة, تخطيط, رسم بياني

تم إنشاء الوصف تلقائياً

From the above graph we can analyze that distribution of accident is more between ten to hundred. When it goes further its decreasing exponentially.

* + 1. A
  1. Severity column analysis & preprocessing:
     1. Convert the Severity to integer
     2. Group data by severity
     3. Sort the result in descending order based on the severity counts
     4. Get the Severity with the highest number of accidents

**Severity with the highest accidents count: 2**

صورة تحتوي على نص, لقطة شاشة, مستطيل, ميدان/ مربع

تم إنشاء الوصف تلقائياً**Where the Number of accidents of severity degree: 4756722**

* + 1. Plot the severity class frequency:

From the graph we conclude that the class of severity = 2 is the most frequent class.

* + 1. Calculate average severity per city :
       1. Result:

صورة تحتوي على نص, لقطة شاشة, قائمة طعام, الخط

تم إنشاء الوصف تلقائياً

* + - 1. Histogram plot:

صورة تحتوي على نص, لقطة شاشة, تخطيط, رسم بياني

تم إنشاء الوصف تلقائياً

From the above graph, **most of the cities have average severity = 2**

* 1. Start time analysis
     1. Convert "Start\_Time" and "End\_Time" columns to datetime format
     2. Hour:
        1. Extract the hour from the "Start\_Time" column
        2. Show the distribution of number of accidents per hours

صورة تحتوي على لقطة شاشة, رسم بياني, تخطيط, نص

تم إنشاء الوصف تلقائياً

* + 1. Days:
       1. Extract the day from the "Start\_Time" column
       2. Show the distribution of number of accidents per days of week

صورة تحتوي على نص, لقطة شاشة, رسم بياني, تخطيط

تم إنشاء الوصف تلقائياً

* + 1. Months:
       1. Extract the month from the "Start\_Time" column
       2. Plot histogram of the 'Day\_Of\_Week' column

صورة تحتوي على نص, لقطة شاشة, رسم بياني, تخطيط

تم إنشاء الوصف تلقائياً

Initially, there is a high number of accidents, which gradually decreases as the months progress. However, towards the end of the year, there is a slight increase in the number of accidents again.

* + 1. Years:
       1. Repeat as above we get:

صورة تحتوي على نص, لقطة شاشة, تخطيط, خط

تم إنشاء الوصف تلقائياً

From the graph we can analyze that accidents are increasing every year but we can see 2023 have the lowest accident recorded. Thats really unusual. We need to see the data of 2023.

صورة تحتوي على نص, لقطة شاشة, رسم بياني, تخطيط

تم إنشاء الوصف تلقائياً

The reason for the seemingly lower number of accidents in 2023 is primarily due to data availability. The dataset used for analysis only includes data up to March 2023, which means it doesn't cover the entire year. As a result, the observed decrease in accidents for 2023 may not accurately reflect the true accident rate for the entire year, as data for the later months is missing. Therefore, any conclusions drawn about accident trends in 2023 should be approached with caution, keeping in mind the incomplete data for that year.

* 1. Start Latitude and Start Langitude
     1. We plot a scatter plot to distribute the locations of accidents geographically:

صورة تحتوي على نص, خريطة

تم إنشاء الوصف تلقائياً

* 1. Analyzing **weather** columns:
     1. ### 1.Are there more accidents in warmer or colder areas?
        1. Define temperature bins and labels
        2. Use 'when' function to assign temperature categories based on the bins
        3. Add the temperature category column to the DataFrame
        4. Group by temperature category and count the number of accidents
        5. Results

+--------------------+---------+

|Temperature\_Category|Accidents|

+--------------------+---------+

| Cold| 1800430|

| Warm| 1393745|

| Moderate| 2462664|

+--------------------+---------+

* + - 1. Convert the result to Pandas DataFrame for plotting

صورة تحتوي على رسم بياني, لقطة شاشة, الرسومات, دائرة

تم إنشاء الوصف تلقائياً

* + 1. Frequency of Weather Conditions
       1. Group by weather condition and count the frequency of each condition
       2. Get top 10 most frequent weather conditions:

صورة تحتوي على نص, لقطة شاشة, رسم بياني, تخطيط

تم إنشاء الوصف تلقائياً

* + - 1. Impact of Weather on Accident Severity
         1. Group data by weather condition and calculate mean severity
         2. Plot the values:

صورة تحتوي على نص, لقطة شاشة, التصميم

تم إنشاء الوصف تلقائياً

Light blowing snow is a weather condition characterized by light snowfall accompanied by strong winds that cause the snow to be lifted from the ground and blown around. This weather condition can significantly reduce visibility and create hazardous driving conditions. Overall, it poses significant hazards to drivers and can contribute to the severity of accidents by reducing visibility, creating slippery road conditions, and increasing the likelihood of collisions.

## Extracting insights from data

📄🌟Here is a summary & conclusion of the EDA phase insights we got:

* 1. The dataset only includes data for 49 states.
  2. Miami has the highest number of reported accidents.
  3. The frequency of accidents per city follows an exponential decrease pattern.
  4. Only 13.5% of cities have more accidents than the average.
  5. Accidents are most common between 8 am to 10 am and 3 pm to 6 pm, suggesting a higher likelihood during peak commuting hours.
  6. Weekdays show a higher number of accidents compared to weekends.
  7. The year 2023 has the lowest number of reported accidents, likely due to data availability only until March 2023.
  8. Coastal areas experience higher accident rates compared to inland regions.
  9. The top 10 most frequent weather conditions for accidents are: Fair, Mostly Cloudy, Cloudy, Clear, Partly Cloudy, Overcast, Light Rain, Scattered Clouds, Light Snow, and Fog.
  10. Analyzing the relationship between weather conditions and accident severity revealed the following insights:
      + Light Blowing Snow has the highest average severity at 3.67.
      + Patches of Fog / Windy, Heavy Freezing Rain / Windy, and Light Fog have average severities of 3.14, 3.00, and 3.00, respectively.
      + Partial Fog / Windy, Heavy Thunderstorms and Snow, Light Thunderstorms and Snow, Heavy Ice Pellets, Heavy Blowing Snow, and Drifting Snow / Windy also show significant average severity levels.

## Model/classifier training

# Results & Evaluation

# Trials [including unsuccessful ones]