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

Idea

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

Planned approach

- We will 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.