Project Proposal

Project Title: Analysis of Traffic Accidents in the US
Group ID: 1
Members: Shervan Shahparnia and Nathan Cohn

Project Summary:
This project will identify key risk factors for traffic accidents in the United States and develop a predictive model for accident likelihood. We will use the Kaggle "US Accidents" dataset, which contains over 7.7 million records from 2016 to 2021 and includes details such as accident timestamps, weather conditions, road types, and geographic coordinates. Additionally, we will use complementary road network GIS data. Through exploratory data analysis (EDA) and machine learning techniques, we will assess how factors like time, weather, infrastructure, and nearby points of interest influence accident rates. Our findings are intended to support targeted urban planning and traffic safety strategies and may have potential applications in autonomous vehicle technology.

We will conduct exploratory data analysis to investigate the following themes:

- **Time and Weather Patterns:** We will look at how the number of accidents changes with different times of the day, days of the week, and seasons. We will also study how weather conditions like temperature, rain, and visibility affect both the number of accidents and how severe they are.
- **Road Features and Urban vs. Rural Areas:** We will analyze how different road features, such as road types, speed limits, and the number of lanes, along with intersection details like traffic signals and roundabouts, are related to accident outcomes. We will compare these factors in urban and rural areas to see how they differ.
- **Other Factors and Clustering:** We will examine how nearby points of interest, such as speed bumps, give-way signs, and railway crossings, might influence the chance of an accident. We will also explore how long accidents last in relation to environmental factors and try to find any geographic clusters or hotspots where accidents happen more often.

In parallel, we will develop at least one predictive model to estimate accident likelihood based on these key factors.

Broader Impacts:
Our research has the potential to enhance road safety by identifying high-risk factors and accident-prone locations. These insights can inform traffic authorities on implementing more effective safety measures, such as optimizing traffic signals,

adjusting speed limits, and redesigning hazardous road segments. By integrating road network data, we will provide a more comprehensive understanding of accident risks related to infrastructure.

Moreover, our findings could support the advancement of autonomous driving technologies. Self-driving vehicles rely on predictive models to navigate safely, and our analysis of accident risk factors can help improve their decision-making algorithms. By identifying patterns in high-risk areas, autonomous vehicle developers can refine their systems to avoid potential hazards, making self-driving technology safer and more reliable.

Overall, our project contributes to the future of transportation by improving traffic safety and supporting the integration of autonomous vehicles into urban environments.

Data Sources:

US Accidents Dataset (Kaggle): This dataset contains over 7.7 million traffic accident records with more than 40 columns of information. It was created by PhD students at Ohio State University using data from various traffic APIs and accident data sources.

The dataset that we will be using:
https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?resource=download

The paper of the researchers who created the dataset and their sources/references:
https://arxiv.org/abs/1906.05409

Road Network GIS Data: Geospatial data from state or local transportation departments will be used to analyze road layouts and traffic density. This data includes details on road structure such as road types, speed limits, the number of lanes, and possibly traffic signal locations. Incorporating this information will allow us to better pinpoint accident hotspots and understand how road design affects traffic accidents.

- We may employ the usage of various API for speed limits such as the roads API provided by Google. Additionally, the API has features that we may use to determine the number of lanes for larger highways and roads, especially those in metropolitan areas.

Expected Major Findings:

We anticipate the following key findings from our analysis:

- **Temporal Patterns:** Accident frequency is expected to peak during rush hours and weekends, reflecting increased traffic volume and human factors such as fatigue or distracted driving.
- **Weather Influence:** Severe weather conditions, such as heavy rain, fog, and

snow, will likely correlate with higher accident rates and severity due to reduced visibility and road traction.

- **Road Network Characteristics:** Certain road features, such as intersections lacking traffic signals, high-speed highways, or poorly maintained roads, may contribute to higher accident rates.
- **Accident Hotspots:** Geospatial analysis will help identify clusters of accident-prone areas, allowing us to investigate their underlying causes.
- **Predictive Modeling:** We will develop a machine learning model (e.g., random forest classifier) to estimate accident risk based on time of day, weather conditions, and road attributes. This model could serve as a tool for urban planners and traffic management agencies to predict and mitigate accident risks.