CS 163 Proposal

**Project Proposal**

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

**Project Summary:**We will be focusing on traffic accidents across the United States to find the patterns and risk factors that lead to more accidents. We will use the "US Accidents" dataset from Kaggle, which includes over 7.7 million accident records. This dataset was created by PhD students at Ohio State University in 2019, with references on how they created the dataset using various traffic APIs and other accident data sources. It includes many details such as unique accident IDs, severity ratings, timestamps, GPS coordinates, and other descriptive information regarding each incident. Additionally, it includes weather data such as temperature, wind chill, humidity, pressure, visibility, wind speed, and precipitation. It also includes information about traffic signals, crossings, and other points of interest (POI). The dataset also provides details on daylight conditions through sunrise/sunset times and various twilight periods. We plan to combine this dataset with road network GIS data (which covers the structure of roads, road types, speed limits, number of lanes, and possibly even traffic signal locations) from local transportation departments.

Although we want to develop at least one predictive model, we first thought of some exploratory data analysis questions (EDA). We will investigate:

* How accident frequencies vary by time of day, day of the week, and season.
* The relationship between weather conditions (like temperature, precipitation, and visibility) and both the frequency and severity of traffic accidents.
* How different road network characteristics, such as road types, speed limits, and lane counts, correlate with accident outcomes.
* Whether intersections with features such as traffic signals, crossings, or roundabouts experience fewer or less severe accidents compared to those without these features.
* The differences in accident patterns between urban and rural areas, and how road network density and connectivity influence accident rates.
* The relationship between accident duration (the time from start to end) and factors such as weather conditions or time of day.
* The effect of nearby POIs, like speed bumps, give-way signs, and railway crossings, on the likelihood of an accident.
* Differences in accident frequency and severity between weekdays and weekends.
* How various daylight conditions (daytime, civil twilight, nautical twilight, and astronomical twilight) impact accident occurrences.
* The identification of geographical clusters or hotspots and the underlying factors contributing to these patterns.

**Broader Impacts:**Our project aims to deliver insights that will help improve road safety and support better urban planning. By analyzing the expansive dataset with over 7.7 million traffic accident records along with detailed weather, daylight, and road network data, we expect to identify key risk factors and patterns that contribute to accidents. These findings can help local authorities and traffic management agencies implement targeted safety measures, such as adjusting traffic signal timings, setting appropriate speed limits, or redesigning dangerous road segments. This could lead to fewer accidents and save lives.

In addition, our work could benefit the autonomous driving industry. As self-driving vehicles become more common, a detailed understanding of accident risk factors and high-risk areas can help companies improve their navigation systems and collision avoidance algorithms. This data can guide autonomous driving firms in refining their technology and integrating their vehicles safely into our existing transportation network. Overall, our project not only aims to enhance current road safety but also provides important insights that could shape the future of transportation and the development of self-driving vehicles in our society.

**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 Dat**a: 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 expect to see clear patterns in when accidents occur. For example, accidents might happen more often during rush hours or on weekends, which can help us understand the times when accidents are most likely to occur.
* Our study will probably show that certain weather conditions like heavy rain, fog, or extreme temperatures are closely connected to both the number of accidents and how serious they are.
* By combining road network data with information about POIs, we expect to find specific road features that raise the risk of accidents. For example, intersections without proper traffic signals or areas with many roads close together might turn out to be accident hotspots.
* We also plan to identify geographic clusters where accidents tend to happen, which will give us clues about the causes of these hotspots.
* We will build a predictive model using machine learning techniques. One idea is to use a random forest classifier that estimates the likelihood of accidents based on factors such as time of day, weather conditions, and road characteristics. This model could be a useful tool for traffic management and urban planning.