DSCI 5260: Project Selection

Prof: Javier Rubio-Herrero

Team 10:

Sankeerth Kumar Poojala	-	11556385
Govardhan Reddy Boyella	-	11551545
Nimisha Yerra	-	11521722
Baby Sriteja Narla	-	11550645

Project name: **Data Analysis on road accidents**

Dataset:

The National Highway Traffic Safety Administration (NHTSA) provides a public API called "CrashViewer" that us to access information about motor vehicle crashes in the United States. The data made available through the API includes information such as the date and location of the crash, the type of vehicles involved, the number of people killed or injured, and the contributing factors to the crash. This API can be used to build tools and applications that provide insights into traffic safety, support data-driven decision making, and improve road safety for all users.

Data Source: https://crashviewer.nhtsa.dot.gov/CrashAPI

The "CrashViewer" API provides access to several attributes that describe each crash, including:

Date and Time: The date and time of the crash, including the year, month, day, and time of day.

Location: The location of the crash, including the state, county, city, and the latitude and longitude of the crash site, urban/rural.

Vehicle Information: Information about the vehicles involved in the crash, including the type of vehicle (e.g., passenger car, truck, motorcycle), the make and model, and the age of the vehicle.

People Information: Information about the people involved in the crash, including the number of fatalities, the number of injured people, and the age and gender of each person.

Contributing Factors: Information about the factors that contributed to the crash, including the type of crash, the environmental conditions (e.g., rain, snow), and any human factors (e.g., driver error, distraction).

Preprocessing and exploratory data analysis is done using pandas and scikit-learn to clean the data and structure it for the ease of use.

Question 1

Perform various correlations in terms of EDA and visualizations:

- Highest causes for injuries/fatalities in every region
- Over the years how the causes are trending
- Highway types to number of injuries/fatalities
- Urban vs rural setting
- Vehicle type vs crash severity?
- Age and gender vs crash involvement?

Current Plan:

- ➢ We shall be using various graphs in seaborn for example, Maps for geographical representation of crash data across the nation, scatter plot (e.g.: for plotting a relationship between age and crash severity), heatmaps (for density of the crashes) and more.
- ➢ Holoviews would be used to generate more interactive visualizations that allows us to interact and change the representation of the data as we see fit.
- ➢ Using scikit-learn, prediction models like linear regression, Support Vector Machines and Random Forest algorithms can be designed.

Question 2

How effective (both positive and negative) are the measures/policies put forth by federal or local governing bodies?

Could essentially consider a specific region/time and check the progress.

Current Plan:

- ➢ Trying to estimate the effect of the measures/policies by federal or local governing bodies would be done with the help of collecting the data about the policies introduced for the transportation safety or crash preventions.
- ➢ Comparing that data with the crash data over the years would give a very good idea about the type of impact introduction of that policy had.
- ➢ Measures/Policies being the hypothesis.

Question 3

Who is at fault? Vehicle drivers' negligence (includes drunk driving, not following proper safety precautions) or other reasons such as climate.

Current Plan:

- ➢ Using various attributes that can be considered as drivers' negligence would be compared alongside with the attributes that represent external factors like climate, road type and more.
- ➢ Designing interactive visualizations on these attributes would be helpful to determine any hidden patterns or reasoning.