Sandhya Srinivasa (48018117)
Emily Arcidiacono (47608461)
Chase Vaught (47950218)
Daniel Laureano (47485260)
EMIS 3309

# Project Proposal: Road Accidents

For the project we plan to analyze a dataset that looks at car accidents in the UK from 1979-2015. The dataset was found on Kaggle and the link to the dataset is https://www.kaggle.com/akshay4/road-accidents-incidence. However, the data on Kaggle was fetched from Open Data Platform UK and is being shared under Open Government License. The dataset is made up of data from road accidents that occurred between 1979 to 2015 in Great Britain. Additionally, the dimension of the set is made up of 70 columns and 285331 rows. The features range from vehicle type, the gender of the driver, locations of the accident, weather, and police force on the scene. As a collective, we have decided to trim some of the columns and focus on certain features. Such as the type of transportation, which ranges from motorcycles, taxis, vans, and cars. Other columns we have decided to keep are the number of casualties, age of the driver, casualty type, and casualty severity. Our current response column will be accident severity. However, as the group moves forward and begins to clean the data and see results. The response might change because the group found an interesting answer from a different feature.

The audience for our dataset is individuals who are choosing what type of vehicle to buy and are considering the safety of the vehicle. From the dataset we are trying to understand how the type of car and other features involved in the accident affects the severity of the accident. By letting our response column be the severity of the accident, we will be able to predict the severity of the accident depending on the age of the car, car type, and other factors. Some of the more specific questions we will be answering are listed below.
How does the age of the car affect the severity of the accident?
Is there a relationship between the sex of the driver and the severity of the accident?
How does weather conditions affect accident severity?
Does point of impact affect accident severity?
How does right hand/left hand driving affect accident severity?

We are planning to apply the classification method, LDA, for our statistical learning method. LDA will be used because our dataset's response feature is categorical and also has multiple classes. Another classification method we will apply will be multinomial logistic regression. This learning method also permits the response feature to be categorical and non-binary. The dataset's response feature will be accident severity which is a categorical value because there are 3 possible outcomes for accident severity: 1, 2, 3, which correspond to Fatal, Serious, and Slight. We are going to predict these values based on the type of car and the number of previous accidents the car has had.

One of the potential difficulties in this data set is missing values. Whether the data is missing, out of range, or unclassified, there are some gaps in the data. In many of the columns,

data that is out of range or is missing is replaced with a value of -1. Missing data is a point of error that cannot be fixed. This issue could possibly after the results of our study. Another possible difficulty with this data set is the lack of background information. We know that this data is a cumulation of data from the UK between 1979-2015, but we do not know how this data was collected. Therefore, we do not know the accuracy of the data. The UK has different laws and regulations when it comes to cars and motorcycles, so we must take that into consideration.