# SI618 Data Manipulation and Analysis
## Project 1 Proposal

# 1 Motivation

With the mass production and popularization of automobiles, road accidents have become a big issue in the past decades.

The details of road accidents have long been recorded and stored by the police in many places. Nowadays with the help of data science, we can study the accumulated data, and try to answer question like:

- Which category of road leads to most accidents?
- Which type of collision implies the most severity?
- What atmospheric condition is most likely to cause road accidents?
- ...

# 2 Datasets Description

## 2.1 Accidents in France From 2005 to 2016

The dataset from [1] describes road accidents happening in France from 2005 to 2016. Three of the *csv* files provided, *caracteristics.csv*, *places.csv*, and *users.csv*, will be involved in this project. They respectively provide the basic information, the road conditions, and the victim information of the accidents.

## 2.2 Accidents in the UK From 2005 to 2015

The dataset from [2] describes road accidents happening in the UK from 2005 to 2015. This dataset provide information such as the road types, the casualties, the vehicles, etc.

# 3 Procedures

1. Join *csv* files of the dataset about accidents in France, and try to find out:
   - top accident place,
   - top accident time,
   - road types with most accidents recorded in different atmosphere conditions,
   - ...

   in France from 2005 to 2016.
2. Join *csv* files of the dataset about accidents in the UK, and try to find out:

- top accident place,
- the day of week corresponding to the most fatal accidents,
- the vehicle type corresponding to the most fatal accidents,
- ...

in the UK from 2005 to 2015.

3. Join the datasets about accidents in France and about the accidents in the UK, and try to:
   - find out the vehicle type corresponding to the highest rate of fatal accidents;
   - compare the rate of fatal accidents in the UK and in France;
   - compare the number of accidents in every year in the two countries from 2005 to 2015;
   - ...

# 4 Large-scale Computation Tasks

- With the help of *mrjob*, using *map* and *reduce* to count the number of fatal accidents for vehicle type, and calculate the corresponding fatal rate. We can analyze the safety of each vehicle type based on the results.
- With the help of *Spark*, calculate the fatal number and fatal rate for each road type in each atmospheric conditions in France. We can analyze the influence of weather and road environment on traffic safety.
- With the help of *Spark*, calculate and compare the fatal accident number and fatal rate in each year from 2005 to 2015 in the two countries. We can analyze the result to try to find a trend of the change of fatal rate of accidents.
- ...

# 5 Visualization

- A bar chart showing the fatal rate corresponding to the vehicle types.
- A scatter plot showing the trend of change of fatal rate of accidents in the two countries over time from 2005 to 2015.

# References

[1] "Accidents in France from 2005 to 2016", *Kaggle*. `https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016`

[2] "UK Car Accidents 2005-2015", *Kaggle*. `https://www.kaggle.com/silicon99/dft-accident-data`