Project - NYC Crashes

PCS - UConn, July 2022

PUT YOUR NAME HERE

The NYC motor vehicle collisions crash table contains details on the crashes. The dataset we are using contains the crashes in January, 2022. The real-time full data with documentation is available from NYC Open Data. Here we import the data and create an `hour` variable for the crash time.

```
## This is a subset from January 2022
crash <- read.csv("https://github.com/statds/ids-s22/raw/main/notes/data/nyc_mv_collisions_202201.csv")

## lower case variable names
names(crash) <- tolower(names(crash))

## frequency table by borough
xtabs(~ borough, data = crash)

## Create an hour variable with integer values from 0 to 23
crash$hour <- as.numeric(sub(":[0-9][0-9]", "", crash$crash.time))

table(crash$hour)
hist(crash$hour, breaks=24)
```

Your goal in this project to explore the data and report your findings. Example questions to consider are:

- Check if the number of persons killed is the summation of the number of pedestrians killed, cyclist killed, and motorists killed. From now on, use the number of persons killed as the sum of the pedestrians, cyclists, and motorists killed.
- Construct a cross table for the number of persons killed by the contributing factors of vehicle one. Collapse the contributing factors with a count of less than 100 to "other". Is there any association between the contributing factors and the number of persons killed?
- Create a new variable death which is one if the number of persons killed is 1 or more; and zero otherwise. Construct a cross table for death versus borough. Test the null hypothesis that the two variables are not associated.
- Fit a logistic model with death as the outcome variable and covariates that are available in the data or can be engineered from the data. Example covariates are crash hour, borough, number of vehicles involved, etc. Interprete your results.
- In preparation for the class event at the NYC Open Data Week on March 8, suggest a meaningful question that can be answered by the data (but no need answer it this time; that will be for next week). You may think about comparison with data from other periods, in which case, you will need to download the right data.

Be creative and don't be limited by the examples.