In my project I would use the data set for the traffic violation stops in Montgomery County Police, Maryland (available at https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q). It contains the location, date and the time of the stop, description of the traffic violation, information about the driver (race, gender, DL state), the vehicle (type, make and model), the officer (the district, uniformed or not), as well as the outcome of the stop (citation or warning).

I chose this data set because it can answer many interesting questions. The most important and fascinating is whether the outcome of a traffic stop is solely determined by the violation type, or other factors are also important. For example, is it more likely to get a citation if the driver is out-of-state, or nonwhite, or male? Are the time of day and the day of the week important? Are there police districts that are significantly stricter than the others? Are drivers of more expensive vehicles more or less likely to receive a citation? These are important questions, and the answers obviously can be useful for drivers, but also for the police departments, because it would give them some valuable information about the police officers’ decision-making process.

The first phase of the project would be exploratory: I will look at the data and try to discover as many as possible useful insights. I have already started doing this, and in Fig. 1 I show the distribution of the two possible outcomes (citations or warning) for five different violations. It is clear that for some violations warning is much more likely (for example, speeding in 30 mph zone), whereas in others citation is virtually guaranteed. But there are violations for which the outcome is less certain, and these are the violation that are most interesting to model.

In Fig. 2 I have plotted the proportion of citations for the nine most likely traffic violation stops, grouped by the gender of the driver. Small but persistent effect can be seen: the female drivers have lower probability to receive a citation than the male drivers for the same violation. At the same time I have not found significant effect on the outcome by the state of the driver: in-state and out-of-state have about the same chance of being warned.

The second phase of the project would be modeling. The ideal is be to build a statistical model that can predict the outcome (citation or warning) based on the variables that I identified as important at the first stage. The baseline model I am going to use is with a single predictor – the type of violation, and the goal is to create a model that can predict significantly better that the baseline model. From the exploration I have already done it seems that the gender of the driver could be a significant predictor of the outcome, whereas the state of origin of the driver is not. I plan to try logistic regression, as well as classification tree models.

The third phase would be to constantly update and refine the model by using the dataset API.