# Final Project Proposal

By Wendi Wang

## Individual Final Project Dataset

- **Subject Area / Field of Interest**

  I would like to investigate the potential effects of geographical location on the number and types of traffic violations via data visualization in this final project. Specifically, I would like to verify if some categories of traffic violations are dominant in areas with certain geographical locations. Finally, I hope with the results of the relationship between geographical locations and certain types of traffic violations, people can avoid similar traffic violations in the future.

- **Source of Data & Specific dataset(s)**

  I decide to focus on traffic violations in Montgomery County in the United States. This dataset is publicly released with the link.

- **Dataset Description**

  There are 1015271 rows and 35 columns in this dataset, with 3 numeric variables, 29 categorical variables, and 3 date/time variables. Since the variables are mostly categorical, considering there are 1015271 data entries to be chosen, I plan to use a sample size of around 50000-100000 (5%-10%) of the dataset size. If possible, I would also like to increase the sample size from small to large gradually from approximately 1%-50%, thus drawing a reliable and representative conclusion based on sampled data visualization results. All the information will be stored in a single .csv file called Traffic_Violations.csv.

- **Technology planned to Use**

  I plan to use R as the media to obtain data visualization results between the number/types of traffic violations and geographical locations. As an auxiliary, I will also use Python to derive quantitative approximations of such relationships via interpolations or regression if possible.