# Bike Accidents in Great Britain-Data Mining Project

Nicolas Mavromatis

CSPB 4502

# Description

- Dataset: Bike accidents in Great Britain, 1970-2018
- Includes date/time, speed limit, road/weather conditions, gender, age, etc.
- Over 800,000 data points
- Initial Investigation:
  - Gender/age correlation
  - Road conditions/speed limit/type correlation
  - Year/time correlation
  - Light correlation
  - Weather correlation
- Questions:
  - Is there a specific age range or gender most associated with biking accidents?
  - Do certain road types or speed limit zones correlate with high numbers of accidents?
  - Are certain years or time of day most associated with accidents? Could this be explained by brightness?
  - Do certain seasons, and therefore weather  and road conditions, associate with accidents?
- Goal:
  - After investigation, predict conditions of high accidents in order to avoid them in the future.

# Prior Work

- Study in Denmark, ~350 crashes
- Findings:
  - Winter road maintenance crucial
  - Warm weather: Individual factors affect, no gender difference
  - Elderly (>65 years) lower  number, but higher chance for injury when accident
  - 18% sustain injuries from bike crash
  - Average cost of €1,701 for treatment
  - 82% of crash victims used helmet
  - 79% of crashes in urban area
  - Half in daylight, half at dark or dusk/dawn
  - 24% of crashes on dry surface
  - Half of crashes in Winter
  - Most during rush hour
- Limitation: Small sample size, self reporting, few large-scale studies
- Source: https://www.sciencedirect.com/science/article/pii/S0001457521003845:

# Proposed Work

- Integration: Combine two excel sheets, with different attributes, linked by ID
- Data Cleaning:
  - Convert dates to workable pandas type, then add a season column based on calculation
  - Account for unknown weather and road type values. Number may not be significant, so no work may be necessary. These values may be ignored when considering specific attributes.
- Feature extraction: bin data by attribute, such as time, speed limit, road condition, age, etc. to find interesting patterns. Time will need to be processed to have more approximate bins.
- Analysis: Create a number of scatter and bar plots to find interesting patterns. Maybe other analyses would be useful (TBD as course progresses).
- Summarize: Averages, stddev, quartiles, chi-squared, frequent patterns
- Characterize patterns associated with accidents for future prediction
- Ultimately, goal is to find interesting patterns and answer questions. Do results agree with the previous study, and predictions?

# Tools

- Python-easy to use w/ many integrated libraries
- Pandas-merge and clean data, integration with .csv files, loading/saving data
- Numpy-good for statistical analysis, like average, stdev, chi-squared
- Scipy-builds on numpy, allows for more complex statistical algorithms
- Matplotlib-detailed graphing. Great for investigations and results
- Github-version control
- Others-TBD as work proceeds.

# Evaluation

- Stats: Chi-squared, mean, median, support, confidence, t-test  etc.
  - More knowledge is necessary to understand which statistical methods are most appropriate for data
- Graphs: Evaluate patterns, answer questions. Great for initial investigation
- Comparison of findings in Denmark study
- Insight into safety/prevention
- Characterization and prediction of conditions that lead to accidents
- Suggestion for future studies, ways to reduce accidents
- Discussion of limitations-no helmet data, data from one country