Analysis of US Accidents

Team Members

- Matthew Scott
- Paul Schneider
- Daniel Simonson
- Kevin Qian

Description

- We intend to explore and answer questions about how weather, time of day, precipitation, geolocation, and other external factors lead to an increase of accidents in the US. Our particular interest is finding a correlation between why an accident might be more likely to occur near a certain location vs another, during extreme weather instances vs not, or during daytime rush hour vs nighttime low visibility

Prior Work

- We have found the dataset we plan to use for the project and investigated the data.
- Discussed options for analysis
- Discussed potential questions to ask and explore with regards to our data
  - Do we see a higher rate of accidents with higher precipitation?
  - Which locations have the highest amount of reported accidents within X-miles of each other? Are there any discernible external factors contributing to this increase?
  - What time of day are accidents most likely to occur?

Proposed Work

- Data cleaning:
  - Review and scrub data for values that are null, negative, or otherwise unnecessary data.
- Questions:
  - Create and focus on 2-4 questions to guide our analysis.
- Data integration:
  - Avoid redundant data, columns, etc. through correlation analysis.
- Data Reduction:
  - If specific data is not integral to answering our questions then remove

Dataset

- List of datasets to use: US Accidents (2016 - 2023)
- Where found: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data
  - Data Sourced from local traffic APIs
- Uploading primary dataset to shared Google Drive.
- Country wide car accident data set covering 49 states from Feb 2016 to Mar 2023
- Data was collected through US and State Departments of Transportation, Local Law enforcement and traffic cameras/sensors.
- Roughly 7.7 million accidents were recorded during this window.

List of Tools to Use

- Python
  - Pandas
  - Numpy
  - MatplotLib
- Excel
- Github
- Zoom
- Text Message
- Google Drive
- Tableau or PowerBI

Evaluation

- Present a collection of visualizations that show our correlation analysis.
- Draw conclusions from our results to see if they align with our previous predictions.
- Apply the Apriori Algorithm to conduct a pattern analysis that finds the most similar attribute values present in an accident.
- Conduct a Bayesian Classification to determine the probability of the most similar attribute values found by the Apriori Algorithm.

Citations

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.