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

Questions sought to be answered

- ➔ What driving conditions are associated with the majority of accidents?
- ➔ What driving conditions cause the most severe accidents?
- ➔ How does traffic accident volume changed with time?
- ➔ What locations have the most traffic accidents?

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.

Data

- Main Attribute Features
  - ID
  - Start_Time/End_Time
  - Start_Lat/End_Lat
  - Wind_Speed/Wind_Chill
  - Humidity(%)
  - Pressure(mi)
  - Sunrise_Sunset
  - Visibility(mi)
  - Distance(mi)
  - Temperature(F)
  - Precipitation(in)

Data Preparation Work

- Data cleaning:
  - Reviewed and scrubbed data for values that are null, negative, or otherwise unnecessary data.
  - Converted several attributes from quantitative data to categorical
    - Wind_Speed(mph), Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Precipitation(in), Sunrise_Sunset)
- 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

List of Tools to Use

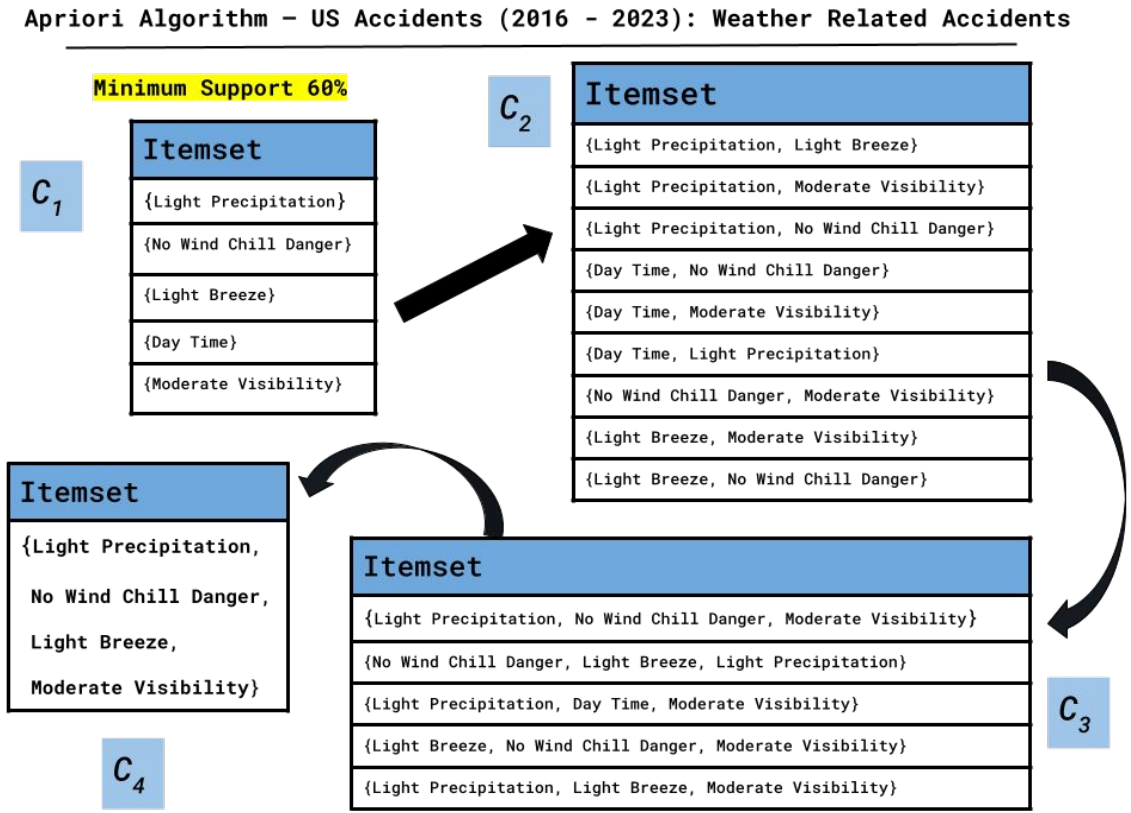
- Python
  - Pandas
  - Numpy
  - MatplotLib
- Excel
- Github
- Zoom
- Text Message
- Google Drive
- Tableau or PowerBI

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.

Frequency Analysis

- 60% of accidents recorded were in periods of light precipitation, no wind chill, light breeze, and moderate visibility.
- A majority of accidents were not heavily influenced by hazardous driving conditions.
- Hazardous driving conditions may actually increase driver awareness and decrease accident frequency.



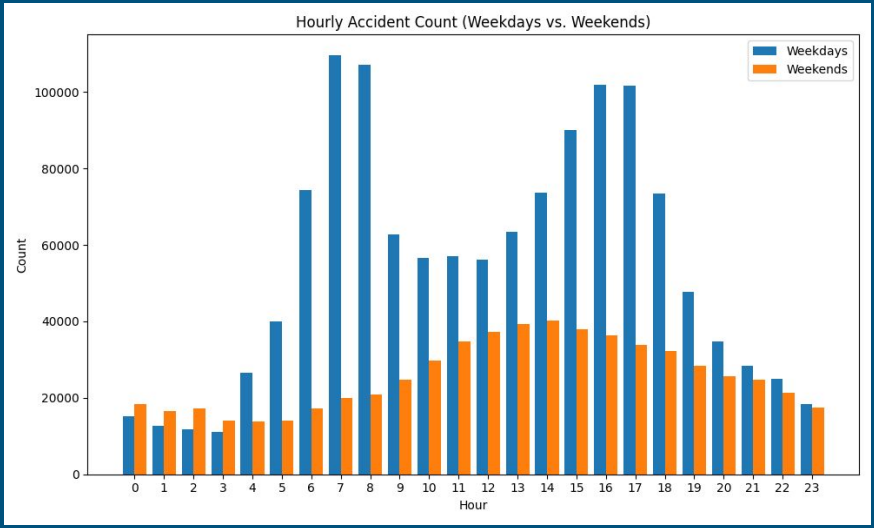
Bayesian Classification

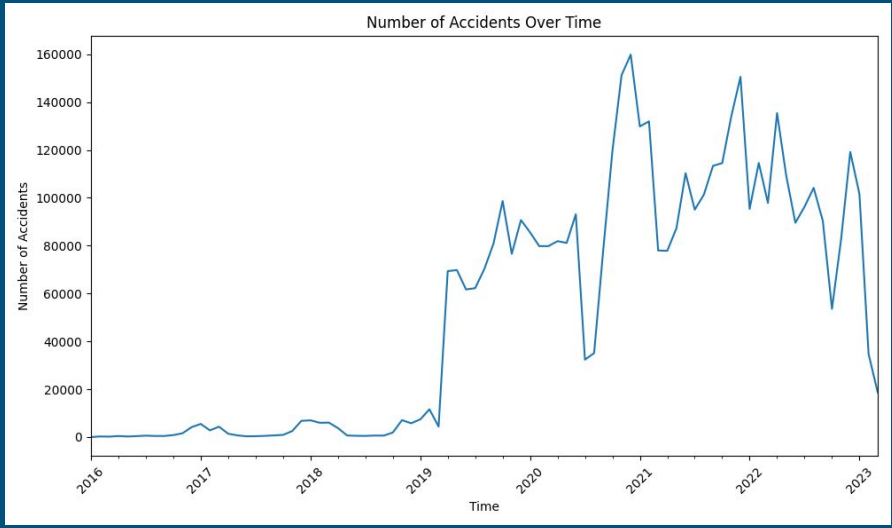
- As driving conditions become more hazardous, the probability of an accident being severe is increased.
- Drawing inferences from the findings, we can conclude that because most accidents are not weather related, a majority are not as severe.

Bayesian Classification:
Hazardous Driving Conditions Predict Accident Severity

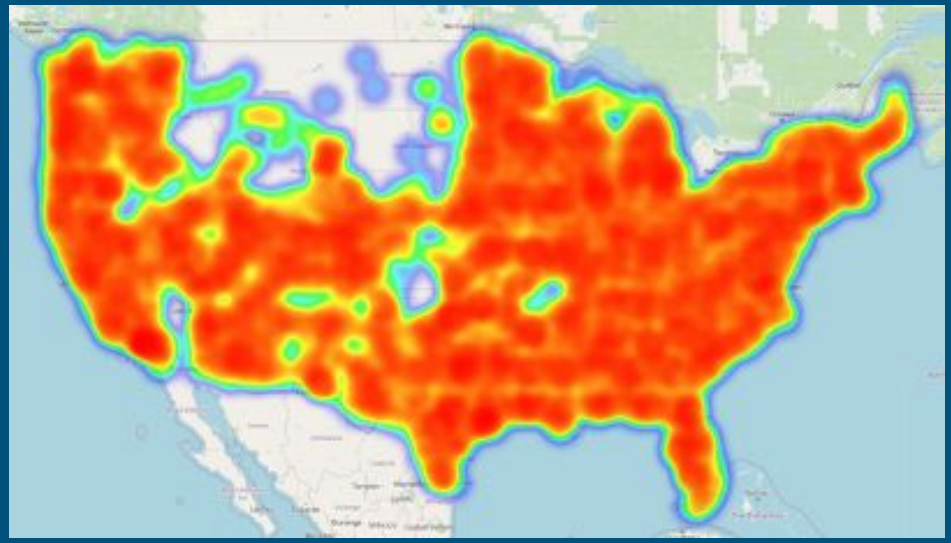
Conditions ↓ / Severity →	Minor Accident	Major Accident	Severe Accident
'Wind Speed' = Strong Breeze 'Visibility' = Poor	0.0840%	0.0347%	0.0485%
'Wind Speed' = Strong Breeze 'Visibility' = Poor 'Time of Day' = Night 'Temperature' = Freezing	0.001814%	0.001239%	0.002026%
'Wind Speed' = Strong Breeze 'Visibility' = Poor 'Temperature' = Freezing 'Time of Day' = Night 'Wind Chill' = Dangerous 'Humidity' = Wet	0.0001262%	0.0001067%	0.0002115%

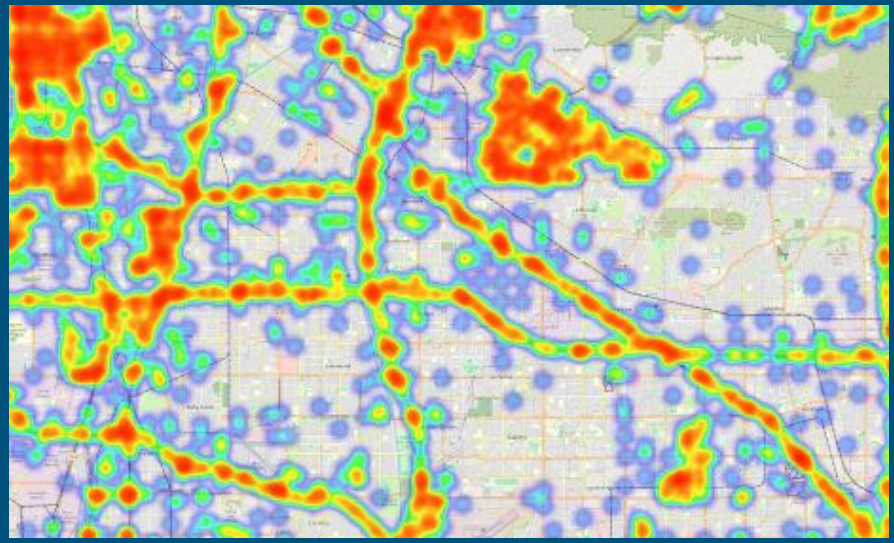
Accidents over time





Heatmap





How these findings can be applied

- To avoid a severe accidents, commuters should not drive in hazardous conditions
- Because most accidents occur in nice weather, there's likely a larger volume of cars on the road when driving conditions are good.
- Commuters should avoid driving on weekdays during peak traffic hours: (7 - 9 am) (4 - 6pm)
- Commuters should avoid driving during holiday seasons
- Commuters should consider using public transport in highly populated areas and local governments should subsidize public transportation.
- There are significantly fewer accidents when there are fewer cars on the road -More public transportation

Citations

- chonyy. "Apriori Python." GitHub, GitHub, https://github.com/chonyy/apriori_python.
- 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.