Predicting Traffic Accident Severity

Zachary Strain
25 September 2020

100 deaths per day in the US

As a result of traffic accidents involving motor vehicles (CDC)

Intro

Predicting traffic accident severity is beneficial to public health and safety

- Traffic accidents involving motor vehicles cause 100 deaths per day in the US
- Cost of related productivity losses and medical care exceeds $75 billion
- Could data on location, weather condition, and points of interest around traffic accidents predict when and where more sever accidents are more likely to occur?
- Such data would empower city officials (i.e. transportation, safety, and zoning departments) to plan better and safer cities, and more effectively deploy city resources in response to traffic accidents.

Dataset Source

- The data used for this project is from a dataset made available through the research from the following papers:
  - Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (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.

About the Dataset

- Collected between February 2016 and June 2020,
- Contains data on approximately 3.5 million traffic accidents:
  - weather conditions (i.e. temperature, precipitation, wind speed, etc.),
  - location information (i.e., coordinates, street address, city, state),
  - points of interest nearby the traffic accident (i.e., crossing, speed bump, station, railway, stop, traffic signal),
- Collected from 49 states in the US; this analysis focuses exclusively on Texas

`Severity` of accidents in Texas

Exploring the outcome variable

- According to the author of the dataset, the `Severity` variable ranks the severity of an accident on a scale of 1 (least severe) to 4 (most severe).

- Understood as the impact of the accident on traffic in terms of the length of delay it causes, i.e., an accident with a severity ranking of 4 caused a longer delay to traffic than an accident with a severity ranking of 3.

Figure 1: Number of Accidents by Severity Value

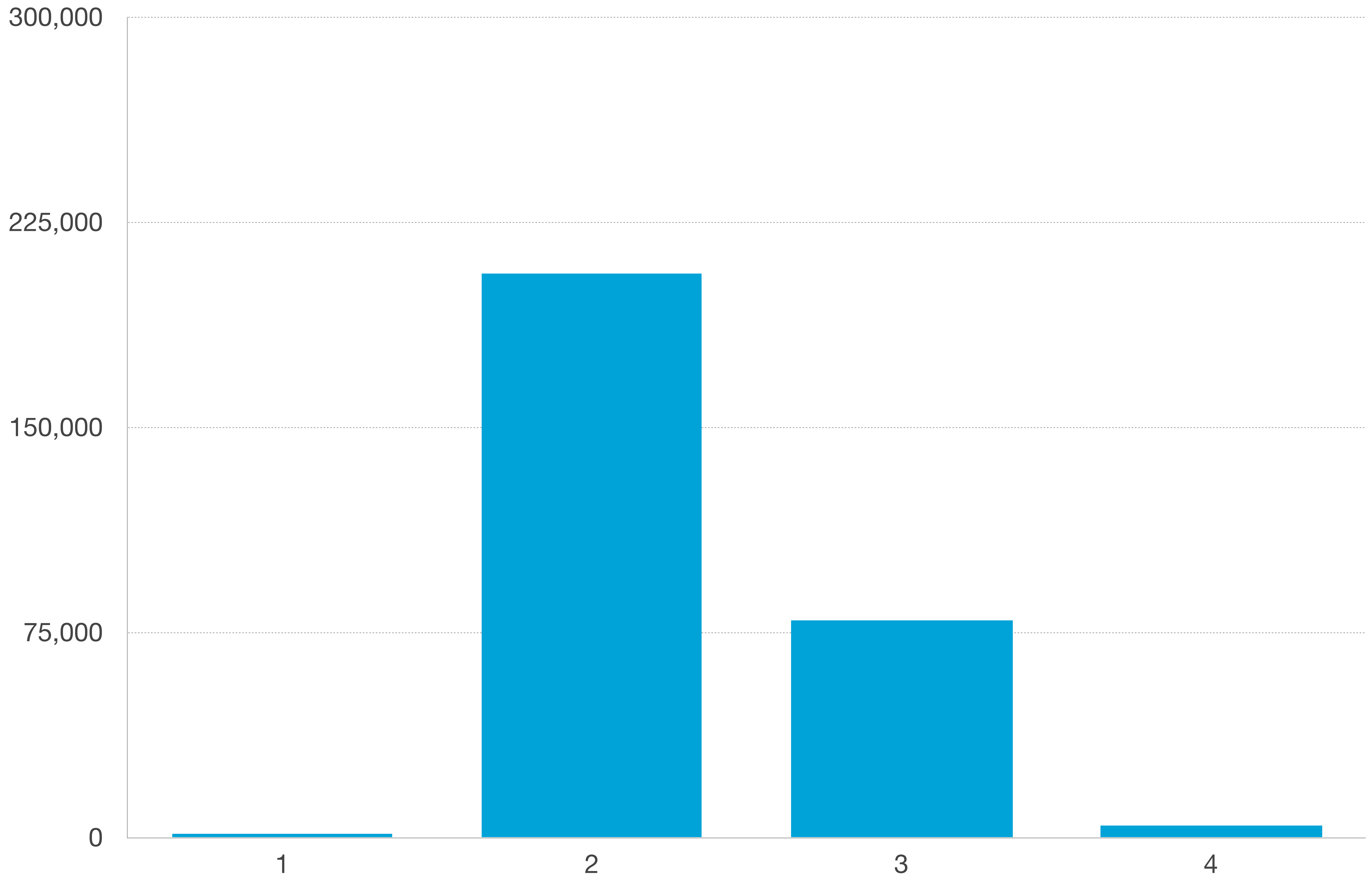
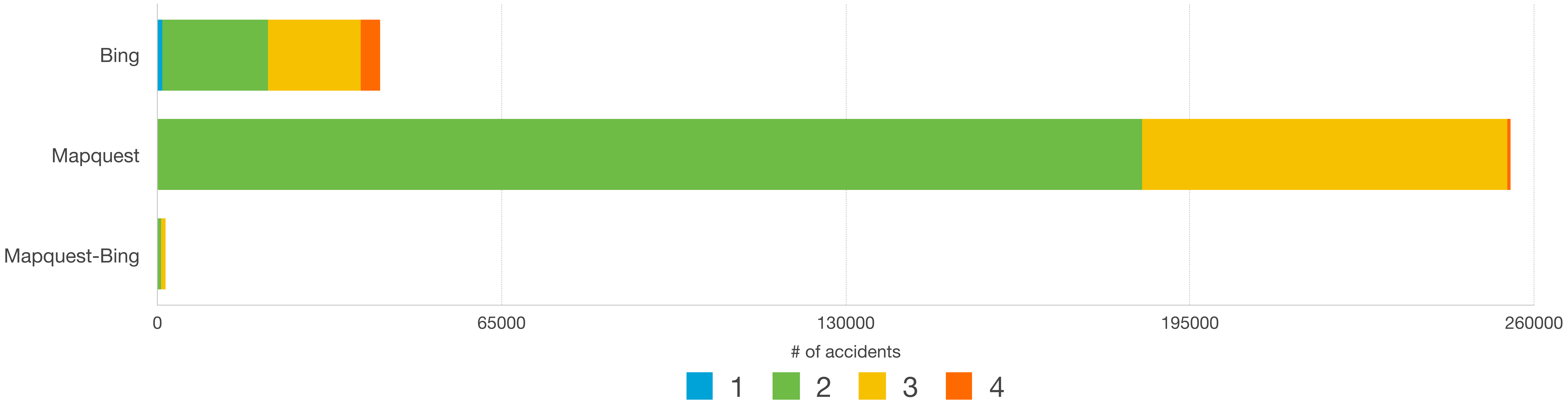


Figure 2: Severity of Traffic Accidents by reporting source



Data sources

Traffic accident data used in the dataset was categorized in to the following sources:

- Bing
- Mapquest
- Bing-Mapquest

Given the more even distribution of data from Bing, analysis focused exclusively on these observations

Frequency and severity of accidents by day

- More accidents happened on weekdays

- But accidents on the weekend were more severe on average

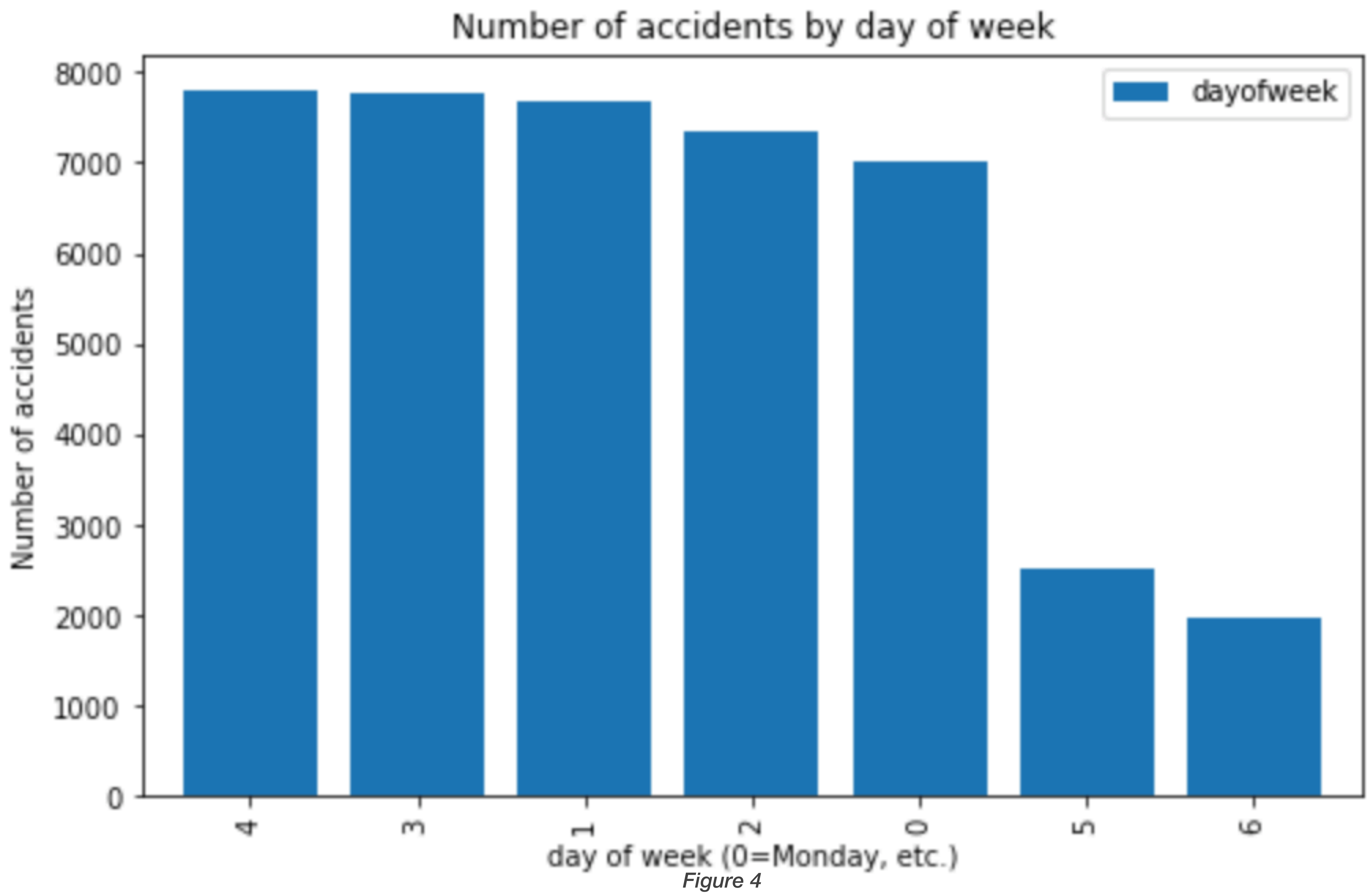
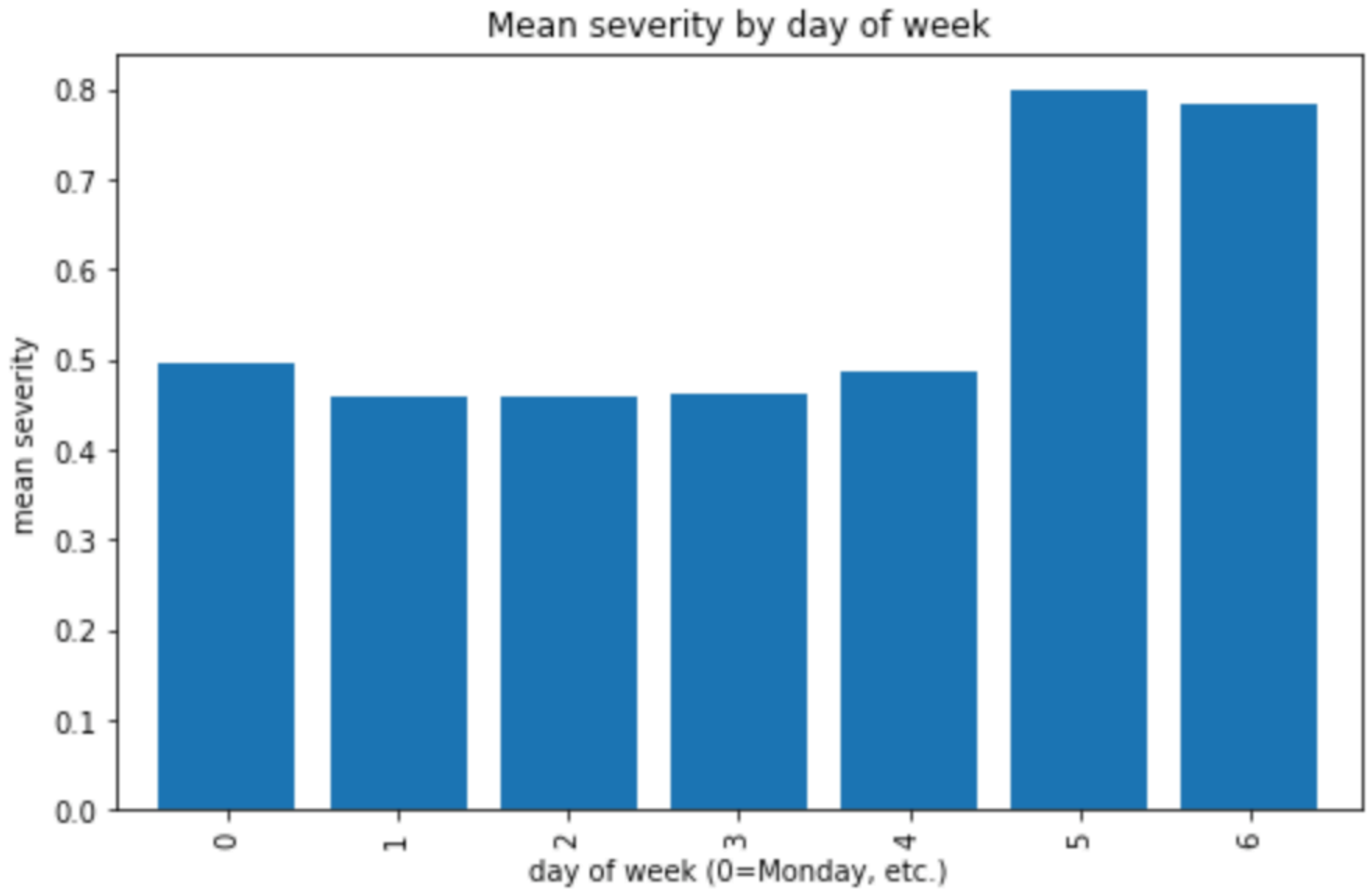


Figure 4



Frequency and severity of accidents by hour

- Most accidents occurred during commuting hours, suggesting that the number of cars on the road correlates with number of accidents

- However, accidents were more severe on average at night.

Figure 5

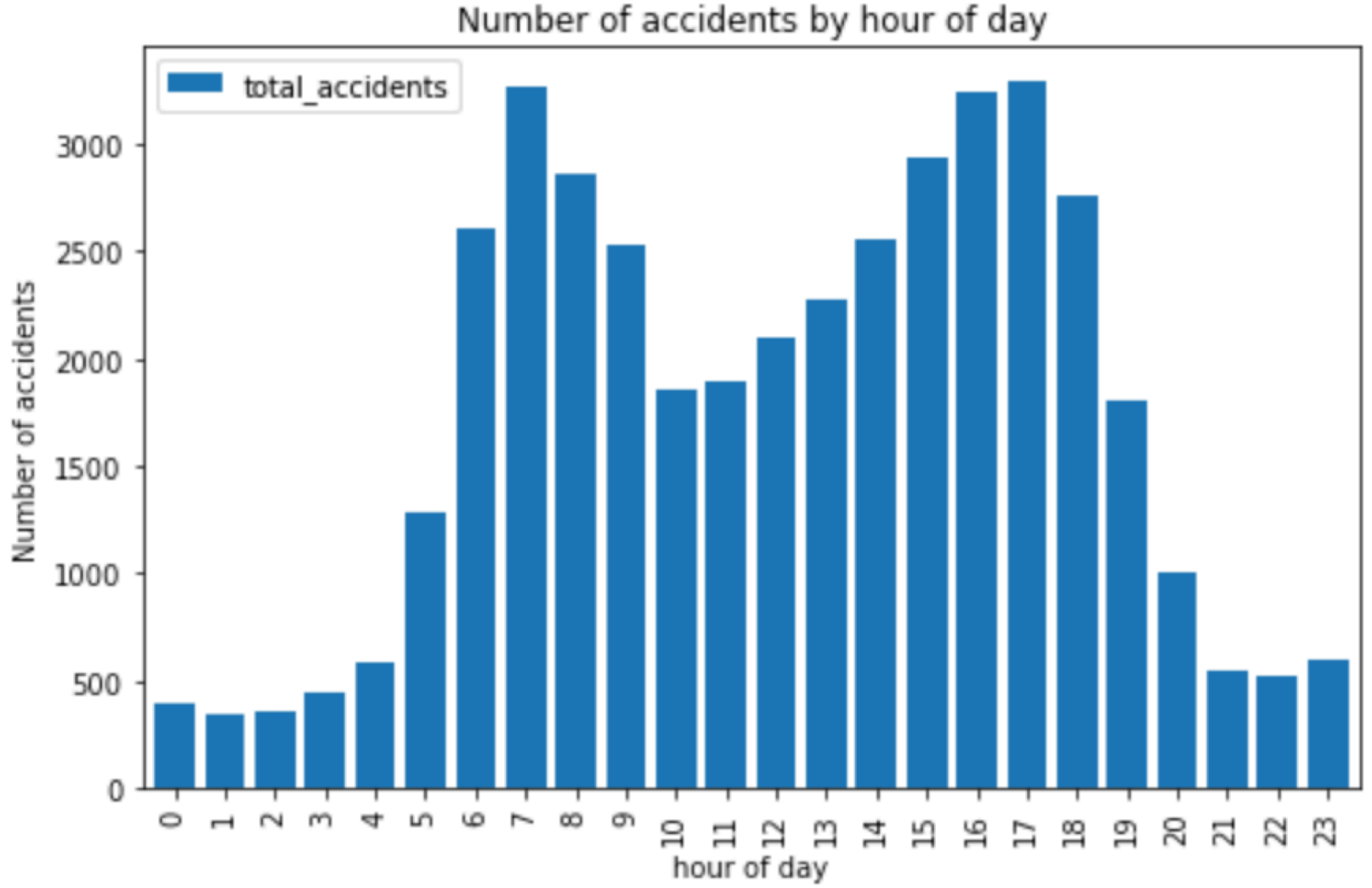
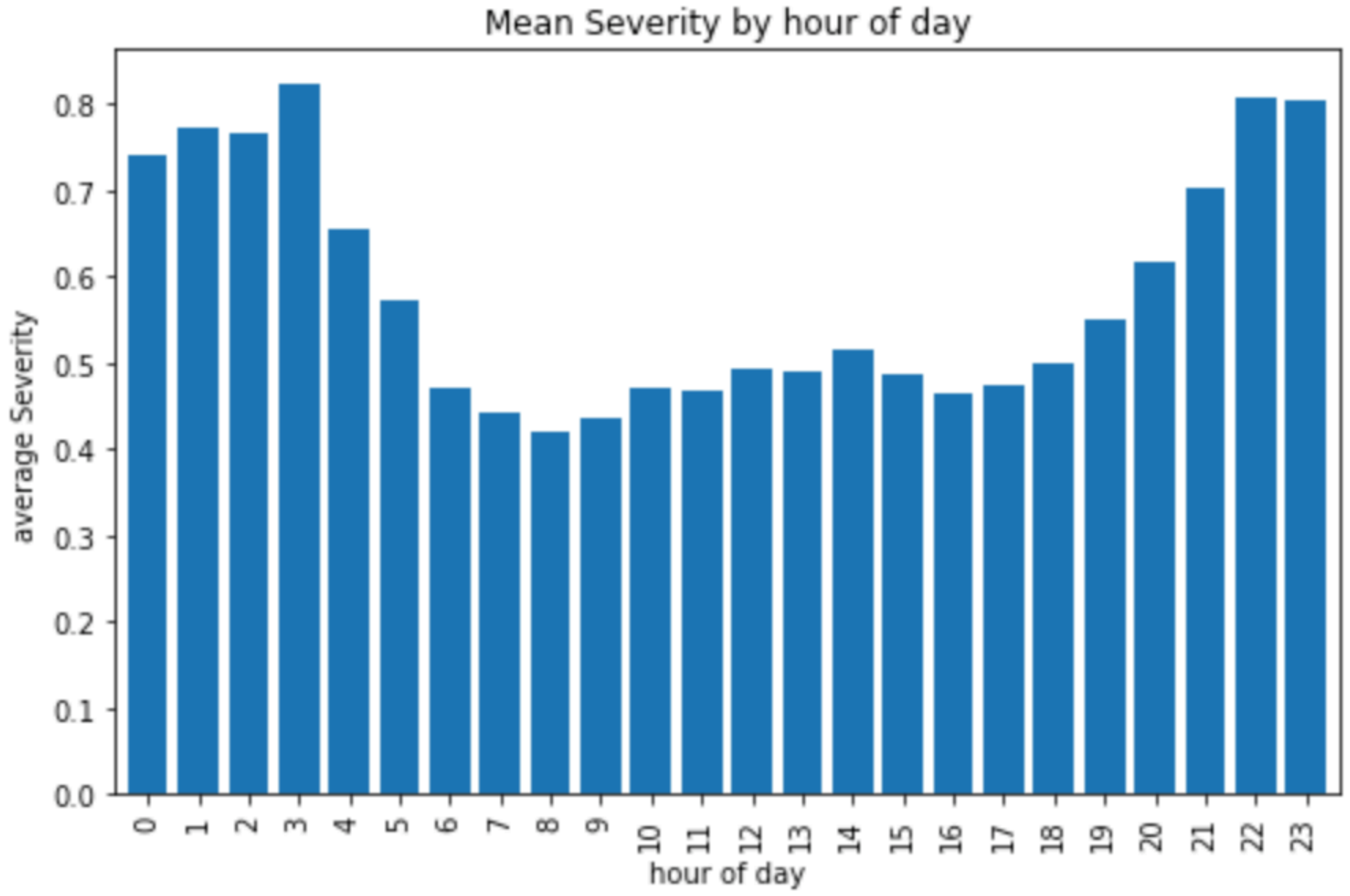
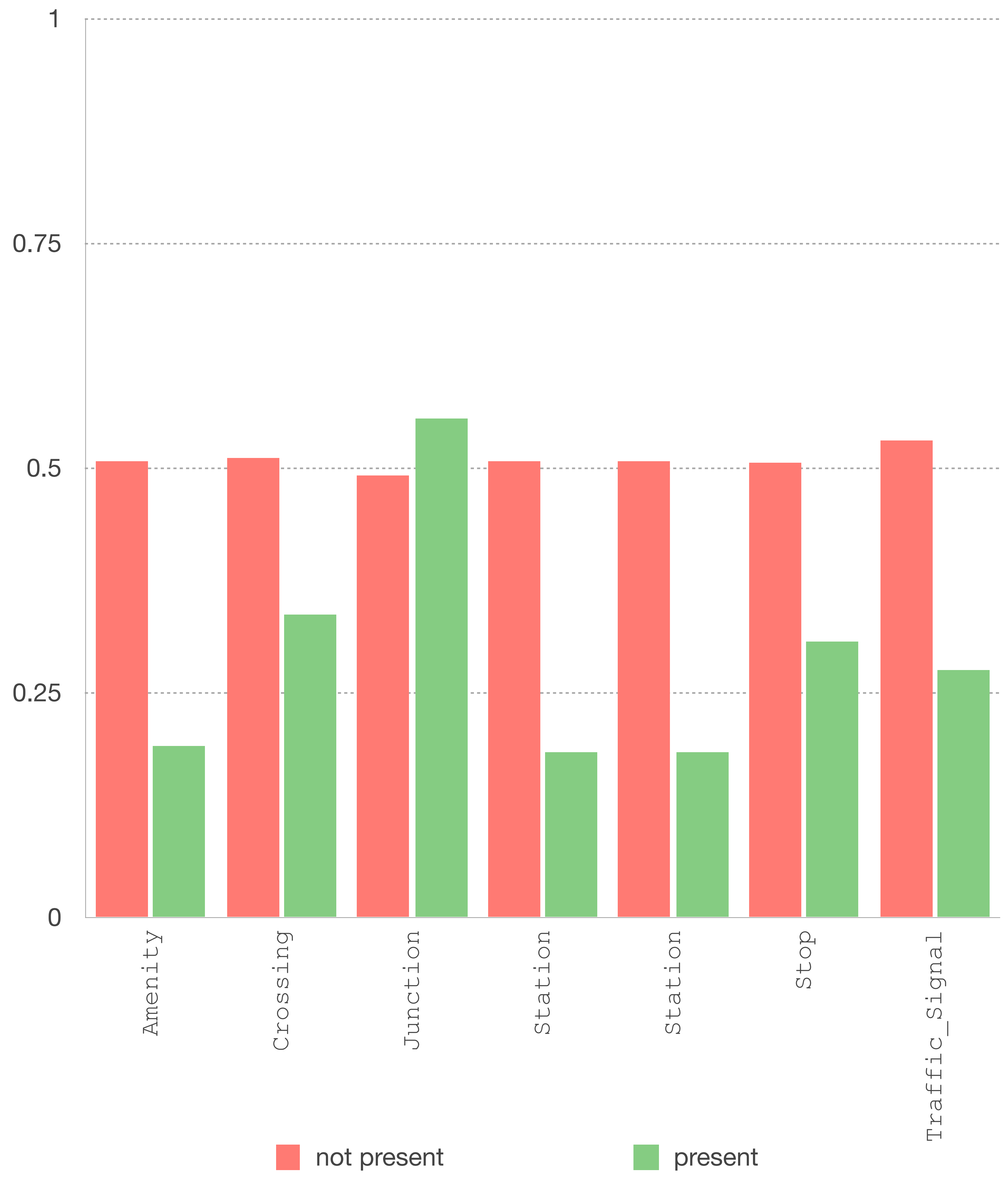


Figure 6



Presence of Points of Interest (POIs)

Figure 7: Mean Severity by presence of POI



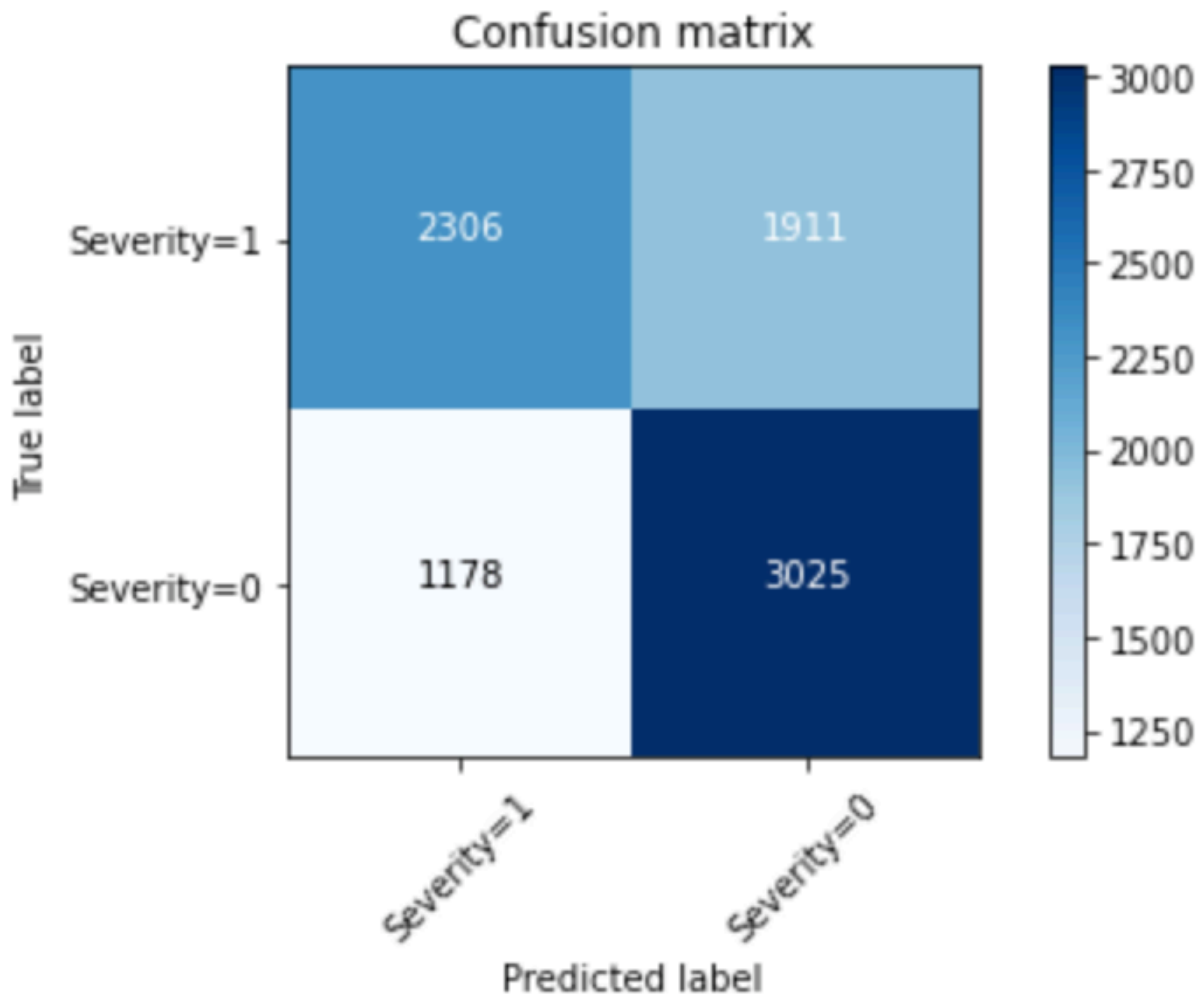
- The mean severity of accidents varied on average depending on the presence of different POIs.
- Far less noticeable differences for weather variables

Predicting Severity

Logistic Regression Model

- This model calculates the probability of an accident having a Severity rating of 0 or 1.
- Model performed better in predicting Severity of 0 than 1.
- Overall evaluation metrics
  - Precision = 0.64
  - Recall = 0.63
  - F1-score = 0.63
  - Log-loss = 0.63

Figure 8



Discussion and Conclusion

- Environmental variables (weather, infrastructure, time), to a limited degree, can help predict the severity of traffic accidents
- Given the evaluation metrics and mostly weak correlation between independent variables and the dependent variable, there is potential to improve the predictive power of this model
- Exploratory analysis suggest that accidents that occur at night and off-peak hours are more severe on average
- Limitations:
  - Single-data source
  - Lack of geographic data analysis
  - Focus on one state: Texas

Infrastructural elements are certainly the ones most within a realm of human control, and for those that have a significant impact on traffic accident severity, city planners/developers, transportation officials, etc., would find more robust results of interest.