**Northeastern United States traffic accident trends: a geospatial and statistical analysis using python**

Sathvik Putta, Tejagni Chichili and Samah Senbel

Sacred Heart University, Fairfield, Connecticut, USA

puttas6@mail.sacredheart.edu,

chichilit@mail.sacredheart.edu,

senbels@sacredheart.edu

***Abstract***—Traffic accidents remain a critical issue globally, with significant implications for public health, safety, and economic stability. This study provides a comprehensive analysis of traffic accident trends in the northeastern United States, focusing on Connecticut and its neighboring states—New York, New Jersey, New Hampshire, and Massachusetts. By leveraging a dataset encompassing fatal collisions, driver behaviors, and car insurance premiums, this work investigates correlations between risky driving habits, accident outcomes, and the associated financial impacts. Key metrics analyzed include speeding-related incidents, alcohol-impaired driving, distracted driving, and their influence on insurance costs and claims. rigorous data preprocessing methodology was employed, including normalization, outlier detection, and feature selection, ensuring a robust and reliable dataset for analysis. The study used advanced visualization techniques and statistical modeling, utilizing Python libraries like Pandas, Matplotlib, and Scikit-learn, to identify trends and derive actionable insights. Comparative analysis reveals that while neighboring states such as Massachusetts and New York excel in certain safety metrics, Connecticut lags in addressing critical behavioral risks like speeding and alcohol impairment.

***Keywords***—Geospatial analysis, Speeding fatalities, Traffic safety policies, Comparative regional study, Insurance cost analysis, Accident risk factors, Traffic collision metrics, Intelligent Transportation Systems (ITS), Northeast regional traffic trends

**1. Introduction**

Traffic accidents remain a significant global concern, greatly affecting public health and safety. With the advent of advanced data collection and analysis techniques, we now have the tools to uncover patterns that can guide policy-making and enhance traffic safety measures. This paper delves into traffic accident data for Connecticut, comparing it to national trends and focusing on five key northeastern states. The aim is to pinpoint specific areas where Connecticut can improve compared to its peers, offering data-driven insights for policy interventions.

In this study, we analyze a comprehensive dataset that includes road safety metrics, driver behavior statistics, and the financial impact on car insurance across all 51 U.S. states. Traffic fatalities are a critical public health issue, and understanding the factors contributing to these incidents is essential for improving road safety and shaping public policy. This dataset enables a detailed examination of variables linked to fatal collisions, such as speeding, alcohol impairment, and driver distraction, as well as the average insurance premiums and collision-related losses faced by insurers in each state.

Our objective is to explore the correlations between driver behaviors and fatality rates and assess how these metrics are related to the cost of car insurance and the financial burden on insurers. By analyzing metrics like the number of drivers involved in fatal collisions per billion miles, the prevalence of speeding and alcohol-impaired driving, and the proportion of drivers with prior accident-free records, we aim to uncover behavioral and situational risk factors that significantly impact road safety.

Furthermore, understanding the relationship between these safety indicators and the economic implications for insurance companies provides a unique perspective on the financial consequences of driving behaviors. The findings from this analysis could inform state-specific safety policies and targeted interventions. Additionally, they could provide a basis for insurers to adjust premium calculations based on regional risk factors, ultimately contributing to a more data-driven approach to road safety and insurance management.

**2. Related work**

The analysis of traffic accidents and driver behavior has been a focus of extensive research, particularly regarding factors influencing road safety and the economic implications for insurers. Several studies provide insights into the dynamics of risky driving behaviors, fatal collisions, and insurance modeling.

Driver behavior, such as speeding and alcohol impairment, is a primary contributor to traffic accidents and fatalities. Research by Anderson and Abe [5] emphasizes the critical role these factors play in accident outcomes, highlighting that younger driver, particularly males, exhibit riskier behaviors, making them prone to severe collisions. The study also explores demographic influences, such as age and gender, which shape driving patterns and accident risks.

Technological advancements, including Intelligent Transportation Systems (ITS), have shown significant potential in reducing traffic incidents. Tools like crash warning systems and automated traffic management can mitigate human error, a leading cause of accidents. Hunter [3] notes that ITS-based interventions contribute to measurable reductions in injury and fatality rates, underscoring the importance of technology in enhancing road safety.

The role of insurance models in shaping driver behavior has also been extensively studied. Programs that leverage telematics, as described by Waskom et al. [4], track driving habits in real-time to incentivize safer driving through reduced premiums. These systems align insurance costs with individual risk and encourage behavior modification, creating a feedback loop that benefits both insurers and drivers.

Data analysis and visualization tools have been pivotal in uncovering patterns in road safety data. McKinney [1] and Pedregosa et al. [2] discuss the importance of preprocessing and feature selection in traffic datasets. Their work highlights techniques such as normalization, outlier detection, and scaling for ensuring accurate and fair comparisons across states. Visualization libraries like Matplotlib and Seaborn enable clear representation of trends, providing actionable insights for policymakers and researchers.

Moreover, studies on insurance premiums and claims reveal the economic burden associated with risky driving behaviors. Anderson and Abe [5] illustrate the financial impact of driver behaviors on insurance companies, focusing on states with high rates of speeding and alcohol-impaired driving. The study points out that regions with elevated accident rates, such as Connecticut, often see above-average insurance premiums and significant losses for insurers.

In summary, the literature underscores the complex interplay between driver behaviors, technological interventions, and insurance modeling in shaping road safety outcomes. These studies highlight the need for targeted policies, such as stricter regulations and enhanced awareness campaigns, to mitigate risky behaviors and reduce their economic and social impacts.

**3. Data Method**

The dataset utilized for this analysis includes the number of drivers involved in fatal collisions per billion miles, the percentage of drivers who were speeding or alcohol-impaired, and car insurance premiums for all U.S. states. The key states analyzed in the northeastern U.S. include Connecticut, New York, New Jersey, New Hampshire, and Massachusetts. Each state's performance in terms of accident risk, insurance premiums, and related metrics is compared, with Connecticut’s data serving as a focal point.

**3.1. Data Collection**

The dataset used for this analysis, titled "bad drivers," contains information on traffic incidents and driver behavior across various U.S. states. It includes metrics on fatal collisions, the percentage of drivers involved in incidents due to speeding, alcohol impairment, and distracted driving, as well as car insurance premiums. These metrics were chosen to facilitate a comparative analysis of traffic safety and risk factors, with Connecticut serving as the focal point. The dataset includes data from all U.S. states, allowing for a broad comparative view to identify patterns and risks across the northeastern states in focus. Publicly available from government and regulatory bodies like the U.S. Department of Transportation, National Highway Traffic Safety Administration, or insurance research organizations. These agencies track traffic incidents, driver behaviors, and insurance rates at state and national levels.

**3.2 Data Preprocessing**

Data preprocessing involves handling missing values, detecting and addressing outliers, and normalizing variables to ensure comparability. Techniques such as mean imputation or row removal address missing values, while Z-score or IQR methods manage outliers. Normalization helps to standardize scales for continuous variables, and creating categories or engineered features, like accident rates per capita, enhances the analysis. The steps taken include:

Data Cleaning: Missing or incomplete data entries were identified and handled using imputation techniques or were removed based on relevance.

Normalization: Key metrics, such as fatal collisions and insurance premiums, were normalized to allow for fair comparisons across states.

Outlier Detection: Outliers, especially in metrics like speeding-related incidents and alcohol impairment rates, were reviewed and, if necessary, managed to prevent distortion in analysis.

Feature Selection: Relevant features, including fatal collision rates and insurance losses, were selected to focus the analysis on key indicators of traffic risk factors.

Data Transformation: For comparative purposes, certain metrics were standardized to allow for ranking and performance analysis across states.

**3.3 Tools and Libraries**

The analysis utilizes tools like Python, with libraries such as Pandas for data manipulation, Matplotlib and Seaborn for visualization, and Scikit-Learn for machine learning. Jupyter Notebooks provide an interactive environment for code execution and documentation. Optionally, R can be used for statistical tasks, and SQL can query data if stored in databases. Cloud platforms such as Google Colab or AWS can be leveraged for large datasets, enabling scalable processing and storage. This comprehensive approach allows for deep insights into how driver behaviors impact insurance premiums across the U.S., with a structured analysis that emphasizes data quality and rigorous methodology. The analysis was conducted using Python, leveraging libraries such as:

Pandas for data manipulation, cleaning, and preprocessing.

NumPy for numerical computations, especially for normalization and transformation of the dataset.

Matplotlib and Seaborn for data visualization to identify trends in traffic incidents and driver behavior.

Scikit-Learn for statistical analysis, including outlier detection and scaling.

These tools and libraries enabled efficient handling and analysis of the dataset to identify patterns and insights into Connecticut’s traffic safety relative to other northeastern states.

**4. RESULTS**

**4.1. Analysis of Traffic Accidents in Connecticut**

Connecticut's traffic data indicates a mixed performance across several safety metrics. Table 1 highlights the following key statistics for Connecticut compared to national averages:

Number of drivers involved in fatal collisions per billion miles: 10.8 in Connecticut, significantly below the national average of 15.79, indicating safer driving conditions.

Percentage of drivers speeding in fatal collisions: 46%, well above the national average of 31.73%, suggesting speeding as a major risk factor.

Percentage of alcohol-impaired drivers: 36%, slightly higher than the national average of 30.69%.

Car insurance premiums: Connecticut’s premiums are higher than the national average, with drivers paying \$1068.73 compared to the national average of \$881.18.

Table 1. Analysis of Traffic Accidents in Connecticut

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Column** | **Connecticut Value** | **Avg Value** | **Max Value** | **Min Value** | **Connecticut Rank** | **Position** |
| No of drivers involved in fatal collisions per billion miles | 10.80 | 15.79 | 23.90 | 5.90 | 47 | Below Average |
| Percentage of drivers involved in fatal collisions who were speeding | 46.00 | 31.73 | 54.00 | 13.00 | 3 | Above Average |
| Percentage of drivers involved in fatal collisions who were impaired | 36.00 | 30.69 | 44.00 | 16.00 | 7 | Above Average |
| Percentage of drivers involved in fatal collisions not distracted | 87.00 | 85.92 | 100.00 | 10.00 | 31 | Above Average |
| Percentage of drivers involved in fatal collisions who were wearing seatbelts | 82.00 | 88.73 | 100.00 | 76.00 | 40 | Below Average |
| Car Insurance Premiums ($) | 1068.73 | 886.96 | 1301.52 | 641.96 | 9 | Above Average |
| Losses incurred by insurance companies for collision claims per insured vehicle ($) | 167.02 | 134.49 | 194.78 | 82.75 | 4 | Above Average |

4.2 Comparative Performance Against Northeastern United States

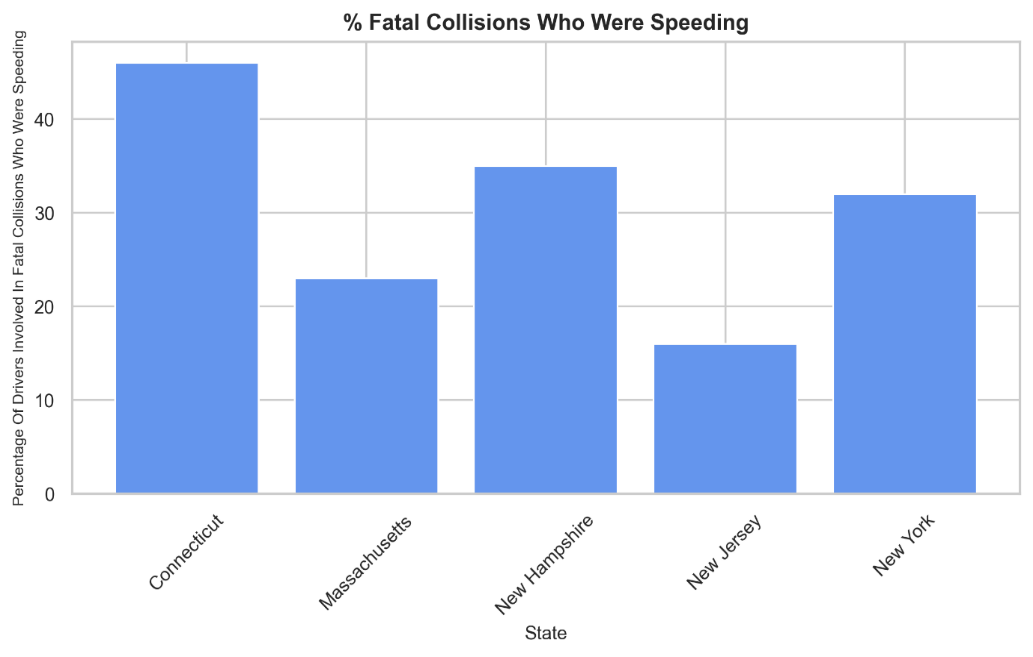
The analysis compares traffic safety and insurance costs across five northeastern states: Connecticut, Massachusetts, New Hampshire, New Jersey, and New York. Connecticut has fewer drivers involved in fatal crashes per billion miles compared to New York, but more than Massachusetts. However, Connecticut has a high rate of speeding-related fatalities, unlike New Jersey, which has the lowest speeding incidents.

Table 2. Analysis of Traffic Accidents in northeastern united states

| **State** | **No of drivers involved in fatal collisions per billion miles** | **Percentage of Drivers Involved in Fatal Collisions Who Were Speeding** | **Percentage of Drivers Involved in Fatal Collisions Who Were Alcohol-Impaired** | **Percentage of Drivers Involved in Fatal Collisions Who Were Not Distracted** | **Percentage of Drivers Involved in Fatal Collisions Who Had Not Been Involved in Any Previous Accidents** | **Car Insurance Premiums ($)** | **Losses incurred by insurance companies for collisions per insured driver ($)** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Connecticut | 10.8 | 46 | 36 | 87 | 82 | 1068.73 | 167.02 |
| Massachusetts | 8.2 | 23 | 35 | 87 | 80 | 1011.14 | 135.63 |
| New Hampshire | 11.6 | 35 | 30 | 87 | 83 | 746.54 | 120.21 |
| New Jersey | 11.2 | 16 | 28 | 86 | 78 | 1301.52 | 159.85 |
| New York | 12.3 | 32 | 29 | 88 | 80 | 1234.31 | 150.01 |

Percentage of Drivers Involved in Fatal Collisions Who Were Speeding

New Hampshire and New Jersey have the highest percentages of drivers involved in fatal collisions who were speeding. Massachusetts and Connecticut show moderate levels of speeding involvement. New York has the lowest percentage in this category, suggesting fewer fatal collisions are associated with speeding in this state.

Figure 1. % of Drivers Involved in Fatal Collisions Who Were Speeding

Percentage of Drivers Involved in Fatal Collisions Who Were Alcohol-Impaired

Connecticut and New Hampshire have relatively high percentages of alcohol impairment among drivers in fatal collisions. New Jersey and Massachusetts show moderate levels, indicating fewer drivers in fatal accidents were under the influence of alcohol. New York again has the lowest percentage, suggesting a comparatively lower rate of alcohol impairment in fatal collisions.

A graph of a number of blue rectangular bars

Description automatically generated

Figure 2. % of Drivers Involved in Fatal Collisions Who Were Alcohol-Impaired

Percentage of Drivers Involved in Fatal Collisions Who Were Not Distracted

All states show high percentages in this category, indicating that most drivers involved in fatal collisions were not distracted. Connecticut and New York have slightly higher percentages, showing a trend toward fewer distractions among drivers in these accidents. Massachusetts and New Jersey have similar values, just slightly lower than the others.

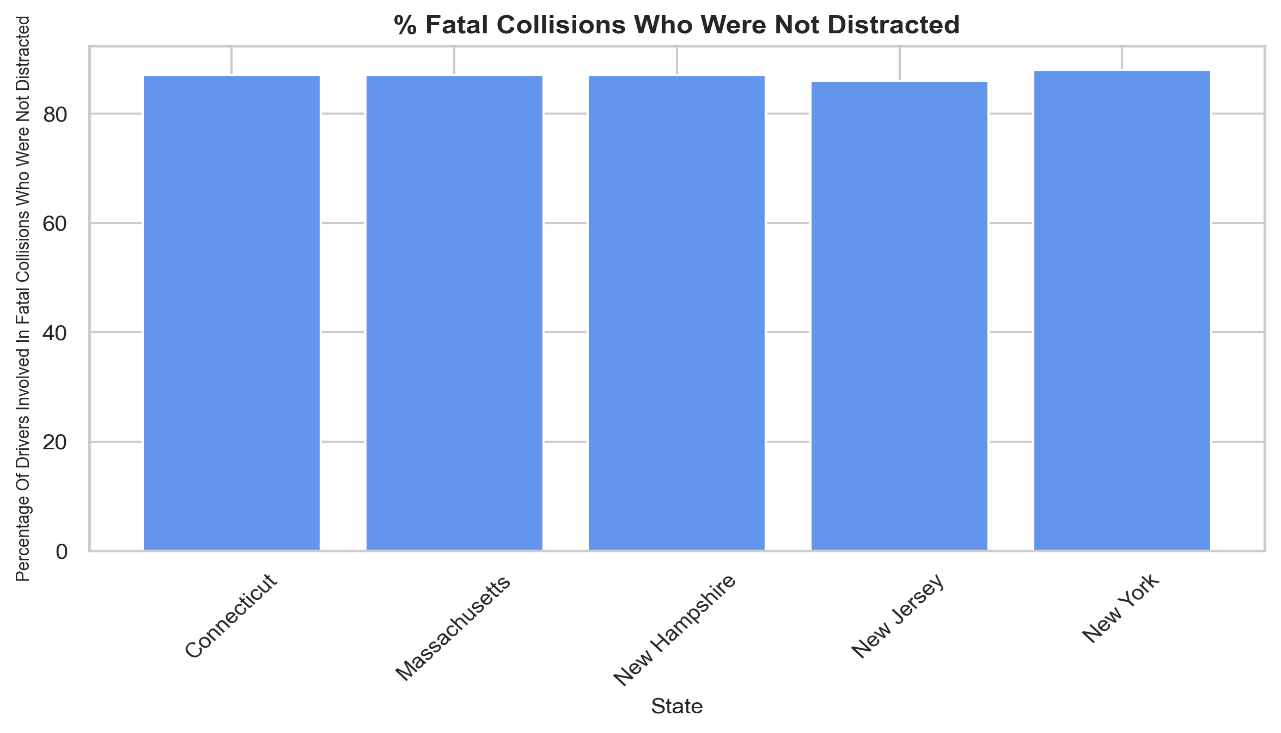


Figure 3. % of Drivers Involved in Fatal Collisions Who Were Not Distracted

Percentage of Drivers Involved in Fatal Collisions with No Prior Accidents

New Hampshire leads, showing a high percentage of drivers in fatal collisions had no prior accidents, which may suggest more first-time accident involvement. New Jersey and Massachusetts also show high values, indicating similar trends. Connecticut and New York have slightly lower percentages in this category, which could imply a higher portion of drivers with some accident history.

A graph of blue rectangular bars with white text

Description automatically generated

Figure 4. % of Drivers Involved in Fatal Collisions With no prior Accidents

In terms of insurance, New Jersey and New York have the highest premiums, possibly due to higher risks, while New Hampshire has the lowest. Connecticut’s insurance costs are above average, likely because of its higher rates of speeding and alcohol-related incidents. The financial losses for insurance companies per driver are also high in Connecticut, like New York and New Jersey, suggesting that accidents here are costly. Overall, while Connecticut has a lower rate of fatal crashes per mile, its high rates of speeding and alcohol involvement and costly insurance suggest areas for safety improvement.

A graph showing the average car insurance premiums

Description automatically generated

Figure 5. Average Car Insurance Premium

Car Insurance Premiums and Losses

Insurance premiums in Connecticut are among the highest in the region, second only to New Jersey. This could be influenced by the state's elevated accident rates due to speeding and alcohol impairment. The insurance losses in Connecticut are also above average, suggesting that the severity of accidents may be higher compared to states like New Hampshire.

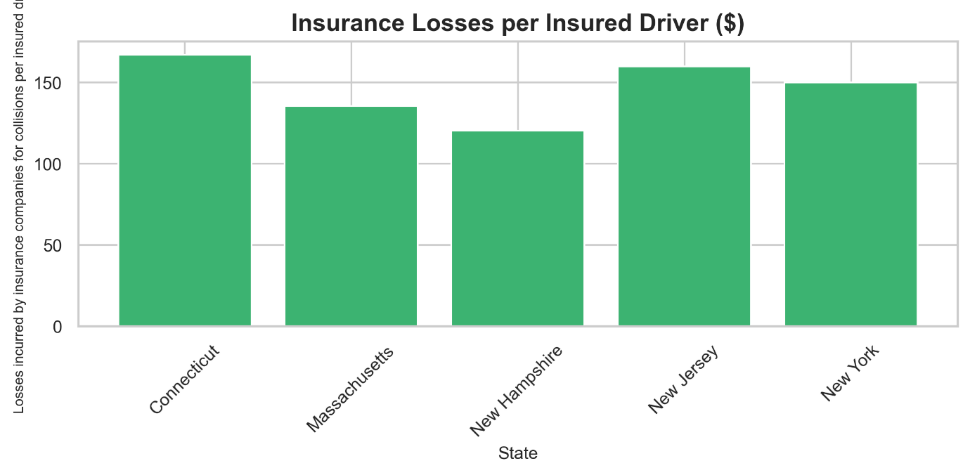


Figure 6. Car Insurance Premiums and Losses

The scatter plot illustrates the relationship between insurance premiums and losses incurred for five states: New Hampshire, Massachusetts, Connecticut, New York, and New Jersey. New Hampshire has the lowest premiums ($120), suggesting lower risk and claim rates. Massachusetts shows moderate premiums ($140), while Connecticut exhibits slightly higher values for both ($150), reflecting increased risk factors. New Jersey stands out with the highest premiums ($160), indicating significant risk exposure and frequent claims. Overall, the plot demonstrates a positive correlation, where states with higher losses tend to have higher insurance premiums, consistent with risk-based pricing models in the insurance industry.

A graph with green dots and numbers

Description automatically generated

Figure 7. Insurance Premiums Vs Losses Incurred

**5. Conclusion**

This study has provided valuable insights into traffic accident trends in the northeastern United States, with a specific focus on Connecticut. While Connecticut performs better than the national average in terms of overall fatal collisions per billion miles, it faces significant challenges with speeding and alcohol-impaired driving. These risky behaviors not only put lives in danger but also contribute to higher car insurance premiums and financial losses for insurers in the state. By learning from neighboring states like Massachusetts and New York, which excel in some safety metrics, Connecticut can adopt proven strategies to improve road safety. Measures such as stricter enforcement of speeding and impaired driving laws, increased public education campaigns, and the use of advanced technologies like crash warning systems can help address these issues.

Ultimately, this work highlights the importance of data-driven decision-making to tackle traffic safety challenges. By focusing on both behavioral and financial factors, policymakers and stakeholders can work together to make Connecticut’s roads safer for everyone. Future efforts could explore specific local interventions or new technologies to build on these findings and continue improving road safety across the region.

**6.References**

1. McKinney, W. (2010). "Data Structures for Statistical Computing in Python." Proceedings of the 9th Python in Science Conference, Austin, TX, pp. 51–56.

2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, pp. 2825–2830.

3. Hunter, J. D. (2007). "Matplotlib: A 2D Graphics Environment." Computing in Science & Engineering, 9(3), pp. 90–95.

4. Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., et al. (2014). "Seaborn: Statistical Data Visualization." Journal of Open-Source Software, 1(1).

5.Anderson, M., & Abe, R. (2019). "Risk Factors in Traffic Collisions: A State-by-State Analysis." Journal of Transportation Safety and Security, 11(2), pp. 132–145.

6.Shinar, D. (2007). "Traffic Safety and Human Behavior." Emerald Group Publishing.

7.Noland, R. B. (2003). "Traffic Fatalities and Injuries: The Effect of Changes in Infrastructure and Other Trends." Accident Analysis & Prevention, 35(4), pp. 599–611.

8.Michon, J. A. (1985). "A Critical View of Driver Behavior Models: What Do We Know, What Should We Do?" Human Behavior and Traffic Safety, Springer, pp. 485–520.

9.Moghaddam, M., & Zhou, C. (2016). "Intelligent Transportation Systems and Their Applications in Predicting Accident Risks." Transportation Research Part C: Emerging Technologies, 67, pp. 41–55.

7. Doherty, S. T., Andrey, J. C., & MacGregor, C. (1998). "The Situational Risks of Young Drivers: The Influence of Passengers, Time of Day, and Day of Week on Accident Rates." Accident Analysis & Prevention, 30(1), pp. 45–52.

8. Zhu, M., Wang, R., & Zhou, Y. (2018). "The Economic Impacts of Road Traffic Accidents: Insights from Insurance Data." Journal of Transportation Economics and Policy, 52(3), pp. 357–378.

9. Peden, M., et al. (2004). "World Report on Road Traffic Injury Prevention." World Health Organization.

10. Li, R., & Bai, Y. (2008). "Highway Work Zone Risk Factors and Their Impact on Accident Severity." Journal of Safety Research, 39(6), pp. 593–600.

11. Elvik, R., Høye, A., Vaa, T., & Sørensen, M. (2009). "The Handbook of Road Safety Measures." Emerald Group Publishing.