DATA C102 SP24 Final Project Team 42

JORDAN G. TAQI-EDDIN, University of California - Berkeley, USA
PETER FLO GRINDE-HOLLEVIK, University of California - Berkeley, USA
CHASE HERTEL, University of California - Berkeley, USA
MING SENN TEO, University of California - Berkeley, USA

Additional Key Words and Phrases: road safety, difference-in-differences, speed limit, Arkansas, Mississippi, accident duration

ACM Reference Format:
Jordan G. Taqi-Eddin, Peter Flo Grinde-Hollevik, Chase Hertel, and Ming Senn Teo. 2024. DATA C102 SP24 Final Project Team 42. 1, 1 (May 2024), 25 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 RESEARCH QUESTIONS & BACKGROUND

1.1.1 *Causal inference – Research Question.* What was the causal impact of Arkansas's 5-mile-per-hour speed limit increase on the number of accidents in the state?

1.1.2 *Causal inference – Background.* Legislative action and public safety in the United States has a co-productive relationship. The nation's motorways serve as an arena through which this interconnected dynamic comes to life. Despite casualties at the hands of drunk drivers being abundant, legal safeguards against such tragedies were barren prior to the 1980s. Federal legislators were well informed on the gravity of the situation, as the U.S. Department of Transportation compiled a report on the threat of driving under the influence in 1968. [16] The report explains the danger of "problem drinkers" whose actions are the driving cause behind the accidents in which they were involved. At the local level, checks were put in place to crack down on driving while intoxicated. [9] However without any national standard, these measures and their enforcement varied from community to community.

By the beginning of the 1980s, yearly alcohol-related traffic fatalities stood at approximately 30,000. [5] One of these unfortunate individuals was the daughter of a woman named Candy Lightner, who was a co-founder of the organization Mothers Against Drunk Driving (MADD). [9] Through protests and public service announcements, Lightner and MADD raised public awareness about the nation's DUI epidemic. Furthermore, these efforts helped get the federal government to enforce DUI regulations in states via incentive mechanisms. Only five years after the creation of MADD in 1980, deaths caused by intoxicated motorists dropped to around 17,000. [8]

DUI enforcement is demonstrative of the disjointed and impactful nature of American traffic laws. However, legislators affect road safety via various other avenues, such as speed limits. Like drunk driving statutes, enforcement of speed

Authors' addresses: Jordan G. Taqi-Eddin, jgte29@berkeley.edu, University of California - Berkeley, 2195 Hearst Ave, Berkeley, California, USA, 94720-1786; Peter Flo Grinde-Hollevik, hollevik@berkeley.edu, University of California - Berkeley, 2195 Hearst Ave, Berkeley, California, USA, 94720-1786; Chase Hertel, chasehertel@berkeley.edu, University of California - Berkeley, 2195 Hearst Ave, Berkeley, California, USA, 94720-1786; Ming Senn Teo, mingsennteo@berkeley.edu, University of California - Berkeley, 2195 Hearst Ave, Berkeley, California, USA, 94720-1786.

limits was historically left up to states' discretion. [17] It was not until the 1970s that the federal government intervened by setting the national speed limit at 55 miles per hour, as a reaction to the OPEC oil embargo. [6] The National Maximum Speed Law of 1974 had an immediate impact on road safety as fatalities dropped from 54,052 to 45,196 in the year after its passing.

However, this was followed by a string of gradual concessions that returned a large part of the autonomy that states had lost with the maximum speed limit standardization. In the 1980s, states were permitted to raise the upper bound of their speed limits to 65 miles per hour with the Surface Transportation and Uniform Relocation Assistance Act of 1987. Then in the ensuing decade, Congress reverted to the pre-1970s standard with the National Highway Designation Act of 1995, giving states complete agency over their traffic laws once more.

Although only 41 states increased their limits in response to the enactment of the 1987 law, all have since raised the limits on both urban interstates and non-interstate roads following the federal legislation eight years later. Today, different limits reflect the heterogeneous natures of the states in which they are enforced, with maximum speed limits ranging from 60 miles per hour in Hawaii to 85 miles per hour on select Texas interstates. [13] [1]

The aftermath of the National Maximum Speed Law demonstrated the clear relationship between speed limit laws and road safety. Additionally, the work of *Friedman et al., 2009* further strengthened this link by concluding that there was an uptick in deaths (17 percent) after control over speed limit laws was returned to state jurisdictions. [6] During the summer of 2020, a new law went into effect in Arkansas that increased the existing maximum speed limit to 75. [12] The insights provided by prior literature and crash data indicate that such an increase could have potentially made Arkansas motorways more dangerous. This notion is supported by the fact that yearly accidents in the state skyrocketed from 991 in 2019 to 3,090 in 2020. [11] Furthermore, in 2021 — the first complete year during which the law, Act 784, was in effect — the crash tally rose even higher to 7,384. Conversely, Arkansas's neighbor to the southeast, Mississippi, did not see such a stark jump in accidents in 2020 or 2021. While the state did have a considerably higher amount of crashes during the aforementioned period, rising from 1984 in 2019 to 3441 in 2021, this increase is nowhere near as severe as that in Arkansas. The two states share many geographic and demographic affinities, as they are located right next to one another and have similar populations — Arkansas: 3,014,348, Mississippi: 2,958,409 (2020 U.S. Census Bureau Population Estimates). [15] However, unlike Arkansas, Mississippi did not follow suit and raise its speed limit to 75 miles per hour, and have a maximum speed limit set at 70. [14] Therefore, the similarities between the two states and the fact that Mississippi's speed limit is at the same level as Arkansas's before Act 784 went into effect gives the unique opportunity to determine the causal effect that the limit increase had on state road safety and is the central point of interest of our empirical investigation.

1.2.1 *Prediction with GLMs and nonparametric methods – Research Question.* What is the impact of road conditions on the duration of traffic accidents, and how does this impact vary across different cities?

1.1.2 *Prediction with GLMs and nonparametric methods – Background.* When a traffic accident occurs, what's generally recommended is parties involved (including witnesses) pull to the side and sort out the situation, either calling law enforcement or exchanging insurance information and filing a claim. This isn't always feasible though, as there are many sophisticated factors that play into a car crash. Our research question wants to dive deeper into these complexities to analyze what specific conditions make an accident last for longer and delay others on the road. Because this won't necessarily be the same in every area, we selected cities where traffic was a significant issue.

For city planning departments, some real-life scenarios where they could implement local policies based on the conclusions drawn from our analysis of the dataset could revolve around issuing different kinds of warnings for drivers. Commissioners could ideate on different ways of notifying residents that an accident has occurred, or is likely to

occur based on the various weather or irregular conditions. They could potentially partner with navigation software services such as Google or Apple Maps to issue localized notifications to residents with the app open within a specified geographic region. As well, there are AMBER Alert type notifications that can be sent out, or temporary signage posted to alert drivers on major thoroughfares. Some use cases that this model can bring about are:

- How often roads should be maintained - infrastructure investment
- Improving emergency response - cleaning up after the accident
- How roads should be monitored - safety legislation of speed limits

2 DATA

In our project, we used the US National Traffic Accident dataset. [11] The dataset is from Kaggle. It's neither a sample or a census. Each row in the dataset represents an accident with its severity, the conditions that the accident occurred in, and where it happened. According to the description of the dataset on the website, the data is collected via multiple APIs that provide streaming traffic incident (or event) data. This might result in a selection bias as the data collected could skew towards more severe accidents that are more likely to be reported. It also could be a convenient sample due to the likelihood of reporting more urban accidents than rural accidents is much higher because of its proximity to the local traffic accident reporters. Given that some of the data from columns like Start_Time and End_Time is usually collected manually, this might result in some measurement errors and we can have some unreliable data. We wish to have multiple columns for various reasons. First would be the insurance involved for the particular accident. Second would be some information on the vehicle that was in the accident like the model and the manufacturer of the vehicle, which can provide us an idea of potential faults from the manufacturer. Third would be the health condition of the driver, which can give us more information on whether they were in control or not. Fourth would be the quality of the road or the maintenance record, so we have an idea of whether the road requires maintenance or not. Lastly, damages to surroundings. This can provide more information on the severity score. In addition to our accident data, we also got county level population estimates from the U.S. Census Bureau. [15] Such estimates were especially helpful in the causal inference sections of our project.

3 EDA AND DATA PREPROCESSING

3.1.1 *Causal inference – Temporal Aggregation.* For our empirical analysis, we aggregated the data in two different ways. When implementing our state-level difference in differences estimator, we aggregated the data monthly, but when refining the granularity of our data down to the county level, we decided to aggregate the data yearly. For all of the fixed effect variables in our data (e.g. "Temperature(F)" or "Junction"), we took their averages when performing our state-by-month and county-by-year aggregations. This means that for binary variables, such as all of our location-based fixed effects, the average returned is a proportion of the accidents that have a one for such a feature. For example, if the mean for the "Roundabout" feature for Arkansas during August 2020 was 0.5, this would mean that half of the accidents that occurred in the state during this month happened at a roundabout.

3.1.2 *Causal inference – Feature Engineering.* There were not many features that we had to manually create for our analyses, as the vast majority of the controls were generated during the temporal aggregation process. The main features that we had to engineer were our treatment and control variables, which respectively were just binary indicators for whether or not a state was Arkansas or the time period of the observation past the enactment date of Act 784, July 1, 2020. To reflect the fact that the law was not in effect for the full year, we set the binary variable in 2020 to be 0.5 for

our county-level investigations since the observations were aggregated yearly. Such a feature engineering technique was similarly applied in *Deschenes et al., 2017.* [4] By multiplying these two features we were able to get our difference in differences feature, *Limit Raised* × *Arkansas*. In addition to the controls created through temporal aggregation and feature engineering, we obtained the population feature for our model by getting the U.S. Census Bureau population estimate during a given year for the state or county of an observation (depending on the scale of the examination). Due to limitations imposed by the temporal granularity of our data, we use the same state and county population estimates for all observations during the same year.

3.2.1 *Prediction with GLMs and nonparametric methods.* In Figure B1, certain trends were observed related to the presence of traffic signals at accident sites and the duration of these accidents. When an accident occurred near a traffic signal, which typically marks the intersection of multiple thoroughfares, the median duration of the accident was approximately 7000 seconds. In contrast, accidents that did not occur near a traffic signal had a shorter median duration, around 5000 seconds. Additionally, accidents at traffic signals exhibited larger outliers in duration, reaching up to 16000 seconds, compared to those without traffic signals, which had outliers up to 14000 seconds. Furthermore, the interquartile range (IQR) for accidents with a traffic signal was wider below the median, suggesting higher variance in durations, whereas for accidents without a traffic signal, the IQR was wider above the median, indicating a different pattern of variance. These observations suggest areas for further investigation into how traffic signals influence accident durations. The visualizations provided are relevant to these research questions as they illustrate the increased duration of accidents near traffic signals, suggesting that the presence of traffic signals at high congestion areas significantly impacts the duration of accidents. This finding motivates the question of how traffic signal-induced congestion affects accident severity and response times, potentially offering a direction for future research and interventions.

In figure B2, across the five cities analyzed, the time of day was observed to play a significant role in the duration of traffic accidents. The median duration of accidents remained consistent at around 6000 seconds, regardless of the time of day. However, during daytime, there was a noticeably higher interquartile range (IQR) above the median, indicating higher variance in accident durations compared to nighttime. Nighttime accidents were fairly evenly split in terms of duration around the median. Notably, the daytime also saw larger outliers in accident duration.

The visualizations of accident durations across different times of day are crucial to the research questions posed, as they highlight the impact of varying traffic conditions throughout the day. By showing that daytime accidents tend to last longer and are more variable in duration, these visualizations suggest that the higher volume of vehicles during the day may contribute to more complex accident scenarios and potentially longer resolution times. This observation motivates further questions regarding how traffic density and visibility at different times of day influence accident outcomes, offering potential avenues for targeted traffic management and safety measures.

In figure B3, cross the analysis of the five cities with the highest incidence of traffic accidents, notable trends and variances in accident duration were observed relative to the presence of traffic signals. Generally, intersections without traffic signals exhibited a higher median duration of accidents compared to those with signals, along with significantly higher outliers, except in Miami and Orlando. Specifically, in Charlotte, the median accident duration was 5000 seconds without traffic signals and 3000 seconds with them, with a notably large interquartile range (IQR) for accidents occurring without traffic signals and outliers reaching close to 15000 seconds. Houston mirrored this trend, showing a large IQR and similar outlier durations, with median times of 4750 seconds without and 4000 seconds with traffic signals. Los Angeles displayed the largest outliers, approaching 20000 seconds, and substantial variance in both scenarios. Miami and Orlando, however, exhibited more even distributions of outliers near 15000 seconds and large IQRs in both

conditions, with median durations of 7500 seconds without and 5500 seconds with traffic signals in Miami, and 7500 seconds without and 5000 seconds with in Orlando.

In figure B4, cross the analysis of the five cities with the highest incidence of traffic accidents, notable trends and variances in accident duration were observed relative to severity levels. We observed that every city has a different trend for accident durations that has a severity level of 2 and 4. For severity level of 2, the most notable one is LA, where the median accident duration was roughly 7500 seconds, has a large interquartile range (IQR) and outliers reaching close to 20000 seconds. For the rest of the cities, the IQR are smaller and has a smaller outlier range. For severity level of 4, Houston mirrored this trend, showing a large IQR and similar but slightly larger outlier durations, with median times of 7500 seconds. For the rest of the cities, the IQR are smaller and has a smaller outlier range.

In figure B5, cross the analysis of the five cities with the highest incidence of traffic accidents, notable trends and variances in accident duration were observed relative to the day of the accident occurred on. Generally, most cities observe similar trend on weekdays where the median accident remain roughly the same for their respective cities. However, in LA, we noticed that there are variations amongst the median accident durations for different days. In Charlotte, we also noticed that there are larger median accident duration for weekends.

The visualizations are particularly relevant to the research questions as they underscore city-specific differences in traffic accident durations, influenced by the presence or absence of traffic signals. Highlighting how outlier durations and IQR ranges vary significantly across cities demonstrates the importance of localized traffic management strategies. These insights motivate further questions about the effectiveness of traffic signals in mitigating accident durations and the potential need for tailored approaches in traffic management and safety protocols in cities with high accident rates. This analysis not only illustrates how accident durations can vary significantly across different urban settings but also provides a basis for investigating the influence of urban infrastructure on traffic safety.

3.2.2 *Prediction with GLMs and nonparametric methods - Feature Engineering.* We decided to work with only the top 5 cities with the most accidents after the pandemic lockdown (any accidents occurring past the year 2020). We filtered the main dataset and converted them into five main cities datasets that we combined for our analyses and model prediction. We created two main features for our analyses and prediction, which are the columns Day and ETA. Day is a categorical column that contains the day that the accident occurred on (weekday/weekend). ETA is a categorical column that describes the level of delay that accidents caused. We set the level of delays that makes the most sense to us on how long and bad the traffic delay can be.

4 CAUSAL INFERENCE — LITERATURE REVIEW

There is extensive literature that has investigated the causal impact of speed limit laws on road safety. In their 2011 paper, *Friedman et al* use a Poisson mixed-regression model to examine the consequence of the 1995 National Highway Designation Act. [6] They look at the monthly tallies of traffic fatalities and injuries during the twenty years following the passing of the legislation. Additionally, to account for confounding covariates, the researchers stratify states into road types and speed limit subgroups. Our research strategy has many similarities with the *Freidman* paper, especially when it comes to the temporal and spatial aggregation schemes. However, we chose to model the monthly accident counts using a negative binomial distribution instead of a Poisson to guard against interspersion, as detailed in *Coxe et al., 2009.* [3]

Additional research design discrepancies between *Friedman* and our empirical investigation arise because of our decision to protect against confounding via a natural experiment. The quasi-experimental method that we chose was a difference-in-differences model, very similar to the one employed in *Card & Krueger, 1993.* [2] In this paper, a DiD

estimator was used to investigate the impact of New Jersey's new minimum wage law on employment, with the control in this model being the neighboring state of Pennsylvania. We take a similar approach to the construction of our difference-in-difference model, by choosing Mississippi as the control for our analysis of the impact of the speed limit increase in Arkansas. Furthermore, we were unable to use a DiD model to examine the impact of the speed limit increase on accident severity as the parallel trend assumption, a key requirement for this type of estimator, did not hold, as seen in the "Monthly Mean Severities" plot of *Appendix A*.

When comparing accident counts at the county level we use a kernel-based matching mechanism similar to that in *Heckman et al., 1998.* [7] Furthermore, when conducting our chi-squared test of independence using contingency tables, we do not make an assumption of normality within the underlying distributions. This is because as the "County Yearly Accident Totals Distributions" in *Appendix A* show, trying to make such a claim would not be consistent with the trends demonstrated in the data, and could severely impact the reliability of our conclusions. However, this is not an issue for us because in *McHugh, 2013* it is stated that for distributions where the data is "seriously skewed or kurtotic," the chi-squared test can be performed without making such an assumption. [10]

5 CAUSAL INFERENCE — STATE-LEVEL REGRESSION MODEL

The regression model for analyzing the effect of legislation and environmental factors on traffic accidents at the state level is defined as follows:

$$Y_{s,t,y} = \beta_0(\text{Limit Raised}_{t,y} \times \text{Arkansas}_s) + \beta_1(\text{Population}_{s,t}) + \lambda_{s,t,y} + \kappa_{s,t,y} + \epsilon_{s,t,y} \tag{1}$$

Where:

- $Y_{s,t,y}$: Total number of accidents in state s during time-period t and year y.
- $\text{Limit Raised}_{t,y}$: Binary indicator for whether the Arkansas Law, Act 784, had gone into effect (i.e., any time after July 1, 2020).
- Arkansas_s: Binary indicator for whether state s is Arkansas.
- $\text{Population}_{s,t}$: Log population estimate of state s during year y.
- $\lambda_{s,t,y}$: Vector of location-based fixed effects for state s during time-period t and year y (e.g., Traffic Signal, Junction, Crossing).
- $\kappa_{s,t,y}$: Vector of weather-based fixed effects for state s during time-period t and year y (e.g., Temperature, Visibility, Wind Speed).
- $\epsilon_{s,t,y}$: Error term.

Note: To see the Causal DAG for the model, refer to *Appendix A*.

Additionally, the distribution of accidents is modeled as:

$$Y_{c,t,y} \mid \beta \sim \text{NegBin}(\exp(\mathbf{x}_i^T \beta)) \tag{2}$$

INTERPRETATION OF RESULTS

The analysis focused on evaluating the impact of Arkansas's Act 784, which increased the speed limit from 70 mph to 75 mph, on traffic accidents across the state. The legislation's intent was likely aimed at improving traffic flow and reducing travel times. However, our findings suggest an unintended consequence of this policy change.

Table 1. Regression Coefficients for β_0 across State-Level Model

Model Description	Coefficient	Std. Err.	P>|z|
No FEs	1.3450	0.211	0.000
No FEs, MA_Smoothed	1.2968	0.211	0.000
Location FEs	1.4048	0.228	0.000
Location FEs, MA_Smoothed	1.3454	0.228	0.000
Weather FEs	1.4162	0.297	0.000
Weather FEs, MA_Smoothed	1.4092	0.297	0.000
Location and Weather FE	1.4110	0.320	0.000
Location & Weather FEs, MA_Smoothed	1.3575	0.320	0.000

Summary of Findings

The regression coefficients for the interaction term, 'Limit Raised × Arkansas', across various model configurations were consistently positive and statistically significant (Table 1). This indicates a robust increase in traffic accidents following the enactment of Act 784. The increase in speed limit appears to correlate with an increase in accidents, which is a common outcome in traffic safety studies where higher speeds tend to increase both the likelihood and severity of accidents.

Detailed Interpretation

This section delves into the implications of the observed coefficients, enhancing the understanding of how different models captured the impact of increasing the speed limit from 70 mph to 75 mph under Act 784.

- **Basic and Smoothed Models**:
  - **No Fixed Effects**: The model without fixed effects yielded a coefficient of $\beta_0 = 1.3450$ (Std. Err. = 0.211, $p < 0.001$), indicating a significant increase in accidents immediately following the policy change.
  - **No Fixed Effects, MA_Smoothed**: Adjusting for smoothing, the impact remained substantial with a coefficient of $\beta_0 = 1.2968$ (Std. Err. = 0.211, $p < 0.001$), reaffirming the initial finding.
- **Location and Weather Fixed Effects**:
  - **Location Fixed Effects**: Incorporation of geographical fixed effects resulted in a coefficient of $\beta_0 = 1.4048$ (Std. Err. = 0.228, $p < 0.001$), suggesting that local geographical factors did not mitigate the increased accident risk.
  - **Weather Fixed Effects**: Accounting for weather conditions, the coefficient was $\beta_0 = 1.4162$ (Std. Err. = 0.297, $p < 0.001$), indicating that meteorological variations also did not influence the accident increase.
- **Combined Fixed Effects Models**:
  - **Location and Weather Fixed Effects**: The most comprehensive model showed a coefficient of $\beta_0 = 1.4110$ (Std. Err. = 0.320, $p < 0.001$), highlighting that even when controlling for both location and weather, the effect of the increased speed limit on accident rates was pronounced and significant.
  - **Location & Weather FEs, MA_Smoothed**: Similar findings were observed with smoothing, where the coefficient was $\beta_0 = 1.3575$ (Std. Err. = 0.320, $p < 0.001$), confirming the robustness of the results across different model specifications.

6 CAUSAL INFERENCE — COUNTY-LEVEL REGRESSION MODEL

The regression model for analyzing the impact of specific legislation and environmental conditions on traffic accidents at the county level is described by the following equation:

$$Y_{c,t,y} = \beta_0(\text{Limit Raised}_{t,y} \times \text{Arkansas}_c) + \beta_1(\text{Population}_{c,t}) + \lambda_{c,t,y} + \kappa_{c,t,y} + \epsilon_{c,t,y} \tag{3}$$

Where:

- $Y_{c,t,y}$: Total number of accidents in county c during time-period t and year y.
- $\text{Limit Raised}_{t,y}$: Binary indicator for whether the Arkansas Law, Act 784, had gone into effect (i.e., any time after July 1, 2020).
- Arkansas_c: Binary indicator for whether county c is in Arkansas.
- $\text{Population}_{c,t}$: Log population estimate of county c during year y.
- $\lambda_{c,t,y}$: Vector of location-based fixed effects for county c during time-period t and year y (e.g., Traffic Signal, Junction, Crossing).
- $\kappa_{c,t,y}$: Vector of weather-based fixed effects for county c during time-period t and year y (e.g., Temperature, Visibility, Wind Speed).
- $\epsilon_{c,t,y}$: Error term.

Note: To see the Causal DAG for the model, refer to *Appendix A*.

Additionally, the distribution of accidents is modeled as:

$$Y_{c,t,y} \mid \beta \sim \text{NegBin}(\exp(\mathbf{x}_i^T \beta)) \tag{4}$$

Table 2. Regression Coefficients for β_0 across County-Level Model

Model Description	Coefficient	Std. Err.	P>|z|
No F.E.	2.6026	0.486	0.000
No F.E., MA_Smoothed	2.5242	0.485	0.000
Location FEs	2.6527	0.530	0.000
Location FEs, MA_Smoothed	2.5726	0.530	0.000
Weather FEs	2.3144	0.620	0.000
Weather FEs, MA_Smoothed	2.1744	0.620	0.000
Location & Weather FEs	2.1715	0.704	0.002
Location & Weather FEs, MA_Smoothed	2.0812	0.704	0.003

Detailed Interpretation of County-Level Model

This section provides a detailed analysis of the regression outcomes at the county level, where Act 784, increasing the speed limit from 70 mph to 75 mph, was also studied for its impact on traffic accidents. Each model's findings are elucidated to interpret how local variations influence the effect of the speed limit increase.

- **Basic and Smoothed Models**:
  - **No Fixed Effects**: The model without any fixed effects reported a significant increase in accidents with a coefficient of $\beta_0 = 2.6026$ (Std. Err. = 0.486, $p < 0.001$), indicating that the immediate effect of the law was a substantial increase in traffic incidents.

  - **No Fixed Effects, MA_Smoothed**: The smoothed model, which helps control for variability, also showed a pronounced effect with a coefficient of $\beta_0 = 2.5242$ (Std. Err. = 0.485, $p < 0.001$).
- **Fixed Effects Models**:
  - **Location Fixed Effects**: Including location-based fixed effects such as traffic signals and junctions resulted in a coefficient of $\beta_0 = 2.6527$ (Std. Err. = 0.530, $p < 0.001$), suggesting that even when accounting for these factors, the increase in accidents remained significant.
  - **Weather Fixed Effects**: When weather conditions were considered, the model produced a coefficient of $\beta_0 = 2.3144$ (Std. Err. = 0.620, $p < 0.001$), still indicating a significant increase in accidents.
- **Comprehensive Fixed Effects Models**:
  - **Location and Weather Fixed Effects**: Combining both sets of fixed effects, the coefficient slightly decreased but remained significant at $\beta_0 = 2.1715$ (Std. Err. = 0.704, $p = 0.002$).
  - **Location & Weather FEs, MA_Smoothed**: The most controlled model further smoothed showed a coefficient of $\beta_0 = 2.0812$ (Std. Err. = 0.704, $p = 0.003$), still supporting the trend seen in simpler models.

These results underline a consistent increase in traffic accidents across all county-level models following the increase in speed limits enacted by Act 784. The uniformity of the impact across different settings and despite various controls strongly points to the direct influence of the policy change on increasing accident risks.

Chi-Squared Test of Independence for Yearly Crash Concentrations

In order to assess the distribution of crash counts in states before and after the imposition of Act 784, a chi-squared test of independence was performed. This test aimed to determine if there were significant changes in the distribution of crash hotspots by comparing yearly aggregated crash counts for 2019 and 2021. The year 2019 represents the last full year before the enactment of Act 784, and 2021 the first full year afterward.

Analysis Methodology. Crash counts were normalized by the total number of crashes in the respective state and year to form a concentration measure:

$$\text{Concentration}_{c,s,y} = \frac{\text{Crash Count}_{c,s,y}}{\text{Total Crashes}_{s,y}} \quad \forall \text{County } c, \text{ in state } s, \text{ during year } y$$

This concentration measure helps in assessing changes in crash distributions irrespective of the total number of crashes. The normalization ensures that the analysis accounts for any general increases or decreases in traffic volumes or reporting practices.

Statistical Test and Results. The assumptions of the chi-squared test, as detailed by McHugh (2013), indicate that the normality of the distribution of crash counts does not affect the validity of the test results. The following results were obtained from the chi-squared test:

- Chi-squared statistic: 0.0
- P-value: 1.0
- Degrees of freedom: 1

Interpretation. The extremely high p-value and a chi-squared statistic of 0 indicate that there is no statistically significant difference in the distribution of crash concentrations between the two years examined. This finding suggests that the increase in crash counts observed in Arkansas was not due to the formation of new crash hotspots but was a statewide

phenomenon. The results imply that the increase in crashes was uniformly distributed across the state, rather than being concentrated in specific areas.

This uniform increase highlights the broader impact of Act 784 across the entire state of Arkansas, underscoring the importance of considering statewide traffic safety measures and interventions in response to changes in speed limit laws.

7 PREDICTION WITH GLMS AND NONPARAMETRIC METHODS — ACCIDENT DURATION FORECASTING

The GLM model for predicting accident duration based on road conditions, time factors, and cities is defined as follows:

$$Y = \beta_0(\text{Severity}) + \beta_1(\text{Day}) + \beta_2(\text{City}) + \beta_3(\text{Time of Day}) + \beta_4(\text{Road Conditions}) + \epsilon \tag{5}$$

Where:

- Y: The level of traffic delay caused by the accident (categorized into five levels based on the length of the accident duration).
- Severity: The level of severity caused by the accident.
- Day: The day of the accident occurred on.
- City: The city which the accident occurred in.
- Time of Day: The time of day which the accident occurred on (Day or Night).
- Road Conditions: Vector of location-based fixed effects (e.g., Amenity, Traffic Signal, Junction, Crossing).
- ϵ: Error term.

GLM JUSTIFICATION AND ASSUMPTIONS

We used the Poisson GLM with a log link function which ensures that the prediction is positive as the ETA(categorized accident duration) cannot be negative. One of the assumptions we have is that after applying the inverse link function that there will be a linear relationship between the response variable and the explanatory variables. There should not be any multicollinearity between the features.

NONPARAMETRIC JUSTIFICATION AND ASSUMPTIONS

We used Random Forest for our nonparametric method as a large portion of our features are categorical, which means that decisions trees will work well and split on these variables. It's good for classification and it's better over decision trees as it reduces overfitting and variance. Our assumption is that there is no missing value from the data provided we removed them during our data cleaning stage.

INTERPRETATION OF RESULTS

The analysis focused on evaluating the impact of severity, road conditions, and variations across the top 5 cities on accident duration. Using the 95% confidence interval and p-value for the coefficients, we will determine whether a feature is significant for our accident duration prediction. As for our model performance evaluation, we split the dataset into a training and test and calculated the RMSE and accuracy of each model and compared them together to get a

sense of how well the models fit the data and generalizes for unseen data. The metrics we used are a 0.5 cutoff for both models in terms of accuracy.

Figure 1: Frequentist GLM Training Result

Generalized Linear Model Regression Results

Dep. Variable:	ETA	No. Observations:	205954
Model:	GLM	Df Residuals:	205927
Model Family:	Poisson	Df Model:	26
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3.2037e+05
Date:	Sun, 05 May 2024	Deviance:	66570.
Time:	23:22:41	Pearson chi2:	5.70e+04
No. Iterations:	100	Pseudo R-squ. (CS):	0.02324
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.1058	0.248	4.451	0.000	0.619	1.593
Amenity[T.True]	0.0258	0.011	2.335	0.020	0.004	0.047
Bump[T.True]	0.0638	0.060	1.060	0.289	-0.054	0.182
Crossing[T.True]	0.0292	0.004	7.828	0.000	0.022	0.037
Give_Way[T.True]	-0.0168	0.021	-0.798	0.425	-0.058	0.024
Junction[T.True]	-0.0563	0.007	-7.655	0.000	-0.071	-0.042
No_Exit[T.True]	-0.0051	0.018	-0.286	0.775	-0.040	0.030
Railway[T.True]	-0.0419	0.013	-3.267	0.001	-0.067	-0.017
Roundabout[T.True]	0.1169	0.165	0.710	0.478	-0.206	0.439
Station[T.True]	0.0546	0.005	11.203	0.000	0.045	0.064
Stop[T.True]	-0.0006	0.007	-0.092	0.927	-0.014	0.013
Traffic_Calming[T.True]	-0.0052	0.038	-0.136	0.891	-0.079	0.069
Traffic_Signal[T.True]	-0.0793	0.004	-18.644	0.000	-0.088	-0.071
Turning_Loop[T.True]	2.069e-15	1.15e-15	1.796	0.072	-1.89e-16	4.33e-15
Severity	-0.1579	0.005	-30.330	0.000	-0.168	-0.148
City_Charlotte	0.1470	0.050	2.951	0.003	0.049	0.245
City_Houston	0.2140	0.050	4.299	0.000	0.116	0.312
City_LA	0.2158	0.050	4.333	0.000	0.118	0.313
City_Miami	0.2742	0.050	5.512	0.000	0.177	0.372
City_Orlando	0.2549	0.050	5.121	0.000	0.157	0.352
Day_Friday	0.1562	0.036	4.385	0.000	0.086	0.226
Day_Monday	0.1513	0.036	4.247	0.000	0.081	0.221
Day_Saturday	0.1856	0.036	5.199	0.000	0.116	0.256
Day_Sunday	0.1927	0.036	5.378	0.000	0.122	0.263
Day_Thursday	0.1373	0.036	3.853	0.000	0.067	0.207
Day_Tuesday	0.1446	0.036	4.056	0.000	0.075	0.214
Day_Wednesday	0.1381	0.036	3.876	0.000	0.068	0.208
Sunrise_Sunset_Day	-0.2460	0.333	-0.738	0.461	-0.899	0.407
Sunrise_Sunset_Night	-0.1751	0.333	-0.525	0.599	-0.829	0.478
Distance	0.0410	0.001	35.499	0.000	0.039	0.043

Summary of Findings

The regression coefficients for some of the road conditions terms, along with severity, days, cities, and distance are statistically significant (Table 3). It seems that the variation for the time of day which the accident occurred on doesn't correlate well with determining the level of delay that accidents caused.

Detailed Interpretation of GLM

According to the result shown above, Amenity, Crossing, Junction, Railway, Station, Traffic_Signal, Severity, Distance, Days, and Cities are significant features for predicting accident duration as their coefficient p-values are below the 0.05 threshold. For the rest of the features, all of the confidence intervals contain 0 and they are not statistically significant due to their p-value being above the 0.05 threshold. Amenity, Crossing, Junction, Railway, Station, Traffic_Signal, Severity, Distance, Days, and Cities serve as important features that would play a major role in predicting accident duration.

Using the 95% confidence interval for every coefficient listed above, we are able to interpret the uncertainty of the model. Ex: For every increase in Distance, we are 95% confident that it will increase the log odds of the level of accident delay by 0.039 and 0.043.

Discussion

Table 3. Model Performance

Model Description	RMSE (Training)	RMSE (Testing)	Accuracy (Training)	Accuracy (Testing)
Frequentist GLM Model	0.9024	0.9049	0.5295	0.5285
Random Forest	0.5931	0.9034	0.8108	0.6378

In this section, we calculated the training RMSE and testing RMSE and the accuracy for both GLM and the random forest model. The Random Forest model performs better than the Frequentist GLM model as it has a higher testing accuracy and a similar testing RMSE to the GLM model. However, we are confident in using the Frequentist GLM model for future datasets as the RMSE for both training and test are relatively small. This indicates that the model works well with unseen data. The deviance of the model is incredibly large, and the log-likelihood doesn't seem that great.

Model Fit

We evaluated how well each model fits the data by their training accuracy. The random forest model best fit the training data based on its accuracy around 81 percent, which beats the training accuracy of the Frequentist GLM model of 53 percent. Provided the RMSE for the GLM is around 0.90 and 0.59 for the Random Forest model, we concluded that they don't fit the data very well as those RMSE are quite large for the range of our dependent variables which is between 0 and 4.

Limitations

The limitations of the GLM model assume that there's no multicollinearity between the features, which is very unlikely as there are always some connections between the road conditions. For the non-parametric model, it's hard to

interpret it. Random Forest causes the outputs to lose its true meaning during the training stage, as it's an ensemble method. It's also hard to foster deeper insights from any analysis as the model is a black box model.

Potential Improvements

To improve our model, we should incorporate more features that can contribute to accident duration. In reality, we would have more information about how the accident occurred. Information such as the vehicle that caused the accident, the numbers of vehicles involved in the accident, and number of deaths can be also main factors that will influence the accident duration. If we have a dataset with more detailed information on the accident that occurred, we might have lower variances and better predictions.

8 CONCLUSION

Summary of Key Findings

For our first research question, we used causal inference. We found that there is a causal relationship between the increase in speed limit and the increase in accidents across the entire state of Arkansas. For our second research question, we used prediction with GLMs and nonparametric methods. We found that Amenity, Crossing, Junction, Railway, Station, Traffic_Signal, Severity, and Distance played a significant role in predicting accident duration.

Generalizability of Results

The analysis of traffic accidents across different cities and days reveals a strong correlation between these factors and accident duration. Additionally, variations in road conditions, such as crossings, railways, traffic signals, and accident severity, played a significant role in influencing accident durations. The distance traveled prior to the accident also emerged as a critical factor for predicting accident duration.

While generalization of these findings is possible, it requires the inclusion of additional features such as vehicle type and the number of vehicles involved to improve prediction accuracy. Relying solely on accident severity provides limited insight due to significant variance in how accidents can occur based on the vehicle itself. Thus, the ability to generalize is present to a small extent but requires more comprehensive data to account for the full range of influencing factors.

Research consistently shows that increasing speed limits leads to a rise in the number of crashes due to higher travel speeds reducing reaction times and increasing the severity of collisions. This causal relationship underscores the need for careful consideration of speed limit policies to prioritize road safety.

Call to Action

Governments should enhance traffic infrastructure, particularly in high-speed limit intersections, by investing in modern traffic systems to minimize accident duration and improve traffic flow. They should also implement data-driven traffic management strategies that utilize traffic accident data to identify high-risk areas and tailor traffic solutions accordingly. Furthermore, strengthening vehicle safety regulations and promoting the adoption of advanced driver-assistance systems (ADAS) will help reduce accident severity and improve overall road safety.

Merging

For our first research question, we merged with the US Census Bureau data to get county-level population estimates. This enabled us to get more granular data to get a better sense of the causal impact of Arkansas's Act 784, which increased the speed limit of 70 to 75 mph on traffic accidents. For our second research question, the dataset has all the information needed so we didn't merge another data source.

Limitations

For our first research question, having the demographics of the drivers in the state and the speed of the vehicle before the accident occurred would help further solidify the results. For our second research question, our dataset lacks certain features that contribute to accident duration. If we have a wider range of features like the number of vehicles involved and the vehicle type, the model would be more generalizable.

Future Studies

For our first research question, we could expand the scale of our causal inference study to a national level scale that encompasses all of the states in the US as in *Friedman et al.* [6]. We could also look at the causal impact of speed limits on the number of deaths using our difference-in-difference framework on a national scale. For our second research question, we could reach a more generalizable conclusion if we had access to a more detailed accident dataset.

REFERENCES

[1] [n. d.]. https://www.txdot.gov/safety/driving-laws/speed-limits/limits.html

[2] David Card and Alan B Krueger. 1993. Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania.

[3] Stefany Coxe, Stephen G West, and Leona S Aiken. 2009. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of personality assessment* 91, 2 (2009), 121–136.

[4] Olivier Deschenes, Michael Greenstone, and Joseph S Shapiro. 2017. Defensive investments and the demand for air quality: Evidence from the NOx budget program. *American Economic Review* 107, 10 (2017), 2958–2989.

[5] James C Fell and Robert B Voas. 2006. Mothers against drunk driving (MADD): the first 25 years. *Traffic injury prevention* 7, 3 (2006), 195–212.

[6] Lee S Friedman, Donald Hedeker, and Elihu D Richter. 2009. Long-term effects of repealing the national maximum speed limit in the United States. *American journal of public health* 99, 9 (2009), 1626–1631.

[7] James J Heckman, Hidehiko Ichimura, and Petra Todd. 1998. Matching as an econometric evaluation estimator. *The review of economic studies* 65, 2 (1998), 261–294.

[8] Barron H Lerner. 2011. Drunk driving, distracted driving, moralism, and public health. *New England journal of medicine* 365, 10 (2011), 879–881.

[9] Patricia Melody Loewit-Phillips. 2013. Mothers against drunk driving (MADD): history and impact. *International Journal of Childbirth Education* 28, 4 (2013), 62.

[10] Mary L McHugh. 2013. The chi-square test of independence. *Biochemia medica* 23, 2 (2013), 143–149.

[11] Sobhan Moosavi. 2023. US Accidents (2016 - 2023) [Dataset]. https://ondefendapp.com/countrywide-automobile-accident-report-pdf

[12] State of Arkansas. 2020. 2020 Arkansas Code, Title 27 - Transportation, Subtitle 4 - Motor Vehicular Traffic, Chapter 51 - Operation Of Vehicles — Rules Of The Road. State Law.

[13] State of Hawaii. 2017. Hawaii Revised Statutes Section 291C-106. https://casetext.com/statute/hawaii-revised-statutes/division-1-government/title-17-motor-and-other-vehicles/chapter-291c-statewide-traffic-code/part-x-speed-restrictions/section-291c-106-speed-limit-for-daniel-k-inouye-highway Speed limit for Daniel K. Inouye Highway.

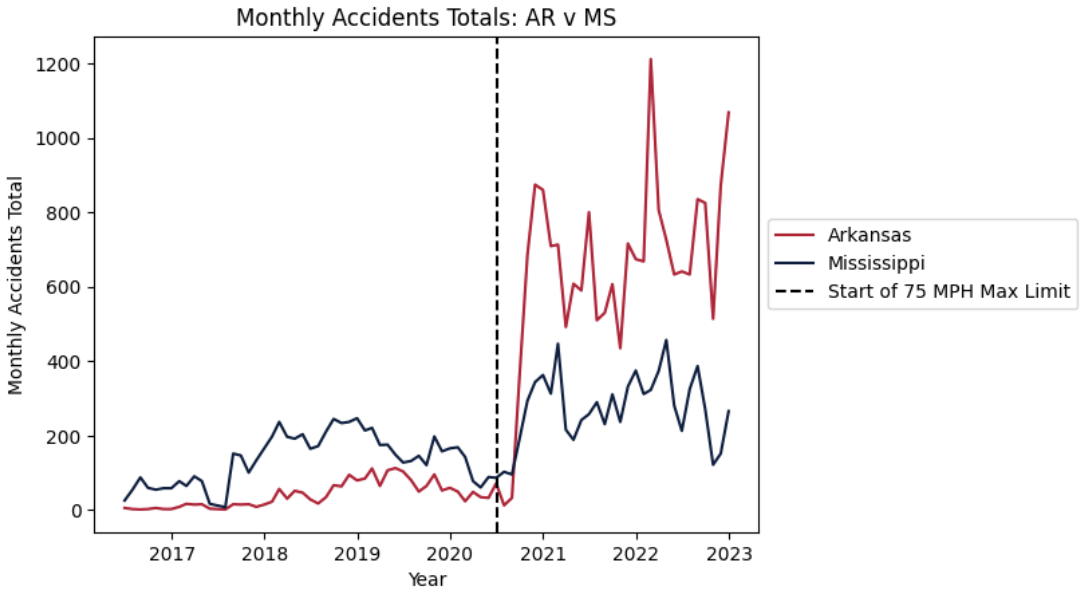
[14] State of Mississippi. 2020. 2020 Mississippi Code, Title 63 - Motor Vehicles and Traffic Regulations, Chapter 3 - Traffic Regulations and Rules of the Road. State Law.

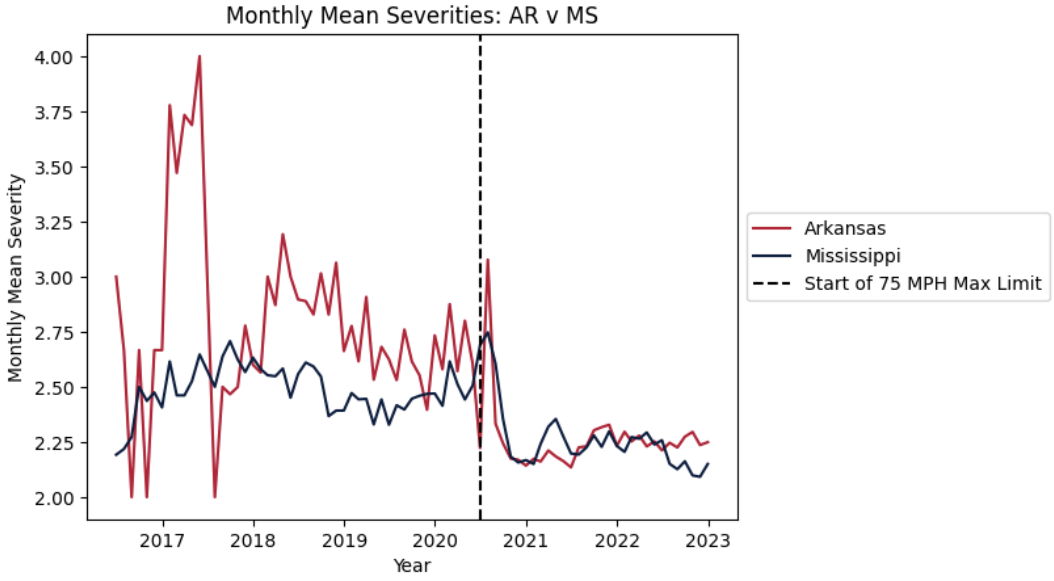
[15] U.S. Census Bureau. 2023. Population Estimates: Housing Unit Estimates for US, States, and Counties [Dataset]. https://catalog.data.gov/dataset/population-estimates-housing-unit-estimates-for-us-states-and-counties.

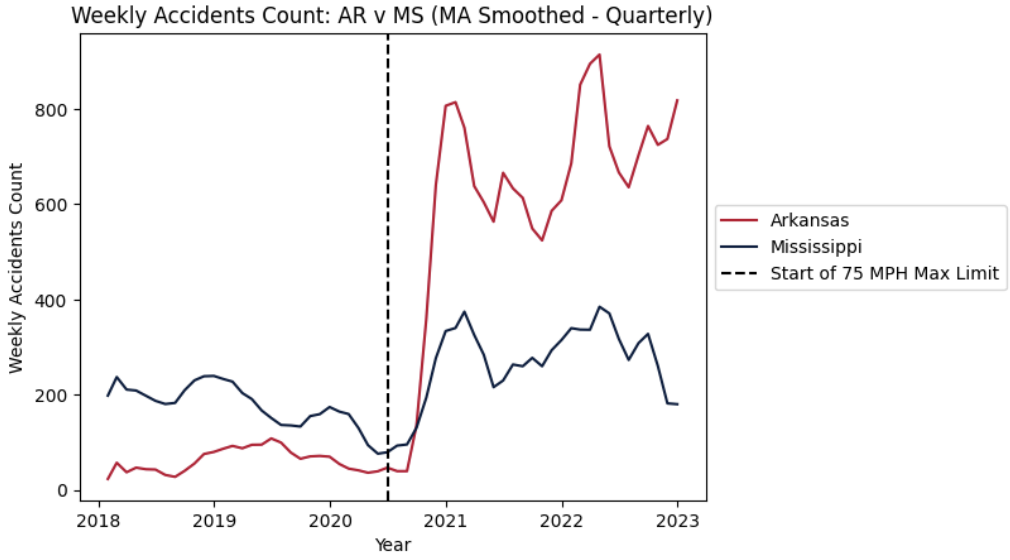
[16] U.S. Department of Transportation. 1968. *1968 Alcohol and Highway Safety.* Technical Report. U. S. Government Printing Office, Washington, DC.

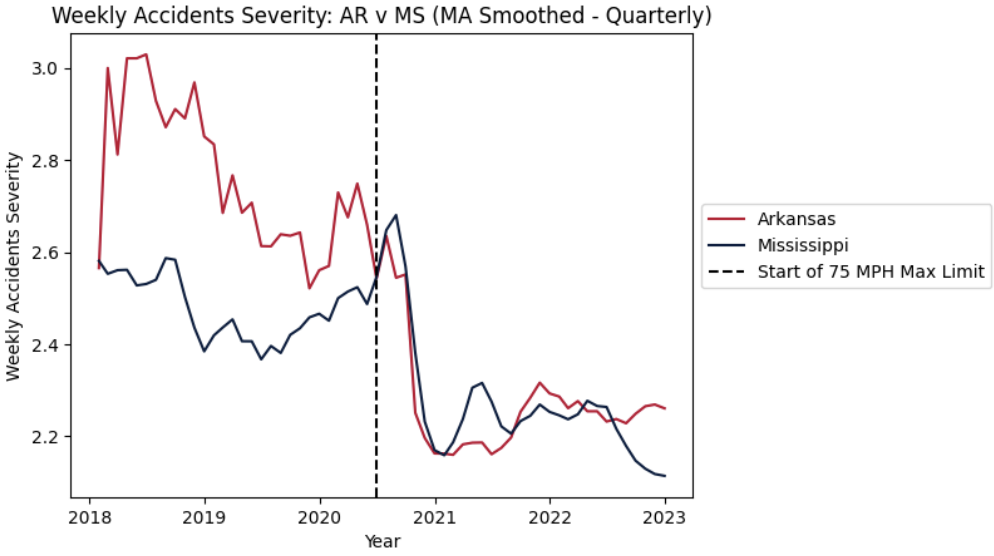
[17] Robert O Yowell. 2005. The evolution and devolution of speed limit law and the effect on fatality rates. *Review of Policy Research* 22, 4 (2005), 501–518.

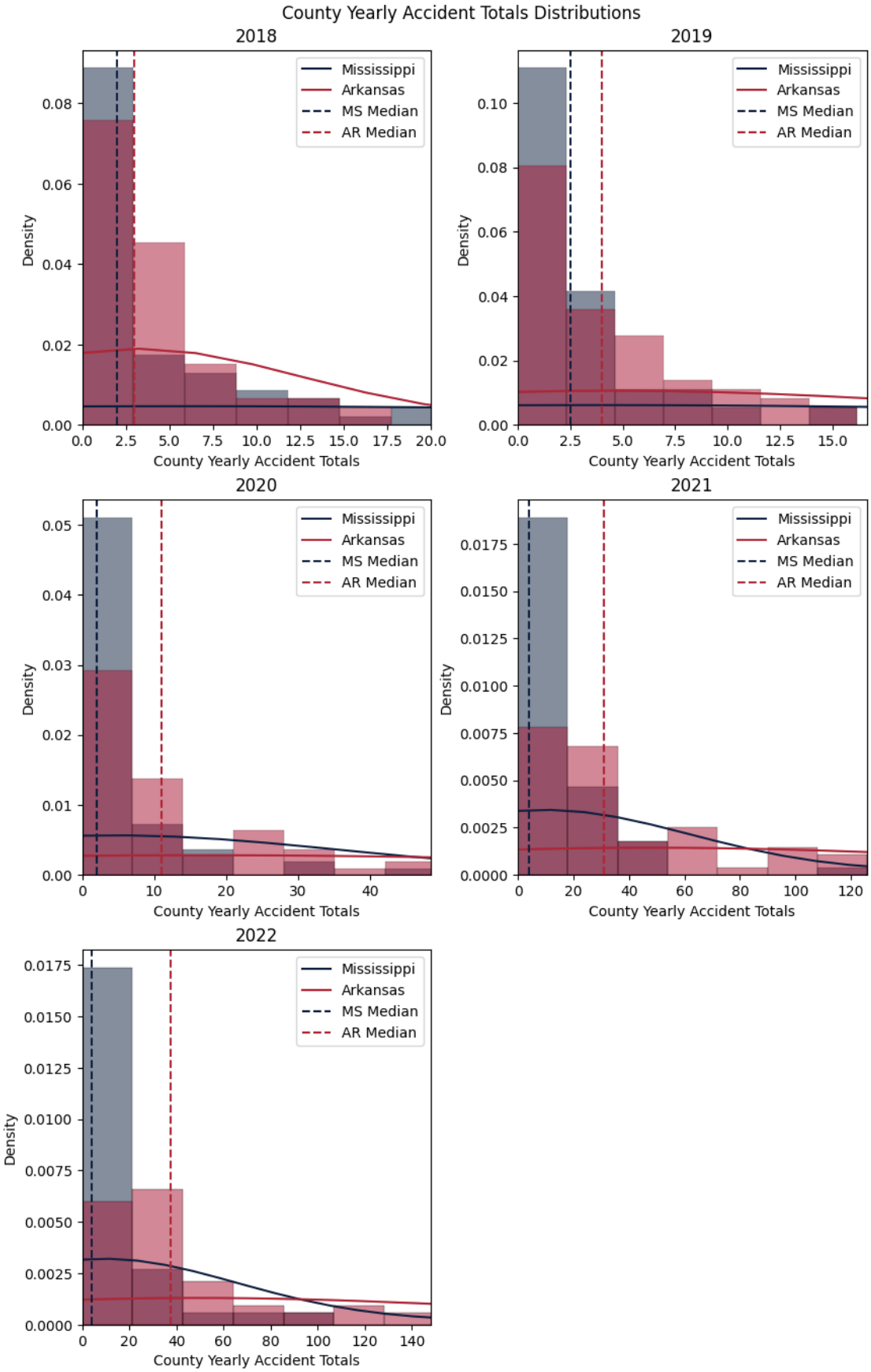
A ACCIDENT DATA VISUALIZATIONS & CAUSAL DAG



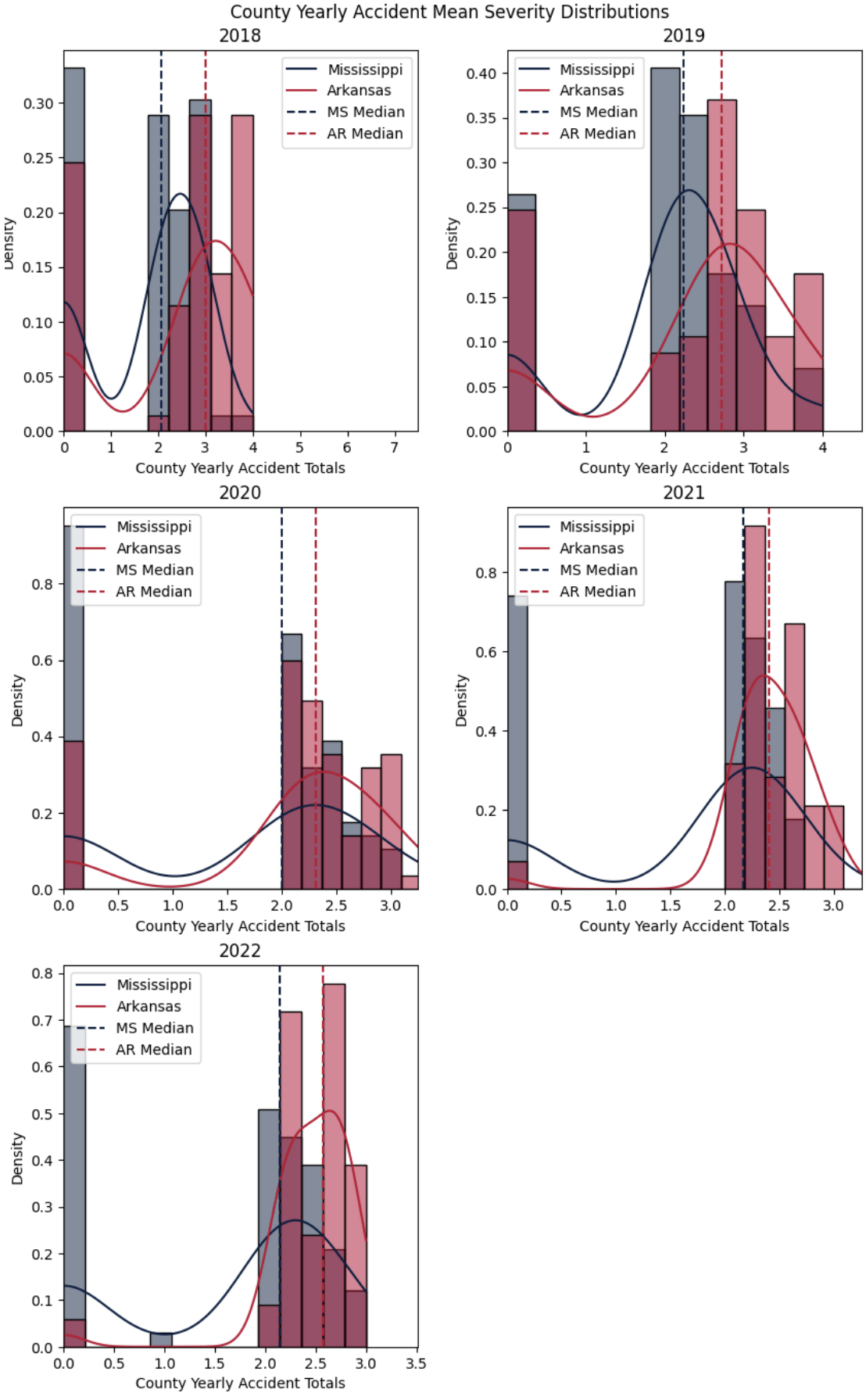



Weekly Accidents Count: AR v MS (MA Smoothed - Quarterly)
Weekly Accidents Count
0
200
400
600
800
2018
2019
2020
2021
2022
2023
Year
Arkansas
Mississippi
Start of 75 MPH Max Limit


Weekly Accidents Severity: AR v MS (MA Smoothed - Quarterly)
Weekly Accidents Severity
2.2
2.4
2.6
2.8
3.0
2018
2019
2020
2021
2022
2023
Year
Arkansas
Mississippi
Start of 75 MPH Max Limit

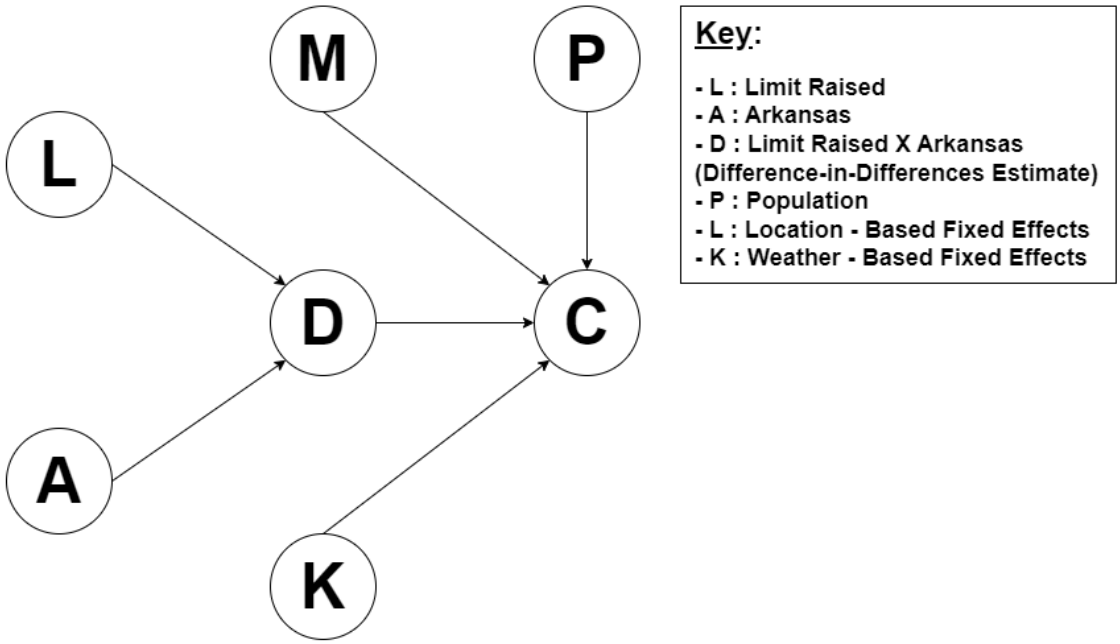


*NOTE: Plots are zoomed in to exclude outliers.



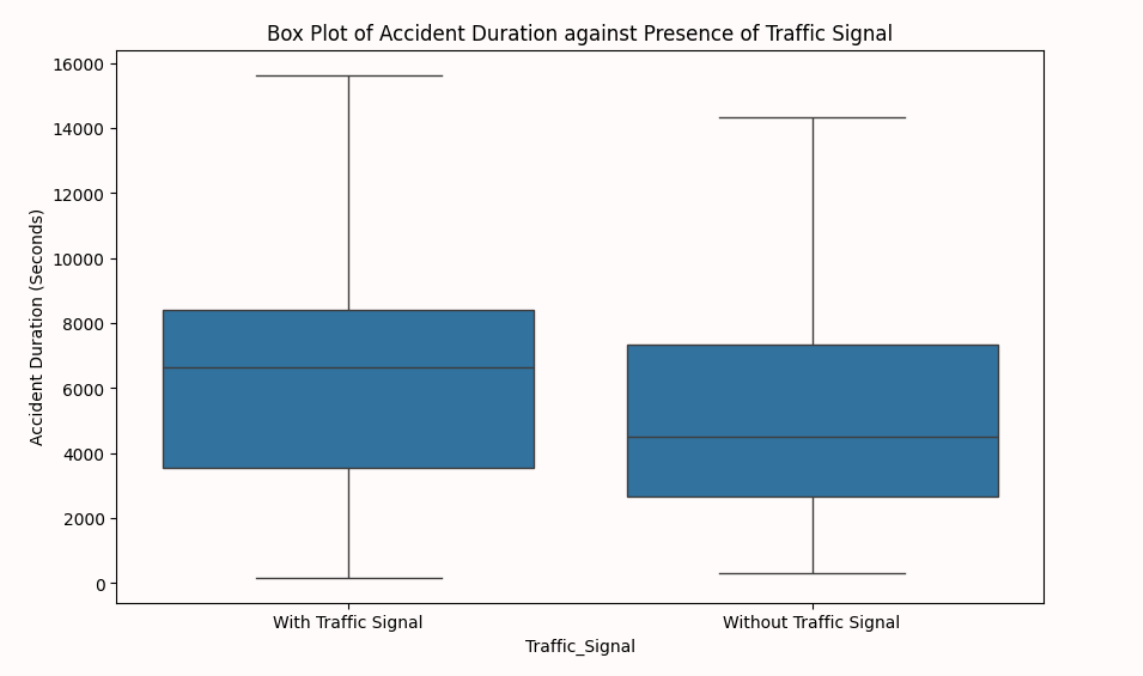
*NOTE: Plots are zoomed in to exclude outliers.

Causal DAG for DID Models



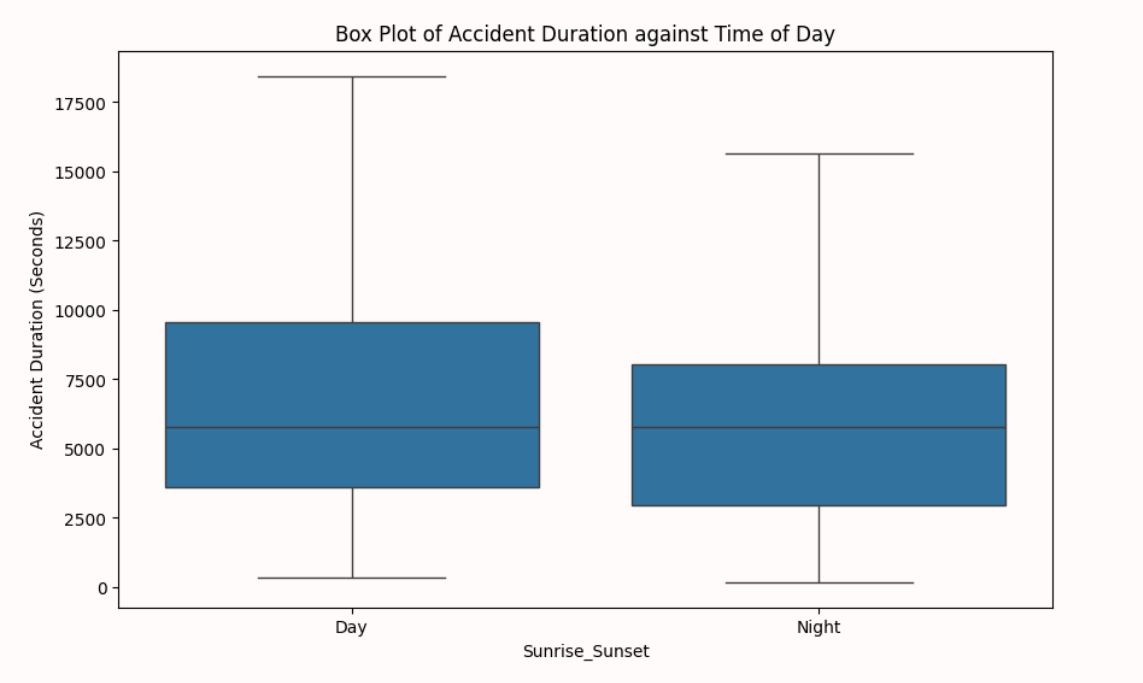
B EDA VISUALIZATIONS

Figure B1



Notes: (a) Aggregated across the 5 cities by Presence of Traffic Signal;

Figure B2



Notes: (a) Aggregated across the 5 cities by Time of Day;

Figure B3

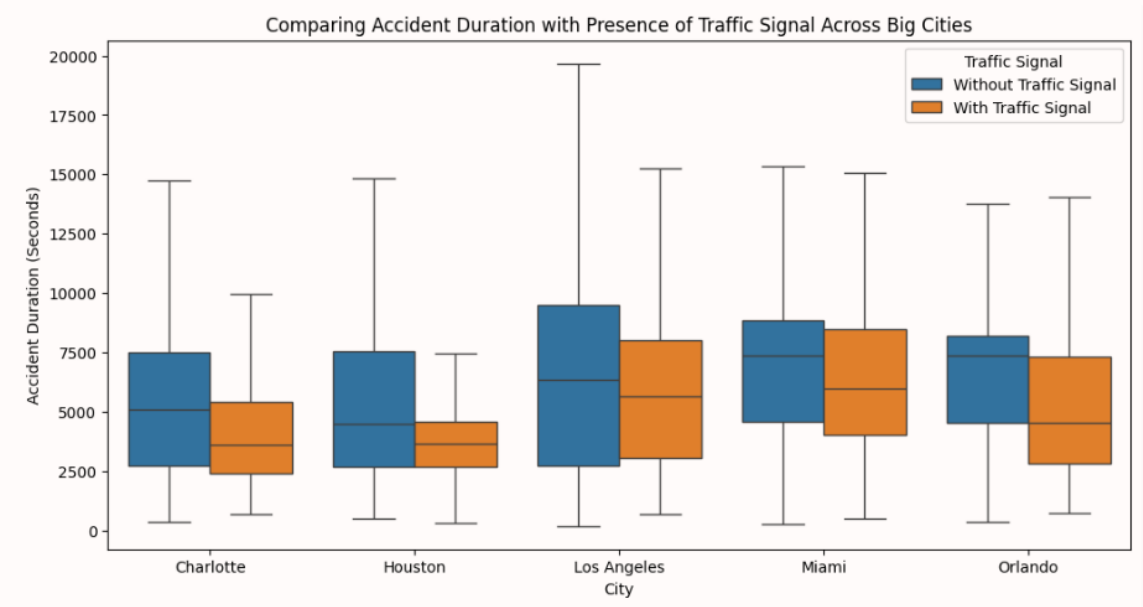

Comparing Accident Duration with Presence of Traffic Signal Across Big Cities
Traffic Signal
Without Traffic Signal
With Traffic Signal
Accident Duration (Seconds)
20000
17500
15000
12500
10000
7500
5000
2500
0
Charlotte
Houston
Los Angeles
Miami
Orlando
City

Figure B4

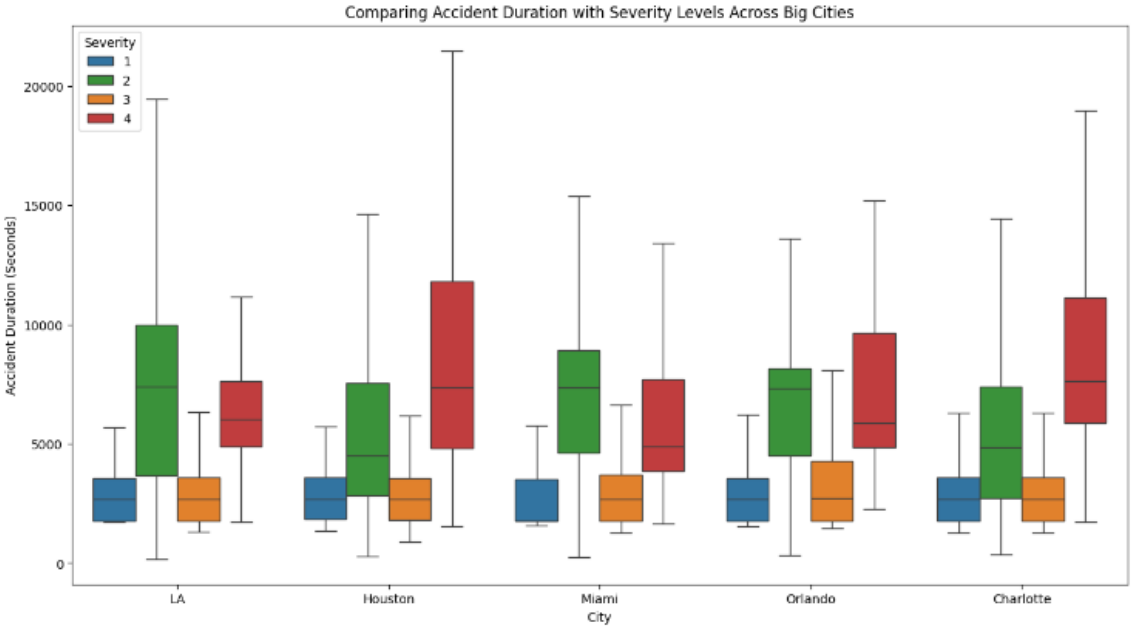
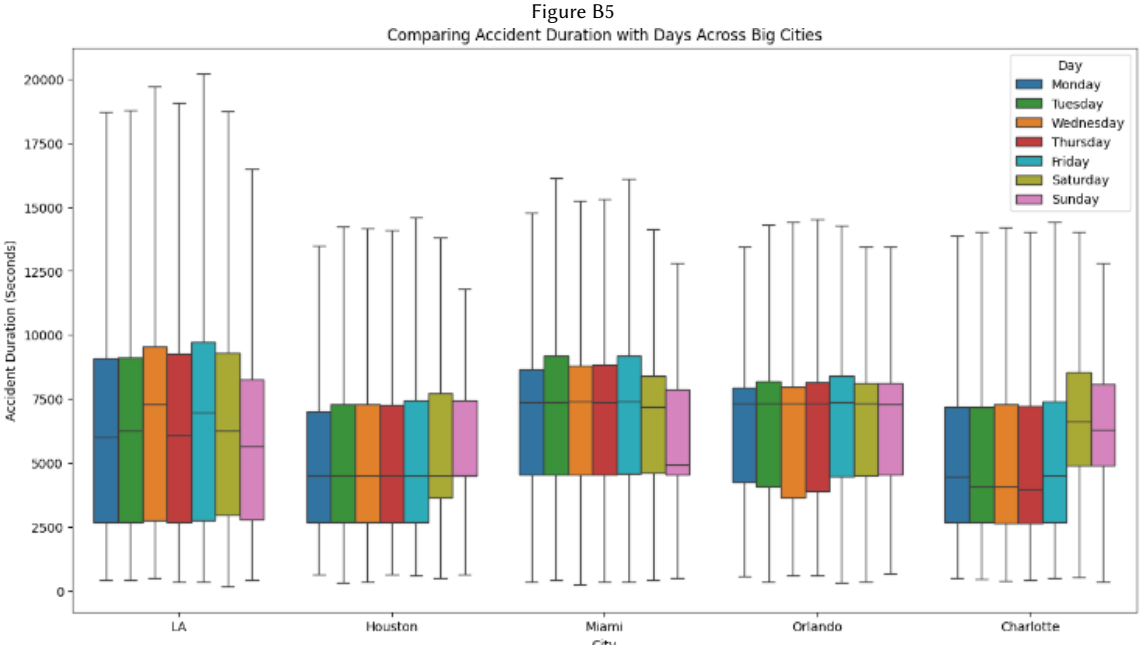

Comparing Accident Duration with Severity Levels Across Big Cities
Severity
1
2
3
4
Accident Duration (Seconds)
20000
15000
10000
5000
0
LA
Houston
Miami
Orlando
Charlotte
City

Figure B5


Comparing Accident Duration with Days Across Big Cities
Day
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Accident Duration (Seconds)
20000
17500
15000
12500
10000
7500
5000
2500
0
LA
Houston
Miami
Orlando
Charlotte