Comparative Analysis of Traffic Accident Counts Across U.S. States and Cities: A Regression-Based Approach

Ahmadou Diallo, Fatima M. Athar, Gulnaz Javadova,
Mehak Zahid and Neerav N. Gala

January 5, 2026

Abstract

This study examines traffic accident counts across the United States at the state, multi-city, and single-city levels using multiple statistical modeling approaches. The exposure-adjusted state-level analysis using Negative Binomial Regression model reveals that structural and environmental factors, rather than individual crash behaviors, account for most of the variation in fatality risk per mile across states. In predicting daily crash counts in multi-city level analysis of New York, San Francisco and Chicago combined, Negative Binomial model outperformed Poisson and Quasi Poisson models, using predictors of road conditions, lighting conditions and time-of-day factors, while weather situation degraded model performance when used in conjunction with the same subset of variables. For the city of Los Angeles, public holidays and accident location emerged as the most influential predictors of accident frequency. Notably, and in contrast to earlier research, the square-root–transformed multiple linear regression model demonstrated a superior empirical fit and higher predictive performance compared with the Poisson model. At the New York City level, we examine how active street construction relates to motor-vehicle crash counts. Using 2023 NYC Open Data and a Negative Binomial count model with a length offset, we find that segments with active construction have about 3.6 times as many crashes per kilometer per month as comparable non-construction segments, highlighting work zones on major corridors as priority targets for safety interventions.

Introduction

Traffic accidents have been a concern since almost the same time as the invention of the automobile in 1886, with the first death from falling out of a steam-powered vehicle recorded in Belfast, Ireland in 1869 [1] and the first pedestrian death involving a car occurring in the UK in 1896 [2].

The early 1900s witnessed a steep increase in the ownership of motor cars, and the increase in accident rates followed not long after. By the 1920s and 30s, death by automobile had become one of the most common cause of mortality in industrialized nations. The apparent lack of infrastructure to accommodate motor vehicles in society along with efforts by advocates like Ralph Nader prompted the first serious efforts at regulation and road safety improvements, which led to landmark legislation like the 1966 National Traffic and Motor Vehicle Safety Act in the United States [3] imposing the addition of safety features in cars.

Since then, safety innovations have only progressed, with additions like airbags and anti-lock braking systems becoming standard car features in the 1990s and stricter drunk driving laws being put into place to prevent alcohol-related crashes. The 21st century sees modern technology being used to bring further safety to vehicles, such as self-driving cars; however, it comes with its added risks such as inaccurate models being used to train the self-driving car computers, resulting in traffic violations or passenger / pedestrian injuries and death [4].

While traffic fatalities have decreased per mile traveled in many developed countries as a result of these efforts, they remain an important public safety concern worldwide with approximately 1.19 million deaths attributed to vehicles in 2021 [5]. In the United States alone, 6.14 million police-reported traffic crashes occurred in 2023, resulting in 37,654 fatalities and 1.7 million injuries [6]. These alarming figures reflect a global growth trend in the number of traffic accidents [7], highlighting the need for systematic analysis.

Over the past decades, researchers have increasingly recognized that traffic crashes are not random phenomena but the result of interrelated behavioral, infrastructural, and environmental factors. A considerable body of research has sought to explain the determinants of crash frequency. Examining road accidents dynamics and improving prediction accuracy [8] may help generate awareness about impacts of driving habits or vehicle features and allow societies to develop effective strategies for reducing frequency and severity of traffic crashes, thus creating a safer world for future generations.

Literature Review

A chronological review of the literature reveals that early foundational studies have played a pivotal role in shaping the statistical modeling of accident frequency. Several studies attempted to model vehicle accidents and high geometric design relationships, concluding that Poisson regression was a preferred method over linear regression for modeling accident data [9]. A study on the relationship between highway geometric factors and truck accidents also found that linear regression inadequately captured this relationship, whereas the Poisson model provided a more accurate representation of truck accident occurrences [10]. However, comparisons of model performance reveal that Poisson models misrepresent accident data with significant overdispersion, highlighting a need for more general probability distributions [9].

Mohamed and Essam [11] analyzed 1,606 traffic accidents that occurred on a principal arterial road in Central Florida over a span of three years. Their study identified several explanatory variables that significantly increased the likelihood of accidents, including high traffic volume, excessive speeding, narrow lane and shoulder widths, a greater number of lanes, urban roadway sections, and reduced median widths. The authors found that female drivers were more crash-prone in heavy traffic and narrow lanes, while male drivers were likelier to speed. Younger and older drivers had higher crash risks, especially in heavy traffic and on curves. The authors initially employed the Poisson regression approach; however, it was deemed unsuitable because the mean and variance of the dependent variables differed significantly, indicating substantial overdispersion in the data. Consequently, the Negative Binomial model was adopted, as it extends the Poisson regression framework by allowing the variance to exceed the mean, thereby providing a better fit for overdispersed count data.

In 1996, Milton and Mannering [12] developed a model to analyze accident frequencies on an arterial roadway in Washington State. Their findings revealed that narrow shoulders, sharp horizontal curves, reduced lane widths, and high traffic volumes all contributed to increased accident frequency. The study also emphasized the limitations of traditional linear regression and highlighted the advantages of Poisson and Negative Binomial regression models. Due to the presence of overdispersion in their data—where the variance exceeded the mean—they adopted the Negative Binomial model to evaluate the influence of roadway geometry and traffic characteristics on annual accident frequencies along principal arterials in Washington State.

Overall, these early empirical studies laid the groundwork for understanding how roadway design and traffic characteristics influence accident counts, motivating the next wave of research into environmental, demographic, and behavioral factors.

According to the National Safety Council (USA), traffic accidents have changed greatly over time. From 1913 to 2023, fatalities from traffic accidents increased by 966%, from 4,200 deaths in 1913 to 44,762 in 2023 [13]. The estimated global cost of road traffic crashes (including injuries, deaths, property damage, and lost productivity) is around USD 3.6 trillion per year (3% of global GDP), according to iRAP's latest data (2024) [14].

With the rising complexities of modeling motor vehicle crashes to provide defensible guidance on how to properly model crash data, Dominique Lord [15] first examined the motor vehicle crash process using theoretical principles and a basic understanding of crash mechanisms. His studies show that the fundamental crash process follows a Bernoulli trial with unequal probabilities of independent events, also known as Poisson trials.

Natalia and Fred [16] conducted an assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents using Indiana highway accident data. The results showed that the presence of approved design exceptions had no statistically significant effect on the average frequency or severity of accidents, suggesting that current procedures for granting design exceptions have been sufficiently rigorous to avoid adverse safety impacts. However, the findings also suggested that the process determining accident frequency does vary between roadway sites with design exceptions and those without.

Jie and Kara [17] conducted research to develop a model that allows researchers to simultaneously model crash outcomes by severity, based on a type of multivariate Poisson (MVP) specification that can be estimated within a Bayesian framework using Gibbs sampling. Crash counts for over 40,000 homogeneous segments of Washington State highways in 1996 were used to estimate the model. As expected, a positive correlation in unobserved factors affecting count outcomes was found to exist across severity levels, resulting in a statistically significant additive latent term.

While these studies focused on geometric and environmental conditions, subsequent research extended the analytical focus to the socioeconomic and demographic dimensions of crashes.

Across multiple studies, socioeconomic status (SES) has emerged as an important demographic factor impacting both likelihood and severity of road traffic

accidents. SES is mostly measured through income, education, occupation and area-level socioeconomic disadvantage. Lower SES has been consistently correlated with higher crash involvement.

While efforts to improve roadway safety have largely focused on improving geometric design enhancements, Sagar et al. (2021) [18] examined how socioeconomic characteristics of driver residence zip code along with other demographic characteristics (age, gender, rurality, traffic convictions) influence road crash counts across Kentucky, US, a state that reports higher than national average for number of crashes. Using logistic regression models, the study found that communities with lower income levels, lower educational attainment, and higher proportions of minority populations experience significantly higher crash frequencies.

In Tehran, Sehat, M et al. (2012) [19] conducted a similar study using logistic regression where they found that individuals who were tenants or had lower-value households and from lower-income were significantly more likely to be involved in road traffic accidents, with males at a higher risk. Traffic accidents were higher for men who were unemployed and with pre-diploma education, while for women, traffic accidents were higher if they were retired, illiterate and divorced/ widowed. The study also found that people in lower house value quartiles had significantly higher risk of being in a traffic accident, even after adjusting for some demographic factors like age, gender, marital status etc.

In Sweden, Hasselberg et al. (2005) [20] used Poisson regression to show that young drivers aged 16–23 from lower socioeconomic backgrounds (measured by parental education, income, and occupation) experience notably higher rates of traffic injuries. Building on this, Hanna et al. (2010) [21] focused on unlicensed young drivers and found that those from low-SES families, particularly in rural areas, were disproportionately involved in crashes with more severe outcomes, compared to peers from high SES families. The study also notes that many unlicensed drivers were old enough to obtain a license, highlighting concerns about licensing accessibility. .

Getachew et all (2024) [22] conducted a systemic review, analyzing 16 studies in Ethiopian drivers. The study highlighted that financial constraints and low education attainment, among other factors, led to limited driving experience, limited awareness of road safety, substance use, poor vehicle maintenance and frequent traffic violations, all of which contributed to incidence of road traffic accidents.

Complementing the demographic literature, other studies have analyzed spatial, behavioral, and psychological aspects influencing crash fatality and driver aggression.

Feng Guo [23] collected driving performance and behavior data using multiple cameras and sensors for participant drivers recruited for up to three years across different age groups. He used a mixed-effects logistic regression model with driver-specific random effects and found that teenage, young adult, and senior drivers are more adversely impacted by secondary-task engagement than middle-aged drivers. Visual–manual distractions affect drivers of all ages, whereas cognitive distractions may have a larger impact on young drivers.

In 2016, a study on drivers from rural areas traveling to urban areas and vice versa determined that it was not just the accident location, but also the driver's lack of familiarity with roads that increased the likelihood of a crash being deadly [24]. Another factor that may increase the probability of a traffic accident being lethal in rural areas is the lower use of seatbelts as observed in a study based on 2014 US traffic accident data, given that all drivers involved were adults [25]. The findings of these studies indicate a possible lack of traffic law enforcement, rule awareness, or even appropriate infrastructure to handle an unexpectedly high inflow of vehicles in a given jurisdiction.

In 2020, Arnau et al created a Poisson regression model on traffic fatality data for England and Wales from 2008 to 2018 to describe the relationship between the number of traffic accidents in select urban areas and their population [26]. The model results showed that the likelihood of an accident occurring grew sublinearly with the population, whereas the likelihood of those accidents causing death decreased sublinearly. Traffic congestion may be an indirect cause of the observed behaviour, as it could cause an increase in drivers' stress levels thereby increasing the probability of a traffic crash happening while reducing its severity.

A survey conducted in 2021 explores road rage as another cause for fatal car crashes and indicates the need for an integrative conceptual framework to understand it, based on a psychological analysis of emotion and emotion regulation [27]. The suggested framework would serve two purposes: it would organize existing research on road rage and reveal key areas that future studies should address. However, since road rage is a subjective concept, conducting empirical studies to validate its causal effect on crash frequency poses challenges in measurement and analysis.

Synthesizing across these studies, a clear methodological evolution emerges in how crash-frequency models have been developed, reflecting the progression from traditional regression frameworks to

more advanced probabilistic and learning-based techniques. Poisson regression remains one of the most widely applied approaches for modeling crash-frequency data compared to linear models [9][28][10] because of its interpretability [29]. However, the restrictive assumption of equal mean and variance often fails in real-world crash data, prompting researchers to adopt more flexible variants such as the Negative Binomial [11][30] and zero-inflated models to account for overdispersion and excess zeros. In addition to these count-based techniques, multivariate regression analyses have also been used to reveal the underlying determinants of accident risk [31]. Over time, crash-frequency modeling has advanced beyond traditional Poisson and Negative Binomial regressions toward hierarchical, Bayesian, random-parameter models designed to address key statistical complexities such as overdispersion, unobserved heterogeneity, temporal correlation, and excess zeros [32]. On the other hand, more recent contributions integrate time-series and machine learning approaches, for instance, combined multifactor regression with ARIMA and LSTM models to forecast traffic accident frequency, achieving prediction accuracies between 94.67% and 97.64% [7]. Collectively, these developments indicate a methodological continuum in which classical count-based models provide the theoretical foundation, while modern learning-based methods extend their predictive capability.

Taken together, the literature highlights the diversity of methods applied to crash-frequency modeling, with a clear evolution toward integrating human, socioeconomic, and behavioral dimensions. Yet, despite these contributions, gaps remain in directly comparing the performance of classical regression models across different sets of explanatory factors. Addressing this gap, the present study undertakes a comparative regression analysis of traffic accident frequency with emphasis on various predictor groups. Aligned with the United Nations Decade of Action for Road Safety 2021–2030, which envisions a 50% reduction in global road traffic deaths and injuries [33], this study aims to explore how modeling choices, predictor variables, spatial scales and environments influence the performance and interpretation of crash frequency Regression models across U.S. geographies. This study aspires to generate actionable insights that support safer road environments and contribute to the global commitment to sustainable and resilient mobility.

Analysis

A hierarchical strategy is adopted to analyze crash frequencies across various US geographies, beginning with ***state-level*** crash modelling to achieve a sense of broad trends and arrive at baseline predictive performance. Then, the analysis will be narrowed down to ***multi-city level*** by picking a combination of three diverse cities of New York, Chicago and San Francisco to capture insights that maybe masked at state level by offering a diverse sample than a single city alone. This will be followed by a more focused analysis of ***single city*** of Los Angeles to capture more localised crash patterns. Finally, two modelling approaches will be tested ***within same city*** of New York, with different variable subsets to reveal how sensitive crash predictions are to choice of inputs.

Methodology

To analyse the contributing factors for fatal crashes in the United States, initially the broader and structural dynamics are studied. Such **state-level analysis** requires adjusting for how much people actually drive, as US states differ dramatically in traffic volume. For example, raw crash counts show that states like Texas, California, and Florida have the greatest fatality count, however this may result from high traffic volume, and does not necessarily indicate high risk in those states. Thus, analyzing crash per mile instead of raw crash counts enables to distinguish states with higher fatality risk and main reasons behind these differences.

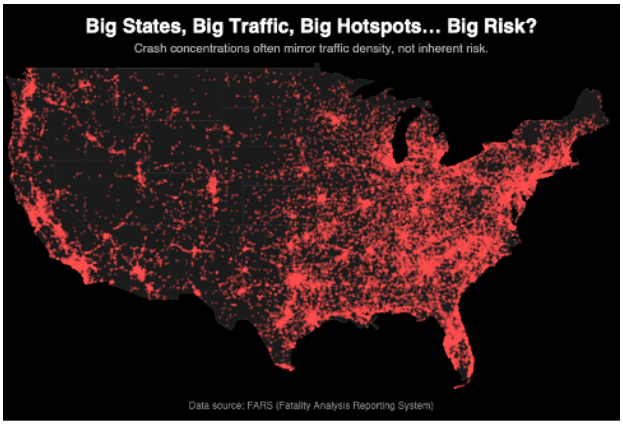

Figure 1: *US Accidents (2023)*

Building on this idea of exposure-adjusted risk, this analysis measures fatal accidents per mile in each state in 2023. Practically, the aggregated fatal crash counts are modelled with Negative Binomial regression model with log of vehicle miles travelled (VMT). When the log of VMT is added as an offset to the model, the raw counts of fatal crashes transform into fatality risk per mile.

To operationalize this model with real data, a nationwide dataset was used from Fatality Analysis Reporting System (FARS)[34]. This dataset reports every fatal crash occurring in the US with detailed informa-

tion related to the accident circumstances, people and vehicles involved. Because exposure and socioeconomic context also influence fatality risk, the Vehicle Miles Traveled (VMT) dataset from FHWA[35] and the income per capita data from the Bureau of Economic Analysis (BEA)[36] were integrated, which provides a structural socioeconomic control.

To ensure comparability across these sources, all three datasets were cleaned, standardized, and merged. The merged dataset then allowed to create a state-level summary where each state has:

- total fatal crashes,
- vehicle miles traveled,
- behavioral characteristics (alcohol, drug, speeding, licensing indicators),
- average driver and average vehicle age,
- environmental context (rurality, lighting, weather), and
- economic indicators (income per capita).

To understand how behavioral and environmental factors impact crash frequency, selected FARS data were grouped into meaningful binary indicators. To measure the impact of driver and vehicle age, the average vehicle/driver age was calculated. Categorical binning of age was avoided due to small sample size. Behavioral variables included the share of drivers with alcohol, drug, speeding, and invalid licencing. Besides, the various weather conditions were aggregated into a single binary variable indicating whether the crash occurred in adverse weather. The share of bad weather related crashes, the share of crashes in rural area, and the share of crashes in nighttime were calculated. Income per capita was used to control for socio-economic differences over states and log of VMT was used as an offset to measure the crash risk per mile.

Descriptive statistics of predictors shows that about 18% of fatal crashes involved alcohol and 11% involved drug use, 22% occurred under adverse weather, and almost 55% took place during nighttime or low-light periods. Around 40% of accidents happened in rural areas while 28% were related to speeding. For around 27% of accidents, any of the drivers had invalid driver licence.

The visual analysis confirmed that big states with highest fatality count, in fact, have highest traffic volume. When fatality counts are normalized by VMT, the ranking changes dramatically and those most 'dangerous' states converge to national averages while states such as Mississippi, South Carolina, Arizona, Kentucky, etc. emerge as the highest-risk environments on a per-mile basis. This insight motivates to study which conditions drive the differences in fatal crash risk per mile in states, and whether these factors are systemic or behaviour-related.

An alternate technique is to move from state level analysis to **multicity level** analysis. One approach adopted predicts crash frequencies using traffic collision data on city level, by clubbing three diverse cities together in one dataset, namely: New York (East) [37], Chicago (Midwest) [38], San Francisco (West) [39]. The three cities provide a diverse mix of urban environments, with unique traffic and road infrastructure as well as traveling behaviors. The cities were deliberately chosen to be geographically distant so as to avoid the risk of spatial autocorrelation where crash counts in similar locations tend to be similar because of factors such as population density, weather conditions etc. Data for Chicago includes all types of motor vehicle crashes reported to the Chicago Police Department during the year 2023. Data for New York includes all motor vehicle crashes reported to the New York Police Department during the year 2023, that included an injury, fatality or property damage above $1000. Data for San Francisco includes all motor vehicle crashes that resulted in injury and were reported to the San Francisco Police Department during the year 2023. SFO data excludes crash data that fall under the jurisdiction of California Highway Patrol, or crash data causing property damage only, as well as non-injury collisions, thus there is some under reporting due to limited SFPD resources. The chosen approach is to use Poisson regression model to predict daily crash frequencies. This approach works best because it is designed for non-negative integer outcomes, handles skewed distributions effectively, and allows inclusion of both categorical and continuous predictors. The goal is to use independent variables of lighting condition, weather situation, crash time and road conditions, for all three cities. Therefore, all chosen predictors will be categorical. Crash data will be aggregated at the daily level per city to reduce random variation and for stability of predictions. This level of aggregation will create the count variable of daily "crashes", which will be our dependent variable. Data cleaning objectives include imputing missing values, correcting for inconsistent formats and recoding categorical variables for consistency. The aggregated data will be randomly split into 80% for training and 20% for testing purpose. In the event that Poisson model's assumptions are violated because of overdispersion, further models will be trained and tested in following sequence: Quasi Poisson, Negative Binomial, Zero Inflated Negative Binomial. Once we arrive at the model that fits the data well, variable selection will be implemented along with cross validation to achieve a model that renders best prediction accuracies in MSE, MAE, MAPE and PM.

After multi city level, a **city level analysis** was

carried out. The study compared OLS and Poisson regression in analyzing daily traffic accident counts in **Los Angeles** for the year 2022, using three primary datasets:

- Traffic Collision Data from the Los Angeles Police Department [40]: The raw dataset was filtered to include only collisions that occurred in 2022. These observations were aggregated to produce a daily accident count, denoted as *Accidents*, which served as the response variable. The police reporting area with the highest number of accidents on each day (*AreaNameTop*), as well as the mean age of victims (*VictimAgeMean*), were extracted as predictor variables.
- 2022 Official Holidays Calendar from StreetsLA [41]: The data set of one binary variable (called *Holiday*) indicating whether a particular day was a public holiday in Los Angeles or not in the calendar year 2022
- Los Angeles International Airport – Passenger Traffic by Terminal data set [42] : The raw data was cleaned and aggregated to get the total number of passengers traveling (called *AirportPassengerCount*) via the Los Angeles International Airport each month in 2022

The traffic collision data were merged with the holiday dataset using the day of the year, followed by a merge with the airport passenger dataset using the month of the year. The resulting analytical dataset contained 365 observations (one per day) with four predictor variables and one response variable.

Univariate and bivariate analyses were conducted to examine distributions and relationships among variables. The dataset was then divided into training (292 observations) and testing (73 observations) subsets using a random seed of 42. Cook's distance was used to assess influential observations, leading to the identification and removal of 20 outliers from the training set, resulting in a final sample size of 272 training observations.

The following models were trained and tested:

- Model1 : Multi linear regression (MLR) model consisting of all four predicting variables and trained on training data including outliers
- Model2 : MLR model consisting of all four predicting variables and trained on training data excluding outliers
- Model3 : MLR model consisting of all four predicting variables and trained on training data excluding outliers. However, a square root tranformation was applied to the response variable
- Model4 : A stepwise forward regression was carried out on model3 and the model was trained using the selected variables
- Model5 : A Poisson regression (PR) consisting of all four predicting variables and trained on training data including outliers
- Model6 : A stepwise forward regression was carried out on model5 and the model was trained using the selected variables

Goodness-of-fit assessments and hypothesis tests were conducted to evaluate the explanatory power of each model. The MLR models were examined for multicollinearity, while the Poisson models were assessed for potential overdispersion.

All models were subsequently evaluated on the testing dataset, and their predictive performance was compared using MSE, MAE, MAPE and PM.

Having compared OLS and Poisson regression for modeling daily traffic accidents in Los Angeles, we now shift to **New York City** and a finer spatial scale, individual road segments, to quantify how much active street construction is associated with higher motor-vehicle crash counts compared with non-construction segments.The main goal of the analysis is to quantify how much active street construction is associated with higher motor vehicle crash counts, after accounting for where in New York City the crashes occur and for differences in segment length and seasonality. The outcome of interest is the number of crashes per road segment per month in 2023, and the key explanatory variable is whether that segment had at least one active construction permit at the time of the crash.

The analysis uses three NYC Open Data sources restricted to calendar year 2023:

- Motor Vehicle Collision: New York Police Department NYPD [37]
- Street Construct Centerline Data: Office of Technology and Innovation (OTI) [43]
- Centerline Data: City's Office of Technology and Innovation (OTI) [44]

The crash file provides crash date, time and location (latitude and longitude), basic characteristics of the event and the borough as coded by NYPD. The centerline data represent the road network as line geometries with attributes such as borough, street names, segment identifiers and segment length. The construction dataset provides permit-level information on road and sidewalk works, including the permitted location, borough, and the start and end dates of the work period. Crashes, street segments, and construction permits were first cleaned and put into the same map coordinate system for New York City. Each crash and each permit was then attached to its nearest street segment using geographic coordinates so that all three sources referred to the same

set of roadway segments. Using the permit start and end dates, a crash on a segment was marked as "in construction" if it happened while a permit was active there. Finally, the data were summarized by street segment and month for 2023, counting how many crashes occurred, how many involved injuries, whether any construction was underway, and how long each segment is (used as a measure of exposure). Before fitting models, the panel was randomly split into a training and a test set using an 80/20 split at the segment-month level, with a fixed random seed for reproducibility. All models were fit on the training data and later checked on the test set. The primary modeling approach was count regression. The starting point was a Poisson regression in which the expected number of crashes per segment-month is modeled as a log-linear function of construction status, borough and month, offset by the logarithm of segment length in kilometers.

Because Poisson regression assumes that the variance equals the mean, the analysis next checked for overdispersion using the Pearson dispersion statistic. The resulting dispersion factor was approximately 29, which is far larger than 1 and even well above the informal threshold of 2 often used in practice. This indicates strong overdispersion and signals that standard errors and p-values from the Poisson model would be misleading. In response, the same specification was re-estimated using a Negative Binomial regression,

$$\log(\mu_i) = x_i^T \beta, \quad \mathrm{Var}(Y_i \mid x_i) = \mu_i(1 + \kappa\mu_i) \quad (1)$$

which includes an extra dispersion parameter to allow the variance of crashes to exceed the mean. Model comparison between Poisson and Negative Binomial was conducted using the Akaike Information Criterion (AIC), and additional diagnostics were performed on the Negative Binomial fit, including residual plots, a global likelihood-ratio test against an intercept-only model, and a Cook's distance check for influential segment-months.

Two statistical approaches - logistic regression and one way ANOVA - were deemed unsuitable for this analysis based on the nature of our response variable and our modeling goals. Logistic regression is suitable for classification problems using binary outcomes, e.g. if a crash occurred or not, whereas our goal is to model crash frequencies, a count variable. Applying logistic regression to a count variable would require converting it to a binary variable, which can sabotage predictive power. Similarly, one-way ANOVA would be relevant for a comparative analysis of mean crash counts across categorical variables e.g. cities or states. It is not suitable for the predictive modeling of crash counts or for handling multiple categorical and continuous variables simultaneously. An extension of ANOVA is Multiple Linear Regression, which can be used for various categorical and numeric variables simultaneously; however, we choose to not consider this because there is a risk of invalid negative predictions and nature of count data which is expected to violate goodness-of-fit assumptions of MLR. Count data is typically skewed, with a non-constant variance, making Poisson regression a more accurate and statistically appropriate choice for this study.

Another approach to looking at New York City traffic data for predicting accident severity is considering a categorical response variable which indicates whether a collision was fatal for one or more of the people involved; binomial logistic regression is the optimal choice for this scenario. The goal of this approach is to look at the importance of considering vehicular information and that of the involved persons in predicting traffic accidents. As safety measures have evolved over time, there are many concerns regarding the quality of the make of cars and the safety equipment, which is why this question has been considered. Data for collisions [45], vehicles [46] and persons [47] are readily available online for downloading and conducting independent analyses.

To enhance the predicting power of both models, forward-backward stepwise regression was further performed to compare the AICs and get a definitive answer on the better of the two models. One aspect of this analysis that may need improvement is the removal of unnecessary features which not only increase the dataset size but also slow down the models especially if they have little to no predictive power another area of improvement is that while data for one year is sufficiently large, it cannot cater to longer cyclical behavior such as any breakthroughs in car manufacturing, changes in traffic laws or even global events such as climate change or political / economical shifts that have a ripple effect.

Results

To analyze crash frequency differences at **state-level**, initially, a Poisson model was explored, with data aggregated on a daily basis. Traffic volume data was divided by 365 to approximate daily VMT. Although the daily Poisson model showed no overdispersion and fit the data well, it was ultimately decided that this model is not suitable for the research question, considering the assumption that factors like rurality, licensing conditions, nighttime driving share, or vehicle age would not meaningfully change from one day to the next. Thus, aggregating accidents into daily observations would simply inflate effects, by introducing noise without adding useful variation.

Given that the study's goal is to explain how and why states differ from one another in crash risk, it was decided that the natural unit of analysis is the state itself. For this reason, the final model was built on the 51 state level (n=51) using Negative Binomial Regression model. However, the main challenge for this approach is model validation. In traditional machine-learning approaches, common model validation methods are randomly splitting data into train/test subsets, k-fold CV, or LOOCV. However, since each observation (i.e. state) in this analysis has distinct and extensively varying demographic or environmental features, holding out one/several states and explaining them based on the model built on other states' data could possibly lead to biased or misleading performance. For this reason, the model is assessed using standard diagnostic checks rather than prediction-focused validation. The aim here is explanatory rather than predictive - to understand what drives differences in fatal crash risk across states, not to forecast future counts.

Thus, for state-level Negative Binomial model, both the deviance goodness-of-fit test (p = 0.128) and Pearson dispersion test (p = 0.144) indicated a good fit. The Pearson dispersion ratio (1.24) reflects mild dispersion (big reduction of dispersion compared to Poisson model at 15.83), and the estimated dispersion parameter θ = 42.6, SE = 9.46) confirms that overdispersion is present but manageable and appropriately handled by the NB model. Overall, the diagnostic results confirm that the model is an adequate fit to state-level fatal crash counts after adjusting for exposure via the log(VMT) offset.

The results of Negative Binomial model at **state level** show that several structural and environmental characteristics play a meaningful role in shaping fatality risk per mile. First, nighttime driving is strongly linked to higher fatality risk. Increasing the nighttime share by 10 percentage points is associated with roughly a 12-13% rise in fatal crash risk. Surprisingly, adverse weather shows an opposite pattern. States with more crashes in rain, snow, fog, or similar conditions actually have lower fatality rates. A 10-percentage-point increase in bad-weather crashes is linked to an 8-9% decrease in fatality risk. This counterintuitive finding possibly reflects behavioral adaptation: drivers tend to slow down during adverse weather, reducing the severity of outcomes.

Income per capita also matters: higher-income states tend to have fewer fatal crashes, possibly because of safer roads, safer cars, or stronger enforcement.

Some other predictors are close to being statistically important. States with more rural roads generally have higher fatality risk, and states with more speeding-related or unlicensed-driver crashes also tend to do worse.

On the other hand, factors like average driver age, alcohol or drug involvement, and average vehicle age are not statistically significant in explaining between-state differences. This does not mean they are unimportant in individual crashes. This does not imply these behaviors are unimportant at the micro level; rather, their variation across states is overshadowed by broader structural conditions.

Overall, the results indicate that state-level fatality risk is shaped more by structural features - rural road exposure, nighttime driving conditions, adverse weather behavior, and socioeconomic context than by individual risky behavior (alcohol, drugs, speeding). This aligns with the exploratory analysis and supports the conclusion that national road-safety improvements must address systemic issues – especially rural road design, lighting infrastructure, and socioeconomic disparities – to meaningfully reduce fatality rates.

Further narrowing our analysis to **multi-city level** of New York, Chicago and San Francisco combined, the intial challenge was cleaning and aggregating the data. For example, road conditions that were described with words like 'snow', 'ice', 'slush' across the datasets were all recoded as 'snow.slush', and road conditions described as 'sand', 'mud', 'dirt', 'muddy' were all recoded as 'mud.dirt'. Similar recoding was performed for other variables of lighting condition, crash time (0-23hour format) and weather situation. For this approach, controlling variables are 'city', to control for regional factors, and 'hour', to control for time-of-day differences. Explanatory variables are 'city', to account for spatial differences, 'hour' to pick rush hour patterns, 'weathersit', to explain crash risk, 'lighting' to explain how visibility influences crash risk and 'road_condition' to explain how road surface traction is associated with crashes. Only time variable had a few missing values which were imputed with median. The aggregation of categorical variables was done by picking the mode or most frequent reported variable condition on daily basis. For example, on a particular day in New York, if most crashes were associated with road condition being 'wet', we will use that to describe all crashes for that day. This reduces noise from within day variability, but we do maintain variability across 365 days of the year and do not lose predictive power, while also simplifying the model that avoids overfitting. Exploratory data analysis revealed that there is huge variability in daily, median crash counts across the three cities, with SFO at approximately 20, Chicago at approximately 300, and New York at approximately 1100. Most crashes occur during peak morning rush hours

of 7 to 8AM and then during mid-day to evening, between 12 noon to 7PM. The crash counts are lowest during late night and early morning, between 8PM to 6AM. The crash counts are only slightly less on weekends, compared to weekdays, with highest median counts reported on Fridays. With regards to weather situation, most crash frequencies can be associated with snow first, rainy weather ranking second, with a large IQR of 800 crashes for both, pointing at the large spread. Crash counts are mostly associated with daylight, with a significant amount of variability, while 'dark-with lights'" appears as the safest lighting condition with zero median, implying that perhaps other factors such as traffic volume play a role in low crash counts, in conjunction with lighting. The highest median crashes of approximately 1100 are reported for roads with snow and slush, while wet and dry roads exhibit the same median and variability of crash counts at approximately 300. Initially, Poisson regression was chosen for this analysis as follows:

$$Crash\ counts = Y_i \sim \text{Poisson}(\lambda_i),$$

$$\lambda_i = \exp(\beta_0 + \beta_1 \text{city}_i + \beta_2 \text{hour}_i + \beta_3 \text{road}_i + \beta_4 \text{light}_i + \beta_5 \text{weathersit}_i)$$

The model is quite optimistic and returned 15 of the 31 predictors as statistically significant at alpha level of 1%. These predictors are intercept, cityNY, citySFO, hour3, hour7, hour14, hour17, hour18, hour21, roadsnow.slush, roadwet, lightdaylight, weathersitcloudy, weathersitrainy and weathersitsnow. For clarity, in this model, the intercept estimate of 5.4 can be interpreted as follows: approximately 221 crashes are expected per day, when all variables are at baseline, which here means for Chicago, at hour0 or midnight, when the road is dry, when the weather is clear and lighting condition is Dark with lights. The overall regression was statistically significant with a p-value of 0, implying that at least one of the predictors significantly explains the variability in crash frequencies. Testing for the subset of weather coefficients revealed that they are jointly statistically significant at p-value of 0. However, this model failed the goodness of fit diagnostics as p-value for both Pearson and Deviance residuals came to 0, while QQ plot and histogram of deviance residuals revealed heavy-tailed distribution. Standardized residuals were examined at a cutoff of 99.995% quantile of normal distribution, identifying 143 out of 1095 observations or 13% as outliers, pointing at heavy tailed distribution, and potential overdispersion or zero inflation. This diagnosis was further confirmed when the overdispersion parameter was computed to be 10.78 which is a clear violation of assumption of poisson distribution, making statistical inference from Poisson model unreliable. This is likely because of unobserved heterogeneity from factors affecting crash counts not included in the model, such as traffic volume, driver behavior, local events etc. This led to testing new modelling approach so that extra variance in the crash counts can be accounted for. First, Quasi Poisson model was fit with all variables, which resulted in 4 out of 31 variables being statistically significant at 1% alpha. This is an improvement compared to Poisson model we fit earlier. However, this model also failed goodness of fit test, returning p-value of 0 with reference to residual deviances. Next, Negative Binomial model was fit with all variables, which resulted in 4 out of 31 variables being statistically significant at 1% alpha, namely intercept, 'cityNY', 'citySFO' and 'light.daylight', same as Quasi model. This model fits the data reasonably well, for an alpha of 1%, the p-value was computed to be 0.11. Visual inspection of deviance residuals via histogram revealed a smooth bell shaped curve and QQ plot highlights less deviation at tails compared to Poisson model, further confirming that the Negative Binomial model is a good fit. For variable selection, we opted for Stepwise Forward/Backward selection because it selects entire factors (with all levels) based on AIC, keeping the model interpretable whereas Lasso/ Ridge regression can shrink some levels of a categorical variable, leading to partial factor selection, complicating interpretability. This is particularly useful because all our chosen predictors are categorical, and small in number. In addition to above, using 'mpath' package in R, we explored glmregNB() to fit Lasso and Elastic Net regularization, and cv.glmregNB to apply cross validation for the original Negative Binomial model, but it was computationally intensive. Fitting and cross-validating over a sequence of lambda values can take a long time (in hours), specially when categorical predictors expand into several dummy variables. Therefore, cross validation through L1 and L2 regularization were not applied, only Stepwise selection was applied. The model with the lowest AIC of 8377.3 discarded the weather variable completely, while the model with highest AIC of 12125 was the intercept-only model. The Stepwise selected model formula is as follows:

$$Crash\ counts = Y_i \sim \text{NegBin}(\lambda_i, \alpha)$$

$$\lambda_i = \exp(\beta_0 + \beta_1 \text{city}_i + \beta_2 \text{hour}_i + \beta_3 \text{road}_i + \beta_4 \text{light}_i) \quad (2)$$

Finally, prediction accuracies were computed on test data, comprising of 20% observations, for all four models fit in this approach, namely Poisson model, Quasi Poisson model, Negative Binomial model and Stepwise Negative Binomial model. Lowest MSE,

MAE, MAPE and PM were reported by the Stepwise Negative Binomial Model. Full model for Negative Binomial is second best in performance, while Quasi Poisson and full Poisson model are same in performance, and are the worst. The prediction accuracies are summarized in Table 1. At **multi-city level**, the Stepwise Negative Binomial is successful because it models the count data appropriately, while also addressing the issue of overdispersion. The prediction accuracies are good. The MSE is approximately 8406 for number of crashes. Since, both New York and Chicago have relatively large number of crashes compared to SFO, we can expect a big number for MSE because of difference in scales across the cities, however, this does not mean that model performs poorly. MAE of approximately 55 can be interpreted as the model's predictions are off by 55 crashes on average. This is acceptable for the data as NY has average daily crashes exceeding 1100, and Chicago has average daily crashes around 300. The error mostly seems to stem from SFO data which has average daily crashes lower than 55. We note that compared to MSE, the MAE is more robust for extreme values and works well with overdispersed data. MAPE shows that on average, predictions deviate by about 20% from observed values. An excellent MAPE is under 10% and a poor MAPE is above 30%, so our measure is within acceptable range. It is more intuitive because it is scale-free, however less accurate for small counts, which is important to note here because SFO has small counts. PM of 0.046 implies that model error is only 4.5% of the total variation in the observed data, this is a very good number as its closer to 1, however this measure is most suited for OLS method compared to chosen negative binomial model.

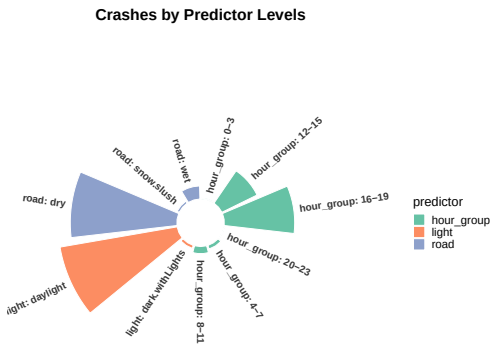

Figure 2: *NY, SFO, Chicago Crashes (2023)*

Moving from multi city to city level analysis of **Los Angeles**, the response variable, *Accidents*, was approximately normally distributed with a slight right skew. The bivariate analysis provided early indications that the variables *Holiday* and *AreaNameTop* (figure 3) might possess explanatory power. In contrast, the quantitative predictors—*VictimAgeMean*, *AirportPassengerCount*, and *Accidents*—exhibited no meaningful correlation with daily accident counts.

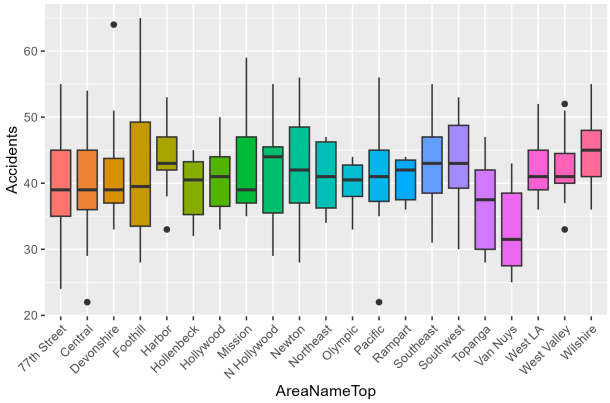

Figure 3: *LA 2022 - AreaNameTop versus Accidents*

The full multiple linear regression model (Model 1) appeared to provide an adequate fit to the training data and demonstrated statistically significant explanatory power according to the F-test. The variable *Holiday*1 was significant at the $\alpha = 0.05$ level and indicated that, holding all other predictors constant, holidays were associated with approximately 8.32 fewer accidents compared to non-holiday weekdays in Los Angeles in 2022. Additionally, *AreaNameTopVanNuys* was statistically significant at the same level, suggesting that the Van Nuys area experienced roughly 8.76 fewer accidents per day relative to the reference area, 77th Street. The model showed no evidence of multicollinearity; however, Cook's distance diagnostics identified 20 influential observations in the training dataset.

Model 2, which was trained on the dataset with outliers removed, exhibited an improved R^2 value relative to Model 1. In this specification, several location indicators—including *AreaNameTopHarbor*, *AreaNameTopSouthwest*, and *AreaNameTopWilshire*—emerged as statistically significant at the $\alpha = 0.05$ threshold, in addition to *Holiday*1 and *AreaNameTopVanNuys*. These results further reinforce the conclusion that both public holidays and geographic location exert a significant influence on daily accident counts.

A Box-Cox test calculated the optimum $\lambda = 0.75$, there a square root transformation was on the response variable in model4, which fit the training data when and showed a slightly improved the R^2.

To perform variable selection, stepwise forward regression was implemented which selected *Holiday* and *AreaNameTop*. Model4 was trained using only these to varibles along with square transformed re-

sponse variable. Model4 showed a good fit as well as a decent $R^2 = 15.8$. The model4 was as follows:

$$\sqrt{\hat{Accidents}} = -0.42 \cdot Holiday + \beta \cdot AreaNameTop + 6.31 \quad (3)$$

where β is the coefficient vector of $AreaNameTop$.

A full Poisson regression model (Model 5) was fitted to the training dataset. The likelihood ratio test indicated statistically significant explanatory power, with both *Holiday* and several *AreaNameTop* categories significant at the $\alpha = 0.05$ level. However, the goodness-of-fit test suggested that the model did not adequately capture the underlying structure of the data. Despite this limitation, the model exhibited very low dispersion, with a dispersion parameter below 2, indicating that overdispersion was not a major concern.

Subsequently, variable selection was conducted using a forward stepwise procedure, which retained only the *Holiday* predictor. Model 6 was then estimated using this reduced set of variables. The likelihood ratio test again indicated significant explanatory power, but the goodness-of-fit assessment remained unsatisfactory. This result may suggest the presence of nonlinear relationships or interaction effects that were not captured by the model specification. Nonetheless, Model 6 also displayed low dispersion, with a dispersion parameter below 2. Here is the equation for model 6:

$$\log E(Accidents|Holiday) = -0.19 \cdot Holiday + 3.72 \quad (4)$$

Based on the prediction accuracies reported in Table 1, the multiple linear regression model with a square-root transformation of the response variable outperformed both Poisson regression models.

Contrary to findings in previous studies, the results for Los Angeles in 2022 indicate that the square-root–transformed multiple linear regression model provided a better empirical fit than the Poisson regression models.

Building on the state-level analysis showing that structural and environmental conditions, we now focus on our **New York City analysis** to investigate how a specific local structural factor active street construction affects crash counts on individual road segments.

After data cleaning and spatial joins, the analytic dataset contains roughly 89k crashes in 2023 with valid coordinates and a matching street segment. When crashes are categorized according to whether they occurred on segments with active construction, around 8% of crashes in a typical month are on construction segments, with the remaining 92% on non-construction segments. Overall, construction segments account for about eight percent of all crashes, which is non-negligible given that they represent only a subset of the street network. Monthly time-series plots show that crash activity follows the expected seasonal pattern, higher during spring and summer, lower in winter. While the share of crashes happening in construction zones remains relatively stable over the year. Mapping the crash points over the street network confirms that both construction and non-construction crashes are concentrated on the busiest arterial roads and highways in each borough, including corridors such as Broadway, Queens Boulevard and major expressways.

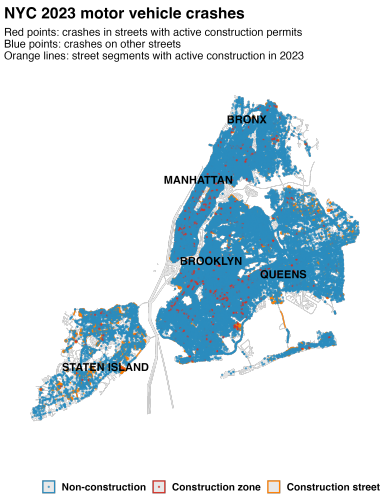

Figure 4: *NYC 2023 Motor vehicle crashes and active constructions*

As shown in Figure 4, maps 2023 crashes over the NYC street network, distinguishing crashes in construction zones, other crashes, and streets with active construction permits. Construction-related crashes are scattered across all boroughs but cluster especially along heavily traveled corridors where both traffic volumes and construction activity are more frequent. The overlay of crashes on the centerline geometry suggests that the geometric joins behave sensibly: crash points snap to nearby segments rather than drifting off the street network, and the spatial clustering seen on the map is consistent with known high-volume routes rather than artefacts of the matching procedure. The Poisson regression confirms a strong positive association between construction status and crash rates, but also highlights why a more flexible model is needed. In the Poisson fit, the coefficient for `in_construction` is about 1.24 and highly statistically significant, corresponding to an estimated rate ratio of approximately exp(1.24), or about 3.5 times as many crashes per kilometer per month on construction segments compared with non-

Table 1: *Prediction Accuracy Results*

Approach	Models	MSE	MAE	MAPE	PM
Multi-city (NY+Chicago+SFO)	Full Poisson	8747.57	55.09	0.207	0.048
	Quasi Poisson	8747.57	55.09	0.207	0.048
	Full Negative Binomial	8505.70	54.86	0.204	0.046
	*Stepwise Negative Binomial	8405.73	54.63	0.203	0.045
City level - Los Angeles	Model 1	31.54	4.12	0.12	0.86
	Model 2	34.47	4.85	0.12	0.94
	Model 3	34.49	4.84	0.12	0.94
	*Model 4	34.20	4.77	0.12	0.93
	Model 5	31.74	4.67	0.12	0.86
	Model 6	31.85	4.59	0.12	0.87
City level - New York	Poisson Regression	40.76	3.43	156.87	-0.02
	Negative Binomial	97.58	6.24	367.55	-0.86
City level - New York	Binomial Regression (1 dataset)	43.59	2.28	167.98	-0.01
	*Binomial Regression (3 datasets)	42.87	2.15	161.46	-0.01

** Chosen model based on prediction accuracies

construction segments of the same length, borough and month. However, the Pearson dispersion factor of roughly 29 indicates that the variance of the crash counts is far larger than the Poisson model allows, so this model is primarily used as a baseline rather than for final inference.

The Negative Binomial model, which uses the same predictors and offset as the Poisson model, provides a much better fit to the data. The estimated coefficient for the construction indicator is about 1.29 with a very small standard error, implying a rate ratio of exp(1.29) = 3.6. In substantive terms, after adjusting for borough, month and segment length, segments with active construction experience a crash rate that is more than three times higher than comparable segments without construction. Borough effects and a few month dummies are also statistically significant, indicating that crash rates are somewhat higher in Brooklyn, Manhattan and Queens relative to the Bronx reference category, and that there are modest seasonal fluctuations even after accounting for construction. The AIC drops from about 199,710 for the Poisson model to about 114,579 for the Negative Binomial model, confirming that allowing for overdispersion substantially improves model fit. Residual diagnostics for the Negative Binomial regression suggest that the model is reasonably well calibrated for this application. Pearson residuals are centered around zero across the range of fitted values, with no strong systematic pattern or curvature, although a small number of segment-months with very high crash counts generate large positive residuals, as expected in a dataset with heavy tails. The dispersion factor is close to one under the Negative Binomial specification, indicating that the extra-dispersion parameter successfully captures the excess variability in the crash counts. A likelihood-ratio test comparing the fitted model to an intercept-only Negative Binomial model produces a test statistic of about 2,399 with a p-value effectively equal to zero, showing that the set of covariates (construction status, borough and month) jointly explain a substantial portion of the variation in segment-month crashes. Cook's distance values are all very small, with no single segment-month exerting undue influence on the estimated coefficients. Taken together, the descriptive and regression results tell a consistent story. Construction zones in New York City are associated with a pronounced increase in crash counts even after accounting for where and when the crashes occur and for differences in segment length. Roughly 8% of all 2023 crashes happen on segments with active construction, and the final Negative Binomial model estimates that, on otherwise similar segments, construction is associated with more than a three-fold increase in the crash rate per kilometer per month. However, test-set prediction metrics (MAE, MAPE and PM) show that both the Poisson and Negative Binomial models have weak predictive accuracy for individual segment months (with the Poisson performing slightly better), so we use the Negative Binomial model primarily to explain average crash-rate differences, such as the impact of construction rather than as a precise forecasting tool.

After creating and training the Binomial Regression models based on the baseline and enhanced datasets, it has been deduced from the model summaries that in the baseline model constructed using only collisions data, the key explanatory variables

are:

- boroughBRONX
- boroughBROOKLYN
- boroughMANHATTAN
- boroughQUEENS

In the enhanced model created using all three datasets (collisions, vehicles and person), the variables with the highest predictive power are

- Motor Vehicle Collision: New York Police Department NYPD [37]
- Street Construct Centerline Data: Office of Technology and Innovation (OTI) [43]
- Centerline Data: City's Office of Technology and Innovation (OTI) [44]

For both models, an α-level of 0.05 has been used.

This means that the baseline model suggests that a fatal accident is more likely to happen in the Bronx, Brooklyn, Manhattan and Queens districts of New York City; the enhanced model, however, indicates that on top of the features mentioned above, any car made before 2010 could contribute to a fatal car crash occurring.

To gain further clarity on this comparison, forward-backward stepwise regression was performed. With an AIC of 2465, the enhanced model is the clear winner. This is further supported by the prediction accuracy values as stated in Table 1.

Explanation of Changes

The **state-level analysis** uses FARS and BEA data instead of the CRSS and US-Accidents datasets to ensure full national coverage and consistent state-level indicators. Some of the initially planned variables were redefined to match the structure and availability of the new datasets.

For **multi-city level** analysis of NY, SFO and Chicago combined, there were no changes in datasets and variables selected. However, dependent variable was changed from monthly crashes to daily crashes. This approach improved analysis by providing higher temporal resolution capturing short term variations in weather changes. This also increased number of data points from 36 (12 observations per city) to 1095 (365 observations per city) improving model stability and predictive power.

On the **Los Angeles city level**, upon further examination, it became evident that the Crash Report Sampling System dataset [48] is provided only in aggregate form, which precludes the extraction of data specific to the Los Angeles region. Consequently, this data set was excluded from the analysis.

On the other hand, the Los Angeles International Airport – Passenger Traffic by Terminal data set [42] was incorporated, as it provides monthly measures of traveler flows through Los Angeles International Airport.

Because the Passenger Traffic by Terminal dataset contains observations only through 2022, the temporal scope of the overall analysis was adjusted accordingly, shifting the reference year from 2023 to 2022.

For **City level - New York City**, the plan was to combine three crash datasets:

- CRSS
- Smoosavi
- NYC Open Data

However, CRSS and Smoosavi only provide national data aggregated by broad regions (e.g., Northeast, Southwest) and do not contain fields that allow me to isolate New York City, so they could not be merged in a NYC-specific way. I therefore revised the plan and replaced CRSS with the NYC DOT Street Construction Permit data and Smoosavi with the NYC Centerline data from the City's Office of Technology and Innovation, harmonizing their different geographic coordinate formats so they could be joined to the NYC crash records.

Upon deeper investigation on the datasets cited for **NYC collisions data vs vehicle + collision data model comparison**, it was determined that the persons data set [49] contained information for the whole state of New York, which could not be joined with the city-level data for vehicles and collisions; crash-level data [50] was also incomplete, adding to complications. Due to these reasons, the data sets were removed from the approach.

The complete datasets were found for both person-level [47] and collision-level data [45] and incorporated into the analysis.

Conclusion

The study aims to identify best predictive model for crash frequencies across various spatial scales in US, using multi-city and single city datasets and to understand explanatory dynamics behind differences among states. Future work could extend this analysis by incorporating different spatial scales such as county-level within states, with new variable subsets by possibly using machine learning approaches for counts such as Random Forest and Gradient Boosting.

It is important to note that, for the purpose of keeping the models interpretable, several variables and locations were omitted, which may limit the

generalization of the results to other locations and contexts.

References

[1] Isabelle Fallon and Desmond O'Neill. "The world's first automobile fatality". In: *Accident Analysis & Prevention* 37.4 (2005), pp. 601–603. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2005.02.002. URL: https://www.sciencedirect.com/science/article/pii/S0001457505000369.

[2] Guinness World Records. "First person killed by a car". In: (2024).

[3] Richard Weingroff. "A Moment in Time: Highway Safety Breakthrough". In: (2021). URL: https://www.fhwa.dot.gov/highwayhistory/moment/highway_safety_breakthrough.cfm.

[5] World Health Organization. *Global status report on road safety 2023*. Licence: CC BY-NC-SA 3.0 IGO. Geneva, Switzerland: World Health Organization, 2023. URL: https://www.who.int/publications/i/item/9789240086517.

[6] National Center for Statistics and Analysis. *Overview of motor vehicle traffic crashes in 2023*. Traffic Safety Facts Research Note DOT HS 813 705. Washington, DC, United States: National Highway Traffic Safety Administration, Apr. 2025. URL: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813705.

[7] D. Y. Lu. "Prediction of road traffic accident quantity: multi factor regression analysis". In: *Advances in Transportation Studies* (2025), pp. 125–136. DOI: 10.53136/979122182056010. URL: https://research.ebsco.com/c/i2q7gb/search/details/qzvpcbxmn5?db=a9h&limiters=None&q=traffic+accidents&searchMode=boolean&isEbscoSignIn=false.

[8] E. Bakış et al. "Prediction of traffic accidents trend with learning methods: a case study for Batman, Turkey". In: *Scientific Reports* 15.1 (2025), pp. 1–22. DOI: 10.1038/s41598-025-11835-9. URL: https://doi.org/10.1038/s41598-025-11835-9.

[9] Shaw-Pin Miaou and Harry Lum. "Modeling vehicle accidents and highway geometric design relationships". In: *Accident Analysis & Prevention* 25.6 (1993), pp. 689–709. ISSN: 0001-4575. DOI: https://doi.org/10.1016/0001-4575(93)90034-T. URL: https://www.sciencedirect.com/science/article/pii/000145759390034T.

[10] S. C. Joshua and N. J. Garber. "Estimating truck accident rate and involvements using linear and Poisson regression models". In: *Transportation Planning and Technology* 15.1 (1990), pp. 41–58. DOI: 10.1080/03081069008717439.

[11] Mohamed A. Abdel-Aty and A.Essam Radwan. "Modeling traffic accident occurrence and involvement". In: *Accident Analysis & Prevention* 32.5 (2000), pp. 633–642. ISSN: 0001-4575. DOI: https://doi.org/10.1016/S0001-4575(99)00094-9. URL: https://www.sciencedirect.com/science/article/pii/S0001457599000949.

[12] John Milton and Fred Mannering. "The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies". In: *Transportation* 25.4 (1998), pp. 395–413. DOI: 10.1023/A:1005095725001.

[13] National Safety Council. *Historical Car Crash Deaths and Rates*. https://injuryfacts.nsc.org/motor-vehicle/historical-fatality-trends/deaths-and-rates/. 2025.

[14] iRAP. *Latest Data Provides New Safety Insights*. https://irap.org/2024/07/latest-data-provides-new-safety-insights/. 2024.

[15] D. Lord, S. Washington, and J. Ivan. "Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory — compares count models for crash frequency". In: *Accident Analysis & Prevention* (2005). URL: https://www.sciencedirect.com/science/article/abs/pii/S0001457504000521.

[16] N. V. Malyshkina and F. L. Mannering. "Empirical assessment of the impact of highway design characteristics on accident frequencies — discusses Poisson/NegBin approaches for accident counts". In: *Accident Analysis & Prevention* (2010). URL: https://www.sciencedirect.com/science/article/abs/pii/S000145750900178X.

[17] J. Ma and K. Kockelman. "Bayesian multivariate Poisson regression for models of injury counts by severity, with an application to intersection crashes — multivariate Poisson for intersection crash counts". In: *Transportation Research Record* (2006). URL: https://www.caee.utexas.edu/prof/kockelman/public_html/TRB06MVPBayesian.pdf.

[18] Shraddha Sagar, Nikiforos Stamatiadis, and Arnold Stromberg. "Effect of Socioeconomic and Demographic Factors on Crash Occurrence". In: *Transportation Research Record* 2675.12 (2021), pp. 80–91.

DOI: 10.1177/03611981211027887. URL: https://journals.sagepub.com/doi/10.1177/03611981211027887.

[19] Mojtaba Sehat et al. "Socioeconomic Status and Incidence of Traffic Accidents in Metropolitan Tehran: A Population-based Study". In: *International Journal of Preventive Medicine* 3.3 (2012), pp. 181–190. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3309632/pdf/IJPVM-3-181.pdf.

[20] Marie Hasselberg and Lucie Laflamme. "Socioeconomic Background and Road Traffic Injuries: A Study of Young Car Drivers in Sweden". In: *Traffic Injury Prevention* 4.3 (2003), pp. 249–254. ISSN: 1538-9588 (Print), 1538-957X (Online). DOI: 10.1080/15389580309882. URL: https://www.tandfonline.com/doi/epdf/10.1080/15389580309882?needAccess=true.

[21] Christina L. Hanna et al. "Road traffic crash circumstances and consequences among young unlicensed drivers: A Swedish cohort study on socioeconomic disparities". In: *BMC Public Health* 10.14 (2010). DOI: 10.1186/1471-2458-10-14. URL: https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-10-14.

[22] Eyob Getachew et al. "Socioeconomic and Behavioral Factors of Road Traffic Accidents among Drivers in Ethiopia: Systematic Review and Meta-Analysis". In: *BMC Public Health* 24.2857 (2024). DOI: 10.1186/s12889-024-20376-1. URL: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-024-20376-1.

[23] Feng Guo et al. "The effects of age on crash risk associated with driver distraction". In: *International Journal of Epidemiology* 46.1 (2017). Peer-reviewed article, pp. 258–265. DOI: 10.1093/ije/dyw234. URL: https://academic.oup.com/ije/article/46/1/258/2418691.

[24] Ilan Shrira and Kenji Noguchi. "Traffic fatalities of drivers who visit urban and rural areas: An exploratory study". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 41 (2016), pp. 74–79. ISSN: 1369-8478. DOI: 10.1016/j.trf.2016.05.003. URL: https://www.sciencedirect.com/science/article/pii/S1369847816300614.

[25] Laurie Beck et al. "Rural and Urban Differences in Passenger-Vehicle–Occupant Deaths and Seat Belt Use Among Adults—United States, 2014". In: *MMWR Surveillance Summaries* 66.SS-17 (2017), pp. 1–13. DOI: 10.15585/mmwr.ss6617a1.

[26] C Cabrera-Arnau, R P Curiel, and S R Bishop. "Uncovering the behaviour of road accidents in urban areas". In: *Royal Society open science* 7.191739 (2020). DOI: 10.1098/rsos.191739. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7211831/.

[27] Johan Bjureberg and James J. Gross. "Regulating road rage". In: *Social and Personality Psychology Compass* 15.3 (2021), e12586. DOI: 10.1111/spc3.12586. URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/spc3.12586.

[28] J. Ivan and P. O'mara. "Prediction of traffic accident rates using Poisson regression". In: *76th annual meeting of the transportation research board*. 970861. 1997.

[29] Md.Kamrul Khan and Md. Tarek Hasan. "A POISSON REGRESSION APPROACH TO MODELING TRAFFIC ACCIDENT FREQUENCY IN URBAN AREAS". In: *American Journal of Interdisciplinary Studies* 3.04 (Dec. 2022), pp. 117–156. DOI: 10.63125/wqh7pd07. URL: https://ajisresearch.com/index.php/ajis/article/view/36.

[30] S. O. Mohammed, M. O. Rahma, and F. Dweiri. "Unravelling the veil of traffic safety: a comprehensive analysis of factors influencing crash frequency across US states". In: *Transportation Safety and Environment* 6.4 (2024). DOI: 10.1093/tse/tdae016. URL: https://doi.org/10.1093/tse/tdae016.

[31] Imran Ashraf et al. "Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis". In: *PLOS ONE* 14.10 (Oct. 2019), pp. 1–29. DOI: 10.1371/journal.pone.0223473. URL: https://doi.org/10.1371/journal.pone.0223473.

[32] Dominique Lord and Fred Mannering. "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives". In: *Transportation Research Part A: Policy and Practice* 44.5 (2010), pp. 291–305. ISSN: 0965-8564. DOI: https://doi.org/10.1016/j.tra.2010.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0965856410000376.

[33] World Health Organization. *Global Plan for Road Safety 2021–2030*. https://cdn.who.int/media/docs/default-source/documents/health-topics/road-traffic-injuries/global-plan-for-road-safety.pdf. 2021.

References - Data Sources

[34] National Highway Traffic Safety Administration. *Fatality Analysis Reporting System (FARS), 2023 National Dataset*.

[35] U.S. Department of Transportation, Federal Highway Administration. *Highway Statistics Series, Table VM-2: Vehicle Miles of Travel by Functional System*. https://www.fhwa.dot.gov/policyinformation/statistics.cfm. 2023.

[36] Bureau of Economic Analysis. *State personal income and related data, 2023*.

[37] New York City Police Department (NYPD). *Motor Vehicle Collisions - Crashes*. 2023, 2025. URL: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95.

[38] Chicago Police Department. *Traffic Crashes – Crashes*. 2025. URL: https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/data_preview.

[39] San Francisco Department of Public Health (SFDPH) and San Francisco Police Department (SFPD). *Traffic Crashes Resulting in Injury*. 2025. URL: https://data.sfgov.org/Public-Safety/Traffic-Crashes-Resulting-in-Injury/ubvf-ztfx.

[40] Los Angeles Police Department (LAPD) OpenData. *Traffic Collision Data from 2010 to Present*. Data last modified 2025-03-14; metadata updated 2025-03-22. 2025. URL: https://catalog.data.gov/dataset/traffic-collision-data-from-2010-to-present.

[41] City of Los Angeles, StreetsLA. *2022 Official Holidays Calendar*. https://streetsla.lacity.org/sites/default/files/2022_OfficialHolidays_Calendar.pdf. Accessed: YYYY-MM-DD. 2022.

[42] Department of Transportation City of Los Angeles. *Los Angeles International Airport – Passenger Traffic By Terminal*. Data last modified 2023-11-30. :contentReference[oaicite:0]index=0. data.lacity.org, 2023. URL: https://data.lacity.org/Transportation/Los-Angeles-International-Airport-Passenger-Traffi/g3qu-7q2u.

[43] New York City Department of Transportation (DOT). *Street Construction Permits (2022–Present)*. Permits for roadway and sidewalk construction in New York City. 2023. URL: https://data.cityofnewyork.us/Transportation/Street-Construction-Permits-2022-Present-/tqtj-sjs8/about_data.

[44] New York City Office of Technology and Innovation (OTI). *NYC Street Centerline (LION)*. Single-line street centerline network for New York City. 2023.

[45] Kaggle. *Motor Vehicle Collisions - Crashes (NYC)*. 2023. URL: https://www.kaggle.com/datasets/adelanseur/motor-vehicle-collisions-crashes.

[46] NYC OpenData. *Motor Vehicle Collisions - Vehicles*. 2023. URL: https://catalog.data.gov/dataset/motor-vehicle-collisions-vehicles.

[47] New York City Police Department (NYPD). *Motor Vehicle Collisions - Person (NYC)*. 2023. URL: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu.

[48] U.S. Department of Transportation, National Highway Traffic Safety Administration. *Crash Report Sampling System (CRSS), 2023 Annual File*. https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system. 2023.

[49] NYS Department of Motor Vehicles (DMV). *Motor Vehicle Crashes - Individual Information: Three Year Window*. 2023. URL: https://data.ny.gov/Transportation/Motor-Vehicle-Crashes-Individual-Information-Three/ir4y-sesj.

[50] New York Police Department. *Motor Vehicle Collisions - Crashes*. 2023. URL: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95.

Appendix: Team Contributions

Table below shows each individual team members' contribution to the analysis.

Analysis Approach	Assignee
Exposure-Adjusted State-Level Modeling of Fatal Crash Frequency	Gulnaz Javadova
Multi-city level analysis for NY, Chicago and SFO combines	Mehak Zahid
Comparison of Multi Linear Regression and Poisson regression in analyzing daily traffic accident counts in Los Angeles for the year 2022	Neerav Nemchand Gala
New York City Traffic accidents in construction vs Non-construction zones	Ahmadou Z Diallo
Single city analysis for NY	Fatima M. Athar

Table 2: *Contributions to Analysis*

Group tasks—such as report formatting, table construction, and related responsibilities—were distributed evenly among all team members.