Applying Big Data Analytics on Motor Vehicle Collision in New York: Insights and Reasons

*

1st Sachin Garg
Data Science Department
CUNY Graduate Center
New York City, USA
Sgarg1@gradcenter.cuny.edu

2nd Garima Goyal
Data Science Department
CUNY Graduate Center
New York City, USA
Ggoyal@gradcenter.cuny.edu

***Abstract*—Motor vehicle collisions pose significant challenges to urban safety and transportation systems. This project employs advanced Big Data Analytics techniques to comprehensively analyze motor vehicle collision accidents in New York City (NYC). The vast dataset, spanning several years, is sourced from NYC's official records, providing a wealth of information on accident locations, types, contributing factors, and severity. The primary objectives of this analysis are to identify patterns, trends, and key factors influencing the frequency and severity of collisions. Leveraging cutting-edge data analytics tools and methodologies, we delve into the intricate details of the dataset. Techniques such as statistical analysis, regression analysis, and machine learning algorithms are applied to unveil hidden insights within large and complex dataset. The report explores temporal trends, examining how collision patterns vary across different times of the day, days of the week, and seasons. Spatial analyses focus on identifying high-risk zones and understanding the impact of geographical factors on collision occurrences. Moreover, the study investigates the influence of contributing factors such as weather conditions, road conditions, and nature-related factors on collision severity.**

I. Introduction

In this comprehensive report, we delve into the intricate analysis of vast sets of big data pertaining to motor vehicle collisions within the dynamic landscape of New York City (NYC). The primary objective is to extract meaningful insights that shed light on the key determinants of accidents, thereby facilitating the identification of precise and targeted prevention strategies. The dataset utilized for this analysis comprises a staggering over 2 million records sourced from open data portals, providing a rich and expansive foundation for our investigation.

Our overarching goals encompass a thorough understanding of the diverse causes underlying these collisions, deciphering intricate patterns that may emerge, pinpointing high-risk areas within the city, and, most crucially, formulating actionable recommendations to curtail the frequency and severity of motor vehicle collisions in NYC.

The enormity of the dataset enables us to conduct a nuanced examination of the multifaceted aspects of motor vehicle collisions, transcending mere statistical scrutiny. Through a meticulous analysis of this information, we aim to contribute significantly to the enhancement of public safety by offering insights that go beyond the surface-level examination of accidents. Our approach involves not only identifying trends but also delving into the root causes, allowing for the development of targeted interventions.

As we progress through this report, we will navigate through the intricacies of the data, unveiling noteworthy findings that will serve as the foundation for informed decision-making. By assimilating and interpreting this wealth of information, we aspire to provide city planners, policymakers, and relevant stakeholders with actionable intelligence that can be leveraged to implement effective preventive measures, ultimately fostering a safer and more secure transportation environment in New York City.

II. Data Collection, Cleaning and Preprocessing

The motor vehicle collision dataset for New York City comprised a voluminous 2040203 rows, encompassing a wealth of information across 29 attributes. These attributes ranged from fundamental details such as date, time, and location to more nuanced elements including the number of persons injured, detailed vehicle specifications, and an array of contributing factors. This expansive dataset, while rich in potential insights, presented inherent challenges that necessitated a rigorous approach to data cleaning and preprocessing.

In the initial phase, our data cleaning efforts were directed towards addressing a variety of issues, outliers, typos, and inconsistencies. Recognizing the importance of cohesive data representation, we executed transformations that consolidated similar variables into unified categories. This strategic move not only streamlined the dataset but also facilitated a more coherent and insightful analysis. Special attention was directed towards factors contributing to accidents and vehicle types, which underwent a standardization process. We observed some typos/misspellings ('Illnes', 'Illness'), ('Cell Phone (hand-Held)', 'Cell Phone (hand-held)') and invalid data such as ('80', '1') in the 'CONTRIBUTING FACTOR' column. There

Fig. 1. Outliners

were some data inconsistencies as well such as ('Fell Asleep', 'Drowsy', Lost Consciousness) which were merged into one category 'Fatigued/ Drowsy/ Sleep/ Unconscious'.This standardization not only enhanced the clarity of the dataset but also laid the groundwork for effective pattern identification, a pivotal aspect of our analytical endeavors.

A fuzzywuzzy library was introduced to encounter the inconsistencies and invalid data in the 'VEHICLE TYPE CODE' columns. Fuzzywuzzy is a Python library that provides a simple way to compare strings and determine their similarity using 'Levenshtein distance'. Although fuzzy-wuzzy was not enough for data cleaning, we performed some manual tasks using '.loc' method for better accuracy. Furthermore, location attributes, including street names, underwent meticulous standardization. This involved the removal of special characters (using 'regular expressions'), the expansion of abbreviations, and the imposition of a consistent format (using 'str.title().str.strip()'). By undertaking these measures, we aimed to eradicate inconsistencies, ensuring a uniform and reliable representation of location data. Such standardization not only contributes to the accuracy of our analyses but also streamlines the interpretation of results for stakeholders and decision-makers.

In essence, the rigorous data cleaning and preprocessing efforts undertaken in this phase not only fortified the dataset against potential distortions but also laid the groundwork for a more robust and insightful analysis of motor vehicle collisions in New York City. The resulting refined dataset sets the stage for meaningful exploration, allowing us to extract nuanced patterns and actionable insights to guide targeted preventive strategies.

III. Dealing with Missing Values

Missing values in a dataset refer to the absence of information or data points for specific variables or observations. These gaps in the dataset can arise for various reasons, such as errors in data collection, incomplete records, or simply because certain information was not available at the time of recording. To remove the inaccuracy and unreliability of the insights derived from the data, it becomes essential to employ techniques such as imputation, where missing values are estimated or filled in based on existing data, to ensure a more comprehensive

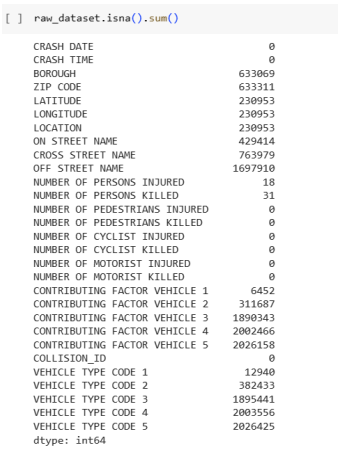

Fig. 2. Missing values

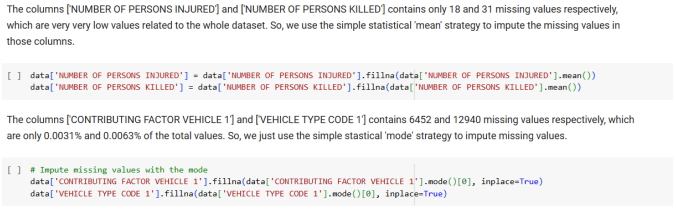

Fig. 3. Not true missing values

and meaningful analysis. Proper handling of missing values is essential to maintain the integrity and reliability of the dataset for meaningful statistical and machine-learning analyses.

Approximately 0.5% of the records exhibited such anomalies, and our meticulous cleanup procedures aimed to rectify these discrepancies to ensure the integrity of the dataset. To mitigate the impact of missing values, we employed a combination of statistical techniques and machine learning algorithms for imputation, ensuring a robust foundation for subsequent analyses. Out-liners, we can observe some typos/misspelling ('Illnes', 'Illness'), ('Cell Phone (hand-Held)', 'Cell Phone (hand-held)'), data inconsistency ('Fell Asleep', 'Drowsy', Lost Conciousness') and invalidation ('80', '1'). We had dealed with these out liners to avoid Null values and efficient dataset.

The columns ['NUMBER OF PERSONS INJURED'] and ['NUMBER OF PERSONS KILLED'] exhibit minimal missing values, with only 18 and 31 instances, respectively, which constitute a negligible proportion of the entire dataset. Consequently, we opted for the straightforward statistical 'mean' strategy to address the missing values in these columns. Similarly, the 'mode' strategy was applied to handle the 0.0031% and 0.0063% of missing values in columns ['CONTRIBUTING FACTOR VEHICLE 1'] and ['VEHICLE TYPE CODE 1']. While the columns ['CONTRIBUTING FACTOR VEHICLE 2/3/4/5'] and ['VEHICLE TYPE CODE 2/3/4/5'] feature a substantial number of missing values, it's essential to note that these aren't genuine missing values. The rationale

is that when only one vehicle is involved in an accident, vehicle codes 2/3/4/5 are irrelevant. Similarly, if the first two vehicles are involved, vehicle codes 3/4/5 are inapplicable, and this pattern continues. Therefore, we categorize these missing values as 'Not Applicable'.

IV. Insights

A. Contributing Factors and Vehicle Types

The visualization Fig 4, depicts prevalent collision types, emphasizing preventable incidents through driver attentiveness and adherence to road rules. The primary collision type is rear-end collisions, constituting 35% of all incidents, followed by sideswipe collisions (15%), turning collisions (10%), and head-on collisions (10%). These collisions, such as rear-end, sideswipe, and turning incidents, frequently result from drivers neglecting their surroundings or failing to signal their intentions. Head-on collisions are often attributed to drivers traversing the wrong side of the road or engaging in illegal passing maneuvers.

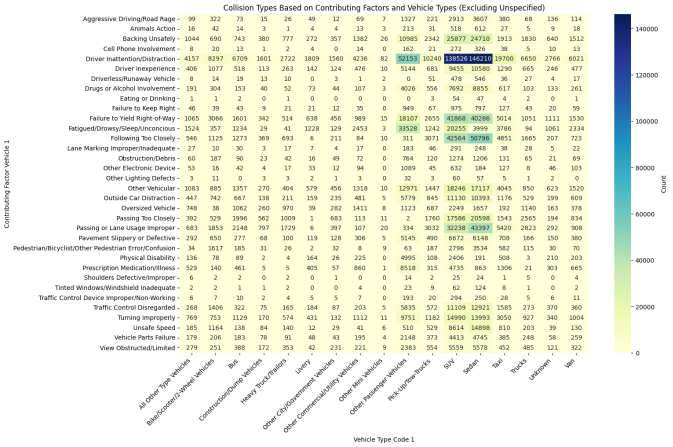

Fig. 4. Contributing Factors Vs Vehicle Types

Additionally, the graph underscores that the leading contributing factors to collisions are within the purview of drivers' choices. Factors such as driver inattention/distraction, following too closely, and failure to yield right-of-way are indicative of decisions made by drivers that can significantly impact collision outcomes.

According to Trottenberg (2019), the surge in roadway fatalities in New York City and across the nation is being fueled by the growing popularity of SUVs and light trucks. According to an article by Coburn in StreetsBlogNYC, the state Department of Motor Vehicles concluded that city residents witnessed a significant 21 percent increase in the ownership of suburbans from 2016 to 2020, while, during the same period, the count of sedans in the five boroughs experienced a notable number of 807,482. Furthermore, our analysis of vehicle types revealed similar trends as both SUVs and sedans shared (60%) involvement in accidents, followed by trucks (20%), and other passenger vehicles (10%). This breakdown sheds light on the prevalent vehicle types contributing to collision incidents.

B. Distribution of Top 10 Contributing Factors

Fig 5 shows that driver inattention/distraction is the leading contributing factor for accidents, accounting for nearly 40% of all accidents. This is followed by traffic control disregarded, passing too closely, backing unsafely, and fatigue/drowsy/sleep/unconscious.

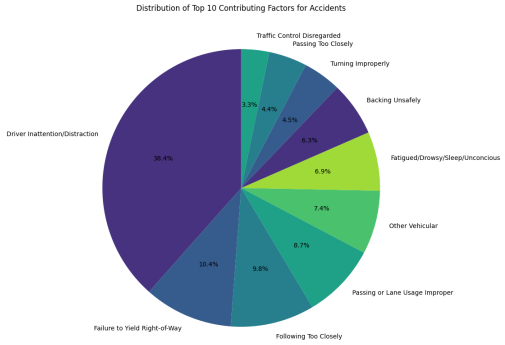

Fig. 5. Top 10 Contributing Factors

According to Forbes Advisor, distracted driving played a significant role in New York City collisions in 2022, with a total of 28,753 vehicles involved in accidents due to driver inattention or distraction (Bieber). The prevalence of smartphone usage while driving, such as texting or using mobile applications, has been a major concern. Studies have shown that engaging in distracted activities diverts the driver's attention from the road, such as looking at other crashes off the road, being distracted from other passengers, etc.

Educational campaigns and stricter enforcement of laws related to mobile phone use while driving are essential components of addressing this issue. Additionally, advancements in technology, like hands-free options and driver-assistance systems, can be explored as potential solutions to mitigate the impact of distracted driving on road safety.

C. Distribution of Accidents Across Boroughs in NYC

In Figure 6, the data reveals a notable disparity in the number of accidents across New York City boroughs, with Brooklyn registering the highest incidence, followed by Queens, Manhattan, the Bronx, and Staten Island. Brooklyn alone accounts for nearly 30% of all accidents, with Queens representing 25%, Manhattan 20%, the Bronx 15%, and Staten Island 10%. Based on the 2020 census data, Brooklyn emerges as the most populous borough in New York City, representing 31% of the total population with 2.7 million residents. Queens follows closely behind, constituting 27% of the population with 2.4 million individuals. Manhattan accounts for 19%, the Bronx for 17%, and Staten Island has the smallest share at 6%. These population statistics underscore a clear correlation – "the higher the population density, the greater the likelihood of accidents occurring".

Moreover, the unique blend of mixed commercial and residential development in Brooklyn and Queens results in a diverse array of vehicles on the road, including cars, trucks,

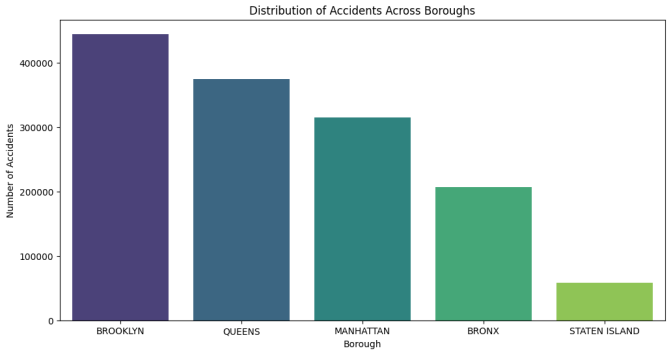

Fig. 6. Accidents Across Boroughs in New York City

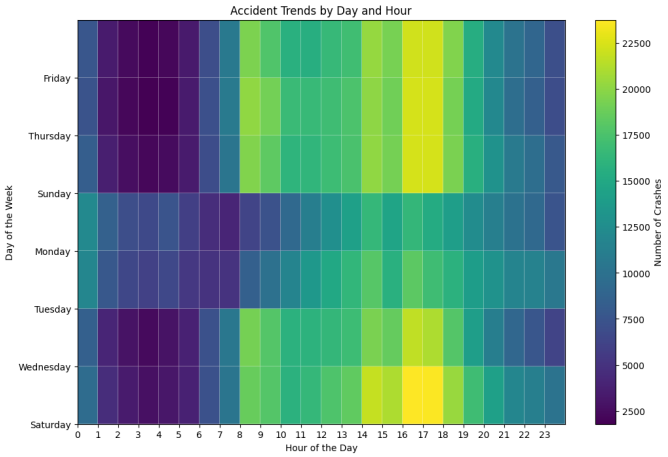

Fig. 7. Accident Trends by Day of Week and Hour of Day

buses, and bicycles. The statistical analysis of accidents reflects the impact of this varied mix, intensifying the complexity of traffic interactions and contributing to an elevated risk of accidents. The intricate transportation networks in these boroughs, characterized by a diversity of road types, intersections, and traffic patterns, correlate with statistical evidence pointing to the challenging road environment that heightens the likelihood of accidents.

Brooklyn, acting as a pivotal nexus for major roads, including expressways and arterial routes, exhibits statistical data that underscores its role as a hub for significant traffic convergence. This contributes to higher accident rates, as reflected in the data. In Queens, statistical indicators demonstrate the impact of economic activity and commercial centers, such as JFK International Airport, amplifying the traffic volume and resulting in statistically higher accident frequencies.

D. Accident Trends by Day of Week and Hour

The observed trends in the graph (fig 7) can be attributed to various factors. Firstly, Fridays and Saturdays stand out as days with higher accident rates, likely due to increased social activities and alcohol consumption during these days. The combination of people being out socializing, and potentially consuming alcohol elevates the risk of accidents during these periods.

Secondly, the afternoon rush hour between 4 pm and 7 pm exhibited an increased risk of accidents, as did Friday and Saturday nights. This heightened risk during the afternoon rush hour can be attributed to the high volume of commuters on the roads, potentially leading to congestion, increased stress, and a greater likelihood of accidents.

Additionally, the elevated accident risk on Friday and Saturday nights is likely linked to individuals driving home from bars and clubs during the late hours. This time frame is associated with factors such as impaired driving due to alcohol consumption and reduced visibility, both of which contribute to a higher likelihood of accidents.

Thirdly, the gradual decrease in the number of accidents throughout the day may be linked to fatigue. Towards the end of the day, people tend to be more tired, and fatigue can impair their ability to drive safely. This fatigue-related decline in alertness and reaction times could explain the diminishing trend in accident numbers as the day progresses.

E. Distribution by Top 10 Dangerous Streets

The comprehensive analysis of total severity (injuries + fatalities) of accidents across the top 10 streets offers vital insights into the distribution of severe cases, necessitating a closer examination of each thoroughfare. Topping the list is Belt Parkway, registering an alarming 8000+ severe accidents. The heightened incidence on the Belt Parkway necessitates an in-depth investigation into factors contributing to the severity, be it high traffic volumes, complex intersections, or specific road conditions. Notably, the Belt Parkway has faced chronic congestion since the 1960s, driven by its crucial role in connecting JFK International Airport to Long Island, Queens, and Brooklyn, including Manhattan. Despite running through a narrow green belt, it traverses densely populated areas, adding complexity to traffic conditions.

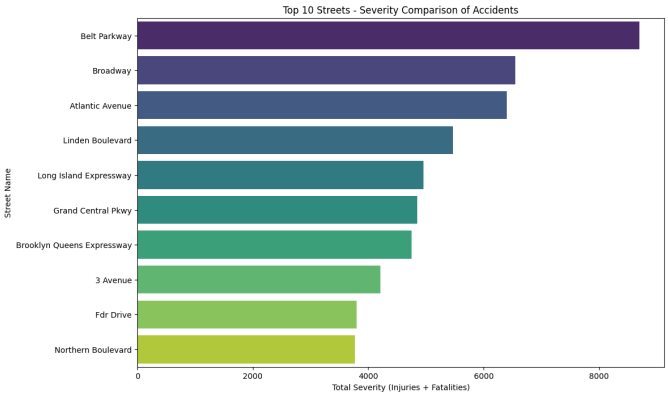

Fig. 8. Top 10 Dangerous Streets Throughout NYC

Following closely are Broadway and Atlantic Avenue, securing the second and third positions with approximately 6700 and 6500 severe cases, respectively. These figures highlight the

pressing need for targeted safety measures on these streets, potentially addressing challenges such as heavy pedestrian traffic, intricate road layouts, or concentrated commercial activities. Broadway, spanning the entire length of Manhattan at 33 miles, naturally incurs a significant number of motor vehicle accidents. The constant updates, renovations, and ongoing construction projects contribute to the complexity of traffic patterns on Broadway, potentially leading to accidents.

Linden Street occupies the fourth spot with around 5700 severe cases, prompting a focused analysis of road design, traffic flow, and potential contributing factors to formulate effective safety interventions.

In conclusion, the statistical analysis, combined with contextual insights, highlights the urgency of tailored safety interventions on these top 10 streets. A comprehensive examination of each street's specific characteristics and challenges is essential to implement effective measures aimed at reducing the severity of accidents and enhancing overall road safety.

F. Number of Accidents by Month and Year in NYC

The heat map below underscores the elevated frequency of motor vehicle collisions in specific months, notably May, June, and October, within New York City. This pattern is influenced by a combination of seasonal, environmental, and societal factors. The spring months of May and June witness improving weather conditions, prompting increased outdoor activities and heightened traffic, thereby elevating the potential for accidents. These months also experience a surge in outdoor events, festivals, and recreational activities, contributing to heightened traffic volumes and accident rates.

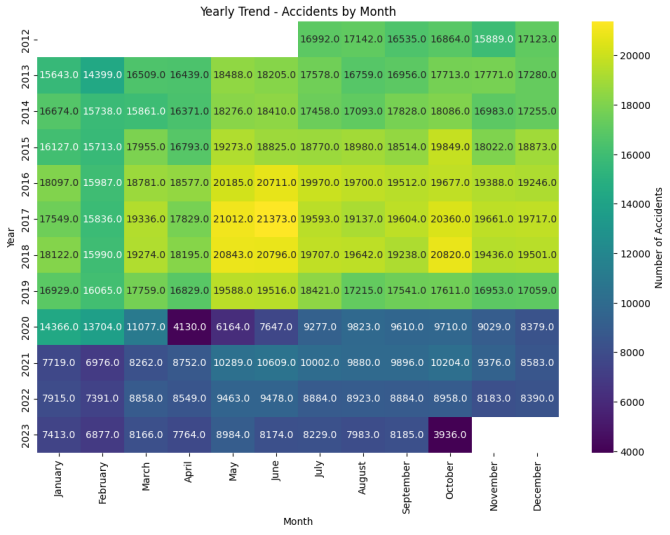

Fig. 9. Accidents Trends by Year and Month

May and June coincide with the conclusion of the school year, resulting in increased school-related activities, and events, and heightened parental involvement in driving their children, further amplifying traffic volumes and accident likelihood. Additionally, the favorable weather during these months leads to heightened construction and road maintenance activities, contributing to traffic congestion and potential accidents in construction zones.

October, situated in the fall, similarly experiences increased traffic due to favorable weather conditions and heightened engagement in outdoor activities and events. Fall-related festivals and activities attract larger crowds, leading to a surge in vehicular traffic and an associated rise in accident rates. Moreover, October marks the commencement of holiday season preparations, contributing to increased shopping, events, and travel, particularly around holidays like Halloween, thereby heightening the likelihood of traffic incidents.

G. Accident Trends by Yearly Basis

The line chart depicting yearly trends in the number of accidents reveals intriguing patterns over the examined period. From summer 2012 to 2020, the line consistently fluctuated between 17,000 to 19,000 accidents annually in NYC, indicating a stable baseline influenced by typical traffic patterns and seasonal variations. However, a notable deviation occurred in March-April 2020, recording a sharp decline to approximately 2700 accidents in NYC. This abrupt reduction aligns with the onset of the COVID-19 pandemic, as lockdowns and stay-at-home orders led to a substantial decrease in vehicular traffic and subsequent accidents.

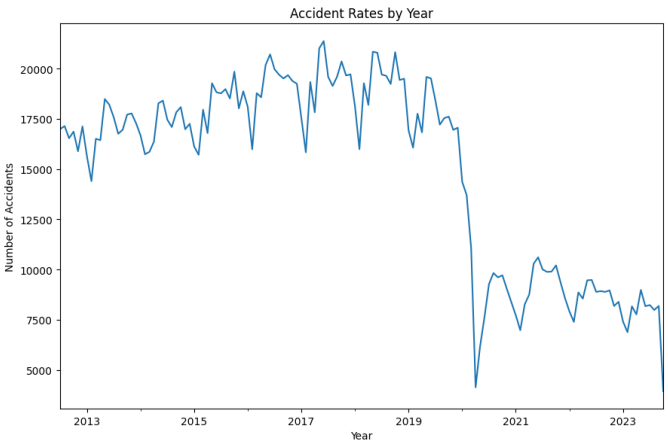

Fig. 10. Accidents Trend by Yearly

Post-October 2020, the line exhibits consistent up-and-down fluctuations, ranging between 7500 to 9500 accidents per year. This period likely reflects the gradual return to normalcy as pandemic-related restrictions eased. The resurgence in accident numbers correlates with increased mobility, the resumption of economic activities, and a return to regular traffic volumes. With more people working from home, there has been a noticeable reduction in daily commuting in this period. Fewer individuals on the roads during typical rush hours contribute to lower traffic volumes and, consequently, a potential decrease in accidents. Dynamic factors such as changes in commuting patterns, business reopenings, and shifts in public behavior contribute to the observed trends.

In summary, the line chart captures the impact of broader societal events, particularly the influence of the COVID-19 pandemic on traffic patterns. The fluctuations in accident numbers underscore the dynamic nature of road safety, emphasizing the importance of adaptive measures and traffic management strategies in response to evolving circumstances.

H. Analyzing Total Number of Deaths on Yearly Basis

In this comprehensive analysis Figure 11, we delve into the intricate landscape of motor vehicle crashes in New York City, focusing specifically on the critical aspect of human toll – fatalities. By scrutinizing trends over the years, we seek valuable insights that not only shed light on the severity of road safety challenges but also pave the way for informed interventions and policy considerations. Despite a decrease in traffic volume on the roads amid the COVID-19 pandemic, the number of fatalities(deaths) resulting from motor vehicle collisions in New York City showed unaffected changes during and post-2020.

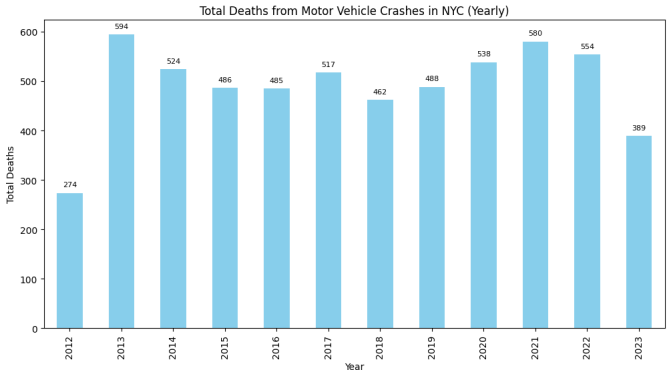

Fig. 11. Total Number of Deaths on Yearly Basis

Crashes resulting in severe or fatal injuries, specifically linked to speeding, experienced a notable rise. This increase is primarily associated with the reduced traffic volume resulting from lockdown measures and the adoption of teleworking, creating conditions conducive to speeding. According to Kim Royster, the transportation chief of the Police Department, drivers seized the opportunity of open roads to speed with their vehicles. ”Visibility is very important when it comes to traffic enforcement, especially for speeding and drag-racing drivers” (Royster). However, it’s crucial to acknowledge that additional factors such as mental health issues like depression, emergency response delays due to increased patient hospitalization and the heightened busyness of emergency vehicles, and other factors may also contribute to high-risk driving behavior during the pandemic.

V. Implications and Recommendations

Addressing highway safety has been a longstanding concern, dating back to the early days of automobiles. Notably, the first recorded automobile accident took place in 1896, predating the birth of anyone in this course.

In 1924, at the National Conference on Street and Highway Safety led by then-Secretary of Commerce Herbert Hoover, the federal government conceptualized a comprehensive approach to enhance highway safety. This included initiatives like traffic safety education, public relations efforts, traffic control, and the development of a standardized vehicle code. The prevailing belief was that, given the technology of the time, highways and vehicles were constructed as well as possible, and the majority of accidents were attributable to ”willful, careless, irresponsible, or incompetent drivers.”

The Vision Zero initiative, launched in 2014 with the aim of reducing crash risks and eliminating motor vehicle fatalities, has led to a decline in traffic deaths, according to Forbes Advisor (Bieber). The Vision Zero establishment includes ”the three E’s” as the standard for a well-rounded traffic safety program, integrating these three E’s, representing Education, Engineering, and Enforcement, into educational materials to address the dynamic interplay between motorists, vehicles, and roadways.

Breaking down the three E’s:

Education: This element focuses on the human aspect, incorporating driver education, improvement programs, and Public Safety Announcements as integral components of the educational strategy for promoting traffic safety.

Engineering: Involves the application of scientific and technological principles to enhance safety. Examples include the implementation of safety features such as Jersey barriers, breakaway signs, rumble strips, crash barrels, and seat belts, all stemming from advancements in traffic safety engineering.

Enforcement: Encompasses the development and execution of traffic safety laws. This involves activities such as the issuance of traffic citations for violations, enforcement of DWI laws, assistance in traffic routing, and the conduct of crash investigations. These efforts are carried out by law enforcement professionals in New York as part of their broader enforcement initiatives.

VI. Conclusion

In conclusion, the utilization of advanced analytical techniques has played a pivotal role in transforming vast data-set into actionable intelligence. By leveraging big data, authorities and policymakers have been able to glean valuable insights that contribute to saving lives, preventing injuries, and promoting road safety. The continuous analysis and incorporation of diverse data-set ensure a holistic evaluation of safety measures, thereby enabling more informed decision-making to further enhance the effectiveness of interventions aimed at improving overall road safety in NYC.

Importantly, this project highlights the immense potential of data-driven solutions for tackling complex urban issues. As more connected devices and sensors are deployed throughout New-York infrastructure, there is an opportunity to integrate even more data sources to uncover hidden patterns and trends. Our visualizations and analytical approach could be expanded to model a wider range of risk factors over time. Bringing additional context around weather, congestion zones, public

transit flows, and land use patterns may reveal further nuances into the interplay of elements that contribute to collisions. Applying similar techniques to other urban domains like public health, the environment, housing, and education can accelerate knowledge discovery across the public sector. Ultimately, nurturing a culture and capability for evidence-based governance promises to uplift communities by optimizing the use of finite resources where they are needed most. The insights gleaned so far are only the tip of the iceberg when it comes to leveraging NYC's data for the public good.

REFERENCES

[1] Bieber, Christy. "NYC Car Accident Statistics In 2023." *Forbes Advisor*, 30 June 2023, https://www.forbes.com/advisor/legal/nyc-car-accident-statistics/. Accessed 18 November 2023.

[2] Coburn, Jesse. "DRIVING US MAD: Number of SUVs and Similar Cars is Up 21 Percent in NYC, TA Reveals" *STREETSBLOGNYC*, 25 May 2021, https://nyc.streetsblog.org/2021/05/25/data-shows-a-dangerous-rise-in-suv-purchases-in-nyc. Accessed 18 November 2023.

[3] Goldbaum, Christina. "Why Emptier Streets Meant an Especially Deadly Year for Traffic Deaths." *The New York Times*, 01 January 2021, https://www.nytimes.com/2021/01/01/nyregion/nyc-traffic-deaths.html. Accessed 18 November 2023.

[4] Liao, Felix H. and Lowry, Michael. "Speeding and Traffic-Related Injuries and Fatalities during the 2020 COVID-19 Pandemic: The Cases of Seattle and New York City." *medRxiv*, 01 November 2021. Accessed 18 November 2023.