Enhancing Road Safety through Data Analysis: A Case Study of Traffic Crashes Resulting in Injury

George Mason University

INFS-580-001 | Prof. Dr. Paul Smith

Ashrith Bhooka Ravinandan

George Mason University

Fairfax, Virginia

abhookar@gmu.edu

Abstract

The increasing number of road accidents that cause injuries has been a major public safety concern in recent years. This research explores the complex dynamics of traffic crashes that result in injuries, using an extensive set of data that includes several aspects of geographic, temporal, and meteorological data. The aim is to unravel the complex relationships between weather, temporal patterns, and particular locations by using cutting-edge statistical models and geospatial analysis techniques. This will enable us to provide a more comprehensive understanding of the factors that influence the frequency and severity of traffic crashes that result in injuries. The study starts with an exploration of the impact of meteorological variables on distinct collision types, utilizing generalized additive models. Building on existing literature (Becker et al., 2022), our study not only confirms the overarching influence of adverse weather conditions but also identifies specific associations between meteorological factors and particular crash scenarios. Rain, snow, sun glare, and high winds emerge as significant contributors, each exhibiting distinctive correlations with various collision types. This depth of specificity is particularly relevant to our central research question regarding the most common weather conditions associated with injury-related traffic incidents.

The temporal aspect of traffic incidents is scrutinized through an analysis of crash data by day of the week and time of day. Aligning with findings from previous studies (Injury Facts, 2023), the research identifies peak periods of risk, revealing that Saturdays and late afternoons witness higher frequencies of injury-related crashes. The implications of these temporal patterns extend beyond mere statistical observations, offering actionable insights for optimizing safety measures during specific days and hours. Furthermore, the research delves into the frequency analysis of Equivalent Property-Damage-Only (EPDO) crashes at intersections, acknowledging the importance of specific locations in shaping collision dynamics. Drawing from the work of Sharafeldin et al. (2023), My study extrapolates insights into location-specific factors influencing the occurrence of injury-related crashes. Pavement friction, roadway functional classification, road surface type, and other intersection-related characteristics emerge as pivotal factors affecting collision frequency. While the focus is on EPDO crashes, the implications extend to injury-related incidents, providing a nuanced understanding of how location-specific attributes contribute to crash occurrence.

The temporal component of traffic occurrences is investigated by analysing crash data by weekday and time of day. In line with prior study (Injury Facts, 2023), my findings identify peak hours of risk, demonstrating that Saturdays and late afternoons had higher rates of injury-related collisions. These temporal patterns have ramifications that go beyond simple statistical findings, providing actionable information for optimising safety measures during specific days and hours.

1. Introduction

In recent years, the escalating frequency of road accidents has raised alarms globally, prompting researchers and policymakers to delve into the intricacies of this pervasive issue. This research embarks on a comprehensive exploration, drawing insights from three pivotal perspectives on road safety. The first facet examines the influence of adverse weather conditions, unraveling how factors such as rain, snow, and fog contribute to increased accident rates. Simultaneously, the study investigates temporal patterns, discerning whether certain times of the day or days of the week are more prone to accidents. Finally, our analysis scrutinizes intersection-related factors, acknowledging the pivotal role these traffic junctures play in accident occurrence.

As we synthesize findings from three distinct reports, a cohesive narrative emerges, shedding light on the interconnectedness of these factors and their collective impact on road safety. By holistically approaching the investigation, we aim to provide nuanced insights that transcend individual circumstances, enabling a more holistic comprehension of road safety dynamics. This integrated understanding, we believe, is paramount for the development of targeted and effective strategies to curtail accidents and enhance overall transportation safety. Through this research, we strive to contribute not only to the academic discourse surrounding road safety but also to the practical realm of policy implementation, fostering a safer environment for road users worldwide.

This project answers the following questions:

- What are the most common weather conditions associated with traffic crashes resulting in injury?
- Are there specific days of the week or times of the day when these accidents are more likely to occur?
- Do certain intersections or road types have a higher frequency of injury-related accidents

2. Related Works

The exploration of road safety is a multifaceted endeavour, often necessitating a comprehensive understanding of various factors influencing traffic accidents. In pursuit of this knowledge, several research reports have delved into distinct aspects, providing valuable insights that contribute to the broader understanding of road safety dynamics. To gain a deeper understanding of the topic I explored three more research reports. The following summarises the research reports and explain how they relate to my research questions.

a. Weather impacts on various types of road crashes: A Quantitative Analysis using generalized additive models.

The purpose of this paper is to use generalised additive models to quantify the joint impact of traffic volume and climatic conditions on the probability of 78 distinct collision types. It considers a range of meteorological factors, such as precipitation, sun glare, and strong winds, and how these affect several kinds of traffic accidents.

This research paper investigates the relationship between meteorological conditions and road crashes in German administrative districts. While my research question focuses on the most common weather conditions associated with traffic crashes resulting in injury, the paper provides valuable insights into the broader impact of various weather conditions on different types of road crashes. Here's an analysis of how the paper relates to my research question:

Confirmation of Weather Effects: The research verifies that unfavourable weather conditions, like rain, snow, glare from the sun, and strong winds, have a considerable influence on a range of different kinds of traffic accidents. The relative risk of various crash types is increased by these meteorological conditions. The analysis offers insights into different kinds of crashes, but it also reaffirms that weather has a significant impact on road safety, including injury-causing incidents.

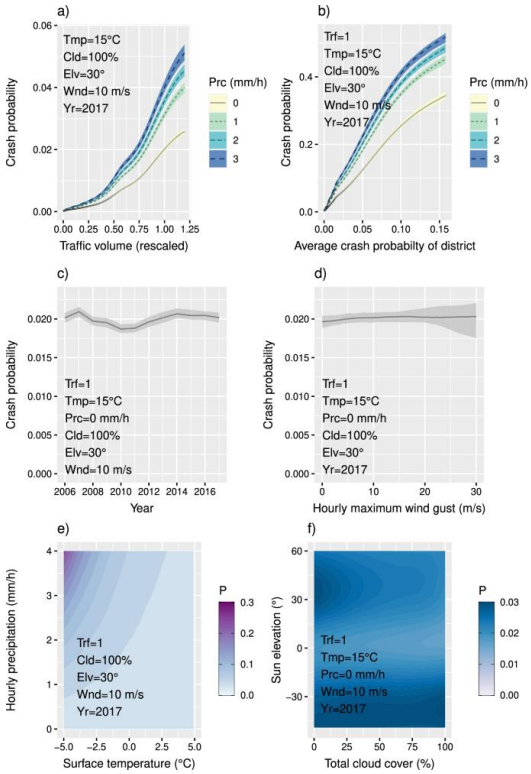


Figure: 1

A generalised linear model estimates the hourly probability of multi-car rear-end crashes, and there are functional links between the predictor variables and the model. Shaded regions represent 95% confidence intervals, which are computed from 100 models fitted with randomly selected training data.

Specific Weather-Crash Associations: To take things a step further, the paper pinpoints particular correlations between different kinds of crashes and certain meteorological circumstances. For instance, it emphasises that rain is more closely linked to single-car crashes than snow, which has the greatest impact on single-truck crashes. This level of specificity is useful for my investigation of accidents that result in injuries.

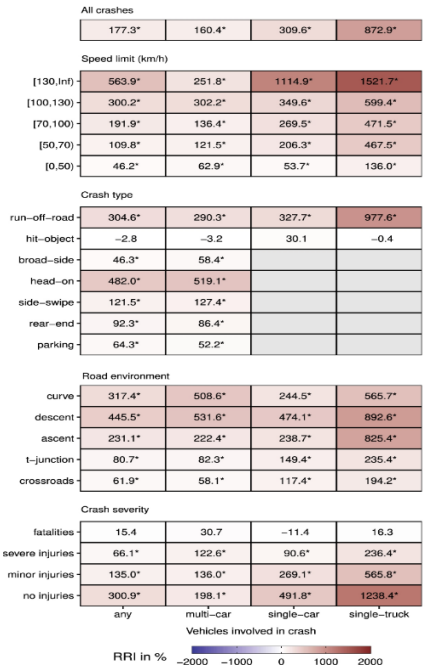


Figure: 2

Relative risk increase (RRI) of crash probabilities in situations with precipitation and negative temperature (Tmp = -3°C and Prc = 1 mm/h) compared to situations without precipitation and positive temperatures (Tmp = +3°C and Prc = 0 mm/h). Significant changes (i. e. more than 95 of 100 models fitted with randomly drawn training data show the same direction of change) are indicated with an asterisk

Sun Glare and Multi-Car Crashes: It is noteworthy that the report found that there is a higher likelihood of multi-car crashes when there is sun glare. It suggests that certain weather conditions may result in crash situations that are more complicated and may involve more than one car.

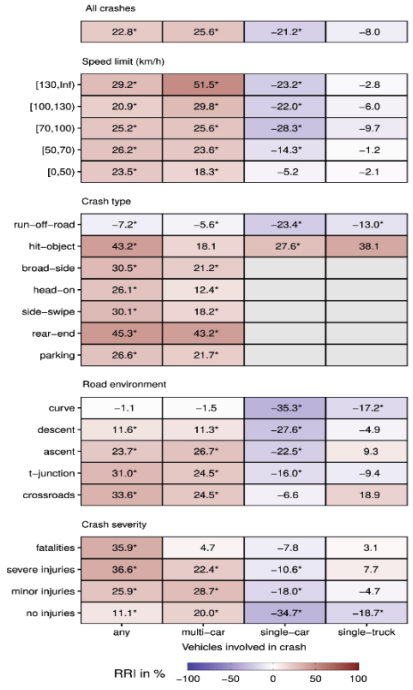


Figure: 3

Relative risk increase (RRI) of crash probabilities in situations with low sun elevation angle and cloud free conditions (Elv = 20 and Cld = 0%) compared to situations with low sun elevation angle and clouded conditions (Elv = 20 and Cls = 100%). Significant changes (i.e. more than 95 of 100 models fitted with randomly drawn training data show the same direction of change) are indicated with an asterisk

High Wind Speeds and Single-Truck Crashes: Another important finding in the paper is the correlation between high wind speeds and single-truck crashes. The correlation between strong winds and accidents involving particular vehicle types, such as trucks.

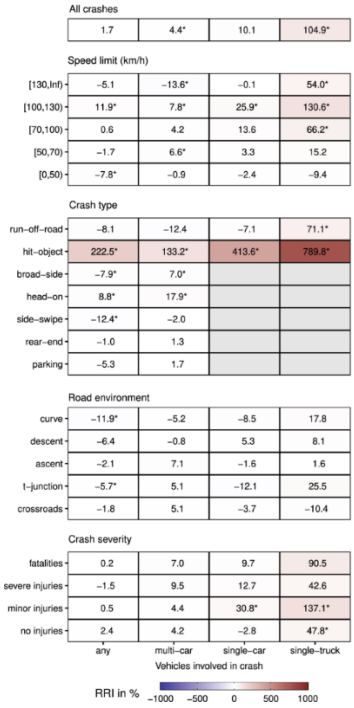


Figure: 4

Relative risk increase (RRI) of crash probabilities in situations with high wind speeds (Wind = 25 m/s) compared to situations with low wind speeds (Wind = 5 m/s). Significant changes (i.e. more than 95 of 100 models fitted with randomly drawn training data show the same direction of change) are indicated with an asterisk

b. Crashes by Time of Day and Day of Week

The report “Crashes by Time of Day and Day of Week” provides insightful information on the trends in car accidents concerning the days of the week and times of day. They also highlight the significance of taking into account mileage changes, certain days and months, and seasonal and monthly trends. The reports provide insights into the timing and frequency of car accidents, which are highly relevant to my research question regarding specific days of the week and times of the day when accidents are more likely to occur.

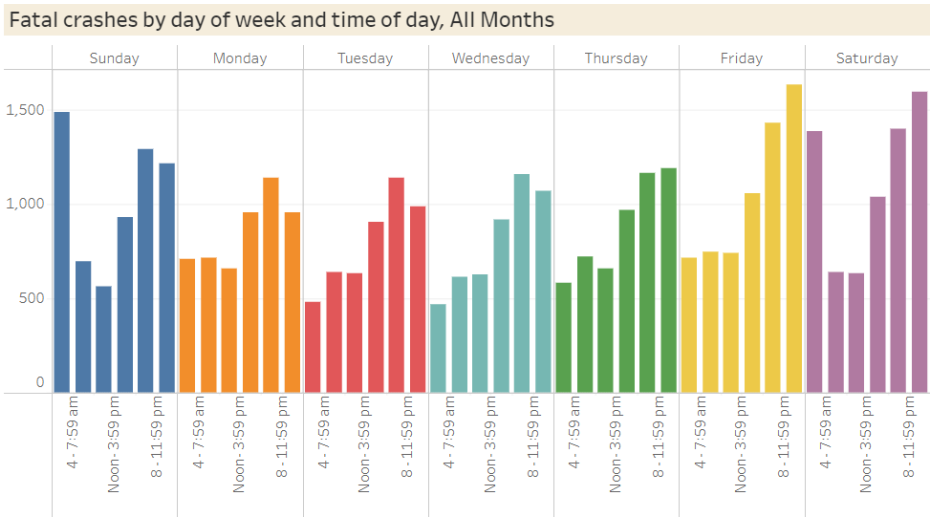


Figure: 5

Days of the Week: According to the research, Saturdays saw the highest number of fatal auto accidents, which were more common on the weekends. This data supports my research question by demonstrating that certain days of the week—especially weekends—are linked to a higher risk of accidents. The patterns in my research region are comparable, It is necessary to implement more safety precautions or interventions on weekends in order to reduce the number of accidents.

Times of the Day: According to the reports, 4 p.m. to 7:59 p.m. was the peak period for both fatal and nonfatal crashes. This supports my research question by demonstrating that accidents are more likely to occur during specific hours of the day.

Seasonal and Monthly Trends: The records additionally put insight into seasonal changes, showing that different seasons have distinct peak times for both fatal and nonfatal crashes. This data is relevant to my research subject, particularly if there are distinct seasons in the location. It suggests that in order to meet the changing risk of accidents, safety measures may need to be modified depending on the season.

Specific Days and Months: Saturday, August 7th, 2021 was found to be the deadliest day. This particular data point emphasises that some days and months may have abnormally high accident frequencies, which is directly related to my study. When putting safety precautions in place, it's critical to take these particular days or months into consideration if such patterns are seen in the research area.

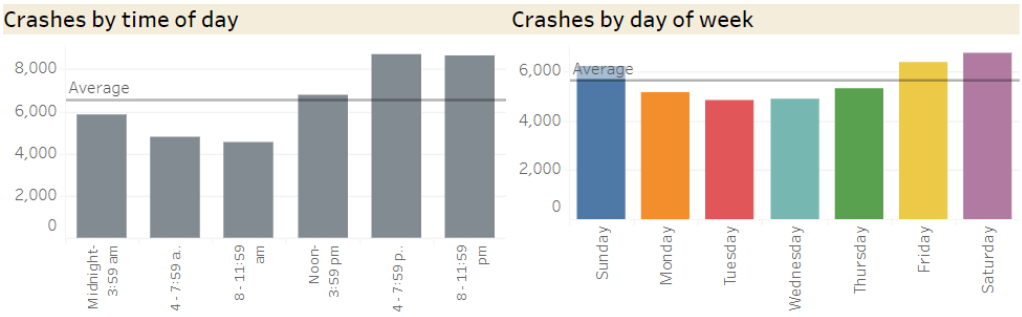


Figure: 6

c. Frequency Analysis of Equivalent Property-Damage-Only (EPDO) Crashes at Intersections.

The research paper titled "Frequency Analysis of Equivalent Property-Damage-Only (EPDO) Crashes at Intersections" by Mostafa Sharafeldin, Khaled Ksaibati, and Ken Gerow, conducted by the Wyoming Technology Transfer Center and the University of Wyoming, focuses on the analysis of factors contributing to the frequency of equivalent property-damage-only (EPDO) crashes at intersections. The relevance of intersection-related collisions, their influence on public health and the economy, and the application of EPDO crash analysis to quantify crash frequency and severity are all addressed in the research paper. It addresses the crucial issue of traffic accidents and their accompanying expenses, emphasising the importance of intersection-related accidents in terms of frequency and severity. The study employs the notion of Equivalent Property-Damage-Only (EPDO) accidents to measure the effect of these events, which assesses the severity of crashes based on their financial implications.

The research paper offers insightful information on the variables affecting the number of collisions at crossings. The findings may be connected to my research topic regarding the increased incidence of accidents at specific intersections or road types, even though it focuses exclusively on EPDO crashes and takes into account factors unique to the state of Wyoming. The focus of the paper on location, roadway type, pavement friction, and other characteristics provides important background information for comprehending the connection between these elements and accident frequency.

Pavement Friction: One significant element influencing the frequency of EPDO crashes at junctions is identified in the article as pavement friction. It has been discovered that increasing pavement friction greatly lowers the frequency of EPDO crashes. The consequence is that junction safety depends on maintaining a sufficient level of pavement friction.

Location and Grade: It is reported that there are less EPDO crashes at urban crossroads than at rural ones. This is probably because rural locations have more complicated traffic patterns and faster travel rates. In addition, there are generally fewer EPDO collisions at crossroads with rising or falling grades, maybe as a result of vehicles being extra cautious on undulating terrain.

Roadway Functional Classification: Compared to local roads, intersections on higher-class roadways, such as interstates and key arterial routes, see more EPDO crashes. In contrast, there are fewer EPDO crashes at crossings on collector routes. These variations can be attributed to factors including traffic volume, vehicle behaviour on various types of roads, and speed.

Road Surface Type: Compared to intersections with asphalt surfaces, those with concrete road surfaces see a higher probability of EPDO collisions. This discrepancy is probably caused by the features of Wyoming's concrete road segments, which include faster speeds and heavier traffic.

Guardrails and Right Shoulder: Research has shown that putting up guardrails at intersections reduces the number of EPDO collisions. Conversely, crossings with paved right shoulders are associated with a higher incidence of EPDO crashes, possibly due to improper overtaking or turning manoeuvres on the shoulder.

Horizontal Curvature: The study indicates that different horizontal curvatures have distinct effects on EPDO accident frequencies depending on the radius. Whereas lower crash frequencies are linked to strong horizontal curves, higher crash frequencies are linked to minor horizontal curves. This is addressed in connection to driving cautiously on road segments that curve.

Crash Severity	Number of Crashes	Crash Costs (USD)	EPDO Counts
Property-damage-only (PDO)	6807	$306,315,000	6807
Suspected minor injury/possible injury/unknown	2140	$385,200,000	8560
Suspected serious injury	144	$84,240,000	1872
Fatal injury	17	$211,905,000	4709

Figure:7- Summary of crash counts, costs, and EPDO frequency.

Continuous Variables				
	Mean	Standard Deviation	Minimum	Maximum
EPDO crash count (response)	13.77	30.271	1	314
Pavement friction	41.48	9.186	18.65	71
Number of lanes	3.481	0.877	2	4
Median width	13.87	21.078	0	120
Right shoulder width	5.296	2.908	0	10
Binary Variable			**Count**	**Percentage**
Type: Four legs or more (1 if yes or 0 otherwise)			1270	79.7
Location: Urban (1 if yes or 0 otherwise)			1319	82.7
Traffic control: Signalized (1 if yes or 0 otherwise)			1298	81.4
Grade: Uphill or downhill (1 if yes or 0 otherwise)			170	10.7
Functional classification: Interstate (1 if yes or 0 otherwise)			22	1.4
Functional classification: Principal arterial (1 if yes or 0 otherwise)			1204	75.5
Functional classification: Minor arterial (1 if yes or 0 otherwise)			180	11.3
Functional classification: Collector (1 if yes or 0 otherwise)			31	1.9
Road surface: Concrete (1 if yes or 0 otherwise)			654	41.0
Guardrail (1 if yes or 0 otherwise)			30	1.9
Median: Depressed (1 if yes or 0 otherwise)			169	10.6
Median: Raised (1 if yes or 0 otherwise)			973	61.0
Paved right shoulder (1 if yes or 0 otherwise)			1450	91.0
Slight horizontal curve (>1500 ft radius) (1 if yes or 0 otherwise)			245	15.4
Heavy horizontal curve (<1500 ft radius) (1 if yes or 0 otherwise)			195	12.2

Figure: 8- Descriptive data statistics.

3. Dataset

The dataset is a comprehensive collection of information related to traffic collisions, with each row representing a specific incident. It encompasses a total of 60 columns and 56,010 rows, with each column offering distinct details about the collisions. Key geographical data is included, such as latitude, longitude, and geocode information, along with identifiers for intersections and segments.

The dataset also captures crucial temporal aspects, including the date and time of each collision, categorized by year, month, day of the week, and time intervals. Location-related information, such as jurisdiction, primary and secondary roads, and the presence of an intersection, further enriches the dataset.

Weather and road conditions at the time of the collisions are documented in columns like weather_1, weather_2, road_surface, and road_cond_1. These details could be pivotal in understanding the influence of environmental factors on collision occurrences.

The dataset delves into collision specifics, covering severity (collision_severity), collision type (type_of_collision), and involved parties (mviw). Information about pedestrian actions during collisions (ped_action) and road surface conditions (road_surface) adds granularity to the analysis.

Details about the parties involved in the collisions, their types, directions of travel, and movements pre-accident are included. Geographic and mapping information, as well as URLs for additional references, are also present.

Timestamps such as data_as_of, data_updated_at, and data_loaded_at provide insights into data maintenance and updates. Moreover, the dataset includes outcome metrics, namely number_killed and number_injured, offering a quantitative measure of the impact of each collision.

	unique_id	cnn_intrsctn_fkey	cnn_sgmt_fkey	case_id_pkey	tb_latitude	tb_longitude	geocode_source	geocode_location	collision_datetime	collision_date	...	party2_move_pre_acc	point	data_as_of	data_updated_at	c
0	15414	23926000	NaN	151003670	37.777857	-122.406436	SFPD-CROSSROADS	CITY STREET	11/18/2015 09:04:00 PM	2015 November 18	...	Proceeding Straight	POINT (-122.40643 37.777855)	11/18/2015 12:00:00 AM	04/26/2023 12:00:00 AM	
1	52640	23969000	NaN	200212108	37.754453	-122.408343	SFPD-INTERIM DB	CITY STREET	03/28/2020 02:35:00 PM	2020 March 28	...	Proceeding Straight	POINT (-122.40834 37.75445)	04/02/2020 12:00:00 AM	04/26/2023 12:00:00 AM	
2	36577	22495000	5408000.0	140836694	37.714040	-122.461135	SFPD-CROSSROADS	CITY STREET	10/04/2014 01:25:00 PM	2014 October 04	...	Proceeding Straight	POINT (-122.461136 37.71404)	10/04/2014 12:00:00 AM	04/26/2023 12:00:00 AM	
3	27582	24726000	9299000.0	160732947	37.791627	-122.402516	SFPD-CROSSROADS	CITY STREET	09/08/2016 02:00:00 PM	2016 September 08	...	Not Stated	POINT (-122.40252 37.791626)	09/08/2016 12:00:00 AM	04/26/2023 12:00:00 AM	
4	14911	23811000	179000.0	180566685	37.779856	-122.394008	SFPD-INTERIM DB	CITY STREET	07/30/2018 11:25:00 AM	2018 July 30	...	NaN	POINT (-122.394005 37.779858)	10/22/2018 12:00:00 AM	04/26/2023 12:00:00 AM	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
56004	60030	20618000	NaN	220716801	37.729211	-122.400834	SFPD-INTERIM DB	CITY STREET	10/18/2022 05:09:00 PM	2022 October 18	...	Proceeding Straight	POINT (-122.40083 37.72921)	02/01/2023 12:00:00 AM	04/27/2023 12:00:00 AM	
56005	59928	23150000	923101.0	220678574	37.737605	-122.475224	SFPD-INTERIM DB	CITY STREET	10/03/2022 04:45:00 PM	2022 October 03	...	Stopped	POINT (-122.47523 37.737606)	02/01/2023 12:00:00 AM	04/27/2023 12:00:00 AM	
56006	21803	25958000	NaN	3497597	37.778329	-122.426641	SFPD-CROSSROADS	CITY STREET	11/18/2007 01:25:00 PM	2007 November 18	...	Proceeding Straight	POINT (-122.42664 37.778328)	11/18/2007 12:00:00 AM	04/26/2023 12:00:00 AM	
56007	59557	21890000	NaN	220814518	37.742295	-122.423286	SFPD-INTERIM DB	CITY STREET	11/27/2022 12:01:00 AM	2022 November 27	...	Proceeding Straight	POINT (-122.42329 37.742294)	02/01/2023 12:00:00 AM	04/27/2023 12:00:00 AM	
56008	47771	27266000	NaN	190581708	37.773210	-122.470163	SFPD-INTERIM DB	CITY STREET	08/08/2019 09:30:00 PM	2019 August 08	...	Proceeding Straight	POINT (-122.47016 37.77321)	08/09/2019 12:00:00 AM	04/26/2023 12:00:00 AM	

56009 rows × 60 columns

Figure: 9

4. Analysis

A. Tools and Methods

Jupyter Notebook:

Jupyter Notebook is an open-source interactive web application that allows users to create and share documents containing live code, equations, visualizations, and narrative text. It supports various programming languages, including Python, R, and Julia, making it versatile for data analysis, scientific research, and education. The interface combines code cells for writing and executing code with markdown cells for text explanations, creating a seamless environment for exploratory data analysis and collaborative work. Jupyter Notebooks have gained widespread popularity in the data science and research communities for their interactive and dynamic nature, fostering a user-friendly and efficient computational experience.

R Studio:

R Studio is a powerful integrated development environment (IDE) for the R programming language. It provides a user-friendly interface for statistical computing and data analysis, catering to both beginners and experienced R users. With features like syntax highlighting, code completion, and a built-in console, R Studio enhances the coding experience. The IDE supports the creation of R scripts, visualizations, and interactive applications, facilitating seamless exploration and communication of data-driven insights. Widely used in academia, industry, and research, R Studio plays a pivotal role in the R ecosystem, fostering efficient development and collaboration in statistical computing and data science projects.

MySQL workbench:

MySQL Workbench is a visual database design and administration tool that streamlines MySQL database development. Offering a comprehensive set of features, it enables users to design, model, and generate databases visually. The platform supports SQL development with syntax highlighting and code completion, fostering efficient query creation. With powerful tools for database administration, including performance monitoring and user management, MySQL Workbench ensures robust database management. Its intuitive interface allows for seamless navigation and collaboration, making it a popular choice for developers and database administrators seeking a unified environment for designing, developing, and maintaining MySQL databases.

Python:

Python, a versatile high-level programming language, emphasizes readability and simplicity, making it ideal for beginners and professionals alike. Known for its clear syntax, extensive standard library, and community support, Python is widely used in web development, data science, artificial intelligence, and automation. Its interpreted nature promotes rapid development, while dynamic typing enhances flexibility. Python's object-oriented and functional programming capabilities contribute to its adaptability across diverse domains. The language's popularity continues to grow due to its ease of learning, expansive ecosystem, and broad applicability, solidifying its role as a go-to language for a range of programming tasks and industries.

NumPy:

NumPy, short for Numerical Python, is a fundamental library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently. NumPy's arrays facilitate advanced mathematical and logical operations, making it essential for scientific and data-intensive applications. The library also integrates seamlessly with other data science tools, enhancing their computational capabilities. With its performance optimizations and versatility, NumPy serves as a cornerstone for tasks like data manipulation, statistical analysis, and machine learning, contributing significantly to Python's standing as a powerful language for scientific computing.

Pandas:

Pandas is a powerful data manipulation and analysis library for Python, offering data structures like DataFrames and Series that simplify working with structured data. Widely used in data science and analytics, Pandas provides tools for cleaning, transforming, and exploring datasets with ease. It seamlessly integrates with other libraries like NumPy, enhancing computational capabilities. Pandas excels in handling missing data, time-series data, and diverse data types, making it indispensable for tasks such as data wrangling and preprocessing. Its intuitive syntax and comprehensive functionality have made Pandas a cornerstone in the Python ecosystem, empowering users to efficiently manage and analyze complex datasets.

Matplotlib:

Matplotlib is a versatile and widely-used 2D plotting library for Python, facilitating the creation of static, animated, and interactive visualizations. With a diverse range of plotting functions, it enables users to generate a variety of charts, graphs, and plots for data representation. Matplotlib's flexibility allows customization of every aspect of a plot, ensuring precise and visually appealing results. Often used in conjunction with NumPy, it is a crucial tool in scientific computing, data analysis, and exploratory data visualization. Its simplicity and extensive documentation make Matplotlib accessible for both beginners and advanced users, solidifying its role in the Python ecosystem.

Seaborn:

Seaborn is a statistical data visualization library in Python built on top of Matplotlib. Known for its aesthetically pleasing and informative statistical graphics, Seaborn simplifies the creation of complex visualizations. It provides a high-level interface for drawing attractive and informative statistical graphics, including heatmaps, violin plots, and scatter plots. Seaborn's integration with Pandas data structures and its concise syntax make it user-friendly for data analysts and scientists. The library excels in showcasing relationships in data and is often used for exploratory data analysis and presentation. With its rich visualization capabilities, Seaborn complements Matplotlib, enhancing the overall plotting experience in Python.

NLTK:

The Natural Language Toolkit (NLTK) is a powerful Python library for natural language processing (NLP) and text analysis. It provides tools and resources for tasks such as tokenization, stemming, part-of-speech tagging, and sentiment analysis. NLTK's extensive collection of corpora and lexical resources supports research and development in linguistics

and computational linguistics. With a user-friendly interface, NLTK facilitates the exploration and processing of textual data, making it a valuable resource for developers, researchers, and educators in the field of NLP. Its modular design and wide-ranging functionality make NLTK a go-to toolkit for various applications in understanding and working with human language.

MultinomialNB:

Multinomial Naive Bayes (MultinomialNB) is a probabilistic classification algorithm commonly used in natural language processing and text classification. It is an extension of the Naive Bayes algorithm designed for discrete data, making it particularly suited for features that represent word counts or term frequencies in text data. MultinomialNB assumes that the features are conditionally independent given the class, simplifying computations. Widely employed in spam filtering, document categorization, and sentiment analysis, MultinomialNB leverages Bayesian probability to make predictions based on the likelihood of observing feature values in each class. Its simplicity, efficiency, and effectiveness contribute to its popularity in text-based machine learning tasks.

B. Data Cleaning

I used Python for data cleaning, employing several steps to enhance the dataset. First, I converted column names to title case for consistency. Then, I applied a function to format text data to title case. Subsequently, I checked for missing values, identified duplicate records, and dropped rows with null values. Finally, I reset the index for better data representation. The following steps were performed in Python for data cleaning:

i) Converting the Column names to Title case

```
In [44]: data.columns = data.columns.str.title()
         data
```

Out[44]:

	Unique_Id	Cnn_Intrsctn_Fkey	Cnn_Sgmt_Fkey	Case_Id_Pkey	Tb_Latitude	Tb_Longitude	Geocode_Source	Geocode_Location	Collision_Datetime	Colli
0	15414	23926000	NaN	151003670	37.777857	-122.406436	SFPD-CROSSROADS	CITY STREET	11/18/2015 09:04:00 AM	No
1	52640	23969000	NaN	200212108	37.754453	-122.408343	SFPD-INTERIM DB	CITY STREET	03/28/2020 02:35:00 PM	2020
2	36577	22495000	5408000.0	140836694	37.714040	-122.461135	SFPD-CROSSROADS	CITY STREET	10/04/2014 01:25:00 PM	201
3	27582	24726000	9299000.0	160732947	37.791627	-122.402516	SFPD-CROSSROADS	CITY STREET	09/08/2016 02:00:00 PM	Sep
4	14911	23811000	179000.0	180566685	37.779856	-122.394008	SFPD-INTERIM DB	CITY STREET	07/30/2018 11:25:00 AM	20
...	...	...	...	...	...	...	...	...	...	
56004	60030	20618000	NaN	220716801	37.729211	-122.400834	SFPD-INTERIM DB	CITY STREET	10/18/2022 05:09:00 PM	202
56005	59928	23150000	923101.0	220678574	37.737605	-122.475224	SFPD-INTERIM DB	CITY STREET	10/03/2022 04:45:00 PM	202
56006	21803	25958000	NaN	3497597	37.778329	-122.426641	SFPD-CROSSROADS	CITY STREET	11/18/2007 01:25:00 PM	No
56007	59557	21890000	NaN	220814518	37.742295	-122.423286	SFPD-INTERIM DB	CITY STREET	11/27/2022 12:01:00 AM	No
56008	47771	27266000	NaN	190581708	37.773210	-122.470163	SFPD-INTERIM DB	CITY STREET	08/08/2019 09:30:00 PM	20

56009 rows × 60 columns

Figure: 10

The code ‘data.columns = data.columns.str.title()’ capitalizes the first letter of each word in the column names, ensuring a consistent title case format. This improves readability and standardizes the naming convention across the dataset, enhancing the overall organization of the data.

ii) Converting all the rows to title case

```
In [45]: def title_case_str(val):
             return val.title() if isinstance(val, str) else val
         data = data.applymap(title_case_str)
         data
```

Out[45]:

	Unique_Id	Cnn_Intrsctn_Fkey	Cnn_Sgmt_Fkey	Case_Id_Pkey	Tb_Latitude	Tb_Longitude	Geocode_Source	Geocode_Location	Collision_Datetime	Colli
0	15414	23926000	NaN	151003670	37.777857	-122.406436	Sfpd-Crossroads	City Street	11/18/2015 09:04:00 Am	No
1	52640	23969000	NaN	200212108	37.754453	-122.408343	Sfpd-Interim Db	City Street	03/28/2020 02:35:00 Pm	2020
2	36577	22495000	5408000.0	140836694	37.714040	-122.461135	Sfpd-Crossroads	City Street	10/04/2014 01:25:00 Pm	201
3	27582	24726000	9299000.0	160732947	37.791627	-122.402516	Sfpd-Crossroads	City Street	09/08/2016 02:00:00 Pm	Sep
4	14911	23811000	179000.0	180566685	37.779856	-122.394008	Sfpd-Interim Db	City Street	07/30/2018 11:25:00 Am	20
...	...	...	...	...	...	...	...	...	...	
56004	60030	20618000	NaN	220716801	37.729211	-122.400834	Sfpd-Interim Db	City Street	10/18/2022 05:09:00 Pm	202
56005	59928	23150000	923101.0	220678574	37.737605	-122.475224	Sfpd-Interim Db	City Street	10/03/2022 04:45:00 Pm	202
56006	21803	25958000	NaN	3497597	37.778329	-122.426641	Sfpd-Crossroads	City Street	11/18/2007 01:25:00 Pm	No
56007	59557	21890000	NaN	220814518	37.742295	-122.423286	Sfpd-Interim Db	City Street	11/27/2022 12:01:00 Am	No
56008	47771	27266000	NaN	190581708	37.773210	-122.470163	Sfpd-Interim Db	City Street	08/08/2019 09:30:00 Pm	20

56009 rows × 60 columns

Figure: 11

The function 'title_case_str' is applied element-wise to the entire dataset using 'data.applymap(title_case_str)'. This function title-cases each string value in the dataset, maintaining the original format for non-string values. It ensures uniformity and readability in string data throughout the dataset.

iii) Checking for missing values

```
In [46]: #Checking for missing values
         completeness_check = data.isnull().sum()
         print("Completeness Check:")
         print(completeness_check)
```

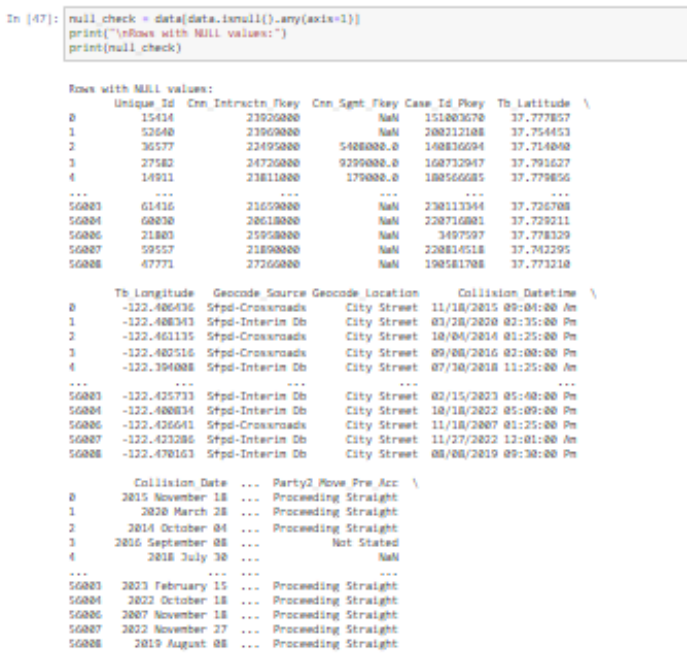
```
Completeness Check:
Unique_Id                        0
Cnn_Intrsctn_Fkey                0
Cnn_Sgmt_Fkey                30810
Case_Id_Pkey                     0
Tb_Latitude                    147
Tb_Longitude                   147
Geocode_Source                   0
Geocode_Location                 0
Collision_Datetime               0
Collision_Date                   0
Collision_Time                  56
Accident_Year                    0
Month                            0
Day_Of_Week                      8
Time_Cat                        51
Juris                            0
Officer_Id                    1064
Reporting_District            6510
Beat_Number                   4369
Primary_Rd                       0
Secondary_Rd                    92
Distance                        79
Direction                        1
Weather_1                        0
Weather_2                      796
Collision_Severity               0
Type_Of_Collision                0
Mviw                             0
Ped_Action                       0
Road_Surface                     0
Road_Cond_1                      0
Road_Cond_2                      0
Lighting                         0
Control_Device                   0
Intersection                     1
Vz_Pcf_Code                      1
Vz_Pcf_Group                   808
Vz_Pcf_Description               0
Vz_Pcf_Link                    678
Number_Killed                    3
Number_Injured                   0
Street_View                    126
Dph_Col_Grp                      0
Dph_Col_Grp_Description          0
Party_At_Fault                5546
Party1_Type                     11
Party1_Dir_Of_Travel            11
Party1_Move_Pre_Acc             10
Party2_Type                   3982
Party2_Dir_Of_Travel          3980
Party2_Move_Pre_Acc           3979
Point                          147
Data_As_Of                       0
Data_Updated_At                  0
Data_Loaded_At                   0
Neighborhoods                  773
Sf Find Neighborhoods          773
Current Police Districts       231
Current Supervisor Districts   192
Analysis Neighborhoods         205
dtype: int64
```

Figure:12

The completeness check reveals the count of missing values in each column. Notable columns with missing values include "Cnn_Sgmt_Fkey," "Tb_Latitude," "Tb_Longitude," "Collision_Time," and

others. Addressing these gaps may involve imputation or further investigation, ensuring data integrity for subsequent analysis.

iv) Checking for Null values



```
In [47]: null_check = data[data.isnull().any(axis=1)]
         print("\nRows with NULL values:")
         print(null_check)

Rows with NULL values:
       Unique_Id  Cnn_Intrsctn_Fkey  Cnn_Sgmt_Fkey  Case_Id_Pkey  Tb_Latitude  \
0          15414           23926000            NaN     151803670    37.777857
1          52640           23969000            NaN     208212108    37.754453
2          36577           23495000      5486000.0     148836094    37.714040
3          27582           24726000      9299000.0     168732947    37.791627
4          14911           23811000       179000.0     180566685    37.779856
...          ...                ...            ...           ...          ...
56003      61416           21659000            NaN     230113544    37.726708
56004      60830           20618000            NaN     228716801    37.729211
56006      21801           25958000            NaN       3497597    37.778329
56007      59557           21890000            NaN     228814518    37.742295
56008      47771           27266000            NaN     190581708    37.773210

       Tb_Longitude   Geocode_Source Geocode_Location      Collision_Datetime  \
0       -122.406436  Sfpd-Crossroads      City Street  11/18/2015 09:04:00 Am
1       -122.408343  Sfpd-Interim Db      City Street  03/28/2020 02:35:00 Pm
2       -122.461135  Sfpd-Crossroads      City Street  10/04/2014 01:25:00 Pm
3       -122.402516  Sfpd-Crossroads      City Street  09/08/2016 02:00:00 Pm
4       -122.394088  Sfpd-Interim Db      City Street  07/30/2018 11:25:00 Am
...             ...              ...              ...                     ...
56003   -122.425733  Sfpd-Interim Db      City Street  02/15/2023 05:40:00 Pm
56004   -122.408834  Sfpd-Interim Db      City Street  10/18/2022 05:09:00 Pm
56006   -122.426641  Sfpd-Crossroads      City Street  11/18/2007 01:25:00 Pm
56007   -122.423286  Sfpd-Interim Db      City Street  11/27/2022 12:01:00 Am
56008   -122.470163  Sfpd-Interim Db      City Street  08/08/2019 09:30:00 Pm

         Collision_Date  ...  Party2_Move_Pre_Acc  \
0      2015 November 18  ...  Proceeding Straight
1         2020 March 28  ...  Proceeding Straight
2       2014 October 04  ...  Proceeding Straight
3     2016 September 08  ...          Not Stated
4          2018 July 30  ...                 NaN
...                 ...  ...                 ...
56003  2023 February 15  ...  Proceeding Straight
56004   2022 October 18  ...  Proceeding Straight
56006  2007 November 18  ...  Proceeding Straight
56007  2022 November 27  ...  Proceeding Straight
56008     2019 August 08  ...  Proceeding Straight
```

Figure: 13

The displayed DataFrame shows rows with at least one NULL value in any column. These rows may need attention during analysis or require imputation to maintain data integrity. The specific columns with missing values can guide further steps in addressing these gaps for a comprehensive dataset.

v) Checking for duplicate values

```
In [48]: duplicate_check = data[data.duplicated(keep='first')]
         print("\nDuplicate Records:")
         print(duplicate_check)

Duplicate Records:
Empty DataFrame
Columns: [Unique_Id, Cnn_Intrsctn_Fkey, Cnn_Sgmt_Fkey, Case_Id_Pkey, Tb_Latitude, Tb_Longitude, Geocode_Source, Geocode_Locatio
n, Collision_Datetime, Collision_Date, Collision_Time, Accident_Year, Month, Day_Of_Week, Time_Cat, Juris, Officer_Id, Reportin
g_District, Beat_Number, Primary_Rd, Secondary_Rd, Distance, Direction, Weather_1, Weather_2, Collision_Severity, Type_Of_Colli
sion, Mviw, Ped_Action, Road_Surface, Road_Cond_1, Road_Cond_2, Lighting, Control_Device, Intersection, Vz_Pcf_Code, Vz_Pcf_Gro
up, Vz_Pcf_Description, Vz_Pcf_Link, Number_Killed, Number_Injured, Street_View, Dph_Col_Grp, Dph_Col_Grp_Description, Party_At
_Fault, Party1_Type, Party1_Dir_Of_Travel, Party1_Move_Pre_Acc, Party2_Type, Party2_Dir_Of_Travel, Party2_Move_Pre_Acc, Point,
Data_As_Of, Data_Updated_At, Data_Loaded_At, Neighborhoods, Sf Find Neighborhoods, Current Police Districts, Current Supervisor
Districts, Analysis Neighborhoods]
Index: []

[0 rows x 60 columns]
```

Figure: 14

The code checks for duplicate records in the DataFrame. The displayed result, an empty DataFrame, indicates that there are no duplicate records based on all columns. This step ensures data integrity by identifying and handling any repeated entries in the dataset.

vi) Dropping Null values

```
In [49]: data = data.dropna()
         data
```

Out[49]:

	Unique_Id	Cnn_Intrsctn_Fkey	Cnn_Sgmt_Fkey	Case_Id_Pkey	Tb_Latitude	Tb_Longitude	Geocode_Source	Geocode_Location	Collision_Datetime	Colli
15	439	20410000	13331000.0	130384415	37.713257	-122.414407	Sfpd-Crossroads	City Street	05/10/2013 02:53:00 Pm	20
24	54003	20364000	2445000.0	200639029	37.714311	-122.408277	Sfpd-Interim Db	City Street	10/22/2020 08:15:00 Pm	20
39	57019	30079000	13154102.0	210833590	37.787227	-122.421798	Sfpd-Interim Db	City Street	12/17/2021 09:49:00 Am	De
44	30729	26629000	3453000.0	6096662	37.786863	-122.434697	Sfpd-Crossroads	City Street	09/27/2012 10:22:00 Am	Sep
59	5080	21964000	3028101.0	3997255	37.733801	-122.435670	Sfpd-Crossroads	City Street	11/21/2008 08:40:00 Am	No
...	...	...	...	...	...	...	...	...	...	...
55996	25812	24655000	6108000.0	6058802	37.787746	-122.405197	Sfpd-Crossroads	City Street	03/01/2012 09:36:00 Am	201
55997	43785	27486000	6071101.0	3538811	37.780568	-122.473965	Sfpd-Crossroads	City Street	12/05/2007 04:32:00 Pm	De
56001	48020	21647000	5087000.0	190541536	37.724707	-122.429379	Sfpd-Interim Db	City Street	07/25/2019 10:34:00 Pm	20
56002	35717	26705000	2785001.0	6049383	37.804143	-122.425134	Sfpd-Crossroads	City Street	10/15/2012 06:00:00 Pm	20
56005	59928	23150000	923101.0	220678574	37.737605	-122.475224	Sfpd-Interim Db	City Street	10/03/2022 04:45:00 Pm	20

16396 rows × 60 columns

Figure: 15

The code removes rows with missing values ('NaN') from the DataFrame, ensuring a cleaner dataset. The resulting DataFrame, 'data', now contains 16,396 rows and 60 columns. This step helps in handling missing or incomplete data, enhancing the dataset's quality for subsequent analysis or modeling.

vii) Resetting the Index

```
In [50]: data = data.reset_index(drop=True)
         data
```

Out[50]:

	Unique_Id	Cnn_Intrsctn_Fkey	Cnn_Sgmt_Fkey	Case_Id_Pkey	Tb_Latitude	Tb_Longitude	Geocode_Source	Geocode_Location	Collision_Datetime	Colli
0	439	20410000	13331000.0	130384415	37.713257	-122.414407	Sfpd-Crossroads	City Street	05/10/2013 02:53:00 Pm	20
1	54003	20364000	2445000.0	200639029	37.714311	-122.408277	Sfpd-Interim Db	City Street	10/22/2020 08:15:00 Pm	202
2	57019	30079000	13154102.0	210833590	37.787227	-122.421798	Sfpd-Interim Db	City Street	12/17/2021 09:49:00 Am	De
3	30729	26629000	3453000.0	6096662	37.786863	-122.434697	Sfpd-Crossroads	City Street	09/27/2012 10:22:00 Am	Sep
4	5080	21964000	3028101.0	3997255	37.733801	-122.435670	Sfpd-Crossroads	City Street	11/21/2008 08:40:00 Am	No
...	...	...	...	...	...	...	...	...	...	...
16391	25812	24655000	6108000.0	6058802	37.787746	-122.405197	Sfpd-Crossroads	City Street	03/01/2012 09:36:00 Am	2012
16392	43785	27486000	6071101.0	3538811	37.780568	-122.473965	Sfpd-Crossroads	City Street	12/05/2007 04:32:00 Pm	De
16393	48020	21647000	5087000.0	190541536	37.724707	-122.429379	Sfpd-Interim Db	City Street	07/25/2019 10:34:00 Pm	20
16394	35717	26705000	2785001.0	6049383	37.804143	-122.425134	Sfpd-Crossroads	City Street	10/15/2012 06:00:00 Pm	20
16395	59928	23150000	923101.0	220678574	37.737605	-122.475224	Sfpd-Interim Db	City Street	10/03/2022 04:45:00 Pm	202

16396 rows × 60 columns

Figure: 16

Here I reset the index of the DataFrame after dropping rows with missing values. The resulting DataFrame, 'data', now has a new index ranging from 0 to 16,395. This step is helpful to ensure a continuous index and facilitates better indexing and referencing of rows in subsequent analyses.

C. Statistical Analysis

I used SQL queries to perform statistical analysis, exploring key metrics such as total incidents, severity distribution, and road-specific details. Statistical summaries cover fatalities, injuries, and collision distances. The analysis aims to enhance understanding of road safety patterns and contribute valuable insights for effective accident prevention strategies."

i) Average and Maximum Number of Injured

This calculates the average and maximum number of injuries recorded in the 'TrafficCrashes' dataset. The 'AvgInjured' result represents the mean number of injuries across all traffic incidents, while 'MaxInjured' indicates the highest number of injuries observed in a single incident. These statistics provide insights into the typical and extreme injury scenarios within the analyzed traffic crash data.

	AvgInjured	MaxInjured
▶	1.2157	19

Figure: 17

ii) Collision Severity Distribution

This examines the distribution of collision severity within the 'TrafficCrashes' dataset. By grouping incidents based on their collision severity, it calculates the count of occurrences for each severity level. The result provides an overview of the distribution of collision severity categories, offering valuable insights into the prevalence of different levels of severity in the analyzed traffic crash data.

	Collision_Severity	CollisionCount
▶	Injury (Other Visible)	4676
	Injury (Complaint Of Pain)	10156
	Fatal	172
	Injury (Severe)	1107

Figure: 18

iii) Average Distance of Collisions

This calculates the average distance of collisions from the 'TrafficCrashes' dataset. By computing the mean distance value, the analysis provides valuable information on the typical separation between vehicles or objects involved in traffic incidents. This metric contributes to understanding the spatial aspects of collisions, aiding in the assessment of potential patterns or trends related to collision distances.

	AvgDistance
▶	116.50108621438768

Figure: 19

iv) Top 5 Primary Roads with Highest Collision Count

This is used to identify and rank the primary roads with the highest collision counts. By aggregating and ordering the collision data based on primary roads, the analysis highlights the areas where traffic incidents are most concentrated. This information is instrumental in

pinpointing locations that may require additional safety measures or interventions to reduce the frequency of collisions.

Primary_Rd	CollisionCount
Mission St	831
Market St	505
19Th Ave	333
Geary Blvd	319
Van Ness Ave	284

Figure: 20

v) Year with Maximum Number of Accidents

This examines the dataset, focusing on the distribution of collision frequencies across different years. By grouping the data based on accident years and ordering them in descending order, the query highlights the year with the maximum number of traffic crashes. This analysis provides insights into the temporal patterns of collisions, aiding in understanding variations and potential factors influencing road safety across different years.

Accident_Year	Frequency
2013	1018

Figure: 21

vi) Summary Statistics

This provides a comprehensive statistical summary of fatalities and injuries recorded in the 'TrafficCrashes' dataset. The analysis includes minimum, maximum, average, and total counts for both the number of individuals killed and injured in traffic incidents. By aggregating this information, the query offers a holistic view of the severity of these incidents, aiding in understanding the overall impact on public safety.

Min_Killed	Max_Killed	Avg_Killed	Total_Killed	Min_Injured	Max_Injured	Avg_Injured	Total_Injured
0	3	0.010924213270436347	176	0	19	1.2157	19586

Figure: 22

D. Visualizations

Here, I used the R programming language for comprehensive visualizations for my dataset. Leveraging R's powerful ggplot2 library, I crafted bar charts showing collision severity, temporal patterns, geographical locations, and a pie chart illustrating the distribution of collision severities.

i) Bar chart for Collision Severity

The bar chart visualizes the distribution of collision severity in a dataset, employing the ggplot2 library in the R programming language. Each bar represents a specific severity category, with the height indicating the corresponding count of collisions. The chart provides a clear overview of the prevalence of different severity levels, aiding in understanding the dataset's characteristics. The title "Collision Severity Distribution" summarizes the purpose of the visualization, while the x-axis labels the severity categories, and the y-axis represents the

count of occurrences. The use of a distinct steel-blue color enhances visibility and comprehension of the presented information.

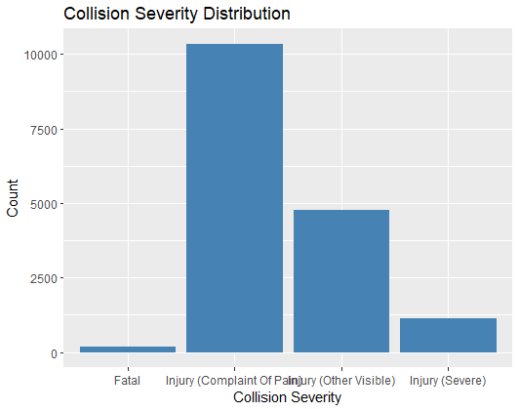


Figure: 23

ii) Time distribution of collisions

The temporal distribution of collisions is visualised using ggplot2 in R as a histogram. Each bar depicts the number of collisions during various time intervals (binwidth = 1 hour), with grey bars underlined in red. The collision timings are shown by the x-axis, while the associated count is represented by the y-axis. The figure, titled "Distribution of Collisions by Time," provides insights into temporal trends, and the tilted x-axis labels make it easier to read. The colour contrast helps to emphasise the visual representation of collision rates throughout the day.

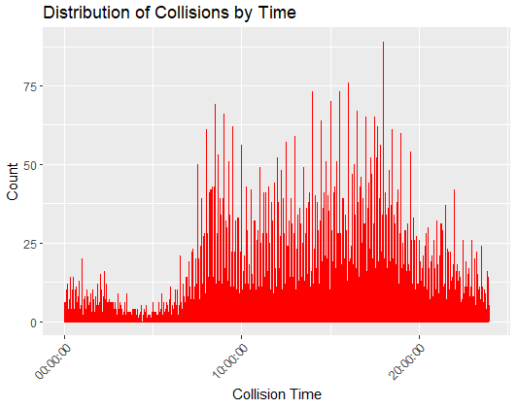


Figure: 24

iii) Scatter plot of Latitude and Longitude

The scatter plot visualizes collision locations on a map using ggplot2 in R. Each point corresponds to a collision's geographical coordinates, with a blue color and 0.5 alpha for enhanced visibility. The x-axis represents longitude, the y-axis denotes latitude, and the title, "Collision Locations on Map," succinctly conveys the chart's purpose. This visualization provides an immediate spatial understanding of where collisions are concentrated, aiding in identifying patterns and potential areas of interest for further analysis or intervention.

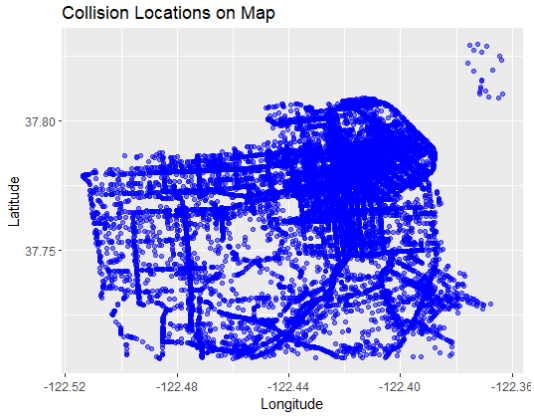


Figure: 25

iv) Bar chart for Day of Week

The bar chart visualizes collisions by day of the week using ggplot2 in R. Each bar corresponds to a specific day, colored in purple for clarity. The x-axis represents the days of the week, ordered chronologically, and the y-axis denotes the count of collisions. Titled "Collisions by Day of Week," this chart offers a comprehensive overview of the dataset's temporal distribution. The rotated x-axis labels enhance readability, and the distinctive color aids in differentiating between days. This visualization facilitates the identification of patterns and trends in collision occurrences throughout the week.

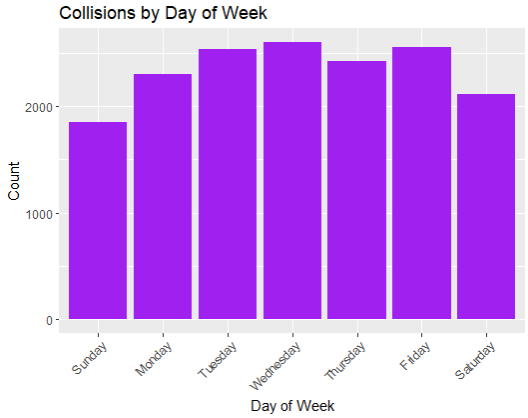


Figure: 26

v) Pie chart for Collision Severity distribution

The pie chart illustrates the distribution of collision severities based on a summary table created using ggplot2 in R. Each segment represents a severity category, with the size proportionate to the count of collisions. Titled "Distribution of Collision Severities," the chart offers a visually intuitive representation of the dataset's severity composition. The color-coded legend clarifies the severity levels, and the radial layout enhances the viewer's perception. The absence of unnecessary elements in the theme_void() setting keeps the focus on the distribution pattern. This visualization aids in quickly grasping the relative frequencies of different collision severities within the dataset.

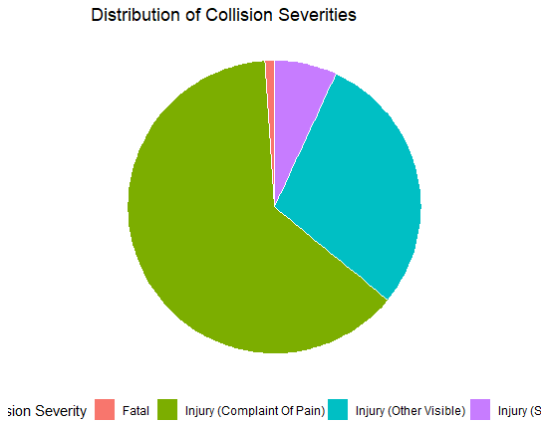


Figure: 27

E: NOIR Analysis

Sno.	Column Name	Noir Type
1	Unique_Id	Nominal
2	Cnn_Intrsctn_Fkey	Nominal
3	Cnn_Sgmt_Fkey	Nominal
4	Case_Id_Pkey	Nominal
5	Tb_Latitude	Ratio
6	Tb_Longitude	Ratio
7	Geocode_Source:	Nominal
8	Geocode_Location	Nominal
9	Collision_Datetime	Interval
10	Collision_Date	Nominal
11	Collision_Time	Interval
12	Accident_Year	Interval
13	Month	Ordinal
14	Day_Of_Week	Ordinal
15	Time_Cat	Nominal
16	Juris	Nominal
17	Officer_Id	Nominal
18	Reporting_District	Nominal
19	Beat_Number	Nominal
20	Primary_Rd	Nominal
21	Secondary_Rd	Nominal
22	Distance	Ratio
23	Direction	Nominal
24	Weather_1	Nominal
25	Collision_Severity	Ordinal
26	Type_Of_Collision	Nominal
27	Mviw	Nominal
28	Ped_Action	Nominal
29	Road_Surface	Nominal
30	Road_Cond_1	Nominal
31	Road_Cond_2	Nominal

32	Lighting	Nominal
33	Control_Device	Nominal
34	Intersection	Nominal
35	Vz_Pcf_Code	Nominal
36	Vz_Pcf_Group	Nominal
37	Vz_Pcf_Description	Nominal
38	Vz_Pcf_Link	Nominal
39	Number_Killed	Ratio
40	Number_Injured	Ratio
41	Street_View	Nominal
42	Dph_Col_Grp	Nominal
43	Dph_Col_Grp_Description	Nominal
44	Party_At_Fault	Nominal
45	Party1_Type	Nominal
46	Party1_Dir_Of_Travel	Nominal
47	Party1_Move_Pre_Acc	Nominal
48	Party2_Type	Nominal
49	Party2_Dir_Of_Travel	Nominal
50	Party2_Move_Pre_Acc	Nominal
51	Point	Ratio
52	Data_As_Of	Interval
53	Data_Updated_At	Interval
54	Data_Loaded_At	Interval
55	Neighborhoods	Nominal
56	Sf Find Neighborhoods	Nominal
57	Current Police Districts	Nominal
58	Current Supervisor Districts	Nominal
59	Analysis Neighborhoods	Nominal

Table: 1

F: NLP Methods

I used natural language processing (NLP) techniques to analyze text data in a dataset of traffic crash records. Using the NLTK library, I performed data preprocessing by converting text to lowercase, removing special characters and stopwords, and applying sentiment analysis with the Sentiment Intensity Analyzer. The "Weather_1" column was selected for analysis. The resulting DataFrame includes sentiment scores and labels (Positive, Negative, or Neutral) for each entry, providing insights into the emotional tone of weather-related descriptions in the dataset. This NLP approach enhances understanding of textual information, contributing valuable context to traffic crash records.

```
      Weather_1  Sentiment_Score Sentiment_Label
0         Clear           0.3818        Positive
1         Clear           0.3818        Positive
2         Clear           0.3818        Positive
3         Clear           0.3818        Positive
4         Clear           0.3818        Positive
...         ...              ...             ...
16391   Raining           0.0000         Neutral
16392    Cloudy           0.0000         Neutral
16393     Clear           0.3818        Positive
16394     Clear           0.3818        Positive
16395     Clear           0.3818        Positive

[16396 rows x 3 columns]
```

Figure: 28

5. Results

i) What are the most common weather conditions associated with traffic crashes resulting in injury?

This question aims to find the relationship between weather conditions and traffic crashes resulting in injuries. The primary focus is on understanding the prevailing weather patterns during such incidents to enhance our knowledge of the environmental factors contributing to road safety issues. Utilizing a dataset containing comprehensive information on various collision scenarios, we seek to identify the most common weather conditions associated with injury-causing traffic accidents.

Weather Conditions Associated with Traffic Crashes Resulting in Injury:

In my analysis of traffic crashes resulting in injury, first, I isolated the relevant data points. Subsequently, I conducted an exploration of the most common weather conditions associated with these injury-related incidents.

```
Weather Conditions Associated with Traffic Crashes Resulting in Injury:
Clear                  13430
Cloudy                  1747
Raining                  848
Not Stated                87
Fog                       65
Other                     32
Wind                      13
Other: Not On Scene        1
Snowing                    1
Name: Weather_1, dtype: int64
```

Figure: 29

The above figure shows the most common weather conditions are clear skies, accounting for 13,430 incidents. Cloudy weather follows with 1,747 cases, while raining conditions are associated with 848 incidents. Notably, there are 87 cases where the weather status is not stated. Other weather conditions, including fog, wind, and snowing, contribute to a smaller number of incidents. It's crucial to address the impact of diverse weather patterns on road safety to enhance preventive measures. The dataset underscores the significance of weather-related factors in understanding and mitigating injuries stemming from traffic accidents.

Visualization: The Python code generates a bar plot using the Seaborn library to illustrate the top 10 weather conditions associated with traffic crashes resulting in injuries. The figure size is set to 12 by 6 inches to ensure clarity. The x-axis displays different weather conditions, while the y-axis indicates the corresponding number of crashes. The color palette “viridis” is applied to enhance visual distinction. The plot is titled "Weather Conditions Associated with Traffic Crashes Resulting in Injury," and the x-axis label is “Weather Condition,” while the y-axis label is “Number of Crashes.” Additionally, the x-axis labels are rotated for better readability. The “tight_layout()” function ensures optimal spacing, and “plt.show()” displays the final plot.

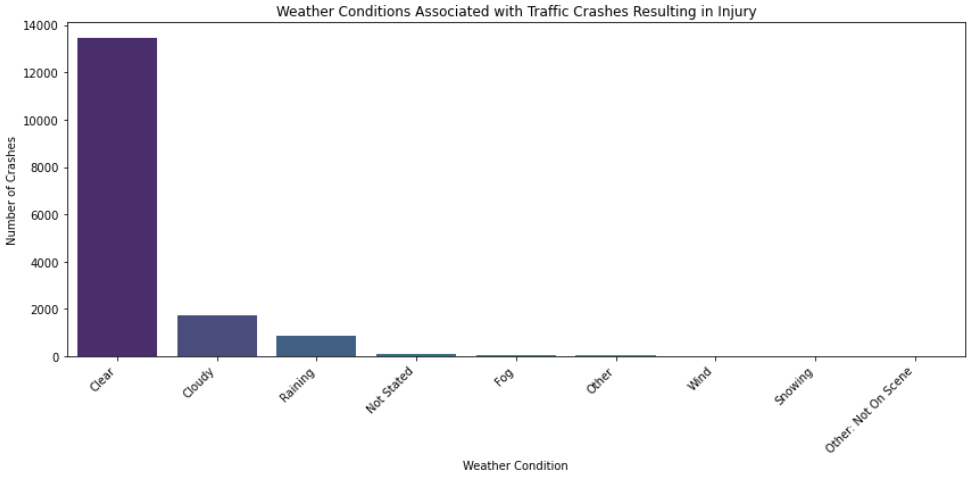


Figure: 30

ii) Are there specific days of the week or times of the day when these accidents are more likely to occur?

The question focuses on understanding if there are specific patterns in the occurrence of accidents based on days of the week and times of the day. Analyzing the dataset, which contains detailed information about traffic accidents in San Francisco, allows for insights into the temporal distribution of these incidents.

Traffic accidents represent a significant concern in urban areas, posing risks to public safety and emphasizing the need for targeted preventive measures. To better address this issue, it is crucial to investigate whether there are discernible patterns in the occurrence of accidents based on specific temporal factors.

During the analysis, the 'Collision_Datetime' column was converted to a datetime format. Subsequently, new columns were created to extract the day of the week and hour of the day from the collision timestamps.

Accidents Across Days of the Week:

```
Accidents Distribution Across Days of the Week:
Wednesday    2602
Friday       2553
Tuesday      2543
Thursday     2427
Monday       2301
Saturday     2118
Sunday       1852
Name: day_of_week, dtype: int64
```

Figure: 31

From the above figure we can see that the distribution of accidents across days revealed that Wednesday recorded the highest incidents (2602), followed closely by Friday (2553) and Tuesday (2543). Conversely, fewer accidents occurred on weekends, with Sunday having the lowest count (1852). This information provides a snapshot of the temporal patterns of accidents, highlighting potential focus areas for traffic safety measures during mid-week periods compared to weekends in San Francisco.

Visualization: The bar plot illustrates the distribution of traffic accidents across days of the week in San Francisco. Displaying a clear visual representation, Wednesday emerges as the day with the highest number of accidents, followed closely by Friday and Tuesday. Conversely, weekends exhibit a lower frequency of accidents, with Sunday having the fewest incidents. This graphical representation enhances the understanding of the weekly variations in accident rates, providing valuable insights for prioritizing safety measures. The ordered presentation of weekdays enhances clarity, emphasizing the mid-week peak in accident occurrences and offering a concise visual summary for stakeholders and policymakers involved in traffic safety initiatives.

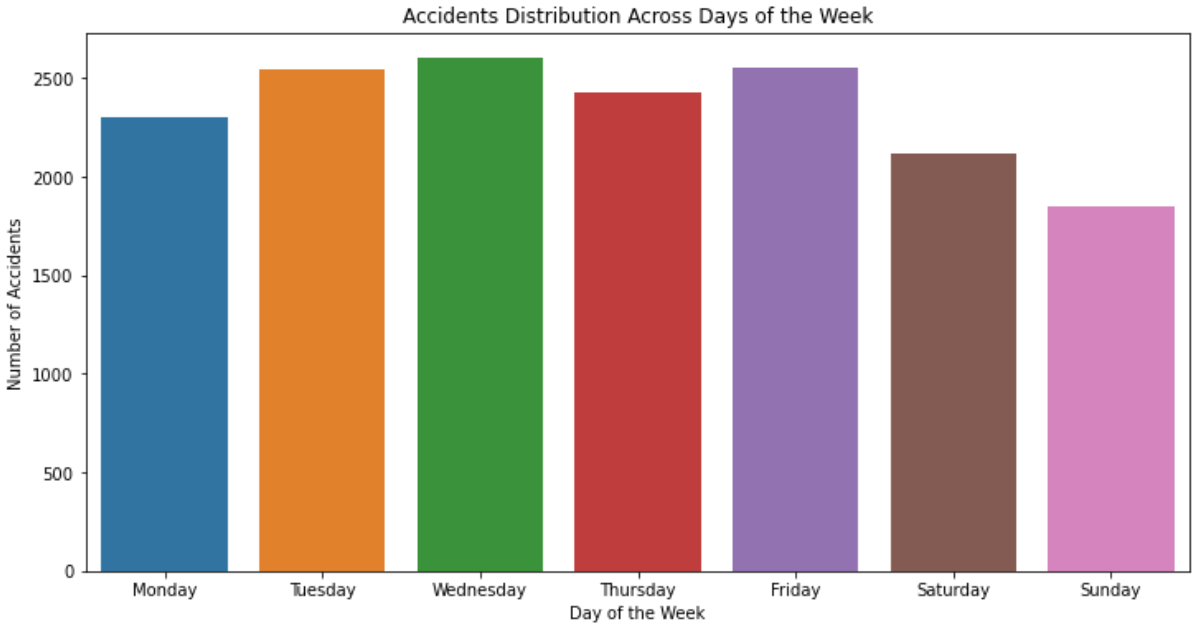


Figure: 32

Accidents Across Times of the Day:

The distribution of traffic accidents across times of the day reveals notable patterns in San Francisco. Examining the 'hour_of_day' column, it is evident that the majority of accidents occur during the late afternoon and early evening hours. Specifically, the highest count is observed during the 5:00 PM hour, suggesting heightened risk during evening commuting hours. Conversely, the early morning hours record the fewest accidents, with the period between 3:00 AM and 5:00 AM exhibiting the lowest frequency. This temporal analysis provides crucial insights into daily accident trends, informing targeted interventions and safety measures, especially during peak traffic hours in the late afternoon.

```
Accidents Distribution Across Times of the Day:
17    1353
18    1242
15    1196
16    1177
14    1064
8     1012
13     925
12     908
9      886
19     867
11     838
10     731
20     649
21     605
7      605
22     510
23     447
0      291
2      250
6      247
1      247
5      137
3      121
4       88
Name: hour_of_day, dtype: int64
```

Figure: 33

Visualization: The bar plot visualizes the distribution of traffic accidents across different hours of the day in San Francisco. Displaying a clear temporal pattern, the peak in accidents occurs during the late afternoon, with 5:00 PM registering the highest frequency. This trend suggests increased vulnerability during evening rush hours. In contrast, the early morning hours, particularly between 3:00 AM and 5:00 AM, exhibit the lowest accident counts. The graphical representation provides a concise overview of daily variations in accident rates, offering valuable insights for targeted safety interventions and traffic management strategies, especially during the identified peak hours of risk in the late afternoon.

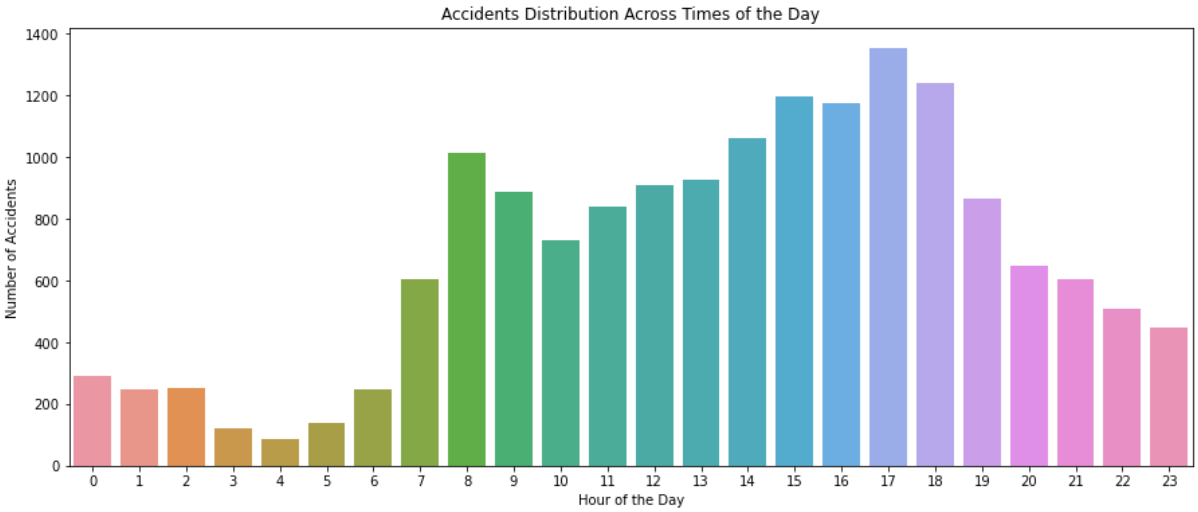


Figure: 34

iii) Do certain intersections or road types have a higher frequency of injury-related accidents

In this analysis, the aim is to investigate the factors contributing to injury-related accidents within the context of intersections and road types. The dataset provided contains detailed information about various accidents, including the location, circumstances, and severity of each incident. Our specific focus is on understanding whether certain intersections or types of roads exhibit a higher frequency of accidents resulting in injuries.

To address the research question, I performed an in-depth analysis of the dataset, exploring patterns and trends related to injury-related accidents. Considering the factors such as the type of collision, road conditions, and characteristics of the intersections involved. By aggregating and summarizing this information, we aim to identify any notable associations between specific intersections or road types and the frequency of injury-related accidents.

Top 10 Intersections with the Highest Frequency of Injury-Related Accidents:

The analysis focuses on injury-related accidents within the dataset, specifically filtering data where collision severity indicates injuries. The top 10 intersections with the highest frequency of such incidents are highlighted. The most common scenario is accidents occurring "Midblock > 20Ft," constituting 9369 cases. "Intersection <= 20Ft" follows with 4048 incidents, emphasizing the significance of accidents at or near intersections. Additionally, "Intersection Rear End <= 150Ft" is notable, suggesting a prevalence of rear-end collisions within close proximity to intersections, total 2807 cases. These findings shed light on critical areas requiring targeted safety interventions and underline the importance of intersection-specific measures to reduce injury-related accidents.

```
Top 10 Intersections with the Highest Frequency of Injury-Related Accidents:
Midblock > 20Ft                  9369
Intersection <= 20Ft             4048
Intersection Rear End <= 150Ft    2807
Name: Intersection, dtype: int64
```

Figure: 35

Visualization: The bar plot visualizes the top 10 intersections with the highest frequency of injury-related accidents. The x-axis represents different intersection types, while the y-axis indicates the corresponding number of accidents. "Midblock > 20Ft" stands out as the most prevalent intersection type, with nearly 9369 injury-related incidents. Following closely are "Intersection <= 20Ft" and "Intersection Rear End <= 150Ft" with 4048 and 2807 accidents, respectively. The visualization provides a clear overview of the intersections posing a higher risk of injury-related accidents, emphasizing the need for targeted safety measures and interventions in these specific scenarios.

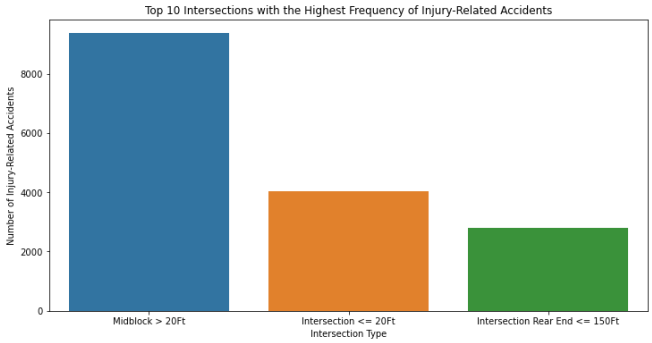


Figure: 36

Road Types with the Highest Frequency of Injury-Related Accidents:

The analysis now extends to road types, specifically focusing on injury-related accidents. The output reveals the top 10 road surface conditions associated with the highest frequency of injuries. "Dry" surfaces predominate with a significant count, implying that accidents on dry roads lead to more injuries. This is followed by "Wet" conditions, indicating that accidents in wet weather contribute to a substantial number of injuries. The findings emphasize the influence of road surface conditions on the severity of accidents and underscore the importance of addressing safety measures tailored to diverse weather and road conditions to mitigate injury risks effectively.

```
Road Types with the Highest Frequency of Injury-Related Accidents:
Dry             14514
Wet              1487
Not Stated        200
Slippery           13
Snowy Or Icy       10
Name: Road_Surface, dtype: int64
```

Figure: 37

Visualization: The bar plot visually represents the top 10 road surface conditions associated with the highest frequency of injury-related accidents. The x-axis displays different road types, while the y-axis indicates the corresponding number of accidents. "Dry" surfaces stand out prominently, with a substantial count, indicating that injury-related accidents are most prevalent on dry roads. The plot also highlights "Wet" conditions as the second-highest contributor to injuries. This visualization provides a clear understanding of the road surface conditions influencing injury-related accidents, emphasizing the need for targeted safety measures, especially during adverse weather conditions, to mitigate the impact of accidents on road users' well-being.

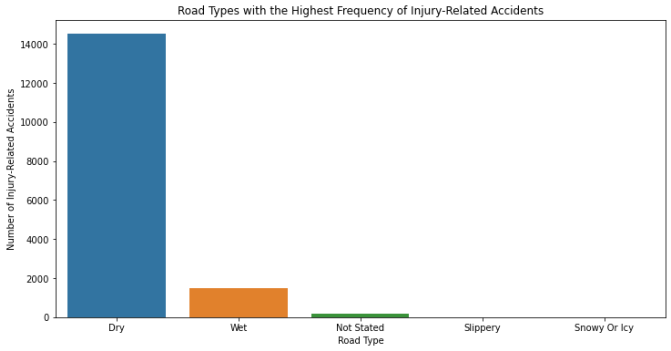


Figure: 38

6. Fitting a Linear Regression Model:

I used a linear regression model to examine the relationship between the severity of traffic collisions (measured numerically) and the resulting number of injuries.

The variable "Collision_Severity" is converted into a numerical format for further analysis. The levels "Fatal," "Injury (Complaint Of Pain)," and "Injury (Other Visible)" are assigned numeric values using the as.numeric(factor()) function. This conversion enables the incorporation of collision severity as a numerical predictor in subsequent statistical analyses. Following this transformation, summary statistics are generated for the dataset using the summary() function. Additionally, a box plot is created using the ggplot2 package to visually explore the distribution of the number of injuries across different collision severity categories. This graphical representation provides an initial insight into the relationship between collision severity and the number of injuries.

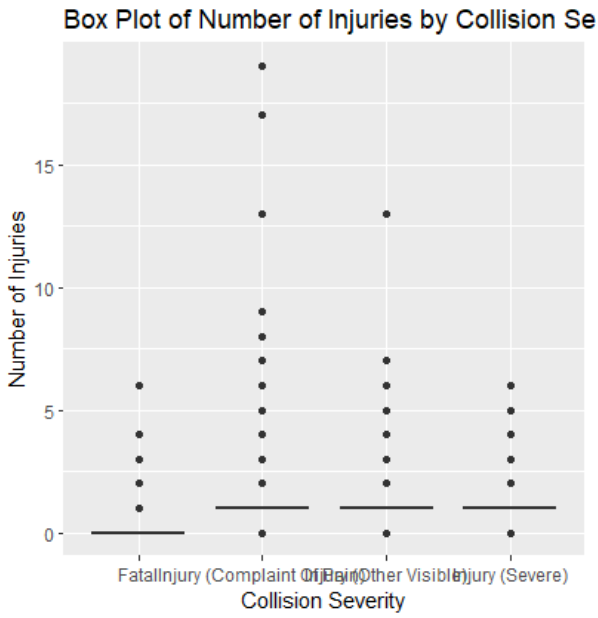


Figure: 39

I built a linear regression model using the lm() function to examine the relationship between the numeric representation of collision severity (Collision_Severity_Num) and the number of injuries (Number_Injured). The regression model aims to quantify the impact of collision severity on the observed variation in the number of injuries. The summary() function provides detailed statistical information about the regression model, including coefficients, significance levels, and goodness-of-fit measures.

```
Call:
lm(formula = Number_Injured ~ Collision_Severity_Num, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2204 -0.2204 -0.2128 -0.2128 17.7872

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.197840   0.026020  46.035   <2e-16 ***
Collision_Severity_Num 0.007504   0.011069   0.678    0.498
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6599 on 15265 degrees of freedom
  (1129 observations deleted due to missingness)
Multiple R-squared:  3.011e-05, Adjusted R-squared:  -3.54e-05
F-statistic: 0.4596 on 1 and 15265 DF,  p-value: 0.4978
```

Figure: 40

The subsequent visualization employs the ggplot2 package to create a scatter plot with a fitted regression line, offering a visual representation of the linear relationship between collision severity and the number of injuries.

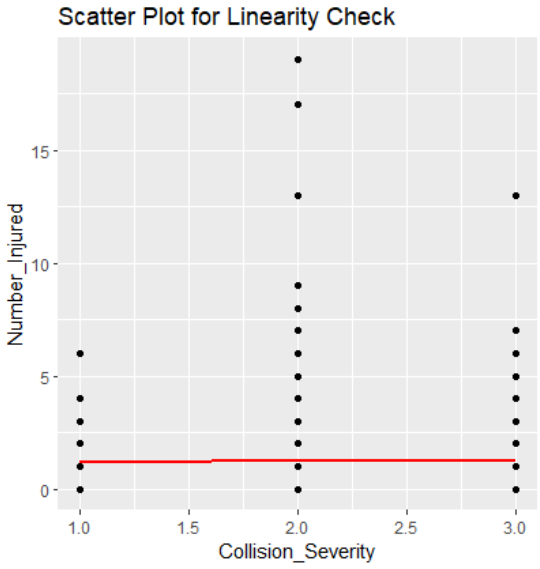


Figure: 41

Predictions are generated using the linear regression model previously built. The "predict()" function is employed to estimate the number of injuries ('Predicted_Number_Injured') based on the model and the existing data in the dataframe ('df'). The newly predicted values are then added as a new column to the dataframe. The 'head()' function is used to display the first few rows of the dataset, specifically focusing on the columns related to the actual number of injuries ('Number_Injured') and the predicted number of injuries ('Predicted_Number_Injured'). This allows for a quick comparison between observed and predicted values, offering insights into

the model's performance in capturing the relationship between collision severity and the number of injuries.

	Number_Injured <dbl>	Predicted_Number_Injured <dbl>
1	1	1.22
2	1	1.21
3	0	1.21
4	2	1.21
5	1	1.21
6	1	1.22

Figure: 42

The Scatter plot depicts the actual number of injuries (Number_Injured) against the predicted values (Predicted_Number_Injured) from the linear regression model. Each point represents an observation, with actual values in black and the model's predictions in blue. The plot visually assesses the model's accuracy by comparing observed and forecasted outcomes. The inclusion of a fitted blue line helps visualize how well the model captures the overall trend. This visualization is crucial for evaluating the predictive performance of the regression model in estimating the number of injuries based on the numeric representation of collision severity.

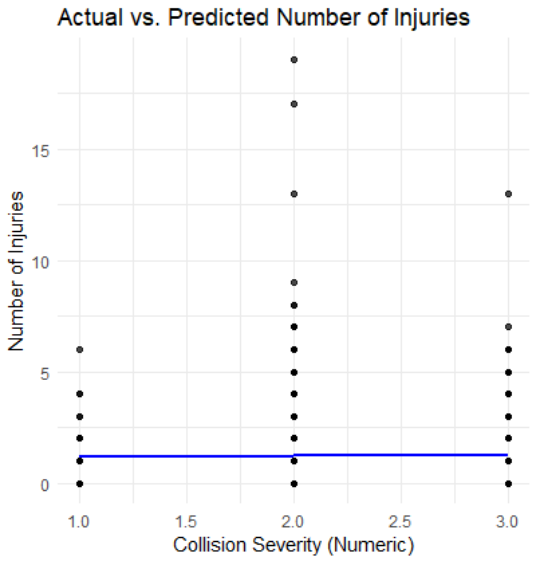


Figure: 43

7. Conclusion

In conclusion, this research employs a multidimensional approach to enhance road safety by analyzing traffic crashes resulting in injuries. This study delves into the intricate interplay of adverse weather conditions, temporal patterns, and intersection-related factors, utilizing advanced statistical models and geospatial analysis techniques. Findings affirm the significant impact of adverse weather, with rain, snow, sun glare, and high winds exhibiting distinct correlations with various collision types. Temporal analysis reveals heightened risk during Saturdays and late afternoons, offering actionable insights for targeted safety measures. The frequency analysis of Equivalent Property-Damage-Only (EPDO) crashes at intersections unravels location-specific factors such as pavement friction, roadway functional classification, road surface type, and more, shedding light on their pivotal role in collision dynamics.

Drawing from related works, the research aligns with and extends existing literature, contributing nuanced insights to the broader discourse on road safety. The dataset, comprising comprehensive information on collisions, serves as a robust foundation for the analyses conducted.

Visualizations crafted using the R programming language provided intuitive representations of collision severity distribution, temporal patterns, geographical concentrations, and day-of-week insights. These visualizations offer stakeholders a clear understanding of the dataset's characteristics, aiding in the identification of patterns and trends crucial for targeted interventions.

Natural language processing techniques were applied to analyse textual data related to weather conditions, revealing clear skies as the most common weather pattern associated with injury-related accidents. This NLP approach added a nuanced layer to the understanding of environmental factors impacting road safety.

The analysis of weather conditions, temporal patterns, and intersection/road types in San Francisco's traffic accidents dataset revealed critical insights. Clear skies were associated with the highest injury-related incidents, emphasizing weather's role in road safety. Mid-week days, particularly Wednesday, showed elevated accident frequencies, with late afternoons presenting peak risks. "Midblock > 20Ft" intersections and "Dry" road surfaces were identified as high-risk scenarios. These findings underscore the need for targeted safety interventions at specific intersections and during identified high-risk conditions to reduce injury-related accidents in the city.

The linear regression model provided a quantitative evaluation of the relationship between collision severity and the number of injuries, offering predictive insights into injury outcomes. The model's performance was visually assessed, providing a comprehensive evaluation of its accuracy in capturing the observed trends.

In summary, this research amalgamated diverse analytical techniques to provide a holistic understanding of the factors influencing traffic accidents in San Francisco. The derived insights have significant implications for policymakers, traffic management authorities, and other stakeholders, facilitating evidence-based decision-making for the improvement of road safety and the reduction of injury-related incidents.

8. Limitations

i) Data Limitations: The accuracy and reliability of the results are contingent on the quality and completeness of the dataset. Incomplete or inaccurate entries, underreporting, or misclassifications may introduce biases and impact the robustness of the analysis.

ii) Temporal and Geographic Scope: The study focused specifically on traffic crash data in San Francisco, and findings may not be universally applicable to other regions with distinct traffic patterns, infrastructure, and socio-economic factors.

iii) Weather Analysis Complexity: While weather conditions were analyzed using NLP, the dataset might not capture the full spectrum of weather-related nuances. Future research could benefit from more sophisticated meteorological data integration for a more nuanced understanding of weather impacts.

iv) Intersection and Road Type Associations: The identified intersections and road types with high injury-related accident frequencies provide insights, but causation cannot be definitively inferred. External factors, such as traffic signal efficiency, road maintenance, or local events, were not explicitly considered and could influence accident rates.

v) Linear Regression Assumptions: The linear regression model assumes a linear relationship between collision severity and the number of injuries. Non-linear relationships or the influence of unobserved variables could impact the model's accuracy.

vi) Generalization of Preventive Measures: While specific intersections and road types were highlighted, caution is needed when generalizing preventive measures. Tailored interventions may be required for different locations, and a one-size-fits-all approach might not be effective.

vii) Temporal Patterns and Causation: The identified temporal patterns do not establish causation. Factors contributing to accidents on specific days or times were not exhaustively examined and may involve complex interplays of variables.

viii) External Factors: External factors such as economic conditions, city development projects, or changes in transportation policies were not considered. These factors could influence traffic patterns and safety outcomes.

9. Future work

i) Multifactorial Analysis: Investigate the interaction of multiple factors simultaneously, such as the combined influence of weather, road conditions, and time of day on accident severity. This could provide a more holistic understanding of the complexities involved in traffic safety.

ii) Machine Learning Models: Explore the application of advanced machine learning models for predictive analysis. Models such as random forests or neural networks could offer more nuanced predictions and uncover non-linear relationships within the dataset.

iii) Temporal Trends and Urban Development: Examine the impact of long-term urban development and infrastructure changes on traffic patterns and safety. Analyzing temporal trends over several years could reveal evolving patterns and inform city planning for safer roadways.

iv) Behavioral Analysis: Integrate behavioral data, such as driver characteristics and driving habits, to understand how human factors contribute to accidents. This could provide insights into preventive measures targeting specific behaviors.

v) Real-time Data Integration: Incorporate real-time data sources, such as live traffic feeds, weather updates, and road closures, to enhance the timeliness and accuracy of safety analyses. This would enable proactive interventions based on current conditions.

vi) Public Health Implications: Explore the public health implications of traffic accidents, considering long-term physical and mental health outcomes for individuals involved. This could inform public health policies and interventions.

vii) Community Engagement: Involve local communities in the research process to gather qualitative insights and community-specific knowledge. Engaging with residents can provide a more nuanced understanding of localized issues and potential solutions.

viii) Comparative Studies: Conduct comparative studies with datasets from different cities or regions to identify commonalities and differences in traffic safety patterns. This could contribute to the development of more universally applicable safety measures.

ix) Intervention Effectiveness: Evaluate the effectiveness of specific safety interventions implemented as a result of previous analyses. Assessing the impact of measures such as improved signage or traffic signal changes can provide valuable feedback for future initiatives.

x) Integration of Autonomous Vehicles: Investigate the impact of autonomous vehicles on traffic safety. Analyzing data from areas with increased autonomous vehicle presence could offer insights into the potential of emerging technologies in reducing accidents.

10. References

Traffic crashes resulting in injury - catalog. (2023b, October 13). https://catalog.data.gov/dataset/traffic-crashes-resulting-in-injury

Becker, N., Rust, H. W., & Ulbrich, U. (2022). Weather impacts on various types of road crashes: a quantitative analysis using generalized additive models. *European Transport Research Review*, *14*(1). https://doi.org/10.1186/s12544-022-00561-2

Car crashes by time of day and day of week - Injury facts. (2023, April 18). Injury Facts. https://injuryfacts.nsc.org/motor-vehicle/overview/crashes-by-time-of-day-and-day-of-week/

Sharafeldin, M., Ksaibati, K., & Gerow, K. G. (2023). Frequency analysis of Equivalent Property-Damage-Only (EPDO) crashes at intersections. *Eng*, *4*(2), 1116–1126. https://doi.org/10.3390/eng4020064