Report: Traffic Collision Data Analysis

Traffic collisions pose significant risks to public safety, requiring continuous monitoring and analysis to enhance road safety measures. Government agencies, city planners, and policymakers must leverage data-driven insights to improve infrastructure, optimize traffic management, and implement preventive measures.

1. Data Preparation

1.1. Loading the dataset

1.1.1. Check for data consistency and ensure all columns are correctly formatted.

- Spark's type inference can lead to inaccuracies and precision loss, especially with large integer IDs.
- This issue affects case_id columns in parties_df, victims_df, and collision_df.
- Define case_id as DecimalType(22,0) for precise storage.
- Implicit inference may cause data loss or unexpected behavior.
- Set columns like party_age and vehicle_year to StringType initially to better handle non-numeric entries and NULL values before casting to IntegerType.

2. Data Cleaning

2.1. Missing Values

2.1.1. Check for Missing Values

- For collisions_df, many columns like reporting_district, caltrans_county, caltrans_district, state_route, postmile, side_of_highway, and pcf_violation_subsection showed a high percentage of missing values (ranging from approximately 59.06% to 72.89%).
- For parties_df, columns like party_sex, party_age, party_sobriety, cellphone_in_use, cellphone_use_type, vehicle_year, vehicle_make, statewide_vehicle_type, chp_vehicle_type_towing, and party_race had a significant percentage of missing values (ranging from approximately 23.15% to 48.72%).
- For victims_df, victim_sex, victim_age, victim_safety_equipment_1, and victim_safety_equipment_2 had missing values over 39%.

2.1.2. Drop Sparse Columns

- For this analysis, after analysing the missing value percentage for each dataset, the threshold was set to 50%.

- For collision_df: The columns reporting_district, caltrans_county, caltrans_district, state_route, postmile, side_of_highway, and pcf_violation_subsection were identified as having more than 50% missing values and were dropped.
- Knowing the importance of some columns for analysis, latitude, longitude, and location_type were explicitly excluded from being dropped, even if sparse.
- For case_id_df, parties_df and victims_df: No columns were dropped under the 50% sparsity threshold.
- Furthermore, I created a list of all the important columns from collision_df and kept only those for further analysis.

2.1.3. Convert Data Type

- Collision dates and process dates are converted to DateType in a specific format.
- All numeric columns are ensured to be converted to integer type.

2.1.4. Handle Missing Values

- Numerical columns: Missing values (NULL) in integer, double, and decimal columns were imputed with 0.
- String columns: Missing string values (NULL or empty strings) were imputed with the string 'Unknown'.
- Timestamp columns: Missing values in collision_time were filled with the earliest existing timestamp in that column (2026-01-14 00:00:00 for collisions_df).

2.2. Fixing Columns

2.2.1. Remove Duplicates

- 35 duplicate records were removed from case_id dataframe.
- 1 and 0 duplicate records were removed from collisions, parties and victims dataframes.

2.2.2. Detect Outliers using IQR

- The Interquartile Range (IQR) method was used to identify outliers in numerical columns.
- case_df: No outliers detected.
- collisions_df: Significant outliers were found in injured_victims (33,313) and longitude (266,742), indicating unusually high injury counts or potentially invalid geographic coordinates.
- parties_df: Outliers were present in party_age (11,263), cellphone_in_use (26,406), party_number_killed (6,889), party_number_injured (15,414), and vehicle_year (190,071), suggesting extreme ages, usage patterns, or vehicle models.
- victims_df: Outliers were detected in victim_age (17,284).

2.2.3. List of numerical columns to check for outliers

- List out all numerical columns by their data types, such as integer, double, and decimal(22,0).

2.3. Handling Outliers

2.3.1. Remove Outliers

- Rows containing values outside these calculated bounds for a given numerical column were removed from the DataFrame.
- Removing outliers can improve model accuracy and stability.

3. Exploratory Data Analysis

3.1. Data Preparation

3.1.1. Classify variables into categorical and numerical.

- case_id_df:
- Categorical Columns: None
- Numerical Columns: ['case_id', 'db_year']
- collision_df:
- Categorical Columns: ['county_location', 'weather_1', 'collision_severity', 'party_count', 'type_of_collision', 'road_surface', 'road_condition_1', 'lighting']
- Numerical Columns: ['case_id', 'jurisdiction', 'county_city_location', 'killed_victims', 'injured_victims', 'latitude', 'longitude']
- parties_df:
- Categorical Columns: ['party_type', 'party_sex', 'party_sobriety', 'direction_of_travel', 'party_safety_equipment_1', 'party_safety_equipment_2', 'financial_responsibility', 'cellphone_use_type', 'other_associate_factor_1', 'movement_preceding_collision', 'vehicle_make', 'statewide_vehicle_type', 'chp_vehicle_type_towing', 'chp_vehicle_type_towed', 'party_race']
- Numerical Columns: ['id', 'case_id', 'party_number', 'at_fault', 'party_age', 'cellphone_in_use', 'party_number_killed', 'party_number_injured', 'vehicle_year']
- victims_df:
- Categorical Columns: ['victim_role', 'victim_sex', 'victim_degree_of_injury', 'victim_seating_position', 'victim_safety_equipment_1', 'victim_safety_equipment_2', 'victim_ejected']
- Numerical Columns: ['id', 'case_id', 'party_number', 'victim_age']

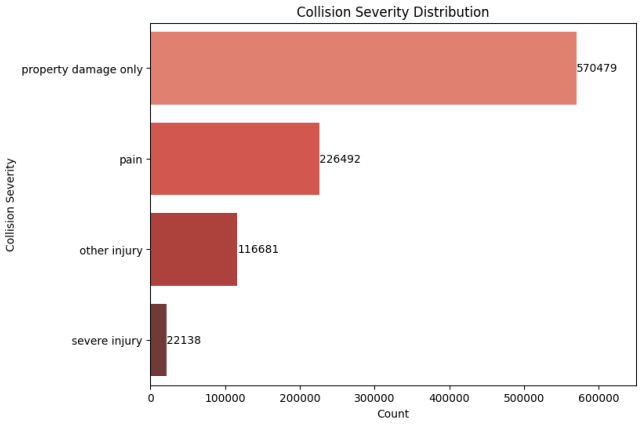
3.1.2. Encode Categorical Variables

- Most machine learning algorithms are designed to operate on numerical input. They cannot directly understand or process text-based categorical labels (e.g., 'clear', 'raining', 'cloudy' for weather).
- Converting these labels into a numerical format allows the algorithms to ingest and process the data.

3.1.3. String Indexing for Categorical Columns.

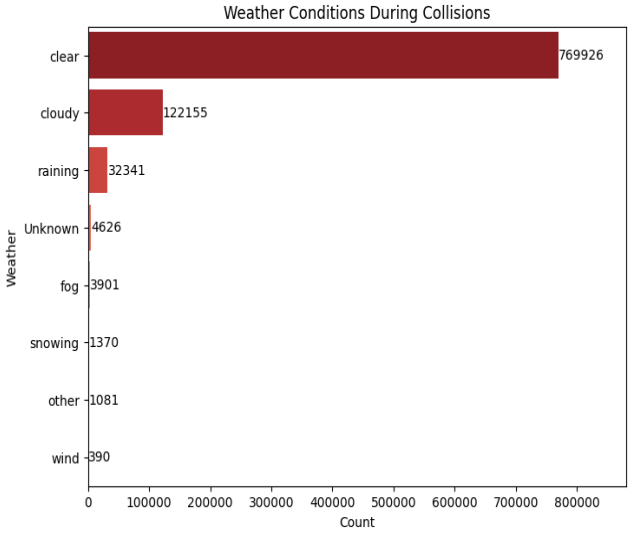
- Used StringIndexer from pyspark.ml library to assign a unique integer to each unique category in a column.
- This encoding is only applied to collision_df to support ML-models in our analysis.

3.2. Analyze the distribution of collision severity.



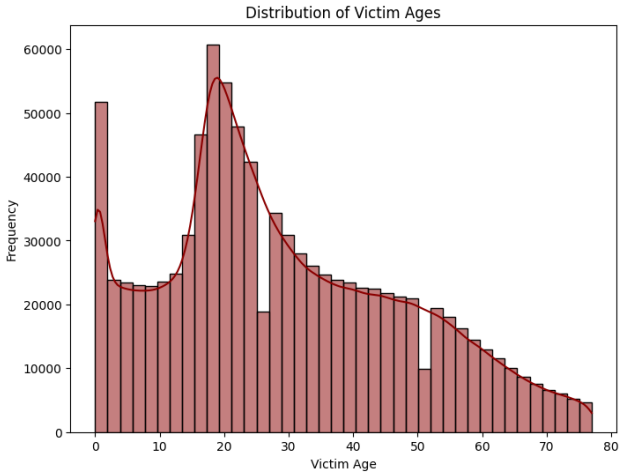
-570,479 incidents that the majority of traffic collisions in the dataset result only in ‘property damage’, without physical injuries.
-The next most common severity is 'pain' with 226,492 incidents, indicating discomfort or minor injuries. 'Other injury' incidents totaled 116,681, pointing to more significant but non-life-threatening injuries.
-The least frequent category is 'severe injury', with 22,138 incidents. Although lower in count, these incidents represent the most critical outcomes in terms of human impact.

3.3. Weather conditions during collisions.



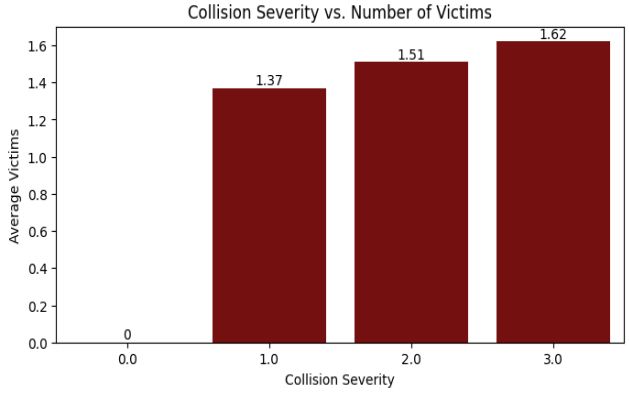
-The overwhelming majority of collisions, 769,926 incidents, occurred under clear weather conditions. It implies that other factors, such as driver behavior, road design, or traffic volume, play a more prominent role in collision occurrence during ideal weather.
- Cloudy weather was a factor in 122,155 collisions, the second most common driving condition. Rain caused 32,341 collisions, and while less than clear or cloudy weather, it still reduces visibility and traction.
- Other hazardous conditions included unknown factors (4,626 collisions), fog (3,901), snow (1,370), unspecified conditions (1,081), and wind (390).

3.4. Victims Age Distribution.



- The histograph clearly shows a very high frequency of victims in the younger age brackets of 17-30. This indicates that individuals in their late teens and 20s are highly represented among collision victims.
- There's also a significant number of victims at age 0-1 (38704). Highlighting the vulnerability of new born childern.
- After age 30 gradual decline in frequency as age increases.

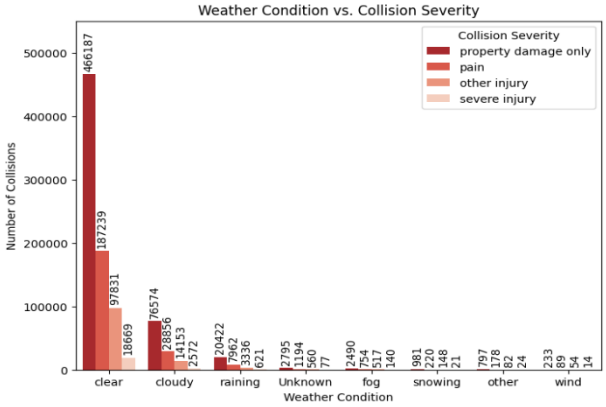
3.5. Collision Severity vs Number of Victims.



- There is a positive relationship between collision severity and number of victims. As collision severity increases, the average number of victims rises accordingly.
-Severe Injury(3.0) Collisions (Avg. 1.62 Victims): These are the most complex and impactful incidents.
-Other Injury(2.0) Collisions (Avg. 1.51 Victims): Involving multiple victims, even minor injuries can affect many people.

- Pain(1.0)-Related Collisions (Avg. 1.37 Victims): These incidents cause physical harm but are not life-threatening.
- Property Damage Only(0.0) Collisions (Avg. 0.0 Victims): As expected, these result in no injuries or fatalities.

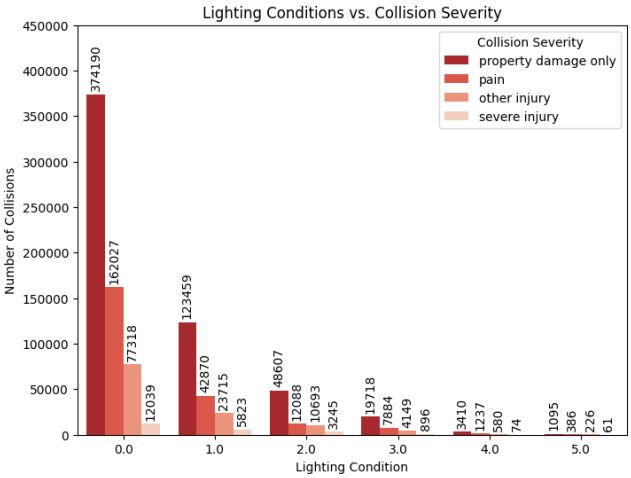
3.6. Weather Conditions vs Collision Severity.



-Weather conditions strongly correlate with collision severity. weather conditions tend to increase the likelihood of injury or severe injury when collisions occur
-The highest number of collisions occurs during clear weather.
-Most collisions under clear conditions result in property damage only or minor injuries.
-This suggests that collision frequency is driven more by exposure (higher traffic volume) than by adverse conditions.

- The next most common weather condition for collisions is 'cloudy,' followed by 'raining.' While the severity distribution in these conditions is similar to clear weather, the counts are lower. Although total collisions in 'raining' weather are fewer, the proportion of injury-related collisions is notable.
- Less Frequent Weather Conditions: Weather conditions like 'fog', 'snowing', and 'wind' lead to fewer collisions but increase incident severity. A notable number of collisions also occurred under 'unknown' weather conditions.

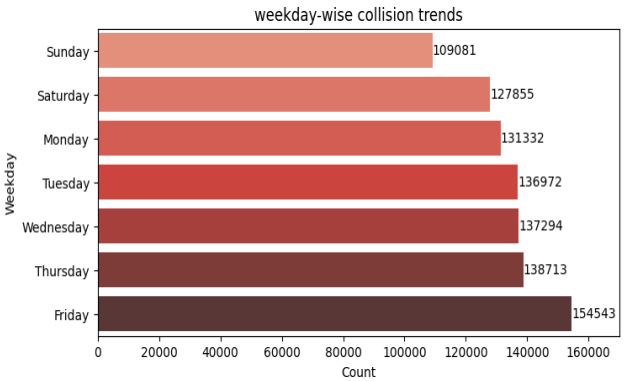
3.7. Lighting conditions vs Collision Severity.



-Lighting conditions strongly correlate with the severity of collisions.
-The majority of collisions occur in daylight(0.0), which is expected because most driving happens during the day.However, many of these collisions result in property damage only, even severity ones.
-Dark with no street lights(1.0), Dark with street lights(2.0) show a higher proportion of injury and severe injury collisions compared to daylight.

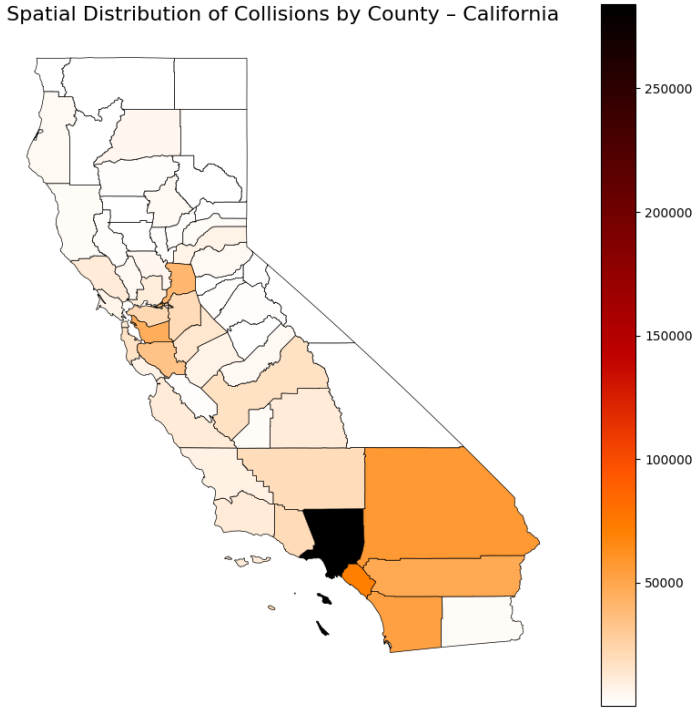
- Collisions at dusk or dawn(3.0) are fewer than in daylight, but the proportion of injury and severe injury is relatively high
- Unknown Lighting(4.0) and Dark with Street Lights Not Functioning(5.0) categories account for a smaller number of collisions

3.8. Weekday-Wise Collision Trends.



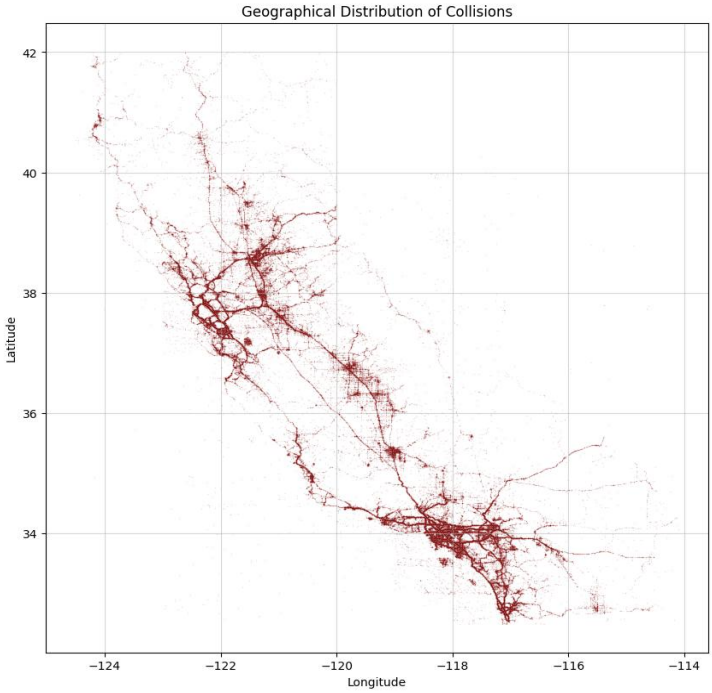
- Collisions generally rise throughout the week, peaking at the week's end..
- Friday consistently records the highest number of collisions(154,543). Increased traffic volume due to weekend travel.
- Monday to Thursday maintain consistent collision counts between 130,000 - 140,000. These days reflect regular weekday commuting patterns
- Saturday Collisions Drop as Sunday records the fewest collisions (109,081). This suggests a reduction in incidents during the weekend holiday.

3.9. Spatial Distribution of Collisions.



- Urbanized and densely populated counties tend to have higher collision counts compared to more rural counties, which appear in lighter shades on the map.
- Los Angeles and its surrounding Southern California counties (Orange, San Bernardino, San Diego, Riverside) appear as darker shades, indicating higher collision densities.
- Other densely populated regions, such as the West Coast Bay Area (Alameda, Santa Clara, Contra Costa, San Mateo, San Francisco), also show significant collision counts.
- Northern California counties are shown in lighter shades, indicating low collision areas in rural regions.

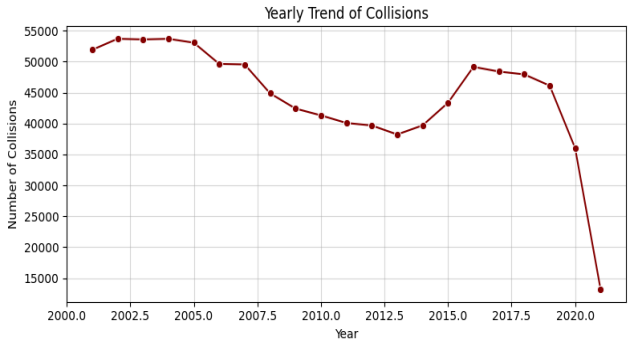
3.10. Collision Analysis by Geography.



- The scatter plot clearly shows dense clusters of red dots (representing collisions) in major urban centers. Specifically, the areas around Los Angeles, the Bay Area (San Francisco, Oakland, San Jose), San Diego, and other metropolitan regions appear significantly darker and more concentrated.
- Beyond just city centers, you can often discern patterns along major freeways and state routes.
- In contrast, rural and less populated regions of California show a much sparser distribution of dots, indicating fewer collision incidents in these areas.

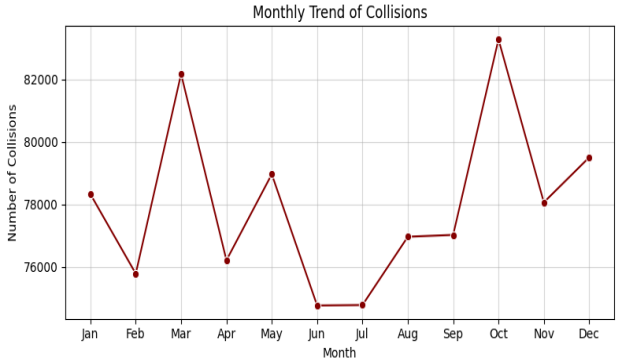
3.11. Collision Trends Over Time.

- Yearly Trend of Collisions



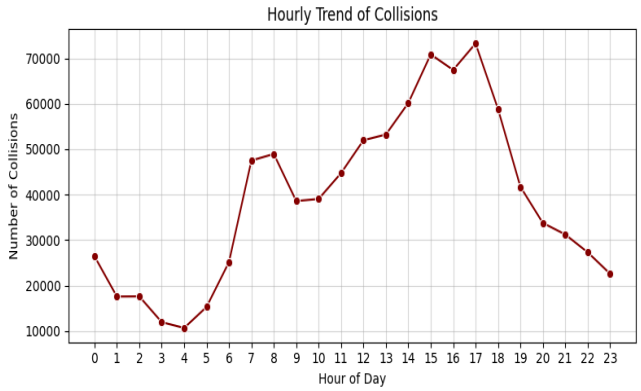
- The line plot of yearly collision counts shows a fluctuating pattern. From 2001 to 2005, collision counts were relatively high (around 51,000-53,000).
- They then saw a significant decline from 2006 to 2013, reaching a low of 38,234 in 2013.
- After 2013, collision counts generally increased again until 2018 (reaching 47,951), before showing a notable dip in 2020(Covid).

- Monthly Trend of Collisions



- The monthly trend generally shows a relatively consistent number of collisions throughout the year
- October tends to be the month with the highest number of collisions (83,274).
- March (82,167), May (78961) and December (79,499) also show higher collision counts.
- June (74,762) and July (74,776) appear to have slightly lower counts.

- Hourly Trend of Collisions



- The most dangerous period is the afternoon/evening rush hour, with a pronounced peak between 3 PM - 6 PM. 5 PM (17:00) records the highest number of collisions (73,255).
- A significant peak occurs in the morning around 7 AM - 8 AM (e.g., 47,496 at 7 AM, 48,939 at 8 AM).
- Collision counts are significantly lower during the very early morning hours (1 AM - 4 AM).

4. ETL Querying

4.1. Loading the Dataset

4.2. Top 5 Counties

county_location	Collision_count
los angeles	284100
orange	72042
san bernardino	56737
san diego	53104
riverside	48686

4.3. Month with Highest Collisions

collision_month	Collision_count
10	83274

The month of October has the highest number of collisions.

4.4. Weather Conditions with Highest Collisions.

weather	weather_indexed	Collision_count
clear	0.0	769926

Clear weather is the most common condition during collisions.

4.5. Fatal Collisions.

percentage_fatalities
0.0

A fatal collision is any collision in which at least one person died. The data indicates that there have been no fatalities, and the number of victims killed is zero.

4.6. Dangerous Time for Collisions.

collision_hour	Collision_count
17	73255

5 PM (17:00) records the highest number of collisions.

4.7. Road Surface Conditions.

road_surface	road_surface_indexed	Collision_count
dry	0.0	845636
wet	1.0	76833
Unknown	2.0	8141
snowy	3.0	4122
slippery	4.0	1048

4.8. Lighting Conditions.

lighting	lighting_indexed	Collision_count
daylight	0.0	625574
dark with street ...	1.0	195867
dark with no street.	2.0	74633

5. Conclusion

Final insights:

- High-Risk Locations: Los Angeles, Orange, San Bernardino, San Diego, and Riverside counties consistently exhibit the highest collision frequencies, acting as primary high-risk zones.
- Peak Accident Times: The most dangerous periods are between 3 PM and 6 PM (afternoon rush hour) on weekdays, with Fridays experiencing the most incidents. October, March, and December also show slightly elevated collision frequencies.
- Vulnerable Road Users (VRUs): Collisions involving pedestrians and bicyclists showed an increasing trend from 2012-2013, peaking around 2018-2019, highlighting specific periods and user groups requiring targeted interventions.
- Environmental Factors: While most collisions occur under seemingly ideal conditions (clear weather, daylight, dry roads), adverse conditions like raining weather, wet road surfaces, and dark lighting (with or without streetlights) significantly increase the risk and severity of injury-related collisions.

Recommendations:

1. Targeted Infrastructure & Enforcement: Implement targeted infrastructure improvements (e.g., enhanced lighting, road design modifications, dedicated bike lanes) and increase law enforcement presence in identified high-risk urban areas and major freeways. These efforts should be particularly focused during peak afternoon rush hours on Fridays and in the high-risk months of October, March, and December.
2. Dynamic Traffic Management: Optimize traffic management by analyzing trends in collision severity, weather conditions, and lighting. This includes making data-driven adjustments to traffic signal timings, improving road design, and implementing intelligent transportation systems for real-time congestion warnings, especially during adverse weather conditions and dark hours.
3. Enhanced Pedestrian and Cyclist Safety: Propose data-driven policy changes such as targeted safety campaigns for vulnerable road users (focused on peak years like 2012-2019), increased enforcement against distracted driving, and comprehensive driver education programs emphasizing VRU safety.
4. Proactive Intervention in High-Risk Zones: Utilize geographical collision density and historical accident data to identify specific hotspots for proactive intervention. This involves deploying additional police patrols, launching public safety campaigns, and conducting detailed traffic engineering studies at frequently occurring collision sites to identify and fix infrastructure deficiencies.
5. Predictive Modeling: Develop predictive models to anticipate collision hotspots and support proactive safety measures. These models can utilize temporal, spatial, and environmental data to forecast collision risks in real-time, enabling authorities to allocate resources more efficiently and intervene before incidents occur.

6. Visualization Integration using Tableau/ PowerBI

https://public.tableau.com/shared/3W8SQZBH6?:display_count=n&:origin=viz_share_link