```
CREATE DATABASE crash_analysis;
USE crash_analysis;

SELECT * FROM news_crashes LIMIT 10;

SHOW COLUMNS FROM news_crashes;


--  ===================  DATA PREPROCESSING  ===================

# 1.Replacing NA with null
UPDATE news_crashes
SET
  age = NULLIF(TRIM(age), 'NA'),
  gender = NULLIF(TRIM(gender), 'NA'),
  `Million Plus City` = NULLIF(TRIM(`Million Plus City`), 'NA'),
  killed = NULLIF(TRIM(killed), 'NA'),
  injured = NULLIF(TRIM(injured), 'NA');


# 2.remove duplicates
#s.no is unique so no duplicates
SELECT COUNT(*) FROM news_crashes;
SELECT COUNT(DISTINCT `S. No.`) FROM news_crashes;

#3.Split Latlong into Latitude and Longitude
ALTER TABLE news_crashes
ADD COLUMN latitude DECIMAL(10, 8),
ADD COLUMN longitude DECIMAL(11, 8);

UPDATE news_crashes
SET
  latitude = TRIM(SUBSTRING_INDEX(Latlong, ',', 1)),
  longitude = TRIM(SUBSTRING_INDEX(Latlong, ',', -1));

ALTER TABLE news_crashes
DROP COLUMN latitude,
DROP COLUMN longitude;

ALTER TABLE news_crashes
ADD COLUMN latitude DECIMAL(12, 9),
ADD COLUMN longitude DECIMAL(12, 9);

UPDATE news_crashes
SET
  latitude = TRIM(SUBSTRING_INDEX(Latlong, ',', 1)),
  longitude = TRIM(SUBSTRING_INDEX(Latlong, ',', -1));

SELECT Latlong, latitude, longitude
FROM news_crashes
LIMIT 10;

#4.Converting Crash Date and Article Date to DATE format

ALTER TABLE news_crashes
ADD COLUMN crash_date_parsed DATE,
ADD COLUMN article_date_parsed DATE;
```

```
UPDATE news_crashes
SET
  `Crash Date` = REPLACE(`Crash Date`, 'Sept', 'Sep'),
  `Article Date` = REPLACE(`Article Date`, 'Sept', 'Sep');


UPDATE news_crashes
SET
  crash_date_parsed = CASE
    WHEN `Crash Date` LIKE '%-%' THEN STR_TO_DATE(`Crash Date`, '%d-%b-
%y')
    ELSE STR_TO_DATE(`Crash Date`, '%e %b %y')
  END,
  article_date_parsed = CASE
    WHEN `Article Date` LIKE '%-%' THEN STR_TO_DATE(`Article Date`,
'%d-%b-%y')
    ELSE STR_TO_DATE(`Article Date`, '%e %b %y')
  END;



SELECT
  `Crash Date`, crash_date_parsed,
  `Article Date`, article_date_parsed
FROM news_crashes
LIMIT 10;

#5: Handle Missing or Null Values & Data Cleaning

SELECT
  COUNT(*) AS total_rows,
  SUM(CASE WHEN killed IS NULL THEN 1 ELSE 0 END) AS killed_null_count,
  SUM(CASE WHEN injured IS NULL THEN 1 ELSE 0 END) AS
injured_null_count,
  SUM(CASE WHEN gender IS NULL OR gender = '' THEN 1 ELSE 0 END) AS
gender_null_count,
  SUM(CASE WHEN 'crash type' IS NULL OR 'crash type' = '' THEN 1 ELSE 0
END) AS crash_type_null_count
FROM news_crashes;

UPDATE news_crashes
SET gender = 'Unknown'
WHERE gender IS NULL OR gender = '';

#6: Remove duplicates if any (based on your unique ID, say S_No)

DELETE t1 FROM news_crashes t1
INNER JOIN news_crashes t2
WHERE
  't1.S. No.' > 't2.S. No.'
  AND 't1.S. No.' = 't2.S. No.';


CREATE or replace VIEW cleaned_crash_data AS
SELECT *
FROM news_crashes
WHERE
    crash_date_parsed IS NOT NULL
    AND YEAR( crash_date_parsed) IN (2022, 2023);




--          =================== KEY PERFORMANCE INDICATORS
===================

-- 1. TOTAL CRASHES REPORTED
SELECT COUNT(*) AS total_crashes FROM cleaned_crash_data;
/****
-- =================== INSIGHT ===================
   A total of 2,888 crashes were reported in the dataset, highlighting
the scale of road safety incidents recorded.
-- ================================================
****/


-- 2. TOTAL FATALITIES AND INJURIES
SELECT
  SUM(killed) AS total_killed,
  SUM(injured) AS total_injured
FROM cleaned_Crash_Data;

/****
-- =================== INSIGHT ===================
-- â□¢ 6,557 people lost their lives
-- â□¢ 7,779 people were injured.
This reinforces the severity of accidents and the need for improved
traffic safety policies.
-- ================================================
****/


-- 3.AVERAGE FATALITIES AND INJURIES PER CRASH
SELECT
  AVG(killed) AS avg_killed_per_crash,
  AVG(injured) AS avg_injured_per_crash
FROM cleaned_crash_data;

/****
-- =================== INSIGHT ===================
  Each crash resulted in an average of 2.27 fatalities and 2.69
injuries, which is significantly high.
-- ================================================
****/


-- 4.AVERAGE REPORTING DELAY
SELECT
    round(AVG(DATEDIFF(article_date_parsed, crash_date_parsed)),2) AS
avg_reporting_delay
FROM cleaned_crash_data;

/****
-- =================== INSIGHT ===================
-- The average reporting delay between the occurrence of a crash and
the publication of the article is 1.32 days.
-- This indicates that news outlets are generally quick to report
accidents, with a minimal delay in coverage.
-- ===================================================
****/





-- =================== TEMPORAL INSIGHTS((When do crashes happen?)
===================

-- 1.CRASH METRICS BY DAY OF THE WEEK

SELECT
    `Crash Day` AS day_of_week,
    YEAR(`crash_date_parsed`) AS crash_year,
    COUNT(*) AS crash_count,
      SUM(killed) AS total_killed,
    SUM(injured) AS total_injured,
    ROUND(sum(killed) / count(*), 2) AS avg_fatalities_per_crash,
    ROUND(sum(injured) / count(*), 2) AS avg_injuries_per_crash
FROM cleaned_crash_data
GROUP BY `Crash Day`,crash_year
ORDER BY count(*) desc;

/****
-- =================== INSIGHT ===================

-- **Sunday consistently has the highest crash volume**, with 261
crashes in 2022 and 245 in 2023 â□□ suggesting heightened risk due to
weekend travel and leisure activity.
-- **Fatalities per crash dropped** on Sundays from 2.33 (2022) to 1.98
(2023), but **injuries per crash rose** from 2.38 to 2.78 â□□
suggesting better survival but persistent severity.
-- **Wednesday in both years shows extreme severity**: highest avg.
fatalities per crash in 2023 (2.64) and injuries (3.45 in 2022, 3.22 in
2023) â□□ despite not being the highest crash volume.
-- **Friday and Tuesday crashes in 2023 show rising severity**, with
increasing injuries per crash (Friday: 3.16, Tuesday: 2.7) â□□ possibly
due to end-of-week fatigue or commute rush.
-- **Thursday has the lowest injuries per crash both years**
(~2.13â□□2.14), indicating possibly less traffic or more urban, lower-
speed crashes.
-- **Monday shows improvement in 2023**: crashes and injuries reduced
slightly, indicating possible enforcement or behavior changes at the
week's start.

-- ===================================================

-- ================= RECOMMENDATION ==============

-- **Prioritize enforcement and awareness on high-risk days**: Sunday,
Wednesday, Friday, and Tuesday â□□ focus on speed control, fatigue
management, and alcohol checks.
```

```
-- **Deploy increased patrol and response teams** on Sundays and
Wednesdays â□□ especially on highways and intercity roads where
severity is highest.
-- **Educate weekend and mid-week drivers** (e.g., long-distance
travelers, truckers) about fatigue, overtaking risks, and defensive
driving.
-- **Use targeted social campaigns** by day: e.g., â□□Safe Sunday
Driveâ□□ for family drivers, â□□Watchful Wednesdayâ□□ for transport
workers.
-- **Pre-position emergency services** for high-fatality days to reduce
time-to-treatment and increase survival.
-- Consider deeper breakdowns: crash type, road type, and vehicle
involved by weekday â□□ to tailor interventions more precisely (e.g.,
are head-on collisions spiking on Wednesdays?).

-- ================================================
****/



-- 2.CRASH METRICS BY MONTH

SELECT
      YEAR(`crash_date_parsed`) AS crash_year,
    MONTH as crash_month,
    COUNT(*) AS crash_count,
      SUM(killed) AS total_killed,
    SUM(injured) AS total_injured,
    ROUND(sum(killed) / count(*), 2) AS avg_fatalities_per_crash,
    ROUND(sum(injured) / count(*), 2) AS avg_injuries_per_crash
FROM cleaned_crash_data
GROUP BY crash_year,crash_month
ORDER BY count(*) desc;

/****
-- =================== INSIGHT ===================

-- ð□□¥ **May is consistently the most dangerous month**:
   -  2022: 143 crashes, 402 deaths (highest fatalities per crash at
      2.81)
   -  2023: 157 crashes, 410 deaths â□□ highest overall fatalities and
      crash volume
-- ð□□□ **Aprilâ□□June 2023 also shows rising severity** (avg
fatalities per crash > 2.4) â□□ likely tied to pre-monsoon long-
distance travel or road conditions.
-- ð□§¨ **November 2023 has the highest fatalities per crash (3.3)**
despite the lowest crash volume â□□ indicating rare but **extremely
deadly crashes**, possibly fog-related or involving heavy vehicles.
-- â□□ï¸□ **Summer months (Aprilâ□□June)** across both years
consistently show high injuries per crash (3+), suggesting high-impact,
high-speed crashes.
-- â□□ï¸□ **Winter months (December, January)** show variable results:
   -  2022 December: High injuries per crash (3.17)
   -  2023 December: Lower volume (63 crashes) but still high injuries
      (3.48) â□□ possibly due to fog, visibility issues, or holiday
      traffic.
```

```
-- ð□□□ **August & September 2023 show low severity**: Avg. injuries
per crash drop to ~1.7â□□2.0 â□□ possibly reflecting safer driving
conditions or more minor crashes.

-- ================================================

-- ================= RECOMMENDATION ==============

-- ð□□□ **Prioritize road safety interventions from April to June**,
particularly around long weekends, school holidays, and heat-related
fatigue risks.
-- ð□□£ï¸□ Focus on **highway safety and long-haul routes** in **May**,
as it consistently leads in fatalities â□□ deploy mobile enforcement,
speed cameras, and rest stop awareness.
-- ð□§□ **Investigate November spike in severity** (3.3 deaths/crash)
â□□ assess visibility issues (fog), road quality, or high-impact crash
types (e.g., multi-vehicle, truck-related).
-- ð□□§ï¸□ Prepare for high injury crashes during pre-monsoon months
(June) â□□ improve ambulance response and roadside trauma readiness.
-- ð□§  Launch **seasonal campaigns**:
   -  Summer: "Stay Sharp, Drive Cool"
   -  Winter: "Slow Down in the Fog"
   -  Holidays: â□□Drive to Arrive â□□ Not Just to Travelâ□□
-- ð□□□ Suggest deeper cross-analysis by:
   -  Crash type (e.g., head-on, hit-and-run)
   -  Time of day (e.g., night visibility in winter)
   -  Vehicle type (e.g., trucks in Novemberâ□□December)

-- ================================================
****/



-- 5. Time taken to report incident
SELECT
    CASE
        WHEN DATEDIFF(article_date_parsed, crash_date_parsed) = 0 THEN
'Same Day'
        WHEN DATEDIFF(article_date_parsed, crash_date_parsed) = 1 THEN
'Next Day'
        WHEN DATEDIFF(article_date_parsed, crash_date_parsed) BETWEEN 2
AND 7 THEN '2-7 Days'
        ELSE '1+ Week'
    END AS reporting_delay_bucket,
    YEAR(`crash_date_parsed`) AS crash_year,
    COUNT(*) AS crash_count,
    SUM(killed) AS total_killed,
    SUM(injured) AS total_injured,
    ROUND(sum(killed) / count(*), 2) AS avg_fatalities_per_crash,
    ROUND(sum(injured) / count(*), 2) AS avg_injuries_per_crash
FROM cleaned_crash_data
GROUP BY reporting_delay_bucket,crash_year
ORDER BY crash_count DESC;

/****
-- ================== INSIGHT ==================
-- â□ ï¸□ **Crashes reported on the same day are the deadliest**, with
extremely high average fatalities and injuries per crash:
   -  2022: 3.89 fatalities / 4.93 injuries per crash
   -  2023: 3.84 fatalities / **5.87 injuries per crash** (highest in
      entire dataset)
-- ð□□° **Same-day reporting likely reflects high-impact, high-
visibility crashes** â□□ e.g., multi-vehicle pileups, celebrity
victims, or mass casualties.
-- ð□□□ **Next-day reporting covers most crashes** (2022: 713 | 2023:
671), with moderate severity (2.2â□□2.23 fatalities/crash) â□□
indicating general road crash coverage follows a 24-hour media cycle.
-- â□□ **Crashes reported with delays of 2â□□7 days show much lower
severity**:
   -  Avg. fatalities: ~1.6â□□1.8 per crash
   -  Avg. injuries: ~1.2â□□1.3 per crash
   -  These may reflect smaller, localized incidents or delayed rural
      reporting.
-- ð□□□ **Very late reports (1+ week)** are extremely rare and low-
severity â□□ likely due to underreporting or data lag in remote areas.

-- ================================================

-- ================= RECOMMENDATION ==============

-- ð□§  **Use reporting delay as a proxy for crash impact and
newsworthiness**:
   -  Crashes reported on the **same day** should be prioritized for
      detailed forensic review or media scrutiny.
   -  These also represent public concern events â□□ useful for
      policymaker attention.
-- ð□□¢ **Encourage timely and standardized crash reporting**,
especially for lower-severity but frequent crashes â□□ to reduce bias
toward only major events.
-- ð□□□ Use this dimension in **dashboard filters** â□□ allow users to
view only high-impact (same-day) crashes or delayed reports for trend
analysis.
-- ð□□□ Recommend further exploration:
   -  Is **reporting delay** correlated with geography (urban vs
      rural)?
   -  Does **vehicle type** or **crash type** (e.g., hit-and-run)
      affect reporting delay?
   -  Are **certain states or media sources** consistently delayed?

-- â□□ Combine this with emergency response data (if available) to
understand how **media speed aligns (or doesn't) with real-time
response**.

-- ================================================
****/





-- =================== DEMOGRAPHIC ANALYSIS ===================

-- 1. Which Demographics Face the Highest Fatality Risk?
```

```
SELECT
    CASE
        WHEN CAST(SUBSTRING_INDEX(Age, ',', 1) AS UNSIGNED) BETWEEN 0
AND 17 THEN '0-17'
        WHEN CAST(SUBSTRING_INDEX(Age, ',', 1) AS UNSIGNED) BETWEEN 18
AND 30 THEN '18-30'
        WHEN CAST(SUBSTRING_INDEX(Age, ',', 1) AS UNSIGNED) BETWEEN 31
AND 45 THEN '31-45'
        WHEN CAST(SUBSTRING_INDEX(Age, ',', 1) AS UNSIGNED) BETWEEN 46
AND 60 THEN '46-60'
        WHEN CAST(SUBSTRING_INDEX(Age, ',', 1) AS UNSIGNED) > 60 THEN
'60+'
        ELSE 'Unknown'
    END AS age_group,
    Gender,
    COUNT(*) AS crash_count,
    SUM(Killed) AS total_killed,
    SUM(Injured) AS total_injured
FROM cleaned_crash_data
WHERE Age IS NOT NULL
  AND Age <> ''
  AND Gender IS NOT NULL
  AND Gender <> ''
GROUP BY age_group, Gender
ORDER BY age_group, crash_count DESC;

/****
-- ================== INSIGHT ==================
-- Young adults (18â□□30) are the most affected group, with male
fatalities (1008)
-- far exceeding all other categories, indicating higher risk-taking
behaviors.
-- Across all age groups, males consistently show higher crash counts
and fatalities than females.
-- "Both" gender cases (family/group travel) in the 31â□□45 and 46â□□60
brackets report high fatalities,
-- suggesting severe crashes involving multiple family members.
-- Children (0â□□17) and elderly (60+) have lower crash counts, but
still show vulnerability
-- due to high casualties relative to exposure (fragility, lack of
protection).
-- ================================================

-- ================ RECOMMENDATION ==============
-- Launch targeted road safety campaigns focusing on **18â□□30 males**
addressing
-- speeding, drink-driving, and night-time travel.
-- Strengthen **child protection measures** like school-zone safety,
helmet use, and child seats.
-- For family/group travel, promote **seatbelt discipline, safe long-
distance driving practices**,
-- and vehicle loading awareness.
-- For elderly travelers, consider awareness on **safe seating, regular
health/vision checks**,
-- and policy support for senior driver licensing.
-- Overall, design **demographic-specific interventions** instead of
one-size-fits-all campaigns.
-- ================================================
****/



-- 2.
SELECT
    `Crash Type`,
    CASE
        WHEN CAST(SUBSTRING_INDEX(Age, ',', 1) AS UNSIGNED) < 18 THEN
'<18'
        WHEN CAST(SUBSTRING_INDEX(Age, ',', 1) AS UNSIGNED) BETWEEN 18
AND 30 THEN '18-30'
        WHEN CAST(SUBSTRING_INDEX(Age, ',', 1) AS UNSIGNED) BETWEEN 31
AND 45 THEN '31-45'
        WHEN CAST(SUBSTRING_INDEX(Age, ',', 1) AS UNSIGNED) BETWEEN 46
AND 60 THEN '46-60'
        ELSE '60+'
    END AS age_group,
    COUNT(*) AS crash_count,
    SUM(killed) AS total_killed,
    SUM(injured) AS total_injured,
    ROUND(SUM(killed) / COUNT(*), 2) AS avg_fatalities_per_crash,
    ROUND(SUM(injured) / COUNT(*), 2) AS avg_injuries_per_crash
FROM cleaned_crash_data
GROUP BY
    `Crash Type`,
    age_group
ORDER BY
    `Crash Type`
   ;

/****
-- =================== INSIGHT ===================
-- Seniors (60+) face the highest fatality and injury rates across most
crash types,
-- especially in **Vehicle Overturn**, **Run Off Road**, and **Head On
Collisions**.
-- Teens (<18) show extreme injury risk in **Vehicle Overturn** and
**Run Off Road** crashes,
-- indicating high severity despite lower crash counts.
-- Young adults (18â□□30) consistently lead in crash frequency,
particularly in
-- **Fixed Object**, **Head On**, and **Hit and Run** categories.
-- Crashes involving **Fixed Objects** and **Parked Vehicles** are more
fatal with age,
-- suggesting vulnerability due to slower reaction times or reduced
awareness.
-- **Hit and Run** and **Rear-End Collisions** are frequent but less
fatal,
-- often involving younger and middle-aged drivers in urban settings.
-- ================================================

-- ================= RECOMMENDATION ==============
-- Target **senior drivers** with regular health checks, license
renewal reviews, and
-- safer road design in high-senior zones.
-- Strengthen **teen driver training** with focus on control, hazard
perception,
-- and crash avoidance for high-risk scenarios.
-- Launch behavior-focused campaigns for **18â□□30 drivers** addressing
speed, distraction,
-- and nighttime risks.
-- Improve infrastructure (guardrails, visibility) to reduce **run-off-
road**
-- and **fixed object** crashes.
-- Use surveillance and penalties to combat **hit and run** incidents,
-- and promote rear-end collision prevention tech in dense traffic
areas.
-- Favor **crash-type and age-specific interventions** over generic
safety messaging.
-- ==================================================
****/




-- =================  GEOGRAPHICAL INSIGHTS  ===================

-- 1. crash metrics by state
SELECT
    State,
    COUNT(*) AS total_crashes,
    YEAR(`crash_date_parsed`) AS crash_year,
   SUM(killed) AS total_killed,
    SUM(injured) AS total_injured,
    ROUND(sum(killed) / count(*), 2) AS avg_fatalities_per_crash,
    ROUND(sum(injured) / count(*), 2) AS avg_injuries_per_crash
FROM cleaned_crash_data
GROUP BY State,crash_year
ORDER BY total_crashes DESC;

/****
-- =================== INSIGHT ===================
-- **Uttar Pradesh** emerges as the deadliest state in both years:
--    â□¢ 2022: 486 deaths, 694 injuries, avg 3.12 deaths/crash
--    â□¢ 2023: 460 deaths, 602 injuries, avg 2.74 deaths/crash
--    It not only has high crash volume but also the **highest average
fatalities per crash** among major states.

-- **Jammu** is the deadliest state *per crash*:
--    â□¢ 2023: Only 18 crashes caused 106 deaths and 175 injuries â□□
avg 5.89 deaths & 9.72 injuries per crash
--    â□¢ 2022: 22 crashes caused 98 deaths and 238 injuries â□□ avg
4.45 deaths & 10.82 injuries per crash
--    These figures suggest extremely severe crash outcomes in the
region.

-- **Manipur, Sikkim, Chhattisgarh, and Odisha** also show extreme
severity per crash despite low crash counts, indicating that even
single incidents in these areas tend to be highly fatal.

-- **Maharashtra, Madhya Pradesh, and Andhra Pradesh** show
consistently high crash volumes and total casualties, but lower average
fatality rates compared to Uttar Pradesh and Jammu.
```

```sql
-- **Himachal Pradesh** and **Odisha** show alarmingly high injury
rates per crash (avg > 5 injuries per crash), which could reflect
dangerous road conditions or delayed emergency response.

-- **Kerala, Delhi, and West Bengal** show relatively lower fatality
and injury rates, possibly reflecting better infrastructure or
reporting patterns.

-- **Most northeastern states (Mizoram, Nagaland, Meghalaya)** have
very few reported crashes but often high severity when crashes do occur
â□□ indicating possible under-reporting or data sparsity.

-- ================================================

-- ================= RECOMMENDATION ==============
-- 1. **Prioritize safety audits in Uttar Pradesh and Jammu** â□□ their
high average fatalities per crash demand urgent infrastructure
upgrades, better enforcement, and medical response.

-- 2. **Focus on High-Severity, Low-Frequency States** (like Manipur,
Sikkim, Mizoram):
--     Even with low crash counts, the severity suggests lack of road
safety systems and delayed response.

-- 3. **Deploy Targeted Interventions in Hilly or Hazard-Prone
States**:
--     States like Himachal Pradesh and Uttarakhand show high severity,
likely due to terrain. Install crash barriers, warning signage, and
improve visibility on curves.

-- 4. **Enhance Trauma Care and Emergency Response**:
--     States with high injuries per crash (e.g., Odisha, Jammu,
Himachal Pradesh) require faster accident response, better first aid
access, and trauma centers.

-- 5. **Use Tableauâ□□s Year Filter with Dual-Axis Map** to show change
in crash metrics per state between 2022 and 2023, highlighting areas
where severity increased or dropped.

-- 6. **Enable Policy Makers to Use Visuals for Budget
Prioritization**:
--     States like UP, Jammu, MP, and Odisha should receive urgent
resource allocation for crash prevention.

-- ================================================
****/


-- 2.High rish crash locations by severity index

SELECT
    Location,
    round(SUM(killed)* 3 + sum(injured) / COUNT(*),1) AS
crash_severity_score
FROM
   cleaned_crash_data
GROUP BY
    Location
```

```
ORDER BY
    crash_severity_score DESC
LIMIT 20;

/****
-- =================== INSIGHT ===================
-- The deadliest location is **Trungal**, with a crash severity score
of **134**, likely indicating a crash with many fatalities and
injuries.
-- Other severe hotspots include **Timari**, **Kotli Jhajjar**, and
**Sohagi Pahad**, each exceeding a score of 80 â□□ suggesting extremely
high-impact crashes.
-- Many locations like **Pimpalkhuta**, **Damta**, and **Desuri** also
show similarly high severity despite fewer total crashes â□□ signaling
that even isolated incidents have deadly outcomes.
-- Urban peripheries and isolated road segments (e.g., **Tirumala Ghat
Rd**, **Narmada Bridge**, **Sohagi Pahad**) appear repeatedly, possibly
due to speed, poor visibility, or lack of barriers.
-- Locations like **Hotel Mirchi** and **Manapparai** hint at crash-
prone zones near commercial establishments or small towns, where
pedestrian/vehicle mixing may be common.
-- This analysis reveals that **crash count alone underrepresents
danger** â□□ severity score gives better prioritization for targeted
intervention.
-- ================================================

-- ================= RECOMMENDATION ==============
-- Conduct **site-specific audits** at the top 20 high-severity
locations to assess infrastructure gaps (sharp curves, signage,
lighting, roadside hazards).
-- Install **warning signage, crash barriers, and speed calming
measures** at isolated or curved road segments.
-- Improve **surveillance** and **response time** at remote and hilly
areas (like **Tirumala Ghat Rd**, **Sohagi Pahad**, **Longsai**) where
fatal crashes are often underreported or delayed.
-- Deploy **geofencing-based alerts**, especially for accident-prone
stretches on highways.
-- Engage local communities to identify behavioral or environmental
issues contributing to repeated crashes at these points.
-- ================================================
****/


-- 3.Crash Trends in Urban vs Rural Areas
SELECT
    CASE
        WHEN `Million Plus City` IS NOT NULL
             AND `Million Plus City` <> ''
             AND `Million Plus City` <> 'Nil'
        THEN 'Urban'
        ELSE 'Rural'
    END AS area_type,
    YEAR(`crash_date_parsed`) AS crash_year,
    COUNT(*) AS crash_count,
    SUM(killed) AS total_killed,
    SUM(injured) AS total_injured,
    ROUND(SUM(killed)/COUNT(*),2) AS avg_fatalities_per_crash,
    ROUND(SUM(injured)/COUNT(*),2) AS avg_injuries_per_crash
```

```
FROM cleaned_crash_data
GROUP BY crash_year,area_type;


/****
-- =================== INSIGHT ===================
-- Crashes in **rural areas** are significantly more **severe** than
those in urban areas across both 2022 and 2023.
-- In 2023, rural areas saw **944 crashes** with **2,506 deaths**,
averaging **2.65 fatalities per crash** â□□ nearly **double the urban
average**.
-- Injury rates also follow a similar pattern â□□ rural crashes average
**3.32 injuries per crash**, compared to just **1.28** in urban areas.
-- Despite having **fewer crashes**, rural zones contribute a
**disproportionately high share** of both fatalities and injuries.
-- The consistency across both years highlights a persistent **urban-
rural disparity** in crash outcomes.

-- ================================================

-- ================= RECOMMENDATION ==============
-- Prioritize **emergency response improvements** in rural areas â□□
better ambulance access, trauma centers, and first responder coverage.
-- Invest in **rural road safety infrastructure** â□□ signage,
lighting, median barriers, and speed control measures.
-- Implement **awareness campaigns** tailored to rural drivers focusing
on safe overtaking, speed limits, and night driving.
-- Explore why crashes are deadlier in rural zones â□□ higher speeds,
lack of enforcement, or poor road conditions â□□ and address root
causes.
-- In dashboards, **highlight rural severity** to inform resource
allocation and policymaking.
-- ================================================
****/


-- 4.
SELECT
    `Million Plus City`,
     #YEAR(`crash_date_parsed`) AS crash_year,
 #round(SUM(killed)* 3 + sum(injured) / COUNT(*),1) AS severity_score
round ((3 * SUM(killed) + SUM(injured) * 1.0) / count(*), 1) as
severity_score

FROM
    cleaned_crash_data
WHERE
     `Million Plus City` <> 'nil'
GROUP BY
      `Million Plus City`
ORDER BY
    severity_score DESC
LIMIT 10;




-- ==================== CRASH CHARACTERISTICS ======================
-- 1.Crash metrics by crash type

SELECT
   `Crash Type`,
    COUNT(*) AS total_crashes,
    YEAR(`crash_date_parsed`) AS crash_year,
   SUM(killed) AS total_killed,
    SUM(injured) AS total_injured,
    ROUND(sum(killed) / count(*), 2) AS avg_fatalities_per_crash,
    ROUND(sum(injured) / count(*), 2) AS avg_injuries_per_crash
FROM cleaned_crash_data
GROUP BY `Crash Type`, crash_year
ORDER BY total_crashes DESC;

/****
-- =================== INSIGHT ===================

-- **Head-On Collisions** remain the deadliest crash type in both
years:
   -  2023: 342 crashes â□□ 912 deaths, 992 injuries (avg: 2.67
      fatalities/crash)
   -  2022: 294 crashes â□□ 805 deaths, 927 injuries (avg: 2.74
      fatalities/crash)
   -  These contribute the highest total fatalities across all types.

-- **Vehicle Overturns** show the highest **severity per crash**:
   -  2022: 3.83 fatalities & 5.71 injuries per crash
   -  2023: 3.16 fatalities & 5.86 injuries per crash
   -  Indicates extremely high risk when overturning occurs.

-- **Run Off Road** crashes, although fewer in number, result in very
high **injury averages**:
   -  2023: 6.48 injuries/crash
   -  2022: 5.71 injuries/crash

-- **Hit and Run** crashes are very frequent:
   -  Combined over 560 incidents across 2 years
   -  Despite slightly lower severity, they indicate a systemic issue
      with accountability.

-- **Fixed Object Collisions** and **With Parked Vehicles** have
increasing severity in 2023 â□□ suggesting potential visibility or
signage issues.

-- **Hit from Side** is the least frequent and least fatal, but showed
increased severity in 2023 (2.24 fatalities vs 1.94 in 2022).

-- ================================================


-- ================= RECOMMENDATION ===============

-- Strengthen road design and signage to reduce **Head-On Collisions**:
   -  Install median barriers, especially on highways and undivided
      roads.
   -  Enforce speed limits and overtaking rules.
```

```
-- Target high-severity crash types (**Vehicle Overturn**, **Run Off
Road**) with:
   -  Better edge barriers
   -  Warning signs on curves, slopes, and poor-condition roads
   -  Public campaigns on safe driving in hilly or rural areas.

-- Address **Hit and Run** through:
   -  Enhanced surveillance (CCTV, dashcams)
   -  Strict legal enforcement
   -  Encouragement of bystander reporting via hotlines or mobile apps

-- Investigate increase in severity of **With Parked Vehicle** and
**Fixed Object Collisions**:
   -  Improve night-time visibility
   -  Restrict dangerous parking practices
   -  Add reflective barriers or bollards

-- Use these crash type insights to drive **policy**, **road
engineering**, and **targeted driver education** campaigns.

-- ==================================================
****/



-- 2. Top Vehicle Combinations in Crashes
SELECT
    `Vehicle 1`,
    `Vehicle/Object 2`,
    CASE
        WHEN `Vehicle/Object 2` = 'Nil' OR `Vehicle/Object 2` IS NULL
THEN 'Single Vehicle Crash'
        ELSE 'Multi Vehicle Crash'
    END AS crash_type,
    COUNT(*) AS crash_count,
    SUM(killed) AS total_killed,
    SUM(injured) AS total_injured
FROM cleaned_crash_data
GROUP BY `Vehicle 1`, `Vehicle/Object 2`, crash_type
ORDER BY crash_count DESC
LIMIT 10;


/****
-- =================== INSIGHT ===================
-- The most dangerous vehicle combination is **Car vs Truck**, with 652
fatalities in 210 crashes â□□ averaging over **3.1 deaths per crash**,
the highest among all pairs.
-- **Bus vs Truck** crashes, although fewer in count (66), caused **268
deaths and 1,027 injuries**, indicating extremely high injury severity
â□□ averaging **15.6 injuries per crash**.
-- **Two Wheeler vs Truck** crashes are the most frequent (342 crashes)
and deadly (542 deaths), revealing extreme vulnerability of two-
wheelers when hit by heavy vehicles.
-- **Pedestrian-involved crashes** (with cars or trucks) also show
alarmingly high fatality and injury counts, especially given that
pedestrians are non-vehicular participants.
```

```
-- **Two Wheeler vs Bus and vs Car** also appear frequently,
reinforcing that two-wheelers are disproportionately at risk across
multiple crash scenarios.
-- Crashes involving **Unidentified** or **Nil (single-vehicle)**
objects show lower injuries/fatalities, but may reflect underreporting
or missing data quality.
-- ==================================================

-- ================= RECOMMENDATION ==============
-- Strictly enforce lane discipline and speed limits for heavy vehicles
(trucks, buses), especially in areas with high two-wheeler and
pedestrian traffic.
-- Mandate **protective infrastructure** like pedestrian overpasses,
crossing signals, and barriers in urban areas.
-- Promote **awareness and training for two-wheeler riders** about
blind spots, overtaking, and safe distances from trucks and buses.
-- Install **vehicle surveillance and crash detection systems** in
high-risk routes to monitor and analyze real-time vehicle interactions.
-- Improve **helmet and seatbelt enforcement**, especially for
vulnerable users like two-wheeler riders and pedestrians.
-- Consider **vehicle restriction zones** in pedestrian-heavy or mixed-
traffic areas to reduce high-severity combo crashes.
-- ==================================================
****/


-- 3.Crash metrics by road type

SELECT
`Road Type`,
    COUNT(*) AS total_crashes,
    YEAR(`crash_date_parsed`) AS crash_year,
   SUM(killed) AS total_killed,
    SUM(injured) AS total_injured,
    ROUND(sum(killed) / count(*), 2) AS avg_fatalities_per_crash,
    ROUND(sum(injured) / count(*), 2) AS avg_injuries_per_crash
FROM cleaned_crash_data
GROUP BY `Road Type`,crash_year
ORDER BY total_crashes DESC;

/****
-- =================== INSIGHT ===================
-- **National Highways (NH)**, despite having fewer crashes than Other
Roads (OR), show **higher average fatalities and injuries per crash**
â□□ reaching **2.54 fatalities and 3.18 injuries per crash in 2023**,
the highest across all road types.
-- **State Highways (SH)** also record significant severity, with
**2.66 fatalities per crash in 2023**, suggesting they are not far
behind NHs in terms of crash impact.
-- **Other Roads (OR)** have the highest number of crashes in both 2022
and 2023, but **lower severity averages** (around 2.04â□□2.09
fatalities/crash and 2.27â□□2.45 injuries/crash), possibly due to lower
speeds or urban congestion.
-- The overall trend indicates that **NH and SH crashes are more
lethal**, while ORs contribute to a **higher volume of incidents**,
albeit with slightly lower severity.
-- The increase in average injuries on OR from 2022 to 2023 (2.27 â□□
2.45) may signal deteriorating road safety or response systems.
```

```
-- ================================================

-- ================= RECOMMENDATION ==============
-- **Prioritize enforcement and safety audits** on **National and State
Highways**, especially focusing on high-speed zones and known
blackspots.
-- Implement **traffic calming measures** (e.g., rumble strips,
signage, speed monitoring) on NH and SH segments prone to severe
crashes.
-- Improve **emergency response infrastructure** (e.g., ambulances,
first responders) on highways to reduce fatality rates post-crash.
-- On **Other Roads**, address volume-based risk by enhancing **road
signage, pedestrian crossings, and lane discipline enforcement**.
-- Integrate **AI-based surveillance and predictive crash analytics**
on NH and SH to detect high-risk patterns and intervene preemptively.
-- Encourage **public awareness campaigns** highlighting the higher
fatality risk of highway travel vs. city or rural roads.
-- ================================================
****/



-- 4.

SELECT
    "Vehicle 1",
    "Vehicle/Object 2",
    "Road Type",
    COUNT(*) AS crash_count,
    SUM(killed) AS total_killed,
    SUM(injured) AS total_injured,
    ROUND(SUM(killed) * 3 + SUM(injured), 2) AS crash_severity_score,
    ROUND(SUM(killed) * 1.0 / COUNT(*), 2) AS avg_fatalities_per_crash,
    ROUND(SUM(injured) * 1.0 / COUNT(*), 2) AS avg_injuries_per_crash
FROM crash_data_table
GROUP BY
    "Vehicle 1",
    "Vehicle/Object 2",
    "Road Type"
ORDER BY
    crash_severity_score DESC
LIMIT 10;

SELECT
        CASE
            WHEN `Vehicle/Object 2` IS NULL OR `Vehicle/Object 2` =
'Nil'
            THEN `Vehicle 1`
            ELSE CONCAT(`Vehicle 1`, ' + ', `Vehicle/Object 2`)
        END AS vehicle_combo,
        COUNT(*) AS crash_count,
        SUM(killed) AS total_killed,
        SUM(injured) AS total_injured,
        ROUND((SUM(killed) + SUM(injured))/COUNT(*),2) AS
avg_casualties_per_crash
    FROM cleaned_crash_data
    GROUP BY vehicle_combo
    order by total_killed desc limit 15;
```

```
/****
-- =================== INSIGHT ===================
-- The **Bus + Gorge** combination is the deadliest in terms of average
casualties per crash, with **30.5 casualties per incident** â□□ this
points to catastrophic crash events, likely involving rollovers or
plunges.
-- **Bus + Truck** crashes, although fewer (66), resulted in **1,027
injuries and 268 deaths**, with an average of **19.6 casualties per
crash**, indicating severe mass-casualty incidents.
-- **SUV + Truck** and **Auto Rickshaw + Truck** combinations have
extremely high average casualties per crash (over 9), suggesting that
smaller or lighter vehicles fare poorly in collisions with heavy
trucks.
-- The **Car + Truck** combination is the most lethal by absolute
fatalities (652 deaths), with an average of **4.79 casualties per
crash**, indicating a common and deadly crash pattern.
-- Two-wheeler combinations (e.g., **Two Wheeler + Truck**, **Two
Wheeler + Bus**, **Two Wheeler + Car**) occur frequently and show
consistently high fatality and injury counts, highlighting the
vulnerability of two-wheeler riders.
-- **Pedestrian-related** combinations, particularly with trucks and
cars, show alarming casualty numbers, underscoring the need for better
pedestrian safety measures.
-- ================================================

-- ================= RECOMMENDATION ==============
-- Implement **speed restrictions and overtaking bans** for heavy
vehicles like trucks and buses on hilly terrain and in urban areas.
-- Introduce **physical barriers, guardrails, and proper lighting** on
roads with gorge-side routes to reduce catastrophic crashes like Bus +
Gorge or Car + Gorge.
-- Enforce **vehicle fitness checks and safety audits** for buses and
trucks to ensure safe mechanical condition.
-- Deploy **intelligent traffic management systems** to monitor heavy
vehicle movement near pedestrian zones and vulnerable road user areas.
-- Promote **protective infrastructure** (e.g., motorcycle lanes,
pedestrian crossings, footbridges) and awareness campaigns targeting
two-wheeler riders and pedestrians.
-- Increase **visibility and enforcement** at known high-risk junctions
involving common combos like Car + Truck or Two Wheeler + Truck.
-- ================================================
****/
```