

Data Dictionary (Updated)

monthly_claims_insurance_driverid.csv

The Monthly Claims Report table contains **major accident claims data**, including details about the policy, claim processing timeline, responsible adjusters, financials, and loss specifics. Each row represents **an individual claim** reported for a major accident within a given month. These records range from Jan 2023- March 2025

Column Name	Data Type	Description
Year of Policy Effect	INTEGER	The year in which the policy was issued or became effective.
Loss Date	DATE	The date when the accident or loss event occurred.
Reported Date	DATE	The date the claim was officially reported to the insurance company.
Lag Time	INTEGER	The number of days between Loss Date and Reported Date .
Client Name	VARCHAR	The name of the insured entity (individual or company).
Claim Number	VARCHAR	A unique identifier assigned to the claim.
Adjuster	VARCHAR	The name or ID of the insurance adjuster handling the claim.
Loss Location	VARCHAR	The specific location where the accident occurred.
State Code	CHAR(2)	The two-letter state abbreviation (e.g., CA for California).
Claims Line	VARCHAR	The type of insurance claim (Commercial Auto).
Loss Cause	VARCHAR	General cause of loss (Auto Accident , Fire , Theft , etc.).
Loss Cause Detail	VARCHAR	Additional details specifying the cause of loss.
Current Claim Status	VARCHAR	The latest status of the claim (Open , Closed , Pending).
Claim Current Close	DATE	The date when the claim was officially closed.
Insured Driver	VARCHAR	The name of the insured driver involved in the accident.

Column Name	Data Type	Description
Loss Description	TEXT	Detailed summary of the accident or loss event.
Gross Paid	DECIMAL(15,2)	Total amount paid by the insurer before recoveries.
Loss Reserve	DECIMAL(15,2)	Estimated amount set aside for future claim payments.
Recoveries	DECIMAL(15,2)	Amount recovered from subrogation, salvage, or other sources.
Net Incurred	DECIMAL(15,2)	The net cost of the claim (Gross Paid + Loss Reserve – Recoveries).
driver_id	INTEGER	The associated driver ID with Motive App

`accident_register_combined_driverid.csv`

The `accident_records` table documents **all minor accident incidents** that occurred in **2025, 2024, and 2023**. It captures details such as **location, severity, involved personnel, claims, and financial impact**. These records range from Jan 2023- March 2025

- **Each row represents a single accident report** recorded within the given timeframe.
- **Primary Key:** Claim Number (unique identifier for each claim).

Column Name	Data Type	Description
Location of Accidents	VARCHAR	The physical location where the accident occurred (e.g., city, site, or address).
Region	VARCHAR	The regional classification associated with the accident location.
DBA	VARCHAR	"Doing Business As" name for the business entity associated with the accident.
General Manager	VARCHAR	The name of the General Manager overseeing the location where the accident occurred.
Date of Accident	DATE	The date (YYYY-MM-DD) when the accident took place.
Time of Accident	TIME	The exact time when the accident occurred.
Date Reported	DATE	The date (YYYY-MM-DD) when the accident was reported.
Fatality	BOOLEAN	Indicates whether the accident resulted in a fatality (true = yes, false = no).
Driver Name	VARCHAR	The name of the driver involved in the accident.

Column Name	Data Type	Description
Copy of State or Insurance Report	BOOLEAN	Indicates whether a state or insurance report is available (true = yes, false = no).
Severity	VARCHAR	The severity level of the accident (e.g., Minor , Moderate , Severe).
Summary of Accident	TEXT	A brief description of how the accident occurred and contributing factors.
Claim Number	VARCHAR	Unique identifier assigned to the claim related to the accident.
Total Incurred	DECIMAL (15 , 2)	The total financial cost associated with the accident (e.g., damages, medical expenses).
Notes	TEXT	Additional details or remarks about the accident.
APMM Recordable	BOOLEAN	Indicates whether the accident is recordable under APMM safety compliance (true = yes, false = no).
driver_id	INTEGER	The associated driver ID with Motive App

data_inspections_w_vehicle_map

The `data_inspections_w_vehicle_map` table records truck inspections at the beginning of the day/ before trips and post-trip to ensure compliance with safety checks. These records range from Jan 2023- March 2025

- Each row represents an individual inspection log.
- **Primary Key:** `inspection_id` (unique identifier for each inspection record).
- **Unique Constraint:** `(vehicle_id, date, inspection_type)` , ensuring a vehicle is inspected only once per type per day.
- **Duplicate Handling:** Drivers may accidentally log multiple inspections on the same day, or omit `inspection_type` .

Column Name	Data Type	Description
id	INTEGER	Auto-incremented unique row identifier (optional primary key alternative).
inspection_id	INTEGER	Unique identifier for the inspection log.
vehicle_id	INTEGER	Unique ID assigned to the vehicle being inspected.
date	DATE	The date when the inspection occurred (YYYY-MM-DD).

Column Name	Data Type	Description
location	VARCHAR	The physical location where the inspection took place.
status	VARCHAR	The inspection status (acceptable , open , resolved).
inspection_type	VARCHAR	Type of inspection performed (pre-trip , post-trip). May be missing in some cases.
driver_id	INTEGER	The unique identifier of the driver performing the inspection.
mechanic_id	INTEGER	Mechanic's ID, if the inspection required mechanic review.
reviewer_id	INTEGER	ID of the reviewer who validates or approves the inspection.
number	VARCHAR	The vehicle's unique number in the company's internal tracking system.
status-2	VARCHAR	The vehicle's operational status (active , deactivated).
make	VARCHAR	The manufacturer or main company of the vehicle.
model	VARCHAR	The specific model of the vehicle.

Notes

- Each inspection is linked to a `vehicle_id` and `date` , ensuring tracking of daily vehicle inspections.
- Duplicate inspections may occur due to drivers mistakenly logging multiple entries. Data validation should account for this.
- Inspection type (`pre-trip` , `post-trip`) may be missing, requiring data cleansing or defaulting.
- Vehicles remain active (`status-2`) unless deactivated, indicating whether the truck is still in service.
- Foreign keys (`driver_id` , `mechanic_id` , `reviewer_id`) should link to the respective user/employee database for consistency.
- This process is not rigorously being implemented; therefore, a lot of missing information

data_idle_events_w_vehicle_map

The `idle_events_log` table records idle events whenever a truck is idling during a trip, within the driver's clock-in and clock-out hours. These records range from Jan 2023- March 2025

- Each row represents a single idle event, capturing start time, end time, and duration.
- Idle events are logged per `event_id` whenever a truck is idling during a trip within clock-in/out hours.
- **Primary Key:** `event_id` (unique identifier for each idle event).

Column Name	Data Type	Description
id	INTEGER	Auto-incremented unique row identifier (optional).
event_id	INTEGER	Unique identifier for each idle event (Primary Key).
start_time	TIMESTAMP	The timestamp when idling began.
end_time	TIMESTAMP	The timestamp when idling ended.
driver_company_id	INTEGER	The driver's company ID (from Motive's system).
driver_id	INTEGER	Unique identifier for the driver.
minutes_idling	INTEGER	Total duration (in minutes) the truck remained idle or the driver was resting.
number	VARCHAR	The vehicle's unique number in the company's internal tracking system.
status-2	VARCHAR	The vehicle's operational status (active , deactivated).
make	VARCHAR	The manufacturer or main company of the vehicle.
model	VARCHAR	The specific model of the vehicle.

data_driving_periods_w_vehicle_map

The `data_driving_periods_w_vehicle_map` table tracks driving sessions for truck drivers, linking them to vehicles and recording trip details such as distance, timestamps, and driving type. These records range from Jan 2023- March 2025

- Each row represents a **single driving period** (a continuous driving session).
- **Primary Key:** `event_id` (unique identifier for each driving period).
- **Unassigned Trips:** If a driver does not log into the Motive mobile app, the trip is recorded under "unassigned" .
- **Trip Classification:** The column `driving_period_type` categorizes trips as `driving` , `pc` (Personal Conveyance), or `ym` (Yard Move).

Column Name	Data Type	Description
id	INTEGER	Auto-incremented unique row identifier (optional).
event_id	INTEGER	Unique identifier for each driving period (Primary Key).
driver_id	INTEGER	Unique identifier for the driver. If unassigned, marked as "unassigned" .
driver_first_name	VARCHAR	The first name of the driver (if signed in).

Column Name	Data Type	Description
driver_last_name	VARCHAR	The last name of the driver (if signed in).
vehicle_id	INTEGER	Unique identifier of the vehicle used in the driving session.
start_date	TIMESTAMP	Start time of the driving period.
end_date	TIMESTAMP	End time of the driving period.
driving_distance	DECIMAL (10 , 2)	Distance driven in miles or kilometers (depending on system settings).
driving_period_type	VARCHAR	Categorization of trip type: <code>driving</code> , <code>pc</code> (Personal Conveyance), or <code>ym</code> (Yard Move).
driver_company_id	INTEGER	The driver's company ID from Motive's system.
minutes_driving	INTEGER	Total time (in minutes) spent driving during the session.
month	INTEGER	The month (same as <code>start_date</code>) when the trip started.
created_at	TIMESTAMP	Timestamp when the driving period record was created.
updated_at	TIMESTAMP	Timestamp when the record was last updated.
unassigned	BOOLEAN	Indicates if the driver was logged in (<code>false</code>) or if the trip was recorded under "unassigned" (<code>true</code>).
not_current	BOOLEAN	Indicates whether the record is outdated (<code>true</code>) or active (<code>null</code>).
number	VARCHAR	The vehicle's unique number in the company's tracking system.
status-2	VARCHAR	The vehicle's operational status (<code>active</code> , <code>deactivated</code>).
make	VARCHAR	The manufacturer or main company of the vehicle.
model	VARCHAR	The specific model of the vehicle.

data_combined_events_w_vehicle_map

The `data_combined_events_w_vehicle_map` table consolidates **safety and operational events** detected by the Motive system. This includes **hazard events, safety violations, coaching interventions, vehicle information, and status tracking**. These records range from Jan 2023- March 2025

- **Each row represents a recorded event** associated with a driver and vehicle.
- **Primary Key:** `event_id` (unique identifier for each event).
- **Contains both driving behavior events and vehicle-specific information.**

Column Name	Data Type	Description
id	INTEGER	Auto-incremented unique row identifier (optional).
event_id	INTEGER	Unique identifier for each event (Primary Key).
type	VARCHAR	The type of event recorded. One of: <code>camera_obstruction</code> , <code>cell_phone</code> , <code>crash</code> , <code>distraction</code> , <code>driver_facing_cam_obstruction</code> , <code>drowsiness</code> , <code>forward_collision_warning</code> , <code>manual_event</code> , <code>near_miss</code> , <code>ran_a_red_light</code> , <code>road_facing_cam_obstruction</code> , <code>seat_belt_violation</code> , <code>speeding</code> , <code>stop_sign_violation</code> , <code>tailgating</code> , <code>unsafe_lane_change</code> .
driver_id	INTEGER	Unique identifier for the driver involved in the event.
driver_first_name	VARCHAR	The first name of the driver.
driver_last_name	VARCHAR	The last name of the driver.
vehicle_id	INTEGER	Unique identifier of the vehicle associated with the event.
coaching_status	VARCHAR	The coaching review status for the event. Options: <code>Pending Review</code> , <code>Coachable</code> , <code>Coached</code> , <code>Uncoachable</code> .
start_date	TIMESTAMP	The time when the driver start driving
severity	VARCHAR	The severity level of the event (<code>Low</code> , <code>Moderate</code> , <code>Severe</code>).
group_id	INTEGER	The group (e.g., department or fleet) associated with the driver.
group_name	VARCHAR	The name of the driver's assigned group.
month	INTEGER	The month (same as <code>start_date</code>) when the event occurred.

Column Name	Data Type	Description
created_at	TIMESTAMP	Timestamp when the event record was created.
updated_at	TIMESTAMP	Timestamp when the record was last updated (typically when <code>coaching_status</code> changes).
max_over_speed_in_kph	DECIMAL (5,2)	Maximum recorded speed over the limit (in kilometers per hour) for speeding violations.
max_over_speed_in_mph	DECIMAL (5,2)	Maximum recorded speed over the limit (in miles per hour) for speeding violations.
number	VARCHAR	The vehicle's unique number in the company's tracking system.
status-2	VARCHAR	The vehicle's operational status (<code>active</code> , <code>deactivated</code>).
make	VARCHAR	The manufacturer or main company of the vehicle.
model	VARCHAR	The specific model of the vehicle.

full_driver_details.csv

The `truck_driver_records` table contains a **list of active and deactivated truck drivers**, along with their licensing, activity status, and device usage details.

- Each row represents a single driver.
- Primary Key: `Driver ID` (unique identifier for each driver).
- Tracks driver status (`active` or `deactivated`).

Column Name	Data Type	Description
Driver	VARCHAR	Full name of the truck driver.
Driver ID	INTEGER	Unique identifier assigned to the driver (Primary Key).
License State	CHAR(2)	The state where the driver's license was issued (e.g., <code>CA</code> , <code>TX</code>).
Groups	VARCHAR	The group(s) or fleet(s) to which the driver is assigned.
Driver Type	VARCHAR	Indicates whether the driver follows <code>ELD</code> (Electronic Logging Device) regulations or is classified as <code>Logs Not Required</code> .
Vehicle Gateway Enabled	BOOLEAN	Indicates whether a vehicle gateway is enabled (<code>true</code> = Yes, <code>false</code> = No).

Column Name	Data Type	Description
Last Activity	TIMESTAMP	The last recorded activity of the driver (e.g., last login, last recorded trip).
Recent Vehicle	VARCHAR	The vehicle ID or description of the last vehicle assigned to the driver.
Vehicle Gateway S/N	VARCHAR	The serial number of the vehicle's logging gateway (if applicable).
Last Drive Time	TIMESTAMP	The timestamp of the last recorded driving session.
Cycle Rule	VARCHAR	The regulatory cycle rule the driver follows (e.g., US 70-hour/8-day , Canada South 70-hour/7-day).
Device Platform	VARCHAR	The type of mobile platform the driver uses for logging (e.g., iOS , Android).
App Version	VARCHAR	The version of the Motive app installed on the driver's device.
App Up To Date	BOOLEAN	Indicates whether the driver's app is updated to the latest version (true = Yes, false = No).
Time Zone	VARCHAR	The driver's configured time zone (e.g., UTC-5 , PST).
driver_status	VARCHAR	The current status of the driver: active or deactivated .

External Data

State Crash Data

- We have on csv for each year (2023-2025) with the following information gathered for MCMIS Crash data includes crashes that are reported by states to the FMCSA through the SAFETYNET computer reporting system.

Field Name	Data Type	Description
Crash ID	String	A unique identifier for each crash record.
Crash Year	Numeric	The year in which the crash occurred.
Quarter-CY	String	Calendar quarter of the crash. Calculated as: Q1 (Jan 01-Mar 31), Q2 (Apr 01-Jun 30), Q3 (Jul 01-Sep 30), Q4 (Oct 01-Dec 31).
Report State	String	The state in which the crash occurred.
Truck Bus Indicator	String	Indicates vehicle type involved: "T" for truck, "B" for bus.

Field Name	Data Type	Description
Fatal Count	Boolean/Numeric	Calculated field indicating whether the crash record involves a fatality.
Non Fatal Count	Boolean/Numeric	Calculated field indicating whether the crash record is non-fatal.
Fatalities	Numeric	Calculated field showing the number of fatalities resulting from the crash.
Injuries	Numeric	Calculated field showing the number of injuries resulting from the crash.
Injury Count	Boolean/Numeric	Calculated field indicating whether the crash resulted in injuries.
Tow Away	Boolean/Numeric	Calculated field indicating whether the crash required vehicle(s) to be towed away.
Federal Recordable	String	"Yes" or "No" - Indicates if crash meets federal criteria: at least 1 fatality, 1 injury requiring immediate medical attention away from the scene, or a vehicle towed due to disability damage.
State Recordable	String	"Yes" or "No" - Indicates if crash meets state reportable criteria.
Vehicles In Accident	Numeric	The total number of vehicles involved in the crash.
City Code	String	Code that translates to city name.
City	String	Name of the city where the crash occurred.
County Code	String	Code that translates to county name.
County Name	String	Name of the county where the crash occurred.
State	String	Name of the state where the crash occurred.
Location	String	Text field with detailed location information.
Trafficway Desc	String	Description of the trafficway where the crash occurred.
Road Surface Condition Desc	String	Description of road surface conditions (e.g., dry, wet, snow, ice).
Weather Condition Desc	String	Description of weather conditions (e.g., snow, rain, no adverse conditions).
Light Condition Desc	String	Description of lighting conditions (e.g., daylight, dark).
First Harmful Event	String	Description of the first harmful event in the crash (e.g., collision involving motor vehicle in transport).

Field Name	Data Type	Description
USDOT Number	String	U.S. Department of Transportation identification number for the carrier.
Carrier Name	String	Name of the carrier involved in the crash.
Carrier Street	String	Street address of the carrier.
Carrier City	String	City of the carrier's address.
Carrier State	String	State of the carrier's address.
Carrier Zip Code	String	ZIP code of the carrier's address.
Carrier Domicile	String	"Yes" or "No" - Indicates if carrier is domiciled in the state where the crash occurred.
Carrier Add Date	Date	Date when the carrier obtained operating authority (Census data).
New Entrant	Boolean/String	Calculated field indicating if the carrier's crash occurred within 18 months of obtaining operating authority.
Mexican Location	String	Carrier address/Mexican neighborhood if applicable.
Interstate Carrier	String	"Yes" or "No" - Indicates if carrier operates as an interstate carrier.
State Number	String	State registration number of the carrier.
State Issuing Number	String	State that assigned the registration number.
Driver Age Category	String	Age category of the driver.
Driver License State	String	State that issued the driver's license.
Valid Driver License	String	"Yes" or "No" - Indicates if the driver had a valid license.
Driver License Class Desc	String	Description of the driver's license class.
Citation Issued Code Desc	String	Indicates if citation was issued: "Yes", "No", or "Warning".
Vehicle Configuration Desc	String	Description of vehicle configuration (e.g., bus, tractor/semi-trailer).
Cargo Body Type Desc	String	Description of cargo body type (e.g., bus, cargo tank).

Field Name	Data Type	Description
Axles	Numeric	Number of axles on the vehicle.
GVWR	Numeric	Gross Vehicle Weight Rating.
GVW Rating Desc	String	Description of the Gross Vehicle Weight Rating.
Vehicle License Number	String	License plate number of the vehicle.
Vehicle License State	String	State that issued the vehicle license.
Hazmat Placard	String	"Yes" or "No" - Indicates if vehicle displays a hazardous materials placard.
Crash Number	String	Report number for the crash.
Crash Date	Date	Date when the crash occurred.
Day of Week Category	String	Category indicating day of the week when the crash occurred.
Crash Time	Time	Time when the crash occurred (in military format).
Time of Day Category	String	Category indicating time of day when the crash occurred.
Crash Seq No	Numeric	Sequence number indicating vehicle order (1st or 2nd) in the crash; value is 1 unless multiple vehicles were involved.
Agency	String	Agency that collected information about the crash.
Record Status	String	Indicates if this is an "Add record" (added to MCMIS and not modified) or a "Change record" (modified since initial addition).
Matched Status	String	Indicates if the US DOT# successfully matched with MCMIS Motor carrier census information. Values include: complete, nonmatch, intratstate, fmcsa hold, potential resolution, to census search, non motor carrier, initial load.
SAFETYNET Input Date	Date	Date when the crash was added to SAFETYNET.
MCMIS Original Upload Date	Date	Date when the crash record was first uploaded from SAFETYNET to MCMIS. May not be populated for older records.
MCMIS Upload Date	Date	Date when the current crash record was uploaded to MCMIS. For change records, this is the date of the

Field Name	Data Type	Description
		change; for add records, this is the original upload date.
Number Days to SAFETYNET	Numeric	Calculated field showing the number of days between crash date and SAFETYNET input date.
Number Days to MCMIS	Numeric	Calculated field showing the number of days between crash date and MCMIS upload date.
Counter	Numeric	Field used to subtotal the number of records.

Precipitation Data

- This dataset provides daily precipitation data for the conterminous United States from January 1, 2023 to January 31, 2025.
- Each row in the dataset represents one zip code and there is one column for each date
- We have three datasets: one for 2023, one for 2024, and one for 2025

Column Name	Data Type	Description
ALAND20	INTEGER	2020 census land area
AWATER20	INTEGER	2020 Census water area
CLASSFP20	VARCHAR	2020 Census Class of a geographic entity
FUNCSTAT20	VARCHAR	2020 Census functional status
INTPTLAT20	VARCHAR	2020 Census latitude of the internal point
MTFCC20	INTEGER	2020 Census Feature Class Code
ZCTA5CE20	STRING	Zip Code
latitude	INTEGER	Latitude
longitude	VARCHAR	Longitude
PRISM_ppt _	VARCHAR	Total precipitation (rain+melted snow) for each Date in mm

Zip Code Data

- This dataset contains the 2020 Census 5-Digit ZIP Code Tabulation Area (ZCTA5) shapefiles from the TIGER/Line series, last updated on January 27, 2024. The data is derived from the U.S. Census Bureau's Master Address File / Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) Database.
- ZCTAs are geographic representations of U.S. Postal Service (USPS) ZIP Code service areas created by the Census Bureau specifically for statistical purposes. These boundaries

are delineated once per decade following each decennial census for the United States, Puerto Rico, and U.S. territories. The ZCTA boundaries in this release reflect those established following the 2020 Census.

- Files:
 - `tl_2020_us_zcta520.cpg` - A code page file that specifies the character encoding used in the attribute table.
 - `tl_2020_us_zcta520.dbf` - The dBASE database file containing attribute data for each ZCTA, such as ZIP code, geographic identifiers, and area measurements.
 - `tl_2020_us_zcta520.prj` - The projection file defining the coordinate system and map projection used for the spatial data.
 - `tl_2020_us_zcta520.shp` - The main shapefile containing the actual geometric data (polygons) for each ZCTA.
 - `tl_2020_us_zcta520.shp.ea.iso.xml` and `tl_2020_us_zcta520.shp.iso.xml` - XML metadata files containing information about the data source, creation, specifications, and other details about the shapefile.
 - `tl_2020_us_zcta520.shx` - The shape index file that allows software to quickly locate the spatial features within the .shp file.
- For more information see [Data.gov](https://data.gov)

US Cities States Data

- Free US Zip Code Database provided by <https://simplemaps.com/data/us-zips>
- All latitude and longitude coordinates for ZCTAs (and geographic relationships) are now based on ZCTA shapes from the 2020 Census.

Column Name	Data Type	Description
city	String	The primary city name associated with the ZIP code, using proper capitalization and spacing.
city_ascii	String	The city name in ASCII format, with special characters removed and standardized for easier processing.
state_id	String	The two-letter state abbreviation (e.g., "NY" for New York) following USPS standards.
state_name	String	The full name of the state (e.g., "New York").
county_fips	String	The Federal Information Processing Standards (FIPS) code for the county, a unique identifier used by the US government.
county_name	String	The full name of the county where the ZIP code is located.

Column Name	Data Type	Description
lat	Numeric	The latitude coordinate of the approximate geographic center of the ZIP code area, in decimal degrees.
lng	Numeric	The longitude coordinate of the approximate geographic center of the ZIP code area, in decimal degrees.
population	Numeric	The estimated population residing within the ZIP code boundaries, based on recent Census or other demographic data.
density	Numeric	Population density, measured in people per square mile or square kilometer.
source	String	The source of the ZIP code data, which may include Census Bureau, USPS, or other demographic sources.
military	Boolean	Indicates whether the ZIP code is primarily associated with a military installation (true/false).
incorporated	Boolean	Indicates whether the area is within an incorporated city or town boundary (true/false).
timezone	String	The time zone in which the ZIP code is located (e.g., "America/New_York").
ranking	Numeric	A relative importance ranking assigned to the ZIP code, potentially based on population or other metrics.
zips	String	The 5-digit ZIP code itself.
id	String/Numeric	A unique identifier assigned to each record in the database.

Fusion Site Regional Data

,

- Information on Fusion Site Geographic footprint

Column Name	Data Type	Description
ID	Text	Unique identifier
Brand	Text	Company brand name (e.g., "Arkansas Portable Toilets")
Region	Text	Geographic region designation (e.g., "Gulf South")
State	Text	U.S. state location (e.g., "Arkansas")

Column Name	Data Type	Description
City	Text	City name (e.g., "Little Rock")
Address	Text	Complete address string
PR	Boolean	Indicator for Portable Restroom services (X = Yes, blank = No)
RR	Boolean	Indicator for Restroom Trailer services (X = Yes, blank = No)
T	Boolean	Unknown service type indicator
Shower	Boolean	Indicator for Shower services (X = Yes, blank = No)
RO	Boolean	Unknown service type indicator
Front Loader	Boolean	Indicator for Front Loader equipment (X = Yes, blank = No)
Septic	Boolean	Indicator for Septic services (X = Yes, blank = No)
Grease	Boolean	Indicator for Grease Trap services (X = Yes, blank = No)
Concrete	Boolean	Indicator for Concrete services (X = Yes, blank = No)
S&B	Boolean	Unknown service type indicator
Conex	Boolean	Indicator for Conex container services (X = Yes, blank = No)
Mobile Offices	Boolean	Indicator for Mobile Office services (X = Yes, blank = No)
Res Storage	Boolean	Indicator for Residential Storage services (X = Yes, blank = No)
Refrigeration	Boolean	Indicator for Refrigeration services (X = Yes, blank = No)
Latitude	Numeric	Geographic latitude coordinate (e.g., 34.768452)
Longitude	Numeric	Geographic longitude coordinate (e.g., -92.277291)
Number	Numeric	Street number portion of address (e.g., 924)
Street	Text	Street name portion of address (e.g., "W 15th St")
Unit Type	Text	Type of unit designation, if applicable
Unit Number	Text	Unit identifier, if applicable
City	Text	City name, parsed from address (e.g., "North Little Rock")
State	Text	State abbreviation (e.g., "AR")
County	Text	County name (e.g., "Pulaski County")
Zip	Numeric	ZIP/Postal code (e.g., 72114)
Country	Text	Country code (e.g., "US")
FUSION SITE	Text	Contact information for regional manager

Appendix

Not all of the data we needed for our analysis was readily available in file formats compatible with our modeling pipeline. We first had to manually process the raw data from original sources to create usable input data for our project.

The following information outlines our workflow from downloading the raw data to preparing the input data for this project. We needed to use GeoPandas specifically to handle the .BIL file format before converting to Spark dataframes, as this geospatial data required specialized processing before it could be incorporated into our main analytical pipeline.

Additionally, we originally downloaded the state crash data in .xlsx format, but this was causing a lot of issues with Databricks and GCC crashing when trying to read in several of these huge files. Instead, we simply converted the sheets we need to a csv and then made one csv per year so we could read it in more easily for the rest of our data pipeline.

State Crash Data From Source

- There is one .xlsx for each state with the following sheets:
 - GENERAL DEFINITIONS
 - MCIS COLUMN NAMES
 - FARS COLUMN NAMES
 - FARS & MCMIS COMMON NAMES
 - {State Code} 2018-2025 MCMIS
 - {State Code} {Year} MCMIS
 - {State Code} 2018-2025 FARS
 - {State Code} {Year} FARS
- We gathered this data from the {State Code} {Year} MCMIS sheets for every state of interest between the years 2023 and 2025, convert it to a csv and then saved all of the data for each year into its own csv with . This process is outlined in more detail in `state_crash_data_raw_to_input.ipynb`. The outputs from this notebook were manually downloaded from the Google Cloud bucket, and directly uploaded into DBFS
- For more information about the Motor Carrier Management Information System (MCMIS) see [FMCSA](#)

Precipitation Data

- This dataset provides daily precipitation data for the conterminous United States from January 1, 2023 to January 31, 2025. The data is spatially gridded at a 4km resolution using the PRISM model, which was developed by Dr. Christopher Daly of the PRISM Climate Group at Oregon State University.
- The PRISM model defines a "day" differently than the standard midnight-to-midnight period. Instead, each "day" represents the 24-hour period ending at 12:00 Greenwich Mean Time

(7:00am Eastern Standard Time).

- The reliability of the data depends on its age. Data older than 6 months is considered "stable" and unlikely to change until a major new PRISM release. Data from 1-6 months ago is "provisional" and likely to change as reporting networks finalize their information. Data from the current month should be treated as "early results" that will certainly change as new reporting stations are added and quality control measures are applied.
- Each day of data includes seven different files:
 1. A CSV file contains a list of all stations that provided input data for the PRISM model run that produced the grid associated with this particular day.
 - Station
 - Name
 - Longitude
 - Latitude
 - Elevation (m)
 - Network
 - station_id
 2. A BIL (Band Interleaved by Line) file serves as the main raster data file, which is a common format for storing spatial raster data.
 3. A PRJ file defines the coordinate system and projection information for the spatial data.
 4. An HDR file contains header information about the data structure and parameters.
 5. An STX file is included with each day's data.
 6. An XML auxiliary file contains metadata for the BIL format.
 7. A TXT file provides additional metadata and information about the dataset.
- The pre-processing for this dataset into the input data tables used in our data pipeline can be found in `weather_data_processing.ipynb`
- For more information see [PRISM](#)