DAMG7370 – FINAL PROJECT
Motor Vehicle collisions/Crashes

DATASETS:

AUSTIN	NYC	CHICAGO
ID	CRASH_DATE	CRASH_RECORD_ID
CRASH_ID	CRASH_TIME	CRASH_DATE_EST_I
CRASH_FATAL_FL	BOROUGH	CRASH_DATE
CASE_ID	ZIP_CODE	POSTED_SPEED_LIMIT
PRIMARY_ADDRESS	LATITUDE	TRAFFIC_CONTROL_DEVICE
RPT_BLOCK_NUM	LONGITUDE	DEVICE_CONDITION
RPT_STREET_NAME	LOCATION	WEATHER_CONDITION
RPT_STREET_SFX	ON_STREET_NAME	LIGHTING_CONDITION
CRASH_SPEED_LIMIT	CROSS_STREET_NAME	FIRST_CRASH_TYPE
ROAD_CONSTR_ZONE_FL	OFF_STREET_NAME	TRAFFICWAY_TYPE
LATITUDE	NUMBER_OF_PERSONS_INJURED	LANE_CNT
LONGITUDE	NUMBER_OF_PERSONS_KILLED	ALIGNMENT
CRASH_SEV_ID	NUMBER_OF_PEDESTRIANS_INJURED	ROADWAY_SURFACE_COND
SUS_SERIOUS_INJRY_CNT	NUMBER_OF_PEDESTRIANS_KILLED	ROAD_DEFECT
POSS_INJRY_CNT	NUMBER_OF_CYCLIST_INJURED	REPORT_TYPE
NON_INJRY_CNT	NUMBER_OF_CYCLIST_KILLED	CRASH_TYPE
UNKN_INJRY_CNT	NUMBER_OF_MOTORIST_INJURED	INTERSECTION_RELATED_I
TOT_INJRY_CNT	NUMBER_OF_MOTORIST_KILLED	NOT_RIGHT_OF_WAY_I
DEATH_CNT	CONTRIBUTING_FACTOR_VEHICLE_1	HIT_AND_RUN_I
UNITS_INVOLVED	CONTRIBUTING_FACTOR_VEHICLE_2	DAMAGE
POINT	CONTRIBUTING_FACTOR_VEHICLE_3	DATE_POLICE_NOTIFIED
MOTOR_VEHICLE_DEATH_COUNT	CONTRIBUTING_FACTOR_VEHICLE_4	PRIM_CONTRIBUTORY_CAUSE
MOTOR_VEHICLE_SERIOUS_INJURY_COUNT	CONTRIBUTING_FACTOR_VEHICLE_5	SEC_CONTRIBUTORY_CAUSE
BICYCLE_DEATH_COUNT	COLLISION_ID	STREET_NO
BICYCLE_SERIOUS_INJURY_COUNT	VEHICLE_TYPE_CODE_1	STREET_DIRECTION

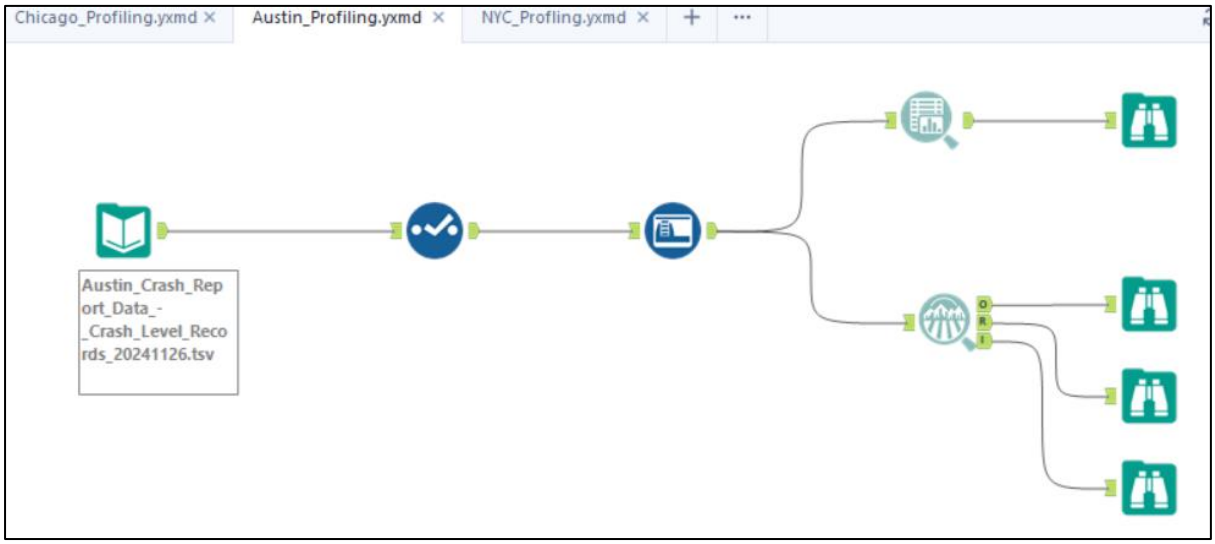
MOTORCYCLE_DEATH_COUNT	VEHICLE_TYPE_CODE_2	STREET_NAME
OTHER_DEATH_COUNT	VEHICLE_TYPE_CODE_3	BEAT_OF_OCCURRENCE
OTHER_SERIOUS_INJURY_COUNT	VEHICLE_TYPE_CODE_4	PHOTOS_TAKEN_I
ONSYS_FL	VEHICLE_TYPE_CODE_5	STATEMENTS_TAKEN_I
PRIVATE_DR_FL		DOORING_I
MICROMOBILITY_SERIOUS_INJURY_COUNT		WORK_ZONE_I
MIRCOMOBILITY_DEATH_COUNT		WORK_ZONE_TYPE
CRASH_TIMESTAMP_US_CENTRAL		WORKERS_PRESENT_I
CRASH_TIMESTAMP		NUM_UNITS
IS_DELETED		MOST_SEVERE_INJURY
IS_TEMPORARY_RECORD		INJURIES_TOTAL
LAW_ENFORCEMENT_FATALITY_COUNT		INJURIES_FATAL
REPORTED_STREET_PREFIX		INJURIES_INCAPACITATING
		INJURIES_NON_INCAPACITATING
		INJURIES_REPORTED_NOT_EVIDENT
		INJURIES_NO_INDICATION
		INJURIES_UNKNOWN
		CRASH_HOUR
		CRASH_DAY_OF_WEEK
		CRASH_MONTH
		LATITUDE
		LONGITUDE
		LOCATION

- **Objective**: To analyse motor vehicle collisions/crashes data from New York City, Austin, and Chicago using advanced data architectural techniques and business intelligence tools.
- **Data Sources**:
  - NYC Open Data
  - Austin Open Data
  - Chicago Data Portal

PART 1: DATA PREPARATION

- **Data Profiling**: Using Alteryx

AUSTIN DATASET



FIELD	TYPE	DESCRIPTION
ID	Number	The unique crash identifier within the Vision Zero crash database
CRASH_ID	Number	Unique identifier for each crash incident.
CRASH_FATAL_FL	Text	Indicates if the crash was fatal ('Y' for Yes, 'N' for No).
CASE_ID	Text	Identifier for the case associated with the crash.
PRIMARY_ADDRESS	Text	Primary address where the crash occurred.
SECONDARY_ADDRESS	Text	Secondary address related to the crash location, if applicable.
RPT_BLOCK_NUM	Text	Reported block number where the crash occurred.
RPT_STREET_NAME	Text	Reported street name where the crash occurred.
RPT_STREET_SFX	Text	Suffix of the reported street name (e.g., St, Ave).
CRASH_SPEED_LIMIT	Number	Speed limit at the location of the crash.
ROAD_CONST_ZONE_FL	Text	Indicates if the crash occurred in a road construction zone ('Y' or 'N').
LATITUDE	Number	Latitude coordinate of the crash location.

LONGITUDE	Number	Longitude coordinate of the crash location.
CRASH_SEV_ID	Number	Severity identifier of the crash.
SUS_SERIOUS_INJRY_CNT	Number	Suspected serious injury count resulting from the crash.
POSS_INJRY_CNT	Number	Possible injury count resulting from the crash.
NON_INJRY_CNT	Number	Non-injury count (individuals involved without injuries).
UNKN_INJRY_CNT	Number	Unknown injury count (injury status not determined).
TOT_INJRY_CNT	Number	Total injury count resulting from the crash.
DEATH_CNT	Number	Total number of fatalities resulting from the crash.
UNITS_INVOLVED	Number	Number of units (vehicles, pedestrians, etc.) involved in the crash.
POINT	Location	Geographical point representing the crash location.
MOTOR_VEHICLE_DEATH_COUNT	Number	Number of motor vehicle occupant fatalities.
MOTOR_VEHICLE_SERIOUS_INJURY_COUNT	Number	Number of motor vehicle occupants with serious injuries.
BICYCLE_DEATH_COUNT	Number	Number of bicyclist fatalities.
BICYCLE_SERIOUS_INJURY_COUNT	Number	Number of bicyclists with serious injuries.
MOTORCYCLE_DEATH_COUNT	Number	Number of motorcyclist fatalities.
OTHER_DEATH_COUNT	Number	Number of fatalities not categorized above.
OTHER_SERIOUS_INJURY_COUNT	Number	Number of serious injuries not categorized above.
ONSYS_FL	Text	On-system flag indicating if the road is part of the state highway system ('Y' or 'N').
PRIVATE_DR_FL	Text	Indicates if the crash occurred on a private drive ('Y' or 'N').
MICROMOBILITY_SERIOUS_INJURY_COUNT	Number	Number of serious injuries involving micromobility devices (e.g., e-scooters).
MIRCOMOBILITY_DEATH_COUNT	Number	Number of fatalities involving micromobility devices.
CRASH_TIMESTAMP_US_CENTRAL	Date & Time	Timestamp of the crash in US Central time.

CRASH_TIMESTAMP	Date & Time	General timestamp of the crash.
IS_DELETED	Text	Indicates if the record has been deleted ('Y' or 'N').
IS_TEMPORARY_RECORD	Text	Indicates if the record is temporary ('Y' or 'N').
LAW_ENFORCEMENT_FATALITY_COUNT	Number	Number of law enforcement fatalities resulting from the crash.
REPORTED_STREET_PREFIX	Text	Prefix of the reported street name (e.g., N, S, E, W).

Data Quality Analysis (AUSTIN)

By the reference of the 5Cs of data

Measure	**Importance**	**Required Insights**
Clean	Ensures that data is free from errors, irrelevant entries, and is formatted correctly.	Check for and remove null values in critical fields like CRASH_ID and LATITUDE.
Consistent	Verifies that data is logically coherent with uniformity across datasets.	Ensure LATITUDE and LONGITUDE values align with valid Austin geographic boundaries.
Comprehensive	Assesses the extent to which data covers all necessary aspects and elements.	Confirm all fields related to injuries and fatalities (e.g., SUS_SERIOUS_INJRY_CNT, DEATH_CNT) are populated.
Confirmed	Validates that data is accurate and verified against reliable sources.	Cross-reference LATITUDE and LONGITUDE with mapping services for accuracy.
Current	Confirms that the dataset is up-to-date and relevant for the intended analysis.	Verify the CRASH_TIMESTAMP_US_CENTRAL field reflects recent crash incidents.

Field Analysis (Austin Dataset)

Field	**Description**	**Analysis**
ID	Unique identifier for each record.	Ensure uniqueness and validate as a non-null numeric field.

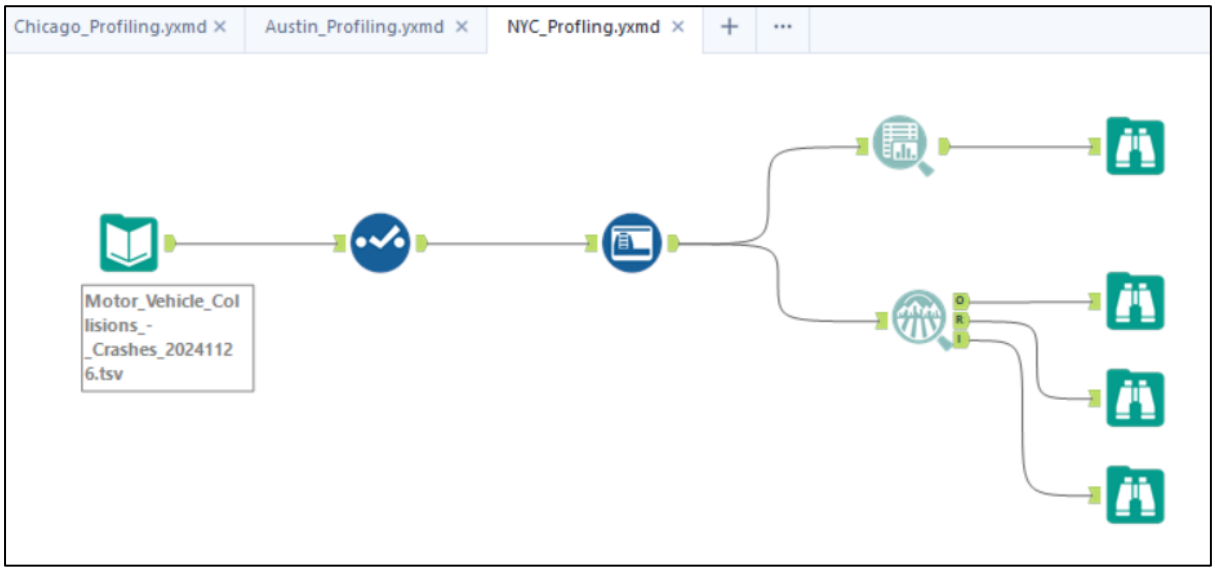
CRASH_ID	Unique identifier for each crash incident.	Validate as a non-null numeric field to maintain integrity across records.
CRASH_FATAL_FL	Indicates if the crash was fatal ('Y' or 'N').	Ensure valid binary values ('Y' or 'N') and check for consistency with DEATH_CNT.
CASE_ID	Identifier for the case associated with the crash.	Verify against official case records to ensure accuracy.
PRIMARY_ADDRESS	Primary address where the crash occurred.	Cross-reference with geographical data to ensure location accuracy.
SECONDARY_ADDRESS	Secondary address related to the crash location, if applicable.	Validate secondary address details for completeness.
RPT_BLOCK_NUM	Reported block number where the crash occurred.	Ensure that block numbers align with reported street data.
RPT_STREET_NAME	Reported street name where the crash occurred.	Validate street names against official mapping tools.
RPT_STREET_SFX	Suffix of the reported street name (e.g., St, Ave).	Standardize suffix values for consistency (e.g., "Street" vs. "St").
CRASH_SPEED_LIMIT	Speed limit at the location of the crash.	Check for reasonable values and flag outliers (e.g., unusually high or low limits).
ROAD_CONST_ZONE_FL	Indicates if the crash occurred in a road construction zone ('Y' or 'N').	Ensure valid binary values ('Y' or 'N') and verify with associated construction zone data.
LATITUDE	Latitude coordinate of the crash location.	Validate coordinates fall within Austin's geographic boundaries.
LONGITUDE	Longitude coordinate of the crash location.	Cross-check coordinates with mapping tools for accuracy.

CRASH_SEV_ID	Severity identifier of the crash.	Validate severity levels correspond to the injury and fatality data.
SUS_SERIOUS_INJRY_CNT	Suspected serious injury count resulting from the crash.	Ensure consistency with TOT_INJRY_CNT and validate against case records.
POSS_INJRY_CNT	Possible injury count resulting from the crash.	Validate that this count logically aligns with total injuries.
NON_INJRY_CNT	Non-injury count (individuals involved without injuries).	Ensure logical alignment with total units involved and injury counts.
UNKN_INJRY_CNT	Unknown injury count (injury status not determined).	Investigate and minimize the frequency of unknown values to improve data completeness.
TOT_INJRY_CNT	Total injury count resulting from the crash.	Confirm alignment with individual injury counts (SUS_SERIOUS_INJRY_CNT, POSS_INJRY_CNT, etc.).
DEATH_CNT	Total number of fatalities resulting from the crash.	Cross-check consistency with fatality counts across different units (e.g., MOTOR_VEHICLE_DEATH_COUNT).
UNITS_INVOLVED	Number of units (vehicles, pedestrians, etc.) involved in the crash.	Validate against individual records of injuries and fatalities for consistency.
POINT	Geographical point representing the crash location.	Ensure POINT corresponds to the LATITUDE and LONGITUDE fields.
MOTOR_VEHICLE_DEATH_COUNT	Number of motor vehicle occupant fatalities.	Validate against DEATH_CNT and check for logical alignment.
MOTOR_VEHICLE_SERIOUS_INJURY_COUNT	Number of motor vehicle occupants with serious injuries.	Ensure this count is consistent with total injuries (TOT_INJRY_CNT).

BICYCLE_DEATH_COUNT	Number of bicyclist fatalities.	Cross-check with DEATH_CNT and validate against case data.
BICYCLE_SERIOUS_INJURY_COUNT	Number of bicyclists with serious injuries.	Ensure alignment with total injuries and injury severity fields.
MOTORCYCLE_DEATH_COUNT	Number of motorcyclist fatalities.	Validate against DEATH_CNT for consistency.
OTHER_DEATH_COUNT	Number of fatalities not categorized above.	Investigate and validate against case records for completeness.
OTHER_SERIOUS_INJURY_COUNT	Number of serious injuries not categorized above.	Ensure logical consistency with total injury counts.
ONSYS_FL	On-system flag indicating if the road is part of the state highway system.	Validate binary values and check for alignment with location data.
PRIVATE_DR_FL	Indicates if the crash occurred on a private drive ('Y' or 'N').	Cross-check for logical consistency with address details.
MICROMOBILITY_SERIOUS_INJURY_COUNT	Number of serious injuries involving micromobility devices (e.g., e-scooters).	Ensure this count is included in total injuries and validate against micromobility records.
MIRCOMOBILITY_DEATH_COUNT	Number of fatalities involving micromobility devices.	Cross-check with DEATH_CNT and investigate completeness.
CRASH_TIMESTAMP_US_CENTRAL	Timestamp of the crash in US Central time.	Validate timestamp accuracy and ensure consistency with CRASH_DATE.
CRASH_TIMESTAMP	General timestamp of the crash.	Ensure consistency with CRASH_TIMESTAMP_US_CENTRAL.
IS_DELETED	Indicates if the record has been deleted ('Y' or 'N').	Verify binary values and investigate records flagged for deletion.

IS_TEMPORARY_RECORD	Indicates if the record is temporary ('Y' or 'N').	Check for logical use of temporary flags and ensure proper updates.
LAW_ENFORCEMENT_FATALITY_COUNT	Number of law enforcement fatalities resulting from the crash.	Validate against DEATH_CNT and cross-check with external law enforcement data.
REPORTED_STREET_PREFIX	Prefix of the reported street name (e.g., N, S, E, W).	Ensure prefixes align with Austin's official street data and validate completeness.

New York City Dataset



FIELD	TYPE	DESCRIPTION
CRASH_DATE	Date	The date on which the crash occurred.
CRASH_TIME	Time	The time at which the crash occurred.
BOROUGH	Text	The borough where the crash took place (e.g., Manhattan, Brooklyn).
ZIP_CODE	Text	The postal code of the crash location.
LATITUDE	Number	The latitude coordinate of the crash location.
LONGITUDE	Number	The longitude coordinate of the crash location.
LOCATION	Text	Combined latitude and longitude in a single field.
ON_STREET_NAME	Text	The street where the crash occurred.
CROSS_STREET_NAME	Text	The nearest cross street to the crash location.
OFF_STREET_NAME	Text	Additional off-street description of the crash location, if available.

NUMBER_OF_PERSONS_INJURED	Number	Total number of persons injured in the crash.
NUMBER_OF_PERSONS_KILLED	Number	Total number of persons killed in the crash.
NUMBER_OF_PEDESTRIANS_INJURED	Number	Total number of pedestrians injured in the crash.
NUMBER_OF_PEDESTRIANS_KILLED	Number	Total number of pedestrians killed in the crash.
NUMBER_OF_CYCLIST_INJURED	Number	Total number of cyclists injured in the crash.
NUMBER_OF_CYCLIST_KILLED	Number	Total number of cyclists killed in the crash.
NUMBER_OF_MOTORIST_INJURED	Number	Total number of motorists injured in the crash.
NUMBER_OF_MOTORIST_KILLED	Number	Total number of motorists killed in the crash.
CONTRIBUTING_FACTOR_VEHICLE_1	Text	Primary contributing factor of the first vehicle involved in the crash.
CONTRIBUTING_FACTOR_VEHICLE_2	Text	Primary contributing factor of the second vehicle involved in the crash.
CONTRIBUTING_FACTOR_VEHICLE_3	Text	Primary contributing factor of the third vehicle involved in the crash, if applicable.
CONTRIBUTING_FACTOR_VEHICLE_4	Text	Primary contributing factor of the fourth vehicle involved in the crash, if applicable.
CONTRIBUTING_FACTOR_VEHICLE_5	Text	Primary contributing factor of the fifth vehicle involved in the crash, if applicable.

COLLISION_ID	Number	A unique identifier for each collision record.
VEHICLE_TYPE_CODE_1	Text	The type of the first vehicle involved in the crash (e.g., sedan, truck).
VEHICLE_TYPE_CODE_2	Text	The type of the second vehicle involved in the crash.
VEHICLE_TYPE_CODE_3	Text	The type of the third vehicle involved in the crash, if applicable.
VEHICLE_TYPE_CODE_4	Text	The type of the fourth vehicle involved in the crash, if applicable.
VEHICLE_TYPE_CODE_5	Text	The type of the fifth vehicle involved in the crash, if applicable.

Data Quality Analysis (New York City)

By the reference of the 5Cs of data

Measure	**Importance**	**Required Insights**
Clean	Ensures that data is free from errors, irrelevant entries, and is formatted correctly.	Check for and remove null or invalid values in critical fields like COLLISION_ID and LOCATION.
Consistent	Verifies that data is logically coherent with uniformity across datasets.	Ensure ZIP_CODE values are valid and align with NYC boroughs.
Comprehensive	Assesses the extent to which data covers all necessary aspects and elements.	Confirm all injury and fatality-related fields (e.g., NUMBER_OF_PERSONS_INJURED, NUMBER_OF_PERSONS_KILLED) are filled.
Confirmed	Validates that data is accurate and verified against reliable sources.	Cross-reference LATITUDE and LONGITUDE with NYC mapping services for location accuracy.
Current	Confirms that the dataset is up-to-date and relevant for the intended analysis.	Ensure CRASH_DATE includes the latest crash data reported in NYC.

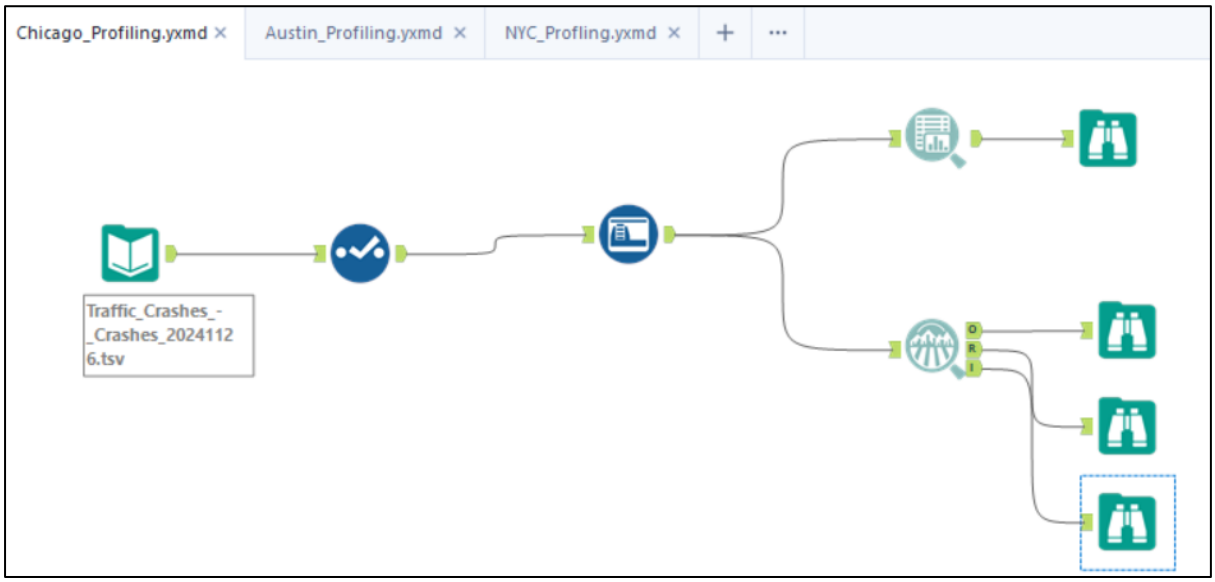
Field Analysis NYC Dataset

Field	Description	Analysis
CRASH_DATE	Date when the crash occurred.	Validate the format and ensure no null values to maintain temporal accuracy.
CRASH_TIME	Time when the crash occurred.	Ensure consistency with CRASH_DATE and validate the time format.
BOROUGH	The borough where the crash occurred (e.g., Manhattan, Brooklyn).	Cross-reference with ZIP_CODE for consistency and validate against NYC borough names.
ZIP_CODE	Postal code of the crash location.	Ensure ZIP codes match NYC regions and validate against BOROUGH.
LATITUDE	Latitude coordinate of the crash location.	Validate that values fall within NYC's geographic boundaries.
LONGITUDE	Longitude coordinate of the crash location.	Ensure values are accurate and correspond to LATITUDE for correct mapping.
LOCATION	Combined latitude and longitude of the crash location.	Validate that LOCATION reflects LATITUDE and LONGITUDE accurately.
ON_STREET_NAME	Name of the street where the crash occurred.	Check for accuracy and completeness using NYC street mapping.
CROSS_STREET_NAME	Name of the cross street near the crash location.	Ensure logical consistency with ON_STREET_NAME.
OFF_STREET_NAME	Off-street description of the crash location, if applicable.	Validate against geographical data for off-street locations.
NUMBER_OF_PERSONS_INJURED	Total number of individuals injured in the crash.	Ensure values align with sum of injuries across pedestrians, motorists, and cyclists.
NUMBER_OF_PERSONS_KILLED	Total number of fatalities in the crash.	Confirm alignment with individual fatality counts across demographics (pedestrians, cyclists, motorists).

NUMBER_OF_PEDESTRIANS_INJURED	Total number of pedestrians injured in the crash.	Validate that pedestrian injuries are correctly categorized and summed.
NUMBER_OF_PEDESTRIANS_KILLED	Total number of pedestrian fatalities in the crash.	Cross-check consistency with NUMBER_OF_PERSONS_KILLED and other injury-related fields.
NUMBER_OF_CYCLIST_INJURED	Total number of cyclists injured in the crash.	Validate consistency with other injury fields and ensure completeness.
NUMBER_OF_CYCLIST_KILLED	Total number of cyclist fatalities in the crash.	Ensure alignment with total fatality fields.
NUMBER_OF_MOTORIST_INJURED	Total number of motorists injured in the crash.	Validate consistency with total injuries and ensure proper categorization.
NUMBER_OF_MOTORIST_KILLED	Total number of motorist fatalities in the crash.	Cross-check alignment with NUMBER_OF_PERSONS_KILLED.
CONTRIBUTING_FACTOR_VEHICLE_1	Primary contributing factor for the first vehicle involved in the crash.	Validate against a predefined list of contributing factors (e.g., speeding, weather).
CONTRIBUTING_FACTOR_VEHICLE_2	Secondary contributing factor for the second vehicle involved, if applicable.	Ensure logical consistency with CONTRIBUTING_FACTOR_VEHICLE_1.
CONTRIBUTING_FACTOR_VEHICLE_3	Contributing factor for the third vehicle involved, if applicable.	Validate data entries and ensure values align with the first and second factors.
CONTRIBUTING_FACTOR_VEHICLE_4	Contributing factor for the fourth vehicle involved, if applicable.	Ensure accuracy and completeness if multiple vehicles are involved.
CONTRIBUTING_FACTOR_VEHICLE_5	Contributing factor for the fifth vehicle involved, if applicable.	Validate logical consistency for entries in multi-vehicle crashes.
COLLISION_ID	Unique identifier for each collision.	Ensure uniqueness and validate against the dataset for data integrity.

VEHICLE_TYPE_CODE_1	Type of the first vehicle involved in the crash.	Standardize vehicle type categories (e.g., sedan, SUV, truck).
VEHICLE_TYPE_CODE_2	Type of the second vehicle involved in the crash, if applicable.	Validate alignment with VEHICLE_TYPE_CODE_1 for logical consistency.
VEHICLE_TYPE_CODE_3	Type of the third vehicle involved in the crash, if applicable.	Ensure accuracy and completeness for multi-vehicle crashes.
VEHICLE_TYPE_CODE_4	Type of the fourth vehicle involved in the crash, if applicable.	Validate data entries for consistency and completeness.
VEHICLE_TYPE_CODE_5	Type of the fifth vehicle involved in the crash, if applicable.	Ensure data quality and logical consistency for crashes involving multiple vehicles.

Chicago Dataset



Field Name	Data Type	Description
CRASH_RECORD_ID	Text	Unique identifier for each crash record.
CRASH_DATE_EST_I	Text	Indicates if the crash date is estimated ('Y' or 'N').
CRASH_DATE	Date & Time	Date and time when the crash occurred.
POSTED_SPEED_LIMIT	Number	Speed limit posted at the crash location.
TRAFFIC_CONTROL_DEVICE	Text	Type of traffic control device present at the crash location.
DEVICE_CONDITION	Text	Condition of the traffic control device.
WEATHER_CONDITION	Text	Weather conditions at the time of the crash.
LIGHTING_CONDITION	Text	Lighting conditions at the time of the crash.
FIRST_CRASH_TYPE	Text	Initial type of collision in the crash sequence.
TRAFFICWAY_TYPE	Text	Layout or type of the trafficway where the crash occurred.
LANE_CNT	Number	Number of lanes in the roadway at the crash location.

ALIGNMENT	Text	Roadway alignment (e.g., straight, curve) at the crash location.
ROADWAY_SURFACE_COND	Text	Condition of the roadway surface at the time of the crash.
ROAD_DEFECT	Text	Any defects present in the roadway at the time of the crash.
REPORT_TYPE	Text	Type of report filed for the crash.
CRASH_TYPE	Text	Classification of the crash type.
INTERSECTION_RELATED_I	Text	Indicates if the crash is related to an intersection ('Y' or 'N').
NOT_RIGHT_OF_WAY_I	Text	Indicates if failure to yield right-of-way was a factor ('Y' or 'N').
HIT_AND_RUN_I	Text	Indicates if the crash was a hit-and-run incident ('Y' or 'N').
DAMAGE	Text	Extent of damage resulting from the crash.
DATE_POLICE_NOTIFIED	Date & Time	Date and time when the police were notified about the crash.
PRIM_CONTRIBUTORY_CAUSE	Text	Primary cause contributing to the crash.
SEC_CONTRIBUTORY_CAUSE	Text	Secondary cause contributing to the crash.
STREET_NO	Number	Street number of the crash location.
STREET_DIRECTION	Text	Street direction (e.g., N, S, E, W) of the crash location.
STREET_NAME	Text	Street name of the crash location.
BEAT_OF_OCCURRENCE	Text	Police beat where the crash occurred.
PHOTOS_TAKEN_I	Text	Indicates if photos were taken at the crash scene ('Y' or 'N').

STATEMENTS_TAKEN_I	Text	Indicates if statements were taken at the crash scene ('Y' or 'N').
DOORING_I	Text	Indicates if the crash involved dooring ('Y' or 'N').
WORK_ZONE_I	Text	Indicates if the crash occurred in a work zone ('Y' or 'N').
WORK_ZONE_TYPE	Text	Type of work zone where the crash occurred.
WORKERS_PRESENT_I	Text	Indicates if workers were present in the work zone ('Y' or 'N').
NUM_UNITS	Number	Number of units (vehicles, pedestrians, etc.) involved in the crash.
MOST_SEVERE_INJURY	Text	Most severe injury reported in the crash.
INJURIES_TOTAL	Number	Total number of injuries reported.
INJURIES_FATAL	Number	Number of fatal injuries reported.
INJURIES_INCAPACITATING	Number	Number of incapacitating injuries reported.
INJURIES_NON_INCAPACITATING	Number	Number of non-incapacitating injuries reported.
INJURIES_REPORTED_NOT_EVIDENT	Number	Number of reported injuries with no evident injury.
INJURIES_NO_INDICATION	Number	Number of individuals with no indication of injury.
INJURIES_UNKNOWN	Number	Number of injuries with unknown status.
CRASH_HOUR	Number	Hour of the day when the crash occurred.
CRASH_DAY_OF_WEEK	Number	Day of the week when the crash occurred.
CRASH_MONTH	Number	Month when the crash occurred.

LATITUDE	Number	Latitude coordinate of the crash location.
LONGITUDE	Number	Longitude coordinate of the crash location.
LOCATION	Location	Combined latitude and longitude of the crash location.

Data Quality Analysis (CHICAGO)

By the reference of the 5Cs of data

Measure	**Importance**	**Required Insights**
Clean	Ensures that data is free from errors, irrelevant entries, and is formatted correctly.	Check for and remove null values in critical fields like CRASH_RECORD_ID and LOCATION.
Consistent	Verifies that data is logically coherent with uniformity across datasets.	Ensure STREET_NAME and STREET_DIRECTION follow consistent naming conventions.
Comprehensive	Assesses the extent to which data covers all necessary aspects and elements.	Confirm all injury and fatality-related fields (e.g., INJURIES_TOTAL, INJURIES_FATAL) are populated.
Confirmed	Validates that data is accurate and verified against reliable sources.	Cross-reference LATITUDE and LONGITUDE with Chicago mapping services to ensure accuracy.
Current	Confirms that the dataset is up-to-date and relevant for the intended analysis.	Verify that CRASH_DATE reflects recent crash data reported in Chicago.

Field Analysis

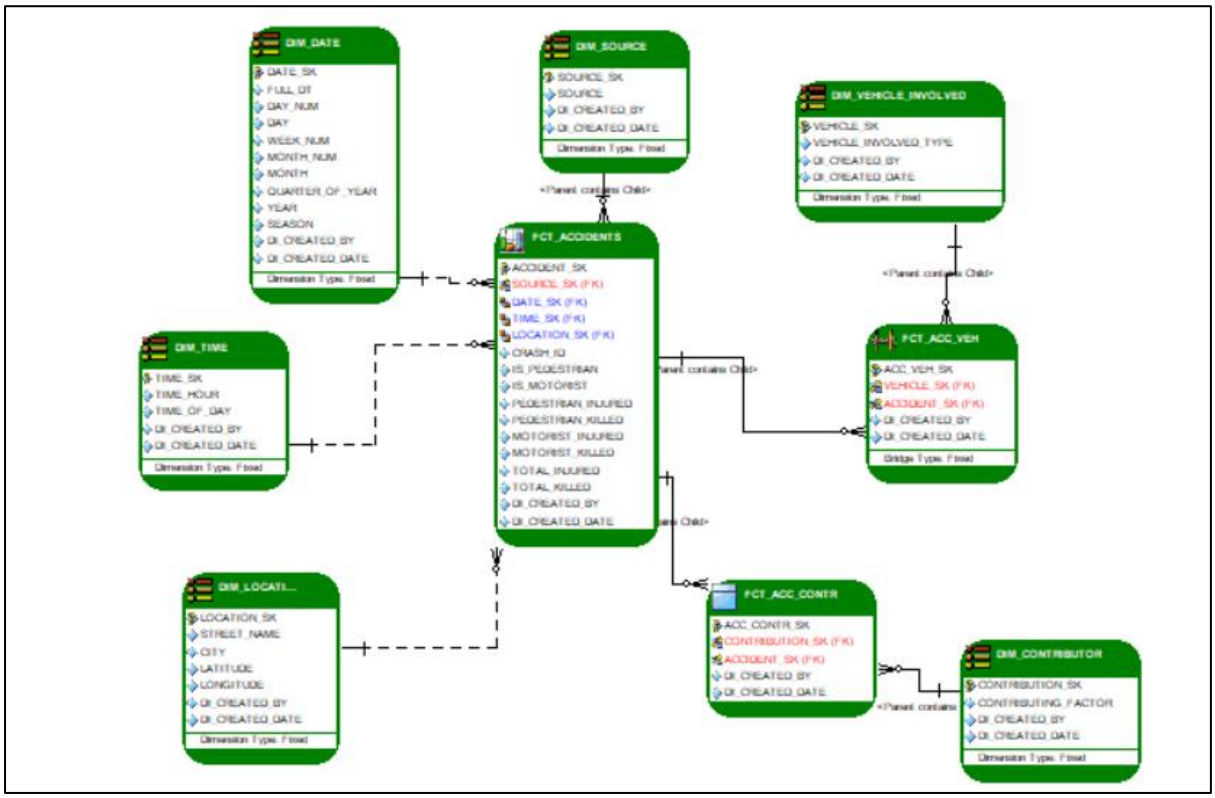
Field	**Description**	**Analysis**
CRASH_RECORD_ID	Unique identifier for each crash record.	Ensure uniqueness and no null values to maintain data integrity.
CRASH_DATE_EST_I	Indicates if the crash date is estimated ('Y' or 'N').	Check for valid binary values ('Y' or 'N') and assess the frequency of estimated dates.
CRASH_DATE	Date and time when the crash occurred.	Validate the date format and ensure consistency with other time-related fields like CRASH_HOUR.

POSTED_SPEED_LIMIT	Speed limit posted at the crash location.	Identify outliers (e.g., unrealistic speed limits) and ensure alignment with Chicago's traffic regulations.
TRAFFIC_CONTROL_DEVICE	Type of traffic control device present at the crash location.	Validate against a predefined list of devices (e.g., stop sign, signal).
DEVICE_CONDITION	Condition of the traffic control device.	Check for valid entries (e.g., functional, not functional) and investigate null values.
WEATHER_CONDITION	Weather conditions at the time of the crash.	Standardize categorical values and address missing data.
LIGHTING_CONDITION	Lighting conditions at the time of the crash.	Ensure valid categories such as daylight, dark, dawn, or dusk.
FIRST_CRASH_TYPE	Initial type of collision in the crash sequence.	Check for logical consistency with other crash details, such as location or severity.
TRAFFICWAY_TYPE	Layout or type of the trafficway where the crash occurred.	Validate categorical values for accuracy and relevance (e.g., one-way, divided).
LANE_CNT	Number of lanes in the roadway at the crash location.	Identify anomalies such as unusually high or missing lane counts.
ALIGNMENT	Roadway alignment (e.g., straight, curve) at the crash location.	Cross-validate with crash type for logical consistency (e.g., sharp curves and single-vehicle crashes).
ROADWAY_SURFACE_COND	Condition of the roadway surface at the time of the crash.	Ensure completeness and validity of surface conditions (e.g., dry, wet, icy).
ROAD_DEFECT	Any defects present in the roadway at the time of the crash.	Investigate non-standard or null values for accuracy.
REPORT_TYPE	Type of report filed for the crash.	Validate against predefined report categories (e.g., driver, officer report).
CRASH_TYPE	Classification of the crash type.	Ensure logical consistency between crash type and

		other factors such as TRAFFICWAY_TYPE.
INTERSECTION_RELATED_I	Indicates if the crash is related to an intersection ('Y' or 'N').	Validate binary entries and cross-reference with location data.
NOT_RIGHT_OF_WAY_I	Indicates if failure to yield right-of-way was a factor ('Y' or 'N').	Ensure consistency with contributory causes (PRIM_CONTRIBUTORY_CAUSE) .
HIT_AND_RUN_I	Indicates if the crash was a hit-and-run incident ('Y' or 'N').	Confirm accuracy and completeness of values, especially in severe crash cases.
DAMAGE	Extent of damage resulting from the crash.	Standardize descriptions of damage and ensure consistency with injury severity.
DATE_POLICE_NOTIFIED	Date and time when the police were notified about the crash.	Check for timely reporting compared to the crash timestamp.
PRIM_CONTRIBUTORY_CAUSE	Primary cause contributing to the crash.	Validate against a predefined list of causes and investigate null values.
SEC_CONTRIBUTORY_CAUSE	Secondary cause contributing to the crash.	Ensure logical alignment with the primary cause.
STREET_NO	Street number of the crash location.	Validate against STREET_NAME and STREET_DIRECTION for consistency.
STREET_DIRECTION	Direction of the street where the crash occurred (e.g., N, S, E, W).	Ensure values match valid street directions for Chicago.
STREET_NAME	Street name of the crash location.	Cross-validate with GIS data for accuracy.
BEAT_OF_OCCURRENCE	Police beat where the crash occurred.	Ensure that values correspond to valid police beats in Chicago.
PHOTOS_TAKEN_I	Indicates if photos were taken at the crash scene ('Y' or 'N').	Validate binary values and assess completeness of this data.
STATEMENTS_TAKEN_I	Indicates if statements were taken at the crash scene ('Y' or 'N').	Ensure consistency with other investigative details (e.g., HIT_AND_RUN_I).

DOORING_I	Indicates if the crash involved dooring ('Y' or 'N').	Confirm binary entries and investigate their frequency.
WORK_ZONE_I	Indicates if the crash occurred in a work zone ('Y' or 'N').	Validate against related fields like WORK_ZONE_TYPE and WORKERS_PRESENT_I.
WORK_ZONE_TYPE	Type of work zone where the crash occurred.	Standardize categories (e.g., construction, maintenance).
WORKERS_PRESENT_I	Indicates if workers were present in the work zone ('Y' or 'N').	Ensure logical consistency with WORK_ZONE_TYPE.
NUM_UNITS	Number of units (vehicles, pedestrians, etc.) involved in the crash.	Validate against injury and damage fields to ensure consistency.
MOST_SEVERE_INJURY	Most severe injury reported in the crash.	Confirm alignment with individual injury counts (e.g., fatal, incapacitating).
INJURIES_TOTAL	Total number of injuries reported.	Validate that this aligns with the sum of all individual injury types.
INJURIES_FATAL	Number of fatal injuries reported.	Ensure consistency with severity and contributory causes.
LATITUDE	Latitude coordinate of the crash location.	Validate against Chicago's geographic boundaries.
LONGITUDE	Longitude coordinate of the crash location.	Cross-validate with GIS tools for accuracy.
LOCATION	Combined latitude and longitude of the crash location.	Ensure this field accurately reflects the LATITUDE and LONGITUDE fields.

Dimensional Model



The provided diagram represents a dimensional data model designed for analyzing motor vehicle accident data. Below is a detailed explanation of each component and its relationships, suitable for inclusion in your project documentation:

1. **Central Fact Table: FCT_ACCIDENTS**
   The **FCT_ACCIDENTS** table is the core of the model, storing the measurable data points related to motor vehicle accidents. It is linked to multiple dimensions to support detailed analysis.
   Attributes of FCT_ACCIDENTS:
   - **Keys:**
     - ACCIDENT_SK: A surrogate key uniquely identifying each accident.
     - Foreign keys (DATE_SK, TIME_SK, LOCATION_SK, SOURCE_SK) link to dimension tables.
   - **Metrics:**
     - PEDESTRIAN_INJURED, PEDESTRIAN_KILLED: Total pedestrians injured or killed.
     - MOTORIST_INJURED, MOTORIST_KILLED: Total motorists injured or killed.
     - CYCLIST_INJURED, CYCLIST_KILLED: Total cyclists injured or killed.
     - TOTAL_INJURED, TOTAL_KILLED: Overall totals for injuries and fatalities.

   The fact table aggregates accident data and connects to supporting dimensions for detailed analysis.

2. **Dimension Tables**
   a) **DIM_DATE**

- **Purpose:** Provides date-related details for accidents.
- **Attributes:**
    - DATE_SK: Unique identifier for each date.
    - DAY, MONTH, YEAR, SEASON: Attributes for temporal analysis, such as accident trends over time or seasonality.

b) DIM_TIME

- **Purpose:** Stores information about the time of accidents.
- **Attributes:**
    - TIME_SK: Unique identifier for each time.
    - TIME_HOUR, TIME_OF_DAY: Enables analysis based on the hour of the day or whether accidents happen more often during specific time periods (e.g., rush hours).

c) DIM_LOCATION

- **Purpose:** Captures geographical details about where accidents occur.
- **Attributes:**
    - LOCATION_SK: Unique identifier for each location.
    - STREET_NAME, CITY, LATITUDE, LONGITUDE: Helps in spatial analysis, such as identifying high-risk areas or mapping accidents geographically.

d) DIM_SOURCE

- **Purpose:** Tracks the origin of accident data.
- **Attributes:**
    - SOURCE_SK: Unique identifier for the data source.
    - SOURCE: Provides metadata about where the data came from (e.g., NYC, Chicago).
- **Usage:** Ensures traceability of the data and supports multi-city accident analysis.

e) DIM_VEHICLE_INVOLVED

- **Purpose:** Stores details about vehicles involved in accidents.
- **Attributes:**
    - VEHICLE_SK: Unique identifier for vehicle types.
    - VEHICLE_INVOLVED_TYPE: Describes the type of vehicle involved (e.g., sedan, truck, motorcycle).

This dimension is used to analyze accidents involving specific vehicle types.

f) DIM_CONTRIBUTOR

- **Purpose:** Identifies factors contributing to accidents.
- **Attributes:**
    - CONTRIBUTION_SK: Unique identifier for each contributing factor.
    - CONTRIBUTING_FACTOR: Describes the reason or cause of the accident (e.g., speeding, adverse weather).

This dimension helps in root cause analysis of accidents.

3. Bridge Tables

a) FCT_ACC_VEH

- **Purpose:** Resolves the many-to-many relationship between accidents and vehicles.
- **Attributes:**
  - ACCIDENT_SK: Foreign key linking to the FCT_ACCIDENTS table.
  - VEHICLE_SK: Foreign key linking to the DIM_VEHICLE_INVOLVED table.
- **Usage:** Tracks multiple vehicles involved in a single accident.

b) FCT_ACC_CONTR

- **Purpose:** Resolves the many-to-many relationship between accidents and contributing factors.
- **Attributes:**
  - ACCIDENT_SK: Foreign key linking to the FCT_ACCIDENTS table.
  - CONTRIBUTION_SK: Foreign key linking to the DIM_CONTRIBUTOR table.
- **Usage:** Tracks multiple causes contributing to a single accident.