

DAMG7370 – FINAL PROJECT

Motor Vehicle collisions/Crashes

DATASETS:

AUSTIN	NYC	CHICAGO	MONTGOMERY
ID	CRASH_DATE	CRASH_RECORD_ID	REPORT_NUMBER
CRASH_ID	CRASH_TIME	CRASH_DATE_EST_I	LOCAL_CASE_NUMB ER
CRASH_FATAL_FL	BOROUGH	CRASH_DATE	AGENCY_NAME
CASE_ID	ZIP_CODE	POSTED_SPEED_LIM IT	ACRS_REPORT_TYPE
PRIMARY_ADDRESS	LATITUDE	TRAFFIC_CONTROL_ DEVICE	CRASH_DATE_TIME
RPT_BLOCK_NUM	LONGITUDE	DEVICE_CONDITION	HIT_RUN
RPT_STREET_NAME	LOCATION	WEATHER_CONDITI ON	ROUTE_TYPE
RPT_STREET_SFX	ON_STREET_NAME	LIGHTING_CONDITI ON	LANE_DIRECTION
CRASH_SPEED_LIMIT	CROSS_STREET_NAM E	FIRST_CRASH_TYPE	LANE_TYPE
ROAD_CONSTR_ZONE_F L	OFF_STREET_NAME	TRAFFICWAY_TYPE	NUMBER_OF_LANES
LATITUDE	NUMBER_OF_PERSO NS_INJURED	LANE_CNT	DIRECTION
LONGITUDE	NUMBER_OF_PERSO NS_KILLED	ALIGNMENT	DISTANCE
CRASH_SEV_ID	NUMBER_OF_PEDEST RIANS_INJURED	ROADWAY_SURFAC E_COND	DISTANCE_UNIT
SUS_SERIOUS_INJRY_CN T	NUMBER_OF_PEDEST RIANS_KILLED	ROAD_DEFECT	ROAD_GRADE
POSS_INJRY_CNT	NUMBER_OF_CYCLIS T_INJURED	REPORT_TYPE	ROAD_NAME
NON_INJRY_CNT	NUMBER_OF_CYCLIS T_KILLED	CRASH_TYPE	CROSS_STREET_NAM E
UNKN_INJRY_CNT	NUMBER_OF_MOTOR IST_INJURED	INTERSECTION_REL ATED_I	OFF_ROAD_DESCRIP TION
TOT_INJRY_CNT	NUMBER_OF_MOTOR IST_KILLED	NOT_RIGHT_OF_WA Y_I	MUNICIPALITY
DEATH_CNT	CONTRIBUTING_FAC TOR_VEHICLE_1	HIT_AND_RUN_I	RELATED_NON_MOT ORIST
UNITS_INVOLVED	CONTRIBUTING_FAC TOR_VEHICLE_2	DAMAGE	AT_FAULT
POINT	CONTRIBUTING_FAC TOR_VEHICLE_3	DATE_POLICE_NOTI FIED	COLLISION_TYPE
MOTOR_VEHICLE_DEATH _COUNT	CONTRIBUTING_FAC TOR_VEHICLE_4	PRIM_CONTRIBUTO RY_CAUSE	WEATHER
MOTOR_VEHICLE_SERIO US_INJURY_COUNT	CONTRIBUTING_FAC TOR_VEHICLE_5	SEC_CONTRIBUTOR Y_CAUSE	SURFACE_CONDITIO N
BICYCLE_DEATH_COUNT	COLLISION_ID	STREET_NO	LIGHT

BICYCLE_SERIOUS_INJURY_COUNT	VEHICLE_TYPE_CODE_1	STREET_DIRECTION	TRAFFIC_CONTROL
MOTORCYCLE_DEATH_COUNT	VEHICLE_TYPE_CODE_2	STREET_NAME	DRIVER_SUBSTANCE_ABUSE
OTHER_DEATH_COUNT	VEHICLE_TYPE_CODE_3	BEAT_OF_OCCURRENCE	NON_MOTORIST_SUBSTANCE_ABUSE
OTHER_SERIOUS_INJURY_COUNT	VEHICLE_TYPE_CODE_4	PHOTOS_TAKEN_I	FIRST_HARMFUL_EVENT
ONSYS_FL	VEHICLE_TYPE_CODE_5	STATEMENTS_TAKEN_I	SECOND_HARMFUL_EVENT
PRIVATE_DR_FL		DOORING_I	JUNCTION
MICROMOBILITY_SERIOUS_INJURY_COUNT		WORK_ZONE_I	INTERSECTION_TYPE
MIRCOMOBILITY_DEATH_COUNT		WORK_ZONE_TYPE	ROAD_ALIGNMENT
CRASH_TIMESTAMP_US_CENTRAL		WORKERS_PRESENT_I	ROAD_CONDITION
CRASH_TIMESTAMP		NUM_UNITS	ROAD_DIVISION
IS_DELETED		MOST_SEVERE_INJURY	LATITUDE
IS_TEMPORARY_RECORD		INJURIES_TOTAL	LONGITUDE
LAW_ENFORCEMENT_FATALITY_COUNT		INJURIES_FATAL	LOCATION
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 analyze 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](#)
 - [Montgomery Data Portal](#)

PART 1: DATA PREPARATION

- **Data Profiling:** Using Alteryx

- **Overview of all 4 datasets:**

The NYC, Austin, Chicago, and Montgomery datasets collectively provide comprehensive insights into vehicular crashes, emphasizing geographic locations, contributing factors, and crash outcomes such as injuries and fatalities. Each dataset captures unique regional attributes: NYC focuses on borough-specific crash data with detailed contributing factors and injuries; Austin highlights micromobility incidents, construction zones, and injury severities; Chicago provides environmental and roadway-specific details like lane counts, alignment, and surface conditions; and Montgomery emphasizes collision types, harmful events, and traffic control data. Together, these datasets enable trend analysis, geospatial mapping, and root cause assessment, offering a rich foundation for improving road safety, urban planning, and policy-making through consistent, clean, comprehensive, and validated data.

- **Field Analysis**

Common Fields Across Datasets

- **Date and Time:** All datasets include fields for the date and time of crashes. Ensuring proper formats and non-null values is critical for temporal analysis.
- **Geographic Information:** Latitude, longitude, and location-related fields provide spatial context. These fields need validation against mapping services for accuracy.
- **Injuries and Fatalities:** Fields like NUMBER_OF_PERSONS_INJURED (NYC), TOT_INJRY_CNT (Austin), INJURIES_TOTAL (Chicago), and similar fields in Montgomery highlight crash severity.
- **Crash Types and Contributing Factors:** Fields like CRASH_TYPE, FIRST_CRASH_TYPE, and CONTRIBUTING_FACTOR_VEHICLE_1 detail the nature and cause of crashes, which are essential for root cause analysis.

Unique Fields

- **Austin:** ROAD_CONST_ZONE_FL and MICROMOBILITY_SERIOUS_INJURY_COUNT reflect construction zones and micromobility incidents.
- **NYC:** BOROUGH and ZIP_CODE provide administrative location details specific to New York City.
- **Chicago:** LANE_CNT, ALIGNMENT, and ROADWAY_SURFACE_COND provide insights into roadway conditions and alignments.
- **Montgomery:** TRAFFIC CONTROL, FIRST HARMFUL EVENT, and INTERSECTION TYPE offer detailed descriptions of the crash setting.

- **Data Quality Analysis**

1. **Clean**

NYC: Ensure no null values in critical fields like COLLISION_ID, LOCATION, and LATITUDE.

Austin: Remove null values in fields such as CRASH_ID, LATITUDE, and LONGITUDE.

Chicago: Validate non-null values for CRASH_RECORD_ID and LOCATION.

Montgomery: Remove invalid or null entries in fields like REPORT NUMBER, CRASH DATE/TIME, and LOCATION.

2. **Consistent**

NYC: Ensure ZIP codes align with boroughs and validate street names against NYC mapping.

Austin: Check latitude and longitude fall within Austin's geographic boundaries.

Chicago: Validate STREET_NAME and STREET_DIRECTION for consistency with Chicago naming conventions.

Montgomery: Ensure consistent formats for ROAD NAME, CROSS-STREET NAME, and LANE DIRECTION.

3. **Comprehensive**

NYC: Confirm injury and fatality fields are fully populated (e.g., NUMBER_OF_PERSONS_INJURED).

Austin: Validate completeness in fields such as SUS_SERIOUS_INJRY_CNT and DEATH_CNT.

Chicago: Ensure all injury-related fields, including INJURIES_TOTAL and INJURIES_FATAL, are populated.

Montgomery: Validate completeness for fields like WEATHER, SURFACE CONDITION, and COLLISION TYPE.

4. **Confirmed**

NYC: Cross-reference LATITUDE and LONGITUDE with NYC mapping services.

Austin: Validate geographic coordinates against official Austin maps.

Chicago: Use GIS tools to confirm LATITUDE and LONGITUDE values fall within Chicago's boundaries.

Montgomery: Verify geographic data against reliable mapping sources.

5. **Current**

NYC: Ensure CRASH_DATE reflects the most recent data.

Austin: Confirm CRASH_TIMESTAMP_US_CENTRAL includes the latest incidents.

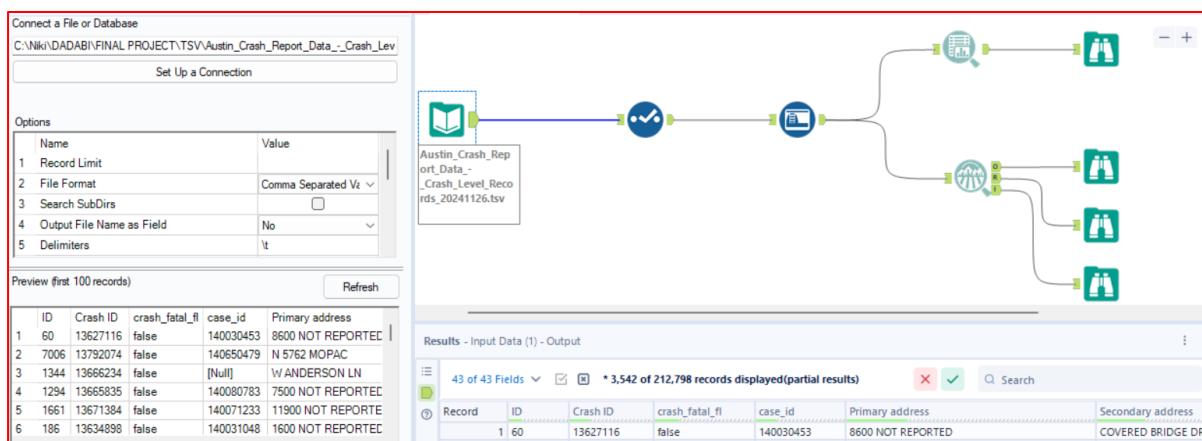
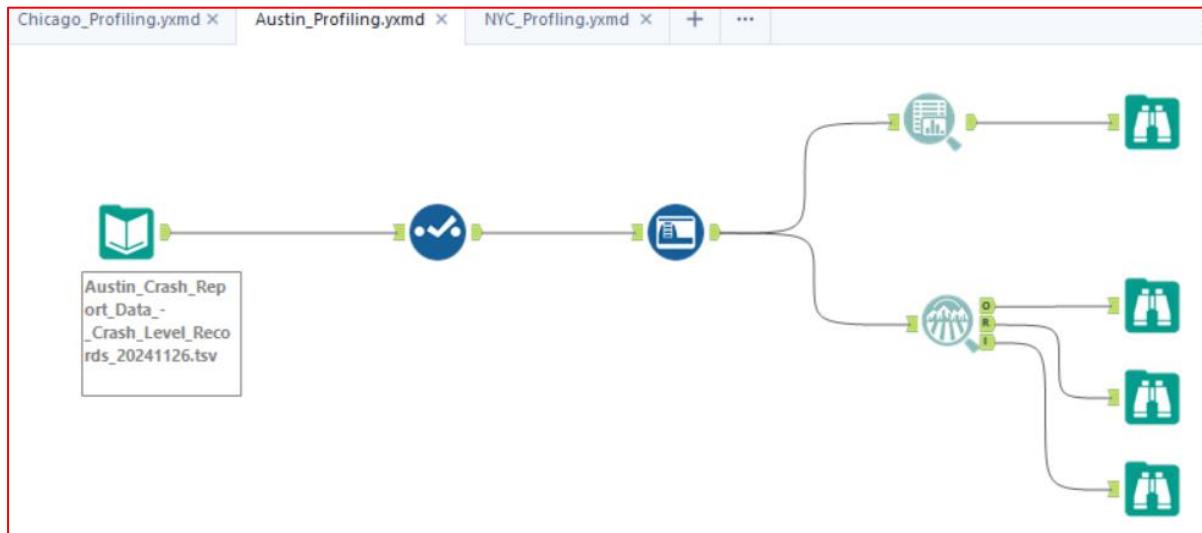
Chicago: Verify CRASH_DATE is up-to-date.

Montgomery: Validate that CRASH DATE/TIME covers recent crash data and remove outdated records.

Consolidated Use Cases

- **Trend Analysis:** Analyze temporal patterns using date and time fields.
- **Geospatial Analysis:** Map crash locations and identify high-risk areas.
- **Severity Assessment:** Evaluate crash outcomes using injury and fatality fields.
- **Root Cause Analysis:** Investigate contributing factors and crash types to identify safety improvement opportunities.

AUSTIN DATASET



Overview of the dataset

The Austin dataset records crash data from the Vision Zero crash database. It contains details on crash locations, injury severity, fatalities, and vehicle types involved. Notable fields include unique identifiers, speed limits, road construction zones, and timestamps. This dataset aids in urban planning, assessing road safety, and identifying crash patterns.

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).

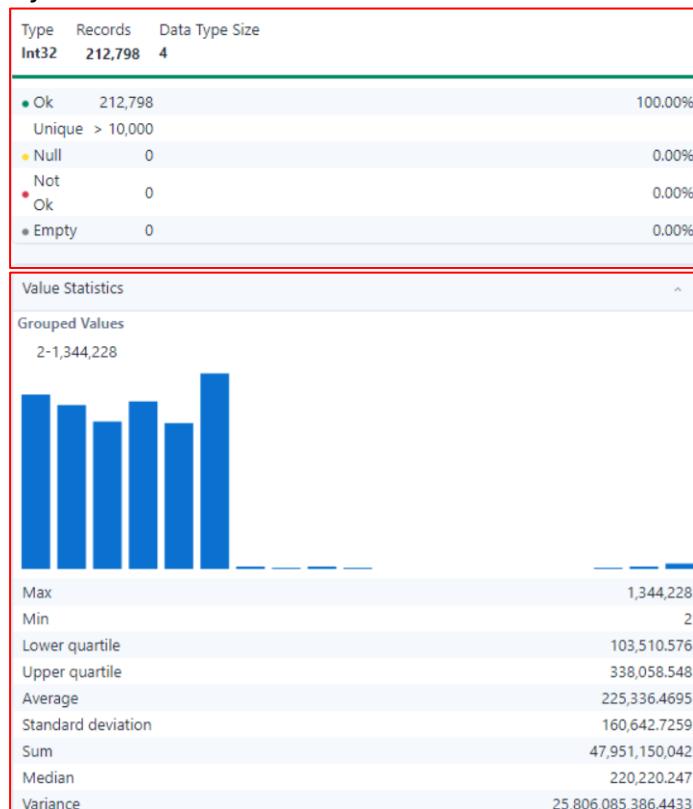
UNITS_INVOLVED	Number of units (vehicles, pedestrians, etc.) involved in the crash.	Validate against individual records of injuries and fatalities for consistency.
POINT	Geographical point representing the crash location.	Ensure POINT corresponds to the LATITUDE and LONGITUDE fields.
MOTOR_VEHICLE_DEATH_COUNT	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 involving fatalities of 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.

Data Observation:

- **ID**

- Summary:

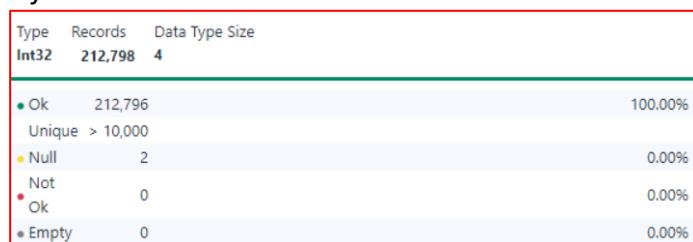


- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **CRASH_ID**

- Summary:

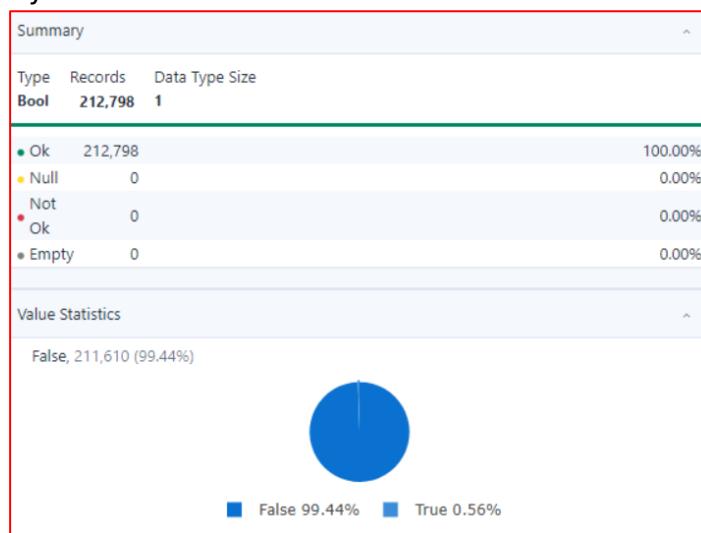




- Observation:
 - Out of 212,798 records, 212,796 are valid (Ok), with 2 null values detected, accounting for 0.00% nullity, which slightly affects data reliability.
 - The field contains over 10,000 unique values, reflecting a diverse and meaningful dataset.
 - No "not-ok" or empty records exist, ensuring most of the data is clean and ready for analysis.

● CRASH_FATAL_FL

- Summary:



- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data integrity and completeness.
 - The data type is Bool with a data type size of 1, confirming its binary nature.
 - The field contains **99.44% False values (211,610 records)** and **0.56% True values (1,188 records)**, reflecting a significant imbalance.

- No missing, not-ok, or invalid entries exist, ensuring the field is clean and ready for analysis.
 - The pie chart visualization highlights the overwhelming majority of False values, suggesting limited variability in this field.
- **CASE_ID**
- Summary:

Summary ▲		
Type	Records	Data Type Size
V_String	212,798	20
● Ok	209,893	98.63%
Unique	> 10,000	
■ Null	2,905	1.37%
Not	0	0.00%
● Ok	0	0.00%
● Empty	0	0.00%
Length Statistics		
Min		1
Max		20
Average		9.10
Shortest Value		1
Longest Value		10000 BLK US HIGHWAY
First Alphanumeric Value		#11-197-1491
Last Alphanumeric Value		tg1243104211602
Blanks		0
Values with Leading Whitespace		0
Values with Trailing Whitespace		0

- Observation:
- Out of 212,798 records, 209,893 (98.63%) are valid, while 2,905 (1.37%) are null, which should be addressed to ensure data completeness.
 - The field contains over 10,000 unique values, demonstrating a diverse and rich dataset.
 - No leading or trailing whitespace was detected, ensuring clean data formatting.
 - No blank values are present, enhancing the field's consistency.

- **PRIMARY_ADDRESS**

- Summary:

Summary ▲			
Type	Records	Data Type	Size
V_String	212,798	54	
● Ok	212,798		100.00%
Unique > 10,000			
■ Null	0		0.00%
Not	0		0.00%
● Ok	0		0.00%
■ Empty	0		0.00%
Length Statistics			
Min	2		
Max	54		
Average	18.20		
Shortest Value	LN		
Longest Value	6900 ED BLUESTEIN BLVD SB TO ED BLUESTEIN BLVD SVRD SB		
First Alphanumeric Value	O O E BEN WHITE TO IH 35 NB RAMP		
Last Alphanumeric Value	ZUNIGA DR DR		
Blanks	0		
Values with Leading Whitespace	0		
Values with Trailing Whitespace	0		

- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability.
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries.
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **SECONDARY_ADDRESS**

- Summary:

Summary ▲			
Type	Records	Data Type	Size
V_String	212,798	64	
● Ok	212,798		100.00%
Unique > 10,000			
■ Null	0		0.00%
Not	0		0.00%
● Ok	0		0.00%
■ Empty	0		0.00%
Length Statistics			
Min	1		
Max	64		
Average	15.90		
Shortest Value	1		
Longest Value	7200 NE CURBLINE OF THE N IH 35 NB SVRD TURNAROUND T...		
First Alphanumeric Value	O 15TH ST		
Last Alphanumeric Value	ZUNIGA DR DR		
Blanks	0		
Values with Leading Whitespace	0		
Values with Trailing Whitespace	0		

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability.
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries.
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.
- **RPT_BLOCK_NUM**
 - Summary:

Summary			
Type	Records	Data Type	Size
V_String	212,798	10	
● Ok	184,176		86.55%
Unique	5,819		2.73%
● Null	28,622		13.45%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			1
Max			10
Average			4.00
Shortest Value			0
Longest Value			9400-11300
First Alphanumeric Value			0
Last Alphanumeric Value			W BEN WHIT
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - Out of 212,798 records, 184,176 (86.55%) are valid, while 28,622 (13.45%) are null, requiring attention to improve data completeness.
 - The field contains 5,819 unique values, reflecting moderate diversity.
 - No missing, not-ok, or empty records exist among the valid entries, ensuring a large portion of the data is usable.

- **RPT_STREET_NAME**

- Summary:

Summary ▲			
Type	Records	Data Type	Size
V_String	212,798	52	
● Ok	212,798		100.00%
Unique	> 10,000		
■ Null	0		0.00%
■ Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics ▲			
Min		1	
Max		52	
Average		10.20	
Shortest Value		1	
Longest Value	MOPAC SOUTH BOUND SERVICE TO NORTH BOUND TURN AR...		
First Alphanumeric Value	0 183 TOLL SB		
Last Alphanumeric Value	ZUNIGA DR		
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

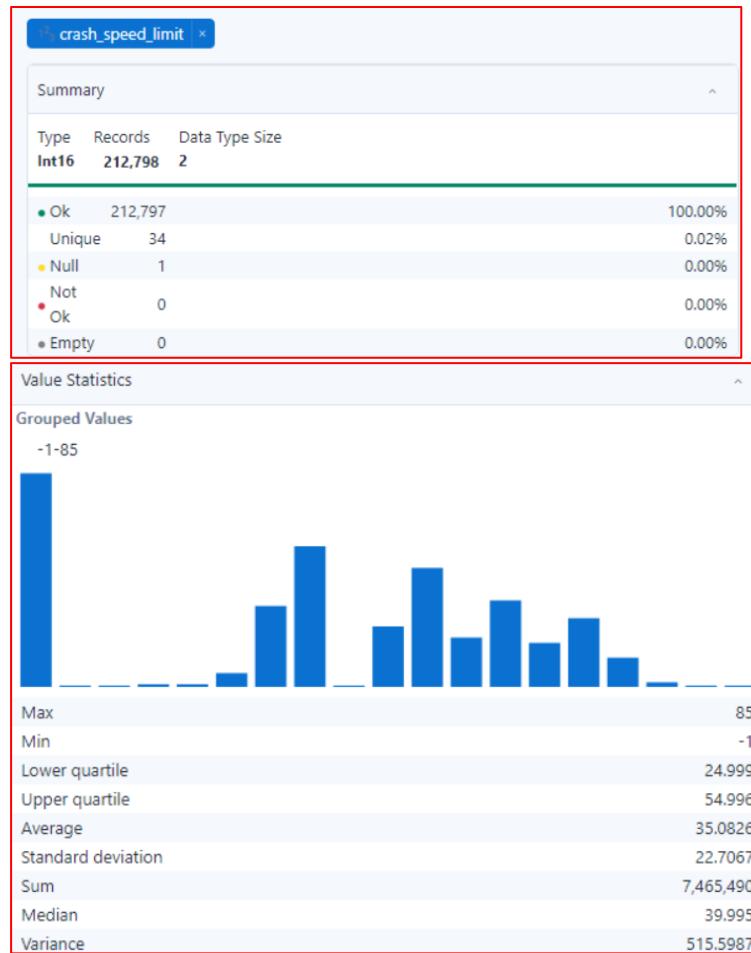
- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability.
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries.
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **RPT_STREET_SFX**

- Summary:

Type	Records	Data Type	Size
String	212,798	4	
● Ok	147,615		69.37%
Unique	20		0.01%
■ Null	65,182		30.63%
■ Not Ok	0		0.00%
● Empty	1		0.00%
Length Statistics ▲			
Min		2	
Max		4	
Average		2.70	
Shortest Value		LN	
Longest Value		EXPY	
First Alphanumeric Value		[Null]	
Last Alphanumeric Value		WAY	
Blanks		1	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

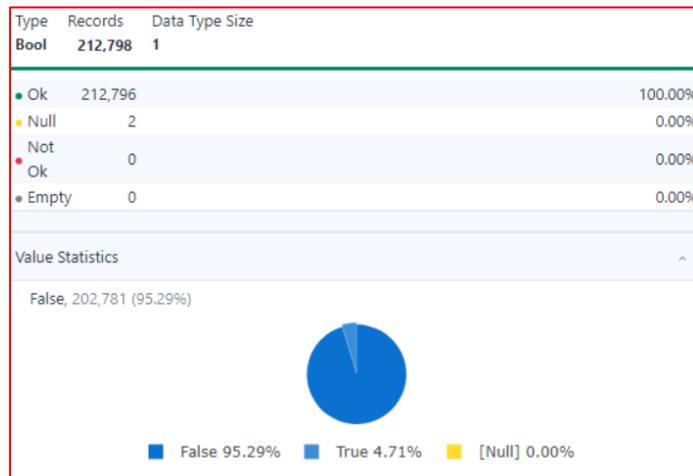
- Observation:
 - Out of 212,798 records, 147,615 (69.37%) are valid, while 65,182 (30.63%) are null, which significantly impacts data completeness and reliability.
 - The field contains only 20 unique values, indicating low variability.
 - No missing or not-ok records exist among the valid entries, ensuring partial data usability
- **CRASH_SPEED_LIMIT**
 - Summary:



- Observation:
 - Out of 212,798 records, 212,797 (100.00%) are valid, with only 1 null value detected, ensuring high data reliability.
 - The field contains 34 unique values, representing a moderately diverse range of crash speed limits.
 - No not-ok or empty records exist, ensuring the field is ready for analysis.

- **ROAD_CONSTR_ZONE_FL**

- Summary:

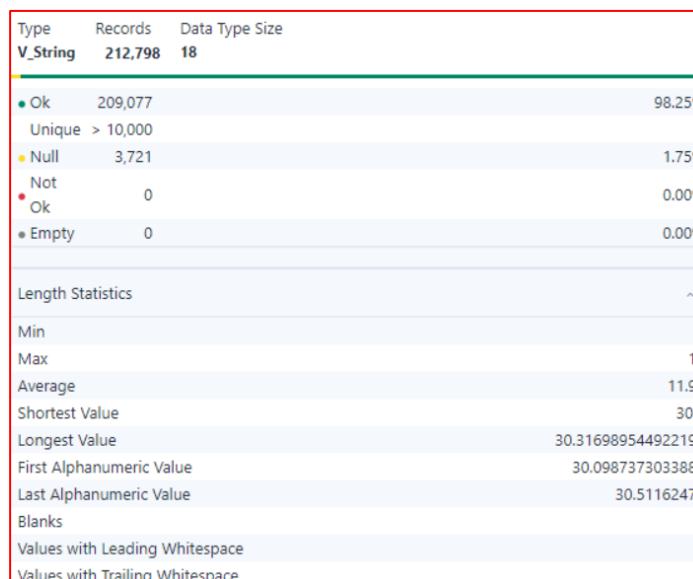


- Observation:

- Out of 212,798 records, 212,796 (100.00%) are valid, with only 2 null values detected, ensuring high data reliability.
 - The field is of Boolean type, with 95.29% of the records (202,781) having a value of False and 4.71% (10,015) having a value of True, indicating a significant imbalance in the distribution.
 - No not-ok or empty records exist, ensuring the field is clean and ready for analysis.
 - The pie chart visualization highlights the dominance of False values, which may suggest a categorical field with a rarely occurring True condition.

- **LATITUDE**

- Summary:



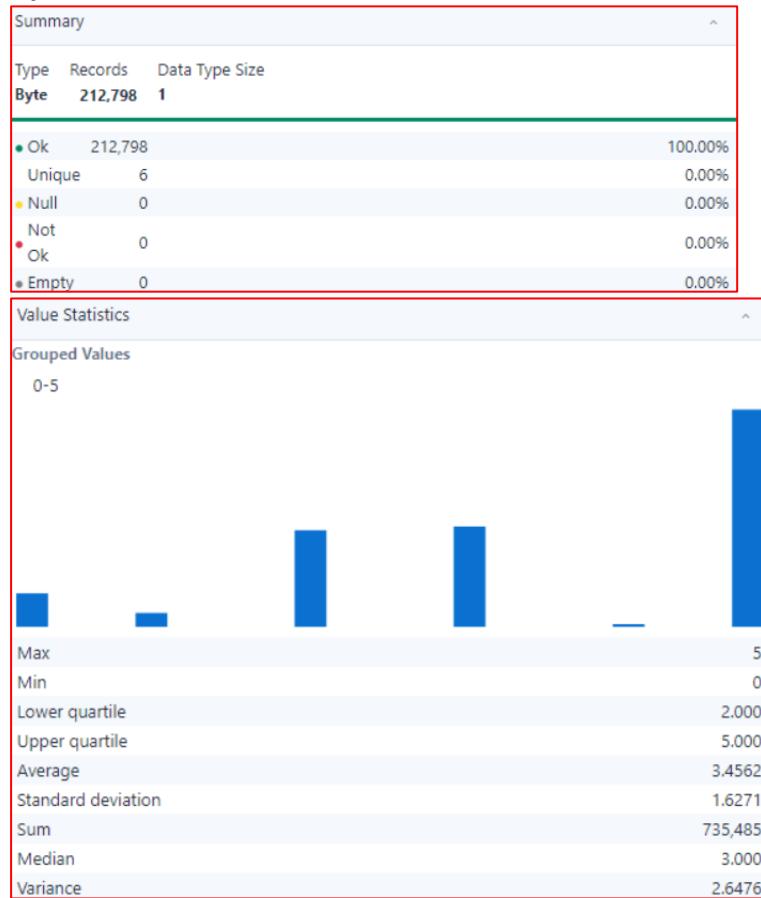
- Observation:
 - Out of 212,798 records, 209,077 (98.25%) are valid, with 3,721 (1.75%) null values detected, requiring attention to improve data completeness.
 - The field contains over 10,000 unique values, indicating a high degree of variability and meaningful data.
 - No not-ok or empty records exist, ensuring the field is mostly clean and ready for analysis.
 - No values contain leading or trailing whitespaces, ensuring clean data formatting.
- **LONGITUDE**
- Summary:

Summary ▲			
Type	Records	Data Type	Size
V_String	212,798	18	
● Ok	209,076		98.25%
Unique	> 10,000		
● Null	3,722		1.75%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics ▲			
Min		6	
Max		18	
Average		12.90	
Shortest Value		-97.66	
Longest Value		-97.70786148611529	
First Alphanumeric Value		-97.57014781	
Last Alphanumeric Value		-97.92717344	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - Out of 212,798 records, 209,076 (98.25%) are valid, while 3,722 (1.75%) null values exist, requiring attention to improve data completeness.
 - The field contains over 10,000 unique values, reflecting high variability and diverse data entries.
 - No not-ok or empty records exist, ensuring the majority of the data is clean and usable.
 - No values contain leading or trailing whitespaces, ensuring clean data formatting.

- **CRASH_SEV_ID**

- Summary:

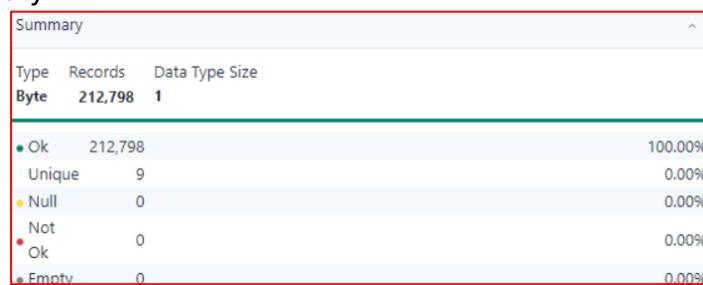


- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability.
 - The field contains 6 unique values ranging from 0 to 5, providing a well-defined categorical dataset.
 - The data distribution shows a median value of 3, with an average of 3.4562, indicating slightly higher occurrences of upper-range values.

- **SUS_SERIOUS_INJRY_CNT**

- Summary:



Value Statistics	
Grouped Values	
0-14	
Max	14
Min	0
Lower quartile	0.000
Upper quartile	0.001
Average	3.4845e-02
Standard deviation	0.2140
Sum	7,415
Median	0.000
Variance	4.5830e-02

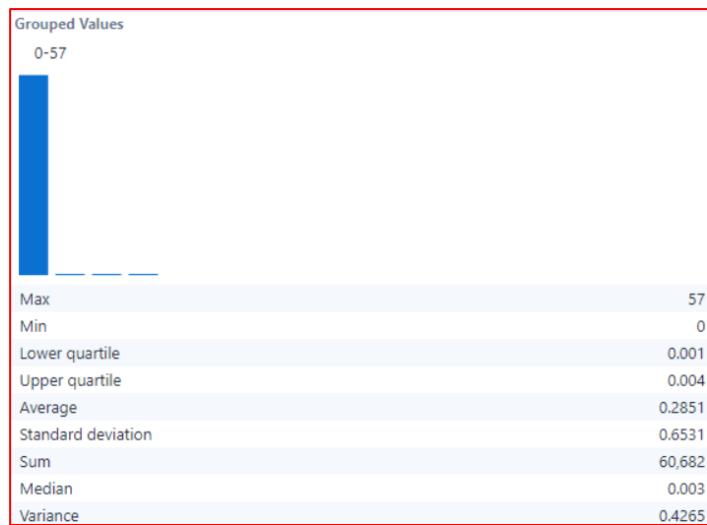
- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability.
- The field contains 9 unique values, all ranging between 0 and 14, indicating a compact and well-defined range.
- The values appear consistent across the dataset, with no indication of formatting or categorization issues.
- If this field represents categorical or ordinal data, the categories are well-defined, but further standardization or binning might improve interpretability.

NONINCAP_INJRY_CNT

- Summary:

Summary			
Type	Records	Data Type	Size
Byte	212,798	1	
<hr/>			
● Ok	212,798		100.00%
Unique	16		0.01%
■ Null	0		0.00%
■ Not	0		0.00%
■ Ok	0		0.00%
■ Empty	0		0.00%

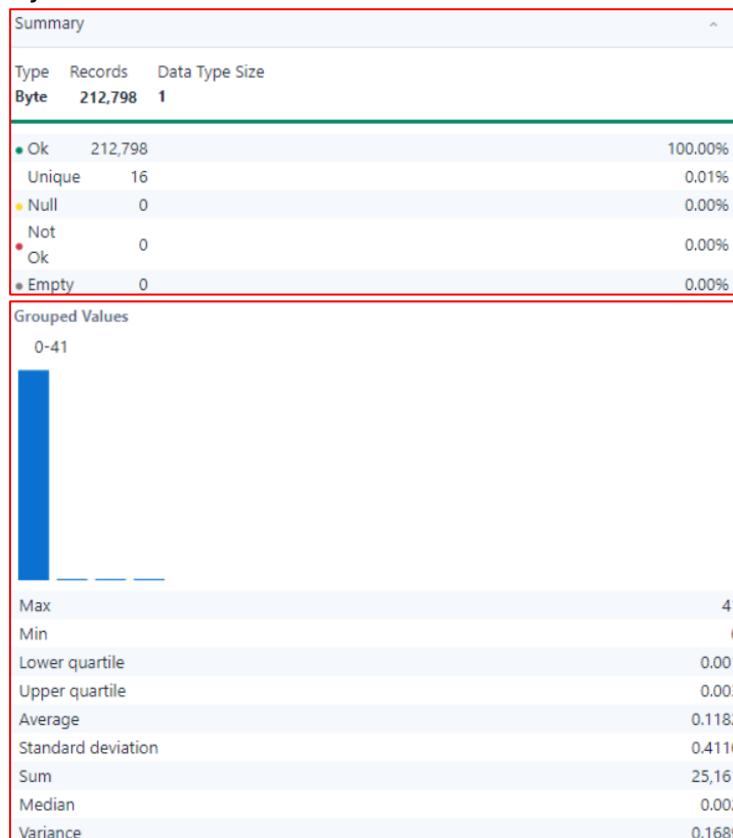


- Observation:

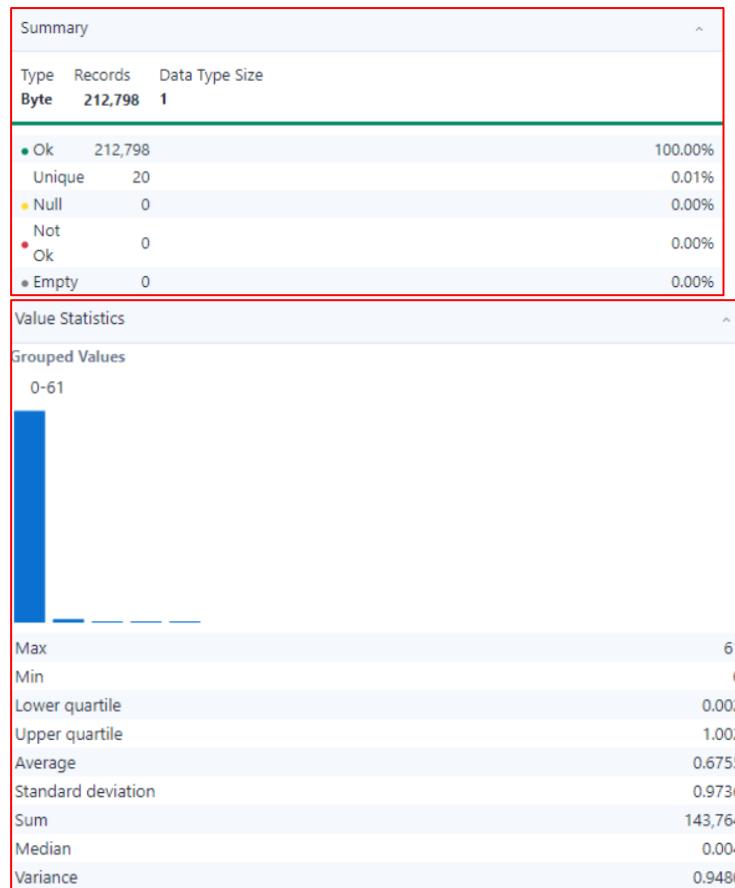
- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring complete data reliability.
- The field contains 16 unique values within the range of 0 to 57, indicating that the data is compact and may represent predefined categories.
- The data appears standardized, as the range is consistent with no formatting or type issues observed.

- **UNKN_INJRY_CNT**

- Summary:



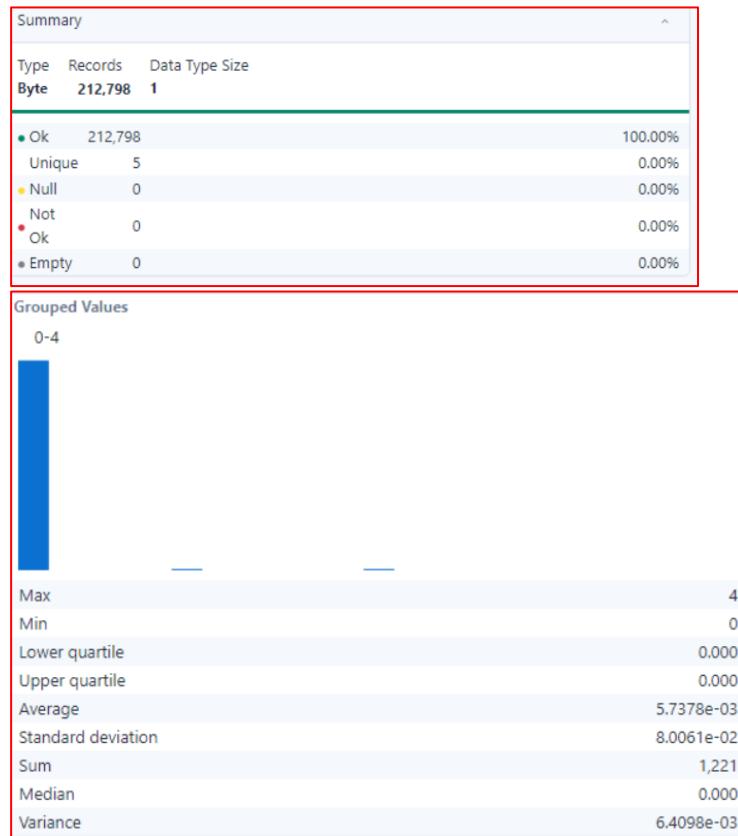
- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring full data reliability.
 - The field contains 16 unique values ranging from 0 to 41, indicating a standardized and compact dataset.
 - The values are consistent with no observed issues in formatting or type, suggesting the data is well-structured.
- **SERIOUS_INJRY_CNT**
 - Summary:



- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring complete data reliability.
 - The field contains 20 unique values within a range of 0 to 61, suggesting a structured and predefined set of categories.
 - The data is consistent and standardized, as no formatting or type-related issues are observed.

- **DEATH_CNT**

- Summary:



- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring full data reliability.
 - The field contains 5 unique values ranging from 0 to 4, reflecting a compact and predefined categorical structure.
 - The data appears consistent and standardized, as no discrepancies or formatting issues are observed.

- **UNITS_INVOLVED**

- Summary:

Summary		
Type	Records	Data Type Size
V_WString	212,798	92
● Ok	212,798	100.00%
Unique	80	0.04%
● Null	0	0.00%
Not	0	0.00%
● Ok	0	0.00%
● Empty	0	0.00%
Length Statistics		
Min		7
Max		92
Average		27.00
Shortest Value		Bicycle
Longest Value		Large passenger vehicle & Motor vehicle – other & Other/Unkn...
First Alphanumeric Value		Bicycle
Last Alphanumeric Value		Pedestrian
Blanks		0
Values with Leading Whitespace		0
Values with Trailing Whitespace		0

- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring complete data reliability.
- The field contains 80 unique values, indicating a diverse yet structured set of categorical entries.
- The shortest value is "Bicycle," while the longest value is "Large passenger vehicle & Motor vehicle – other & Other/Unknown," reflecting a wide range in textual length, though consistently formatted.
- No leading or trailing whitespaces or blank records are present, ensuring the data is clean and free of unnecessary formatting issues.

- **POINT**

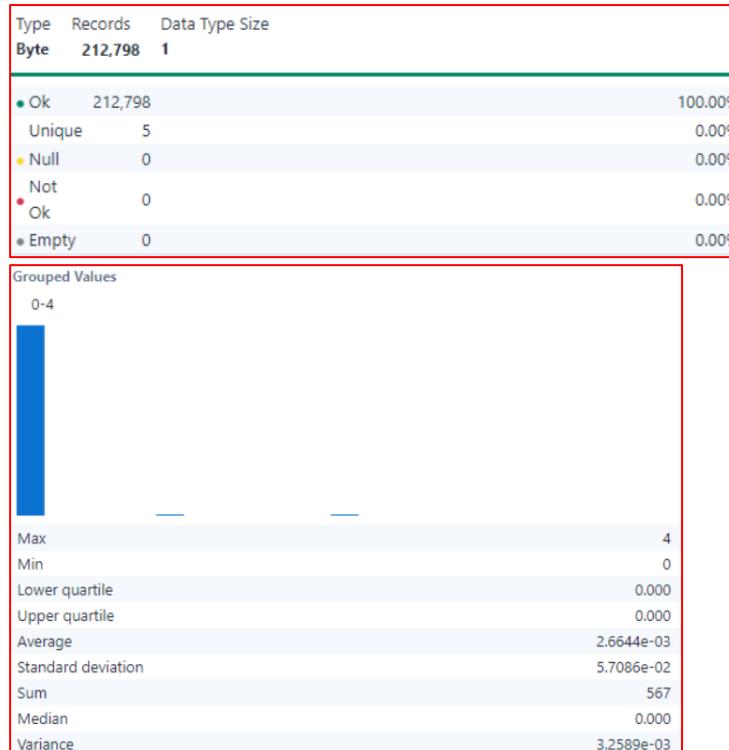
- Summary:

Summary		
Type	Records	Data Type Size
V_String	212,798	45
● Ok	209,076	98.25%
Unique	> 10,000	
● Null	3,722	1.75%
Not	0	0.00%
● Ok	0	0.00%
● Empty	0	0.00%
Length Statistics		
Min		22
Max		45
Average		33.80
Shortest Value		POINT (-97.65954 30.4)
Longest Value		POINT (-97.70786148611529 30.316989544922194)
First Alphanumeric Value		POINT (-97.57014781 30.30911982)
Last Alphanumeric Value		POINT (-97.92717344 30.18349898)
Blanks		0
Values with Leading Whitespace		0
Values with Trailing Whitespace		0

- Observation:
 - A total of 209,076 records (98.25%) are valid, while 3,722 records (1.75%) contain null values, requiring data imputation or exclusion for analysis completeness.
 - The field contains over 10,000 unique values, suggesting a highly granular and detailed dataset.
 - There are no records with leading or trailing whitespaces, blanks, or invalid entries, ensuring clean and consistent data formatting.

- **MOTOR_VEHICLE_DEATH_COUNT**

- Summary:

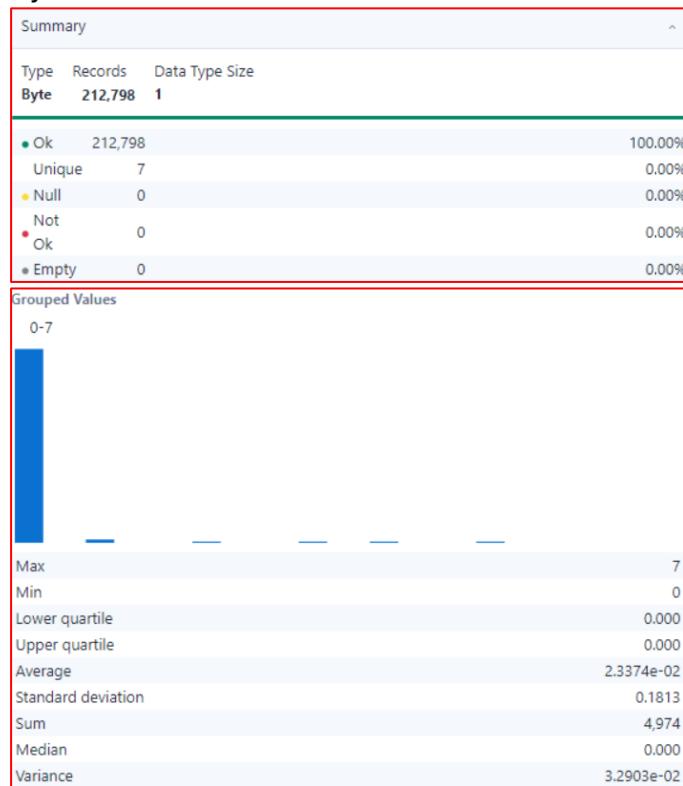


- Observation:

- All 212,798 records are valid with no null, empty, or invalid values detected. This ensures data integrity and reliability.
- The field contains only 5 unique values, suggesting a well-defined and restricted category system.
- The field is clean and ready for analysis, but the distribution suggests its utility might be limited to identifying specific categories or flags with minimal variance

- **MOTOR_VEHICLE_SERIOUS_INJURY_COUNT**

- Summary:

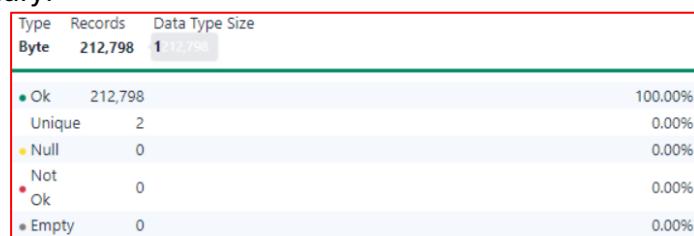


- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability and completeness.
 - The field contains seven unique values, indicating a limited but consistent data range suitable for categorical analysis.
 - The data exhibits consistency and standardization, with no values classified as "Not OK," demonstrating uniformity across the records.

- **BICYCLE_DEATH_COUNT**

- Summary:



Value Statistics	
Grouped Values	
0-1	
Max	1
Min	0
Lower quartile	0.000
Upper quartile	0.000
Average	1.8797e-04
Standard deviation	1.3709e-02
Sum	40
Median	0.000
Variance	1.8793e-04

- Observation:

- All 212,798 records are valid, ensuring no null, empty, or invalid entries.
- The field has two unique values (0 and 1), indicating a binary variable suitable for categorical analysis.
- The dataset shows consistency with clear and standardized values for this binary field.
- The grouped bar chart reveals a significant imbalance between the two categories, which may warrant further analysis depending on the context of the data.

- **BICYCLE_DEATH_COUNT**

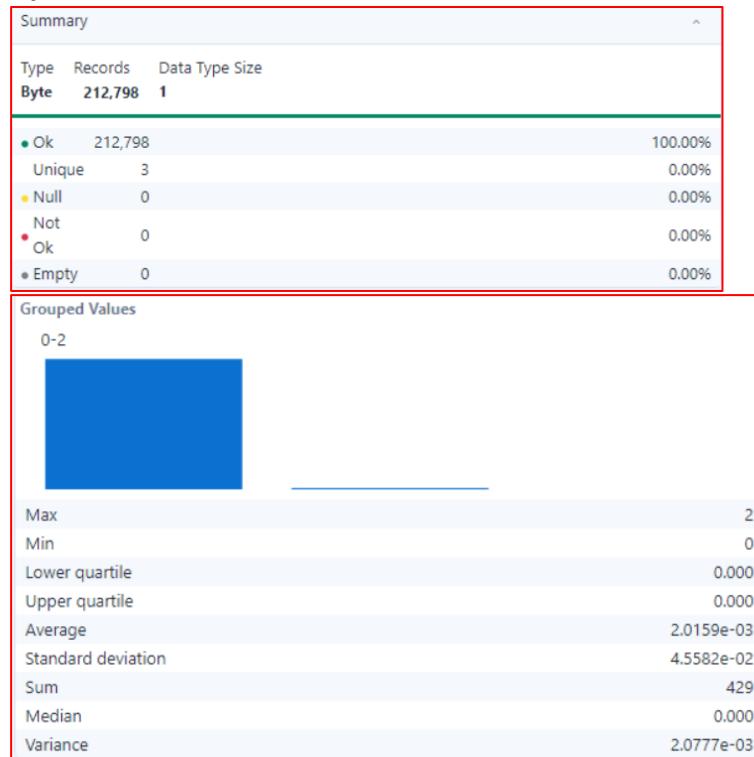
- Summary:

Type	Records	Data Type Size
Byte	212,798	1
● Ok	212,798	100.00%
Unique	5	0.00%
■ Null	0	0.00%
Not	0	0.00%
● Ok	0	0.00%
■ Empty	0	0.00%
Value Statistics		
Grouped Values		
0-14		
Max	14	
Min	0	
Lower quartile	0.000	
Upper quartile	0.001	
Average	2.0018e-03	
Standard deviation	5.3853e-02	
Sum	426	
Median	0.000	
Variance	2.9001e-03	

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring complete and reliable data.
 - The field contains only 5 unique values, indicating a limited categorical range.
 - The data appears consistent with no apparent formatting issues, such as leading or trailing spaces, blanks, or mixed representations.

- **PEDESTRIAN_DEATH_COUNT**

- Summary:

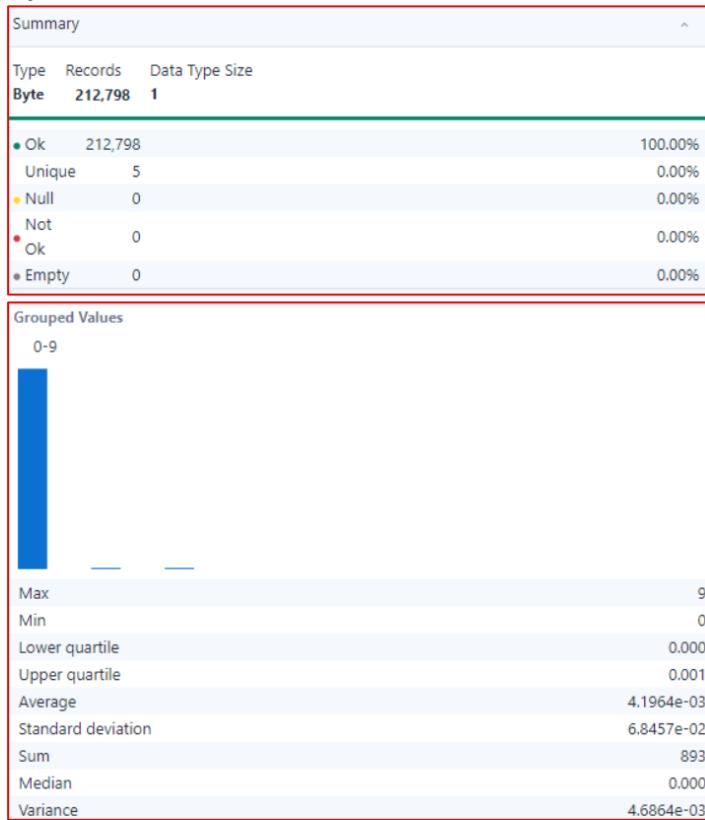


- Observation:

- All 212,798 records are valid, with no null, empty, or invalid entries, indicating complete data reliability.
- The field contains only 3 unique values, suggesting limited variability and potential for binary or categorical classification.
- Given the narrow value range and minimal deviation, the data appears consistent but may lack diversity.
- No apparent issues with data standardization or integrity were observed, making this field ready for categorical or binary analysis.

- **PEDESTRIAN_SERIOUS_INJURY_COUNT**

- Summary:



- Observation:

- All 212,798 records are valid without any null, not-ok, or empty values, ensuring reliability and consistency for analysis.
 - The field comprises 5 unique values, representing limited but distinct categorical data.
 - The data appears to be standardized, and no irregularities or inconsistencies in the formatting were detected during profiling.

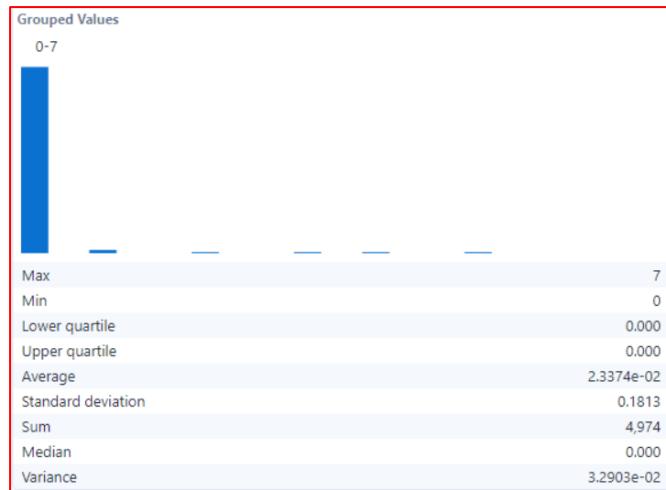
- **MOTORCYCLE_SERIOUS_INJURY_COUNT**

- Summary:

The table shows the distribution of data types: Ok (212,798, 100.00%), Unique (3, 0.00%), Null (0, 0.00%), Not (0, 0.00%), Ok (0, 0.00%), and Empty (0, 0.00%).

Type	Records	Data Type Size
Byte	212,798	1

Category	Value	Percentage
Ok	212,798	100.00%
Unique	3	0.00%
Null	0	0.00%
Not	0	0.00%
Ok	0	0.00%
Empty	0	0.00%

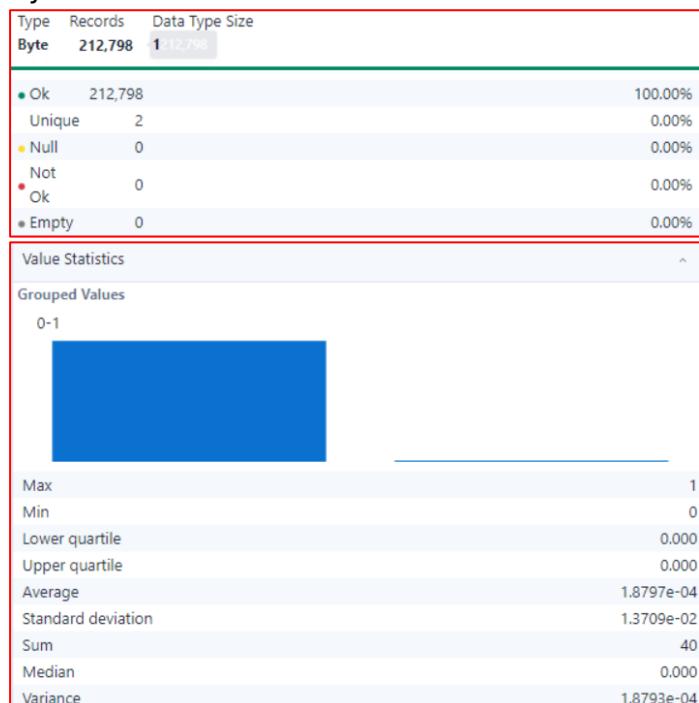


- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data completeness and reliability.
- The field contains 7 unique values, indicating a limited range of categorical data.
- The data is consistent and standardized, with no blank spaces or irregular formatting observed.

- **BICYCLE_DEATH_COUNT**

- Summary:



- Observation:

- All 212,798 records are valid with no null, not-ok, or empty values.
- The field contains 2 unique values, indicating binary classification.
- Values are standardized with consistent formatting.

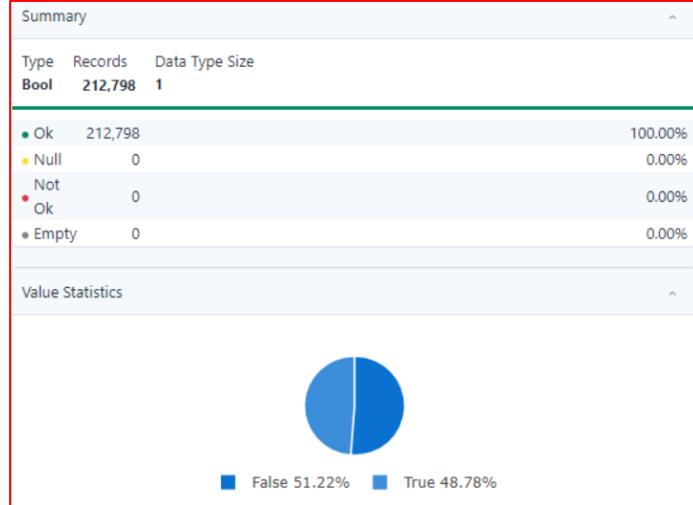
- One category dominates, indicating data imbalance.
- **OTHER_DEATH_COUNT**
- Summary:

Type	Records	Data Type	Size
Byte	212,798	1	
● Ok	212,798		100.00%
Unique	1		0.00%
■ Null	0		0.00%
Not Ok	0		0.00%
● Empty	0		0.00%
Value Statistics			
Grouped Values			
Only one value	0		
Max	0		
Min	0		
Lower quartile	0.000		
Upper quartile	0.000		
Average	0		
Standard deviation	0		
Sum	0		
Median	0.000		
Variance	0		

- Observation:
 - All 212,798 records are valid with no null, not-ok, or empty values.
 - The field contains only one unique value, indicating no variability.
 - The value is standardized across all records.
 - Statistical measures like variance, standard deviation, and average are zero due to the lack of diversity in values.
 - The field may not provide analytical insights due to its uniformity.
- **OTHER_SERIOUS_INJURY_COUNT**
- Summary:

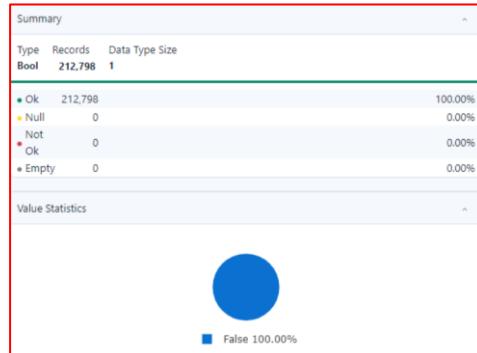
Type	Records	Data Type	Size
Byte	212,798	1	
● Ok	212,798		100.00%
Unique	3		0.00%
■ Null	0		0.00%
Not Ok	0		0.00%
● Empty	0		0.00%
Value Statistics			
Grouped Values			
0-3	3		
Max	3		
Min	0		
Lower quartile	0.000		
Upper quartile	0.000		
Average	4.2293e-05		
Standard deviation	8.3957e-03		
Sum	9		
Median	0.000		

- Observation:
 - All 212,798 records are valid with no null, not-ok, or empty values.
 - The field contains three unique values, ensuring some variability in the data.
 - The field is consistent and standardized, suitable for categorical or indicator analysis.
- **ONSYS_FL**
- Summary:



- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values, ensuring data completeness.
 - The field contains a binary distribution of values: 51.22% as False and 48.78% as True.
 - The data is well-structured and standardized, as there are only two unique values (True/False).
 - The distribution indicates a near-even split, which could be meaningful depending on the context of the field's application.
 - No inconsistencies or formatting issues are observed, making the data ready for analysis.

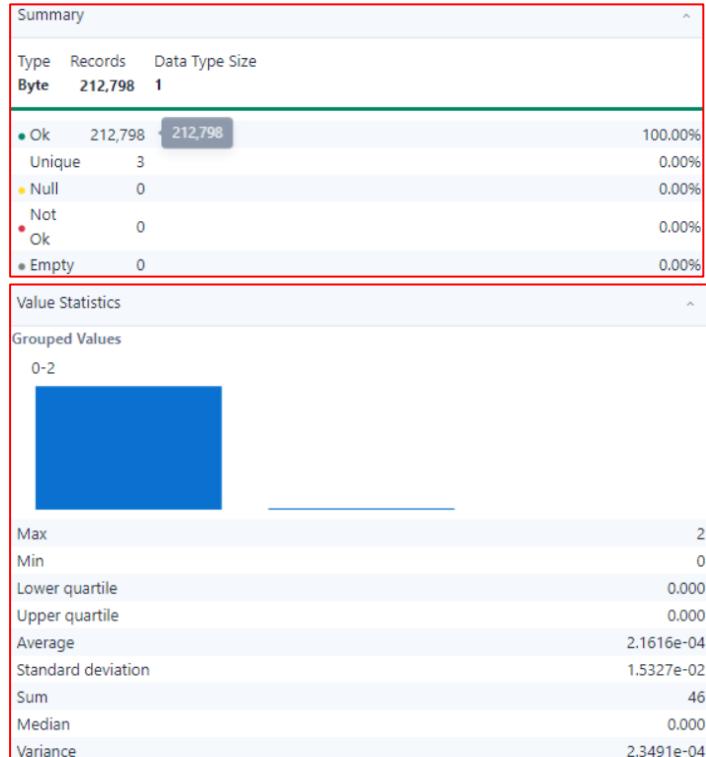
- **PRIVATE_DR_FL**
- Summary:



- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values, ensuring data completeness.
 - The field contains a single unique value: False, which accounts for 100% of the data.
 - The lack of variation in the data indicates it may not be informative for analysis unless the context justifies the uniformity.
 - The field is standardized and consistent, with no formatting issues or anomalies.
 - Further exploration may be needed to understand why all values are False and if this aligns with the intended dataset purpose.

- **MICROMOBILITY_SERIOUS_INJURY_COUNT**

- Summary:



- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring completeness and data reliability.
- The field contains 3 unique values, which indicates low variance and simplicity in categorization.
- The grouped values range from 0 to 2, with all records distributed across these values.
- The data does not contain any standardization issues, as the unique values are consistent across all records.
- The low variance and standard deviation suggest uniformity in data distribution, suitable for analysis without further preprocessing.

- **MICROMOBILITY_DEATH_COUNT**

- Summary:

Type	Records	Data Type	Size
Byte	212,798	1	
<hr/>			
● Ok	212,798		100.00%
Unique	2		0.00%
■ Null	0		0.00%
Not	0		0.00%
● Ok	0		0.00%
■ Empty	0		0.00%
<hr/>			
Grouped Values			
0-1			
			
Max		1	
Min		0	
Lower quartile		0.000	
Upper quartile		0.000	
Average		2.8195e-05	
Standard deviation		5.3099e-03	
Sum		6	
Median		0.000	
Variance		2.8195e-05	

- Observation:

- All 212,798 records are valid, ensuring completeness and data reliability.
 - The field contains 2 unique values, indicating low variance and simplicity in categorization.
 - The grouped values range from 0 to 1, with all records distributed across these values.
 - The data does not contain any standardization issues, as the unique values are consistent across all records.
 - The low variance and standard deviation suggest uniformity in data distribution, suitable for analysis without further preprocessing.

- **CRASH_TIMESTAMP(US/CENTRAL)**

- Summary:

Summary ▲		
Type	Records	Data Type
String	212,798	22
<hr/>		
● Ok	212,798	100.00%
Unique	> 10,000	
■ Null	0	0.00%
Not	0	0.00%
● Ok	0	0.00%
■ Empty	0	0.00%
<hr/>		
Length Statistics ▲		
Min	22	
Max	22	
Average	22.00	
Shortest Value	01/03/2014 08:32:00 AM	
Longest Value	01/03/2014 08:32:00 AM	
First Alphanumeric Value	01/01/2010 01:12:00 AM	
Last Alphanumeric Value	12/31/2023 12:51:00 PM	
Blanks	0	
Values with Leading Whitespace	0	
Values with Trailing Whitespace	0	

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.
 - The bar chart displays a relatively even distribution across grouped ranges, confirming a balanced representation of values

- **CRASH_TIMESTAMP**

- Summary:

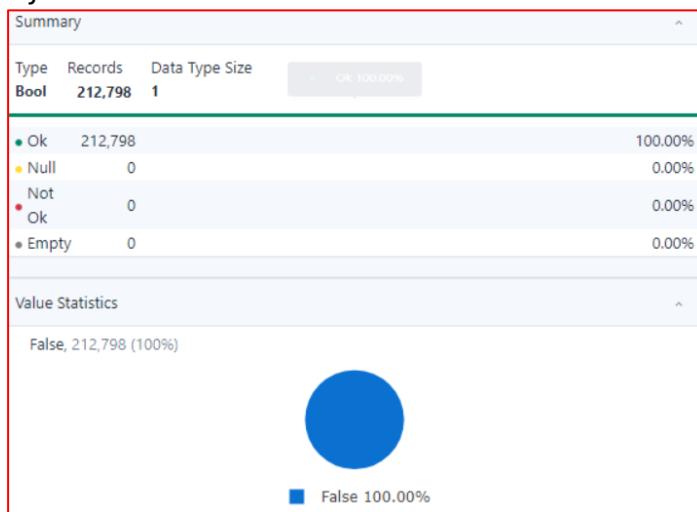
Summary ▲			
Type	Records	Data Type	Size
String	212,798	22	
● Ok	212,798		100.00%
Unique	> 10,000		
■ Null	0		0.00%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
 Length Statistics			
Min			22
Max			22
Average			22.00
Shortest Value			01/03/2014 02:32:00 PM
Longest Value			01/03/2014 02:32:00 PM
First Alphanumeric Value			01/01/2010 06:06:00 PM
Last Alphanumeric Value			12/31/2023 12:59:00 PM
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- All 212,798 records are valid, ensuring completeness and data reliability.
- The field contains over 10,000 unique values, indicating high variability and richness in the dataset.
- The date range spans from 01/01/2010 00:12:00 AM to 12/31/2023 12:51:00 PM, covering an extensive period.
- No null, empty, or invalid values are present, ensuring consistency and readiness for analysis.
- The absence of blanks and trailing or leading whitespace confirms standardized data formatting.

- **IS_DELETED**

- Summary:

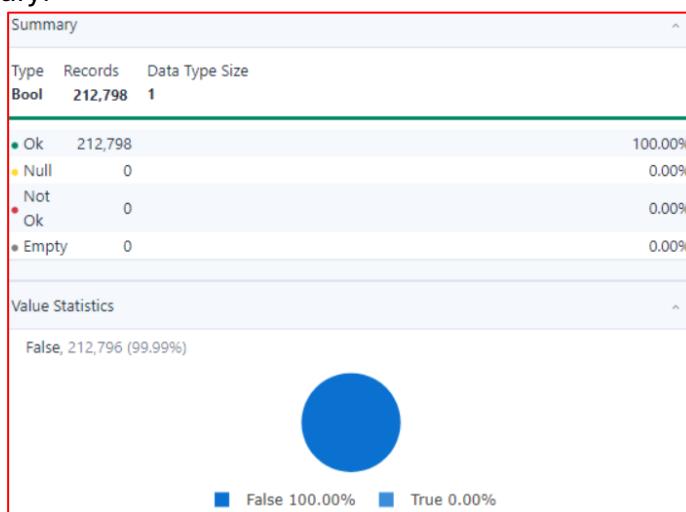


- Observation:

- All 212,798 records are valid, ensuring completeness and data reliability.
 - The field contains a single unique value (False), indicating no variability in the data.
 - No null, empty, or invalid values are detected, ensuring consistency.
 - The lack of variability suggests this field does not provide distinguishing information and may not contribute significantly to data analysis.

- **IS_TEMPORARY_RECORD**

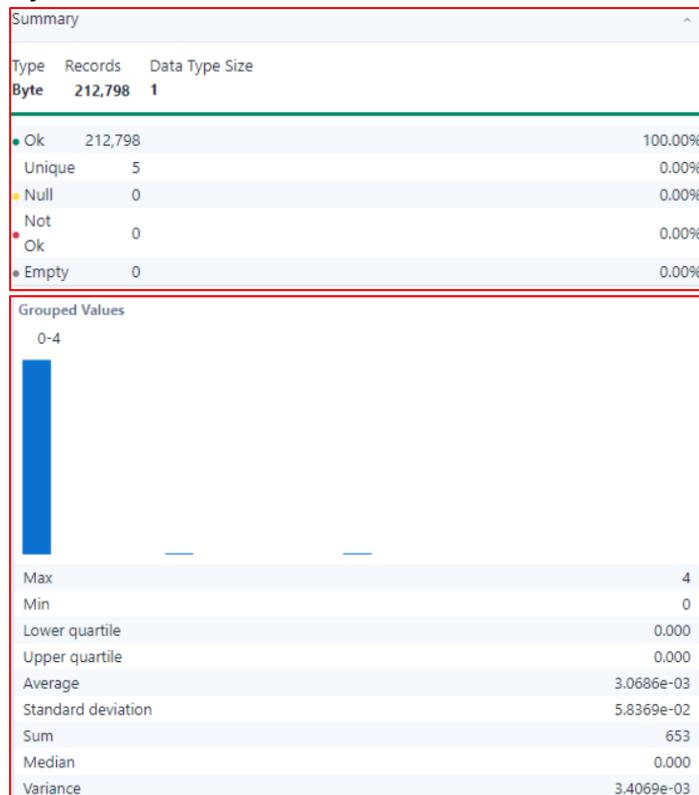
- Summary:



- Observation:
 - All 212,798 records are valid, ensuring completeness and data reliability.
 - The field predominantly contains the value False (99.99%), with an almost negligible presence of True.
 - No null, empty, or invalid values are detected, indicating consistent data quality.
 - The high dominance of False suggests limited variability, which may impact the field's significance in data analysis.

- **LAW_ENFORCEMENT_FATALITY_COUNT**

- Summary:

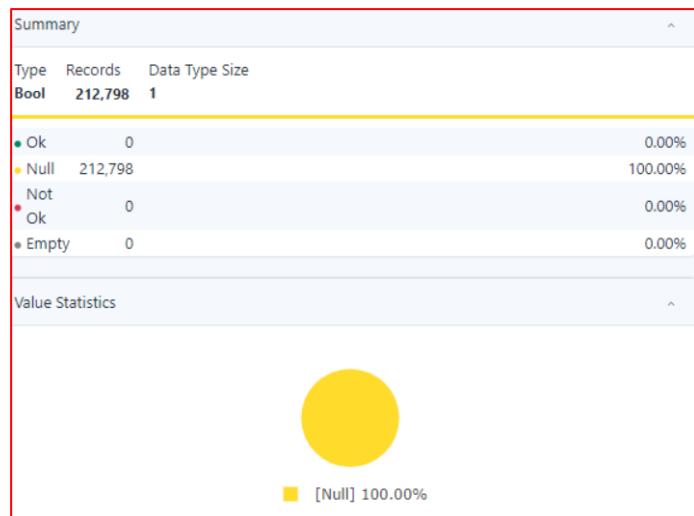


- Observation:

- All 212,798 records are complete with no null or empty values.
- The field contains 5 unique values, indicating low variability.
- All values follow a uniform format, ensuring data consistency.
- Data is well-structured and consistent, requiring no additional preprocessing for analysis.

- **REPORTED_STREET_PREFIX**

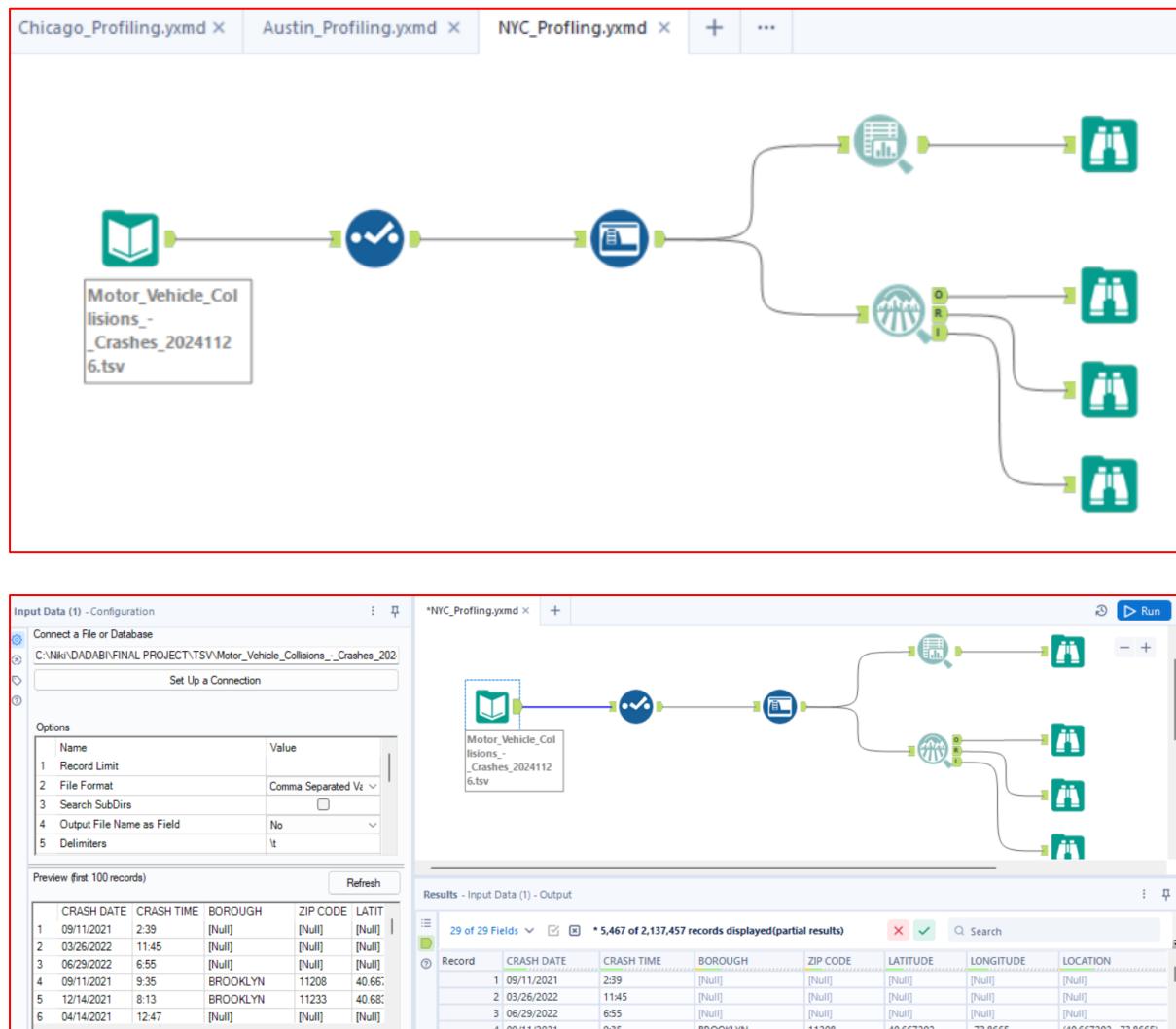
- Summary:



- Observation:

- The entire dataset (212,798 records) contains null values, indicating missing information in this field.
- There are no valid, non-null entries present in the field.
- The field is defined as Boolean but lacks any valid true or false entries.
- This field requires immediate attention—data input needs correction or imputation strategies to handle the null values.
- The field, as it stands, is not usable for analysis without addressing the completeness issue.

NYC Dataset



- **Overview of the dataset**

The Chicago dataset provides crash-related information with a focus on environmental, roadway, and driver-related factors. It includes fields such as weather conditions, lighting, traffic control devices, road alignment, and injury details. This dataset supports safety assessments and strategies to reduce crash incidents.

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 (NYC)**

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.

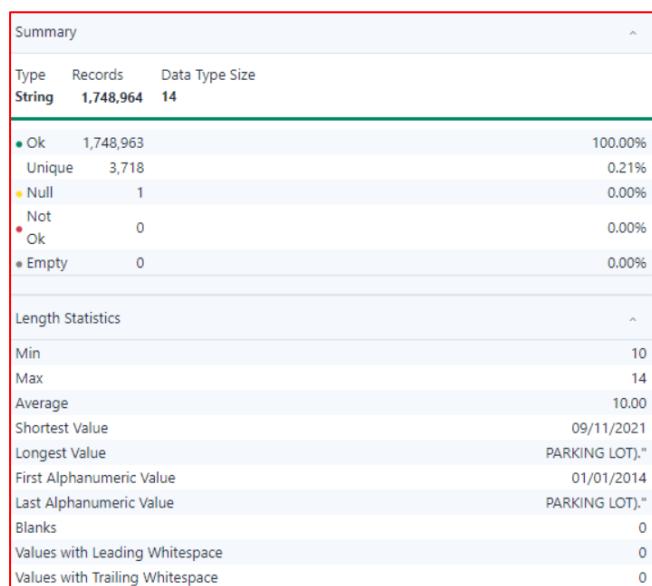
NUMBER_OF_MOTORIST_KILLED	Total number of motorist fatalities in the crash.	Cross-check alignment with NUMBER_OF_PERSONS_KILLED.
CONTRIBUTING_FACTOR_VEHICLE_1	Primary contributing factor for the first vehicle involved in the crash.	Validate against a predefined list of contributing factors (e.g., speeding, 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	Validate data entries for consistency and completeness.

	in the crash, if applicable.	
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.

Data Observation:

- **CRASH_DATE**

- Summary:



- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
- The field contains over 10,000 unique values, indicating diverse and meaningful entries
- No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **CRASH_DATE**

- Summary:

Summary			
Type	Records	Data Type	Size
String	1,748,964	5	
• Ok	1,748,963		100.00%
Unique	1,441		0.08%
• Null	1		0.00%
Not	0		0.00%
• Ok	0		0.00%
• Empty	0		0.00%
Length Statistics			
Min			1
Max			5
Average			4.70
Shortest Value			0
Longest Value			11:45
First Alphanumeric Value			0
Last Alphanumeric Value			9:59
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
- The field contains over 10,000 unique values, indicating diverse and meaningful entries
- No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **BOROUGH**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	13	
• Ok	1,191,661		68.14%
Unique	6		0.00%
• Null	557,303		31.86%
Not	0		0.00%
• Ok	0		0.00%
• Empty	0		0.00%
Length Statistics			
Min			1
Max			13
Average			7.40
Shortest Value			0
Longest Value			STATEN ISLAND
First Alphanumeric Value			0
Last Alphanumeric Value			STATEN ISLAND
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **ZIP_CODE**

- Summary:

Summary		
Type	Records	Data Type Size
String	1,748,964	5
● Ok	1,191,401	68.12%
Unique	234	0.01%
■ Null	557,536	31.88%
Not	0	0.00%
● Ok	0	0.00%
● Empty	27	0.00%
Length Statistics		
Min	1	
Max	5	
Average	5.00	
Shortest Value	0	
Longest Value	11208	
First Alphanumeric Value	[Null]	
Last Alphanumeric Value	11697	
Blanks	27	
Values with Leading Whitespace	0	
Values with Trailing Whitespace	0	

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **LATITUDE**

- Summary:

Summary ▲		
Type	Records	Data Type Size
Double	1,748,964	8
● Ok	1,552,529	88.77%
Unique	> 10,000	
■ Null	196,435	11.23%
Not	0	0.00%
● Ok	0	0.00%
● Empty	0	0.00%

Value Statistics	
Grouped Values	
0-43.34	
Max	43.3444
Min	0
Lower quartile	40.666
Upper quartile	40.770
Average	40.6305
Standard deviation	1.9560
Sum	63,080,090.7189
Median	40.721
Variance	3.8260

- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
- The field contains over 10,000 unique values, indicating diverse and meaningful entries
- No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **LONGITUDE**

- Summary:

Summary		
Type	Records	Data Type Size
Double	1,748,964	8
● Ok	1,552,529	88.77%
Unique	> 10,000	
■ Null	196,435	11.23%
● Not	0	0.00%
● Ok	0	0.00%
● Empty	0	0.00%

Value Statistics	
Grouped Values	
-201.35-0	
Max	0
Min	-201.3599
Lower quartile	-73.974
Upper quartile	-73.865
Average	-73.7569
Standard deviation	3.7041
Sum	-114,509,851.5047
Median	-73.926
Variance	13.7206

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **LOCATION**

- Summary:

Summary ▲			
Type	Records	Data Type	Size
String	1,748,964	25	
● Ok	1,552,529		88.77%
Unique	> 10,000		
■ Null	196,435		11.23%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics ▲			
Min		1	
Max		25	
Average		22.70	
Shortest Value		0	
Longest Value		(40.8359271, -73.9029039)	
First Alphanumeric Value		(0.0, 0.0)	
Last Alphanumeric Value		0	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **ON_STREET_NAME**

- Summary:

Summary ▲			
Type	Records	Data Type Size	
V_String	1,748,964	32	
● Ok	1,374,460	78.59%	
Unique	> 10,000		
■ Null	374,489	21.41%	
Not	0	0.00%	
● Ok	0	0.00%	
● Empty	15	0.00%	
Length Statistics			
Min	1		
Max	32		
Average	14.10		
Shortest Value	0		
Longest Value	WILLIAMSBURG BRIDGE OUTER ROADWA		
First Alphanumeric Value	[Null]		
Last Alphanumeric Value	estfarms road		
Blanks	15		
Values with Leading Whitespace	0		
Values with Trailing Whitespace	0		

- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
- The field contains over 10,000 unique values, indicating diverse and meaningful entries
- No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **CROSS_STREET_NAME**

- Summary:

Summary ▲			
Type	Records	Data Type Size	
V_WString	1,748,964	32	
● Ok	1,066,685	60.99%	
Unique	> 10,000		
■ Null	682,262	39.01%	
Not	0	0.00%	
● Ok	0	0.00%	
● Empty	17	0.00%	
Length Statistics			
Min	1		
Max	32		
Average	13.00		
Shortest Value	1		
Longest Value	CROSS BRONX EXPRESSWAY EXTENSION		
First Alphanumeric Value	[Null]		
Last Alphanumeric Value	♦ ST 138 STREET		
Blanks	17		
Values with Leading Whitespace	0		
Values with Trailing Whitespace	0		

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.
- **OFF_STREET_NAME**
- Summary:

Summary ▲		
Type	Records	Data Type Size
V_String	1,748,964	40
● Ok	302,468	17.29%
Unique	> 10,000	
■ Null	1,446,467	82.70%
Not		
● Ok	0	0.00%
● Empty	29	0.00%
Length Statistics		
Min		3
Max		40
Average		23.40
Shortest Value		LIE
Longest Value	26-45 brooklyn queens expressway wes	
First Alphanumeric Value		[Null]
Last Alphanumeric Value		woodhaven boulevard
Blanks		29
Values with Leading Whitespace		0
Values with Trailing Whitespace		0

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **NUMBER_OF_PERSONS_INJURED**

- Summary:

Summary			
Type	Records	Data Type Size	
V_String	1,748,964	11	
● Ok	1,748,945		100.00%
Unique	29		0.00%
■ Null	19		0.00%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			1
Max			11
Average			1.00
Shortest Value			2
Longest Value			Unspecified
First Alphanumeric Value			0
Last Alphanumeric Value			Unspecified
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **NUMBER_OF_PERSONS_KILLED**

- Summary:

Summary			
Type	Records	Data Type Size	
Byte	1,748,964	1	
● Ok	1,748,931		100.00%
Unique	6		0.00%
■ Null	33		0.00%
Not			
● Ok	0		0.00%
● Empty	0		0.00%

Value Statistics	
Grouped Values	
0-8	
Max	8
Min	0
Lower quartile	0.000
Upper quartile	0.001
Average	1.3837e-03
Standard deviation	3.9223e-02
Sum	2,420
Median	0.000
Variance	1.5384e-03

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.
- **NUMBER_OF_PEDESTRIANS_INJURED**
 - Summary:

Summary			
Type	Records	Data Type	Size
Byte	1,748,964	1	
● Ok	1,748,962		100.00%
Unique	13		0.00%
Null	2		0.00%
Not Ok	0		0.00%
Empty	0		0.00%

Value Statistics	
Grouped Values	
0-27	
Max	27
Min	0
Lower quartile	0.001
Upper quartile	0.002
Average	5.2667e-02
Standard deviation	0.2359
Sum	92,114
Median	0.001
Variance	5.5669e-02

- Observation:

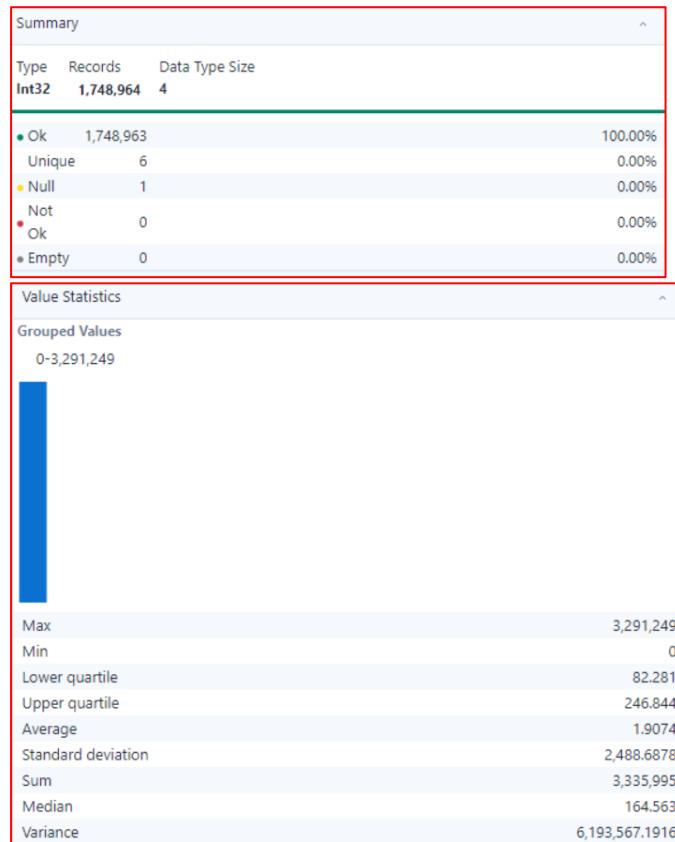
- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
- The field contains over 10,000 unique values, indicating diverse and meaningful entries
- No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **NUMBER_OF_PEDESTRIANS_KILLED**

- Summary:

Type	Records	Data Type Size
Byte	1,748,964	1
● Ok	1,748,962	100.00%
Unique	4	0.00%
■ Null	2	0.00%
Not Ok	0	0.00%
● Empty	0	0.00%
Value Statistics		
Grouped Values		
0-6		
Max	6	
Min	0	
Lower quartile	0.000	
Upper quartile	0.000	
Average	6.9069e-04	
Standard deviation	2.6853e-02	
Sum	1,208	
Median	0.000	
Variance	7.2109e-04	

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.
- **NUMBER_OF_CYCLISTS_INJURED**
- Summary:



- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **NUMBER_OF_CYCLIST_KILLED**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	17	
● Ok	1,748,963		100.00%
Unique	4		0.00%
■ Null	1		0.00%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min	1		
Max	17		
Average	1.00		
Shortest Value	0		
Longest Value	PASSENGER VEHICLE		
First Alphanumeric Value	0		
Last Alphanumeric Value	PASSENGER VEHICLE		
Blanks	0		
Values with Leading Whitespace	0		
Values with Trailing Whitespace	0		

- Observation:

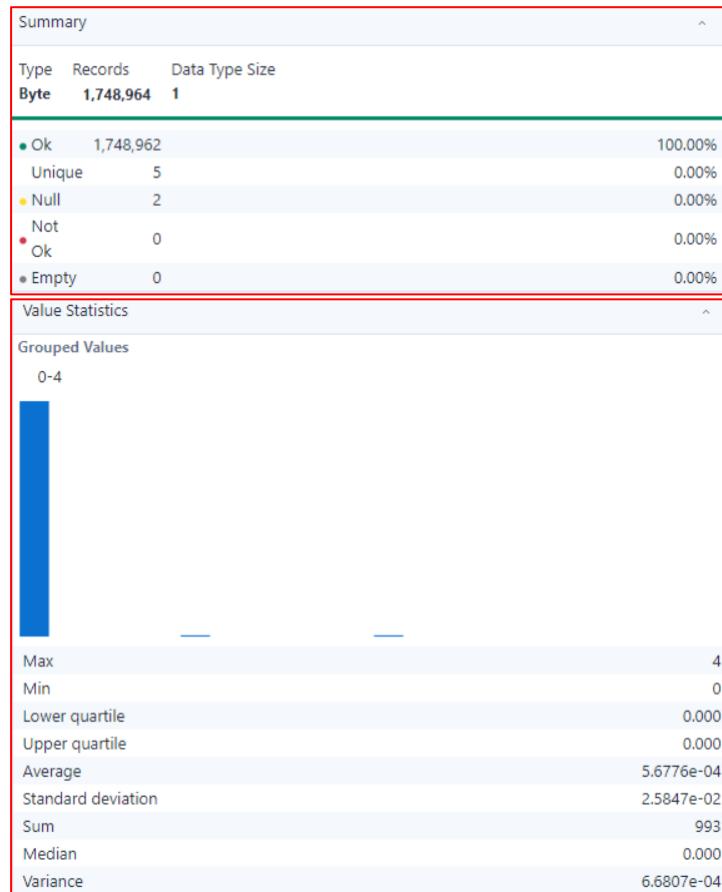
- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **NUMBER_OF_MOTORIST_INJURED**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	17	
● Ok	1,748,963		100.00%
Unique	29		0.00%
■ Null	1		0.00%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min	1		
Max	17		
Average	1.00		
Shortest Value	2		
Longest Value	PASSENGER VEHICLE		
First Alphanumeric Value	0		
Last Alphanumeric Value	PASSENGER VEHICLE		
Blanks	0		
Values with Leading Whitespace	0		
Values with Trailing Whitespace	0		

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.
- **NUMBER_OF_MOTORIST_KILLED**
- Summary:



- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **CONTRIBUTING_FACTOR_VEHICLE_1**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	53	
● Ok	1,743,256		99.67%
Unique	61		0.00%
■ Null	5,708		0.33%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			1
Max			53
Average			19.70
Shortest Value			1
Longest Value		Pedestrian/Bicyclist/Other Pedestrian Error/Confusion	
First Alphanumeric Value			1
Last Alphanumeric Value		Windshield Inadequate	
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **CONTRIBUTING_FACTOR_VEHICLE_2**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	53	
● Ok	1,482,082		84.74%
Unique	61		0.00%
■ Null	266,882		15.26%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			1
Max			53
Average			13.10
Shortest Value			1
Longest Value		Pedestrian/Bicyclist/Other Pedestrian Error/Confusion	
First Alphanumeric Value			1
Last Alphanumeric Value		Windshield Inadequate	
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.
- **CONTRIBUTING_FACTOR_VEHICLE_3**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	53	
● Ok	122,551		7.01%
Unique	51		0.00%
● Null	1,626,413		92.99%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			1
Max			53
Average			11.70
Shortest Value			1
Longest Value	Pedestrian/Bicyclist/Other Pedestrian Error/Confusion		
First Alphanumeric Value			1
Last Alphanumeric Value	View Obstructed/Limited		
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **CONTRIBUTING_FACTOR_VEHICLE_4**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	43	
● Ok	27,373		1.57%
Unique	38		0.00%
■ Null	1,721,591		98.43%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			5
Max			43
Average			11.50
Shortest Value			Glare
Longest Value		Traffic Control Device Improper/Non-Working	
First Alphanumeric Value		Accelerator Defective	
Last Alphanumeric Value		View Obstructed/Limited	
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
- The field contains over 10,000 unique values, indicating diverse and meaningful entries
- No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **CONTRIBUTING_FACTOR_VEHICLE_5**

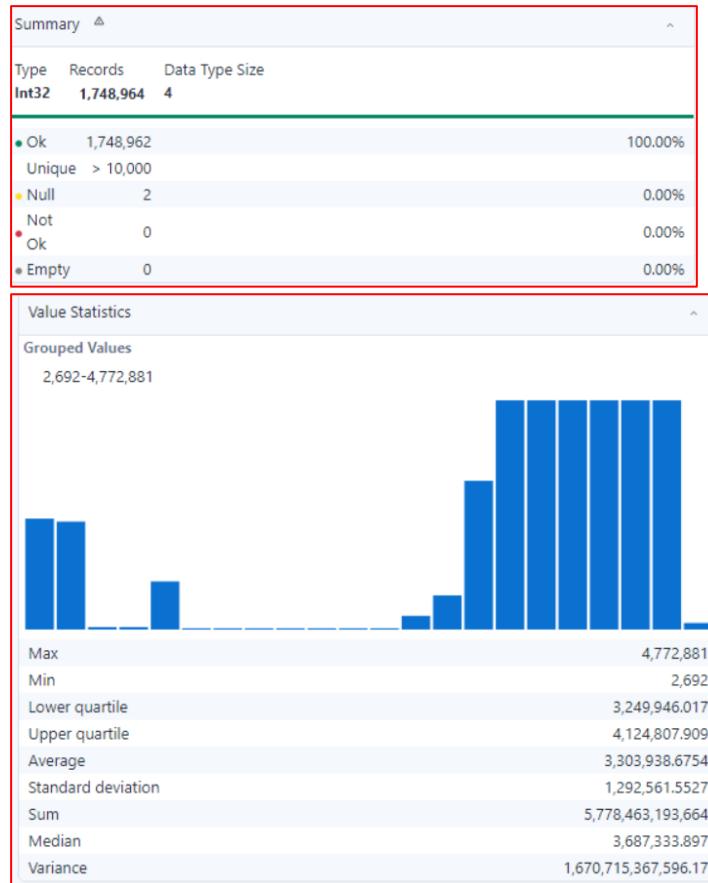
- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	43	
● Ok	7,352		0.42%
Unique	28		0.00%
■ Null	1,741,612		99.58%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			5
Max			30
Average			11.50
Shortest Value		Glare	
Longest Value		Reaction to Uninvolved Vehicle	
First Alphanumeric Value		Aggressive Driving/Road Rage	
Last Alphanumeric Value		Unspecified	
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **COLLISION_ID**

- Summary:



- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **VEHICLE_TYPE_CODE_1**

- Summary:

Summary			
Type	Records	Data Type	Size
V_WString	1,748,964	38	
● Ok	1,737,733		99.36%
Unique	1,429		0.08%
■ Null	11,231		0.64%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		1	
Max		38	
Average		16.90	
Shortest Value		.	
Longest Value		Enclosed Body - Nonremovable Enclosure	
First Alphanumeric Value		"lime mope	
Last Alphanumeric Value		♦MBU	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **VEHICLE_TYPE_CODE_2**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	38	
● Ok	1,421,982		81.30%
Unique	1,582		0.09%
■ Null	326,982		18.70%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		1	
Max		38	
Average		16.10	
Shortest Value		.	
Longest Value		Enclosed Body - Nonremovable Enclosure	
First Alphanumeric Value		(ceme	
Last Alphanumeric Value		yw	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.
- **VEHICLE_TYPE_CODE_3**
- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	35	
● Ok	118,155		6.76%
Unique	225		0.01%
■ Null	1,630,809		93.24%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			2
Max			35
Average			17.60
Shortest Value			PK
Longest Value		Station Wagon/Sport Utility Vehicle	
First Alphanumeric Value		2 dr sedan	
Last Alphanumeric Value		yello	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **VEHICLE_TYPE_CODE_4**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	35	
● Ok	26,418		1.51%
Unique	88		0.01%
■ Null	1,722,546		98.49%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		2	
Max		35	
Average		17.90	
Shortest Value		PK	
Longest Value		Station Wagon/Sport Utility Vehicle	
First Alphanumeric Value		2 dr sedan	
Last Alphanumeric Value		van	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **VEHICLE_TYPE_CODE_5**

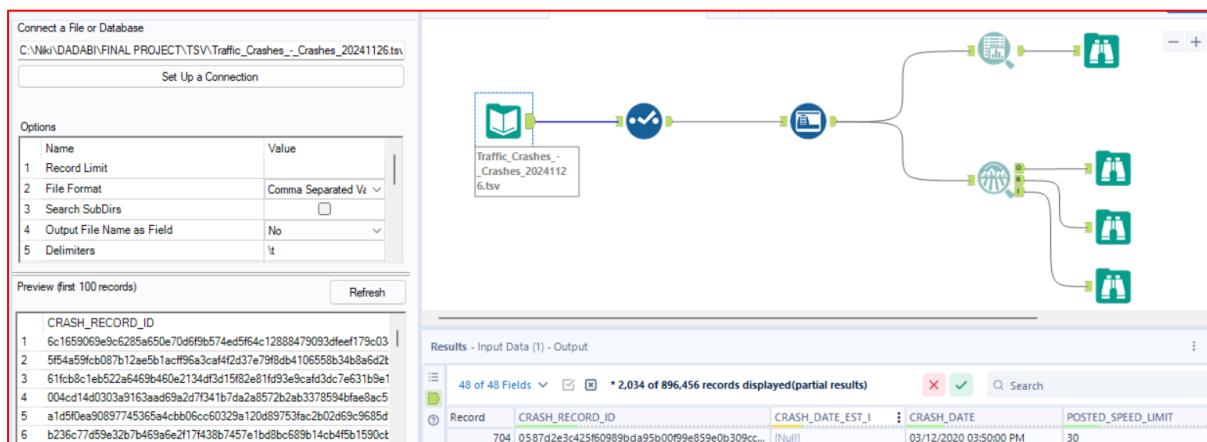
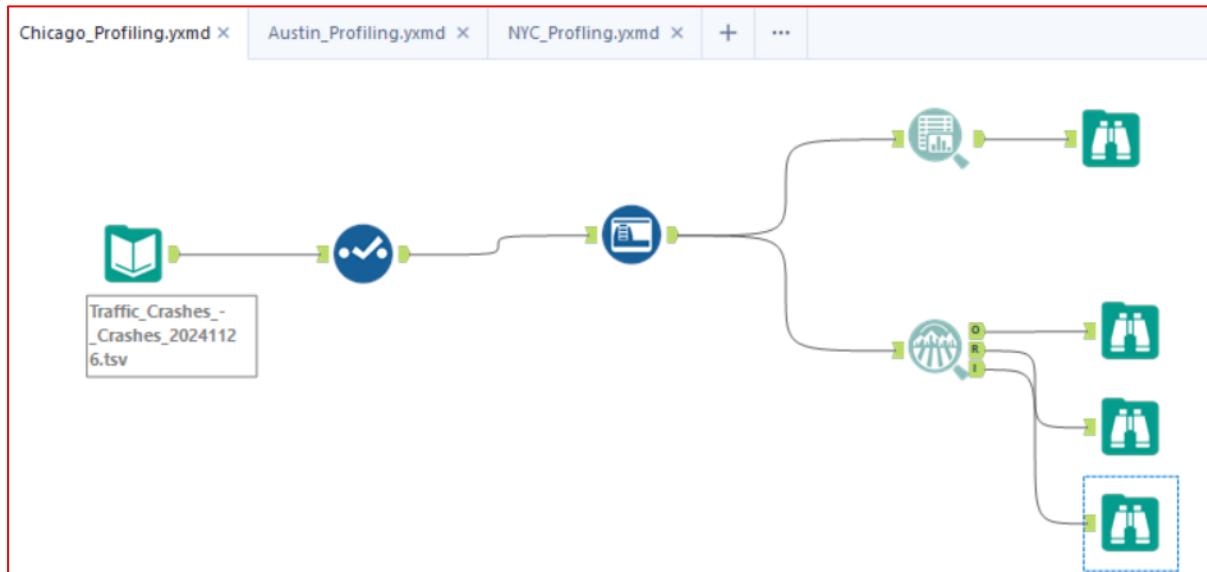
- Summary:

Summary			
Type	Records	Data Type	Size
V_String	1,748,964	35	
● Ok	7,128		0.41%
Unique	61		0.00%
■ Null	1,741,836		99.59%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		2	
Max		35	
Average		18.10	
Shortest Value		C3	
Longest Value		Station Wagon/Sport Utility Vehicle	
First Alphanumeric Value		2 dr sedan	
Last Alphanumeric Value		van	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
- The field contains over 10,000 unique values, indicating diverse and meaningful entries
- No missing or not-ok records exist, ensuring completeness and readiness for analysis.

Chicago Dataset



- **Overview of the dataset**

The Chicago dataset provides crash-related information with a focus on environmental, roadway, and driver-related factors. It includes fields such as weather conditions, lighting, traffic control devices, road alignment, and injury details. This dataset supports safety assessments and strategies to reduce crash incidents.

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.

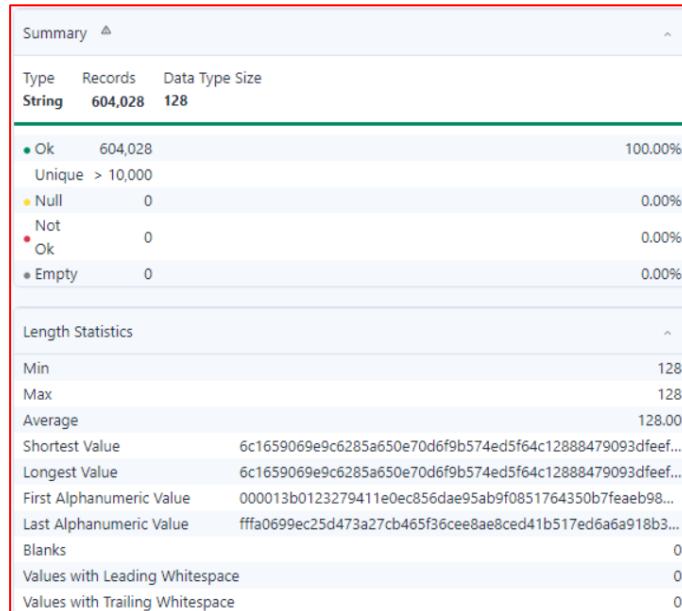
LIGHTING_CONDITION	Lighting conditions at the time of the crash.	Ensure valid categories such as daylight, dark, dawn, or dusk.
FIRST_CRASH_TYPE	Initial type of collision in the crash sequence.	Check for logical consistency with other crash details, such as location or severity.
TRAFFICWAY_TYPE	Layout 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.

Data Observation:

- CRASH_RECORD_ID
 - Summary:



- Observation:
 - All 604,028 records are valid, with no null, empty, or invalid values detected, ensuring data reliability.
 - The field contains over 10,000 unique values, indicating diverse and meaningful entries.
 - No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **CRASH_DATE_EST_I**

- Summary:

Summary			
Type	Records	Data Type	Size
String	604,028	1	
● Ok	44,967		7.44%
Unique	2		0.00%
● Null	559,061		92.56%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			1
Max			1
Average			1.00
Shortest Value			Y
Longest Value			Y
First Alphanumeric Value			N
Last Alphanumeric Value			Y
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- Only 44,967 records (7.44%) are valid, indicating a limited amount of usable data.
- 559,061 records (92.56%) are null, highlighting significant data gaps.
- The field contains only 2 unique values, limiting its analytical utility.
- No empty or not-ok records exist, ensuring consistency among valid entries.
- All records have a fixed length of 1 character, with no blanks, leading, or trailing whitespace.
- The high null rate raises concerns about data quality and may require further investigation.

- **CRASH_DATE**

- Summary:

Type	Records	Data Type Size
String	604,028	22
<hr/>		
● Ok	604,028	100.00%
Unique	> 10,000	
■ Null	0	0.00%
■ Not	0	0.00%
● Ok	0	0.00%
■ Empty	0	0.00%
<hr/>		
Length Statistics		
Min	22	
Max	22	
Average	22.00	
Shortest Value	08/18/2023 12:50:00 PM	
Longest Value	08/18/2023 12:50:00 PM	
First Alphanumeric Value	01/01/2016 01:00:00 AM	
Last Alphanumeric Value	12/31/2023 12:47:00 PM	
Blanks	0	
Values with Leading Whitespace	0	
Values with Trailing Whitespace	0	

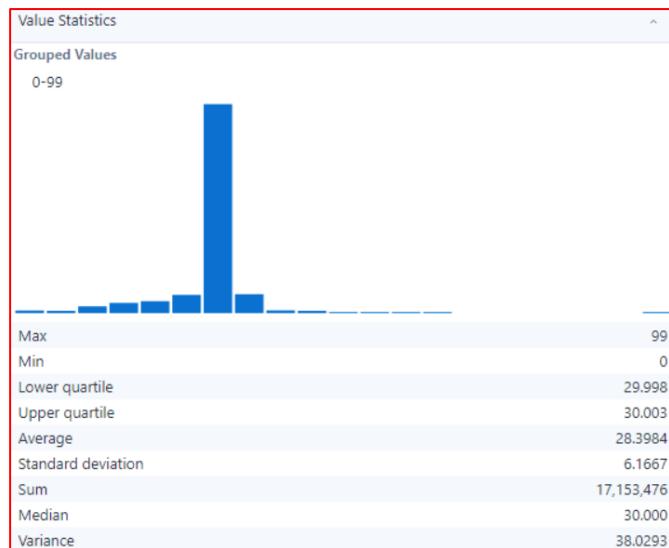
- Observation:

- All 604,028 records (100%) are valid, with no null, empty, or invalid values detected, ensuring complete data reliability.
 - The field contains over 10,000 unique values, indicating a high level of diversity and richness in the dataset.
 - All values have a consistent length of 22 characters, with no blanks, leading, or trailing whitespace.
 - The first alphanumeric value is "01/01/2016 01:00:00 AM," and the last is "12/31/2023 12:47:00 PM," indicating a broad temporal range.
 - The dataset is clean and well-prepared for analysis without any data quality concerns.

- **POSTED_SPEED_LIMIT**

- Summary:

Type	Records	Data Type Size
Byte	604,028	1
<hr/>		
● Ok	604,028	100.00%
Unique	45	0.01%
■ Null	0	0.00%
■ Not	0	0.00%
● Ok	0	0.00%
■ Empty	0	0.00%



- Observation:

- All 604,028 records (100%) are valid, with no null, empty, or invalid values, ensuring high data quality.
- The field contains 45 unique values, representing a low diversity of entries (0.01%).
- The dataset is ready for analysis, with consistent formatting and no data quality concerns.

- **TRAFFIC_CONTROL_DEVICE**

- Summary:

Summary

Type	Records	Data Type	Size
V_String	604,028		24
● Ok	604,028		100.00%
Unique	19		0.00%
Null	0		0.00%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			5
Max			24
Average			12.30
Shortest Value			OTHER
Longest Value			PEDESTRIAN CROSSING SIGN
First Alphanumeric Value			BICYCLE CROSSING SIGN
Last Alphanumeric Value			YIELD
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - All 604,028 records (100%) are valid, with no null, empty, or invalid values detected.
 - The field contains 19 unique values, indicating moderate variability in the data.
 - The shortest value is "OTHER," and the longest is "PEDESTRIAN CROSSING SIGN."
 - No records have blanks, leading, or trailing whitespace, ensuring data cleanliness.
 - The dataset is well-structured and ready for further analysis.

- **DEVICE_CONDITION**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	24	
● Ok	604,028		100.00%
Unique	8		0.00%
■ Null	0		0.00%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			5
Max			24
Average			13.80
Shortest Value			OTHER
Longest Value			WORN REFLECTIVE MATERIAL
First Alphanumeric Value			FUNCTIONING IMPROPERLY
Last Alphanumeric Value			WORN REFLECTIVE MATERIAL
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- All 604,028 records (100%) are valid, with no null, empty, or invalid values detected.
- The field contains 8 unique values, indicating low variability in the dataset.
- The shortest value is "OTHER," and the longest is "WORN REFLECTIVE MATERIAL."
- No records have blanks, leading, or trailing whitespace, ensuring data cleanliness.
- The dataset is highly consistent and ready for analysis.

- **WEATHER_CONDITION**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	24	
● Ok	604,028		100.00%
Unique	12		0.00%
■ Null	0		0.00%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		4	
Max		24	
Average		5.30	
Shortest Value		SNOW	
Longest Value		BLOWING SAND, SOIL, DIRT	
First Alphanumeric Value		BLOWING SAND, SOIL, DIRT	
Last Alphanumeric Value		UNKNOWN	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- All 604,028 records (100%) are valid, with no null, empty, or invalid values detected.
- The field contains 12 unique values, indicating moderate variability.
- The shortest value is "SNOW," and the longest is "BLOWING SAND, SOIL, DIRT."
- No records have blanks, leading, or trailing whitespace, ensuring data consistency.
- The dataset is clean and suitable for further analysis.

- **LIGHTING_CONDITION**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	22	
● Ok	604,028		100.00%
Unique	6		0.00%
■ Null	0		0.00%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		4	
Max		22	
Average		10.80	
Shortest Value		DUSK	
Longest Value		DARKNESS, LIGHTED ROAD	
First Alphanumeric Value		DARKNESS	
Last Alphanumeric Value		UNKNOWN	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - All 604,028 records (100%) are valid, with no null, empty, or invalid values detected.
 - The field contains 6 unique values, indicating low variability in the data.
 - The shortest value is "DUSK," and the longest is "DARKNESS, LIGHTED ROAD."
 - No records have blanks, leading, or trailing whitespace, ensuring data consistency.
 - The dataset is clean, structured, and ready for analysis.

- **FIRST_CRASH_TYPE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	28	
Ok	604,028		100.00%
Unique	18		0.00%
Null	0		0.00%
Not Ok	0		0.00%
Empty	0		0.00%
 Length Statistics			
Min			5
Max			28
Average			13.50
Shortest Value			ANGLE
Longest Value			SIDESWIPE OPPOSITE DIRECTION
First Alphanumeric Value			ANGLE
Last Alphanumeric Value			TURNING
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- All 604,028 records (100%) are valid, with no null, empty, or invalid values detected.
- The field contains 36 unique values, indicating low variability.
- 139,375 records (23.07%) are valid, while 464,653 records (76.93%) are null, impacting data completeness.
- The dataset has notable null values, which may need to be addressed for analysis.

- **TRAFFICWAY_TYPE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	31	
● Ok	604,028		100.00%
Unique	20		0.00%
■ Null	0		0.00%
Not	0		0.00%
● Ok	0		0.00%
■ Empty	0		0.00%
Length Statistics			
Min			4
Max			31
Average			14.10
Shortest Value			RAMP
Longest Value			DIVIDED - W/MEDIAN (NOT RAISED)
First Alphanumeric Value			ALLEY
Last Alphanumeric Value			Y-INTERSECTION
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- All 604,028 records (100%) are valid, with no null, empty, or invalid values detected.
 - The field contains 20 unique values, indicating moderate variability.
 - The shortest value is "RAMP," and the longest is "DIVIDED - W/MEDIAN (NOT RAISED)."
 - No records have blanks, leading, or trailing whitespace, ensuring data cleanliness.
 - The dataset is clean, structured, and ready for analysis.

- **LANE_CNT**

- Summary:

Summary			
Type	Records	Data Type	Size
Int32	604,028	4	
● Ok	139,375		23.07%
Unique	36		0.01%
■ Null	464,653		76.93%
Not	0		0.00%
● Ok	0		0.00%
■ Empty	0		0.00%

Grouped Values	
0-1,191,625	
Max	1,191,625
Min	0
Lower quartile	20.286
Upper quartile	60.857
Average	16.3682
Standard deviation	3,490.1977
Sum	2,281,330
Median	40.572
Variance	12,181,480.0228

- Observation:

- All 604,028 records are analyzed, with 139,375 (23.07%) valid values and 464,653 (76.93%) null values, indicating a significant amount of missing data.
- The field contains 36 unique values, representing low variability.
- Null values account for a significant portion, which may require handling before analysis.

- **ALIGNMENT**

- Summary:

Summary		
Type	Records	Data Type Size
String	604,028	21
• Ok	604,028	100.00%
Unique	6	0.00%
• Null	0	0.00%
Not Ok	0	0.00%
• Empty	0	0.00%
Length Statistics		
Min		12
Max		21
Average		17.90
Shortest Value		CURVE, LEVEL
Longest Value		STRAIGHT ON HILLCREST
First Alphanumeric Value		CURVE ON GRADE
Last Alphanumeric Value		STRAIGHT ON HILLCREST
Blanks		0
Values with Leading Whitespace		0
Values with Trailing Whitespace		0

- Observation:

- All 604,028 records (100%) are valid, with no null, empty, or invalid values detected.
- The field contains 6 unique values, indicating low variability.
- Values range from 12 to 21 characters in length, with an average length of 17.9.

- The shortest value is "CURVE, LEVEL," and the longest value is "STRAIGHT ON HILLCREST."
 - No records have blanks, leading, or trailing whitespace, ensuring data cleanliness.
 - The dataset is clean, consistent, and ready for analysis.
- **ROADWAY_SURFACE_COND**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	15	
● Ok	604,028		100.00%
Unique	7		0.00%
■ Null	0		0.00%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			15
Average			3.70
Shortest Value			DRY
Longest Value			SAND, MUD, DIRT
First Alphanumeric Value			DRY
Last Alphanumeric Value			WET
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- All 604,028 records (100%) are valid, with no null, empty, or invalid values detected.
- The field contains 7 unique values, indicating low variability.
- Values range from 3 to 15 characters in length, with an average length of 3.7.
- The shortest value is "DRY," and the longest value is "SAND, MUD, DIRT."
- No records have blanks, leading, or trailing whitespace, ensuring data consistency.
- The dataset is clean, structured, and ready for analysis.

- **ROAD_DEFECT**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	17	
● Ok	604,028		100.00%
Unique	7		0.00%
■ Null	0		0.00%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			5
Max			17
Average			9.50
Shortest Value			OTHER
Longest Value			DEBRIS ON ROADWAY
First Alphanumeric Value			DEBRIS ON ROADWAY
Last Alphanumeric Value			WORN SURFACE
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- All 604,028 records (100%) are valid, with no null, empty, or invalid values detected.
 - The field contains 7 unique values, indicating low variability.
 - The shortest value is "OTHER," and the longest value is "DEBRIS ON ROADWAY."
 - No records have blanks, leading, or trailing whitespace, ensuring data consistency.
 - The dataset is clean, structured, and ready for analysis.

- **REPORT_TYPE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	26	
● Ok	586,949		97.17%
Unique	3		0.00%
■ Null	17,079		2.83%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			7
Max			26
Average			18.10
Shortest Value			AMENDED
Longest Value			NOT ON SCENE (DESK REPORT)
First Alphanumeric Value			AMENDED
Last Alphanumeric Value			ON SCENE
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - 586,949 records (97.17%) are valid, while 17,079 records (2.83%) are null, indicating minimal missing data.
 - The field contains 3 unique values, representing very low variability.
 - The shortest value is "AMENDED," and the longest value is "NOT ON SCENE (DESK REPORT)."
 - No records have blanks, leading, or trailing whitespace, ensuring data consistency.
 - The dataset is mostly complete and suitable for analysis, with minimal null values to address.
- **CRASH_TYPE**
 - Summary:

Summary		
Type	Records	Data Type Size
V_String	604,028	32
Ok	604,028	100.00%
Unique	2	0.00%
Null	0	0.00%
Not	0	0.00%
Ok	0	0.00%
Empty	0	0.00%
Length Statistics		
Min	22	
Max	32	
Average	24.70	
Shortest Value	NO INJURY / DRIVE AWAY	
Longest Value	INJURY AND / OR TOW DUE TO CRASH	
First Alphanumeric Value	INJURY AND / OR TOW DUE TO CRASH	
Last Alphanumeric Value	NO INJURY / DRIVE AWAY	
Blanks	0	
Values with Leading Whitespace	0	
Values with Trailing Whitespace	0	
 - Observation:
 - All 604,028 records (100%) are valid, with no null, empty, or invalid values detected.
 - The field contains 2 unique values, indicating very low variability.
 - The shortest value is "NO INJURY / DRIVE AWAY," and the longest value is "INJURY AND / OR TOW DUE TO CRASH."
 - No records have blanks, leading, or trailing whitespace, ensuring data consistency.
 - The dataset is clean, structured, and ready for analysis.

- **INTERSECTION_RELATED_I**

- Summary:

Summary			
Type	Records	Data Type	Size
String	604,028	1	
● Ok	138,295		22.90%
Unique	2		0.00%
■ Null	465,733		77.10%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		1	
Max		1	
Average		1.00	
Shortest Value		Y	
Longest Value		Y	
First Alphanumeric Value		N	
Last Alphanumeric Value		Y	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- 138,295 records (22.90%) are valid, while 465,733 records (77.10%) are null, indicating a significant portion of missing data.
 - The field contains 2 unique values, representing minimal variability.
 - All valid values have a fixed length of 1 character.
 - No records have blanks, leading, or trailing whitespace, ensuring data consistency.
 - The high null percentage suggests the need for handling missing data before analysis.

- **NOT_RIGHT_OF_WAY_I**

- Summary:

Summary			
Type	Records	Data Type	Size
String	604,028	1	
● Ok	27,951		4.63%
Unique	2		0.00%
■ Null	576,077		95.37%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		1	
Max		1	
Average		1.00	
Shortest Value		Y	
Longest Value		Y	
First Alphanumeric Value		N	
Last Alphanumeric Value		Y	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - 604,028 records with 27,951 valid entries and 576,077 null values (95.37% null).
 - Two unique values ('Y' and 'N'), each of length 1.
 - No blanks, leading, or trailing whitespaces detected.
 - Data requires preprocessing due to the high null proportion.
- **HIT_AND_RUN_I**
 - Summary:

Summary			
Type	Records	Data Type	Size
String	604,028	1	
● Ok	189,090		31.30%
Unique	2		0.00%
■ Null	414,938		68.70%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		1	
Max		1	
Average		1.00	
Shortest Value		Y	
Longest Value		Y	
First Alphanumeric Value		N	
Last Alphanumeric Value		Y	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - The dataset contains 604,028 records, with 189,090 (31.30%) valid entries and 414,938 (68.70%) null values.
 - The valid entries contain 2 unique string values ('Y' and 'N'), representing binary data.
 - The shortest and longest string values are both 'Y' and 'N', ensuring uniform length.
 - No blanks, leading whitespace, or trailing whitespace were found in the valid records, indicating well-formatted data.
 - The high null proportion (68.70%) suggests the need for further investigation or preprocessing for effective use of this field.

- **DAMAGE**

- Summary:

Summary			
Type	Records	Data Type	Size
String	604,028	13	
● Ok	604,028		100.00%
Unique	3		0.00%
■ Null	0		0.00%
Not	0		0.00%
● Ok	0		0.00%
■ Empty	0		0.00%
Length Statistics			
Min		11	
Max		13	
Average		11.60	
Shortest Value		OVER \$1,500	
Longest Value		\$501 - \$1,500	
First Alphanumeric Value		\$500 OR LESS	
Last Alphanumeric Value		OVER \$1,500	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 604,028 records, all valid without any null, blank, or empty values, ensuring data completeness.
- The field contains 3 unique values: \$500 OR LESS, \$501 - \$1,500, and OVER \$1,500, representing categorical ranges.
- The shortest value is OVER \$1,500, and the longest value is \$501 - \$1,500.
- Both leading and trailing whitespaces are absent, ensuring clean formatting.
- The distinct and well-defined categories suggest the data is consistent and standardized for further analysis.

- **DATE_POLICE_NOTIFIED**

- Summary:

Summary ▲			
Type	Records	Data Type	Size
String	604,028	22	
● Ok	604,028		100.00%
Unique	> 10,000		
■ Null	0		0.00%
Not	0		0.00%
● Ok	0		0.00%
■ Empty	0		0.00%
Length Statistics			
Min		22	
Max		22	
Average		22.00	
Shortest Value		08/18/2023 12:55:00 PM	
Longest Value		08/18/2023 12:55:00 PM	
First Alphanumeric Value		01/01/2016 01:00:00 PM	
Last Alphanumeric Value		12/31/2023 12:51:00 PM	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - The dataset contains 604,028 records, all valid with no null, blank, or empty values, ensuring data completeness and reliability.
 - The field contains more than 10,000 unique values, indicating a high degree of variability in the data.
 - The shortest and longest values are 08/18/2023 12:55:00 PM, representing timestamp data.
 - No leading or trailing whitespaces are present, ensuring clean data formatting.
 - The dataset appears well-structured and standardized for temporal analysis or further processing.

- **PRIM_CONTRIBUTORY_CAUSE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	80	
● Ok	604,028		100.00%
Unique	40		0.01%
■ Null	0		0.00%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			6
Max			80
Average			23.70
Shortest Value			ANIMAL
Longest Value			OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIG...
First Alphanumeric Value			ANIMAL
Last Alphanumeric Value			WEATHER
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - The dataset contains 604,028 records, all valid with no null, blank, or empty values, ensuring data completeness.
 - There are 40 unique values, representing categorical diversity.
 - The shortest value is "ANIMAL," and the longest value is "OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIG..." indicating diverse textual descriptions.
 - No leading or trailing whitespaces are present, ensuring clean data formatting.
 - The dataset is well-structured and standardized, suitable for categorical analysis or classification.

- **SEC_CONTRIBUTORY_CAUSE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	80	
● Ok	604,028		100.00%
Unique	40		0.01%
■ Null	0		0.00%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			6
Max			80
Average			19.50
Shortest Value			ANIMAL
Longest Value			OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIG...
First Alphanumeric Value			ANIMAL
Last Alphanumeric Value			WEATHER
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- The dataset consists of 604,028 records, all valid with no null, blank, or empty values, ensuring data integrity.
- There are 40 unique values, indicating a variety of categorical options.
- The shortest value is "ANIMAL," while the longest is "OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIG...".
- No leading or trailing whitespaces exist, ensuring the data is clean and well-formatted.
- The dataset is suitable for text analysis or categorical classification tasks.

- **STREET_NO**

- Summary:

Summary ▲			
Type	Records	Data Type	Size
Int32	604,028	4	
● Ok	604,028		100.00%
Unique	> 10,000		
■ Null	0		0.00%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%

Value Statistics	
Grouped Values	
0-451,100	
Max	451,100
Min	0
Lower quartile	1,246.311
Upper quartile	5,588.061
Average	3,684.0747
Standard deviation	2,904.3305
Sum	2,225,284,285
Median	3,189.525
Variance	8,435,135.8095

- Observation:

- The dataset consists of 604,028 records, all valid with no null, blank, or empty values, ensuring data reliability.
- The field contains more than 10,000 unique values, indicating high variability and potential for detailed analysis.
- The dataset is consistent and clean, requiring minimal preprocessing for analytical tasks

- **STREET_DIRECTION**

- Summary:

Summary		
Type	Records	Data Type Size
String	604,028	1
● Ok	604,025	100.00%
Unique	4	0.00%
■ Null	3	0.00%
Not Ok	0	0.00%
● Empty	0	0.00%
Length Statistics		
Min		1
Max		1
Average		1.00
Shortest Value		W
Longest Value		W
First Alphanumeric Value		E
Last Alphanumeric Value		W
Blanks		0
Values with Leading Whitespace		0
Values with Trailing Whitespace		0

- Observation:

- The dataset consists of 604,028 records, with 604,025 valid entries and 3 null values, ensuring high data reliability.

- The field contains 4 unique values, indicating limited categorical variability.
- The shortest and longest values are both 1 character long, with examples such as "W."
- No blanks, leading, or trailing whitespaces exist, ensuring the data is clean and formatted.
- The dataset is suitable for simple categorical or directional analyses.

- STREET_NAME**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	31	
• Ok	604,027		100.00%
Unique	1,596		0.26%
Null	1		0.00%
Not			
• Ok	0		0.00%
• Empty	0		0.00%
Length Statistics			
Min		4	
Max		31	
Average		10.70	
Shortest Value		82ND	
Longest Value		MICHIGAN AVE 175 E CHESTNUT AVE	
First Alphanumeric Value		100TH DR	
Last Alphanumeric Value		ZEMKE RD	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

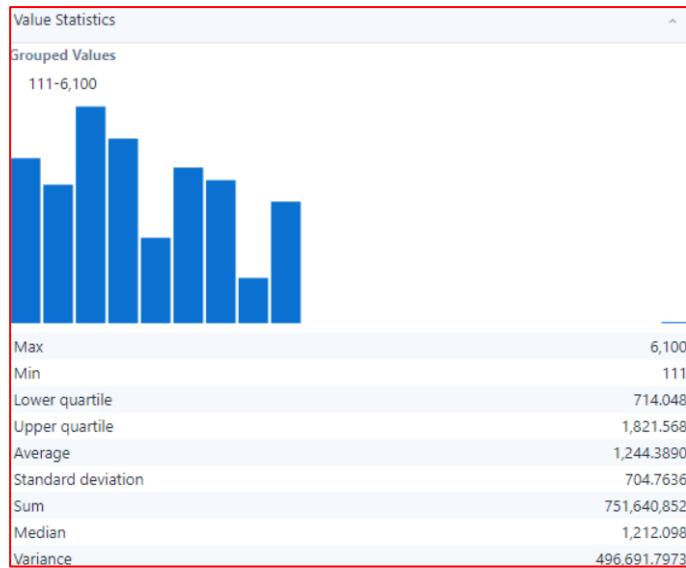
- Observation:

- The dataset consists of 604,028 records, with 604,027 valid entries and 1 null value, ensuring high data integrity.
- There are 1,596 unique values, indicating moderate variability in the data.
- The shortest value is "82ND," and the longest is "MICHIGAN AVE 175 E CHESTNUT AVE."
- No blanks, leading, or trailing whitespaces exist, ensuring the data is clean and well-formatted.
- The dataset is suitable for address-related analysis or location-based categorization.

- BEAT_OF_OCCURANCE**

- Summary:

Summary			
Type	Records	Data Type	Size
Int16	604,028	2	
• Ok	604,024		100.00%
Unique	276		0.05%
Null	4		0.00%
Not			
• Ok	0		0.00%
• Empty	0		0.00%



- Observation:
 - The dataset consists of 604,028 records, with 604,024 valid entries and 4 null values.
 - The field contains 276 unique values, representing a diverse range of data points.
 - This dataset is appropriate for statistical analysis and modeling, given its numeric distribution and variability.

● **PHOTOS_TAKEN_I**

- Summary:

The figure displays a summary table for the 'PHOTOS_TAKEN_I' field with a red border. It includes sections for Type, Records, Data Type, Size, and various status counts:

Type	Records	Data Type	Size
String	604,028	1	
● Ok	8,058		1.33%
Unique	2		0.00%
■ Null	595,970		98.67%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			1
Max			1
Average			1.00
Shortest Value			Y
Longest Value			Y
First Alphanumeric Value			N
Last Alphanumeric Value			Y
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - The dataset consists of 604,028 records, with 8,058 valid entries and 595,970 null values, indicating a high percentage (98.67%) of missing data.

- Only 2 unique values exist within the valid entries, making it a binary or categorical dataset.
- There are no leading or trailing whitespaces, ensuring clean formatting.
- The significant proportion of null values suggests that this field may not be critical or requires imputation strategies for analysis.

STATEMENT_TAKEN_I

- Summary:

Summary			
Type	Records	Data Type	Size
String	604,028	1	
● Ok	13,706		2.27%
Unique	2		0.00%
● Null	590,322		97.73%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		1	
Max		1	
Average		1.00	
Shortest Value		Y	
Longest Value		Y	
First Alphanumeric Value		N	
Last Alphanumeric Value		Y	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset consists of 604,028 records, with 13,706 valid entries and 590,322 null values, accounting for 97.73% missing data.
- There are only 2 unique valid values, making it a binary or categorical dataset.
- No leading or trailing whitespaces are present, indicating well-formatted data.
- The high percentage of null values suggests the field may not be critical or requires imputation or alternative strategies for effective analysis.

- **DOORING_I**

- Summary:

Summary			
Type	Records	Data Type	Size
String	604,028	1	
● Ok	1,879		0.31%
Unique	2		0.00%
■ Null	602,149		99.69%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		1	
Max		1	
Average		1.00	
Shortest Value		Y	
Longest Value		Y	
First Alphanumeric Value		N	
Last Alphanumeric Value		Y	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset consists of 604,028 records, with only 1,879 valid entries, accounting for 0.31% of the dataset, and 602,149 null values, making up 99.69%.
- There are 2 unique valid values, indicating binary or categorical data.
- No leading or trailing whitespaces are present, ensuring the data is clean.
- The extremely high percentage of null values highlights the need for imputation, removal, or strategic handling during analysis.

- **WORK_ZONE_I**

- Summary:

Summary			
Type	Records	Data Type	Size
String	604,028	1	
● Ok	3,427		0.57%
Unique	2		0.00%
■ Null	600,601		99.43%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		1	
Max		1	
Average		1.00	
Shortest Value		Y	
Longest Value		Y	
First Alphanumeric Value		N	
Last Alphanumeric Value		Y	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - The dataset consists of 604,028 records, with 3,427 valid entries (0.57%) and 600,601 null values (99.43%).
 - There are 2 unique valid values, indicating binary or categorical data.
 - No leading or trailing whitespaces exist, ensuring clean data formatting.
 - The high percentage of null values indicates a significant data sparsity issue that requires attention during preprocessing.

- **WORK_ZONE_TYPE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	604,028	12	
Ok	2,639		0.44%
Unique	4		0.00%
Null	601,389		99.56%
Not Ok	0		0.00%
Empty	0		0.00%
Length Statistics			
Min			7
Max			12
Average			10.90
Shortest Value		UTILITY	
Longest Value		CONSTRUCTION	
First Alphanumeric Value		CONSTRUCTION	
Last Alphanumeric Value		UTILITY	
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- The dataset consists of 604,028 records, with 2,639 valid entries (0.44%) and 601,389 null values (99.56%).
- There are 4 unique valid values, suggesting limited categorical data.
- The shortest value is "UTILITY," and the longest value is "CONSTRUCTION."
- No leading or trailing whitespaces exist, ensuring clean data formatting.
- The high percentage of null values indicates significant sparsity in the dataset, which should be addressed during data processing.

- **WORKERS_PRESENT_I**

- Summary:

Summary			
Type	Records	Data Type	Size
String	604,028	1	
● Ok	898		0.15%
Unique	2		0.00%
■ Null	603,130		99.85%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		1	
Max		1	
Average		1.00	
Shortest Value		Y	
Longest Value		Y	
First Alphanumeric Value		N	
Last Alphanumeric Value		Y	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

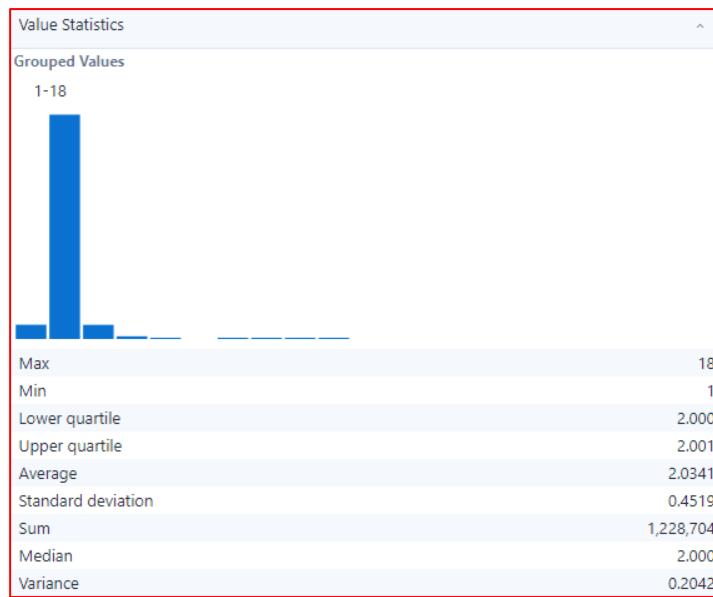
- Observation:

- The dataset consists of 604,028 records, with 898 valid entries (0.15%) and 603,130 null values (99.85%).
 - There are 2 unique valid values, indicating minimal categorical variation.
 - No leading or trailing whitespaces exist, ensuring clean and consistent data formatting.
 - The extremely high percentage of null values highlights sparsity, requiring careful consideration during data processing or analysis.

- **NUM_UNITS**

- Summary:

Summary			
Type	Records	Data Type	Size
Byte	604,028	1	
● Ok	604,028		100.00%
Unique	15		0.00%
■ Null	0		0.00%
Not			
● Ok	0		0.00%
● Empty	0		0.00%



- Observation:
 - The dataset contains 604,028 records, all valid with no null, blank, or empty values, ensuring high data integrity.
 - There are 15 unique values, offering limited variance but suitable for categorical analysis.
 - This dataset is clean and ready for analysis involving grouped or categorical numerical values.

- **MOST_SEVERE_INJURY**

- Summary:

Summary			
Type	Records	Data Type	Size
String	604,028	24	
● Ok	602,601		99.76%
Unique	5		0.00%
● Null	1,427		0.24%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			5
Max			24
Average			22.90
Shortest Value			FATAL
Longest Value			NONINCAPACITATING INJURY
First Alphanumeric Value			FATAL
Last Alphanumeric Value			REPORTED, NOT EVIDENT
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

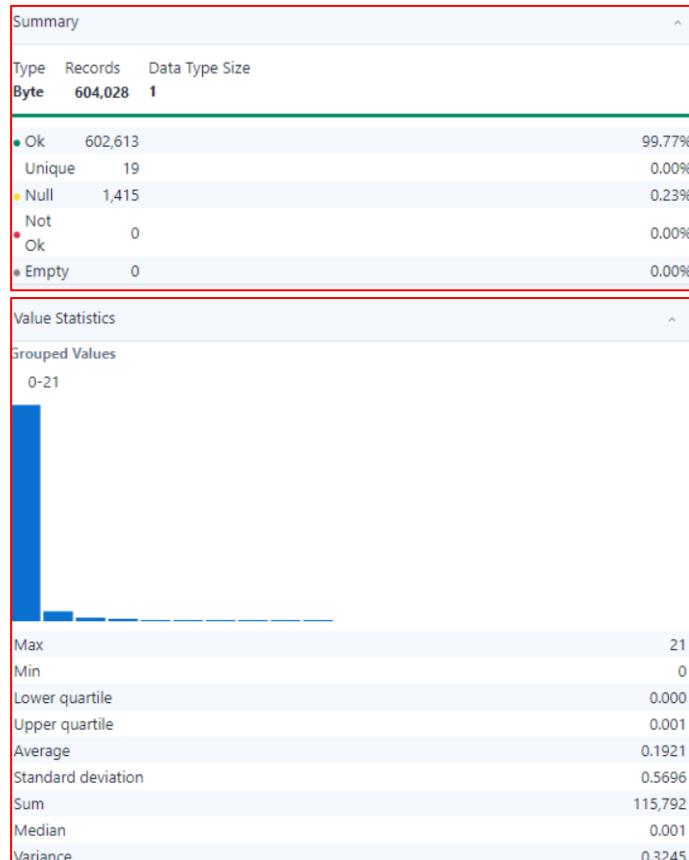
- The dataset consists of 604,028 records, with 99.76% valid and 0.24% null values (1,427 records).
- There are 5 unique values, providing limited categorical diversity.

The shortest value is "FATAL," while the longest value is "NONINCAPACITATING INJURY."

- No leading or trailing whitespaces or blank entries are present, ensuring clean formatting.
- The dataset is structured and suitable for text-based categorical analysis, with minor handling required for null values.

- **INJURY_TOTAL**

- Summary:

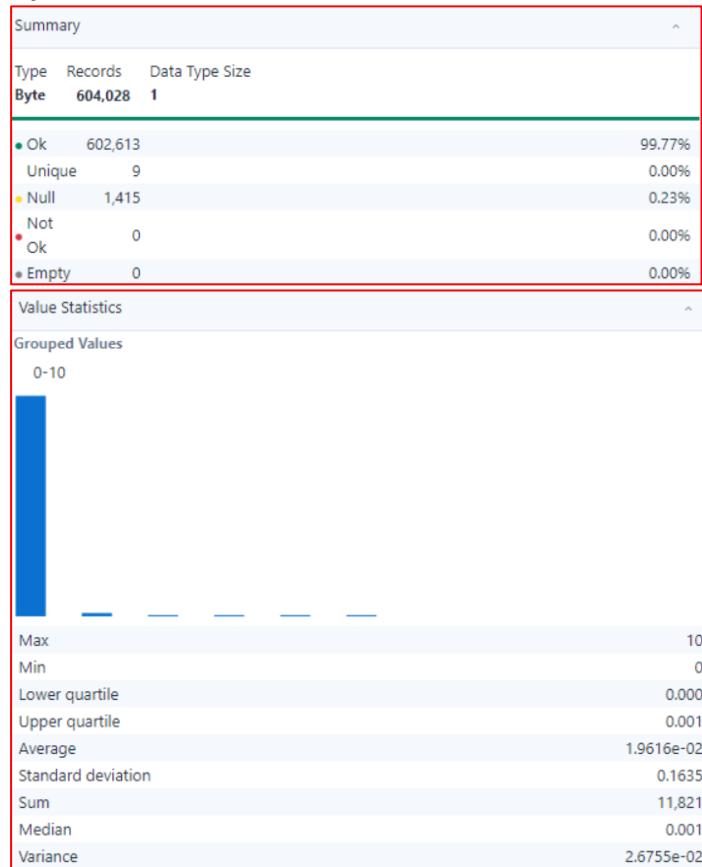


- Observation:

- The dataset contains 604,028 records, of which 602,613 (99.77%) are valid and 1,415 (0.23%) are null.
- There are 19 unique values present in the dataset, indicating moderate categorical diversity.
- Data integrity is maintained with no empty or incorrect records, and the dataset is suitable for statistical analysis or categorical modeling.

- **INJURIES_INCAPICITATING**

- Summary:

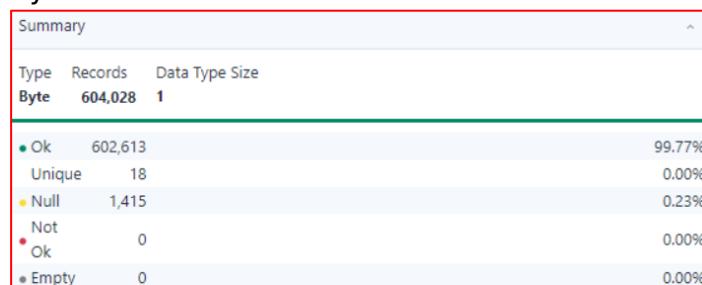


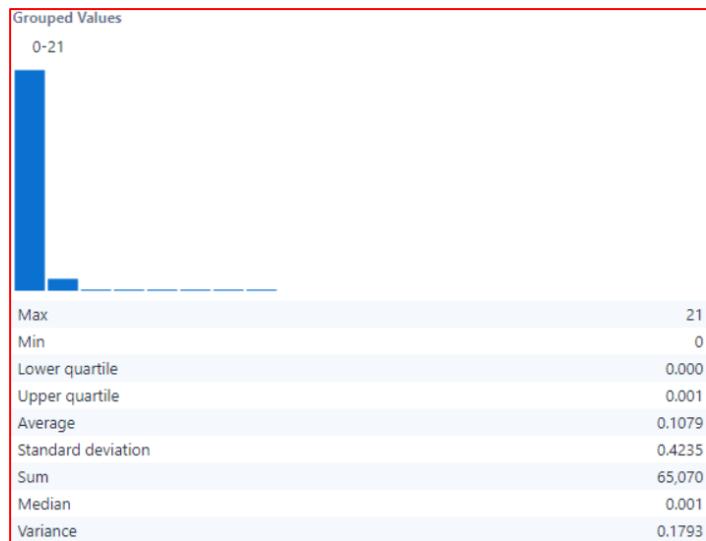
- Observation:

- The dataset contains 604,028 records, with 602,613 valid entries (99.77%) and 1,415 null entries (0.23%).
 - There are 9 unique values in the dataset, indicating limited variability in the data.
 - The dataset is clean and suitable for analysis, but the null values should be handled appropriately for accurate results.

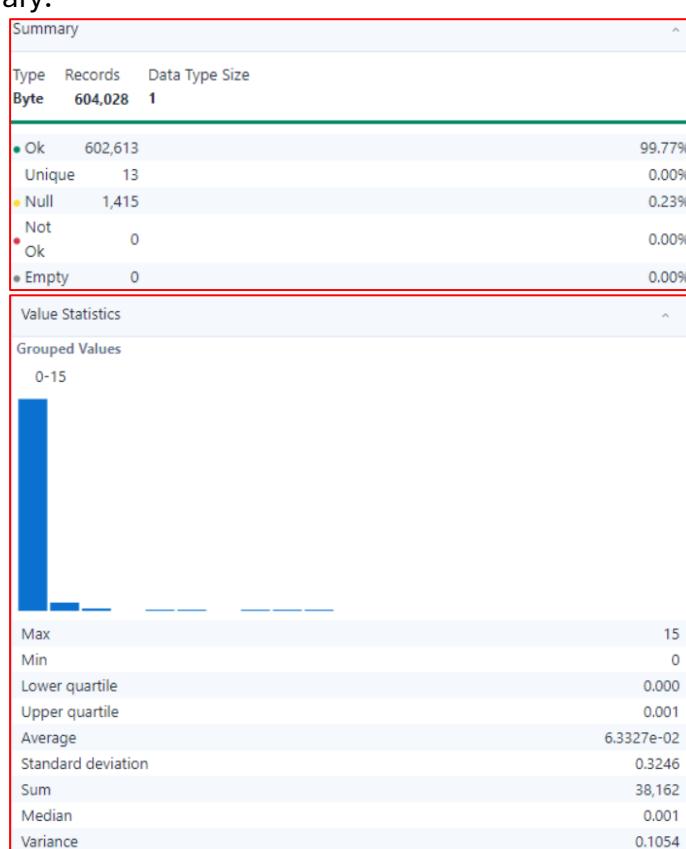
- **INJURIES_NON_INCAPICITATING**

- Summary:

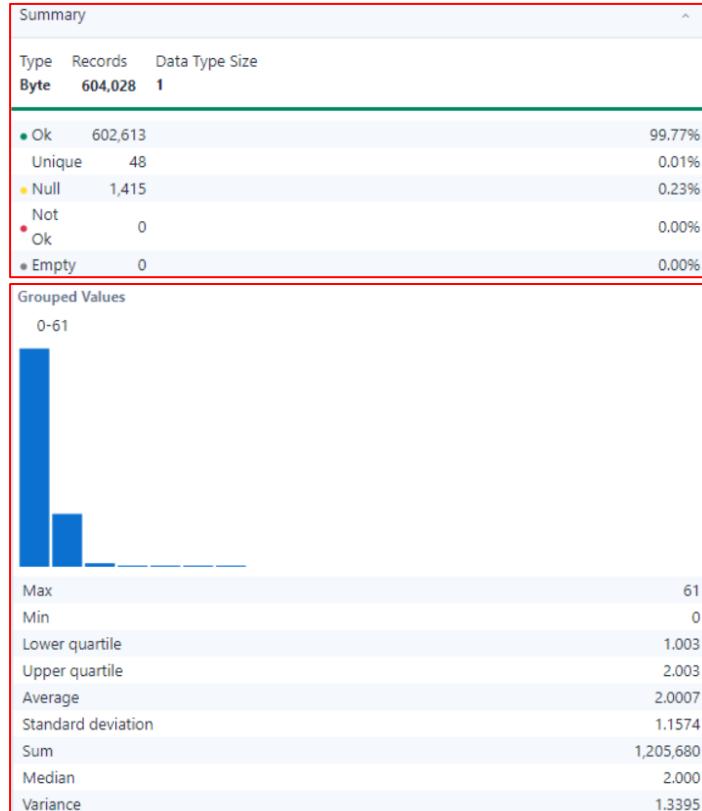




- Observation:
 - The dataset contains 604,028 records with 99.77% valid entries and 0.23% null values.
 - There are 18 unique values observed, indicating moderate categorical diversity.
 - The variance of 0.1793 further supports the minimal spread of values.
- **INJURIES_REPORTED_NON_EVIDENT**
 - Summary:



- Observation:
 - The dataset contains 604,028 records, with 99.77% valid records and 0.23% null values (1,415 records).
 - There are 13 unique values, indicating a limited range of categories.
 - The dataset appears clean and ready for categorical or statistical analysis.
- **INJURIES_NO_INDICATION**
- Summary:



- Observation:
 - The dataset contains 604,028 records, with 99.77% valid records and 0.23% null values (1,415 records).
 - There are 48 unique values, indicating moderate diversity in categories.
 - The dataset appears ready for further categorical or numerical analysis.

- **INJURIES_UNKNOWN**

- Summary:

Type	Records	Data Type	Size
Byte	604,028	1	
● Ok	602,613		99.77%
Unique	1		0.00%
● Null	1,415		0.23%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Value Statistics			
Grouped Values			
Only one value	0		
Max		0	
Min		0	
Lower quartile		0.000	
Upper quartile		0.000	
Average		0	
Standard deviation		0	
Sum		0	
Median		0.000	
Variance		0	

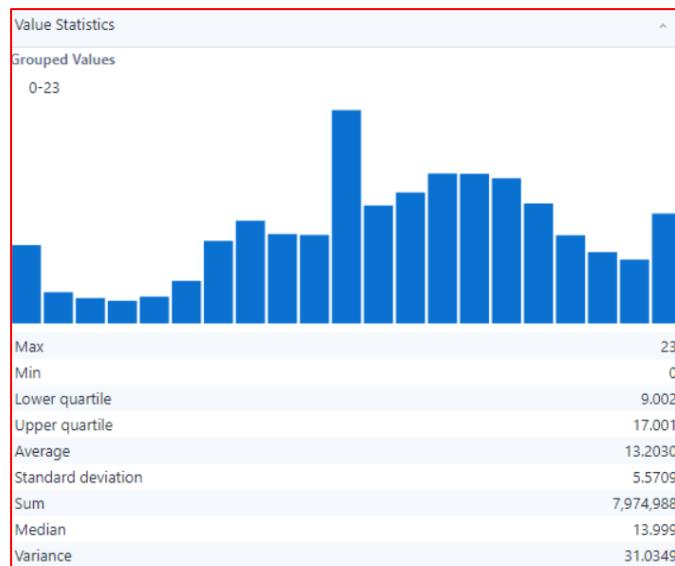
- Observation:

- The dataset consists of 604,028 records, with 99.77% valid and 0.23% null values (1,415 records).
 - There is only one unique value in the dataset, indicating no variability in categorical data.
 - The dataset does not provide meaningful numerical distribution and requires contextual understanding for further use.

- **CRASH_HOUR**

- Summary:

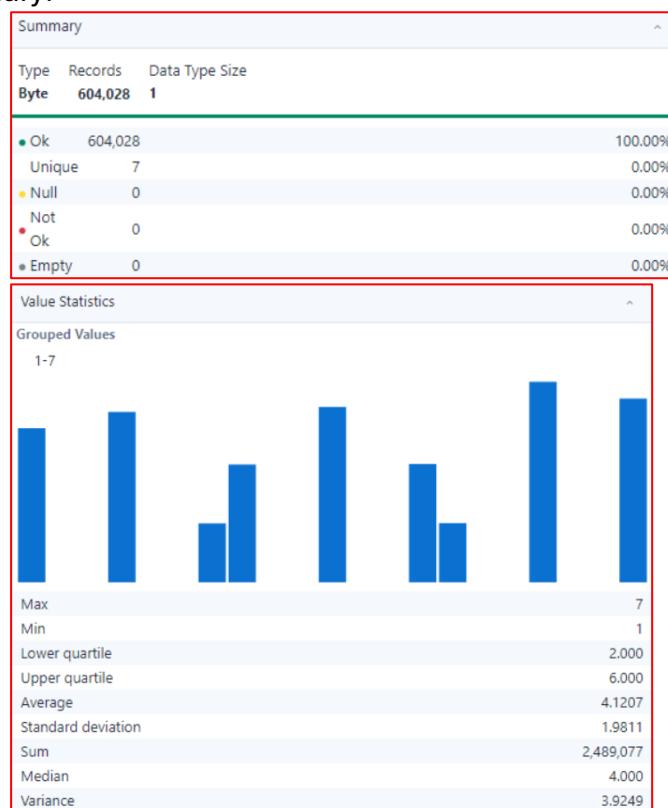
Summary			
Type	Records	Data Type	Size
Byte	604,028	1	
● Ok	604,028		100.00%
Unique	24		0.00%
● Null	0		0.00%
Not			
● Ok	0		0.00%
● Empty	0		0.00%



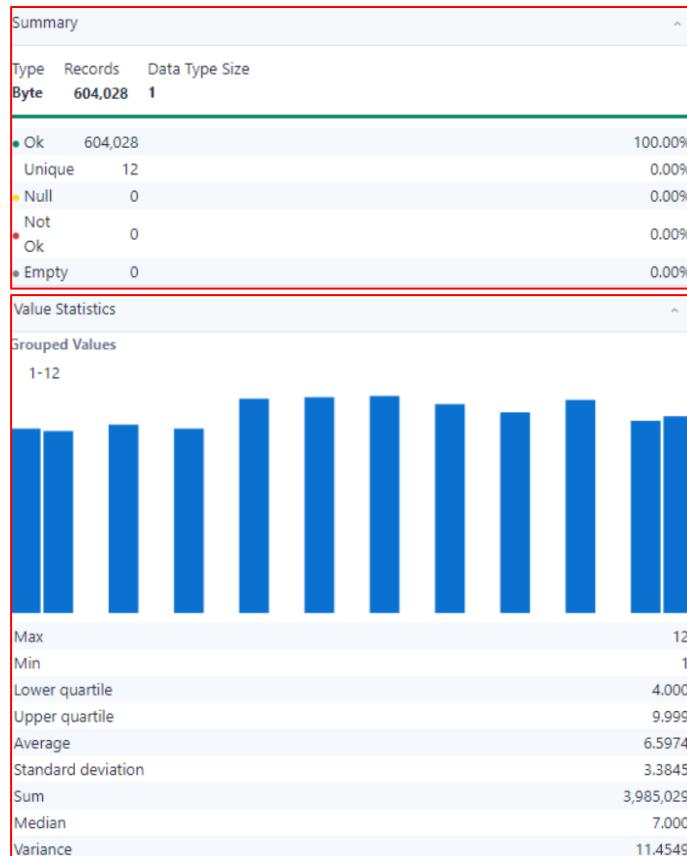
- Observation:
 - The dataset contains 604,028 records, all valid with no null or empty values, ensuring complete data integrity.
 - There are 24 unique values, indicating a moderate range of categorical options.
 - The variance is 31.0349, indicating a reasonable distribution suitable for analysis.

● CRASH_DAY_OF_WEEK

- Summary:



- Observation:
 - The dataset contains 604,028 records, all valid without null, empty, or blank values, ensuring data integrity.
 - There are 7 unique values, indicating a limited range of categories.
 - The dataset is clean and ready for numerical or categorical analysis.
- **CRASH_MONTH**
- Summary:



- Observation:
 - The dataset contains 604,028 records, all valid with no null values, ensuring data completeness.
 - There are 12 unique values, indicating a moderate level of categorical diversity.
 - The dataset appears well-structured and suitable for further categorical or numerical analysis.

- **LONGITUDE**

- Summary:

Summary ▲		
Type	Records	Data Type Size
Double	604,028	8
● Ok	597,925	98.99%
Unique	> 10,000	
● Null	6,103	1.01%
Not	0	0.00%
● Ok	0	0.00%
● Empty	0	0.00%

Value Statistics		
Grouped Values		
-87.93-0		
Max	0	
Min	-87.9361	
Lower quartile	-87.721	
Upper quartile	-87.634	
Average	-87.6718	
Standard deviation	0.7877	
Sum	-52,421,167.7203	
Median	-87.674	
Variance	0.6205	

- Observation:

- The dataset contains 604,028 records, with 98.99% valid records and 1.01% null values (6,103 records).
 - There are over 10,000 unique values, indicating a high degree of granularity.
 - The dataset is suitable for numerical analysis and contains mostly complete data with minimal null values.

- **LATITUDE**

- Summary:

Summary ▲		
Type	Records	Data Type Size
Double	604,028	8
● Ok	597,925	98.99%
Unique	> 10,000	
● Null	6,103	1.01%
Not	0	0.00%
● Ok	0	0.00%
● Empty	0	0.00%

Value Statistics	
Grouped Values	
0-42.02	
Max	42.0227
Min	0
Lower quartile	41.783
Upper quartile	41.924
Average	41.8542
Standard deviation	0.3847
Sum	25,025,674.6072
Median	41.875
Variance	0.1480

- Observation:

- The dataset consists of 604,028 records, with 98.99% valid records and 1.01% null values (6,103 records).
- The field contains more than 10,000 unique values, indicating high variability.
- The variance is 0.1480, indicating a concentrated spread of values within the range.

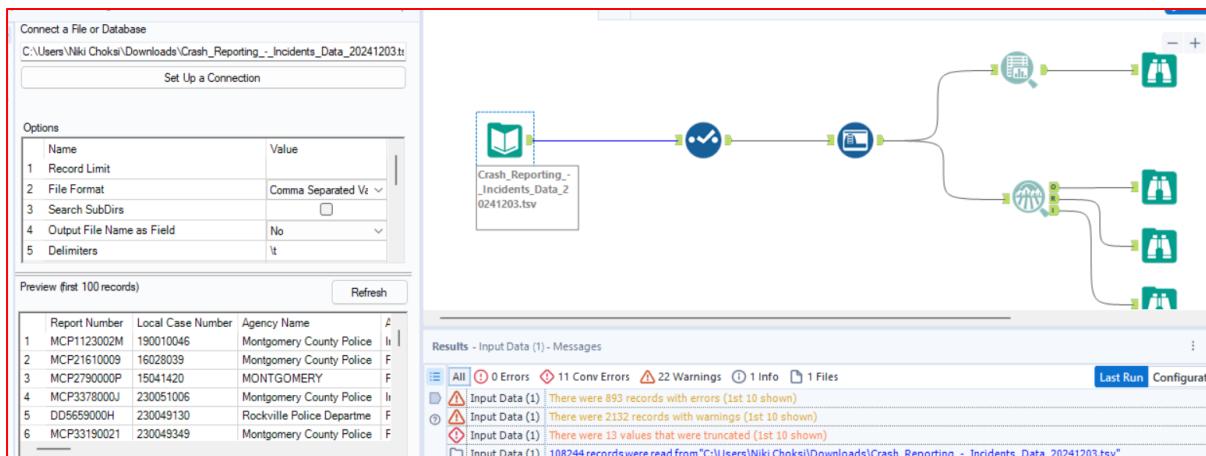
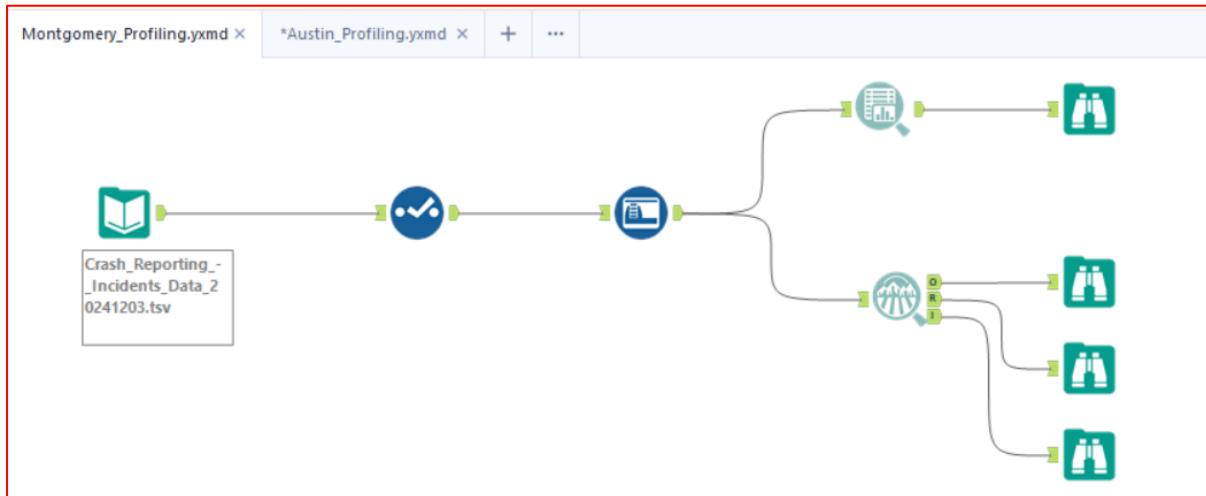
- LOCATION

- Summary:

Summary ▲		
Type	Records	Data Type Size
String	604,028	40
● Ok	597,925	98.99%
Unique	> 10,000	
● Null	6,103	1.01%
Not	0	0.00%
Ok	0	0.00%
● Empty	0	0.00%
Length Statistics		
Min		11
Max		40
Average		39.80
Shortest Value		POINT (0 0)
Longest Value		POINT (-87.665902342962 41.854120262952)
First Alphanumeric Value		POINT (-87.524587386649 41.703272315652)
Last Alphanumeric Value		POINT (0 0)
Blanks		0
Values with Leading Whitespace		0
Values with Trailing Whitespace		0

- Observation:
 - The dataset consists of 604,028 records, with 98.99% valid records and 1.01% null values (6,103 records).
 - The field contains more than 10,000 unique values, indicating high variability.
 - The shortest value is "POINT (0 0)," and the longest value is "POINT (-87.665902342962 41.854120262952)."
 - No leading or trailing whitespaces or blanks are present, ensuring clean data.
 - The dataset is suitable for geospatial or coordinate-based analysis.

Montgomery Dataset



Overview of the Dataset:

The Montgomery dataset contains crash data with details about road conditions, lane types, and involved parties. It highlights unique identifiers, crash types, and substance abuse factors. The dataset is useful for understanding crash causation, improving road infrastructure, and enhancing traffic safety.

Field Name	Data Type	Description
REPORT_NUMBER	Text	Unique identifier assigned to each crash report.
LOCAL_CASE_NUMBER	Text	Identifier used by the local agency for internal tracking of the case.
AGENCY_NAME	Text	Name of the law enforcement agency that reported the incident.

ACRS_REPORT_TYPE	Text	Type of report generated by the Automated Crash Reporting System (ACRS).
CRASH_DATE_TIME	Floating Timestamp	Date and time when the crash occurred.
HIT_RUN	Text	Indicates whether the crash was a hit-and-run incident.
ROUTE_TYPE	Text	Classification of the road where the crash occurred (e.g., interstate, state road).
LANE_DIRECTION	Text	Direction of the lane where the crash occurred (e.g., northbound, southbound).
LANE_TYPE	Text	Specific type of lane involved in the crash (e.g., through lane, turn lane).
NUMBER_OF_LANES	Text	Total number of lanes in the roadway at the crash location.
DIRECTION	Text	General compass direction of travel at the time of the crash.
DISTANCE	Number	Distance from a reference point to the crash location.
DISTANCE_UNIT	Text	Unit of measurement for the distance (e.g., feet, miles).
ROAD_GRADE	Text	Grade or slope of the road at the crash site (e.g., level, uphill).
ROAD_NAME	Text	Name of the road where the crash occurred.
CROSS_STREET_NAME	Text	Name of the nearest cross-street to the crash location.
OFF_ROAD_DESCRIPTION	Text	Description of the crash location if it

		occurred off the main roadway.
MUNICIPALITY	Text	Name of the municipality where the crash occurred.
RELATED_NON_MOTORIST	Text	Indicates involvement of non-motorists (e.g., pedestrians, cyclists) in the crash.
AT_FAULT	Text	Identifies the party at fault in the crash.
COLLISION_TYPE	Text	Describes the type of collision (e.g., rear-end, sideswipe).
WEATHER	Text	Weather conditions at the time of the crash.
SURFACE_CONDITION	Text	Condition of the road surface during the crash (e.g., dry, wet).
LIGHT	Text	Lighting conditions at the time of the crash (e.g., daylight, dark).
TRAFFIC_CONTROL	Text	Type of traffic control device present at the crash location (e.g., stop sign, traffic signal).
DRIVER_SUBSTANCE_ABUSE	Text	Indicates whether driver substance abuse was a factor in the crash.
NON_MOTORIST_SUBSTANCE_ABUSE	Text	Indicates whether substance abuse by a non-motorist was a factor in the crash.
FIRST_HARMFUL_EVENT	Text	The initial event during the crash that caused injury or damage.
SECOND_HARMFUL_EVENT	Text	The subsequent event during the crash that caused injury or damage.
JUNCTION	Text	Indicates if the crash occurred at or near a junction (e.g., intersection, interchange).

INTERSECTION_TYPE	Text	Type of intersection where the crash occurred, if applicable.
ROAD_ALIGNMENT	Text	Alignment of the road at the crash location (e.g., straight, curve).
ROAD_CONDITION	Text	Overall condition of the road at the time of the crash.
ROAD_DIVISION	Text	Describes the division of the road (e.g., divided, undivided).
LATITUDE	Number	Latitude coordinate of the crash location.
LONGITUDE	Number	Longitude coordinate of the crash location.
LOCATION	Location	Combined geographic information (latitude and longitude) of the crash site.

- **Data Quality Analysis(MONGOMERY)**

By the reference of the 5Cs of data

Measure	Importance	Required Insights
Clean	Ensures that the data is free from errors, irrelevant entries, and is formatted correctly.	Check for and remove null or invalid values in critical fields like REPORT NUMBER, CRASH DATE/TIME, and LOCATION.
Consistent	Verifies that data is logically coherent with uniformity across datasets.	Ensure fields like ROAD NAME and CROSS-STREET NAME follow consistent naming conventions and standard formats.
Comprehensive	Assesses the extent to which data covers all necessary aspects and elements.	Confirm all mandatory fields, such as AT FAULT, COLLISION TYPE, WEATHER, and SURFACE CONDITION, are fully populated for comprehensive analysis.
Confirmed	Validates that data is accurate and verified against reliable sources.	Cross-reference LATITUDE and LONGITUDE with

		reliable mapping services to verify location accuracy.
Current	Confirms that the dataset is up-to-date and relevant for the intended analysis.	Ensure CRASH DATE/TIME reflects recent crash incidents and identify outdated records for potential exclusion or updating.

- **Field Analysis**

Field Name	Description	Analysis
REPORT_NUMBER	Unique identifier assigned to each crash report.	Ensure uniqueness and no null values to maintain data integrity.
LOCAL_CASE_NUMBER	Identifier used by the local agency for internal tracking of the case.	Validate for uniqueness and completeness to ensure accurate internal tracking.
AGENCY_NAME	Name of the law enforcement agency that reported the incident.	Cross-validate entries against known law enforcement agencies to ensure accuracy.
ACRS_REPORT_TYPE	Type of report generated by the Automated Crash Reporting System (ACRS).	Validate against predefined report types for consistency.
CRASH_DATE_TIME	Date and time when the crash occurred.	Ensure proper timestamp format and alignment with other

		time-related fields.
HIT_RUN	Indicates whether the crash was a hit-and-run incident.	Validate binary entries (e.g., 'Y' or 'N') and assess for null values.
ROUTE_TYPE	Classification of the road where the crash occurred (e.g., interstate, state road).	Standardize categories and cross-reference with local road classification schemes.
LANE_DIRECTION	Direction of the lane where the crash occurred (e.g., northbound, southbound).	Ensure valid directional entries (e.g., northbound, southbound) and logical consistency with location.
LANE_TYPE	Specific type of lane involved in the crash (e.g., through lane, turn lane).	Validate values against standard lane type categories (e.g., through lane, turn lane).
NUMBER_OF_LANES	Total number of lanes in the roadway at the crash location.	Identify anomalies such as excessively high or missing lane counts.
DIRECTION	General compass direction of travel at the time of the crash.	Ensure valid compass directions and cross-check for logical consistency.

		with lane data.
DISTANCE	Distance from a reference point to the crash location.	Check for realistic values and ensure consistency with the unit of measurement .
DISTANCE_UNIT	Unit of measurement for the distance (e.g., feet, miles).	Validate against expected units (e.g., feet, miles) and ensure consistency with DISTANCE
ROAD_GRADE	Grade or slope of the road at the crash site (e.g., level, uphill).	Cross-validate values (e.g., level, uphill) with other road condition factors.
ROAD_NAME	Name of the road where the crash occurred.	Cross-reference with GIS or official mapping data for accuracy.
CROSS_STREET_NAME	Name of the nearest cross-street to the crash location.	Validate against intersection data and GIS mapping for accuracy.
OFF_ROAD_DESCRIPTION	Description of the crash location if it occurred off the main roadway.	Ensure valid and complete descriptions of off-road crash locations.

MUNICIPALITY	Name of the municipality where the crash occurred.	Validate against known municipality names in the region.
RELATED_NON_MOTORIST	Indicates involvement of non-motorists (e.g., pedestrians, cyclists) in the crash.	Confirm logical consistency with other crash factors and standardize categories.
AT_FAULT	Identifies the party at fault in the crash.	Ensure completeness and logical alignment with other contributory cause fields.
COLLISION_TYPE	Describes the type of collision (e.g., rear-end, sideswipe).	Cross-check for logical consistency with crash details (e.g., lane type, road alignment).
WEATHER	Weather conditions at the time of the crash.	Standardize weather condition values and handle missing data.
SURFACE_CONDITION	Condition of the road surface during the crash (e.g., dry, wet).	Validate surface conditions (e.g., dry, wet, icy) for accuracy and completeness.
LIGHT	Lighting conditions at the time of the crash (e.g., daylight, dark).	Ensure valid lighting condition entries (e.g.,

		daylight, dark, dawn).
TRAFFIC_CONTROL	Type of traffic control device present at the crash location (e.g., stop sign, traffic signal).	Validate against a predefined list of traffic control devices (e.g., stop sign, traffic signal).
DRIVER_SUBSTANCE_ABUSE	Indicates whether driver substance abuse was a factor in the crash.	Investigate null values and validate against crash details for logical consistency.
NON_MOTORIST_SUBSTANCE_ABUSE	Indicates whether substance abuse by a non-motorist was a factor in the crash.	Validate values and assess frequency of reported cases.
FIRST_HARMFUL_EVENT	The initial event during the crash that caused injury or damage.	Ensure logical alignment with crash type and location data.
SECOND_HARMFUL_EVENT	The subsequent event during the crash that caused injury or damage.	Validate consistency with the first harmful event and other related crash fields.
JUNCTION	Indicates if the crash occurred at or near a junction (e.g., intersection, interchange).	Ensure binary entries ('Y' or 'N') and cross-check with INTERSECTION_TYPE
INTERSECTION_TYPE	Type of intersection where the crash occurred, if applicable.	Validate against known intersection types and

		ensure alignment with location data.
ROAD_ALIGNMENT	Alignment of the road at the crash location (e.g., straight, curve).	Check for logical consistency between road alignment and crash type.
ROAD_CONDITION	Overall condition of the road at the time of the crash.	Ensure completeness and validate against predefined condition categories.
ROAD_DIVISION	Describes the division of the road (e.g., divided, undivided).	Confirm valid entries (e.g., divided, undivided) and logical consistency with road layout.
LATITUDE	Latitude coordinate of the crash location.	Validate within acceptable geographic boundaries.
LONGITUDE	Longitude coordinate of the crash location.	Cross-check with LATITUDE for logical location accuracy
LOCATION	Combined geographic information (latitude and longitude) of the crash site.	Ensure this field accurately reflects the latitude and longitude values.

Data Observation:

- **REPORT_NUMBER**

- Summary:

Summary ▲			
Type	Records	Data Type	Size
V_String	108,242	62	
● Ok	108,136		99.90%
Unique	> 10,000		
■ Null	100		0.09%
Not	0		0.00%
● Ok	0		0.00%
● Empty	6		0.01%
Length Statistics ▲			
Min		1	
Max		62	
Average		11.00	
Shortest Value		6	
Longest Value	(DRIVE LOCATED BETWEEN PARKING LOT AND ""KISS AND RID...		
First Alphanumeric Value		[Null]	
Last Alphanumeric Value	the Montgomery County Community College"		
Blanks		6	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset has 108,242 records, with 99.90% valid and 0.09% null values (100 records).
 - There are over 10,000 unique values, indicating a diverse dataset.
 - There are 6 blanks, with no values containing leading or trailing whitespaces.
 - The dataset appears clean and ready for further analysis.

- **AGENCY_NAME**

- Summary:

Summary			
Type	Records	Data Type	Size
String	108,242	25	
● Ok	107,139		98.98%
Unique	39		0.04%
■ Null	1,103		1.02%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics ▲			
Min		3	
Max		25	
Average		22.00	
Shortest Value		MD"	
Longest Value	Rockville Police Departme		
First Alphanumeric Value	20866"		
Last Alphanumeric Value	Takoma Park Police Depart		
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - The dataset contains 108,242 records, with 98.98% valid and 1.02% null values (1,103 records).
 - There are 39 unique values, indicating limited categorical diversity.
 - There are no blanks, and no values have leading or trailing whitespaces.
 - The dataset is mostly clean and suitable for categorical analysis.
- **ACRS_REPORT_TYPE**
- Summary:

Summary		
Type	Records	Data Type Size
String	108,242	22
● Ok	107,592	99.40%
Unique	16	0.01%
■ Null	650	0.60%
Not	0	0.00%
● Ok	0	0.00%
● Empty	0	0.00%
 Length Statistics		
Min	4	
Max	22	
Average	17.80	
Shortest Value	BOTH	
Longest Value	MACHINE OPERATOR/RIDER	
First Alphanumeric Value	BICYCLIST	
Last Alphanumeric Value	Unknown	
Blanks	0	
Values with Leading Whitespace	0	
Values with Trailing Whitespace	0	

- Observation:
 - The dataset contains 108,242 records, with 99.40% valid and 0.60% null values (650 records).
 - There are 16 unique values, indicating limited diversity in categories.
 - There are no blanks, and no values have leading or trailing whitespaces.
 - The dataset appears clean and ready for categorical analysis.
- **CRASH_DATE/TIME**
- Summary:

Summary		
Type	Records	Data Type Size
V_String	108,242	72
● Ok	107,848	99.64%
Unique > 10,000		
■ Null	394	0.36%
Not	0	0.00%
● Ok	0	0.00%
● Empty	0	0.00%
 Length Statistics		
Min	3	
Max	72	
Average	19.60	
Shortest Value	N/A	
Longest Value	Divided, Depressed Median, Divided, Flush Median (greater tha...	
First Alphanumeric Value	01-01-2015 00:30	
Last Alphanumeric Value	Unknown	
Blanks	0	
Values with Leading Whitespace	0	
Values with Trailing Whitespace	0	

- Observation:
 - The dataset contains 108,242 records, with 99.64% valid and 0.36% null values (394 records).
 - The unique value count exceeds 10,000, indicating a high diversity in data.
 - There are no blanks, and no values have leading or trailing whitespaces.
 - The dataset appears clean and suitable for text-based or categorical analysis.

- HIT/RUN

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	29	
● Ok	106,117		98.04%
Unique	91		0.08%
● Null	2,125		1.96%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		2	
Max		29	
Average		2.20	
Shortest Value		No	
Longest Value		Sideswipe, Opposite Direction	
First Alphanumeric Value		38.97488807	
Last Alphanumeric Value		Yes	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 98.04% valid and 1.96% null values (2,125 records).
- There are 91 unique values, indicating moderate diversity in the dataset.
- There are no blanks, and no values have leading or trailing whitespaces.
- The dataset is mostly complete and suitable for categorical or descriptive analysis.

- **ROUTE_TYPE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	25	
● Ok	93,220		86.12%
Unique	90		0.08%
■ Null	15,022		13.88%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		3	
Max		25	
Average		11.90	
Shortest Value		N/A	
Longest Value		SAME DIRECTION RIGHT TURN	
First Alphanumeric Value		-76.93824826	
Last Alphanumeric Value		Unknown	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 86.12% valid values and 13.88% null values (15,022 records).
- There are 90 unique values, indicating moderate variability in the dataset.
- There are no blank values, and no values have leading or trailing whitespaces.
- The dataset shows some missing data but is otherwise suitable for text analysis.

- **LANE_TYPE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	45	
● Ok	19,774		18.27%
Unique	163		0.15%
■ Null	88,468		81.73%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		3	
Max		45	
Average		10.20	
Shortest Value		N/A	
Longest Value		Lane 1, Left Turn Lane, Not Applicable, Other	
First Alphanumeric Value		ACCELERATION LANE	
Last Alphanumeric Value		Unknown	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - The dataset contains 108,242 records, with 19,774 valid records (18.27%) and 88,468 null values (81.73%).
 - There are 163 unique values in the dataset, which represent various categories.
 - There are no blanks or values with leading or trailing whitespace, indicating clean data.

- **NUMBER_OF_LANES**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	104	
● Ok	95,473		88.20%
Unique	102		0.09%
Null	12,769		11.80%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		1	
Max		104	
Average		1.10	
Shortest Value		2	
Longest Value	Not Suspect of Alcohol Use, Not Suspect of Drug Use, Not Susp...		
First Alphanumeric Value		0	
Last Alphanumeric Value	Unknown, Unknown		
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 95,473 valid records (88.20%) and 12,769 null values (11.80%).
- There are 102 unique values, indicating a diverse set of categories.
- The grouped values have a minimum length of 1 character and a maximum length of 104 characters.
- The average length of values is 1.10 characters.
- The shortest value is "2," while the longest value is "Not Suspect of Alcohol Use, Not Suspect of Drug Use, Not Susp..."
- The first alphanumeric value is "0," and the last is "Unknown, Unknown."
- There are no blanks or values with leading or trailing whitespace, indicating clean data formatting.

- **DIRECTION**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	104	
● Ok	93,864		86.72%
Unique	22		0.02%
● Null	14,378		13.28%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			104
Average			4.60
Shortest Value			N/A
Longest Value	Not Suspect of Alcohol Use, Not Suspect of Drug Use, Not Susp...		
First Alphanumeric Value	ALCOHOL CONTRIBUTED		
Last Alphanumeric Value	West		
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- The dataset contains 108,242 records, with 86.72% valid records and 13.28% null values (14,378 records).
 - There are 22 unique values.
 - There are no blank values.
 - No values contain leading or trailing whitespace.

- **DISTANCE_UNIT**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	49	
● Ok	95,449		88.18%
Unique	23		0.02%
● Null	12,793		11.82%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			49
Average			4.00
Shortest Value			N/A
Longest Value	Other Fixed Object (wall, building, tunnel, etc.)		
First Alphanumeric Value	BACKING		
Last Alphanumeric Value	UNKNOWN		
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - The dataset contains 108,242 records, with 88.18% valid records and 11.82% null values (12,793 records).
 - There are 23 unique values.
 - There are no blank values.
 - No values contain leading or trailing whitespace.

- **ROAD_GRADE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	26	
● Ok	93,826		86.68%
Unique	34		0.03%
■ Null	14,416	14,416	13.32%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			26
Average			6.60
Shortest Value			N/A
Longest Value		Motor Vehicle in Transport	
First Alphanumeric Value		BACKING	
Last Alphanumeric Value		Uphill	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 93,826 valid records (86.68%) and 14,416 null records (13.32%).
- There are 34 unique values in the dataset, indicating a moderate range of categories.
- There are no blank values, leading whitespaces, or trailing whitespaces in the dataset.

- **ROAD_NAME**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	43	
● Ok	92,424		85.39%
Unique	4,415	4,415	4.08%
■ Null	15,818		14.61%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		3	
Max		43	
Average		13.30	
Shortest Value		N/A	
Longest Value	SPUR TO SAM EIG HWY FR N/B GREAT SENECA HWY		
First Alphanumeric Value	100 BLK PHILADELPHIA		
Last Alphanumeric Value	ZION RD		
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 92,424 valid records (85.39%) and 15,818 null records (14.61%).
 - There are 4,415 unique values, representing a high level of variability in the data.
 - There are no blank values, leading whitespaces, or trailing whitespaces in the dataset.

- **CROSS_STREET_NAME**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	81	
● Ok	88,380		81.65%
Unique	7,272		6.72%
■ Null	19,862		18.35%
Not	0		0.00%
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		2	
Max		81	
Average		13.60	
Shortest Value		NA	
Longest Value	RAMP 1 FR MD 187 NB TO RAMP 8 (TO IS495) RAMP 7 FR RAM...		
First Alphanumeric Value	10TH AVE		
Last Alphanumeric Value	ZION RD		
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - The dataset has 108,242 records with 81.65% valid entries and 18.35% null values.
 - There are 7,272 unique values.
 - No blanks or whitespace issues are present in the data.

- **OFF_ROAD_DESCRIPTION**

- Summary:

Summary ▲			
Type	Records	Data Type	Size
V_String	108,242	187	
● Ok	13,244		12.24%
Unique	> 10,000		
● Null	94,775		87.56%
Not			
● Ok	223		0.21%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			187
Average			44.80
Shortest Value			N/A
Longest Value	Parking Lot Way UNIT 1 STRUCK UNIT TWO WHEN REVERSING ...		
First Alphanumeric Value	"Capital One" Bank parking lot behind 8315 Georgia Ave , Entra...		
Last Alphanumeric Value	yard of 7507 brookville rd		
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			223

- Observation:
 - The dataset contains 108,242 records, with 12.24% valid entries and 87.56% null values.
 - There are more than 10,000 unique values.
 - There are no blank values, but there are 223 trailing whitespaces.

- **MUNICIPALITY**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	19	
● Ok	85,139		78.66%
Unique	21		0.02%
■ Null	23,103		21.34%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			19
Average			3.90
Shortest Value			N/A
Longest Value		CHEVY CHASE VILLAGE	
First Alphanumeric Value		BROOKVILLE	
Last Alphanumeric Value		WASHINGTON GROVE	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 78.66% valid entries and 21.34% null values.
 - There are 21 unique values.
 - There are no blank values, leading whitespaces, or trailing whitespaces.

- **RELATED_NON_MOTORIST**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	86	
● Ok	6,422		5.93%
Unique	557		0.51%
■ Null	101,820		94.07%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			5
Max			86
Average			10.40
Shortest Value		OTHER	
Longest Value		Other Pedestrian (person in a building, skater, personal conveya...	
First Alphanumeric Value		38.743373	
Last Alphanumeric Value		Wheelchair (non-electric)	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 6,422 valid records (5.93%) and 101,820 null records (94.07%).

- There are 557 unique values, indicating a moderate level of variability in the data.
- The grouped value lengths range from 1 to 86 characters.
- The maximum value length is 86 characters, while the minimum is 1 character.
- There are no blank values, no leading whitespaces, and no trailing whitespaces in the dataset.

- **AT_FAULT**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	12	
● Ok	106,965		98.82%
Unique	800		0.74%
● Null	1,277		1.18%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			4
Max			12
Average			6.20
Shortest Value			BOTH
Longest Value			-77.00126667
First Alphanumeric Value			-76.93133354
Last Alphanumeric Value			UNKNOWN
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- The dataset contains 108,242 records, with 106,965 valid records (98.82%) and 1,277 null records (1.18%).
- There are 800 unique values, indicating a moderate level of variability in the data.
- The grouped value lengths range from 4 to 12 characters.
- There are no blank values, no leading whitespaces, and no trailing whitespaces in the dataset.

- **COLLISION_TYPE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	29	
● Ok	107,002		98.85%
Unique	862		0.80%
■ Null	1,240		1.15%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			29
Average			16.50
Shortest Value			N/A
Longest Value		Sideswipe, Opposite Direction	
First Alphanumeric Value		(38.743373, -77.54699707)	
Last Alphanumeric Value		Unknown	
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- The dataset contains 108,242 records, with 107,002 valid records (98.85%) and 1,240 null records (1.15%).
- There are 862 unique values, representing a moderate level of variability in the data.
- There are 0 blank values, 0 leading whitespaces, and 0 trailing whitespaces in the dataset.

- WEATHER

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	33	
● Ok	106,472		98.36%
Unique	327		0.30%
■ Null	1,770		1.64%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			33
Average			5.20
Shortest Value			N/A
Longest Value		Freezing Rain Or Freezing Drizzle	
First Alphanumeric Value		(38.94984167, -77.11566667)	
Last Alphanumeric Value		WINTRY MIX	
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - The dataset contains 108,242 records, with 106,472 valid records (98.36%) and 1,770 null records (1.64%).
 - There are 327 unique values, representing a relatively low level of variability in the data.
 - The longest value is "Freezing Rain Or Freezing Drizzle", and the shortest value is "N/A".
 - There are 0 blank values, 0 leading whitespaces, and 0 trailing whitespaces in the dataset.

- **SURFACE_CONDITION**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	27	
● Ok	93,521		86.40%
Unique	58		0.05%
■ Null	14,721		13.60%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			27
Average			3.10
Shortest Value			DRY
Longest Value		(39.03950625, -76.99425189)	
First Alphanumeric Value		(38.96613833, -77.002425)	
Last Alphanumeric Value			Wet
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - The dataset contains 108,242 records, with 106,472 valid records (98.36%) and 1,770 null records (1.64%).
 - There are 327 unique values, representing a relatively low level of variability in the data.
 - There are 0 blank values, 0 leading whitespaces, and 0 trailing whitespaces in the dataset.

- **LIGHT**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	24	
● Ok	106,168		98.08%
Unique	17		0.02%
● Null	2,074		1.92%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		3	
Max		24	
Average		9.70	
Shortest Value		N/A	
Longest Value		DARK -- UNKNOWN LIGHTING	
First Alphanumeric Value		DARK -- UNKNOWN LIGHTING	
Last Alphanumeric Value		Unknown	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 106,168 valid records (98.08%) and 2,074 null records (1.92%).
- There are 17 unique values, representing a relatively low level of variability in the data.
- The first alphanumeric value is "DARK -- UNKNOWN LIGHTING", and the last alphanumeric value is "Unknown".

- **TRAFFIC_CONTROL**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	96	
● Ok	105,106		97.10%
Unique	67		0.06%
● Null	3,136		2.90%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min		3	
Max		96	
Average		11.00	
Shortest Value		N/A	
Longest Value		Other Pavement Marking (excluding edgelines, centerlines, or la...	
First Alphanumeric Value		Bicycle Crossing Sign	
Last Alphanumeric Value		Yield Sign	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - The dataset contains 108,242 records, with 105,106 valid records (97.10%) and 3,136 null records (2.90%).
 - There are 67 unique values, representing a relatively low level of variability in the data.
 - The first alphanumeric value is "Bicycle Crossing Sign", and the last alphanumeric value is "Yield Sign".

- **DRIVER_SUBSTANCE_ABUSE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	254	
● Ok	105,510		97.48%
Unique	100		0.09%
■ Null	2,732		2.52%
Not Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			254
Average			18.90
Shortest Value			N/A
Longest Value	Not Suspect of Alcohol Use, Not Suspect of Drug Use, Not Sus...		
First Alphanumeric Value	ALCOHOL CONTRIBUTED		
Last Alphanumeric Value	Unknown, Unknown, Unknown, Unknown, Unknown, Unknown, ...		
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- The dataset contains 108,242 records, with 105,510 valid records (97.48%) and 2,732 null records (2.52%).
- There are 100 unique values, representing a relatively low level of variability in the data.
- There are 0 blank values, 0 leading whitespaces, and 0 trailing whitespaces in the dataset.

- **NON_MOTORIST_SUBSTANCE_ABUSE**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	157	
● Ok	5,892		5.44%
Unique	26		0.02%
■ Null	102,350		94.56%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			157
Average			15.00
Shortest Value			N/A
Longest Value	Not Suspect of Alcohol Use, Not Suspect of Drug Use, Not Susp...		
First Alphanumeric Value	ALCOHOL CONTRIBUTED		
Last Alphanumeric Value	Unknown, Unknown		
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:

- There are 5,892 valid records, representing 5.44% of the total.
- There are 102,350 null records, representing 94.56% of the total.
- There are 26 unique values, indicating a relatively low level of variability in the data.
- The longest value is "Not Suspect of Alcohol Use, Not Suspect of Drug Use, Not Susp...", and the shortest value is "N/A".
- There are 0 blank values, 0 leading whitespaces, and 0 trailing whitespaces in the dataset.

- **FIRST_HARMFUL_EVENT**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	54	
● Ok	106,168		98.08%
Unique	65		0.06%
■ Null	2,074		1.92%
Not			
● Ok	0		0.00%
● Empty	0		0.00%
Length Statistics			
Min			3
Max			54
Average			13.50
Shortest Value	Strikes Object at Rest from Motor Vehicle In Transport		
Longest Value	ANIMAL		
First Alphanumeric Value	Utility Pole/Light Support		
Last Alphanumeric Value	N/A		
Blanks			0
Values with Leading Whitespace			0
Values with Trailing Whitespace			0

- Observation:
 - There are 106,168 valid records, representing 98.08% of the total.
 - There are 2,074 null records, representing 1.92% of the total.
 - There are 65 unique values, indicating a relatively low level of variability in the data.
 - The longest value is "Strikes Object at Rest from Motor Vehicle in Transprt" and the shortest value is "N/A".
 - There are 0 blank values, 0 leading whitespaces, and 0 trailing whitespaces in the dataset.

- **SECOND_HARMFUL_EVENT**

- Summary:

Summary			
Type	Records	Data Type	Size
V_String	108,242	76	
• Ok	100,133		92.51%
Unique	65		0.06%
• Null	8,109		7.49%
Not Ok	0		0.00%
• Empty	0		0.00%
Length Statistics			
Min		3	
Max		76	
Average		5.80	
Shortest Value		N/A	
Longest Value	Struck by Falling, Shifting Cargo Or Anything Set in Motion by ...		
First Alphanumeric Value	ANIMAL		
Last Alphanumeric Value	Utility Pole/Light Support		
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 100,133 valid records (92.51%) and 8,109 null records (7.49%).
- There are 65 unique values, representing a moderate level of variability in the data.
- The longest value is "Struck by Falling, Shifting Cargo Or Anything Set in Motion by ANIMAL," and the shortest value is "N/A."
- There are no blank values, but there are 0 values with leading whitespaces and 0 values with trailing whitespaces in the dataset.

- **JUNCTION**

- Summary:

Summary			
Type	Records	Data Type Size	
V_String	108,242	67	
● Ok	93,241	86.14%	
Unique	22	0.02%	
■ Null	15,001	13.86%	
Not			
● Ok	0	0.00%	
● Empty	0	0.00%	
Length Statistics			
Min		3	
Max		67	
Average		13.30	
Shortest Value		N/A	
Longest Value	Other Location Not Listed Above Within an Interchange Area (...		
First Alphanumeric Value		ALLEY	
Last Alphanumeric Value		UNKNOWN	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 93,241 valid records (86.14%) and 15,001 null records (13.86%).
 - There are 22 unique values, representing a low level of variability in the data. The grouped value lengths range from 4 to 67 characters.
 - The longest value is "Other Location Not Listed Above Within an Interchange Area....", and the shortest value is "N/A."
 - There are no blank values, but there are 0 values with leading whitespaces and 0 values with trailing whitespaces in the dataset.

- **INTERSECTION_TYPE**

- Summary:

Summary			
Type	Records	Data Type Size	
V_String	108,242	25	
● Ok	87,847	81.16%	
Unique	12	0.01%	
■ Null	20,395	18.84%	
Not			
● Ok	0	0.00%	
● Empty	0	0.00%	
Length Statistics			
Min		3	
Max		25	
Average		11.50	
Shortest Value		N/A	
Longest Value	Roundabout/Traffic Circle		
First Alphanumeric Value	Angled/Skewed		
Last Alphanumeric Value	Y-INTERSECTION		
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 87,847 valid records (81.16%) and 20,395 null records (18.84%).
- There are 12 unique values, representing a low level of variability in the data. The grouped value lengths range from 3 to 25 characters.
- The longest value is "Roundabout/Traffic Circle Angled/Skewed," and the shortest value is "N/A."
- There are no blank values, but there are 0 values with leading whitespaces and 0 values with trailing whitespaces in the dataset.

- **ROAD_ALIGNMENT**

- Summary:

Type	Records	Data Type Size
V_String	108,242	33
Ok	93,475	86.36%
Unique	13	0.01%
Null	14,767	13.64%
Not Ok	0	0.00%
Empty	0	0.00%
Length Statistics		
Min	3	
Max	33	
Average	8.30	
Shortest Value	N/A	
Longest Value	Curve Left, Curve Right, Straight	
First Alphanumeric Value	CURVE LEFT	
Last Alphanumeric Value	UNKNOWN	
Blanks	0	
Values with Leading Whitespace	0	
Values with Trailing Whitespace	0	

- Observation:

- The dataset contains 108,242 records, with 93,475 valid records (86.36%) and 14,767 null records (13.64%).
- There are 13 unique values, representing a low level of variability in the data.
- The longest value is "Curve Left, Curve Right, Straight", and the shortest value is "N/A."
- There are no blank values, but there are 0 values with leading whitespaces and 0 values with trailing whitespaces in the dataset.

- **ROAD_CONDITION**

- Summary:

Summary			
Type	Records	Data Type Size	
V_String	108,242	24	
● Ok	91,532	84.56%	
Unique	22	0.02%	
■ Null	16,710	15.44%	
Not			
● Ok	0	0.00%	
● Empty	0	0.00%	
Length Statistics			
Min		3	
Max		24	
Average		9.80	
Shortest Value		N/A	
Longest Value		OBSTRUCTION NOT SIGNALLED	
First Alphanumeric Value		FOREIGN MATERIAL	
Last Alphanumeric Value		View Obstructed	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:

- The dataset contains 108,242 records, with 91,532 valid records (84.56%) and 16,710 null records (15.44%).
- There are 22 unique values, representing a low level of variability in the data.
- The longest value is "OBSTRUCTION NOT SIGNALLED FOREIGN MATERIAL", and the shortest value is "N/A."
- There are no blank values, but there are 0 values with leading whitespaces and 0 values with trailing whitespaces in the dataset.

- **ROAD_DIVISION**

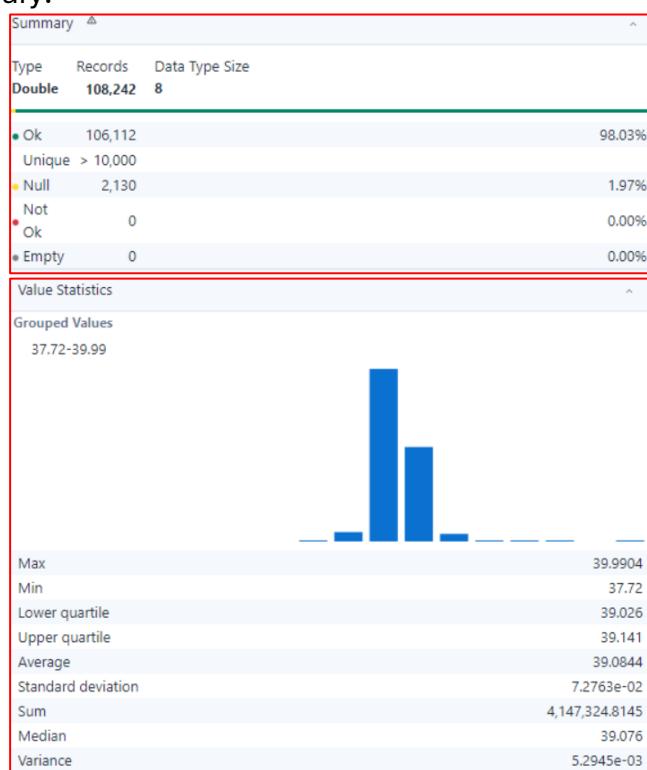
- Summary:

Summary			
Type	Records	Data Type Size	
V_String	108,242	125	
● Ok	93,476	86.36%	
Unique	33	0.03%	
■ Null	14,766	13.64%	
Not			
● Ok	0	0.00%	
● Empty	0	0.00%	
Length Statistics			
Min		3	
Max		125	
Average		32.00	
Shortest Value		N/A	
Longest Value		Divided, Flush Median (greater than 4ft wide), Divided, Raised ...	
First Alphanumeric Value		, Not Divided	
Last Alphanumeric Value		Unknown	
Blanks		0	
Values with Leading Whitespace		0	
Values with Trailing Whitespace		0	

- Observation:
 - The dataset contains 108,242 records, with 93,476 valid records (86.36%) and 14,766 null records (13.64%).
 - There are 33 unique values, representing a low level of variability in the data.
 - The longest value is "Divided, Flush Median (greater than 4ft wide), Divided, Raised ..., Not Divided", and the shortest value is "N/A."
 - There are no blank values, but there are 0 values with leading whitespaces and 0 values with trailing whitespaces in the dataset.

- **LATITUDE**

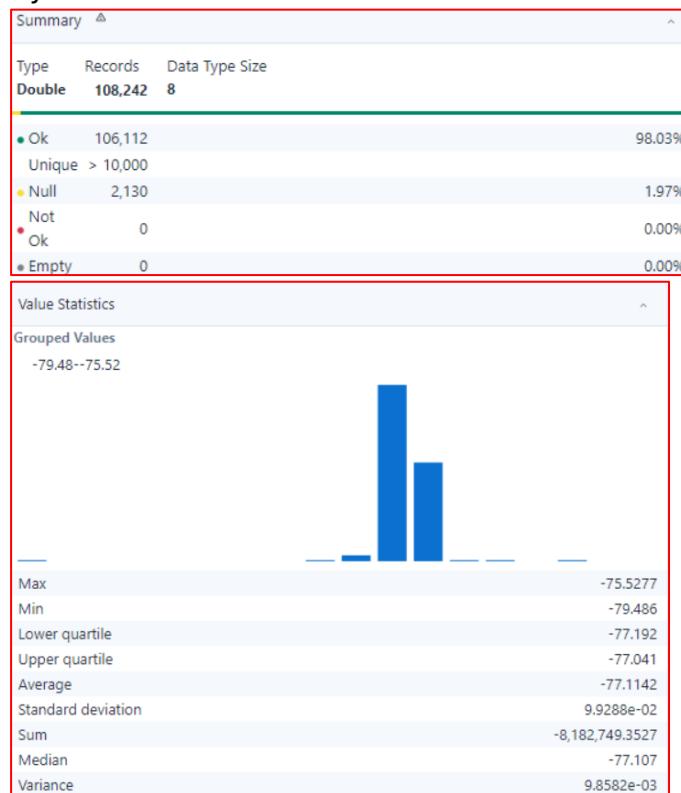
- Summary:



- Observation:
 - The dataset contains 108,242 records with a Data Type Size of 8.
 - There are 106,112 valid ("Ok") records, representing 98.03% of the total, and more than 10,000 unique values.
 - There are 2,130 null records, representing 1.97% of the total. There are no "Not Ok" or "Empty" records.

- **LONGITUDE**

- Summary:



- Observation:

- The dataset contains 108,242 records, with 92,424 valid records (85.39%) and 15,818 null records (14.61%).
 - There are 4,415 unique values, representing a high level of variability in the data.
 - The longest value is "SPUR TO SAM EIG HWY FR N/B GREAT SENECA HWY," and the shortest value is "N/A."
 - There are no blank values, leading whitespaces, or trailing whitespaces in the dataset.

- **LOCATION**

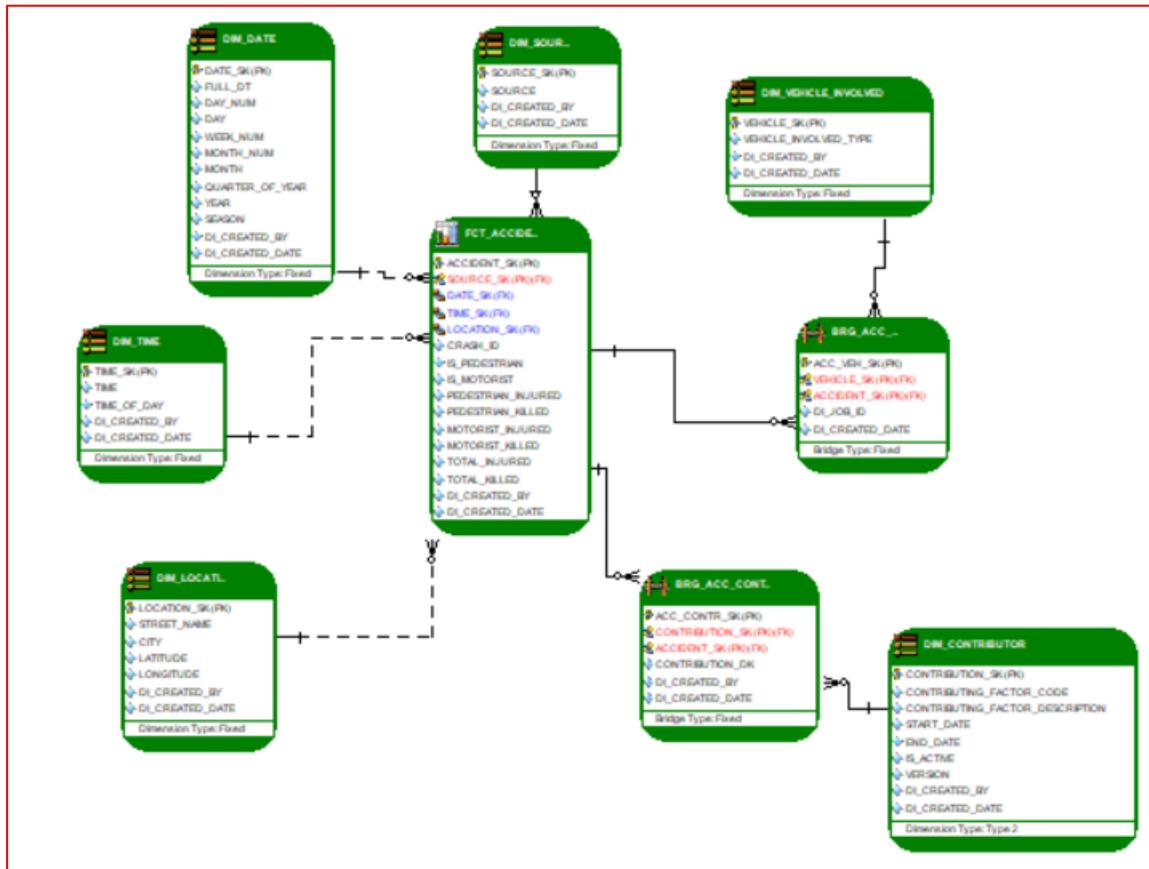
- Summary:

Type	Records	Data Type Size
String	108,242	27
● Ok	106,112	98.03%
Unique	> 10,000	
● Null	2,130	1.97%
● Not	0	0.00%
● Ok	0	0.00%
● Empty	0	0.00%
Length Statistics		
Min	13	
Max	27	
Average	25.80	
Shortest Value	(39.2, -77.4)	
Longest Value	(39.11311333, -77.05759167)	
First Alphanumeric Value	(37.72, -79.48)	
Last Alphanumeric Value	(39.990414, -76.989719)	
Blanks	0	
Values with Leading Whitespace	0	
Values with Trailing Whitespace	0	

- Observation:

- The dataset contains 108,242 records, with 106,112 valid records (98.03%) and 2,130 null records (1.97%).
 - There are over 10,000 unique values, indicating a high level of variability in the data.
 - There are no blank values, values with leading whitespace, or values with trailing whitespace in the dataset.

ER Diagram



This ER diagram models a **Traffic Accident Reporting System** with a star schema approach suitable for data warehousing. The schema is designed to analyze traffic accidents with detailed dimensional attributes and fact tables capturing accident metrics. The model includes **fact tables** and **dimension tables** to support analytical queries.

Tables and Relationships

1. Fact Table: FCT_ACCIDENT

- **Description:** Captures the main metrics of traffic accidents and links to dimension tables.
- **Primary Key:** ACCIDENT_SK, SOURCE_SK, TIME_SK, DATE_SK, LOCATION_SK.
- **Attributes:**
 - TOTAL_REGISTERED: Number of registered vehicles.
 - REGISTERED_INJURED: Injuries involving registered vehicles.
 - REGISTERED_KILLED: Fatalities involving registered vehicles.
 - MOTORIST_INJURED: Number of motorists injured.
 - MOTORIST_KILLED: Number of motorists killed.
 - TOTAL_INJURED: Total number of injuries.
 - TOTAL_KILLED: Total number of fatalities.
 - DL_CREATED_BY, DL_CREATED_DATE: Metadata for data lineage.

- **Relationships:**
 - Linked to DM_DATE, DM_TIME, DM_LOCATION, DM_SOURCE, and BRG_ACC_CONTRIBUTOR tables for contextual data.

2. Dimension Table: DM_DATE

- **Description:** Contains date-related attributes for temporal analysis.
- **Primary Key:** DATE_SK.
- **Attributes:**
 - FULL_DT: Full date value.
 - DAY_NAME, DAY_NUM, WEEK_NUM: Details for day, week, and day number.
 - MONTH_NAME, MONTH_NUM: Month details.
 - QUARTER_OF_YEAR, YEAR, SEASON: Year and seasonal breakdowns.
 - DL_CREATED_BY, DL_CREATED_DATE: Metadata for data lineage.
- **Relationships:**
 - Links to FCT_ACCIDENT via DATE_SK.

3. Dimension Table: DM_TIME

- **Description:** Contains time-related attributes for hourly and time-of-day analysis.
- **Primary Key:** TIME_SK.
- **Attributes:**
 - TIME_OF_DAY: Specific time of the accident.
 - DL_CREATED_BY, DL_CREATED_DATE: Metadata for data lineage.
- **Relationships:**
 - Links to FCT_ACCIDENT via TIME_SK.

4. Dimension Table: DM_LOCATION

- **Description:** Stores geographic details for accidents.
- **Primary Key:** LOCATION_SK.
- **Attributes:**
 - STREET_NAME, CITY, LATITUDE, LONGITUDE: Geographic information.
 - DL_CREATED_BY, DL_CREATED_DATE: Metadata for data lineage.
- **Relationships:**
 - Links to FCT_ACCIDENT via LOCATION_SK.

5. Dimension Table: DM_SOURCE

- **Description:** Stores data source details for reporting.
- **Primary Key:** SOURCE_SK.
- **Attributes:**
 - SOURCE: The source of the accident data.
 - DL_CREATED_BY, DL_CREATED_DATE: Metadata for data lineage.
- **Relationships:**
 - Links to FCT_ACCIDENT via SOURCE_SK.

6. Bridge Table: BRG_ACC_VEHICLE

- **Description:** Captures details of vehicles involved in accidents.
- **Primary Key:** ACC_VEH_SK, VEHICLE_SK, ACCIDENT_SK.
- **Attributes:**
 - VEHICLE_SK: Links to DM_VEHICLE_INVOLVED.
 - DL_CREATED_BY, DL_CREATED_DATE: Metadata for data lineage.
- **Relationships:**
 - Links to DM_VEHICLE_INVOLVED via VEHICLE_SK.
 - Links to FCT_ACCIDENT via ACCIDENT_SK.

7. Dimension Table: DM_VEHICLE_INVOLVED

- **Description:** Stores details of the vehicles involved in accidents.
- **Primary Key:** VEHICLE_SK.
- **Attributes:**
 - VEHICLE_INVOLVED_TYPE: Type of vehicle involved.
 - DL_CREATED_BY, DL_CREATED_DATE: Metadata for data lineage.
- **Relationships:**
 - Links to BRG_ACC_VEHICLE via VEHICLE_SK.

8. Bridge Table: BRG_ACC_CONTRIBUTOR

- **Description:** Captures contributing factors for accidents.
- **Primary Key:** ACC CONTR_SK, CONTRIBUTION_SK, ACCIDENT_SK.
- **Attributes:**
 - CONTRIBUTION_ID: Unique identifier for contributing factors.
 - DL_CREATED_BY, DL_CREATED_DATE: Metadata for data lineage.
- **Relationships:**
 - Links to DM_CONTRIBUTOR via CONTRIBUTION_SK.
 - Links to FCT_ACCIDENT via ACCIDENT_SK.

9. Dimension Table: DM_CONTRIBUTOR

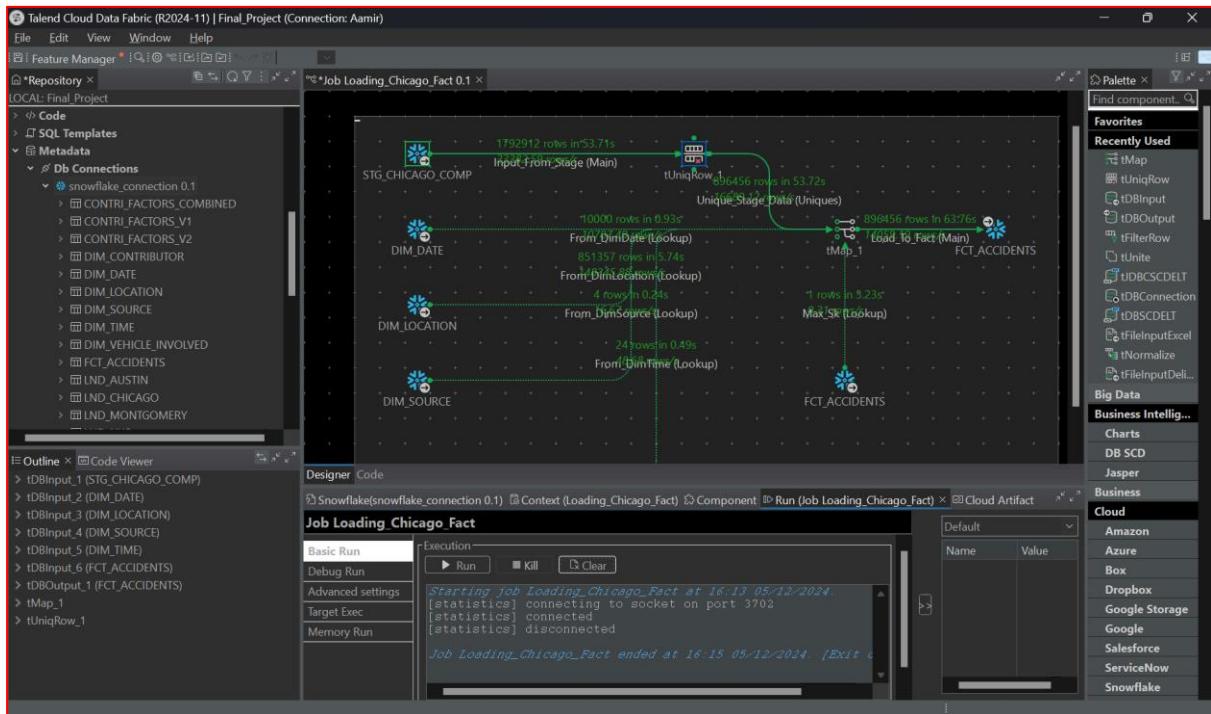
- **Description:** Stores details about contributing factors.
- **Primary Key:** CONTRIBUTION_SK.
- **Attributes:**
 - CONTRIBUTING_FACTOR_CODE, CONTRIBUTING_FACTOR_DESCRIPTION: Categorical codes and descriptions.
 - START_DATE, END_DATE: Validity of the factor.
 - IS_ACTIVE: Indicates if the factor is active.
 - DL_CREATED_BY, DL_CREATED_DATE: Metadata for data lineage.
- **Relationships:**
 - Links to BRG_ACC_CONTRIBUTOR via CONTRIBUTION_SK.

Key Features

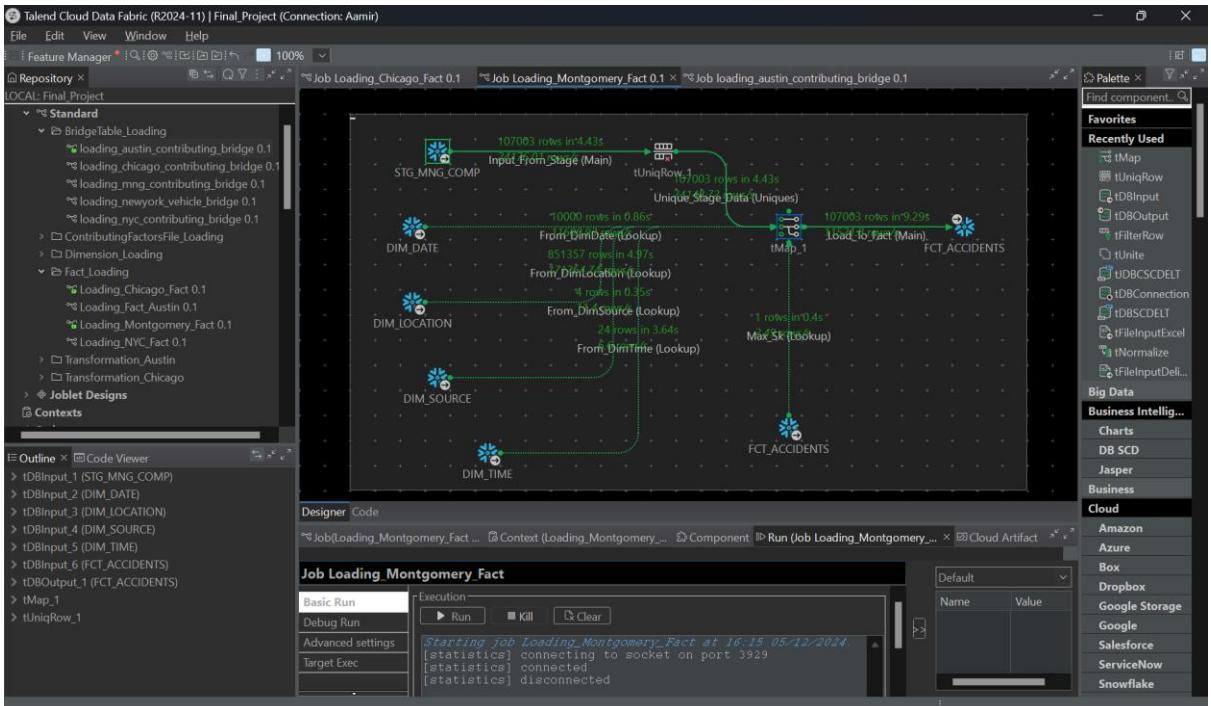
- **Star Schema Design:** Optimized for analytical queries with a central fact table and multiple dimension tables.
- **Bridge Tables:** Allows for many-to-many relationships (e.g., vehicles and contributing factors).
- **Temporal Analysis:** Enables detailed date and time breakdowns.
- **Geographic Analysis:** Captures location-specific attributes for spatial insights.
- **Clean Data Structure:** Ensures referential integrity with primary and foreign keys.

TALEND IMPLEMENTATION

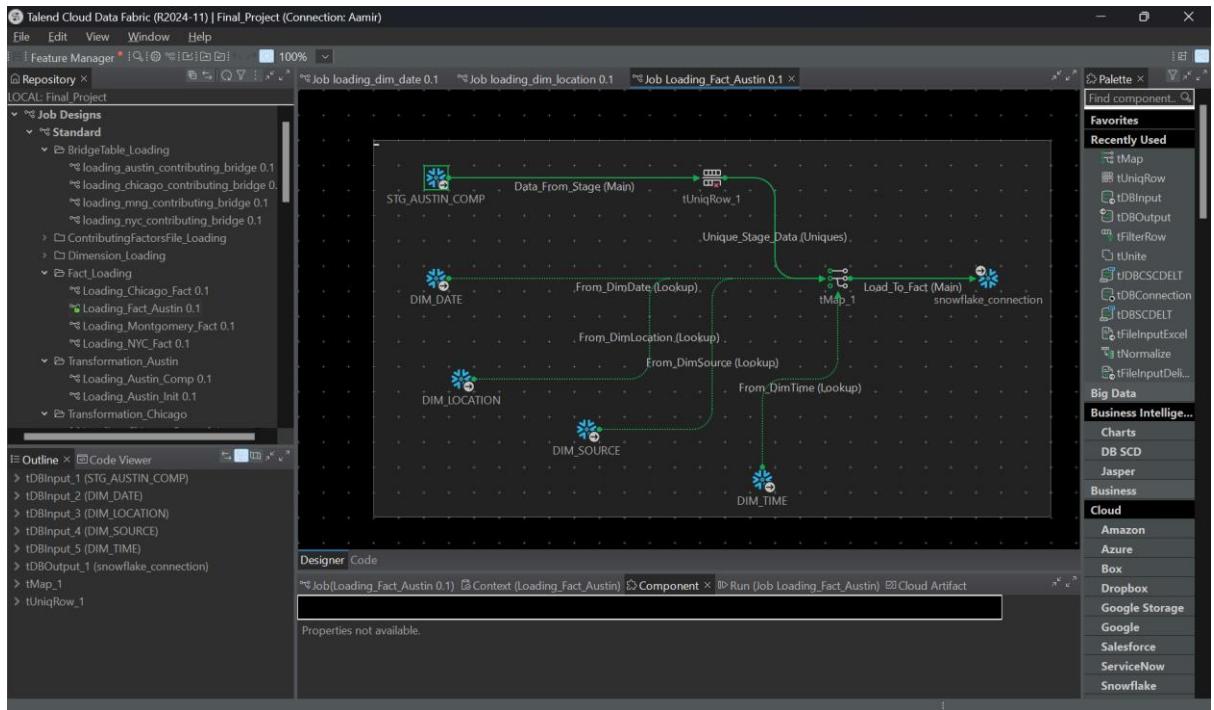
○ CHICAGO FACT LOAD



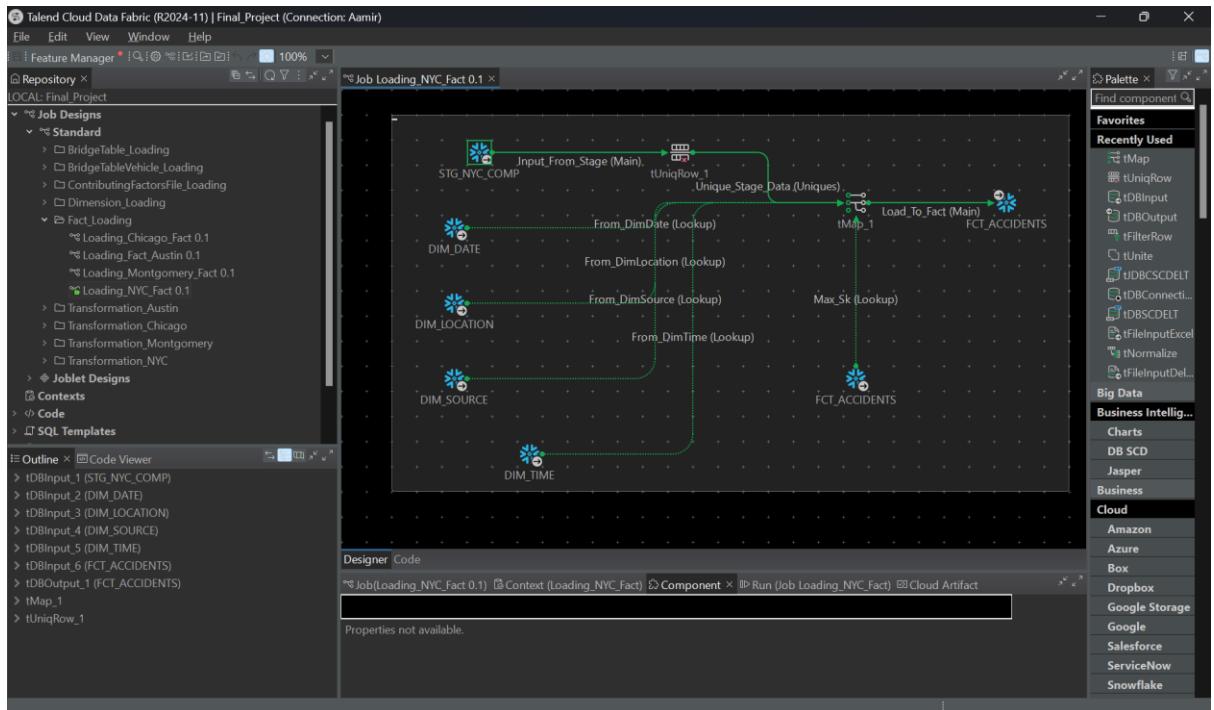
○ MONTGOMERY FACT LOAD



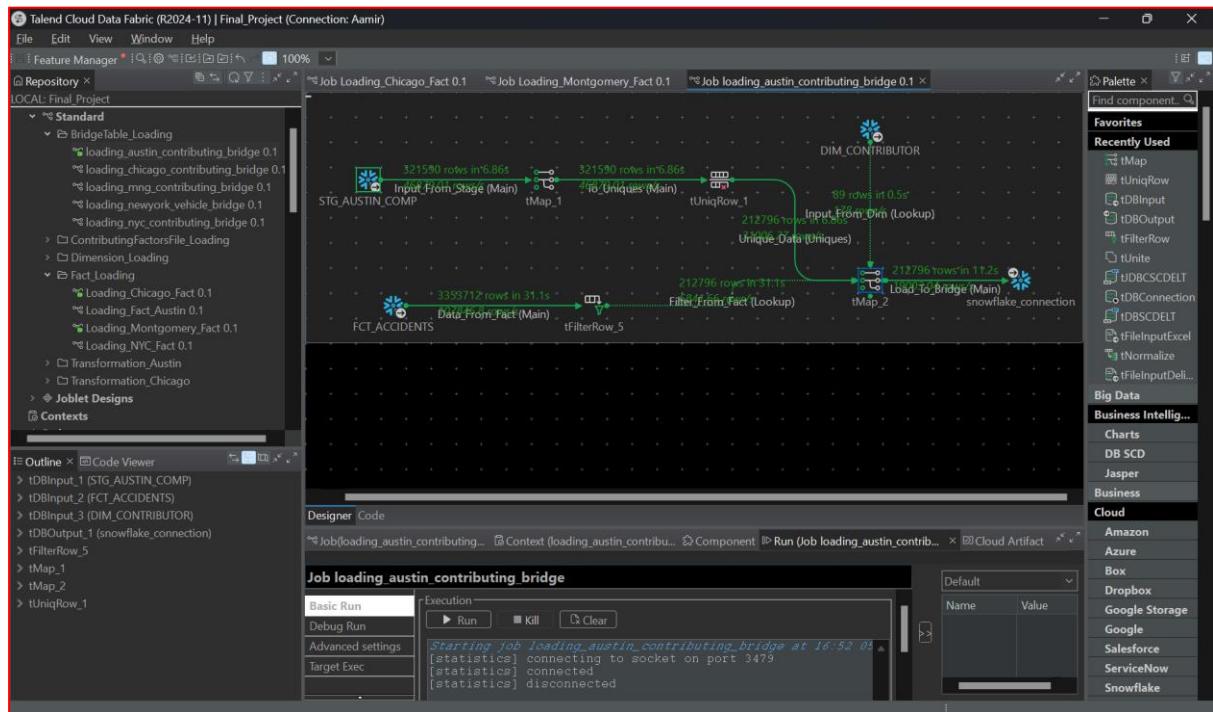
○ AUSTIN FACT LOAD



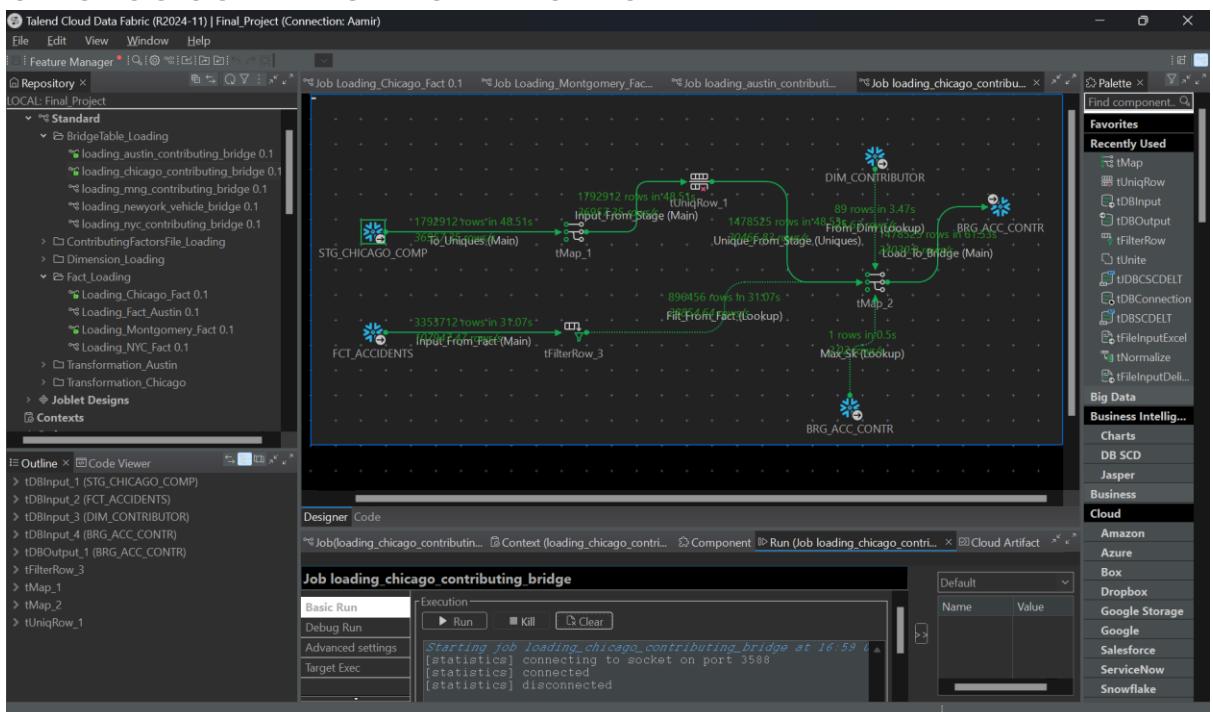
○ NYC FACT LOAD



○ AUSTIN CONTRIBUTING BRIDGE LOAD



○ CHICAGO CONTRIBUTING BRIDGE LOAD

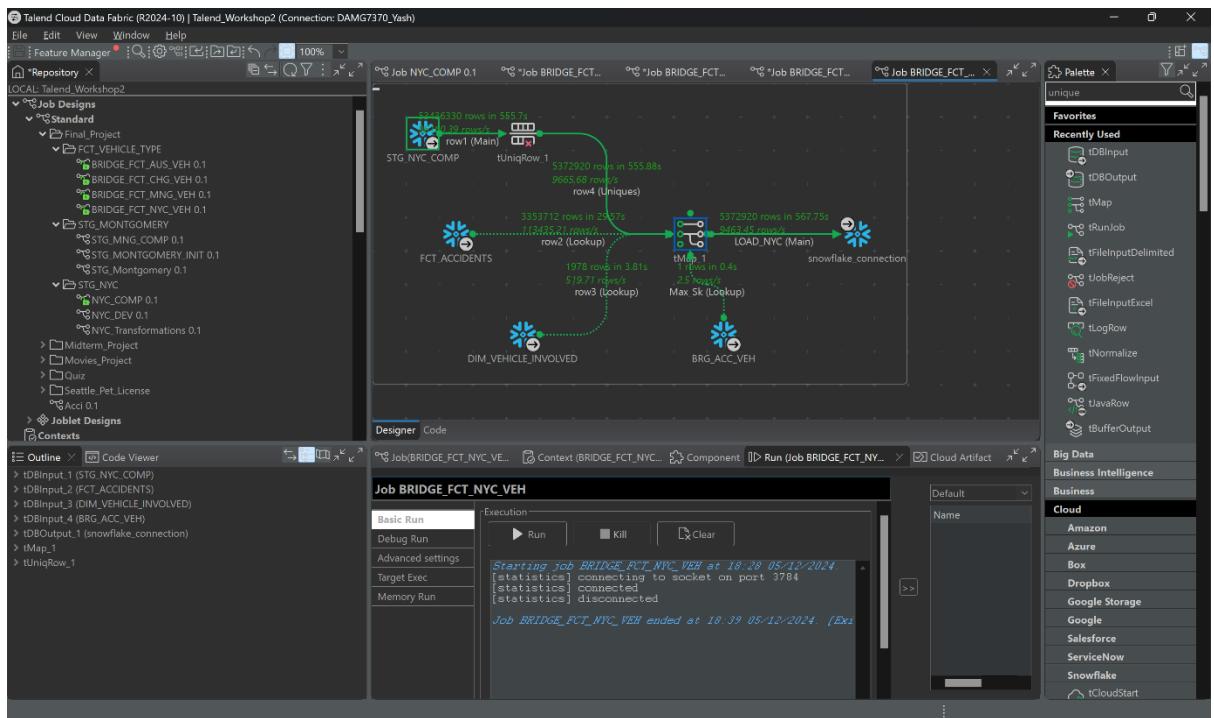


SQL QUERY

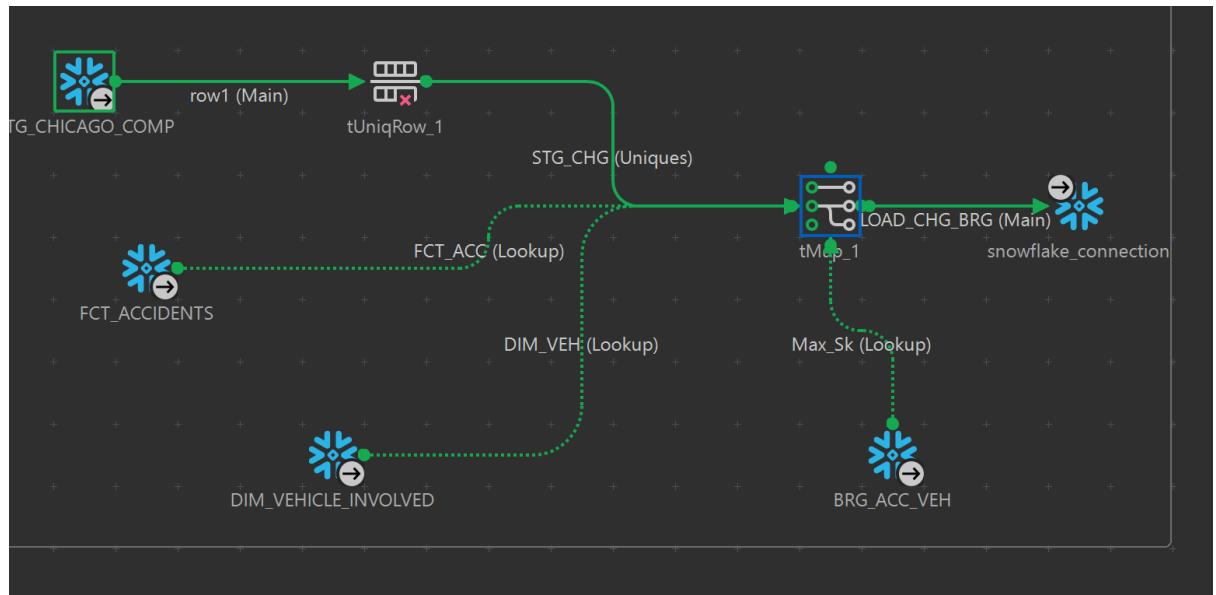
```
ROAD_ACC_FIN.PUBLIC <--> Settings
SELECT COUNT(*) FROM BRG_ACC CONTRA;
```

The screenshot shows a SQL query results page. The query is `SELECT COUNT(*) FROM BRG_ACC CONTRA;`. The result is a single row with the column `COUNT(*)` and value `4433268`.

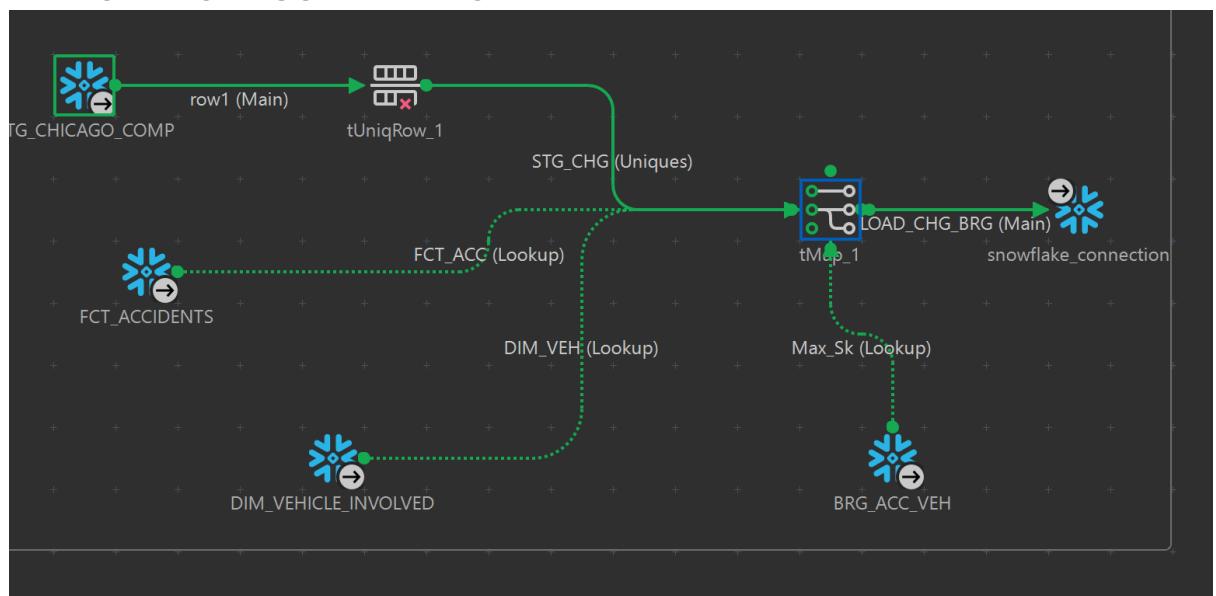
○ BRIDGE FACT NYC VEHICLE



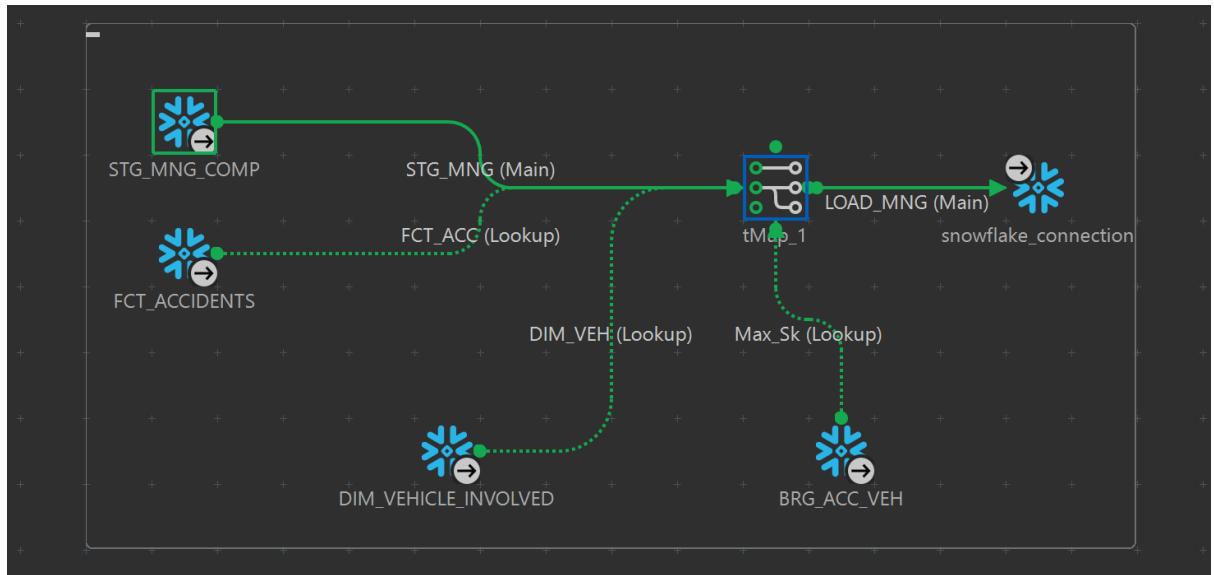
- **BRIDGE FACT CHICAGO VEHICLE**



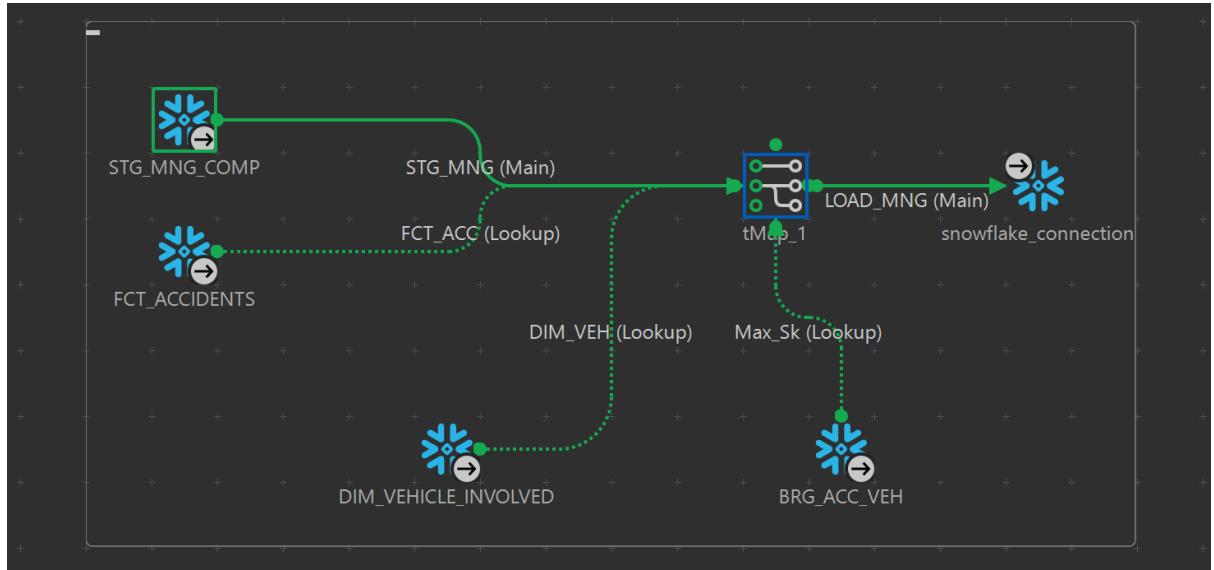
- **BRIDGE FACT AUSTIN VEHICLE**



○ BRIDGE FACT MONTGOMERY VEHICLE



○ BRIDGE FACT NYC VEHICLE



SQL QUERY

```

ROAD_ACC_FIN.PUBLIC  Settings

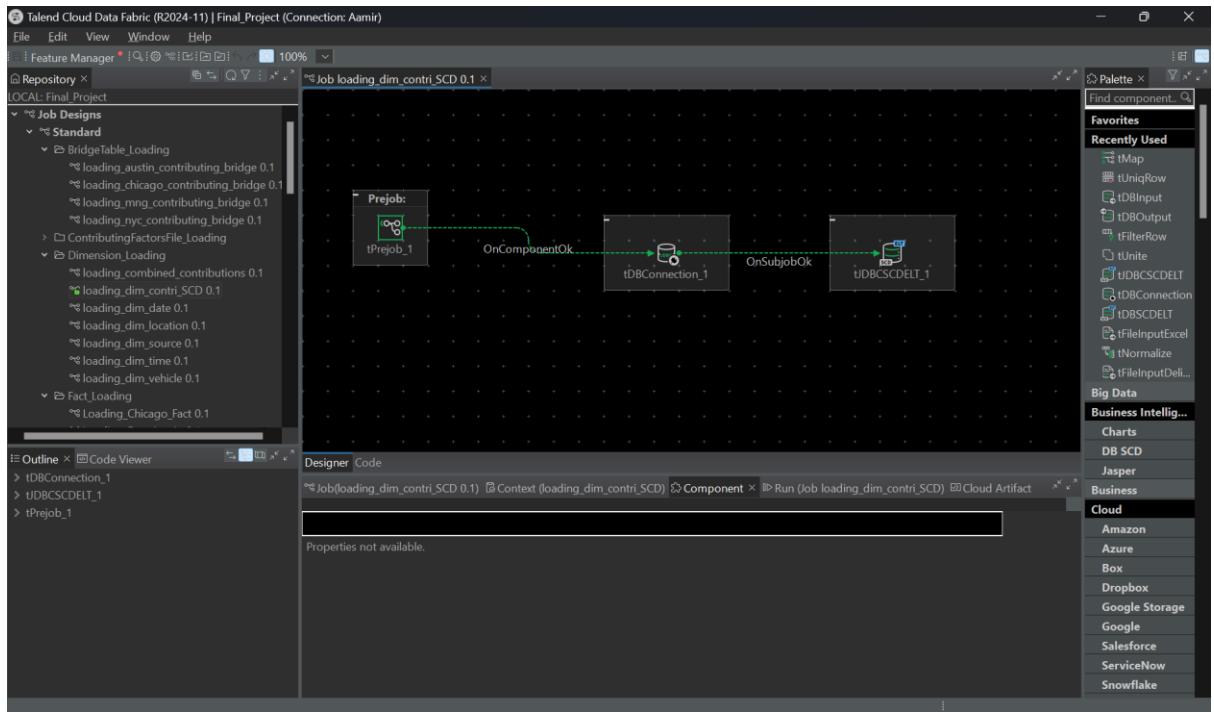
SELECT COUNT(*) FROM BRG_ACC_VEH;

results  ~ Chart

COUNT(*)
4433268

```

○ DIM CONTRIBUTING SCD



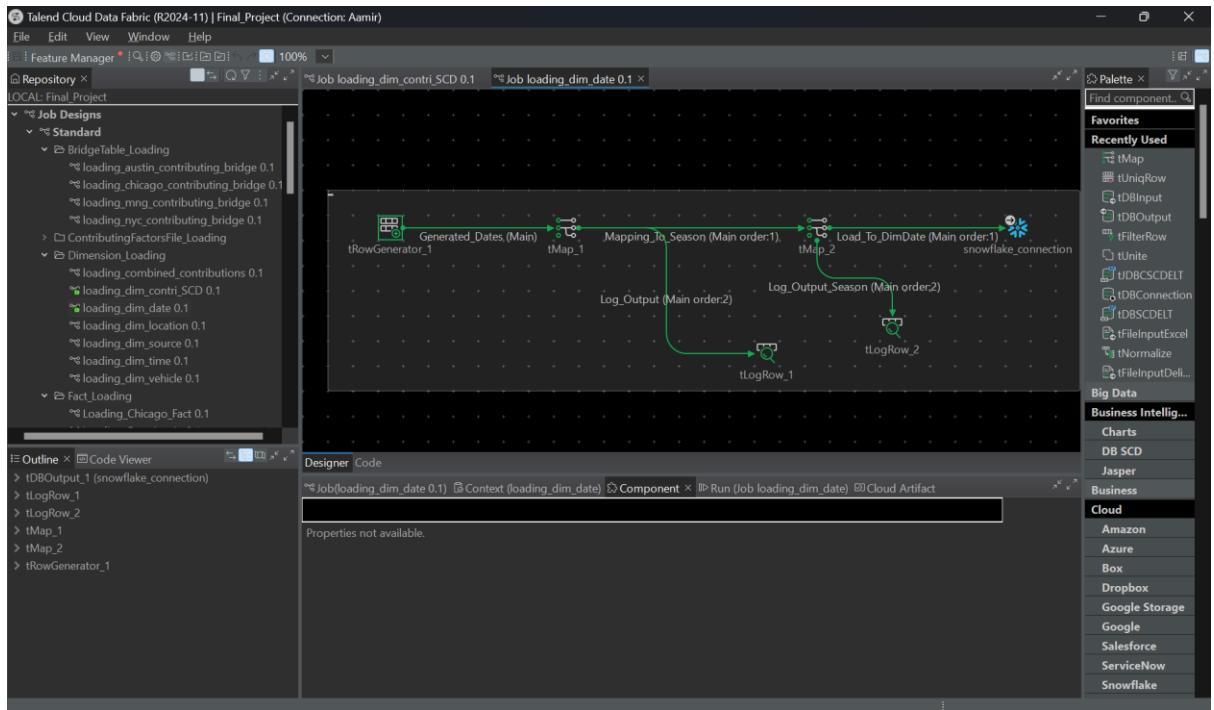
SQL QUERY

```
SELECT COUNT(*) FROM DIM_CONTRIBUTOR;
```

The results of the SQL query are displayed below:

COUNT(*)
89

○ DIM DATE LOAD

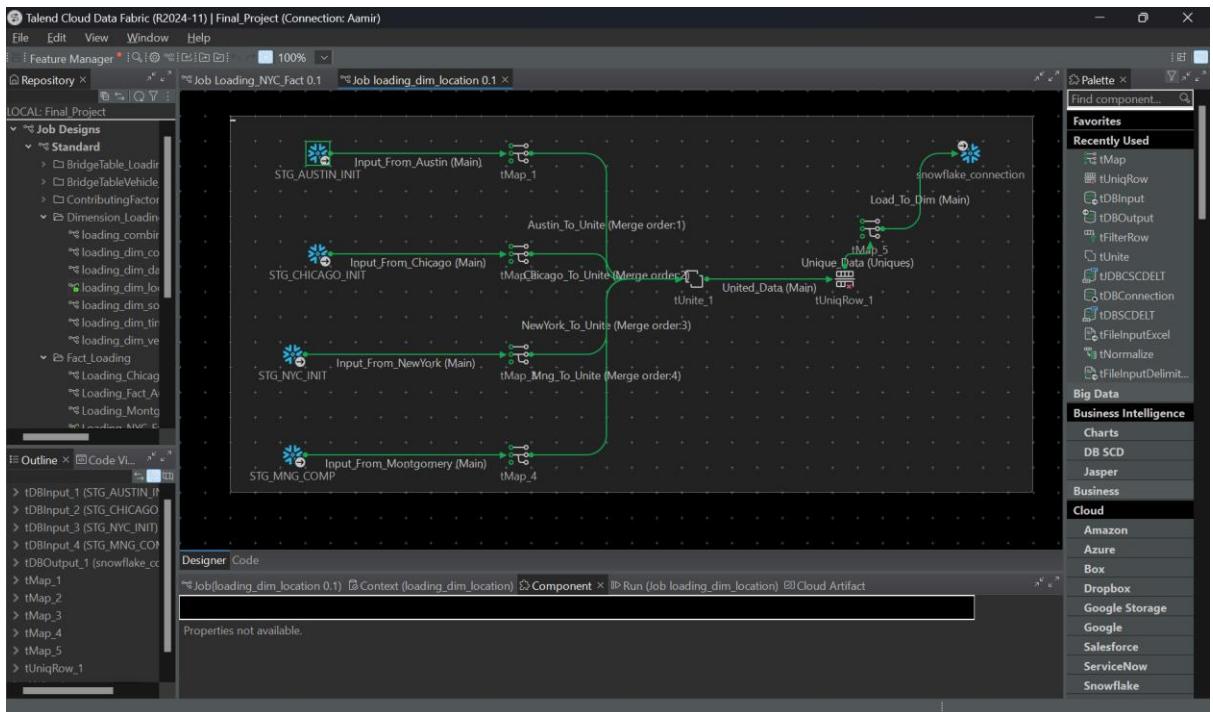


○ SQL QUERY

The screenshot shows the Talend Cloud Data Fabric Results interface. It displays the results of a SQL query: "SELECT COUNT(*) FROM DIM_DATE;". The results show a single row with a COUNT(*) value of 10000. The interface includes tabs for Results and Chart, and a settings menu at the top.

COUNT(*)
10000

○ DIM LOCATION



○ SQL QUERY

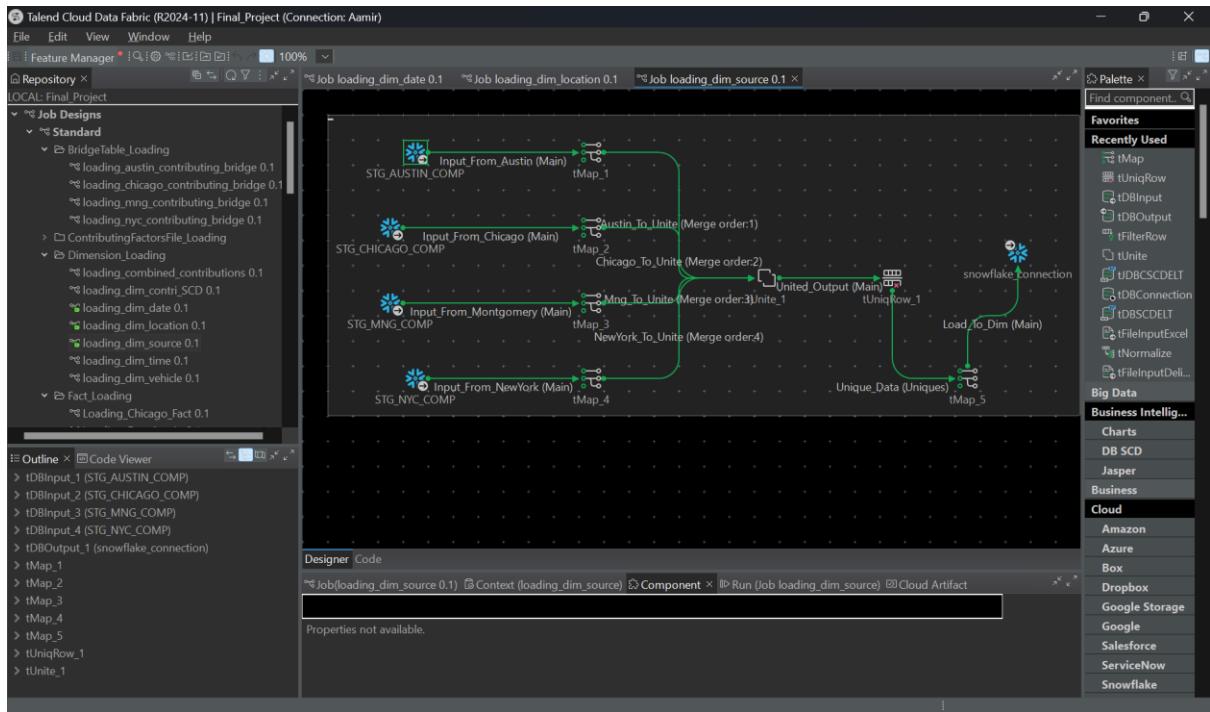
The screenshot shows the results of a SQL query in a database browser. The query is:

```
SELECT COUNT(*) FROM DIM_LOCATION;
```

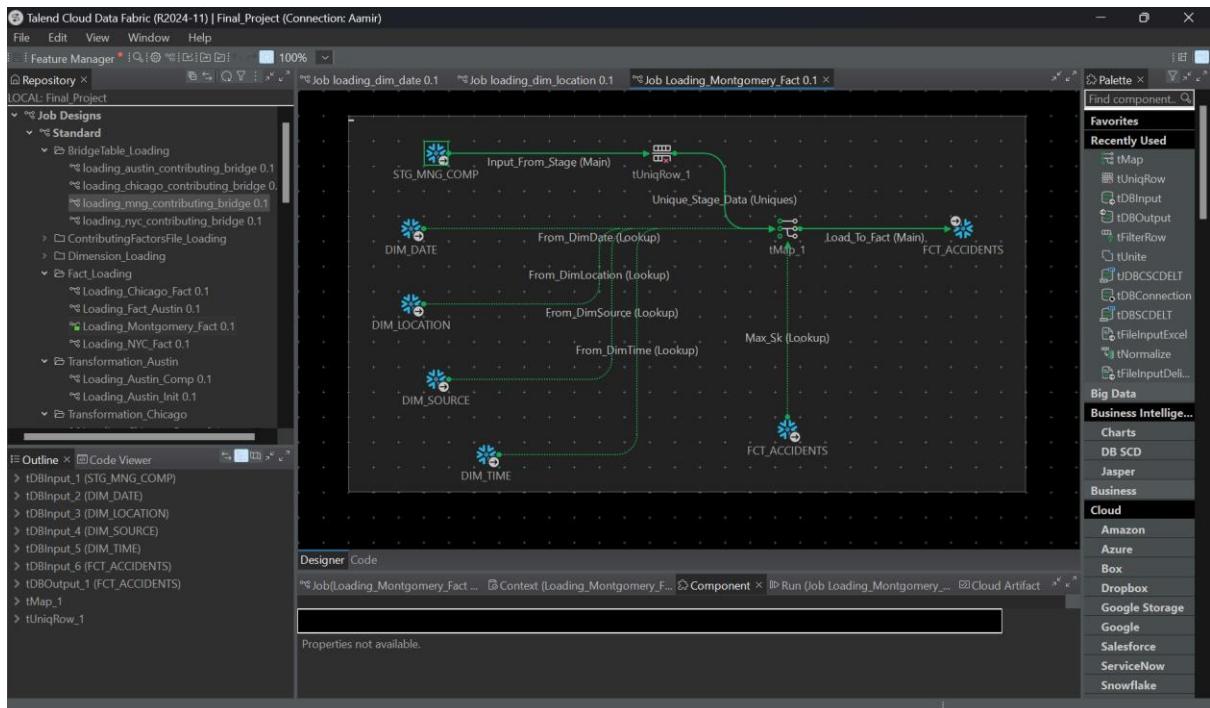
The results table shows one row with the value 851357.

COUNT(*)
851357

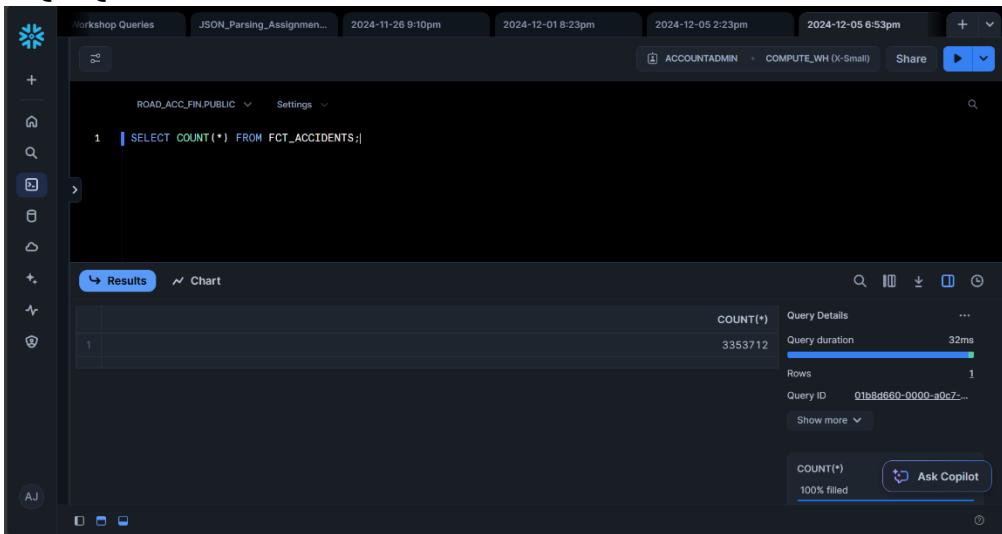
○ DIM SOURCE LOAD



○ MONTGOMERY ACCIDENT FACT



○ SQL QUERY



A screenshot of a Snowflake query results page. The query is:

```
SELECT COUNT(*) FROM FCT_ACCIDENTS;
```

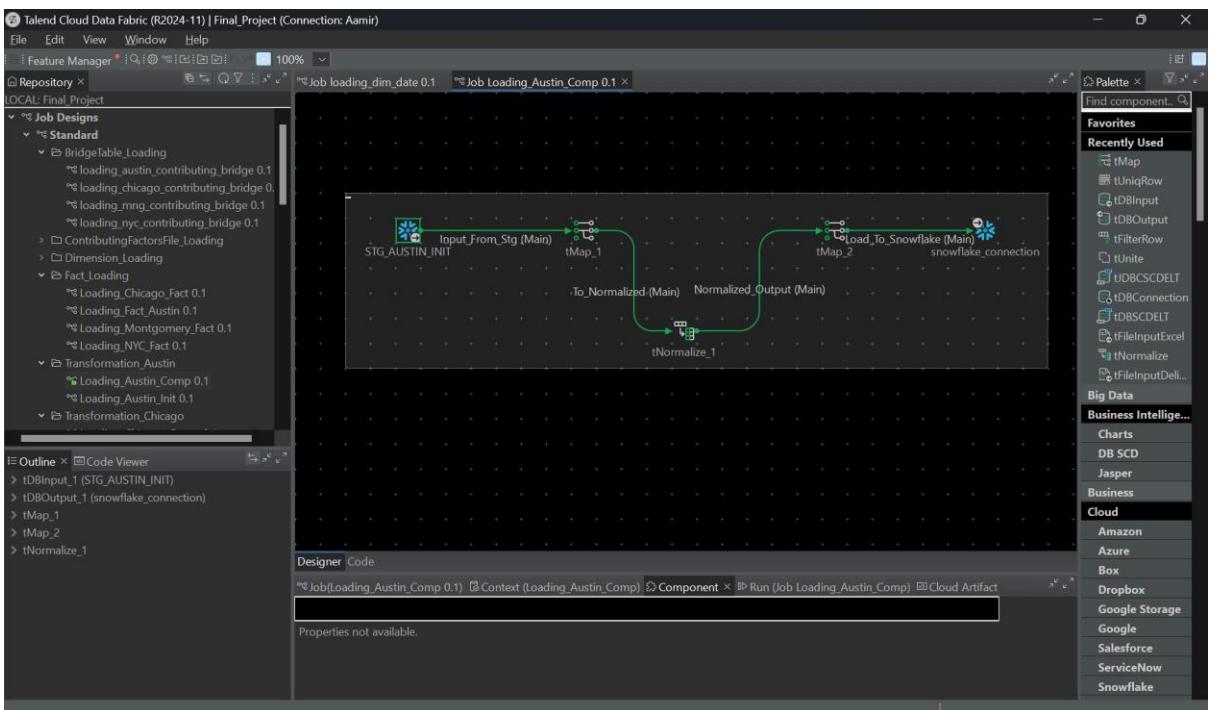
The results table shows one row with the value 3353712. The Query Details panel indicates a query duration of 32ms and 1 row.

COUNT(*)
3353712

Query Details

- Query duration: 32ms
- Rows: 1
- Query ID: 01b8d660-0000-a0c7-...

○ AUSTIN COMP STAGE LOAD

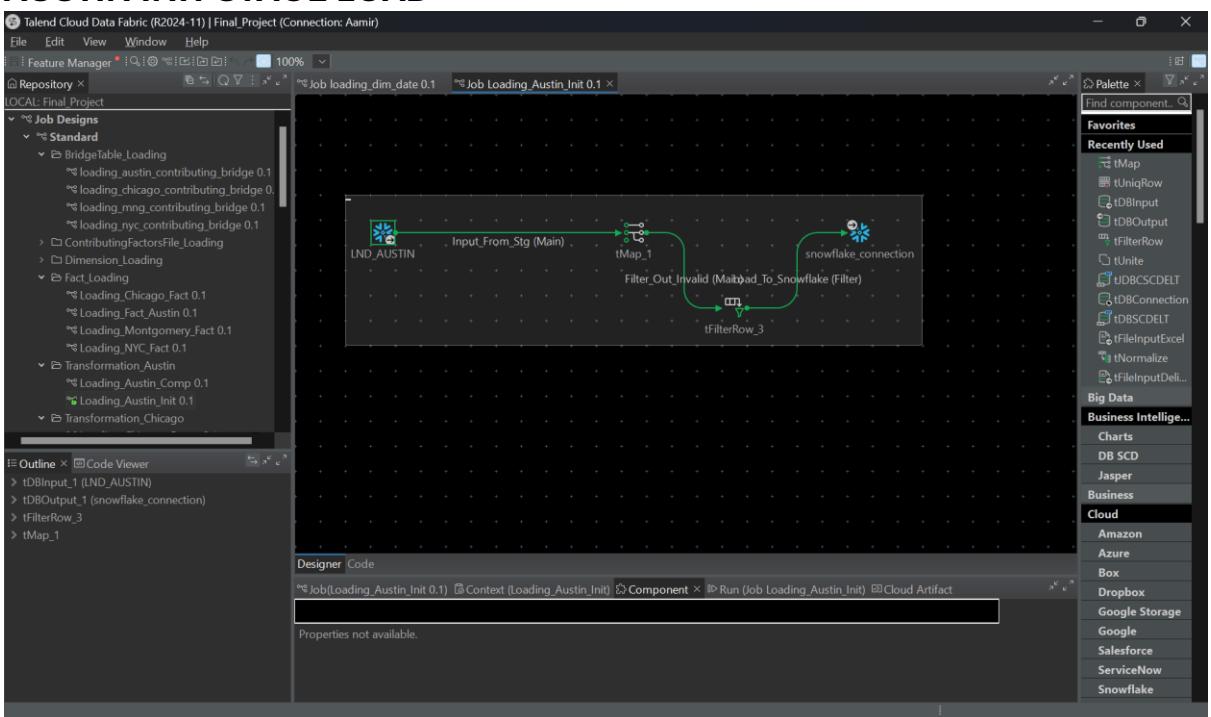


○ SQL QUERY

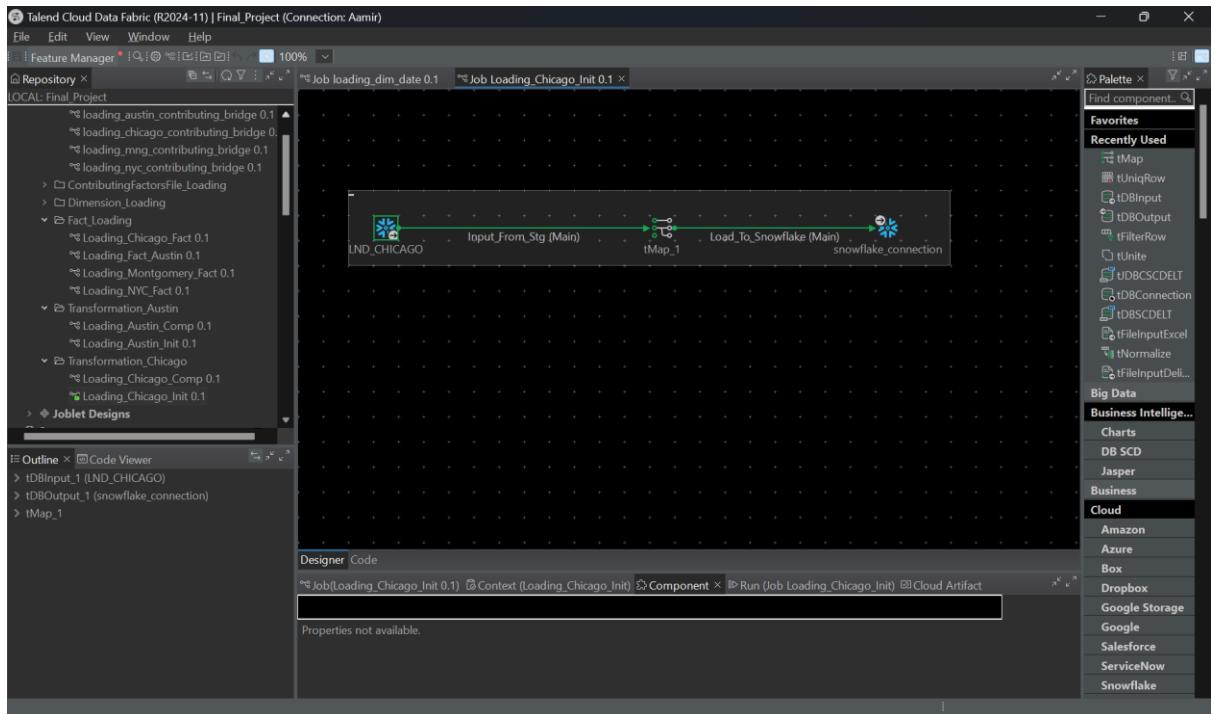
The screenshot shows a SQL query results interface. At the top, it says "ROAD_ACC_FIN.PUBLIC" and "Settings". Below that is the SQL command: "SELECT COUNT(*) FROM STG_AUSTIN_COMP;". At the bottom, there are two tabs: "Results" (which is selected) and "Chart". The results table has one row with the value "321590".

COUNT(*)
321590

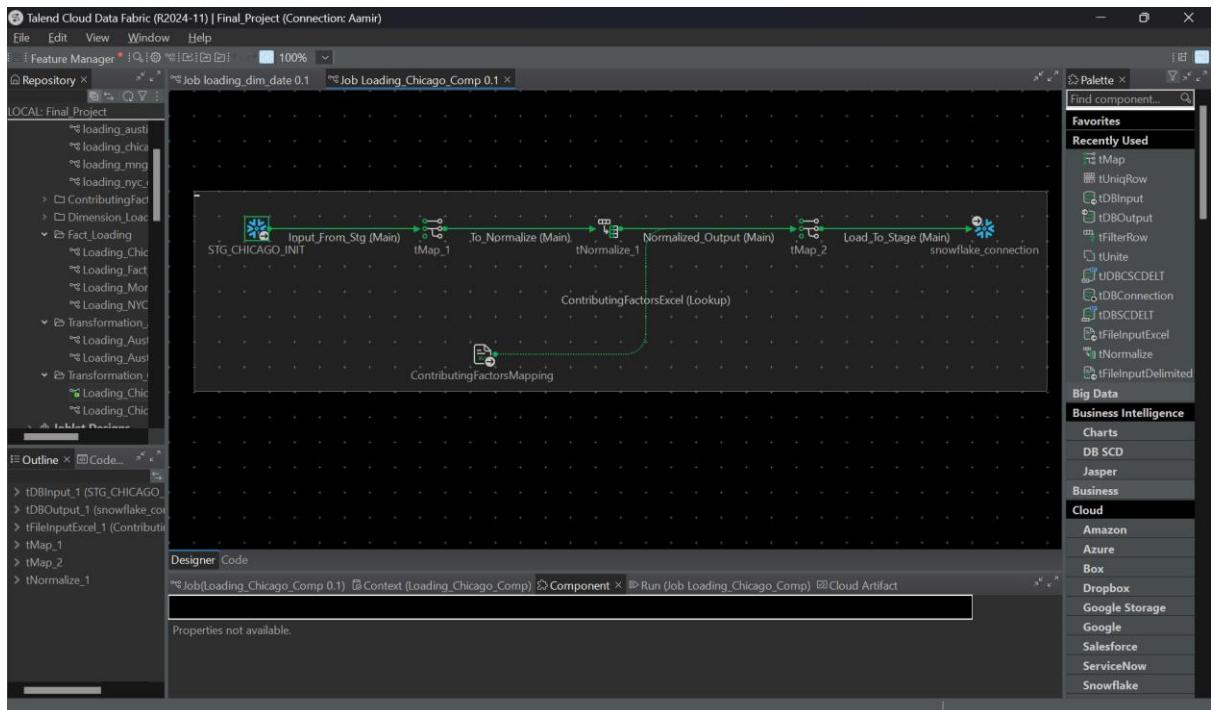
○ AUSTIN INIT STAGE LOAD



○ CHICAGO INIT STAGE LOAD



○ CHICAGO COMP STAGE LOAD



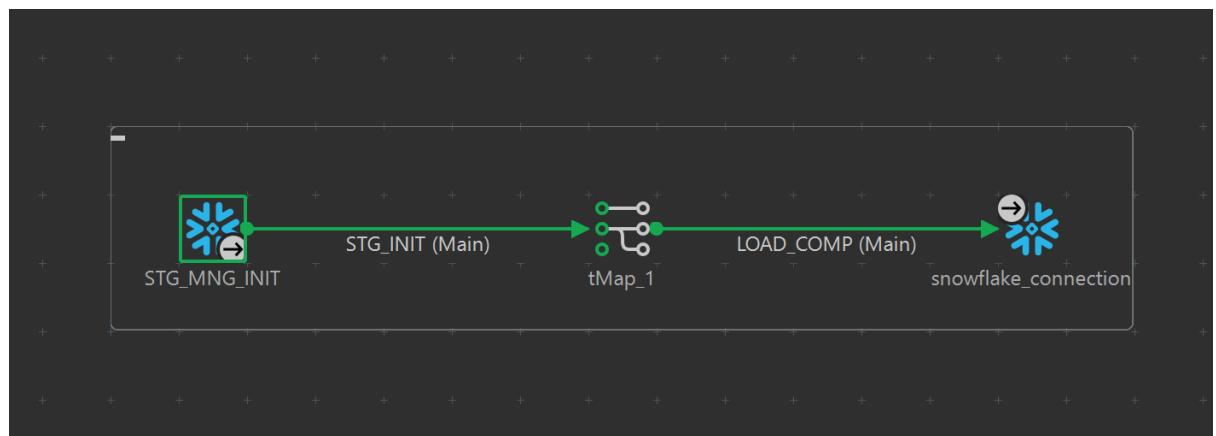
SQL QUERY

```
SELECT COUNT(*) FROM STG_CHICAGO_COMP;
```

Results Chart

COUNT(*)
1792912

- **MONTGOMERY COMP STAGE LOAD**



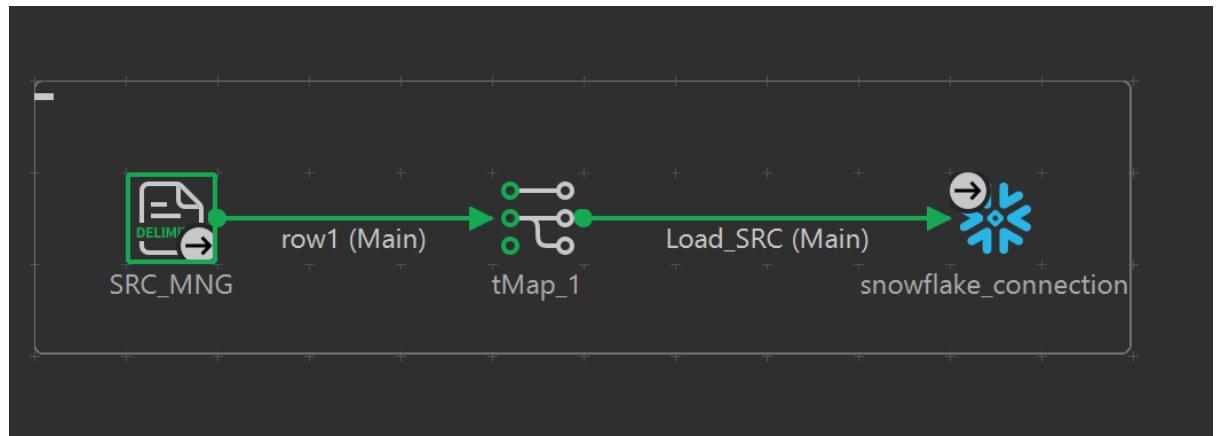
- **SQL QUERY**

```
ROAD_ACC_FIN.PUBLIC ▾ Settings ▾  
SELECT COUNT(*) FROM STG_MNG_COMP;
```

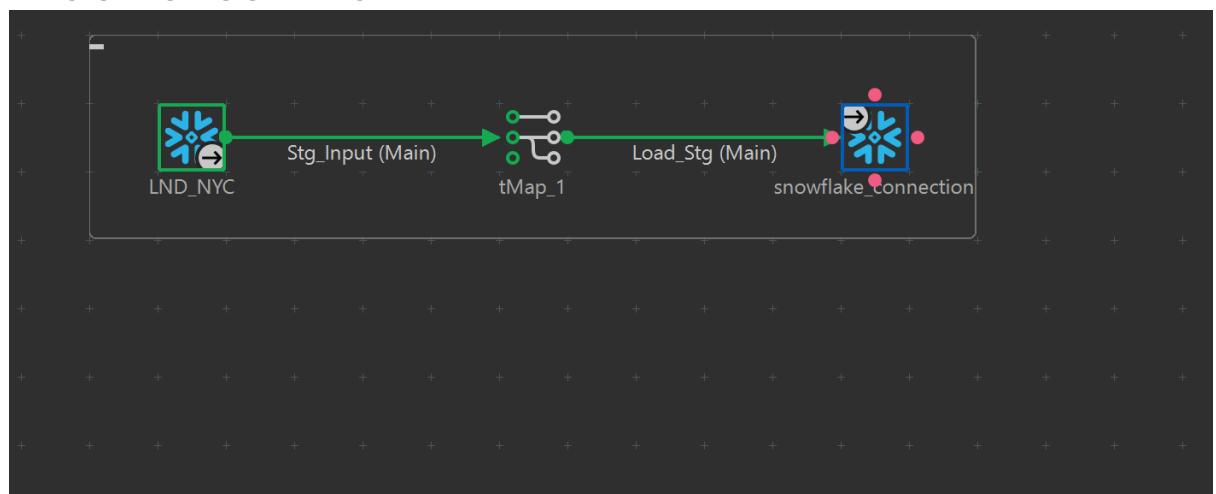
Results Chart

COUNT(*)
107003

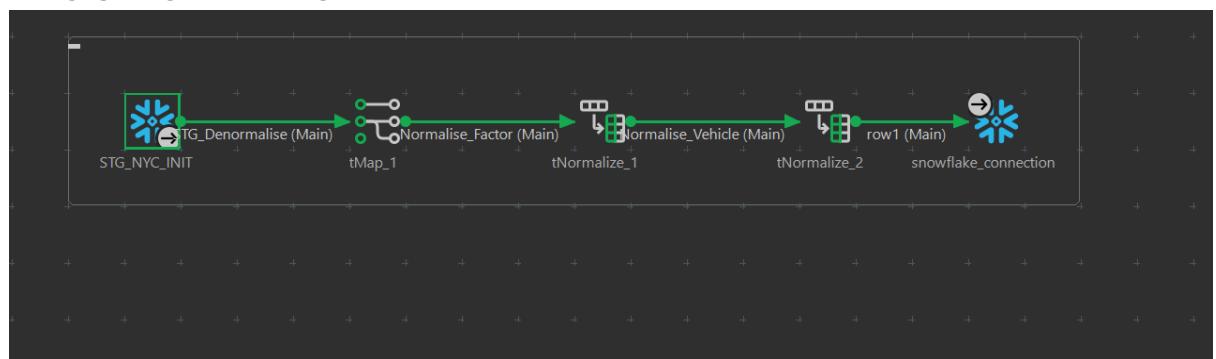
- **MONTGOMERY INIT STAGE LOAD**



- **NYC STAGE COMP LOAD**



- **NYC STAGE DEV LOAD**



- **NYC STAGE INIT LOAD**

