DAMG7370

DESIGNING ADVANCED DATA ARCHITECTURES FOR BUSINESS INTELLIGENCE

FINAL PROJECT - Motor Vehicle Collisions

Under: Prof. Naveen Kuragayala

GROUP - 2

Aamir Jawadwala
Yash Khavnekar
Niki Choksi

DATASETS:

AUSTIN	NYC	CHICAGO	MONTGOMERY
ID	CRASH_DATE	CRASH_RECORD_ID	REPORT_NUMBER
CRASH_ID	CRASH_TIME	CRASH_DATE_EST_I	LOCAL_CASE_NUMBER
CRASH_FATAL_FL	BOROUGH	CRASH_DATE	AGENCY_NAME
CASE_ID	ZIP_CODE	POSTED_SPEED_LIMIT	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_CONDITION	ROUTE_TYPE
RPT_STREET_SFX	ON_STREET_NAME	LIGHTING_CONDITION	LANE_DIRECTION
CRASH_SPEED_LIMIT	CROSS_STREET_NAME	FIRST_CRASH_TYPE	LANE_TYPE
ROAD_CONSTR_ZONE_FL	OFF_STREET_NAME	TRAFFICWAY_TYPE	NUMBER_OF_LANES
LATITUDE	NUMBER_OF_PERSONS_INJURED	LANE_CNT	DIRECTION
LONGITUDE	NUMBER_OF_PERSONS_KILLED	ALIGNMENT	DISTANCE
CRASH_SEV_ID	NUMBER_OF_PEDESTRIANS_INJURED	ROADWAY_SURFACE_COND	DISTANCE_UNIT
SUS_SERIOUS_INJRY_CNT	NUMBER_OF_PEDESTRIANS_KILLED	ROAD_DEFECT	ROAD_GRADE
POSS_INJRY_CNT	NUMBER_OF_CYCLIST_INJURED	REPORT_TYPE	ROAD_NAME
NON_INJRY_CNT	NUMBER_OF_CYCLIST_KILLED	CRASH_TYPE	CROSS_STREET_NAME
UNKN_INJRY_CNT	NUMBER_OF_MOTORIST_INJURED	INTERSECTION_RELATED_I	OFF_ROAD_DESCRIPTION
TOT_INJRY_CNT	NUMBER_OF_MOTORIST_KILLED	NOT_RIGHT_OF_WAY_I	MUNICIPALITY
DEATH_CNT	CONTRIBUTING_FACTOR_VEHICLE_1	HIT_AND_RUN_I	RELATED_NON_MOTORIST
UNITS_INVOLVED	CONTRIBUTING_FACTOR_VEHICLE_2	DAMAGE	AT_FAULT
POINT	CONTRIBUTING_FACTOR_VEHICLE_3	DATE_POLICE_NOTIFIED	COLLISION_TYPE
MOTOR_VEHICLE_DEATH_COUNT	CONTRIBUTING_FACTOR_VEHICLE_4	PRIM_CONTRIBUTORY_CAUSE	WEATHER
MOTOR_VEHICLE_SERIOUS_INJURY_COUNT	CONTRIBUTING_FACTOR_VEHICLE_5	SEC_CONTRIBUTORY_CAUSE	SURFACE_CONDITION
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, Chicago and Montgomery 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.

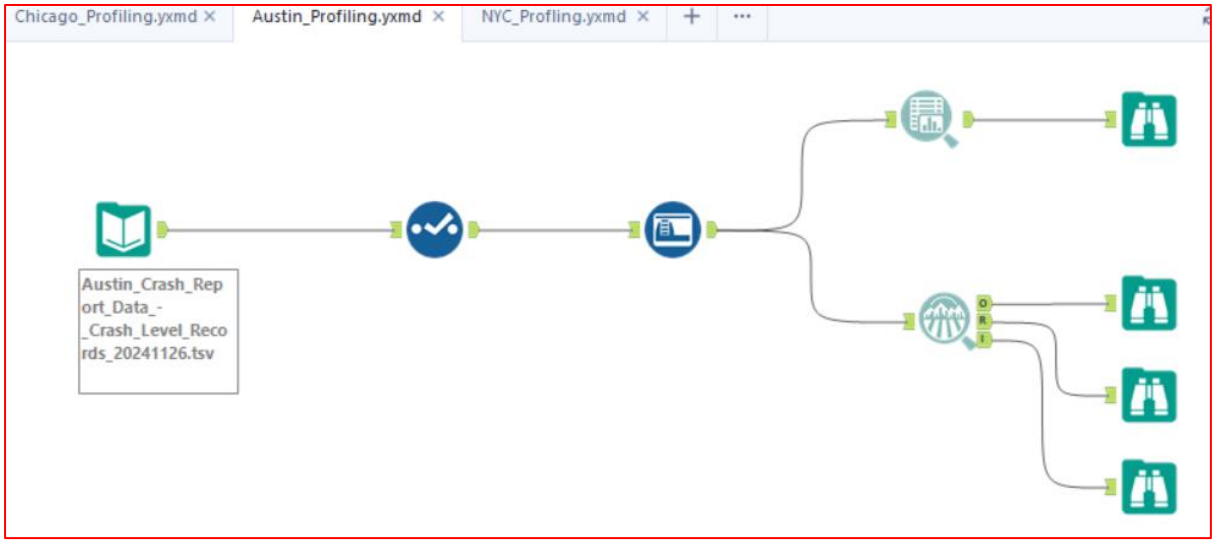
4. **Confirmed**
   NYC: Cross-reference LATITUDE and LONGITUDE with NYC mapping services.
   Austin: Validate geographic coordinates against official Austin maps.
   Chicago: Use GIS tools to confirm LATITUDE and LONGITUDE values fall within Chicago's boundaries.
   Montgomery: Verify geographic data against reliable mapping sources.

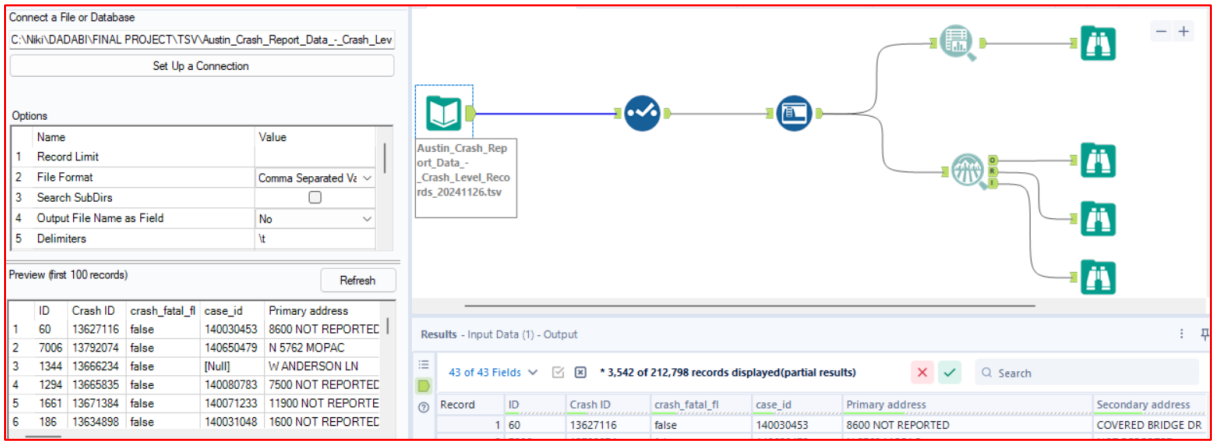
5. **Current**
   NYC: Ensure CRASH_DATE reflects the most recent data.
   Austin: Confirm CRASH_TIMESTAMP_US_CENTRAL includes the latest incidents.
   Chicago: Verify CRASH_DATE is up-to-date.
   Montgomery: Validate that CRASH DATE/TIME covers recent crash data and remove outdated records.

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	Confirm all fields related to injuries and fatalities (e.g.,

	necessary aspects and elements.	SUS_SERIOUS_INJRY_CNT, DEATH_CNT) are populated.
Confirmed	Validates that data is accurate and verified against reliable sources.	Cross-reference LATITUDE and LONGITUDE with mapping services for accuracy.
Current	Confirms that the dataset is up-to-date and relevant for the intended analysis.	Verify the CRASH_TIMESTAMP_US_CENTRAL field reflects recent crash incidents.

Field Analysis (Austin Dataset)

Field	Description	Analysis
ID	Unique identifier for each record.	Ensure uniqueness and validate as a non-null numeric field.
CRASH_ID	Unique identifier for each crash incident.	Validate as a non-null numeric field to maintain integrity across records.
CRASH_FATAL_FL	Indicates if the crash was fatal ('Y' or 'N').	Ensure valid binary values ('Y' or 'N') and check for consistency with DEATH_CNT.
CASE_ID	Identifier for the case associated with the crash.	Verify against official case records to ensure accuracy.
PRIMARY_ADDRESS	Primary address where the crash occurred.	Cross-reference with geographical data to ensure location accuracy.
SECONDARY_ADDRESS	Secondary address related to the crash location, if applicable.	Validate secondary address details for completeness.
RPT_BLOCK_NUM	Reported block number where the crash occurred.	Ensure that block numbers align with reported street data.
RPT_STREET_NAME	Reported street name where the crash occurred.	Validate street names against official mapping tools.
RPT_STREET_SFX	Suffix of the reported street name (e.g., St, Ave).	Standardize suffix values for consistency (e.g., "Street" vs. "St").

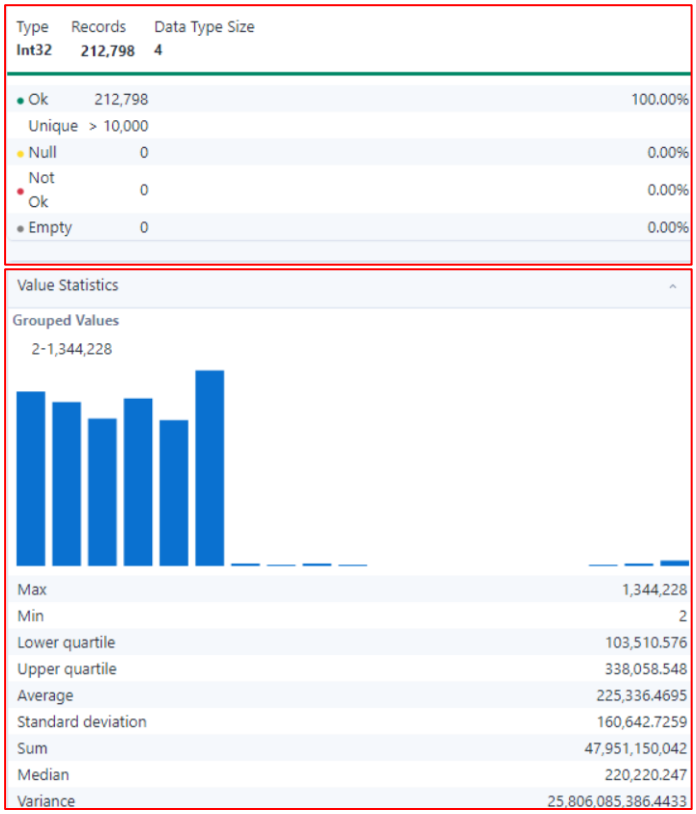
CRASH_SPEED_LIMIT	Speed limit at the location of the crash.	Check for reasonable values and flag outliers (e.g., unusually high or low limits).
ROAD_CONST_ZONE_FL	Indicates if the crash occurred in a road construction zone ('Y' or 'N').	Ensure valid binary values ('Y' or 'N') and verify with associated construction zone data.
LATITUDE	Latitude coordinate of the crash location.	Validate coordinates fall within Austin's geographic boundaries.
LONGITUDE	Longitude coordinate of the crash location.	Cross-check coordinates with mapping tools for accuracy.
CRASH_SEV_ID	Severity identifier of the crash.	Validate severity levels correspond to the injury and fatality data.
SUS_SERIOUS_INJRY_CNT	Suspected serious injury count resulting from the crash.	Ensure consistency with TOT_INJRY_CNT and validate against case records.
POSS_INJRY_CNT	Possible injury count resulting from the crash.	Validate that this count logically aligns with total injuries.
NON_INJRY_CNT	Non-injury count (individuals involved without injuries).	Ensure logical alignment with total units involved and injury counts.
UNKN_INJRY_CNT	Unknown injury count (injury status not determined).	Investigate and minimize the frequency of unknown values to improve data completeness.
TOT_INJRY_CNT	Total injury count resulting from the crash.	Confirm alignment with individual injury counts (SUS_SERIOUS_INJRY_CNT, POSS_INJRY_CNT, etc.).
DEATH_CNT	Total number of fatalities resulting from the crash.	Cross-check consistency with

		fatality counts across different units (e.g., MOTOR_VEHICLE_DEATH_COUNT).
UNITS_INVOLVED	Number of units (vehicles, pedestrians, etc.) involved in the crash.	Validate against individual records of injuries and fatalities for consistency.
POINT	Geographical point representing the crash location.	Ensure POINT corresponds to the LATITUDE and LONGITUDE fields.
MOTOR_VEHICLE_DEATH_COUNT	Number of motor vehicle occupant fatalities.	Validate against DEATH_CNT and check for logical alignment.
MOTOR_VEHICLE_SERIOUS_INJURY_COUNT	Number of motor vehicle occupants with serious injuries.	Ensure this count is consistent with total injuries (TOT_INJRY_CNT).
BICYCLE_DEATH_COUNT	Number of bicyclist fatalities.	Cross-check with DEATH_CNT and validate against case data.
BICYCLE_SERIOUS_INJURY_COUNT	Number of bicyclists with serious injuries.	Ensure alignment with total injuries and injury severity fields.
MOTORCYCLE_DEATH_COUNT	Number of motorcyclist fatalities.	Validate against DEATH_CNT for consistency.
OTHER_DEATH_COUNT	Number of fatalities not categorized above.	Investigate and validate against case records for completeness.
OTHER_SERIOUS_INJURY_COUNT	Number of serious injuries not categorized above.	Ensure logical consistency with total injury counts.
ONSYS_FL	On-system flag indicating if the road is part of the state highway system.	Validate binary values and check for alignment with location data.
PRIVATE_DR_FL	Indicates if the crash occurred on a private drive ('Y' or 'N').	Cross-check for logical consistency with address details.

MICROMOBILITY_SERIOUS_INJURY_COUNT	Number of serious injuries involving micromobility devices (e.g., e-scooters).	Ensure this count is included in total injuries and validate against micromobility records.
MIRCOMOBILITY_DEATH_COUNT	Number of fatalities involving micromobility devices.	Cross-check with DEATH_CNT and investigate completeness.
CRASH_TIMESTAMP_US_CENTRAL	Timestamp of the crash in US Central time.	Validate timestamp accuracy and ensure consistency with CRASH_DATE.
CRASH_TIMESTAMP	General timestamp of the crash.	Ensure consistency with CRASH_TIMESTAMP_US_CENTRAL.
IS_DELETED	Indicates if the record has been deleted ('Y' or 'N').	Verify binary values and investigate records flagged for deletion.
IS_TEMPORARY_RECORD	Indicates if the record is temporary ('Y' or 'N').	Check for logical use of temporary flags and ensure proper updates.
LAW_ENFORCEMENT_FATALITY_COUNT	Number of law enforcement fatalities resulting from the crash.	Validate against DEATH_CNT and cross-check with external law enforcement data.
REPORTED_STREET_PREFIX	Prefix of the reported street name (e.g., N, S, E, W).	Ensure prefixes align with Austin's official street data and validate completeness.

Data Observation:

- **ID**
  - Summary:



Type	Records	Data Type Size
Int32	212,798	4

Ok	212,798	100.00%
Unique	> 10,000	
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

Value Statistics

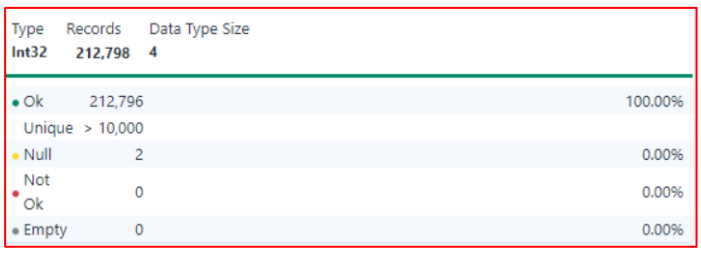
Grouped Values

2-1,344,228

Max	1,344,228
Min	2
Lower quartile	103,510.576
Upper quartile	338,058.548
Average	225,336.4695
Standard deviation	160,642.7259
Sum	47,951,150,042
Median	220,220.247
Variance	25,806,085,386.4433

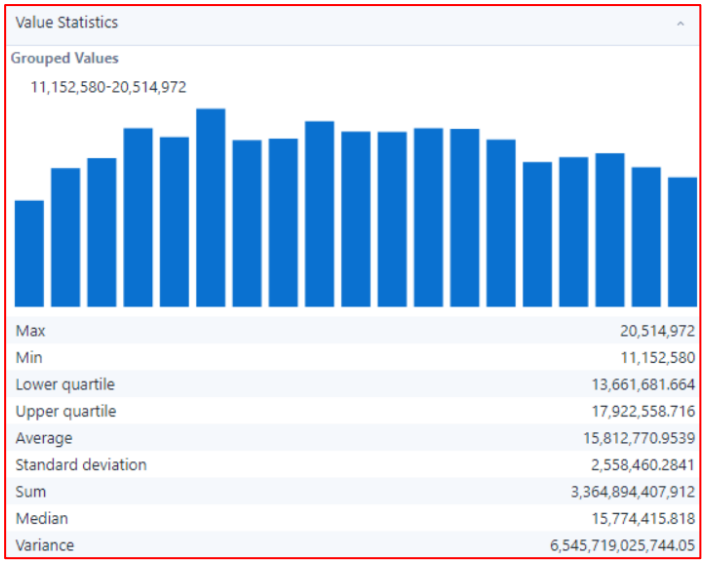
  - Observation:
    - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
    - The field contains over 10,000 unique values, indicating diverse and meaningful entries
    - No missing or not-ok records exist, ensuring completeness and readiness for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

- **CRASH_ID**
  - Summary:



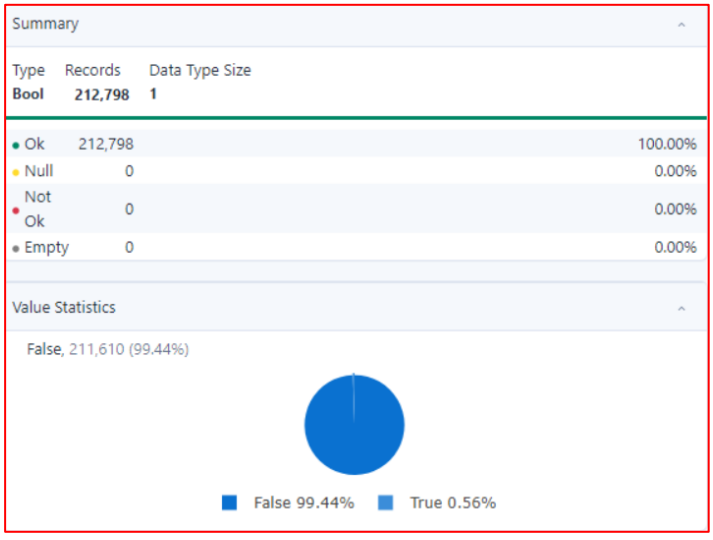
Type	Records	Data Type Size
Int32	212,798	4

Ok	212,796	100.00%
Unique	> 10,000	
Null	2	0.00%
Not Ok	0	0.00%
Empty	0	0.00%



- 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.

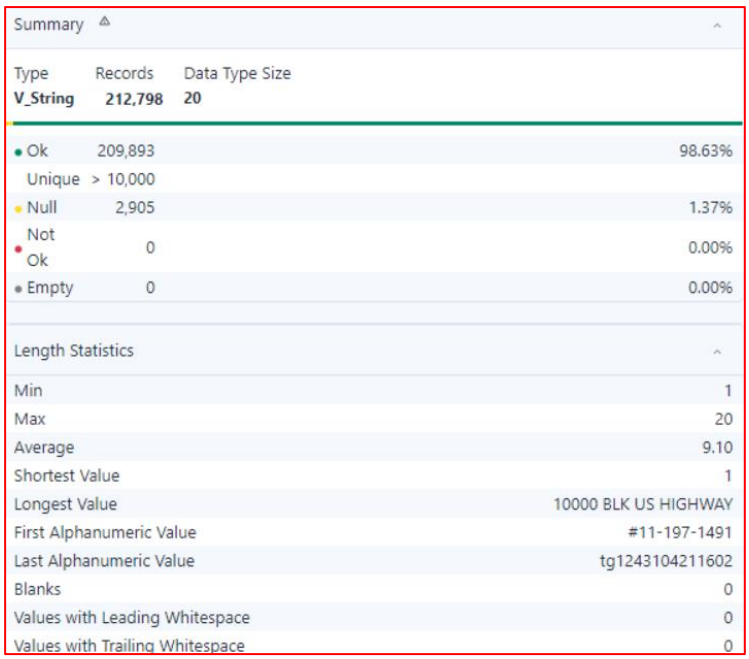
- **CRASH_FATAL_FL**
  - Summary:



- 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

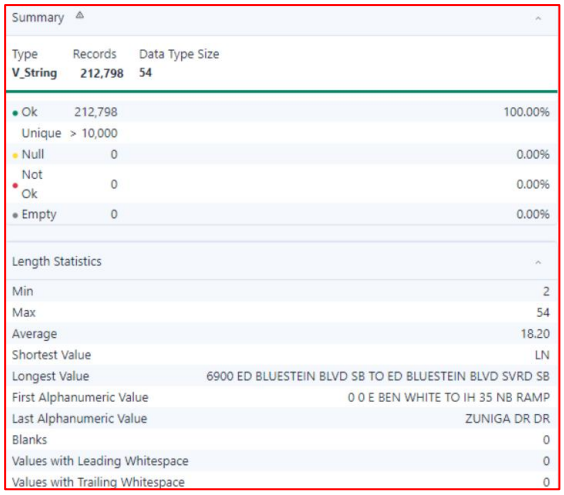
Ok	209,893	98.63%
Unique	> 10,000	
Null	2,905	1.37%
Not Ok	0	0.00%
Empty	0	0.00%

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.

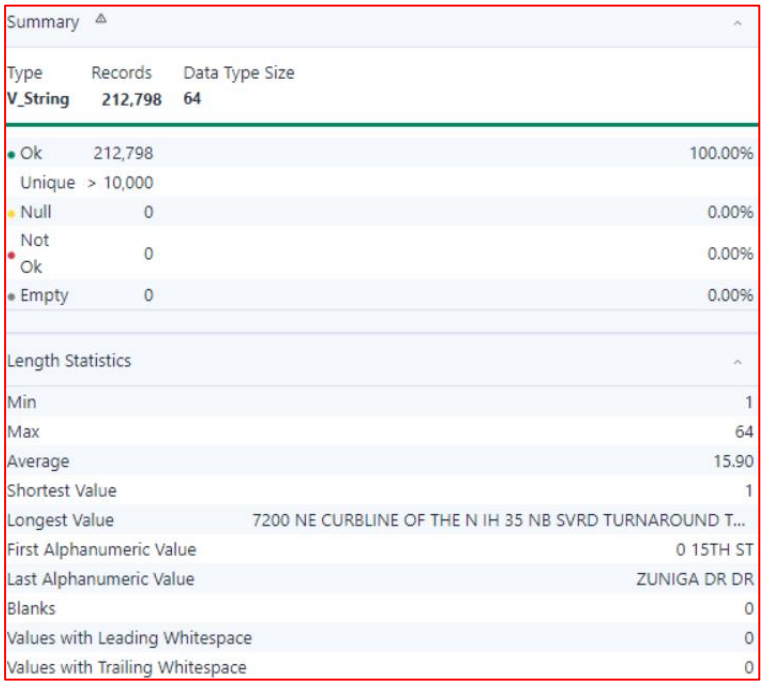
- **PRIMARY_ADDRESS**
  - Summary:



- 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**

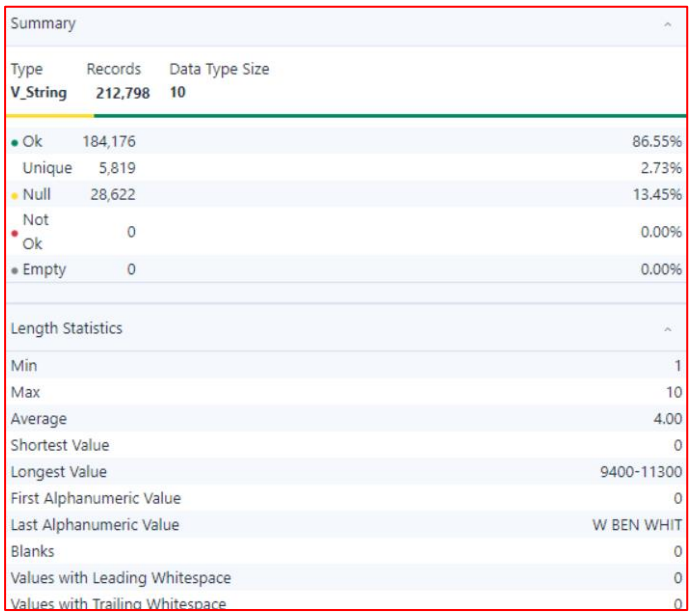
  - Summary:



  - Observation:
    - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability.

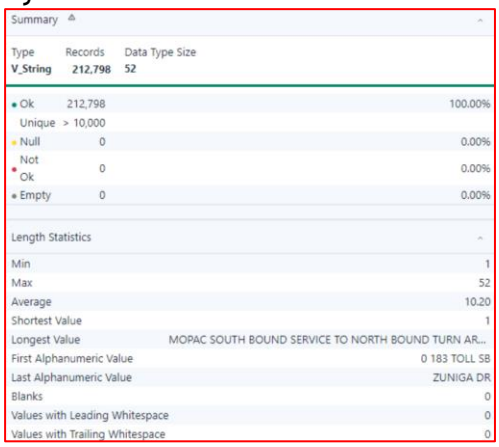
- The field contains over 10,000 unique values, indicating diverse and meaningful entries.
- No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **RPT_BLOCK_NUM**
  - Summary:



  - 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.

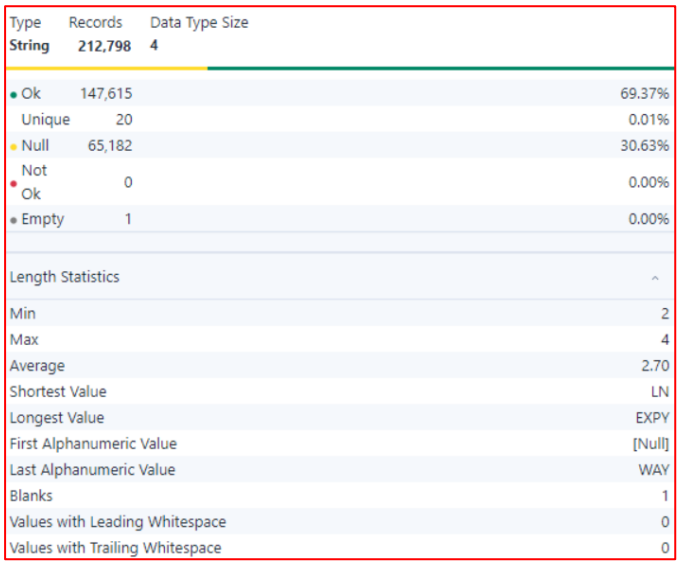
- **RPT_STREET_NAME**
  - Summary:



- 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

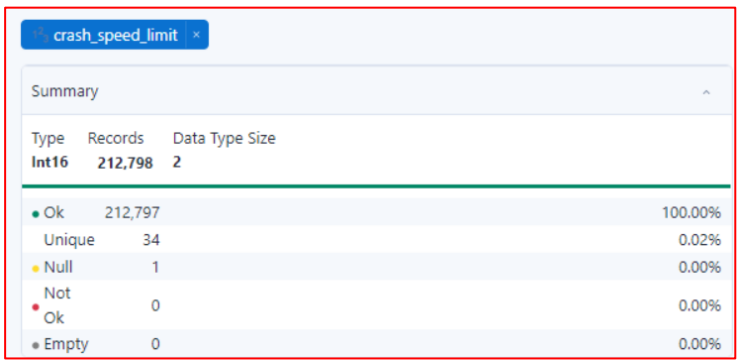
Ok	147,615	69.37%
Unique	20	0.01%
Null	65,182	30.63%
Not Ok	0	0.00%
Empty	1	0.00%

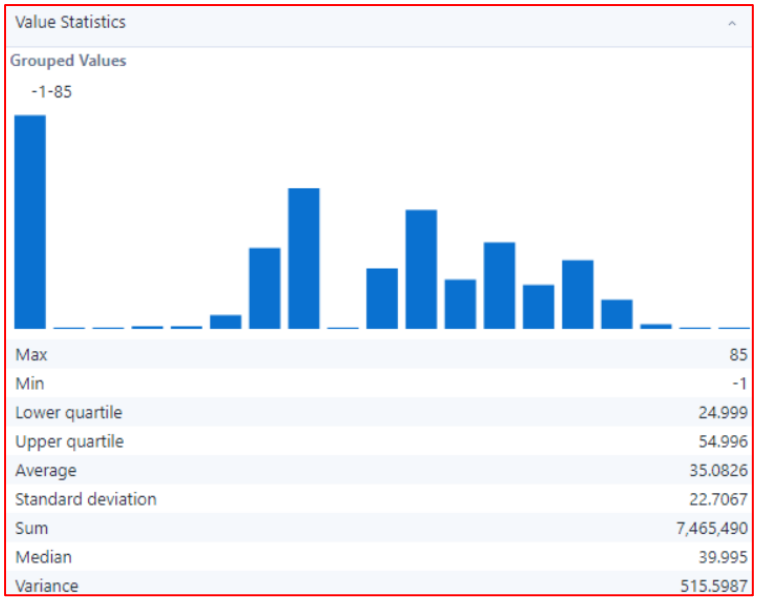
Length Statistics	
Min	2
Max	4
Average	2.70
Shortest Value	LN
Longest Value	EXPY
First Alphanumeric Value	[Null]
Last Alphanumeric Value	WAY
Blanks	1
Values with Leading Whitespace	0
Values with Trailing Whitespace	0

  - 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**

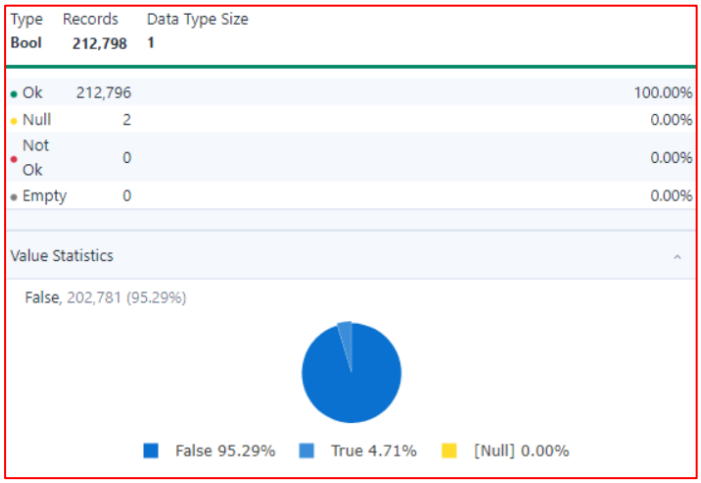
  - Summary:





- 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.

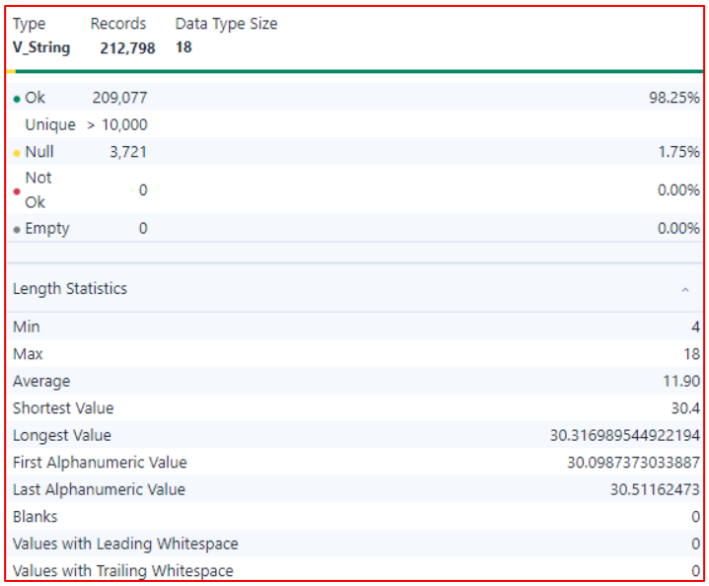
- **ROAD_CONSTR_ZONE_FL**
  - Summary:



  - 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:



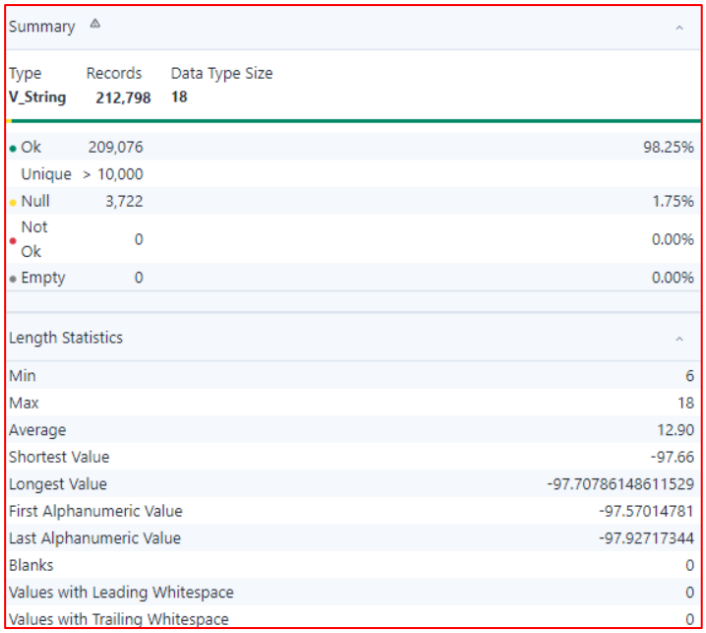
Type	Records	Data Type Size
V_String	212,798	18

Ok	209,077	98.25%
Unique	> 10,000	
Null	3,721	1.75%
Not Ok	0	0.00%
Empty	0	0.00%

Length Statistics	
Min	4
Max	18
Average	11.90
Shortest Value	30.4
Longest Value	30.316989544922194
First Alphanumeric Value	30.0987373033887
Last Alphanumeric Value	30.51162473
Blanks	0
Values with Leading Whitespace	0
Values with Trailing Whitespace	0

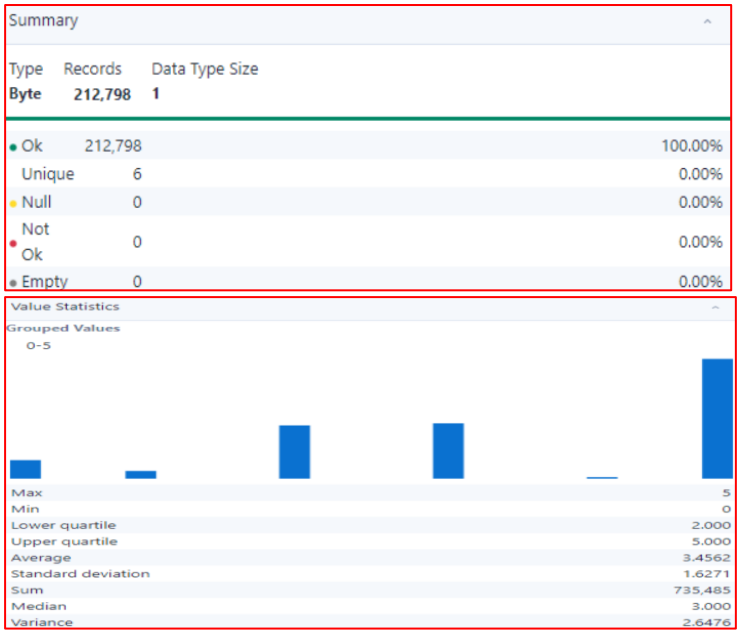
  - Observation:
    - Out of 212,798 records, 209,077 (98.25%) are valid, with 3,721 (1.75%) null values detected, requiring attention to improve data completeness.
    - The field contains over 10,000 unique values, indicating a high degree of variability and meaningful data.
    - No not-ok or empty 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:



- 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.

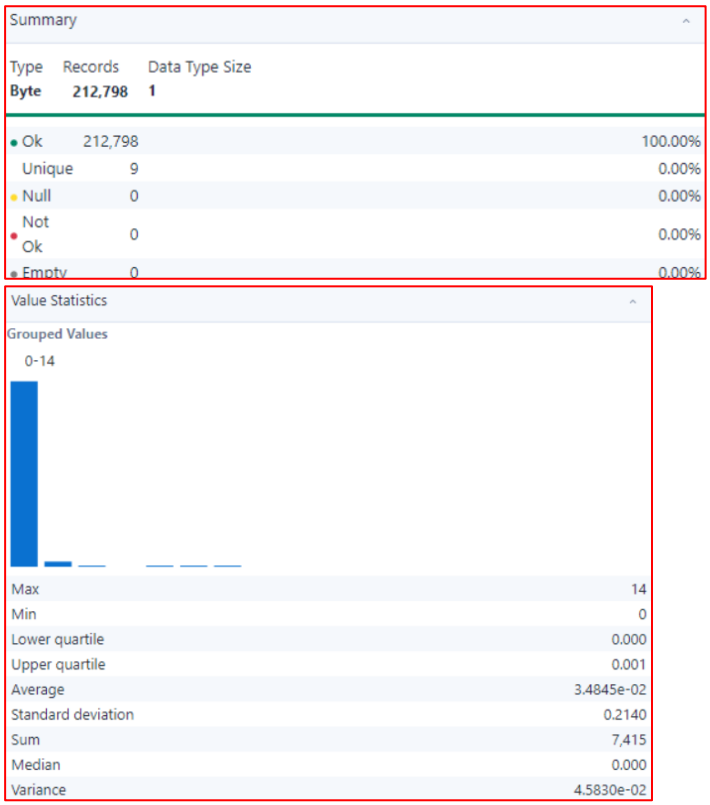
- **CRASH_SEV_ID**
  - Summary:



- 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**

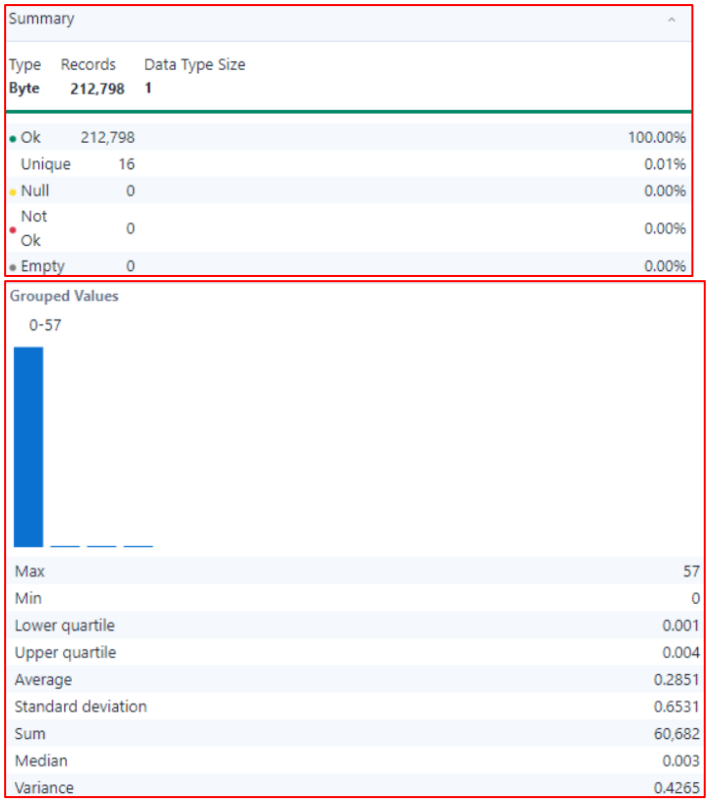
  - Summary:



  - 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.

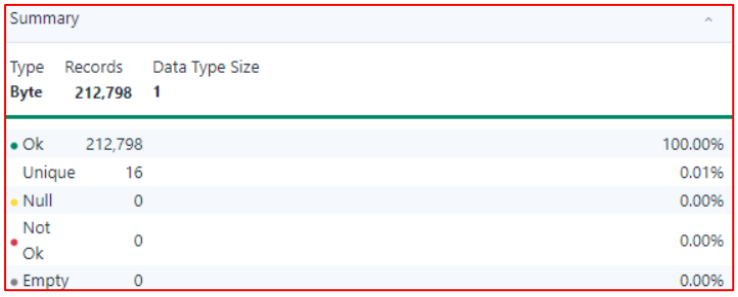
- **NONINCAP_INJRY_CNT**

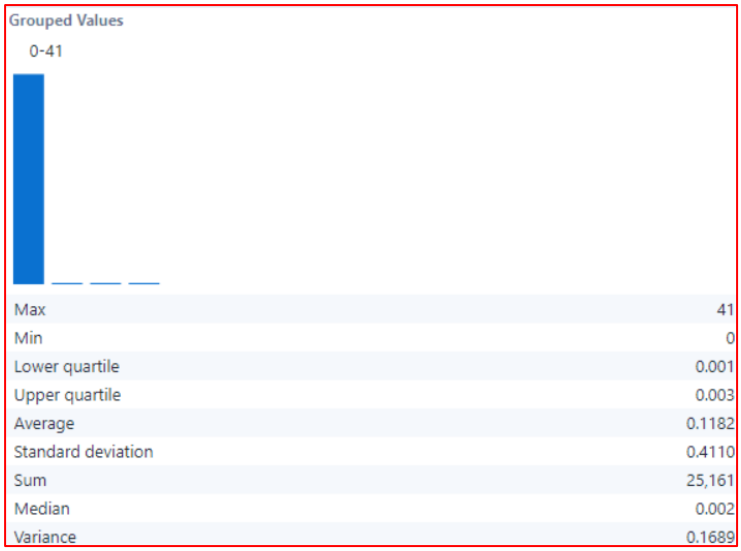
  - Summary:



- 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:

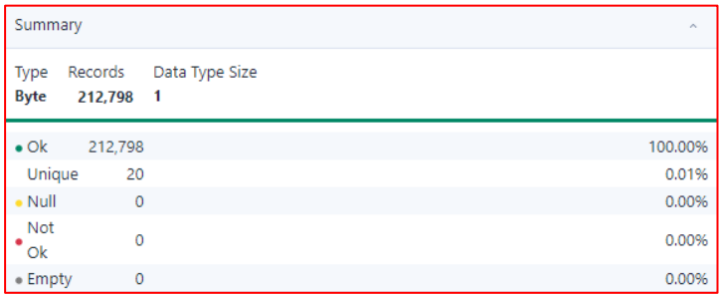


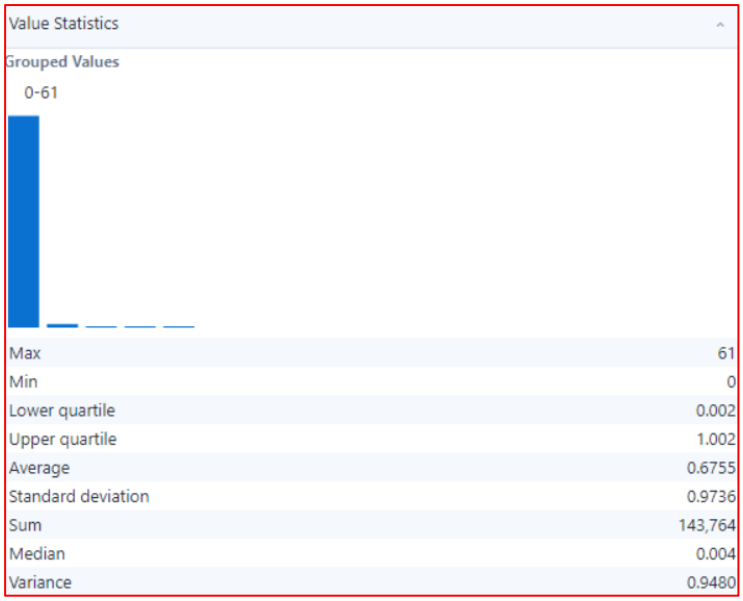


Max	41
Min	0
Lower quartile	0.001
Upper quartile	0.003
Average	0.1182
Standard deviation	0.4110
Sum	25,161
Median	0.002
Variance	0.1689

- 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.

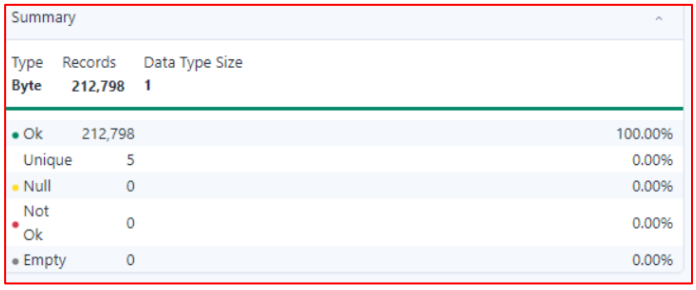
- **SERIOUS_INJRY_CNT**
  - Summary:

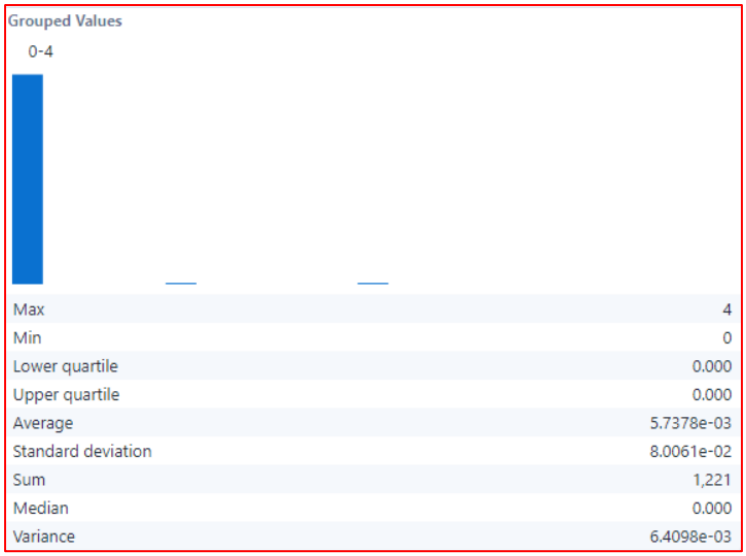




- Observation:
  - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring complete data reliability.
  - The field contains 20 unique values within a range of 0 to 61, suggesting a structured and predefined set of categories.
  - The data is consistent and standardized,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

- **DEATH_CNT**
  - Summary:

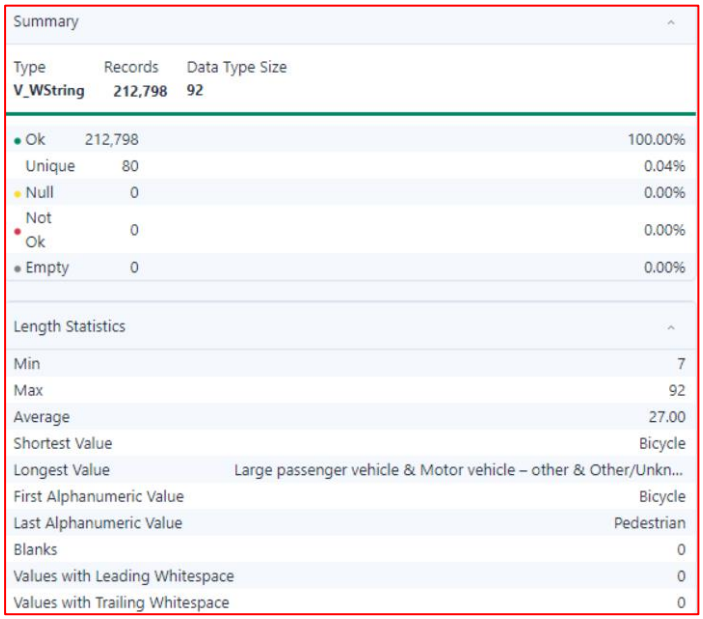




- Observation:
  - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring full data reliability.
  - The field contains 5 unique values ranging from 0 to 4, reflecting a compact and predefined categorical structure.
  - The data appears consistent and standardized, as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

- **UNITS_INVOLVED**

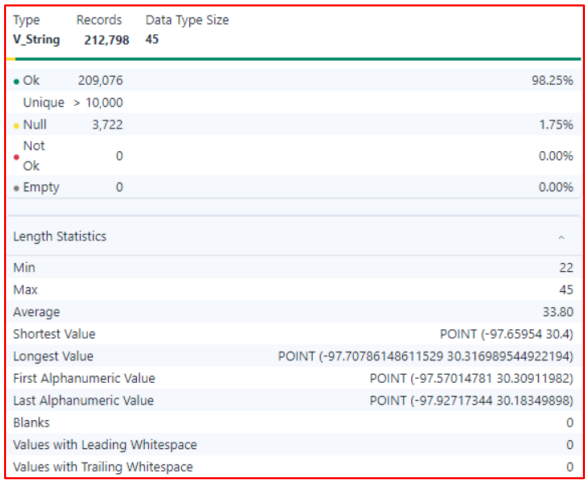
  - Summary:



  - 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:



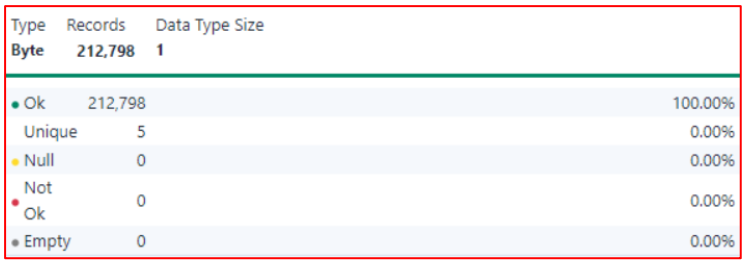
Type	Records	Data Type Size
V_String	212,798	45

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	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

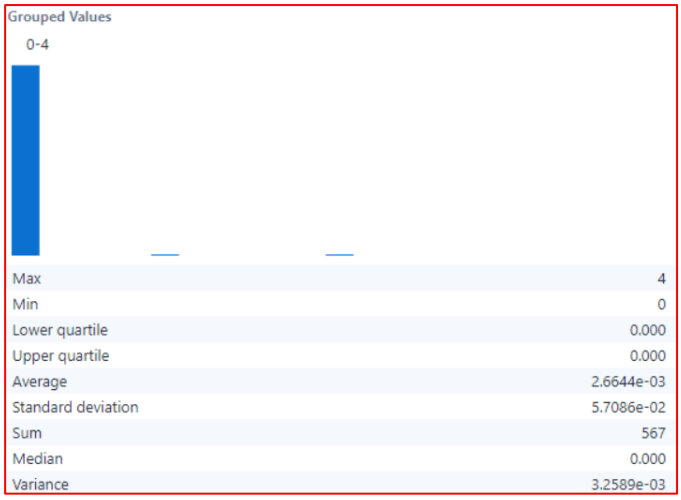
  - Observation:
    - A total of 209,076 records (98.25%) are valid, while 3,722 records (1.75%) contain null values, requiring data imputation or exclusion for analysis completeness.
    - The field contains over 10,000 unique values, suggesting a highly granular and detailed dataset.
    - There are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

- **MOTOR_VEHICLE_DEATH_COUNT**
  - Summary:



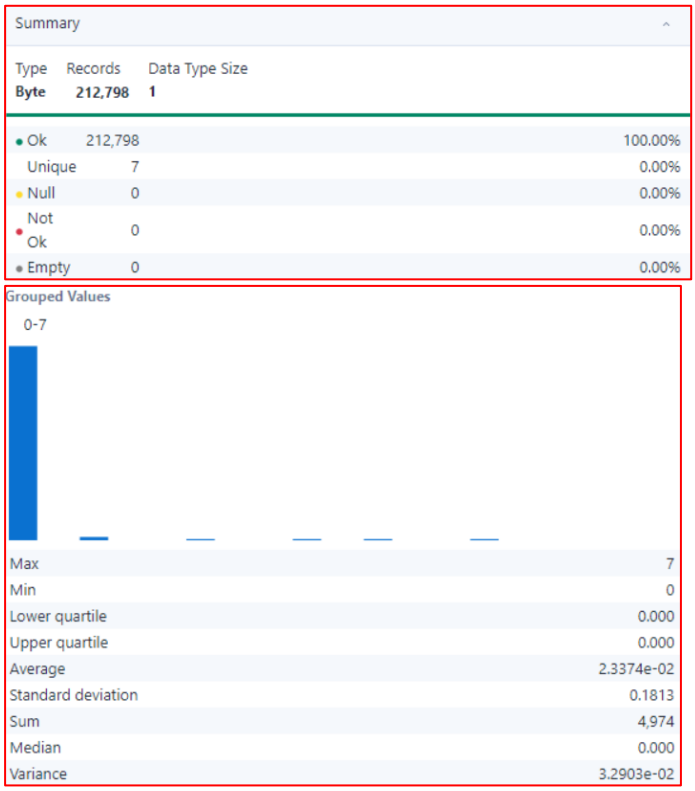
Type	Records	Data Type Size
Byte	212,798	1

Ok	212,798	100.00%
Unique	5	0.00%
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%



- 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

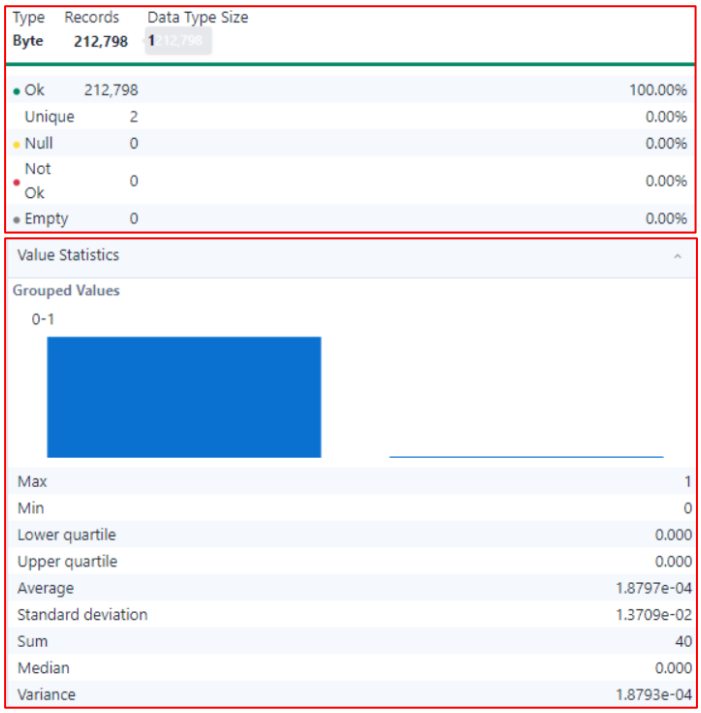
- **MOTOR_VEHICLE_SERIOUS_INJURY_COUNT**
  - Summary:



- 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:



Type	Records	Data Type Size
Byte	212,798	1

Ok	212,798	100.00%
Unique	2	0.00%
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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 Ok	0	0.00%
Empty	0	0.00%

Value Statistics

Grouped Values

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:

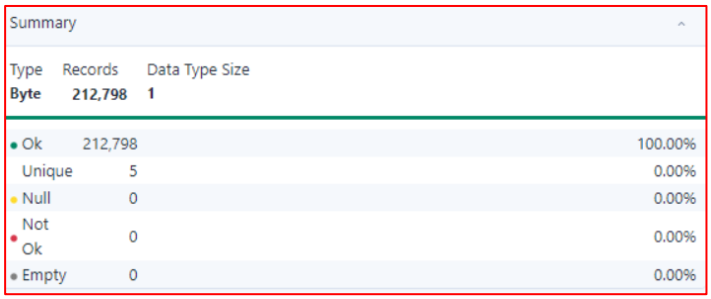
Summary

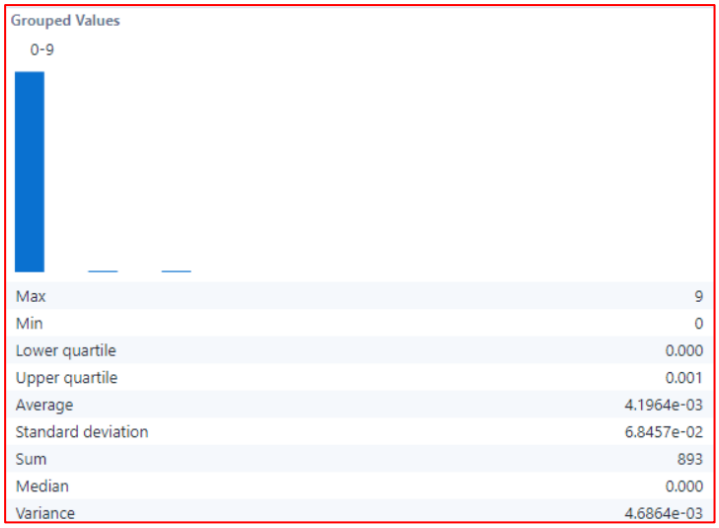
Type	Records	Data Type Size
Byte	212,798	1

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

- 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:

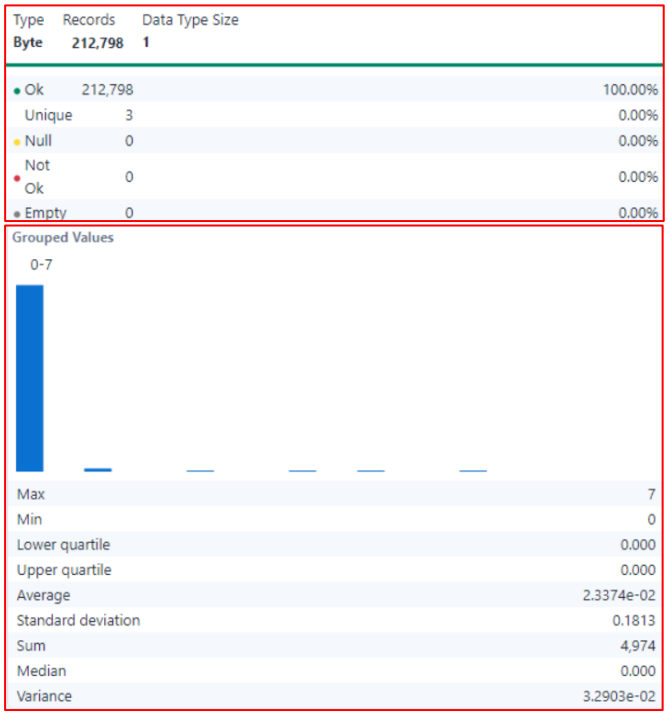




- Observation:
  - All 212,798 records are valid without any null, not-ok, or empty values, ensuring reliability and consistency for analysis.
  - The field comprises 5 unique values, representing limited but distinct categorical data.
  - The data appears to be standardized, and no irregularities or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

- **MOTORCYCLE_SERIOUS_INJURY_COUNT**

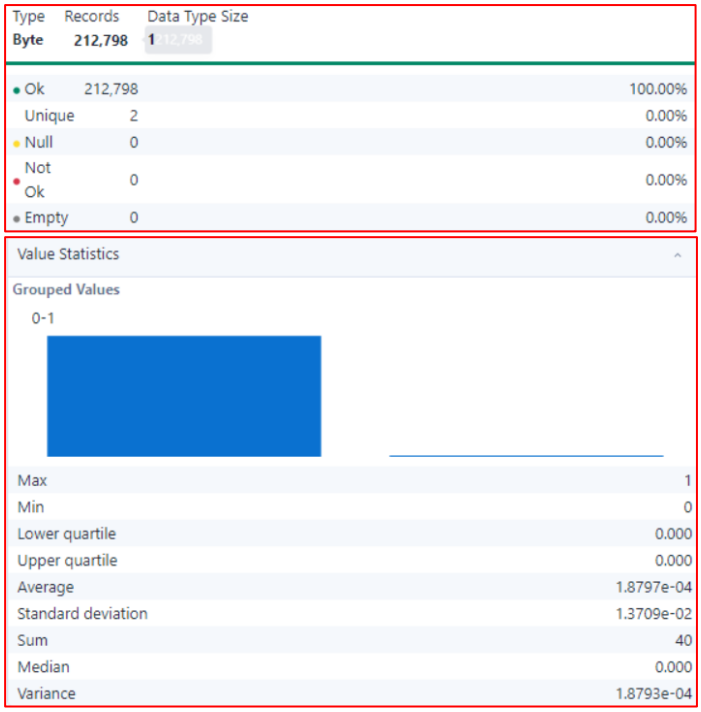
  - Summary:



  - 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:



Type	Records	Data Type Size
Byte	212,798	1

Ok	212,798	100.00%
Unique	2	0.00%
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

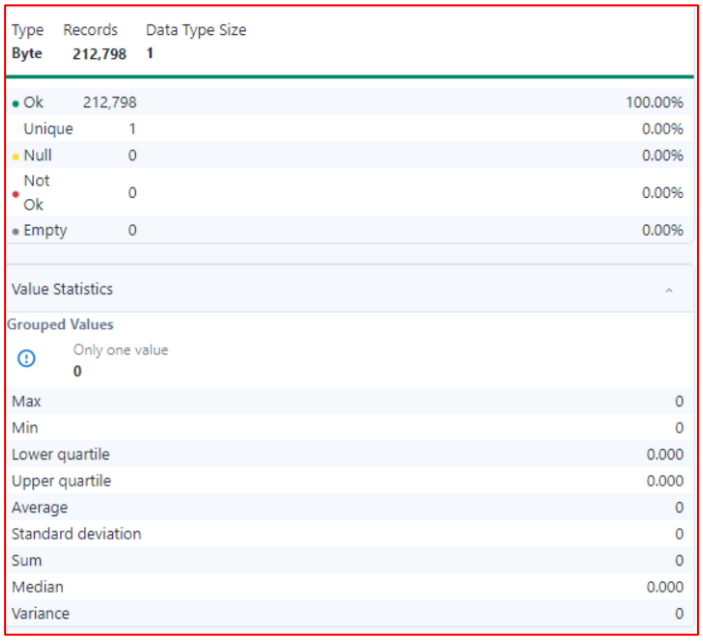
Value Statistics

Grouped Values

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 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

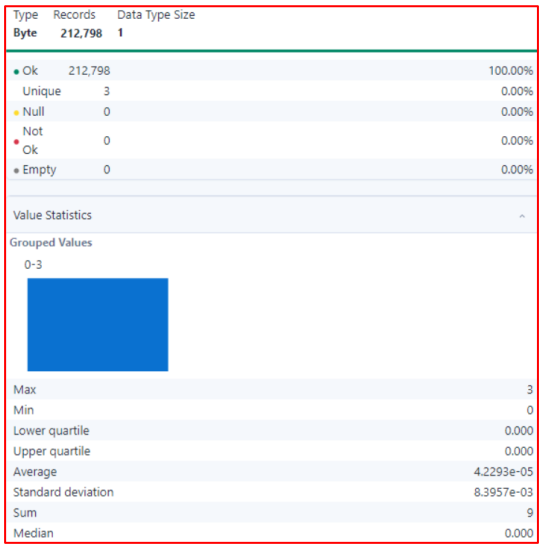
0

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:

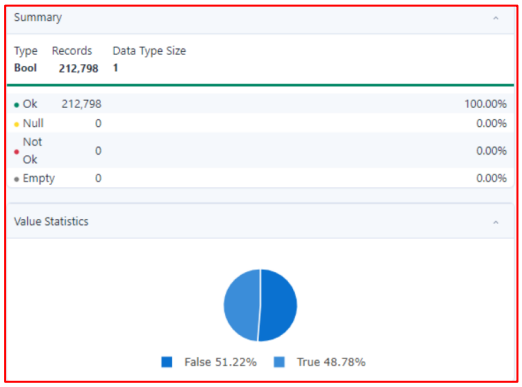
- All 212,798 records are valid with no null, not-ok, or empty values.
- The field contains only one unique value, indicating no variability.
- The value is standardized across all records.
- Statistical measures like variance, standard deviation, and average are zero due to the lack of diversity in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

- **OTHER_SERIOUS_INJURY_COUNT**
  - Summary:



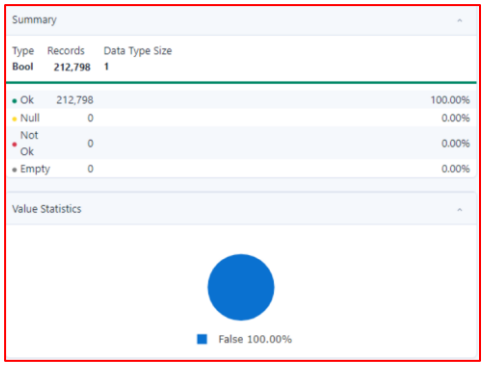
  - Observation:
    - All 212,798 records are valid with no null, not-ok, or empty values.
    - The field contains three unique values, ensuring some variability in the data.
    - The field is consistent and standardized, suitable for categorical or indicator analysis.

- **ONSYS_FL**
  - Summary:



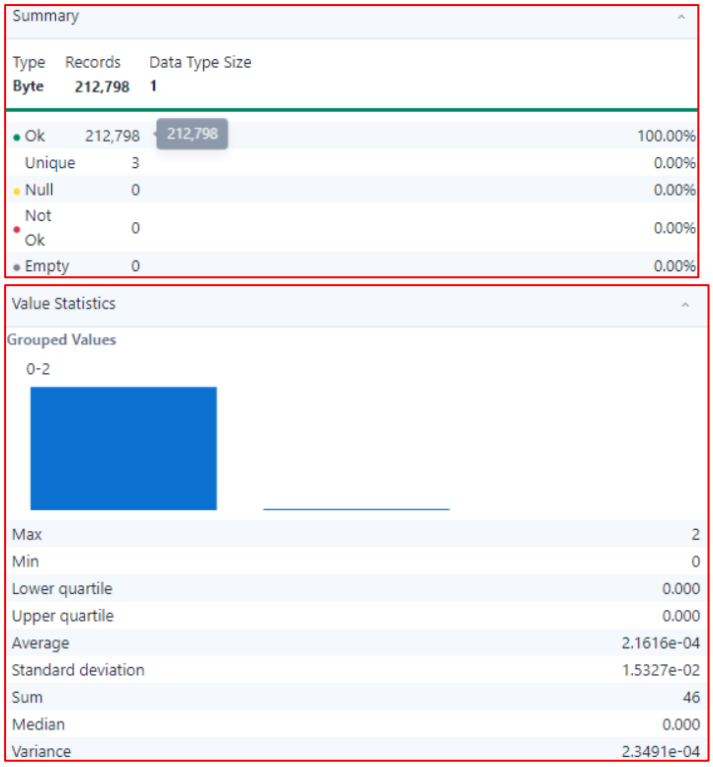
- Observation:
  - All 212,798 records are valid, with no null, empty, or invalid values, ensuring data completeness.
  - The field contains a binary distribution of values: 51.22% as False and 48.78% as True.
  - The data is well-structured and standardized, as there are only two unique values (True/False).
  - 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:



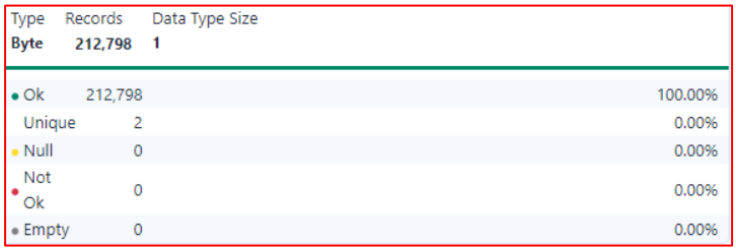
  - Observation:
    - All 212,798 records are valid, with no null, empty, or invalid values, ensuring data completeness.
    - The field contains a single unique value: False, which accounts for 100% of the data.
    - The lack of variation in the data indicates it may not be informative for analysis 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:



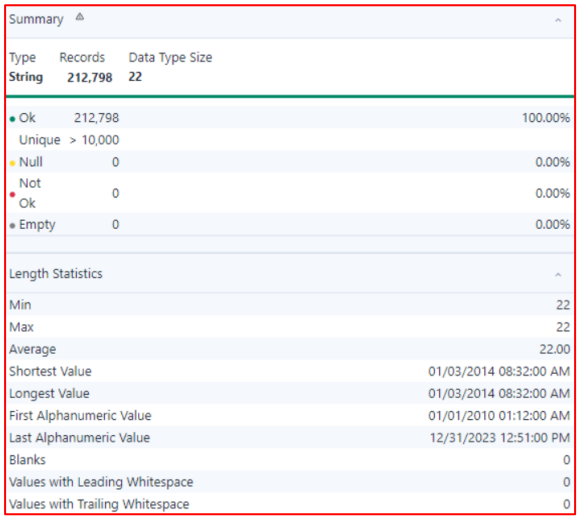
- Observation:
  - All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring completeness and data reliability.
  - The field contains 3 unique values, which indicates low variance and simplicity in categorization.
  - The grouped values range from 0 to 2, with all records 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:



- 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.

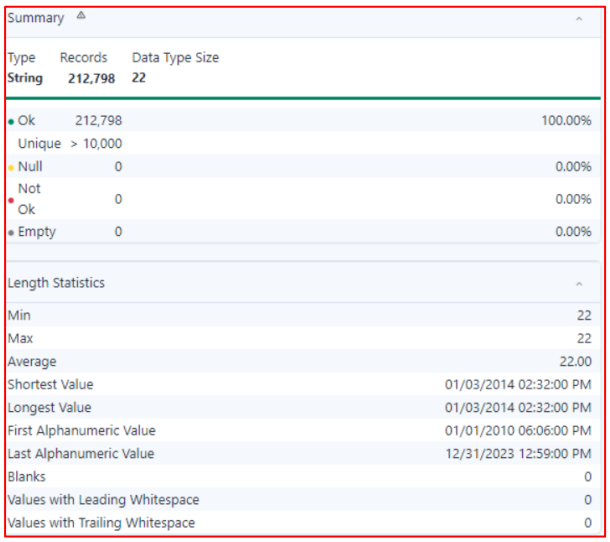
- **CRASH_TIMESTAMP(US/CENTRAL)**
  - Summary:



  - 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

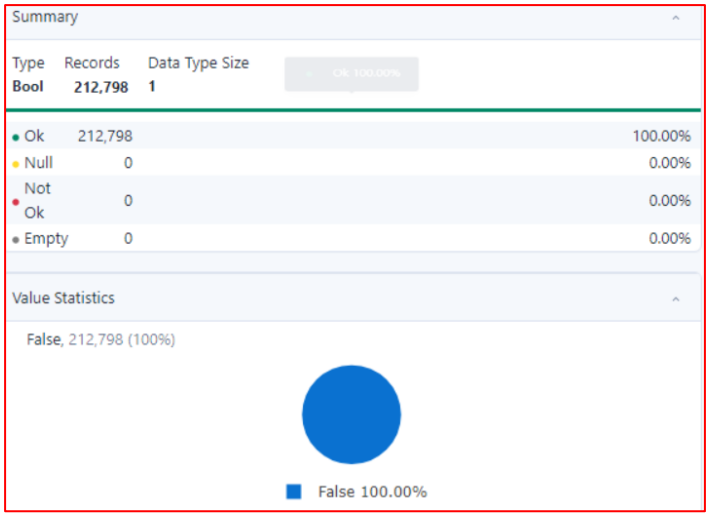
Ok	212,798	100.00%
Unique	> 10,000	
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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

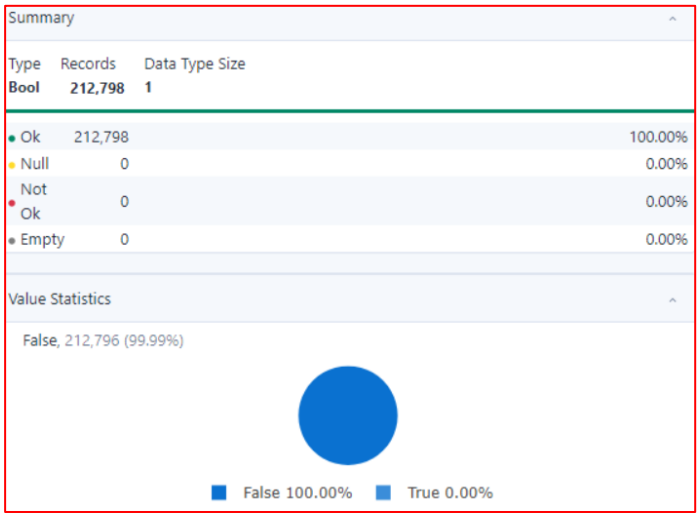
  - Observation:
    - All 212,798 records are valid, ensuring completeness and data reliability.
    - The field contains over 10,000 unique values, indicating high variability and richness in the dataset.
    - The date range spans from 01/01/2010 00:12:00 AM to 12/31/2023 12:51:00 PM, covering an extensive 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.

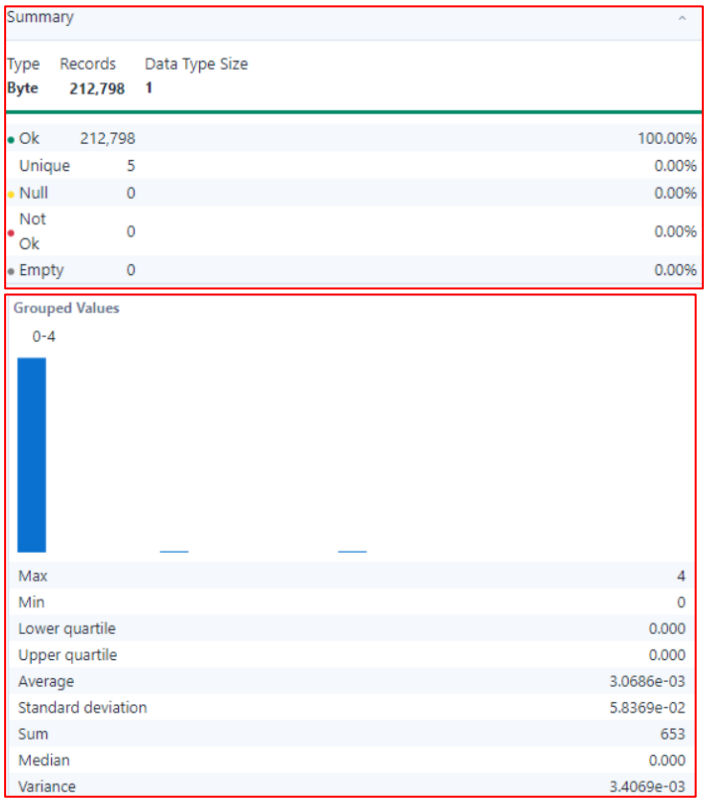
- **IS_TEMPERORY_RECORD**
  - Summary:



- 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.

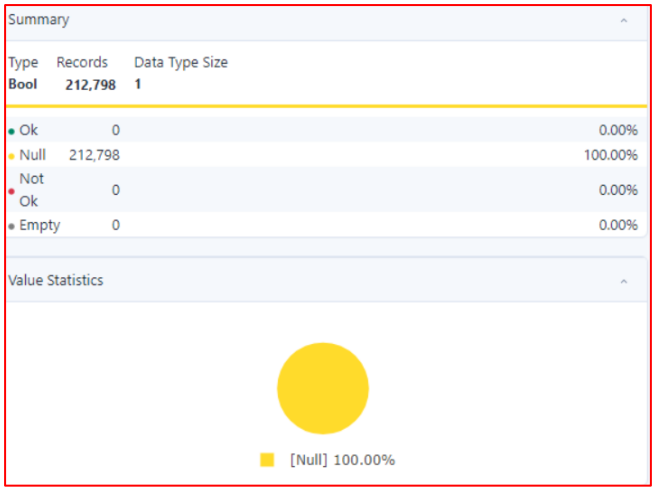
- The high dominance of False suggests limited variability, which may impact the field's significance in data analysis.

- **LAW_ENFORCEMENT_FATALITY_COUNT**
  - Summary:



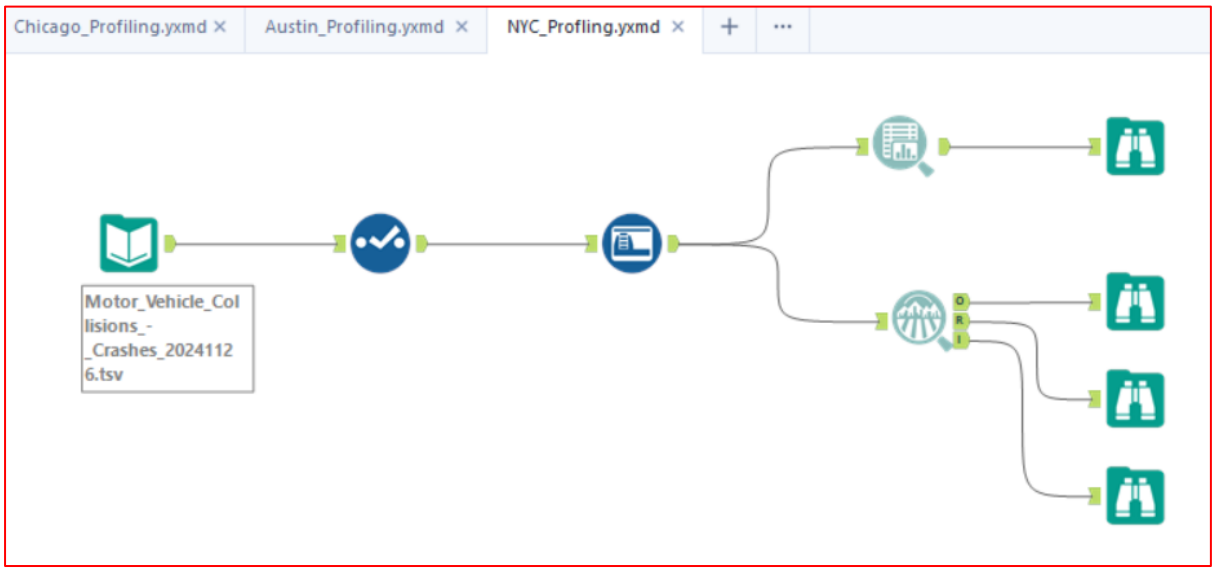
  - Observation:
    - All 212,798 records are complete with no null or empty values.
    - The field contains 5 unique values, indicating low variability.
    - All values follow a uniform format, ensuring data consistency.
    - Data is well-structured and consistent, requiring no additional preprocessing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

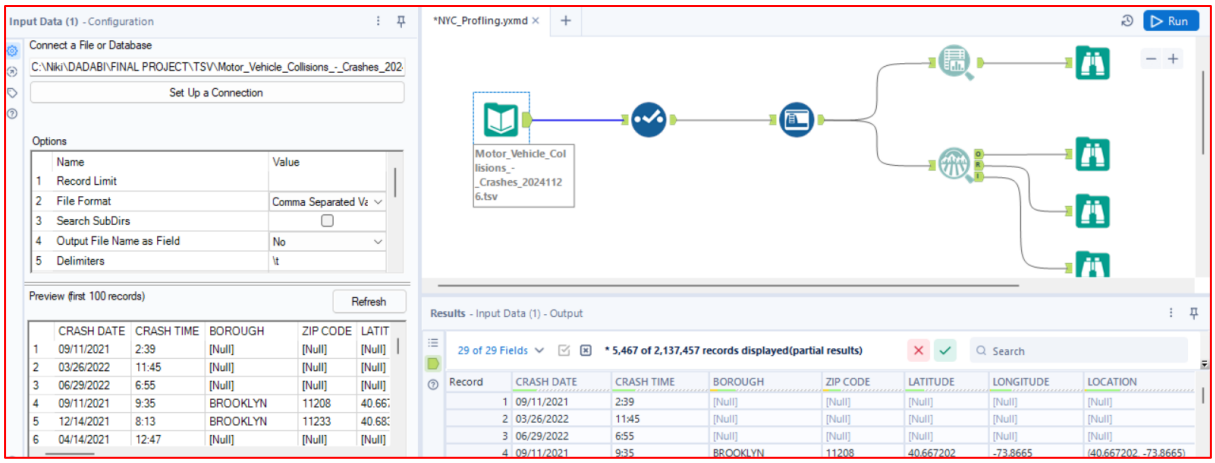
- **REPORTED_STREET_PREFIX**
  - Summary:



- 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**

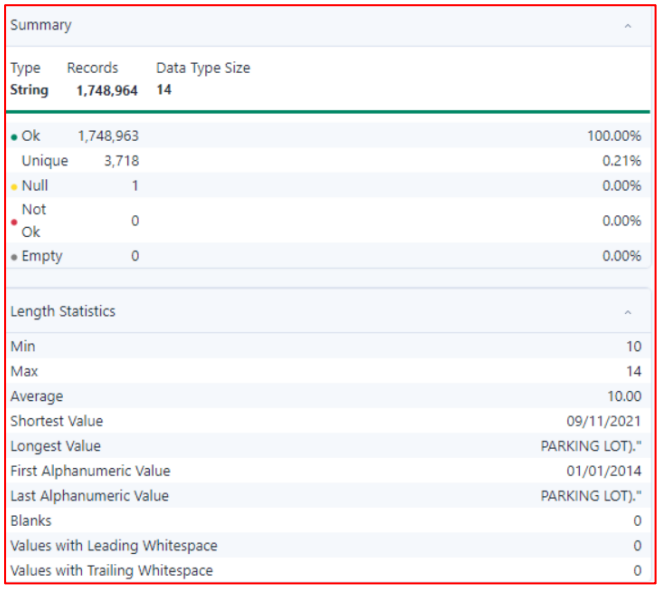
Field	Description	Analysis
CRASH_DATE	Date when the crash occurred.	Validate the format and ensure no null values to maintain temporal accuracy.
CRASH_TIME	Time when the crash occurred.	Ensure consistency with CRASH_DATE and validate the time format.
BOROUGH	The borough where the crash occurred (e.g., 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	Validate against a predefined list of contributing factors (e.g., speeding, weather).

	involved in the crash.	
CONTRIBUTING_FACTOR_VEHICLE_2	Secondary contributing factor for the second vehicle involved, if applicable.	Ensure logical consistency with CONTRIBUTING_FACTOR_VEHICLE_1.
CONTRIBUTING_FACTOR_VEHICLE_3	Contributing factor for the third vehicle involved, if applicable.	Validate data entries and ensure values align with the first and second factors.
CONTRIBUTING_FACTOR_VEHICLE_4	Contributing factor for the fourth vehicle involved, if applicable.	Ensure accuracy and completeness if multiple vehicles are involved.
CONTRIBUTING_FACTOR_VEHICLE_5	Contributing factor for the fifth vehicle involved, if applicable.	Validate logical consistency for entries in multi-vehicle crashes.
COLLISION_ID	Unique identifier for each collision.	Ensure uniqueness and validate against the dataset for data integrity.
VEHICLE_TYPE_CODE_1	Type of the first vehicle involved in the crash.	Standardize vehicle type categories (e.g., sedan, SUV, truck).
VEHICLE_TYPE_CODE_2	Type of the second vehicle involved in the crash, if applicable.	Validate alignment with VEHICLE_TYPE_CODE_1 for logical consistency.
VEHICLE_TYPE_CODE_3	Type of the third vehicle involved in the crash, if applicable.	Ensure accuracy and completeness for multi-vehicle crashes.
VEHICLE_TYPE_CODE_4	Type of the fourth vehicle involved in the crash, if applicable.	Validate data entries for consistency and completeness.
VEHICLE_TYPE_CODE_5	Type of the fifth vehicle involved in the crash, if applicable.	Ensure data quality and logical consistency for crashes involving multiple vehicles.

Data Observation:

- **CRASH_DATE**
  - Summary:

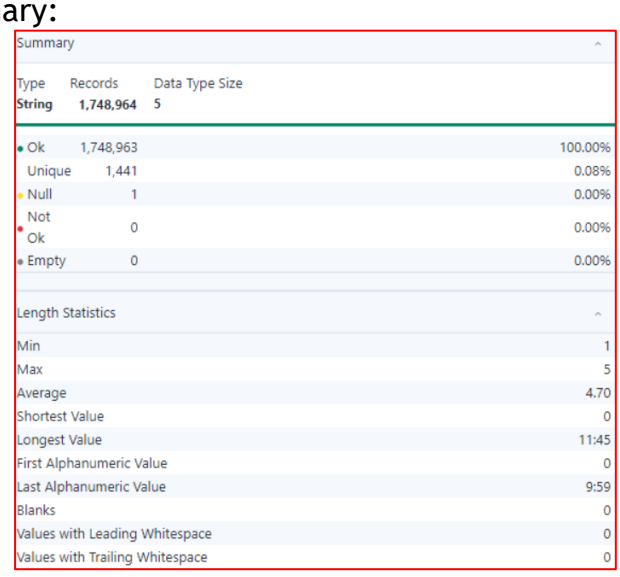


Summary		
Type	Records	Data Type Size
String	1,748,964	14
Ok	1,748,963	100.00%
Unique	3,718	0.21%
Null	1	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

Length Statistics	
Min	10
Max	14
Average	10.00
Shortest Value	09/11/2021
Longest Value	PARKING LOT)."
First Alphanumeric Value	01/01/2014
Last Alphanumeric Value	PARKING LOT)."
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.

- **CRASH_DATE**
  - Summary:



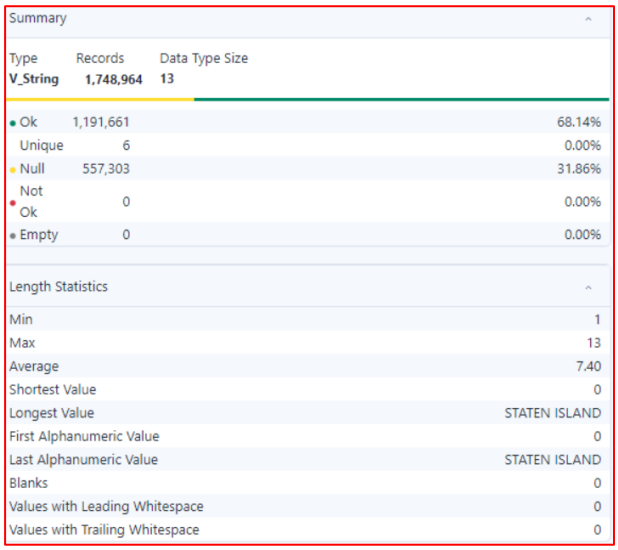
Summary		
Type	Records	Data Type Size
String	1,748,964	5
Ok	1,748,963	100.00%
Unique	1,441	0.08%
Null	1	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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 Ok	0	0.00%
Empty	0	0.00%

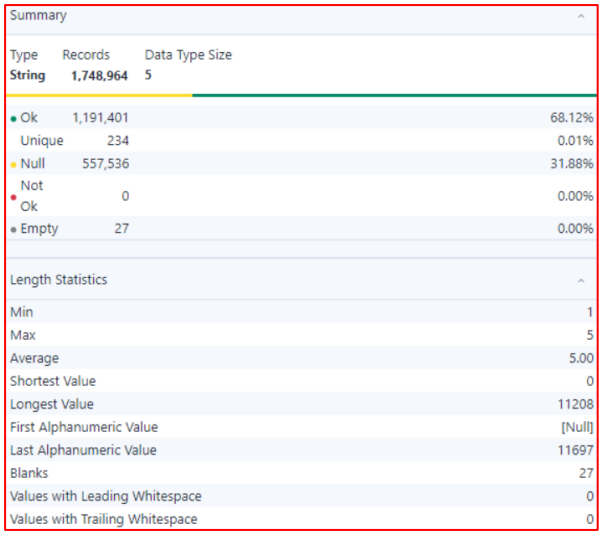
Length Statistics

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**

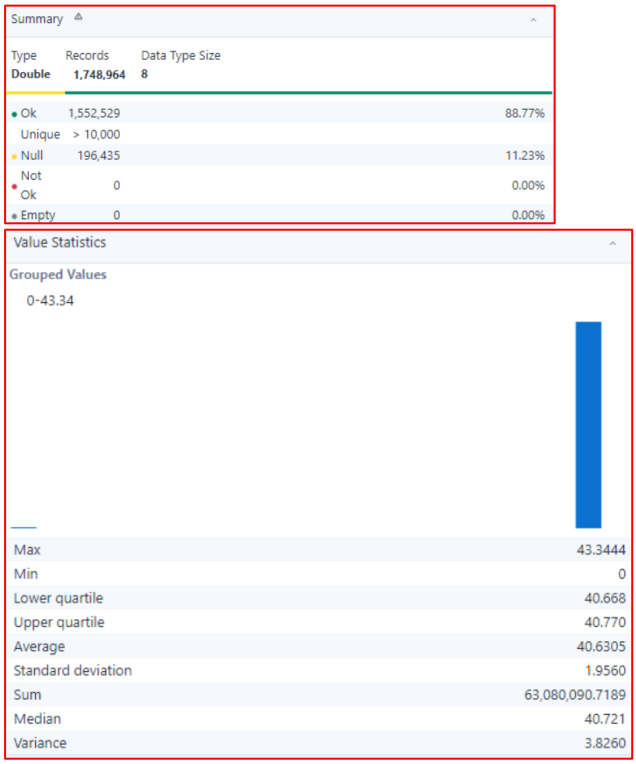
  - Summary:



- 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**

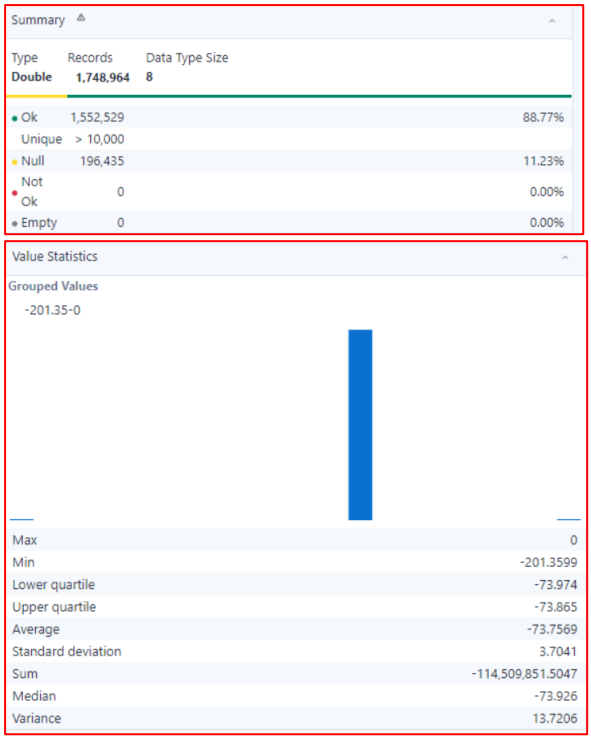
  - Summary:



  - Observation:

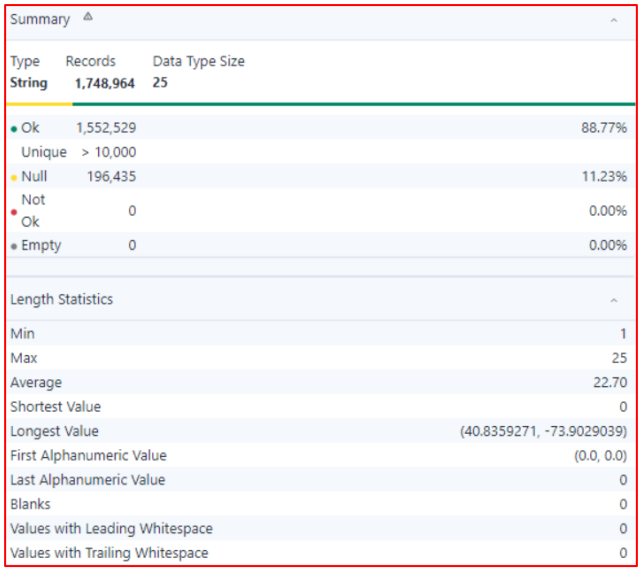
- All 212,798 records are valid, with no null, empty, or invalid values detected, ensuring data reliability
- The field contains over 10,000 unique values, indicating diverse and meaningful entries
- No missing or not-ok records exist, ensuring completeness and readiness for analysis.

- **LONGITUDE**
  - Summary:



  - 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

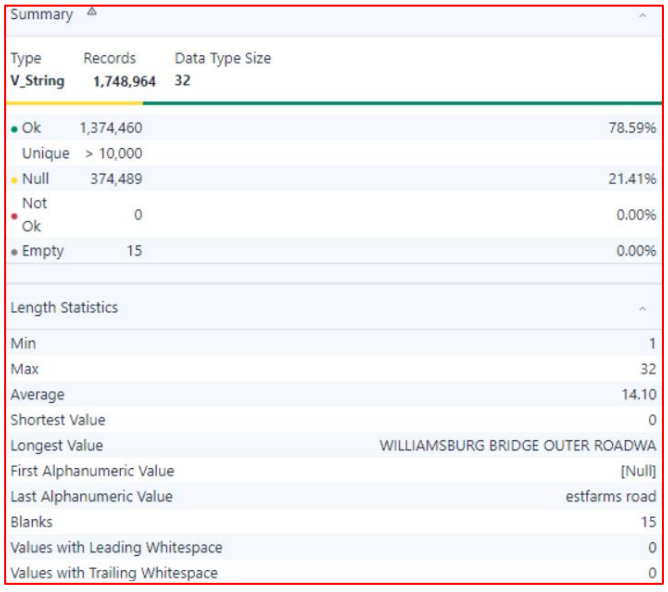
Ok	1,552,529	88.77%
Unique	> 10,000	
Null	196,435	11.23%
Not Ok	0	0.00%
Empty	0	0.00%

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 Ok	0	0.00%
Empty	15	0.00%

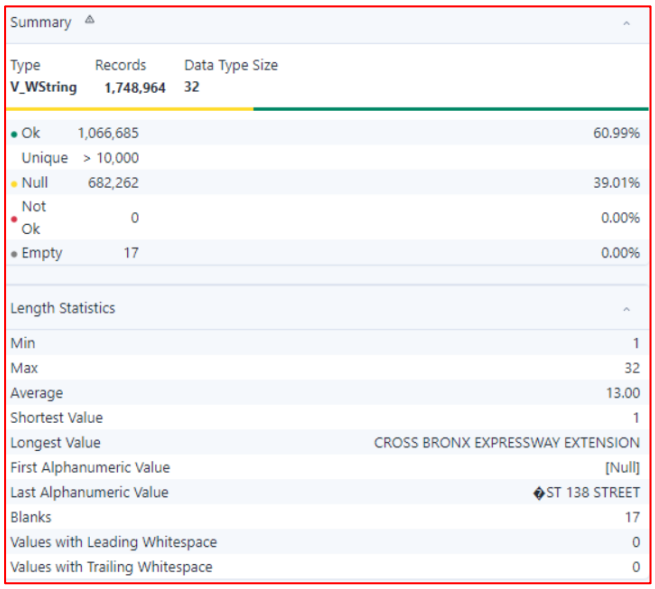
Length Statistics

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:

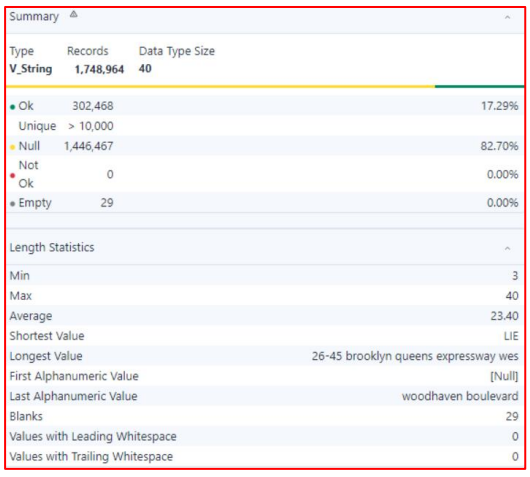


Summary		
Type	Records	Data Type Size
V_WString	1,748,964	32
Ok	1,066,685	60.99%
Unique	> 10,000	
Null	682,262	39.01%
Not Ok	0	0.00%
Empty	17	0.00%

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:

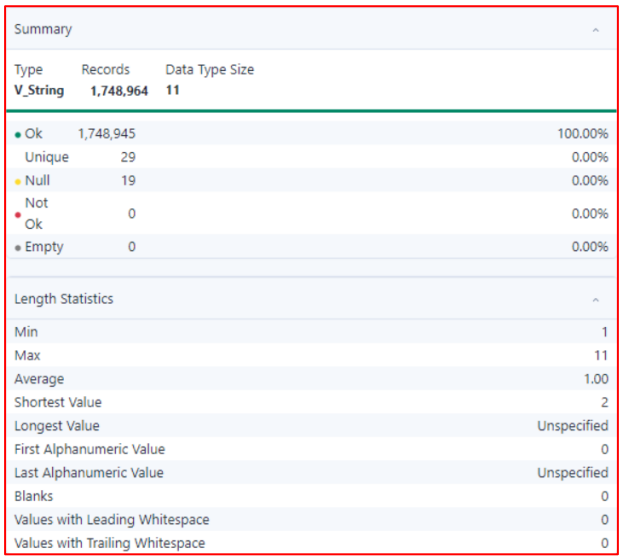


Summary		
Type	Records	Data Type Size
V_String	1,748,964	40
Ok	302,468	17.29%
Unique	> 10,000	
Null	1,446,467	82.70%
Not Ok	0	0.00%
Empty	29	0.00%

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

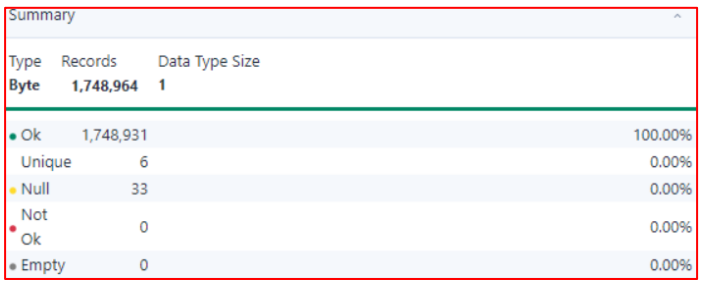
Ok	1,748,945	100.00%
Unique	29	0.00%
Null	19	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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.

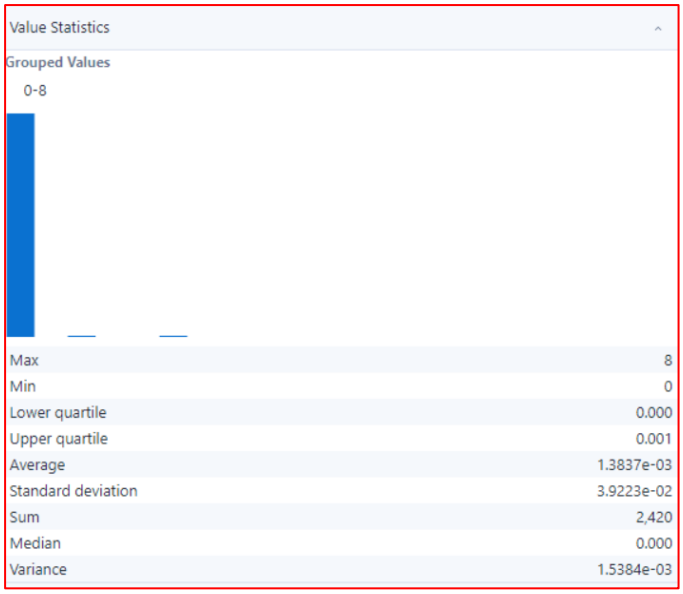
- **NUMBER_OF_PERSONS_KILLED**
  - Summary:



Summary

Type	Records	Data Type Size
Byte	1,748,964	1

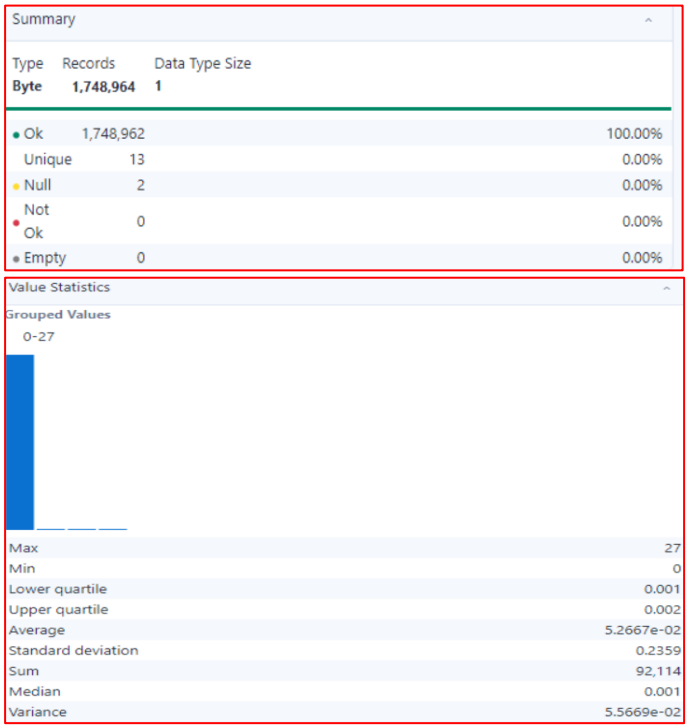
Ok	1,748,931	100.00%
Unique	6	0.00%
Null	33	0.00%
Not Ok	0	0.00%
Empty	0	0.00%



- 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**

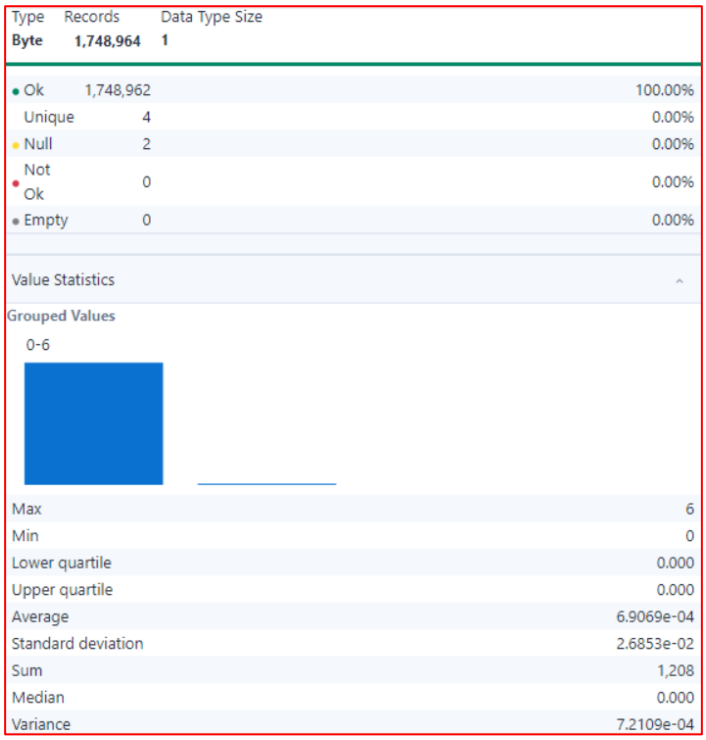
  - Summary:



- 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**

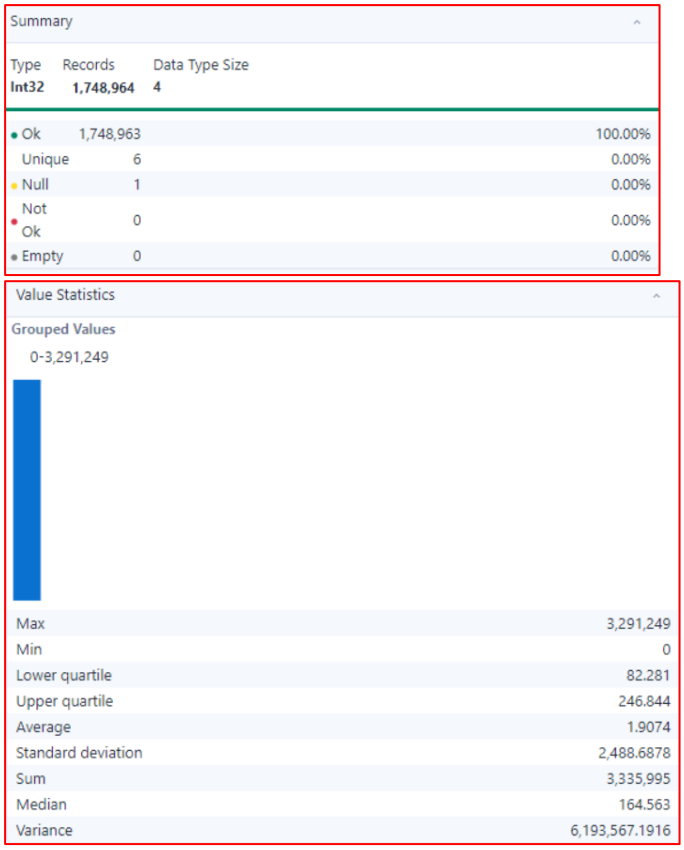
  - Summary:



  - 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**

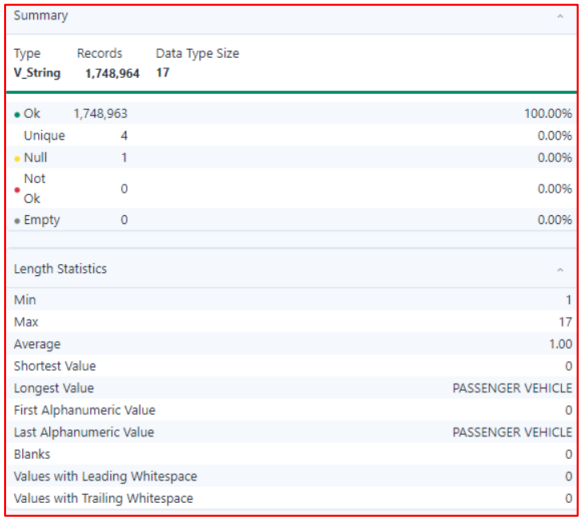
  - Summary:



- 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.

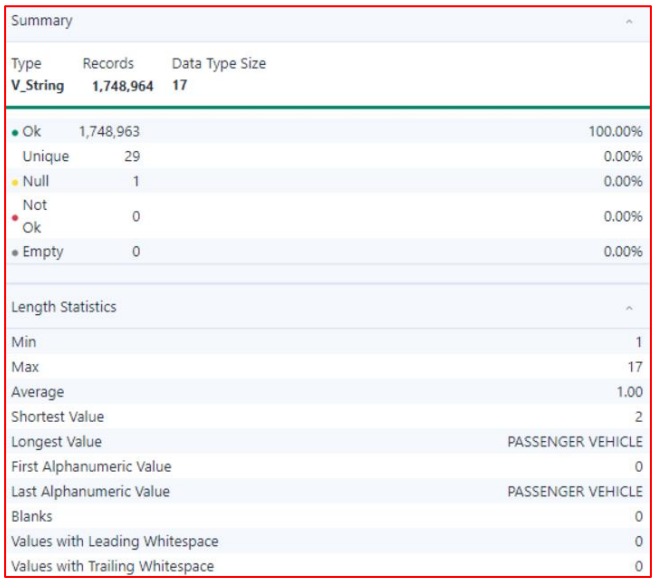
- **NUMBER_OF_CYCLIST_KILLED**

  - Summary:



- 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

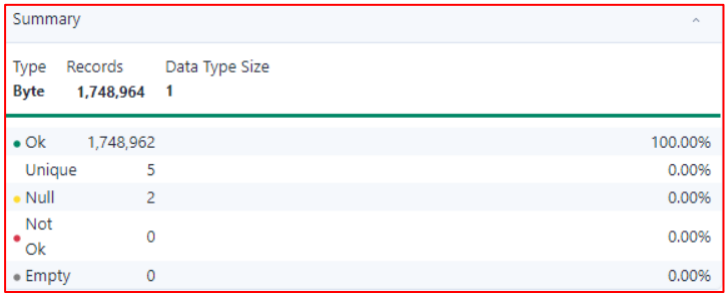
Ok	1,748,963	100.00%
Unique	29	0.00%
Null	1	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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.

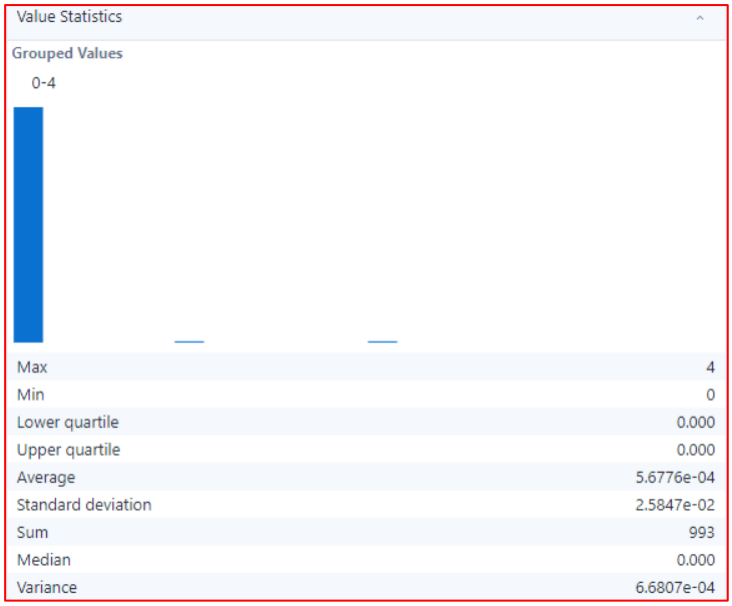
- **NUMBER_OF_MOTORIST_KILLED**
  - Summary:



Summary

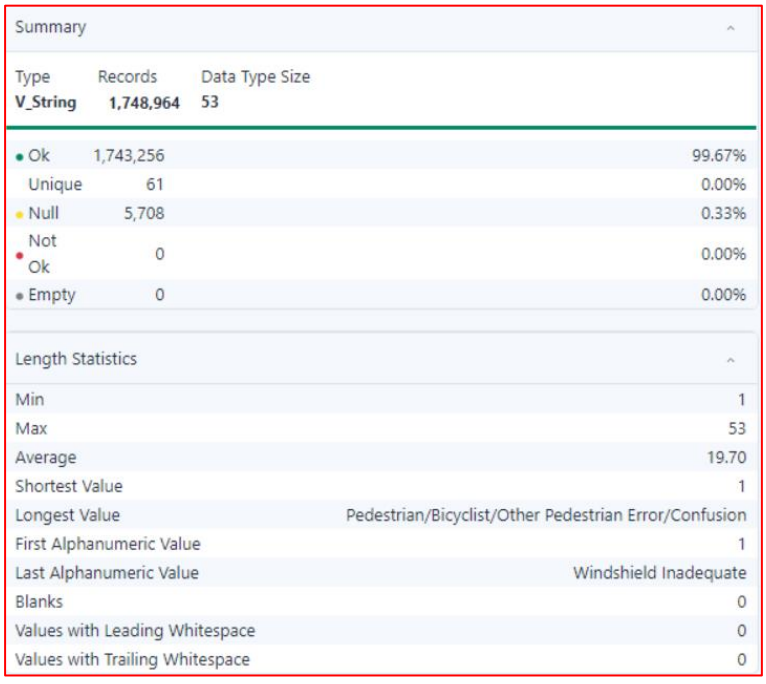
Type	Records	Data Type Size
Byte	1,748,964	1

Ok	1,748,962	100.00%
Unique	5	0.00%
Null	2	0.00%
Not 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 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.

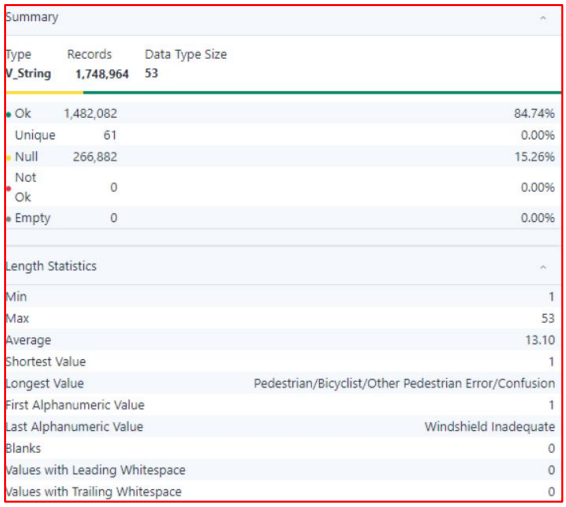
- **CONTRIBUTING_FACTOR_VEHICLE_1**
  - Summary:



  - 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

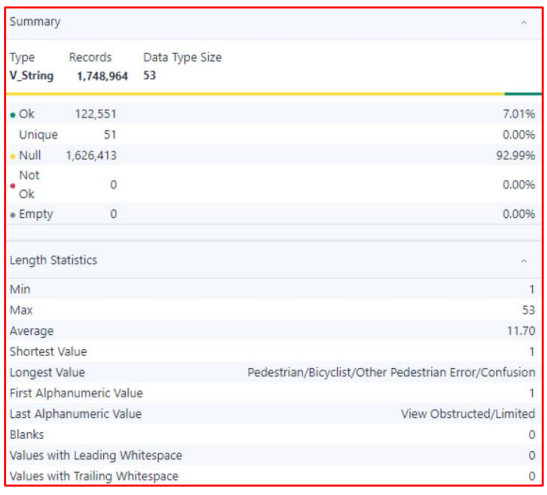
Ok	1,482,082	84.74%
Unique	61	0.00%
Null	266,882	15.26%
Not Ok	0	0.00%
Empty	0	0.00%

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 Ok	0	0.00%
Empty	0	0.00%

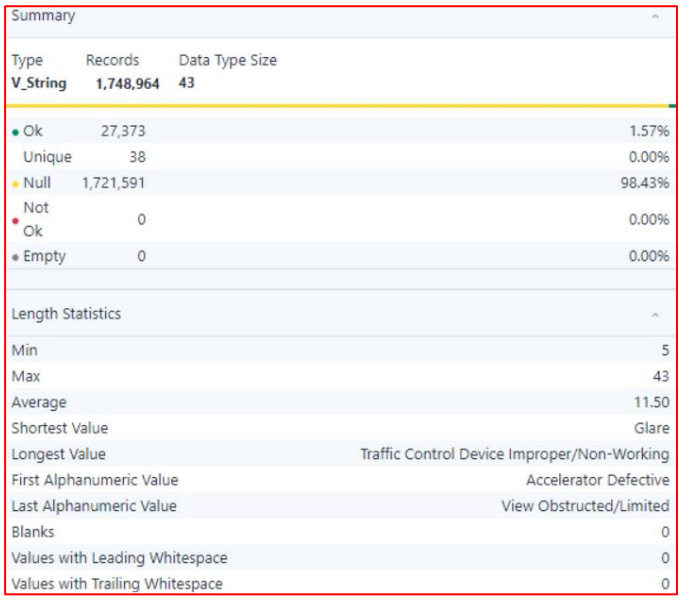
Length Statistics

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:



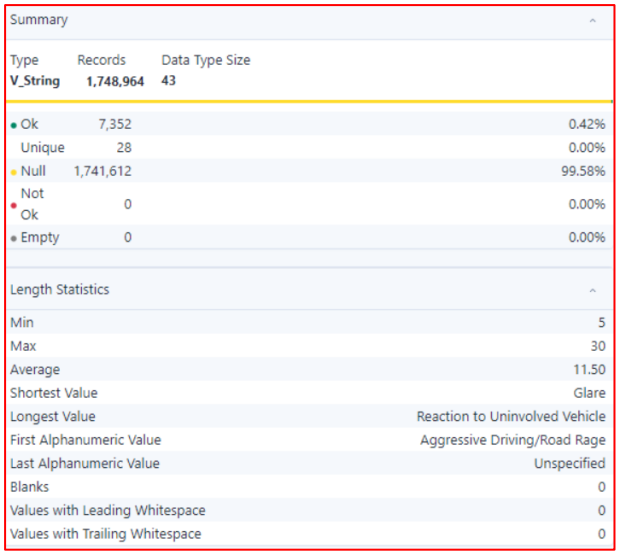
Summary		
Type	Records	Data Type Size
V_String	1,748,964	43
Ok	27,373	1.57%
Unique	38	0.00%
Null	1,721,591	98.43%
Not Ok	0	0.00%
Empty	0	0.00%

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

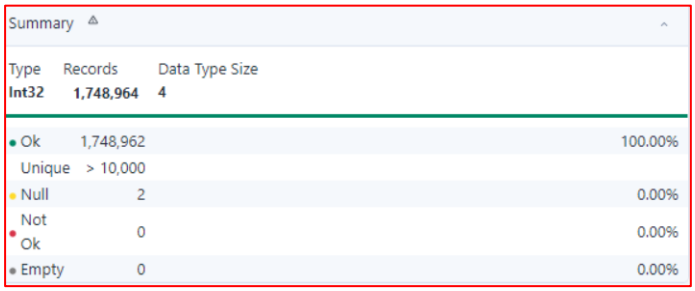
Ok	7,352	0.42%
Unique	28	0.00%
Null	1,741,612	99.58%
Not Ok	0	0.00%
Empty	0	0.00%

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.

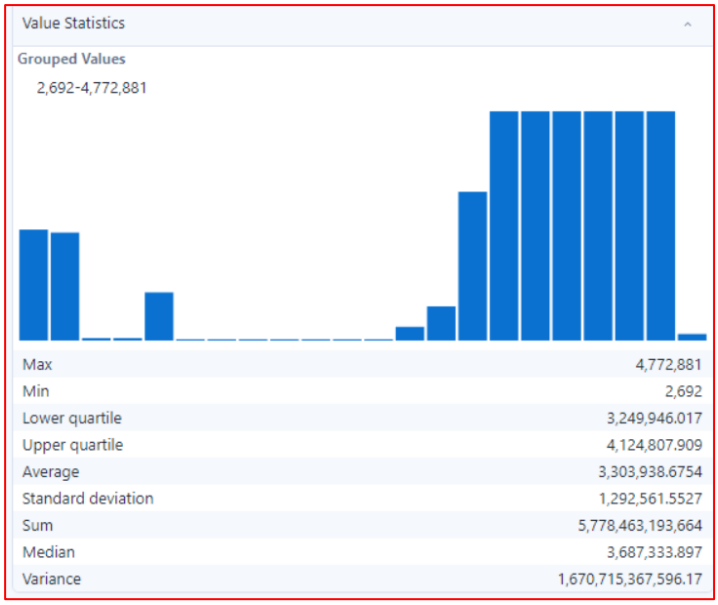
- **COLLISION_ID**
  - Summary:



Summary

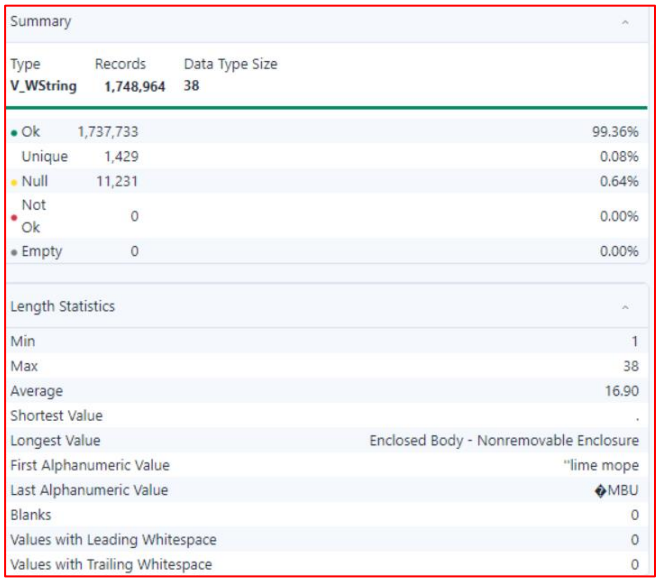
Type	Records	Data Type Size
Int32	1,748,964	4

Ok	1,748,962	100.00%
Unique	> 10,000	
Null	2	0.00%
Not 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 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.

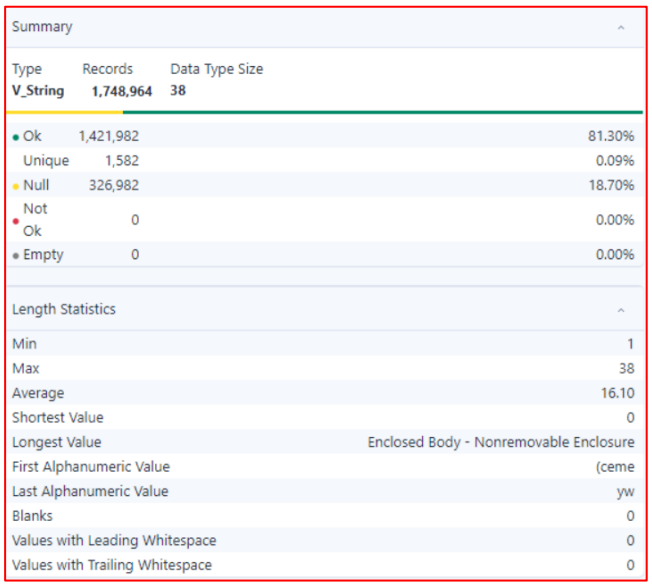
- **VEHICLE_TYPE_CODE_1**
  - Summary:



  - 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

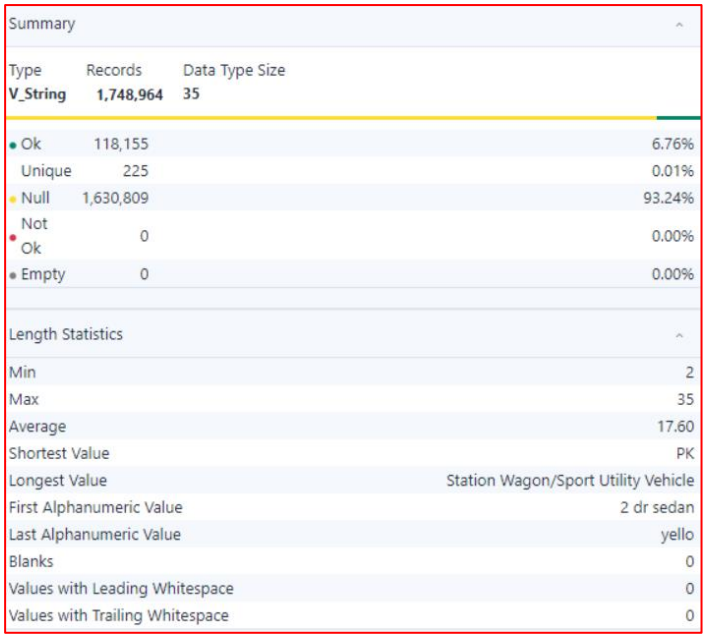
Ok	1,421,982	81.30%
Unique	1,582	0.09%
Null	326,982	18.70%
Not Ok	0	0.00%
Empty	0	0.00%

Length Statistics

Min	1
Max	38
Average	16.10
Shortest Value	0
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

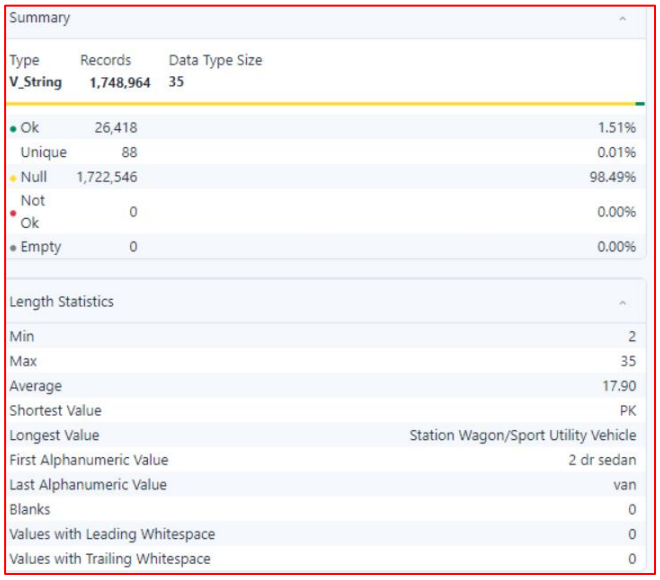
Ok	118,155	6.76%
Unique	225	0.01%
Null	1,630,809	93.24%
Not Ok	0	0.00%
Empty	0	0.00%

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%

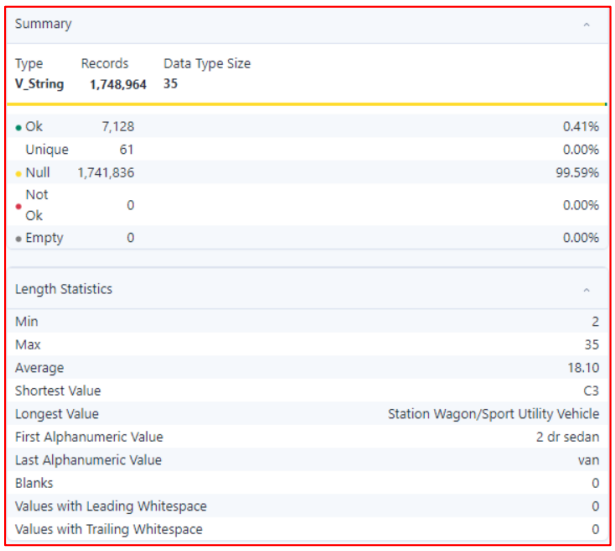
Length Statistics

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:

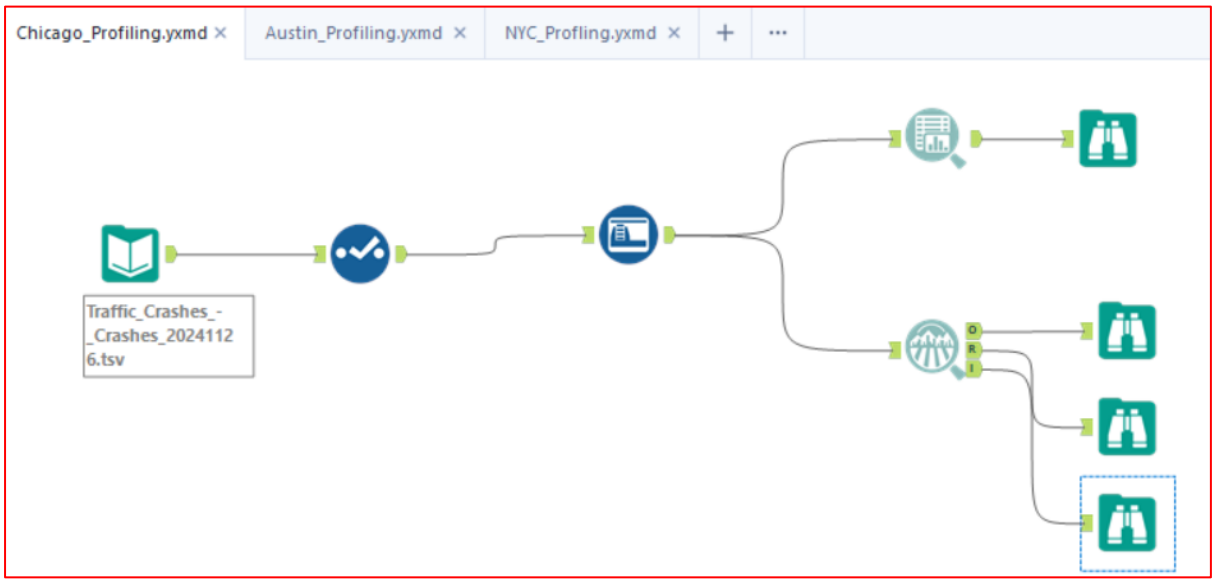


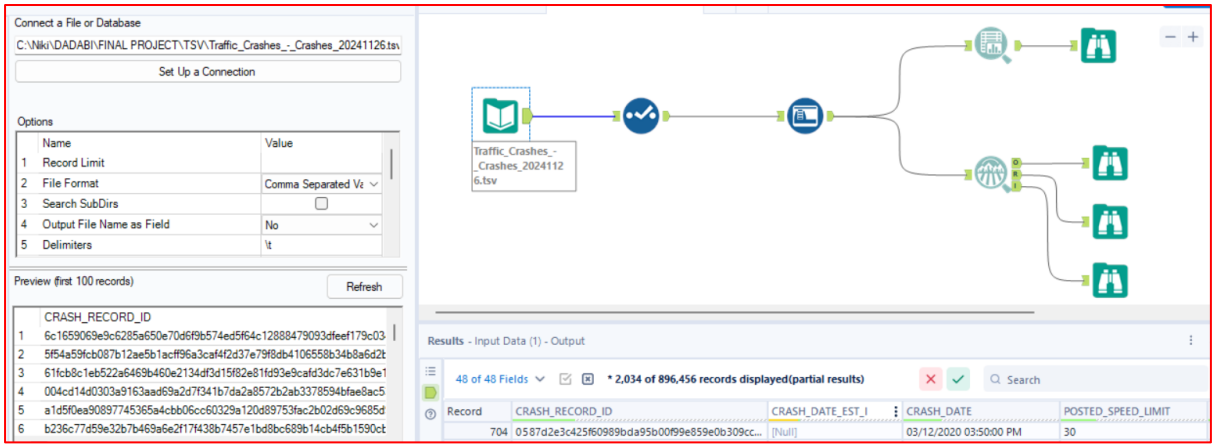
Summary		
Type	Records	Data Type Size
V_String	1,748,964	35
Ok	7,128	0.41%
Unique	61	0.00%
Null	1,741,836	99.59%
Not Ok	0	0.00%
Empty	0	0.00%

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).

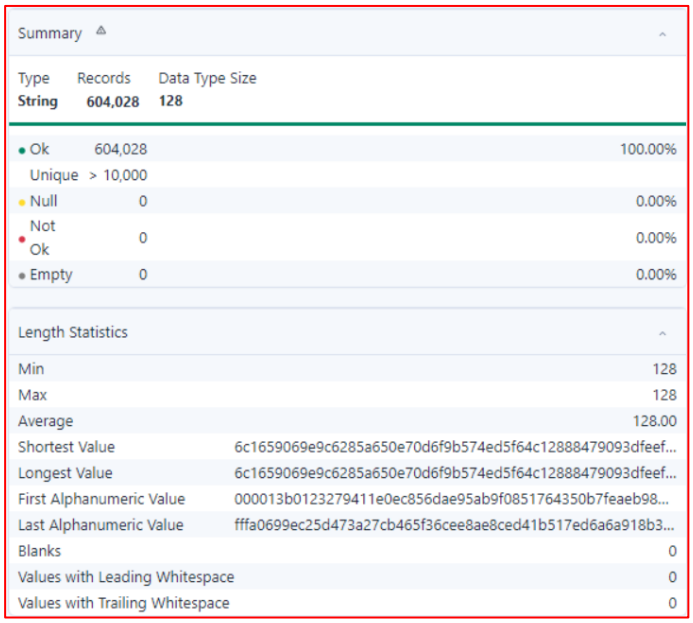
LANE_CNT	Number of lanes in the roadway at the crash location.	Identify anomalies such as unusually high or missing lane counts.
ALIGNMENT	Roadway alignment (e.g., straight, curve) at the crash location.	Cross-validate with crash type for logical consistency (e.g., sharp curves and single-vehicle crashes).
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:



Summary

Type	Records	Data Type Size
String	604,028	128

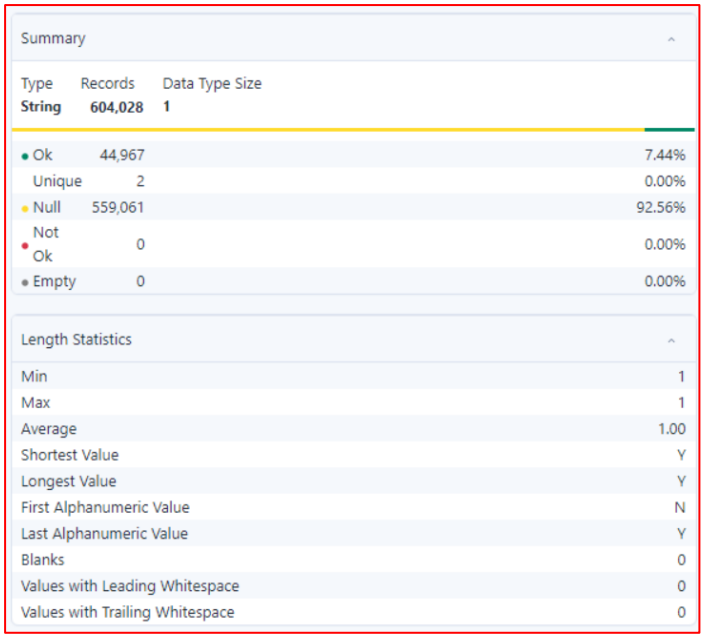
Ok	604,028	100.00%
Unique	> 10,000	
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

Length Statistics

Min	128
Max	128
Average	128.00
Shortest Value	6c1659069e9c6285a650e70d6f9b574ed5f64c12888479093dfeef...
Longest Value	6c1659069e9c6285a650e70d6f9b574ed5f64c12888479093dfeef...
First Alphanumeric Value	000013b0123279411e0ec856dae95ab9f0851764350b7feaeb98...
Last Alphanumeric Value	fffa0699ec25d473a27cb465f36cee8ae8ced41b517ed6a6a918b3...
Blanks	0
Values with Leading Whitespace	0
Values with Trailing Whitespace	0

  - 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

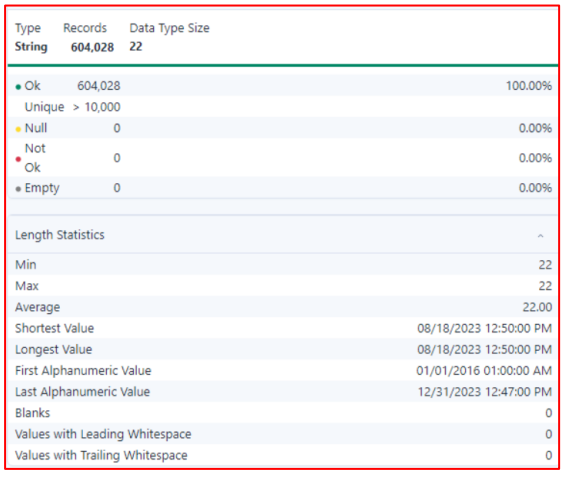
Ok	44,967	7.44%
Unique	2	0.00%
Null	559,061	92.56%
Not Ok	0	0.00%
Empty	0	0.00%

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

Ok	604,028	100.00%
Unique	> 10,000	
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

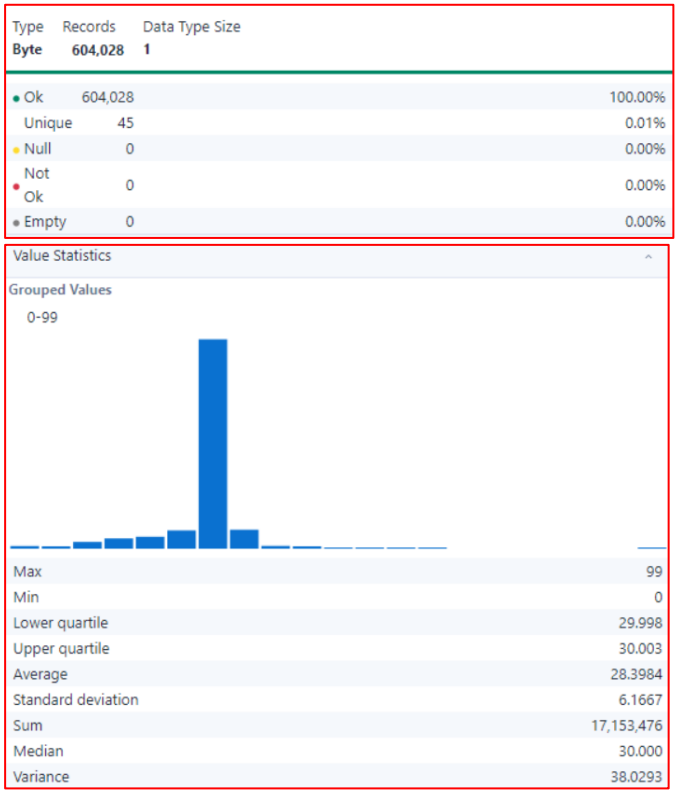
Length Statistics

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:

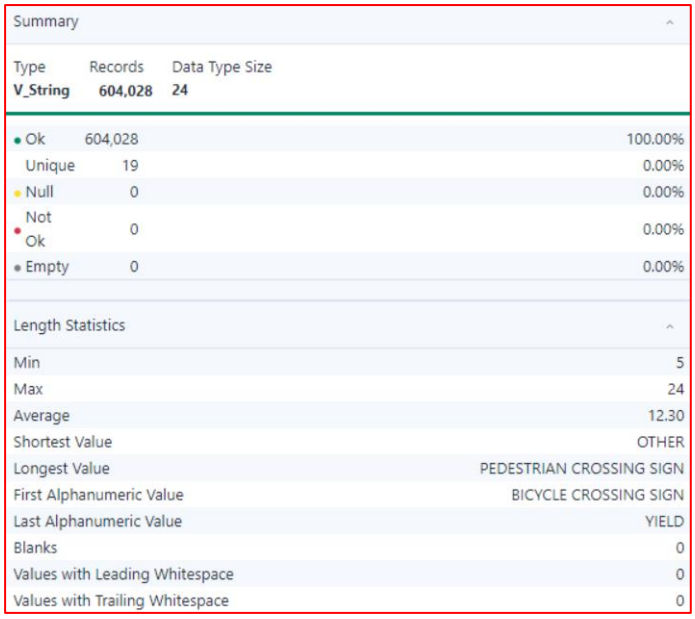
- All 604,028 records (100%) are valid, with no null, empty, or invalid values detected, ensuring complete data reliability.
- The field contains over 10,000 unique values, indicating a high level of diversity and richness in the dataset.
- All values have a consistent length of 22 characters, with no 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:



  - 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

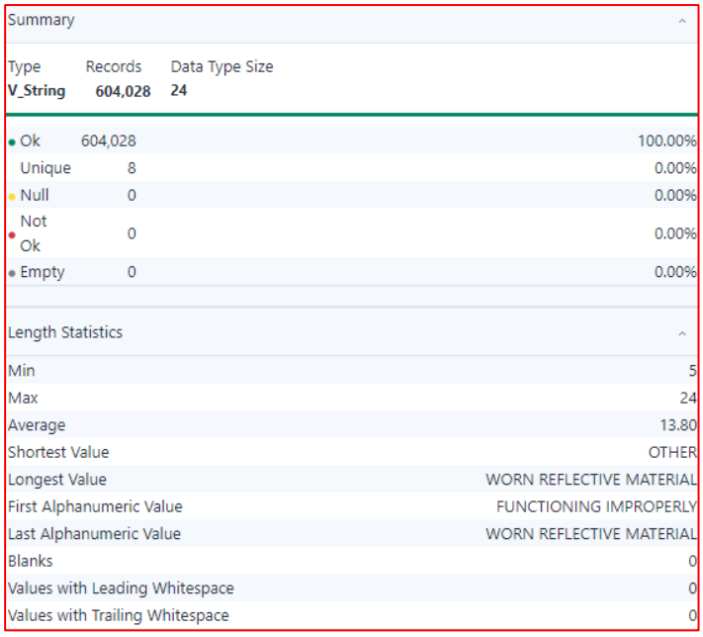
Ok	604,028	100.00%
Unique	19	0.00%
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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

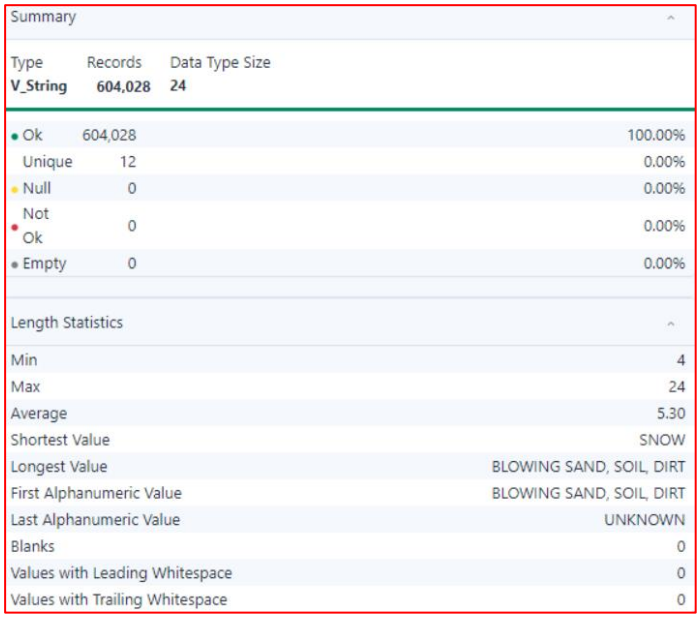
Ok	604,028	100.00%
Unique	8	0.00%
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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 Ok	0	0.00%
Empty	0	0.00%

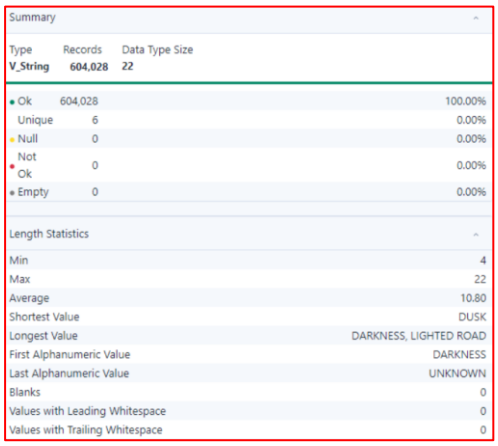
Length Statistics

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**

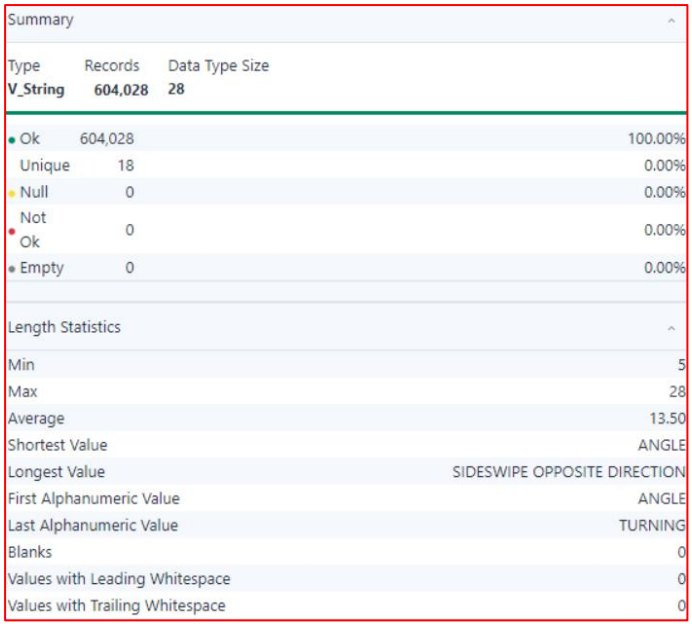
  - Summary:



  - 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

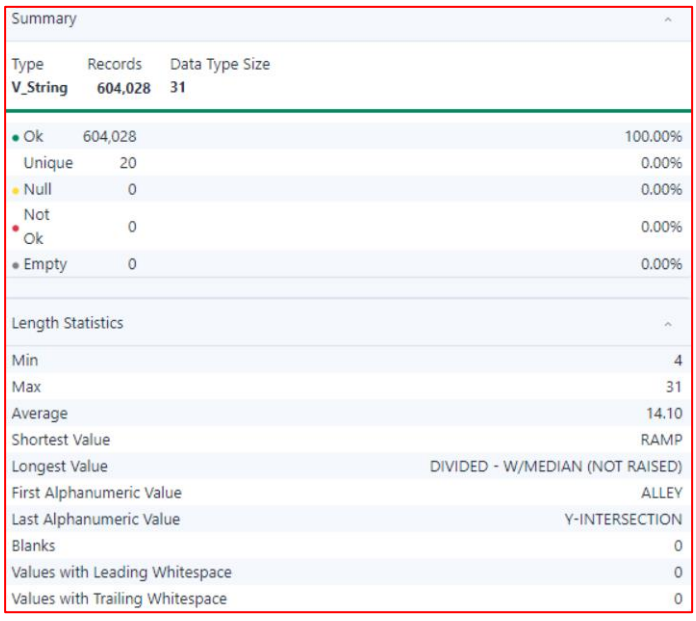
Ok	604,028	100.00%
Unique	18	0.00%
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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

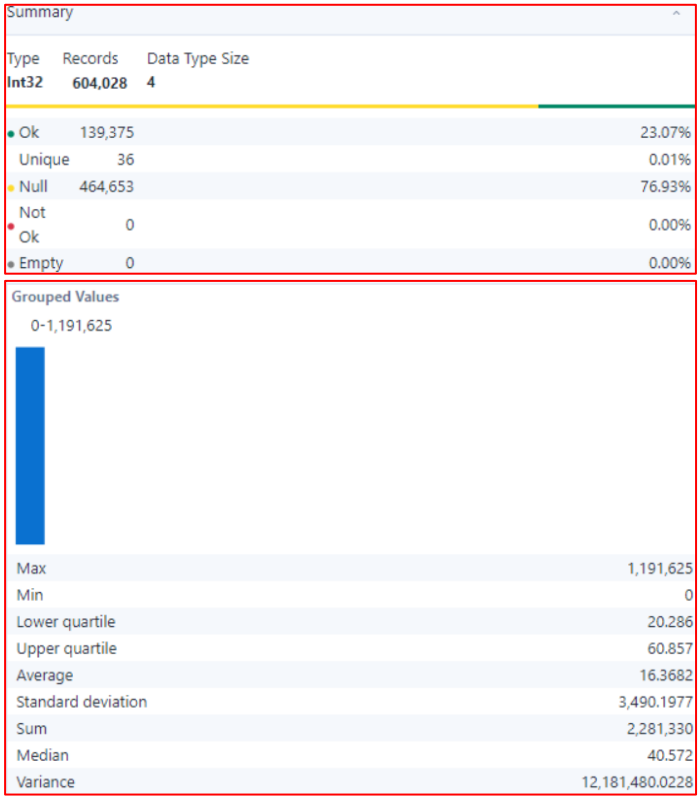
Ok	604,028	100.00%
Unique	20	0.00%
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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:



  - 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%

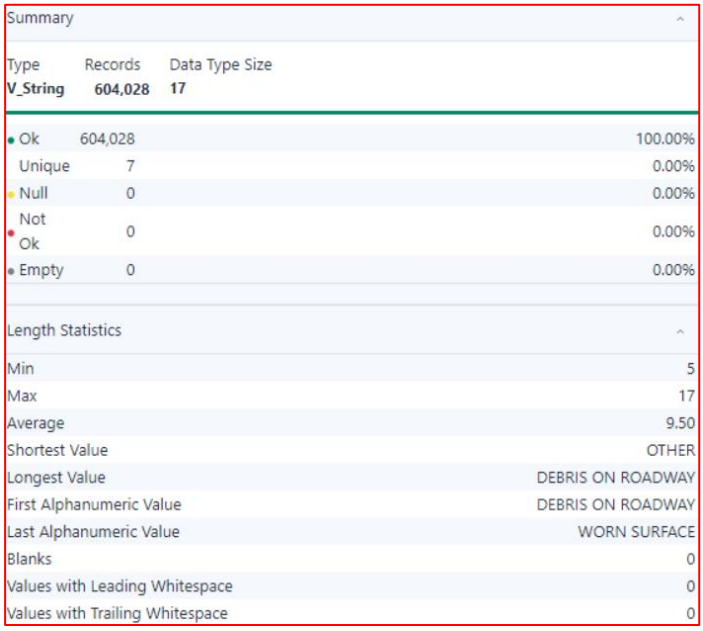
Length Statistics

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%

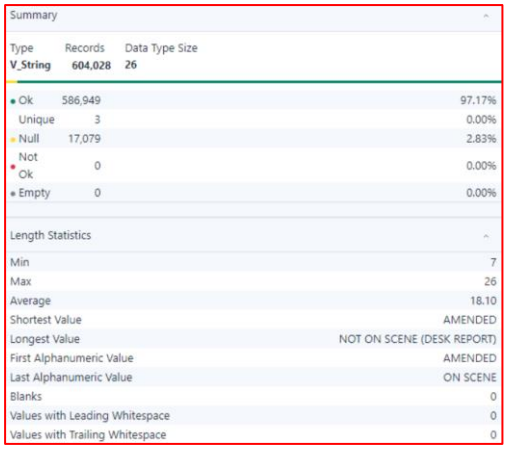
Length Statistics

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**

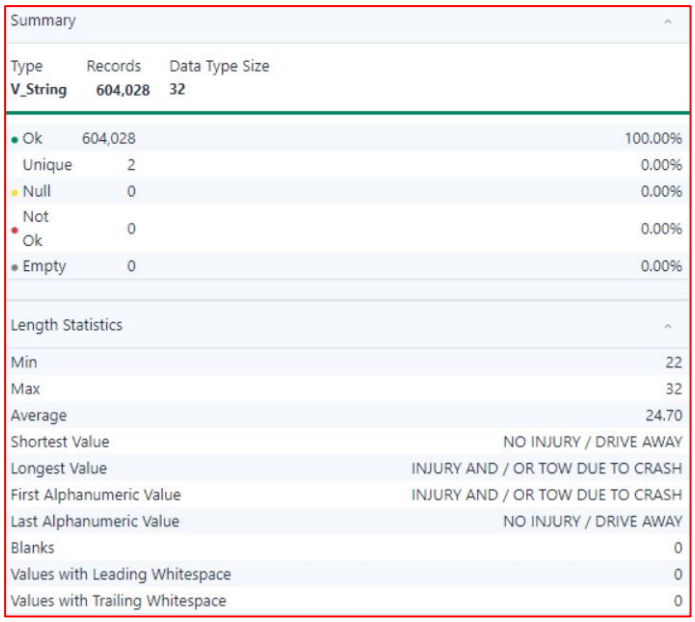
  - Summary:



- 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:

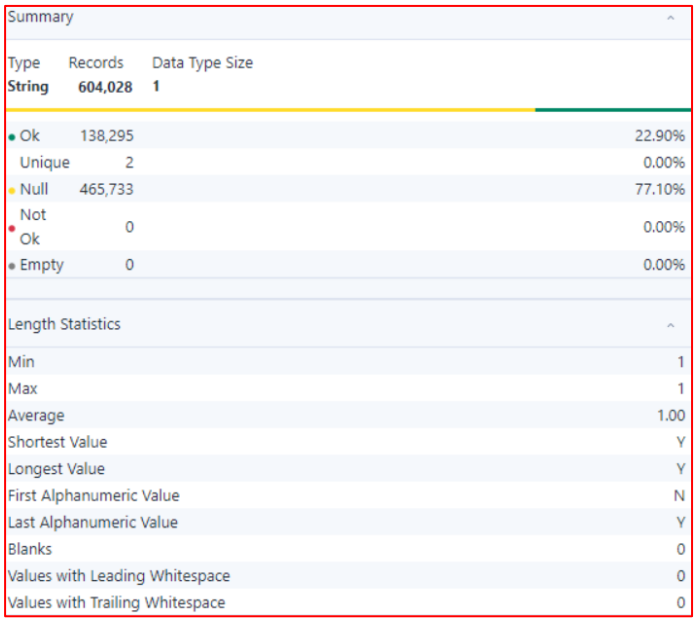


  - Observation:
    - All 604,028 records (100%) are valid, with no null, empty, or invalid values detected.
    - The field contains 2 unique values, indicating very low variability.

- 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 Ok	0	0.00%
Empty	0	0.00%

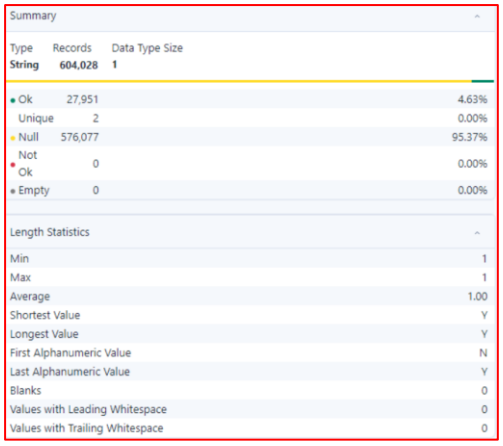
Length Statistics

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.

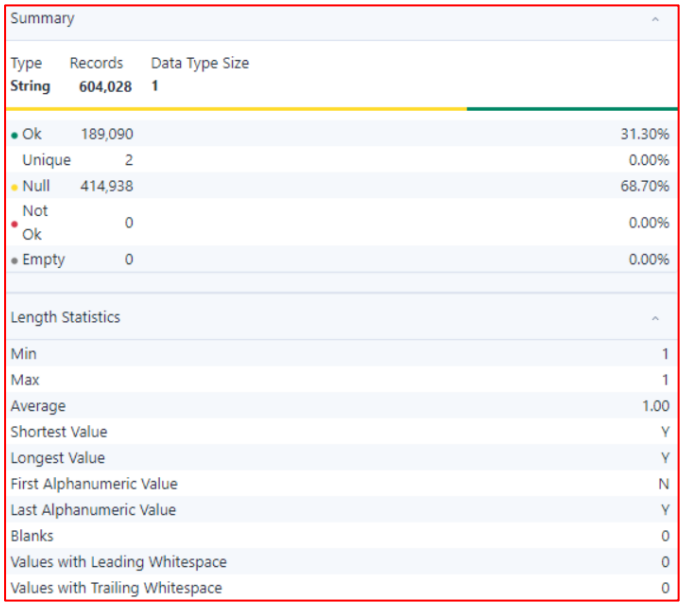
- **NOT_RIGHT_OF_WAY_I**

  - Summary:



- 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.

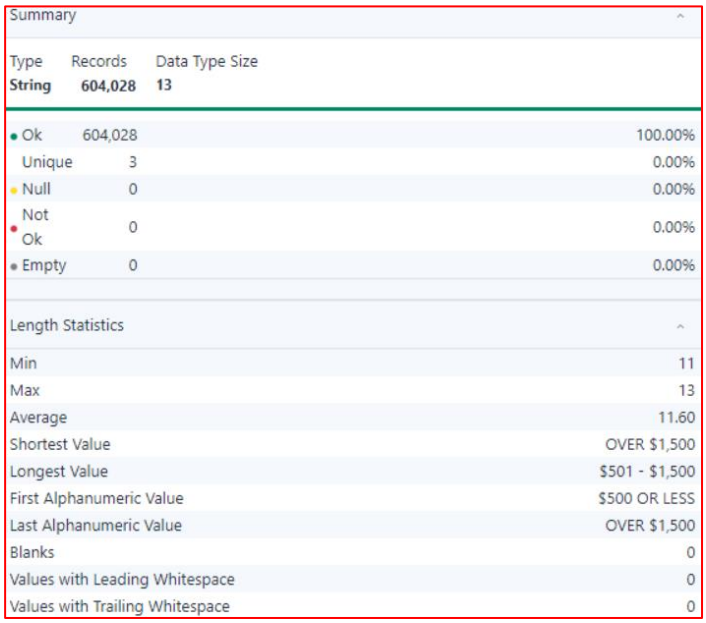
- **HIT_AND_RUN_I**
  - Summary:



  - 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

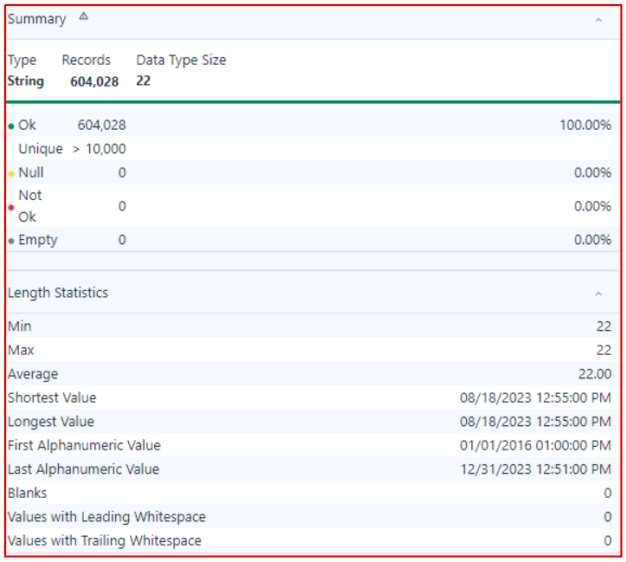
Ok	604,028	100.00%
Unique	3	0.00%
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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

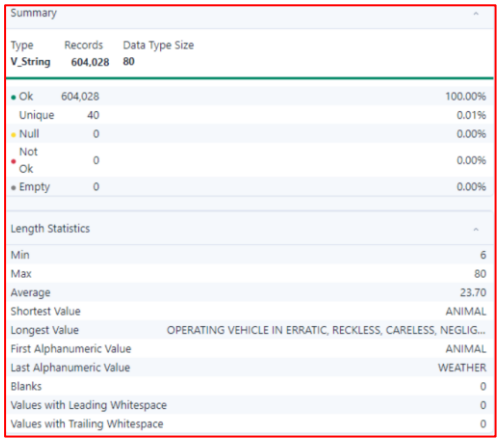
  - Observation:
    - The dataset contains 604,028 records, all valid without any null, blank, or empty values, ensuring data completeness.
    - The field contains 3 unique values: $500 OR LESS, $501 - $1,500, and OVER $1,500, representing categorical ranges.
    - The shortest value is OVER $1,500, 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:



- 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.

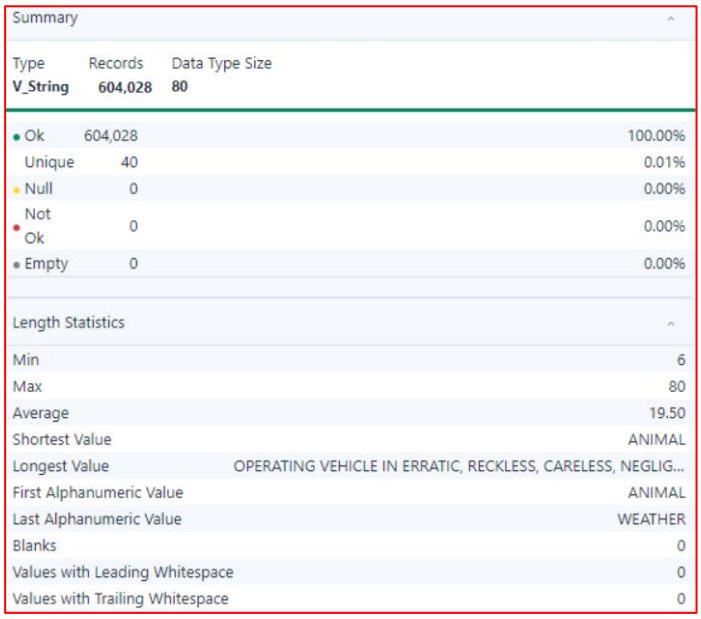
- **PRIM_CONTRIBUTORY_CAUSE**
  - Summary:



  - 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

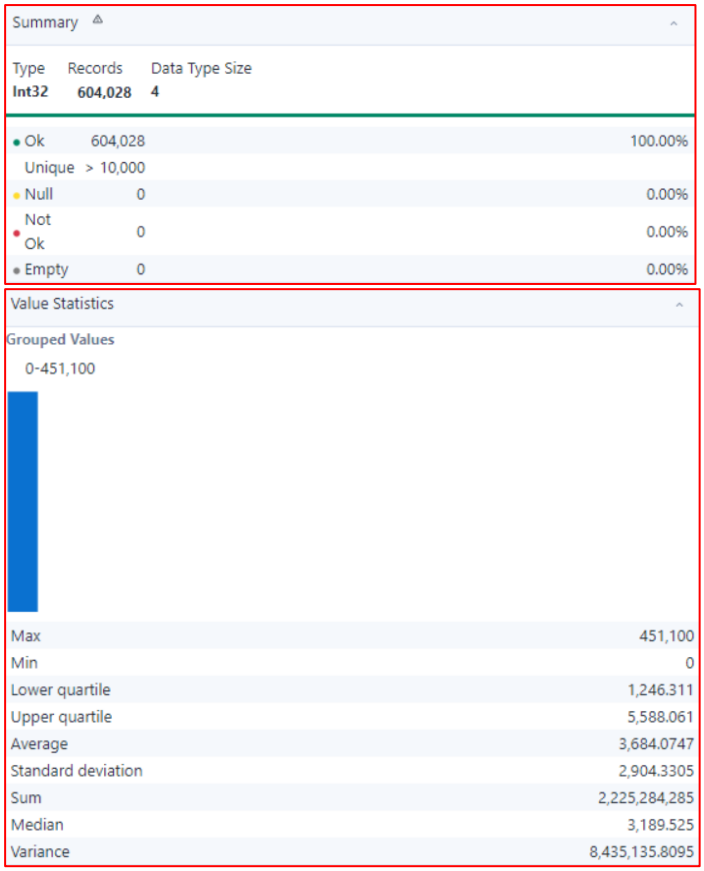
Ok	604,028	100.00%
Unique	40	0.01%
Null	0	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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

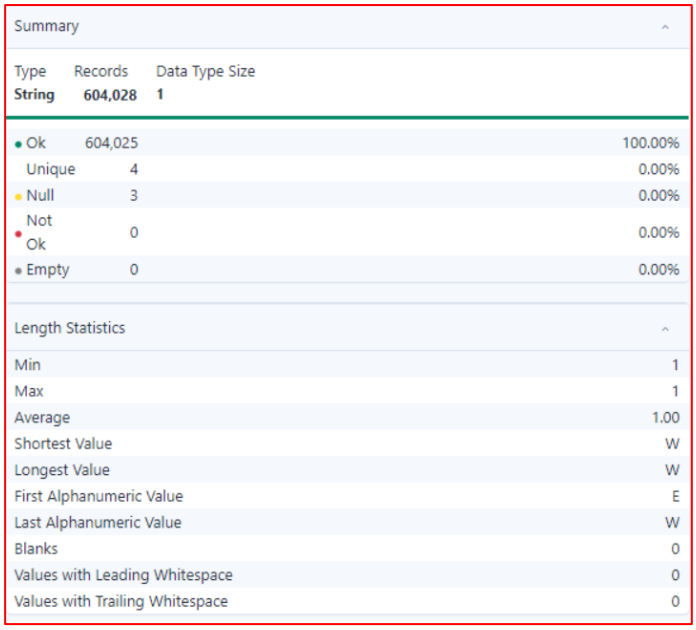
  - Observation:
    - The dataset consists of 604,028 records, all valid with no null, blank, or empty values, ensuring data integrity.
    - There are 40 unique values, indicating a variety of categorical options.
    - The shortest value is "ANIMAL," while the longest is "OPERATING VEHICLE IN ERRATIC, 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:



- 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

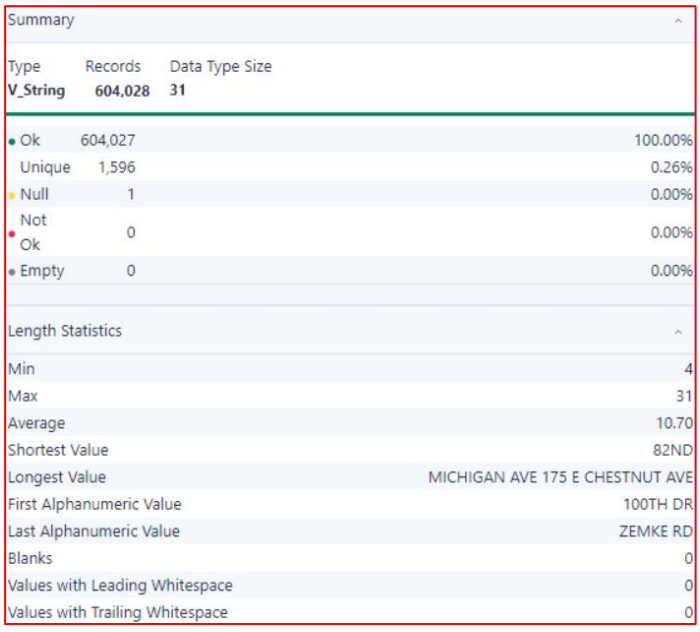
Ok	604,025	100.00%
Unique	4	0.00%
Null	3	0.00%
Not Ok	0	0.00%
Empty	0	0.00%

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%

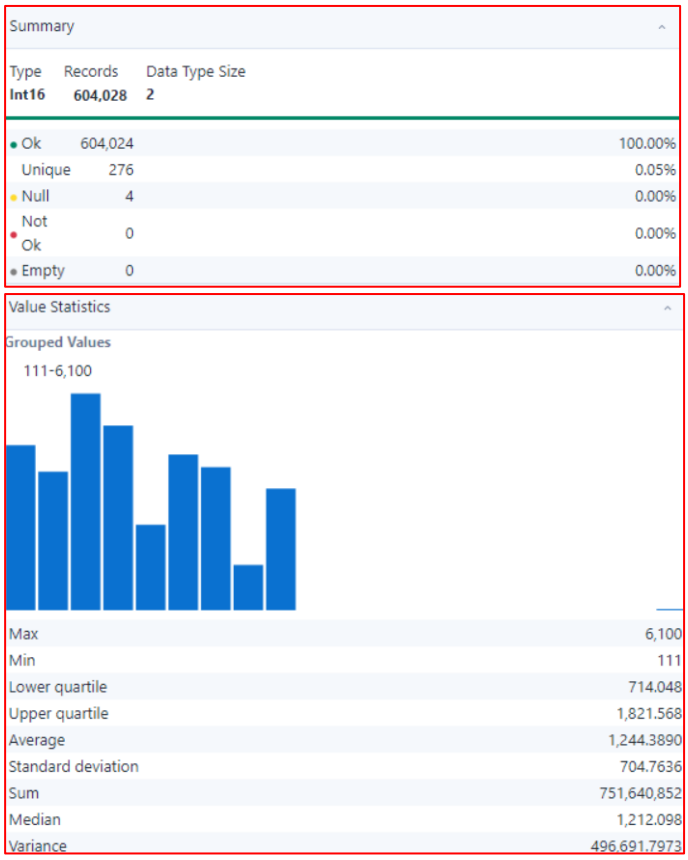
Length Statistics

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**

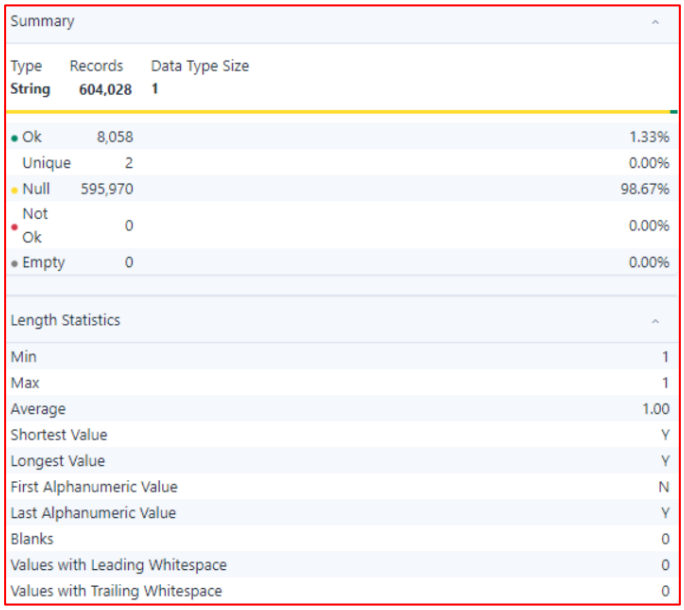
  - Summary:



  - 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:



Summary

Type	Records	Data Type Size
String	**604,028**	**1**

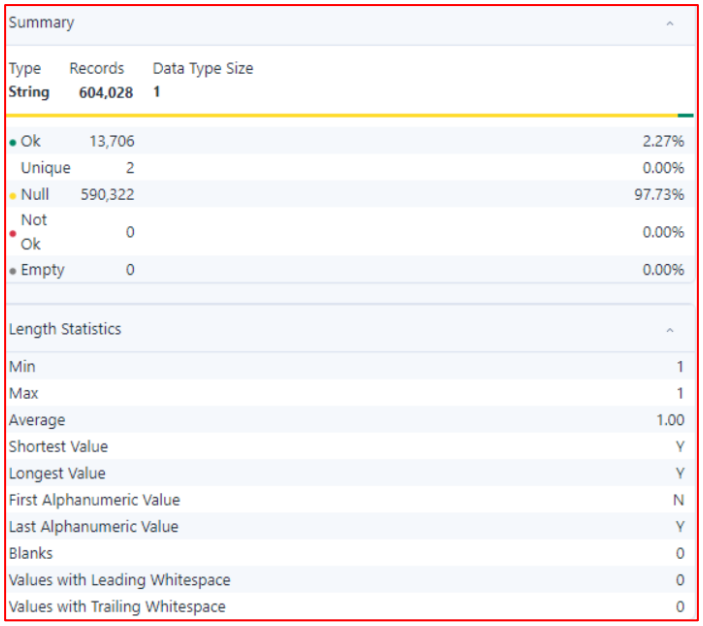
Ok	8,058	1.33%
Unique	2	0.00%
Null	595,970	98.67%
Not Ok	0	0.00%
Empty	0	0.00%

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

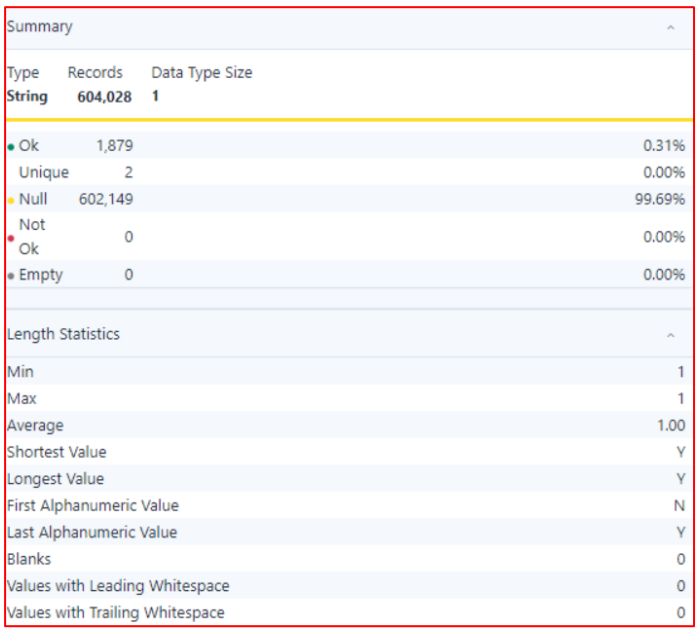
Ok	13,706	2.27%
Unique	2	0.00%
Null	590,322	97.73%
Not Ok	0	0.00%
Empty	0	0.00%

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%

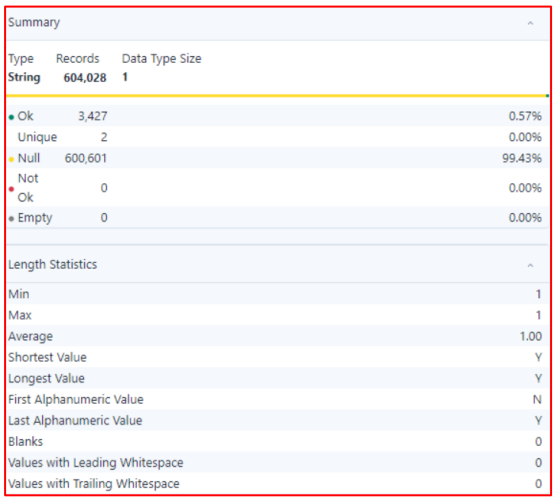
Length Statistics

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%

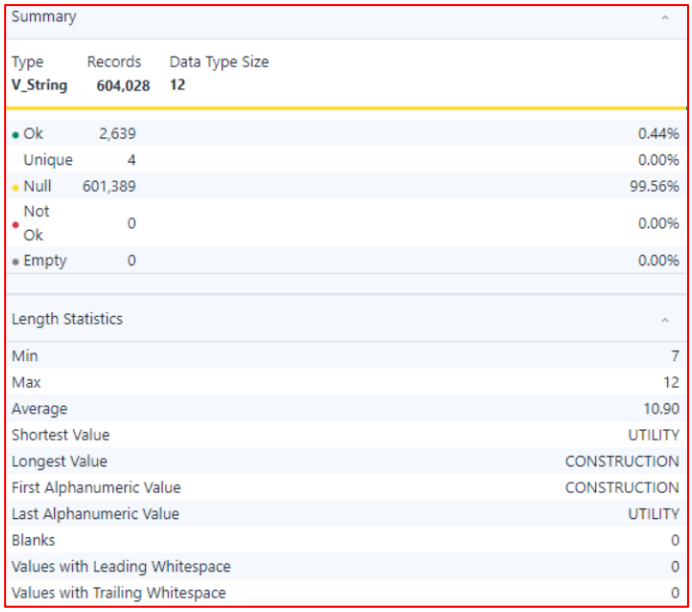
Length Statistics

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

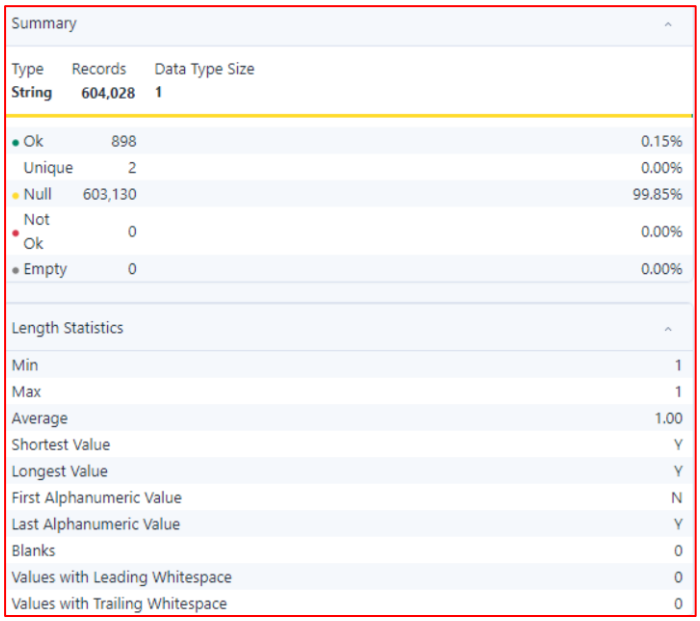
Ok	2,639	0.44%
Unique	4	0.00%
Null	601,389	99.56%
Not Ok	0	0.00%
Empty	0	0.00%

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%

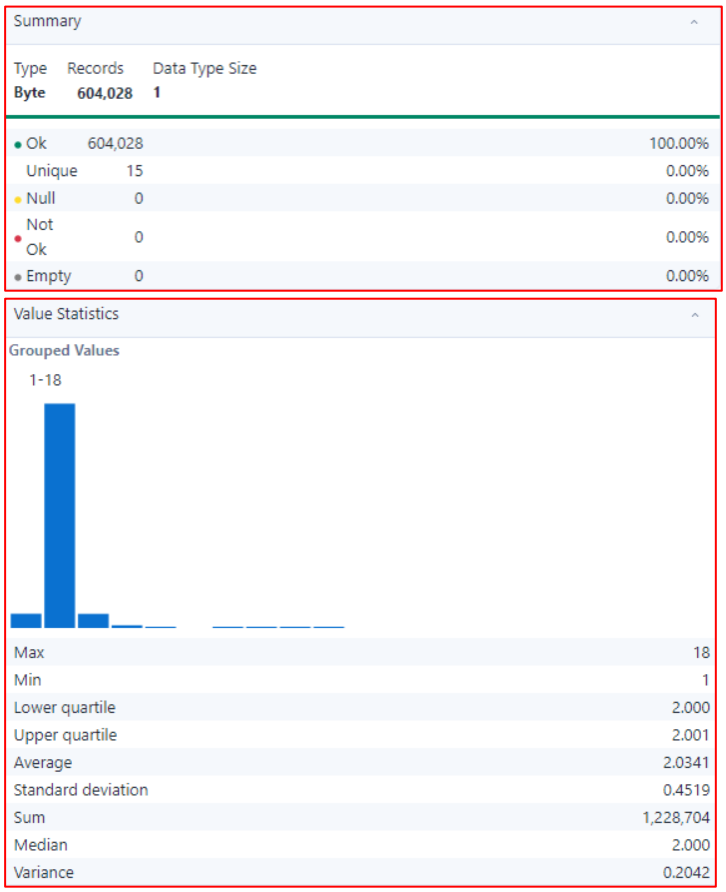
Length Statistics

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**

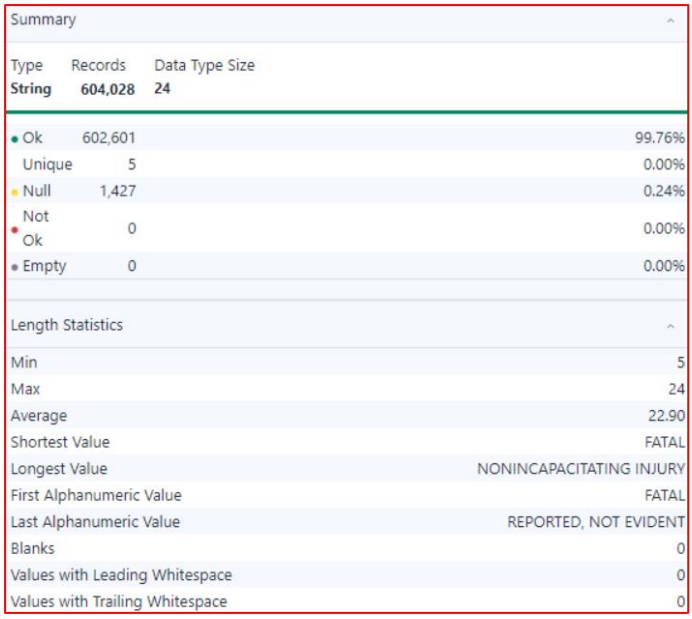
  - Summary:



  - 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

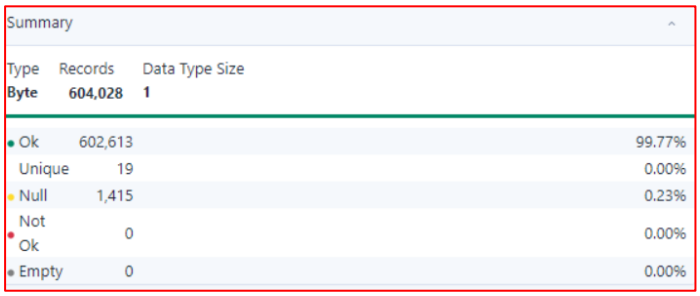
Ok	602,601	99.76%
Unique	5	0.00%
Null	1,427	0.24%
Not Ok	0	0.00%
Empty	0	0.00%

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.

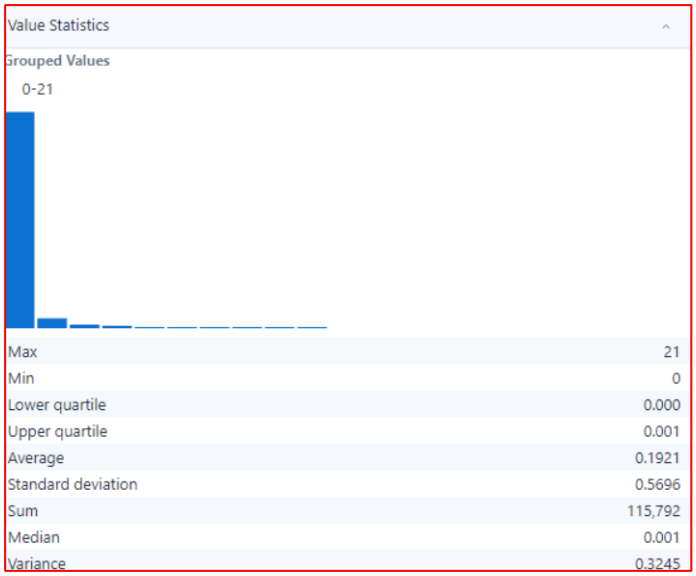
- **INJURY_TOTAL**
  - Summary:



Summary

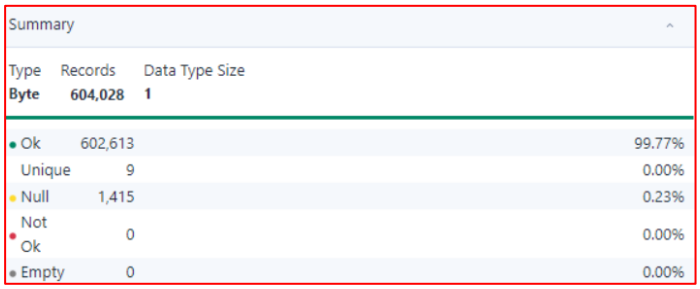
Type	Records	Data Type Size
Byte	604,028	1

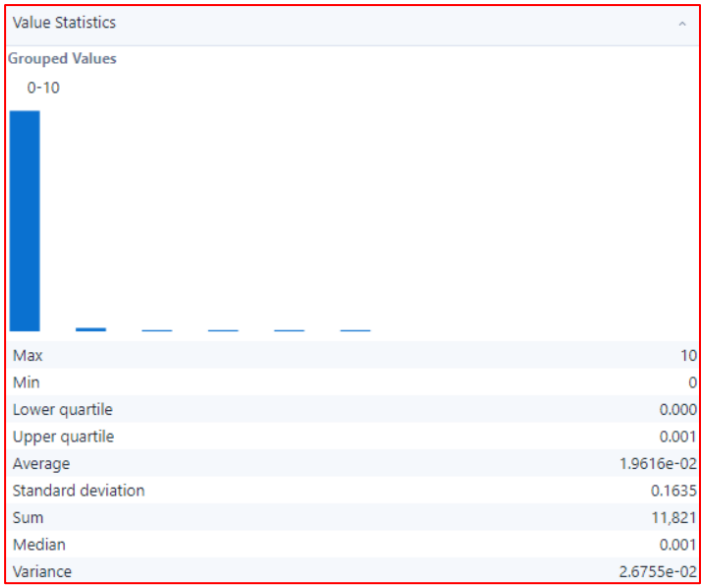
Ok	602,613	99.77%
Unique	19	0.00%
Null	1,415	0.23%
Not Ok	0	0.00%
Empty	0	0.00%



- 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:

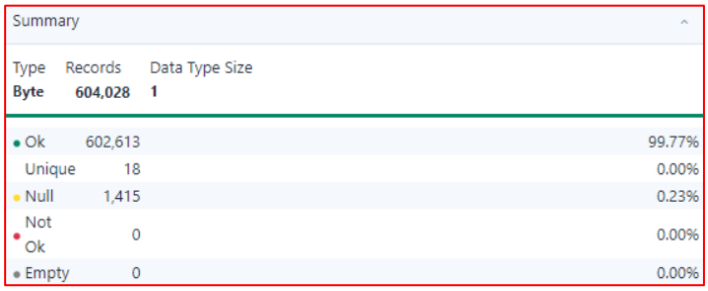




- Observation:
  - The dataset contains 604,028 records, with 602,613 valid entries (99.77%) and 1,415 null entries (0.23%).
  - There are 9 unique values in the dataset, indicating limited variability in the data.
  - The dataset is clean and suitable for analysis, but the null values should be handled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

- **INJURIES_NON_INCAPICITATING**

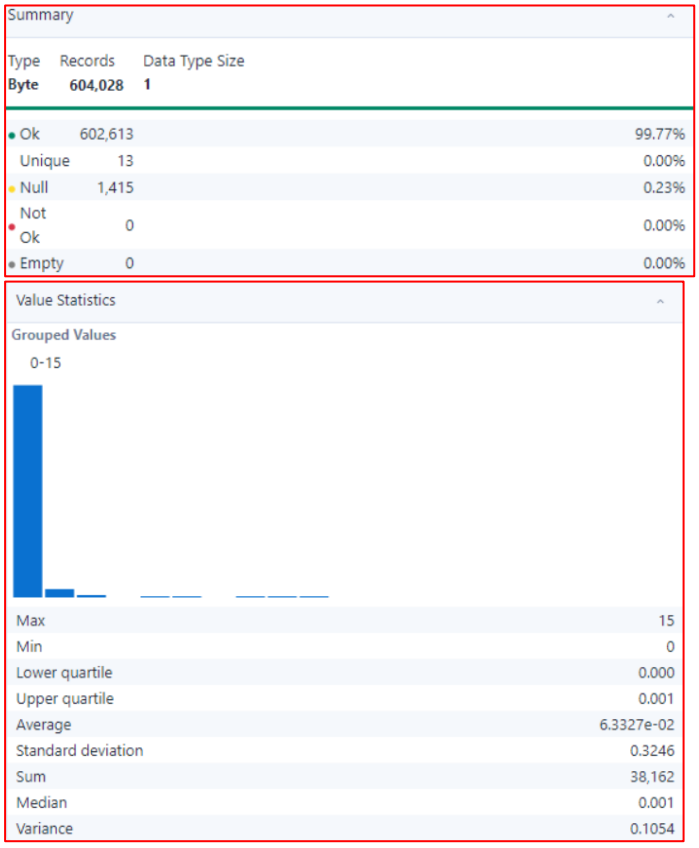
  - Summary:



- 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**

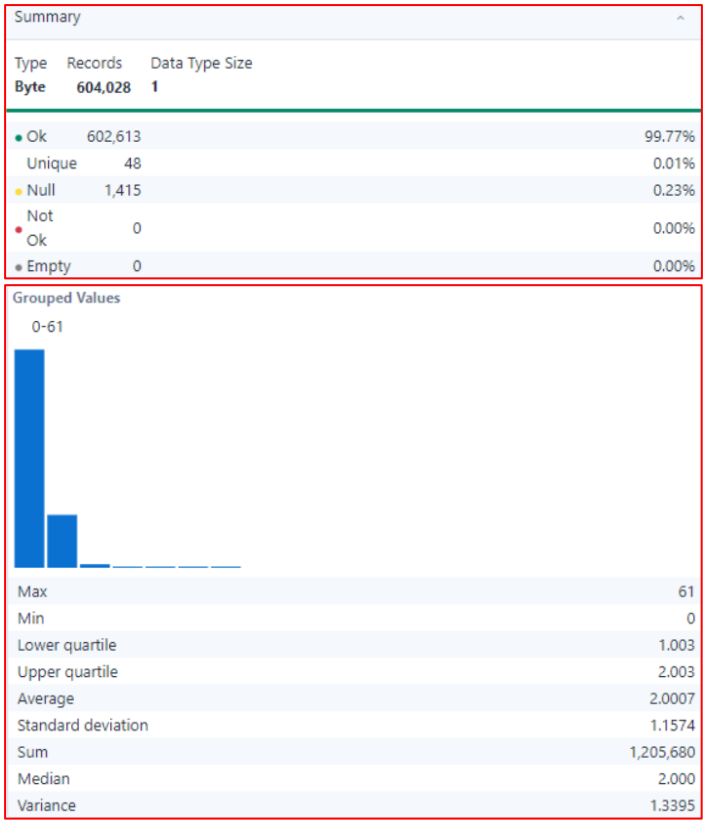
  - Summary:



- 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.

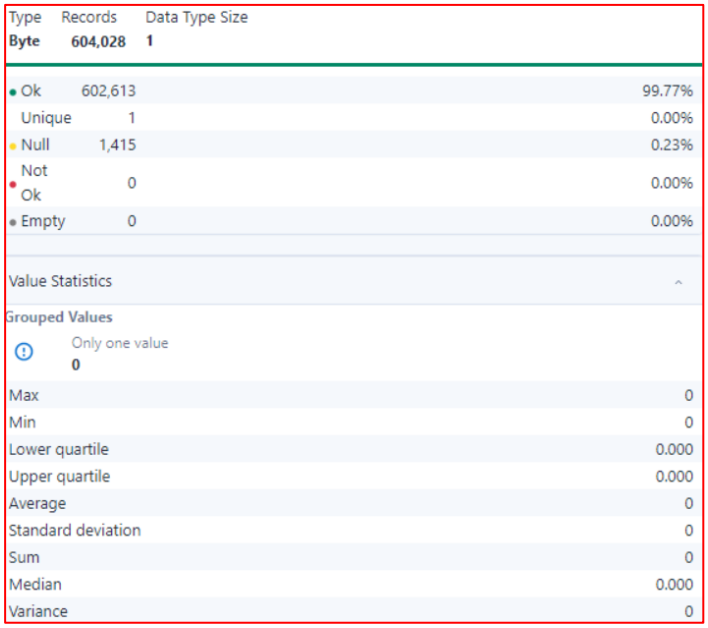
- **INJURIES_NO_INDICATION**

  - 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

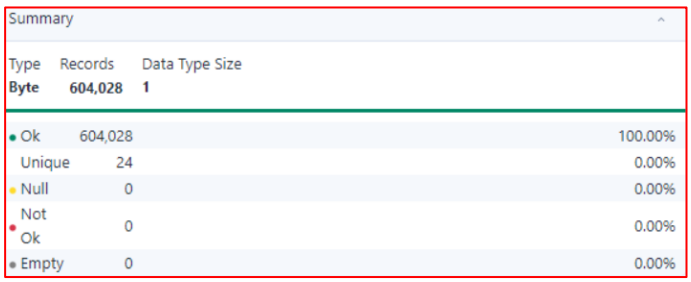
Grouped Values

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.

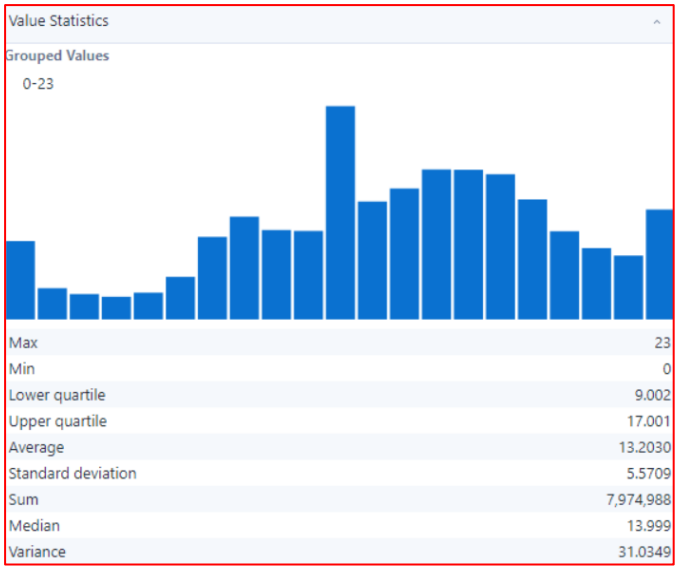
- **CRASH_HOUR**
  - Summary:



Summary

Type	Records	Data Type Size
Byte	604,028	1

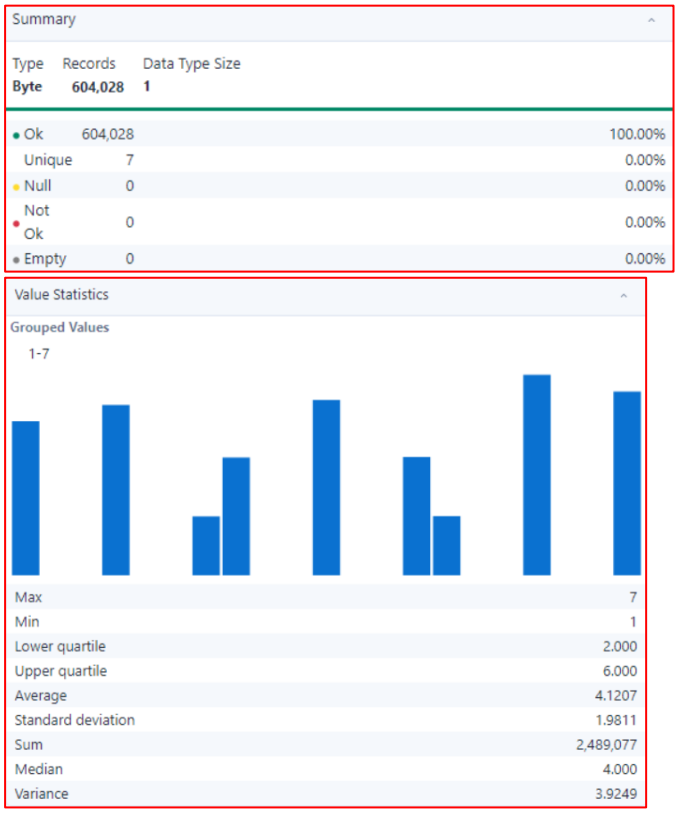
Ok	604,028		100.00%
Unique	24		0.00%
Null	0		0.00%
Not Ok	0		0.00%
Empty	0		0.00%



- 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.

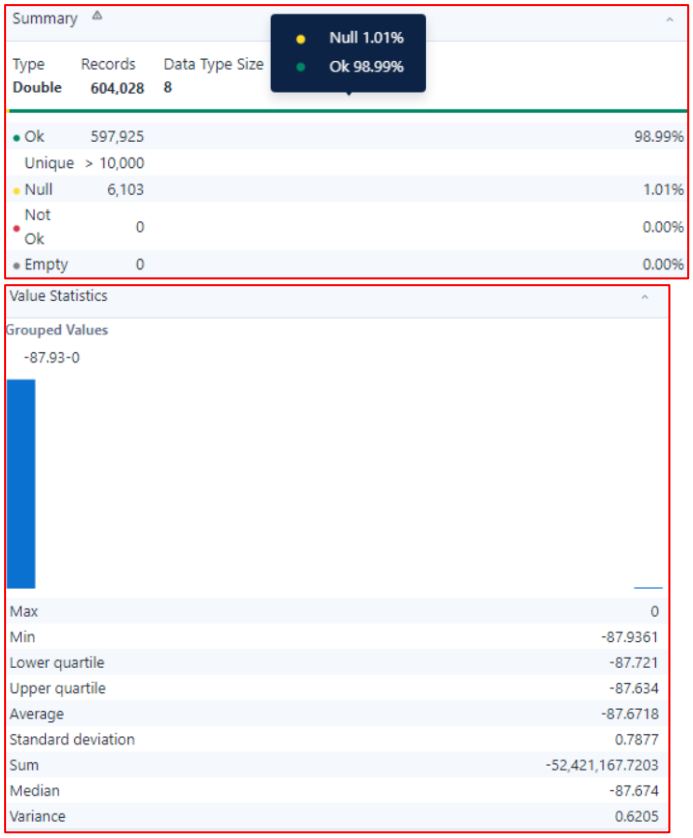
- **CRASH_MONTH**

  - 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**

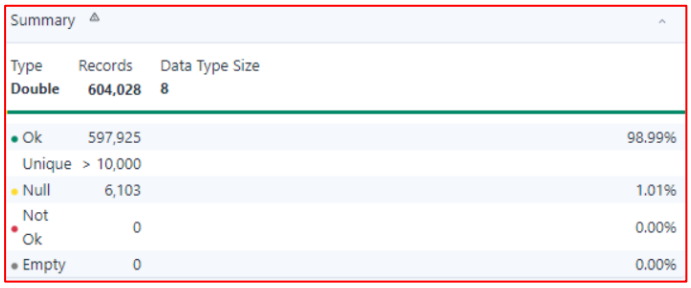
  - Summary:

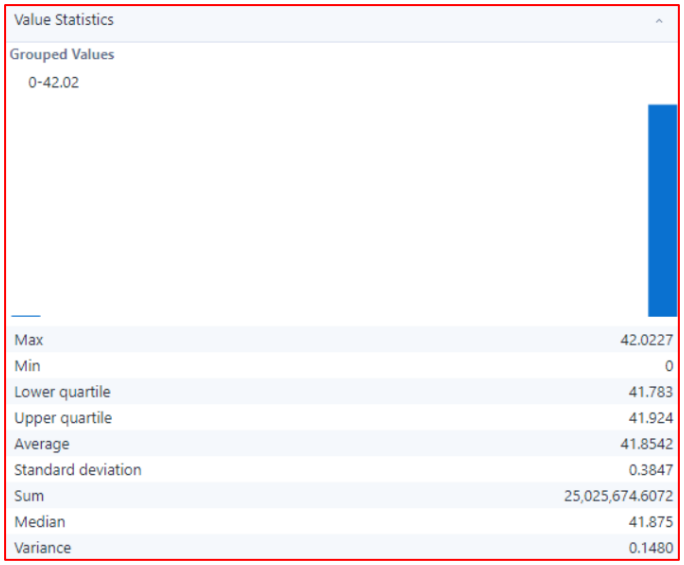


- Observation:
  - The dataset contains 604,028 records, with 98.99% valid records and 1.01% null values (6,103 records).
  - There are over 10,000 unique values, indicating a high degree of granularity.
  - The dataset is suitable for numerical analysis and contains mostly complete data with minimal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

- **LATITUDE**

  - Summary:

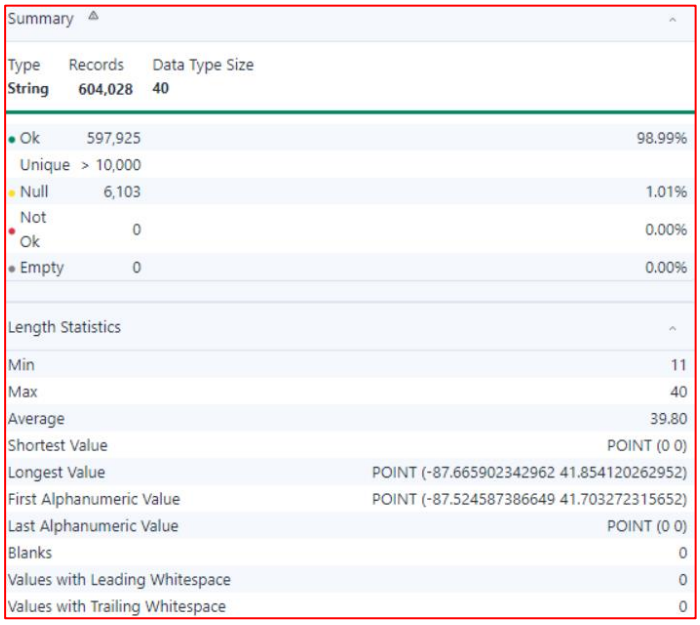




- Observation:
  - The dataset consists of 604,028 records, with 98.99% valid records and 1.01% null values (6,103 records).
  - The field contains more than 10,000 unique values, indicating high variability.
  - The variance is 0.1480, indicating a concentrated spread of values within the range.

- **LOCATION**

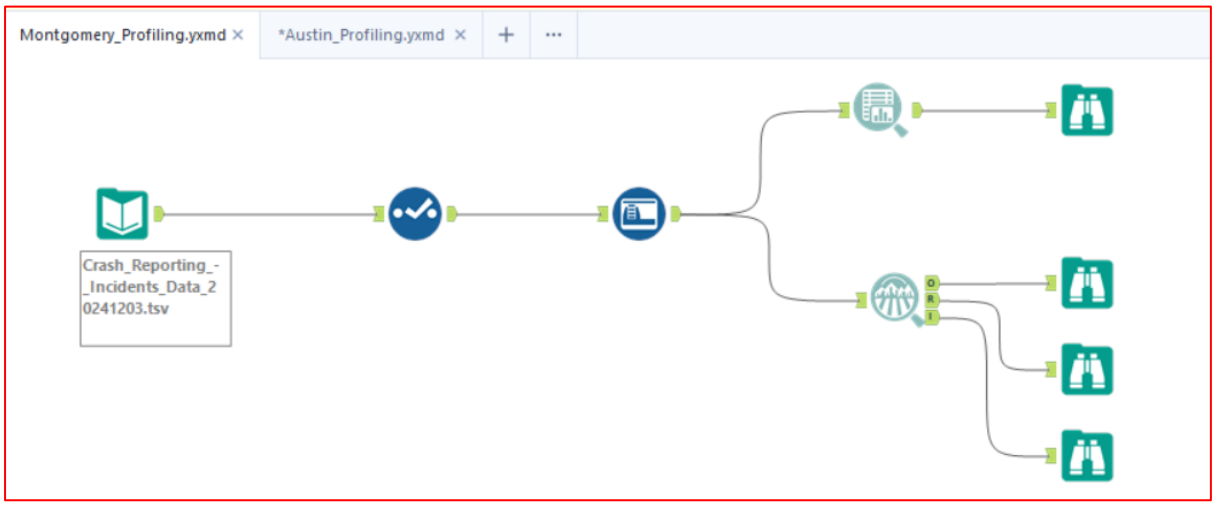
  - Summary:

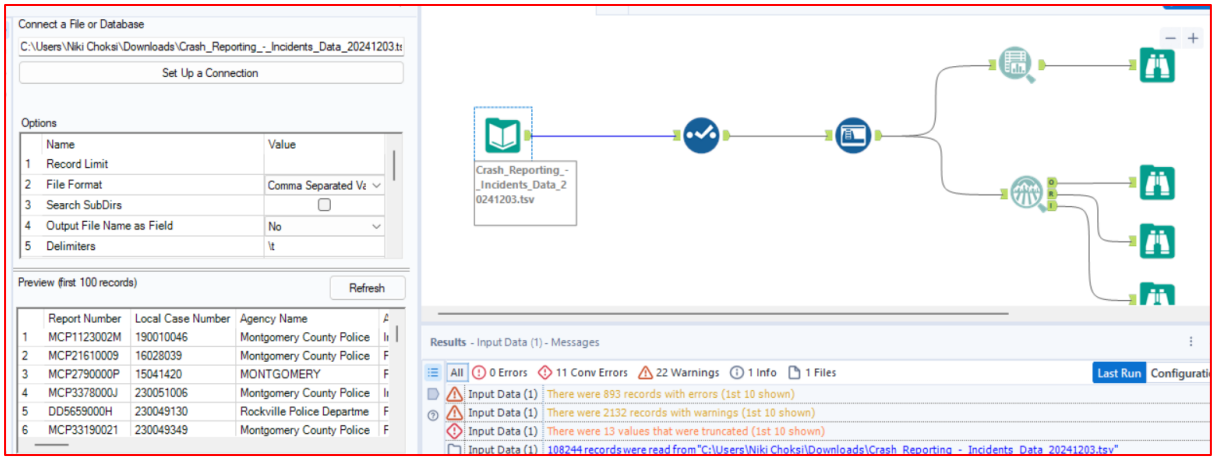


  - Observation:
    - The dataset consists of 604,028 records, with 98.99% valid records and 1.01% null values (6,103 records).

- 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

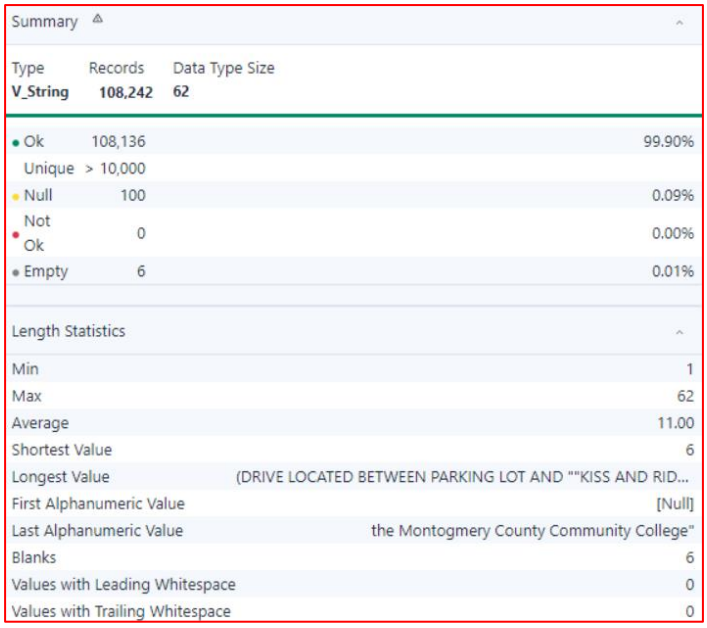
		names in the region.
RELATED_NON_MOTORIST	Indicates involvement of non-motorists (e.g., pedestrians, cyclists) in the crash.	Confirm logical consistency with other crash factors and standardize categories.
AT_FAULT	Identifies the party at fault in the crash.	Ensure completeness and logical alignment 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 INTESECTION 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

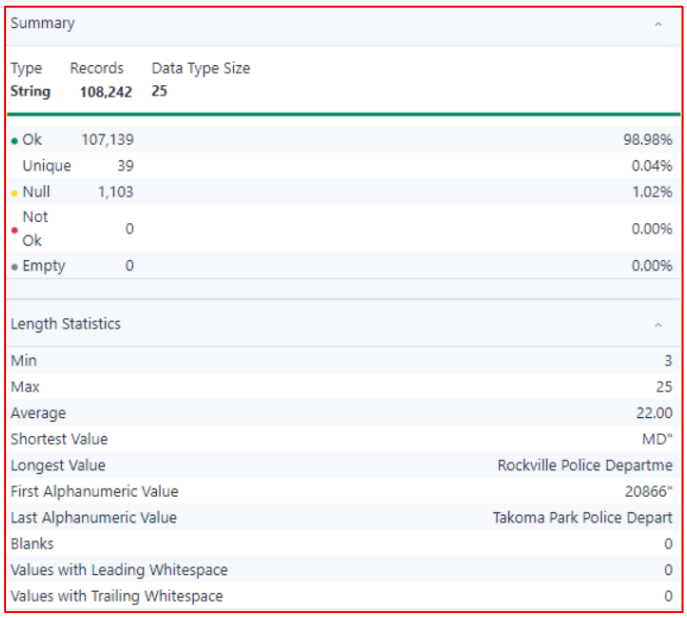
Ok	108,136	99.90%
Unique	> 10,000	
Null	100	0.09%
Not Ok	0	0.00%
Empty	6	0.01%

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 Montogmery 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 Ok	0	0.00%
Empty	0	0.00%

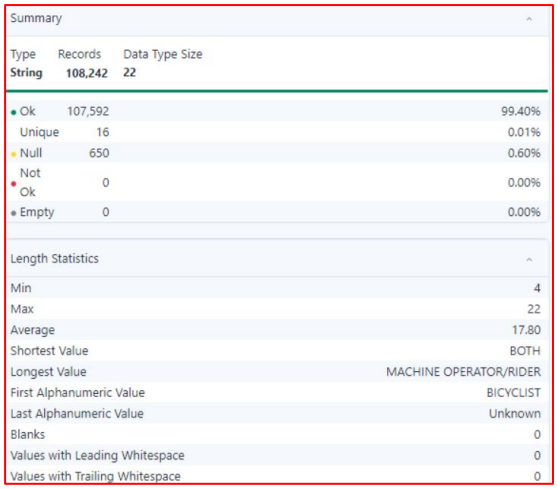
Length Statistics

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

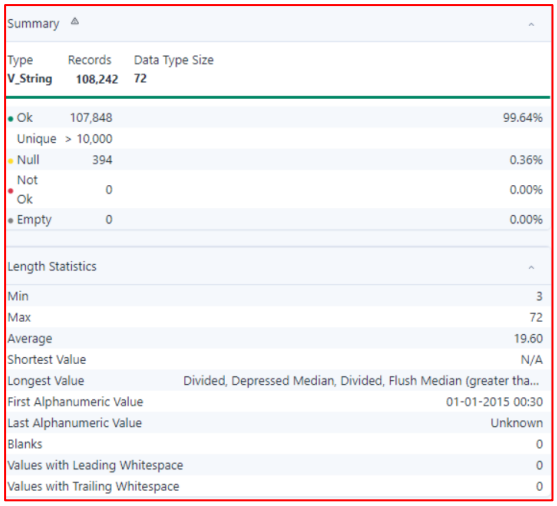
Ok	107,592	99.40%
Unique	16	0.01%
Null	650	0.60%
Not Ok	0	0.00%
Empty	0	0.00%

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

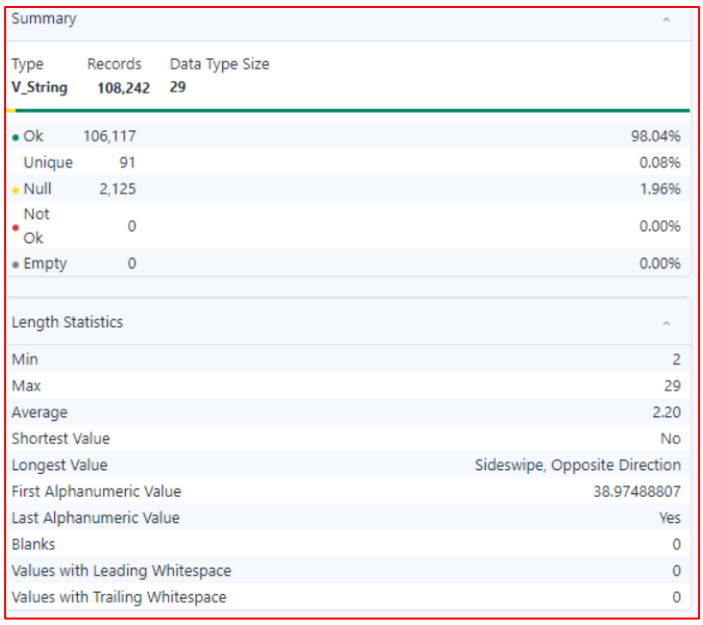
  - Observation:
    - The dataset contains 108,242 records, with 99.40% valid and 0.60% null values (650 records).
    - There are 16 unique values, indicating limited diversity in categories.
    - There are no blanks, and no values have leading or trailing whitespaces.
    - The dataset appears clean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

- **CRASH_DATE/TIME**
  - Summary:



- 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.

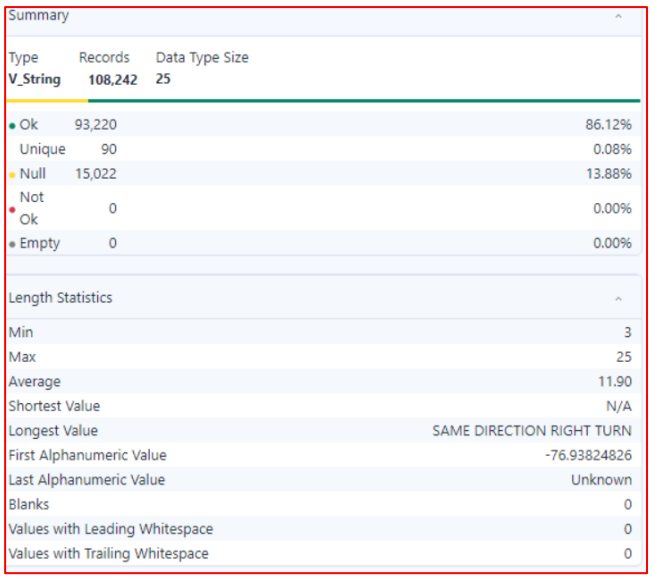
- **HIT/RUN**
  - Summary:



  - 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

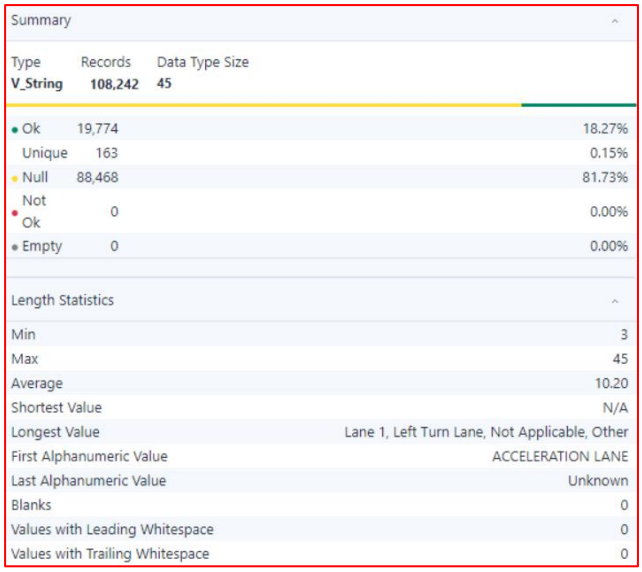
Ok	93,220	86.12%
Unique	90	0.08%
Null	15,022	13.88%
Not Ok	0	0.00%
Empty	0	0.00%

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

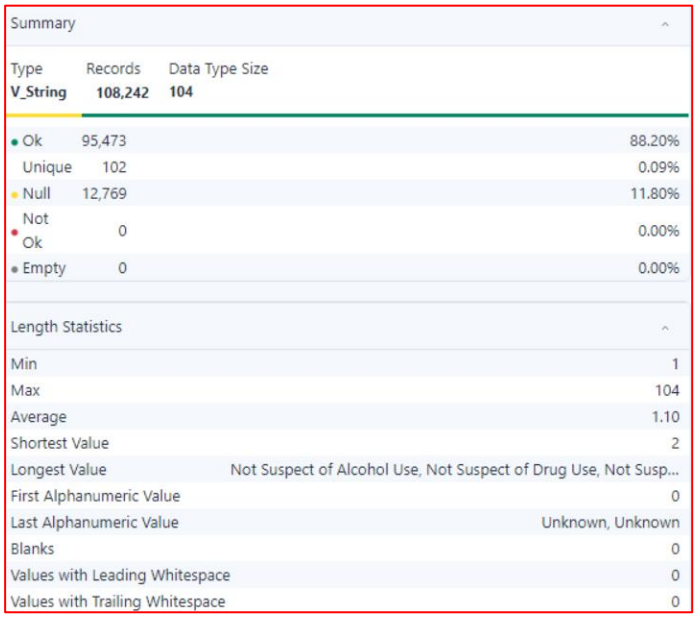
Ok	19,774	18.27%
Unique	163	0.15%
Null	88,468	81.73%
Not Ok	0	0.00%
Empty	0	0.00%

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%

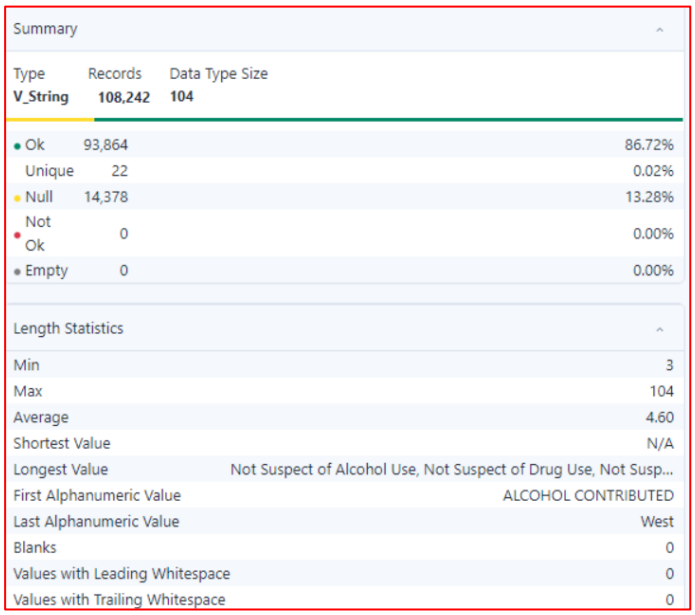
Length Statistics

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

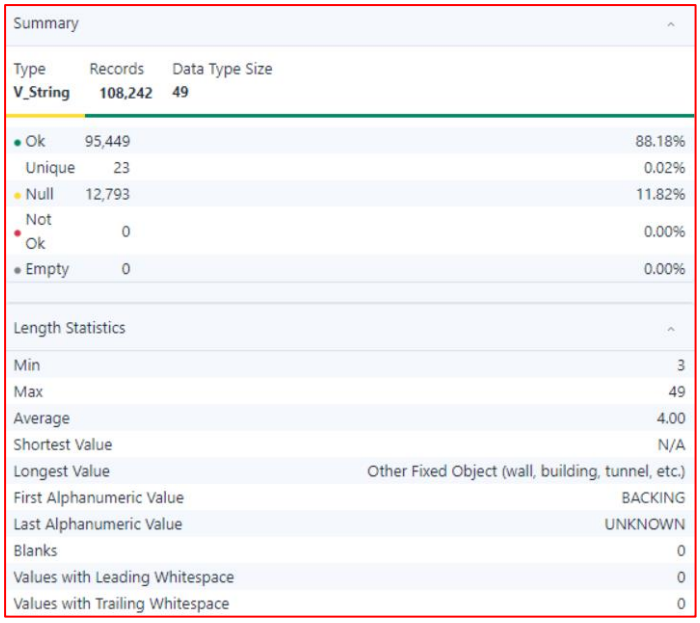
Ok	93,864	86.72%
Unique	22	0.02%
Null	14,378	13.28%
Not Ok	0	0.00%
Empty	0	0.00%

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

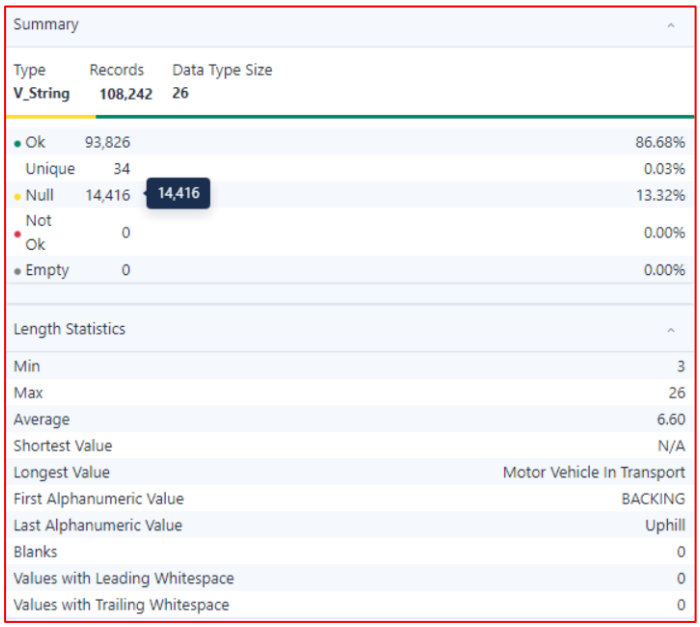
Ok	95,449	88.18%
Unique	23	0.02%
Null	12,793	11.82%
Not Ok	0	0.00%
Empty	0	0.00%

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	13.32%
Not Ok	0	0.00%
Empty	0	0.00%

14,416

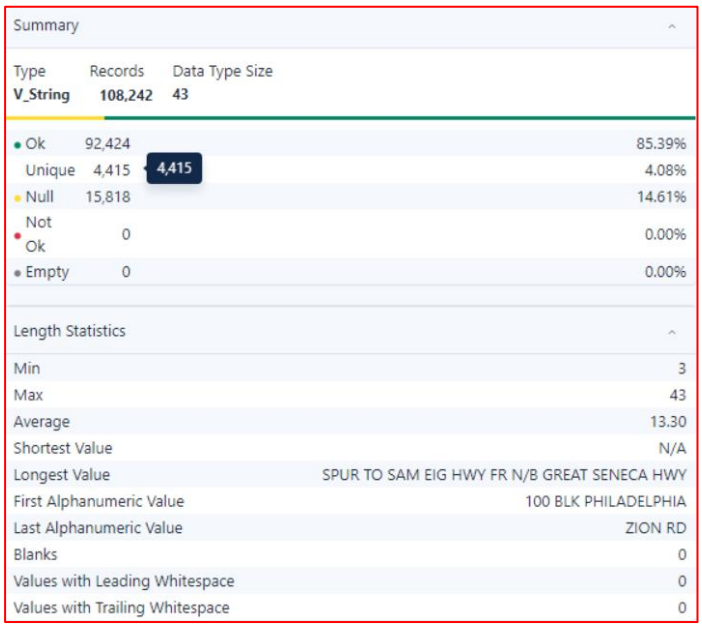
Length Statistics

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:

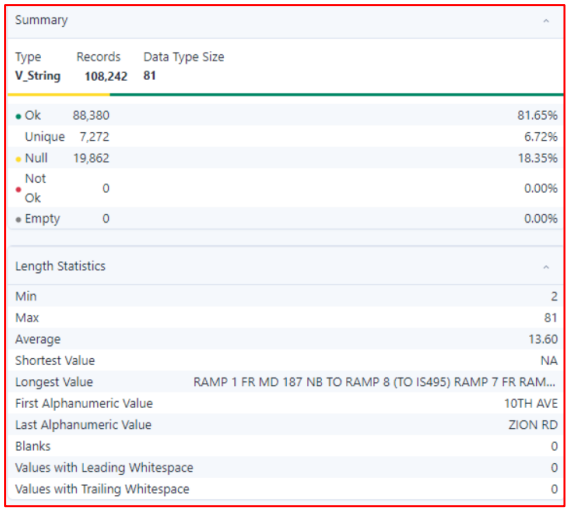


Summary			
Type	Records	Data Type Size	
V_String	108,242	43	
Ok	92,424		85.39%
Unique	4,415	4,415	4.08%
Null	15,818		14.61%
Not Ok	0		0.00%
Empty	0		0.00%

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

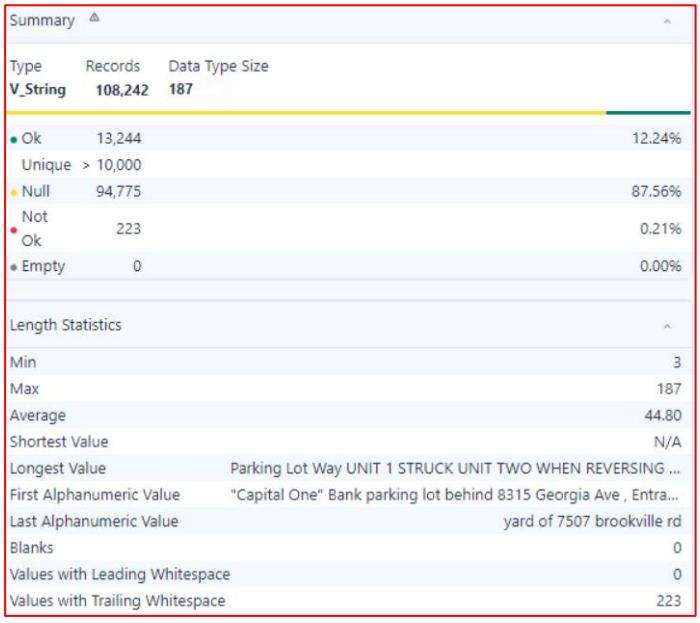
Ok	88,380	81.65%
Unique	7,272	6.72%
Null	19,862	18.35%
Not Ok	0	0.00%
Empty	0	0.00%

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

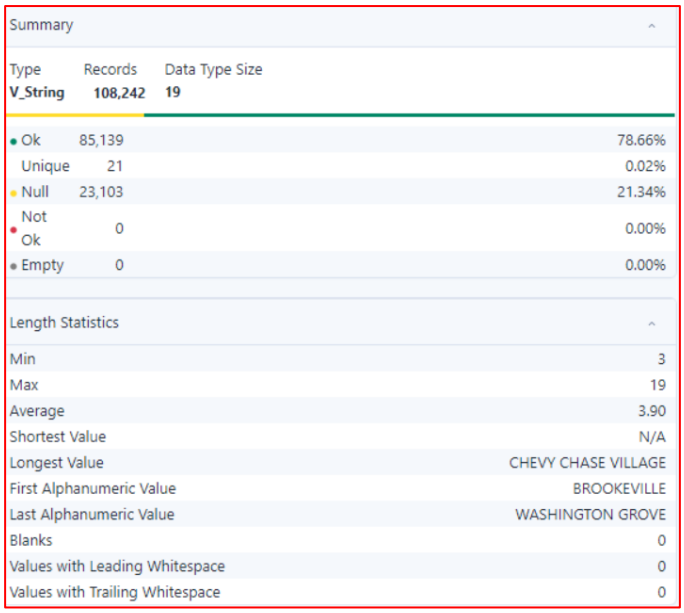
Ok	13,244	12.24%
Unique	> 10,000	
Null	94,775	87.56%
Not Ok	223	0.21%
Empty	0	0.00%

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:

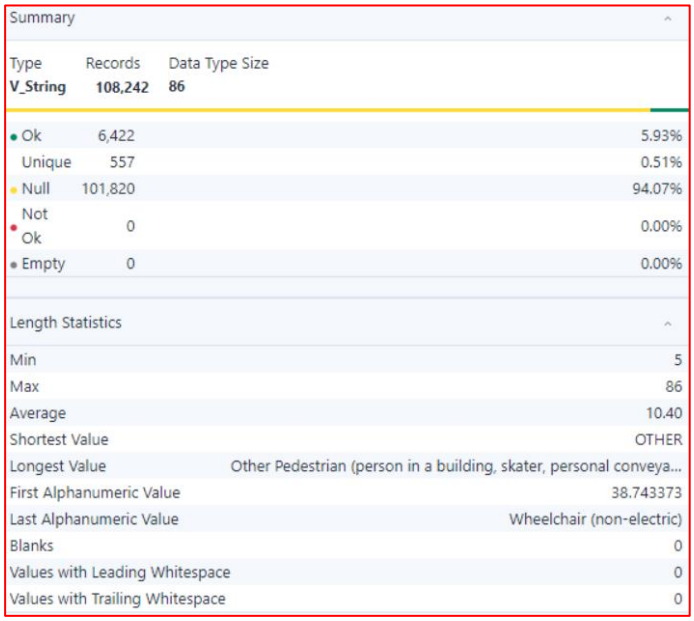


Summary		
Type	Records	Data Type Size
V_String	108,242	19
Ok	85,139	78.66%
Unique	21	0.02%
Null	23,103	21.34%
Not Ok	0	0.00%
Empty	0	0.00%

Length Statistics	
Min	3
Max	19
Average	3.90
Shortest Value	N/A
Longest Value	CHEVY CHASE VILLAGE
First Alphanumeric Value	BROOKEVILLE
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:

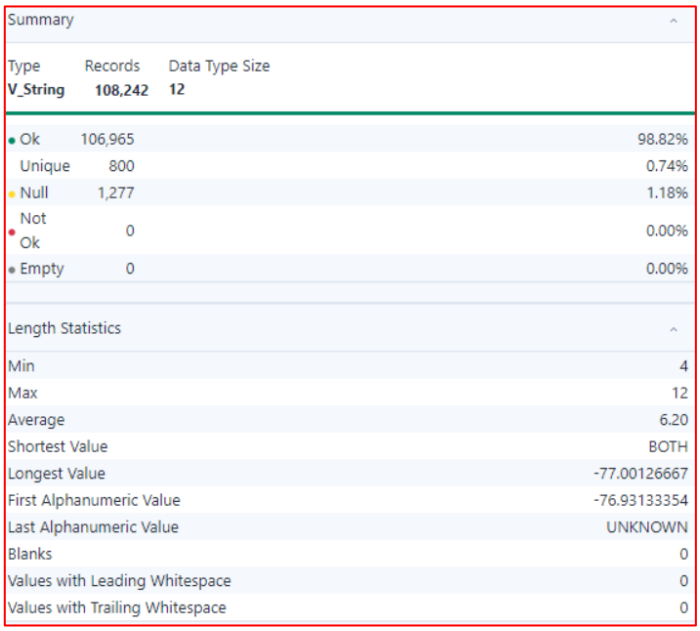


Summary		
Type	Records	Data Type Size
V_String	108,242	86
Ok	6,422	5.93%
Unique	557	0.51%
Null	101,820	94.07%
Not Ok	0	0.00%
Empty	0	0.00%

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:

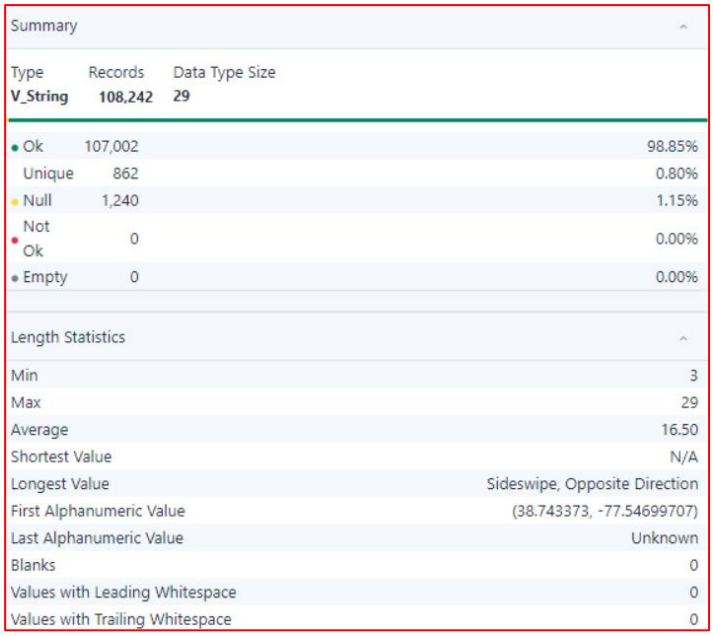


Summary		
Type	Records	Data Type Size
V_String	108,242	12
Ok	106,965	98.82%
Unique	800	0.74%
Null	1,277	1.18%
Not Ok	0	0.00%
Empty	0	0.00%

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

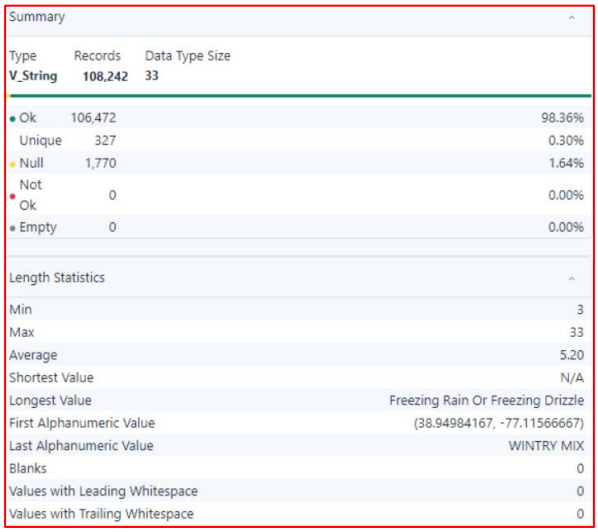
Ok	107,002	98.85%
Unique	862	0.80%
Null	1,240	1.15%
Not Ok	0	0.00%
Empty	0	0.00%

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%

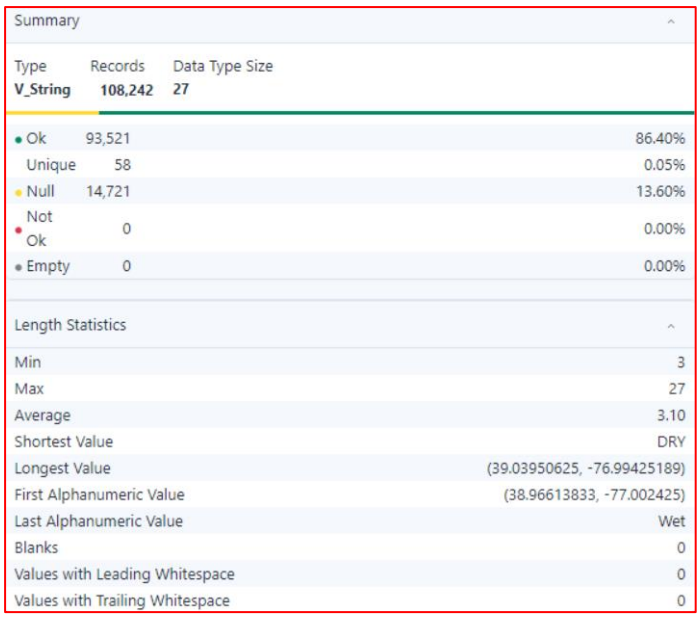
Length Statistics

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

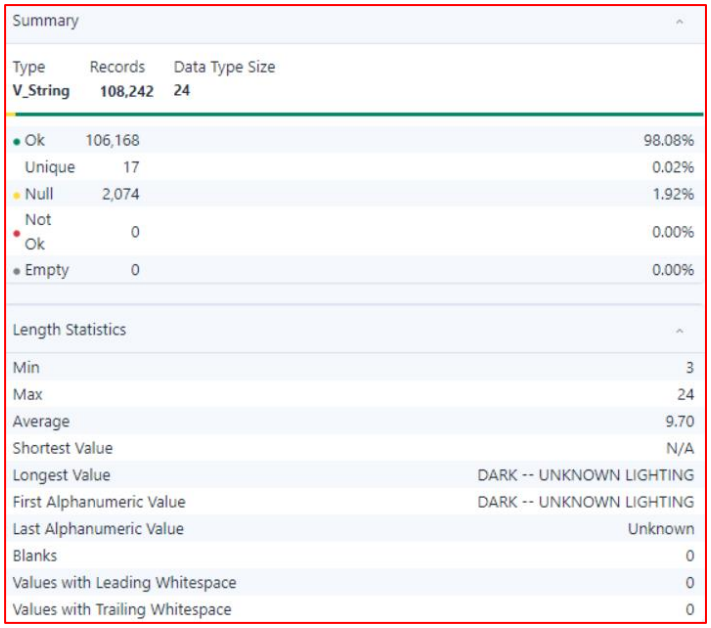
Ok	93,521	86.40%
Unique	58	0.05%
Null	14,721	13.60%
Not Ok	0	0.00%
Empty	0	0.00%

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

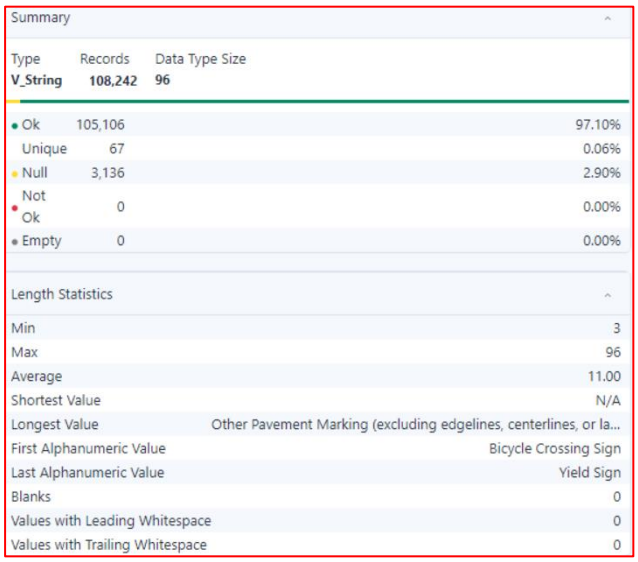
Ok	106,168	98.08%
Unique	17	0.02%
Null	2,074	1.92%
Not Ok	0	0.00%
Empty	0	0.00%

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%

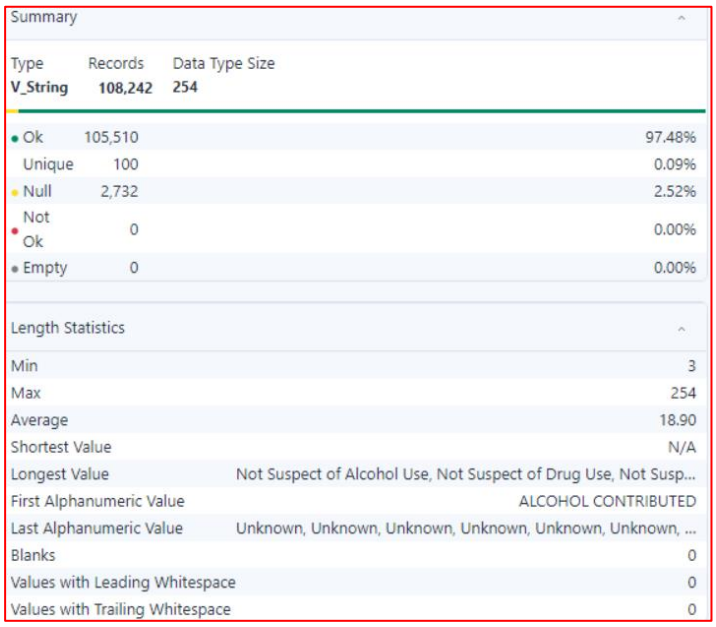
Length Statistics

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

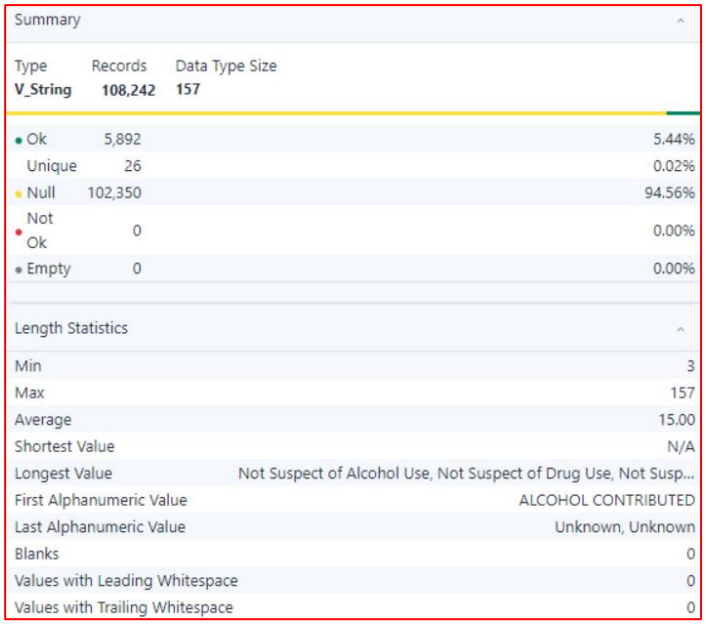
Ok	105,510	97.48%
Unique	100	0.09%
Null	2,732	2.52%
Not Ok	0	0.00%
Empty	0	0.00%

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 Susp...
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

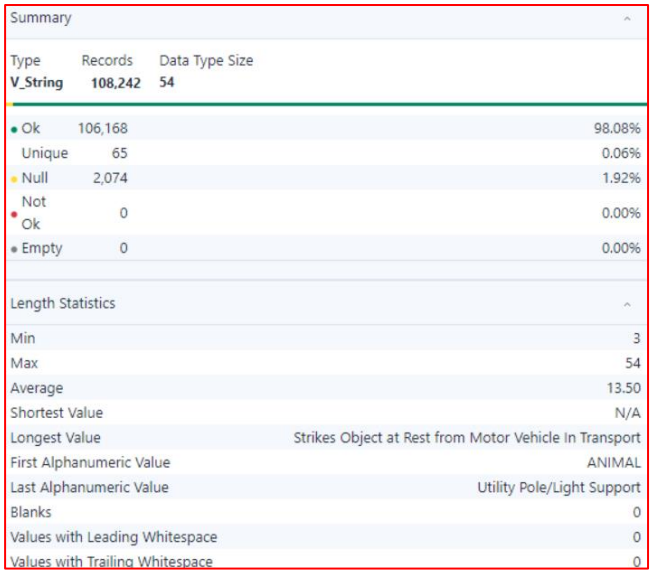
Ok	5,892	5.44%
Unique	26	0.02%
Null	102,350	94.56%
Not Ok	0	0.00%
Empty	0	0.00%

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%

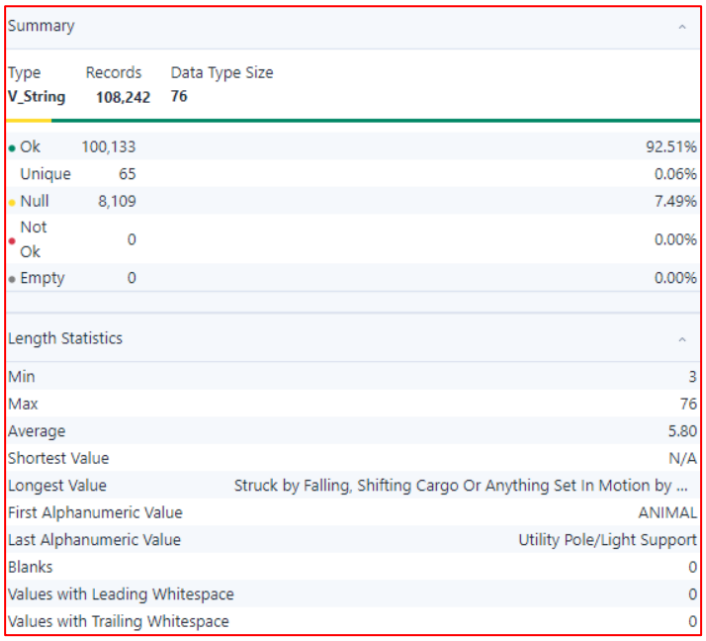
Length Statistics

Min	3
Max	54
Average	13.50
Shortest Value	N/A
Longest Value	Strikes Object at Rest from Motor Vehicle in Transport
First Alphanumeric Value	ANIMAL
Last Alphanumeric Value	Utility Pole/Light Support
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 Transport" 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

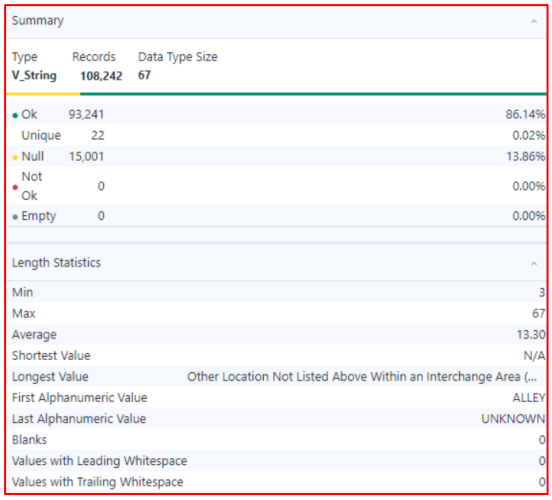
Ok	100,133	92.51%
Unique	65	0.06%
Null	8,109	7.49%
Not Ok	0	0.00%
Empty	0	0.00%

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

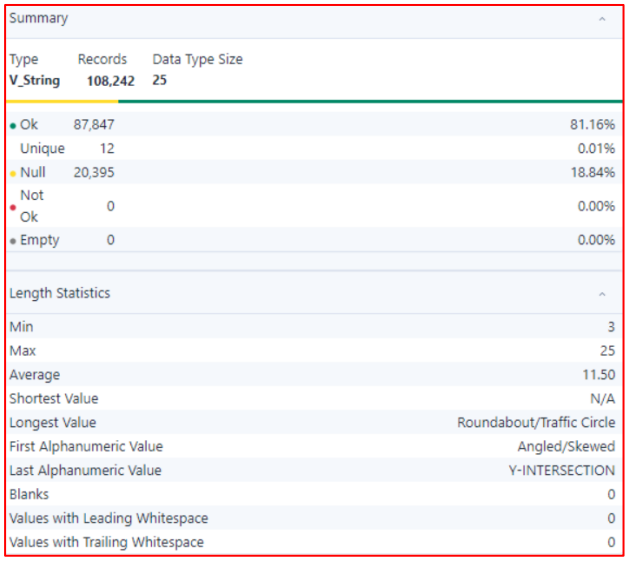
  - Observation:
    - The dataset contains 108,242 records, with 100,133 valid records (92.51%) and 8,109 null records (7.49%).
    - There are 65 unique values, representing a moderate level of variability in the data.
    - The longest value is "Struck by Falling, Shifting Cargo Or Anything Set in Motion 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:



- 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:



  - 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:

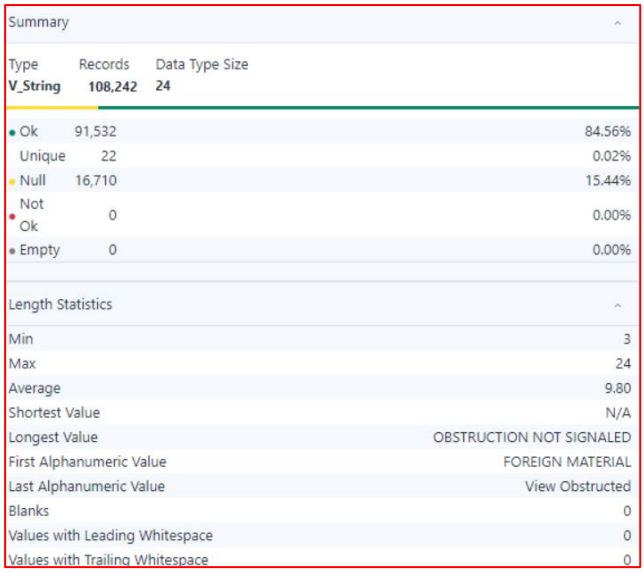
Type	Records	Data Type Size
V_String	108,242	33

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

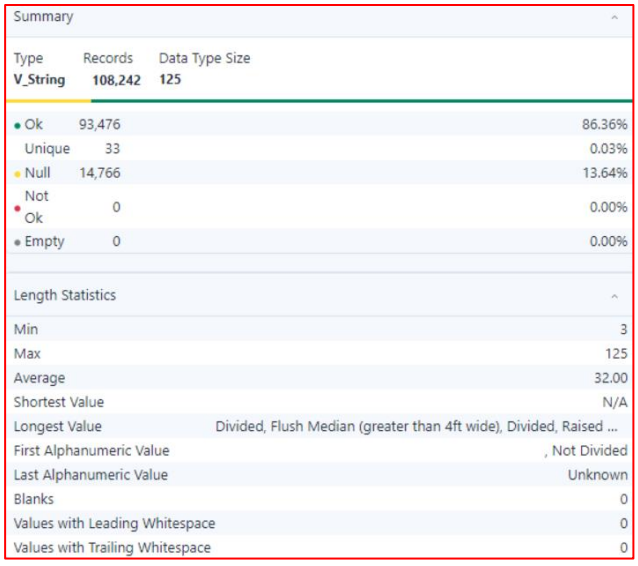
Ok	91,532	84.56%
Unique	22	0.02%
Null	16,710	15.44%
Not Ok	0	0.00%
Empty	0	0.00%

Length Statistics

Min	3
Max	24
Average	9.80
Shortest Value	N/A
Longest Value	OBSTRUCTION NOT SIGNALED
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 SIGNALED 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%

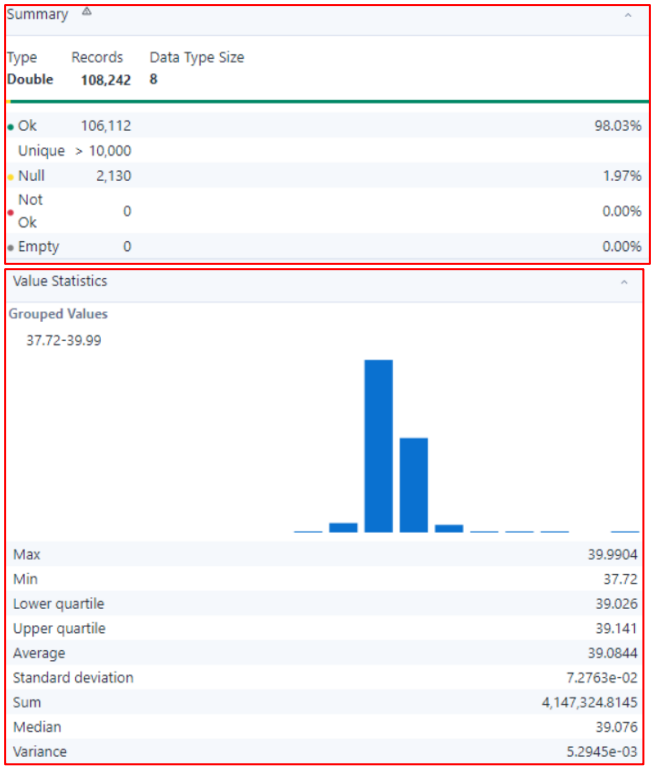
Length Statistics

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%).

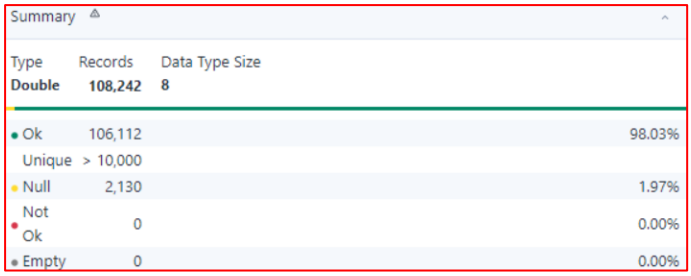
- There are 33 unique values, representing a low level of variability in the data.
- The longest value is "Divided, Flush Median (greater than 4ft wide), Divided, Raised ..., Not Divided", and the shortest value is "N/A."
- There are no blank values, but there are 0 values with leading whitespaces and 0 values with trailing whitespaces in the dataset.

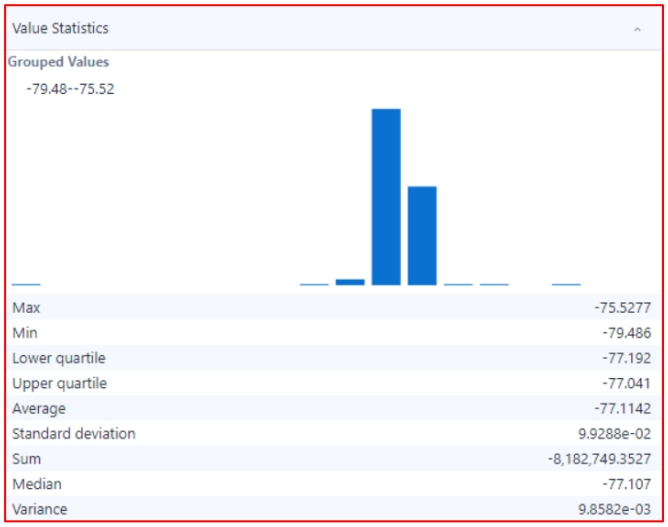
- **LATITUDE**
  - Summary:



  - 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:

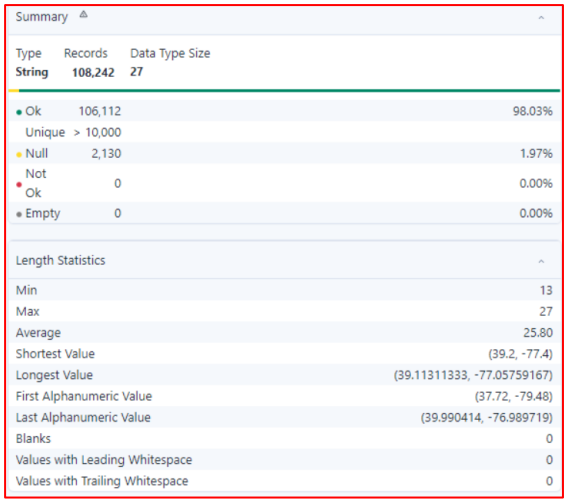




- Observation:
  - The dataset contains 108,242 records, with 92,424 valid records (85.39%) and 15,818 null records (14.61%).
  - There are 4,415 unique values, representing a high level of variability in the data.
  - The longest value is "SPUR TO SAM EIG HWY FR N/B GREAT SENECA HWY," and the shortest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

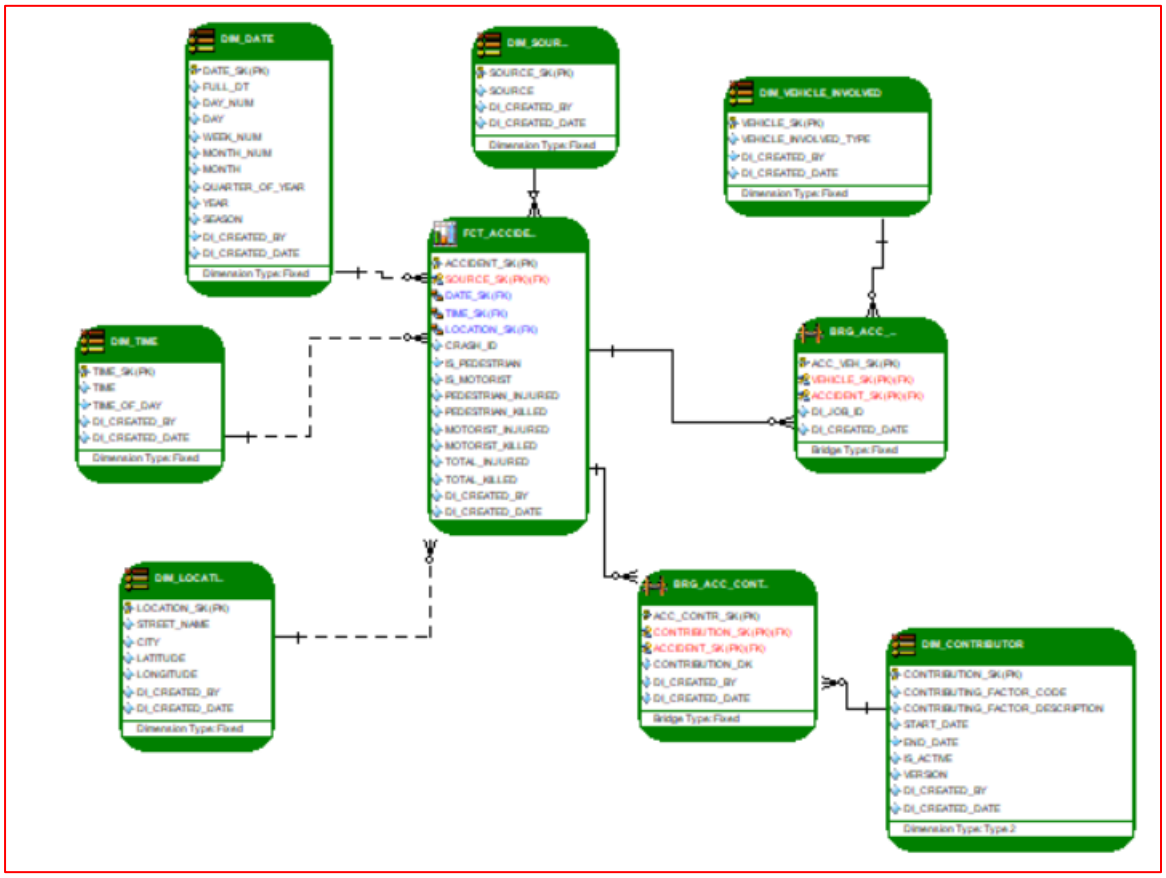
- **LOCATION**

  - Summary:



- 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.

DIMENSION MODEL



This above 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_ACCIDENTS**
    - **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: DIM_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: DIM_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: DIM_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: DIM_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_VEH**
   - **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: DIM_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_CONTR**
   - **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: DIM_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.

AZURE DATA FACTORY IMPLEMENTATION

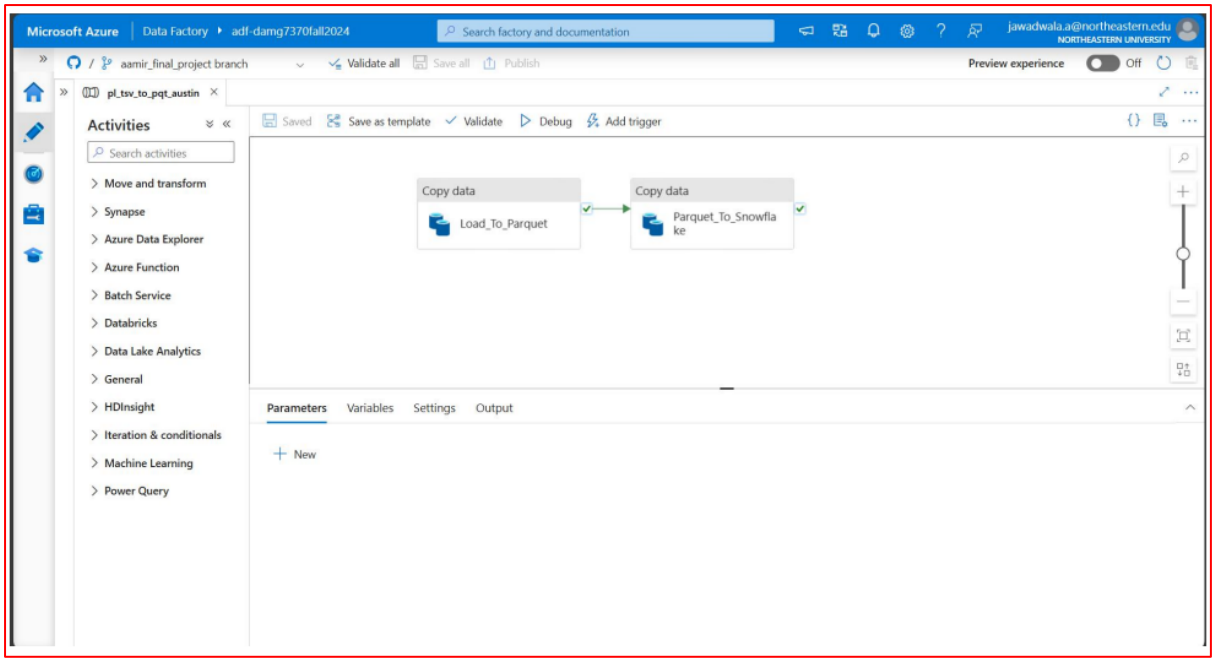
The staging process in Azure Data Factory (ADF) involves the following key steps and components based on the images and workflow provided:

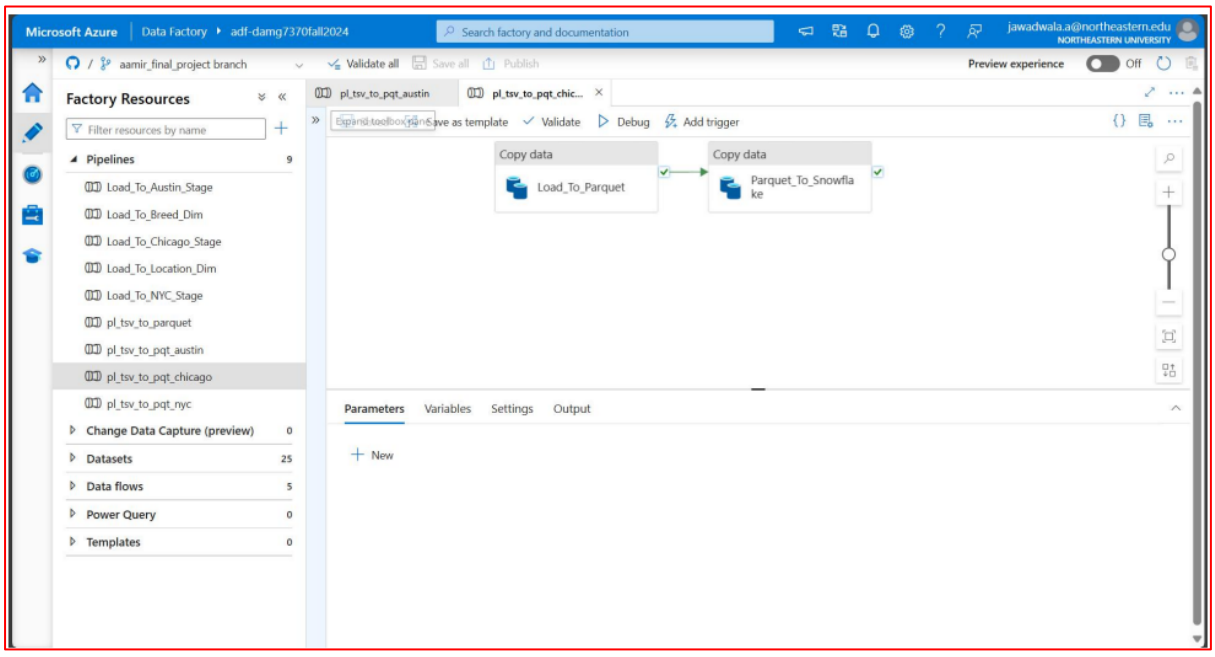
1. **Data Ingestion and Transformation:**
    - Data is extracted from different sources, such as TSV or CSV files, and loaded into staging areas in formats like Parquet to standardize the data structure.
    - Derived columns or transformations (like updating columns) are applied in data flows to clean and preprocess the data.

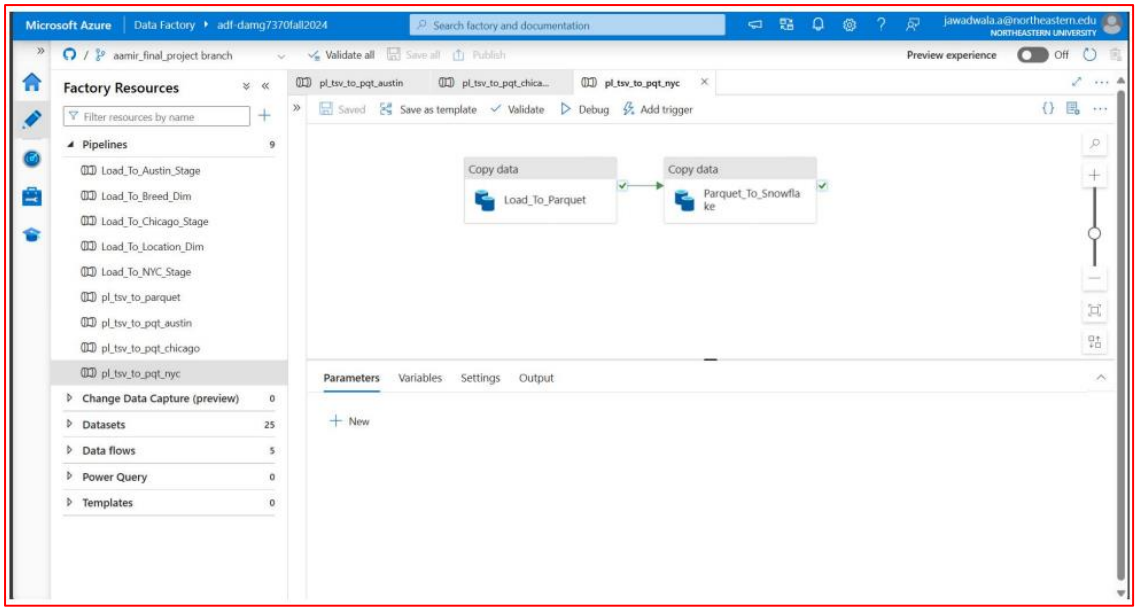
2. **Load to Staging:**
    - Data is moved to staging tables in the cloud environment (e.g., Snowflake) using pipelines.
    - Pipelines like Load_To_Austin_Stage, Load_To_Chicago_Stage, Load_To_NYC_Stage and Load_To_Montgomery standardize the process for each location.
    - The pipelines follow a consistent pattern of ingesting, transforming, and writing to Snowflake.

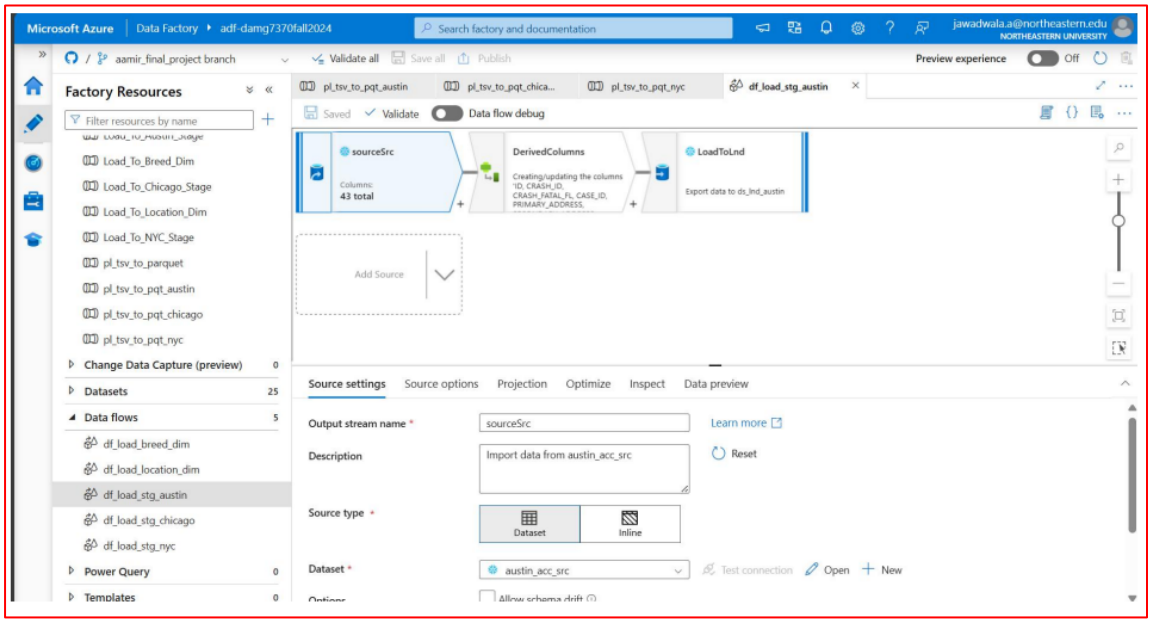
3. **Data Mapping and Cleaning:**
    - Components like tMap and tNormalize (visible in Snowflake workflows) perform tasks such as mapping source columns to target schema and normalization of data for consistency.

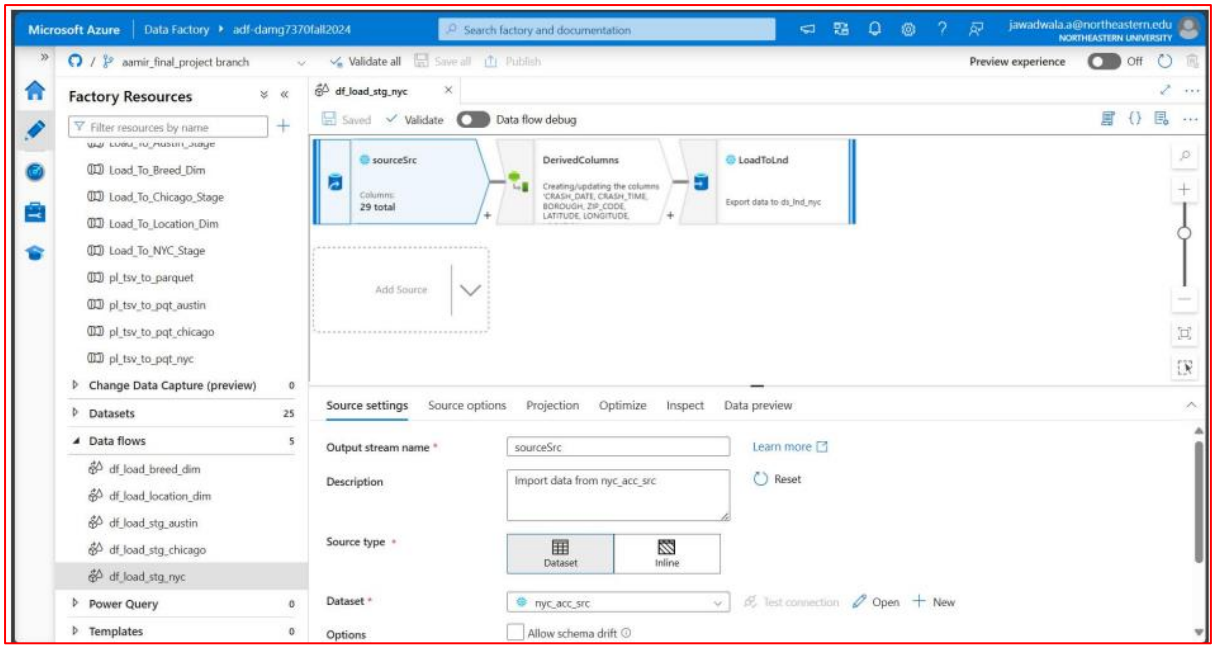
4. **Tools Used:**
    - The pipelines in Azure Data Factory are triggered with activities that connect the data source, perform transformations, and output to destinations.
    - Each city (Austin, Chicago, NYC, Montgomery) has specific data flows configured for their unique data structure but follow a unified process of staging and cleaning.

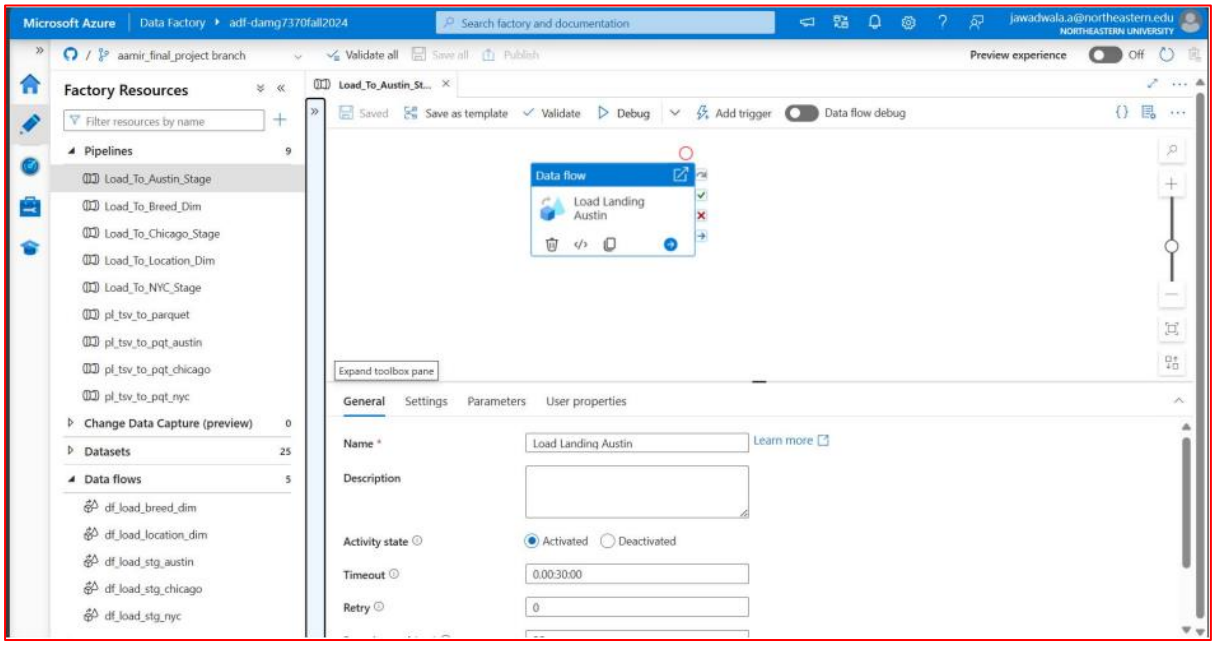

Microsoft Azure
Data Factory
adf-damg7370fall2024
Search factory and documentation
jawadwala.a@northeastern.edu
NORTHEASTERN UNIVERSITY
aamir_final_project branch
Validate all
Save all
Publish
Preview experience
Off
pl_tsv_to_pqt_austin
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Saved
Save as template
Validate
Debug
Add trigger
Copy data
Load_To_Parquet
Copy data
Parquet_To_Snowflake
Parameters
Variables
Settings
Output
New

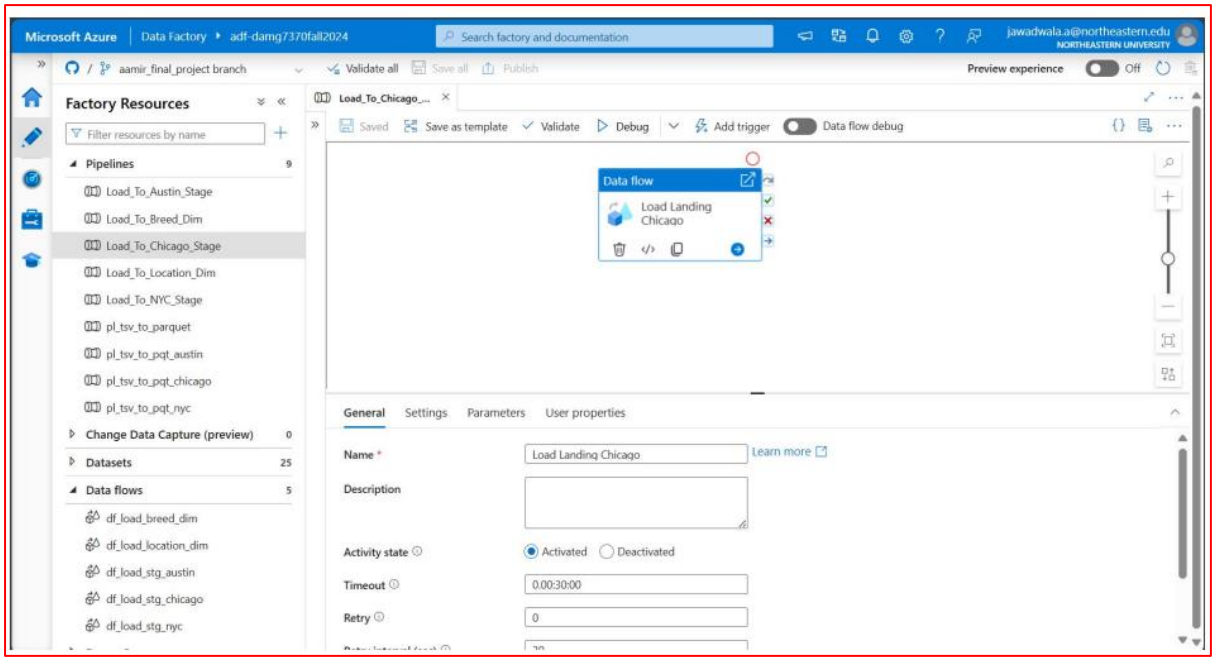

Microsoft Azure
Data Factory
adf-damg7370fall2024
Search factory and documentation
jawadwala.a@northeastern.edu
NORTHEASTERN UNIVERSITY
aamir_final_project branch
Validate all
Save all
Publish
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines
9
Load_To_Austin_Stage
Load_To_Breed_Dim
Load_To_Chicago_Stage
Load_To_Location_Dim
Load_To_NYC_Stage
pl_tsv_to_parquet
pl_tsv_to_pqt_austin
pl_tsv_to_pqt_chicago
pl_tsv_to_pqt_nyc
Change Data Capture (preview)
0
Datasets
25
Data flows
5
Power Query
0
Templates
0
pl_tsv_to_pqt_austin
pl_tsv_to_pqt_chic...
Save as template
Validate
Debug
Add trigger
Copy data
Load_To_Parquet
Copy data
Parquet_To_Snowflake
Parameters
Variables
Settings
Output
New

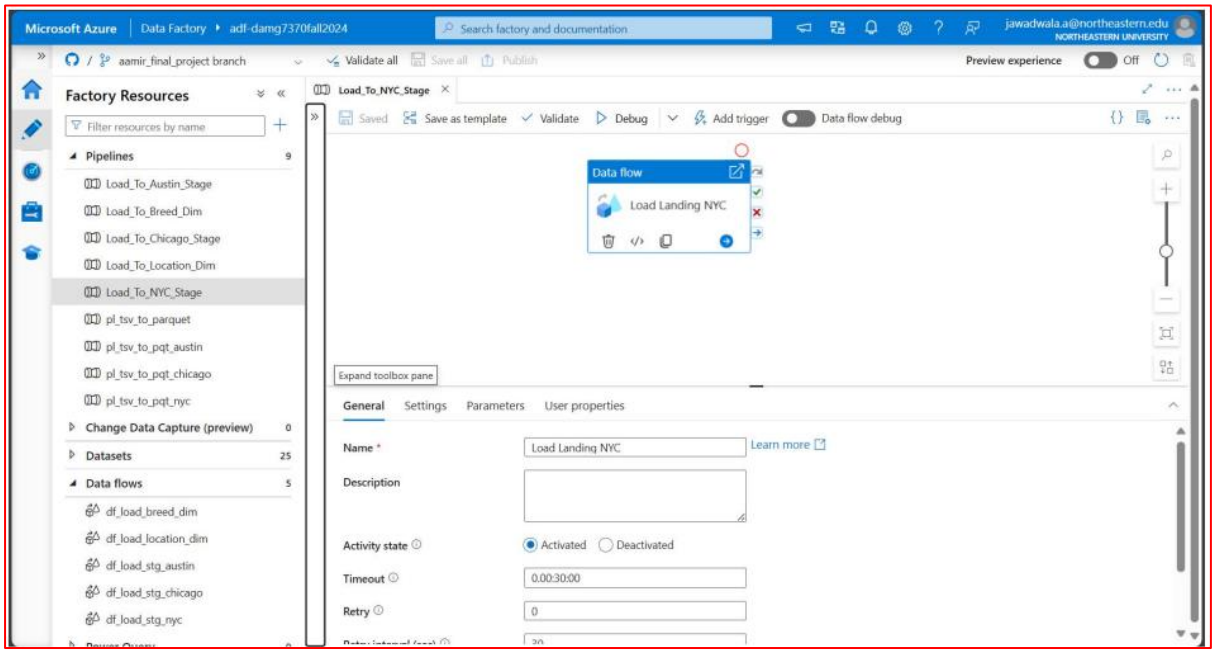

Microsoft Azure
Data Factory
adf-damg7370fall2024
Search factory and documentation
jawadwala.a@northeastern.edu
NORTHEASTERN UNIVERSITY
aamir_final_project branch
Validate all
Save all
Publish
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 9
Load_To_Austin_Stage
Load_To_Breed_Dim
Load_To_Chicago_Stage
Load_To_Location_Dim
Load_To_NYC_Stage
pl_tsv_to_parquet
pl_tsv_to_pqt_austin
pl_tsv_to_pqt_chicago
pl_tsv_to_pqt_nyc
Change Data Capture (preview) 0
Datasets 25
Data flows 5
Power Query 0
Templates 0
pl_tsv_to_pqt_austin
pl_tsv_to_pqt_chica...
pl_tsv_to_pqt_nyc
Saved
Save as template
Validate
Debug
Add trigger
Copy data
Load_To_Parquet
Copy data
Parquet_To_Snowflake
Parameters
Variables
Settings
Output
New


Microsoft Azure
Data Factory
adf-damg7370fall2024
Search factory and documentation
jawadwala.a@northeastern.edu
NORTHEASTERN UNIVERSITY
aamir_final_project branch
Validate all
Save all
Publish
Preview experience
Off
Factory Resources
Filter resources by name
Load_To_Breed_Dim
Load_To_Chicago_Stage
Load_To_Location_Dim
Load_To_NYC_Stage
pl_tsv_to_parquet
pl_tsv_to_pqt_austin
pl_tsv_to_pqt_chicago
pl_tsv_to_pqt_nyc
Change Data Capture (preview) 0
Datasets 25
Data flows 5
df_load_breed_dim
df_load_location_dim
df_load_stg_austin
df_load_stg_chicago
df_load_stg_nyc
Power Query 0
Templates 0
pl_tsv_to_pqt_austin
pl_tsv_to_pqt_chica...
pl_tsv_to_pqt_nyc
df_load_stg_austin
Saved
Validate
Data flow debug
sourceSrc
Columns:
43 total
DerivedColumns
Creating/updating the columns
ID, CRASH_ID,
CRASH_FATAL_FL, CASE_ID,
PRIMARY_ADDRESS,
LoadToLnd
Export data to ds_lnd_austin
Add Source
Source settings
Source options
Projection
Optimize
Inspect
Data preview
Output stream name *
sourceSrc
Learn more
Description
Import data from austin_acc_src
Reset
Source type *
Dataset
Inline
Dataset *
austin_acc_src
Test connection
Open
New


Microsoft Azure
Data Factory
adf-damg7370fall2024
Search factory and documentation
jawadwala.a@northeastern.edu
NORTHEASTERN UNIVERSITY
aamir_final_project branch
Validate all
Save all
Publish
Preview experience
Off
Factory Resources
Filter resources by name
Load_To_Breed_Dim
Load_To_Chicago_Stage
Load_To_Location_Dim
Load_To_NYC_Stage
pl_tsv_to_parquet
pl_tsv_to_pqt_austin
pl_tsv_to_pqt_chicago
pl_tsv_to_pqt_nyc
Change Data Capture (preview) 0
Datasets 25
Data flows 5
df_load_breed_dim
df_load_location_dim
df_load_stg_austin
df_load_stg_chicago
df_load_stg_nyc
Power Query 0
Templates 0
df_load_stg_nyc
Saved
Validate
Data flow debug
sourceSrc
Columns:
29 total
DerivedColumns
Creating/updating the columns
'CRASH_DATE, CRASH_TIME,
BOROUGH, ZIP_CODE,
LATITUDE, LONGITUDE,
LoadToLnd
Export data to ds_lnd_nyc
Add Source
Source settings
Source options
Projection
Optimize
Inspect
Data preview
Output stream name *
sourceSrc
Learn more
Description
Import data from nyc_acc_src
Reset
Source type *
Dataset
Inline
Dataset *
nyc_acc_src
Test connection
Open
New
Options
Allow schema drift


Microsoft Azure
Data Factory
adf-damg7370fall2024
Search factory and documentation
jawadwala.a@northeastern.edu
NORTHEASTERN UNIVERSITY
aamir_final_project branch
Validate all
Save all
Publish
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 9
Load_To_Austin_Stage
Load_To_Breed_Dim
Load_To_Chicago_Stage
Load_To_Location_Dim
Load_To_NYC_Stage
pl_tsv_to_parquet
pl_tsv_to_pqt_austin
pl_tsv_to_pqt_chicago
pl_tsv_to_pqt_nyc
Change Data Capture (preview) 0
Datasets 25
Data flows 5
df_load_breed_dim
df_load_location_dim
df_load_stg_austin
df_load_stg_chicago
df_load_stg_nyc
Load_To_Austin_St...
Saved
Save as template
Validate
Debug
Add trigger
Data flow debug
Data flow
Load Landing
Austin
Expand toolbox pane
General
Settings
Parameters
User properties
Name *
Load Landing Austin
Learn more
Description
Activity state
Activated
Deactivated
Timeout
0.00:30:00
Retry
0


Microsoft Azure
Data Factory
adf-damg7370fall2024
Search factory and documentation
jawadwala.a@northeastern.edu
NORTHEASTERN UNIVERSITY
aamir_final_project branch
Validate all
Save all
Publish
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 9
Load_To_Austin_Stage
Load_To_Breed_Dim
Load_To_Chicago_Stage
Load_To_Location_Dim
Load_To_NYC_Stage
pl_tsv_to_parquet
pl_tsv_to_pqt_austin
pl_tsv_to_pqt_chicago
pl_tsv_to_pqt_nyc
Change Data Capture (preview) 0
Datasets 25
Data flows 5
df_load_breed_dim
df_load_location_dim
df_load_stg_austin
df_load_stg_chicago
df_load_stg_nyc
Load_To_Chicago_...
Saved
Save as template
Validate
Debug
Add trigger
Data flow debug
Data flow
Load Landing Chicago
General
Settings
Parameters
User properties
Name *
Load Landing Chicago
Learn more
Description
Activity state
Activated
Deactivated
Timeout
0.00:30:00
Retry
0


Microsoft Azure
Data Factory
adf-damg7370fall2024
Search factory and documentation
jawadwala.a@northeastern.edu
NORTHEASTERN UNIVERSITY
aamir_final_project branch
Validate all
Save all
Publish
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 9
Load_To_Austin_Stage
Load_To_Breed_Dim
Load_To_Chicago_Stage
Load_To_Location_Dim
Load_To_NYC_Stage
pl_tsv_to_parquet
pl_tsv_to_pqt_austin
pl_tsv_to_pqt_chicago
pl_tsv_to_pqt_nyc
Change Data Capture (preview) 0
Datasets 25
Data flows 5
df_load_breed_dim
df_load_location_dim
df_load_stg_austin
df_load_stg_chicago
df_load_stg_nyc
Load_To_NYC_Stage
Saved
Save as template
Validate
Debug
Add trigger
Data flow debug
Data flow
Load Landing NYC
Expand toolbox pane
General
Settings
Parameters
User properties
Name *
Load Landing NYC
Learn more
Description
Activity state
Activated
Deactivated
Timeout
0.00:30:00
Retry
0

TALEND IMPLEMENTATION

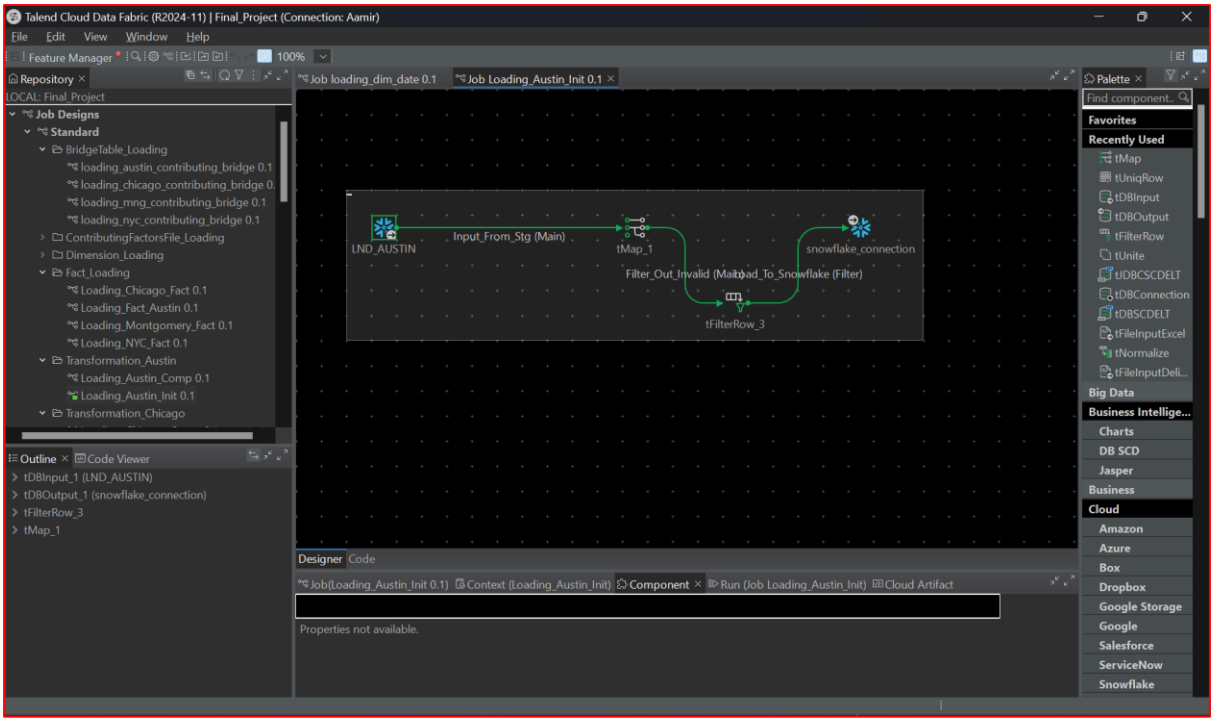
- **AUSTIN INIT STAGE LOAD**

Overview:

This Talend workflow processes the initial Austin data load, performing essential data transformations and filtering before loading the cleaned and normalized data into Snowflake. It ensures data quality and prepares it for integration into downstream analytical systems.

Key Workflow Components:

1. Input Source:
    - STG_AUSTIN_INIT:
        - Raw data from the Austin staging environment, including accident details, contributing factors, and associated entities.
2. Data Transformation:
    - tMap_1:
        - Maps fields to align the raw data with the normalized schema.
        - Applies transformations to split concatenated fields, rename columns, or adjust formats as needed.
3. Data Normalization:
    - Normalize_1:
        - Standardizes data formats and resolves any nested or inconsistent data structures.
        - Breaks down complex fields into structured, individual attributes.
4. Data Mapping for Loading:
    - tMap_2:
        - Prepares the normalized data for final insertion into the Snowflake table.
        - Maps each field to its respective column in the target schema.
5. Data Loading:
    - tDBOutput_1:
        - Loads the cleaned and normalized data into Snowflake for further use in analytics and reporting.



- **AUSTIN COMP STAGE LOAD**

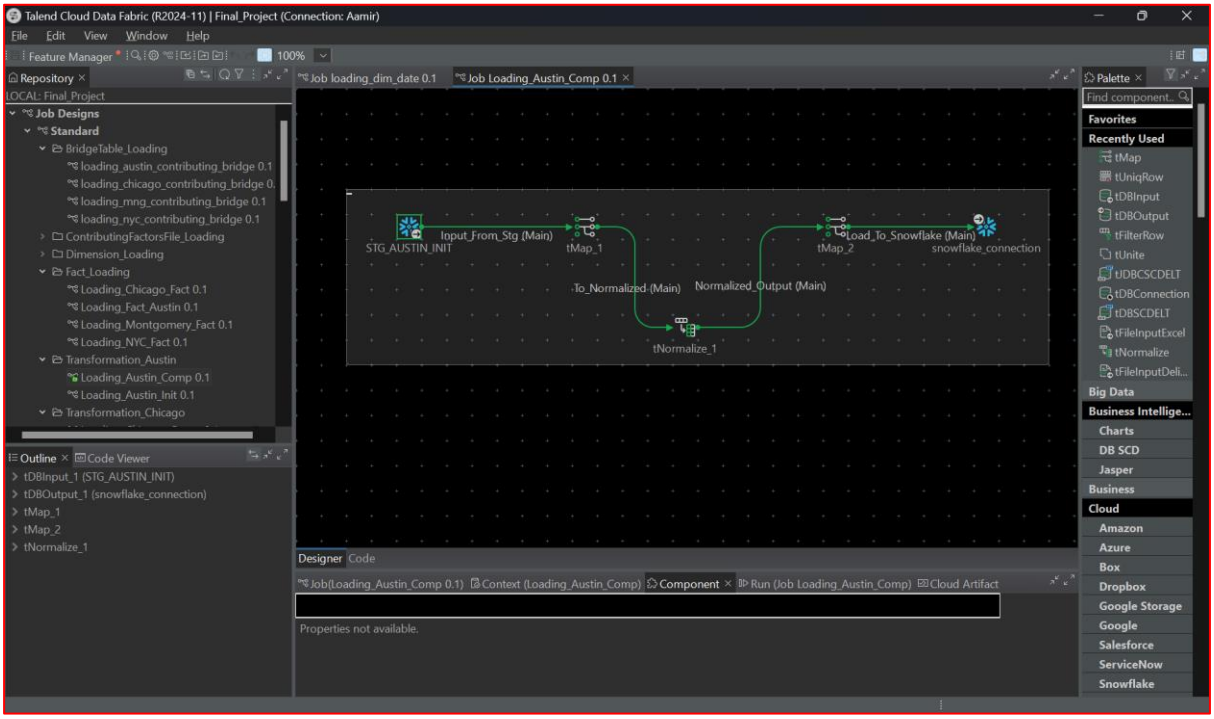
Overview:

The Austin Complete Stage Load workflow processes cleaned and partially normalized accident data for Austin, performs final transformations, and loads it into Snowflake. This ensures data readiness for integration into analytical pipelines and supports detailed reporting.

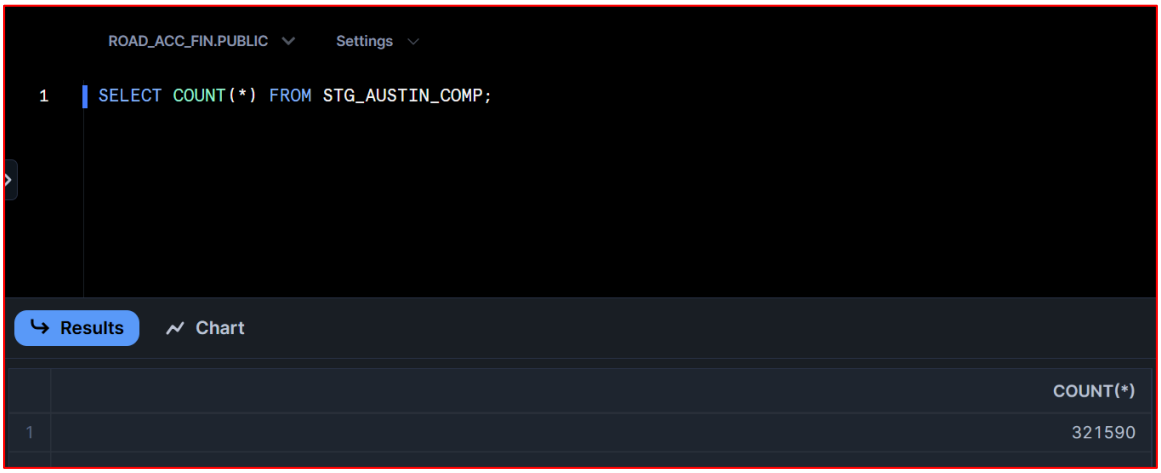
Key Workflow Components:

1. Input Source:
   - STG_AUSTIN_INIT:
     - Input data from the cleaned initial staging table for Austin.
2. Data Transformation:
   - tMap_1:
     - Maps fields from the initial staging table to the target schema.
     - Applies additional transformations, such as column renaming and value standardization.
3. Normalization Process:
   - Normalize_1:
     - Further normalizes the data for consistency, such as:
       - Splitting concatenated fields (e.g., combining make and model into separate columns).
       - Standardizing date and numeric formats.
4. Data Mapping for Final Load:
   - tMap_2:
     - Prepares the transformed data for insertion into Snowflake, aligning with the final schema.
5. Data Loading:

- tDBOutput_1:
  - Loads the transformed and normalized data into the Austin Complete Stage Table in Snowflake.



- **SQL QUERY**



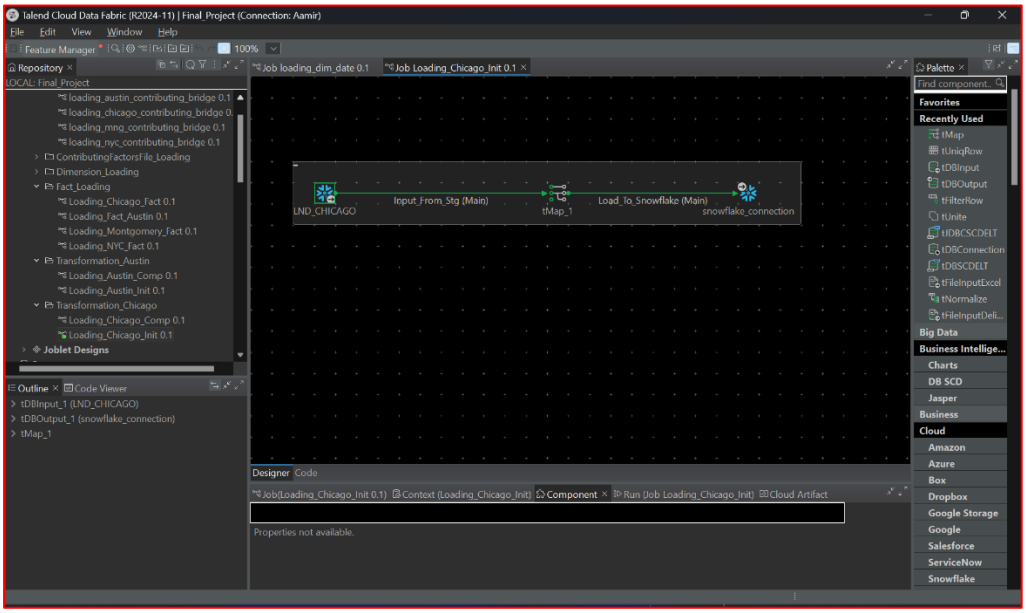
- **CHICAGO INIT STAGE LOAD**

Overview:

The Chicago Initial Stage Load workflow ingests raw accident data for Chicago from the landing zone, performs initial transformations, and loads the data into Snowflake. This process ensures the data is clean and ready for further processing and integration.

Key Workflow Components:

1. Input Source:
    - LND_CHICAGO:
        - Raw accident data from the landing zone containing attributes such as accident details, location, vehicles involved, and contributing factors.
2. Data Transformation:
    - tMap_1:
        - Maps the fields in the raw dataset to the initial staging schema.
        - Standardizes formats and handles minor field-level transformations such as renaming or reformatting date and numeric values.
3. Data Loading:
    - tDBOutput_1:
        - Loads the transformed data into Snowflake's Chicago Initial Stage Table, preserving data integrity and ensuring readiness for normalization and deeper transformations.



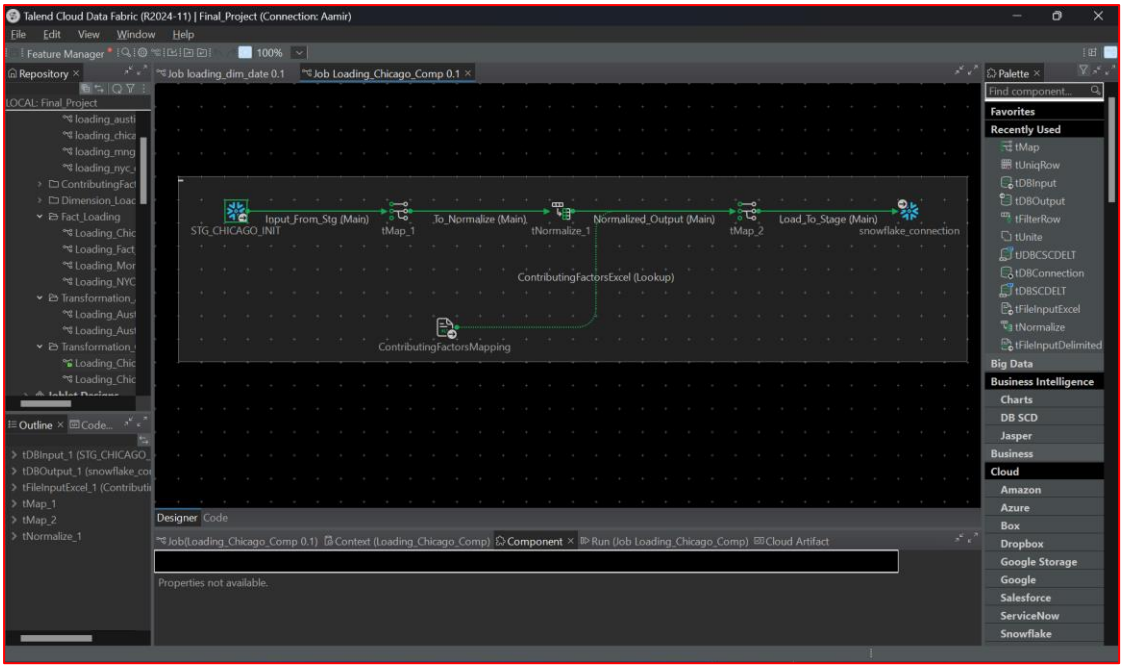
- **CHICAGO COMP STAGE LOAD**

  Overview:

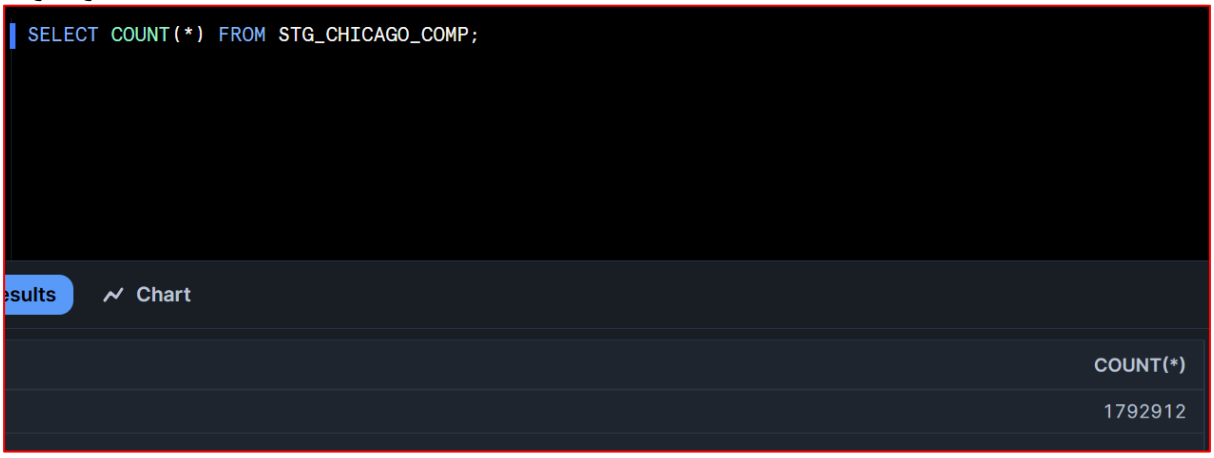
  This workflow finalizes the processing of Chicago accident data by normalizing, mapping, and loading it into the Chicago Complete Stage Table in Snowflake. It ensures the data is ready for analytics and adheres to the defined schema.

  Key Workflow Components:

  1. Input Source:
     - STG_CHICAGO_INIT:
       - Cleaned and validated data from the initial staging process.
  2. Data Transformation:
     - tMap_1:
       - Maps the staging data to the normalized schema.
       - Aligns fields and performs additional transformations, such as column standardization.
  3. Data Normalization:
     - Normalize_1:
       - Finalizes the normalization process:
         - Splits nested fields into individual attributes.
         - Standardizes values (e.g., ensuring consistency across vehicle details or contributing factors).
  4. Data Enrichment:
     - ContributingFactorsMapping:
       - Uses an external Excel file or lookup table to map contributing factors to standardized values.
       - Ensures consistency and accuracy in contributing factor details.
  5. Data Mapping for Final Stage:
     - tMap_2:
       - Prepares the normalized and enriched data for loading into Snowflake.
       - Maps data to the complete stage schema.
  6. Data Loading:
     - tDBOutput_1:
       - Loads the finalized data into Snowflake's Chicago Complete Stage Table.



SQL QUERY



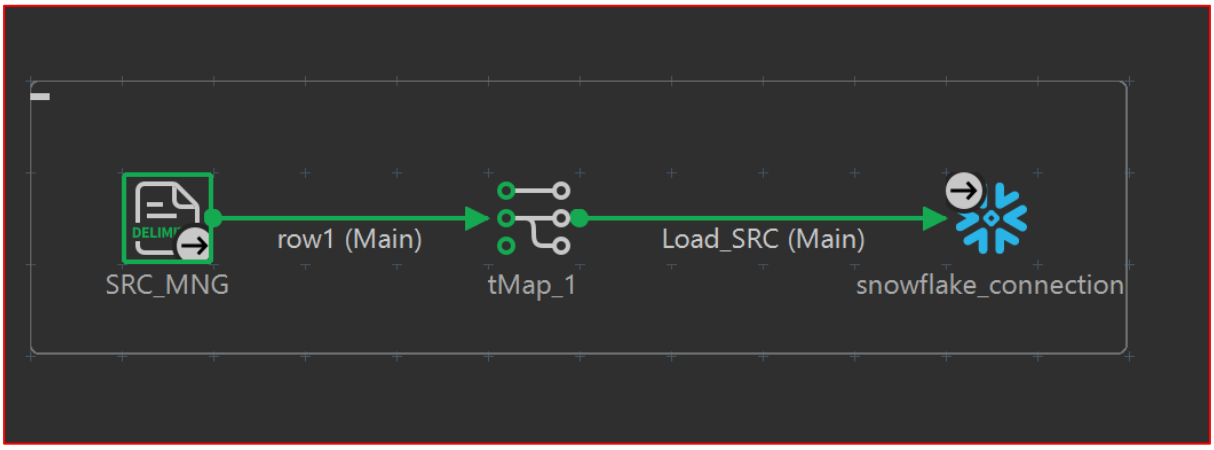
- **MONTGOMERY INIT STAGE LOAD**

  The Montgomery Initial Stage Load workflow ingests raw accident data for Montgomery from a delimited source file, applies essential transformations, and loads the processed data into the Montgomery Initial Stage Table in Snowflake. This step ensures raw data is prepared for further transformations and analysis.

  Key Workflow Components:

  1. Input Source:
     - SRC_MNG:
       - Raw accident data in a delimited file format sourced from Montgomery's data pipeline.
  2. Data Transformation:
     - row1 (Main):
       - Reads the raw data row by row for transformation.
     - tMap_1:

        - Maps raw input fields to the staging schema.
        - Performs basic transformations, such as:
            - Standardizing date formats.
            - Aligning field names with Snowflake schema requirements.
3. Data Loading:
    - Load_SRC:
        - Inserts the transformed data into the Montgomery Initial Stage Table in Snowflake.



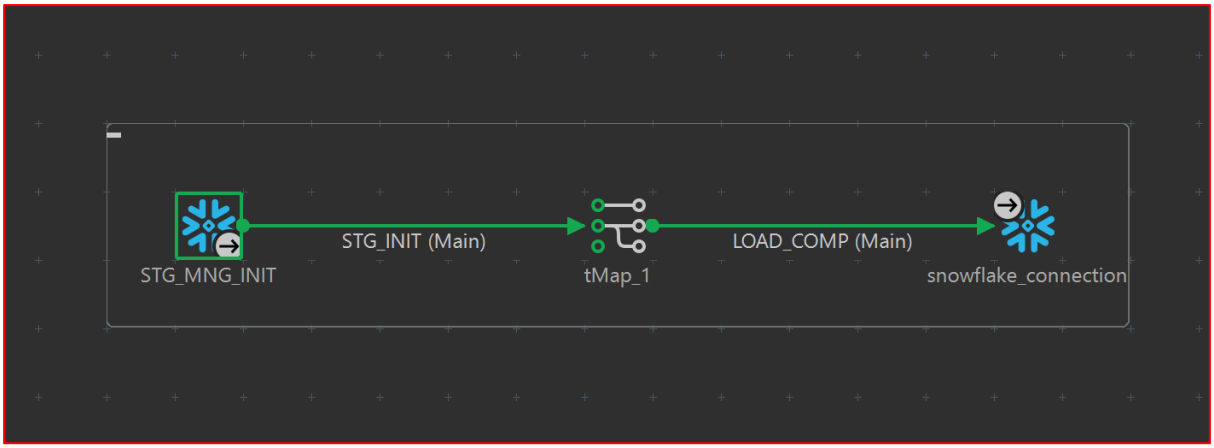
- **MONTGOMERY COMP STAGE LOAD**

Overview:

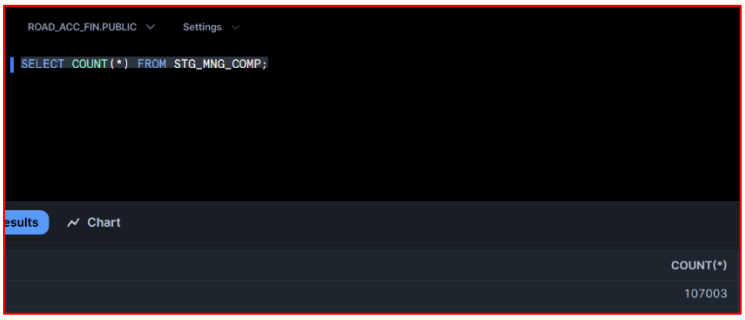
The Montgomery Complete Stage Load workflow processes data from the initial staging table, performs additional transformations, and loads the finalized data into the Montgomery Complete Stage Table in Snowflake. This step ensures the data is fully normalized and ready for analytical use.

Key Workflow Components:

1. Input Source:
    - STG_MNG_INIT:
        - Data from the initial Montgomery staging table after initial validation and cleaning.
2. Data Transformation:
    - tMap_1:
        - Maps fields from the initial staging schema to the complete stage schema.
        - Applies advanced transformations, such as:
            - Deriving additional attributes.
            - Normalizing field values (e.g., standardizing date formats, aligning contributing factors).
3. Data Loading:
    - LOAD_COMP:
        - Inserts the transformed and enriched data into the Montgomery Complete Stage Table in Snowflake.



- **SQL QUERY**

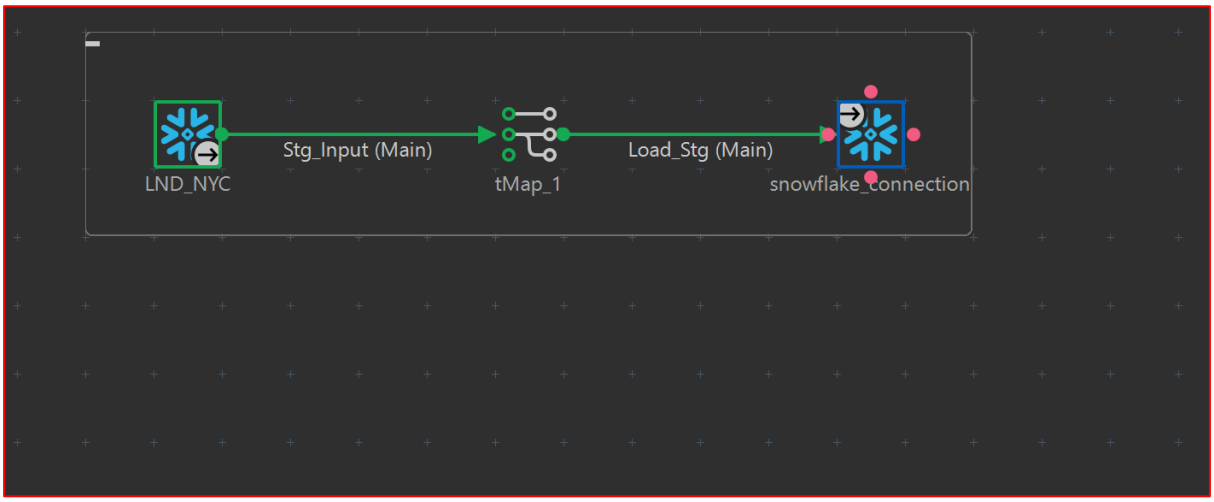


- **NYC STAGE INIT LOAD**

Overview:
The NYC Initial Stage Load workflow ingests raw accident data for New York City, performs basic transformations, and loads the processed data into the NYC Initial Stage Table in Snowflake. This step ensures raw data is prepared for further processing while maintaining data integrity.

Key Workflow Components:

1. Input Source:
   - LND_NYC:
     - Raw accident data from the landing zone, containing fields such as accident details, location, contributing factors, and vehicle information.
2. Data Transformation:
   - tMap_1:
     - Maps the raw input fields to the staging table schema.
     - Applies initial transformations, such as:
       - Aligning field names.
       - Standardizing formats for dates, numeric values, and text.
3. Data Loading:
   - Load_Stg:
     - Inserts the transformed data into the NYC Initial Stage Table in Snowflake.



- **NYC STAGE DEV LOAD**

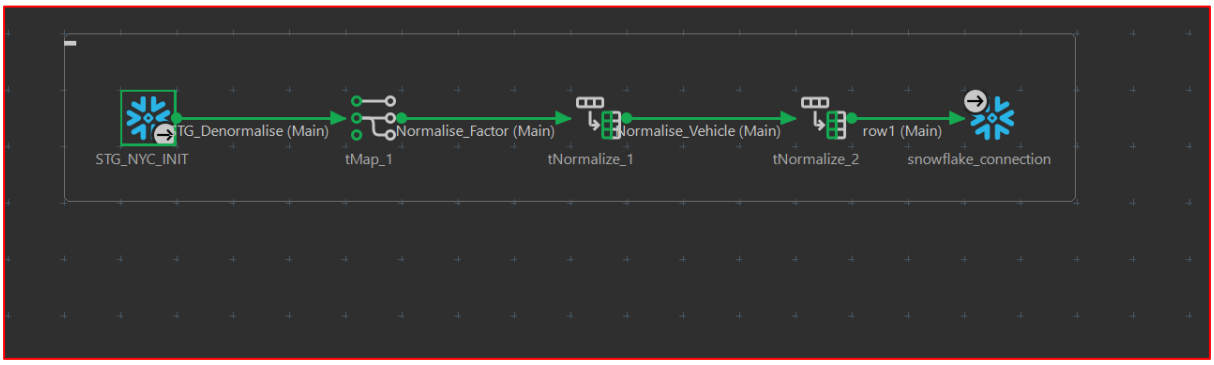
  Overview:

  The NYC Development Load workflow processes normalized data for New York City by further transforming and preparing it for advanced analytical use. This workflow integrates detailed normalization steps for contributing factors and vehicle data, ensuring consistency and completeness before loading the final output into Snowflake.

  Key Workflow Components:

  1. Input Source:
     - STG_NYC_INIT:
       - Initial staging table containing cleaned and partially normalized NYC accident data.
  2. Data Transformation:
     - tMap_1:
       - Maps denormalized input data to the development schema.
       - Organizes data fields for further normalization.
  3. Normalization Steps:
     - tNormalize_1 (Normalize Factors):
       - Extracts and normalizes contributing factors.
       - Splits multiple factors into individual rows for analytical flexibility.
     - tNormalize_2 (Normalize Vehicles):
       - Processes vehicle data by separating make, model, and year.
       - Ensures consistent formatting for vehicle-related attributes.
  4. Data Aggregation:
     - row1:
       - Consolidates normalized and enriched data into the final schema for Snowflake.
  5. Data Loading:
     - snowflake_connection:

- Loads the fully normalized and transformed data into Snowflake's NYC Development Table.



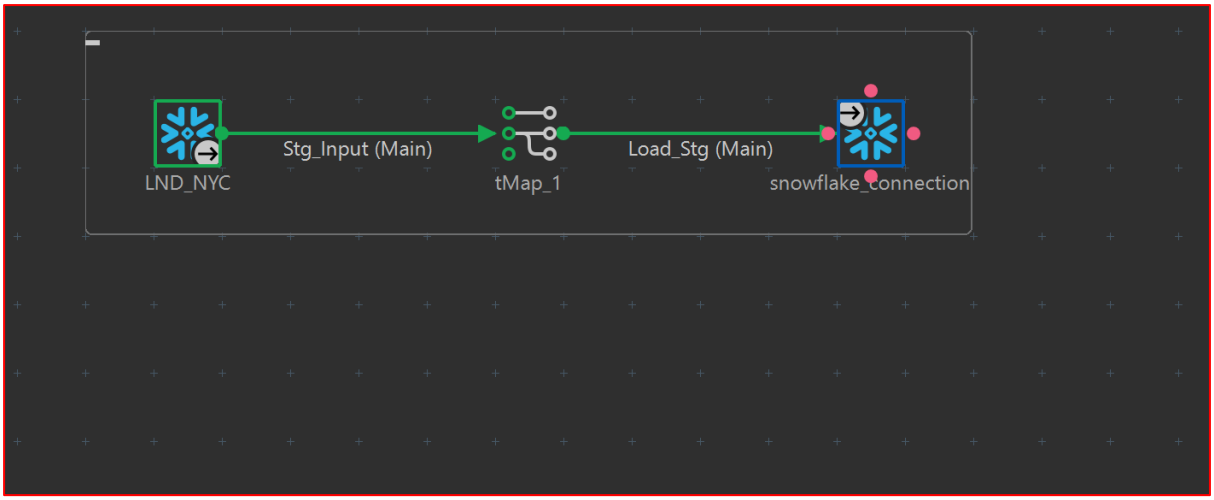
- **NYC STAGE COMP LOAD**

Overview:

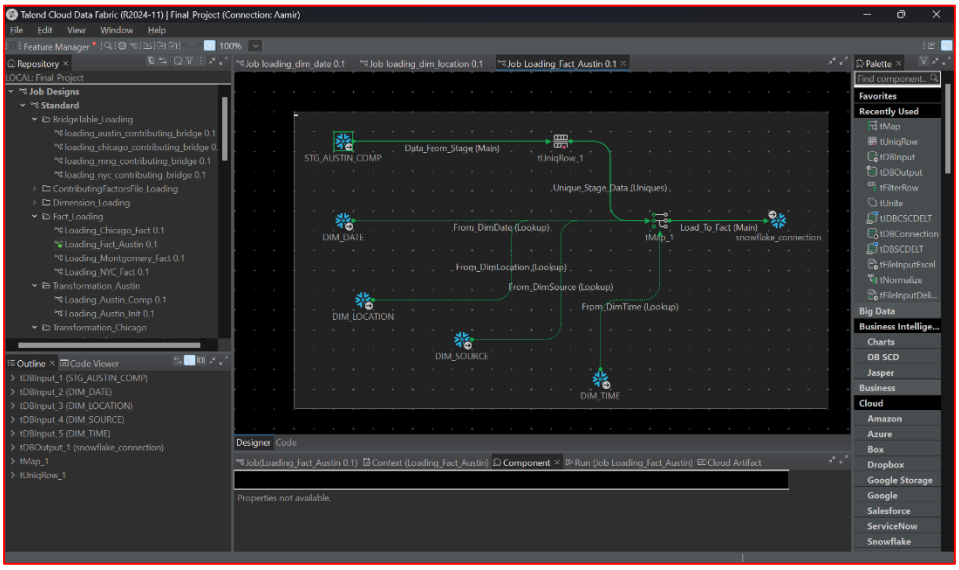
The NYC Complete Stage Load workflow processes data from the initial NYC staging table, performs additional transformations, and loads the normalized and structured data into the NYC Complete Stage Table in Snowflake. This ensures the data is ready for comprehensive analysis and reporting.

Key Workflow Components:

1. Input Source:
   - LND_NYC:
     - Data from the NYC landing zone, containing raw accident records.
2. Data Transformation:
   - tMap_1:
     - Maps input fields to align with the target schema of the complete stage table.
     - Applies any necessary transformations to ensure schema consistency.
3. Data Loading:
   - Load_Stg:
     - Loads the transformed data into Snowflake's NYC Complete Stage Table

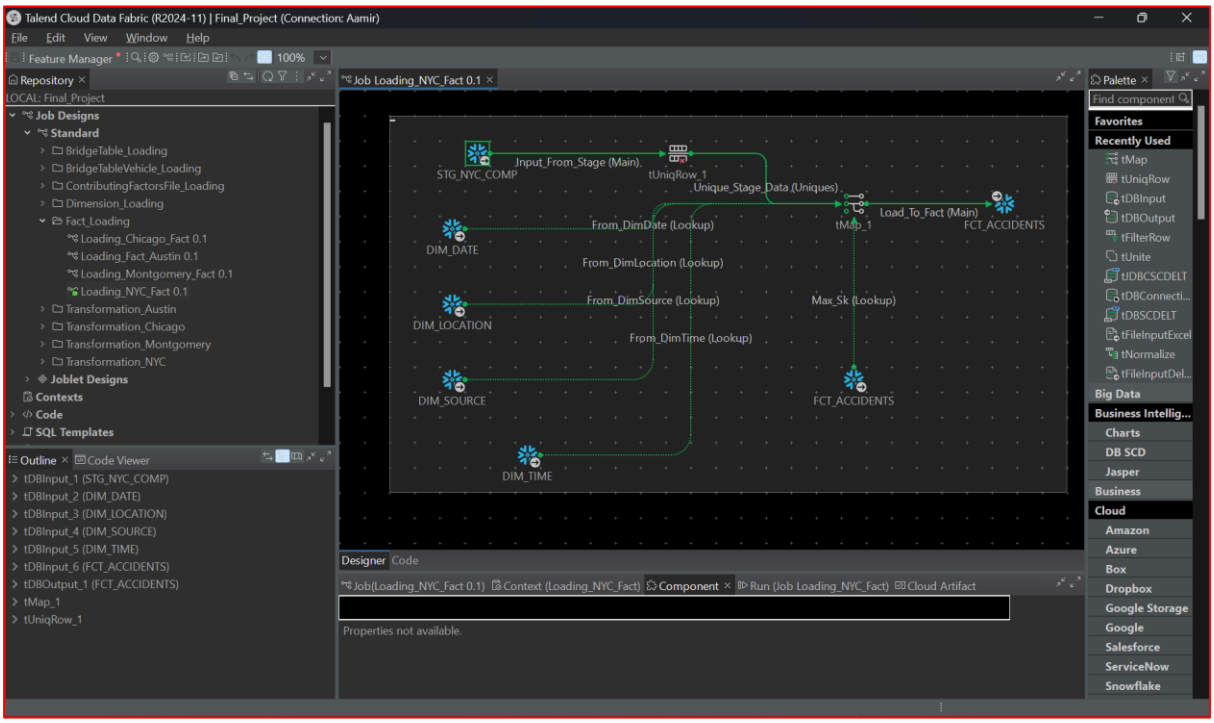


- **AUSTIN FACT LOAD**
  - **Overview:** This process loads data from the staging area (STG_AUSTIN_COMP) into the fact table (FCT_ACCIDENTS) by joining with dimension tables for date (DIM_DATE), location (DIM_LOCATION), source (DIM_SOURCE), and time (DIM_TIME).
  - **Key Steps:**
    - Data Extraction: Data is read from the staging table.
    - Deduplication: Unique rows are identified using the UniqRow component.
    - Dimension Lookups:
      Match keys are fetched from dimensions using lookups (From_DimDate, From_DimLocation, From_DimSource, From_DimTime).
    - Data Transformation: The transformation component (tMap) maps staging fields to target fact table fields.
    - Data Loading: The Load_To_Fact component loads the transformed data into FCT_ACCIDENTS using Snowflake.

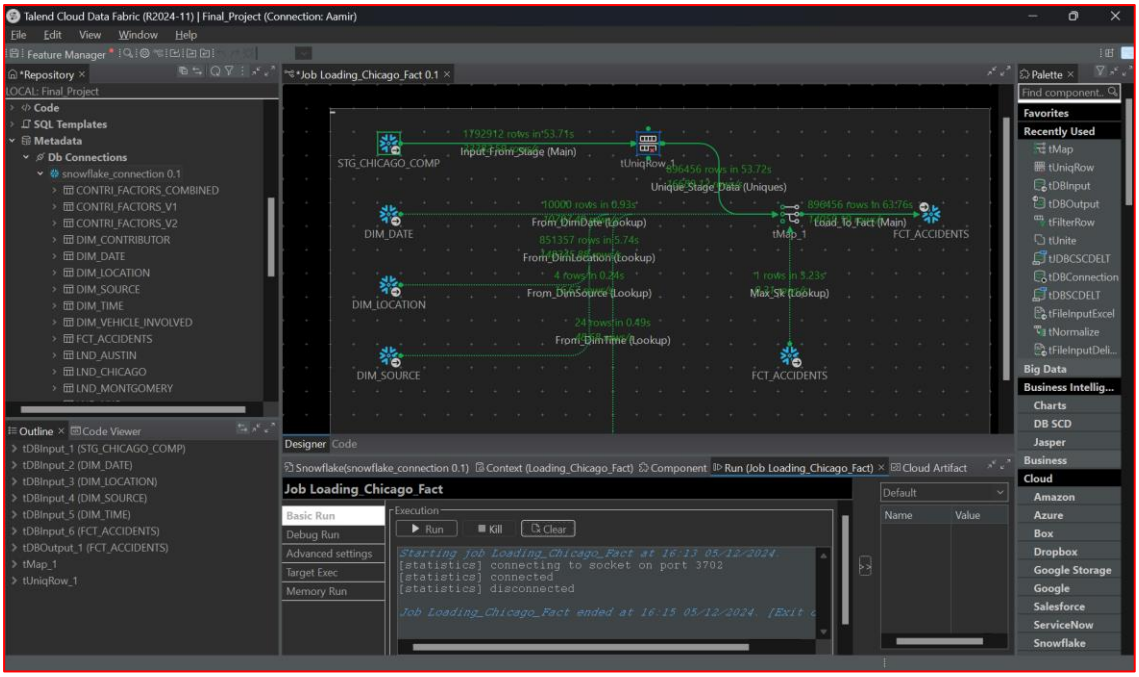


- **NYC FACT LOAD**
  - **Overview:**
    Like Austin, the process ingests data from STG_NYC_COMP into FCT_ACCIDENTS while performing lookups on dimension tables.

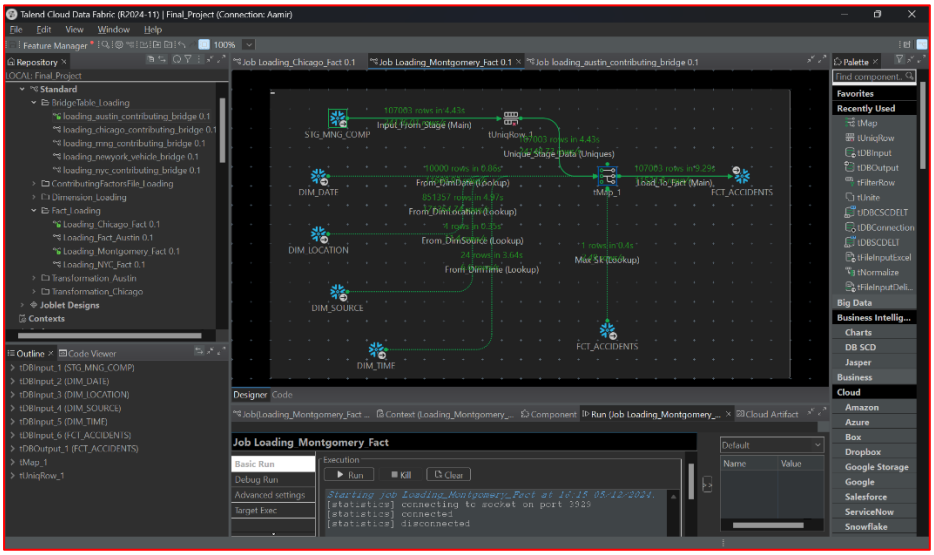
  - **Key Steps:**
    - Data Cleansing: Unique rows from the staging table are extracted.
    - Dimension Integration: Lookups from DIM_DATE, DIM_LOCATION, DIM_SOURCE, and DIM_TIME ensure foreign key relationships.
    - Aggregation/Key Matching:
    - The Max_Sk lookup ensures the highest key consistency in sequential loads.
    - Fact Table Load: Transformed data is inserted into FCT_ACCIDENTS.



- **CHICAGO FACT LOAD**
  - **Overview:** The job extracts Chicago crash data from STG_CHICAGO_COMP and loads it into FCT_ACCIDENTS.
  - **Key Steps:**
    - Row Deduplication: The UniqRow component ensures no duplicate entries are processed.
    - Lookups: Dimensions (DIM_DATE, DIM_LOCATION, DIM_SOURCE, DIM_TIME) enrich the data for fact table integration.
    - Data Mapping: Transformations using tMap ensure correct alignment of source columns with the fact table schema.
    - Data Validation and Load: Final data is loaded into the fact table after validation via Snowflake connections.



- **MONTGOMERY FACT LOAD**
- **Overview:** This job integrates Montgomery crash data from STG_MNG_COMP to FCT_ACCIDENTS.
- **Key Steps:**
  - Extraction and Deduplication: Staging data is processed to remove duplicates.
  - Foreign Key Population: Dimension tables (DIM_DATE, DIM_LOCATION, DIM_SOURCE, DIM_TIME) are used to populate foreign keys.
  - Transformation and Load:
    - Data fields are mapped to fact table fields via tMap.
    - Processed data is inserted into FCT_ACCIDENTS.

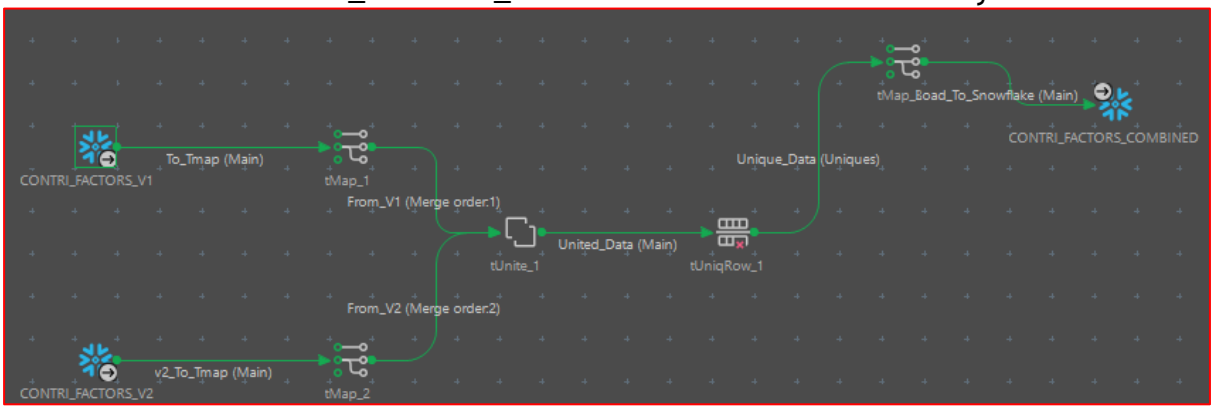


- **DIM COMBINED CONTRIBUTIONS**

Overview:
This Talend workflow integrates two datasets of contributing factors, CONTRI_FACTORS_V1 and CONTRI_FACTORS_V2, into a unified dimension table CONTRI_FACTORS_COMBINED. The process ensures data consolidation, transformation, deduplication, and loading into Snowflake for further use.

Key Workflow Components:

1. Input Sources:
   - CONTRI_FACTORS_V1 and CONTRI_FACTORS_V2: Two different datasets representing contributing factors from various sources.
2. Data Transformation:
   - tMap_1 and tMap_2:
     - Standardize the schema and format of both datasets to align with the target combined table structure.
     - Map relevant fields from the input datasets to a unified schema.
3. Data Consolidation:
   - tUnite_1:
     - Merges the outputs of the two datasets (From_V1 and From_V2) into a single data stream.
     - Maintains the source integrity while preparing data for deduplication.
4. Data Deduplication:
   - tUniqRow_1:
     - Ensures that only unique records are retained in the final dataset.
     - Removes duplicate rows based on specified keys (e.g., factor ID, description).
5. Data Mapping and Loading:
   - tMap_Board_To_Snowflake:
     - Maps the cleaned and consolidated data to the schema of the CONTRI_FACTORS_COMBINED table.
     - Prepares data for final insertion into Snowflake.
   - Snowflake Integration:
     - The combined and deduplicated data is loaded into the CONTRI_FACTORS_COMBINED table for further analysis.



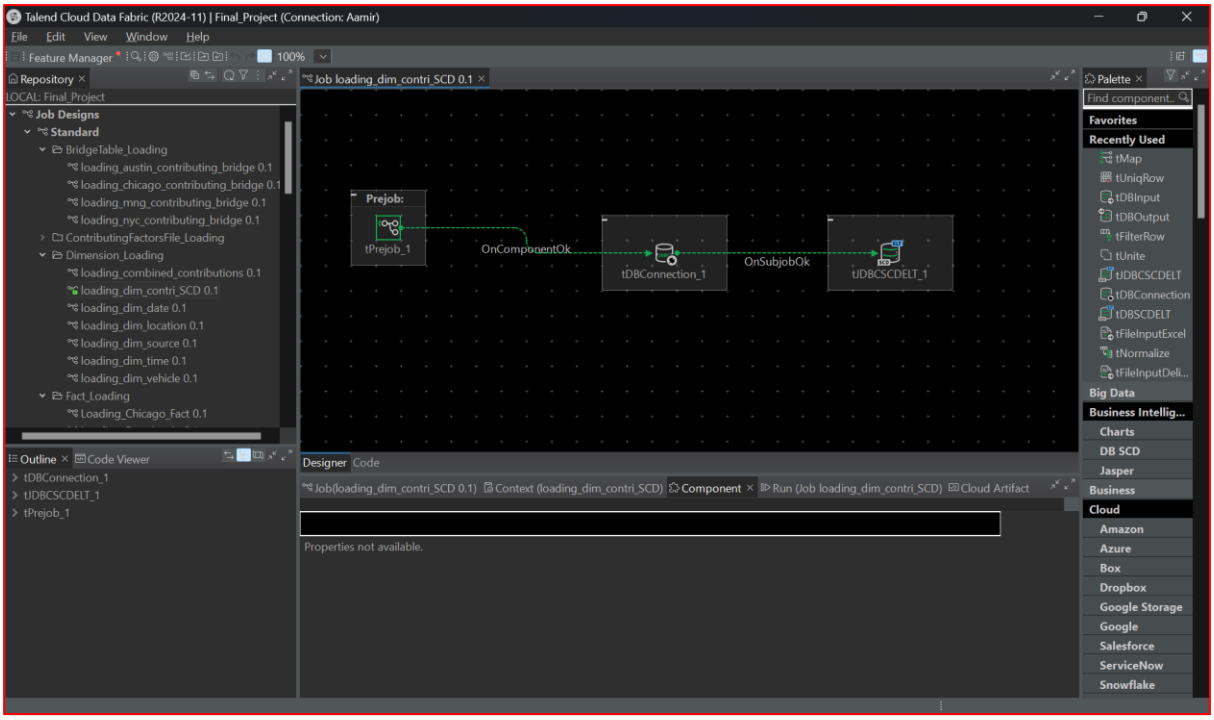
- **DIM CONTRIBUTING SCD**

  Overview:

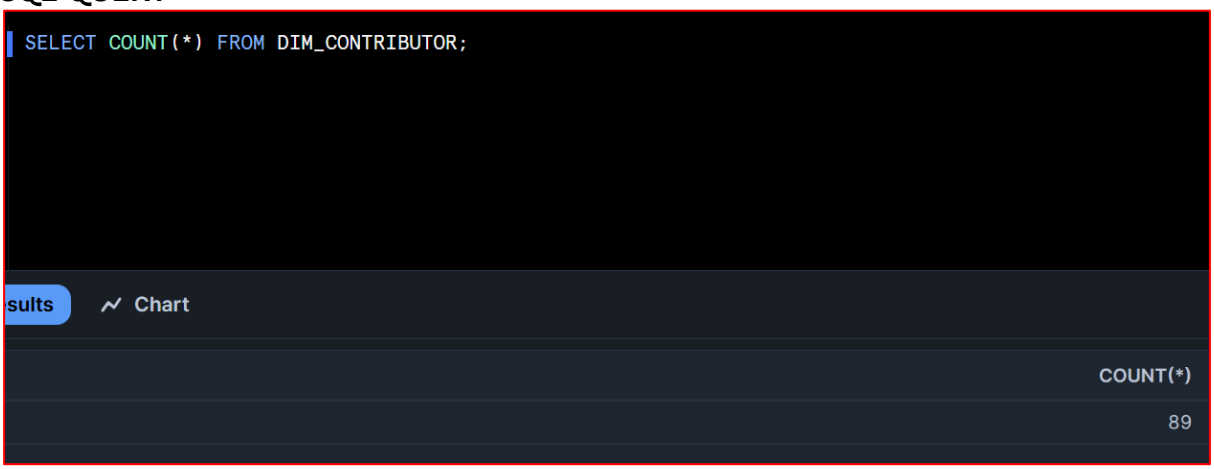
  This Talend workflow handles the loading of the DIM_CONTRIBUTING table, implementing Slowly Changing Dimension (SCD) techniques to manage changes in contributing factors data over time. SCD allows historical data tracking while maintaining current data for analysis.

  Key Workflow Components:

  1. Prejob Initialization:
     - tPrejob_1:
       - Sets up the environment and connections required for the workflow.
       - Ensures database connections and configurations are initialized before proceeding to the main loading process.
  2. Database Connection:
     - tDBConnection_1:
       - Establishes a connection to the target database (e.g., Snowflake).
       - Enables subsequent components to interact with the database for SCD operations.
  3. SCD Implementation:
     - tDBSCDelt_1:
       - Manages the SCD logic for the DIM_CONTRIBUTING table.
       - Compares incoming data with existing records in the dimension table to identify:
         - New Records: Inserts new rows for contributing factors not already present in the table.
         - Updated Records: Updates existing rows if any attribute values change. Historical data is retained, and the existing record is flagged as inactive.
         - Unchanged Records: Maintains records without modification if no changes are detected.



SQL QUERY



- **DIM DATE LOAD**

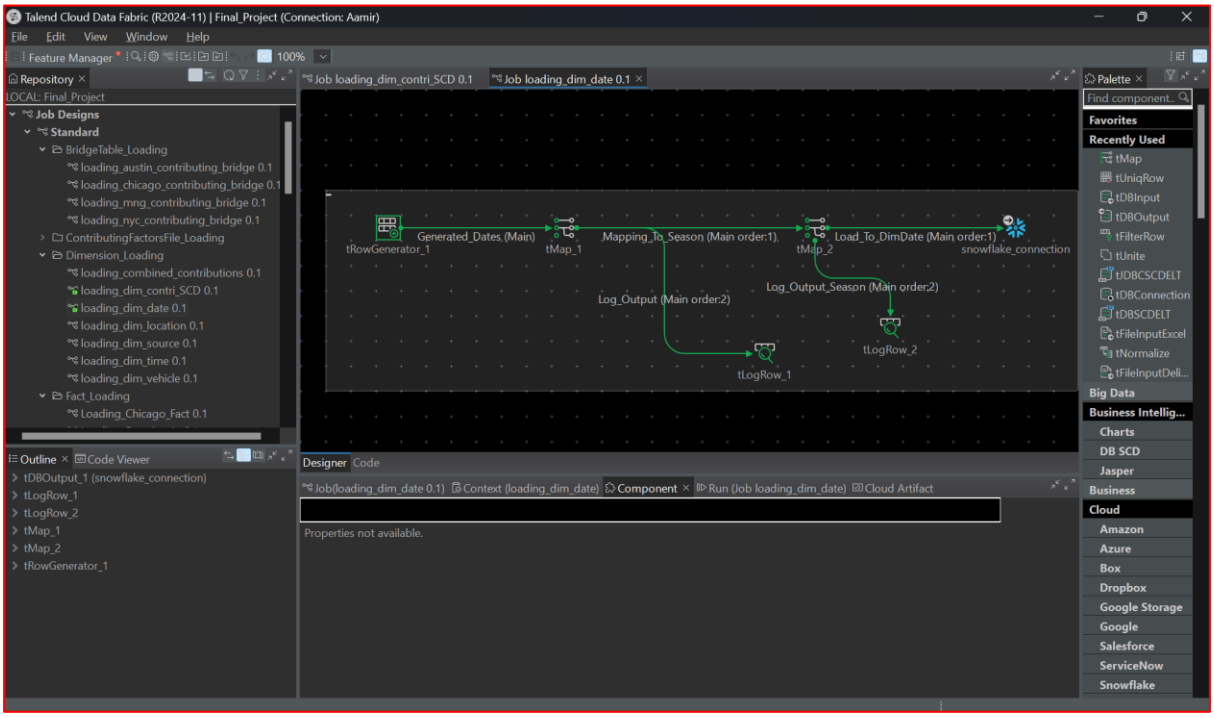
Overview:
This Talend workflow handles the loading process for the DIM_DATE table, which is critical for supporting temporal analysis in the data warehouse. The workflow generates a comprehensive set of dates, maps them to seasonal attributes, and loads the processed data into Snowflake.

Key Workflow Components:

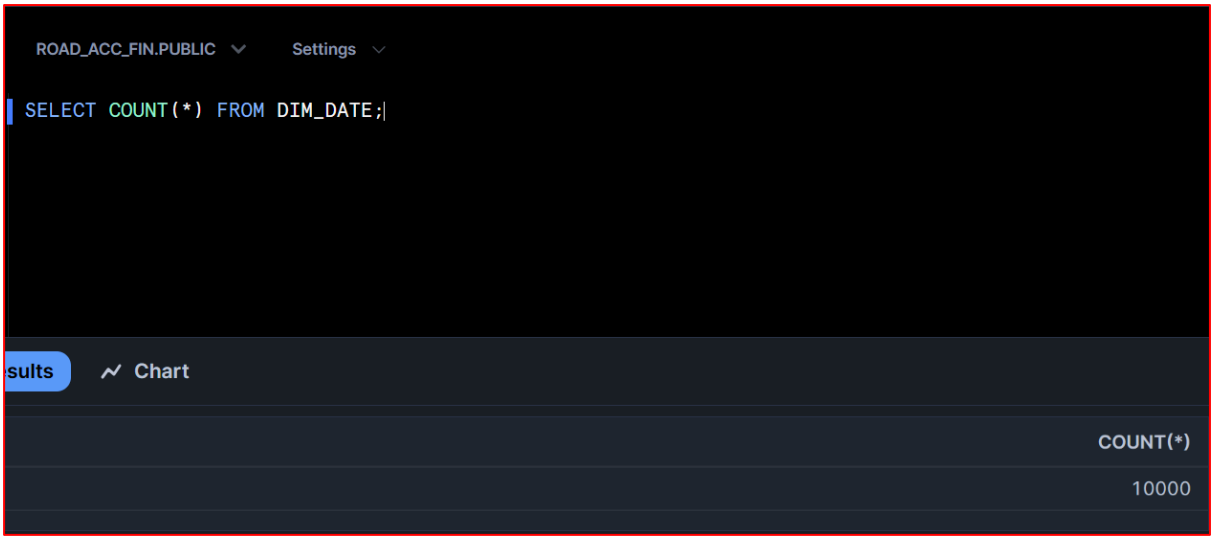
1. Date Generation:
    - tRowGenerator_1:

- Generates a continuous range of dates starting from a predefined start date (e.g., 2020-01-01) to an end date (e.g., 2030-12-31).
- Outputs a dataset with basic date fields such as date_value and day_of_week.

2. Mapping to Seasons:
   - tMap_1:
     - Enriches the generated dates with seasonal attributes (e.g., winter, spring, summer, fall) based on date ranges.
     - Maps additional attributes such as is_weekend, quarter, and fiscal_year.
3. Log Outputs:
   - tLogRow_1 and tLogRow_2:
     - Log intermediate results for debugging and verification:
       - tLogRow_1: Logs the raw date generation output.
       - tLogRow_2: Logs the final enriched dataset before loading.
4. Data Loading:
   - tDBOutput_1:
     - Loads the processed and enriched data into the DIM_DATE table in Snowflake.
     - Ensures that the dimensional table is populated with a robust and well-structured date dataset.



- **SQL QUERY**



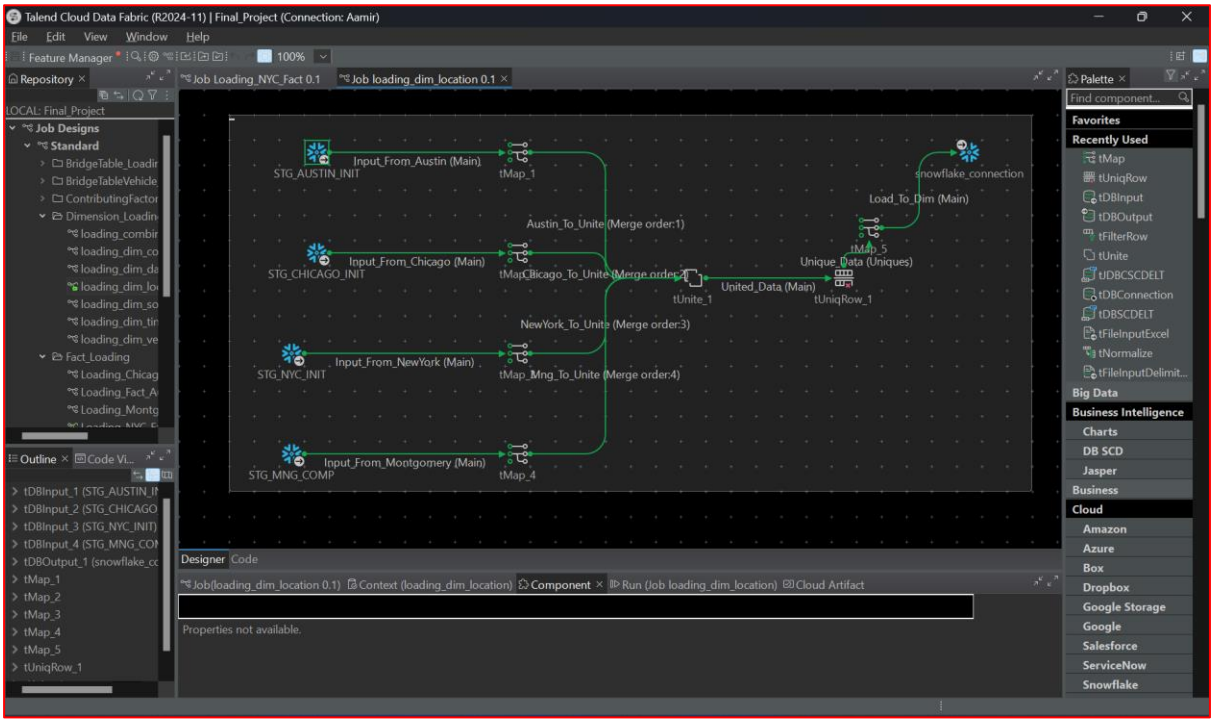
- **DIM LOCATION**

  Overview:
  This Talend workflow consolidates location data from multiple cities (Austin, Chicago, New York, and Montgomery) into a unified DIM_LOCATION table. The workflow ensures deduplication, proper transformation, and loading into Snowflake for analytical purposes.

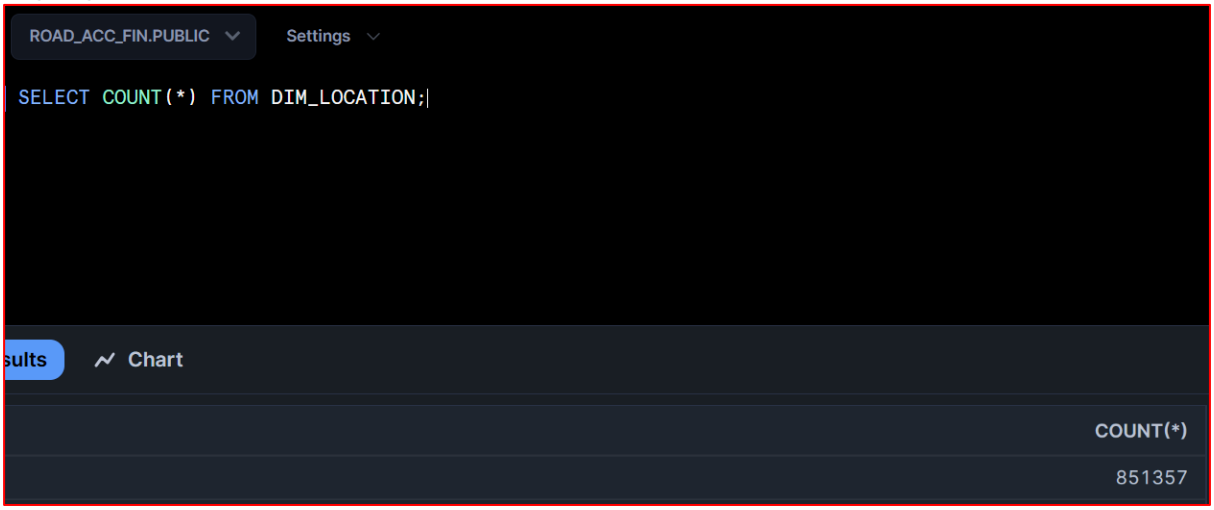
  Key Workflow Components:
  1. Input Sources:
     - STG_AUSTIN_INIT: Location data from Austin.
     - STG_CHICAGO_INIT: Location data from Chicago.
     - STG_NYC_INIT: Location data from New York.
     - STG_MNG_COMP: Location data from Montgomery.
  2. Data Transformation:
     - tMap_1 to tMap_4:
       - Standardize and map location fields (e.g., city, state, zip_code) from individual staging datasets to a unified format.
       - Prepare data for merging into a single dataset.
  3. Data Consolidation:
     - tUnite_1:
       - Combines the standardized datasets from Austin, Chicago, New York, and Montgomery into a single dataset (United_Data).
       - Maintains source-specific integrity during the merge process.
  4. Data Deduplication:
     - tUniqRow_1:
       - Identifies and removes duplicate rows from the consolidated dataset based on unique keys (e.g., location_id, city, state).
  5. Data Loading:
     - tMap_5:

- Maps the deduplicated dataset to the schema of the DIM_LOCATION table.

- tDBOutput_1:
  - Loads the processed and deduplicated data into the Snowflake DIM_LOCATION table.



- **SQL QUERY**

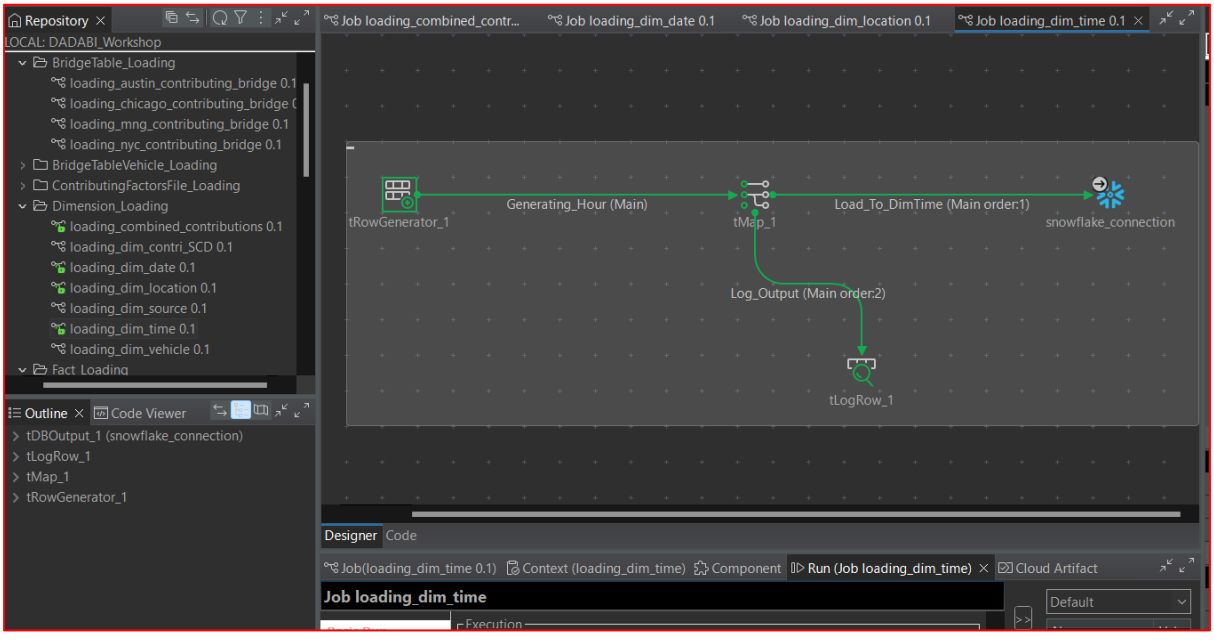


- **DIM TIME**

Overview:
This Talend workflow is designed to populate the DIM_TIME table, which serves as a critical component for temporal analysis in the data warehouse. The workflow generates hourly time intervals, processes the data to include relevant time attributes, and loads it into Snowflake.

Key Workflow Components:

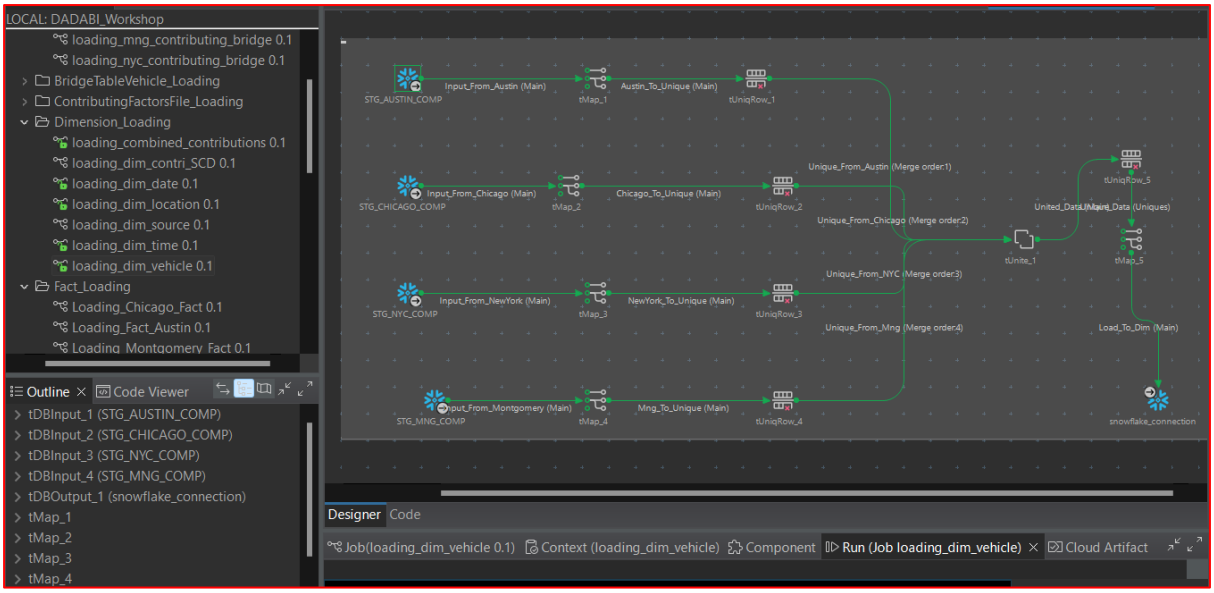
1. Time Generation:
   - tRowGenerator_1:
     - Generates a dataset containing time values in hourly intervals (e.g., 00:00, 01:00, ..., 23:00).
     - Creates a sequence of hours for a complete 24-hour cycle.
2. Data Transformation:
   - tMap_1:
     - Enriches the generated time data with additional attributes such as:
       - hour: The hour of the day in 24-hour format.
       - is_am_pm: Indicates whether the time is AM or PM.
       - time_period: Categorizes the time into periods such as morning, afternoon, evening, or night.
3. Log Output:
   - tLogRow_1:
     - Logs the enriched time data for debugging and verification.
4. Data Loading:
   - tDBOutput_1:
     - Loads the processed time data into the DIM_TIME table in Snowflake.
     - Ensures the table is populated with accurate and comprehensive time-related attributes.



- **DIM VEHICLE**

  Overview:
  This Talend workflow consolidates vehicle data from multiple cities (Austin, Chicago, New York, and Montgomery) into a unified DIM_VEHICLE table. The process includes data transformation, deduplication, and loading into Snowflake, ensuring clean and comprehensive vehicle-related data.

  Key Workflow Components:
  1. Input Sources:
     - STG_AUSTIN_COMP: Vehicle data from Austin.
     - STG_CHICAGO_COMP: Vehicle data from Chicago.
     - STG_NYC_COMP: Vehicle data from New York.
     - STG_MNG_COMP: Vehicle data from Montgomery.
  2. Data Transformation:
     - tMap_1 to tMap_4:
       - Standardizes and maps vehicle fields (e.g., vehicle_type, manufacturer, model_year) from individual city datasets to a unified schema.
  3. Data Deduplication:
     - tUniqRow_1 to tUniqRow_4:
       - Ensures unique vehicle data is retained within each city dataset before merging.
  4. Data Consolidation:
     - tUnite_1:
       - Combines deduplicated datasets from all cities (Unique_From_Austin, Unique_From_Chicago, Unique_From_NYC, Unique_From_Montgomery) into a single dataset (United_Data).
  5. Final Deduplication:
     - tUniqRow_5:
       - Removes duplicates from the consolidated dataset to ensure the final data is clean and unique.
  6. Data Loading:
     - tMap_5:
       - Maps the consolidated and deduplicated data to the schema of the DIM_VEHICLE table.
     - tDBOutput_1:
       - Loads the processed data into the DIM_VEHICLE table in Snowflake.



- **DIM SOURCE LOAD**

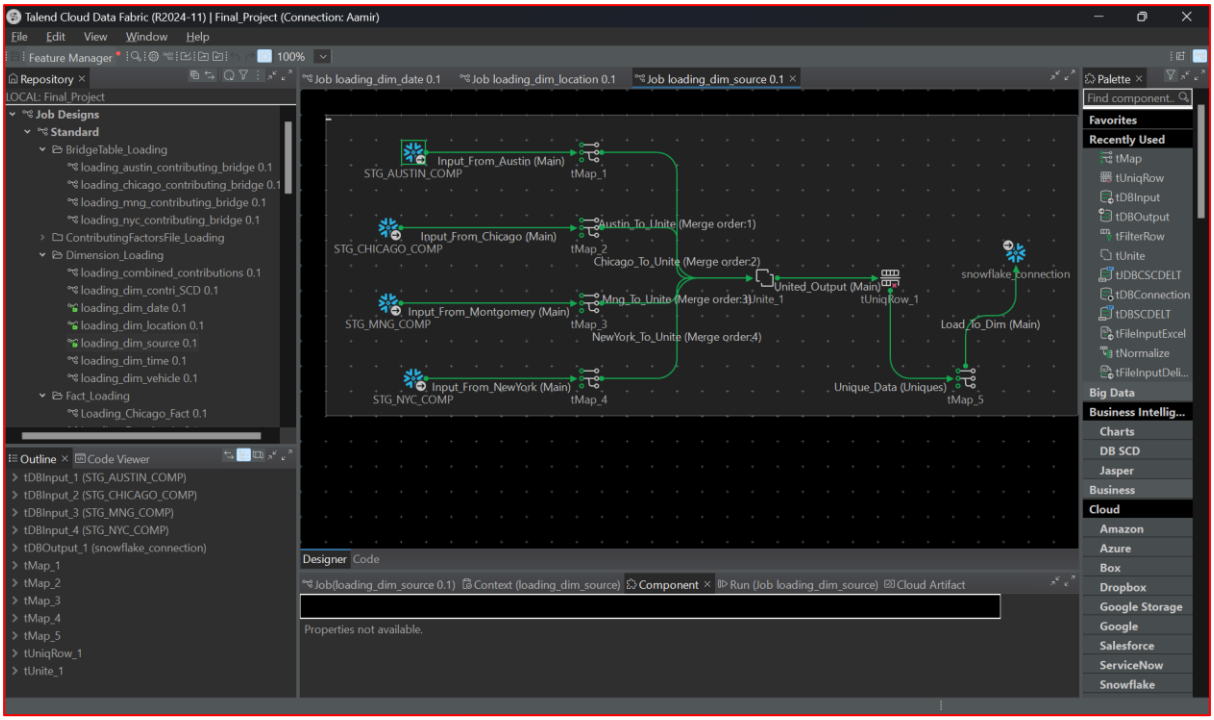
Overview:
This Talend workflow loads data into the DIM_SOURCE table, which stores information about data sources from multiple cities (Austin, Chicago, New York, and Montgomery). The workflow includes standardization, consolidation, deduplication, and loading into Snowflake for efficient querying and integration with fact tables.

Key Workflow Components:

1. Input Sources:
   - STG_AUSTIN_COMP: Source data from Austin.
   - STG_CHICAGO_COMP: Source data from Chicago.
   - STG_MNG_COMP: Source data from Montgomery.
   - STG_NYC_COMP: Source data from New York.
2. Data Transformation:
   - tMap_1 to tMap_4:
     - Standardizes and maps source attributes such as source_name, source_type, and description from each staging dataset to a common schema.
3. Data Consolidation:
   - tUnite_1:
     - Combines the standardized outputs from all cities (Austin_To_Unite, Chicago_To_Unite, Mongomery_To_Unite, NewYork_To_Unite) into a single dataset.
4. Data Deduplication:
   - tUniqRow_1:
     - Removes duplicate rows based on unique keys (e.g., source_id or source_name) to ensure the data is clean and non-redundant.
5. Data Loading:
   - tMap_5:

- Maps the deduplicated data to the final schema of the DIM_SOURCE table.

- tDBOutput_1:
  - Loads the processed data into the Snowflake DIM_SOURCE table.



- **AUSTIN CONTRIBUTING BRIDGE LOAD**

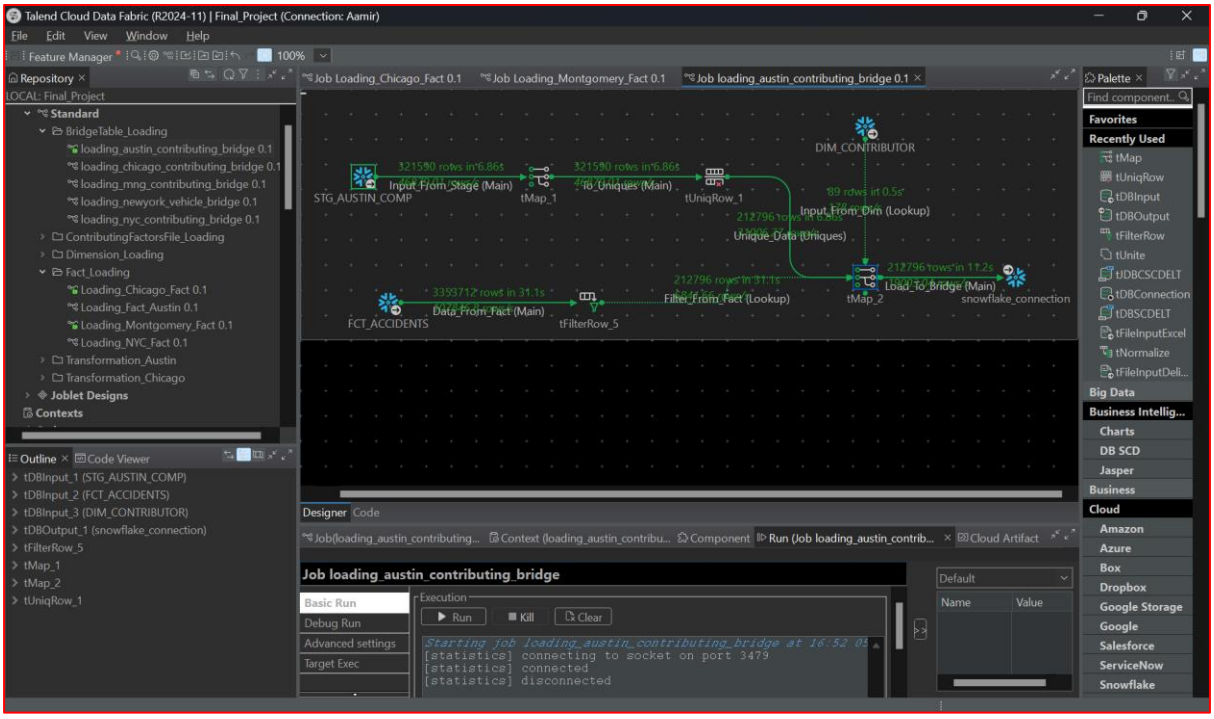
Overview:
This Talend workflow loads data into the Austin Contributing Bridge Table, which links the fact table (FCT_ACCIDENTS) and the dimension table (DIM_CONTRIBUTOR) for the Austin dataset. The bridge table enables many-to-many relationships between accidents and contributing factors, ensuring a robust analytical model.

Key Workflow Components:

1. Input Sources:
   - STG_AUSTIN_COMP: Staging data for Austin, containing raw accident and contributing factor details.
   - FCT_ACCIDENTS: Fact table with accident records.
   - DIM_CONTRIBUTOR: Dimension table containing contributing factors.
2. Data Transformation:
   - tMap_1:
     - Filters and maps relevant fields from the Austin staging data to align with the bridge table schema.
     - Ensures data consistency by aligning IDs from fact and dimension tables.
3. Data Deduplication:

- tUniqRow_1:
  - Ensures unique relationships between accident records and contributing factors.
  - Removes duplicate combinations of accident_id and contributor_id.

4. Fact and Dimension Lookups:
   - tMap_2:
     - Links accident records from FCT_ACCIDENTS with contributing factors from DIM_CONTRIBUTOR using lookups.
     - Populates foreign keys required for the bridge table.
5. Data Loading:
   - tDBOutput_1:
     - Loads the processed data into the Austin-specific Contributing Bridge Table in Snowflake.
     - Maintains many-to-many relationships between accidents and contributing factors.



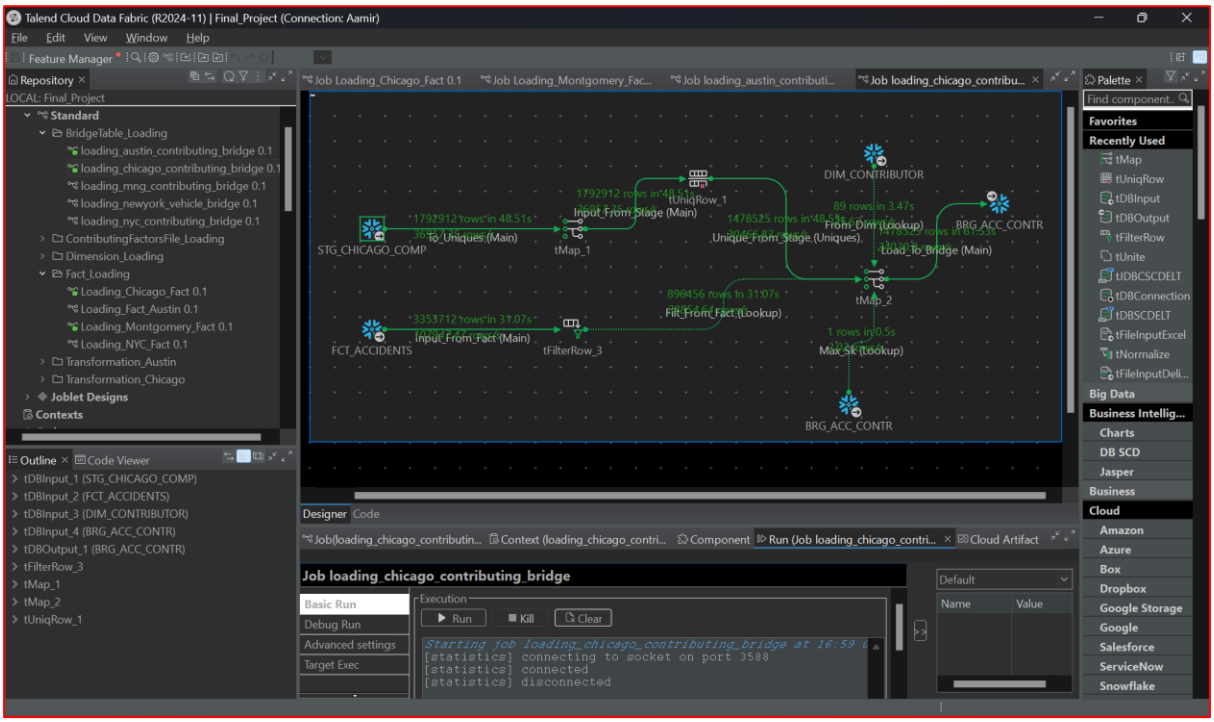
- **CHICAGO CONTRIBUTING BRIDGE LOAD**

Overview:
This Talend workflow loads data into the Chicago Contributing Bridge Table, linking the fact table (FCT_ACCIDENTS) with the dimension table (DIM_CONTRIBUTOR) for the Chicago dataset. The bridge table facilitates many-to-many relationships between accidents and contributing factors, enabling detailed analysis of factors influencing accidents.

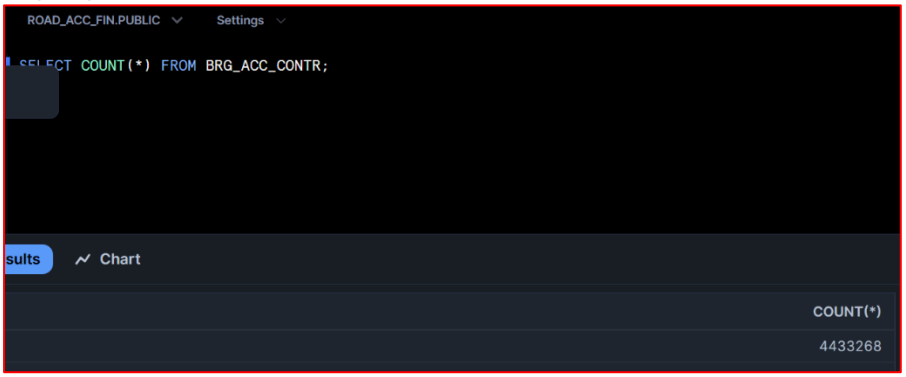
Key Workflow Components:
1. Input Sources:

- STG_CHICAGO_COMP: Staging data containing raw accident and contributing factor details for Chicago.
- FCT_ACCIDENTS: Fact table with detailed accident records.
- DIM_CONTRIBUTOR: Dimension table containing contributing factor details.

2. Data Transformation:
   - tMap_1:
     - Maps relevant fields from the Chicago staging data to align with the bridge table schema.
     - Standardizes data and ensures compatibility with the FCT_ACCIDENTS and DIM_CONTRIBUTOR tables.
3. Data Deduplication:
   - tUniqRow_1:
     - Removes duplicate relationships between accidents and contributing factors.
     - Ensures unique combinations of accident_id and contributor_id.
4. Fact and Dimension Lookups:
   - tMap_2:
     - Performs lookups on FCT_ACCIDENTS and DIM_CONTRIBUTOR to fetch primary and foreign keys.
     - Populates the bridge table with keys and additional fields such as contributor_role.
5. Data Loading:
   - tDBOutput_1:
     - Loads the processed and deduplicated data into the Chicago-specific Contributing Bridge Table in Snowflake.
     - Ensures a robust many-to-many relationship for analytical purposes.



SQL QUERY



- **MONTGOMERY CONTRIBUTING BRIDGE LOAD**

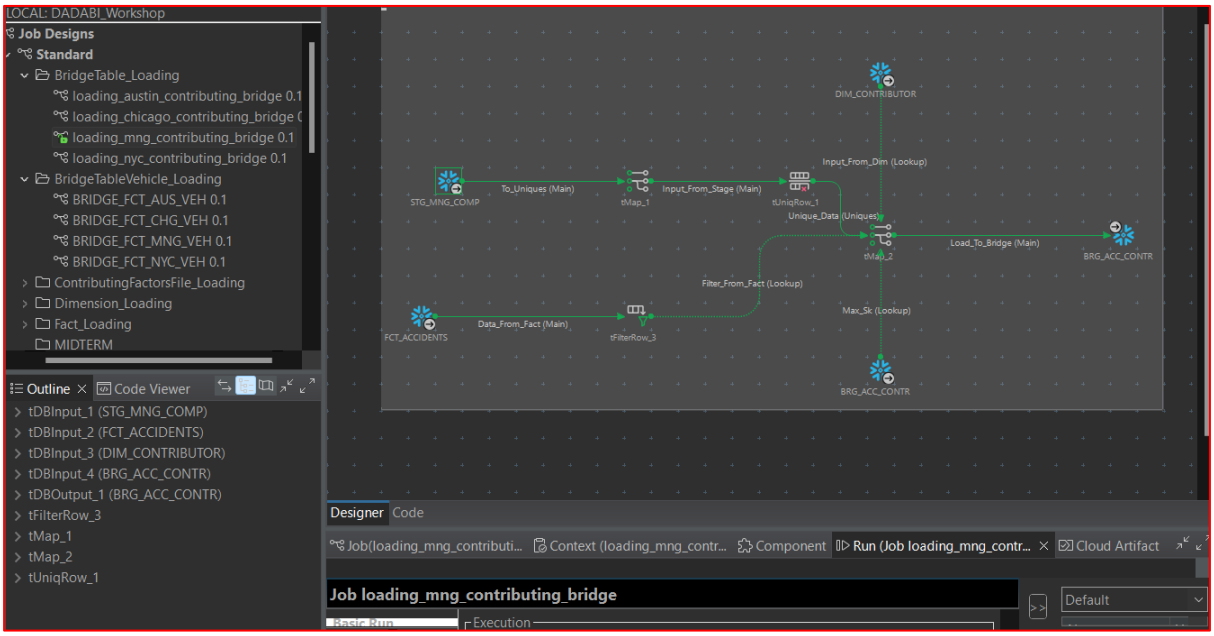
Overview:

This Talend workflow loads data into the Chicago Contributing Bridge Table, linking the fact table (FCT_ACCIDENTS) with the dimension table (DIM_CONTRIBUTOR) for the Chicago dataset. The bridge table facilitates many-to-many relationships between accidents and contributing factors, enabling detailed analysis of factors influencing accidents.

Key Workflow Components:

1. Input Sources:
   - STG_CHICAGO_COMP: Staging data containing raw accident and contributing factor details for Chicago.
   - FCT_ACCIDENTS: Fact table with detailed accident records.
   - DIM_CONTRIBUTOR: Dimension table containing contributing factor details.
2. Data Transformation:
   - tMap_1:
     - Maps relevant fields from the Chicago staging data to align with the bridge table schema.
     - Standardizes data and ensures compatibility with the FCT_ACCIDENTS and DIM_CONTRIBUTOR tables.
3. Data Deduplication:
   - tUniqRow_1:
     - Removes duplicate relationships between accidents and contributing factors.
     - Ensures unique combinations of accident_id and contributor_id.
4. Fact and Dimension Lookups:
   - tMap_2:
     - Performs lookups on FCT_ACCIDENTS and DIM_CONTRIBUTOR to fetch primary and foreign keys.
     - Populates the bridge table with keys and additional fields such as contributor_role.
5. Data Loading:
   - tDBOutput_1:

- Loads the processed and deduplicated data into the Chicago-specific Contributing Bridge Table in Snowflake.
- Ensures a robust many-to-many relationship for analytical purposes.

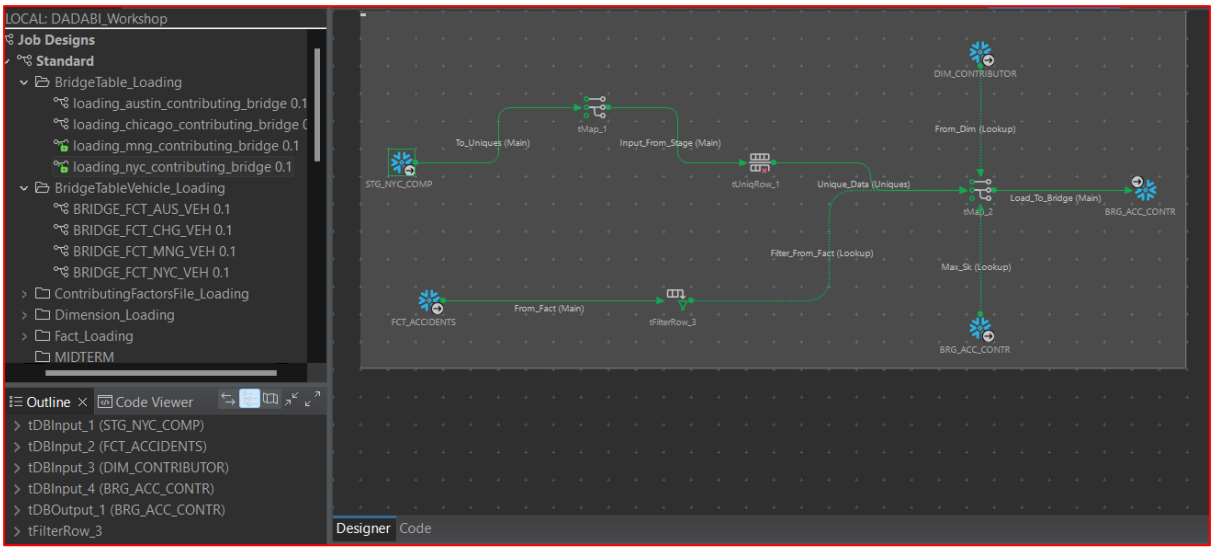


- **NYC CONTRIBUTING BRIDGE LOAD**

Overview:
This Talend workflow processes data for the NYC Contributing Bridge Table, linking the fact table (FCT_ACCIDENTS) and the dimension table (DIM_CONTRIBUTOR). The bridge table establishes many-to-many relationships between accidents and contributing factors for New York City, enabling advanced analytical capabilities.

Key Workflow Components:
1. Input Sources:
   - STG_NYC_COMP: Staging data containing raw accident and contributing factor details for NYC.
   - FCT_ACCIDENTS: Fact table with accident records for NYC.
   - DIM_CONTRIBUTOR: Dimension table containing details of contributing factors.
2. Data Transformation:
   - tMap_1:
     - Maps fields from the NYC staging dataset to align with the schema of the bridge table.
     - Prepares the data for subsequent lookups and deduplication.
3. Data Deduplication:
   - tUniqRow_1:
     - Removes duplicate relationships between accident_id and contributor_id.
     - Ensures that each accident-contributing factor relationship is unique.

4. Fact and Dimension Lookups:
    - tMap_2:
        - Links accident records from FCT_ACCIDENTS with contributing factors from DIM_CONTRIBUTOR using lookups.
        - Populates the bridge table with primary and foreign keys and supplementary attributes such as contributor_role.
5. Data Loading:
    - tDBOutput_1:
        - Loads the processed and deduplicated data into the NYC-specific Contributing Bridge Table in Snowflake.
        - Maintains robust many-to-many relationships for analysis.



- **BRIDGE FACT AUSTIN VEHICLE**

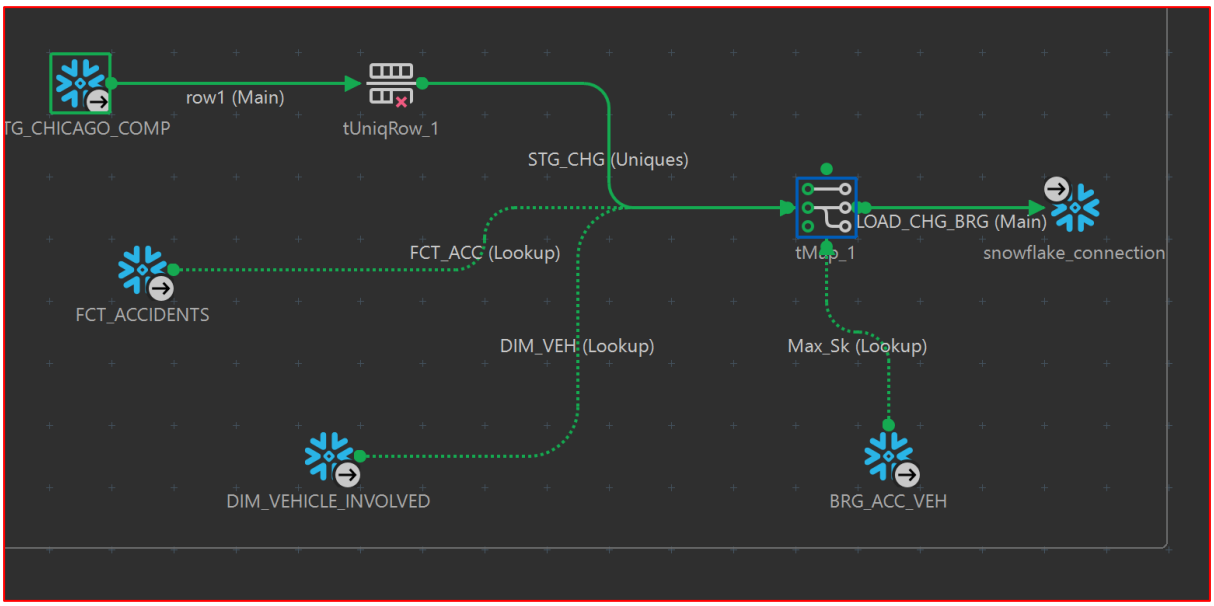
Overview:

This Talend workflow creates a Bridge Table for Austin Vehicle Involvement, which establishes many-to-many relationships between accidents in the FCT_ACCIDENTS table and vehicles in the DIM_VEHICLE table. This bridge table enables analysis of accidents involving multiple vehicles.

Key Workflow Components:

1. Input Sources:
    - STG_CHICAGO_COMP: Staging data containing accident and vehicle information.
    - FCT_ACCIDENTS: Fact table for accident records.
    - DIM_VEHICLE_INVOLVED: Dimension table containing details of vehicles involved in accidents.
2. Data Deduplication:
    - tUniqRow_1:
        - Removes duplicate rows from the staging data to ensure unique combinations of accidents and vehicles.
3. Fact and Dimension Lookups:
    - FCT_ACC:

        - Performs a lookup on the fact table to retrieve accident_id for each record.
    - DIM_VEH:
        - Performs a lookup on the vehicle dimension table to retrieve vehicle_id for each record.
4. Data Transformation:
    - tMap_1:
        - Maps the cleaned and enriched data to align with the bridge table schema.
        - Ensures relationships between accidents and vehicles are correctly represented.
5. Data Loading:
    - tDBOutput_1:
        - Loads the processed data into the Austin Vehicle Bridge Table in Snowflake.
        - Maintains many-to-many relationships between accidents and vehicles.



- **BRIDGE FACT CHICAGO VEHICLE**

Overview:

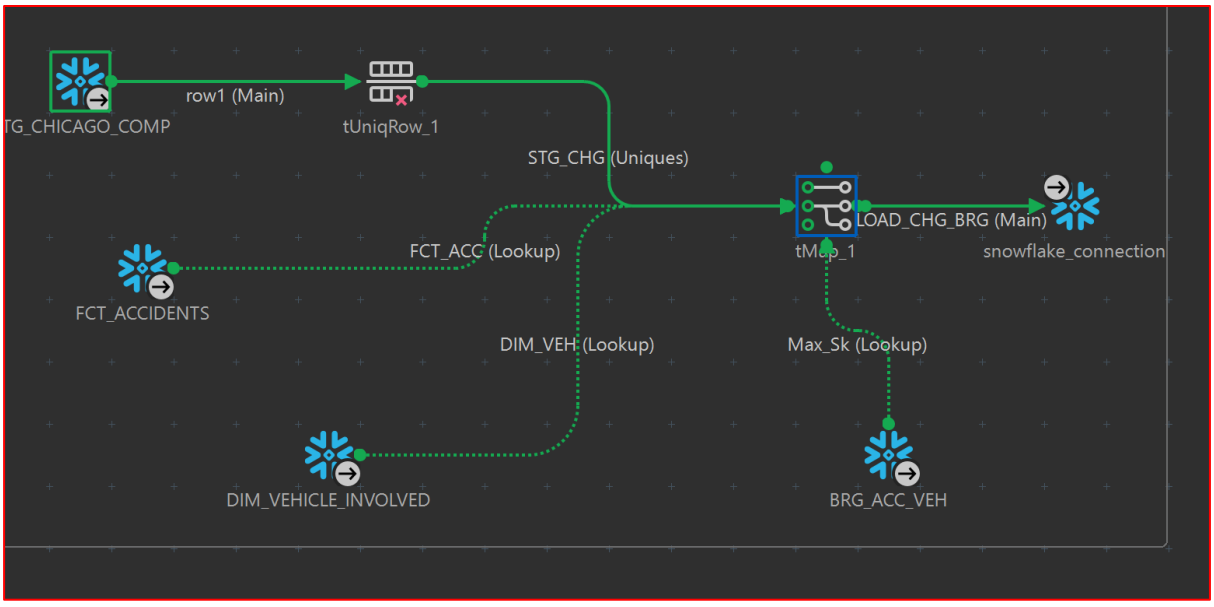
This Talend workflow loads data into the Bridge Table for Chicago Vehicle Involvement, linking the accident fact table (FCT_ACCIDENTS) with the vehicle dimension table (DIM_VEHICLE_INVOLVED). This bridge table manages many-to-many relationships between accidents and vehicles involved in Chicago, enabling detailed analysis.

Key Workflow Components:

1. Input Sources:
    - STG_CHICAGO_COMP: Staging data for Chicago containing accident and vehicle information.
    - FCT_ACCIDENTS: Fact table containing accident records.

- DIM_VEHICLE_INVOLVED: Dimension table containing vehicle details.

2. Data Deduplication:
   - tUniqRow_1:
     - Ensures unique relationships between accident_id and vehicle_id by removing duplicate rows from the staging data.
3. Fact and Dimension Lookups:
   - FCT_ACC:
     - Performs a lookup to retrieve accident_id from the fact table for the respective accident.
   - DIM_VEH:
     - Looks up vehicle_id from the vehicle dimension table using relevant vehicle attributes.
4. Data Transformation:
   - tMap_1:
     - Maps the deduplicated data to the schema of the bridge table, including attributes such as vehicle_role.
5. Data Loading:
   - tDBOutput_1:
     - Inserts the processed data into the Chicago-specific Vehicle Bridge Table in Snowflake, ensuring clean and structured data for analysis.

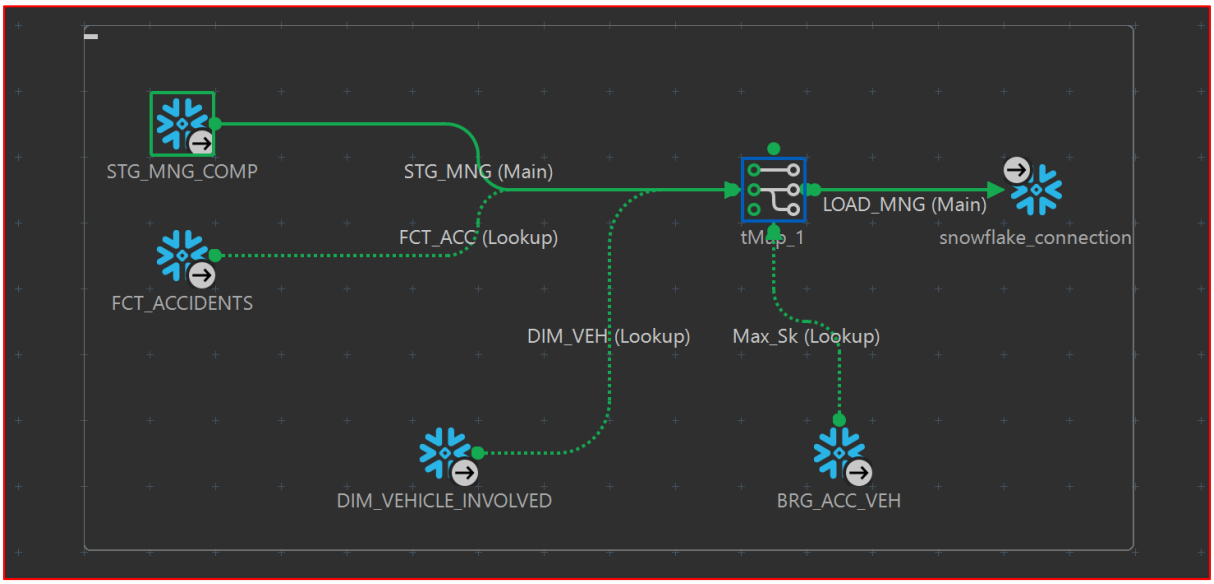


- **BRIDGE FACT MONTGOMERY VEHICLE**

Overview:
This Talend workflow creates the Bridge Table for Montgomery Vehicle Involvement, linking the fact table (FCT_ACCIDENTS) and the dimension table (DIM_VEHICLE_INVOLVED). The bridge table supports many-to-many relationships between accidents and vehicles involved in Montgomery, allowing for granular analysis of accident data.

Key Workflow Components:

1. Input Sources:
    - STG_MNG_COMP: Staging data containing accident and vehicle information for Montgomery.
    - FCT_ACCIDENTS: Fact table containing accident records.
    - DIM_VEHICLE_INVOLVED: Dimension table containing details of vehicles involved in accidents.
2. Fact and Dimension Lookups:
    - FCT_ACC:
        - Retrieves accident_id from the fact table for matching accident records.
    - DIM_VEH:
        - Looks up vehicle_id from the vehicle dimension table using vehicle attributes.
3. Data Transformation:
    - tMap_1:
        - Maps the staging data fields to the bridge table schema.
        - Enriches the data with attributes such as vehicle_role and ensures compatibility with Snowflake.
4. Data Loading:
    - tDBOutput_1:
        - Inserts the processed and enriched data into the Montgomery Vehicle Bridge Table in Snowflake.



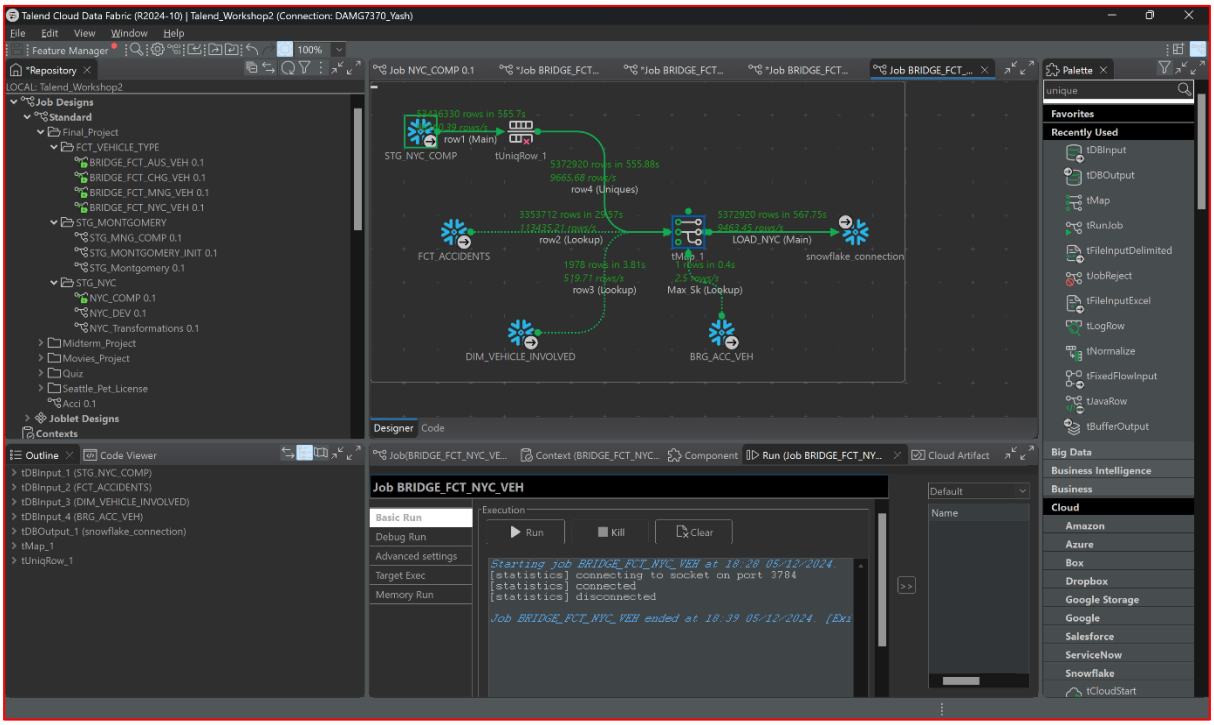
- **BRIDGE FACT NYC VEHICLE**

Overview:

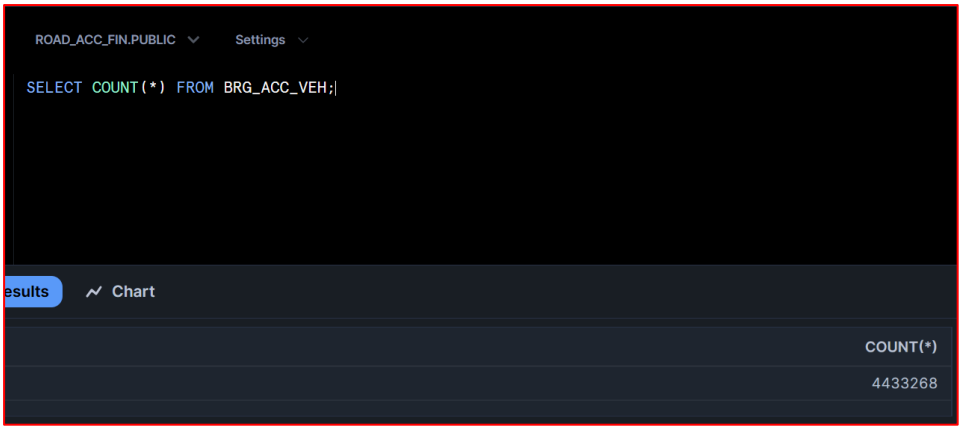
This Talend workflow processes data for the Bridge Table for NYC Vehicle Involvement, linking the accident fact table (FCT_ACCIDENTS) with the vehicle dimension table (DIM_VEHICLE_INVOLVED). The bridge table enables a many-to-many relationship between accidents and vehicles in New York City, facilitating granular analysis of vehicle involvement.

Key Workflow Components:

1. Input Sources:
    - STG_NYC_COMP: Staging data containing accident and vehicle details for NYC.
    - FCT_ACCIDENTS: Fact table with detailed accident records.
    - DIM_VEHICLE_INVOLVED: Dimension table containing vehicle-related data.
2. Data Deduplication:
    - tUniqRow_1:
        - Ensures unique combinations of accident_id and vehicle_id by removing duplicate rows from the staging data.
3. Fact and Dimension Lookups:
    - FCT_ACC:
        - Matches accident data with FCT_ACCIDENTS to retrieve accident_id.
    - DIM_VEHICLE_INVOLVED:
        - Maps vehicle data to retrieve the corresponding vehicle_id.
4. Data Transformation:
    - tMap_1:
        - Maps the deduplicated staging data to the bridge table schema.
        - Includes additional attributes such as vehicle_role for analytical use.
5. Data Loading:
    - tDBOutput_1:
        - Loads the enriched and deduplicated data into the NYC Vehicle Bridge Table in Snowflake.

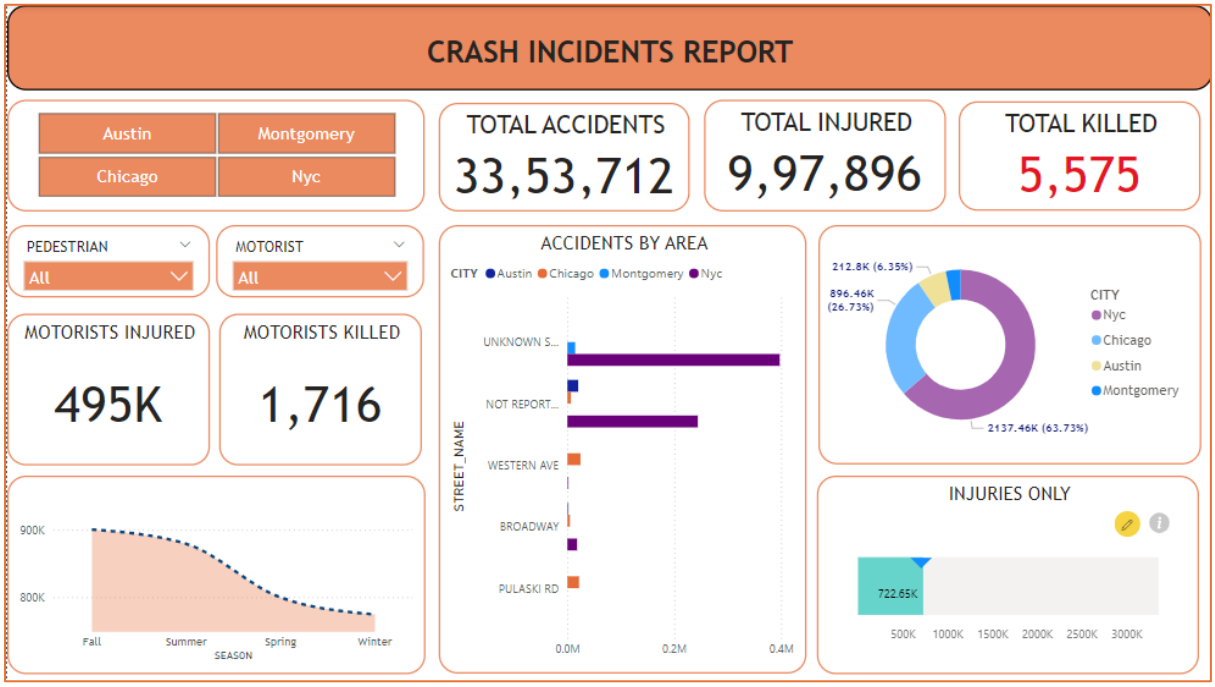


SQL QUERY


ROAD_ACC_FIN.PUBLIC
Settings
SELECT COUNT(*) FROM BRG_ACC_VEH;
esults
Chart
COUNT(*)
4433268

VISUALIZATION (POWER BI IMPLEMENTATION)

DASHBOARD 1



The Crash Incidents Report Dashboard provides a high-level overview of traffic accidents across multiple cities, highlighting key metrics such as total accidents, injuries, fatalities, and city-specific patterns. The interactive filters allow users to explore the data for pedestrians, motorists, and specific cities.

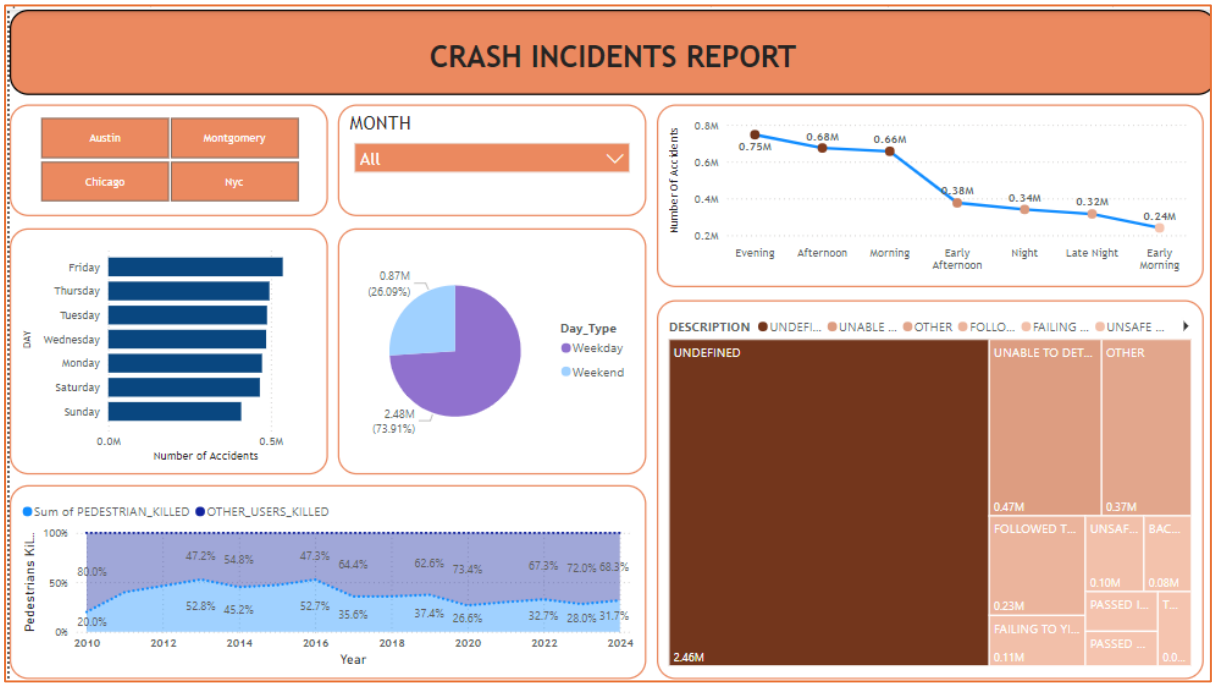
Key Visual Components:

1. City Selection Filter:
    - Buttons: Austin, Montgomery, Chicago, NYC.
    - Functionality: Enables users to toggle between cities to view location-specific data.
2. Key Metrics:
    - Total Accidents (33,53,712):
        - Displays the total number of recorded traffic accidents across all selected cities.
    - Total Injured (9,97,896):
        - Represents the total count of individuals injured in traffic incidents.
    - Total Killed (5,575):
        - Indicates the total fatalities caused by accidents.
3. Motorists Statistics:
    - Motorists Injured (495K):
        - Highlights the count of injured motorists.
    - Motorists Killed (1,716):
        - Displays the number of fatalities among motorists.
4. Accidents by Area (Bar Chart):

- X-axis: Accident count by street names (Unknown St, Not Reported, Western Ave, Broadway, Pulaski Rd).
- Y-axis: City names (Austin, Chicago, NYC, Montgomery).
- Insights: Identifies streets with the highest number of incidents in each city.

5. Accidents Distribution (Pie Chart):
    - Breakdown by City:
        - NYC (6.35%), Chicago (26.73%), Austin (63.73%), Montgomery.
    - Purpose: Shows the proportional distribution of total accidents by city.
6. Seasonal Trends (Area Chart):
    - Seasons: Fall, Summer, Spring, Winter.
    - Metric: Total number of accidents per season.
    - Insight: Helps in identifying accident trends across seasons.
7. Injuries Only (Bar Chart):
    - Metric: Distribution of injuries by severity levels.
    - Visualization Goal: Highlights the count of minor to major injuries without fatalities.

DASHBOARD 2



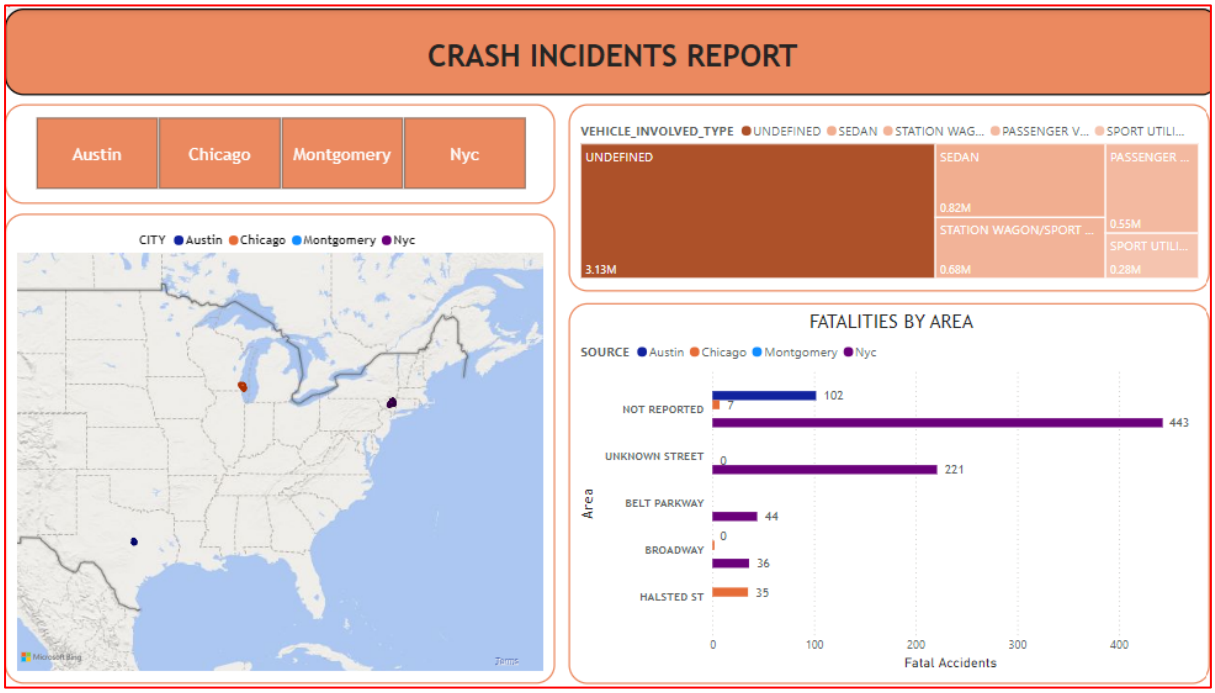
This dashboard provides a detailed analysis of crash incidents segmented by time, day, and other contributing factors. It helps users identify patterns in accident occurrences based on temporal and descriptive factors.

Key Visual Components:

1. City and Month Filters:
    - City Selector: Austin, Montgomery, Chicago, NYC.
    - Month Filter: Dropdown to analyze data for a specific month or all months combined.
2. Day of the Week Analysis (Bar Chart):
    - X-axis: Days of the week (Sunday to Saturday).
    - Y-axis: Number of accidents.
    - Insights:
        - Friday has the highest number of accidents, while Sunday reports fewer incidents.
3. Accidents by Day Type (Pie Chart):
    - Day Categories:
        - Weekday (73.91%).
        - Weekend (26.09%).
    - Purpose: Highlights the distribution of accidents between weekdays and weekends.
4. Accidents by Time of Day (Line Chart):
    - X-axis: Time periods (Evening, Afternoon, Morning, etc.).
    - Y-axis: Number of accidents.
    - Insights:
        - Evening has the highest number of accidents.
        - Accidents significantly decline in the early morning hours.
5. Pedestrian vs. Other User Fatalities (Area Chart):
    - Time Period: 2010–2024.
    - Data Breakdown:

        - Pedestrian fatalities as a percentage of total fatalities.
        - Trend of other user fatalities over the years.
    - Insights:
        - Pedestrian fatalities show a steady increase in percentage over the years.
6. Contributing Factors (Treemap):
    - Breakdown of Factors:
        - Undefined accounts for the largest portion of accidents.
        - Other categories include "Unable to Determine," "Unsafe Following," "Passing Traffic," etc.
    - Purpose: Visualizes the proportion of accidents attributed to specific causes.

DASHBOARD 3



This dashboard provides insights into the distribution of vehicle types involved in accidents, fatalities by specific areas, and city-level visualizations of accident occurrences. It helps to identify high-risk areas and vehicle types contributing to accidents.

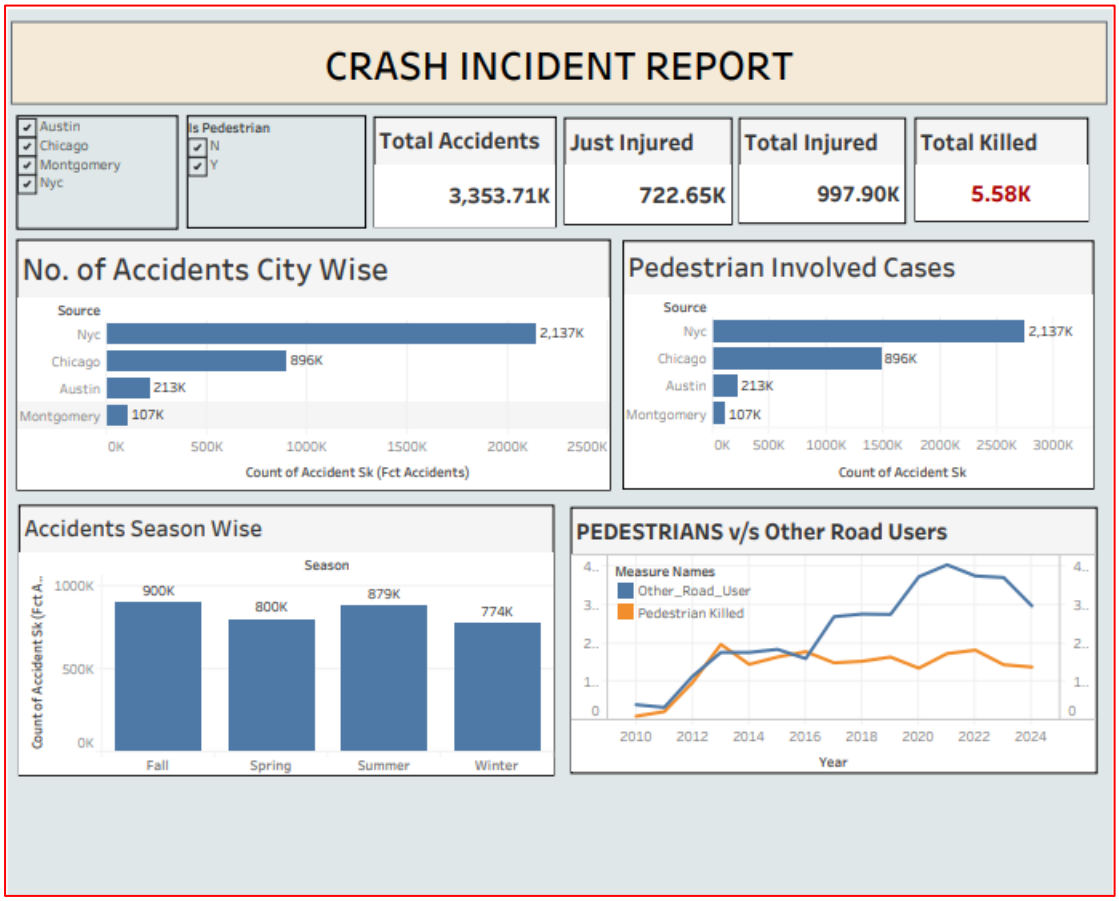
Key Visual Components:

1. City Selection Filter:
    - Buttons: Austin, Chicago, Montgomery, NYC.
    - Functionality: Filters all visualizations to show data specific to the selected city.
2. Vehicle Types (Treemap):
    - Breakdown:
        - Categories include Undefined, Sedan, Station Wagon/SUV, Passenger Vehicle, and Sport Utility.
    - Insights:
        - The majority of vehicle types involved in accidents are categorized as Undefined (3.13M), indicating potential data gaps.
        - Sedans and SUVs are also significant contributors.
3. Fatalities by Area (Bar Chart):
    - X-axis: Number of fatal accidents.
    - Y-axis: Areas, such as Not Reported, Unknown Street, Belt Parkway, Broadway, and Halsted St.
    - Insights:
        - Not Reported areas and Unknown Streets contribute significantly to fatalities.
        - Specific streets like Broadway and Belt Parkway also show notable fatality counts.
4. Geographic Visualization (Map):

- Cities: Austin, Chicago, Montgomery, NYC.
- Markers: Indicate accident locations with size and color intensity representing the number of incidents.
- Insights:
    - Highlights accident hotspots in each city for spatial analysis.

VISUALIZATION (TABLEAU IMPLEMENTATION)

DASHBOARD 1



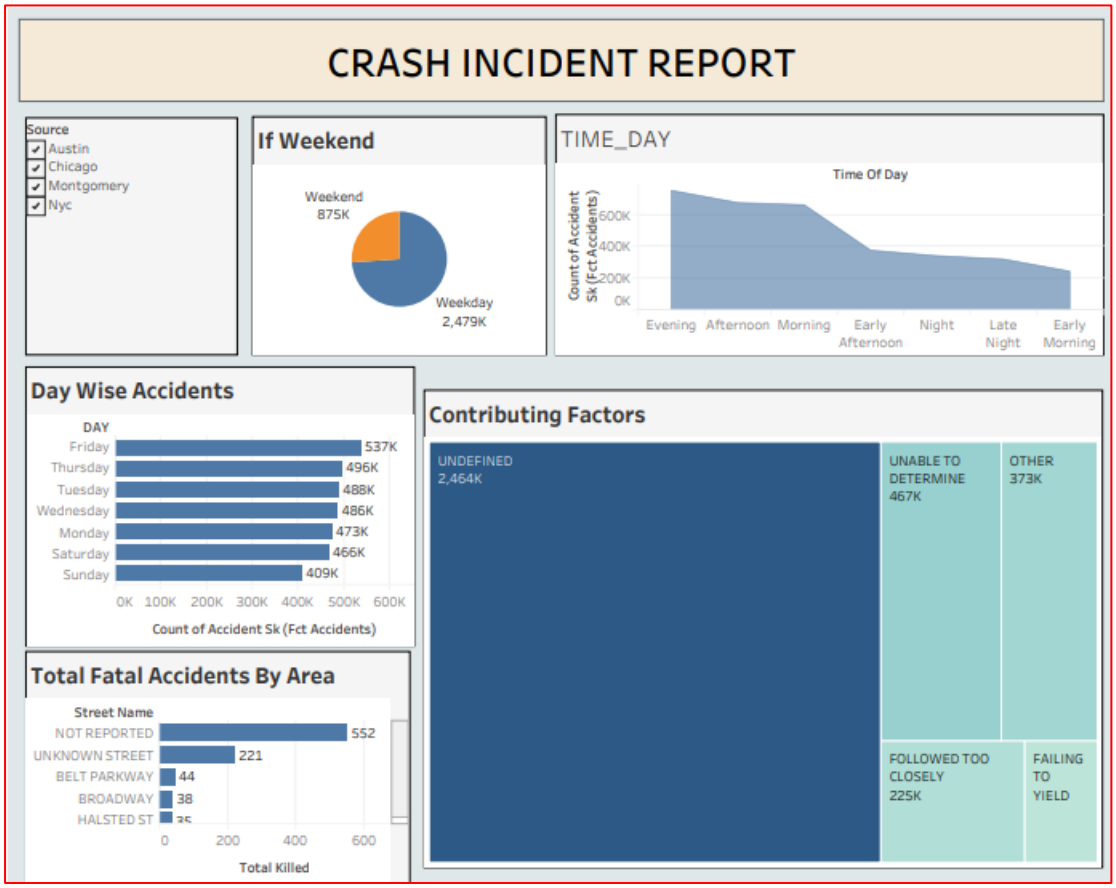
This Tableau dashboard provides a dynamic and interactive analysis of traffic accidents across selected cities, focusing on key metrics such as total accidents, injuries, fatalities, and pedestrian involvement. It enables users to filter data by city and pedestrian involvement for granular insights.

Key Visual Components:

1. City and Pedestrian Filters:
    - City Selector: Austin, Chicago, Montgomery, NYC.
    - Pedestrian Filter: Options for including or excluding pedestrian-related cases.
2. Key Metrics (Summary Cards):
    - Total Accidents: 3.353M.
    - Just Injured: 722.65K.
    - Total Injured: 997.90K.
    - Total Killed: 5.58K (highlighted for emphasis).
3. Number of Accidents City-Wise (Bar Chart):
    - X-axis: Count of accidents.
    - Y-axis: Cities (NYC, Chicago, Austin, Montgomery).
    - Insights:
        - NYC reports the highest number of accidents (2.137M).

        - Montgomery has the lowest (107K).
4. Pedestrian Involved Cases (Bar Chart):
    - Similar to the city-wise accident chart but focuses only on cases involving pedestrians.
    - NYC and Chicago dominate the count of pedestrian-related incidents.
5. Accidents Season-Wise (Bar Chart):
    - X-axis: Seasons (Fall, Spring, Summer, Winter).
    - Y-axis: Number of accidents.
    - Insights:
        - Fall sees the highest number of accidents (900K), while Winter has the lowest (774K).
6. Pedestrians vs. Other Road Users (Line Chart):
    - X-axis: Year (2010–2024).
    - Y-axis: Number of fatalities.
    - Lines:
        - Pedestrian Killed: Represents fatalities involving pedestrians.
        - Other Road Users: Represents fatalities of non-pedestrian road users.
    - Insights:
        - Pedestrian fatalities show a steady increase over the years.
        - Non-pedestrian fatalities peak in recent years but exhibit a slight decline post-2020.

DASHBOARD 2



This dashboard provides insights into the temporal distribution of accidents, contributing factors, and fatalities by area. It enables stakeholders to identify patterns related to time, day, and causal factors of incidents.

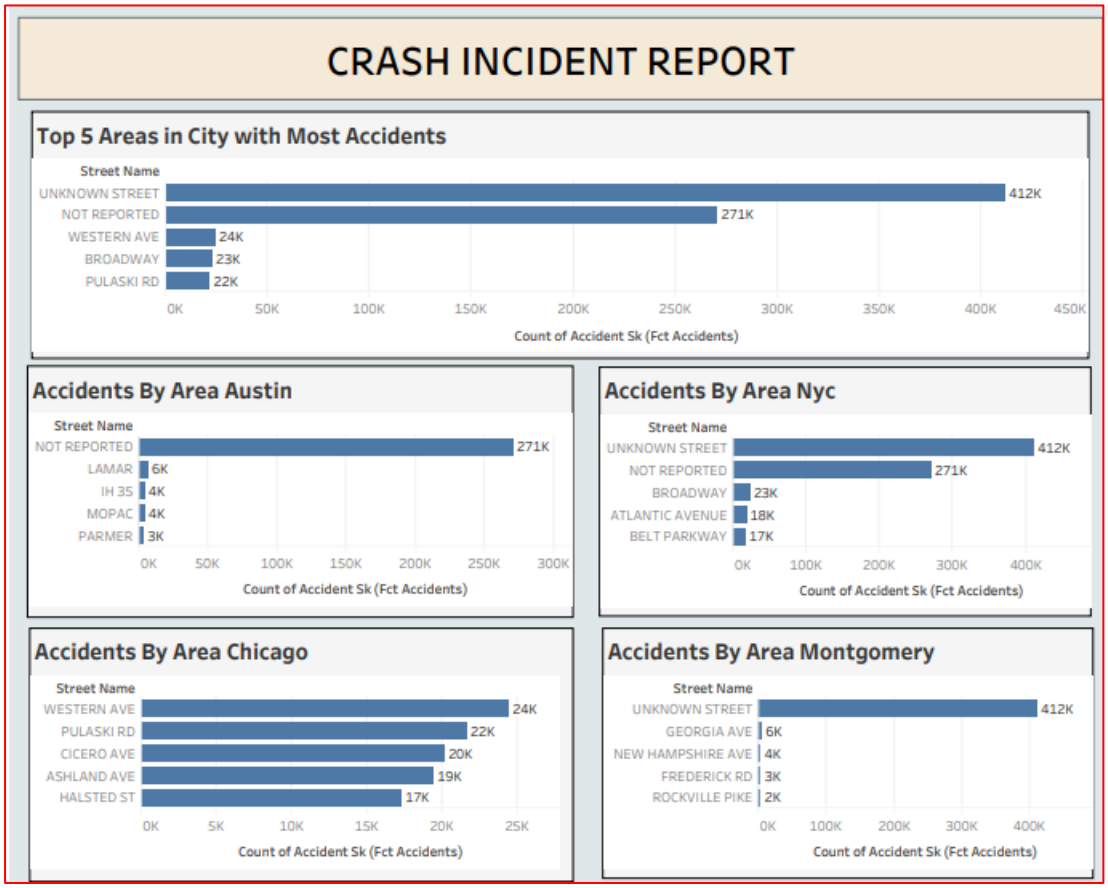
Key Visual Components:

1. City Filter:
    - Allows users to filter data by city (Austin, Chicago, Montgomery, NYC) for targeted analysis.
2. If Weekend (Pie Chart):
    - Weekend: Accidents occurring during weekends (875K, 26%).
    - Weekday: Accidents during weekdays (2.479M, 74%).
    - Insights:
        - Most accidents occur on weekdays, indicating higher traffic volume on working days.
3. Time of Day Analysis (Line Chart):
    - X-axis: Time of day (Evening, Afternoon, Morning, etc.).
    - Y-axis: Count of accidents.
    - Insights:
        - Accidents peak during Evening and drop significantly during Early Morning.
4. Day-Wise Accidents (Bar Chart):
    - X-axis: Number of accidents.
    - Y-axis: Days of the week.

- Insights:
  - Friday reports the highest number of accidents (537K).
  - Sunday has the least (409K).

5. Total Fatal Accidents by Area (Bar Chart):
   - X-axis: Number of fatalities.
   - Y-axis: Areas (e.g., Not Reported, Unknown Street, Belt Parkway, etc.).
   - Insights:
     - Not Reported accounts for the majority of fatalities (552).
     - Specific streets like Unknown Street and Belt Parkway also have significant fatalities.
6. Contributing Factors (Treemap):
   - Breakdown:
     - Undefined: Largest segment (2.464M), indicating data quality issues.
     - Other factors include:
       - Unable to Determine (467K).
       - Followed Too Closely (225K).
       - Failing to Yield and other smaller factors.
   - Purpose:
     - Highlights the need for better reporting on accident causes.

DASHBOARD 3



This dashboard provides an in-depth analysis of accident occurrences across the top five areas within each city, helping stakeholders identify high-risk streets and prioritize safety measures.

Key Visual Components:

1. Top 5 Areas in City with Most Accidents (Bar Chart):
    - X-axis: Count of accidents.
    - Y-axis: Streets.
    - Insights:
        - Unknown Street leads with 412K accidents.
        - Not Reported follows with 271K, indicating significant data gaps.
2. City-Specific Area Analysis (Bar Charts):
    - Separate bar charts for Austin, NYC, Chicago, and Montgomery, displaying the top accident-prone areas in each city.
    - Accidents by Area - Austin:
        - Not Reported: 271K.
        - Other streets like Lamar (6K) and IH 35 (4K) contribute significantly.
    - Accidents by Area - NYC:
        - Unknown Street: 412K accidents.
        - Broadway and Belt Parkway report 23K and 17K respectively.
    - Accidents by Area - Chicago:

        - Western Ave and Pulaski Rd lead with 24K and 22K accidents respectively.
        - Ashland Ave and Halsted St also show significant accident counts.
    - Accidents by Area - Montgomery:
        - Unknown Street dominates with 412K.
        - Georgia Ave (6K) and New Hampshire Ave (4K) follow.

SQL QUERIES FOR BUSINESS REQUIREMENTS

-- How many accidents occurred in NYC, Austin, Chicago and Montogomery? --

```
SELECT COUNT(*) AS NO_OF_ACCIDENTS, ds.SOURCE
FROM FCT_ACCIDENTS fct
INNER JOIN DIM_SOURCE ds ON fct.SOURCE_SK = ds.SOURCE_SK
GROUP BY ds.SOURCE;
```

/* Which areas in the city had the greatest number of accidents
Top 3 areas in each city */

```
WITH ACCIDENTS_RANKING AS (
SELECT dl.STREET_NAME, ds.SOURCE,
COUNT(fct.ACCIDENT_SK) AS NUMBER_OF_ACCIDENTS,
ROW_NUMBER() OVER (PARTITION BY ds.SOURCE ORDER BY
COUNT(fct.ACCIDENT_SK) DESC) AS RANKORDER
FROM FCT_ACCIDENTS fct
JOIN DIM_LOCATION dl ON fct.LOCATION_SK = dl.LOCATION_SK
JOIN DIM_SOURCE ds on fct.SOURCE_SK = ds.SOURCE_SK
GROUP BY dl.STREET_NAME, ds.SOURCE
)  SELECT STREET_NAME,SOURCE,NUMBER_OF_ACCIDENTS
FROM ACCIDENTS_RANKING
WHERE RANKORDER <= 3;
```

/* How many accidents resulted in just injuries? */

```
-- Overall --
SELECT COUNT(*) AS INJURIES_FROM_ACCIDENTS
FROM FCT_ACCIDENTS
WHERE TOTAL_INJURED > 0 AND TOTAL_KILLED = 0;

-- By City --
SELECT COUNT(*) AS INJURIES_FROM_ACCIDENTS, dl.CITY AS City
FROM FCT_ACCIDENTS fct
INNER JOIN DIM_LOCATION dl ON fct.LOCATION_SK = dl.LOCATION_SK
WHERE fct.TOTAL_INJURED > 0 AND fct.TOTAL_KILLED = 0
GROUP BY dl.CITY;
```

/* How often are pedestrians involved in accidents? */

```
SELECT COUNT(*) AS PEDESTRIAN_ACCIDENT
FROM FCT_ACCIDENTS WHERE IS_PEDESTRIAN = 'Y';
```

/* When do most accidents happen?

```
/* seasonality report */
SELECT dd.SEASON, COUNT(*) AS NO_OF_ACCIDENTS
FROM FCT_ACCIDENTS fct
INNER JOIN DIM_DATE dd ON fct.DATE_SK = dd.DATE_SK
GROUP BY dd.SEASON;
```

/* How many motorists are injured or killed in accidents?*/

```
-- Overall --
SELECT
SUM(MOTORIST_INJURED) AS NO_OF_MOTORISTS_INJURED,
SUM(MOTORIST_KILLED) AS NO_OF_MOTORISTS_KILLED
FROM FCT_ACCIDENTS;

-- By City --
SELECT
SUM(MOTORIST_INJURED) AS NO_OF_MOTORISTS_INJURED,
SUM(MOTORIST_KILLED) AS NO_OF_MOTORISTS_KILLED,
dl.CITY
FROM FCT_ACCIDENTS fct
INNER JOIN DIM_LOCATION dl ON fct.LOCATION_SK = dl.LOCATION_SK
GROUP BY dl.CITY;
```

/* Which top 5 areas in 4 cities have the most fatal number of accidents? */

```
WITH ACCIDENTS_RANKING AS (
SELECT dl.STREET_NAME, ds.SOURCE AS CITY,
COUNT(fct.ACCIDENT_SK) AS TOTAL_FATAL_ACCIDENTS,
ROW_NUMBER() OVER (PARTITION BY ds.SOURCE ORDER BY
COUNT(fct.ACCIDENT_SK) DESC) AS RANKORDER
FROM FCT_ACCIDENTS fct
JOIN DIM_LOCATION dl ON fct.LOCATION_SK = dl.LOCATION_SK
JOIN DIM_SOURCE ds on fct.SOURCE_SK = ds.SOURCE_SK
WHERE fct.TOTAL_KILLED > 0
GROUP BY dl.STREET_NAME, ds.SOURCE
)
SELECT STREET_NAME, CITY, TOTAL_FATAL_ACCIDENTS
FROM ACCIDENTS_RANKING
WHERE RANKORDER <= 5;
```

/* Time based Analysis of Accidents */

```
-- Day --
SELECT dd.DAY, COUNT(*) AS NO_OF_ACCIDENTS
FROM FCT_ACCIDENTS fct
INNER JOIN DIM_DATE dd ON fct.DATE_SK = dd.DATE_SK
GROUP BY dd.DAY
ORDER BY NO_OF_ACCIDENTS;

-- Time --
SELECT dt.TIME_HOUR, COUNT(*) AS NUMBER_OF_ACCIDENTS
FROM FCT_ACCIDENTS fct
INNER JOIN DIM_TIME dt ON fct.TIME_SK = dt.TIME_SK
GROUP BY dt.TIME_HOUR
ORDER BY dt.TIME_HOUR;

-- By WeekDay --
SELECT COUNT(*) AS WEEKDAY_ACCIDENTS
FROM FCT_ACCIDENTS fct
INNER JOIN DIM_DATE dd ON fct.DATE_SK = dd.DATE_SK
```

```
WHERE dd.DAY NOT IN ('Saturday', 'Sunday');

-- By Weekend --
SELECT COUNT(*) AS WEEKDAY_ACCIDENTS
FROM FCT_ACCIDENTS fct
INNER JOIN DIM_DATE dd ON fct.DATE_SK = dd.DATE_SK
WHERE dd.DAY IN ('Saturday', 'Sunday');

-- Fatality Analysis --
SELECT
SUM(PEDESTRIAN_KILLED) AS PEDESTRIANS_FATALITIES,
(SUM(TOTAL_KILLED) - SUM(PEDESTRIAN_KILLED)) AS OTHER_USERS_FATALITIES
FROM FCT_ACCIDENTS;

-- What are the most common factors involved in accidents? --
SELECT COUNT(*) AS TOTAL_COUNT, dc.DESCRIPTION
FROM BRG_ACC_CONTR brg
INNER JOIN FCT_ACCIDENTS fct ON fct.ACCIDENT_SK = brg.ACCIDENT_SK
INNER JOIN DIM_CONTRIBUTOR dc ON brg.CONTRIBUTOR_SK = dc.CONTRIBUTOR_SK
GROUP BY dc.DESCRIPTION
ORDER BY TOTAL_COUNT DESC
LIMIT 10;
```