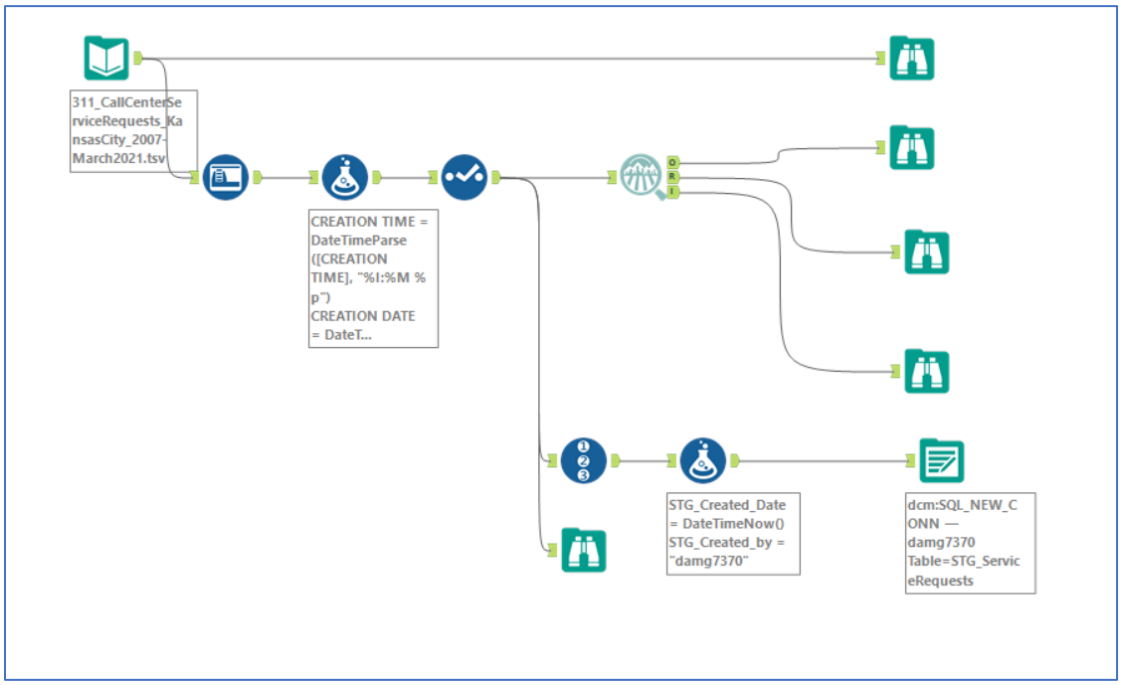
Alteryx Workflow:



Workflow Explanation

In this Alteryx workflow, the Input Data tool is initially connected to a Browse tool, providing a quick way to inspect the raw data without any processing. Simultaneously, the Input Data tool is connected to the Auto Field configuration feature, which automatically adjusts the data types of the data in the dataset to appropriate types based on content analysis. But there were some columns whose data types weren't accurately converted so a combination of select tool and formula tool was used to change the datatypes for columns Creation Date(date), closed date(date), days to close(int), zip code(int).

The workflow then incorporates a Field Summary tool, connected to the Select tool output, generating comprehensive reports on the dataset's statistical properties. These reports can be conveniently viewed using the connected Browse tools, offering insights into the data distribution.

Moreover, the Auto Field configuration feature is linked to a Formula tool, enhancing the dataset by adding two key metadata columns 'STG_Created_Date' and 'STG_Created_By' before the final output. These metadata columns capture the date and user responsible for creating the staging table. This information is invaluable for tracking and auditing purposes, providing insights into when and by whom the staging table was last modified. The browse tool connected to the output anchor of the auto field tool provides a way to inspect data after datatype conversions.

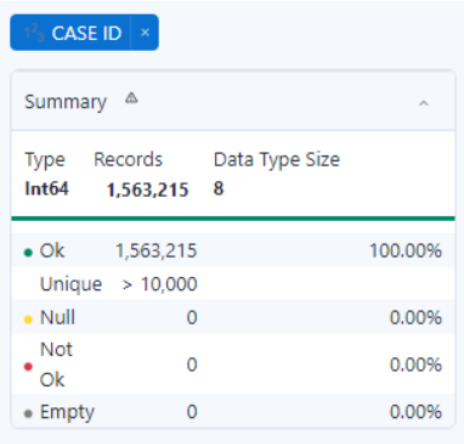
The enhanced dataset is then directed to the Output Data tool, facilitating storage in a SQL Server database. This final step completes the workflow, ensuring that the enriched data is stored for future analysis, reporting, or integration with other systems.

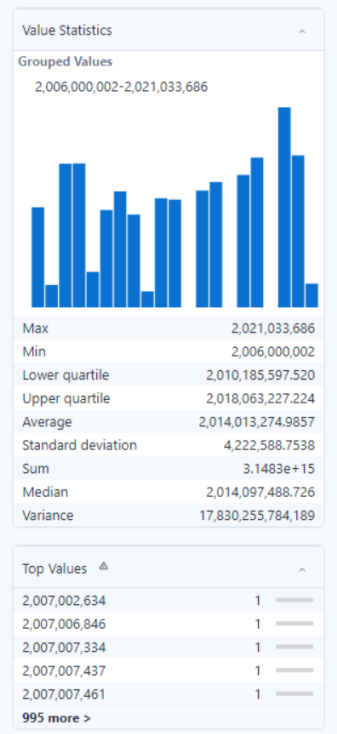
Tool Descriptions

1. **Input Data Tool:** The Input Data tool is a versatile and brings data into the Alteryx workflow. The data source here is a TSV file containing Kansas City Call Center Service Requests and the input data tool reads and imports data from the file into the Alteryx workflow.
2. **Auto Field Tool**: It is a powerful tool in Alteryx that helps automate the process of setting field. It automatically detects the data types of the fields in the dataset.
3. **Select Tool**: It is a tool used to change the datatypes of columns in the dataset.
4. **Formula Tool**: It is used to create, modify and manipulate data by applying various formulas and expressions to existing fields or creating new ones.
5. **Field Summary Tool**: It is a tool used to generate summary statistics for the fields in the dataset and provides a quick and convenient way to obtain statistical measures.
    - O Anchor (Outliers): The "o" output anchor provides information about potential outliers in the data and it includes details about values that are considered extreme or abnormal.
    - r Anchor (Results): The "r" output anchor contains the main summary statistics results for each field. This includes basic statistics such as count, mean, minimum, maximum, and standard deviation and provides a comprehensive overview of the distribution of numeric and categorical variables.
    - i Anchor (Interquartile Range): The "i" output anchor focuses on statistics related to the interquartile range (IQR). The interquartile range is a measure of statistical dispersion, representing the range between the first quartile (Q1) and the third quartile (Q3). This anchor provides insights into the middle 50% of the data distribution.
6. **Browse Tool**: The Browse tool is used for data exploration and visualization.
    - The browse tool connected directly to the Output Anchor of the Input Data tool provides a quick way to inspect the data without further processing.
    - The browse tool connected to the field summary allows to visually inspect the summary statistics for outliers ("o" anchor), overall results ("r" anchor), and interquartile range ("i" anchor).
    - The browse tool connected to the output anchor of the auto field tool provides a way to inspect data after datatype conversions.
7. **Record ID Tool** : It is used to generate a unique identifier for each record in the dataset.
8. **Output Data Tool** : Is used to write or export data from the workflow to an external data source, here SQL Server database. It enables to save, store or publish the results of the data processing or analysis.

PROBLEMS IN THE DATASET

1. CASEID:

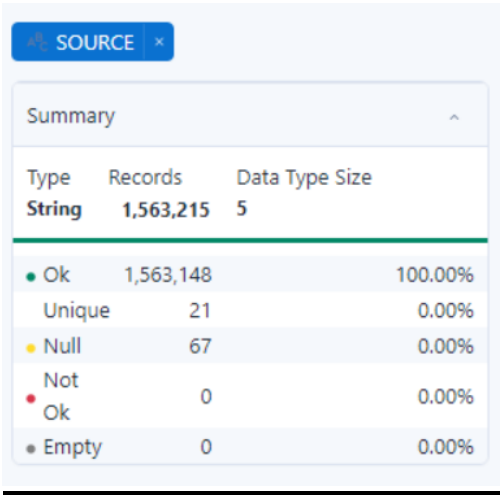


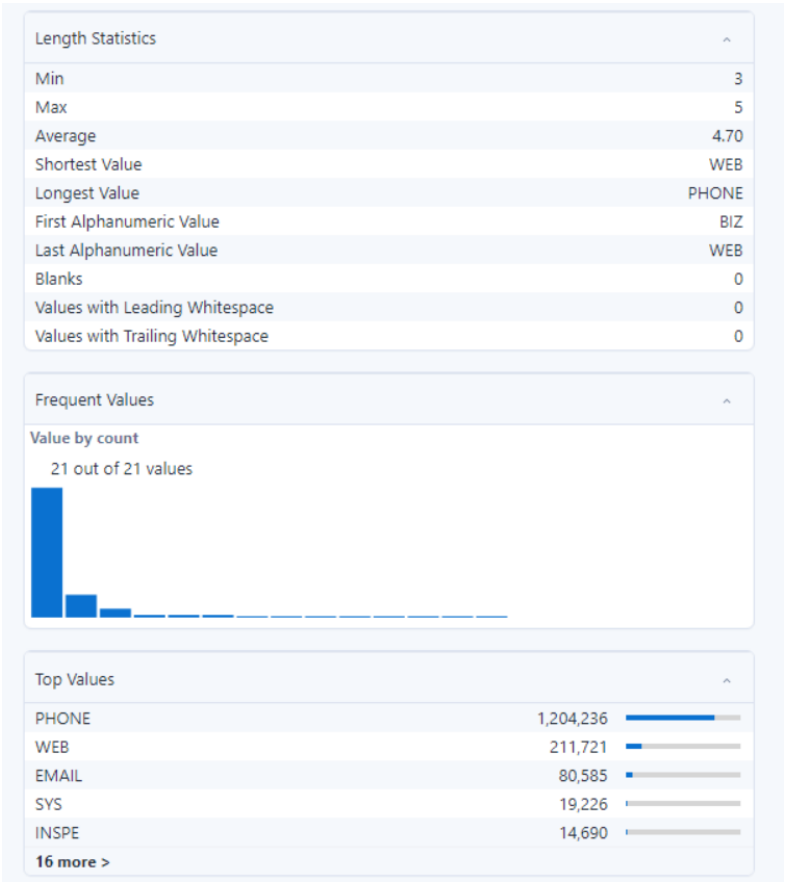


- The CASE ID column **was of type string and has been converted to Int64 and has a size of 8 meaning it stores 64-bit integer values and is now appropriate for IDs**
- There are **no null values** indicating that the column is fully populated and there are no missing entries
- All records are marked as Ok suggesting there are no immediate data quality flags raised
- The value range spans from 2,006,000,002 to 2,021,033,686.
- The **values listed only occur once which supports the uniqueness of CASE ID**

- There are no visible inconsistencies in the format or range of the CASE ID values

2. SOURCE:

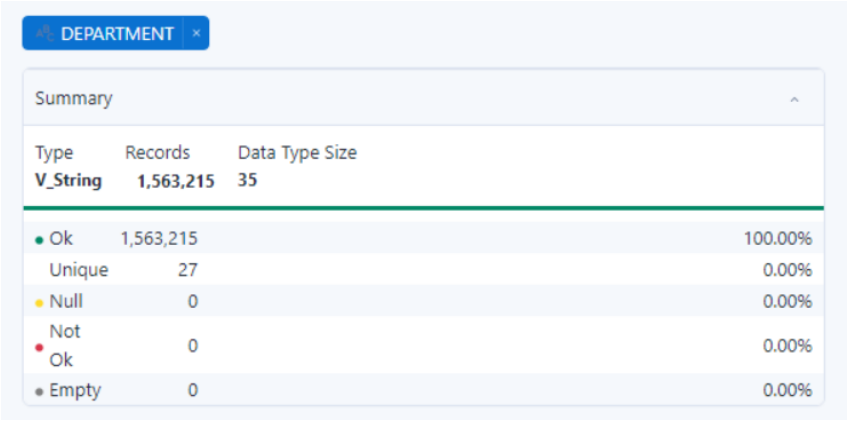


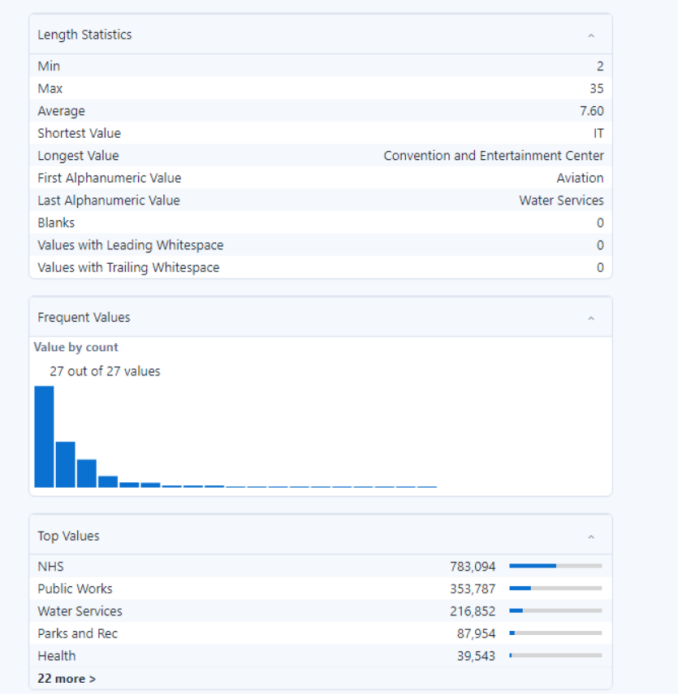


- **There are 67 null entries in the dataset, which are missing values that need to be cleaned either by removing the rows containing null values, replacing all null values with a specific value or N/A or by using the formula tool and applying a specific formula.**

- The length of the strings varies between 3 and 5 characters, with an average length of 4.70.
- The shortest value is WEB and the longest value is PHONE.
- There are **no blanks, or values with leading or trailing whitespace, which is good for consistency.**
- The column seems to have a consistent set of values.
- **No special characters are identified in the top values, and the profiling tool doesn't indicate any issues with special characters.**

3. DEPARTMENT:

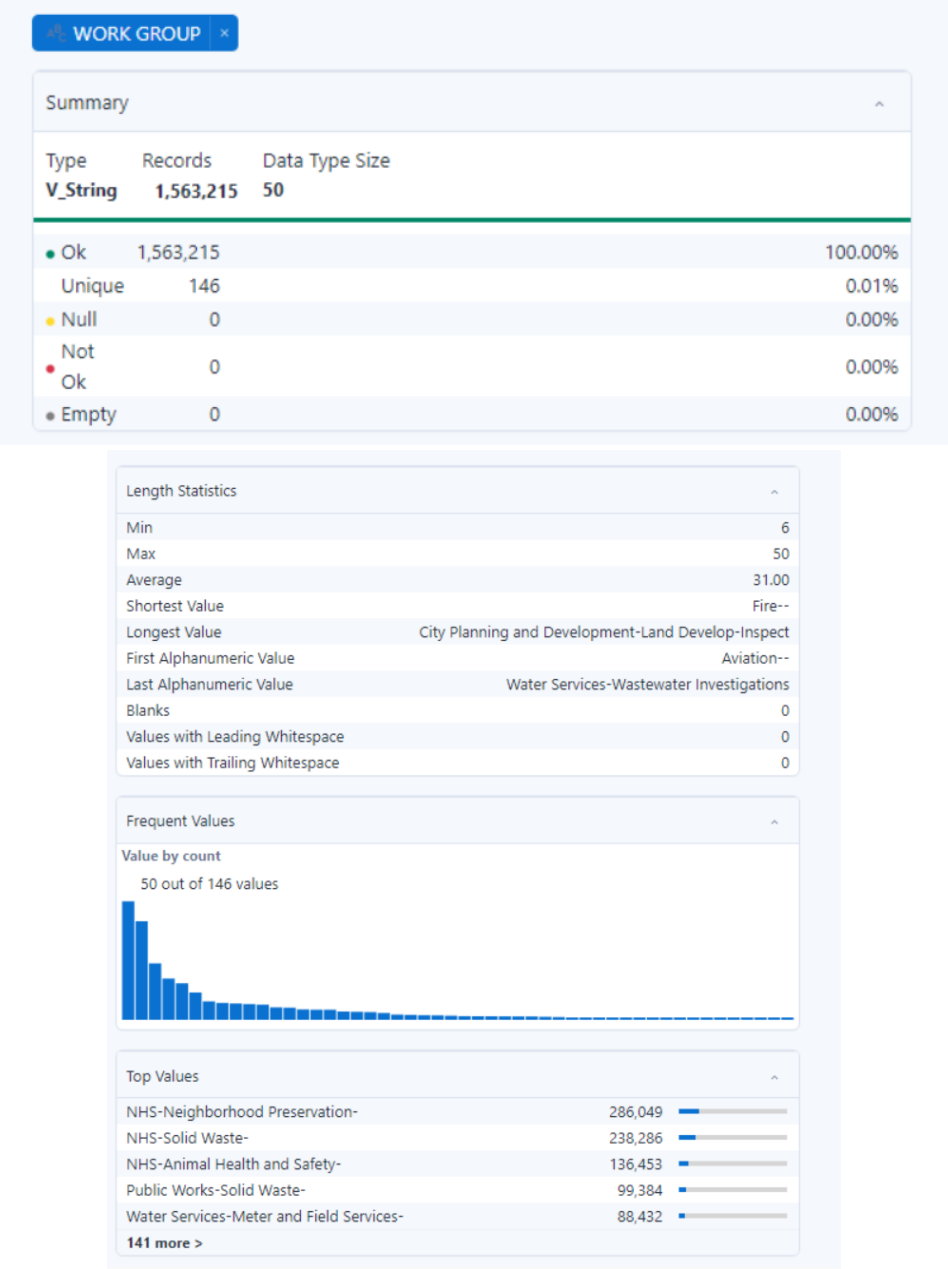




- The column is of type V_String with a maximum character **size of 35, which is suitable for textual department names.**
- All 1,563,215 records are marked as Ok with **no null values** indicating there are no immediate data quality flags for missing values.

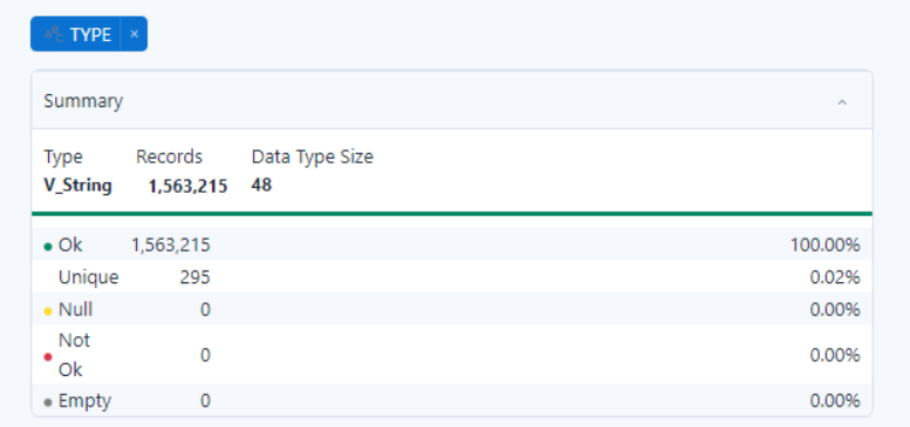
- There are **27 unique values in the column and is reasonable as there would be a limited number of departments within an organization**.
- **The department names contain abbreviations** like IT, NHS and needs to be standardized for data consistency
- **We need to have a standard naming convention for department names to improve consistency. The formula tool can be used because in some instances the department names are abbreviations like IT and some are full names.**

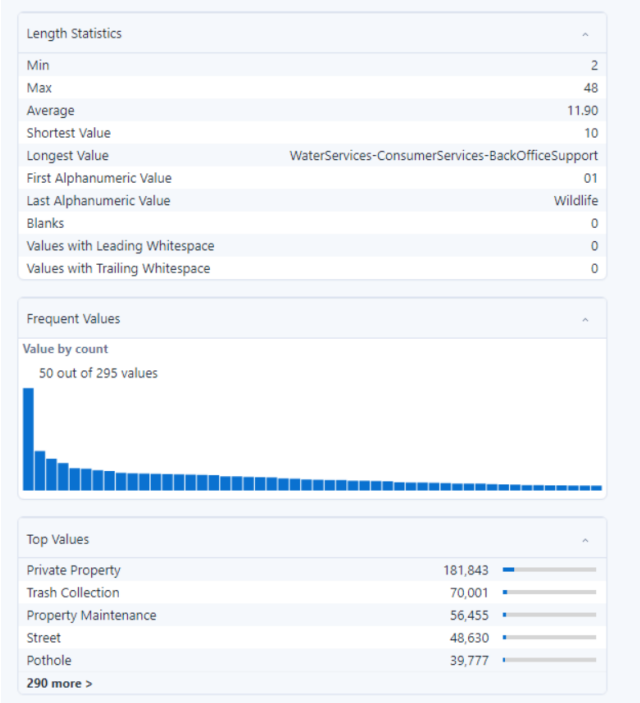
4. WORK GROUP:



- The column is of type V_String with a data type size of 50 which is considerably a **high size for just work groups or names**
- All 1,563,215 records are marked as Ok with **no null or empty values**
- The length of work group names varies from 6 to 50, with an average length of 31.00.
- **The variation in the length of work group names could be due to inconsistent naming conventions where some work groups have prefix and others don't and this can be again standardized using the formula tool.**
- **' - 'are present** in this column **indicating presence of special characters. These can be removed to improve consistency**
- Moreover some of the **workgroup names include the department names as prefix and type as suffix** this needs to be **cleaned as well using formula tool Trim(), Replace()**

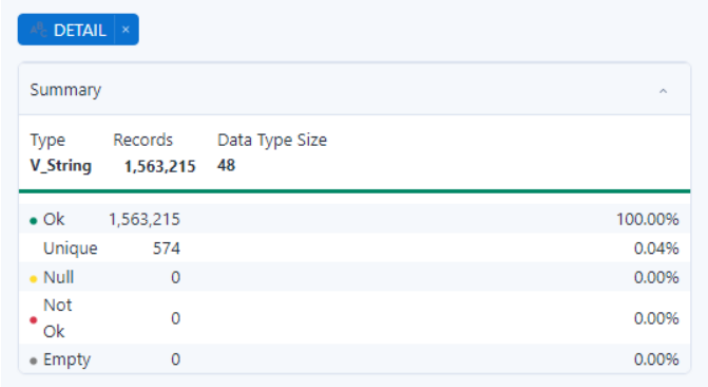
5. **TYPE:**

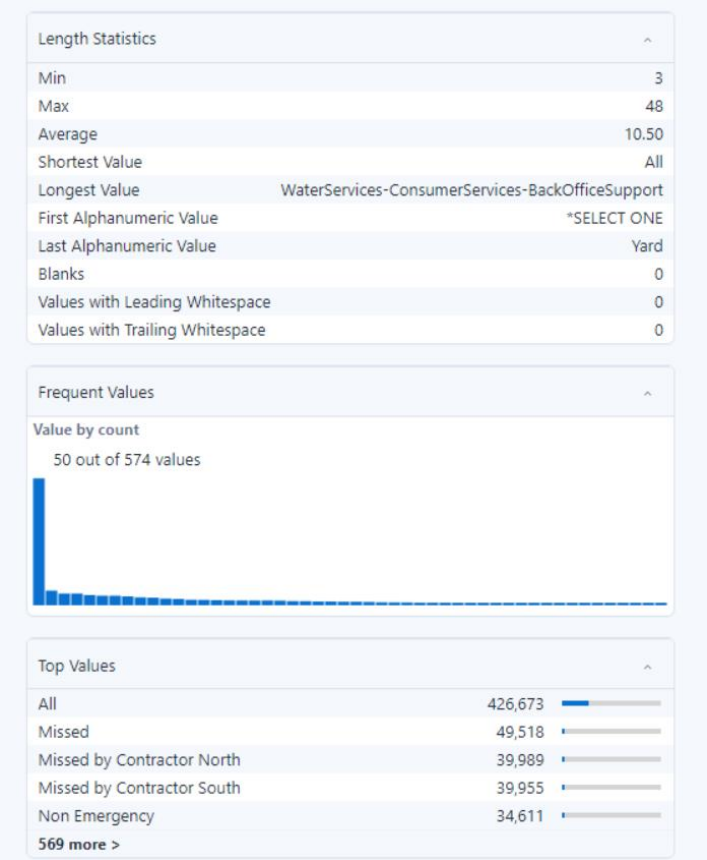




- The column is of type V_String with a data type size of 48 which is high for a column containing only type details. It shows there are some inconsistencies in naming.
- All 1,563,215 records are marked as Ok with **no null or empty values**
- The length of values ranges from 2 to 48 characters, with an average length of 11.90.
- **The longest value is WaterServices-ConsumerServices-BackOfficeSupport. ' - ' is present in this column indicating presence of special characters. These can be removed to improve consistency**
- **Some Types have detailed description which should probably be values of the next column DETAIL and not TYPE. This must be fixed to improve consistency**
- Moreover there are **numerical values present in this field which is in-consistent for a field which contains types or characters**
- Join tool can be used to compare the TYPE field against a standard list and correct discrepancies.

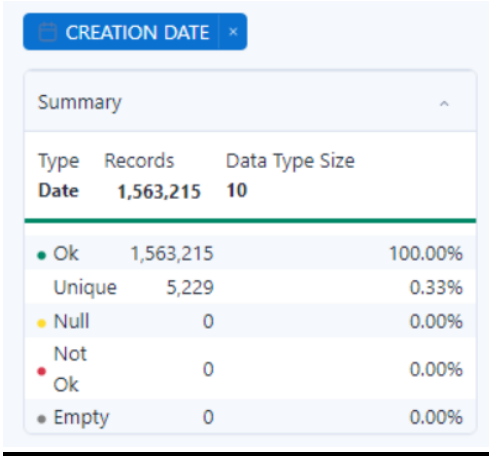
6. DETAIL:

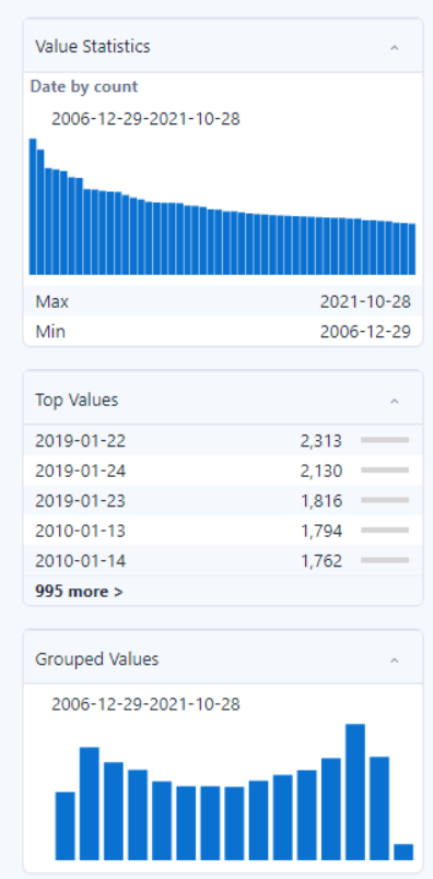




- The column is of type V_String with a data type size of 48
- All 1,563,215 records are marked as Ok **with no null or empty values**, indicating complete data with no immediate missing entries.
- There are 574 unique values in the column, indicating a considerable variety of details.
- The shortest is All and longest values is WaterServices-ConsumerServices-BackOfficeSupport. **We need to check if the values like 'All' and ' *Select One '** are meaningful**.**
- **The length of the values varies widely indicating inconsistent detail granularity and the need for standardization**
- **The use of '-' , '/' , '+' , '*' suggests special characters are used to separate different parts of the values. This needs to be cleaned-formula tool can be used to replace or remove unnecessary special characters.**
- **Join tool can be used to compare the DETAIL field against a master list and correct discrepancies.**

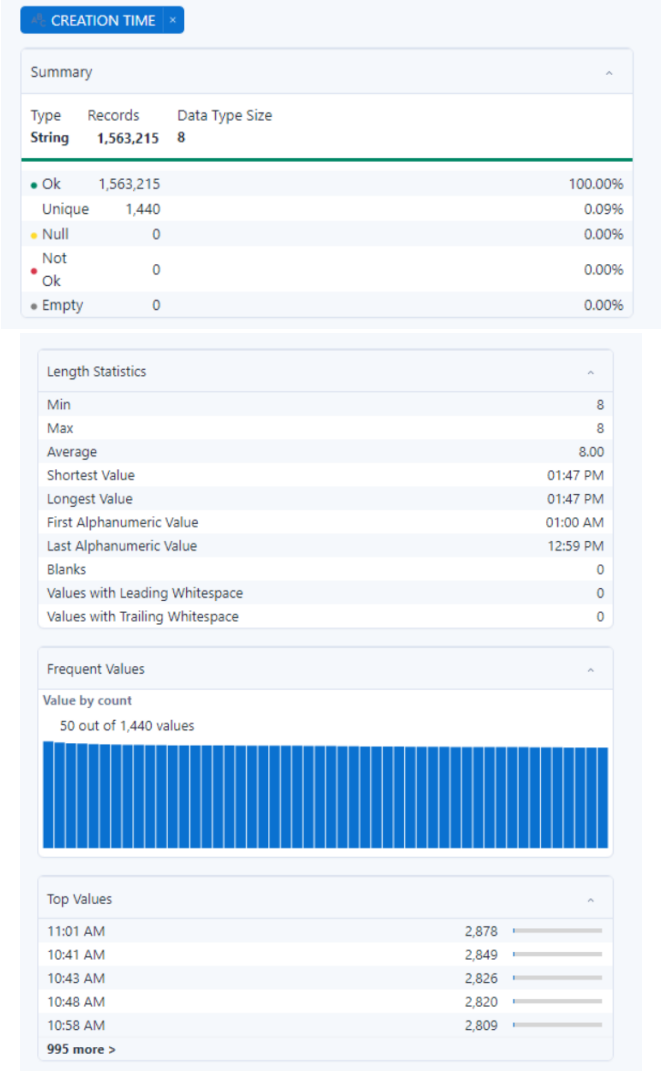
7. CREATION DATE:





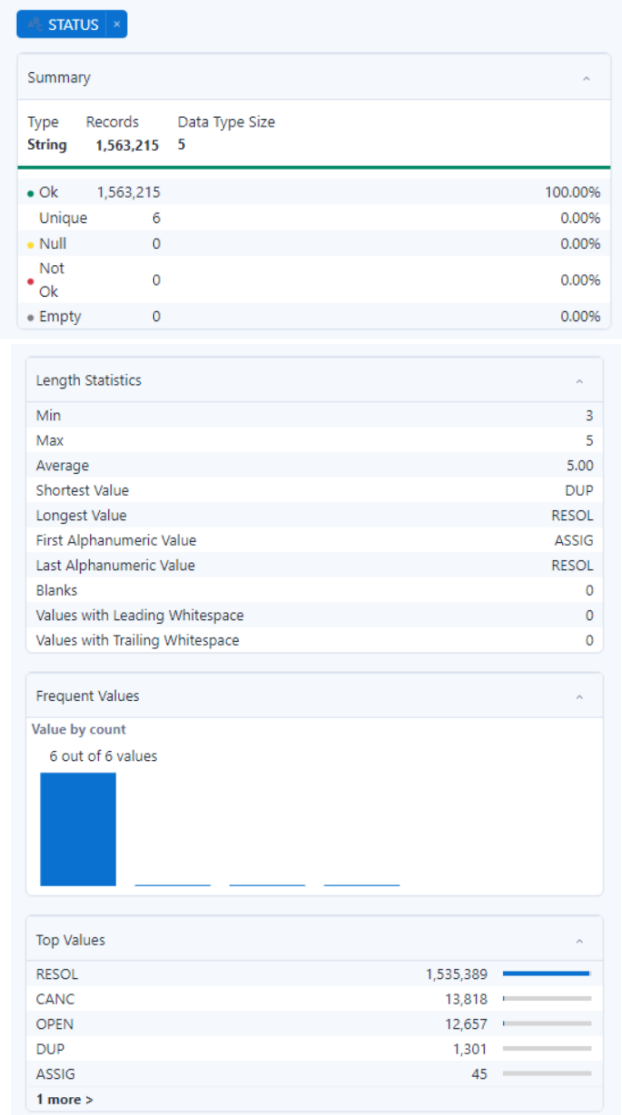
- The column was of **type String and has been converted to Date data type which is now consistent**.
- There are 1,563,215 records, all marked as Ok indicating **no missing or empty values** and no records flagged as problematic.
- Any invalid dates should be corrected or removed.

8. CREATION TIME:



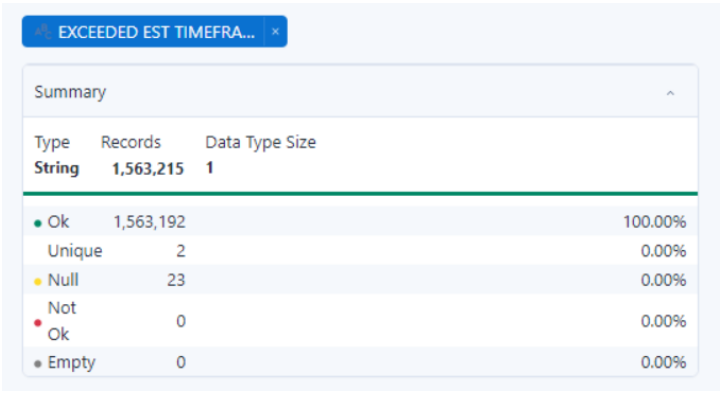
- The column is of type String with a data type size of 8, **which is inconsistent with a time format and needs to be converted to time data type. This has been done using the formula tool.**
- There are 1,563,215 records, all marked as Ok with **no null or not ok values**
- **Need to normalize and standardize to single time zone and single format 12 or 24 hr format to ensure consistency across the dataset.**

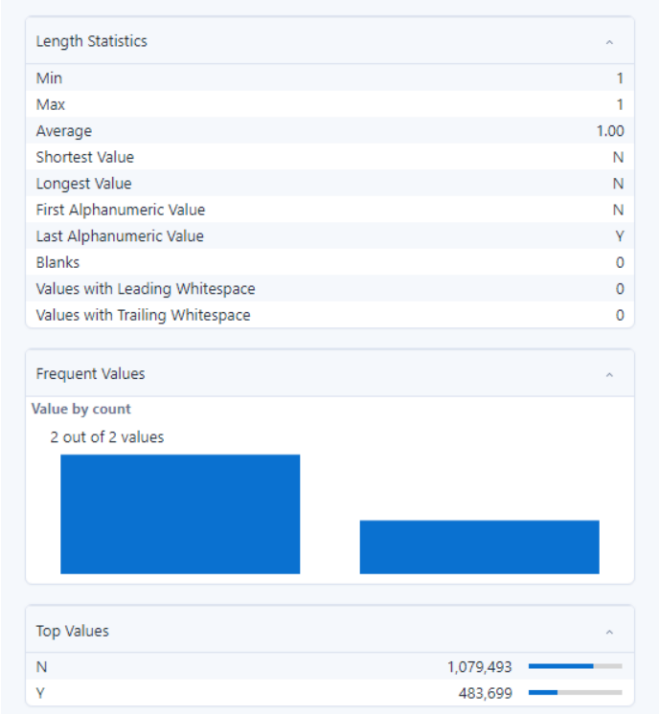
9. STATUS:



- The column is of type String with a data type size of 5
- There are 1,563,215 records, all marked as Ok with **no null, not ok, or empty values**
- **The use of abbreviations like RESOL and ASSIG can be avoided and it may be beneficial to use full descriptions rather than abbreviations for clarity. The Find Replace tool can be used to replace abbreviations.**
- There are no special characters or leading/trailing whitespace in the values, which is good for consistency.

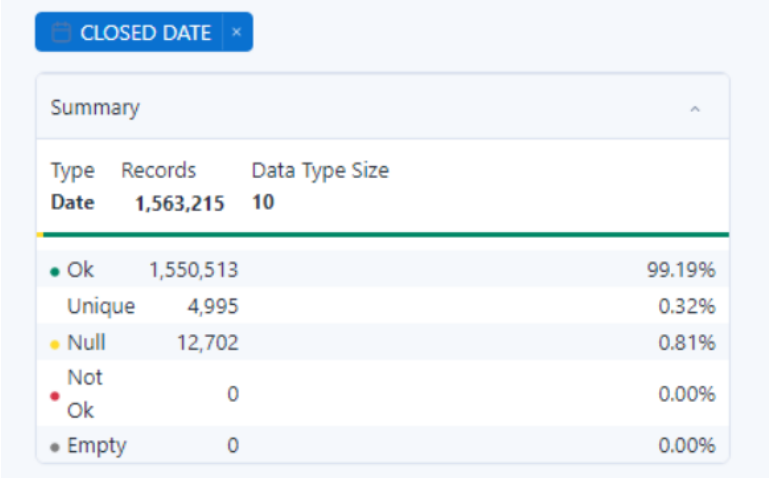
10. EXCEEDED EXT TIMEFRAME:

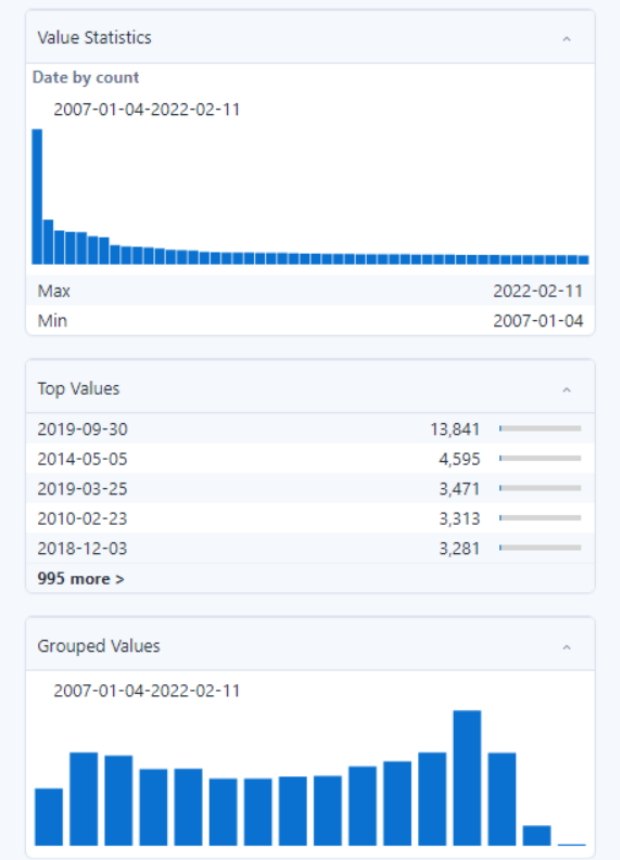




- There are **23 null entries in the dataset**, which are missing values that need to be cleaned either by removing the rows containing null values, replacing all null values with a specific value or N/A or by using the formula tool and applying a specific formula.
- The column is of type String with a data type size of 1 and since the values in the column are Y and N it indicates **better consistency and quality**.

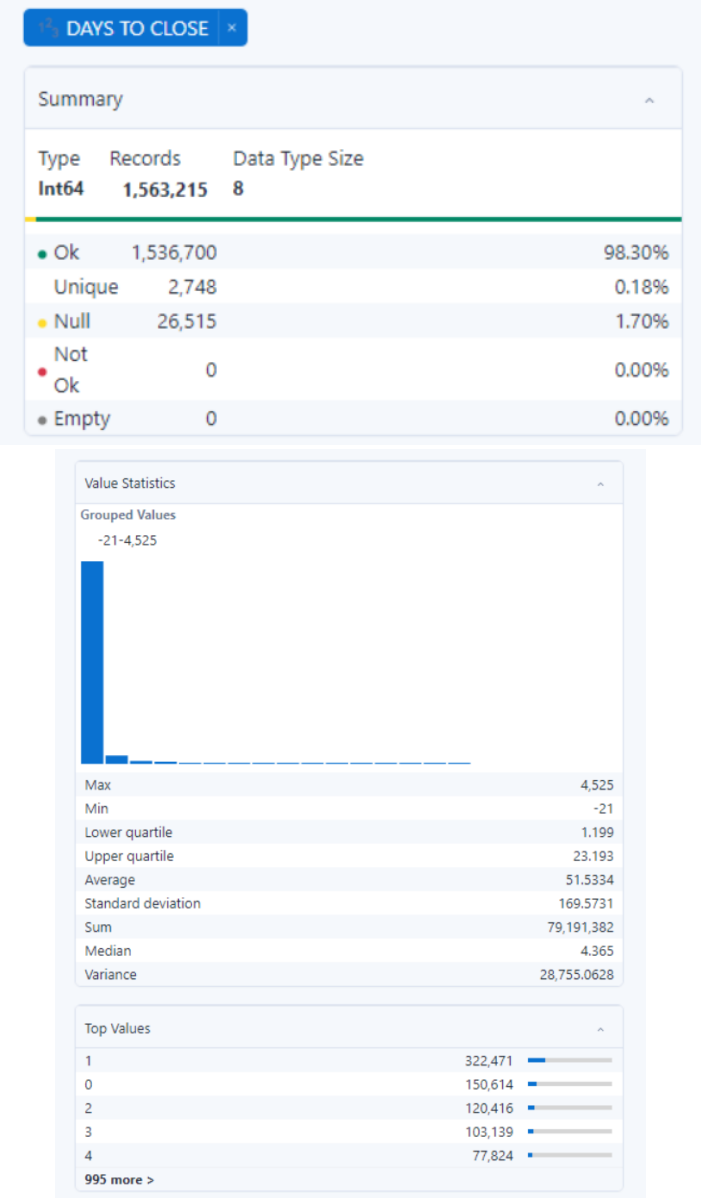
11. CLOSED DATE:





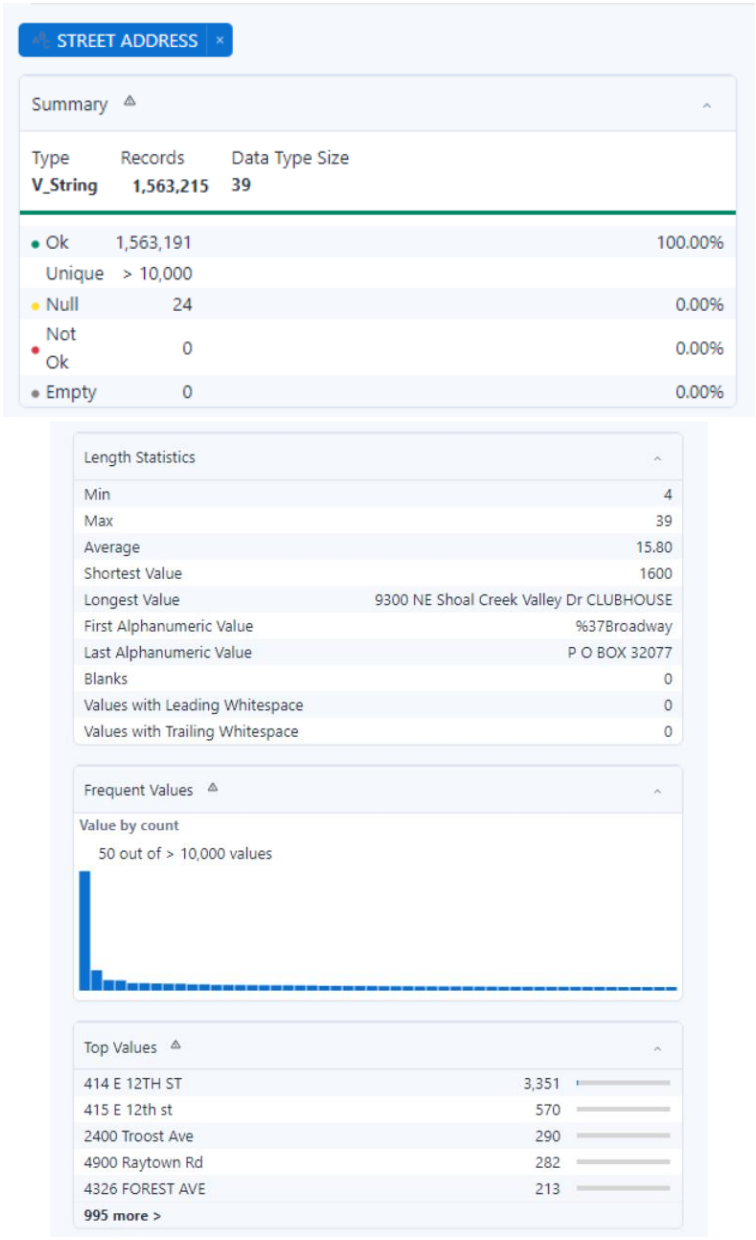
- The column **was of type String which has been converted to Date datatype using the formula tool and is now consistent.**
- **There are 12,702 null entries in the dataset, which are missing values that need to be cleaned either by removing the rows containing null values, replacing all null values with a specific value or N/A or by using the formula tool and applying a specific formula.**

12. DAYS TO CLOSE:



- The column **was of type string and has been converted to Int with a data type** size of 8 which is appropriate for numerical data that represents no of days taken to close a request.
- There are 26,515 null values which indicates that a significant number of records are missing and **needs to be either imputed or the missing rows if insignificant be removed or use Formula tool to assign a particular value or N/A to the missing values**
- The minimum value is -21 which suggests there are entries with **negative values indicating potential errors in data entry because no of days taken to complete a request can never be negative.**
- The maximum value is 4,525, which may indicate some cases took a very long time to close or there are outliers, **this can be confirmed by calculating the difference between creation and closed dates**

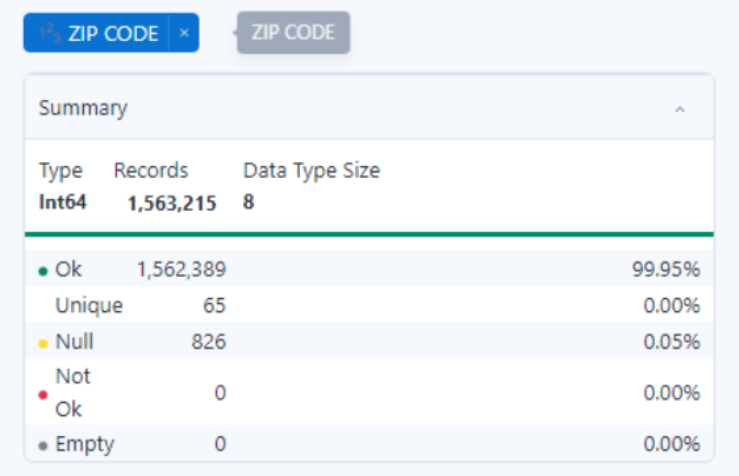
13. STREET ADDRESS:

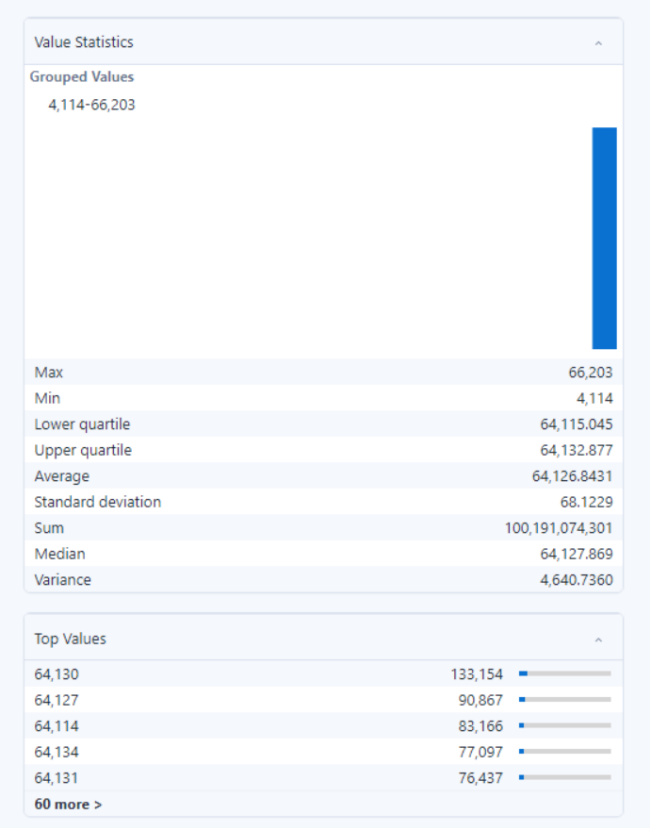


- The column is of type V_String with a data type size of 39 indicating that the field is intended for variable-length string data**.**
- **There are 24 null values** indicating a small number of records are missing causing inconsistency and data quality issues, **so the missing values needs to be imputed or using the formula tool be given a specific value or marked as N/A**
- For consistent formatting and standardization**, the address needs to be split into separate components -street number, street name, PO BOX for more structured analysis.**
- **We can also use a standard format for abbreviations and ensure consistency throughout the dataset.**

- The shortest addresses is 1600 indicating **incomplete address so we need to ensure they are complete and accurate**
- **There are special characters ‘ % ‘ present in this field and needs to be cleaned.**

14. **ZIP CODE:**

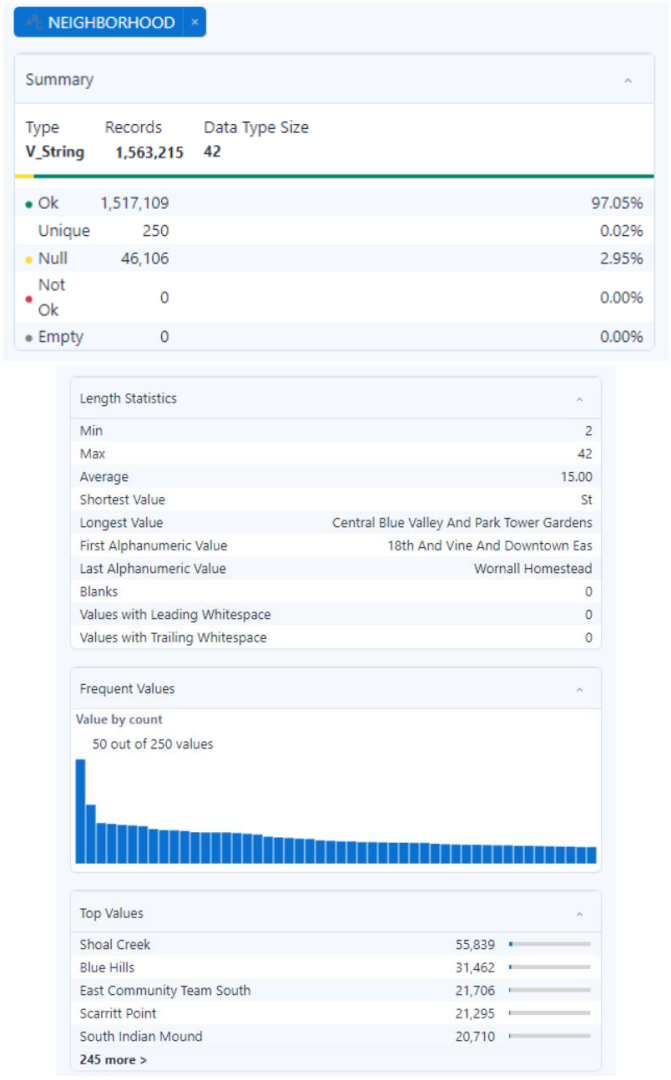




- The column was of type string and has been converted to Int with a data type size of 8 which is **now consistent with the format of zipcodes**

- There are 826 null values, indicating missing data for some records. We **need to impute the missing zip codes** if possible, or exclude these records from analyses that require complete zip code information or use the formula tool to assign a specific value or N/A
- The minimum zip code value is 4114 and the maximum is 66203 which **might not be correct given standard zip code ranges suggesting data entry errors and needs to be corrected**. We need to **validate the zip codes against a known list of valid codes and correct any that are outside standard ranges**

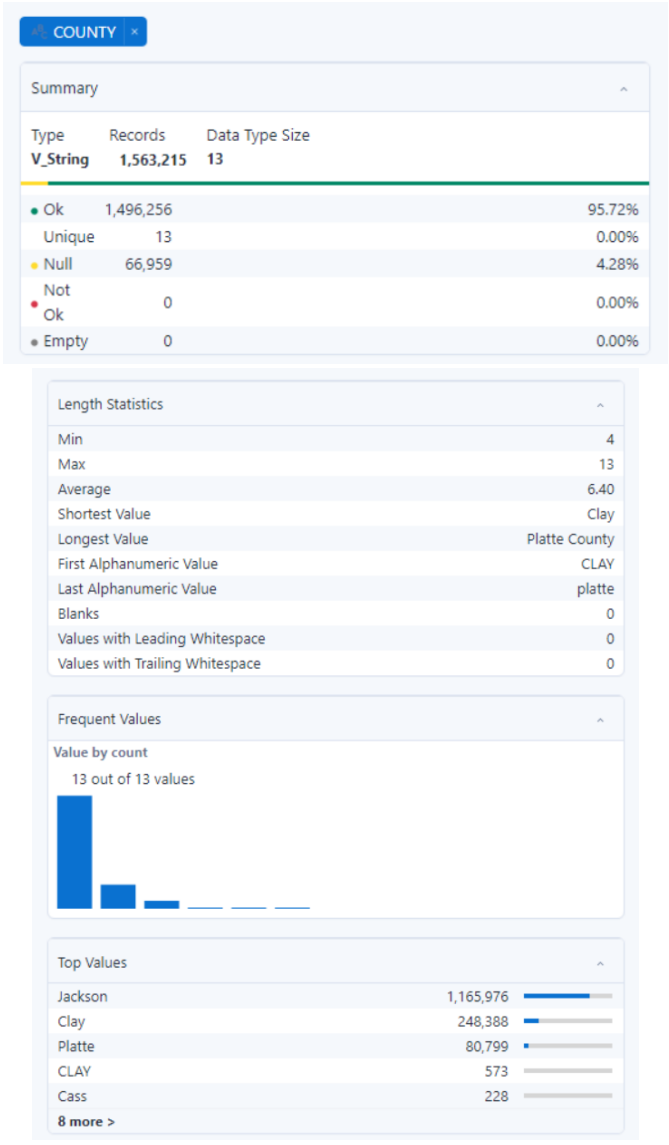
15. NEIGHBORHOOD:



- The column is of type V_String with a data type size of 42 which is maybe a **high size for a textual data that includes neighborhood names** and can be corrected.
- There are 46,106 null values, indicating that a significant portion of the dataset is missing neighborhood information. **These values need to be imputed based on information from street addresses or zip codes or if not necessary this column could be omitted**.

- The **shortest values is St, which is definitely not a neighborhood name showcasing data entry errors and hence quality issues**. This needs to be fixed.
- The **longest values is Central Blue Valley And Park Tower Gardens, which again does not indicate a neighborhood more like a street address**. This needs to be fixed and corrected as well.
- We can **standardize the neighborhood names** by using a reference list of official neighborhood names to ensure consistency.

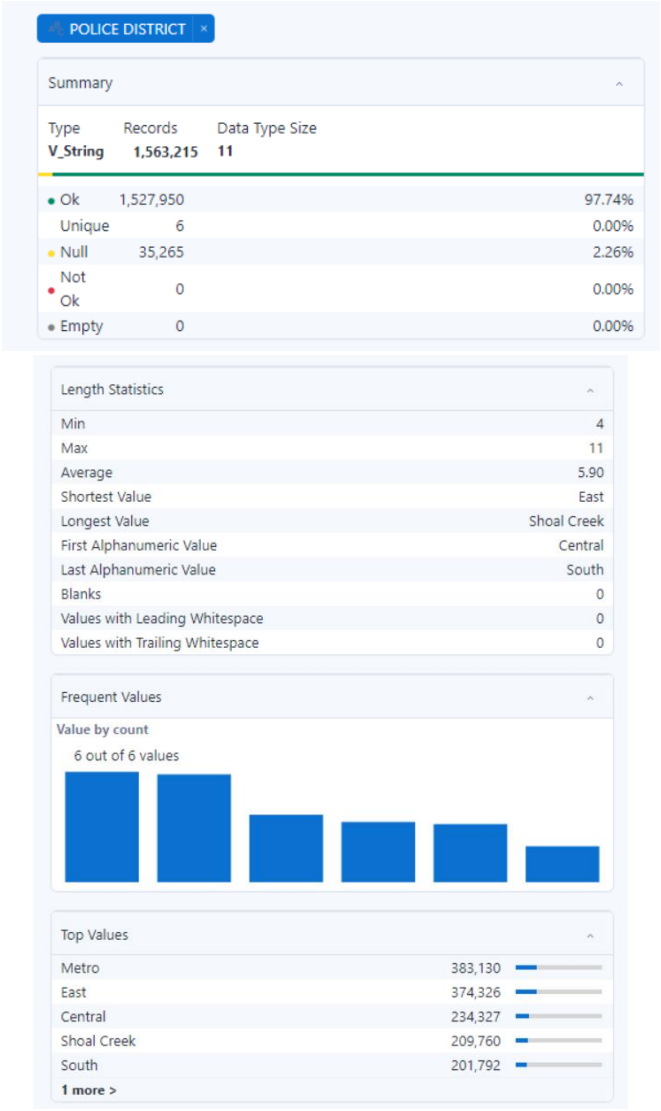
16. COUNTY:



- The column is of type V_String with a data type size of 13 which is appropriate for storing county names.

- There are 66,959 null values, indicating a notable number of records are missing county information. **If they can be determined from data zip codes we can reduce the no of missing values and increase data quality**
- The shortest county name is Clay and the longest is Platte County**. The shows that some values are abbreviated and some use full forms of county names or suffixed as county which is unnecessary in this case. We can trim the values to increase consistency.**
- Moreover, **there are spelling mismatches in the same county names and some values are capitalized and some aren't and the same value Platte is written as Platte and Platte County**. We can standardize the county names to use consistent capitalization and naming conventions.

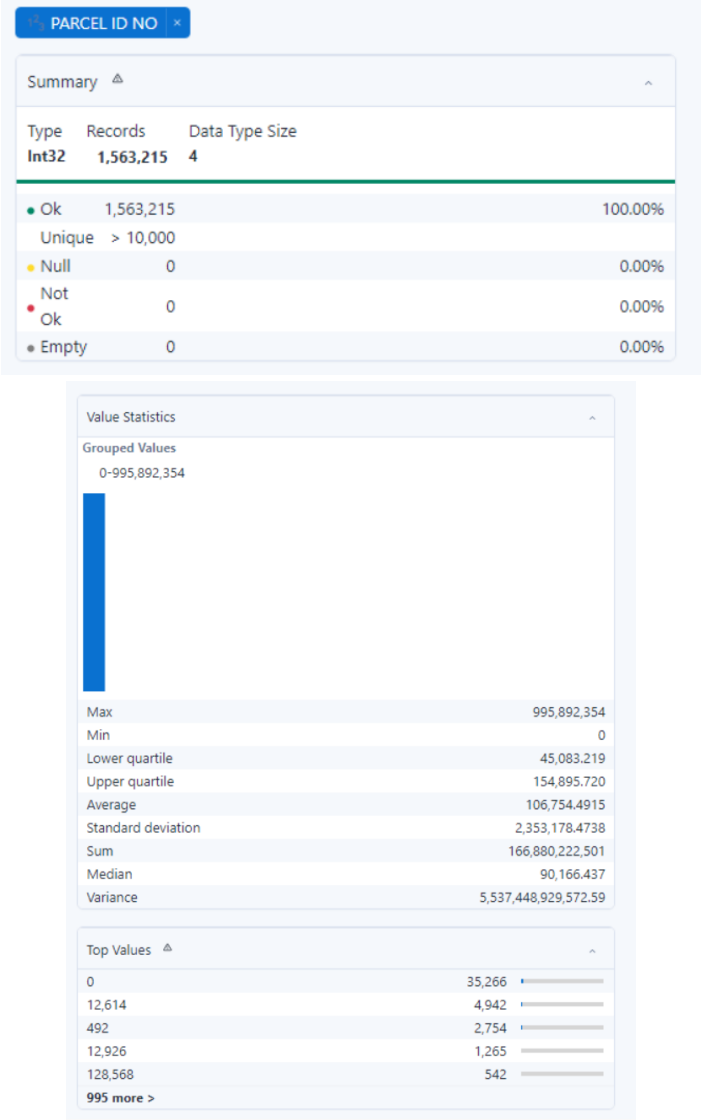
17. POLICE DISTRICT:



- The column is of type V_String with a data type size of 11 which is suitable for storing text-based district names.

- There are **35,265 null values** suggesting that a big portion of the dataset is missing police district information. **If it's possible to get the police district from data such as addresses or zip codes we can impute the missing values**
- Also since the other unique values present in the columns are direction based **we might have to check if Shoal Creek is a police district.**

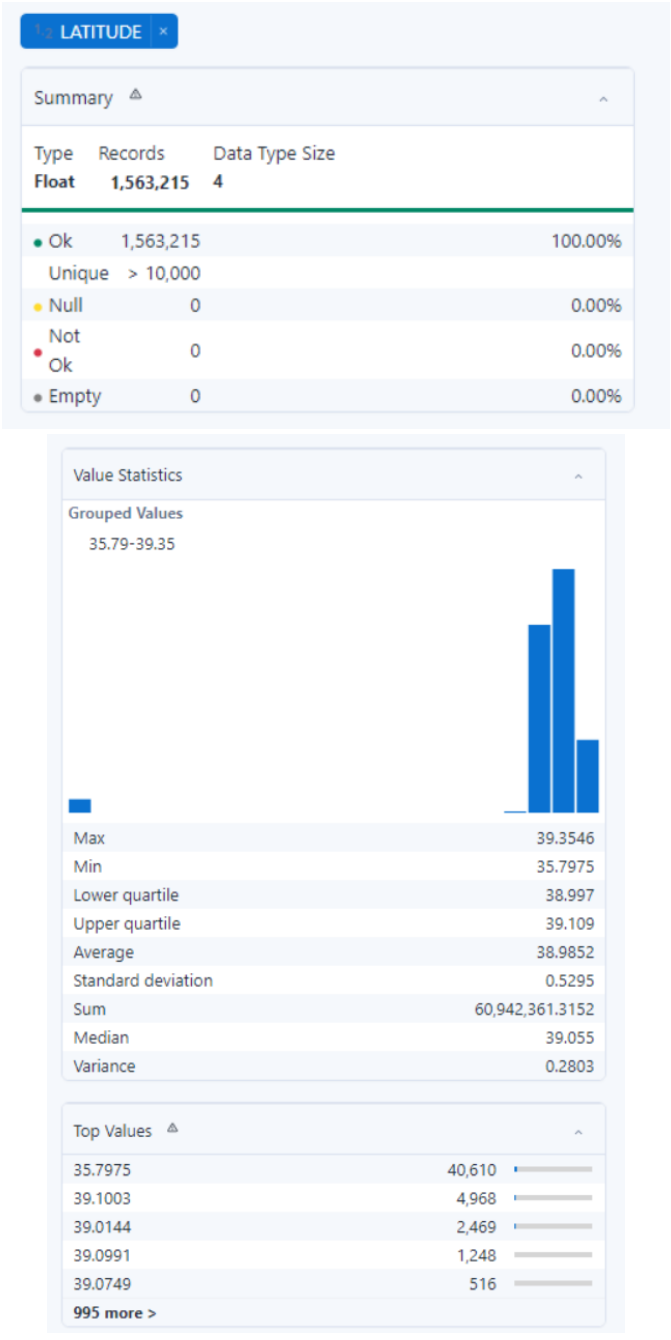
18. PARCEL ID NO:



- The **column was converted to data type Int32 with a data type size of 4** and holds integer values
- There are 1,563,215 records all marked as Ok suggesting that there are **no missing or null values**.
- The **minimum value is 0 which is ambiguous and invalid considering that parcel id's cannot have a value of 0** and the maximum value is 995892354 which could be considered a valid parcel ID number.

- **The repeated occurrence of 0 could be an issue if it's being used incorrectly as a placeholder and represents missing or unknown data.**

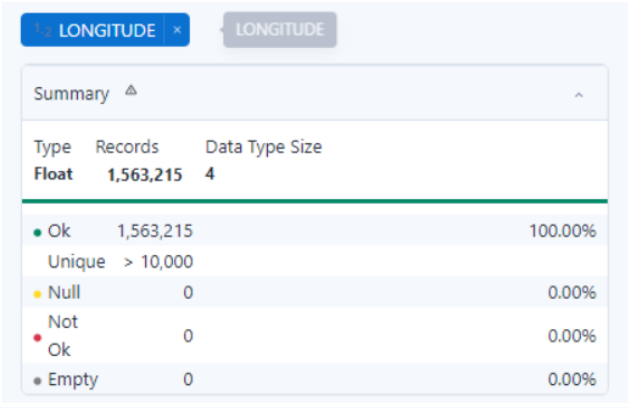
19. **LATITUDE:**

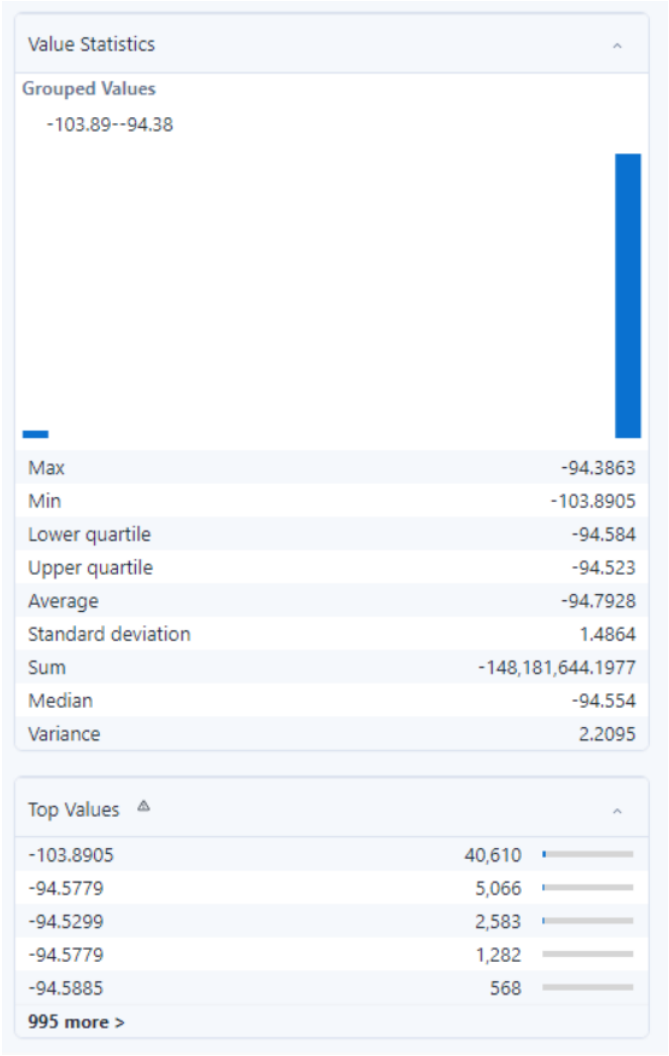


- The column was converted to datatype float, which is now **appropriate for storing latitude values as they require precision and include decimal places**.
- All 1,563,215 records have latitude values indicating that **there are no missing or null entries in this column.**

- We need to **verify that the latitude values fall within the expected geographical bounds for the dataset's region**. Normally latitude values should be between -90 and 90.
- We need to check the number of decimal places is consistent across the dataset.

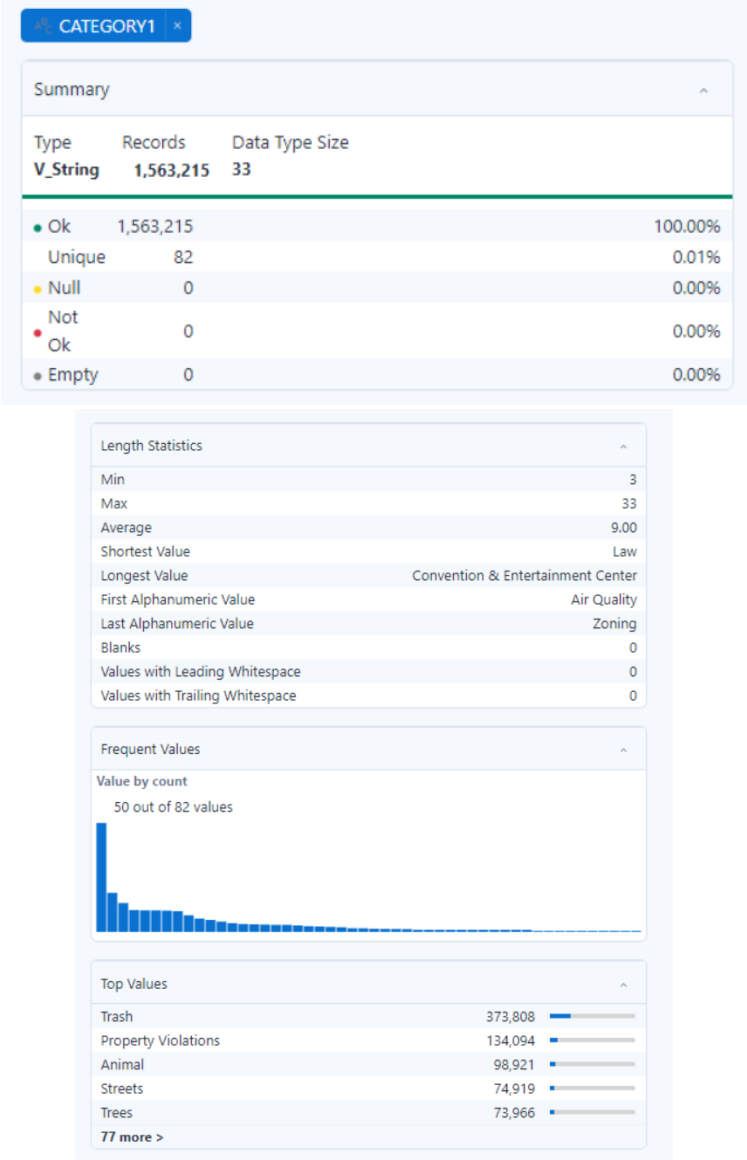
20. <u>LONGITUDE:</u>





- The column was converted to datatype float, which is now **appropriate for storing latitude values as they require precision and include decimal places**.
- There are **no missing or null values** in this column.
- Special character present in this column is **'-' but is consistent with longitude values**
- The longitude values range from -103.89 to -94.38. This range is appropriate for locations within a specific longitude range

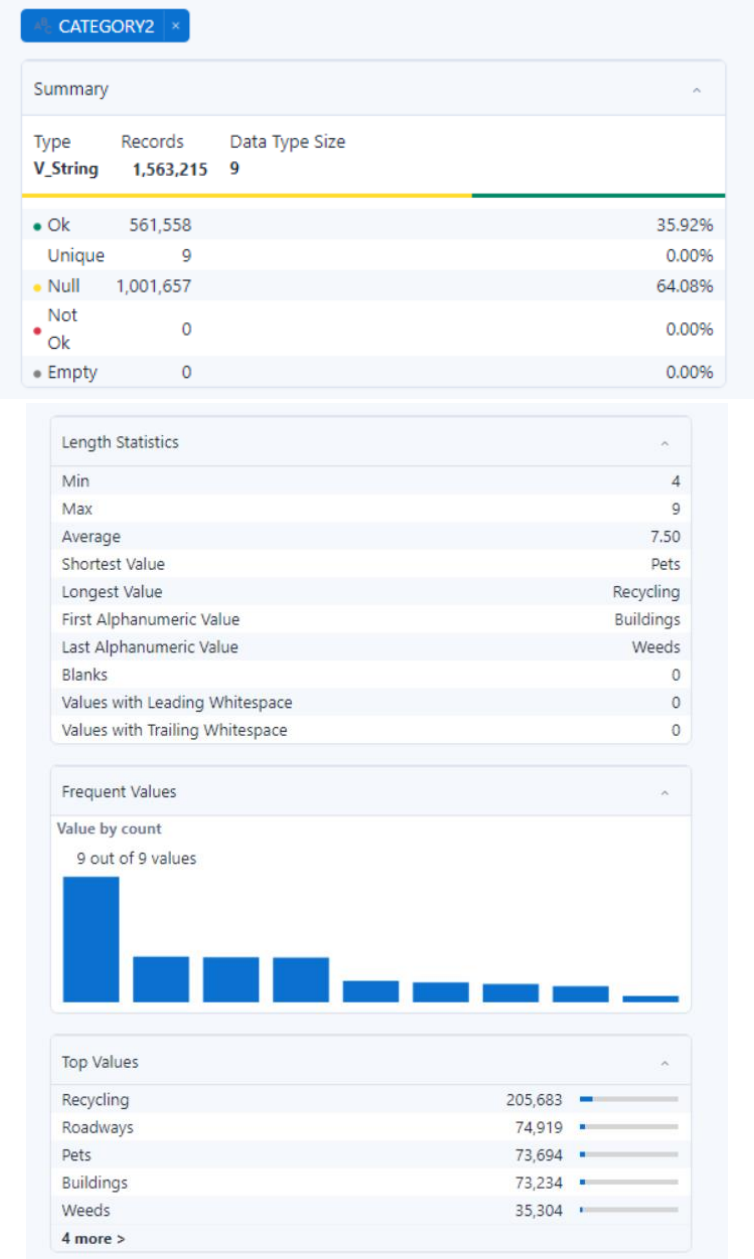
21. CATEGORY1:



- The column is of type V_String with a data type size of 33 **which could be slightly on the higher side for storing category names or descriptions**.
- There are 1,563,215 records all marked as Ok which means **there are no missing or null values in this column.**
- **'&' is a special character used in the values of the columns representing AND**

- **There is inconsistency in the naming convention, values contain both 'and' , '&'**
  ex: Convention & Entertainment Center, Boulevard Parks and Rec
- There is also no standardizations in the categories. There are categories for example water, water leak, storm water, animal, animals, street, streets which could represent the same things, so **we need to ensure the categories are standardized, not repeated and have no spelling variations.**

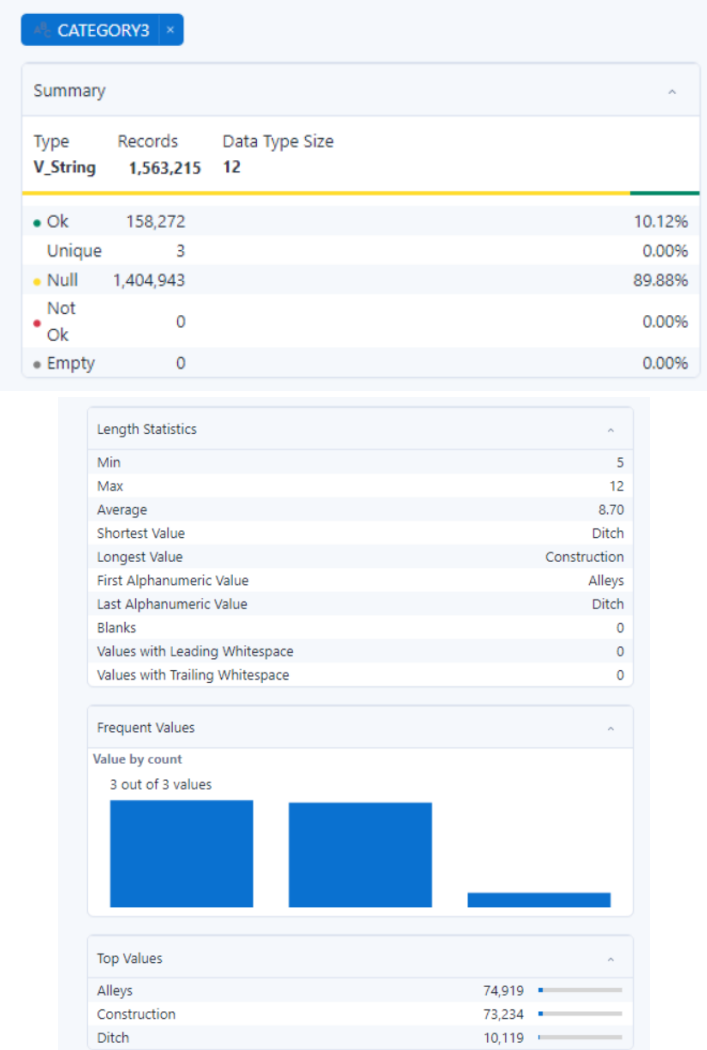
22. CATEGORY2:



- The column is of type V_String with a data type size of 9 which is appropriate for storing short categorical data.

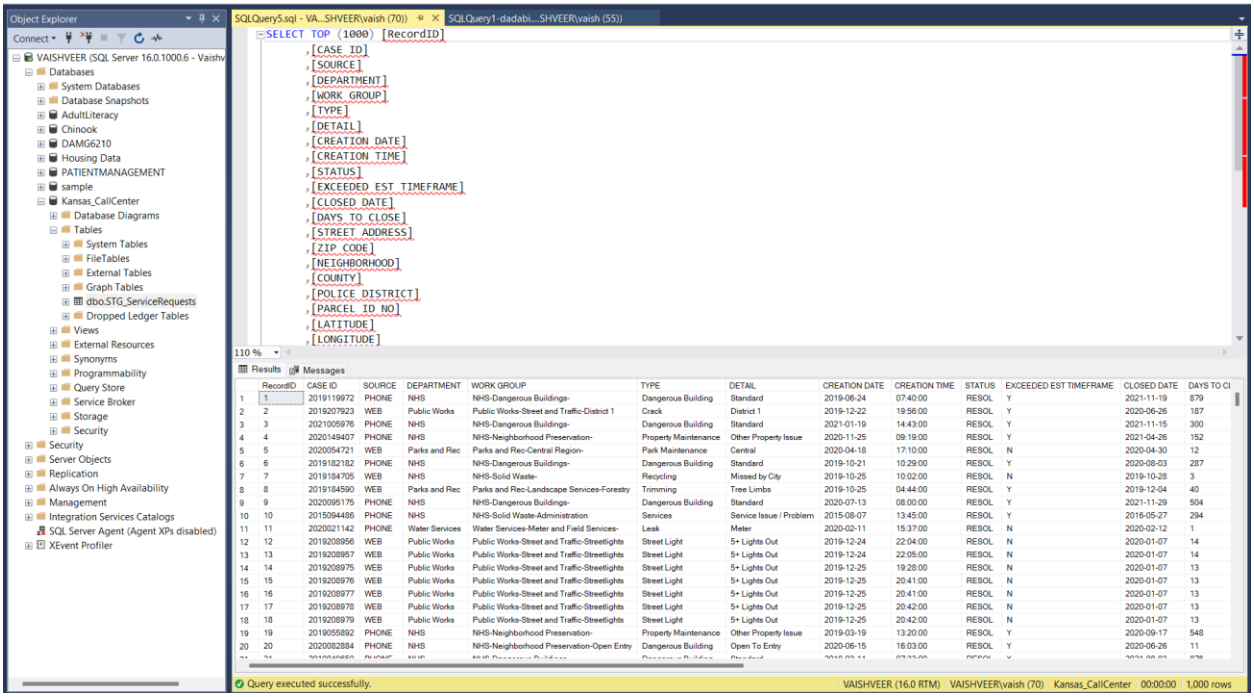
- A significant portion of the records - **1,001,657 or about 64.08%, are null which suggests that this category may be optional or not applicable in many cases. We can replace the null values with a specific value or N/A using the formula tool.**

23. <u>CATEGORY3:</u>

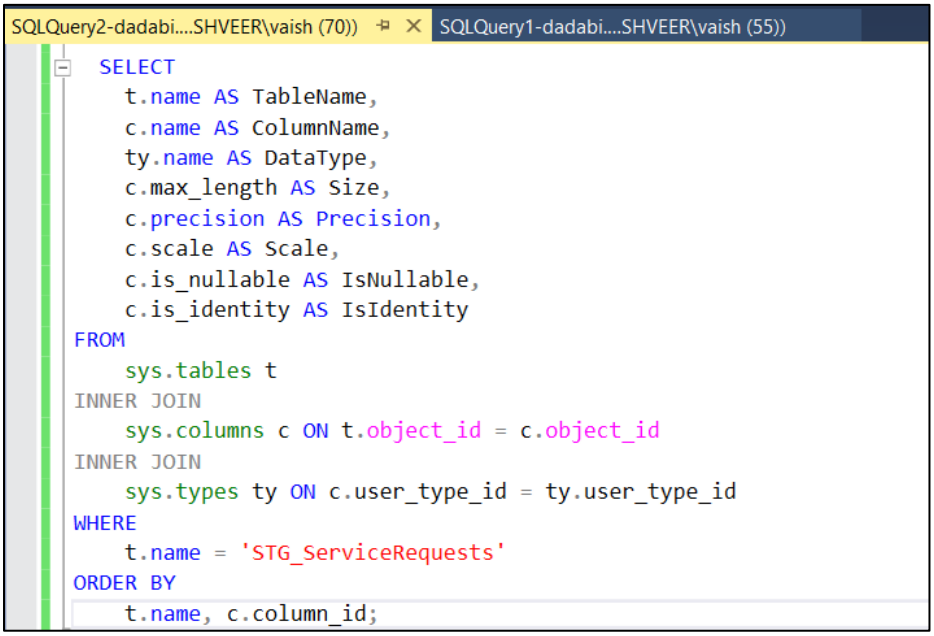


- The column is of type V_String with a data type size of 12
- There are **1,404,943 or about 89.88%, are null. This implies that this category may be optional or not applicable in many cases. We can replace the null values with a specific value or N/A using the formula tool.**

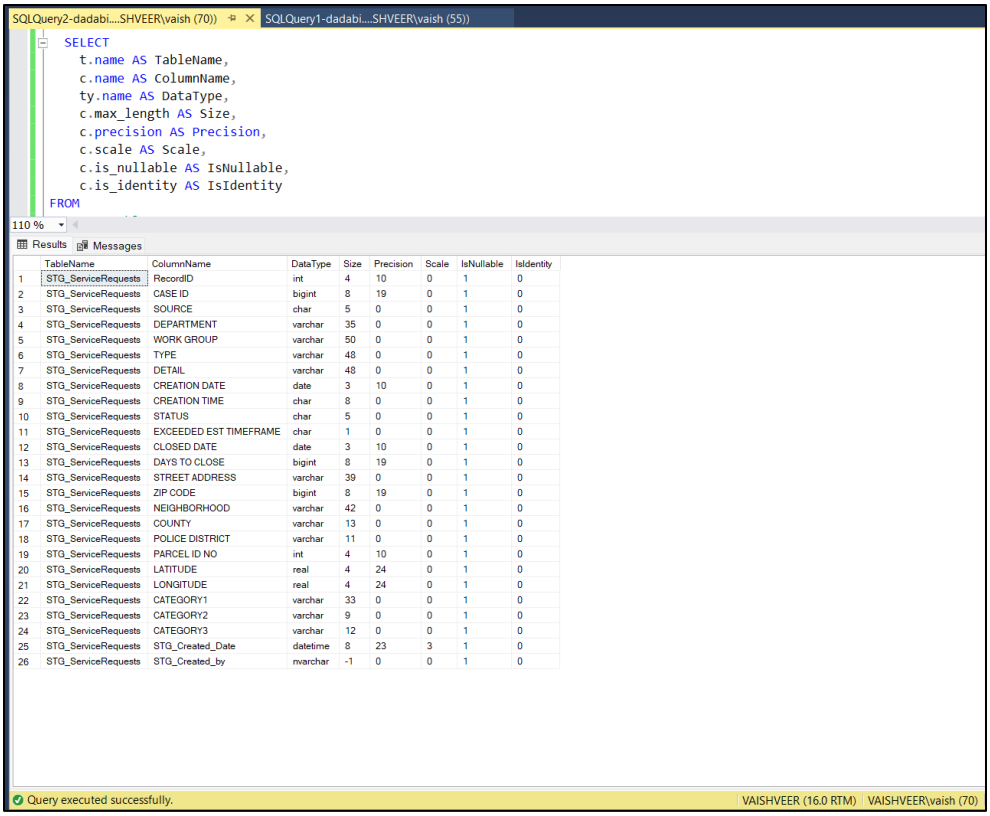
DATA POPULATED IN SQL SERVER



QUERY TO VIEW ALL COLUMNS AND THEIR DATATYPES




```
SELECT
    t.name AS TableName,
    c.name AS ColumnName,
    ty.name AS DataType,
    c.max_length AS Size,
    c.precision AS Precision,
    c.scale AS Scale,
    c.is_nullable AS IsNullable,
    c.is_identity AS IsIdentity
FROM
    sys.tables t
INNER JOIN
    sys.columns c ON t.object_id = c.object_id
INNER JOIN
    sys.types ty ON c.user_type_id = ty.user_type_id
WHERE
    t.name = 'STG_ServiceRequests'
ORDER BY
    t.name, c.column_id;
```



```
SELECT
    t.name AS TableName,
    c.name AS ColumnName,
    ty.name AS DataType,
    c.max_length AS Size,
    c.precision AS Precision,
    c.scale AS Scale,
    c.is_nullable AS IsNullable,
    c.is_identity AS IsIdentity
FROM
```

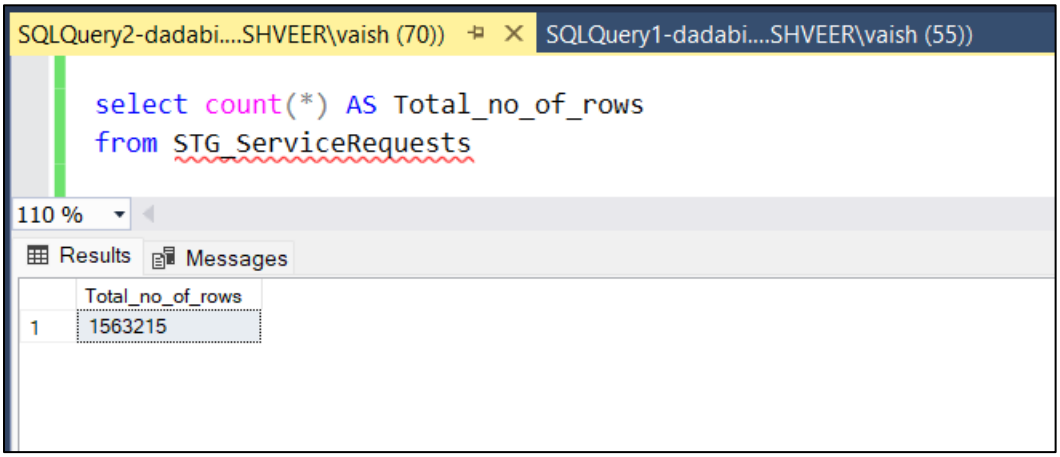
110 %

Results | Messages

	TableName	ColumnName	DataType	Size	Precision	Scale	IsNullable	IsIdentity
1	STG_ServiceRequests	RecordID	int	4	10	0	1	0
2	STG_ServiceRequests	CASE ID	bigint	8	19	0	1	0
3	STG_ServiceRequests	SOURCE	char	5	0	0	1	0
4	STG_ServiceRequests	DEPARTMENT	varchar	35	0	0	1	0
5	STG_ServiceRequests	WORK GROUP	varchar	50	0	0	1	0
6	STG_ServiceRequests	TYPE	varchar	48	0	0	1	0
7	STG_ServiceRequests	DETAIL	varchar	48	0	0	1	0
8	STG_ServiceRequests	CREATION DATE	date	3	10	0	1	0
9	STG_ServiceRequests	CREATION TIME	char	8	0	0	1	0
10	STG_ServiceRequests	STATUS	char	5	0	0	1	0
11	STG_ServiceRequests	EXCEEDED EST TIMEFRAME	char	1	0	0	1	0
12	STG_ServiceRequests	CLOSED DATE	date	3	10	0	1	0
13	STG_ServiceRequests	DAYS TO CLOSE	bigint	8	19	0	1	0
14	STG_ServiceRequests	STREET ADDRESS	varchar	39	0	0	1	0
15	STG_ServiceRequests	ZIP CODE	bigint	8	19	0	1	0
16	STG_ServiceRequests	NEIGHBORHOOD	varchar	42	0	0	1	0
17	STG_ServiceRequests	COUNTY	varchar	13	0	0	1	0
18	STG_ServiceRequests	POLICE DISTRICT	varchar	11	0	0	1	0
19	STG_ServiceRequests	PARCEL ID NO	int	4	10	0	1	0
20	STG_ServiceRequests	LATITUDE	real	4	24	0	1	0
21	STG_ServiceRequests	LONGITUDE	real	4	24	0	1	0
22	STG_ServiceRequests	CATEGORY1	varchar	33	0	0	1	0
23	STG_ServiceRequests	CATEGORY2	varchar	9	0	0	1	0
24	STG_ServiceRequests	CATEGORY3	varchar	12	0	0	1	0
25	STG_ServiceRequests	STG_Created_Date	datetime	8	23	3	1	0
26	STG_ServiceRequests	STG_Created_by	nvarchar	-1	0	0	1	0

Query executed successfully. VAISHVEER (16.0 RTM) VAISHVEER\vaish (70)

ROW COUNT USING SQL QUERY: A total of 1563215 rows are present in the table which is the same count as the count in Alteryx



```
select count(*) AS Total_no_of_rows
from STG_ServiceRequests
```

110 %

Results | Messages

	Total_no_of_rows
1	1563215