Food Inspections Assignments (Final Project)

GROUP 7: Je Sai Kailash Pulipati (002339774)
Deepthi Ramesh (002339269)
Geetika Barla (002372565)

DATASET OVERVIEW:

Chicago Food Inspections Dataset

This dataset contains information from food inspections conducted in the City of Chicago, beginning January 1, 2010, to the present. These inspections are carried out by the Chicago Department of Public Health's (CDPH) Food Protection Program using a standardized inspection process. All inspection results are entered into a centralized database and reviewed by a licensed Environmental Health Practitioner (LEHP) certified by the State of Illinois.

Each record in the dataset typically includes:

- Establishment details (e.g., name, address, license number)
- Inspection date and type (e.g., complaint, routine, license)
- Inspection result (Pass, Pass with Conditions, or Fail)
- Violations identified during the inspection (if any), with relevant codes and descriptions

Dallas Food Establishment Inspection Dataset

This dataset provides historical inspection data for food establishments in Dallas. It has been **sunset** and is **no longer being updated**.

The dataset captures key information from food safety inspections, including:

- Name of the establishment
- Physical location (address)
- Date the inspection was conducted
- Overall inspection score
- Point deductions for individual violations observed

DATASET FILES AND THEIR CONTENTS:

CHICAGO DATASET

Chicago dataset contains approximately 290,000 rows and 17 columns, where each row represents a food inspection. The unique identifier for each inspection is the Inspection ID.

Column Name	Description	Datatype
Inspection ID	Unique identifier for the dataset	Number
DBA Name	Doing business as	Text
AKA Name	Also known as	Aka_name
Lincense#	Business license number issued to the establishment	Number

E 11: E		T
Facility Type	Type of facility (e.g., restaurant, school, grocery store)	Text
Risk	Risk level classification (e.g., Risk 1 - High, Risk 2 - Medium, Risk 3 - Low)	Text
Address	Full address of the establishment	Text
City	City where the facility is located	Text
State	State abbreviation (typically "IL" for Illinois)	Text
#Zip	ZIP code of the establishment location	Number
Inspection Date	Date and type of inspection occured	Floating Timestamp
Inspection Type	Type of inspection conducted	Text
Results	Outcome of insepctions	Text
Violation	Details of violation observed	Text
Latitude	Geographic coordinate latitude	Number
Longitude	Geographic coordinate longitude	Number
Location	Location of restaurant	Location

DALLAS DATASET

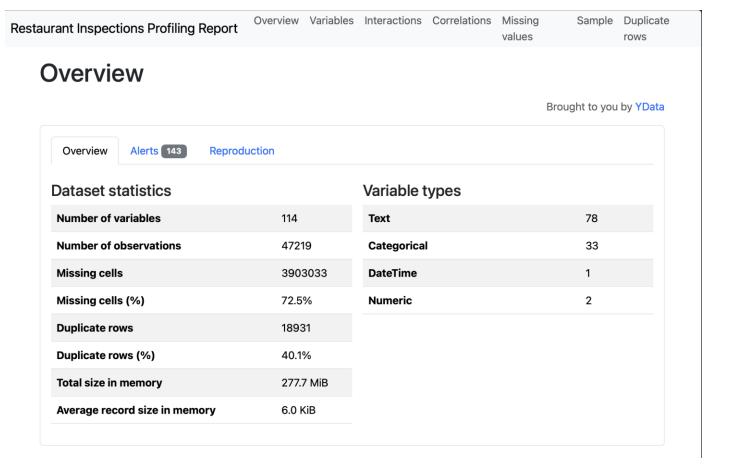
Dallas Dataset contains approximately 79,000 rows and 114 columns, where each row represents a facility inspection.

Column	Description	Datatype
Restaurant Name	Name of the food establishment	Text
Inspection Type	Type of inspection conducted (e.g., routine, complaint)	Text
Inspection Date	Date when the inspection was performed	Date
Inspection Score	Numeric score assigned during the inspection	Number
Street Number	Street number of the establishment's address	Text

	T	
Street Name	Street name of the establishment's address	Text
Street Direction	Street direction (e.g., N, S, E, W)	Text
Street Type	Street type (e.g., Ave, Blvd, St)	Text
Street Unit	Unit number or identifier (if applicable)	Text
Street Address	Full street address of the establishment	Text
Zip Code	ZIP code of the establishment's location	Text
Violation Description 1- 25	Description of the specific violation observed during the inspection	Text
Violation Points 1-25	Points deducted for the specific violation	Number
Violation Details 1-25	The type of violation associated with the enforcement action	Text
Violaton Memo 1-25	Any additional comments about the enforcement action.	Text
Inspection Month	Month in which inspection was	Text
Inspection Year	Year in which inspection was conducted	Text
Lat Long Location	Denotes a location point on a longitude line (perpendicular to the equator) and latitude line (parallel to the equator)	Location

DATA PROFILING AND ANALYSIS:

DALLAS DATASET:



The dataset comprises a large volume of restaurant inspection records and was profiled using YData's profiling tool. The data set includes over 79,000 rows and 114 columns, with each row representing a unique inspection of a food facility in Dallas.

General Statistics

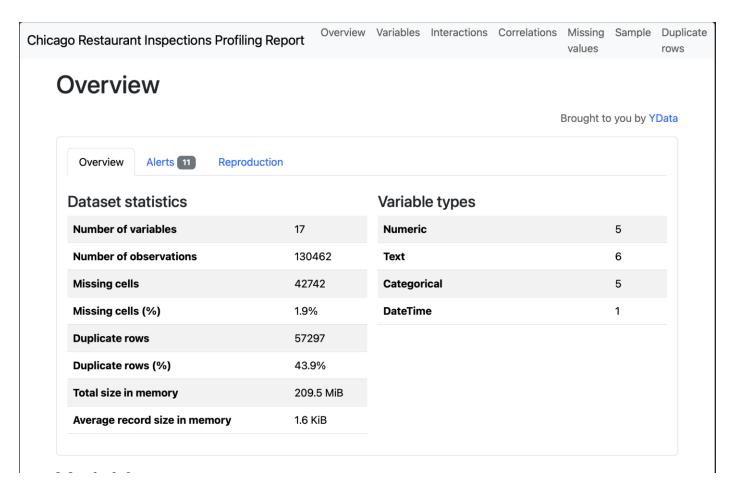
- Total Records: ~79,000
- Columns: 114
- Missing Data: A notable proportion of the columns contains missing values. However, critical columns such as restaurant name and inspection date are mostly complete.
- Duplicates: Minimal to no duplicate rows detected.

Data Quality and Types

- The dataset consists of:
- · Categorical fields such as Restaurant Name, City, and Inspection Type
- Numerical fields like Inspection Score and Violation Points
- Date fields like Inspection Date
- Inconsistencies were detected in fields like ZIP codes (some stored as floats with .0) and city names, requiring standardization.

Key Findings

- High Cardinality: Some columns like Violation Description and Facility Address show high uniqueness, suggesting they should be parsed or categorized if used in models.
- Violation Parsing: Violation-related data is embedded in pipe (|) and comma-separated formats and needs to be exploded or parsed into structured fields for analysis.
- Whitespace and Formatting Issues: Trailing and leading whitespaces are present in some text fields, which can be cleaned using standard text-cleaning techniques.



General Statistics

• Total Records: 14,692

• Columns: 17

• Missing Data: 32,537 cells (~12.86% of total values)

•. Duplicate Rows: 21 (≈0.14%)

Data Quality and Types

- Column Types:
- Categorical: 8
- Text (String):
- Numeric: 3
- Datetime: 2

Cuisine & Inspection Data

- CUISINE DESCRIPTION:
- Most frequent entries: American, Chinese, Pizza, Latin, and Mexican cuisines.
- INSPECTION DATE:
- Complete column.
- Covers inspections from January 2016 to May 2018.
- SCORE:

Ranges from 0 to 188.

Average score ≈ 15.45 .

Contains outliers that may require treatment.

Key Findings

• Missing Data:

Columns like GRADE, VIOLATION CODE, and VIOLATION DESCRIPTION have significant null values.

• Outliers:

Detected in numeric fields like SCORE.

• Violation Information:

Stored in pipe-separated format; needs parsing into Violation Code, Violation Description, and Violation Comment.

• Whitespace & Formatting:

Presence of leading/trailing whitespace in text fields.

ZIP codes stored with .0 require formatting.

GRAIN:

One row per individual violation recorded during an inspection event at a restaurant-location on a specific date

Each row in the FACT_INSPECTION table represents a single violation associated with a specific restaurant's inspection at a particular location and on a specific date. If multiple violations occur in one inspection, there will be multiple rows for that inspection — one for each violation

Explanation:

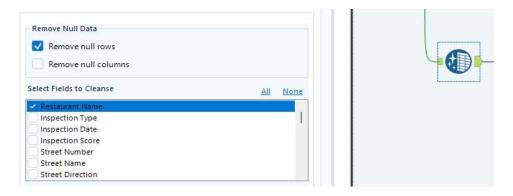
Each row in the FACT_INSPECTION table represents a single violation associated with a specific restaurant's inspection at a particular location and on a specific date. If multiple violations occur in one inspection, there will be multiple rows for that inspection — one for each violation

DATA CLEANING:

Performed data cleaning in alteryx for both Dallas and Chicago datasets

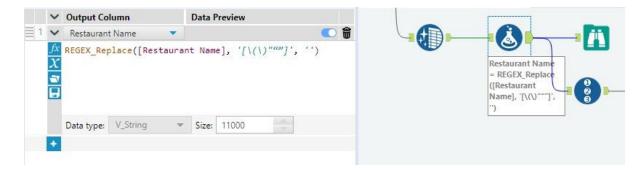
Dallas Dataset Cleaning:

1. Removed Null Records:



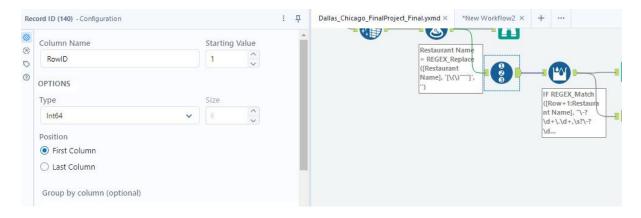
Eliminated all rows where the restaurant name was missing.

2. Cleaned Restaurant Names:



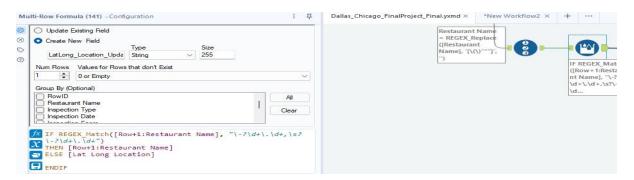
Used REGEX to remove unwanted characters from the Restaurant Name field.

3. Created Unique Identifiers:



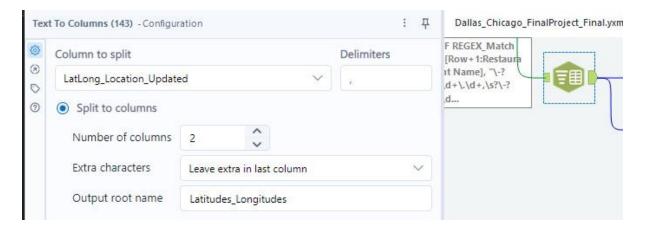
Added a new column for Row ID or Record ID for unique identification of each entry.

4. Extracted Embedded Numbers:



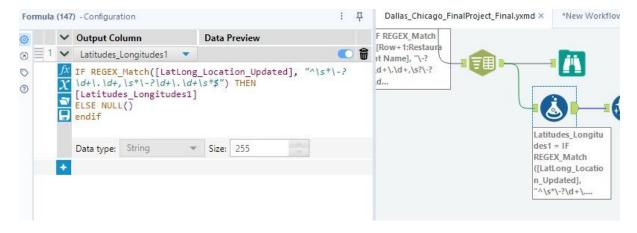
Removed numeric values embedded in restaurant names and reassigned them appropriately to latitude and longitude fields.

5. Split Comma-Separated Values:



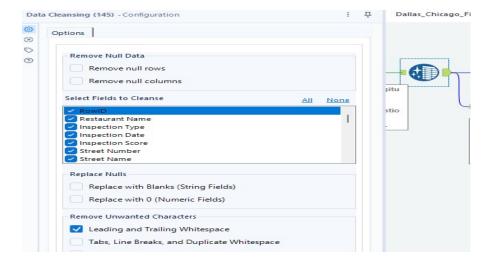
Split columns that contained comma-separated values into multiple separate columns for better structure.

6. Latitude and Longitude Columns:



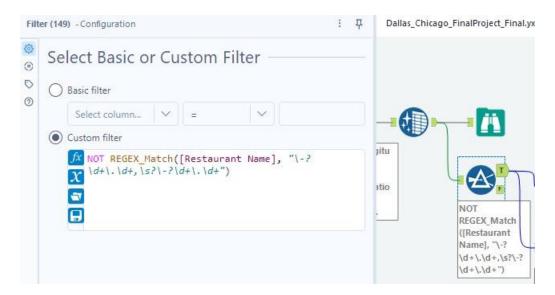
Ensured separate columns were created for both Latitude and Longitude.

7. Whitespace Removal:



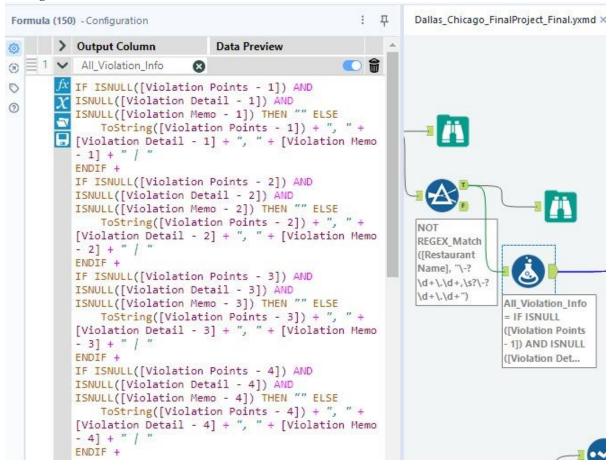
Used a data cleaning tool to remove leading and trailing white spaces from all relevant text fields.

8. Filtered Non-Standard Names:



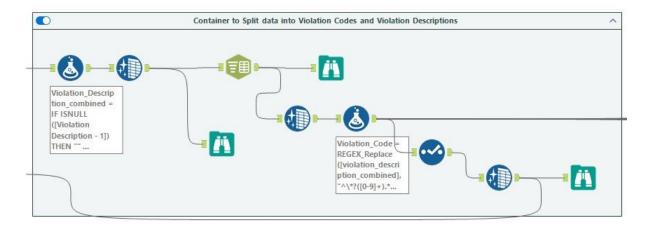
Used the Filter tool to exclude restaurant names containing numbers, retaining only standard textual names.

9. Merged Violation Data



Used the Formula tool to combine multiple violation sets (up to 25 per row), where each set is pipe (|) separated and within the set individual violations are comma-separated.

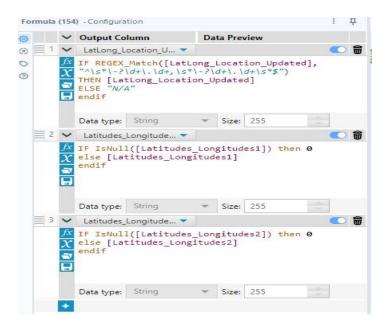
10. Parsed Violation Details:



Split Violation Code, Violation Description, and Violation Comment into individual, dedicated columns.

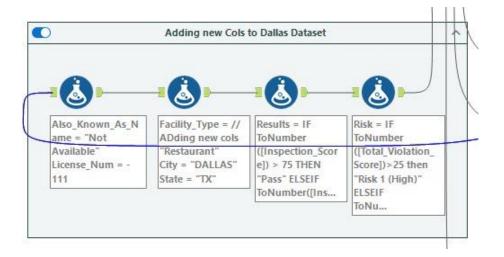
11. Handled Missing Data:





Replaced all null values across every column with N/A to maintain consistency.

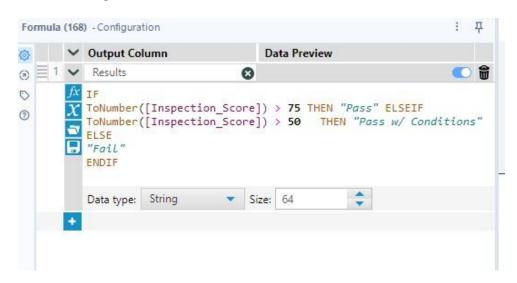
12. Added New Columns:



Used the Formula tool to add columns such as:

AKA Name, License Number, Facility Type, City, State, Default values were provided where applicable.

13. Derived Inspection Result:



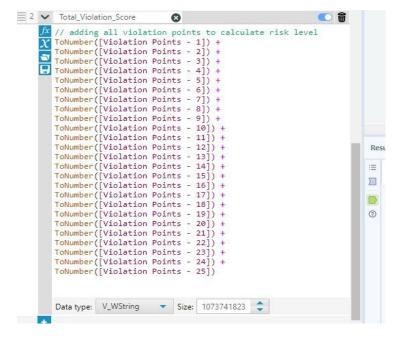
Created a Result column based on Inspection Scores.

14. Calculated Risk Level:



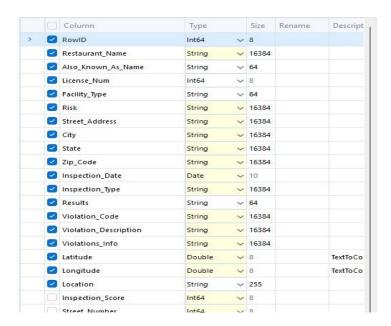
Added a Risk column, with levels determined by the Total Violation Score per row.

15. Computed Total Violation Score:



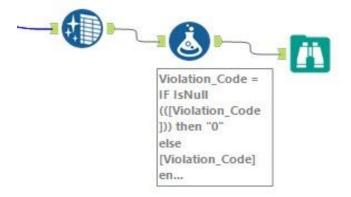
Introduced a new column that sums up all violations associated with a row.

16. Standardized Data Types:



Converted column data types as necessary to ensure consistency and proper formatting.

17. Final Whitespace Cleanup:



Performed a final pass to eliminate any residual white space in text-based columns.

These were the data cleaning done in the dallas dataset, to make sure that the data does not have any problems and is fit to perform future analaysis, based on our requirements.

Chicago Dataset Cleaning:

1. Removed Null Inspection Records:



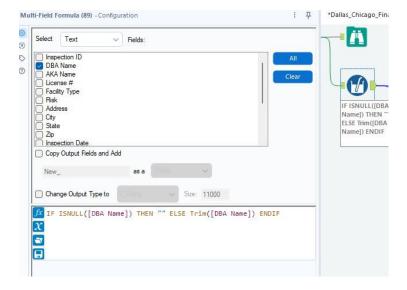
Dropped all rows where the Inspection ID was null.

2. Removed Duplicate Inspections:



Removed duplicated rows based on the Inspection ID field to ensure uniqueness.

3. Trimmed Extra Spaces:



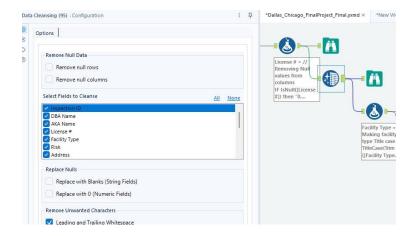
Used a Multi-Field Formula tool to remove leading and trailing spaces from the DBA Name (Doing Business As) column

4. Handled Missing Values in Other Columns:



Replaced or removed null values across other columns as needed.

5. Whitespace Cleanup:



Stripped leading and trailing whitespaces from all relevant fields.

6. Split Risk Column:

Separated the combined All Risk column into three individual columns: Risk 1, Risk 2, and Risk 3.

7 . Standardized City Names:

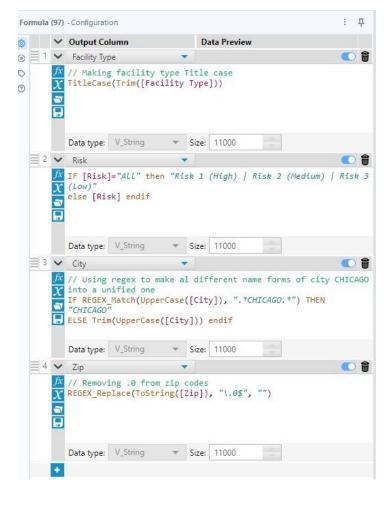
Applied Regex to normalize various representations of "Chicago" into a unified city name.

8. Formatted ZIP Codes:

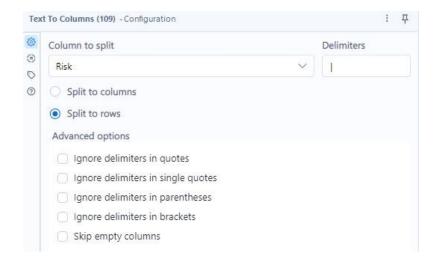
Removed trailing .0 from ZIP code entries for consistency.

9. Formatted Facility Types:

Converted all entries in the Facility Type column to Title Case.

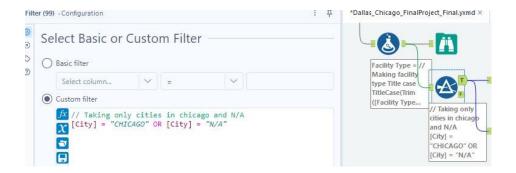


10. Flattened Risk Values:



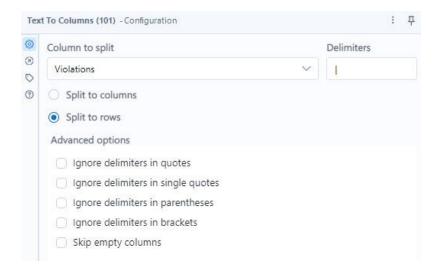
Used the Split to Rows tool to explode pipe-separated risks into individual rows.

11. Filtered for Chicago Only:



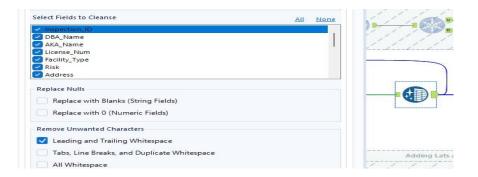
Removed records from cities other than Chicago using the Filter tool.

12. Split Violations into Rows:



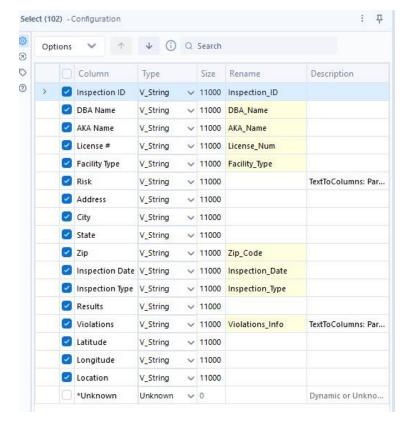
Used Text to Columns to split pipe-separated violation entries into multiple rows.

13. Final Whitespace Cleanup:



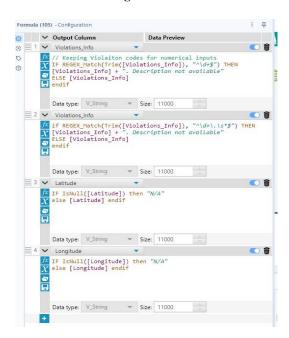
Performed another pass to eliminate any residual white space in text fields.

14. Cleaned Column Names:



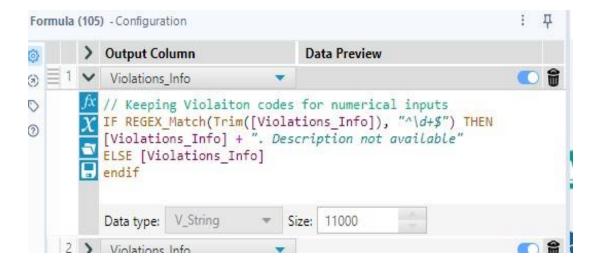
Used the Select tool to strip spaces from column headers.

15. Handled Missing Coordinates:



Used the Formula tool to replace null values in Latitude and Longitude with N/A.

16. Filled Missing Violation Descriptions:



For violation codes with only numeric values and no description, added "Description not available".

17. Data Type Conversion:

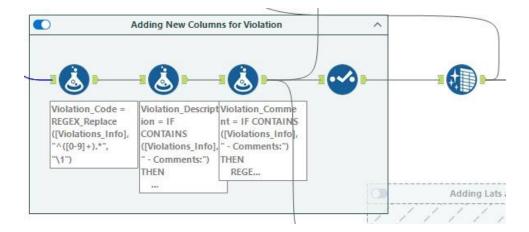
Changed column data types to appropriate formats.

18. Split Violation Details

Parsed violation information into three separate columns Violation Code, Violation Description and Violation Comment

19. Final White Space Removal:

Ensured all text-based columns are free of unnecessary spaces.



FURTHER STEPS:

- We began with data profiling and cleaning to assess data quality and address inconsistencies.
- After that, we created raw tables and uploaded them to the Snowflake data warehouse (Bronze layer) using SnowSQL commands.



• We then connected Databricks to Snowflake and loaded the stage tables into the Silver layer using Databricks.

```
df_dallas = spark.read.parquet("/FileStore/tables/Dallas_Cleaned.parquet")
           df_chicago = spark.read.parquet("/FileStore/tables/Chicago_Cleaned.parquet")
   Data Preview
D ~
        1 df_dallas.printSchema()
           df dallas.show(5)
           df_chicago.printSchema()
      |-- RowID: long (nullable = true)
       -- Restaurant_Name: string (nullable = true)
      -- Also_Known_As_Name: string (nullable = true)
       -- Risk: string (nullable = true)
       -- Street_Address: string (nullable = true)
       -- State: string (nullable = true)
       -- Zip_Code: string (nullable = true)
       -- Inspection_Type: string (nullable = true)
       -- Violations_Info: string (nullable = true)
       -- Longitude: double (nullable = true)
```

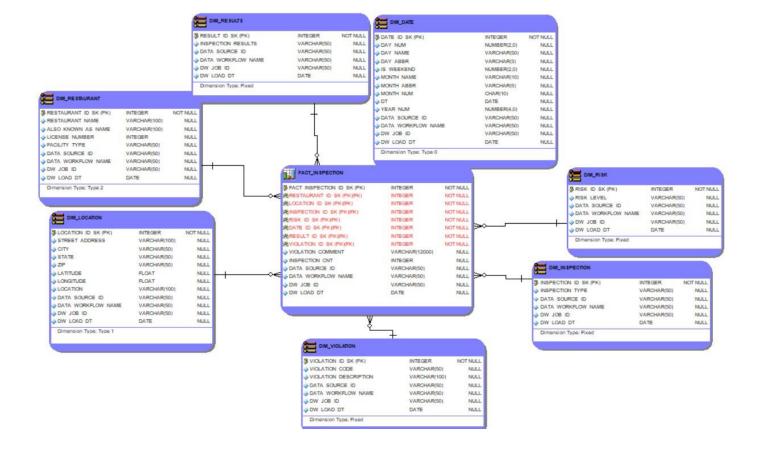
```
Snowflake Connection

1     sfOptions = {
2         "sfURL" : "HYQMYKK-SXA08708.snowflakecomputing.com",
3         "sfPassword" : "snowflake123#",
4         "sfPassword" : "snowflake123#",
5         "sfDatabase" : "CD_DB",
6         "sfSchema" : "CD_SCHEMA",
7         "sfWarehouse" : "CD_WH",
8         "sfRole": "CD_ROLE"
9     }

1     # Write Dallas dataset to Snowflake
2     df_dallas.write \
3          .format("snowflake") \
4          .options(**sfOptions) \
5          .option("dbtable", "STAGE_DALLAS_INSPECTIONS") \
6          .mode("overwrite") \
7          .save()
8     # Write Chicago dataset to Snowflake
10     df_chicago.write \
11          .format("snowflake") \
12          .options(**sfOptions) \
13          .option("dbtable", "STAGE_CHICAGO_INSPECTIONS") \
14          .mode("overwrite") \
15          .save()
16
```

• Designed the dimensional data model using ER/Studio

Dimensional Model Diagram:

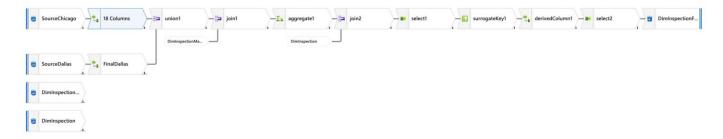


- Generated the physical data model based on the previously created logical model.
- We generated SQL scripts to create dimension and fact tables in the Snowflake database.
- We then Loaded dimension and fact tables into Snowflake using Azure Data Factory (Gold layer).

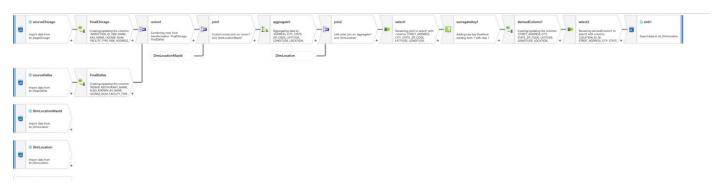
LOADING DIMENSIONS AND FACTS IN SNOWFLAKE USING ADF:

LOADING DIMENSIONS:

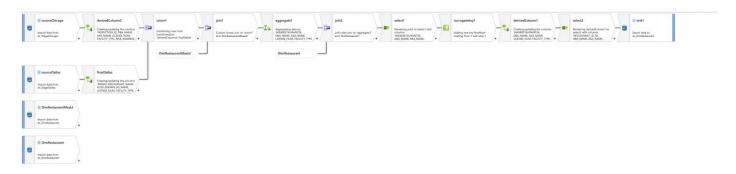
DIM INSPECTION



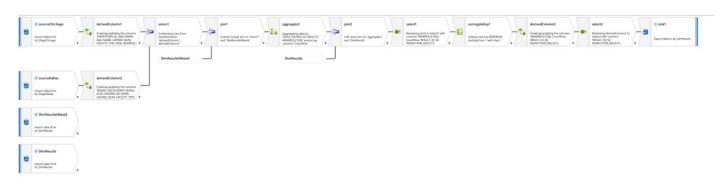
DIM_LOCATION



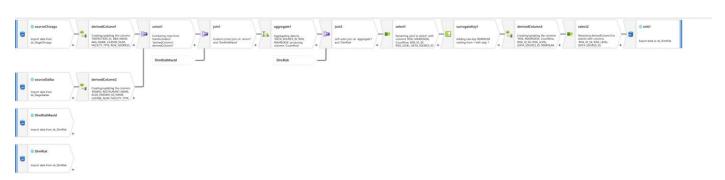
DIM_RESTAURANT



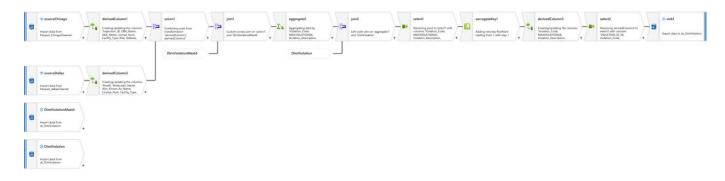
DIM_RESULT:



DIM_RISK:

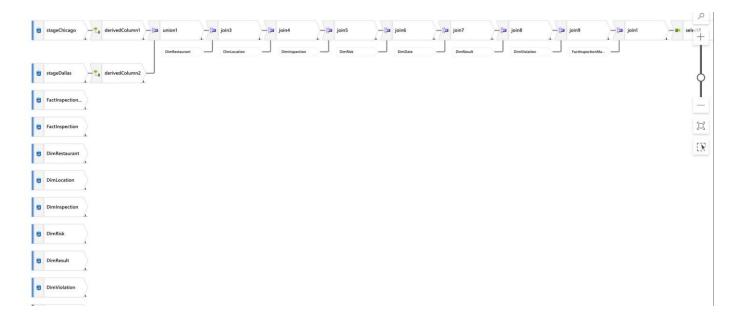


DIM_VIOLATION



LOADING FACT

FACTINSPECTION:



LOADING DIM AND FACTS PIPELINE:



• We then created the necessary visualizations using PowerBI

