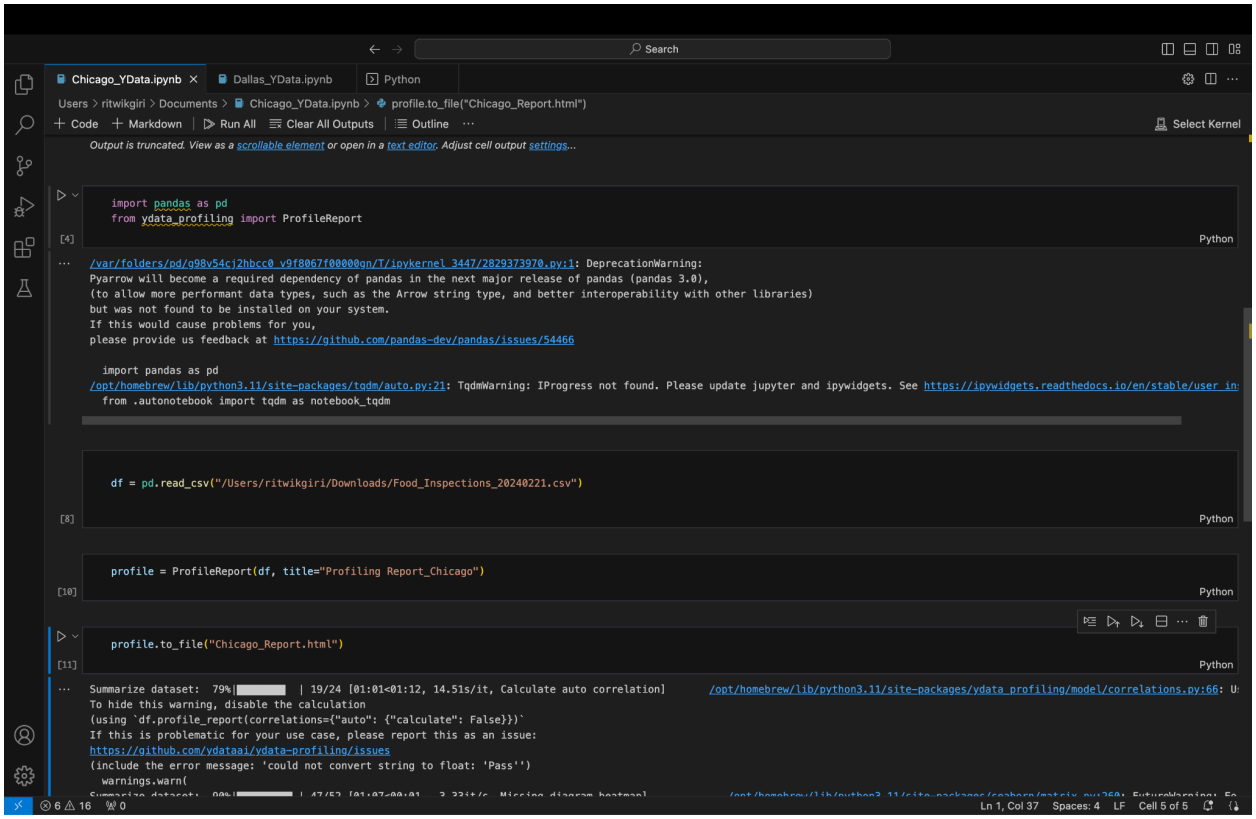
Midterm Project Report

Data Profiling: Chicago



Handling Multi-Valued Attributes and Data Type Conversion:

Multi-Valued Attributes Handling:

Identification: Through dataset examination, we can pinpoint columns containing multi-valued attributes. These are usually in columns listing items separated by delimiters, like violation codes or service categories.

Evaluation: Understand the importance of each multi-valued attribute within your analysis framework. Determine if it's necessary to disaggregate these attributes for detailed analysis or if they can be analyzed as a whole.

Transformation Strategies: For columns requiring disaggregation, select a suitable method. Options include:

Column Splitting: If values are few and uniform, splitting into distinct columns may work.
Row Expansion: To analyze values individually, transform each value into a separate row, replicating other column data for context.

Aggregation: Summarizing these attributes could be useful, particularly for compliance or violation data, through methods like counting occurrences or statistical analysis.

Dataset Structure: Our dataset doesn't have traditional multi-valued attributes but rather distributes violation information across various columns in a 'wide' format.

Violation Data Handling: To analyze violations, restructure the data so each violation has its own row, maintaining a link to its inspection.

Data Type Conversion:

- Inspection Date: Convert string-based dates to datetime format for easier chronological analysis.
- Numeric Data: Inspection scores and street numbers are already integers. For columns like inspection year, convert to integers or maintain as is, based on analysis needs. Latitude and longitude should be split into numeric columns for geospatial analysis. Convert any numeric values stored as strings to their appropriate numeric types.
- Categorical Data: Convert columns representing categories into Pandas' categorical type for efficiency and faster category operations.

Conversion Processes:

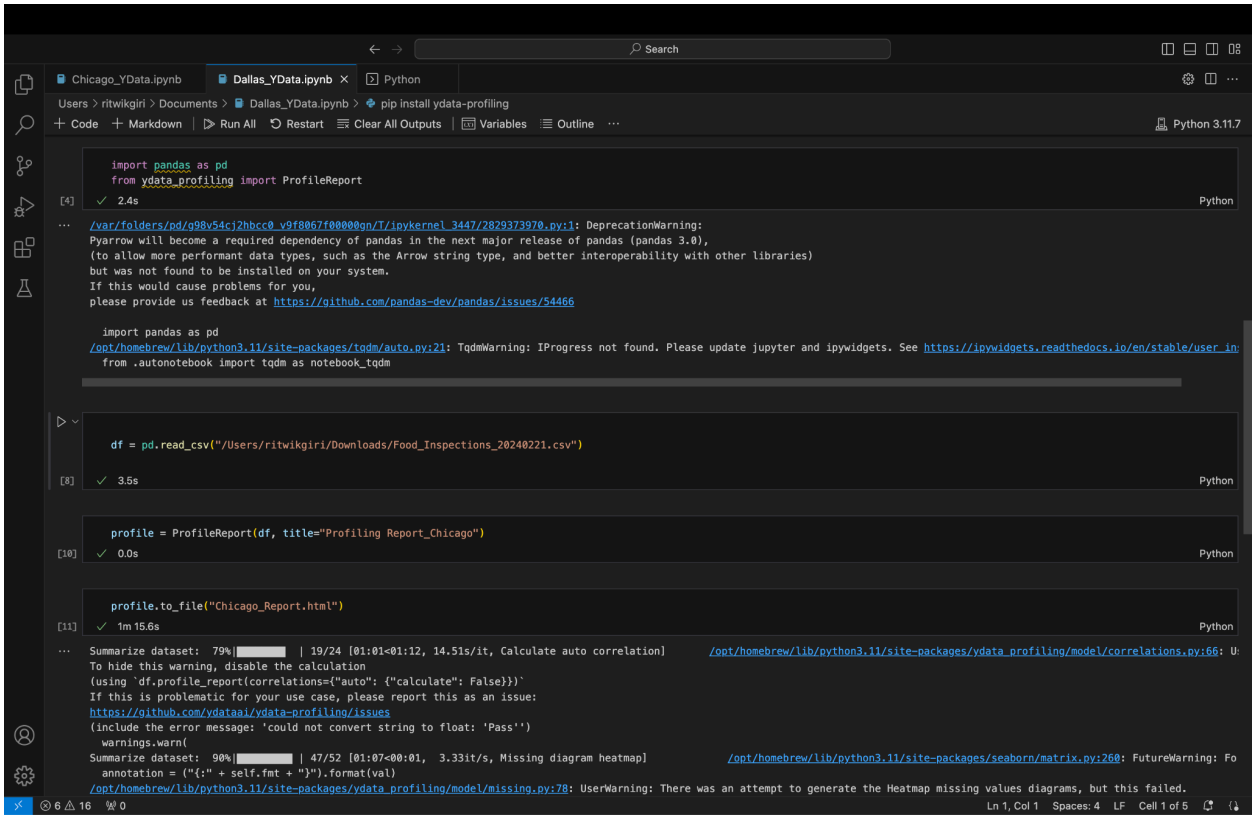
Date and Time: Change any date or time columns from strings to Python's datetime format to enhance time-series analysis and manipulation.

Numerical Conversion: Adjust columns with numeric values stored as strings to the correct numeric data types for accurate analysis.

Categorical Conversion: Change columns with categorical data to Pandas' categorical type to save on memory and simplify category-based analysis.

Post-Conversion Validation: Conduct consistency checks post-transformation to ensure data integrity and correctness, using data validation techniques or summary statistics to verify the success of the conversions.

Data Profiling: Dallas



Handling Multi-Valued Attributes:

For columns with potential multi-valued attributes, such as violation lists or service offerings, the handling strategy will vary based on the analysis objectives:

Column Division: If the data within these attributes is uniform and limited, dividing it into distinct columns may facilitate easier analysis.

Row Expansion: When each value needs individual consideration, expand this data across multiple rows. Ensure to replicate other relevant column data for comprehensive context.

Data Summarization: Sometimes, summarizing the data (e.g., counting occurrences or detecting specific keywords) might offer more analytical value than detailed individual analysis.

Data Type Adjustments:

Adapting data types is crucial for enhancing data usability and accuracy in analysis, particularly for typical columns in inspection datasets:

Inspection Date:
Initial Type: Often stored as a string.
Adjustment: Convert to datetime to enable detailed time-series investigations.

Inspection Score:
Initial Type: May appear as a string due to data inconsistencies.
Adjustment: Convert to a numerical format (integer or float) to normalize data formatting.

Address Information (e.g., Street Number, Name):
Initial Type: Stored as strings.
Adjustment: Maintain as strings but standardize formatting and casing for uniformity.

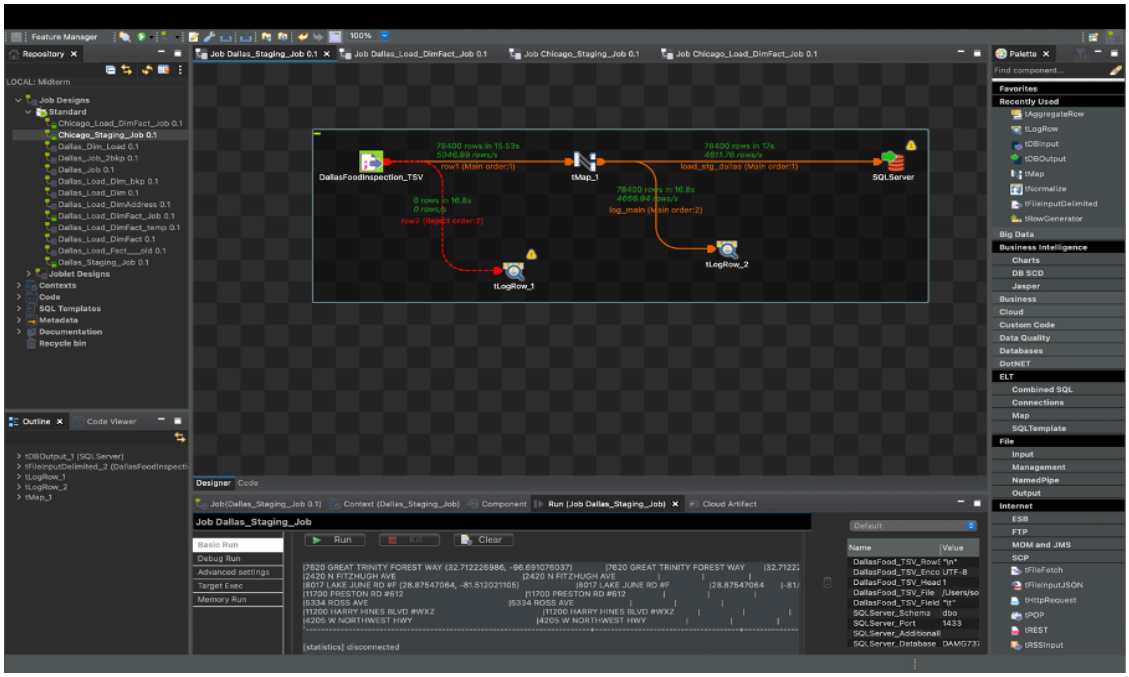
Violation Codes/Points:
Initial Type: Possibly mixed or strings, influenced by missing data or textual annotations.
Adjustment: Convert points to numerical formats when feasible. For violation codes with categorical or multi-valued characteristics, strategize their representation (e.g., dividing into columns or binary encoding) tailored to your analytical needs.

Categorical Columns (e.g., Inspection Type):
Initial Type: String format.
Adjustment: Change to categorical data type in pandas to boost efficiency and support category-based analyses.

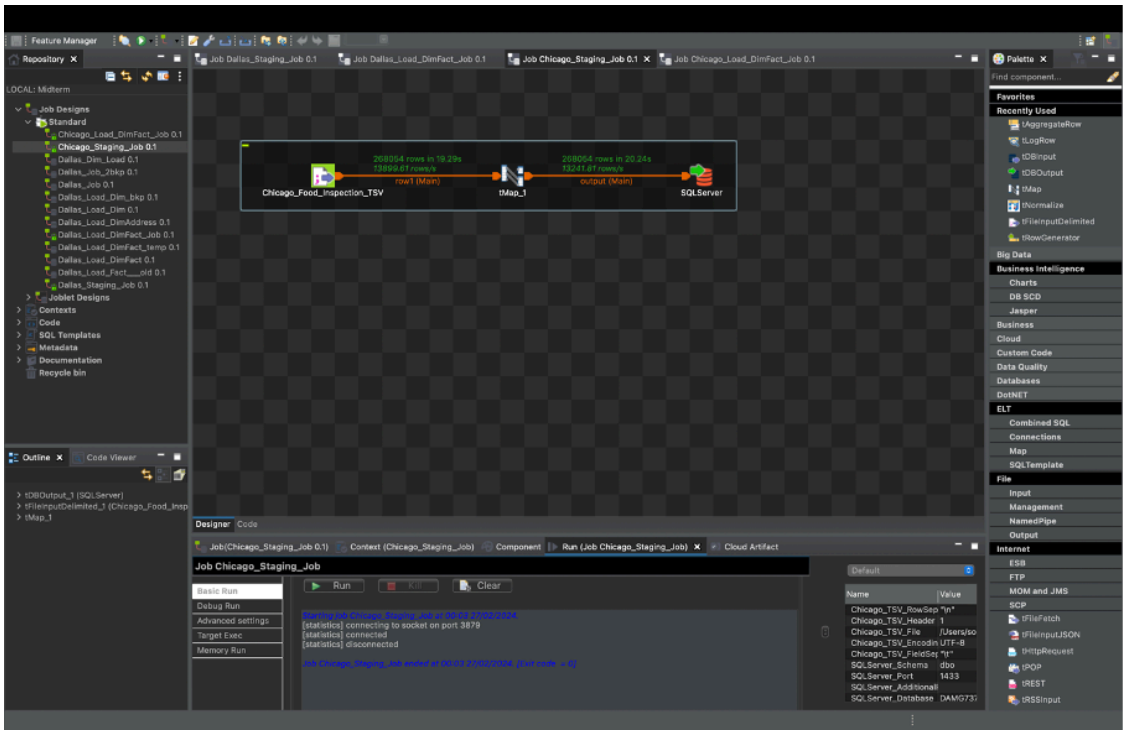
Geolocation Data (Latitude, Longitude):
Initial Type: Strings or a mixed format.
Adjustment: Convert to a double precision format to ensure precision for mapping or spatial analysis purposes.

Data Staging

Dallas data staging:



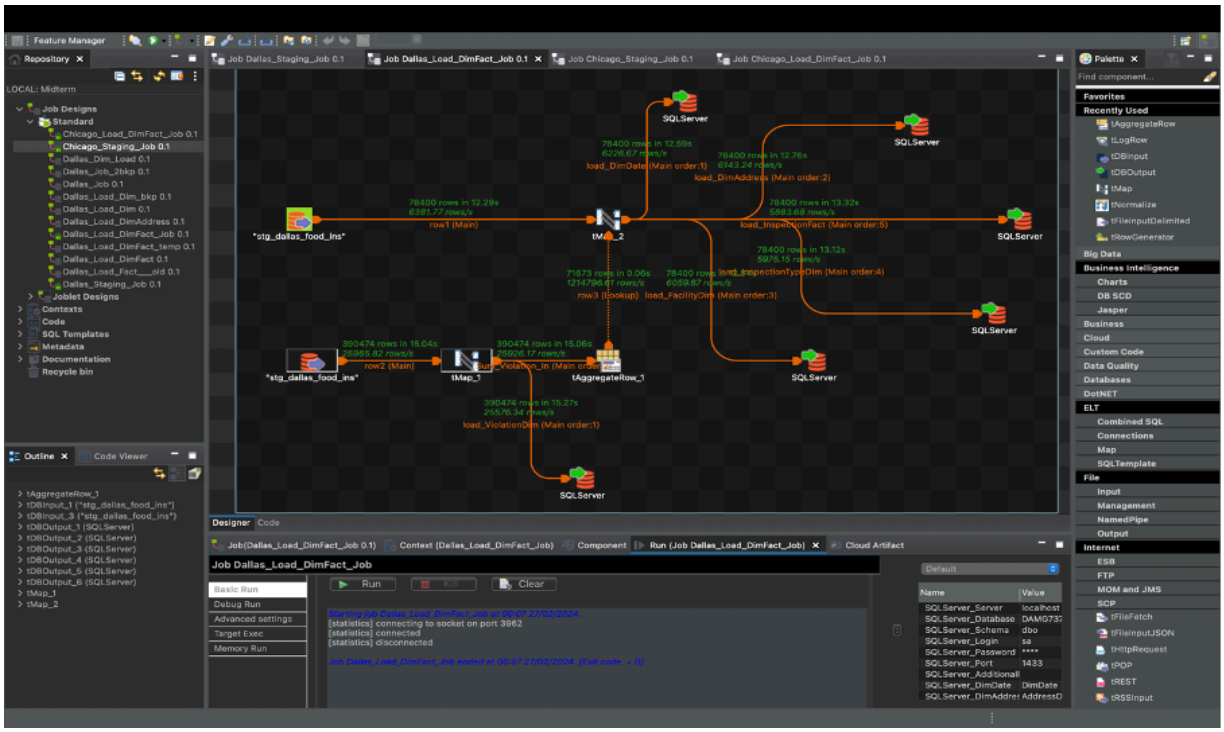
Chicago data staging:



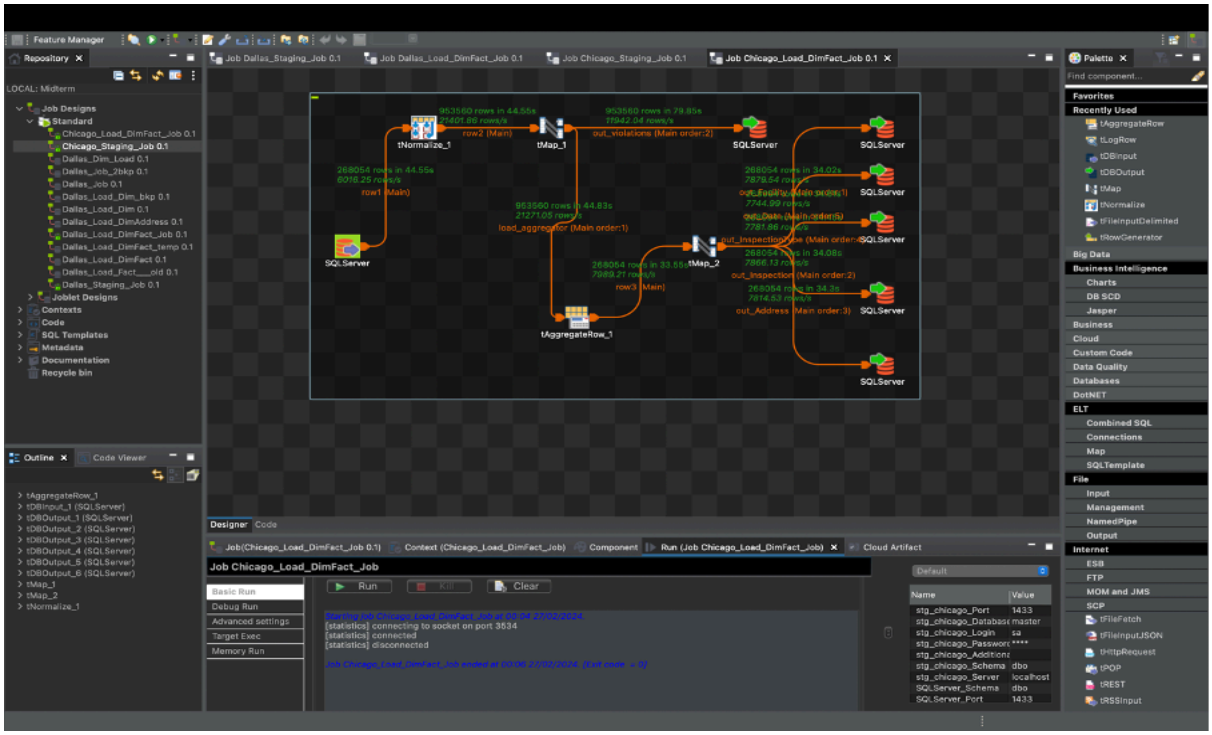
Dimensional Model

Talend Workflow

Dallas data loading:



Chicago data loading:

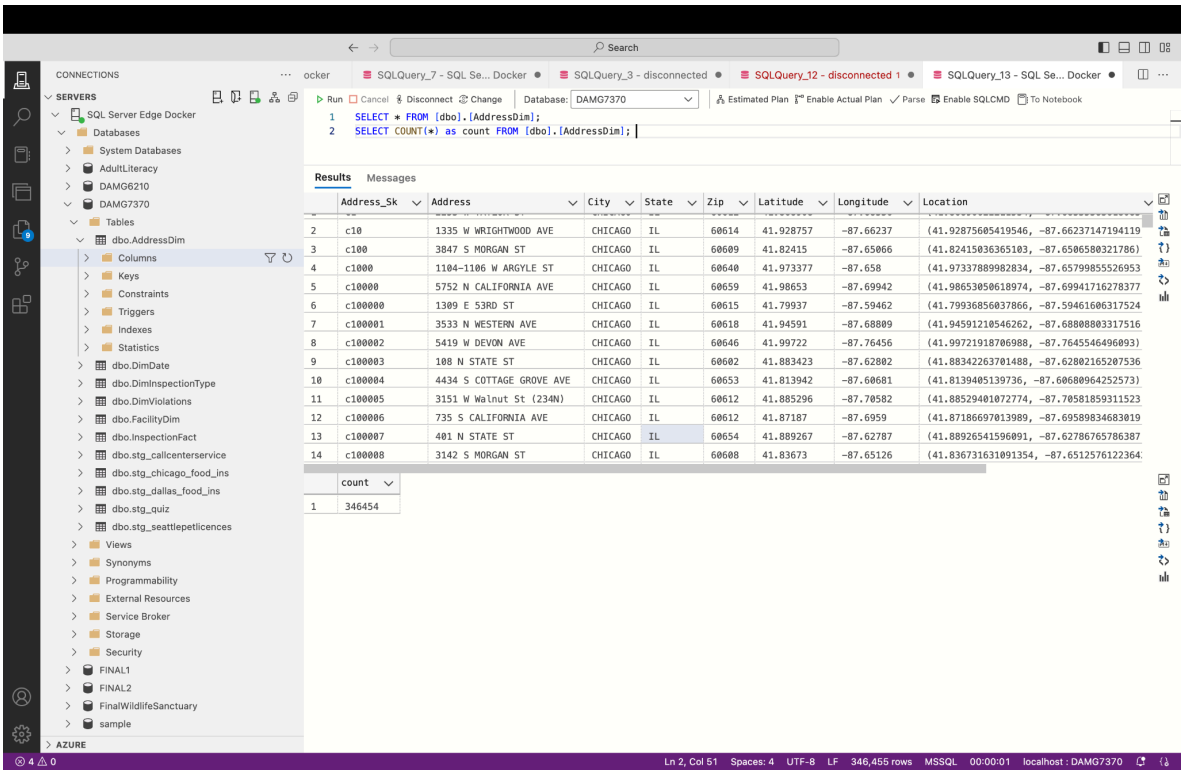


Data Architecture

Dimensions

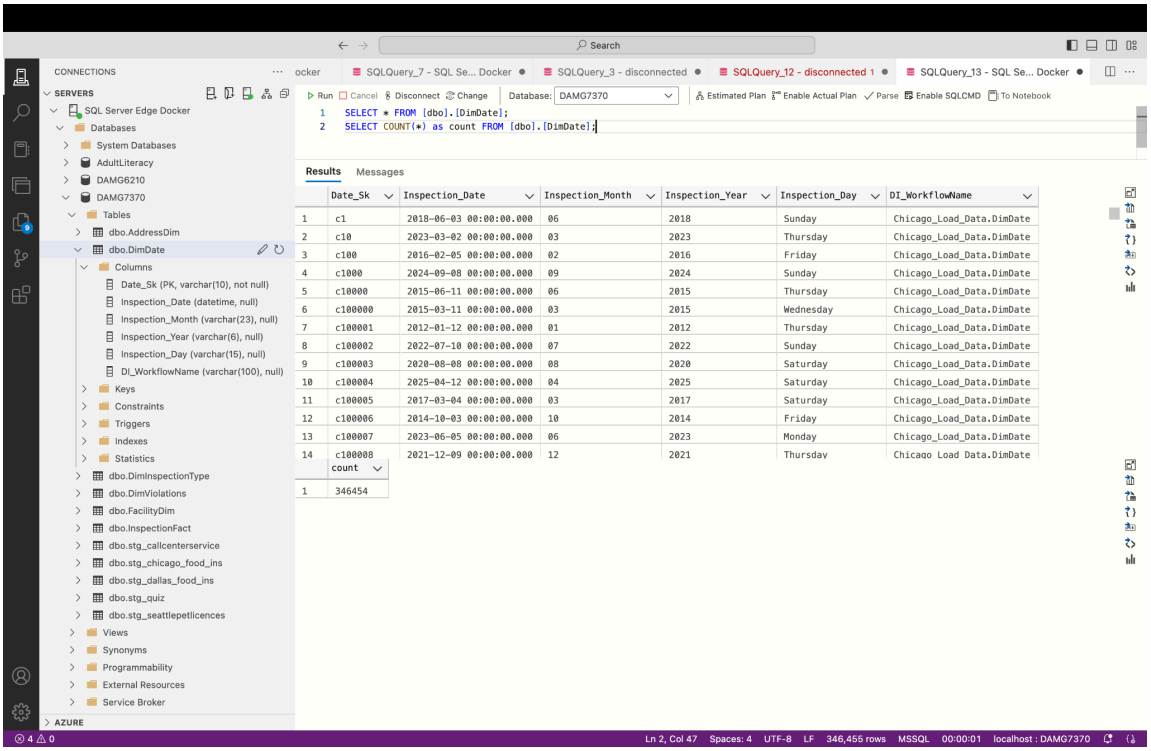
1. AddressDim

```
CREATE TABLE [dbo].[AddressDim] (
    [Address_Sk]       VARCHAR (10)  NOT NULL,
    [Address]          VARCHAR (200) NULL,
    [City]             VARCHAR (100) NULL,
    [State]            VARCHAR (10)  NULL,
    [Zip]              VARCHAR (10)  NULL,
    [Latitude]         VARCHAR (80)  NULL,
    [Longitude]        VARCHAR (80)  NULL,
    [Location]         VARCHAR (200) NULL,
    [DI_WorkflowName] VARCHAR (100) NULL,
    PRIMARY KEY CLUSTERED ([Address_Sk] ASC)
);
```



2. DimDate

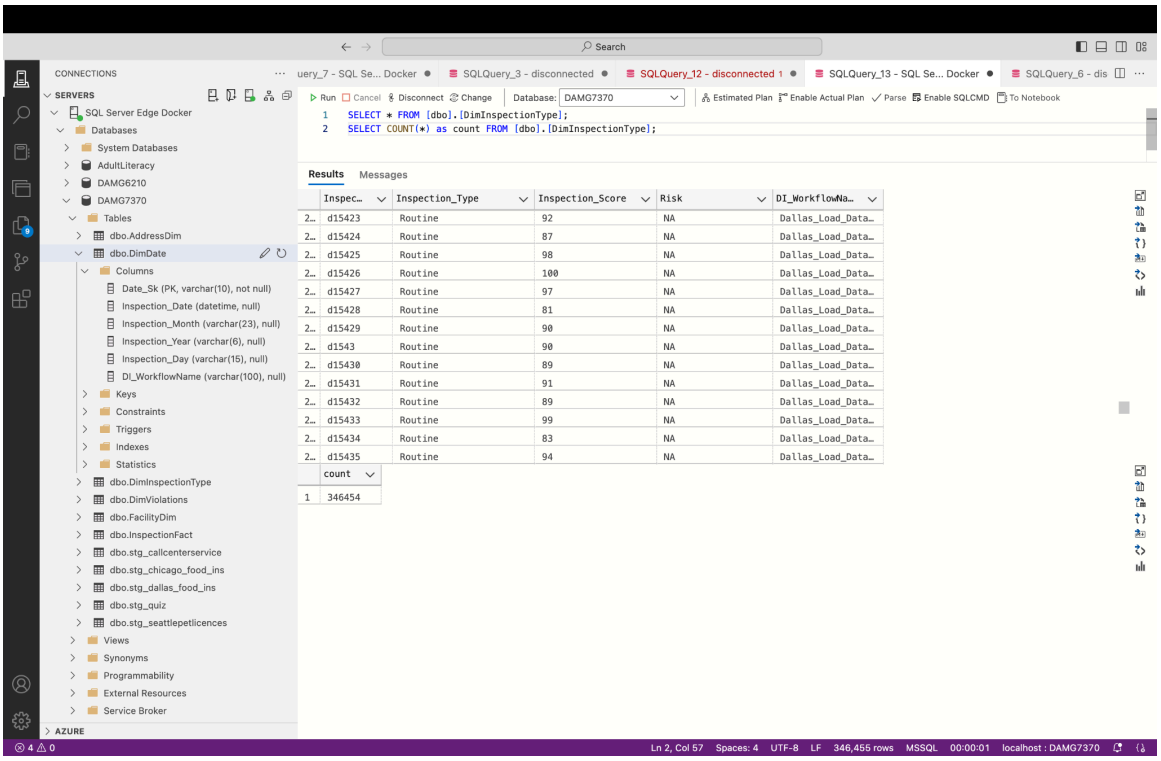
```
CREATE TABLE [dbo].[DimDate] (
    [Date_Sk]           VARCHAR (10)  NOT NULL,
    [Inspection_Date]   DATETIME      NULL,
    [Inspection_Month] VARCHAR (23)  NULL,
    [Inspection_Year]   VARCHAR (6)   NULL,
    [Inspection_Day]    VARCHAR (15)  NULL,
    [DI_WorkflowName]   VARCHAR (100) NULL,
    PRIMARY KEY CLUSTERED ([Date_Sk] ASC)
);
```



3. DimInspectionType

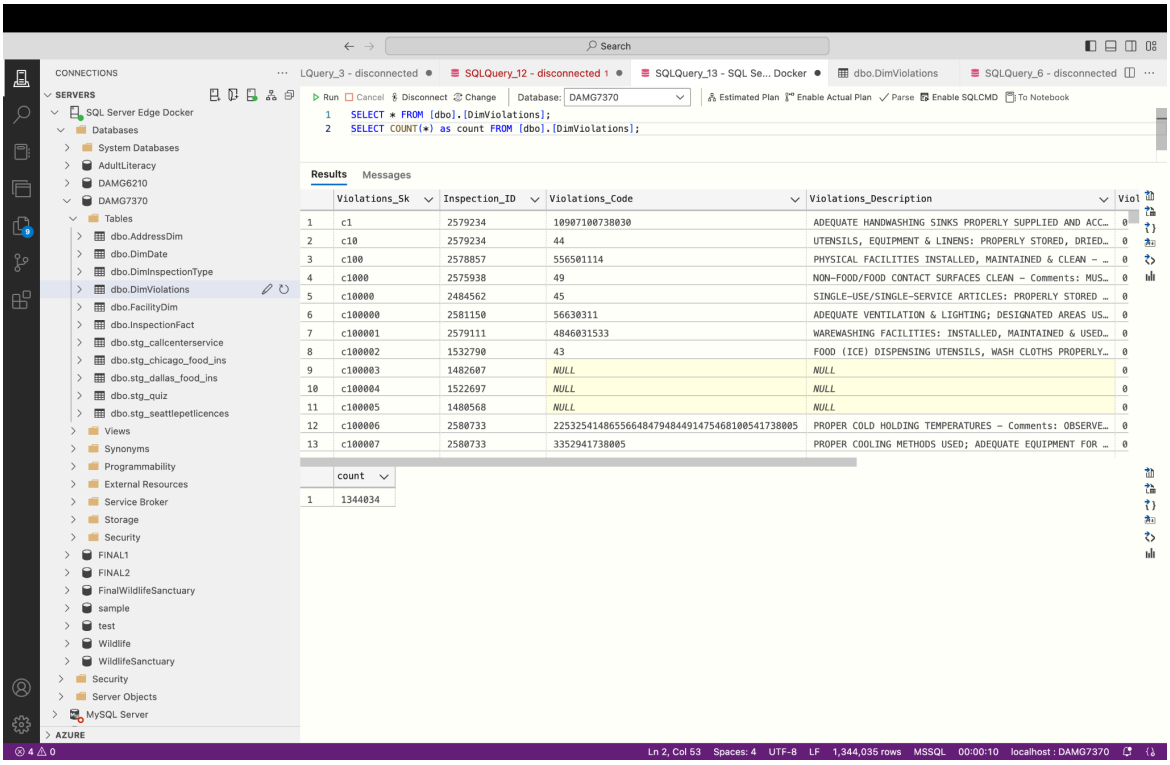
```
CREATE TABLE [dbo].[DimInspectionType] (
    [InspectionTypeID_Sk] VARCHAR (10)  NOT NULL,
    [Inspection_Type]     VARCHAR (50)  NULL,
    [Inspection_Score]    INT           NULL,
    [Risk]                VARCHAR (15)  NULL,
    [DI_WorkflowName]     VARCHAR (100) NULL,
    PRIMARY KEY CLUSTERED ([InspectionTypeID_Sk] ASC)
```

```
);
```



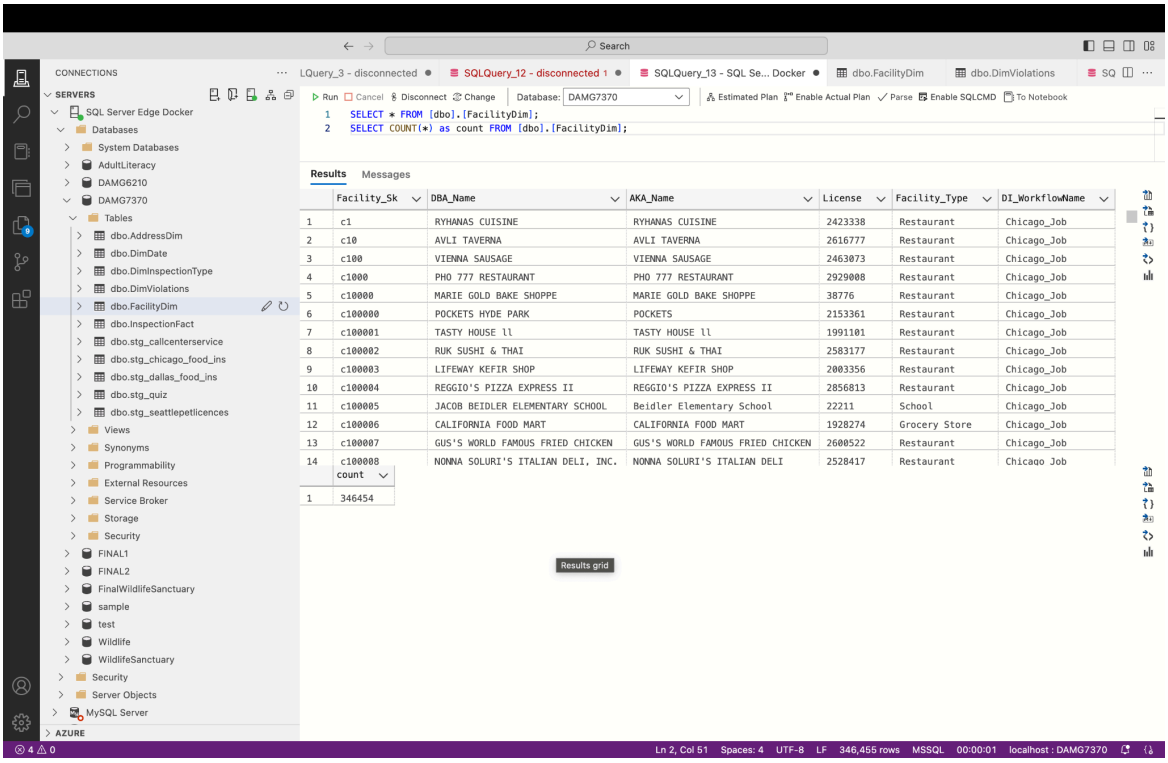
4. DimViolations

```
CREATE TABLE [dbo].[DimViolations] (
   [Violations_Sk]           VARCHAR (10)  NOT NULL,
   [Inspection_ID]           VARCHAR (10)  NULL,
   [Violations_Code]         VARCHAR (MAX) NULL,
   [Violations_Description] VARCHAR (MAX) NULL,
   [Violation_Points]        INT           NULL,
   [DI_CreateDate]           DATE          NULL,
   [DI_WorkflowName]         VARCHAR (100) NULL,
   PRIMARY KEY CLUSTERED ([Violations_Sk] ASC)
);
```



5. FacilityDim

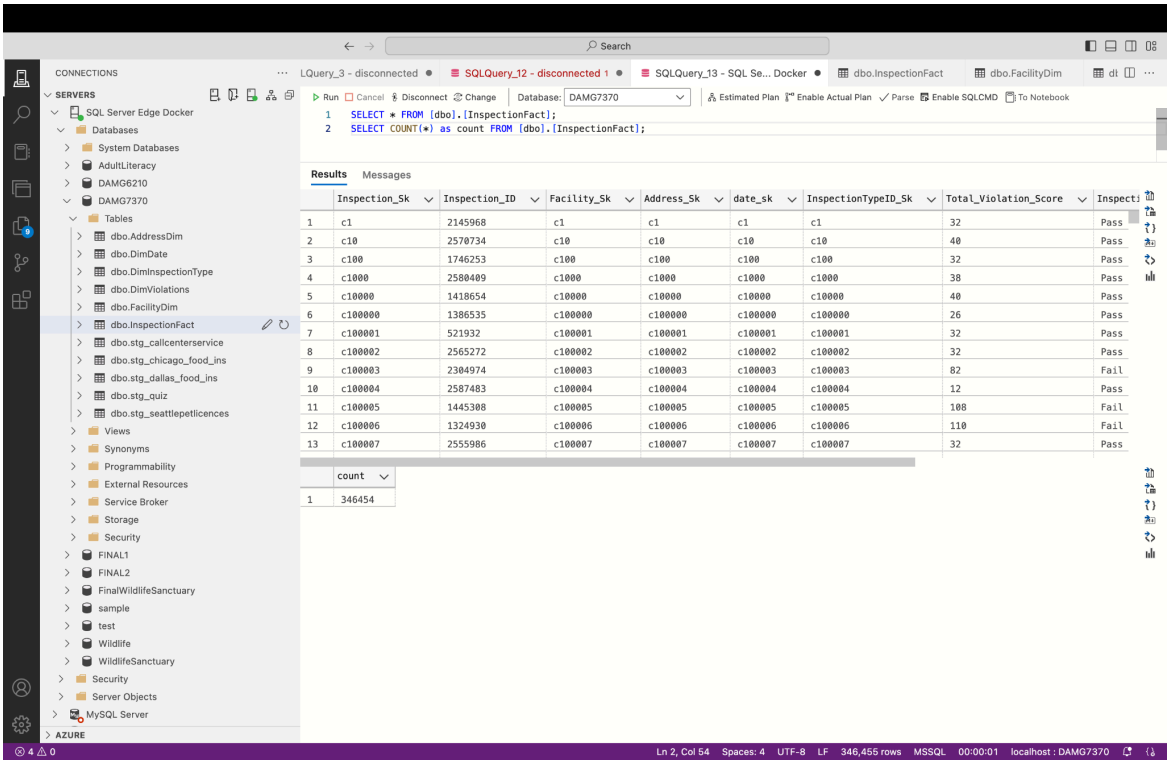
```
CREATE TABLE [dbo].[FacilityDim] (
   [Facility_Sk]      VARCHAR (10)  NOT NULL,
   [DBA_Name]         VARCHAR (80)  NULL,
   [AKA_Name]         VARCHAR (80)  NULL,
   [License]          INT           NULL,
   [Facility_Type]    VARCHAR (50)  NULL,
   [DI_WorkflowName] VARCHAR (100) NULL,
   PRIMARY KEY CLUSTERED ([Facility_Sk] ASC)
);
```



Fact

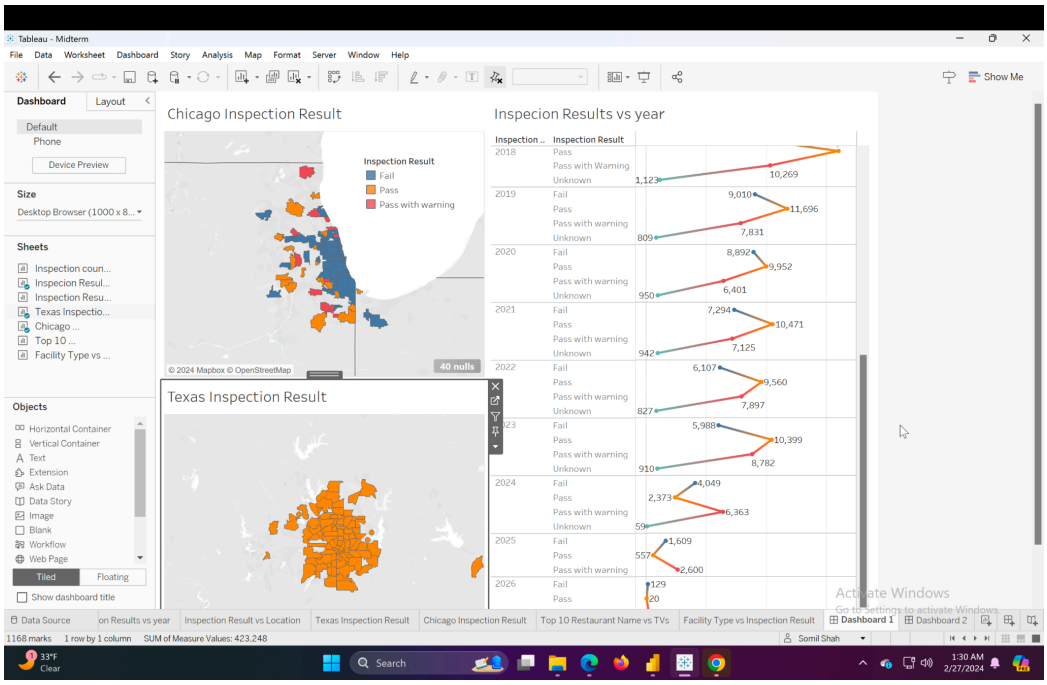
1. InspectionFact

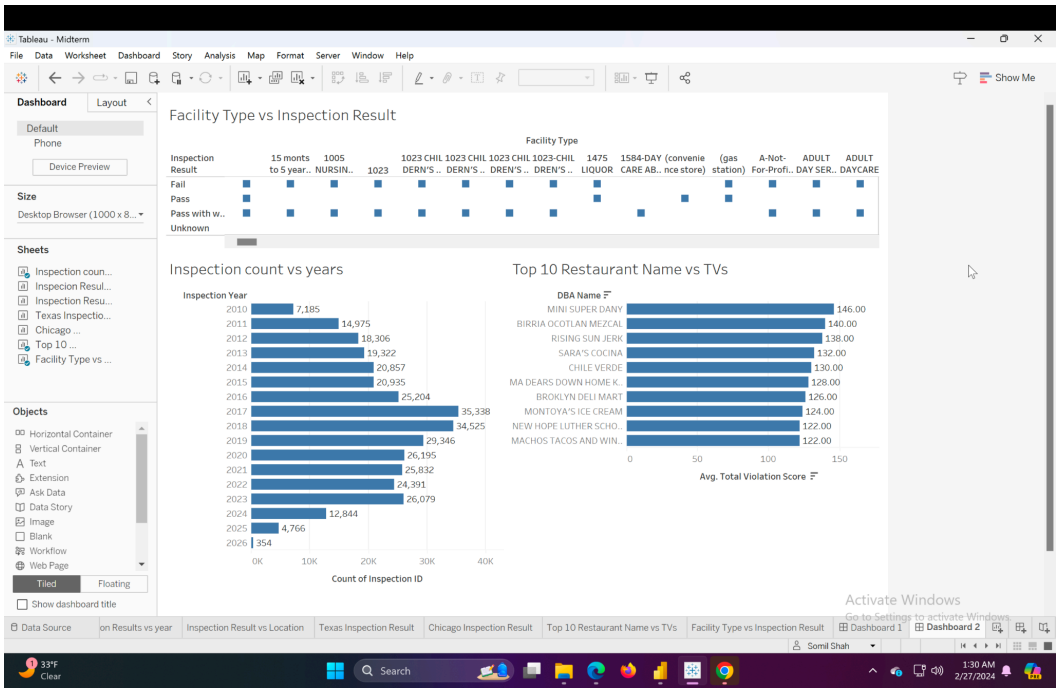
```
CREATE TABLE [dbo].[InspectionFact] (
    [Inspection_Sk]          VARCHAR (10)  NOT NULL,
    [Inspection_ID]          VARCHAR (10)  NULL,
    [Facility_Sk]            VARCHAR (10)  NOT NULL,
    [Address_Sk]             VARCHAR (10)  NOT NULL,
    [date_sk]                VARCHAR (10)  NULL,
    [InspectionTypeID_Sk]    VARCHAR (10)  NOT NULL,
    [Total_Violation_Score] INT            NULL,
    [Inspection_Result]      VARCHAR (100) NULL,
    [DI_WorkflowName]        VARCHAR (100) NULL,
    PRIMARY KEY CLUSTERED ([Inspection_Sk] ASC)
);
```



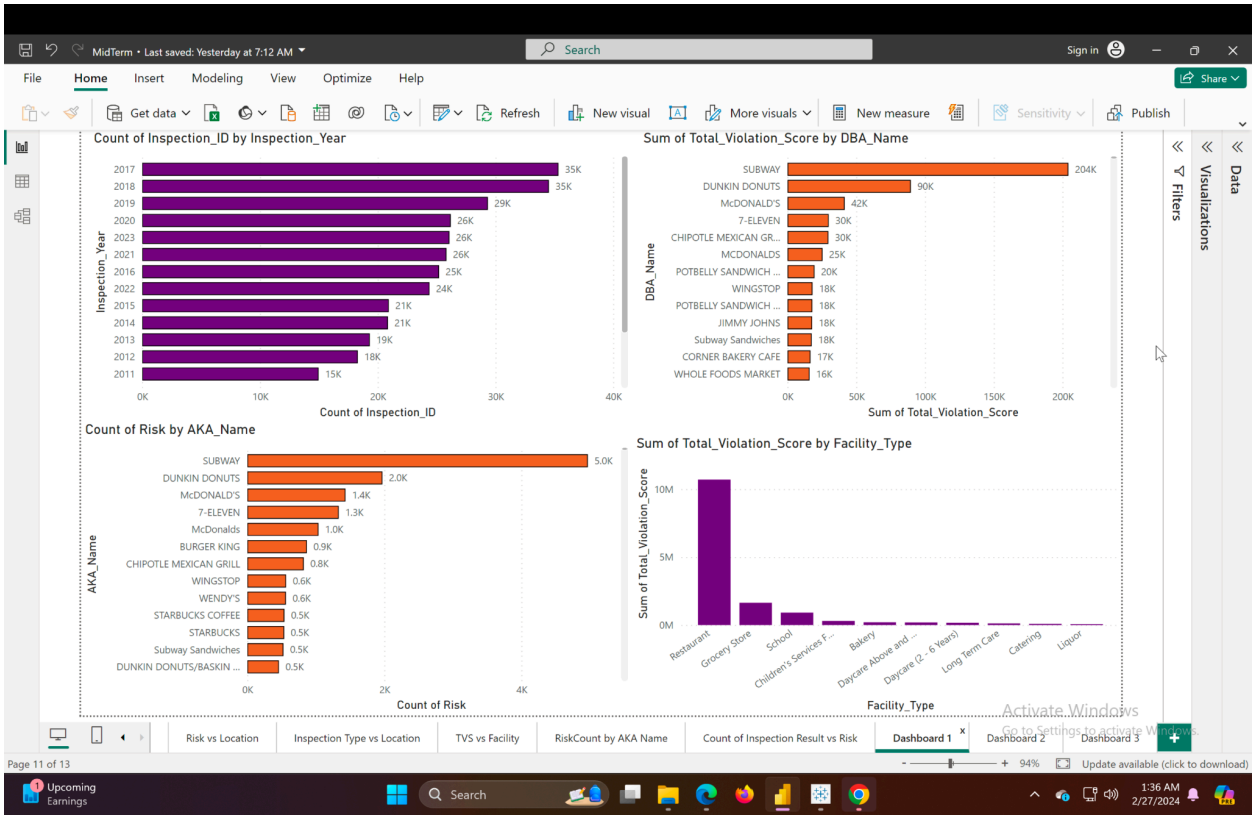
6. BI Visualization

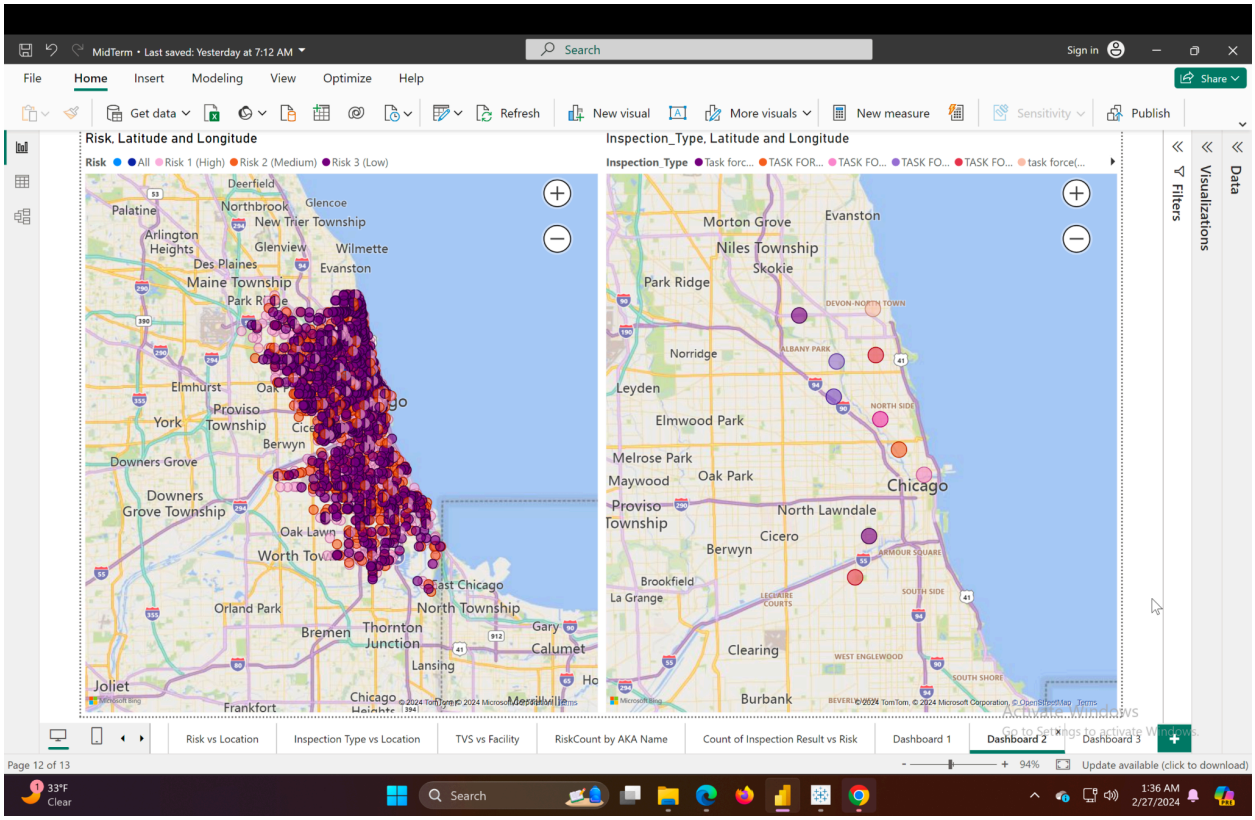
Tableau Dashboards:

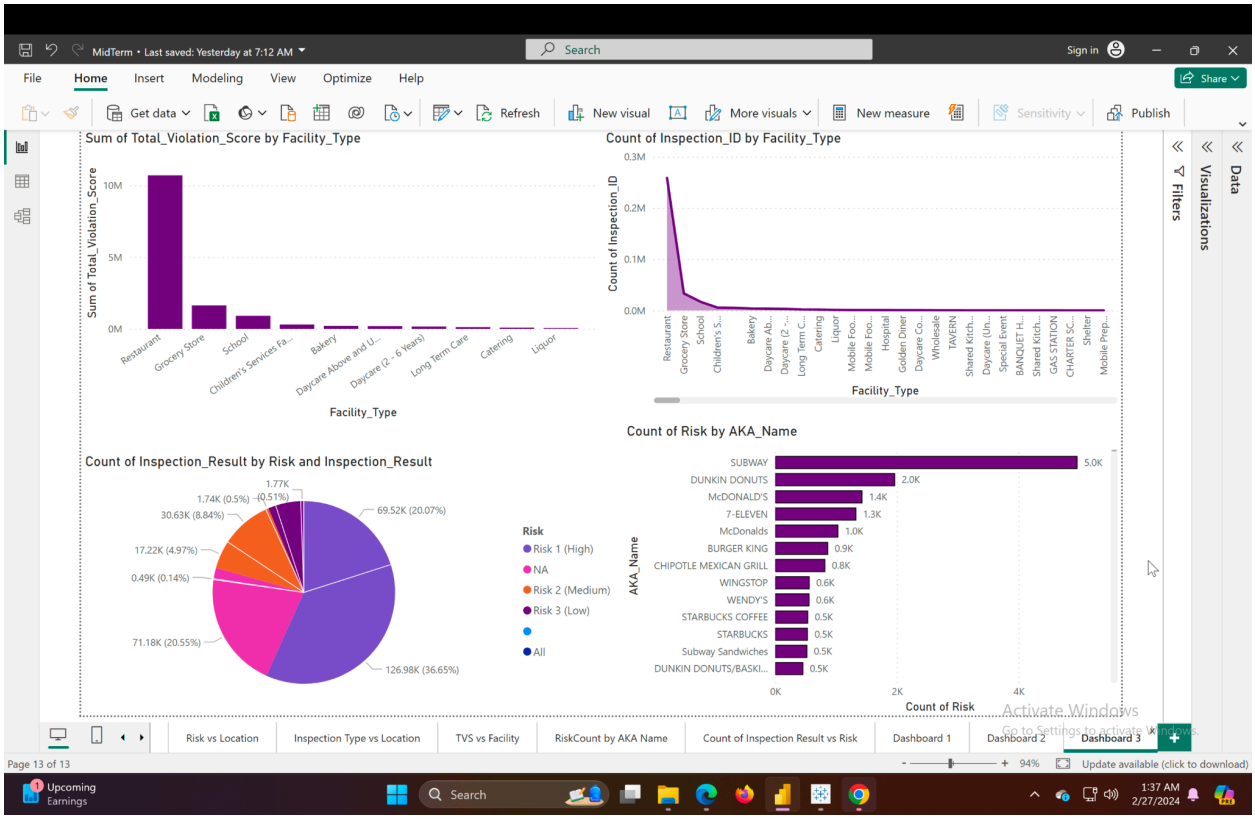




PowerBI Dashboard

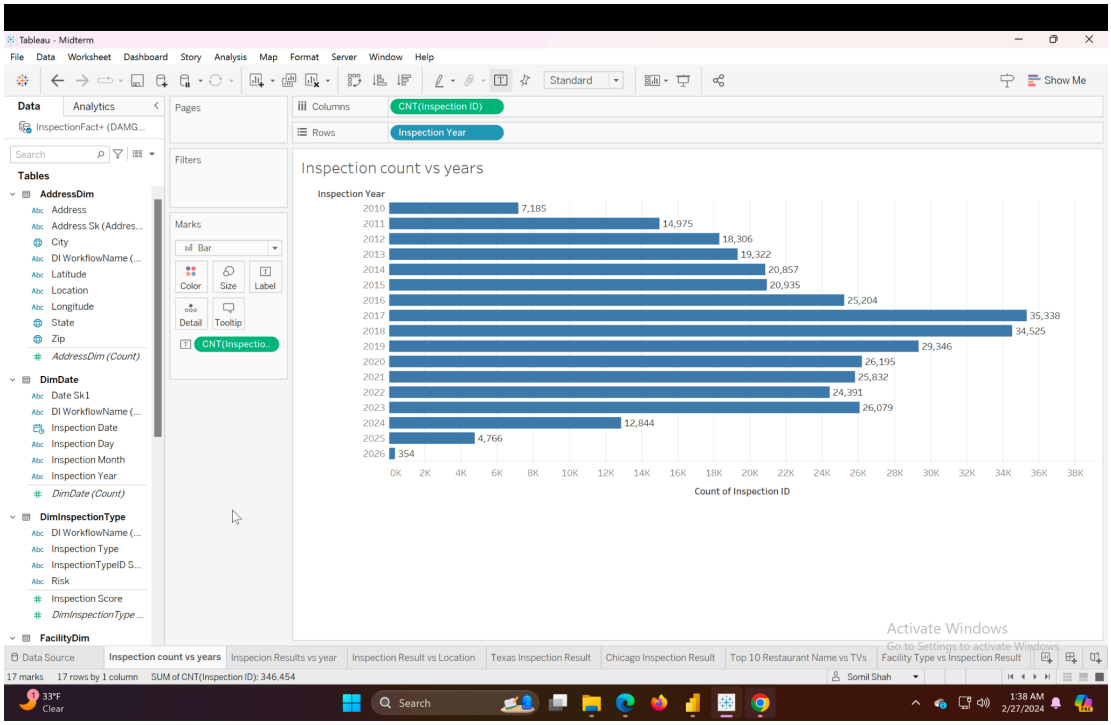


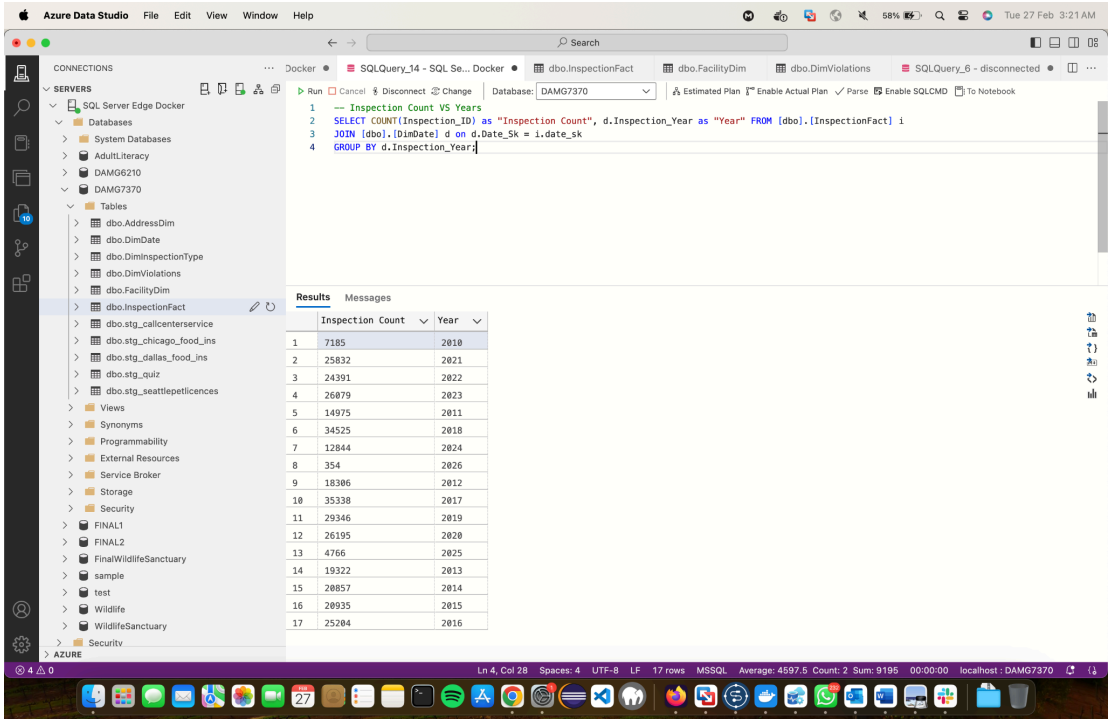

MidTerm • Last saved: Yesterday at 7:12 AM
Search
Sign in
File
Home
Insert
Modeling
View
Optimize
Help
Share
Get data
Refresh
New visual
More visuals
New measure
Sensitivity
Publish
Risk, Latitude and Longitude
Risk All Risk 1 (High) Risk 2 (Medium) Risk 3 (Low)
Inspection_Type, Latitude and Longitude
Inspection_Type Task forc... TASK FOR... TASK FO... TASK FO... TASK FO... task force(...
Filters
Visualizations
Data
Risk vs Location
Inspection Type vs Location
TVS vs Facility
RiskCount by AKA Name
Count of Inspection Result vs Risk
Dashboard 1
Dashboard 2
Dashboard 3
Page 12 of 13
94%
Update available (click to download)
33°F
Clear
Search
1:36 AM
2/27/2024


MidTerm • Last saved: Yesterday at 7:12 AM
Search
Sign in
File
Home
Insert
Modeling
View
Optimize
Help
Share
Get data
Refresh
New visual
More visuals
New measure
Sensitivity
Publish
Sum of Total_Violation_Score by Facility_Type
Sum of Total_Violation_Score
Facility_Type
Count of Inspection_ID by Facility_Type
Count of Inspection_ID
Facility_Type
Count of Inspection_Result by Risk and Inspection_Result
Risk
Risk 1 (High)
NA
Risk 2 (Medium)
Risk 3 (Low)
All
Count of Risk by AKA_Name
AKA_Name
Count of Risk
Filters
Visualizations
Data
Risk vs Location
Inspection Type vs Location
TVS vs Facility
RiskCount by AKA Name
Count of Inspection Result vs Risk
Dashboard 1
Dashboard 2
Dashboard 3
Page 13 of 13
94%
Update available (click to download)
Upcoming
Earnings
Search
1:37 AM
2/27/2024

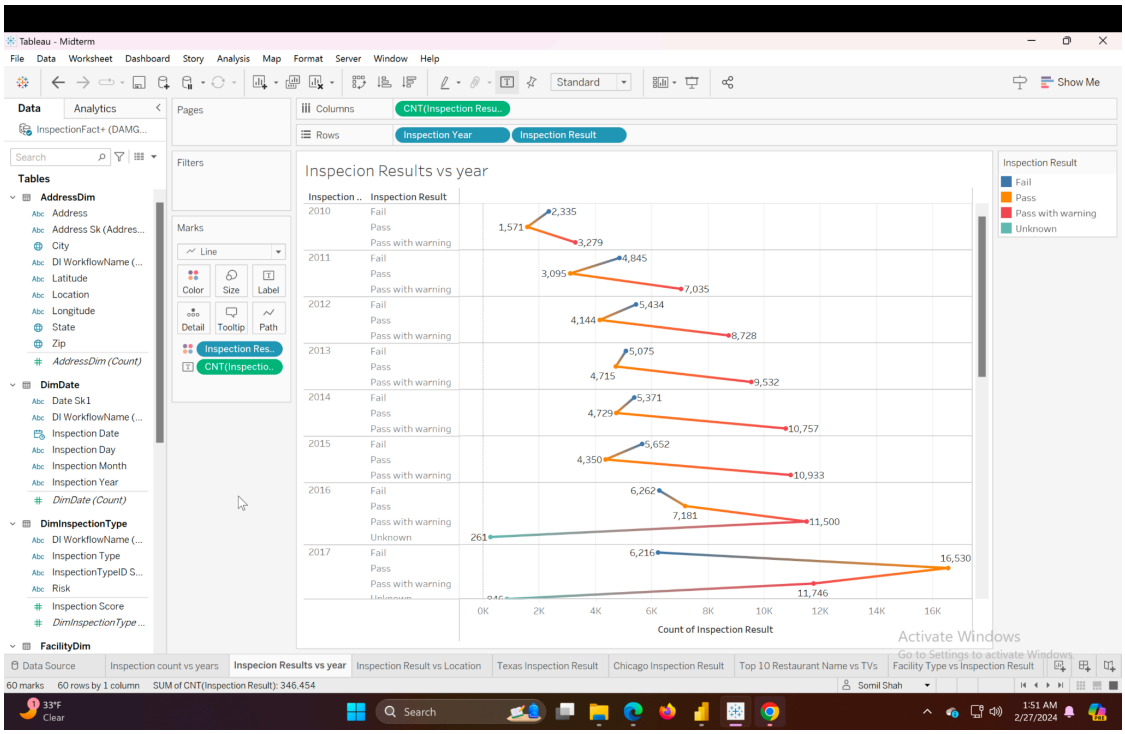
7. Testing

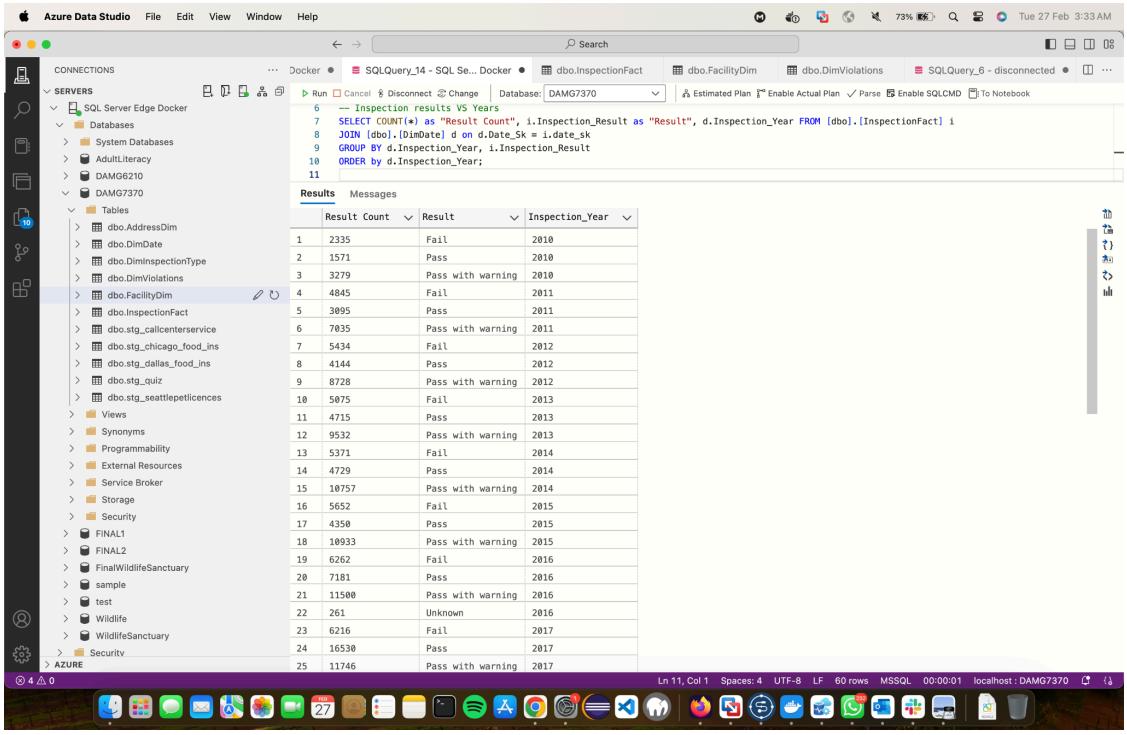
1. Inspection count vs years





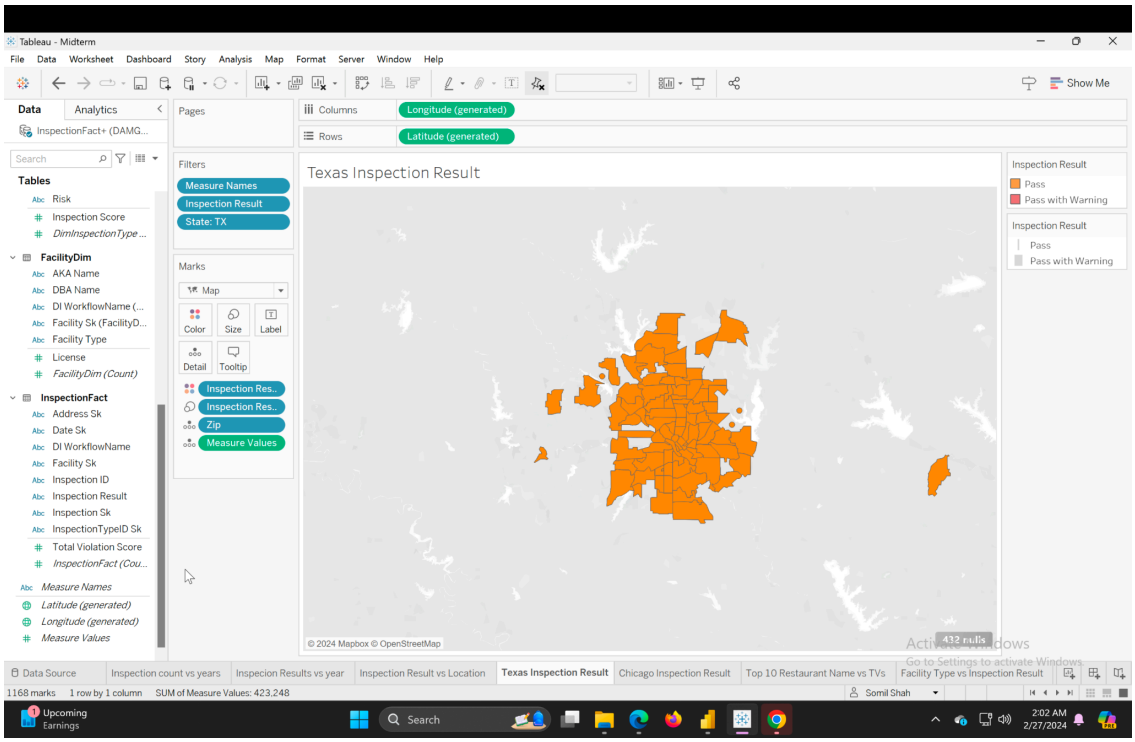
2. Inspection Results vs year

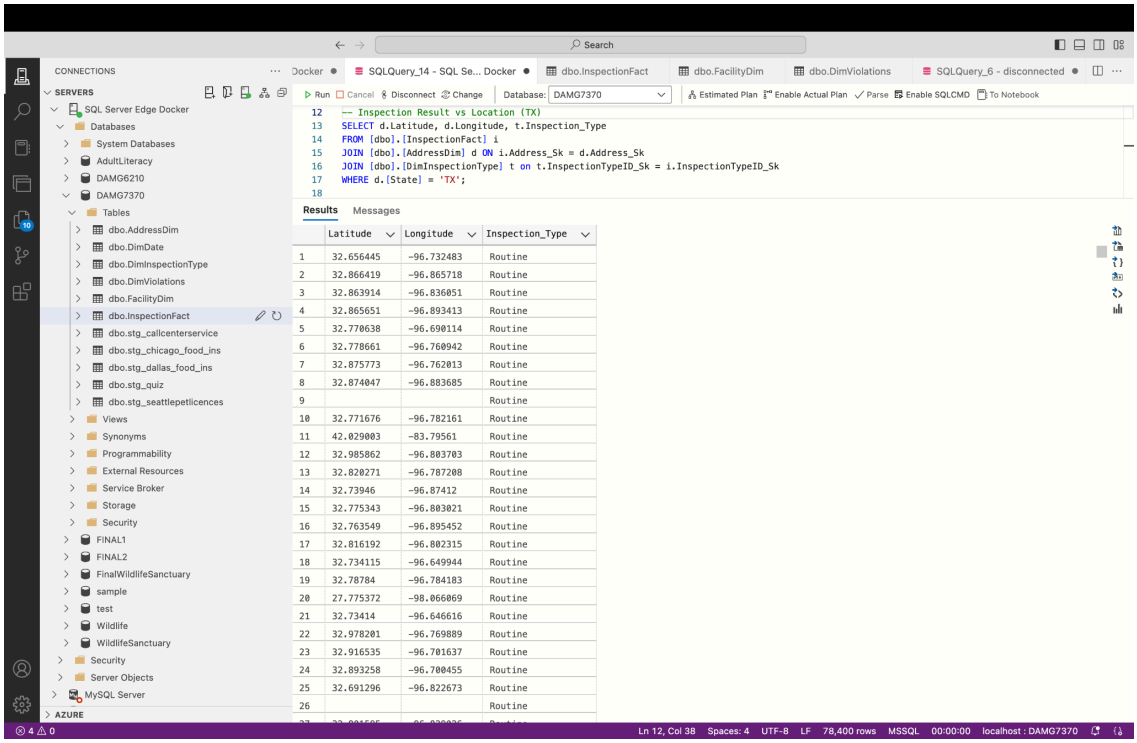




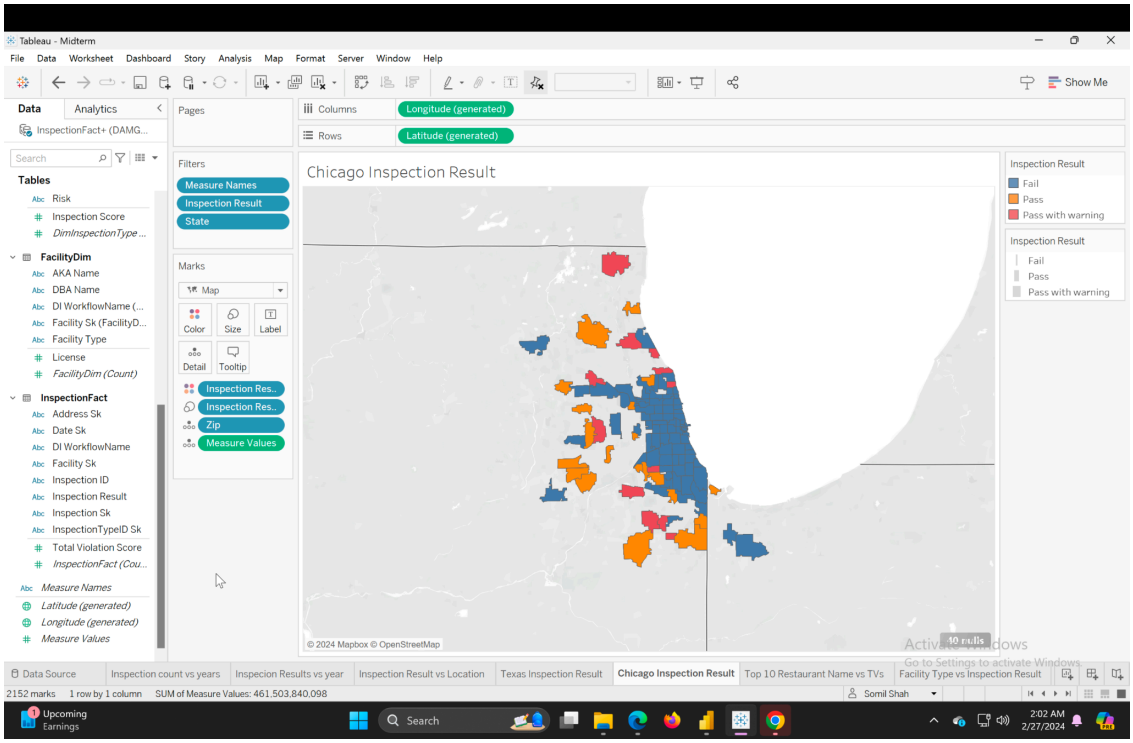
	Result Count	Result	Inspection_Year
1	2335	Fail	2010
2	1571	Pass	2010
3	3279	Pass with warning	2010
4	4845	Fail	2011
5	3095	Pass	2011
6	7035	Pass with warning	2011
7	5434	Fail	2012
8	4144	Pass	2012
9	8728	Pass with warning	2012
10	5075	Fail	2013
11	4715	Pass	2013
12	9532	Pass with warning	2013
13	5371	Fail	2014
14	4729	Pass	2014
15	10757	Pass with warning	2014
16	5652	Fail	2015
17	4350	Pass	2015
18	10933	Pass with warning	2015
19	6262	Fail	2016
20	7181	Pass	2016
21	11500	Pass with warning	2016
22	261	Unknown	2016
23	6216	Fail	2017
24	16530	Pass	2017
25	11746	Pass with warning	2017

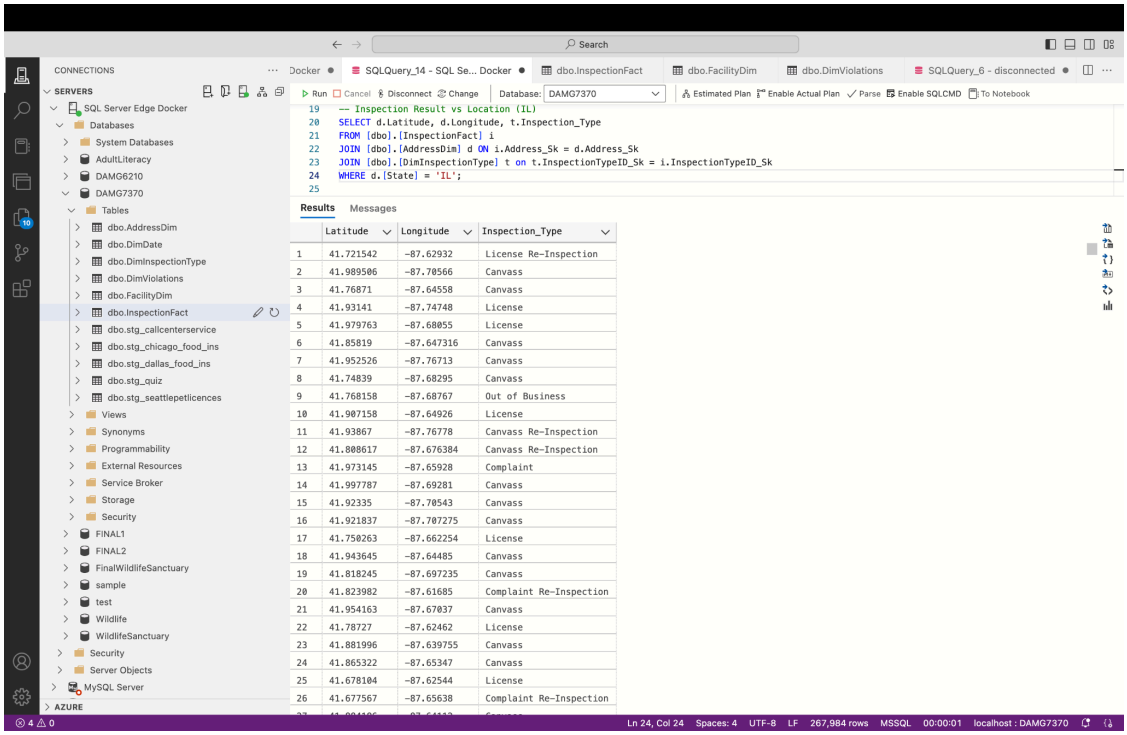
3. Inspection Result VS Location (Dallas)



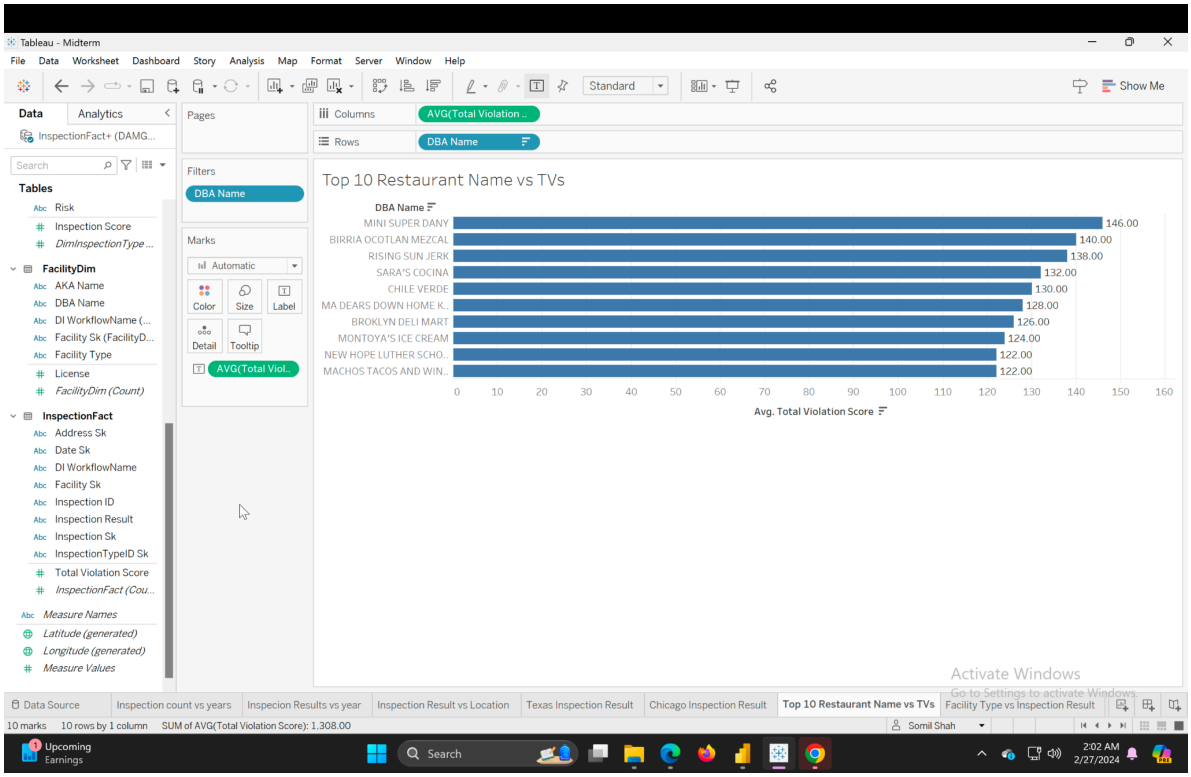


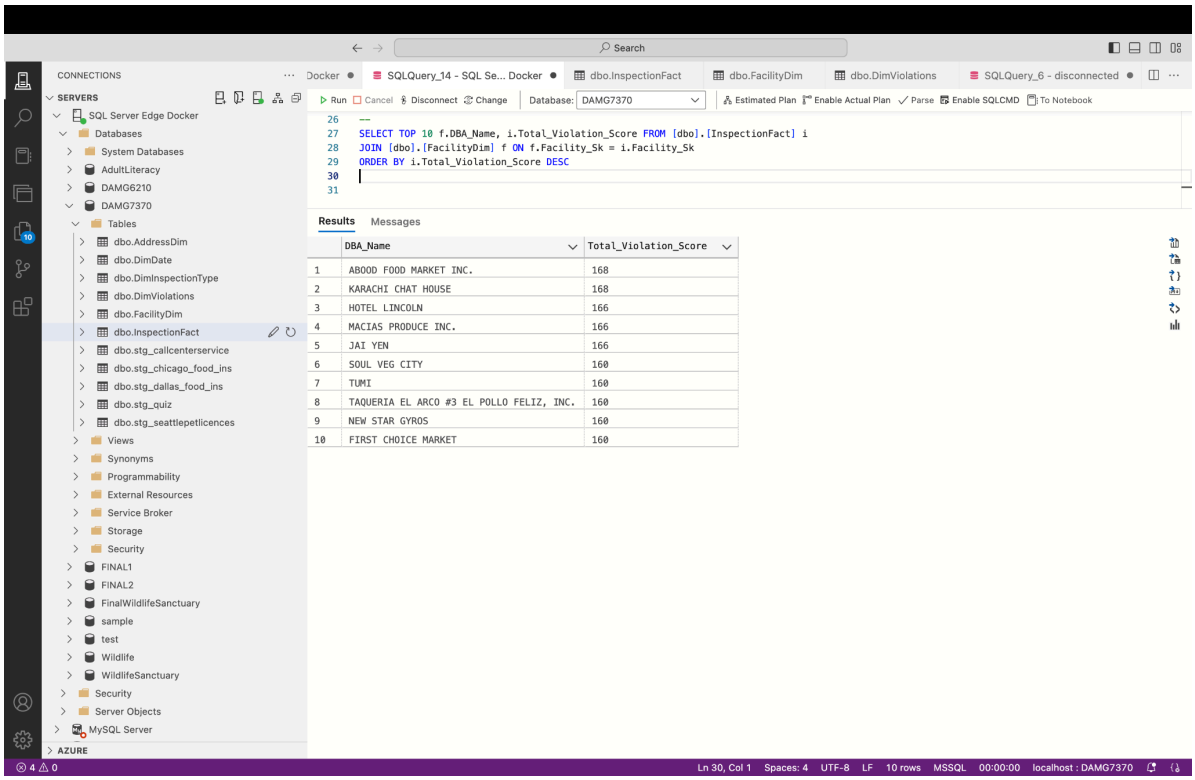
4. Inspection Result VS Location (Chicago)





5. Top 10 Restaurant Name vs TVS





6. Facility Type vs Inspection Result