DAMG 7370: Designing Advanced Data Architecture for Business Intelligence

-----------------------------------------------------------------------------------------------------------------

Darshana Rokade

NUID: 002705752

-----------------------------------------------------------------------------------------------------------------

Implementation:

1. Use Alteryx to profile your TSV data

A screenshot of a computer

Description automatically generated

------------------------------------------------------------------------------------------------------------------------------------

2. Make observations and write down the problems that you see with the data like missing values, duplicates, date formats, data conversions, special characters (if any to be identified) and inconsistencies. (For this implementation you can skip the cleaning part as its not covered)Using Alteryx populate data into Microsoft SQL database OR Azure SQL database OR local MySQL database. (If you have any issues connecting to MySQL ask in the slack for help)

A screenshot of a computer

Description automatically generated

Observations and Problems in the dataset:

**1. First, as per naming conventions should be followed:**

* Changing the column name from CASE ID -> Case\_ID, WORK GROUP -> Work\_Group, etc

|  |  |
| --- | --- |
| Current Column Name | Suggested Column name(As per naming convention) |
| CASE ID | case\_id |
| SOURCE | source |
| DEPARTMENT | department |

**2. Missing Values:**

* The ‘CATEGORY2’ and ‘CATEGORY3’ column has numerous `[Null]` values, indicating missing data.

**3. Data types:**

* All the fields’ data types should be changed as per the data.
* For example: case\_id can be treated as primary key, since it’s an integer value, data type should be changed to int.

A screenshot of a computer

Description automatically generated

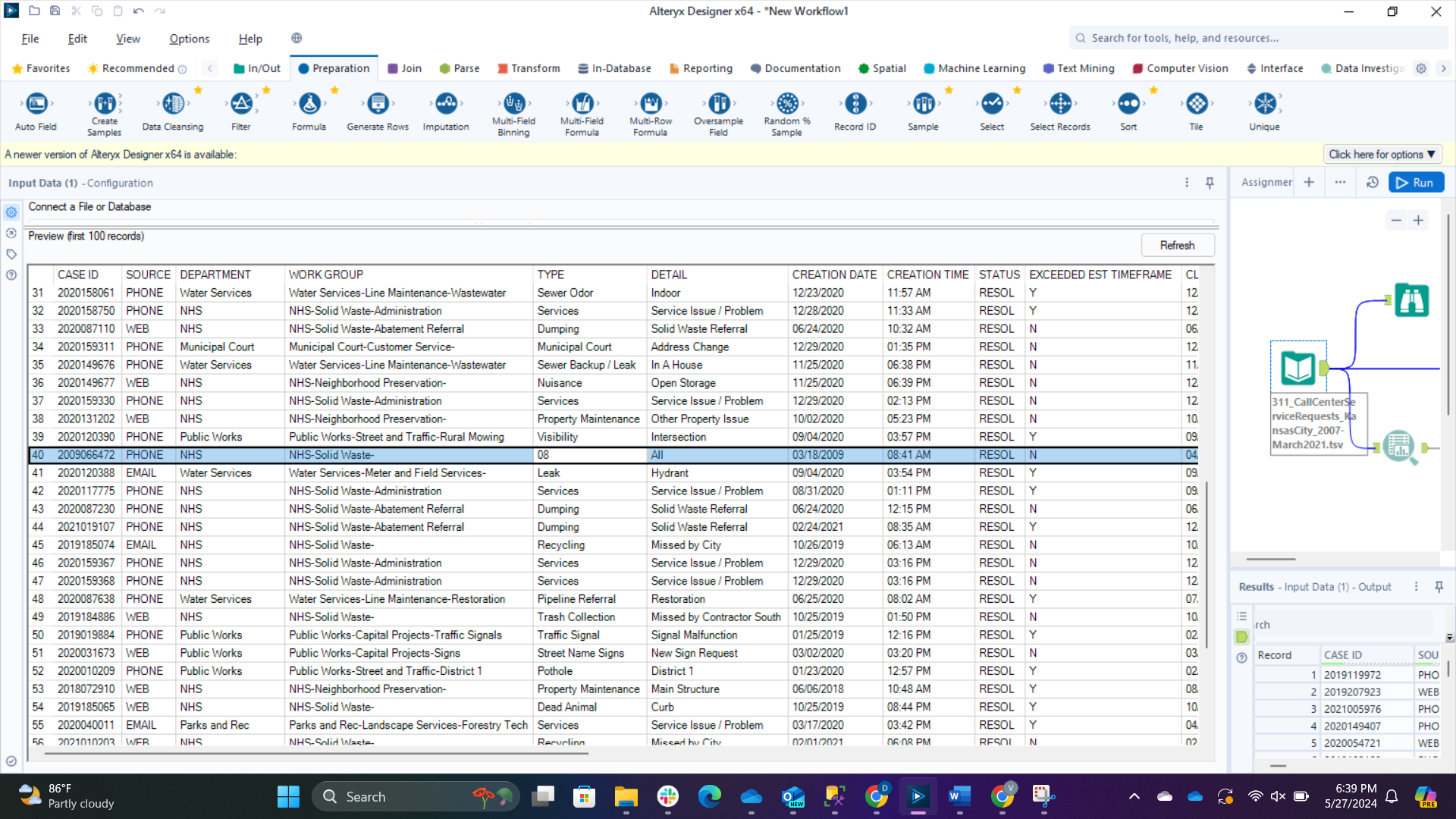
**4. Date Formats:**

* ‘CREATION TIME’ is in ‘HH:MM AM/PM’ format, which needs to be converted to a proper time format.

**5. Data Conversions:**

* ‘DAYS TO CLOSE’ should be verified as a numeric field.
* Latitude and Longitude should be in float format.
* Categorical data such as ‘SOURCE’, ‘DEPARTMENT’, ‘STATUS’, etc., should be converted to appropriate categorical types.

**6. Inconsistencies:**



**Recommendations for Cleaning the Data:**

**1. Remove Duplicates:**

* Use `drop\_duplicates` to remove any duplicate rows.

**2. Convert Date and Time:**

* Convert ‘CREATION DATE’ and ‘CLOSED DATE’ to datetime objects.
* Convert ‘CREATION TIME’ to a time format.

(Possible) Combine ‘CREATION DATE’ and ‘CREATION TIME’ to one column as ‘creation\_timestamp’

**3. Data Type Conversions:**

* Convert ’DAYS TO CLOSE’, ‘Latitude’, and ‘Longitude’ to appropriate numeric types.
* Convert categorical fields to category types.

After Changing Data type:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

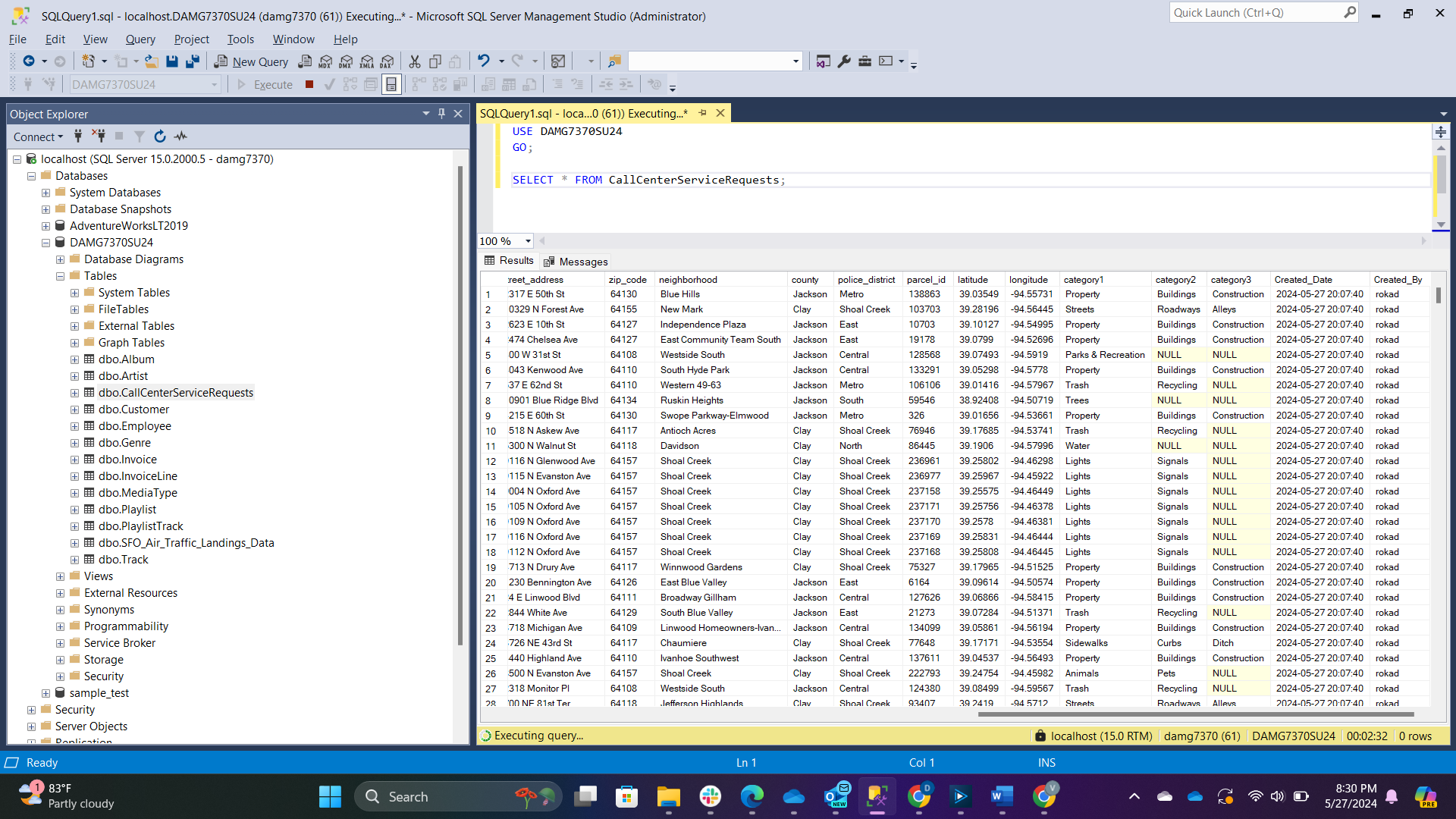
Description automatically generated

------------------------------------------------------------------------------------------------------------------------------------

3. Input CSV dataset -> Identify data issues and document it (ignore the methods to clean data as its not covered) -> Stage Table in Azure or Local SQL server, and MySQL

A computer screen shot of a computer screen

Description automatically generated



Query:A computer screen shot of a computer screen

Description automatically generated

Output:A computer screen with a screen showing a program

Description automatically generated with medium confidence

Row Count:

A screenshot of a computer

Description automatically generated