**Group 16 : Nishita Shewale, Ritik Gupta, Siddharth Pawar, Sakshi Aade**

**Midterm Project**

**Alteryx : Data Profiling Report (Part 1)  
Dataset :** [**https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/data\_preview**](https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/data_preview)

**Description :** Inspections data of restaurants and other food establishments in Chicago.

**Number of attributes : 17,**

**Record Count : 267613,**

**Tool used for concluding the above info : “Browse Tool”**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

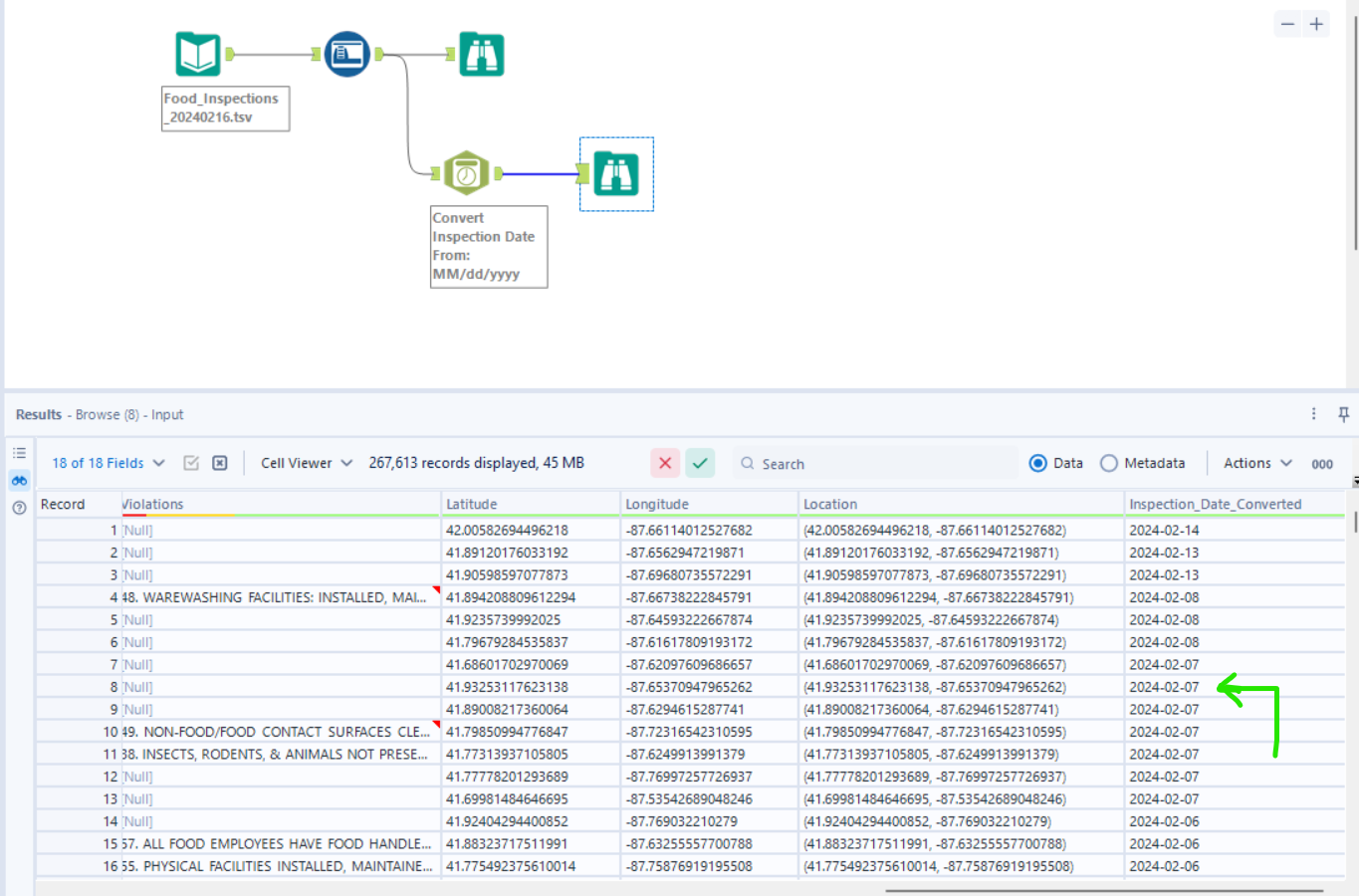
**Issue with the dataset :**  
1. All attributes are in “String” datatype

In order to fix this, we are using the **“Auto-field” tool,** which yields us the following result

A screenshot of a computer

Description automatically generated

2. Now, we need to **convert the Inspection date column to “Date” datatype from String** for which I will use the **“Date-Time Parse” tool.**



This tool helped us convert the Inspection date column from String Datatype to “Date” datatype (which will help us drill the date up and down in visualizations).

3. Other columns to fix

**Convert state to “char” datatype with limit 2**

**Convert Zip to “char” datatype with limit 5**

**Convert Latitude and longitude to “decimal” datatype instead of string.**

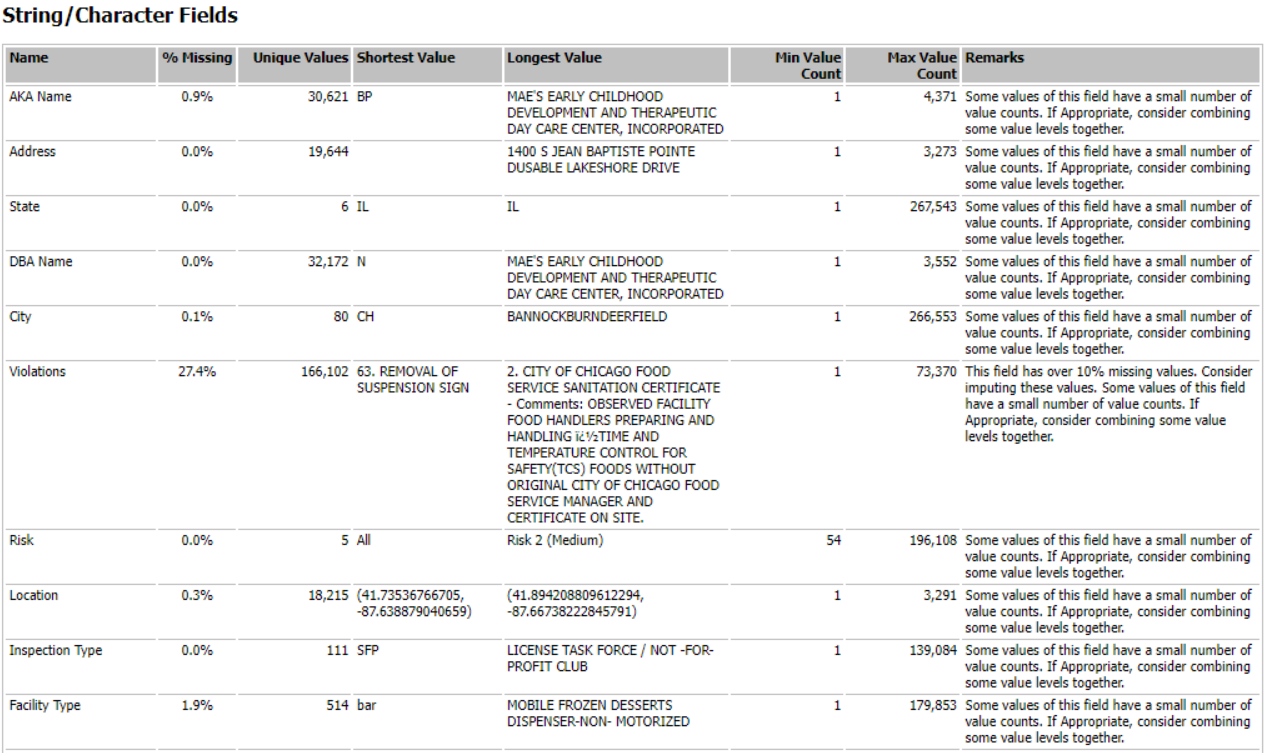
All this can be done using the **“select tool”,** fixing the datatypes and later select only the columns required. (wherein we won`t select the previous inspection date column which was in String format)

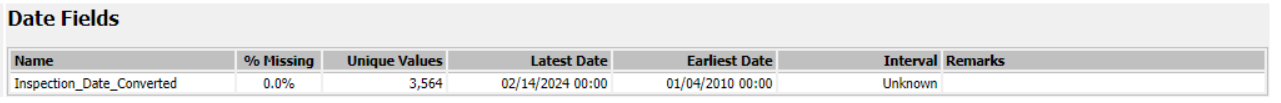
Select tool usage:

A screenshot of a computer

Description automatically generated

4. After using the select tool, in order to see an accumulated report of the null values, min-max length/value, inconsistencies in the data, we used the **“Field-Summary Tool” which helped us generate the following conclusions.**





A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**5. Null values :**

**AKA Name, City, Locations, Facility Type, Violations, Latitude and longitude are the 7 columns having null values in them ; Violations being the attribute with the most null values and city being the attribute with the least null values.**

We can also see the attribute specific inconsistencies in the Browse profile tab on the left.

Example : Inconsistencies in “Violations” Column

A screenshot of a data report

Description automatically generated

6. The null values in any column can be filled using imputation methods using the **“imputation tool”** and the “not-ok” values can be cleansed using the **“data-cleaning” tool** in Alteryx.

7. Further, we used the **“Formula Tool”** in order to add 2 more columns to the staged table (DI\_Created\_Date, DI\_WorkfileName) : Process shown in the screenshot below.

A screenshot of a computer

Description automatically generated

**Screenshot of whole Alteryx Workflow:**

**A screenshot of a computer

Description automatically generated**

**Runtime:**

**A screenshot of a computer

Description automatically generated**

**Row count verification after staging the table in SQL Server: 267613**

**A screenshot of a computer

Description automatically generated**

**DDL script for Stage table (Chicago Dataset) :**

A screenshot of a computer

Description automatically generated

**Alteryx : Data Profiling Report (Part 2)  
Dataset :** [**https://www.dallasopendata.com/Services/Restaurant-and-Food-Establishment-Inspections-Octo/dri5-wcct/data\_preview**](https://www.dallasopendata.com/Services/Restaurant-and-Food-Establishment-Inspections-Octo/dri5-wcct/data_preview)

**Description :** Intended to communicate the name of establishment, the physical location of the establishment, the date the inspection was conducted, the overall score for the inspection, and the point deduction for the individual violations in Dallas City.

**Number of attributes : 114,**

**Record Count : 78400,**

**Tool used for concluding the above info : “Browse Tool”**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Issue with the dataset :**  
1. All attributes are in “String” datatype

In order to fix this, we are using the **“Auto-field” tool,** which yields us the following result

A screenshot of a computer

Description automatically generated

2. Now, we need to **convert the Inspection date column to “Date” datatype from String** for which I will use the **“Date-Time Parse” tool.**

A screenshot of a computer

Description automatically generated

This tool helped us convert the Inspection date column from String Datatype to “Date” datatype (which will help us drill the date up and down in visualizations).

3. Other columns to fix

**Convert Zip to “char” datatype with limit 5**

**Rename Inspection Year column to Inspection\_Financial\_Year**

All this can be done using the **“select tool”,** fixing the datatypes and later select only the columns required. (wherein we won`t select the previous inspection date column which was in String format, also we can disregard the column inspection month as we are covering it in the inspection\_date\_converted column)

Select tool usage:

A screenshot of a computer

Description automatically generated

*Observations :*

*Inspection records are stored in the form of 4 points related to a violation ; Violation Points, Violation Detail, Violation Memo, Violation Description. Every inspection has been provided to store a maximum of 25 violations to its record*.

4. After using the select tool, in order to see an accumulated report of the null values, min-max length/value, inconsistencies in the data, we used the **“Field-Summary Tool” which helped us generate the following conclusions.**

**5. Null values :**

**Street Number, Inspection Score, Inspection\_Date\_Converted, Restaurant Name, Zipcode, Inspection\_Financial\_year, Street Name, Street Address**  are the only columns who do not have any null values in them. Apart from these columns, there are a lot of inconsistencies and null values in others (especially the attributes related to violations, since not every inspection record has the same number of recorded violations and data related to it).

We can also see the attribute specific inconsistencies in the Browse profile tab on the left.

6. The null values in any column can be filled using imputation methods using the **“imputation tool”** and the “not-ok” values can be cleansed using the **“data-cleaning” tool** in Alteryx.

7. Further, we used the **“Formula Tool”** in order to add 2 more columns to the staged table (DI\_Created\_Date, DI\_WorkfileName) : Process shown in the screenshot below.

A screenshot of a computer

Description automatically generated

**Screenshot of whole Alteryx Workflow:**

**A screenshot of a computer

Description automatically generated**

**Runtime:**

**A screenshot of a computer

Description automatically generated**

**Row count verification after staging the table in SQL Server: 78400**

A screenshot of a computer

Description automatically generated

**DDL script for Stage table (Dallas Dataset) :**

A screenshot of a computer

Description automatically generated

**We also performed YData-Profiling on both the datasets in order to explore how it works :**

[**https://northeastern.sharepoint.com/:u:/r/sites/DADABI692/Shared%20Documents/General/ydata-profiling\_Chicago.html?csf=1&web=1&e=6Qj3Zg**](https://northeastern.sharepoint.com/:u:/r/sites/DADABI692/Shared%20Documents/General/ydata-profiling_Chicago.html?csf=1&web=1&e=6Qj3Zg)

[**https://northeastern.sharepoint.com/:u:/r/sites/DADABI692/Shared%20Documents/General/ydata-profiling\_Dallas.html?csf=1&web=1&e=vVrYVS**](https://northeastern.sharepoint.com/:u:/r/sites/DADABI692/Shared%20Documents/General/ydata-profiling_Dallas.html?csf=1&web=1&e=vVrYVS)

**How do we plan to Merge Data?**

The bulk of the fields, including Business Name, Facility Type, Final Inspection Result, Total Violation Score, Address, Inspection ID, ZipCode, Latituide, Longitude, DI Created Date, Inspection Date, and DI Workflow File Name, that are associated with the Chicago and Dallas Dataset post transformation were attempted to match.   
While staging the table from Alteryx, we retrieved the file location information for DI\_Created\_Date and Di\_Workflow\_Filename in order to obtain those common columns. Latitude, longitude, inspection date, address, and zip code were previously included in the two databases.   
We computed Total Violation Score and Final Inspection Result based on the given conditions and matched them between the two datasets. Conditions where :

**A] Calculating Total Violation Score for Inspection**  
   1. Dallas:  
           Total Violation Score for a Inspection= Sum of all Violation Points for each inspection  
   2.Chicago:  
           if **Violations** is null, score always 0, if any violation then 2\*count of violations  
           if **Result** is Failed it should be 70  
           else if Result is Pass w/conditions it should be 40  
           else if Result is Pass it should be 0  
           else if Result is any Other it should be 0  
           if Risk is High: 30,  
                  Low: 10,  
                  Medium : 20  
                  Any Other: 0  
          Total Violation Score = Summation of all mentioned above. If **Total Violation Score** > 100 then Total Violation Score=100  
**B] Calculating Inspection-Result**  
    Total Violation Score Range  
         60 above Fail  
         below 60 and above 30: Pass with Warning,  
         30 and below: Pass  
         any other value: Other

The column "facility type" was already existing in the Chicago data, but there was only one unique value for it, therefore we hardcoded "restaurant" for all of the fields in the Dallas dataset. Since neither dataset contained any null entries for Restaurant Name and DBA Name respectively, we transformed them both to "business\_name" for both the datasets.