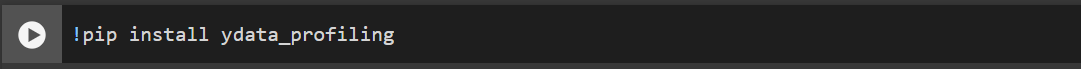
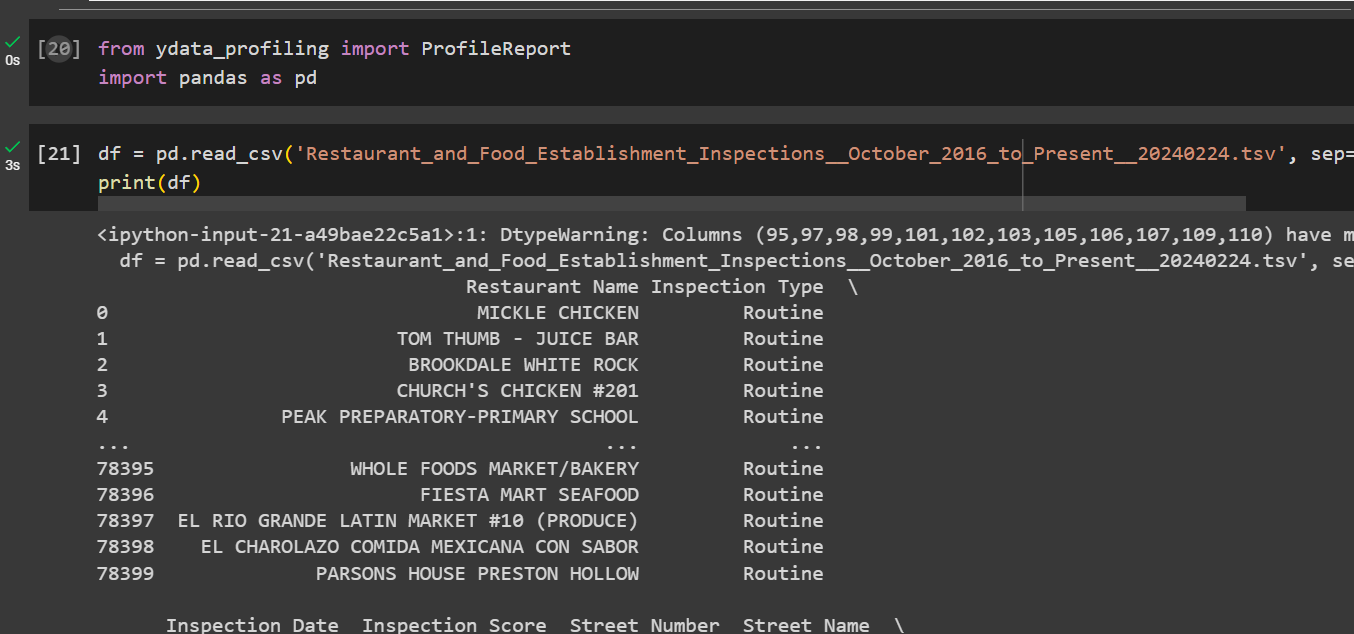
**Y Data Profiling Dallas**

1)



* Initially started with installing the Y data profiling python library.

2.)   
  
  
  
Initialized with the profile report by importing pandas library, then gave the command to read the dataset with tabulation as separator.  
  
  
  
  
  
3.) **To handle multivalued attributes**:-

For columns that may contain multi-valued attributes (e.g., lists of violations or services), the approach will depend on the specific analysis needs:

Splitting into Multiple Columns: If the multi-valued data is consistent and limited in number, you might split these values into separate columns for easier analysis.

Expanding into Multiple Rows: For analysis that requires treating each value independently, consider expanding the multi-valued data into multiple rows, ensuring each row retains a copy of the other columnar data for context.

Aggregating Values: In some cases, it might be more useful to aggregate multi-valued data into a summary form, such as counting the number of values or identifying the presence of specific keywords.

4.) **Data Type Conversion :-**

A typical approach for data type conversions, based on common columns found in inspection datasets, would include:

Inspection Date:

Current Type: Likely string (object in pandas).

Conversion: To datetime for enabling time-series analysis and operations.

Inspection Score:

Current Type: May be read as string due to inconsistencies in the data.

Conversion: To numeric type (int or float), ensuring that all values are consistent and properly formatted.

Address Information (Street Number, Street Name, etc.):

Current Type: String (object).

Conversion: Generally, these would remain as strings, but ensure consistency in formatting and case.

Violation Codes/Points:

Current Type: Could be mixed or string due to missing values or text notes.

Conversion: Numeric where possible for points. For violation codes that are inherently categorical or multi-valued, consider how best to represent these for your analysis (e.g., split into multiple columns or encode as separate binary columns).

Categorical Columns (Inspection Type, etc.):

Current Type: String (object).

Conversion: Convert to categorical type in pandas for efficiency and to facilitate analysis based on categories.

Geolocation Data (Latitude, Longitude):

Current Type: String or mixed.

Conversion: To Double, ensuring accurate representation for any mapping or spatial analysis.