Midterm Team Project

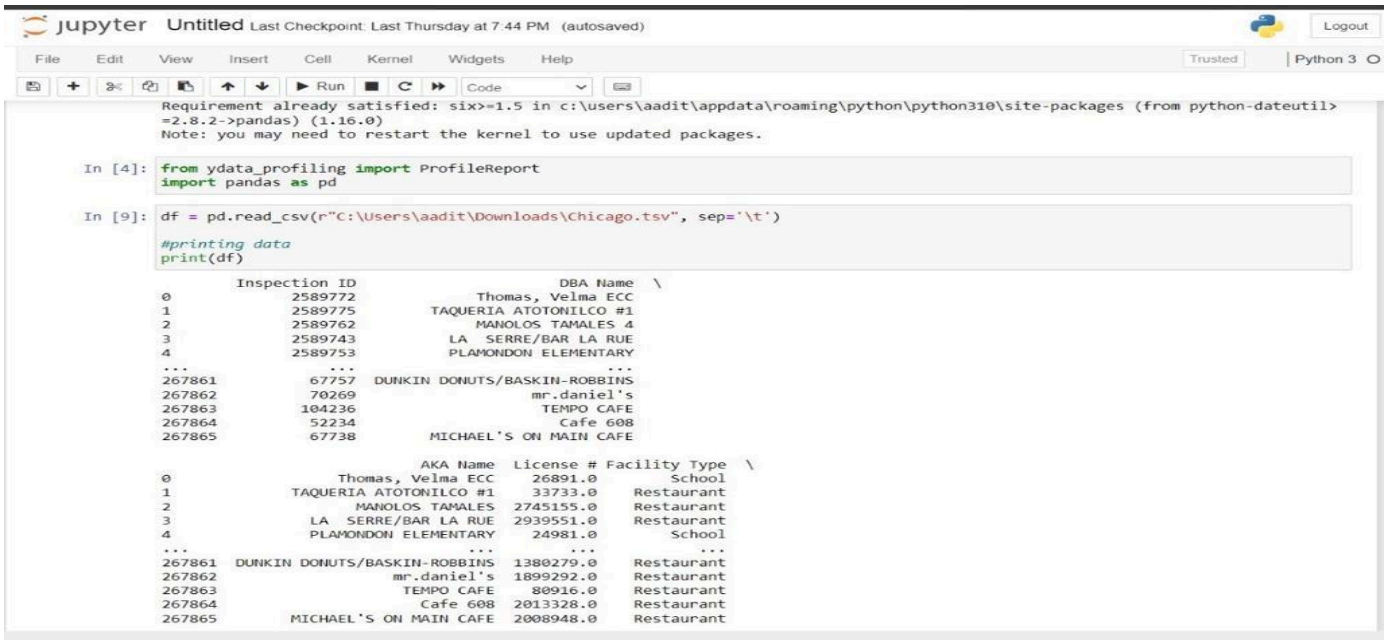
Designing Advanced Data Architectures for Business Intelligence

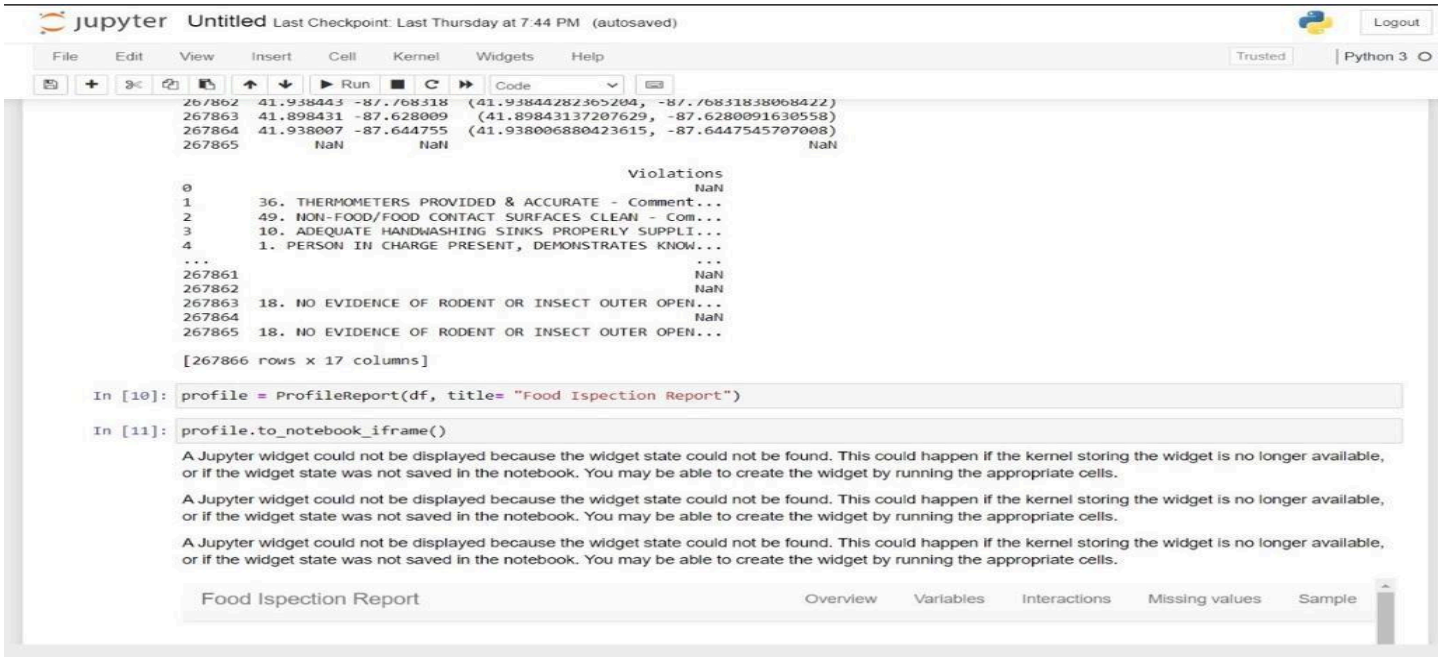
Group 13
Christina Daniel-
Nitish Belagali- 002645514
Shivpriya Mane- 002850626
Shruti Dhamdhere- 002812094

PART 1:

Ydata Profiling:

Food Inspection Report for Chicago:





jupyter Untitled Last Checkpoint: Last Thursday at 7:44 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

Overview

Overview Alerts 5 Reproduction

Dataset statistics

Number of variables	17
Number of observations	267866
Missing cells	84184
Missing cells (%)	1.8%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	34.7 MiB
Average record size in memory	136.0 B

Variable types

Numeric	5
Text	8
Categorical	3
DateTime	1

Variables

jupyter Untitled Last Checkpoint: Last Thursday at 7:44 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

Variables

Inspection ID

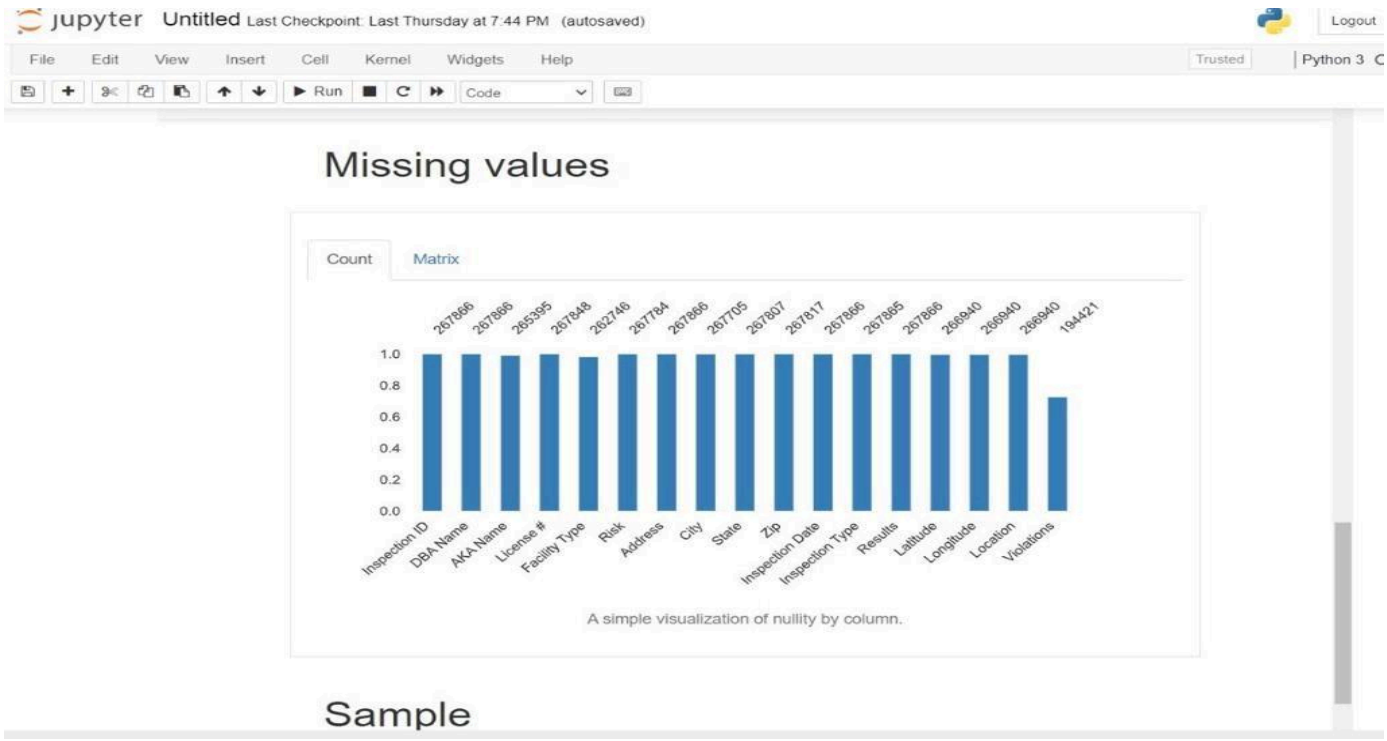
Inspection ID

Real number (ℝ)

UNIQUE

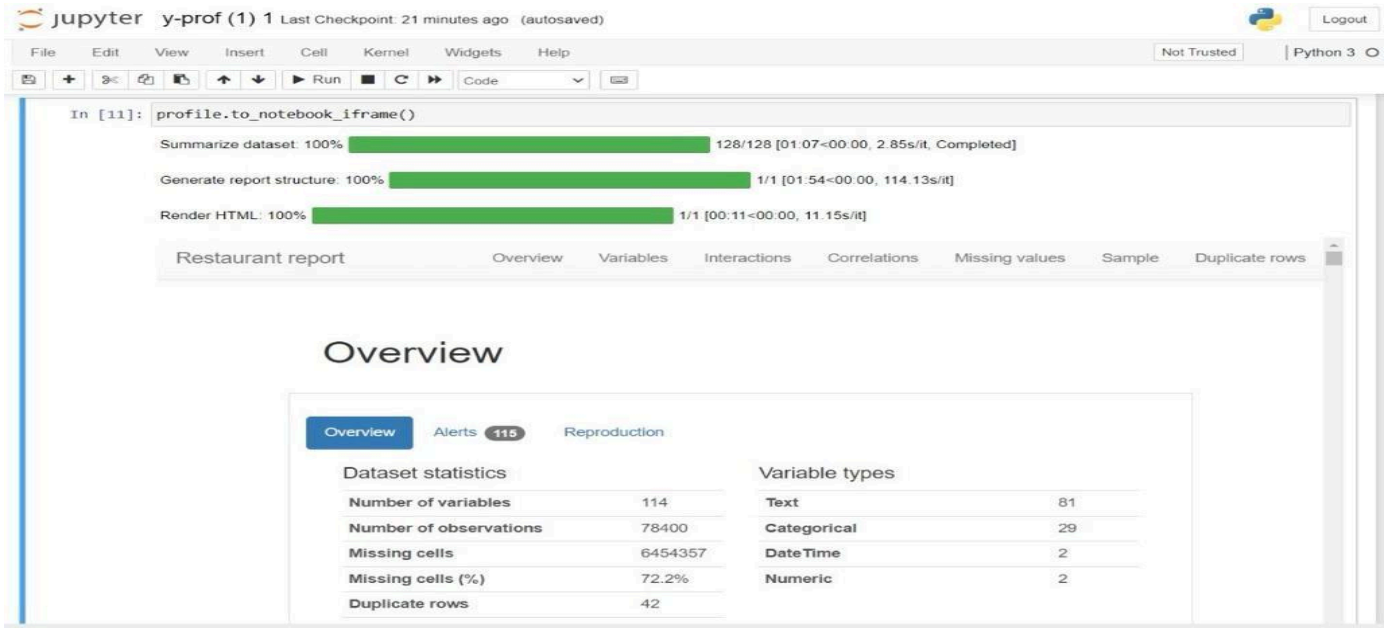
Distinct	267866	Minimum	44247
Distinct (%)	100.0%	Maximum	2589775
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1730790	Memory size	2.0 MiB

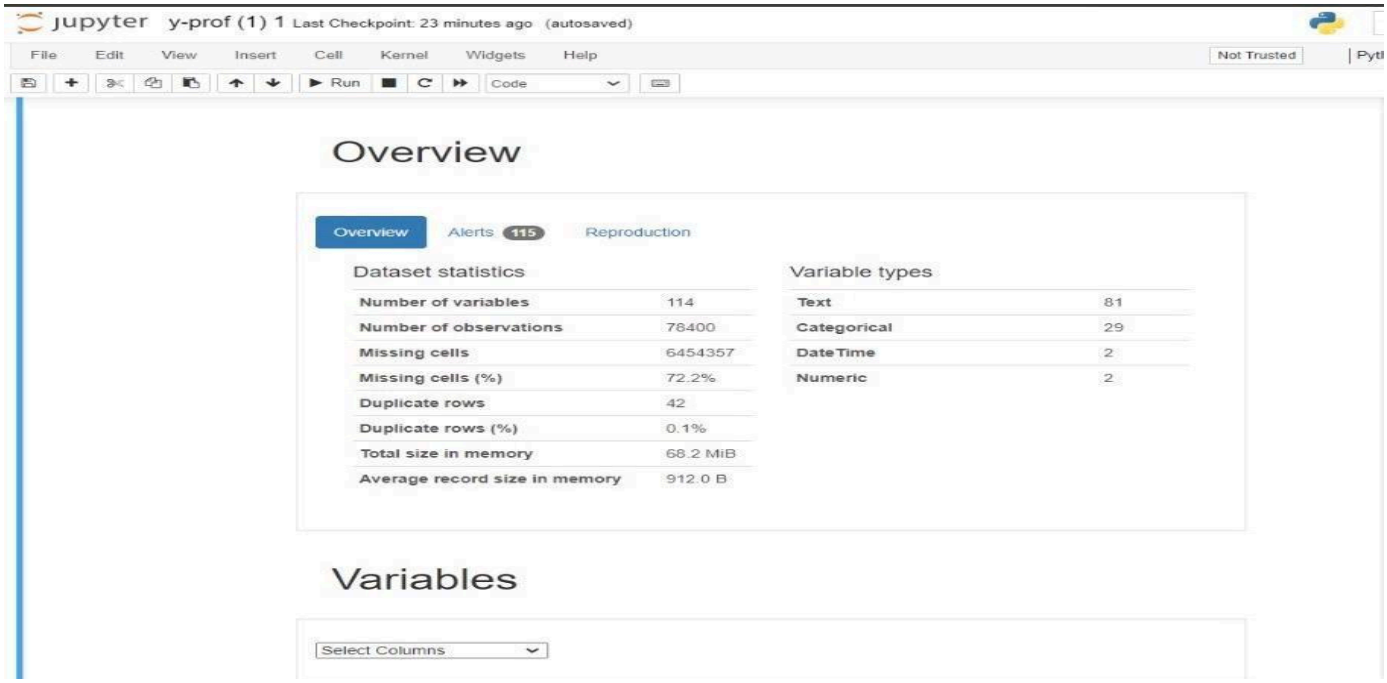
More details



In this report, we undertook a thorough data profiling process, analyzing missing values, distinct values, duplicate rows, and inappropriate data formats, alongside a meticulous review of column details. This comprehensive examination not only ensures data quality by identifying and rectifying issues like missing values, duplicate rows, and improper data formats but also enhances the reliability and accuracy of our analysis. Moreover, through this profiling, we pinpointed relevant attributes and measures crucial for our dimensional modeling efforts. These insights will serve as the foundation for the design and implementation phases, facilitating a robust and effective dimensional model.

YData Profiling to get Restaurant report:





jupyter y-prof (1) 1 Last Checkpoint: 21 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

Run Code

Restaurant report Overview Variables Interactions Correlations Missing values Sample Duplicate rows

Variables

Select Columns

Restaurant Name

Text

Distinct	9136
Distinct (%)	11.7%
Missing	11
Missing (%)	< 0.1%
Memory size	612.6 KiB

More details

Inspection Type

jupyter y-prof (1) 1 Last Checkpoint: 23 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

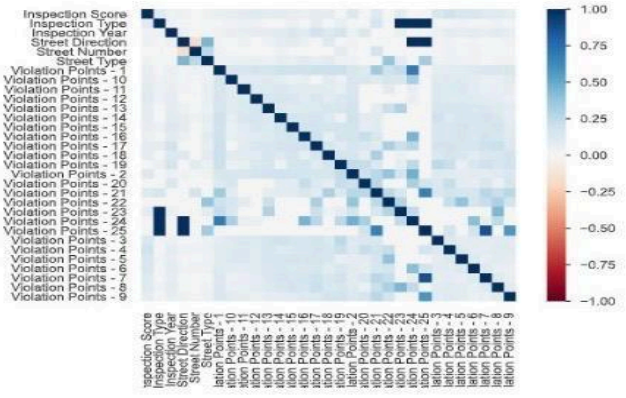
Not Trusted | Python 3

Run Code

Correlations

Auto

Heatmap Table



In [ ]:

jupyter y-prof (1) 1 Last Checkpoint: 23 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

Run Code

Missing values

Count Matrix Heatmap

A simple visualization of nullity by column.

Sample

First rows Last rows

In [ ]:

jupyter y-prof (1) 1 Last Checkpoint: 23 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

Run Code

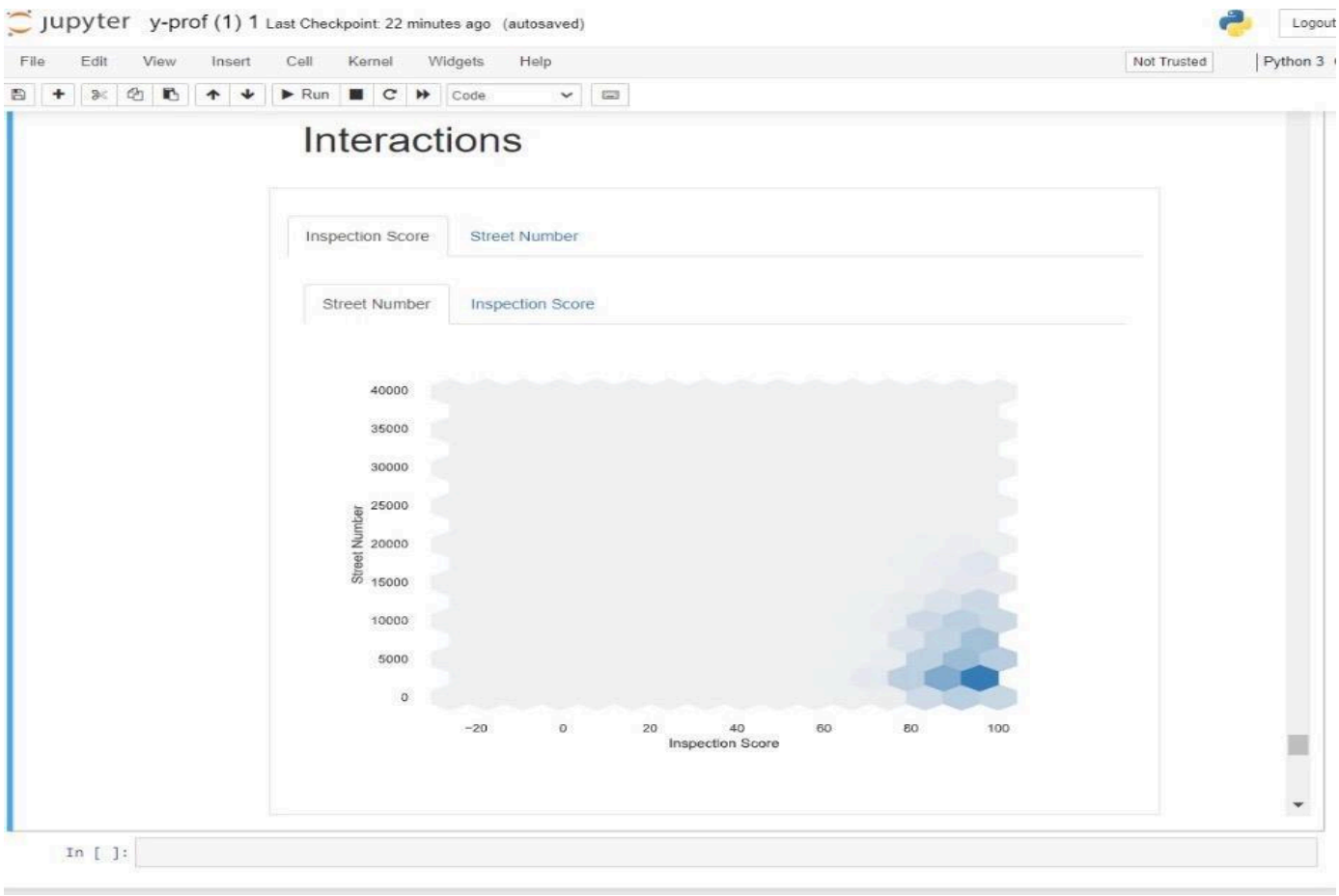
Sample

First rows Last rows

	Restaurant Name	Inspection Type	Inspection Date	Inspection Scc
0	MICKLE CHICKEN	Routine	10/30/2019	100
1	TOM THUMB - JUICE BAR	Routine	04/08/2020	100
2	BROOKDALE WHITE ROCK	Routine	07/29/2020	97
3	CHURCH'S CHICKEN #201	Routine	09/14/2020	100
4	PEAK PREPARATORY-PRIMARY SCHOOL	Routine	04/25/2017	98
5	SAMS CLUB #6482 - DEMO ROOM	Routine	04/05/2021	100
6	AMC THEATRES NORTHPARK 15 (BAR)	Routine	04/09/2021	100
7	SAM'S CLUB #8282 - DEMO ROOM	Routine	11/14/2021	100
8	UCHI DALLAS	Routine	03/11/2020	98
9	UNREFINED BAKERY	Routine	03/23/2020	100

Duplicate rows

In [ ]:



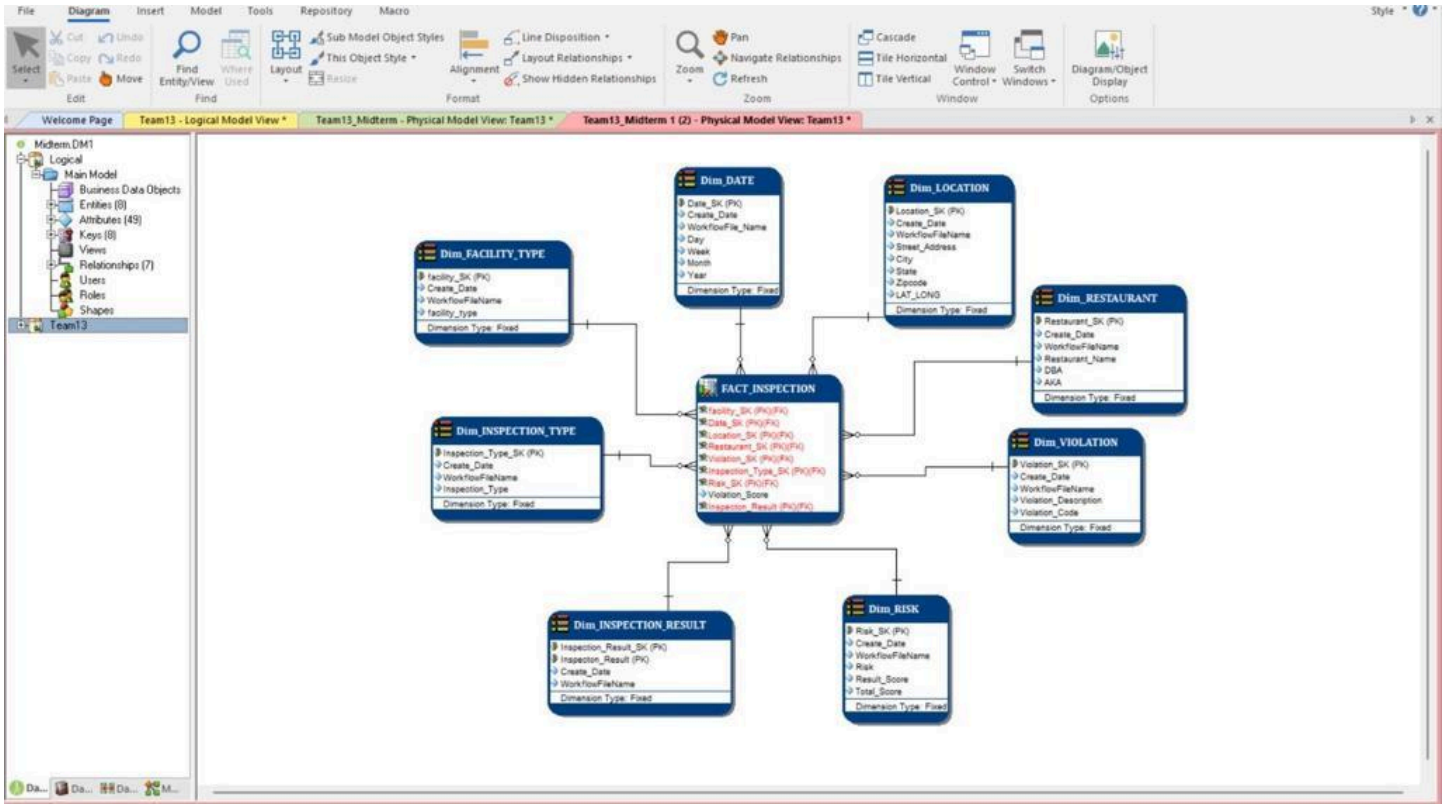
The report presents a comprehensive analysis encompassing various facets of data profiling. We systematically examined missing values, distinct values, duplicate rows, and inappropriate data formats, in addition to conducting a thorough review of column details. This meticulous examination ensures the quality and integrity of our data, laying a solid foundation for subsequent analysis.

Our thorough review of missing values, duplicate rows, and inappropriate data formats is pivotal in ensuring data quality. Identifying and addressing these issues at the outset enhances the reliability and accuracy of our analytical endeavors.

Moreover, our data profiling efforts were instrumental in delineating Dimensions and Facts essential for dimensional modeling. By pinpointing relevant attributes and measures early on, we streamline the subsequent design and implementation phases of our dimensional model. This proactive approach not only facilitates efficient data analysis and reporting but also underscores our commitment to delivering robust insights derived from high-quality data.

Part 2:

Designing dimensional model:



DDL SQL script:

```
/*

* TABLE: Dim_DATE

*/



CREATE TABLE Dim_DATE(

    Date_SK              int              NOT NULL,

    Create_Date          date             NOT NULL,

    WorkflowFile_Name    varchar(100)     NOT NULL,
```

```
   Day                  varchar(10)     NOT NULL,
   Week                 varchar(10)     NOT NULL,
   Month                varchar(10)     NOT NULL,
   Year                 varchar(10)     NOT NULL,
   CONSTRAINT PK7 PRIMARY KEY NONCLUSTERED (Date_SK)
)

go



IF OBJECT_ID('Dim_DATE') IS NOT NULL
   PRINT '<<< CREATED TABLE Dim_DATE >>>'
ELSE
   PRINT '<<< FAILED CREATING TABLE Dim_DATE >>>'
go


/*
* TABLE: Dim_FACILITY_TYPE
*/

CREATE TABLE Dim_FACILITY_TYPE(
```

```
   facility_SK          int             NOT NULL,
   Create_Date          date            NOT NULL,
   WorkflowFileName     varchar(100)    NOT NULL,
   facility_type        varchar(100)    NULL,
   CONSTRAINT PK5 PRIMARY KEY NONCLUSTERED (facility_SK)
)


go




IF OBJECT_ID('Dim_FACILITY_TYPE') IS NOT NULL
   PRINT '<<< CREATED TABLE Dim_FACILITY_TYPE >>>'
ELSE
   PRINT '<<< FAILED CREATING TABLE Dim_FACILITY_TYPE >>>'
go


/*
* TABLE: Dim_INSPECTION_RESULT
*/


CREATE TABLE Dim_INSPECTION_RESULT(
```

```
   Inspection_Result_SK    int             NOT NULL,
   Inspecton_Result        char(10)        NOT NULL,
   Create_Date             date            NOT NULL,
   WorkflowFileName        varchar(100)    NOT NULL,
   CONSTRAINT PK11_1 PRIMARY KEY NONCLUSTERED (Inspection_Result_SK, Inspecton_Result)
)

go



IF OBJECT_ID('Dim_INSPECTION_RESULT') IS NOT NULL
   PRINT '<<< CREATED TABLE Dim_INSPECTION_RESULT >>>'
ELSE
   PRINT '<<< FAILED CREATING TABLE Dim_INSPECTION_RESULT >>>'
go


/*
* TABLE: Dim_INSPECTION_TYPE
*/

CREATE TABLE Dim_INSPECTION_TYPE(
```

```
   Inspection_Type_SK    int             NOT NULL,
   Create_Date           date            NOT NULL,
   WorkflowFileName      varchar(100)    NOT NULL,
   Inspection_Type       varchar(100)    NULL,
   CONSTRAINT PK11 PRIMARY KEY NONCLUSTERED (Inspection_Type_SK)
)

go



IF OBJECT_ID('Dim_INSPECTION_TYPE') IS NOT NULL
   PRINT '<<< CREATED TABLE Dim_INSPECTION_TYPE >>>'
ELSE
   PRINT '<<< FAILED CREATING TABLE Dim_INSPECTION_TYPE >>>'
go


/*
* TABLE: Dim_LOCATION
*/

CREATE TABLE Dim_LOCATION(
```

```
   Location_SK          int               NOT NULL,
   Create_Date          date              NOT NULL,
   WorkflowFileName     varchar(100)      NOT NULL,
   Street_Address       varchar(100)      NOT NULL,
   City                 varchar(50)       NULL,
   State                varchar(50)       NULL,
   Zipcode              varchar(50)       NULL,
   LAT_LONG             decimal(50, 0)    NULL,
   CONSTRAINT PK8 PRIMARY KEY NONCLUSTERED (Location_SK)
)


go



IF OBJECT_ID('Dim_LOCATION') IS NOT NULL
   PRINT '<<< CREATED TABLE Dim_LOCATION >>>'
ELSE
   PRINT '<<< FAILED CREATING TABLE Dim_LOCATION >>>'
go

/*
```

```
* TABLE: Dim_RESTAURANT

*/


CREATE TABLE Dim_RESTAURANT(

   Restaurant_SK       int             NOT NULL,

   Create_Date         date            NOT NULL,

   WorkflowFileName    varchar(100)    NOT NULL,

   Restaurant_Name     varchar(100)    NULL,

   DBA                 varchar(100)    NULL,

   AKA                 varchar(100)    NULL,

   CONSTRAINT PK2 PRIMARY KEY NONCLUSTERED (Restaurant_SK)

)


go



IF OBJECT_ID('Dim_RESTAURANT') IS NOT NULL

   PRINT '<<< CREATED TABLE Dim_RESTAURANT >>>'

ELSE

   PRINT '<<< FAILED CREATING TABLE Dim_RESTAURANT >>>'

go
```

```
/*
* TABLE: Dim_RISK
*/

CREATE TABLE Dim_RISK(
   Risk_SK             int                NOT NULL,
   Create_Date         date               NOT NULL,
   WorkflowFileName    varchar(100)       NOT NULL,
   Risk                numeric(100, 0)    NULL,
   Result_Score        numeric(10, 0)     NULL,
   Total_Score         numeric(10, 0)     NULL,
   CONSTRAINT PK20 PRIMARY KEY NONCLUSTERED (Risk_SK)
)

go

IF OBJECT_ID('Dim_RISK') IS NOT NULL
   PRINT '<<< CREATED TABLE Dim_RISK >>>'
ELSE
```

```
   PRINT '<<< FAILED CREATING TABLE Dim_RISK >>>'
go


/*
* TABLE: Dim_VIOLATION
*/


CREATE TABLE Dim_VIOLATION(
   Violation_SK             int              NOT NULL,
   Create_Date              date             NOT NULL,
   WorkflowFileName         varchar(100)     NOT NULL,
   Violation_Description    varchar(1000)    NULL,
   Violation_Code           int              NULL,
   CONSTRAINT PK10 PRIMARY KEY NONCLUSTERED (Violation_SK)
)


go



IF OBJECT_ID('Dim_VIOLATION') IS NOT NULL
   PRINT '<<< CREATED TABLE Dim_VIOLATION >>>'
```

```
ELSE
   PRINT '<<< FAILED CREATING TABLE Dim_VIOLATION >>>'
go

/*
* TABLE: FACT_INSPECTION
*/

CREATE TABLE FACT_INSPECTION(
   facility_SK           int         NOT NULL,
   Date_SK               int         NOT NULL,
   Location_SK           int         NOT NULL,
   Restaurant_SK         int         NOT NULL,
   Violation_SK          int         NOT NULL,
   Inspection_Type_SK    int         NOT NULL,
   Risk_SK               int         NOT NULL,
   Violation_Score       int         NULL,
   Inspecton_Result      char(10)    NOT NULL,
   CONSTRAINT PK12 PRIMARY KEY NONCLUSTERED (facility_SK, Date_SK, Location_SK, Restaurant_SK,
Violation_SK, Inspection_Type_SK, Risk_SK, Inspecton_Result)
)
```

```
go



IF OBJECT_ID('FACT_INSPECTION') IS NOT NULL
   PRINT '<<< CREATED TABLE FACT_INSPECTION >>>'
ELSE
   PRINT '<<< FAILED CREATING TABLE FACT_INSPECTION >>>'
go

/*
* TABLE: FACT_INSPECTION
*/

ALTER TABLE FACT_INSPECTION ADD CONSTRAINT RefDim_FACILITY_TYPE1
   FOREIGN KEY (facility_SK)
   REFERENCES Dim_FACILITY_TYPE(facility_SK)
go

ALTER TABLE FACT_INSPECTION ADD CONSTRAINT RefDim_DATE2
   FOREIGN KEY (Date_SK)
   REFERENCES Dim_DATE(Date_SK)
```

```
go


ALTER TABLE FACT_INSPECTION ADD CONSTRAINT RefDim_LOCATION3
   FOREIGN KEY (Location_SK)
   REFERENCES Dim_LOCATION(Location_SK)
go


ALTER TABLE FACT_INSPECTION ADD CONSTRAINT RefDim_RESTAURANT4
   FOREIGN KEY (Restaurant_SK)
   REFERENCES Dim_RESTAURANT(Restaurant_SK)
go


ALTER TABLE FACT_INSPECTION ADD CONSTRAINT RefDim_VIOLATION5
   FOREIGN KEY (Violation_SK)
   REFERENCES Dim_VIOLATION(Violation_SK)
go


ALTER TABLE FACT_INSPECTION ADD CONSTRAINT RefDim_INSPECTION_TYPE6
   FOREIGN KEY (Inspection_Type_SK)
   REFERENCES Dim_INSPECTION_TYPE(Inspection_Type_SK)
go
```

```
ALTER TABLE FACT_INSPECTION ADD CONSTRAINT RefDim_RISK7

   FOREIGN KEY (Risk_SK)

   REFERENCES Dim_RISK(Risk_SK)

go



ALTER TABLE FACT_INSPECTION ADD CONSTRAINT RefDim_INSPECTION_RESULT8

   FOREIGN KEY (Inspection_Type_SK, Inspecton_Result)

   REFERENCES Dim_INSPECTION_RESULT(Inspection_Result_SK, Inspecton_Result)

go
```

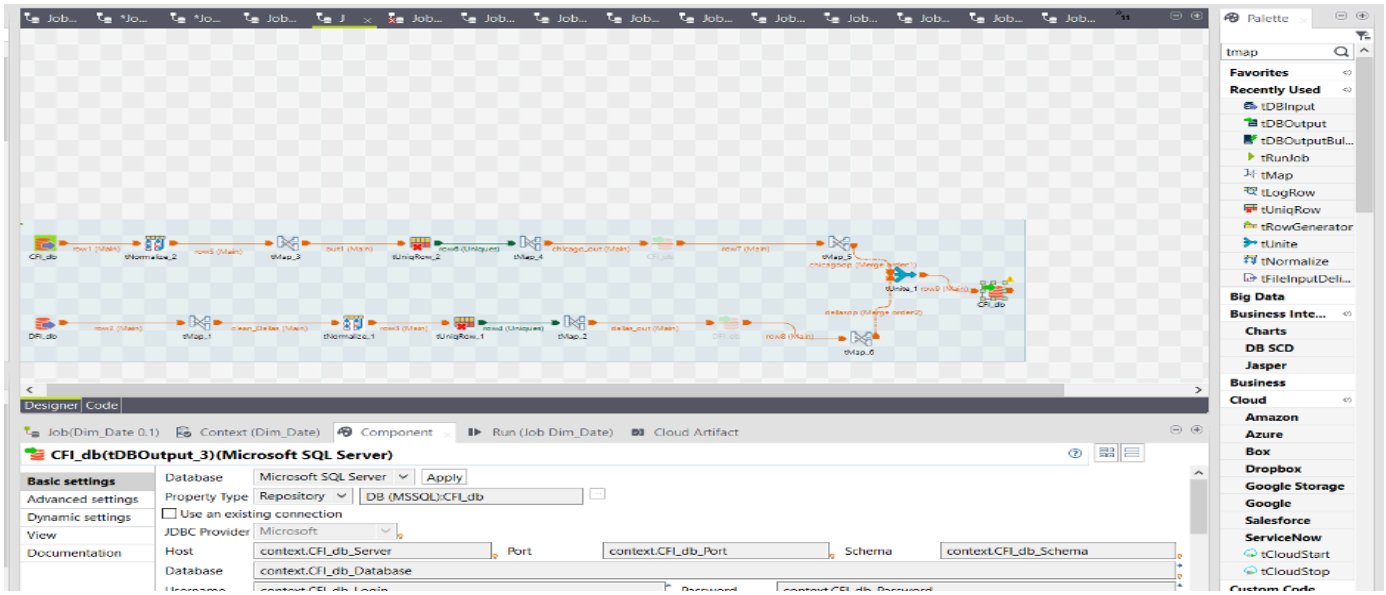
Schema in database:

Please attach ss

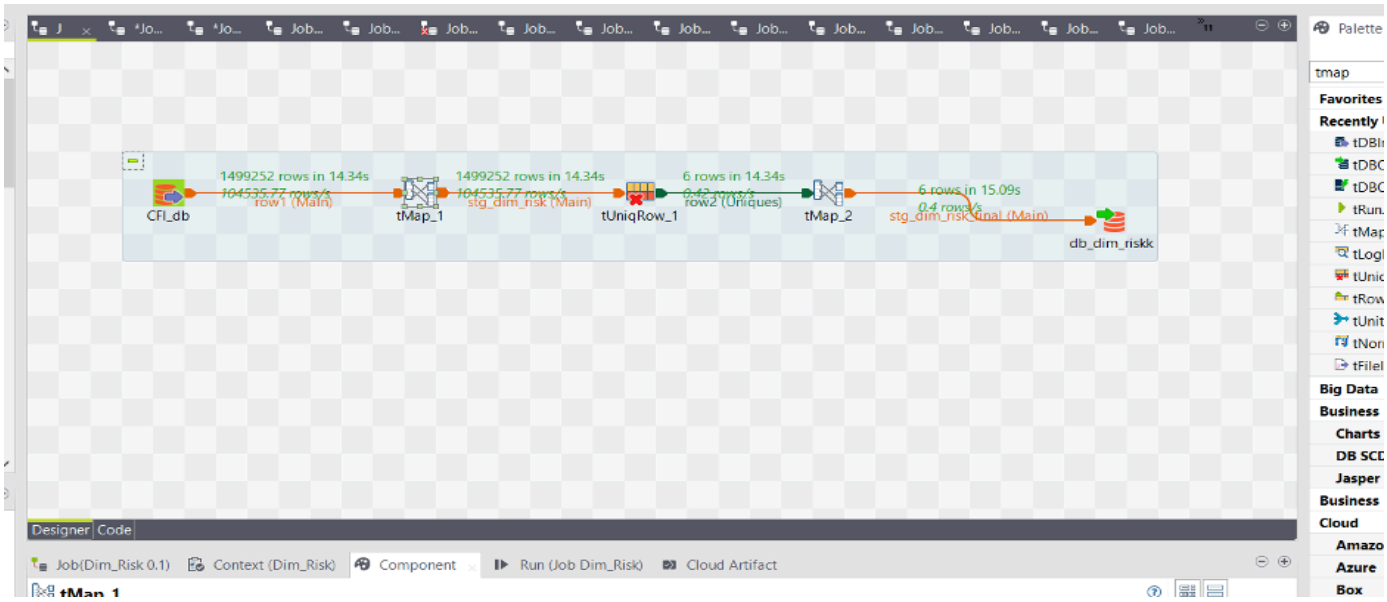
Part 3:

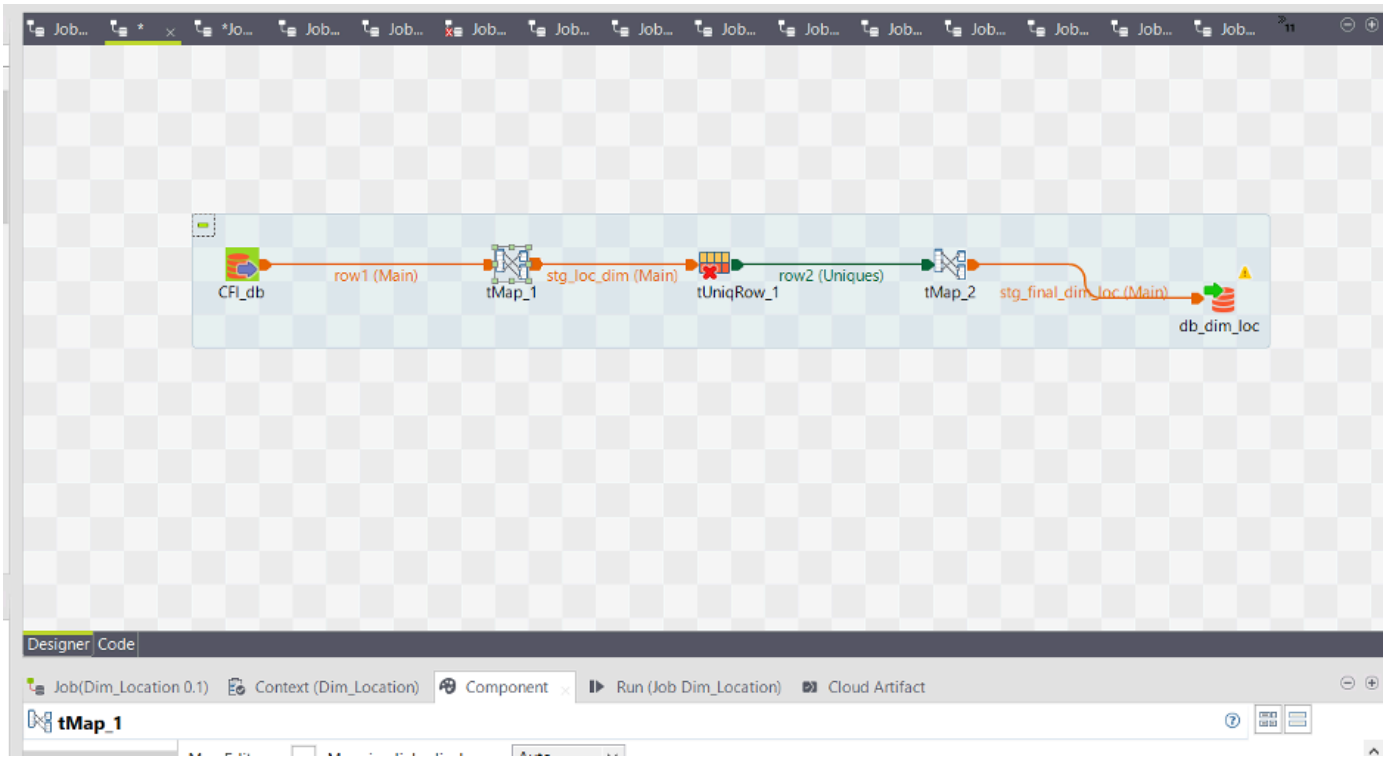
Dimensional Modeling using Talend:

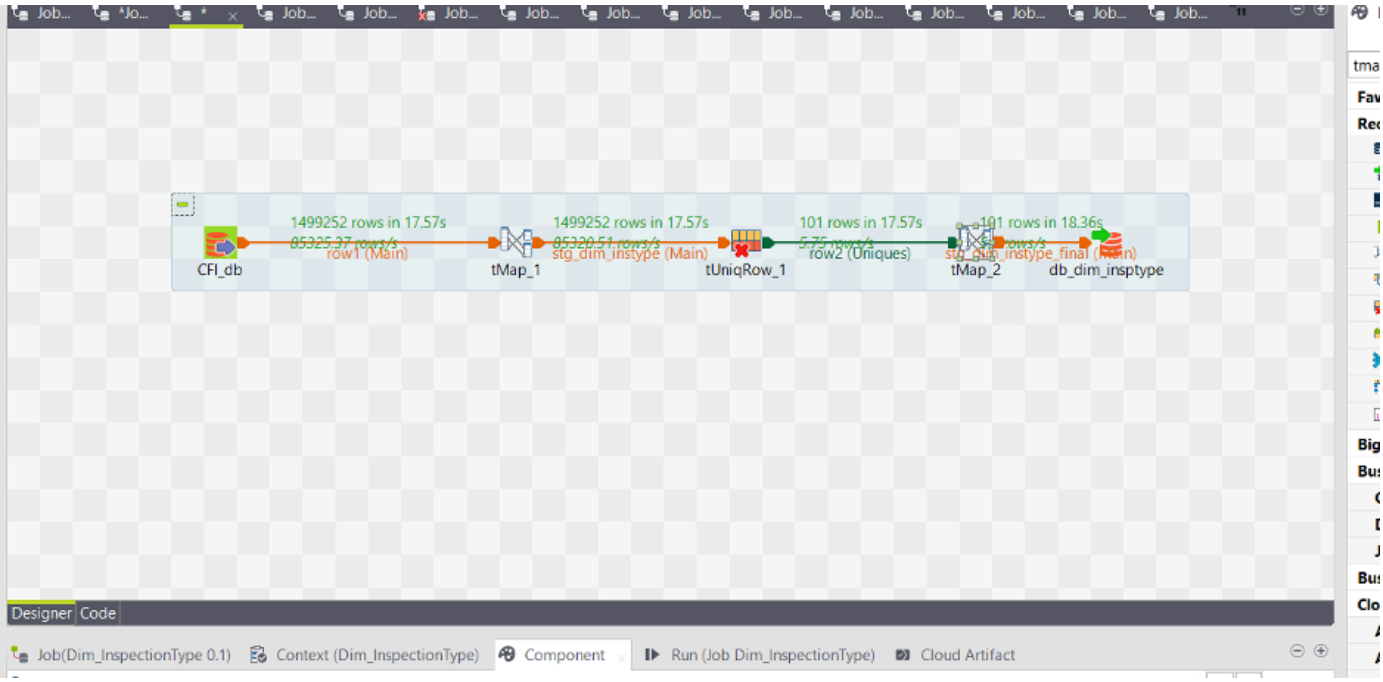
Uniting both datasets:

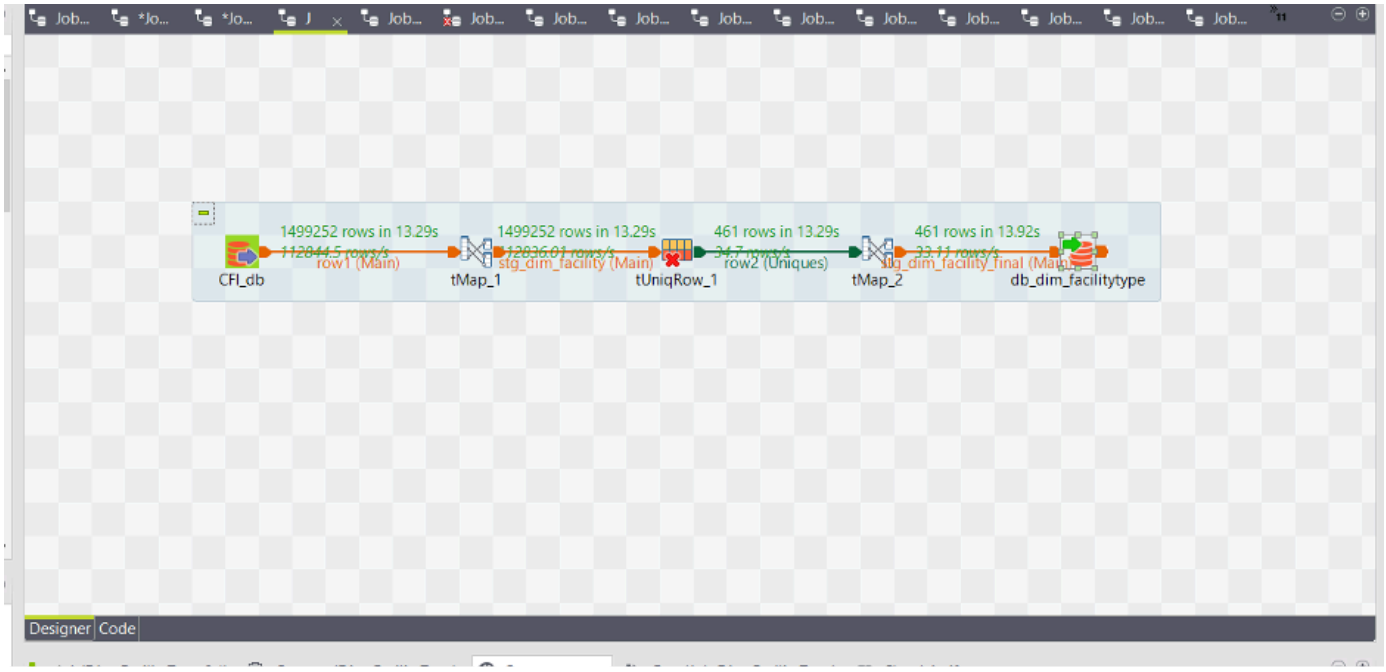


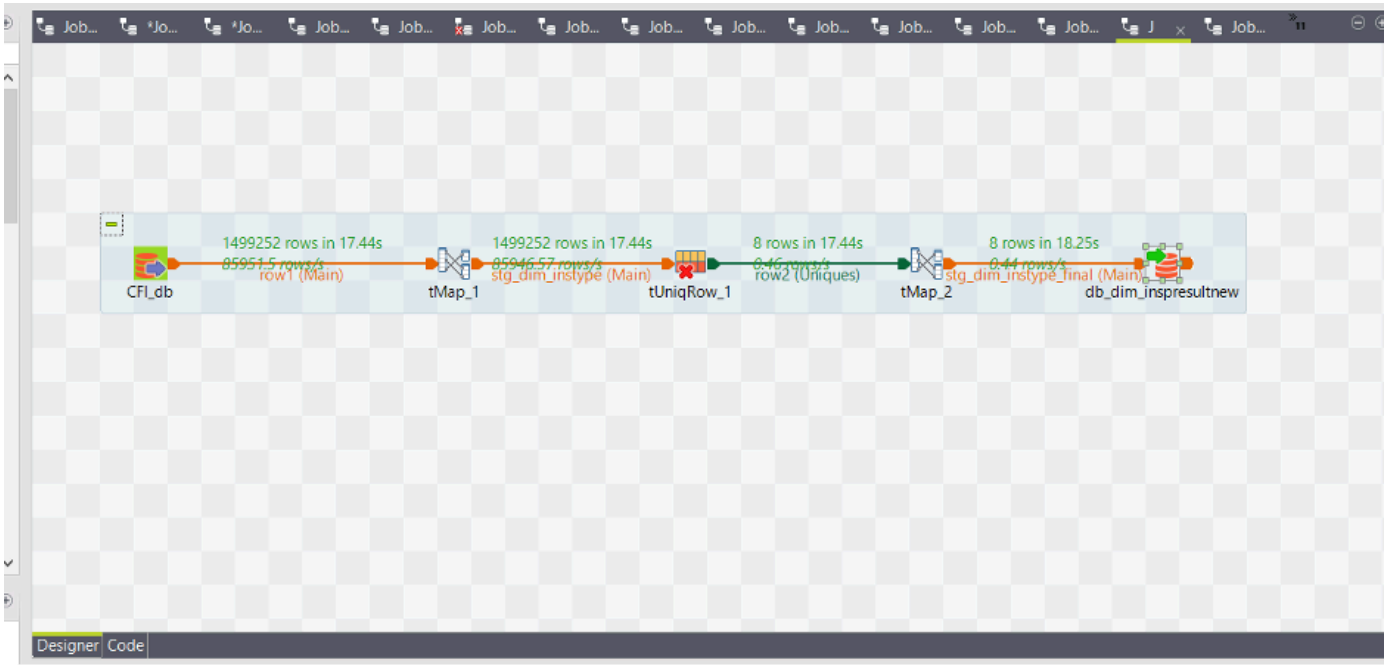
Loading Dimensions:



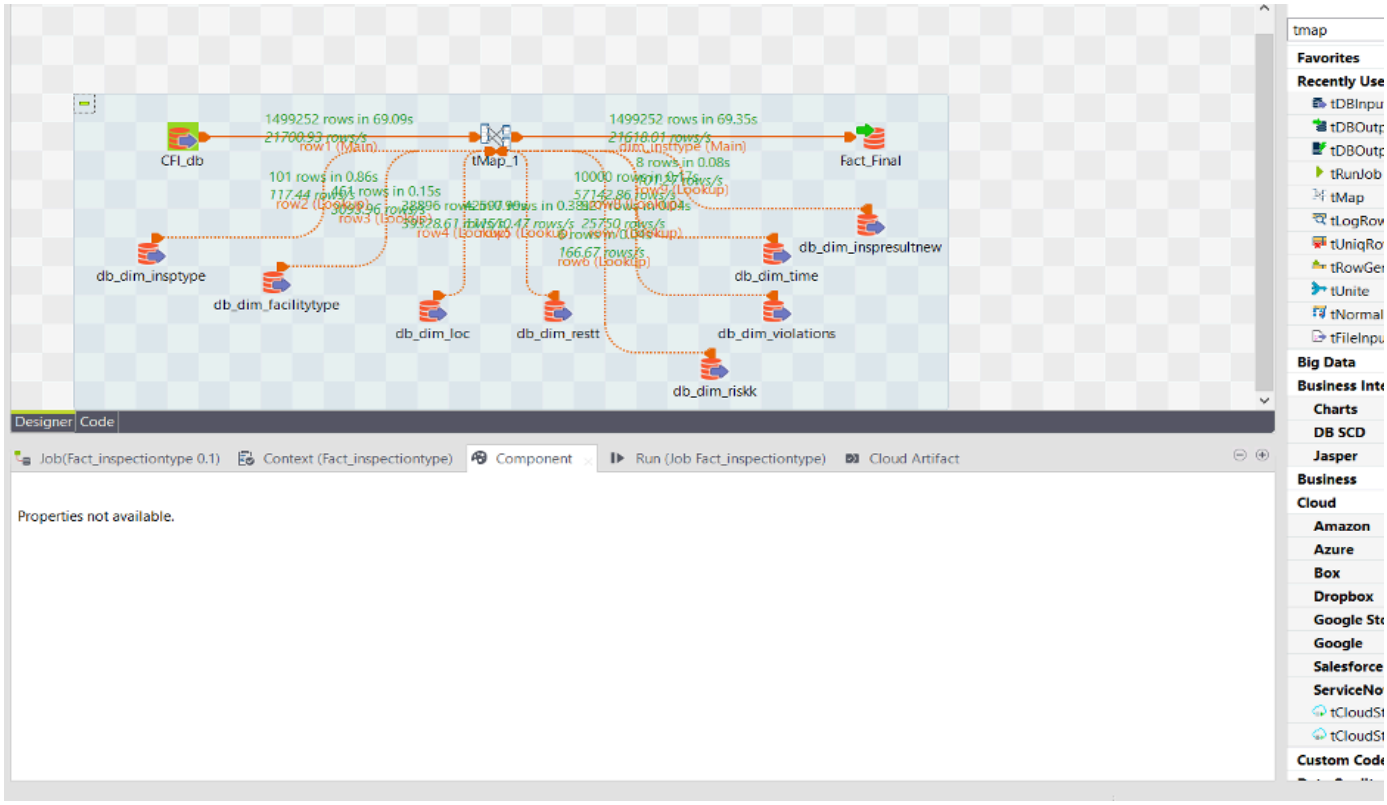

CFI_db
row1 (Main)
tMap_1
stg_loc_dim (Main)
tUniqRow_1
row2 (Uniques)
tMap_2
stg_final_dim_loc (Main)
db_dim_loc
Designer
Code
Job(Dim_Location 0.1)
Context (Dim_Location)
Component
Run (Job Dim_Location)
Cloud Artifact
tMap_1


1499252 rows in 17.57s
85325.37 rows/s
row1 (Main)
1499252 rows in 17.57s
85320.51 rows/s
stg_dim_instype (Main)
101 rows in 17.57s
5.75 rows/s
row2 (Uniques)
101 rows in 18.36s
5.5 rows/s
stg_dim_instype_final (Main)
CFI_db
tMap_1
tUniqRow_1
tMap_2
db_dim_insptype
Designer
Code
Job(Dim_InspectionType 0.1)
Context (Dim_InspectionType)
Component
Run (Job Dim_InspectionType)
Cloud Artifact


1499252 rows in 13.29s
112844.5 rows/s
row1 (Main)
1499252 rows in 13.29s
12836.01 rows/s
stg_dim_facility (Main)
461 rows in 13.29s
34.7 rows/s
row2 (Uniques)
461 rows in 13.92s
33.11 rows/s
stg_dim_facility_final (Main)
CFI_db
tMap_1
tUniqRow_1
tMap_2
db_dim_facilitytype
Designer
Code

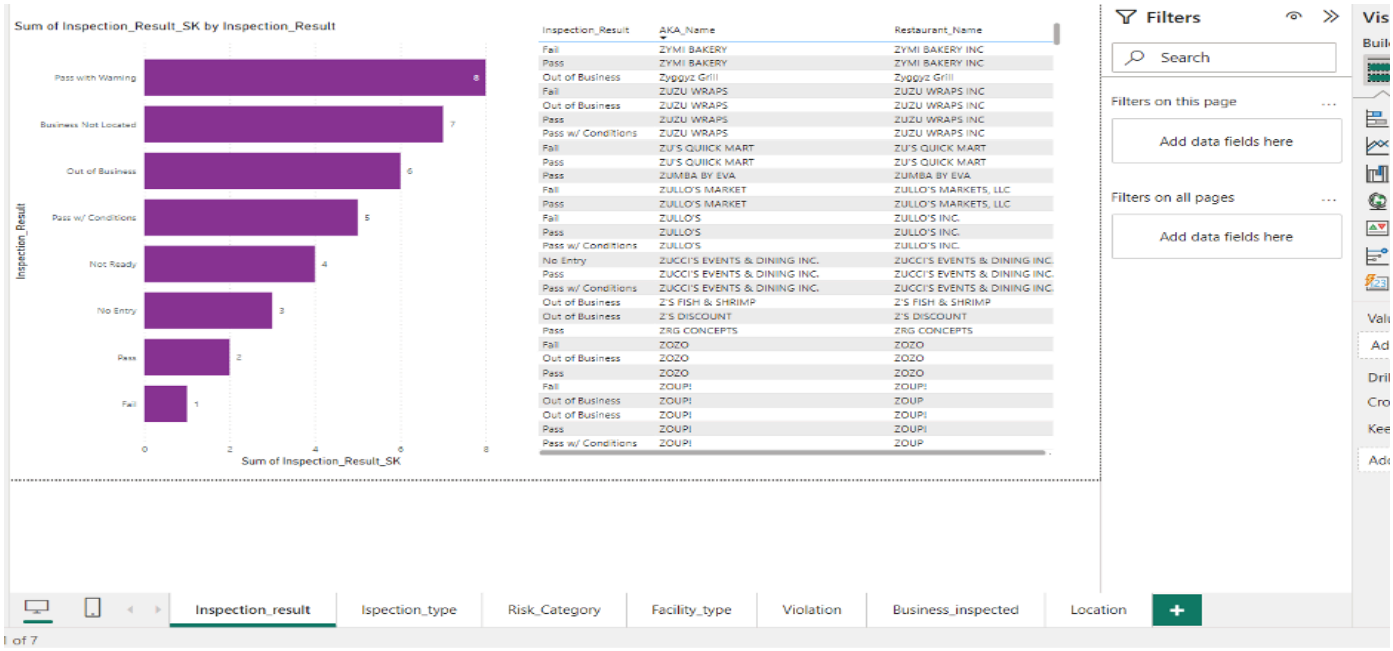

1499252 rows in 17.44s
85951.5 rows/s
row1 (Main)
1499252 rows in 17.44s
85946.57 rows/s
stg_dim_instype (Main)
8 rows in 17.44s
0.46 rows/s
row2 (Uniques)
8 rows in 18.25s
0.44 rows/s
stg_dim_instype_final (Main)
CFI_db
tMap_1
tUniqRow_1
tMap_2
db_dim_inspresultnew
Designer
Code

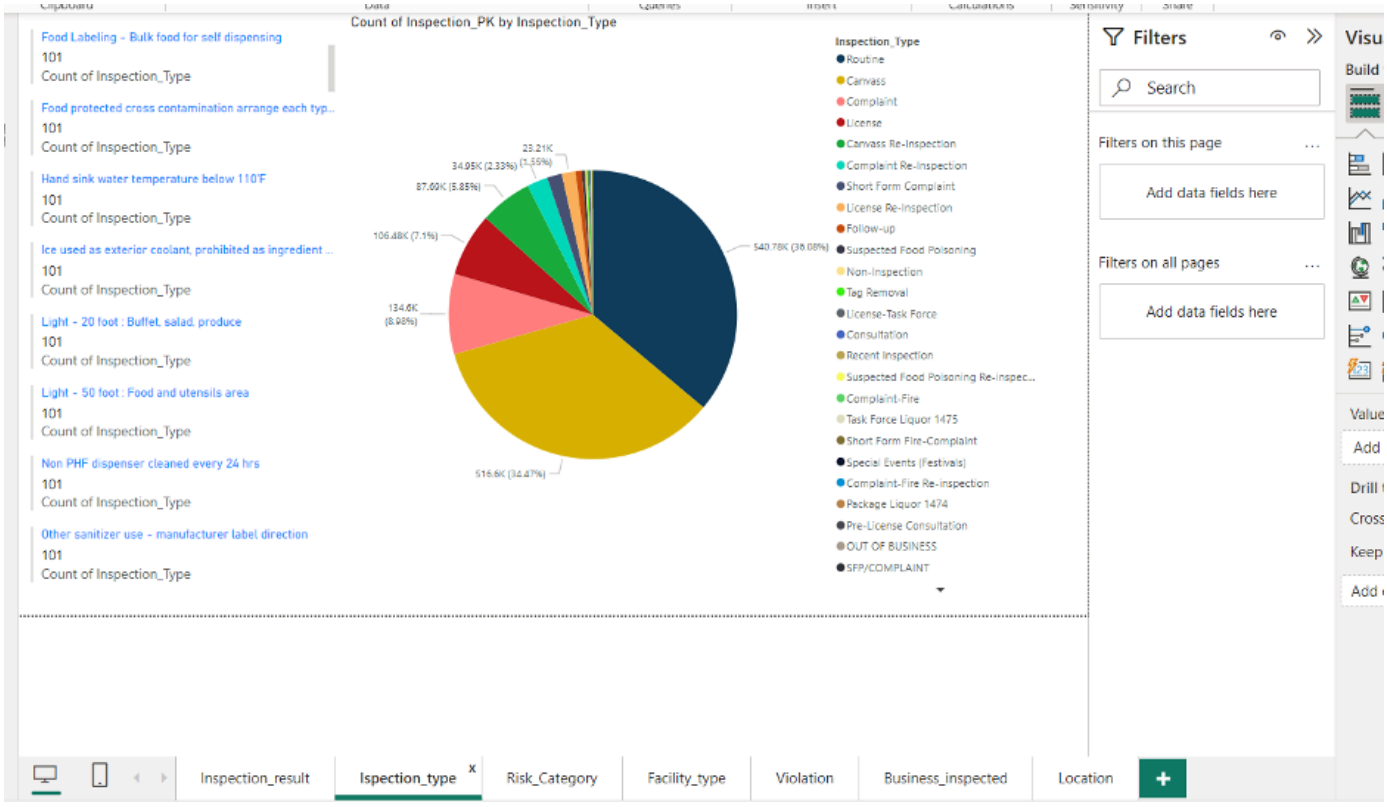
Loading Fact table:



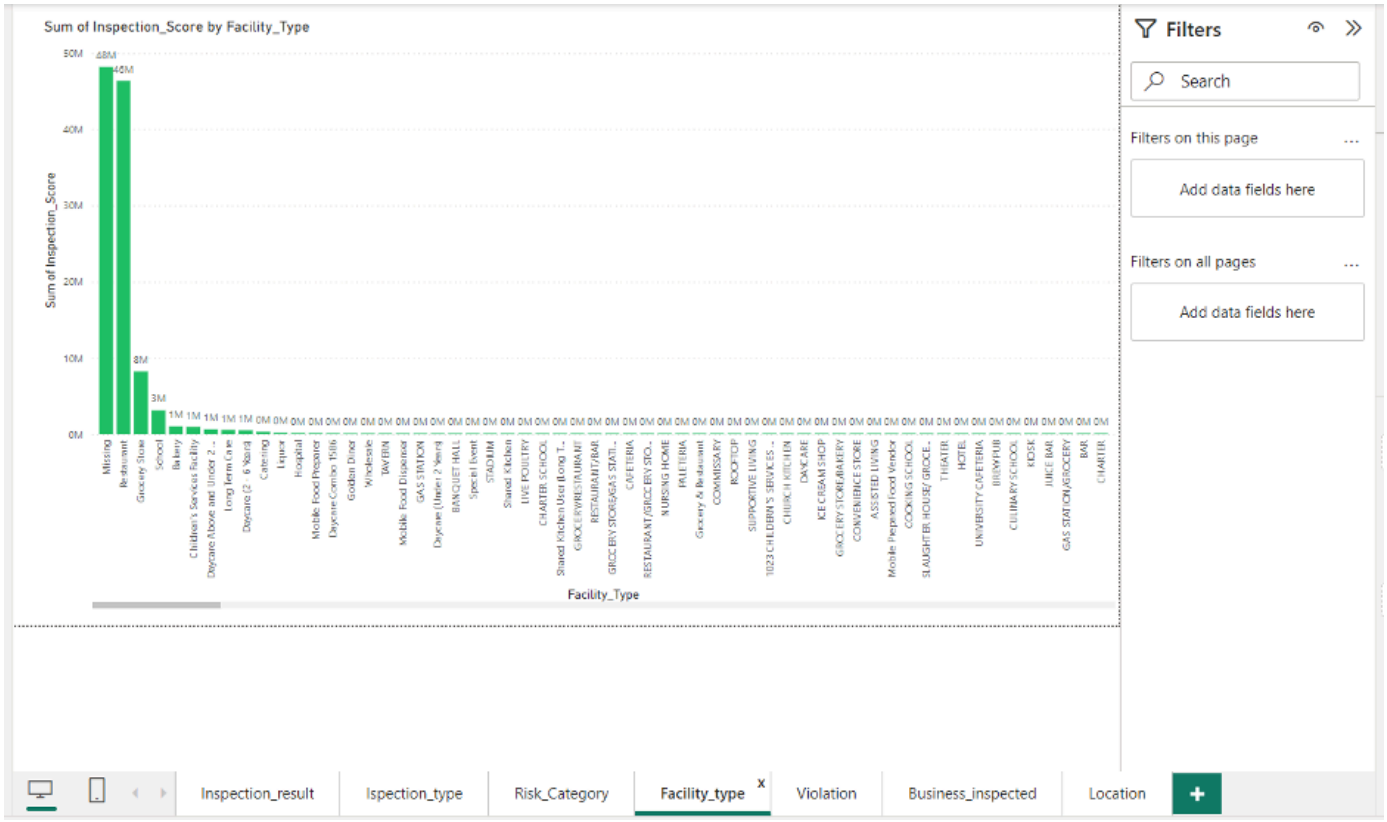
Part 4:

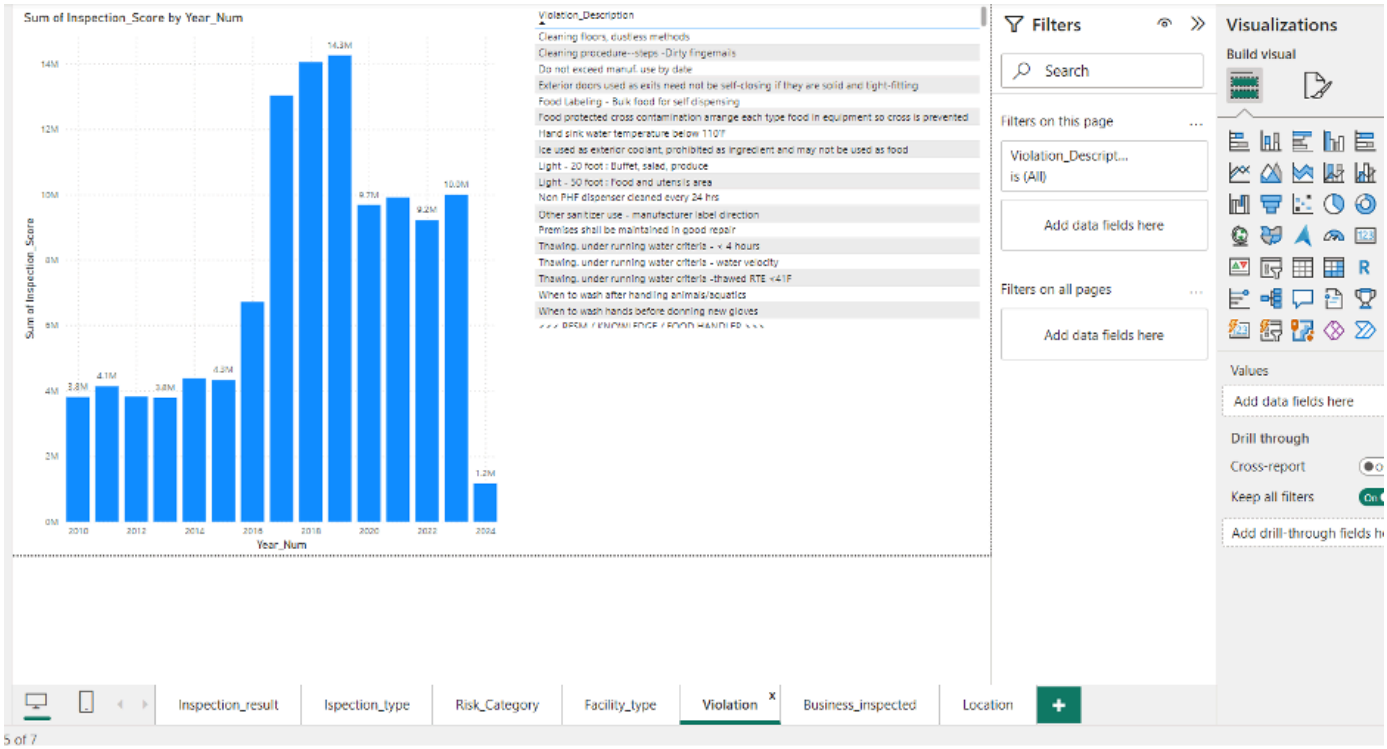
Power BI Dashboards:

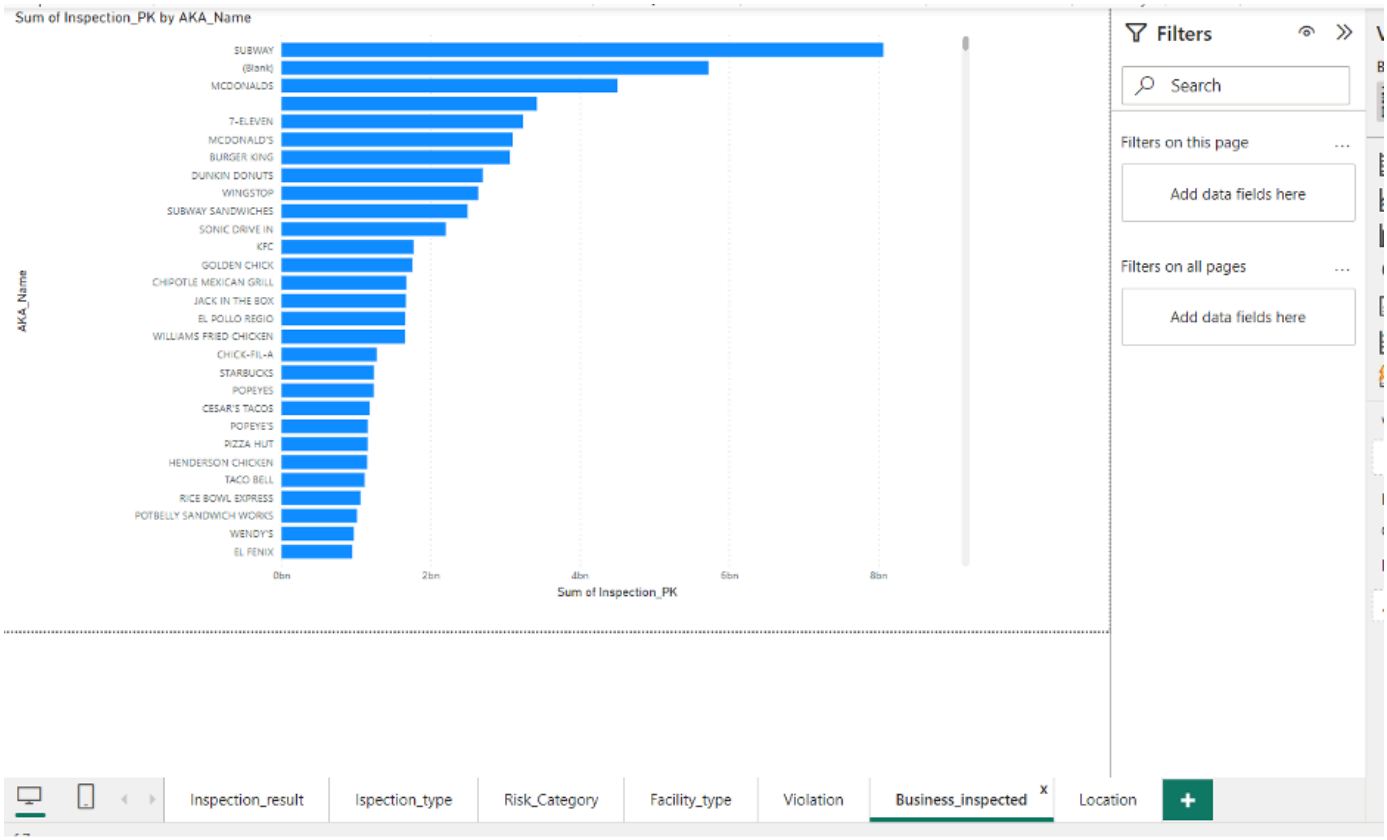



Food Labeling - Bulk food for self dispensing
101
Count of Inspection_Type
Food protected cross contamination arrange each typ...
101
Count of Inspection_Type
Hand sink water temperature below 110°F
101
Count of Inspection_Type
Ice used as exterior coolant, prohibited as ingredient ...
101
Count of Inspection_Type
Light - 20 foot : Buffet, salad, produce
101
Count of Inspection_Type
Light - 50 foot : Food and utensils area
101
Count of Inspection_Type
Non PHF dispenser cleaned every 24 hrs
101
Count of Inspection_Type
Other sanitizer use - manufacturer label direction
101
Count of Inspection_Type
Count of Inspection_PK by Inspection_Type
23.21K (1.55%)
34.95K (2.33%)
87.69K (5.85%)
106.48K (7.1%)
134.6K (8.98%)
540.78K (36.08%)
516.6K (34.47%)
Inspection_Type
Routine
Canvass
Complaint
License
Canvass Re-Inspection
Complaint Re-Inspection
Short Form Complaint
License Re-Inspection
Follow-up
Suspected Food Poisoning
Non-Inspection
Tag Removal
License-Task Force
Consultation
Recent Inspection
Suspected Food Poisoning Re-Inspec...
Complaint-Fire
Task Force Liquor 1475
Short Form Fire-Complaint
Special Events (Festivals)
Complaint-Fire Re-Inspection
Package Liquor 1474
Pre-License Consultation
OUT OF BUSINESS
SFP/COMPLAINT
Filters
Search
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Inspection_result
Ispection_type
Risk_Category
Facility_type
Violation
Business_inspected
Location


Sum of Risk_SK by Risk
0.8M
0.6M
0.4M
0.2M
0.0M
Sum of Risk_SK
0.73M
0.65M
0.11M
0.00M
High
Medium
Low
All
Risk
Sum of Risk_SK
Filters
Search
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Values
Drill through
Cross-report
Keep all filters
Inspection_result
Ispection_type
Risk_Category
Facility_type
Violation
Business_inspected
Location


Sum of Inspection_Score by Facility_Type
Sum of Inspection_Score
Facility_Type
Filters
Search
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Inspection_result
Ispection_type
Risk_Category
Facility_type
Violation
Business_inspected
Location


Sum of Inspection_Score by Year_Num
Sum of Inspection_Score
Year_Num
Violation_Description
Cleaning floors, dustless methods
Cleaning procedure--steps -Dirty fingernails
Do not exceed manuf. use by date
Exterior doors used as exits need not be self-closing if they are solid and tight-fitting
Food Labeling - Bulk food for self dispensing
Food protected cross contamination arrange each type food in equipment so cross is prevented
Hand sink water temperature below 110F
Ice used as exterior coolant, prohibited as ingredient and may not be used as food
Light - 20 foot : Buffet, salad, produce
Light - 50 foot : Food and utensils area
Non PHF dispenser cleaned every 24 hrs
Other sanitizer use - manufacturer label direction
Premises shall be maintained in good repair
Thawing, under running water criteria - < 4 hours
Thawing, under running water criteria - water velocity
Thawing, under running water criteria -thawed RTE <41F
When to wash after handling animals/aquatics
When to wash hands before donning new gloves
Filters
Search
Filters on this page
Violation_Descript...
is (All)
Add data fields here
Filters on all pages
Add data fields here
Visualizations
Build visual
Values
Add data fields here
Drill through
Cross-report
Keep all filters
Add drill-through fields h
Inspection_result
Ispection_type
Risk_Category
Facility_type
Violation
Business_inspected
Location


Sum of Inspection_PK by AKA_Name
SUBWAY
(Blank)
MCDONALDS
7-ELEVEN
MCDONALD'S
BURGER KING
DUNKIN DONUTS
WINGSTOP
SUBWAY SANDWICHES
SONIC DRIVE IN
KFC
GOLDEN CHICK
CHIPOTLE MEXICAN GRILL
JACK IN THE BOX
EL POLLO REGIO
WILLIAMS FRIED CHICKEN
CHICK-FIL-A
STARBUCKS
POPEYES
CESAR'S TACOS
POPEYE'S
PIZZA HUT
HENDERSON CHICKEN
TACO BELL
RICE BOWL EXPRESS
POTBELLY SANDWICH WORKS
WENDY'S
EL FENIX
AKA_Name
0bn
2bn
4bn
6bn
8bn
Sum of Inspection_PK
Filters
Search
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Inspection_result
Ispection_type
Risk_Category
Facility_type
Violation
Business_inspected
Location

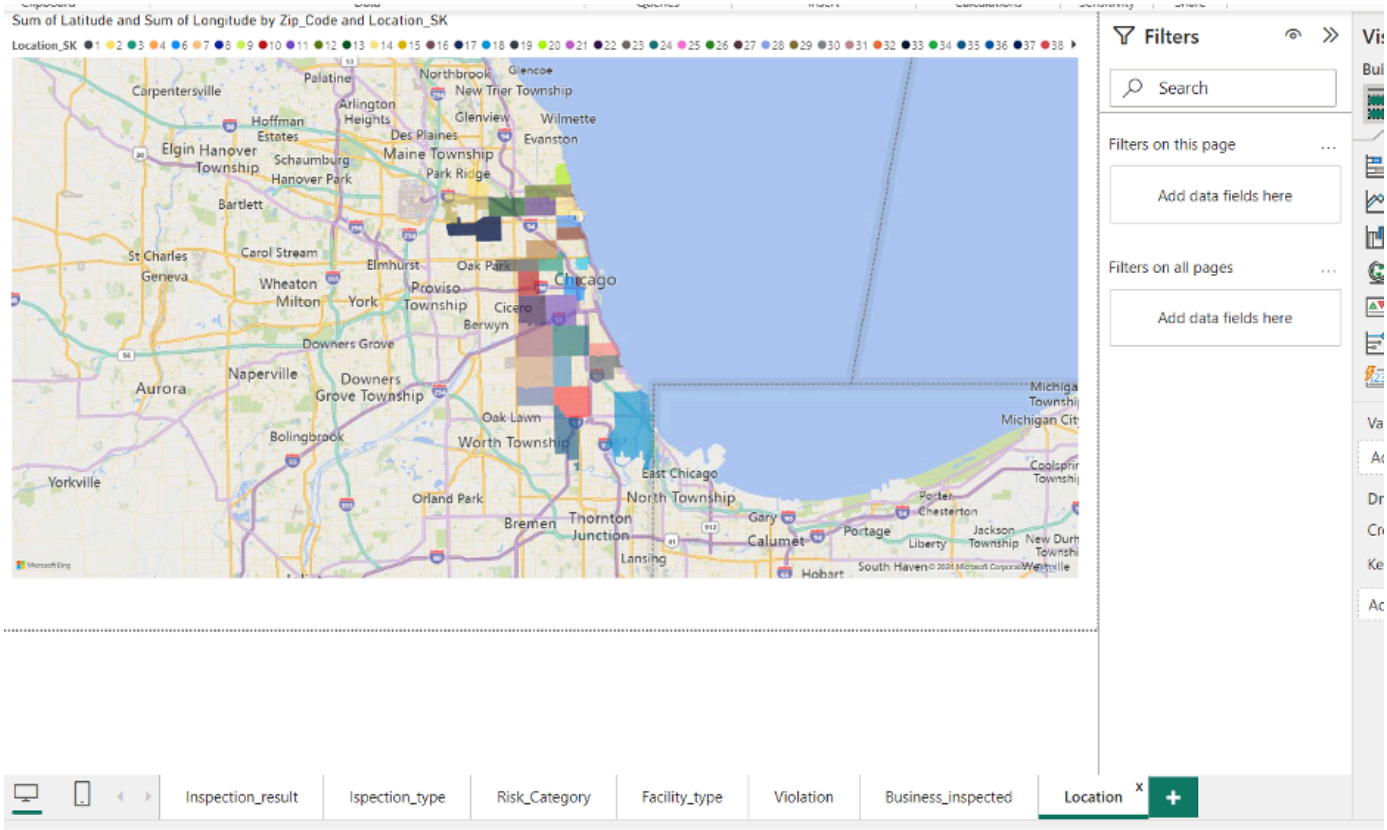
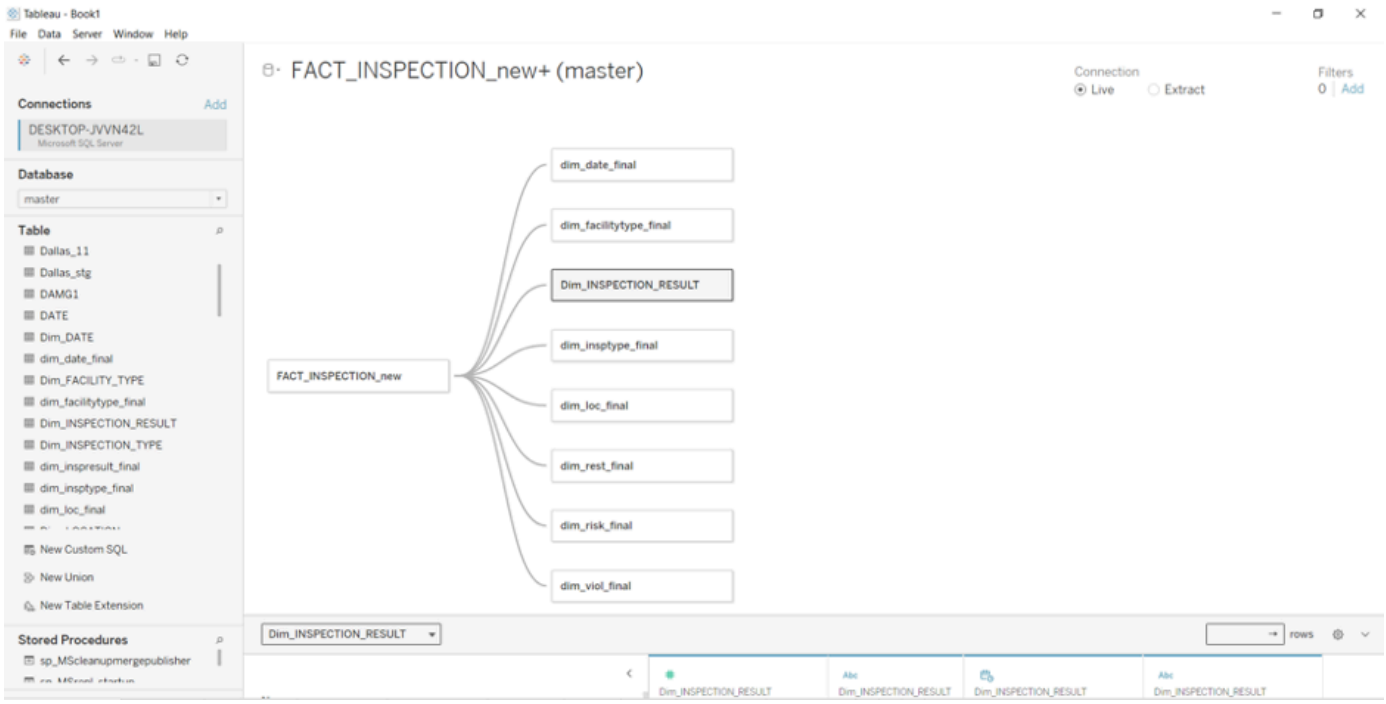
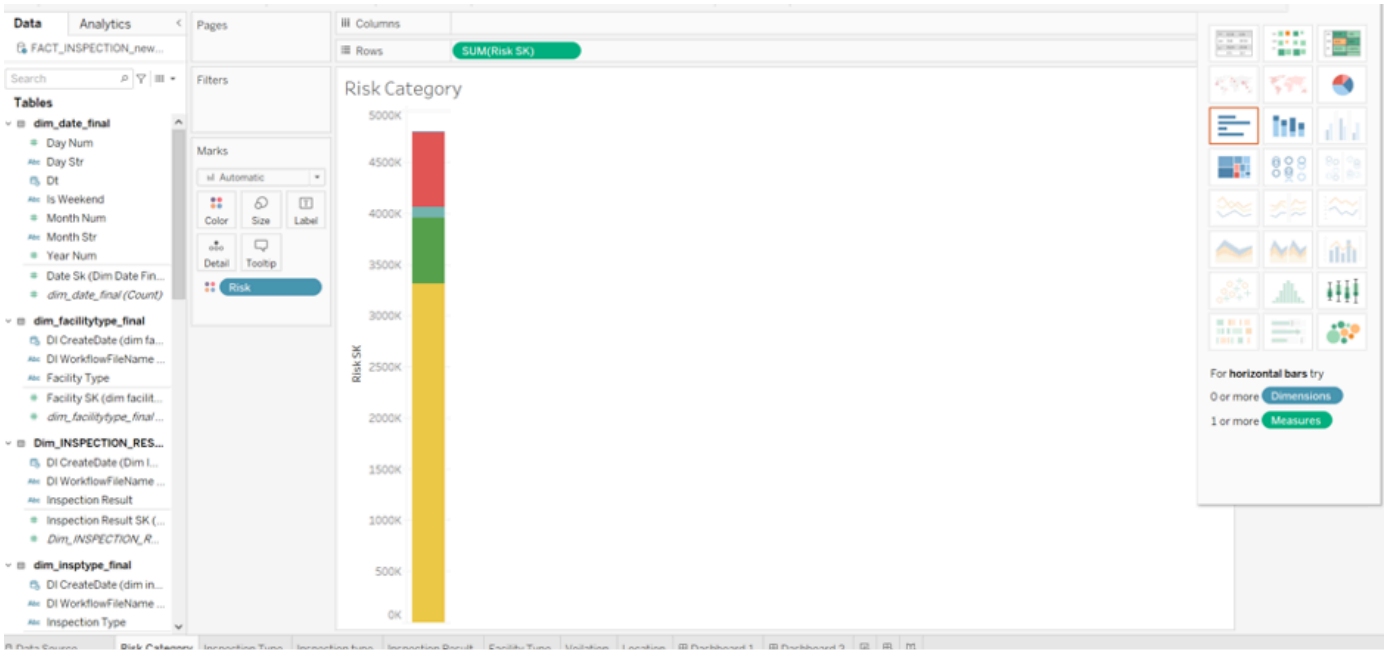
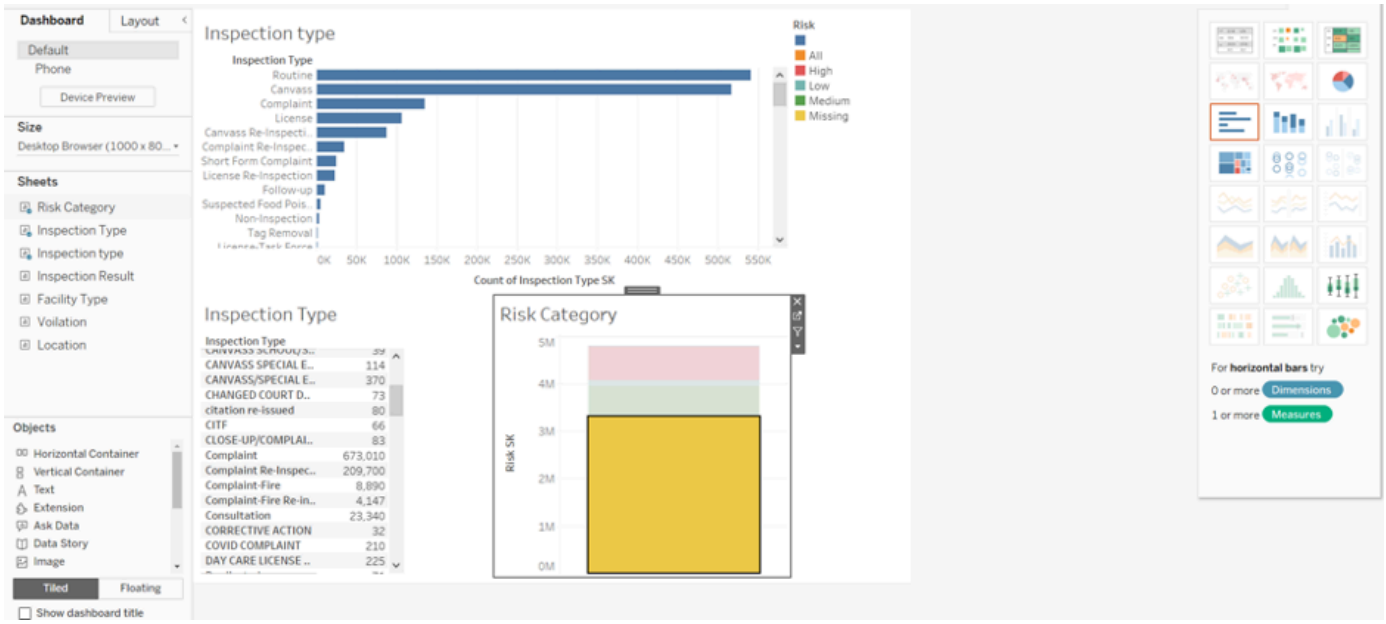

Sum of Latitude and Sum of Longitude by Zip_Code and Location_SK
Location_SK
Carpentersville
Palatine
Northbrook
Glencoe
New Trier Township
Arlington Heights
Hoffman Estates
Glenview
Wilmette
Evanston
Elgin
Hanover Township
Schaumburg
Hanover Park
Des Plaines
Maine Township
Park Ridge
Bartlett
St Charles
Geneva
Carol Stream
Elmhurst
Oak Park
Chicago
Wheaton
Milton
York
Proviso Township
Cicero
Berwyn
Downers Grove
Naperville
Aurora
Downers Grove Township
Oak Lawn
Worth Township
Bolingbrook
Yorkville
Orland Park
Bremen
Thornton Junction
Lansing
East Chicago
North Township
Gary
Calumet
Portage
Porter
Chesterton
Liberty
Jackson Township
Hobart
South Haven
Filters
Search
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Inspection_result
Ispection_type
Risk_Category
Facility_type
Violation
Business_inspected
Location

Tableau Dashboards:






Dashboard
Layout
Default
Phone
Device Preview
Size
Desktop Browser (1000 x 80...
Sheets
Risk Category
Inspection Type
Inspection type
Inspection Result
Facility Type
Voilation
Location
Objects
Horizontal Container
Vertical Container
Text
Extension
Ask Data
Data Story
Image
Tiled
Floating
Show dashboard title
Inspection type
Inspection Type
Routine
Canvass
Complaint
License
Canvass Re-Inspecti..
Complaint Re-Inspec..
Short Form Complaint
License Re-Inspection
Follow-up
Suspected Food Pois..
Non-Inspection
Tag Removal
0K 50K 100K 150K 200K 250K 300K 350K 400K 450K 500K 550K
Count of Inspection Type SK
Risk
All
High
Low
Medium
Missing
Inspection Type
Inspection Type
CANVASS SPECIAL E.. 114
CANVASS/SPECIAL E.. 370
CHANGED COURT D.. 73
citation re-issued 80
CITF 66
CLOSE-UP/COMPLAI.. 83
Complaint 673,010
Complaint Re-Inspec.. 209,700
Complaint-Fire 8,890
Complaint-Fire Re-in.. 4,147
Consultation 23,340
CORRECTIVE ACTION 32
COVID COMPLAINT 210
DAY CARE LICENSE .. 225
Risk Category
Risk SK
5M
4M
3M
2M
1M
0M
For horizontal bars try
0 or more Dimensions
1 or more Measures