Food Inspections Assignment
A Data Analyst Journey

- Dealing with Sonoma County Department of Health Services Food Facility inspection
  - California State Related data

- Understand the **content** & **schema** of the file. Highly recommended to review each and every field manually and understand the details and nature of data as how its stored.

Food Inspections – CA

- [Food Facility Inspections | Open Data | Sonoma County (ca.gov)]
  - This data set is about food inspection from Sonoma county, California.
  - The above link provides more details about Dataset
  - It also provides catalog/metadata about the data
    - Provided the same in form of screen shots.
  - Review data manually and digest its contents to create correct data model
    - More important is the Violation codes and descriptions AND Location field

About this Dataset

Mute Dataset

Updated
October 11, 2023

Data Last Updated	Metadata Last Updated
October 11, 2023	May 28, 2019

Date Created
September 26, 2014

Views	Downloads
12.6K	4,378

Data Provided by	Dataset Owner
Department of Health Services	OpenData

Contact Dataset Owner

Department

Department	Health Services

Focus Area

Focus	Safe, Healthy, and Caring Community

Topics

Category	Health
Tags	food, restaurant, inspection, health

Licensing and Attribution

License	Public Domain
Source Link	http://sonomacounty.ca.gov/Health-Services/

Columns in this Dataset

Column Name	Description	Type
BusinessId	The unique identifier Health Services uses to identify a busine...	Plain Text
Name	Food facility name	Plain Text
Address		Plain Text
City		Plain Text
State		Plain Text
ZipCode		Plain Text
PhoneNumber		Plain Text
InspectionId	The unique identifier Health Services uses to identify an inspe...	Plain Text
Date	The date of the inspection.	Date & Time
InspectionType		Plain Text
ViolationCodes	Codes used by Health Services to identify a violation type.	Plain Text
ViolationDescriptions	The description of the violation(s) found during an inspection.	Plain Text
Location	The address where the inspection was performed.	Location

Food Inspections – CA

- **Business Requirements**
  - Include a new column in appropriate table as **Violation category** and derive the value of this column based on below logic
    - If violation description has text called Minor (ignore case when searching) then assign the category as **MINOR**
    - If violation description has text called Major (ignore case when searching) then assign the category as **MAJOR**
    - Any description that doesn't fall on any of the above criteria then assign the category as **OTHER**
    - **Hint:** Violation code starting with **K** is not a violation hen they fall under **OTHER** category
  - Include a new column in appropriate table called **Violation Score** and derive the value of this column based on below logic
    - minor violation gets 5 points
    - major violation gets 10 points
    - other violation gets 0 points
    - If score exceeds 100 points then store as 100. (More the score means Bad results)
      - Example if an inspection had 5 violations of which 2 are MINOR and 3 are MAJOR then score is 5+5+10+10+10 = 40
  - Include a new column in appropriate table called **Inspection results** and derive the value of this column based on below logic
    - If violation score is between 0 and 60 is **PASS**
    - violation score > 60 is **FAIL**
  - Location filed
    - It contains both Address and Latitude and Longitude. However, While loading to Integration schema just load Latitude and Longitude
    - **Note**: You will use this field to plot the inspections on MAP

Food Inspections – CA

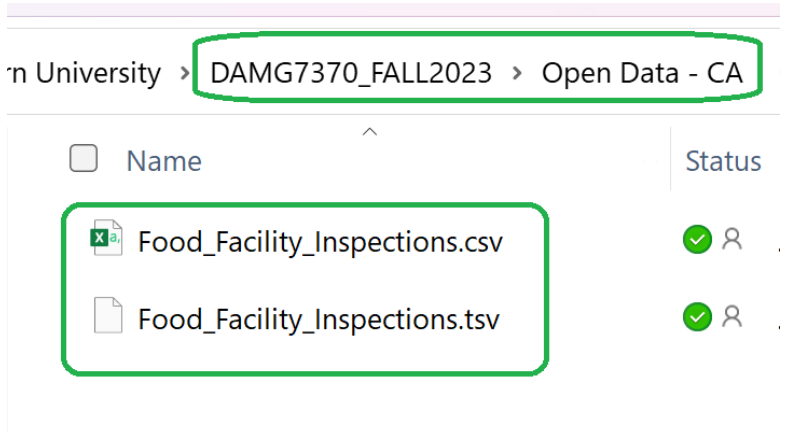
- **Business Requirements**
  - **Reports for dashboards on PBI and Tableau**
    - How many food inspections over Year/Quarter/Month/Weekend/Weekday/Day
    - Number of food inspections over time for below criteria
      - Pass vs fail
      - violation category (Major vs Minor Vs Other)
    - Food establishments inspected
      - Top ten most inspected restaurant(s) (Year wise , City wise)
      - Top ten with worst results (year wise)
      - Top 10 violation codes on inspections
    - Number of Restaurants City wise
    - Most number of violation area wise
    - Map food inspections and find which area had most number of inspections

Food Inspections – CA

- Deliverables (Detailed explanation on each part in subsequent slides)
  - Part 1
    - Get data
    - Perform data profiling
    - Load data into Stage table (stg_)
  - Part 2
    - Design Dimensional Model using ER studio or Navicat
    - Create DDL SQL script for MS SQL Server and MySQL DB
  - Part 3
    - Load dimensional model tables using Stage table
  - Part 4
    - Create BI dashboards

Food Inspections – CA

- **Deliverables: Part 1**
  - Get data from downloaded files via shared one drive.
    - It has 2 file formats (Tab separated and Comma separated)
    - You can use any of the format.
    - If the file from One Drive is having issues, then Download the file directly from below link
      - [Food Facility Inspections | Open Data | Sonoma County (ca.gov)]



  - Load data into Stage table(s) (stg_)
    - MySQL and SQL Server
    - Follow Staging Guidelines standards. All standard audit columns must be present
      - DI_CreatedDate – DateTime when row is loaded
      - DI_WorkflowFileName – Filename that you use to load the data
      - DI_Workflow_ProcessID – Workflow / Job ID
  - Perform data profiling
    - Purpose of profile to identify appropriate data types
    - Identify min and max values so data truncation can be avoided
    - Understand the data appropriately so that you can apply data cleansing methods
      - Violation codes and Descriptions are stored in Pipe delimited format with same column and this needs to be normalized to store in right format. Design your model appropriately.
  - Submit
    - Screenshot of Alteryx workflows
    - Record job start and end times and document the total time each workflow/job takes to complete the process
    - Relevant DDL script of both databases

Food Inspection – CA

- **Deliverables: Part 2**
  - Identify Dimensions & Facts
    - Based on the data analysis Clearly identify Facts and Dimensions
    - Define the Grain based on requirements
    - Explain how you will be handling the Inspection codes for Integration schema loading
    - Make sure to identify Surrogate keys, Relationships and associations
    - Include all standard schema and audit columns for every entity
    - Must contain DATE Dimension table and its SK should be of format YYYYMMDD
    - Appropriate Datatypes must be identified
      - Date to be stored as DATE (If it contains time then DateTime)
      - Text to be stored as VARCHAR
      - Numbers are NUMERIC

  - Create a Dimensional Data Model
    - ER/Studio for windows machines
    - Navicat for MAC users

  - Create DDL for databases
    - MySQL
    - SQL Server

  - Submit
    - Screenshots for each of the above
    - ER Studio /Navicat file, DDL scrips
    - ER model in PDF or JPG format

Food Inspection – CA

- **Deliverables: Part 3**
  - Create data preparation workflow(s) to load data into Integration Schema
    - i.e., dimensional model using Talend
  - Load data by running the Talend Jobs
  - Clearly document the purpose and business logic under every activity documentation section
  - Use Repository for all inputs and outputs so that schema datatypes can propagate correctly
  - Don’t run the Talend jobs without creating the Stage table
  - Don’t run the Talend jobs without creating the Facts and Dimension tables
  - Appropriate meaningful names should be used
    - Job names
    - Variable names
    - Context names
- Submit
  - Screenshots of each of the above
  - Talend Job export
  - Table row counts
  - Record Execution times
  - Export talend documentation export and submit the documentation zip file

Food Inspection – CA

- Deliverables: Part 4
  - Create BI dashboards to answer Business questions (as provided in previous slides)
    - Power BI Desktop published in PBI Service using SQL Server database
    - Tableau Desktop published in Tableau Online using MySQL database

- Suggestions
  - Understand all the deliverables and ask questions with in 2days
  - Complete the ER modeling and get is reviewed ASAP to avoid last minute issues
  - Make sure you are able to connect to the data file
  - Make sure you are able to connect to databases