DAV6100: A project on analyzing spill incidents in New York State

Group 7:

Lakshmikar Reddy Polamreddy

Banty Phagotra

Sandeep

Yashwanth

Yeshiva University

Agenda

- Overview
- Project Flow
- Data Profile
- AWS architecture
- ETL process
- Data visualization and deliverables
- Project Milestones & Timeline
- Team Responsibilities
- Challenges

Overview

Motivation:

- Accidental releases of petroleum and/or other hazardous materials occur throughout New York State. Even small releases have the potential to endanger public health and contaminate groundwater, surface water, and soils.

Project Objectives:

- To identify the major sources and contributing factors for these spill occurrences
- To analyze if there is any relationship between the number of spill incidents and the income of the people in those counties.
- To analyze if there is any relationship between the number of spill incidents and the number of manufacturing establishments in those counties.

Data Warehouse:

- Developed OLAP star schema data model for data analysis
- Utilized AWS cloud services

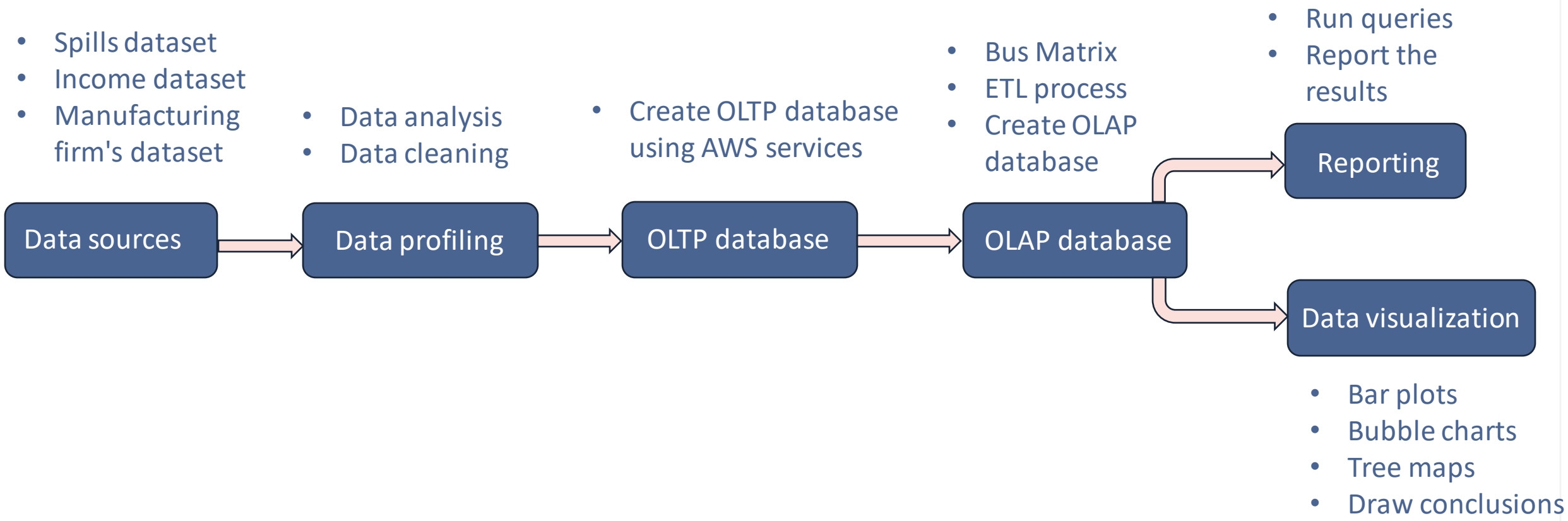
Data sources:

- DATA.NY.GOV
- DATA.CENSUS.GOV

EDA and Data visualization:

- Python libraries – matplotlib and seaborn
- Tableau

Project Flow

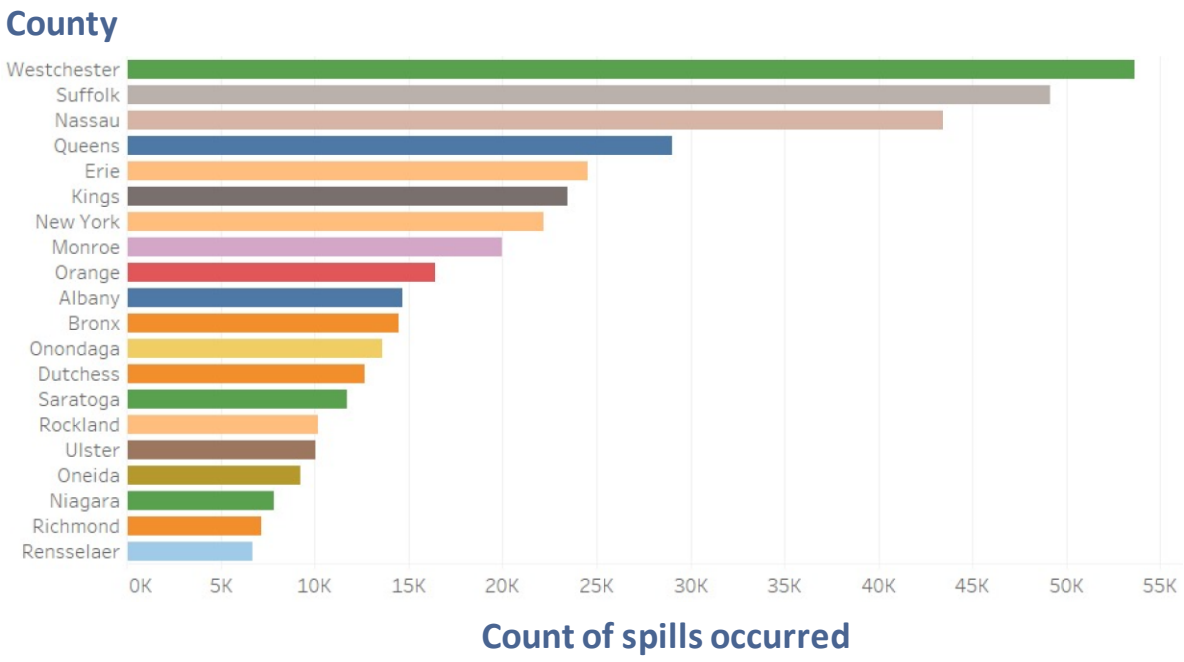


Data Profile: Spill incidents of NY

Dataset Summary

Source of Information	https://data.ny.gov/Energy-Environment/Spill-Incidents/u44d-k5fk
Number of Records	534K
Frequency of updates	Daily
Data type and structure	CSV
Number of columns	20
Size	90.6MB
Granularity	Incident level

Below bar plot shows the top 20 leading counties in NY state in terms of the count of spill incidents occurred



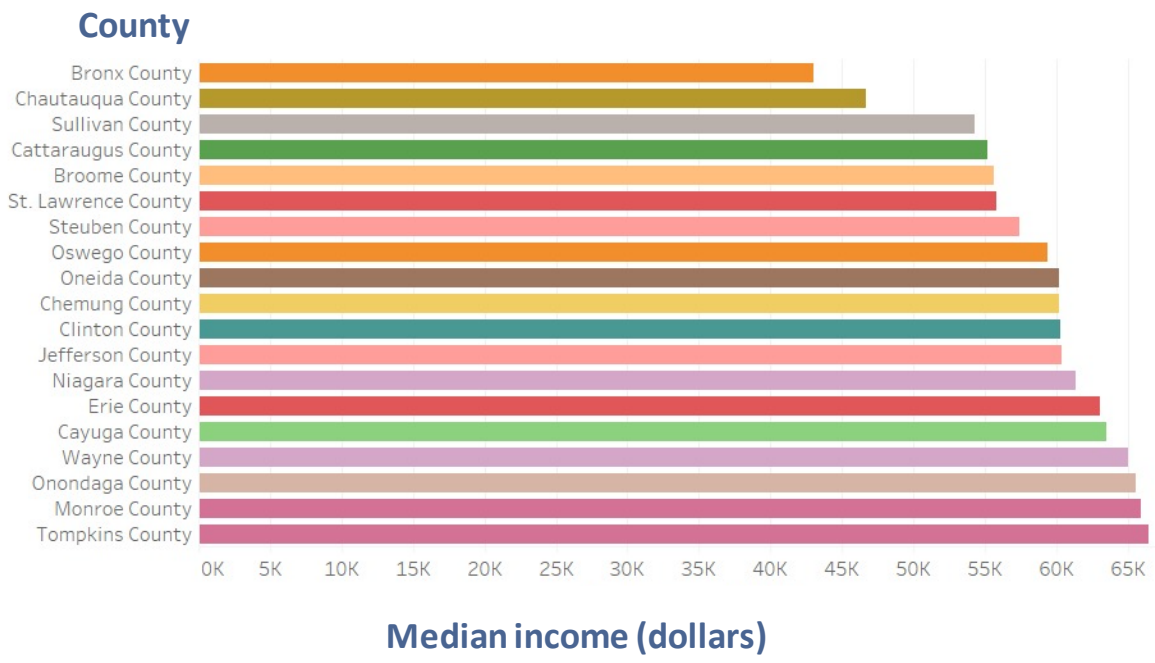
Descriptive Statistics

Data Profile: County-wise Income details in NY

Dataset Summary

Source of Information	https://data.census.gov/table
Number of Records	17
Frequency of updates	5 years
Data type and structure	CSV
Number of columns	153
Granularity	County

Below bar plot shows bottom 20 counties in NY state in terms of the medium income of households



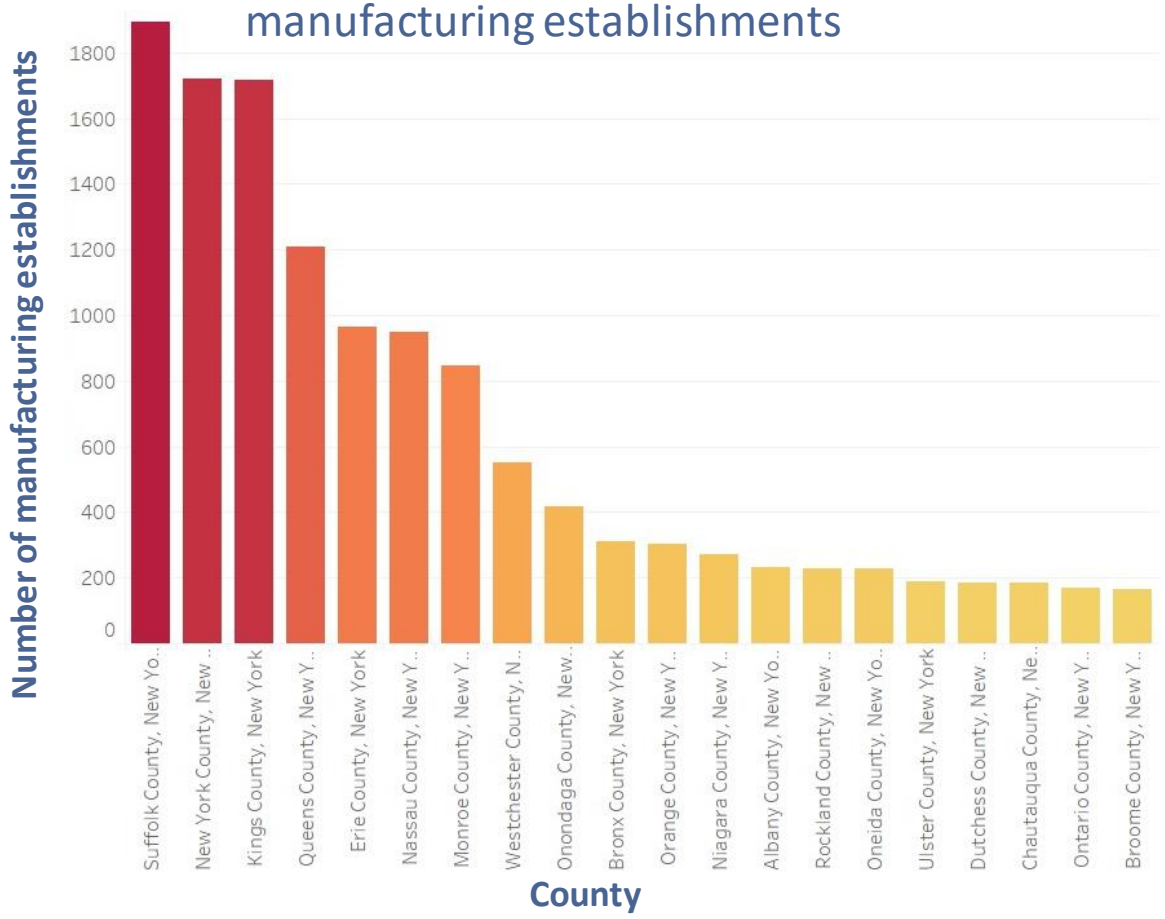
Descriptive Statistics

Data Profile: County-wise Manufacturing firms in NY

Dataset Summary

Source of Information	https://data.census.gov/table
Number of Records	500
Frequency of updates	5 years
Data type and structure	CSV
Number of columns	6
Granularity	County

Below bar plot shows top 20 counties in NY state in terms of the total number of manufacturing establishments



Descriptive Statistics

Additional information: Column details of 3 datasets

#	Column	Non-Null Count	Dtype
0	Spill Number	534454 non-null	int64
1	Program Facility Name	534448 non-null	object
2	Street 1	534324 non-null	object
3	Street 2	41603 non-null	object
4	Locality	533336 non-null	object
5	County	534454 non-null	object
6	ZIP Code	50002 non-null	object
7	SWIS Code	534454 non-null	int64
8	DEC Region	534454 non-null	int64
9	Spill Date	534302 non-null	object
10	Received Date	533977 non-null	object
11	Contributing Factor	534454 non-null	object
12	Waterbody	45847 non-null	object
13	Source	534454 non-null	object
14	Close Date	523978 non-null	object
15	Material Name	534454 non-null	object
16	Material Family	534454 non-null	object
17	Quantity	534454 non-null	float64
18	Units	433186 non-null	object
19	Recovered	534454 non-null	float64

Spill incidents in NY state

#	Column	Non-Null Count	Dtype
0	Total	152 non-null	object
1	Less than $10,000	152 non-null	object
2	$10,000 to $14,999	152 non-null	object
3	$15,000 to $24,999	152 non-null	object
4	$25,000 to $34,999	152 non-null	object
5	$35,000 to $49,999	152 non-null	object
6	$50,000 to $74,999	152 non-null	object
7	$75,000 to $99,999	152 non-null	object
8	$100,000 to $149,999	152 non-null	object
9	$150,000 to $199,999	152 non-null	object
10	$200,000 or more	152 non-null	object
11	Median income (dollars)	152 non-null	object
12	Mean income (dollars)	152 non-null	object
13	Household income in the past 12 months	152 non-null	object
14	Family income in the past 12 months	152 non-null	object
15	Nonfamily income in the past 12 months	152 non-null	object

dtypes: object(16)
memory usage: 20.2+ KB

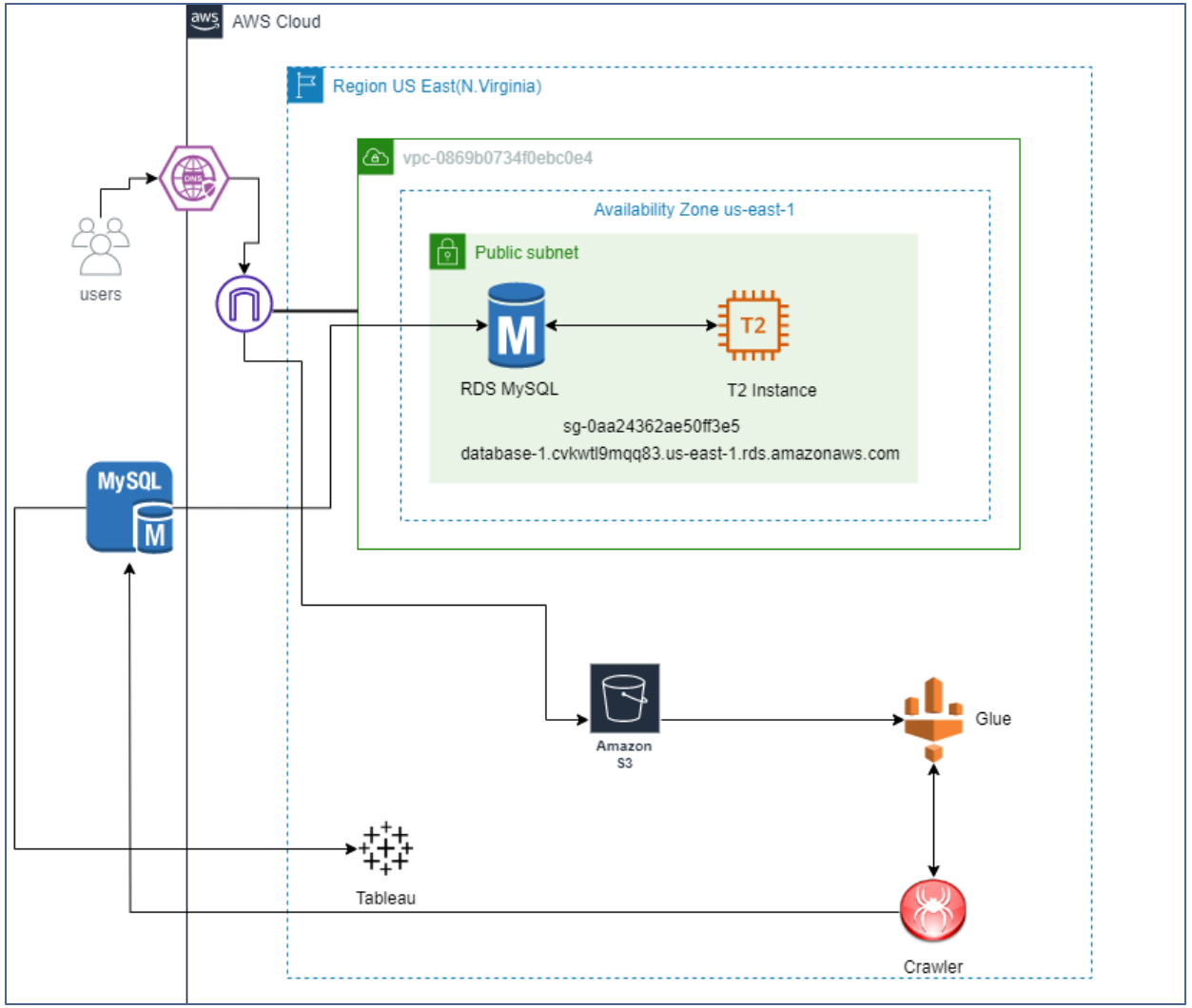
County-wise income details of NY state

#	Column	Non-Null Count	Dtype
0	Geographic Area Name (NAME)	500 non-null	object
1	2017 NAICS code (NAICS2017)	500 non-null	object
2	Meaning of NAICS code (NAICS2017_LABEL)	500 non-null	object
3	Meaning of Employment size of establishments code (EMPSZFE_LABEL)	500 non-null	object
4	Year (YEAR)	500 non-null	int64
5	Number of establishments (ESTAB)	500 non-null	object

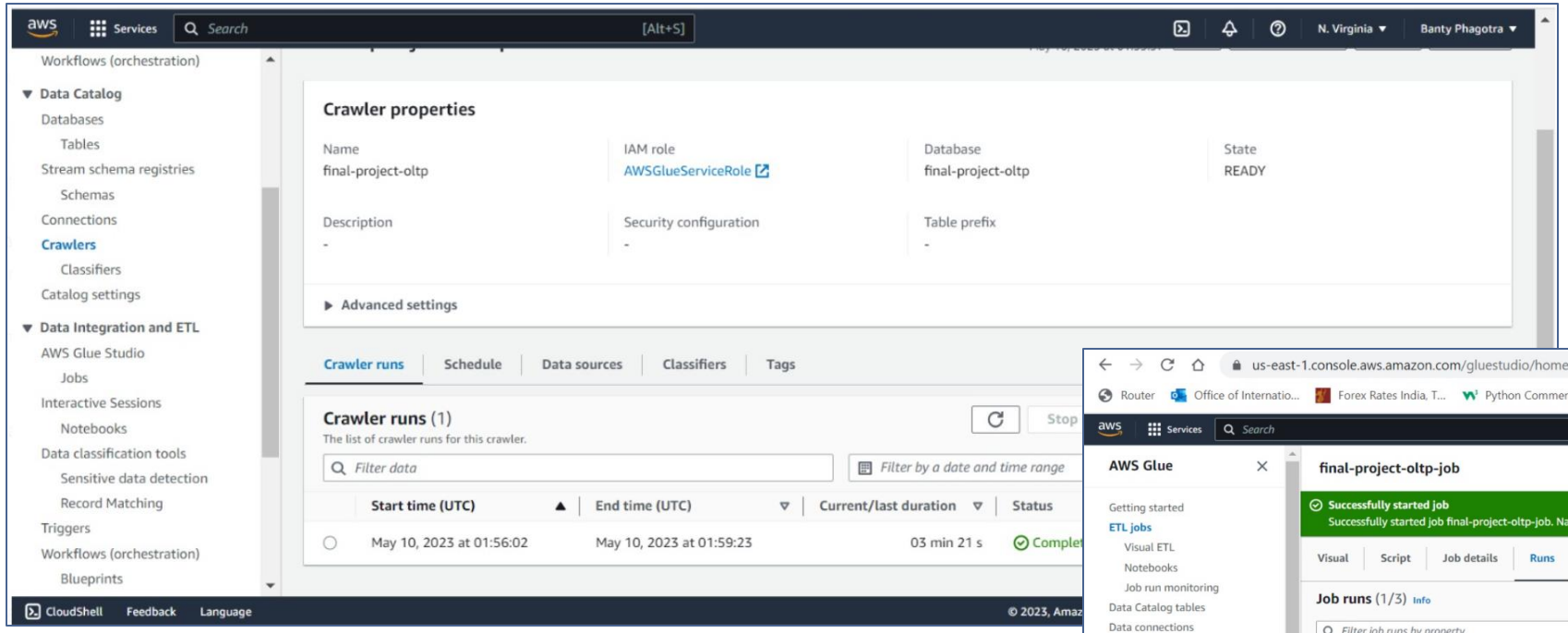
dtypes: int64(1), object(5)
memory usage: 23.6+ KB

Manufacturing establishments of NY state

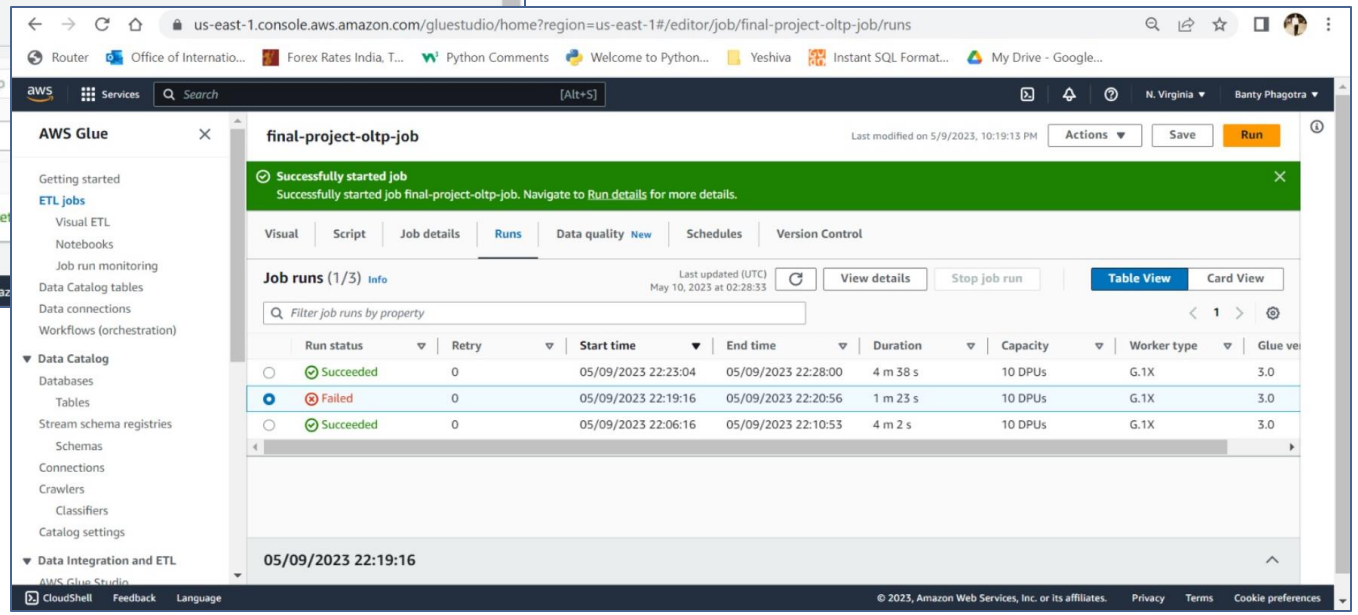
AWS Architecture



OLTP Database ETL process through AWS



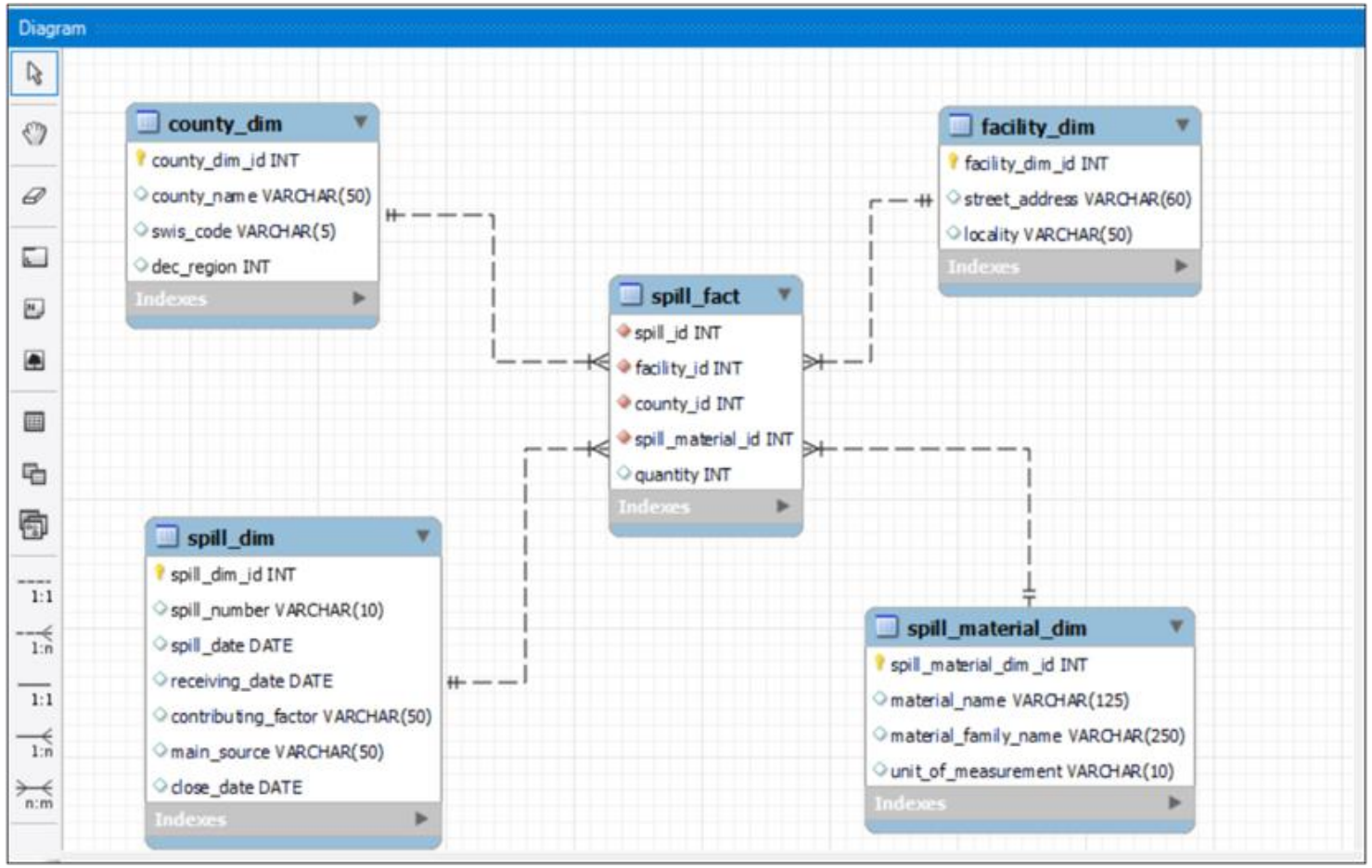
AWS Crawler



AWS Glue job

ETL process and ER Diagram

- First created OLTP database using AWS services
- Then created OLAP database
- Star schema using surrogate keys
- 4 Dimension tables
- 1 Fact table
- Created stored procedures and triggers



ETL process for OLAP

MySQL Workbench

database2

File Edit View Query Database Server Tools Scripting Help

oltp* olap* spill_incidents county_dim facility_dim spill_fact spill_dim

Don't Limit

```
-- Inserting Data into facility_dim table from OLTP Spill_incident table using Insert Statment
insert into final_project_olap.facility_dim
(street_address, locality)
select street_address, locality
from final_project_oltp.spill_incidents ;

-- Inserting Data into spill_dim table from OLTP Spill_incident table using Insert Statment
insert into final_project_olap.spill_dim
(spill_number, spill_date, receiving_date, contributing_factor, main_source, close_date)
select spill_number, spill_date, receiving_date, contributing_factor, main_source, close_date
from final_project_oltp.spill_incidents ;

-- Inserting Data into spill_material_dim table from OLTP Spill_incident table using Insert Statment
insert into final_project_olap.spill_material_dim
(material_name, material_family_name, unit_of_measurement)
select material_name, material_family, unit_of_measurement
from final_project_oltp.spill_incidents ;

-- Inserting Data into spill_fact table from OLTP Spill_incident table using Insert Statment

delete from final_project_olap.spill_fact ;
```

Query Completed

Insert statements sample

Stored procedure and Triggers

```
DELIMITER //

CREATE PROCEDURE final_project_olap.UpdateSpillIncident()
BEGIN
    -- Update county_dim
    UPDATE final_project_olap.county_dim c
    JOIN final_project_oltp.spill_incidents i ON i.county_name = c.county_name
    SET c.swis_code = i.swis_code,
        c.dec_region = i.dec_region
    WHERE i.county_name IS NOT NULL;

    -- Update facility_dim
    UPDATE final_project_olap.facility_dim f
    JOIN final_project_oltp.spill_incidents i ON i.street_address = f.street_address
    SET f.locality = i.locality
    WHERE i.street_address IS NOT NULL;

    -- Update spill_dim
    UPDATE final_project_olap.spill_dim d
    JOIN final_project_oltp.spill_incidents i ON i.spill_number = d.spill_number
    SET d.spill_date = i.spill_date,
        d.receiving_date = i.receiving_date,
```

Stored procedure sample

MySQL Workbench

database2

File Edit View Query Database Server Tools Scripting Help

oltp* olap* spill_incidents county_dim facility_dim spill_fact spill_dim

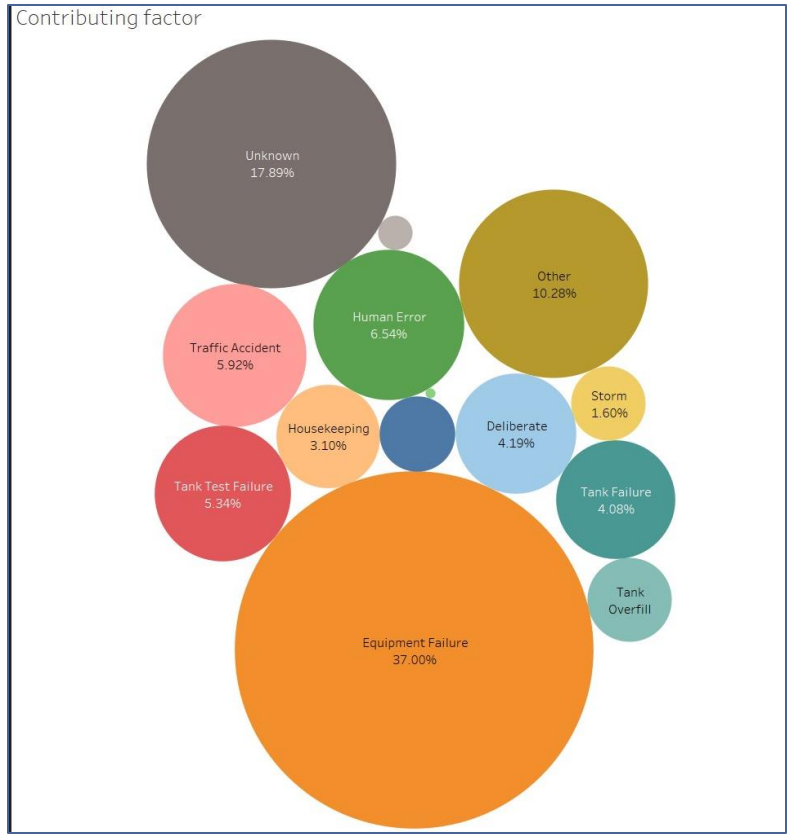
Don't Limit

```
-- After insert Trigger to update rows in all the dimension and fact table in OLAP Schema
DELIMITER $$
CREATE TRIGGER `update_olap_schema` AFTER
 INSERT ON `final_project_oltp`.`spill_incidents` FOR EACH ROW
BEGIN
    INSERT INTO `final_project_olap`.`spill_dim`
    (`spill_number`, `spill_date`, `receiving_date`, `contributing_factor`, `main_source`, `close_date`)
    VALUES (NEW.`spill_number`, NEW.`spill_date`, NEW.`receiving_date`, NEW.`contributing_factor`, NEW.`main_source`, NEW.`close_date`);

    INSERT INTO `final_project_olap`.`facility_dim` (`street_address`, `locality`)
    VALUES (NEW.`street_address`, NEW.`locality`);

    INSERT INTO `final_project_olap`.`county_dim` (`county_name`, `swis_code`, `dec_region`)
    VALUES (NEW.`county_name`, NEW.`swis_code`, NEW.`dec_region`);

    INSERT INTO `final_project_olap`.`spill_material_dim` (`material_name`, `material_family_name`, `unit_of_measurement`)
    VALUES (NEW.`material_name`, NEW.`material_family`, NEW.`unit_of_measurement`);

    SET @spill_dim_id = LAST_INSERT_ID();
    SET @facility_dim_id = LAST_INSERT_ID();
    SET @county_dim_id = LAST_INSERT_ID();
    SET @spill_material_dim_id = LAST_INSERT_ID();
```

Trigger sample

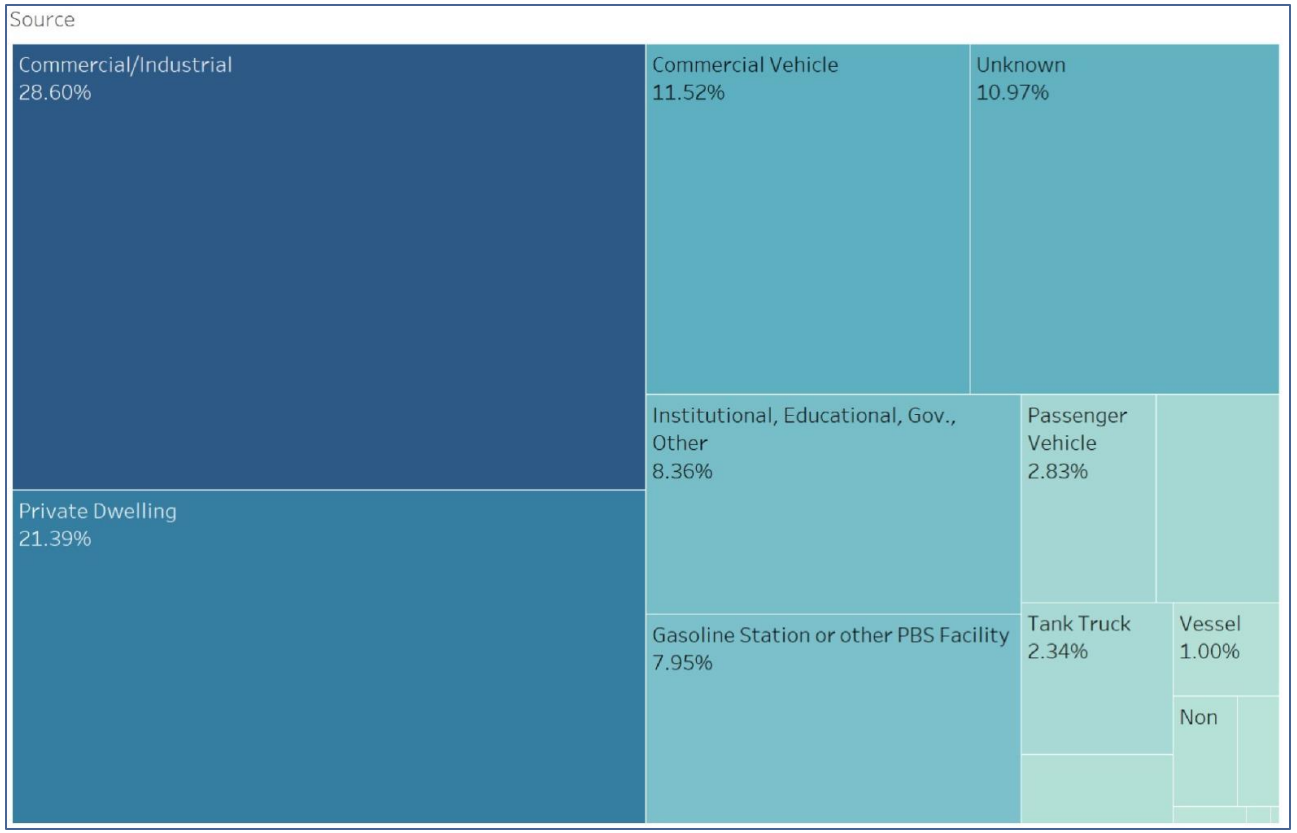
Data visualization and deliverables

Cause of spill



- Equipment failure is the major cause of spill and accounts for 37% of the total spill occurrences
- The cause of spill is unknown in 18%(appx) of the incidents

Source of spill

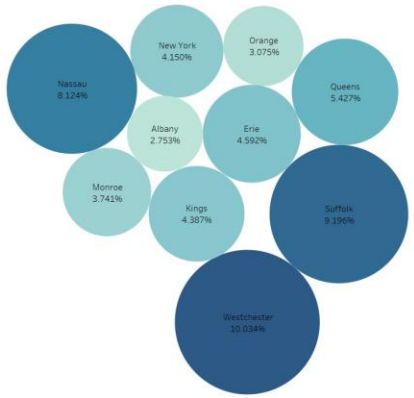


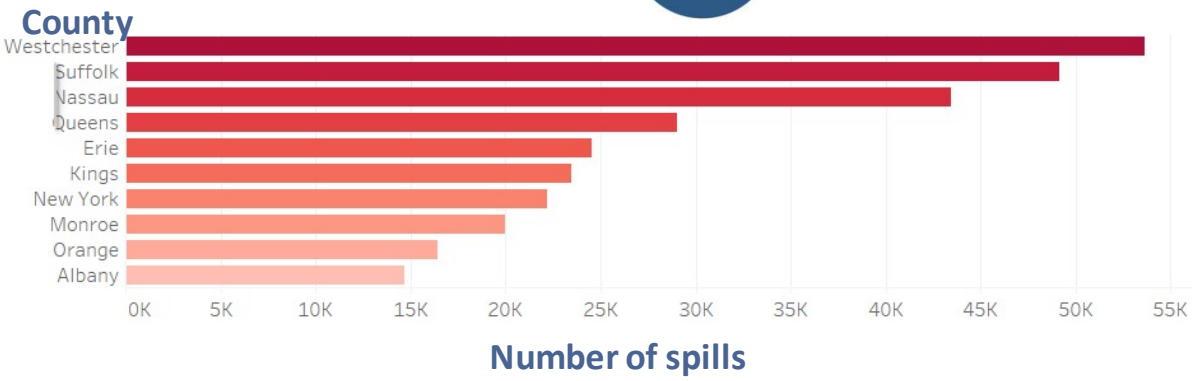
- Commercial/Industrial facilities and private dwellings are the major sources and responsible for 50% of the spill occurrences
- The source of spill is unknown in 11%(appx) of the incidents

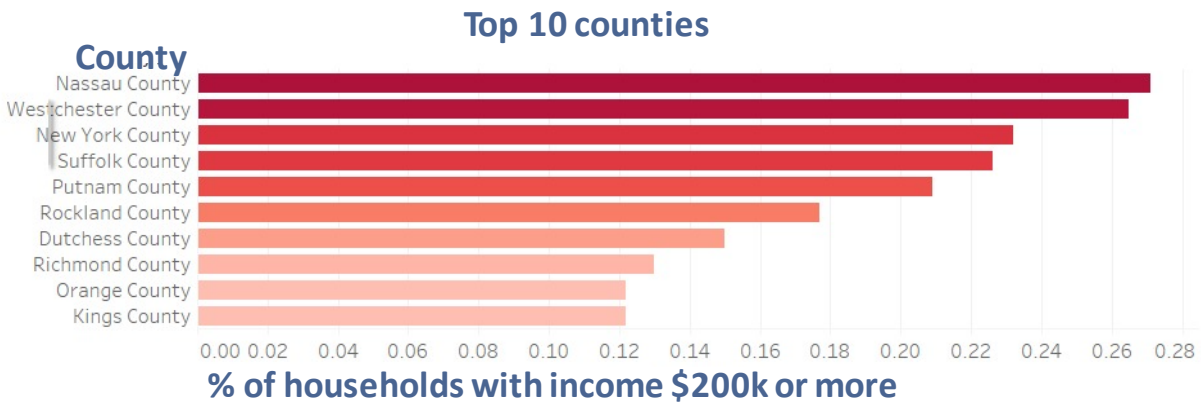
Data visualization and deliverables

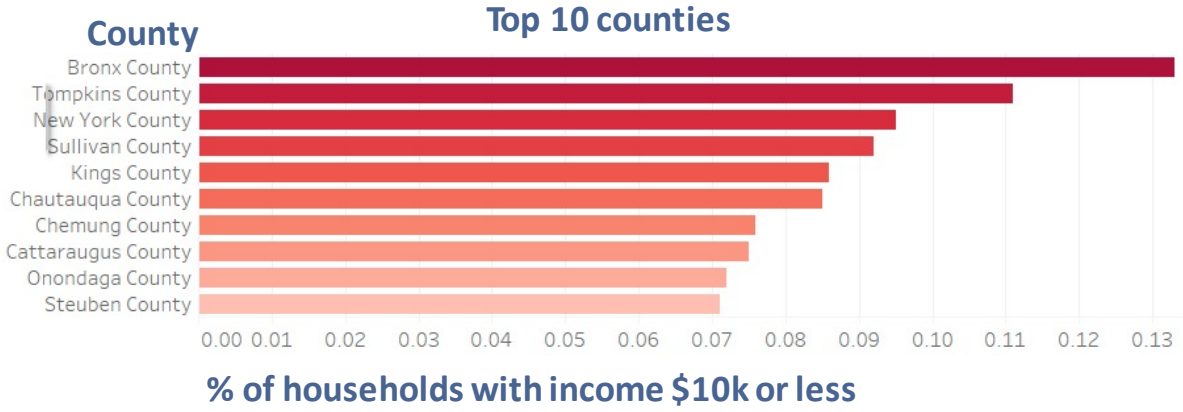
County-wise analysis of the relationship between the number of spill occurrences and the income level of people

10 counties account for more than 55% of the total spills





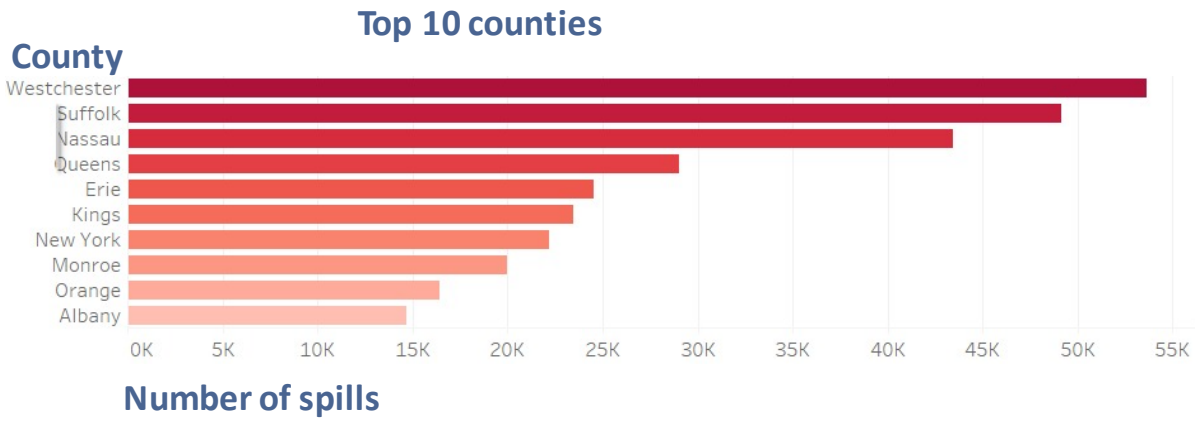


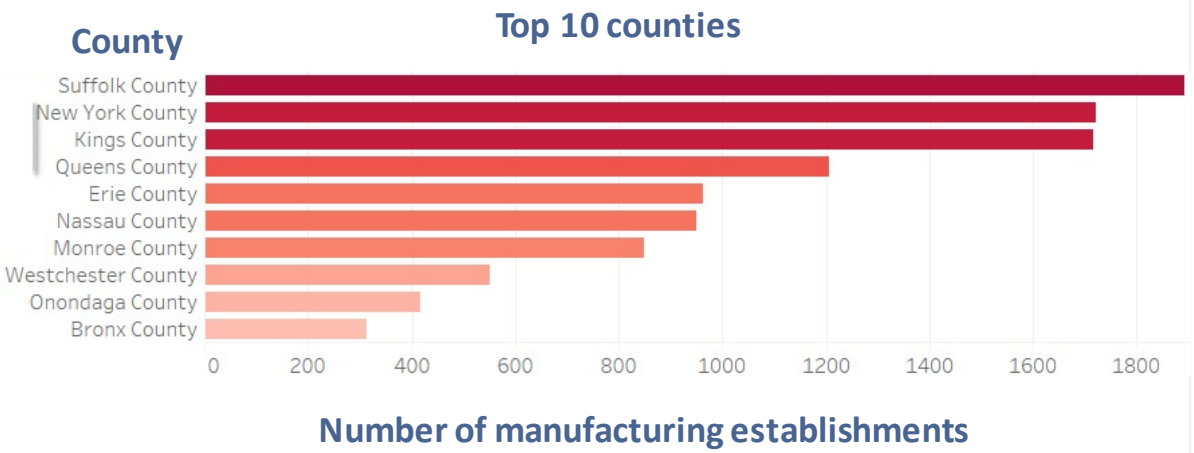


- More spills occurred in counties with more wealthy people (6 out of 10 counties match here)
- Less spills occurred in counties with less wealthy people
- There is no strong relationship between number of spills and the income level of people.
- Our assumption is false

Data visualization and deliverables

County-wise analysis of the relationship between the number of spill occurrences and the number of manufacturing firms





- More spills occurred in counties with a greater number of manufacturing establishments
- Our assumption is true

Project Milestones & Timeline

03-05-2023

Planning
- Present initial project proposal
- Modify the proposal
- Collect required datasets
- Assign roles and responsibilities with due dates

04-03-2023

Analysis
- Data profiling in Python

04-23-2023

Data model
- Develop conceptual data model
- Develop physical data model

05-03-2023

EDA & Data visualization
- EDA in python
- Data visualization in Tableau

05-12-2023

Deliver & Present
- Present final solution to class
- Determine peer score

Team Responsibilities

GROUP 7

Lakshmikar Reddy Polamreddy

- Project management and planning
- To gather datasets and prepare project proposal
- Data profiling, EDA and Data visualization in Python and Tableau
- To prepare final reports and slide deck

Banty Phagotra

- To support in preparing project proposal
- To develop conceptual and physical data model
- To create OLTP and OLAP databases in AWS
- ETL process

Sandeep

- Data visualization in Tableau

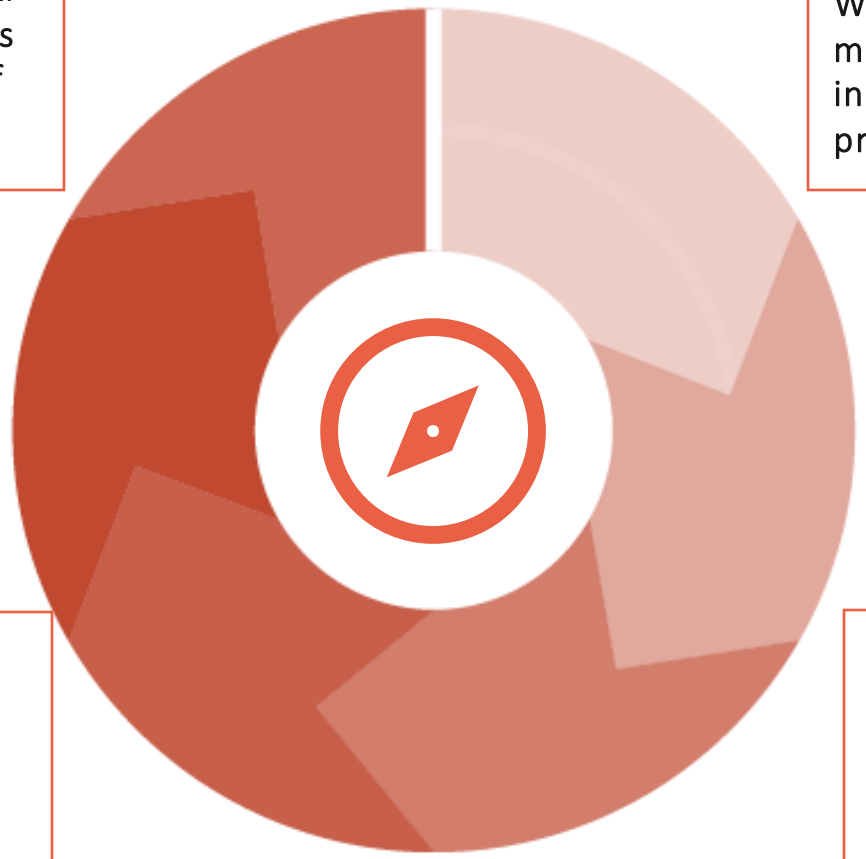
Assumptions

1

We assume that more spill incidents occur in counties where the income level of people is less

3

We assume that the number of manufacturing establishments increase in the same proportion as that of 2017



2

We assume that more spill incidents occur in counties with large number of industrial establishments

4

We assume that there is negligible change in the medium income of the households from 2021

Challenges

Granularity of spills dataset(incident-wise) and income and manufacturing datasets (county-wise) are different, we could not integrate the latter dataset with the former in star schema. So, we handle them separately for analysis and visualization

Created OLTP model in AWS but found challenging to create OLAP model using AWS Glue

Faced challenges to find the relationship between spill incidents and income details under various categories

Thank you!