SCHOOL OF COMPUTER TECHNOLOGY

APPLIED A.I. SOLUTIONS DEVELOPMENT

BIG DATA TOOLS AND TECHNOLOGIES

FINAL REPORT

PROFESSOR - NIZAR ALI

Submitted By:

Krutagna Joshi - 101513552

Vidit Suri - 101513592

Manik Kapil - 101447365

Riya Gupta - 101504556

Vithushan Vijayananthan - 101494734

Date: July 11, 2024

Our project, "Interpreting US Hospitalizations using Several Causal Factors," aims to identify and quantify key factors influencing hospitalization and fatality rates. By using postal codes as a common key, we seek to establish correlations with various factors, including fast food availability, housing conditions, and other socio-economic variables.

To achieve this, we merge all the datasets using PySpark to handle the large volume of data efficiently. We then create visualizations in Power BI to provide clear and actionable insights. The datasets are uploaded to various databases, including Hive, MySQL, SQL Server, and MariaDB, using a virtual machine for seamless data integration and accessibility. This approach ensures robust data handling and comprehensive analysis across multiple platforms.

Below are the datasets utilized in this analysis:

Database of Economic Incentives

The dataset contains information about various economic incentives provided to different recipients across New York State. Each entry includes details such as the recipient's name, project name, description, address, county, region, industry, and start date. The dataset also lists the type and amount of assistance awarded, the total public-private investment, disbursements to-date, project status, and compliance status. Additionally, it includes job creation and retention commitments and records the jobs created and retained to-date

Datafiniti Fast Food Restaurants

This dataset includes information on various fast food restaurants. It captures essential attributes such as the restaurant's name, address, city, state, postal code, latitude, longitude, and phone number. The dataset also provides details on the restaurant's type, cuisine, and menu, as well as ratings and reviews from customers. This comprehensive data is useful for analyzing the distribution, popularity, and customer satisfaction of fast food chains across different regions.

Home Fire Smoke Alarm Installs

This dataset documents smoke alarm installation activities conducted between September 2014 and March 2015. It includes the installation date, location (address, city, state, zip code), and the number of smoke alarms installed. The dataset also records the organization responsible for the installation and any additional comments or notes related to each installation event. This information is crucial for assessing the effectiveness and coverage of smoke alarm installation programs aimed at enhancing home safety.

Homes.com Properties Dataset

The dataset provides detailed information on properties listed on Homes.com. It includes attributes such as property ID, address, city, state, zip code, price, number of bedrooms and bathrooms, square footage, lot size, and year built. Additional details include property type, status (for sale, sold, etc.), and descriptions. This dataset is valuable for real estate market analysis, helping to identify trends in property values, availability, and other market dynamics.

Medical Services

This dataset comprises records of various medical services provided across different locations. Each entry contains information about the service provider, service type, address, city, state, zip code, and contact details. The dataset may also include details on the range of services offered, operational hours, and any specializations. This comprehensive data helps in understanding the distribution and accessibility of medical services, aiding in healthcare planning and resource allocation.

SITE HCC FCT DET

This dataset includes information relevant to healthcare facilities, focusing on factors such as facility type, location, and contact details. It may also capture specifics about the services offered, operational hours, and any accreditation or certification statuses. This dataset is crucial for analyzing the infrastructure of healthcare facilities, identifying gaps in service provision, and planning improvements to healthcare access and quality.

Crime Data

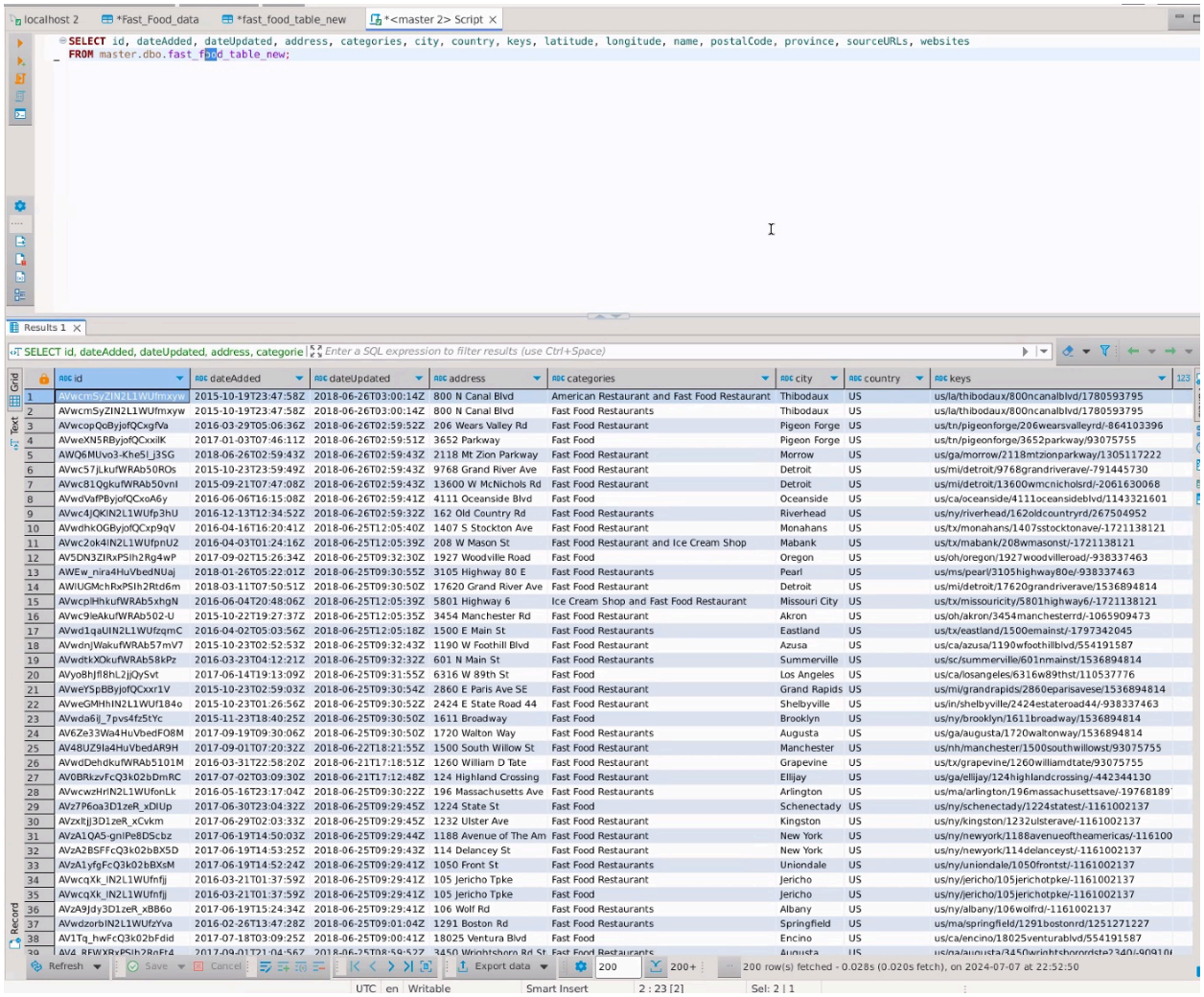
The dataset likely contains historical information related to various categories, potentially including demographics, economic indicators, or other time-sensitive data points. This dataset is essential for longitudinal studies, enabling researchers to track changes over time, compare past and present conditions, and predict future trends based on historical patterns.

The datasets were uploaded into different databases with the following configurations:

- **Hadoop/Hive**:
    - **Database Name**: health_centres
    - **Table Name**: my_table
    - **Details**: This table contains comprehensive data related to health centers, potentially including attributes like center names, addresses, services offered, and patient demographics. This dataset is crucial for analyzing healthcare accessibility and service distribution.

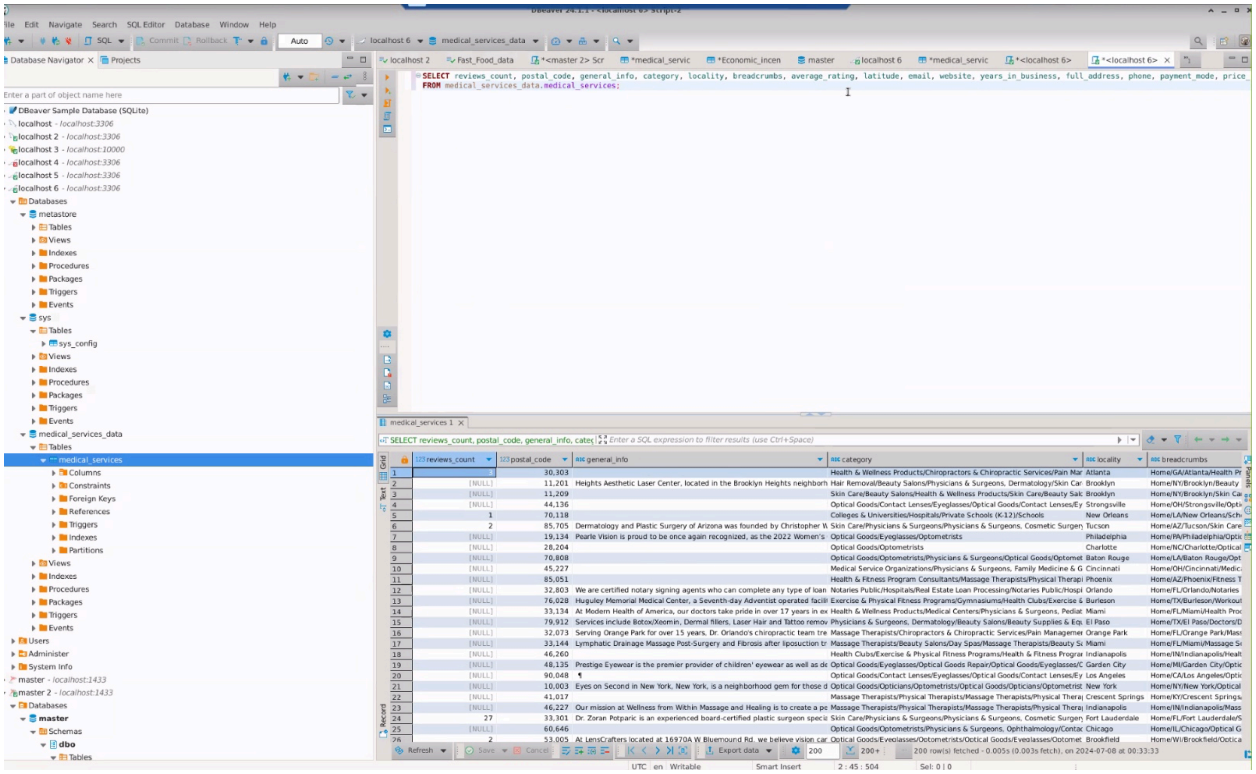
SQL Server:

- **Database**: Datafiniti_Fast_Food_Restaurants
- **Details**: This dataset includes detailed information on various fast food restaurants. It captures essential attributes such as the restaurant's name, address, city, state, postal code, latitude, longitude, and phone number. Additionally, it provides details on the restaurant's type, cuisine, menu, ratings, and customer reviews. This comprehensive data is useful for analyzing the distribution, popularity, and customer satisfaction of fast food chains across different regions.



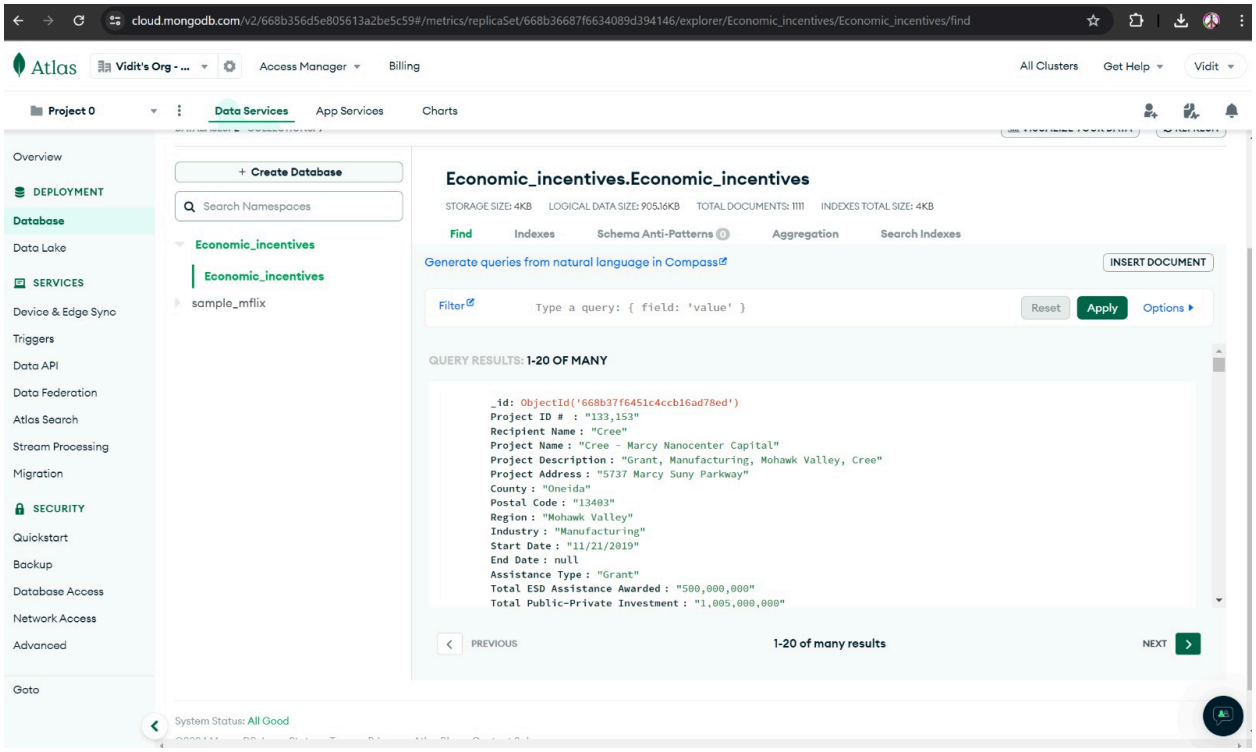
MariaDB:

- **Database**: Medical Services
- **Details**: This dataset comprises records of various medical services provided across different locations. Each entry contains information about the service provider, service type, address, city, state, zip code, and contact details. The dataset may also include details on the range of services offered, operational hours, and any specializations. This comprehensive data helps in understanding the distribution and accessibility of medical services, aiding in healthcare planning and resource allocation.



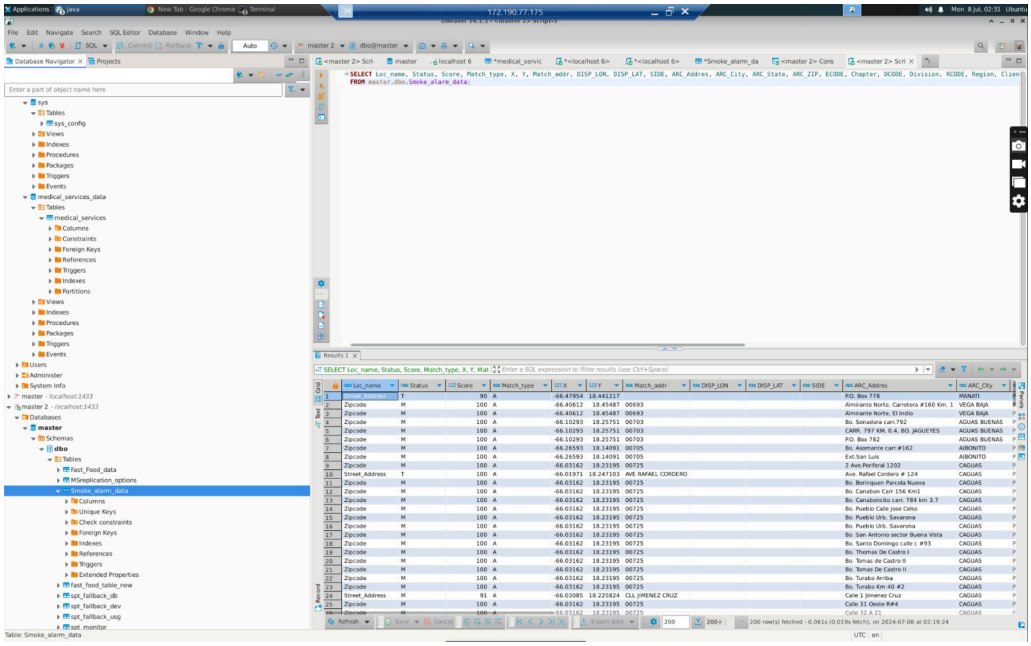
MongoDB:

- **Database**: Economic Incentives
- **Format**: JSON
- **Details**: The dataset contains information about various economic incentives provided to different recipients across New York State. Each entry includes details such as the recipient's name, project name, description, address, county, region, industry, and start date. The dataset also lists the type and amount of assistance awarded, the total public-private investment, disbursements to-date, project status, and compliance status. Additionally, it includes job creation and retention commitments and records the jobs created and retained to-date. This information is essential for analyzing the impact of economic incentives on local economies and job markets.



SQL Server:

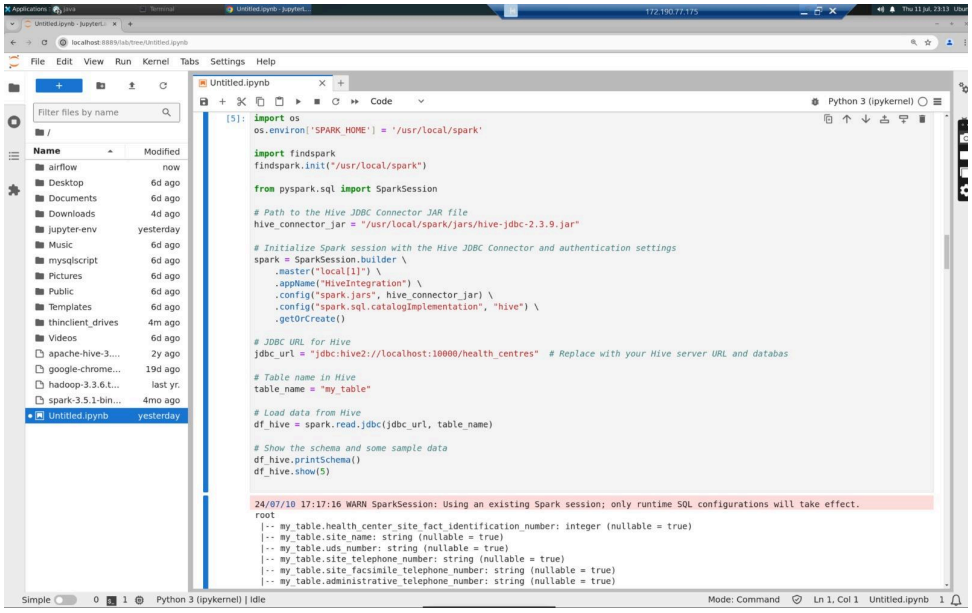
- **Database**: smoke alarm
- **Details**: This dataset documents smoke alarm installation activities conducted between September 2014 and March 2015. It includes the installation date, location (address, city, state, zip code), and the number of smoke alarms installed. The dataset also records the organization responsible for the installation and any additional comments or notes related to each installation event. This information is crucial for assessing the effectiveness and coverage of smoke alarm installation programs aimed at enhancing home safety.

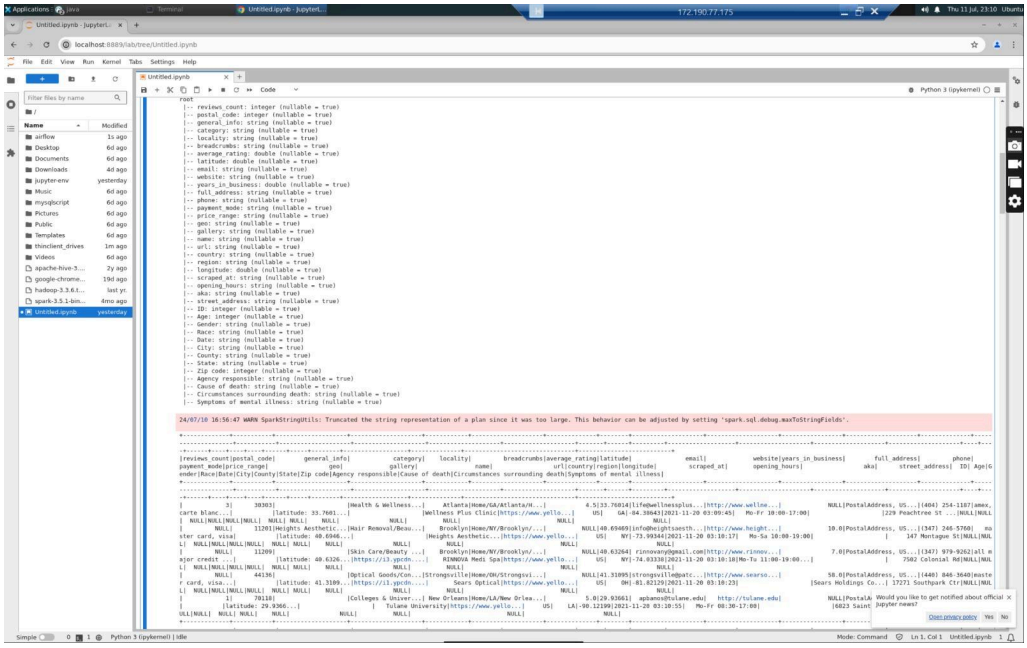


Below we have connected to our databases using jupyter notebook. We set up a spark session to connect to each of the databases.

Connecting to Hive Database:

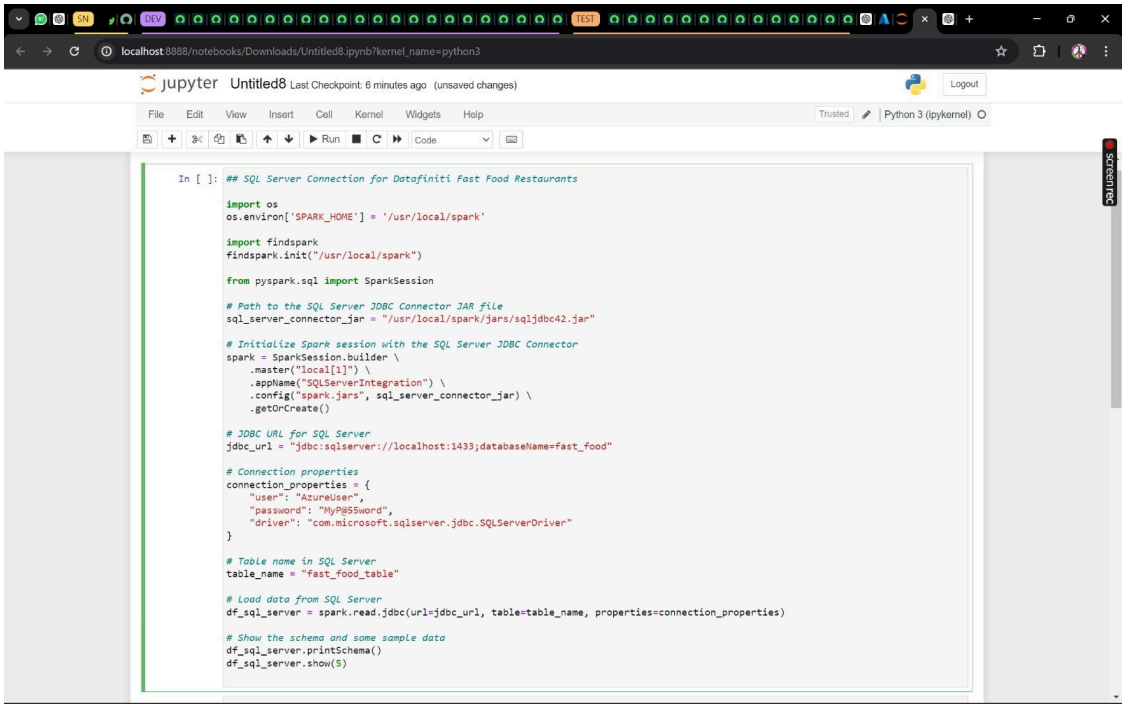
- Set the Spark environment with findspark and initialize a SparkSession with Hive integration.
- Specify the Hive JDBC connector JAR file and the JDBC URL for the Hive database (health_centres).
- Load data from the Hive table (my_table), print its schema, and display sample data.





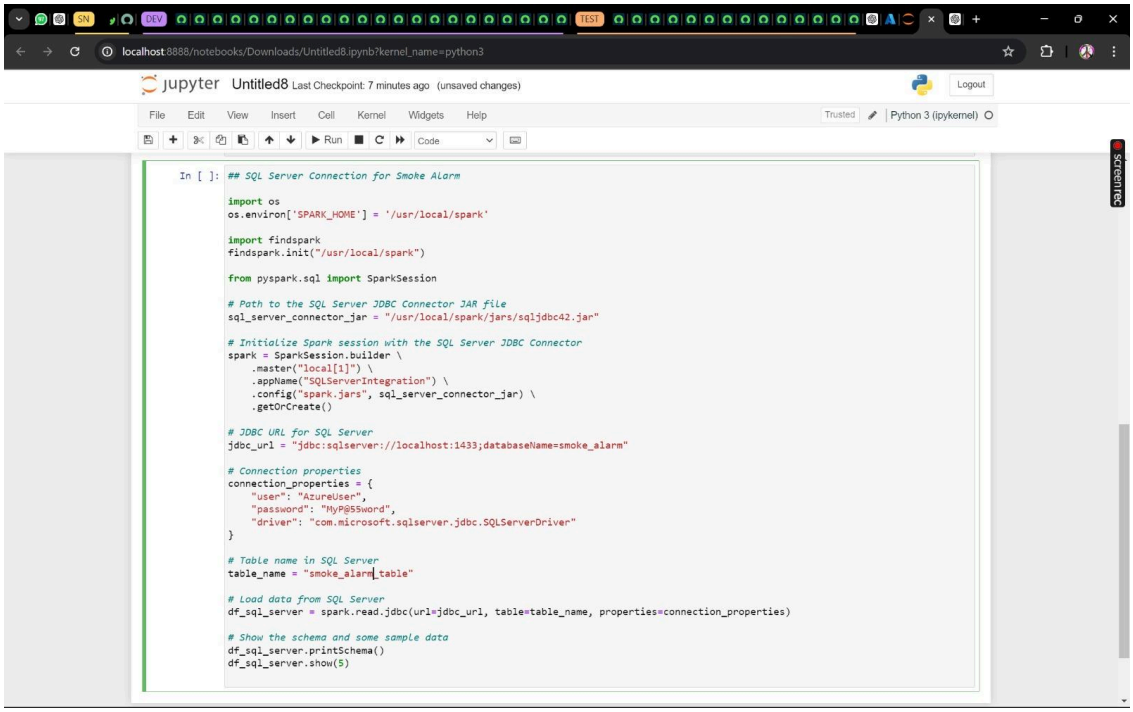
Connecting to SQL Server Database (Datafiniti Fast Food Restaurants):

- Set the Spark environment with findspark and initialize a SparkSession with SQL Server integration.
- Specify the SQL Server JDBC connector JAR file and the JDBC URL for the SQL Server database (fast_food).
- Load data from the SQL Server table (fast_food_table), print its schema, and display sample data.

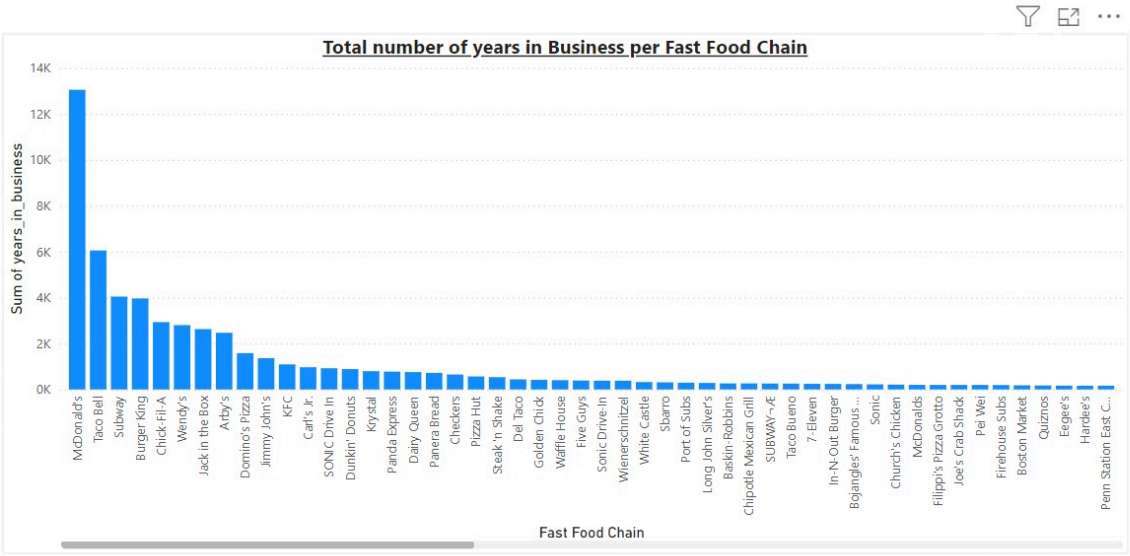


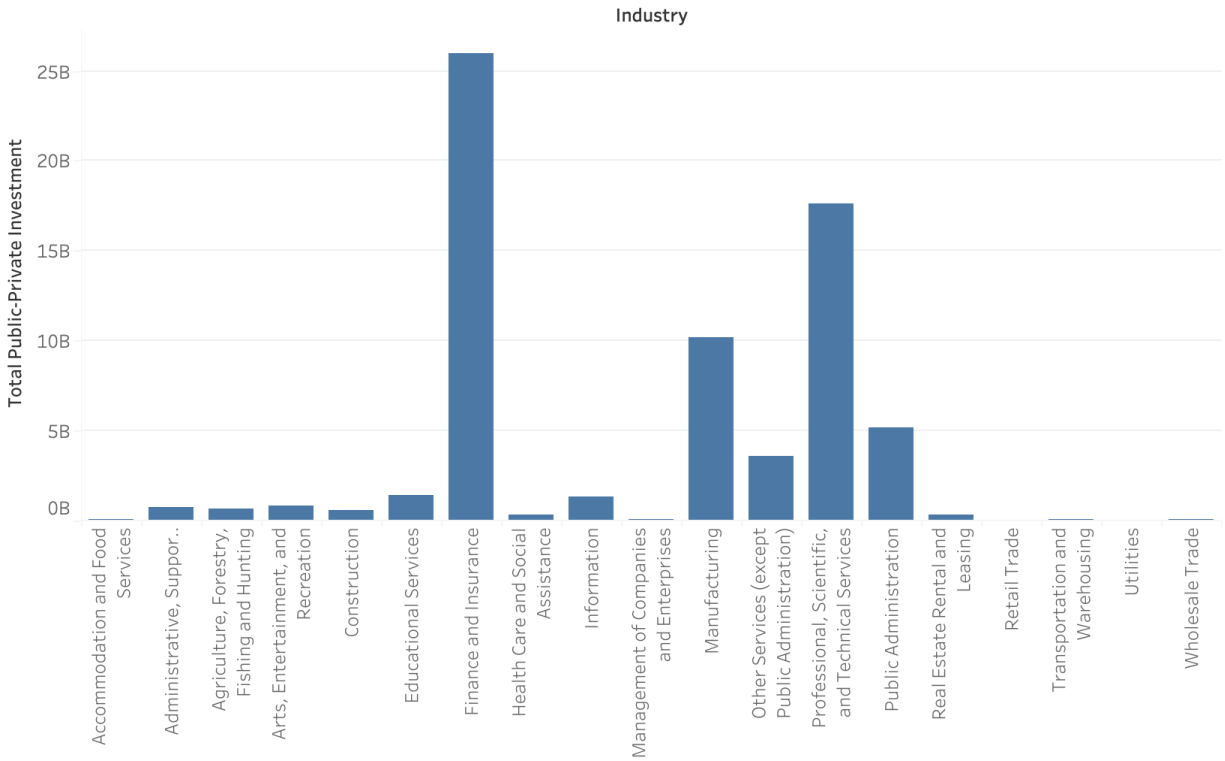
Connecting to SQL Server Database (Smoke Alarm):

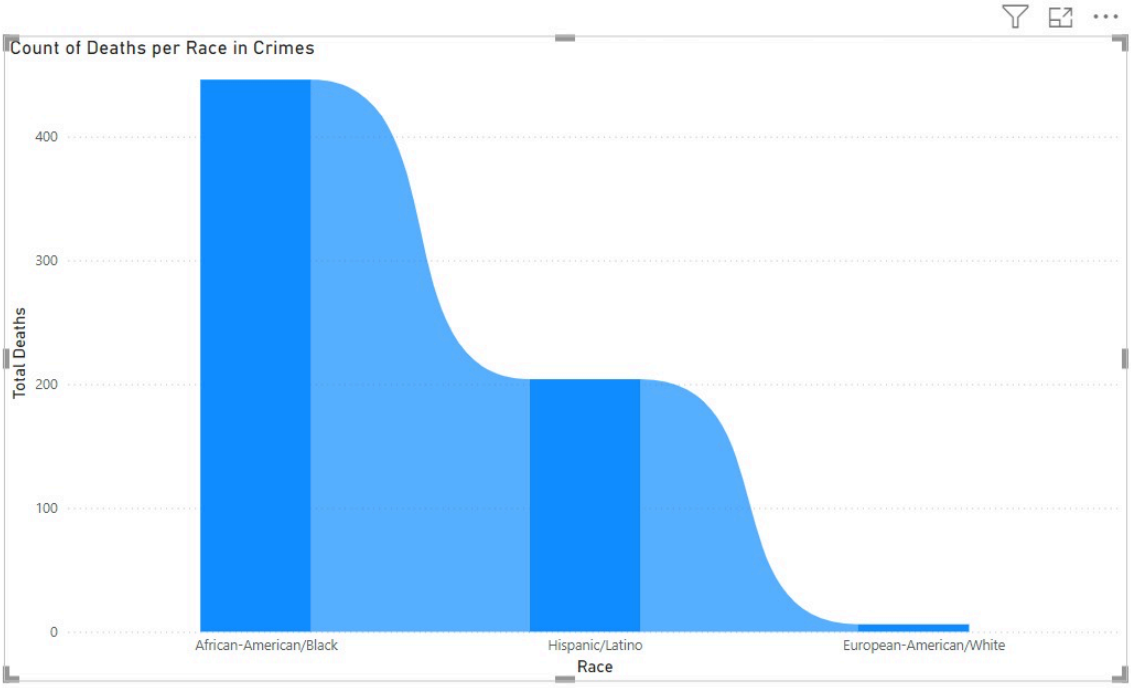
- Set the Spark environment with findspark and initialize a SparkSession with SQL Server integration.
- Specify the SQL Server JDBC connector JAR file and the JDBC URL for the SQL Server database (smoke_alarm).
- Load data from the SQL Server table (smoke_alarm_table), print its schema, and display sample data.

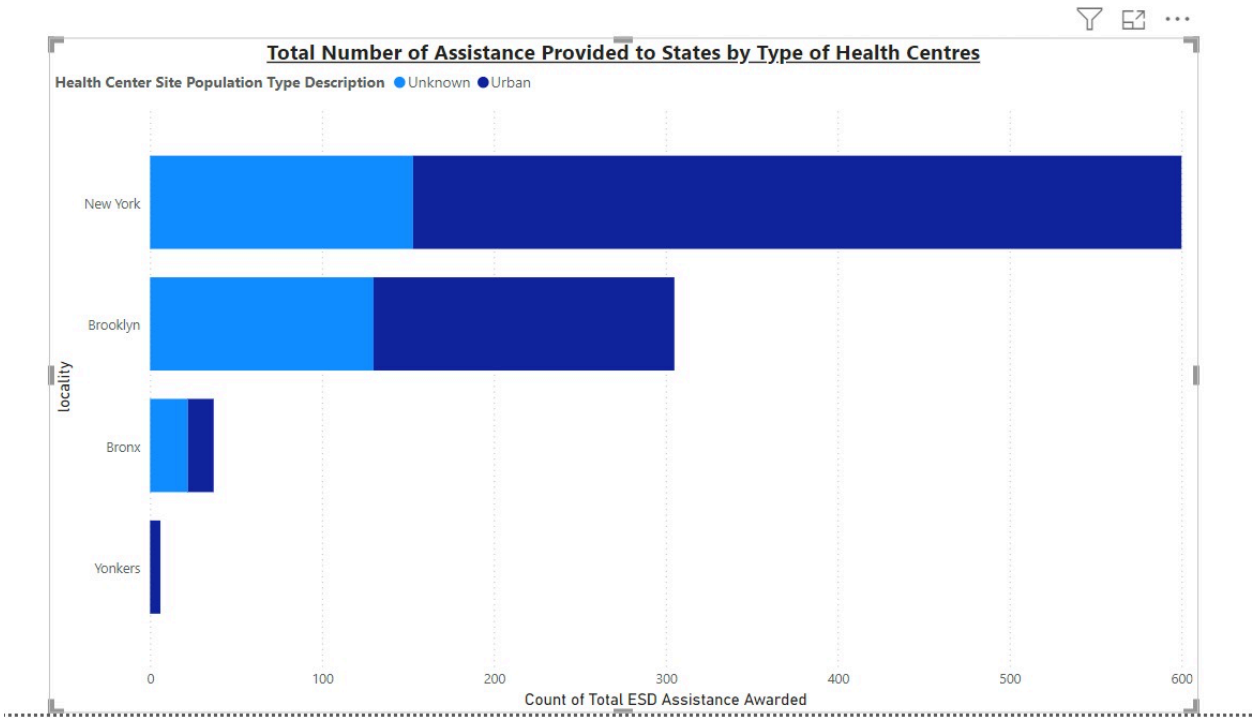


From our merged dataset, we came up with the following visuals:




Industry
Total Public-Private Investment
25B
20B
15B
10B
5B
0B
Accommodation and Food Services
Administrative, Suppor..
Agriculture, Forestry, Fishing and Hunting
Arts, Entertainment, and Recreation
Construction
Educational Services
Finance and Insurance
Health Care and Social Assistance
Information
Management of Companies and Enterprises
Manufacturing
Other Services (except Public Administration)
Professional, Scientific, and Technical Services
Public Administration
Real Estate Rental and Leasing
Retail Trade
Transportation and Warehousing
Utilities
Wholesale Trade


Count of Deaths per Race in Crimes
Total Deaths
400
300
200
100
0
African-American/Black
Hispanic/Latino
European-American/White
Race


Total Number of Assistance Provided to States by Type of Health Centres
Health Center Site Population Type Description Unknown Urban
New York
Brooklyn
Bronx
Yonkers
locality
0
100
200
300
400
500
600
Count of Total ESD Assistance Awarded