École Supérieure privée d'ingénierie et de technologie

Data Engineering Project Report
Rapport

SAFE

Élèves :
Achref BAABOURA
Emna NKHILI
Souleima GHARBI
Med Aziz MAATOUG
Khawla GRIRA
Ahmed Aziz AMMAR

Enseignant :
Nardine HANFI
Mariem GLAA

2024 - 2025

Table des matières

"The only limit to our realization of tomorrow will be our doubts of today."

– Franklin D. Roosevelt

Acknowledgments and Reflections :

We are incredibly proud of how far we have come and what we have accomplished. This project represents the fruit of countless hours of hard work, dedication, and determination. It serves as a testament to our perseverance and teamwork throughout this journey.

We would also like to express our heartfelt gratitude to :
Miss Nardine Hanfi and **Miss Mariem Glaa** for their unwavering support and encouragement. Your belief in us has been a source of inspiration and motivation, and we are truly grateful for your guidance.

Abstract

There are significant challenges in the current food safety landscape :

Customers : Customers often lack clear and reliable information, relying instead on online reviews that are frequently biased, vague, or insufficient.

Restaurant Owners : Restaurant owners face inefficiencies in managing their resources as they are often unaware of the actual problems identified during inspections, making proactive improvements difficult.

Health Inspectors : Health inspectors struggle with outdated systems and processes, which complicates the effective prioritization of high-risk establishments.

These limitations collectively hinder all stakeholders from making well-informed decisions, preventing the achievement of an optimal level of food safety.

FIGURE 1

Context and Dataset Insights

Throughout this project, we primarily focused on US's restaurant inspection data (mainly California State), which plays a central role in assessing food safety. In the United States, health inspections assign a score to establishments based on their compliance with food safety standards. These scores determine the grade of an establishment :

— **A grade :** Score of 90 or higher.
— **B grade :** Score between 80 and 89.
— **C grade :** Score below 80.

Establishments with a score below 90 may face penalties or closures, making the grading system a crucial component in ensuring food safety.

The dataset we worked with provided key information such as :

— **Facility Type :** Type of establishment (e.g., restaurant, store).
— **Risk :** Assigned risk level based on inspection findings.
— **Capacity :** The seating capacity of the establishment.

This data allowed us to understand how various factors influence restaurant safety and helped us focus on predicting inspection outcomes.

FIGURE 2

1 Project Context

1.1 Introduction

The **SAFE Project** is designed to revolutionize food safety management by addressing the pressing challenges faced by customers, restaurant owners, and health inspectors. Leveraging a structured, data-driven approach through the **CRISP-DM methodology**, SAFE ensures efficient execution and tangible results for all key stakeholders.

Food safety remains a critical concern across the restaurant industry. *Customers* often lack access to reliable and clear information about restaurant hygiene, relying instead on biased or incomplete online reviews. *Restaurant owners*, on the other hand, struggle to pinpoint the exact issues highlighted during inspections, limiting their ability to proactively address problems and improve operations. Meanwhile, *health inspectors* face inefficiencies due to outdated systems, which hinder their ability to prioritize high-risk establishments and optimize their inspections.

SAFE bridges these gaps by introducing an innovative, integrated platform that enhances transparency, improves decision-making, and streamlines inspection processes. By transforming fragmented systems into a cohesive solution, SAFE empowers all stakeholders to make informed decisions, driving significant improvements in food safety standards.

1.2 Proposed Solution

The SAFE Project provides a comprehensive, data-driven solution that converts existing challenges into actionable opportunities through the following key components :

- **Intuitive Platform :** A user-friendly interface consolidating essential food safety data into a clear, accessible format for all stakeholders.
- **Interactive Dashboard :** Tailored Key Performance Indicators (KPIs) and insights for restaurant owners, enabling quick, data-backed decisions to enhance safety and compliance.
- **Inspector Tools :** Advanced inspection prioritization powered by predictive machine learning models, improving efficiency and focusing efforts on high-risk establishments.

By integrating **transparency**, **analytics** and **automation** SAFE transforms food safety management into a proactive, intelligent process. It empowers stakeholders with reliable tools, actionable insights, and streamlined workflows to meet the evolving demands of the restaurant industry.

Figure 3 – SAFE : Revolutionizing Food Safety Management

2 Concept and Definitions

2.1 Core Definitions

- **CRISP-DM framework :**

A standard model for data science projects, structured into six key phases : understanding objectives, exploring data, preparing data, modeling, evaluating results, and deployment.

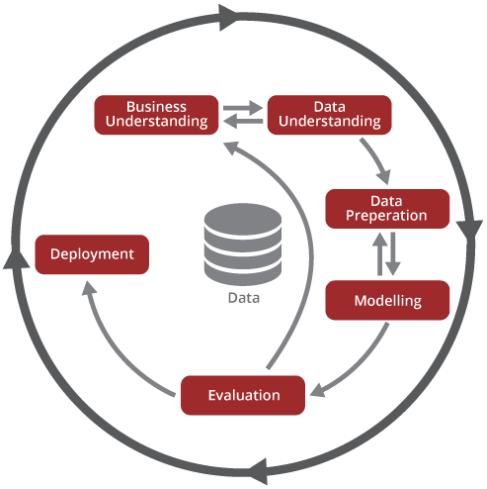


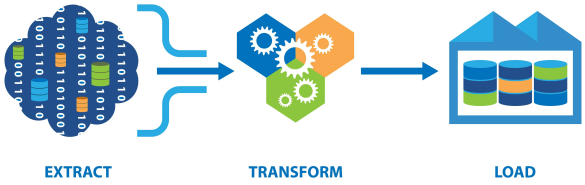
FIGURE 4 – The CRISP-DM Framework

- **Machine Learning :** Machine learning (ML) is a branch of artificial intelligence that focuses on using data and algorithms to enable systems to learn from experience and improve their performance over time, imitating human-like pattern recognition and decision-making processes

In the Context of 'Safe' it was the application of predictive models to classify restaurants based on their risk levels or customer satisfaction..etc

- **Extraction, Transformation, and Loading (ETL) process :** ETL of data into a structured warehouse, including :

— **Data Cleaning :** Managing missing values and removing duplicates.
— **Transformation :** Normalization, data encoding, and the creation of new columns based on specific rules.
— **Loading :** Storing data in a data warehouse for advanced analysis.



2.2 Tools Overview

- **SQL Server Integration Services(SSIS) :**

Used to automate the ETL process, integrating features such as grouping, type conversions, and conditional splits.

- **SQL Server Management Studio (SSMS) :**

An environment used to manage and query data stored in the warehouse.

- **Power BI :**

A visualization tool for designing interactive dashboards.

- **Scikit Learn (Python and Machine Learning framework)**

Used for developing machine learning models, including logistic regression, random forest, K-nearest neighbors (KNN), support vector machine (SVM), etc.

3 Data Preparation and Warehousing

3.1 ETL Process and Data Cleaning

In this phase of the project, the focus was on preparing, transforming, and structuring the data to ensure its readiness for analysis and model development. Given the complex nature of the raw data, an efficient ETL process was implemented using SSIS.

Loading Data and Transformation Processes After completing the extraction phase, the transformed data was systematically processed and loaded into the data warehouse. This stage ensured the data was structured, clean, and ready for downstream tasks such as dashboard creation and machine learning modeling.

Key transformation techniques were applied during this phase using SSIS tools :

- **Derived Columns :**

New columns were created by deriving values from existing ones, enabling more meaningful data analysis. For example, calculated fields such as totals or categorical labels were generated

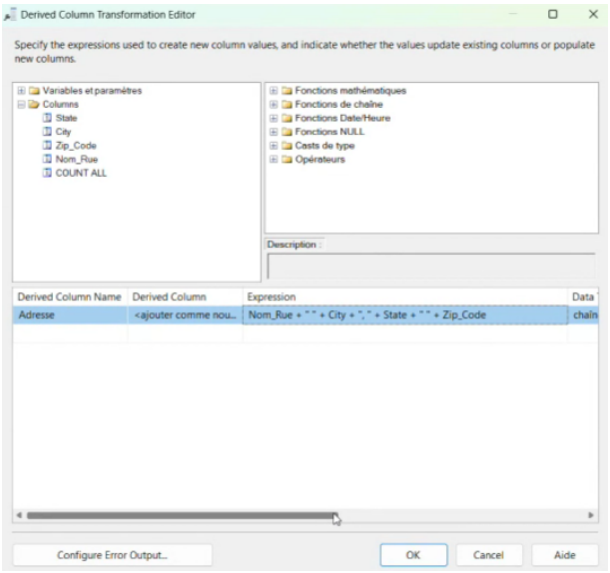


FIGURE 5 – Derived Column Screenshot 1 from VS (SSIS)

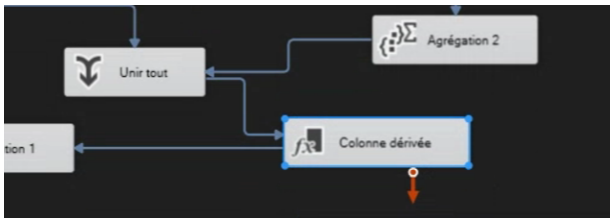


FIGURE 6 – Derived Column Screenshot 2 from VS (SSIS)

- **Aggregation :**

Duplicate records were identified and removed to ensure data uniqueness and consistency. This step ensured the data warehouse contained only distinct records for more accurate analysis

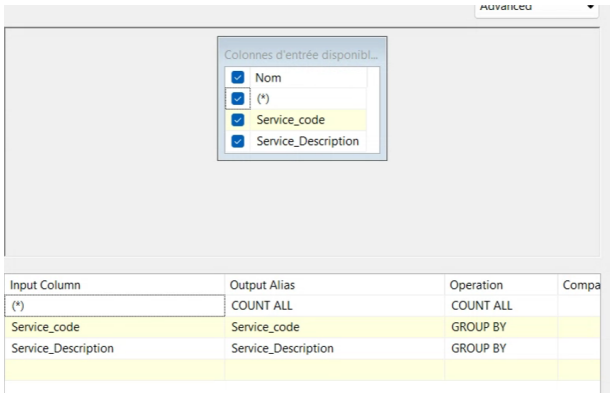


FIGURE 7 – Screenshot of Aggregation from Visual Studio (SSIS)

- **Union All :**

Data from multiple sources or tables was combined into a single dataset to ensure completeness and eliminate fragmentation. Data Type Conversion :

Data types were standardized to align with the schema of the data warehouse. This step ensured compatibility and avoided processing errors.

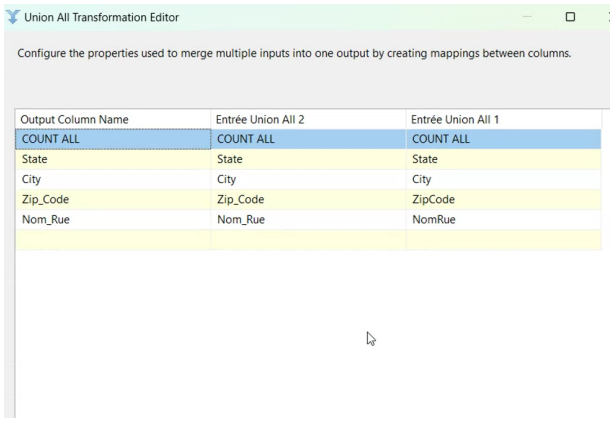


FIGURE 8 – Union All Screenshot from SSIS

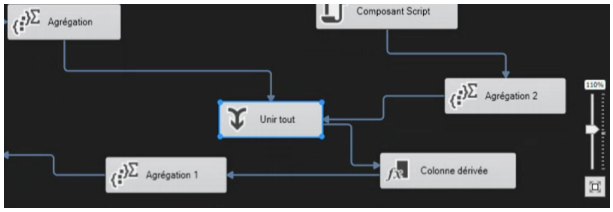


FIGURE 9 – Union All Screenshot from SSIS

- **Lookup :**

Used to match source data with reference data (usually from a dimension table) to ensure data integrity. It verifies that records in the fact table have corresponding matches in the dimension table, ensuring that foreign keys (FK) in the fact table are replaced with the correct primary keys (PK) from the dimension

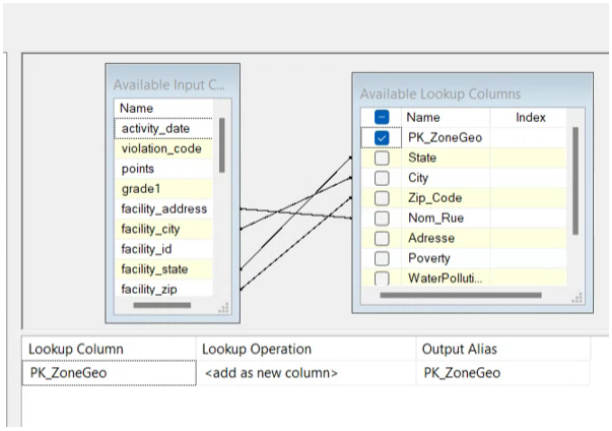


FIGURE 10 – Lookup Screenshot

Through these transformations, the ETL process not only cleaned and organized the data but also added contextual value, setting the foundation for robust dashboards and predictive machine learning models.

3.2 Dimentional Modelings

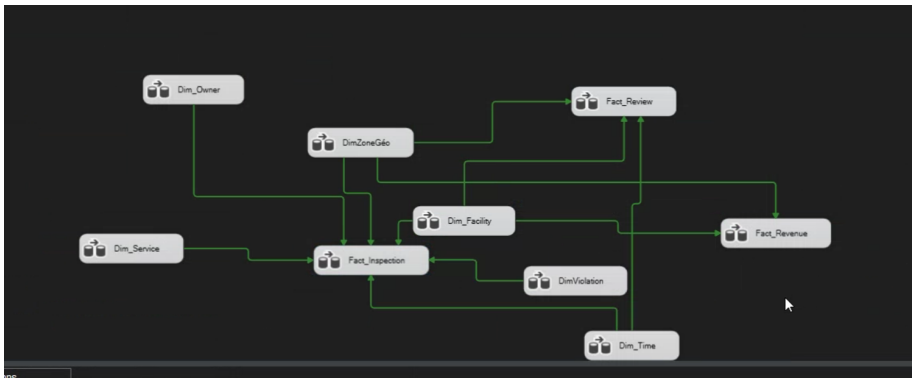


FIGURE 11 – General Structure Screenshot from SSIS

Dimensional modeling is a critical step in the design of a data warehouse, enabling efficient data retrieval and analysis The schema presented in the image follows the Fact Constellation Schema design it allows for multiple fact tables.

- **Fact-Inspection :**

Holds facility inspection-related metrics.

- **Fact-Review :**

Contains restaurant review-related metrics.

- **Fact-Revenue :**

Stores revenue data, and restaurant performance.

Each of these fact tables uses similar dimension tables (like Dim-Facility, Dim-Time, etc.), but they are separate facts for different aspects of the business.

Shared Dimension Tables :

- **Dim-Owner :**

Describes owners, potentially relevant for both inspections and reviews.

- **Dim-Service :**
Describes the services, which could be relevant for both reviews and revenue.
- **Dim-Facility :**
Represents facilities, which are relevant for all three facts (inspection, review, and revenue).
- **Dim-Violation :**
represents violation in facilities.
- **Dim-Time :**
Used across all facts to provide time-based context.
- **Dim-ZoneGéo :**
Used to describe geographical zones related to facilities, which could impact all three facts (inspections, reviews, and revenue).

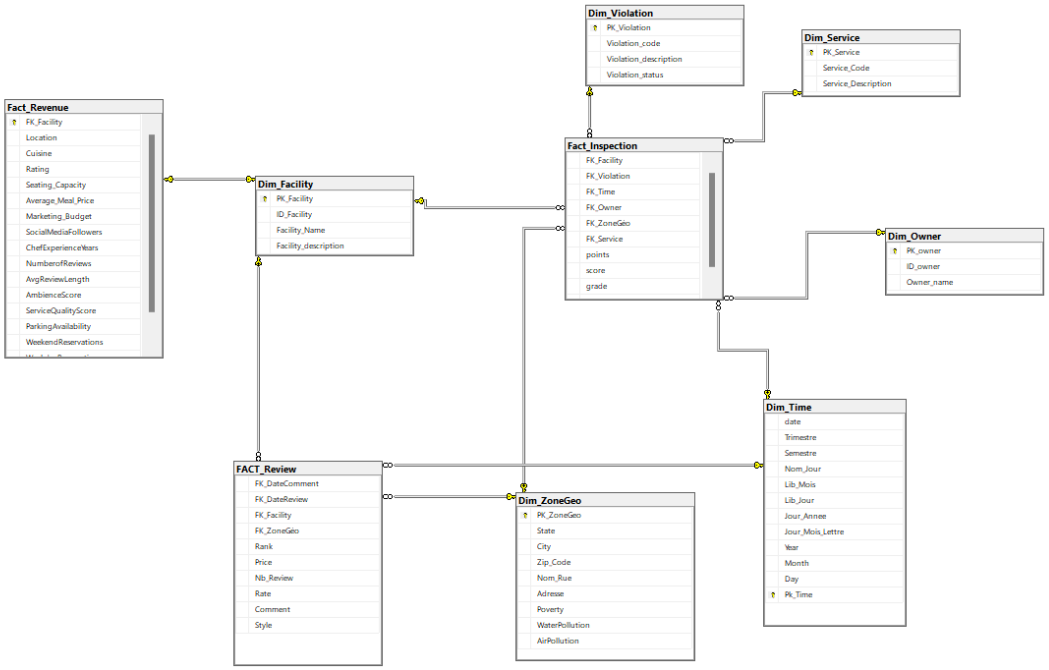


FIGURE 12 – Model of Data Warehouse

3.3 Understanding the Entire Process : A Short Final Summary

The process began with loading resources from a flat file and partitioning the data into distinct staging areas, including geo, facility, and time zones. In the subsequent dimension implementation phase, we extracted the partitioned data from the staging areas, followed by data cleaning and transformation to ensure consistency and quality. Lastly, in the fact table, we integrated the cleaned and transformed data from the dimensions, ensuring alignment for analysis. This methodical approach optimizes data flow, maintains data integrity, and provides a solid foundation for generating accurate and actionable insights.

4 Dashboard Development

4.1 General Introduction

The dashboard development phase is a critical component of our project, designed to provide actionable insights to key stakeholders : **clients, inspectors,** and **restaurant owners**.

By leveraging interactive **Power BI dashboards**, this phase ensures that data is presented in an intuitive, visually engaging manner, empowering decision-makers to make informed choices.

Each dashboard has been tailored to meet the specific objectives, key performance indicators (KPIs), and requirements of its respective decision-maker.

Machine learning models were also integrated into certain dashboards to enhance predictive capabilities and proactive decision-making.

4.2 Decision Makers

The dashboards cater to three primary decision-makers :

— **Client**
— **Inspector**
— **Restaurant Owner**

Each decision-maker has unique objectives and KPIs, which have been translated into individual dashboard designs to deliver targeted insights.

4.3 Dashboards Design

4.3.1 Client Dashboard

Objective :

The client dashboard provides customers with the information needed to make informed dining choices based on inspection scores, risk levels, and reviews. It fosters transparency and trust in the restaurant selection process.

A- KPIs :

1. **Service Quality by Cuisine type**

   Purpose : To compare the popularity of cuisines based on customer preferences and dining trends.

   Threshold :

```
DAX

Popularity_Goal_Numeric =
SWITCH(
    TRUE(),
    [Rank_By_Popularity] <= 3, 1,  -- "Excellent"
    [Rank_By_Popularity] <= 5, 2,  -- "Average"
    3                              -- "Poor"
)
```

Values :

```
DAX

Rank_By_Popularity =
RANKX(
    ALL('Fact_Revenue'[Cuisine]),
    [OverallPopularity],
    ,
    DESC,
    Dense
)
```

Trend Axis : Cuisine.

2. **Service Quality by Restaurant**

Prupose : To rank restaurants based on combined service quality and ambiance scores.

Threshold :

```
DAX

ServiceQualityIndex =
0.7 * AVERAGE(Fact_Revenue[ServiceQualityScore]) +
0.3 * AVERAGE(Fact_Revenue[AmbienceScore])
```

Goal :

```
DAX

ServiceQualityGoal =
SWITCH(
    TRUE(),
    [ServiceQualityIndex] >= 9, 1, -- "Excellent"
    [ServiceQualityIndex] >= 7, 2, -- "Acceptable"
    3                              -- "Poor"
)
```

Trend Axis : Facility Name

3. **Average Score Over Time**

Purpose : To analyze restaurant performance trends over time.

Threshold : Maximum score for a given period.

Value : Average score for restaurants

Trend Axis : Year.

4. **High-Risk Establishments**

Purpose : To monitor the number of high-risk facilities annually and ensure client safety.

Threshold : Target-High-Risk = 1000.

Value :

```
DAX

High_Risk_Establishments =
COUNTROWS(FILTER(Dim_Facility, Dim_Facility[risk_level] = "HIGH RISK"))
```

Trend Axis : Year.

B- Measures :

— **Average Chef Experience :**

Provides insights into the culinary expertise of establishments.

— **Count of Violations :**

Offers an overview of the frequency of health code violations across establishments.

— **Count of Facilities :**

Provides insights into the total number of establishments included in the analysis.

— **Customer Satisfaction :**

```
Rating_Class =
SWITCH(
    TRUE(),
    'Fact_Revenue'[Rating] <= 3, "Unsatisfactory",
    'Fact_Revenue'[Rating] > 3 && 'Fact_Revenue'[Rating] <= 4, "Satisfactory",
    'Fact_Revenue'[Rating] > 4, "Highly Satisfactory"
)
```

— **Service Quality Classification :**

```
ServiceQualityGoal2 =
SWITCH(
    TRUE(),
    [ServiceQualityIndex] >= 9, "Excellent",
    [ServiceQualityIndex] >= 7, "Acceptable",
    "Poor"
)
```

C- Individual Dashboards in Power BI :

1. **Restaurant with Best Service Quality**

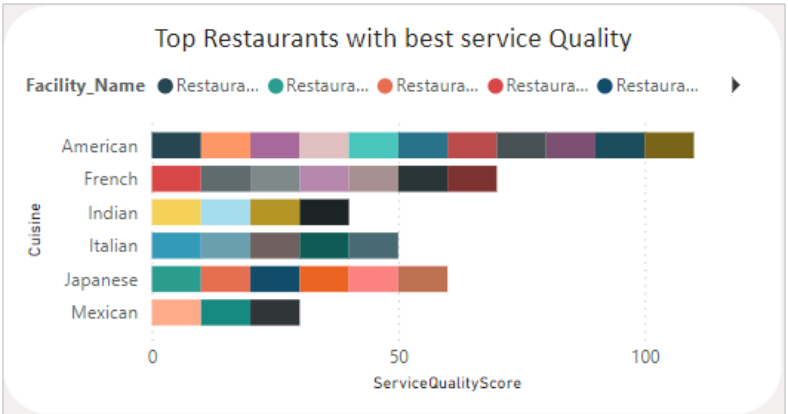


Chart Type : Bar Chart
X-Axis : Service Quality Score
Y-Axis : Cuisine
Prupose : Highlights cuisines offering the best service quality.

2. **Facility Rating by Location**

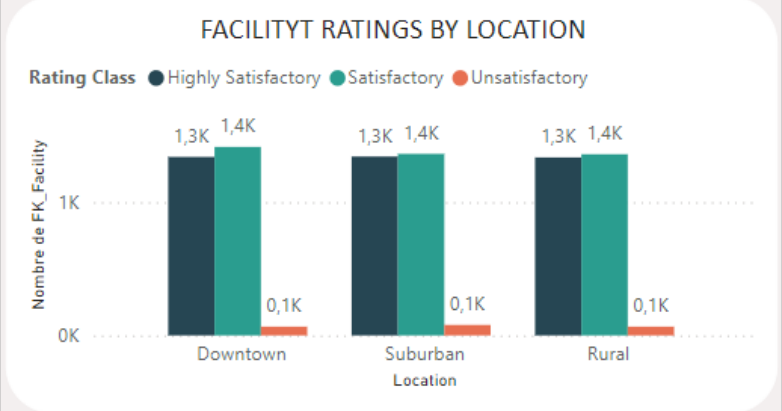


Chart Type : Grouped Histogram
X-Axis : Location (Downtown, Suburban, Rural)
Y-Axis : Number of Facilities
Legend : Rating Class (Highly Satisfactory, Satisfactory, Unsatisfactory)
Prupose : Visualizes facility distribution by location and rating.

3. **Reservation Trends by Location**

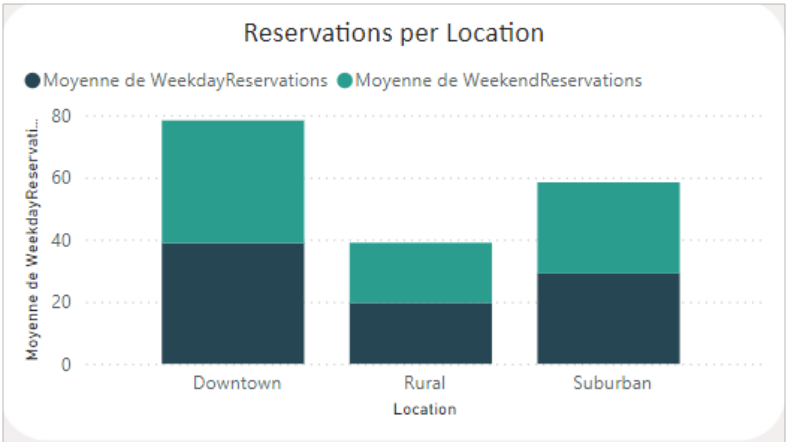
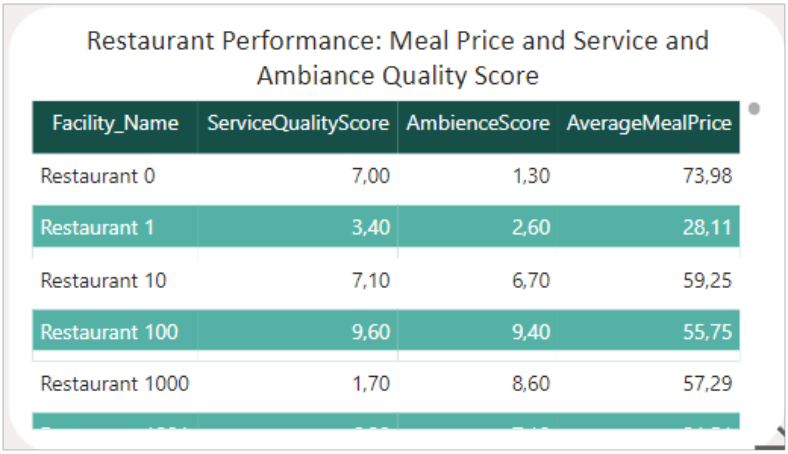


Chart Type : Stacked Line Chart
X-Axis : Location
Y-Axis : Average Weekday and Weekend Reservations
Prupose : Tracks reservation trends by location.

4. **Restaurant Performance Matrix**



Facility_Name	ServiceQualityScore	AmbienceScore	AverageMealPrice
Restaurant 0	7,00	1,30	73,98
Restaurant 1	3,40	2,60	28,11
Restaurant 10	7,10	6,70	59,25
Restaurant 100	9,60	9,40	55,75
Restaurant 1000	1,70	8,60	57,29

Columns : Meal Price, Service Quality, Ambiance Quality
Rows : Restaurant Name
Prupose : Evaluates restaurant performance metrics in a compact format.

5. **Top 5 Facilities by Total Score**

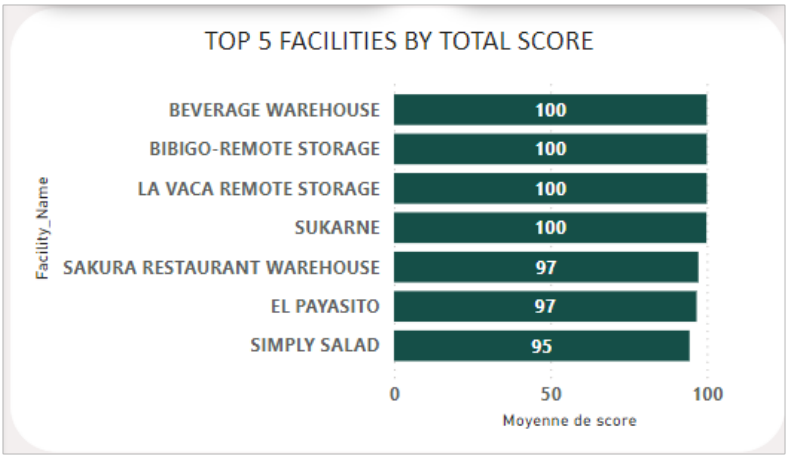


Chart Type : Bar Chart
X-Axis : Average Score
Y-Axis : Facility Name (Filtered to Top 5)
Prupose : Identifies high-performing facilities.

6. **Facility Risk Level Distribution by Location**

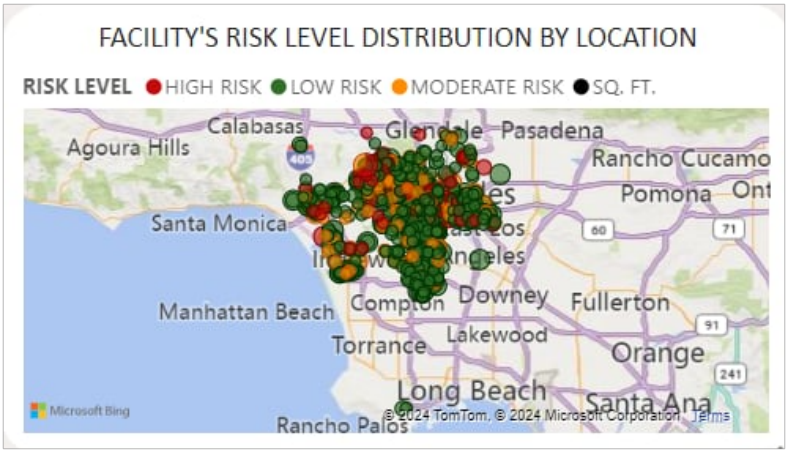


Chart Type : Map
Legend : Risk Level
Bubble Size : Violation Rate
Prupose : Visualizes the geographical distribution of risk levels.

7. **Scores Over Time**

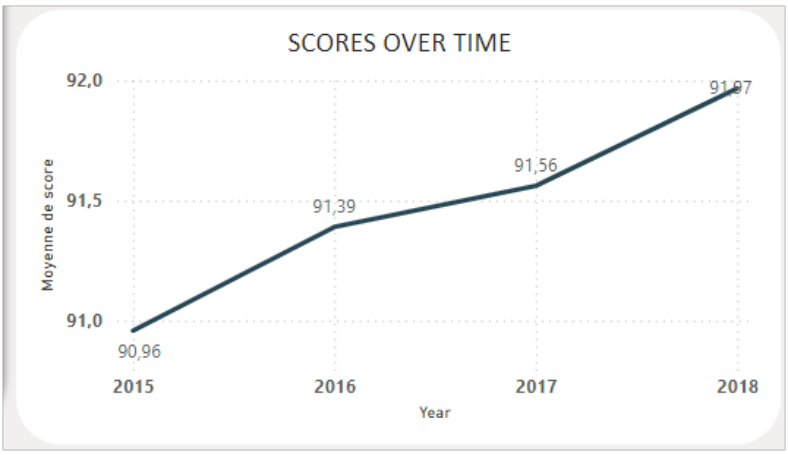


Chart Type : Line Chart
X-Axis : Year
Y-Axis : Average Score
Prupose : Tracks average inspection scores year-over-year.

8. **Inspection Scores by Establishment Type and Capacity**

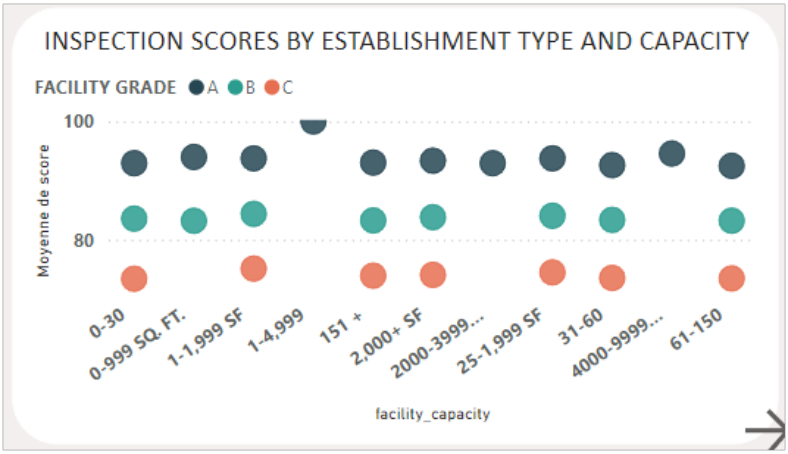


Chart Type : Scatter Plot
X-Axis : Facility Capacity
Y-Axis : Average Score
Legend : Grade (A, B, C)
Bubble Size : Number of Violations
Prupose : Analyzes the relationship between capacity, scores, and violations.

D- Filters :
— **Location :** Filter by location type (Downtown, Suburban, Rural).
— **Parking Availability :** Yes/No.
— **Style :** Filter by cuisine style (e.g., American, Indian, Italian).
— **Facility Address :** Filter by specific facility address.
— **Facility Type :** Caterer, Food Market, Restaurant, etc.
— **Facility Risk Level :** High, Low, Moderate.
E-Challenges :
— Ensuring data consistency across multiple filters and measures.
— Balancing dashboard complexity with user-friendliness.

4.3.2 Inspector Dashboard

Objective :
The Inspector Dashboard provides inspectors with the necessary insights to monitor inspection scores, violations, and environmental factors. It identifies facilities needing immediate attention and those at risk of closure to prioritize inspection resources effectively.
A- KPIs :

1. **Inspection Status :** Helps inspectors understand overall restaurant quality.

```
DAX

KPI_Status =
IF(
    AVERAGE(Fact_Inspection[score]) < 90,
    "⚠ Warning: Low Score",  -- Statut d'avertissement
    "✓ Score is Healthy"     -- Statut sain
)
```

Relation : Used in a gauge visualization to depict the average inspection score.

2. **Violation Status :** Helps inspectors detect the number of violations overall.

```
DAX

ViolationStatus =
IF(
    AVERAGE(Dim_Quality[number_of_violations]) < 25,
    "⚠️ Danger: High number of violations 🚨",
    "✅ Number of violations within an acceptable range"
)
```

3. **Score Mean Analysis :**
   — **Value :** Average of Fact-Inspection[score].
   — **Minimum Value :** Minimum score in Fact-Inspection[score].
   — **Maximum Value :** Maximum score in Fact-Inspection[score].

B- Measures :

— **Number of High-Risk Establishments :**

```
High_Risk_Establishments =
COUNTROWS(
    FILTER(
        Dim_Facility,
        Dim_Facility[risk_level] = "HIGH RISK"
    )
)
```

— **Compliance Rate :**

```
Compliance_Rate =
DIVIDE(
    CALCULATE(
        COUNT(Fact_Inspection[grade]),
        Fact_Inspection[grade] = "C"
    ),
    COUNT(Fact_Inspection[grade]),
    0
)
```

— **Total Inspections :**

```
Number_of_Inspections = COUNT(Fact_Inspection[FK_Time])
```

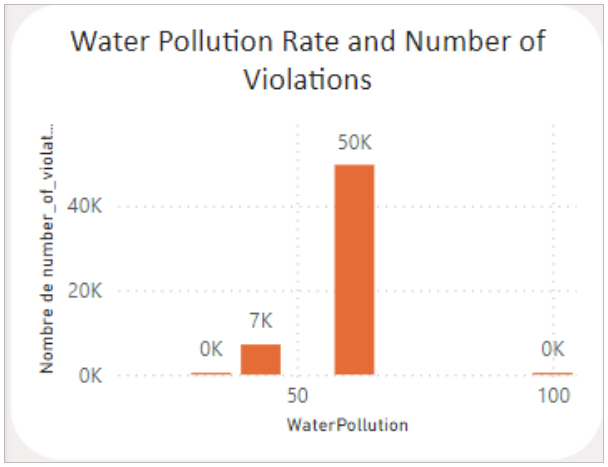
— **Poverty Rate :** Represents poverty percentages by geographic location.
— **Water Pollution Rate :**Displays pollution levels as reported in Dim-Quality.
— **Air Quality Rate :** Represents air pollution levels as reported in Dim-Quality.

— **Number of violations :**

```
Total_Violations = SUM('Dim_Quality'[number_of_violations])
```

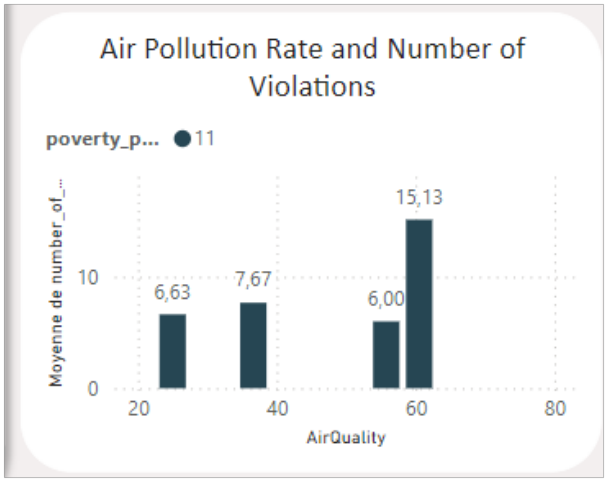
C- Individual Dashboards in Power BI :

1. **Water Pollution Rate vs. Number of Violations**



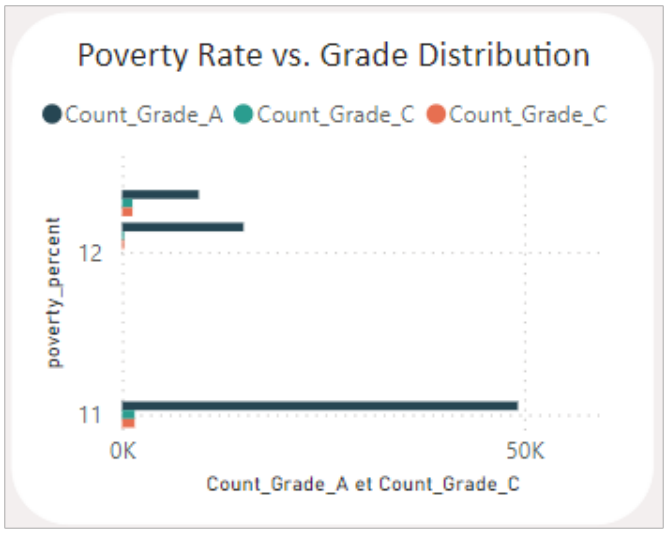
Visualization : Grouped Histogram
X-Axis : Water Pollution Rate
Y-Axis : Number of Violations
Purpose : Shows how water pollution levels relate to the number of violations in facilities.

2. **Air Pollution Rate vs Number of Violations**



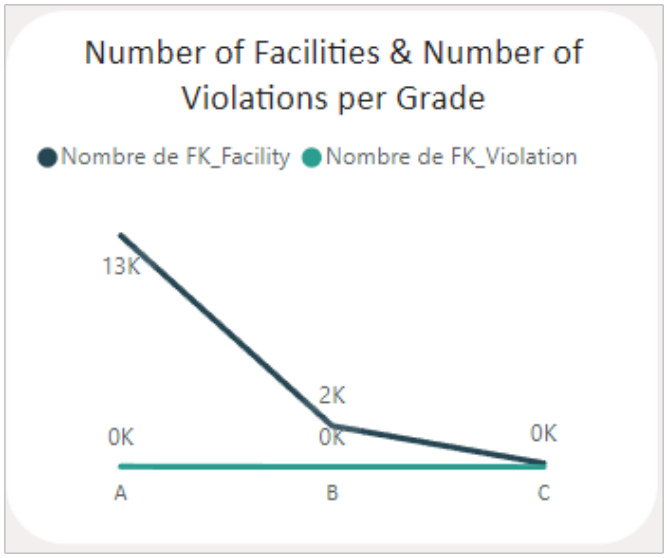
Visualization : Grouped Histogram
X-Axis : Air Quality Rate
Y-Axis : Number of Violations
Legend : Poverty Percent
Purpose : Displays the connection between air pollution levels and the number of violations.

3. **Poverty Rate vs Grade Distribution**



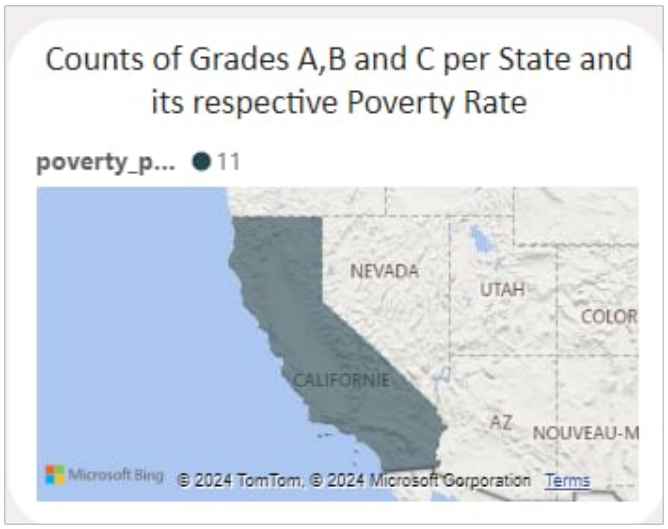
Visualization : Grouped Bar Chart
X-Axis : Water Pollution
Y-Axis : Count of Grades (A, B, C)
Purpose : Shows how poverty rates affect the grades (A, B, C) of facilities

4. **Facilities vs Violations Per Grade**



Visualization : Line Chart
X-Axis : Grades (A, B, C)
Y-Axis : Number of Facilities and Violations
Purpose : Tracks the number of facilities and violations for each grade (A, B, C).

5. **Counts of Grades by State and Poverty Rate**



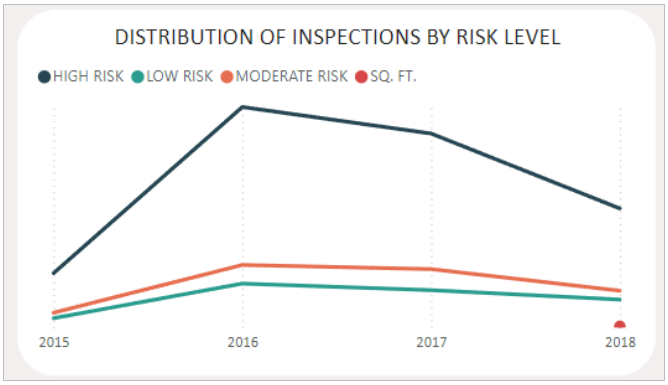
Visualization : Choropleth Map

Location : State

Legend : Poverty Percent

Purpose : Visualizes the distribution of facility grades by state and poverty rate.

6. **Distribution of Inspections by Risk Level**



Visualization : Line Chart

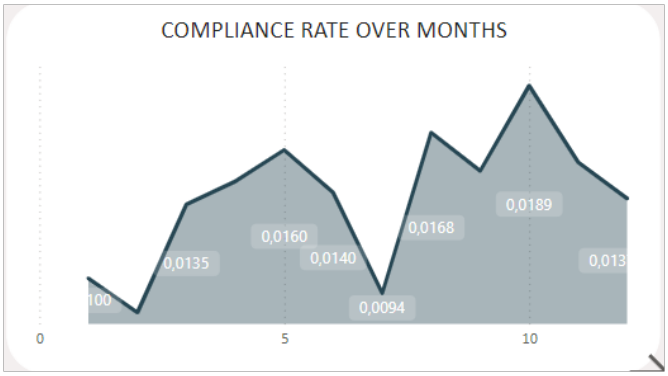
X-Axis : Year

Y-Axis : Number of Inspections

Legend : Risk Levels

Purpose : Tracks the number of inspections based on the risk level of the facilities over time

7. **Compliance Rate Over Months**

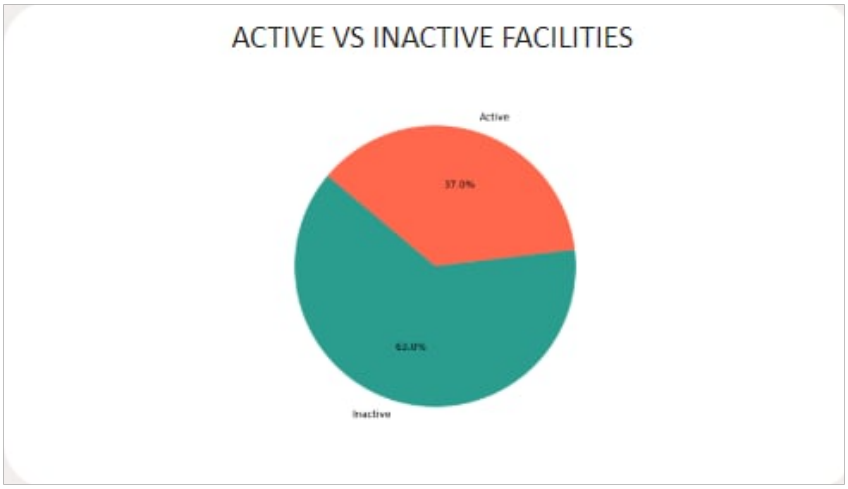


Visualization : Stacked Area Chart
X-Axis : Month
Y-Axis : Compliance Rate
Purpose : Shows how compliance rates change over the months.

8. **Python Visualization 1 : Facilities Needing Immediate Attention**

Purpose : A clustering model predicts facilities requiring urgent inspections. The visualization shows the Top 5 facilities identified by the model.

9. **Python Visualization 2 : Facilities Likely to Close**



Purpose : A clustering model predicts facilities at risk of closure. The visualization shows the distribution of facilities across active and inactive categories.

D- Filters :

— **Grade :** Filter by grade (A, B, C).
— **Number of Violations :** Filter facilities based on the count of violations.
— **Facility Name :** Filter by specific facility names.
— **Facility Address :** Filter by location or specific addresses.
— **Year :** Filter inspection data by year.

D- Challenges :

Predictive Model Accuracy :

The clustering models used to predict high-risk facilities or those likely to close may not always be accurate, leading to misidentification of facilities needing urgent attention.

4.3.3 Owner Dashboard

Objective :

The owner dashboard is aimed at providing business owners with insights into their establishment's compliance, customer feedback, and revenue trends.

It helps them identify areas of improvement and maintain a competitive edge.

A- KPIs :

1. **Violation Diversity**

Value :

```
Mediane_Violation_Diversity =
MEDIANX(
    SUMMARIZE(
        Fact_Inspection,
        Fact_Inspection[FK_Facility],
        "Diversity",
        CALCULATE(
            DISTINCTCOUNT(Dim_Violation[Violation_code])
        )
    ),
    [Diversity]
)
```

Trend Axis :

```
Violation_Diversity =
CALCULATE(
    DISTINCTCOUNT(Dim_Violation[Violation_code]),
    FILTER(
        Fact_Inspection,
        Fact_Inspection[FK_Facility] = EARLIER(Fact_Inspection[FK_Facility])
    )
)
```

2. **Number of Inspections**
   Value :

```
Number_of_Inspections = COUNT(Fact_Inspection[FK_Time])
```

   Trend Axis : Yearly inspections grouped by Fact-Inspection[FK-Time].
   Threshold :

```
Nombre_AT_RISK =
CALCULATE(
    COUNTA('Fact_Inspection'[Risk_of_Closure]),
    'Fact_Inspection'[Risk_of_Closure] IN { "At Risk" }
)
```

3. **Monthly Trend of Reviews**
   Value : Total number of reviews (grouped monthly).
   Threshold : 15,000 reviews.
   Trend Axis : Monthly analysis.
4. **Monthly Trend of Ratings (Target)**
   Value :

```
Max_Rate = MAX(Fact_Inspection[Rating])
```

   Trend Axis : Ratings grouped by month.
   Threshold : Average monthly rating.

B- Measures :

— **Maximum Rank :**

```
Max_Rank = MAX(Dim_Facility[Rank])
```

— **Price Level :** Measure to calculate average price levels across facilities.

— **Number of Moderate Risk Facilities :**

```
Number_Moderate_Risk_Facilities =
COUNTROWS(
    FILTER(
        Fact_Inspection,
        Fact_Inspection[Risk_Level] = "Moderate Risk"
    )
)
```

— **Number of Low Risk Facilities :**

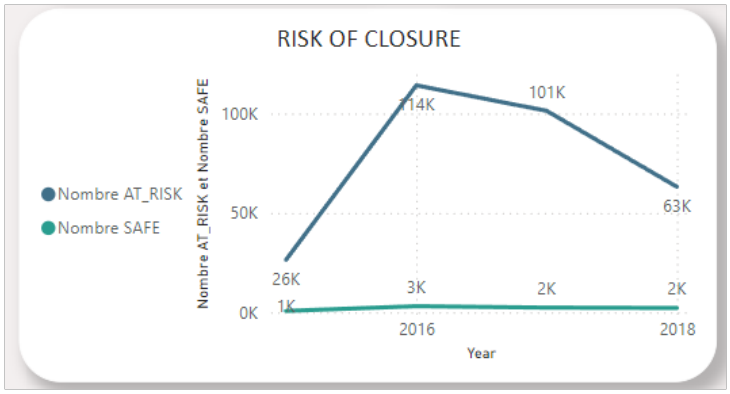
```
Number_Low_Risk_Facilities =
COUNTROWS(
    FILTER(
        Fact_Inspection,
        Fact_Inspection[Risk_Level] = "Low Risk"
    )
)
```

— **Number of High Risk Facilities :**

```
Number_High_Risk_Facilities =
COUNTROWS(
    FILTER(
        Fact_Inspection,
        Fact_Inspection[Risk_Level] = "High Risk
```

C- Individual Dashboards in Power BI :

1. **Risk of Closure :**



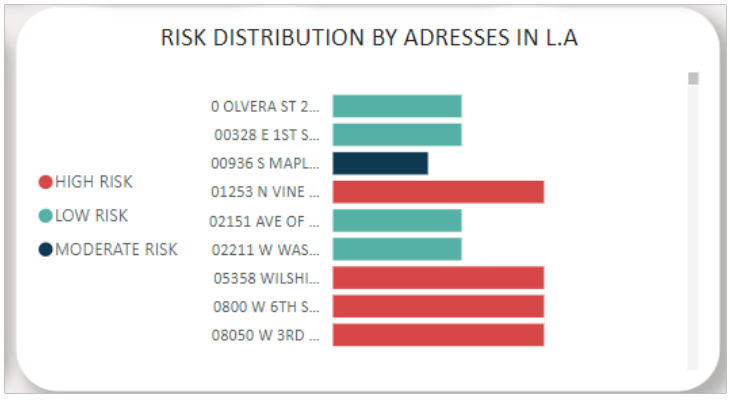
Visualization : Line Chart

X-Axis : Year

Y-Axis : Number of facilities at risk, Number of safe facilities

Purpose : Shows how many facilities are at risk of closure and how many are safe, over time.

2. **Risk Distribution by Addresses in L.A :**



Visualization : Stacked Bar Chart

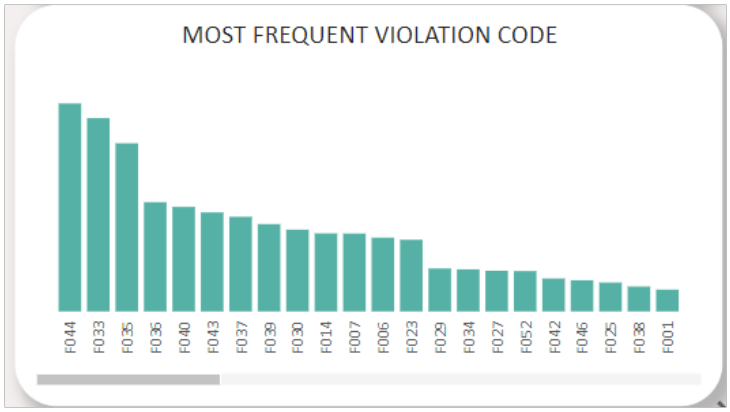
Y-Axis : Addresses

X-Axis : Number of facilities

Legend : Risk Levels

Purpose : Displays the risk levels of facilities by their addresses in Los Angeles, helping owners focus on high-risk areas.

3. **Most Frequent Violation Code :**



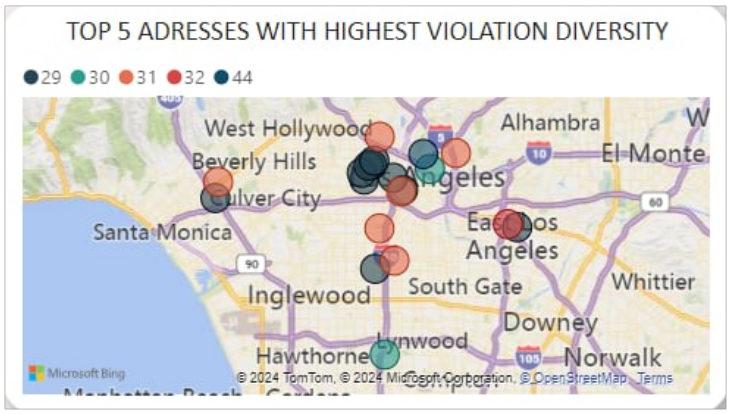
Visualization : Grouped Histogram

X-Axis : Violation codes

Y-Axis : Sum of violations

Purpose : Shows the most common violations, helping owners understand where they need to improve.

4. **Top 5 Addresses with Highest Violation Diversity :**



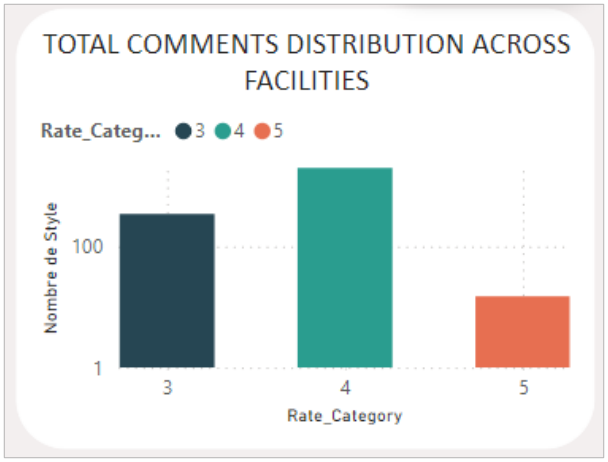
Visualization : Map

Location : Addresses

Legend : Violation Diversity

Purpose : Highlights the top 5 locations with the most different types of violations, helping owners address complex issues.

5. **Total Comments Distribution Across Facilities :**



Visualization : Stacked Histogram

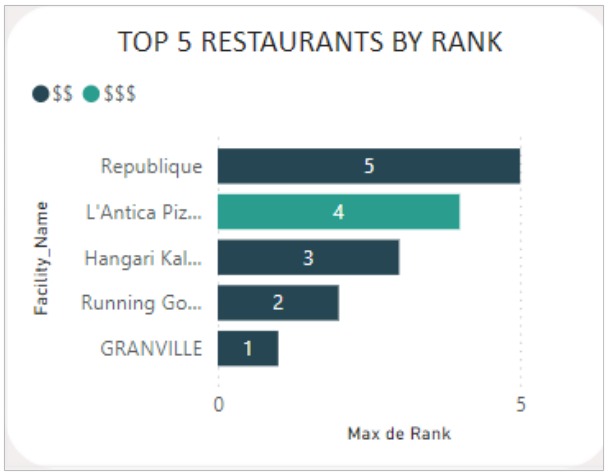
X-Axis : Rate category

Y-Axis : Number of facilities

Legend : Rate category

Purpose : Displays how facilities are rated based on customer feedback, helping owners see which ones need attention.

6. **Top 5 Restaurants by Rank :**



Visualization : Stacked Bar Chart

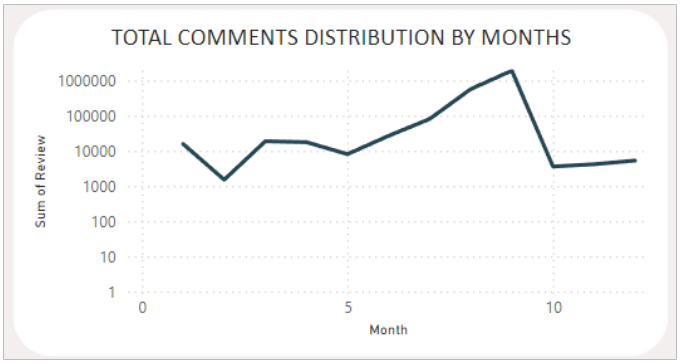
Y-Axis : Facility name

X-Axis : Maximum rank

Legend : Price

Purpose : Shows the top 5 facilities by rank and price, allowing owners to assess their best-performing locations.

7. **Total Comments Distribution by Months :**



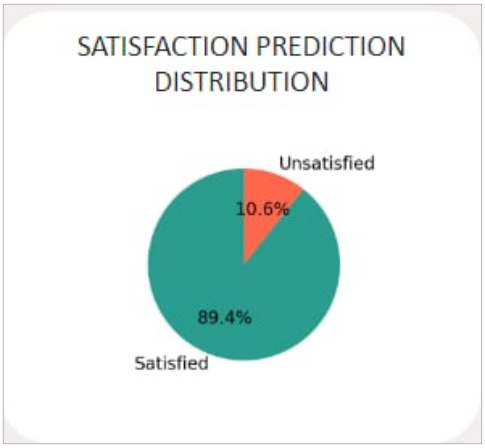
Visualization : Line Chart

X-Axis : Month

Y-Axis : Sum of reviews

Purpose : Tracks the number of customer reviews each month, helping owners monitor customer engagement over time.

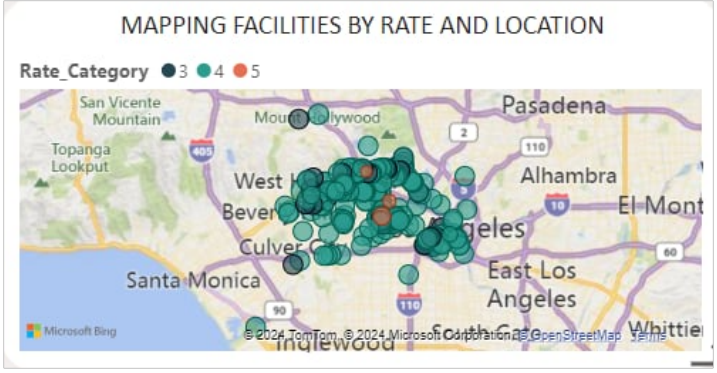
8. **Python Visualization : Satisfaction Prediction Distribution :**



Purpose : An NLP model predicts customer satisfaction based on comments.

Output : Pie chart showing satisfied vs. unsatisfied clients.

9. **Mapping Facilities by Rate and Location :**



Visualization : Map

Location : Addresses

Legend : Rate category

Bubble Size : Number of price levels

Purpose : Shows facilities by their rating and location, helping owners understand performance across different areas.

C- Filters :

— Facility Name : Filter by facility name.

— Facility Address : Filter by specific facility addresses.

— Year : Filter by year of inspection.

D- Challenges :

Integration of NLP Model for Satisfaction Prediction :

Integrating an NLP model to predict customer satisfaction can be complex, especially when dealing with unstructured text data (e.g., comments). Training and maintaining the model requires ongoing effort.

4.4 Model Integration :

The deployment and monitoring of predictive models are crucial to integrating AI solutions into the overall system and ensuring their accuracy and reliability over time. This section outlines the steps required to deploy and monitor predictive models effectively within a Power BI context.

-Steps for Deployment :

1. **Model Training and Saving :** Models, after being trained, are saved in a .pkl format, ready for deployment.
2. **Power BI Setup :** Power BI Desktop is configured to use Python scripts for loading the saved models, including both the sentiment analysis model and the clustering model.
3. **Python Script Integration :** A Python visual in Power BI allows for the vectorization of customer reviews and the generation of sentiment predictions using the pre-trained sentiment analysis model, as well as for clustering facility data to identify at-risk establishments.
4. **Real-Time Predictions :** As new data is ingested into Power BI, predictions from both the sentiment analysis and clustering models are made and visualized dynamically, enabling businesses to monitor customer sentiment and facility risk continuously.
5. **Scheduled Refresh :** The integration also supports scheduled refreshes, ensuring that the latest data is always processed and predictions from both models are updated regularly.

4.5 User Management

In Power BI, user roles enable row-level security (RLS), ensuring that each user sees only the data relevant to them. The dashboard is designed for three roles : **Client**, **Owner**, and **Inspector**, each with access to specific pages :

— **Client** : Access to *Client* and *Client2*.
— **Owner** : Access to *Owner* and *Owner2*.
— **Inspector** : Access to *Inspector* and *Inspector2*.
— **Common Page** : *Home*, accessible by all roles.

Each page is tailored to the user's role, showing relevant visualizations. For example, the *Client* page shows customer data, while the *Owner* and *Inspector* pages focus on operational and inspection metrics, respectively.

Upon login, the *Home* page displays role-specific buttons. For instance, when a **Client** logs in, they will see buttons for the *Client* and *Client2* pages. Clicking on these buttons redirects the user to the corresponding page. Similarly, **Owner** and **Inspector** will see buttons for their respective pages, such as *Owner*, *Owner2*, *Inspector*, and *Inspector2*.

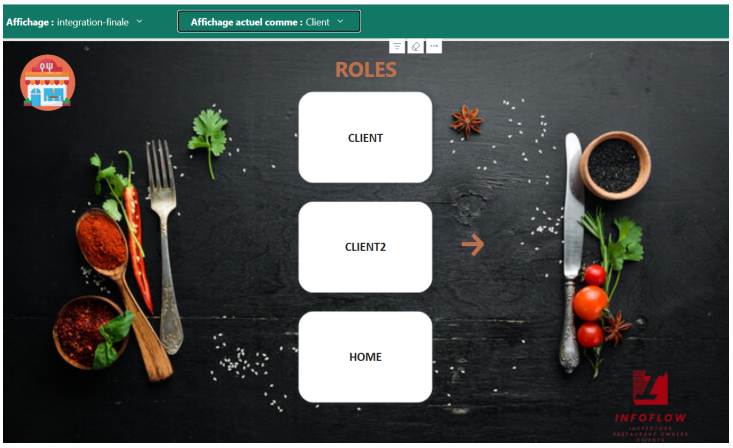


FIGURE 13 – Home Page

To manage access, we created a table in the SSMS data warehouse named `Pages`, which includes :

— `id_page` : Auto-incremented page ID.
— `PageName` : Page name.
— `Client, Inspector, Owner` : Boolean columns determining access.

	Column Name	Data Type	Allow Nulls
🔑	Page_ID	int	☐
	PageName	varchar(255)	☑
	Client	varchar(50)	☑
	Inspector	varchar(50)	☑
	Owner	varchar(50)	☑

FIGURE 14 – Structure of Table Pages

For instance :

— *PageName* = *Home* : Accessible by all roles (`Client = true, Inspector = true, Owner = true`).
— *PageName* = *Client2* : Accessible only by Clients (`Client = true, Inspector = false, Owner = false`).

In Power BI, we configured roles for **Owner**, **Client**, and **Inspector**, mapping them to the corresponding columns in the `Pages` table.

This ensures only the appropriate users can view specific pages, optimizing security and performance.

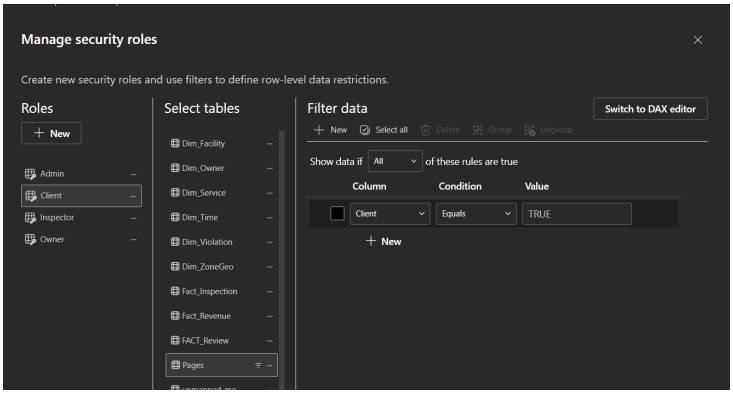


FIGURE 15 – User Management Roles in Power BI

By implementing role-based access control (RBAC), we ensure a personalized and secure experience for users, improving both security and dashboard performance.

5 Machine learning Models

5.1 Model Development

5.1.1 Natural Language Processing (NLP) Model :

Sentiment analysis plays a crucial role in understanding and interpreting customer feedback. By categorizing sentiments into positive, negative, businesses can gain valuable insights into customer opinions and behaviors. This section outlines the development and integration of an NLP model that performs sentiment analysis on customer reviews to provide actionable insights for improving products and services.

Data Collection : The dataset consisted of 26,767 reviews collected from the restaurant's online feedback system. Each review included textual feedback and a star rating.

Data Preprocessing : Reviews were preprocessed through several steps :

1-Text Cleaning : Special characters, punctuation, and stopwords were removed from the reviews to focus on the most relevant information. Negation words (e.g., not, never) were retained to preserve sentiment nuances.

2-Tokenization : The text was broken down into individual tokens (words), allowing for easier analysis.

3-Vectorization : The cleaned reviews were converted into numerical representations using a Term Frequency - Inverse Document Frequency (TF-IDF) vectorizer.

FIGURE 16 – Commentaire Positive

FIGURE 17 – Commentaire Negative

Model Development :

To build an accurate sentiment analysis model, several machine learning models were evaluated.

The models tested include :

— Logistic Regression : A statistical model used for binary classification tasks, well-suited for problems where the output is one of two possible classes (positive or negative).

— Random Forest Classifier : An ensemble learning method that uses multiple decision trees to improve prediction accuracy.

— Support Vector Classifier (SVC) : A robust algorithm that finds the optimal hyperplane to separate data into classes, known for its ability to handle high-dimensional data.

— KNN : A non-parametric algorithm that classifies a new point based on the majority class of its neighbors.

— Multinomial Naive Bayes (NB) : A probabilistic model used for classification tasks, especially useful for text classification problems.

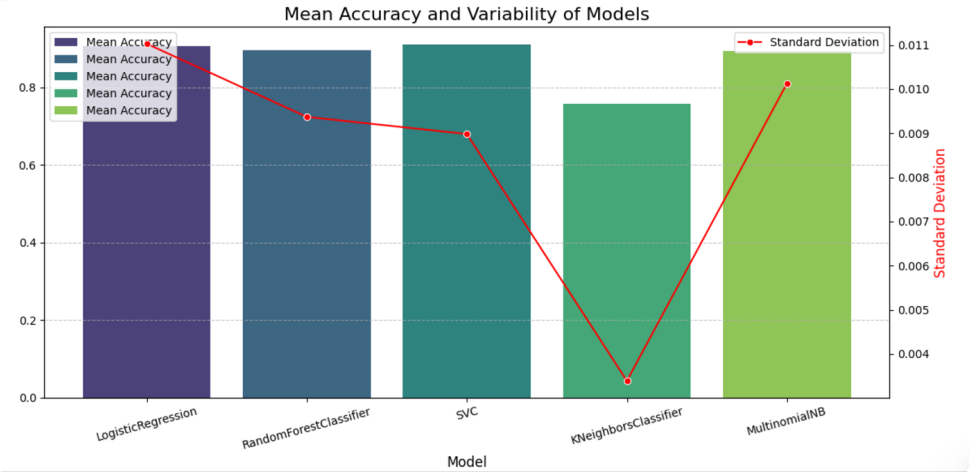


FIGURE 18 – Évaluation des performances des modèles de traitement du langage naturel (NLP).

Each model was trained on the same dataset, and their accuracy was evaluated using cross-validation. After comparing the results, SVC was selected as the best-performing model based on its accuracy in predicting customer sentiments.

5.1.2 Classification Model 1 : Predicting Establishment Risk

Predicting closure risk is vital for proactive management. This section outlines the development of a machine learning model to predict whether establishments are "Safe" or "At Risk."

Data Collection :

The dataset includes 313,675 health inspections with key variables :

— **Violation-Diversity :** Diversity of recorded violations.

— **RISK :** Assigned risk level (0 : High, 1 : Low, 2 : Moderate).

— **Score :** Performance evaluation during inspections.

— **Facility-Type :** Type of establishment (e.g., Restaurant, Store).

Data Pre-processing :

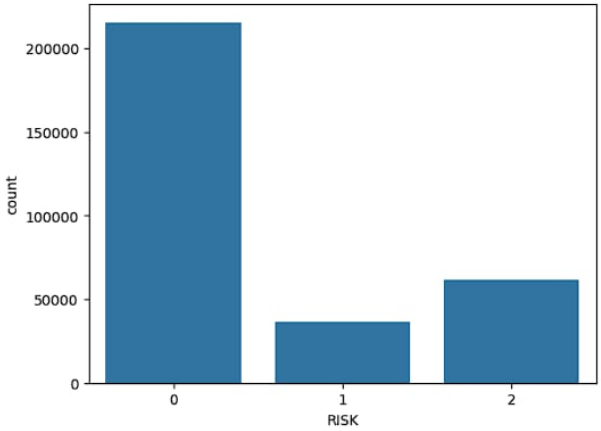


FIGURE 19 – Before Downsampling

Steps to prepare data for model training :

— **Data Cleaning :** Removal of missing values, duplicates, and outliers.

— **Encoding Categorical Data :** Numeric encoding for categorical columns like `Facility-Type`.

— **Normalization :** Standard scaling for numerical variables.

— **Downsampling :** Reduced the number of "Safe" cases to balance the dataset, as "At Risk" cases were fewer.

Model Development :

Machine learning algorithms tested :

— **Random Forest,SVM, SVC, KNN.**

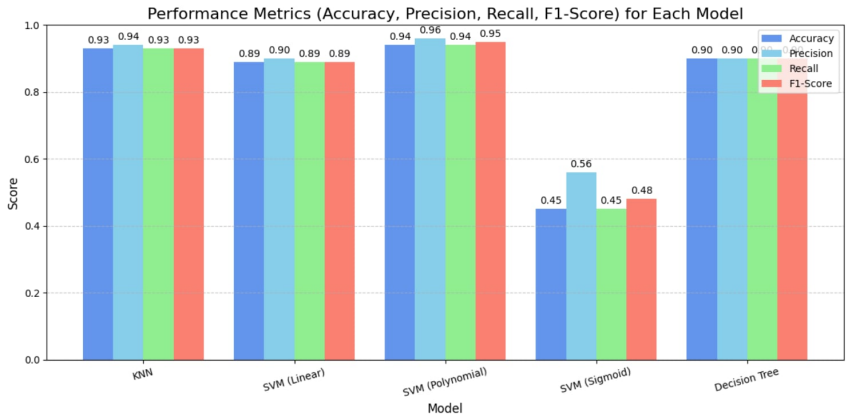


FIGURE 20 – Comparison Between Models

The SVM with polynomial kernel achieved the highest accuracy of **94%**, optimized using GridSearchCV.

Model Evaluation :

Model evaluation metrics :

— **Accuracy, Recall, F1-Score, Confusion Matrix.**

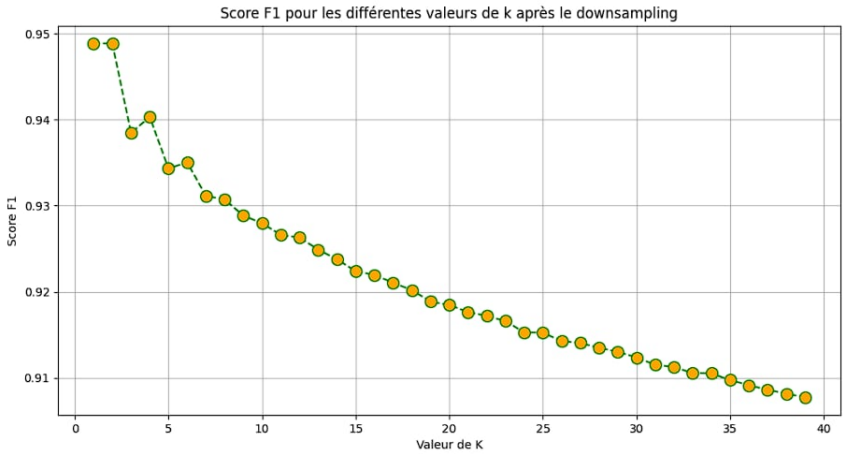


FIGURE 21 – F1 Score

5.1.3 Classification Model 2 : Predicting Restaurant Grades

This model predicts restaurant grades based on inspection and environmental data from Nevada, California, and New York.

Data Collection :

We gathered data on restaurant inspections, along with state-level environmental factors :

- **Poverty Rate :** Percentage of the population below the poverty line.
- **Water Quality :** Rating of water quality per state.
- **Air Quality :** Rating of air quality per state.

Data Pre-processing :

The dataset was cleaned and merged, including :

- **Facility Number of Violations :** Total violations recorded.
- **State Information :** Including poverty rate, water, and air quality.

We simplified the target variable by merging grades B and C into a single class and encoding :

- **A grade :** 0
- **B/C grade :** 1

Data Balancing :

Undersampling was applied to balance the dataset due to fewer B/C grades.

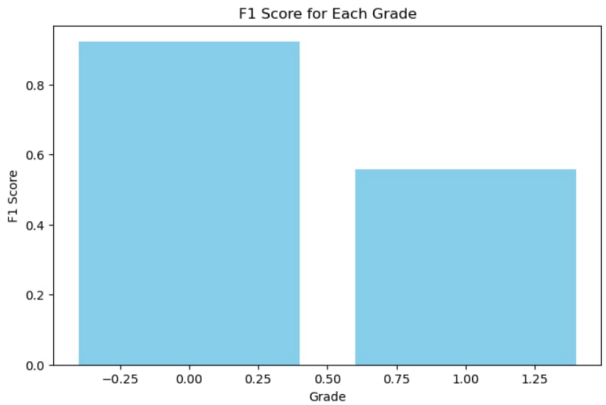


FIGURE 22 – F1 Score

Model Development :

We tested the following models :

— **KNN :** Optimized with k = 25.

— **SVM :** A robust classifier.

— **Polynomial SVM :** For non-linear relationships.

The KNN model achieved **91% accuracy** using F1-score as the performance metric.

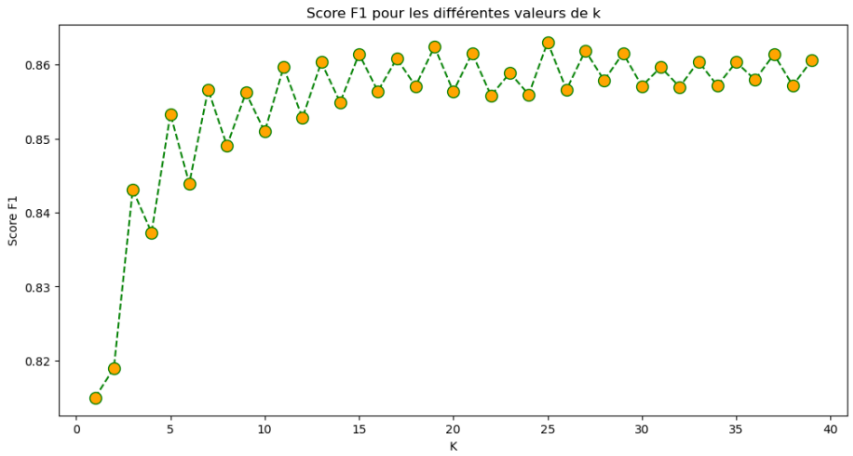


FIGURE 23 – F1 Score Best K

Evaluation :

The model was evaluated using :

— **Accuracy :** Correct predictions for grades.

— **F1-Score :** A balanced performance metric.

The model demonstrated high accuracy in predicting restaurant grades based on the available features.

Next Steps :

This classification model was later integrated into the website, allowing inspectors to prioritize facilities for inspection. By predicting restaurant grades based on inspection data and environmental factors, the model helps identify high-risk establishments, enabling inspectors to focus on those that may require more urgent attention. This ensures a more efficient allocation of resources and improves the overall effectiveness of the inspection process.

5.1.4 Clustering Model

The clustering model plays a critical role in predicting which facilities (restaurants and markets) are most likely to close soon based on their inspection scores, capacity, type, and health risk levels (high, moderate, low). This model focuses on active facilities (program status = active) and aims to identify those that might transition to inactive status due to poor performance metrics.

Objective :

The primary objective of this model is to identify facilities that are likely to become inactive due to poor performance based on inspection scores, capacity, facility type, and health risk levels.

Data Pre-processing : The dataset initially contained missing values and inconsistencies. The following cleaning steps were performed :

- **Handling Missing Values** :
  - Missing values in key columns such as score, risk, facility type, and capacity were addressed using logical rules.
  - Rows with missing target features were dropped.
- **Regex-Based Feature Extraction** :
  - Regex patterns were used to extract facility type, capacity, and risk from the 'pe-description' column.

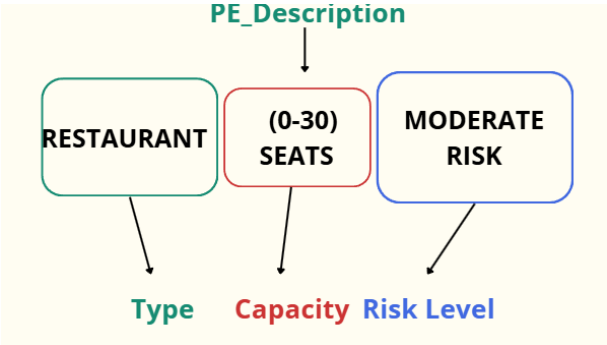


FIGURE 24 – How Risk, Capacity and Type were extracted

- **Removed Irrelevant Columns** :
  - Columns unrelated to the clustering process were dropped.
- **Verified Data Types** :
  - Ensured all features had appropriate data types (e.g., score as a float).

Feature Transformation and Encoding :

- **Label Encoding** : Categorical features (risk, facility type) were encoded using Label Encoder.
- **Feature Scaling** : Numerical features (score) were scaled using StandardScaler to ensure fair weighting in clustering.

Clustering Algorithm :

Algorithm Used : K-Means Clustering K-Means was chosen for its efficiency in grouping data points based on similarity. The silhouette score was used to evaluate the quality of clusters. The optimal number of clusters was determined using the Elbow Method. The model utilized **2 clusters**, providing a clear distinction between different groups of facilities based on their characteristics.

Exploratory Data Analysis : To gain further insights into the data before clustering, various visualization techniques were employed :

- **K-Means Visualization after PCA** :
  - K-Means analysis divided the data into two distinct clusters :
    - **Cluster 0 (purple)** : Concentrated primarily around negative and near-zero scores after PCA dimensionality reduction, indicating poor performance.
    - **Cluster 1 (yellow)** : Composed of points with positive scores, indicating a better overall profile based on principal components.

FIGURE 25 – Visualisation

— **Dendrogram (Hierarchical Clustering)** :
  — Provides a hierarchical visualization of the clustering process. Each merge represents a group of data points being clustered together at a specific distance threshold.
  — This helps in deciding the optimal number of clusters for K-Means by analyzing the largest vertical gaps between merges.

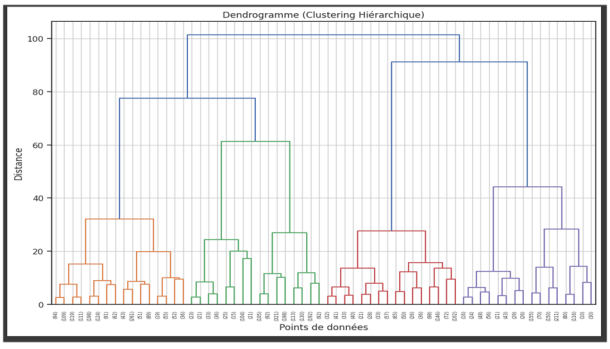


FIGURE 26 – Dendogram Visualization

— **Score Distribution** :
  — Boxplots showed :
    — Cluster 0 had lower median scores, with wider variability and several outliers.
    — Cluster 1 had higher median scores, with a narrower spread and fewer outliers.
  — Facilities in **Cluster 1** generally perform better than those in **Cluster 0**.
— **Risk Distribution** :
  — Facilities in **Cluster 0** exhibited lower risk values with consistent distribution.
  — Facilities in **Cluster 1** had higher risk values with greater variability.
  — **Cluster 1** contains facilities requiring closer inspection due to higher risks.
— **Facility Type Distribution** :
  — The distribution of facility types was fairly similar across both clusters, indicating that facility type does not strongly differentiate them.

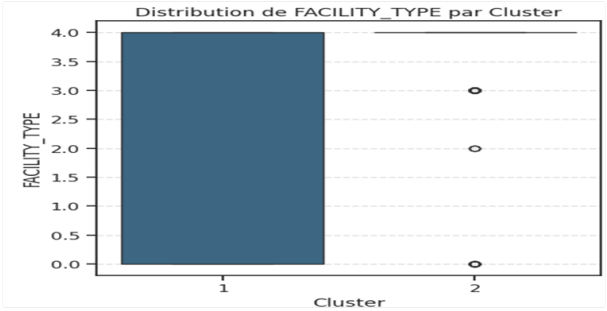


FIGURE 27 – Distribution

— **Capacity Distribution** :
  — Facilities in **Cluster 0** tended to have lower capacities (smaller establishments), while **Cluster 1** contained larger facilities with more variability in size.

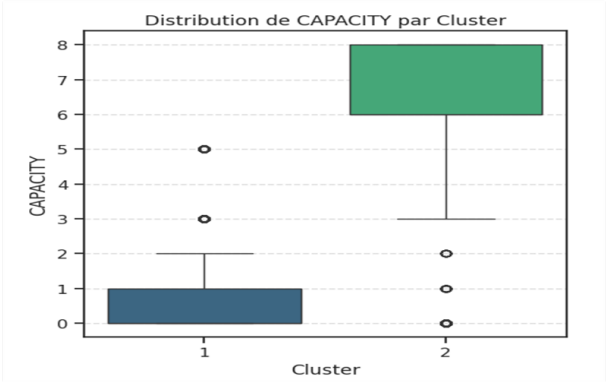


FIGURE 28 – Capacity Distribution

Cluster Profiles

— **Cluster 0** :
  — Lower performance scores and greater variability ; generally smaller establishments with consistent low-risk levels.
  — Represents smaller, low-risk facilities needing support to improve their inspection scores.

— **Cluster 1** :
  — Higher performance scores with less variability ; larger establishments exhibiting higher and more variable risks.
  — Represents larger, higher-risk facilities that perform better overall but require targeted interventions to reduce risk variability.

Conclusion : The clustering analysis reveals distinct profiles between smaller, low-risk facilities needing support to improve their performance, and larger, higher-risk facilities that perform better overall but require enhanced management strategies to mitigate risk variability.

5.1.5 Regression Model

Predicting restaurant revenue is essential for supporting restaurant owners in making informed business decisions, optimizing resource allocation, and improving financial per-

formance. This section outlines the development and integration of a machine learning model designed to predict restaurant revenue.

Data Collection The dataset used in this study was extracted from a data warehouse containing information about restaurant operations and reservations. Key features included :

- **Seating_Capacity :** The number of seats available in the restaurant.
- **Average_Meal_Price :** The average cost of a meal at the restaurant.
- **Marketing_Budget :** The allocated budget for marketing activities.
- **WeekendReservations :** The number of reservations made during weekends.
- **WeekdayReservations :** The number of reservations made during weekdays.
- **Revenue :** The target variable, representing the restaurant's revenue.

These variables were selected for their significant correlations with revenue, as identified during the exploratory data analysis phase.

Data Pre-processing Before training the model, the dataset underwent a series of pre-processing steps to ensure its quality and readiness for analysis :

- **Data Cleaning :** Missing values and duplicates were handled appropriately.
- **Standardization :** Numerical features, excluding the target variable **Revenue**, were standardized using a scaler to ensure uniform feature scales.

Exploratory Data Analysis (EDA) The distribution of the target variable **Revenue** was examined. The revenue data exhibited a positively skewed distribution, resembling a normal distribution but with a longer tail toward higher values. This insight informed the choice of models, as some algorithms may require transformations to handle skewness effectively.

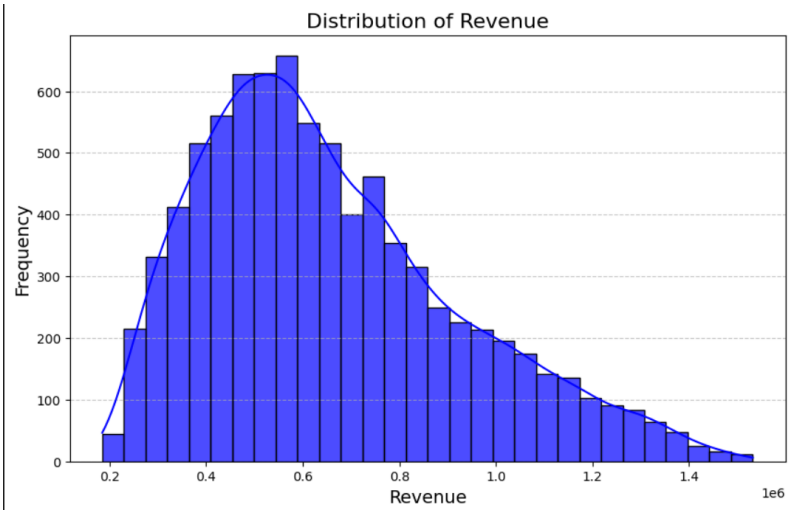


FIGURE 29 – Répartition des revenus des restaurants.

Model Development To identify the best model for predicting **Revenue**, three machine learning algorithms were compared :

- **Ridge Regression :** A linear regression model with regularization to prevent overfitting.

— **Random Forest Regressor :** An ensemble learning method using multiple decision trees to improve prediction accuracy.
— **Gradient Boosting Regressor :** A sequential ensemble method optimizing weak learners to minimize errors.

Hyperparameter Tuning **GridSearchCV** was employed to find the optimal hyperparameters for each model. This process involved evaluating combinations of parameters using cross-validation, ensuring robust performance estimates.

Model Evaluation Models were assessed based on two primary metrics :
— **R^2 Score :** Measures the proportion of variance in the target variable explained by the features. Higher values indicate better fit.
— **Root Mean Squared Error (RMSE) :** Represents the average magnitude of prediction errors. Lower values indicate better model performance.

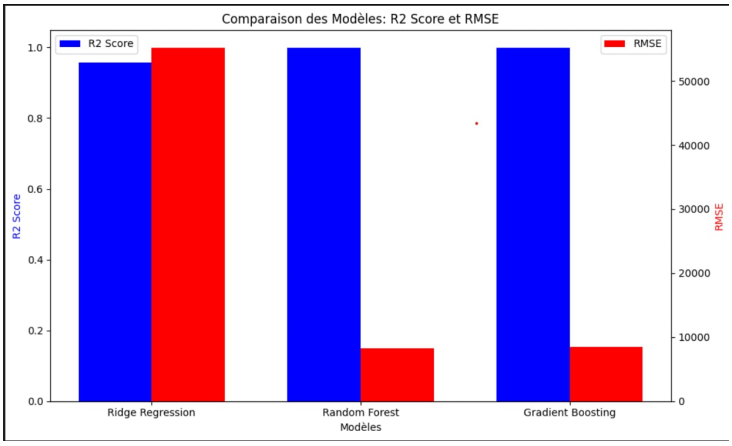


FIGURE 30 – Model Comparison

The **Gradient Boosting Regressor** achieved the best performance, demonstrating the highest R^2 score and the lowest **RMSE**. This model effectively captured complex, non-linear relationships in the data, making it the most suitable choice for predicting restaurant revenue.

5.2 Deployment and Monitoring

Website Overview

The website is a role-based platform designed to serve the specific needs of three user groups : **restaurant owners**, **inspectors**, and **clients**. Each user role is directed to a dedicated site area upon login, providing tailored functionalities and user experiences.

- Restaurant Owners
Restaurant owners can manage their facilities through a user-friendly interface.
— They have access to a Home page summarizing key insights and updates.
— A List Facilities page allows owners to view, add, update, or delete their facilities.

— Additionally, owners can leverage the integrated **Revenue Prediction Model** to estimate their restaurants' financial performance based on relevant business metrics.

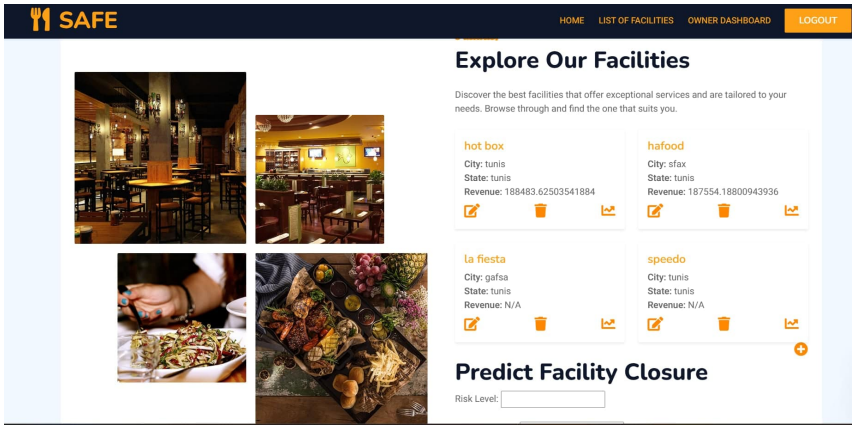


FIGURE 31 – Restaurant Owners Home Page

- Clients

Clients can explore and engage with facilities through intuitive pages.

— They can browse a **List Facilities** page to review and rate their experiences.

— Through the same page, clients can submit reviews for specific facilities.

— An **NLP Model** is integrated to analyze client feedback, categorizing comments as positive or negative to help facilities understand and address customer sentiment.

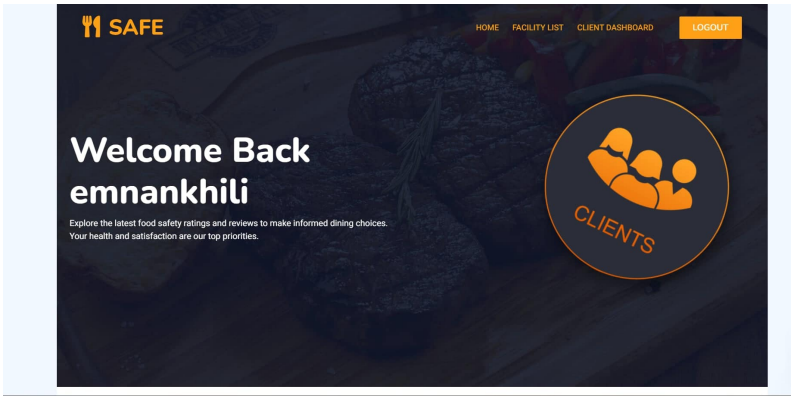


FIGURE 32 – Clients Home Page

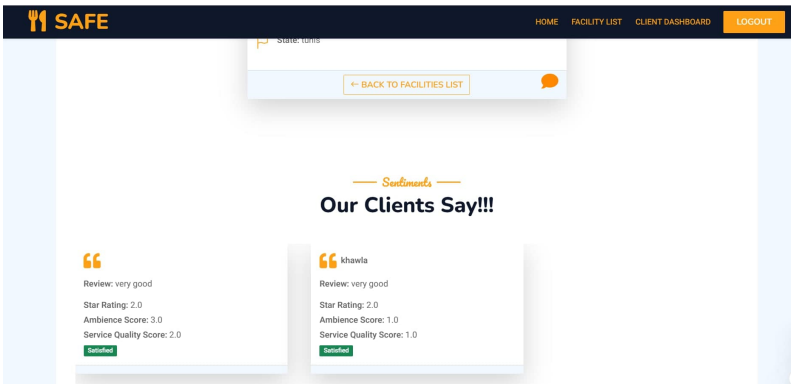


FIGURE 33 – Clients's Review

- Inspectors

Inspectors are equipped with tools to manage inspections effectively.

— They can add inspection records, specifying details such as the facility name.

— The platform integrates a **Classification Model** to predict the **Health Grade** (A, B, or C) of a facility during an inspection based on the entered data.

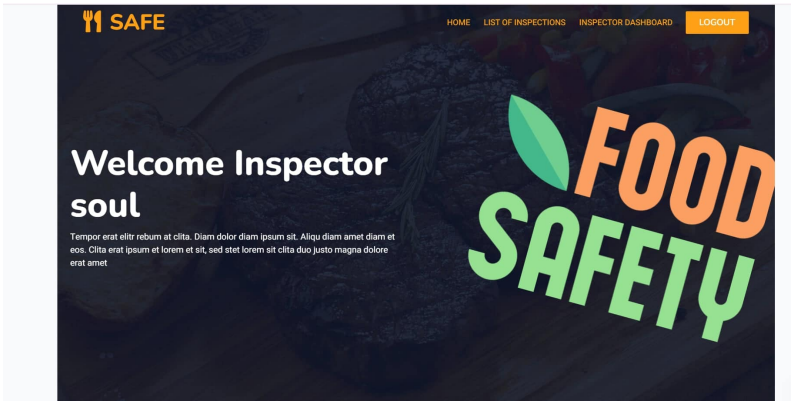


FIGURE 34 – Inspectors Home Page

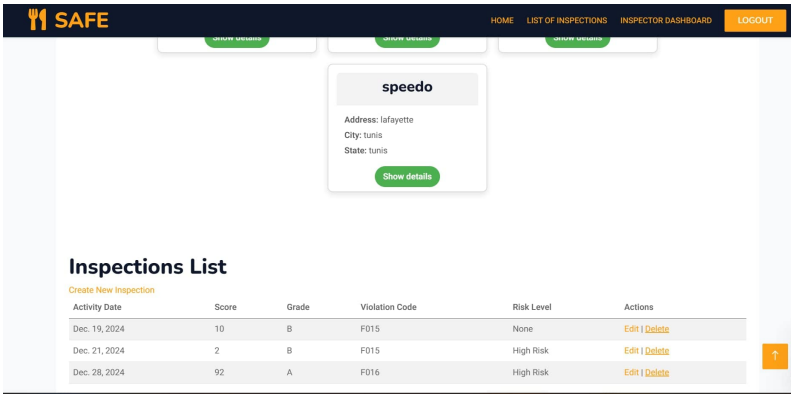


FIGURE 35 – List Inspections

Deployment and Integration

The deployment phase involved integrating the developed machine learning models into different platforms to meet the specific objectives of our project.

Website Integration :

Models such as classification, NLP, and regression were seamlessly incorporated into our web application, which is designed to support restaurant owners, inspectors, and clients. The deployment process involved saving the trained models as .pkl files and utilizing Django to integrate them into the website backend. The website allows restaurant owners to predict revenue and analyze operational metrics through user-friendly forms and dashboards. Inspectors and clients can also interact with insights derived from these models to make informed decisions.

Power BI Integration for Clustering Model :

The clustering model, aimed at identifying facilities at risk of closure based on inspection scores, capacity, type, and risk levels, was deployed in Power BI. This model supports the Inspector perspective by enabling visual analysis of facility profiles. The trained clustering model was saved as a .pkl file. Using Python visuals in Power BI, the cleaned features were loaded, and the model was executed within the Python script editor to generate cluster-based visualizations. These visualizations provide actionable insights for inspectors, such as distinguishing between low-performing, low-risk facilities and high-performing, high-risk facilities, helping prioritize interventions and resources.

6 Conclusion

Bridging Gaps in Food Safety Systems

This project has tackled key challenges in the restaurant inspection ecosystem by developing predictive solutions that benefit customers, restaurant owners, and health inspectors alike.

Comparative Analysis : Existing Tools vs. Our Solution

While platforms like *Yelp*, *Zomato*, and *UberEats* focus on customer reviews and ratings, and health department dashboards offer access to inspection reports, these solutions often lack integration. Additionally, business intelligence tools provide metrics for restaurant owners but fail to address predictive analytics for proactive management. None of these systems combine predictive capabilities with user-centric features for all stakeholders.

Our Innovation

Our platform goes beyond existing solutions by integrating machine learning models to :

- Predict restaurant grades based on inspection data and environmental factors.
- Identify high-risk establishments for proactive interventions.

This approach uniquely merges historical data with predictive analytics, empowering stakeholders with actionable insights.

Impact and Future Potential

By addressing gaps in current systems, our project demonstrates how data-driven insights can revolutionize food safety. Health inspectors can prioritize inspections efficiently, restaurant owners can track performance and risks in real time, and customers gain access to reliable, data-backed evaluations.

As a foundation for future advancements, our project paves the way for expanding datasets, refining algorithms, and integrating more comprehensive features into real-world applications.

"This project embodies our vision for a smarter, safer, and more transparent food safety ecosystem."

Next Plans

Expanding Data Sources and Enhancements

Our immediate goal is to extend the data used in this project. We plan to integrate additional datasets from more states and possibly local inspection reports to improve the accuracy and generalization of our models. This will allow us to fine-tune predictions and offer a more comprehensive analysis for stakeholders.

Model Optimization and Features

In the next phase, we will focus on optimizing our machine learning models by experimenting with advanced algorithms such as ensemble methods, boosting techniques, and deep learning models. We also plan to incorporate more granular data, such as customer sentiment from reviews, to enhance the accuracy of restaurant grade predictions and risk assessments.

User Interface and Experience Improvements

An integral part of the next steps involves refining the user interface of our platform. We aim to provide a seamless and intuitive experience for health inspectors, restaurant owners, and customers, with improved data visualization tools for inspection reports, grades, and predictive analytics.

Real-World Deployment

As we move forward, our long-term goal is to deploy this platform in a real-world setting. We plan to partner with regulatory agencies, restaurant chains, and local health departments to implement the system, allowing real-time data processing and prediction to optimize restaurant inspections and safety protocols.

Team InfoFlow : A Journey of Innovation

As Team **InfoFlow**, we are proud of the journey we have undertaken in this project. From its inception to the final implementation, our collaborative efforts and passion for innovation have shaped the development of this platform. We believe that this project reflects the essence of teamwork, creativity, and the pursuit of excellence. Moving forward, we remain committed to pushing the boundaries of technology and contributing to impactful solutions in the field of restaurant inspections and beyond.

FIGURE 36 – SAFE : Revolutionizing Food Safety Management

7 References

1. **Poverty Rate Dataset**
   Dias, M. (2021). *County-Level Poverty Dataset*. Kaggle. Retrieved from `https://www.kaggle.com/datasets/matheusdias1996/county-level-poverty`
2. **Air Quality and Water Quality Dataset**
   Socioeconomic Data and Applications Center. (2021). *Air Quality and Water Quality Dataset*. Kaggle. Retrieved from `https://www.kaggle.com/datasets/bigironsphere/socioeconomic-data-and-applications-center`
3. **Restaurant Reviews Dataset**
   The Devastator. (2022). *UberEats Restaurant Dataset (Over 100,000 US Restaurants)*. Kaggle. Retrieved from `https://www.kaggle.com/datasets/thedevastator/the-ubereats-restaurant-dataset-over-100000-us-r`
4. **Machine Learning Techniques**
   Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
   James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning : With Applications in R*. Springer.
   Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273–297.
   Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
5. **Data Preprocessing and Balancing Techniques**
   Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
   Chawla, N. V., Kegelmeyer, W. P., & Hall, L. O. (2002). SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
6. **U.S. Food Inspection Standards**
   FDA Food Code. (2017). *Food Code 2017*. U.S. Department of Health & Human Services.