Align Guard Internship

DATA ANALYTICS REPORT

Topic: Food Quality Assessment

Submitted By

Blessy Louis

Blessylouis2002@gmail.com

05th June 2024

Table Of Contents

INTRODUCTION

Ensuring the quality and safety of food served in restaurants is a critical responsibility of public health authorities. Foodborne illnesses can have serious consequences for consumers, including severe health complications and even fatalities. To mitigate these risks, the Food Inspection Department conducts regular inspections of various food facilities in the city. These inspections are aimed at enforcing compliance with health regulations and identifying potential hazards in food handling and preparation practices.

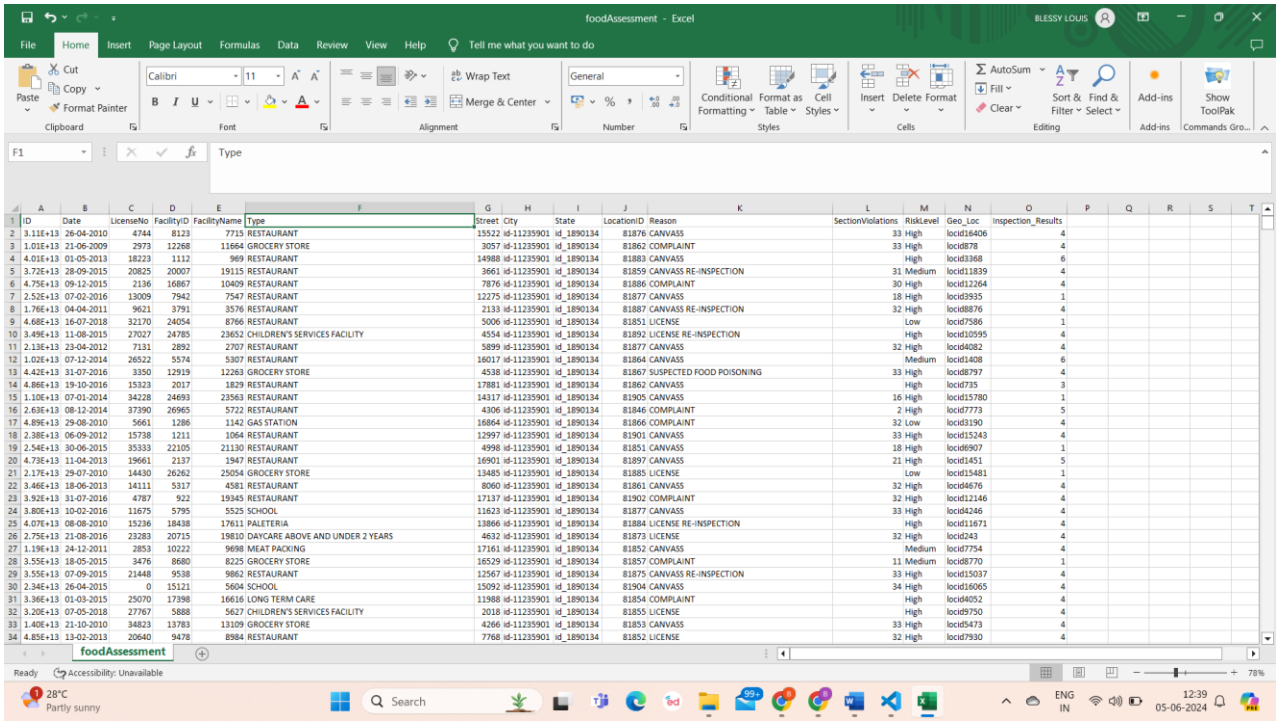
Over time, the department has accumulated a substantial amount of data from these inspections. This data includes information about the facilities inspected, the reasons for inspections, the types and frequencies of violations, and the outcomes of the inspections. By leveraging this historical data, it is possible to develop predictive models that can identify which facilities are more likely to fail future inspections.

AIM & OBJECTIVES

The primary aim of this analysis is to leverage historical food inspection data to develop predictive models that can identify facilities at high risk of failing future inspections. By accurately predicting inspection outcomes, the Food Inspection Department can optimize its inspection schedules and resource allocation, thereby improving food safety standards and protecting public health.

The objective of this analysis is to leverage historical food inspection data to develop predictive models that can accurately identify facilities at high risk of failing future inspections. This involves a comprehensive exploration and cleaning of the dataset, including handling missing values, duplicates, and data type inconsistencies, as well as removing irrelevant columns and outliers. Descriptive analysis will be conducted to generate summary statistics and visualizations to understand the distributions and correlations of key features such as risk levels, reasons for inspections, and inspection results. The analysis will include the preparation of data for modelling by encoding categorical variables and normalizing numerical features, followed by the development and evaluation of machine learning models to predict inspection outcomes. The ultimate goal is to provide actionable insights that will enable the Food Inspection Department to optimize inspection schedules and resource allocation, prioritize high-risk facilities, and enhance food safety, thereby protecting public health and ensuring efficient use of resources.

DATA DESCRIPTION



ID	Date	LicenseNo	FacilityID	FacilityName	Type	Street	City	State	LocationID	Reason	SectionViolations	RiskLevel	Geo_Loc	Inspection_Results
3.11E+13	26-04-2010	4744	8123	7715	RESTAURANT	15522	id-11235901	id_1890134	81876	CANVASS	33	High	locid16406	4
1.01E+13	21-06-2009	2973	12268	11664	GROCERY STORE	3057	id-11235901	id_1890134	81862	COMPLAINT	33	High	locid878	4
4.01E+13	01-05-2013	18223	1112	969	RESTAURANT	14988	id-11235901	id_1890134	81883	CANVASS		High	locid3368	6
3.72E+13	28-09-2015	20825	20007	19115	RESTAURANT	3661	id-11235901	id_1890134	81859	CANVASS RE-INSPECTION	31	Medium	locid11839	4
4.75E+13	09-12-2015	2136	16867	10409	RESTAURANT	7876	id-11235901	id_1890134	81886	COMPLAINT	30	High	locid12264	4
2.52E+13	07-02-2016	13009	7942	7547	RESTAURANT	12275	id-11235901	id_1890134	81877	CANVASS	18	High	locid3935	1
1.76E+13	04-04-2011	9621	3791	3576	RESTAURANT	2133	id-11235901	id_1890134	81887	CANVASS RE-INSPECTION	32	High	locid8876	4
4.68E+13	16-07-2018	32170	24054	8766	RESTAURANT	5006	id-11235901	id_1890134	81851	LICENSE		Low	locid7586	1
3.49E+13	11-08-2015	27027	24785	23652	CHILDREN'S SERVICES FACILITY	4554	id-11235901	id_1890134	81892	LICENSE RE-INSPECTION		High	locid10595	4
2.13E+13	23-04-2012	7131	2892	2707	RESTAURANT	5899	id-11235901	id_1890134	81877	CANVASS	32	High	locid4082	4
1.02E+13	07-12-2014	26522	5574	5307	RESTAURANT	16017	id-11235901	id_1890134	81864	CANVASS		Medium	locid1408	6
4.42E+13	31-07-2016	3350	12919	12263	GROCERY STORE	4538	id-11235901	id_1890134	81867	SUSPECTED FOOD POISONING	33	High	locid8797	4
4.86E+13	19-10-2016	15323	2017	1829	RESTAURANT	17881	id-11235901	id_1890134	81862	CANVASS		High	locid735	3
1.10E+13	07-01-2014	34228	24693	23563	RESTAURANT	14317	id-11235901	id_1890134	81905	CANVASS	16	High	locid15780	1
2.63E+13	08-12-2014	37390	26965	5722	RESTAURANT	4306	id-11235901	id_1890134	81846	COMPLAINT	2	High	locid7773	5
4.89E+13	29-08-2010	5661	1286	1142	GAS STATION	16864	id-11235901	id_1890134	81866	COMPLAINT	32	Low	locid3190	4
2.38E+13	06-09-2012	15738	1211	1064	RESTAURANT	12997	id-11235901	id_1890134	81901	CANVASS	33	High	locid15243	4
2.54E+13	30-06-2015	35333	22105	21130	RESTAURANT	4998	id-11235901	id_1890134	81851	CANVASS	18	High	locid6907	1
4.73E+13	11-04-2013	19661	2137	1947	RESTAURANT	16901	id-11235901	id_1890134	81897	CANVASS	21	High	locid1451	5
2.17E+13	29-07-2010	14430	26262	25054	GROCERY STORE	13485	id-11235901	id_1890134	81885	LICENSE		Low	locid15481	1
3.46E+13	18-06-2013	14111	5317	4581	RESTAURANT	8060	id-11235901	id_1890134	81861	CANVASS	32	High	locid4676	4
3.92E+13	31-07-2016	4787	922	19345	RESTAURANT	17137	id-11235901	id_1890134	81902	COMPLAINT	32	High	locid12146	4
3.80E+13	10-02-2016	11675	5795	5525	SCHOOL	11623	id-11235901	id_1890134	81877	CANVASS	33	High	locid4246	4
4.07E+13	08-08-2010	15236	18438	17611	PALETERIA	13866	id-11235901	id_1890134	81884	LICENSE RE-INSPECTION		High	locid11671	4
2.75E+13	21-08-2016	23283	20715	19810	DAYCARE ABOVE AND UNDER 2 YEARS	4632	id-11235901	id_1890134	81873	LICENSE	32	High	locid243	4
1.19E+13	24-12-2011	2853	10222	9698	MEAT PACKING	17161	id-11235901	id_1890134	81852	CANVASS		Medium	locid7754	4
3.55E+13	18-05-2015	3476	8680	8225	GROCERY STORE	16529	id-11235901	id_1890134	81857	COMPLAINT	11	Medium	locid8770	1
3.55E+13	07-09-2015	21448	9538	9862	RESTAURANT	12567	id-11235901	id_1890134	81875	CANVASS RE-INSPECTION	33	High	locid15037	4
2.34E+13	26-04-2015	0	15121	5604	SCHOOL	15092	id-11235901	id_1890134	81904	CANVASS	34	High	locid16065	4
3.36E+13	01-03-2015	25070	17398	16616	LONG TERM CARE	11988	id-11235901	id_1890134	81854	COMPLAINT		High	locid4052	4
3.20E+13	07-05-2018	27767	5888	5627	CHILDREN'S SERVICES FACILITY	2018	id-11235901	id_1890134	81855	LICENSE		High	locid9750	4
1.40E+13	21-10-2010	34823	13783	13109	GROCERY STORE	4266	id-11235901	id_1890134	81853	CANVASS	33	High	locid5473	4
4.85E+13	13-02-2013	20640	9478	8984	RESTAURANT	7768	id-11235901	id_1890134	81852	LICENSE	32	High	locid7930	4

The dataset used for this analysis comprises historical records of food inspections conducted by the Food Inspection Department. Each record corresponds to a specific inspection performed at a food facility within the city. Below is a detailed description of each feature in the dataset:

ID

The `ID` feature is a unique integer identifier assigned to each inspection record. It ensures that each inspection can be uniquely distinguished within the dataset, which is crucial for data management and analysis.

Date

The `Date` feature captures the date on which the inspection was conducted, stored in DateTime format. This allows for temporal analysis to identify trends and patterns over time, such as seasonal variations in inspection results or changes in compliance over the years.

LicenseNo

The `LicenseNo` feature is a de-identified license number assigned to each facility. This integer identifier helps track inspections of the same facility over time while preserving the confidentiality of the facilities.

FacilityID

The `FacilityID` is a unique de-identified identifier for each facility, represented as an integer. It ensures that each facility can be uniquely identified across different inspection records, facilitating longitudinal studies and trend analysis for individual facilities.

FacilityName

The `FacilityName` feature contains the encoded name of the facility being inspected. This string identifier is important for recognizing specific facilities, especially when combined with location-based features, to analyze inspection outcomes for individual establishments.

Type

The `Type` feature describes the type of facility being inspected, such as a restaurant, food truck, or grocery store. Represented as a string, this feature is useful for categorizing and comparing inspection results across different types of food establishments.

Street

The `Street` feature provides the encoded street address of the facility. This string information is essential for location-based analysis, helping to identify geographical trends in inspection results at a more granular level.

City

The `City` feature encodes the city where the facility is located. This string data is valuable for broader geographical analysis, allowing the identification of city-level trends and patterns in food safety compliance.

State

The `State` feature encodes the state where the facility is located, represented as a string. This feature is important for state-level regulatory compliance and trend analysis, providing insights into regional variations in food safety practices.

LocationID

The `LocationID` is an encoded integer feature representing the specific location of the facility. This enhances spatial analysis by providing a more precise location identifier, which can be used in conjunction with other geographical features.

Reason

The `Reason` feature captures the primary reason for conducting the inspection, such as routine inspections or follow-ups on complaints. Represented as a string, this feature is critical for understanding the context and triggers for inspections, which can influence the outcomes.

SectionViolations

The `SectionViolations` feature quantifies the number of sections of food safety laws violated by the facility. This integer value provides a numerical measure of compliance, indicating the extent of violations observed during the inspection.

RiskLevel

The `RiskLevel` feature indicates the level of risk posed to consumers by the facility, categorized as high, medium, low, or uncertain. This categorical variable is essential for assessing and prioritizing facilities based on the potential health risks they pose.

Geo_Loc

The `Geo_Loc` feature provides a de-identified geographical location of the facility, typically represented as a string. This feature aids in spatial analysis, allowing the mapping of inspection results to specific areas without compromising privacy.

Inspection_Results

The `Inspection_Results` feature records the outcome of the inspection, with possible values including 'FACILITY CHANGED', 'FAIL', 'FURTHER INSPECTION REQUIRED', 'INSPECTION OVERRULED', 'PASS', 'PASS (CONDITIONAL)', and 'SHUT-DOWN'. This categorical variable is the primary target for predictive analysis, helping to identify patterns and predictors of different inspection outcomes.

These features collectively provide a comprehensive dataset for analyzing food safety inspections, enabling the development of predictive models to identify high-risk facilities and optimize inspection schedules.

TARGET AUDIENCE/CUSTOMERS

The primary target audience for this analysis includes several key stakeholders involved in food safety and public health. They are as follows:

1. **Food Inspection Department**

   The immediate target audience is the Food Inspection Department itself. The insights derived from this analysis can help the department optimize their inspection schedules by identifying facilities that are more likely to fail inspections. This enables more efficient allocation of resources, focusing on high-risk establishments and improving overall food safety standards.

2. **Public Health Officials**

   Public health officials can utilize the findings from this analysis to develop strategies and policies aimed at reducing foodborne illnesses. By understanding the patterns and predictors of inspection outcomes, they can implement targeted interventions and educational programs for facilities that pose higher risks.

3. **Regulatory Agencies**

   Regulatory agencies at local, state, and federal levels are another critical audience. The analysis can inform regulatory updates and compliance enforcement strategies, ensuring that food safety laws are effectively implemented and monitored across different jurisdictions.

4. **Restaurant and Food Service Operators**

   Owners and operators of restaurants and food service establishments can benefit from this analysis by understanding common areas of non-compliance and taking proactive measures to improve their food safety practices. This can help them avoid penalties and ensure the safety of their customers.

5. **General Public**

   The general public, especially consumers, are an indirect audience. The ultimate goal of the Food Inspection Department is to protect public health, and by improving the effectiveness of food inspections, this analysis helps ensure that consumers have access to safe and healthy food options.

Each of these groups has a vested interest in the outcomes of food safety inspections, and this analysis provides actionable insights that can help them fulfill their respective roles in maintaining and enhancing food safety standards.

QUESTIONS TO BE ANSWERED

I. **What are the top facilities that are inspected frequently and what are the reasons for the same?**

II. **What are the primary reasons for an inspection to take place?**

III. **What are the Risk Levels involved in different types of Faculties?**

IV. **What are the Risks involved for various Section Violations?**

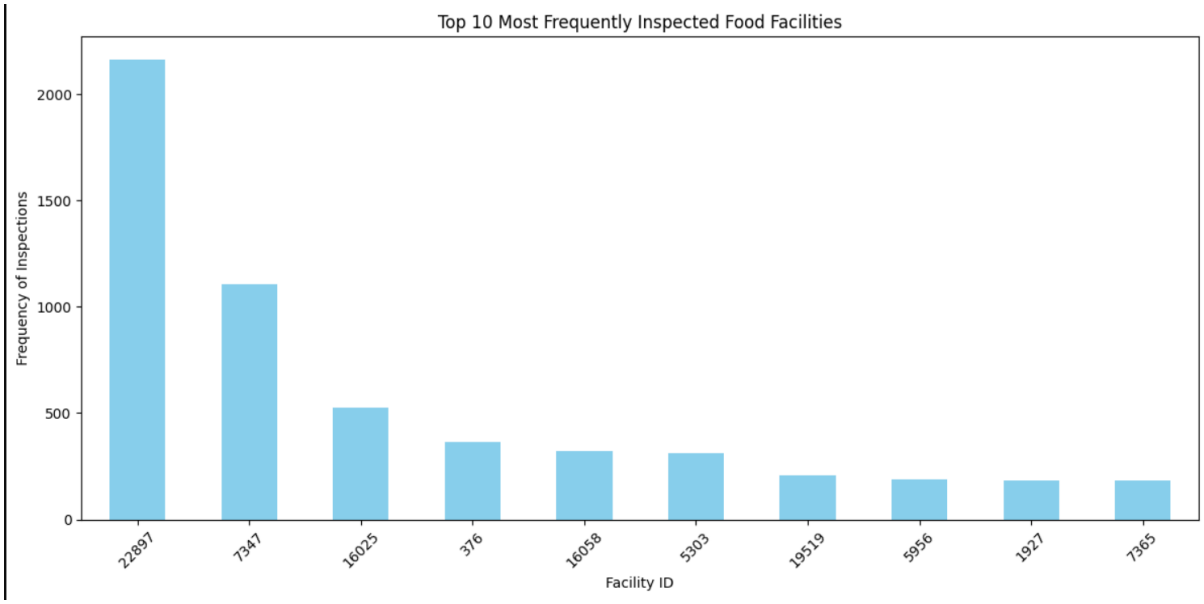
V. **What are the top most frequently observed violations?**

VI. **Is there a correlation between the number of violations and inspection outcome**

VII. **Are certain geographic areas or regions more prone to food safety violations?**

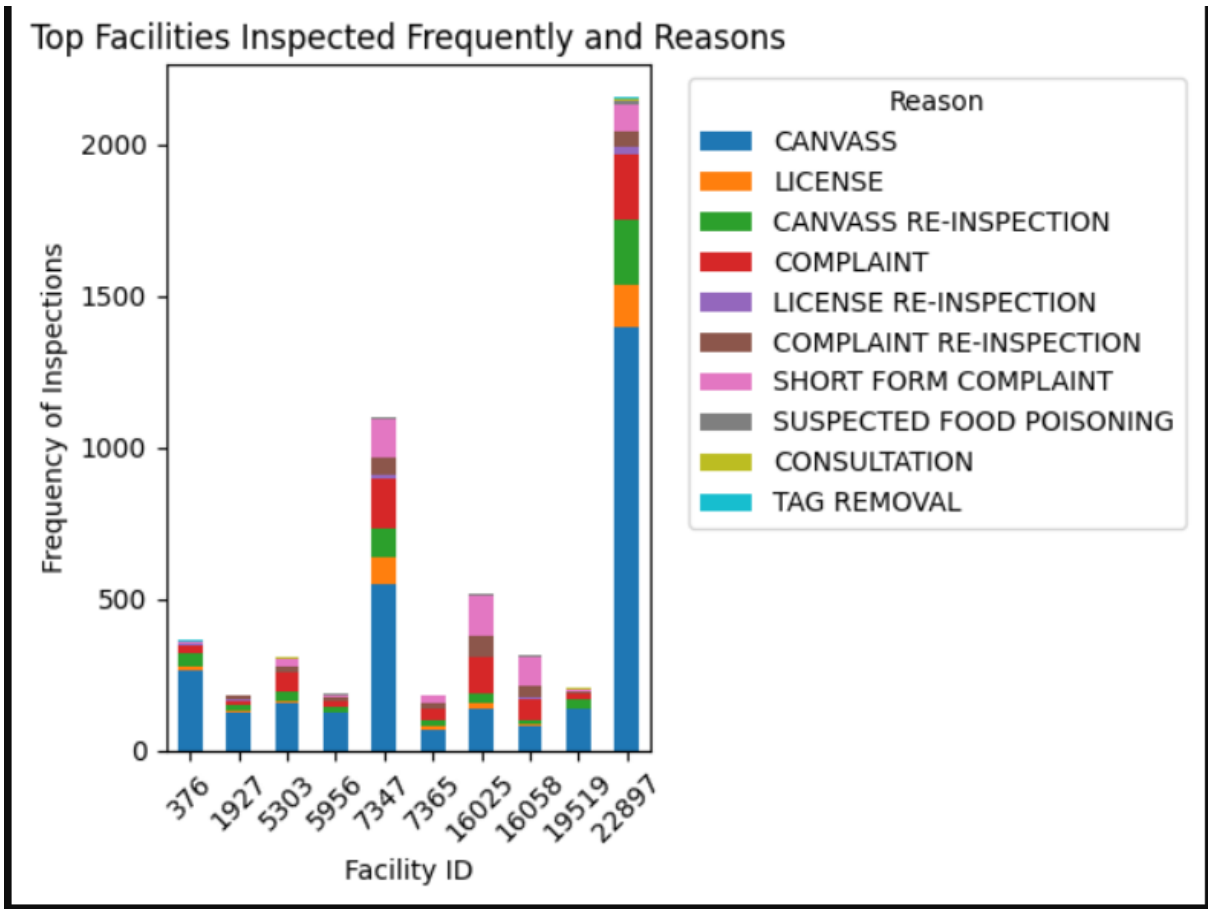
Let's find Solutions for each of these Questions

I. Solution:



This graph shows the top 10 most frequently Inspected Food Facilities, we clearly see the Facility with the ID **22897** has been inspected the greatest number of times i.e. a count of more the 2000 times over a couple of years, which followed by Facility with ID **7347** with a total count of 1100 plus times.

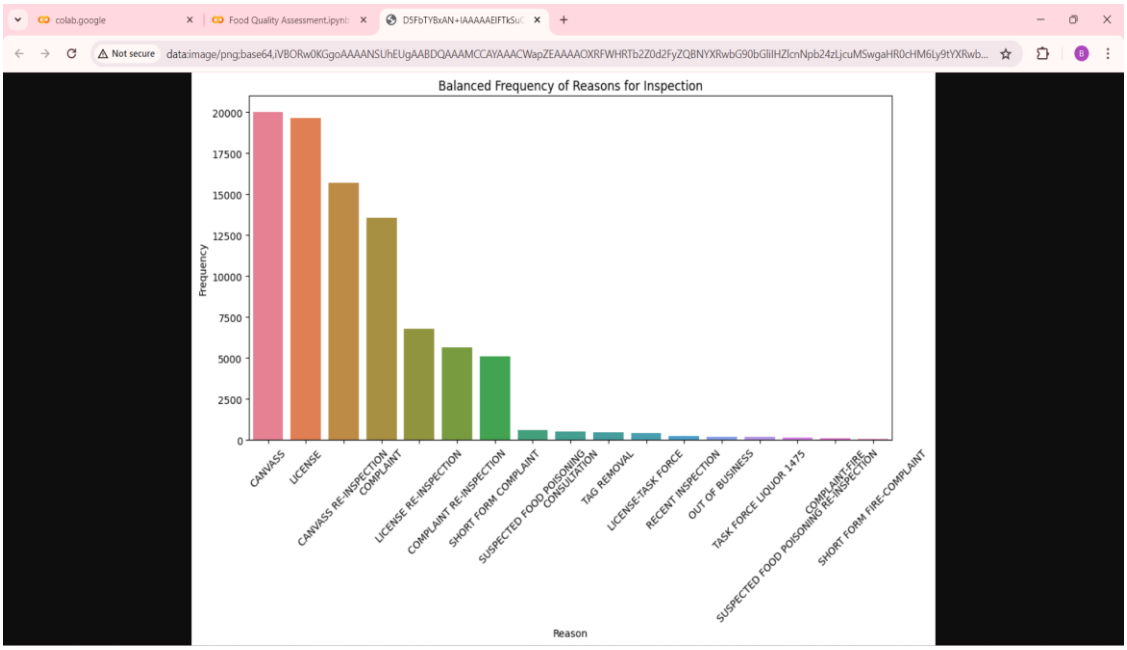
The question that arises here is that what would be the reason for this increased number of inspections in these facilities



The solution for the above question can be found here, we can see that the facility Id **22897,** that recorded the highest number of inspections, is mainly because of canvass**,** we see that almost 60% of the times the inspection is done as a part of the canvassing, this could be because the facility is located in a region where canvassing takes place very often. The next reason would be complaints, we see that at most 20 -30 % id covered by complaints, so this leads to a conclusion that, Inspections are also taking place in this facility because of the complaints received.

We also see in this graph that most of the inspections that is 70-80% of the inspections are conduction as a part of canvassing, which we can conclude that, Food Department officials are regularly monitoring the quality of food in different facilities.

II. Solution:

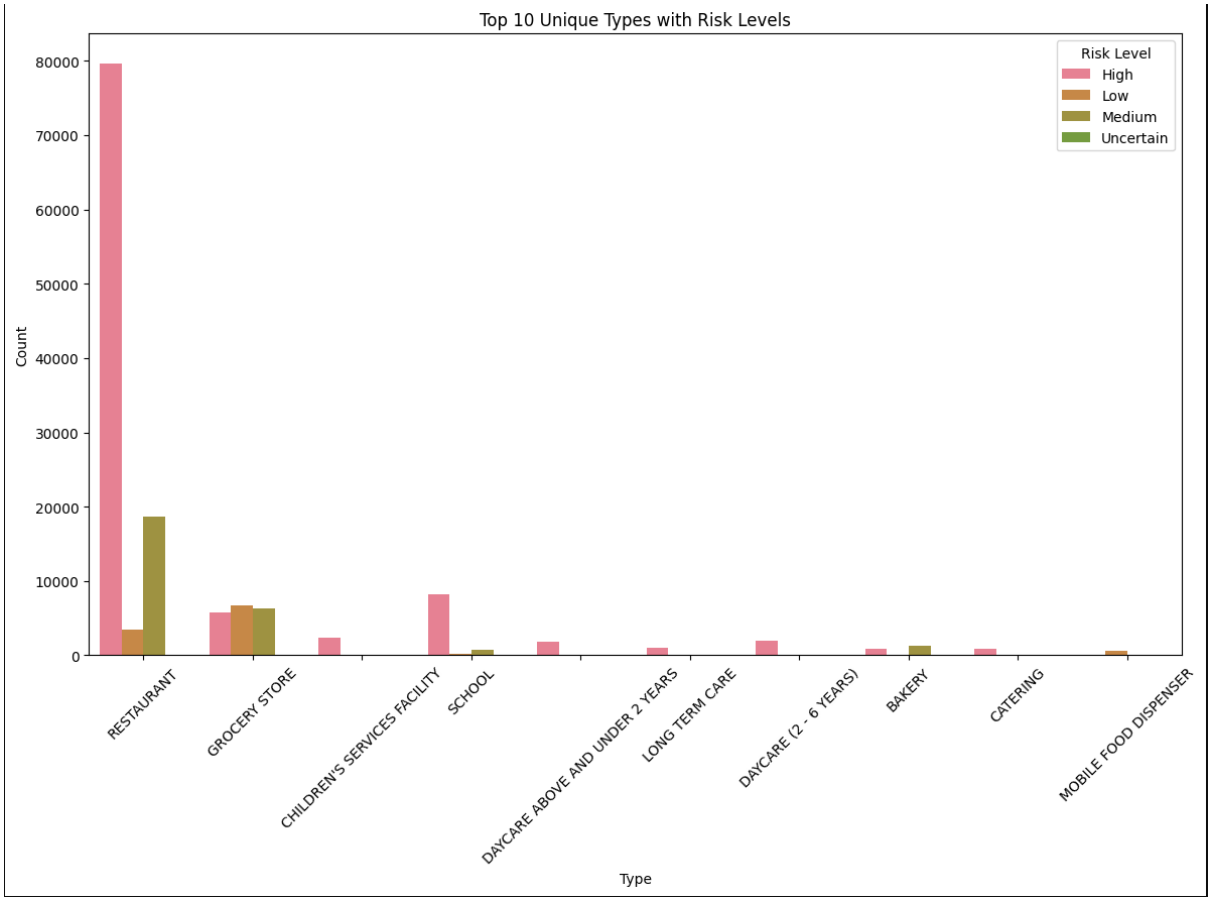


From the above graph, we see that the distribution of various reasons for inspection.

We see that Canvass has the most occurrences with a total occurrence of 19500, followed by License with a second highest occurrence of 18000, we also see that complaints received also occurring quite frequently in the dataset.

We can conclude that, one of the primary reasons for inspections to take place, are Canvassing, Licenses, Canvass Re-inspection, Complaints.

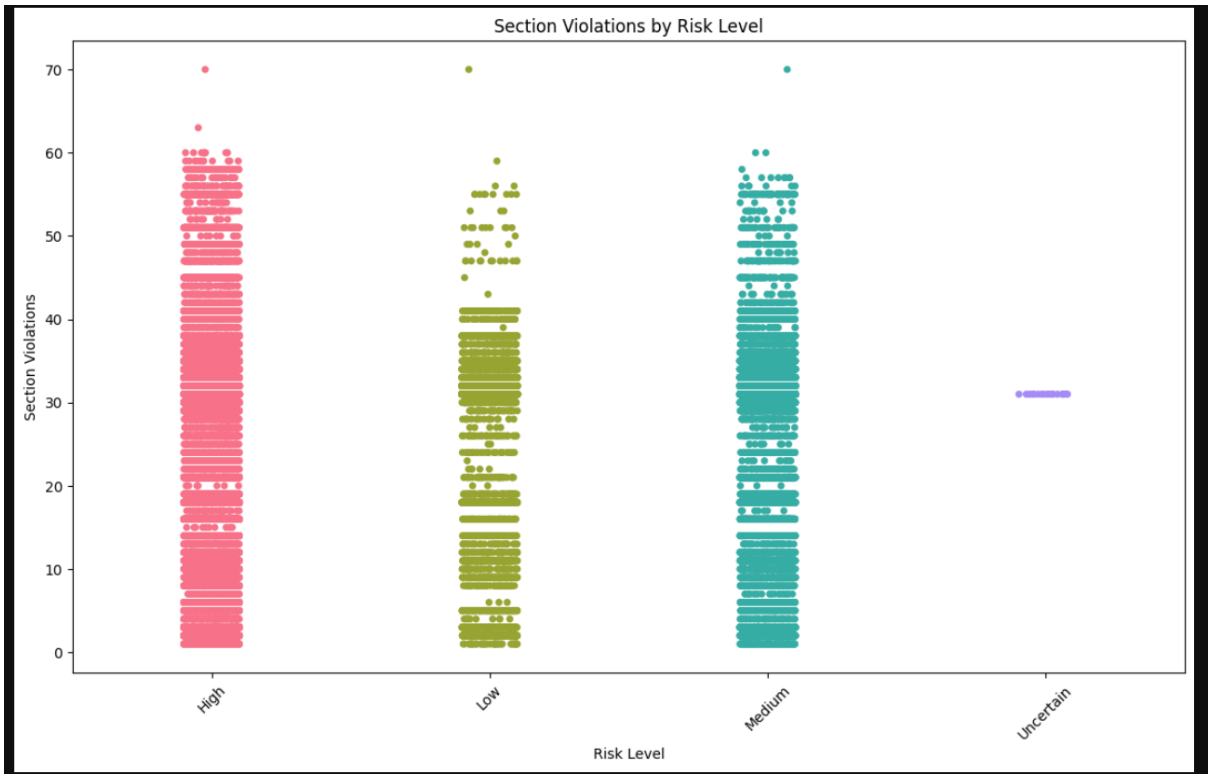
III. Solution



The above graph shows the Type – wise risk level, here we see that the restaurant type facilities are showing an extremely high risk, a small amount of medium risk as well. Now, this prediction could be expressed in this way because, the dataset is completely biased, with maximum data points having **Restaurants,** and also most of the inspection results are determined as **High Risk.** This bias could mainly be because , the food department officials are constantly inspecting Restaurants for a specific reason .

We can conclude from the graph that, in consideration of the bias involved, almost 60% of the Restaurants are said to be at High or medium risk. The second Highest occurrence of type of facility is **Grocery Store**, from the graph we can understand that, **grocery stores** are likely to be in either **low or medium risk.**

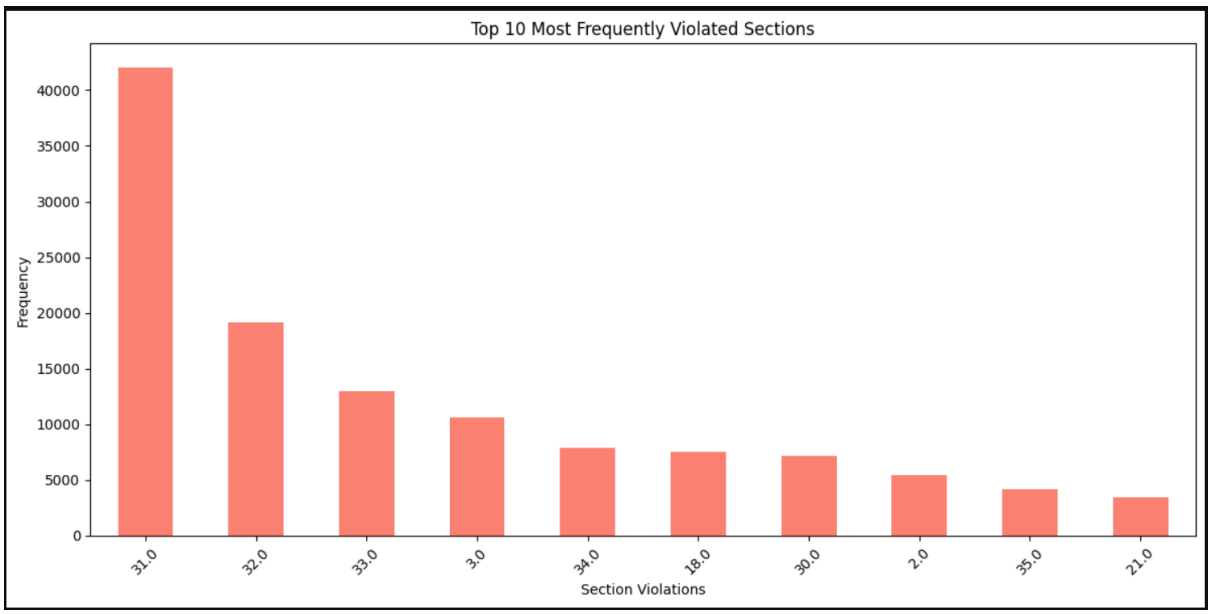
IV. Solution



The above graph depicts distribution of various sections violated, under each risk level. We can also put this into another words, like how much risk is involved in various Sections that are violated, that helps us to create an awareness among the public to ensure safety.

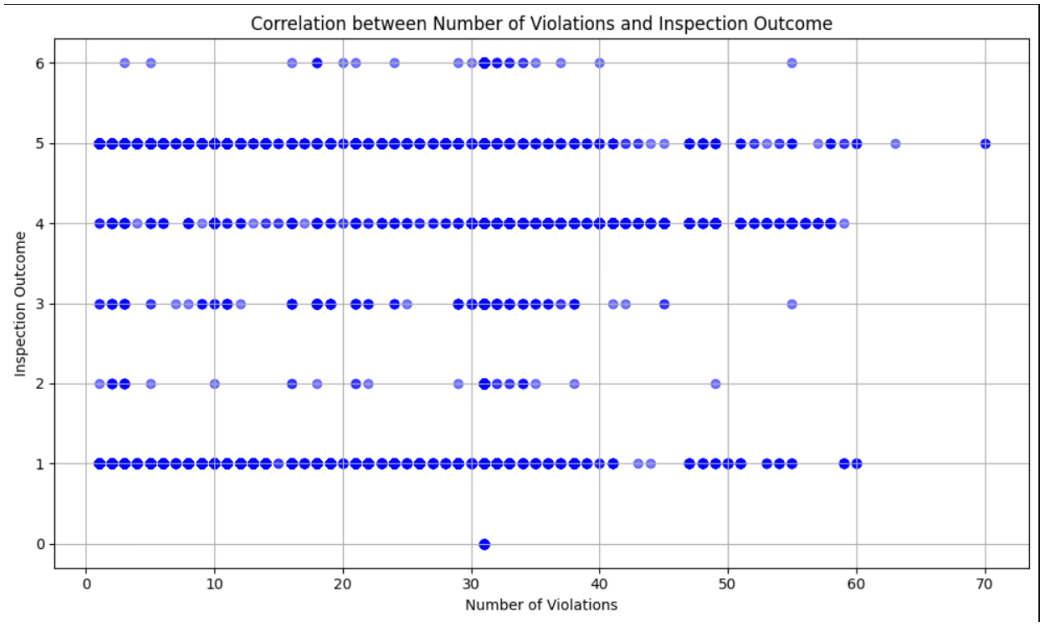
From this graph we see that, the Risk level is uncertain for Sections in 30, suggesting that the officials cannot determine the risk level if the section 30 alone is violated. We also see the if the sections between 1 to 45 are violated, we can say that, it is at a high risk, since these areas are densely populated with data points. In the case of medium, we see a similar pattern as High, but the variations can be determined, based on the number of sections violated. Similarly, we see a spread in the data points also suggesting a lenient pattern.

V. Solution



The above graph, helps us to understand the top 10 most frequently violated sections. We clearly see that Section number 31, has a highest occurrence with a total count of more than 40,000. The second highly violated section is section number 32, with a total violation of 18,000. These sections are ones that are densely populated in the previous graph, suggesting that these sections could cause a high risk when being violated.

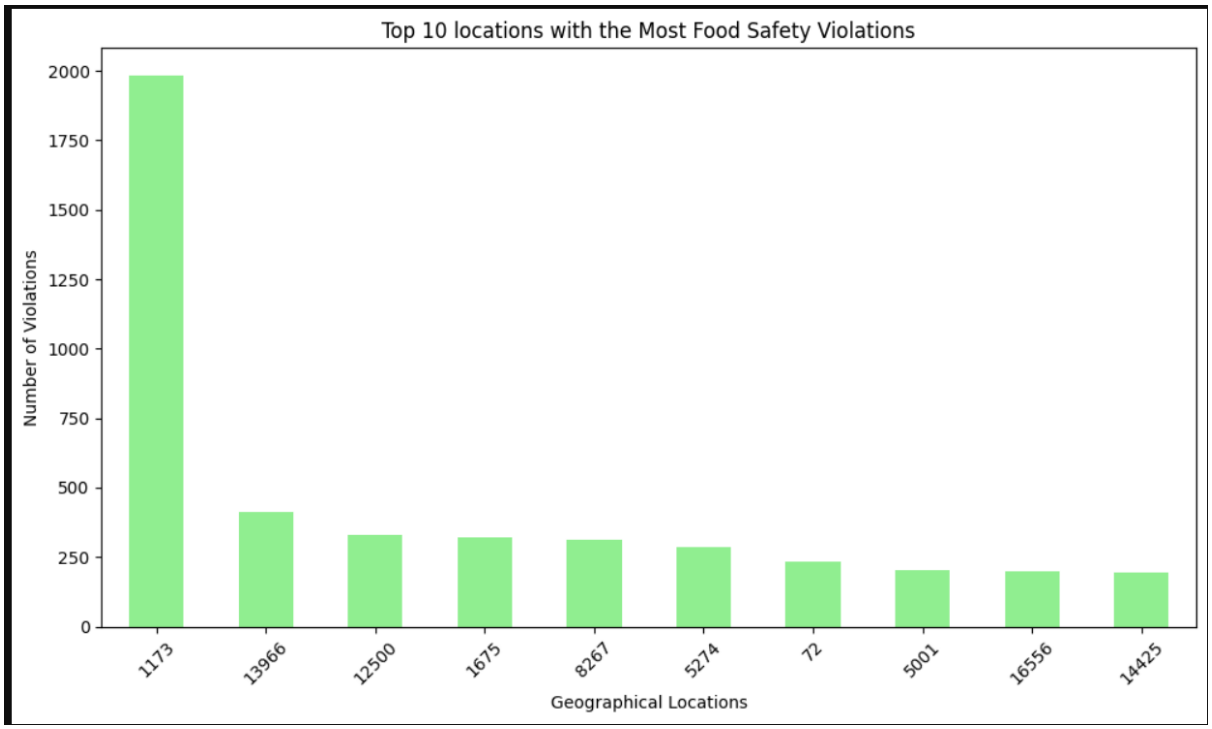
VI. Solution:



From the above graph we see that, there is some pattern of relation between the number of violations and the Inspection outcome. We see that the result 0 which implies “Facility

Changed", has only one violation that is section 31, now this could also be because of the presence of bias in the dataset. We see that the result outcomes 1,4,5 i.e. (Fail, Pass, Pass (Conditionally)) have a similar pattern, suggesting that the results may not be dependent on the number of violations made, and there could be other factors that influence the inspection outcome. Considering the bias in the dataset, we can conclude that, there is no possible relationship between the number of violations made to their corresponding results.

VII. Solution



This graph explains us the top 10 locations with most food safety violations. We can see that the street holding the ID number **1173**, has the highest number of safety violations made, whereas the other streets have a minimal and similar amount of safety violations, again suggesting that there could be some possible bias, because of which we have these results. Overall, we can say that at most 65% of the safety violations have been encountered in the street with ID 1173. This piece of information, can be helpful for the general public, to be cautious when buying food of any other form of eatables from these specified locations. This also suggests that the Food Ministry department has to focus on taking strict safety measures, in see regions that show higher violations of safety rules.

CONCLUSION

This comprehensive analysis of food safety inspection data offers valuable insights for multiple stakeholders involved in ensuring public health and regulatory compliance. Leveraging historical inspection data, the analysis aimed to predict high-risk facilities, prioritize inspections effectively, and enhance food safety standards. The dataset provided detailed information on various aspects of food inspections, including unique identifiers, inspection dates, facility attributes, violation details, risk levels, and inspection outcomes. Through exploratory data analysis (EDA) and predictive modelling, the analysis identified key factors influencing inspection outcomes, common violations, and geographical patterns of non-compliance.

The analysis revealed the top 10 most frequently inspected food facilities, providing insights into the reasons for inspection and highlighting areas requiring targeted interventions. Additionally, an examination of the top 10 most frequently violated sections pinpointed areas of non-compliance, emphasizing the need for improvements in compliance and enforcement efforts. The K-Nearest Neighbours (KNN) algorithm was employed for predictive modelling, demonstrating promising performance in accurately classifying inspection outcomes. However, further optimization and fine-tuning of the model may be necessary to enhance its performance and generalizability across diverse datasets and operational contexts.

Overall, this analysis serves as a foundation for informed decision-making aimed at enhancing food safety practices, protecting public health, and ensuring compliance with regulatory standards across diverse food establishments and geographic regions. Continued monitoring, enforcement, and collaboration among stakeholders are crucial for sustained improvements in food safety practices and regulatory compliance. Further research and development in predictive modelling techniques hold promise for advancing food safety practices and regulatory compliance in the future.