ARCHIVED:Restaurant Inspections Scores(2016-2019)-San Francisco

Project-4 Team-11
Sezer Bozoglan
Amy Hanks

March 20 2025

PROPOSAL

-Analyzing and Predicting Health Inspection Scores for Food Establishments in San Francisco
-Objective: to analyze and predict health inspection scores and risk categories for food establishments in San Francisco.
-We will build machine learning models that predict: Inspection Scores based on violations and other relevant data. Risk Categories of food establishments based on violation data.

Potential Models:

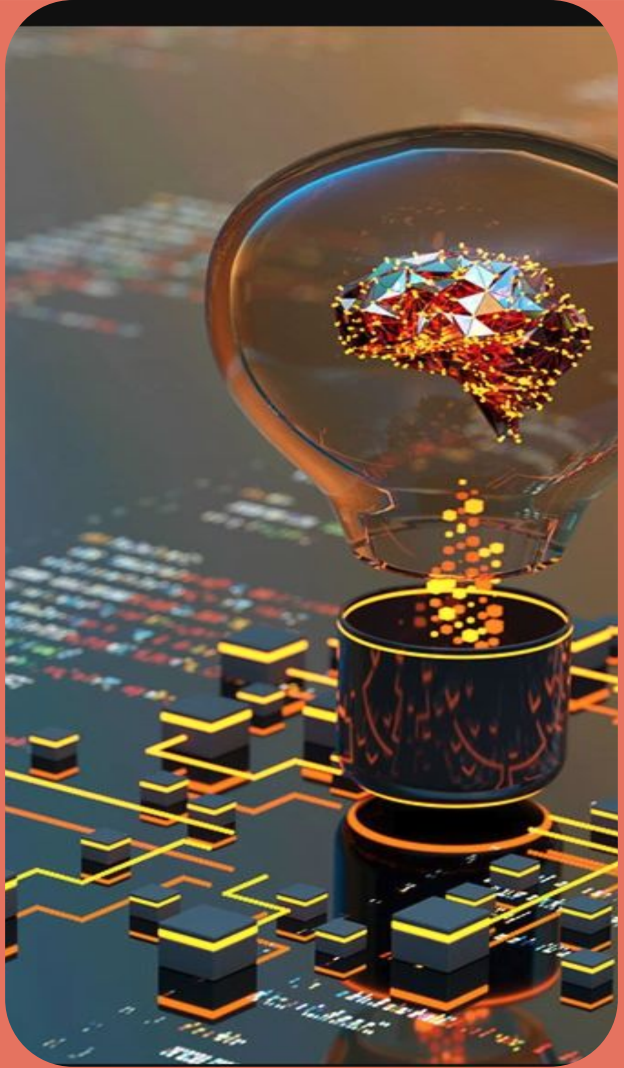
For predicting the inspection score: Regression models like Logistic Regression, Random Forest Classifier, and Support Vector Machines (SVM) will be used to predict the risk level of each establishment.
For classifying risk categories: Classification models

EXPECTED OUTCOMES

-Predictive models that can help businesses and health officials forecast inspection scores and risk categories.

-Data-driven insights that identify which violations are most likely to contribute to poor inspection scores or high-risk categories.

-A user-friendly interface that allows users to input violation data and receive predictions about the establishment's inspection score or risk category.



Release Timeline

WEEK 1 |

- ➔ Collect and clean the dataset using Spark.
- ➔ Perform initial data analysis to understand the distribution of inspection scores, types of violations, and risk categories.
- ➔ Start exploring patterns between violations and scores/risk categories.
- ➔ Divide the dataset into appropriate training and testing sets.

WEEK 2 |

- ➔ Implement and train regression models to predict inspection scores.
- ➔ Implement and train classification models to predict risk categories.
- ➔ Evaluate model performance using accuracy (for classification) and mean squared error (MSE) (for regression).
- ➔ Visualize model results using Matplotlib and Seaborn.

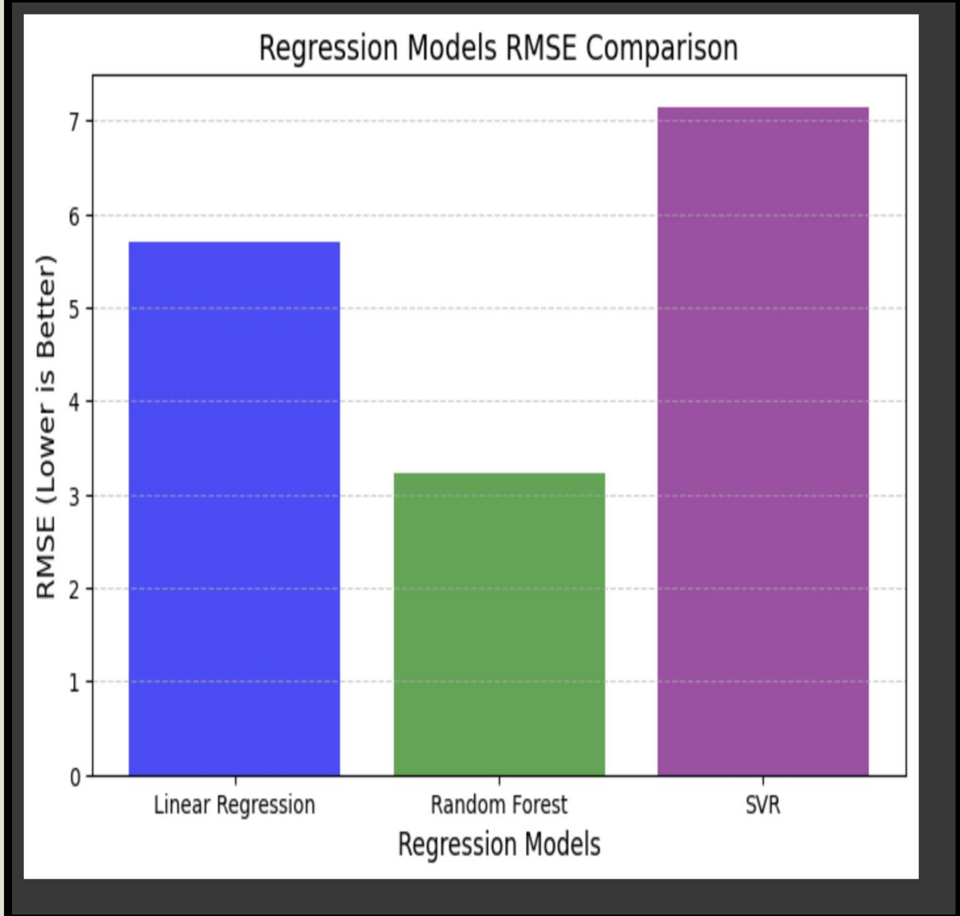
Final Deliverables |

- ➔ Model Documentation: Overview of models used, training, evaluation, and hyperparameters.
- ➔ Analysis Report: Insights into how violations and risk categories correlate with inspection scores.
- ➔ Presentation: A final presentation summarizing findings and models.

Linear Regression RMSE: **5.702782186389993**
Random Forest Regressor RMSE: **3.2317354955442164**
SVR RMSE: **7.140268092985499**

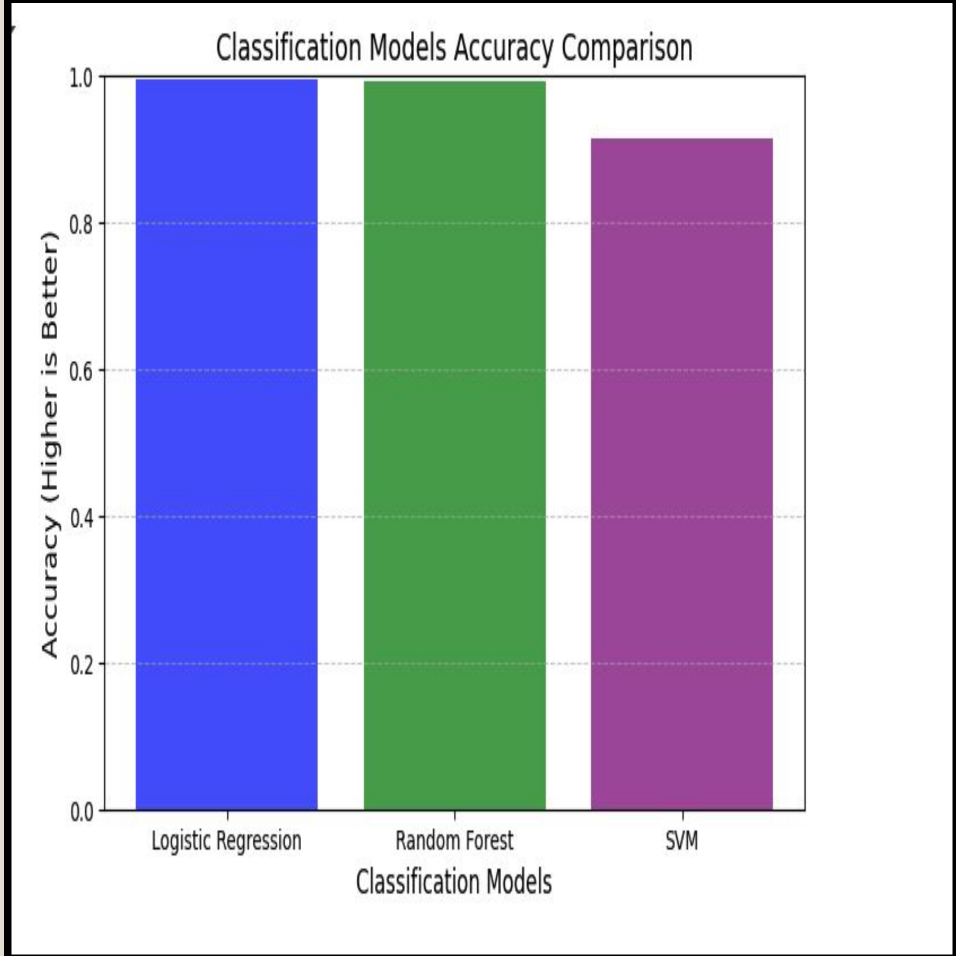
- The Random Forest model has the **lowest RMSE (~3.23)**, meaning it has the best prediction accuracy.
- **A lower RMSE** means the model's predictions are closer to the actual inspection scores.
- The Linear Regression model performs worse than Random Forest, with an **RMSE of 5.70.**
- **The SVR model** has the highest RMSE, indicating that it struggles to predict inspection scores accurately.

=RMSE (**Root Mean Squared Error**)



Logistic Regression Accuracy: **0.9942363112391931**
Random Forest Classifier Accuracy: **0.9923150816522575**
SVM Classifier Accuracy: **0.9125840537944284**

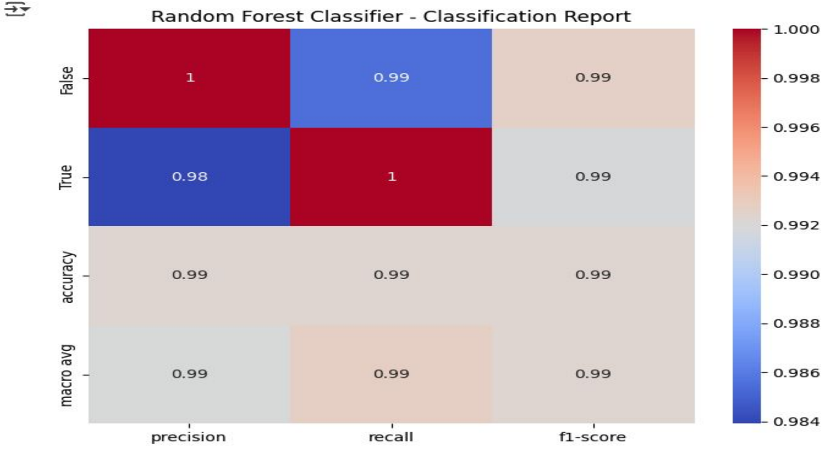
- Logistic Regression performs **slightly better than Random Forest**, achieving **99.42% accuracy**.
- The Random Forest model has **almost the same accuracy** as Logistic Regression.
- Since Random Forest is a **non-linear ensemble model**, it confirms that the features provide strong classification power.
- The **SVM model has noticeably lower accuracy** (~91.25%) compared to the other two.
- **Logistic Regression & Random Forest both perform exceptionally well (~99% accuracy).**
- **SVM lags behind at 91.25%, suggesting it may not be the best model for this dataset.**



- **Logistic Regression & Random Forest** performed **almost perfectly**, meaning they rarely misclassified risk categories.
- **SVM had a higher misclassification rate**, meaning it struggled more to differentiate between risk categories.
- **Precision and Recall are near 1.00 for Logistic Regression and Random Forest**, indicating **very few false positives and false negatives**.

- **Logistic Regression & Random Forest are the top-performing models.**
- **Random Forest's high precision and recall make it highly reliable for classifying risk category**.

```
Index(['business_id', 'business_name', 'business_address',
       'business_postal_code', 'business_latitude', 'business_longitude',
       'inspection_id', 'inspection_date', 'inspection_score',
       'inspection_type', 'violation_id', 'violation_description',
       'risk_category', 'date', 'time'],
      dtype='object')
Logistic Regression Accuracy: 0.9942363112391931
[[546   6]
 [  0 489]]
              precision    recall  f1-score   support

       False       1.00      0.99      0.99       552
        True       0.99      1.00      0.99       489

    accuracy                           0.99      1041
   macro avg       0.99      0.99      0.99      1041
weighted avg       0.99      0.99      0.99      1041

Random Forest Accuracy: 0.9923150816522575
[[544   8]
 [  0 489]]
              precision    recall  f1-score   support

       False       1.00      0.99      0.99       552
        True       0.98      1.00      0.99       489

    accuracy                           0.99      1041
   macro avg       0.99      0.99      0.99      1041
weighted avg       0.99      0.99      0.99      1041

SVM Accuracy: 0.9125840537944284
[[500  52]
 [ 39 450]]
              precision    recall  f1-score   support

       False       0.93      0.91      0.92       552
        True       0.90      0.92      0.91       489

    accuracy                           0.91      1041
   macro avg       0.91      0.91      0.91      1041
weighted avg       0.91      0.91      0.91      1041
```



544 cases were correctly classified as "False" (True Negatives).
489 cases were correctly classified as "True" (True Positives).
8 cases were False Positives (incorrectly classified as "True"). **0 False Negatives** (meaning the model never missed a "True" case).

- ❖ **Very high accuracy:** Nearly all predictions are correct.
- ❖ **Only 8 misclassifications (False Positives).**
- ❖ **No False Negatives**, meaning it never missed a "True" case.

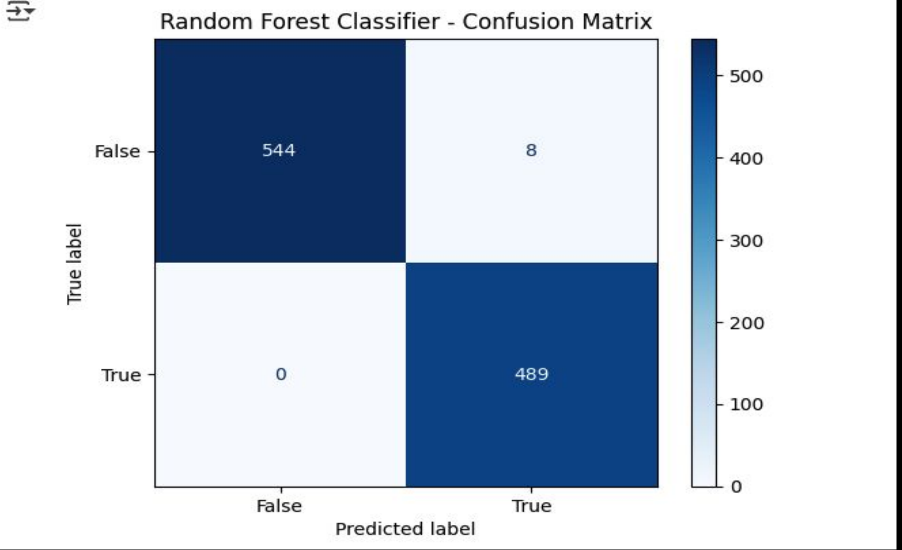
★ **Best model for accuracy and reliability.**

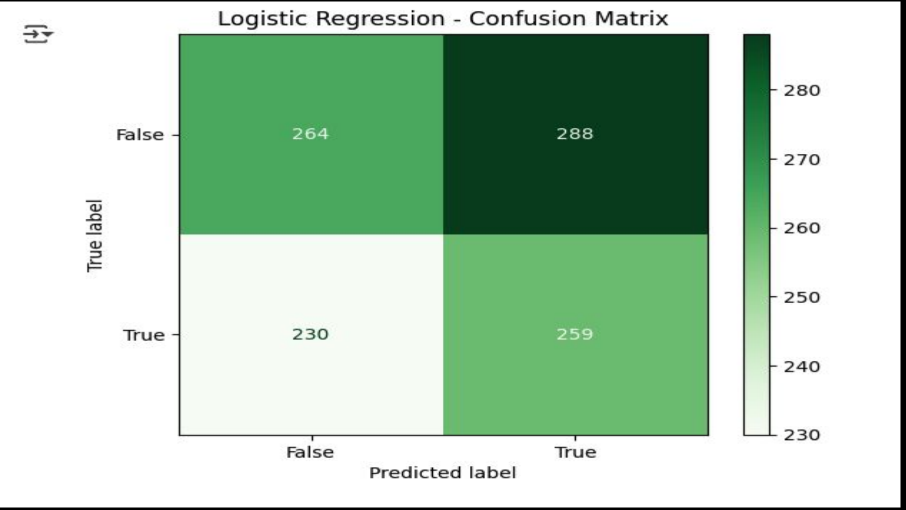
264 cases were correctly classified as "False" (True Negatives).
259 cases were correctly classified as "True" (True Positives).
288 cases were False Positives (wrongly predicted as "True").
230 cases were False Negatives (missed actual "True" cases).

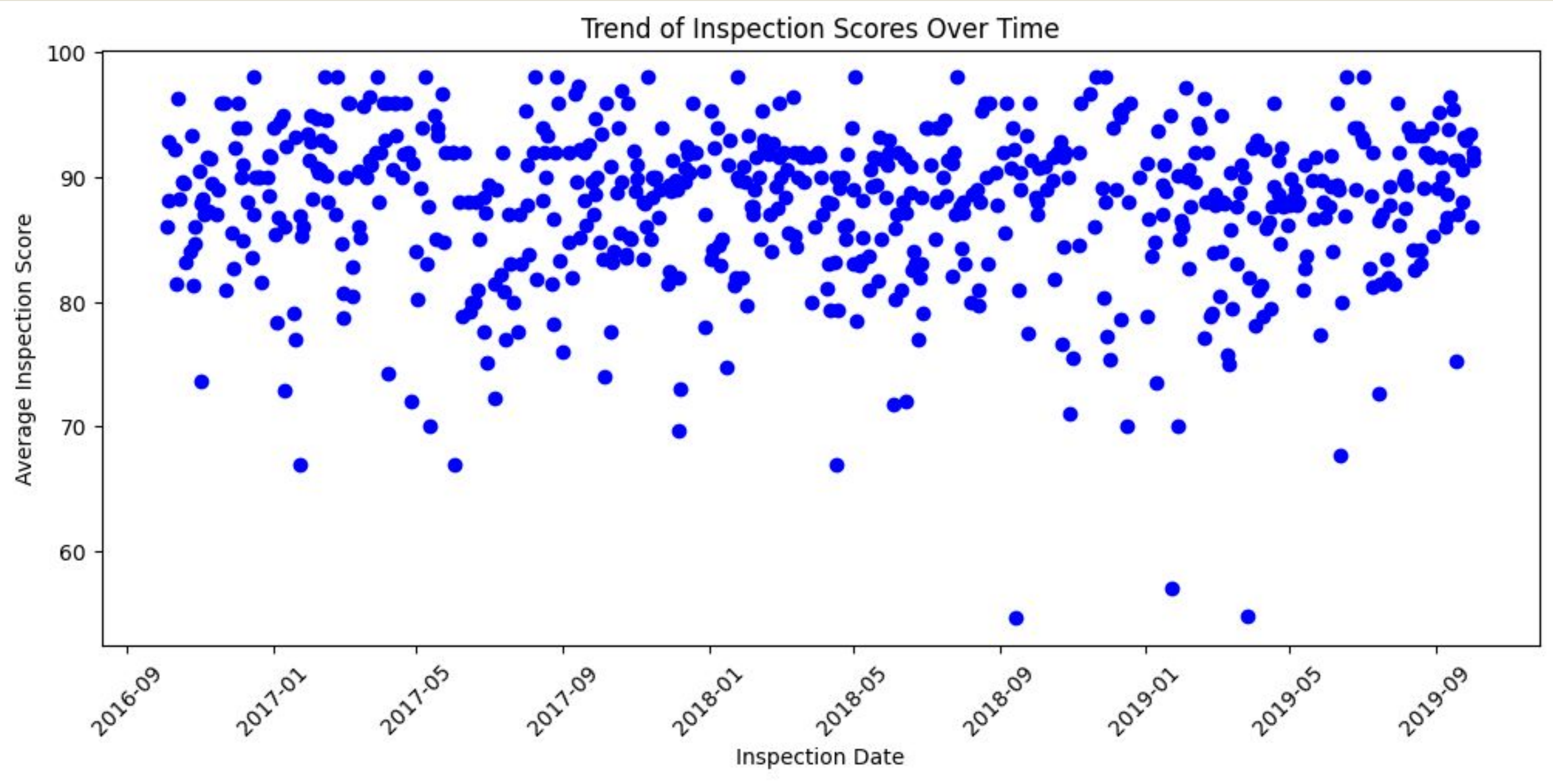
- **Higher misclassification rate:** A lot of False Positives (**288**) and False Negatives (**230**).
- **Accuracy is significantly lower than Random Forest.**
- **The model struggles to separate "True" and "False" labels correctly.**

⚠️ **Not a good choice for reliable predictions. Consider tuning hyperparameters or using a different model.**

Model	True Negatives	True Positives	False Positives	False Negatives	Overall Performance
Random Forest (🔵)	544	489	8	0	✅ Excellent accuracy (Minimal errors)
Logistic Regression (🟢)	264	259	288	230	⚠️ Lower accuracy (High misclassification)







- Most inspection scores are clustered between **80 and 100**, indicating that businesses **generally maintain good inspection scores** over time.
- There are **occasional sharp drops** in inspection scores, where some businesses score below **70 or even 60**.
- These **dips may indicate incidents of violations, seasonal trends, or stricter inspections** during certain periods.

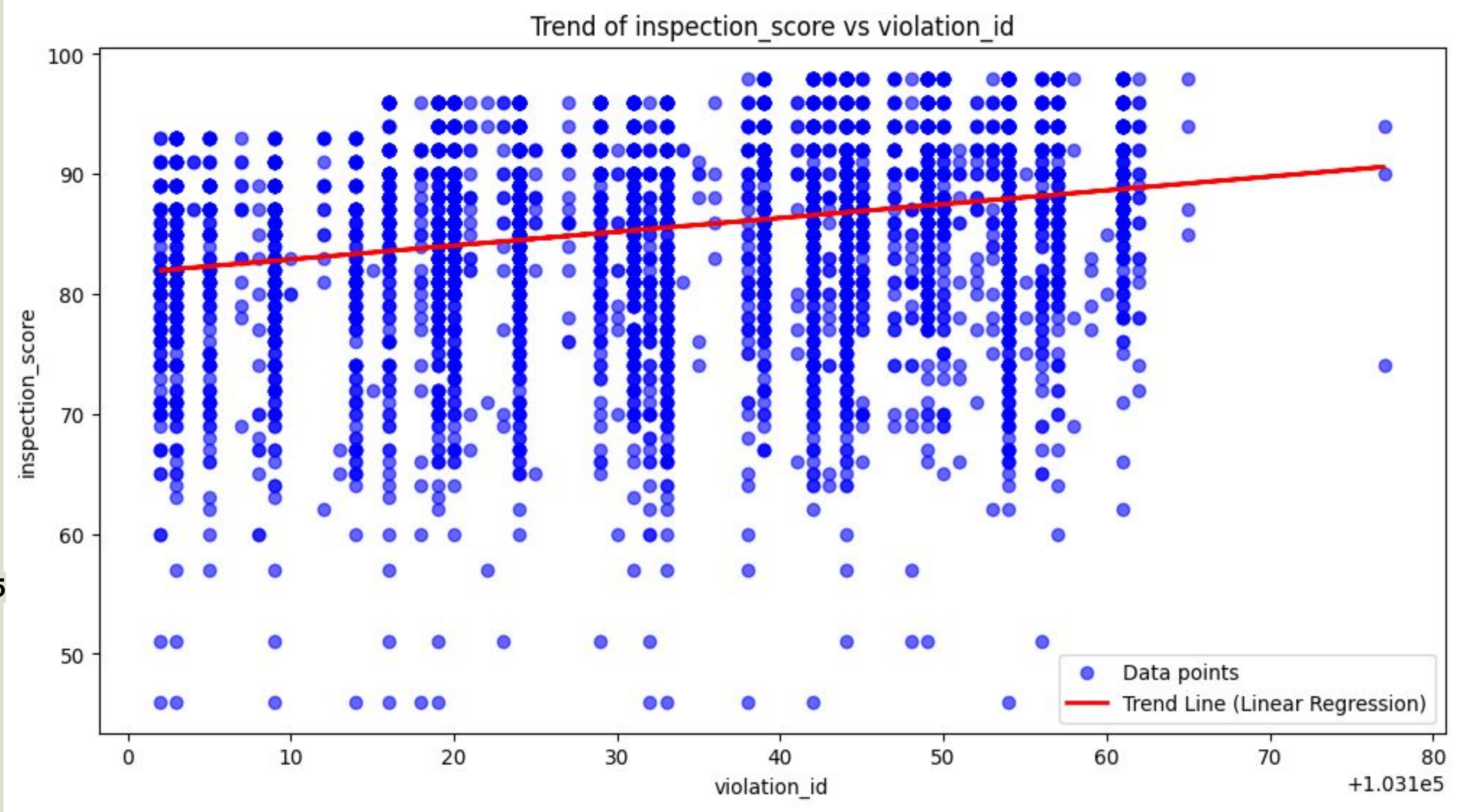
Linear Regression Model Statistics:

Coefficient:
0.1151362338317824

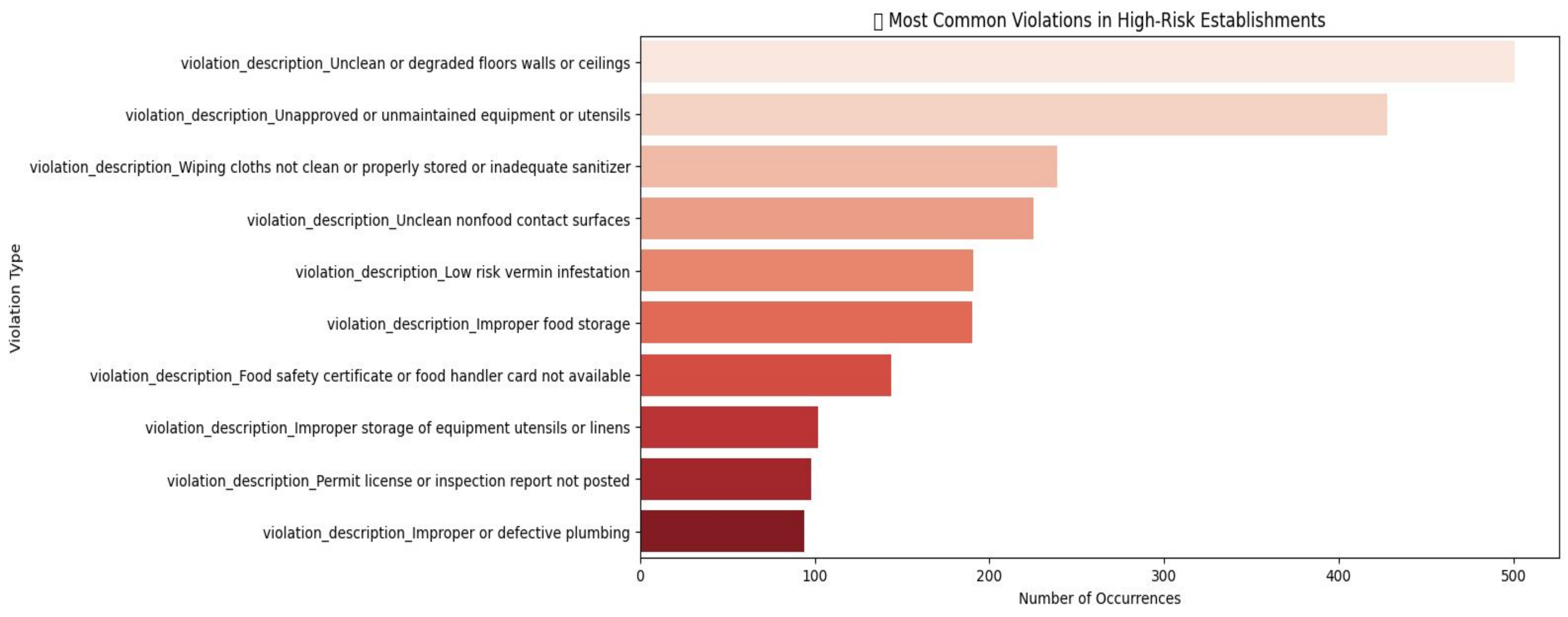
R-squared:
0.05348250733340554

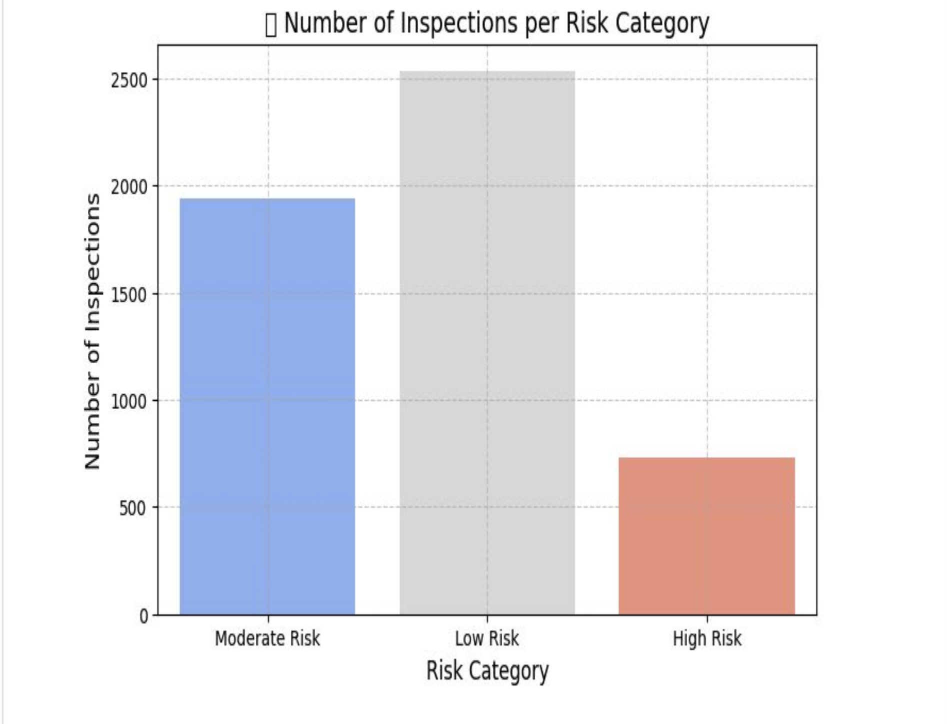
P-value:
3.774774598058409e-64

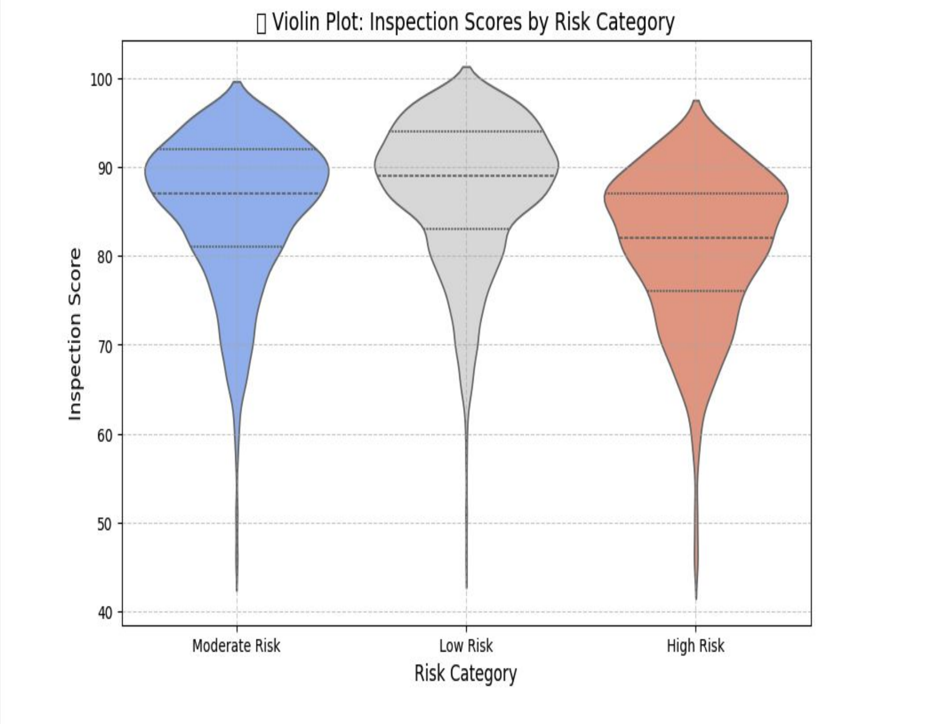
Standard Error:
0.0067149618402496105



High-Risk Common Violations







The higher number of inspections for low-risk businesses could indicate that these establishments comply well with regulations, leading to more routine inspections.

High-risk businesses being inspected less frequently may indicate either a lower number of high-risk establishments or stricter oversight that leads to closure after violations.

The higher the risk category, the more variation in inspection scores, indicating inconsistent compliance in moderate and high-risk businesses. Low-risk businesses maintain a stable, high inspection score, reinforcing that they follow safety and hygiene standards effectively. High-risk businesses show a broader range of scores, which suggests that some high-risk businesses meet compliance while others fail inspections significantly.

Thank You