Restaurants Inspections and Closures

1. Introduction

Inspections are one of the most important factors in a restaurant's ability to stay safe and open to the public, as well as being able to sustain a high success rate. In order to understand how inspections influence a restaurant's operational status, this project takes a look at data gathered from the state of Pennsylvania. The analysis uses key features that are used to predict inspection outcomes and provide insights on how to improve food safety and public health.

2. Data and Preprocessing

Datasets Used: Two datasets, from data.gov, were used for this project: 'Restaurants Inspection Descriptions.csv' and 'Status of Restaurants Post-Inspection.csv'. The datasets include every food establishment in the state of Pennsylvania, addresses, inspections received, types of inspections, grades given, and a plethora of other information.

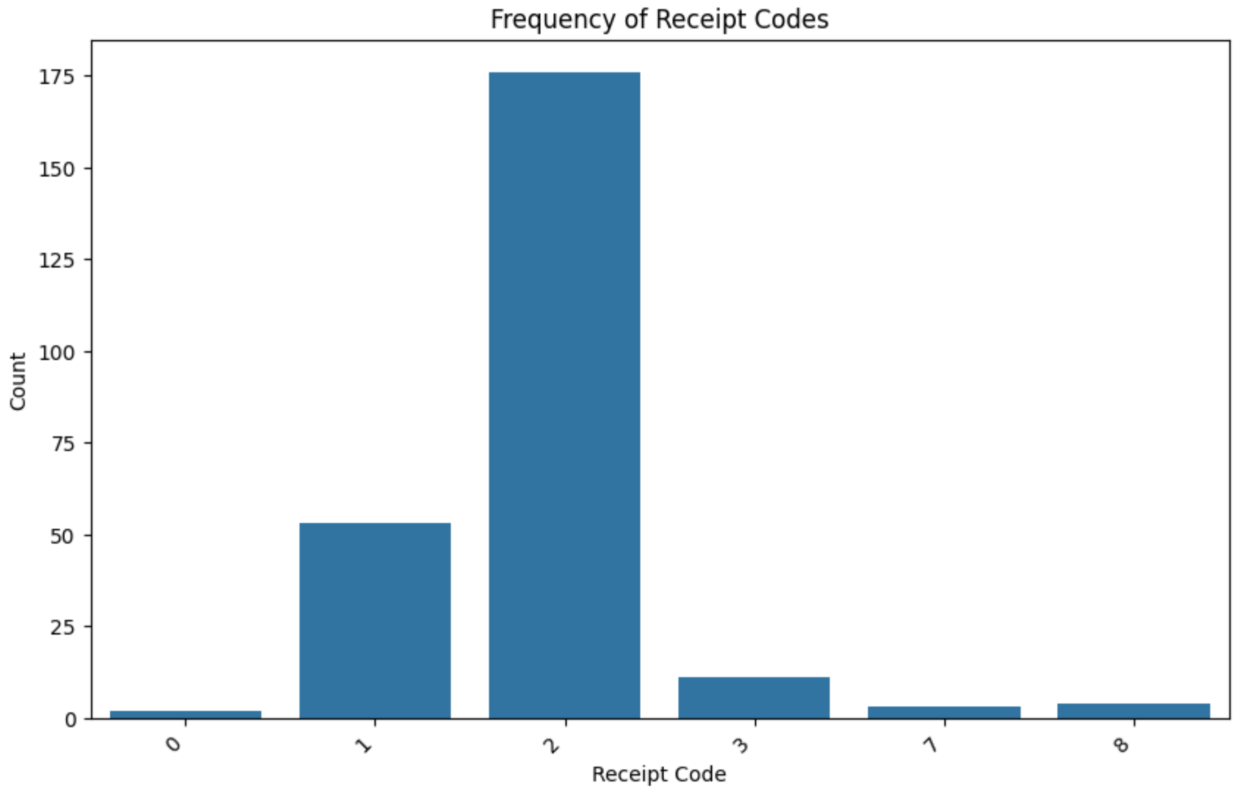
Merging: After uploading these datasets, they were merged based on the 'facility_name' column, due to it being a commonality. This created a comprehensive dataset, which made the data cleaning much easier.

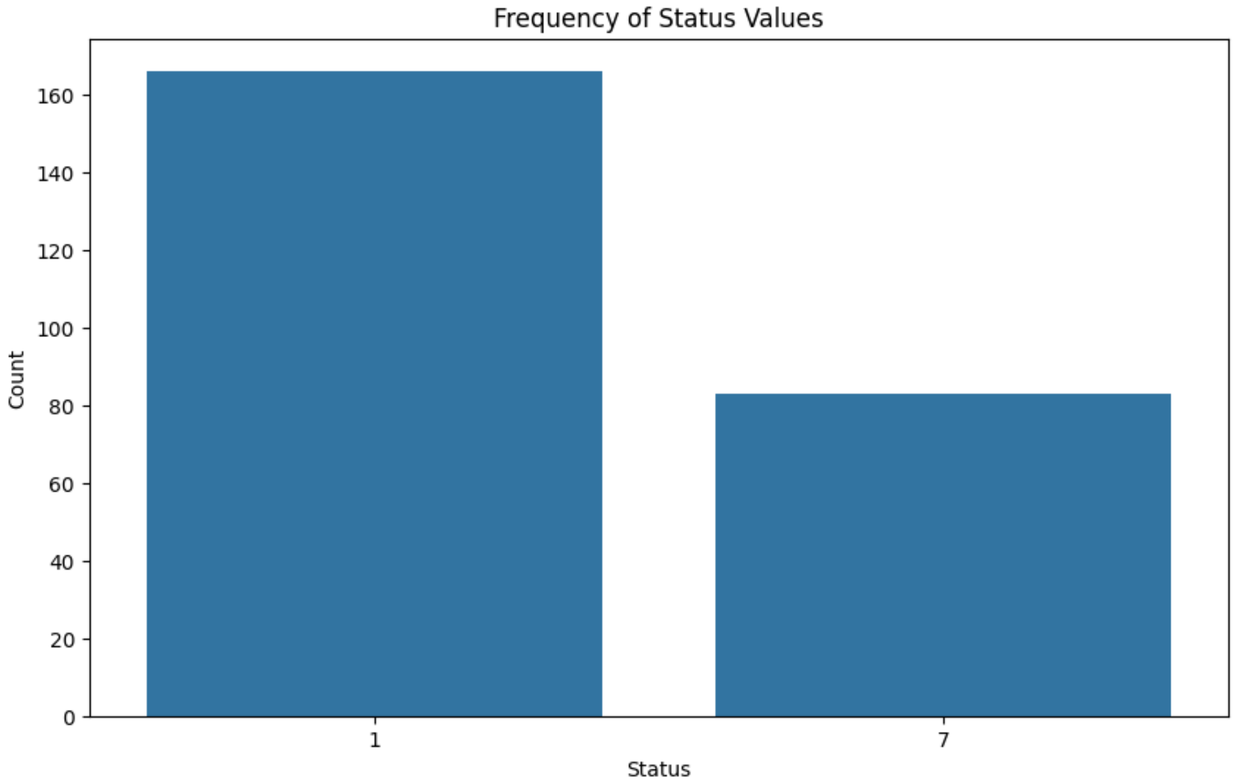
Data Cleaning: Before any data was analyzed, unnecessary columns were dropped. Because the dataset was very large, a sampled data frame of 250 samples was created to prevent the data from crashing the system. Missing values were handled by dropping rows that had a high percentage of missing values and imputation with the mode of the 'low', 'medium', and 'high' columns. Duplicated rows were then dropped and removed as well. A mask was placed on samples with a 'status' of 9, due to a value of 7 meaning the restaurant remained open and a 1 meaning it closed. A 9 is meaningless in this instance.

3. Exploratory Data Analysis (EDA)

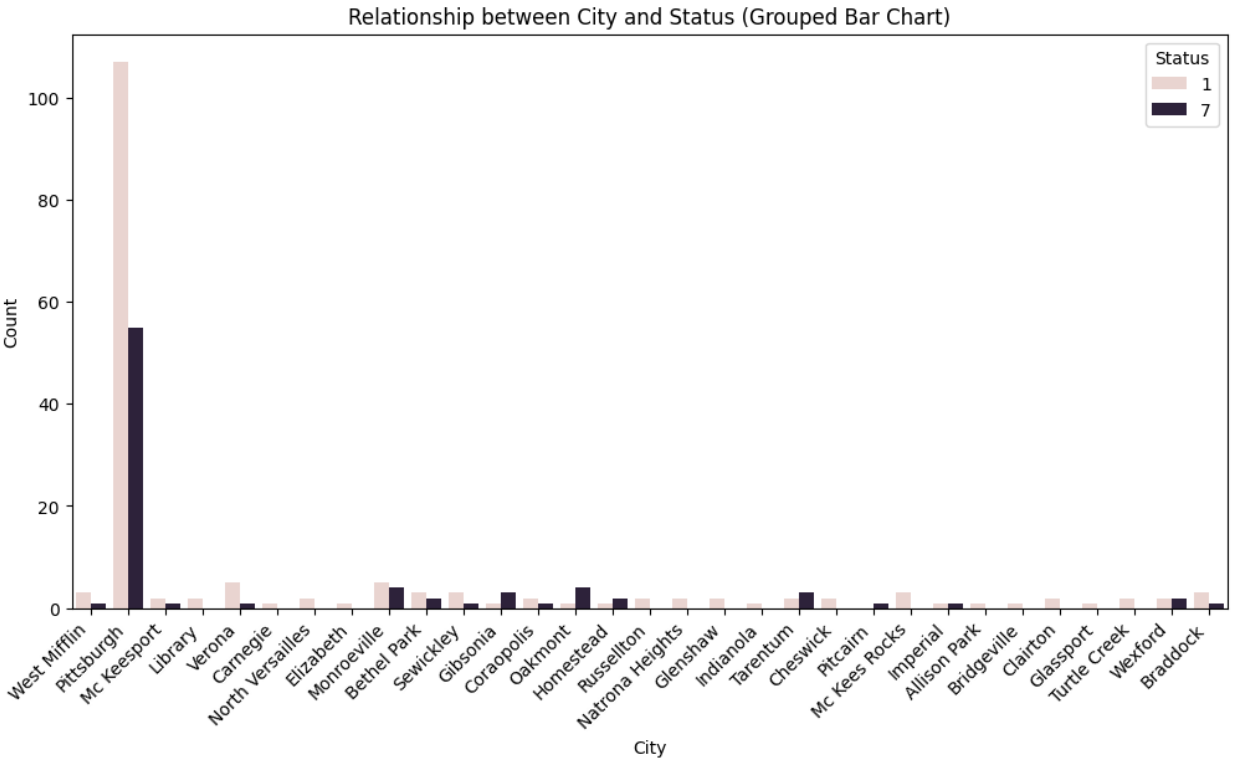
Data Transformation: Before a majority of the data was to be analyzed completely, categorical features, like 'purpose', 'facility_name', 'city_x', among others, were converted into numerical features using Label Encoding. This made it possible to use the data in machine learning modeling later on.

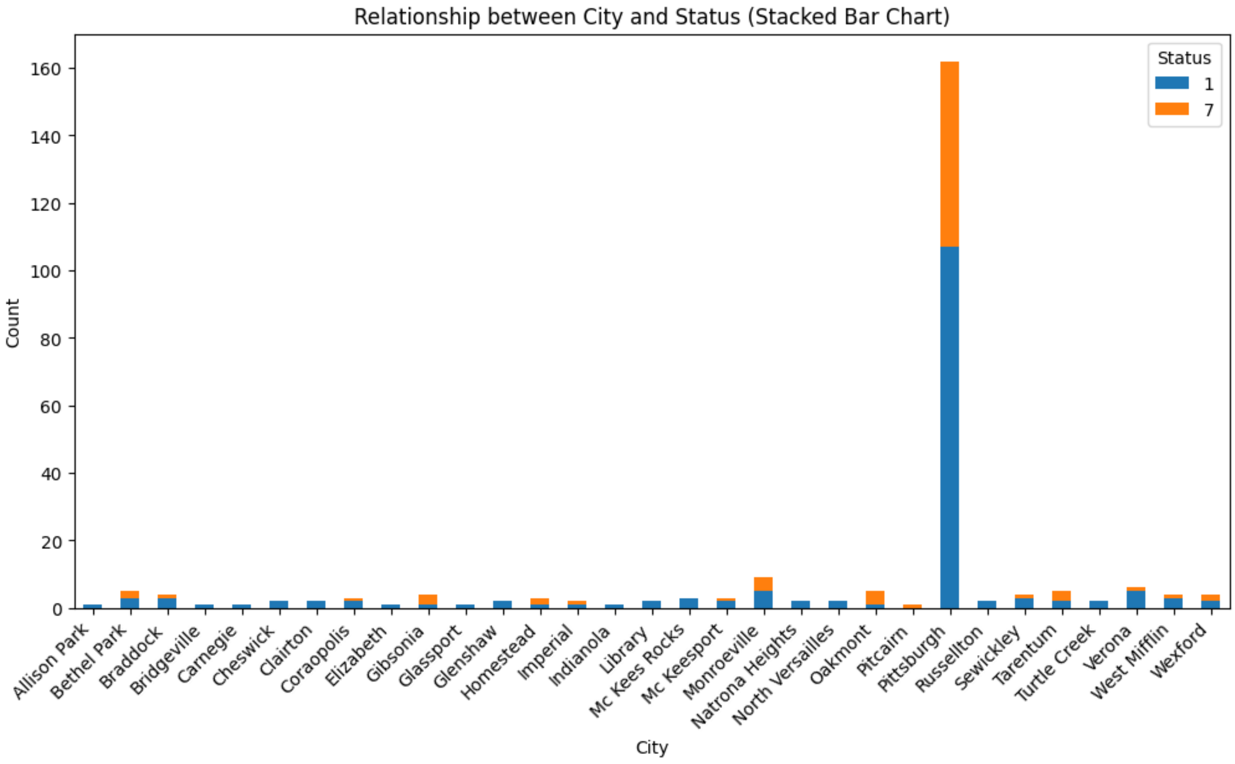
Exploration: To initially understand the data, descriptive statistics were calculated. To visualize frequency distribution among key variables, count plots were made, two of which are seen below. These displayed frequencies of Receipt Codes, as well as Status Values. The Receipt Codes chart displays numbers, 1-7, given to each sample following the first inspection. A score of 7 is passing, while a 1 requires a re-inspection at a future date. Samples receiving a low number far outweighed ones with a high number, indicating a common trend amongst the samples. The Status chart shows that far more samples received a 1 rather than a 7, with a respective count of 166 to 83.



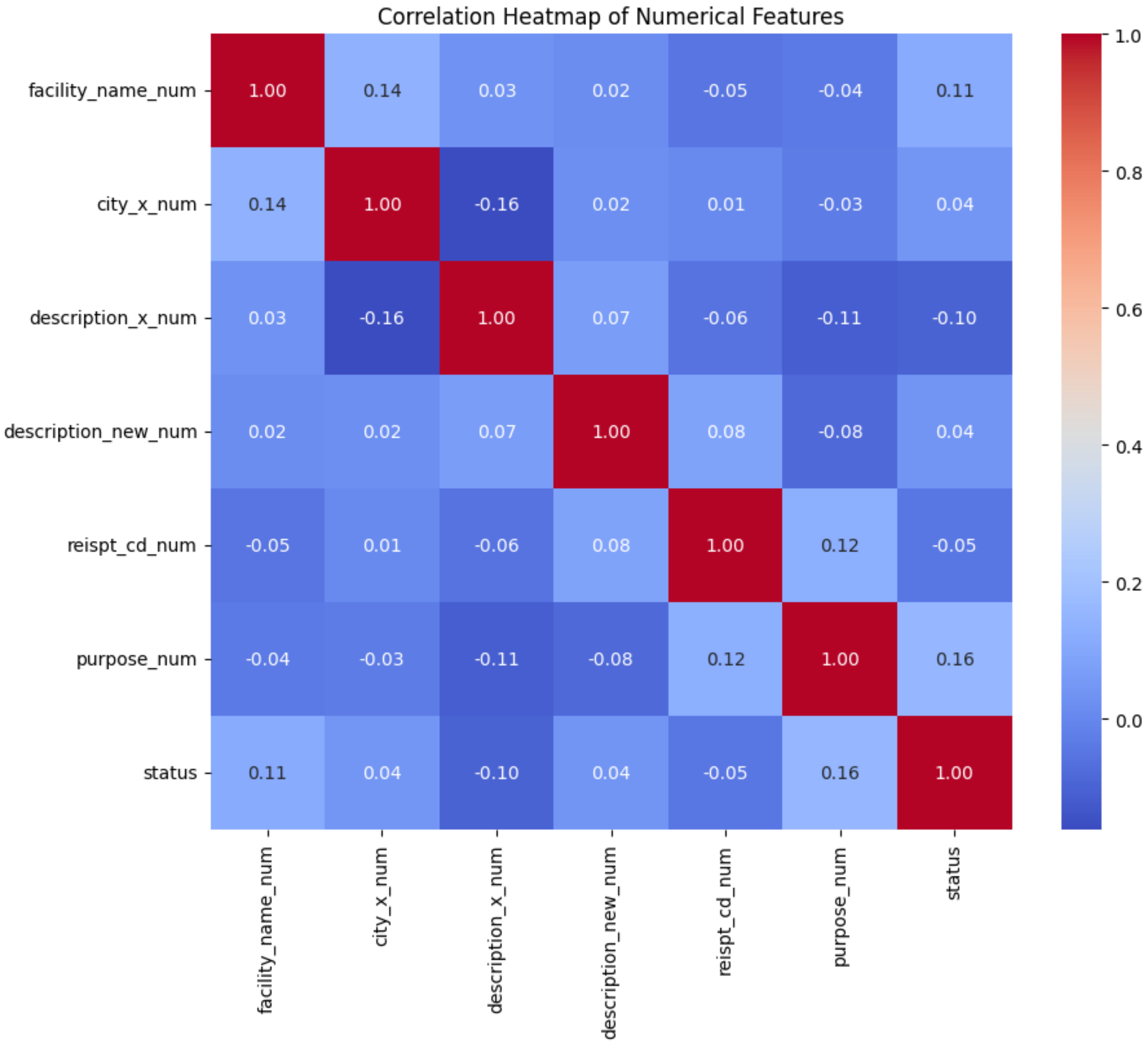


Relationships between variables were examined using box plots, line plots, as well as grouped and stacked bar charts. Looking at the grouped and stacked bar charts (see below), which examines the relationship between cities and status, we can see that a majority of samples reside in Pittsburgh. We can see throughout all of the cities, however, just how much more scores of 1 were received than 7, indicating a plethora of samples with poor health and safety conditions.





A correlation heatmap was generated to identify potential multicollinearity. There is a positive correlation between Facility Names and their status. Cities have a positive correlation with status as well.



3. Modeling

For this project, logistic regression was chosen because of its ability to predict categorical outcomes, such as restaurant status. Relevant features were selected and the target variable was set as ‘status’, due it being the primary focus of the project. The dataset was split into training and testing sets. The Logistic Regression model was trained on the training set. Grid search with cross-validation was implemented to improve the model’s hyperparameters for better performance.

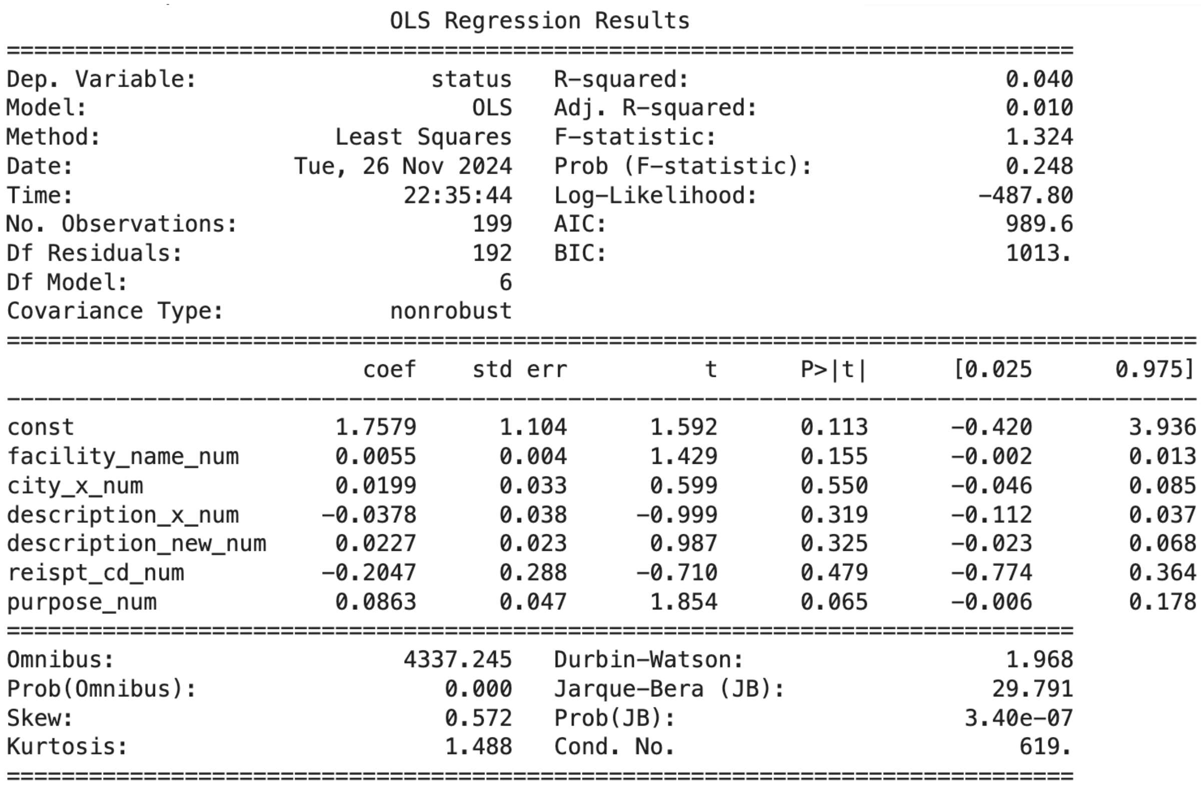
4. Model Evaluation

The model’s performance was assessed on the testing set using metrics such as accuracy, F1-score, precision, recall, mean-squared error (MSE), and root mean-squared error (RMSE).

In addition to this, the coefficients and p-values of the model parameters were analyzed to see how they impacted the predictions and their significance. Multiple Linear Regression was used as an additional way to check the relationship between each feature and the status of the samples.

5. Findings

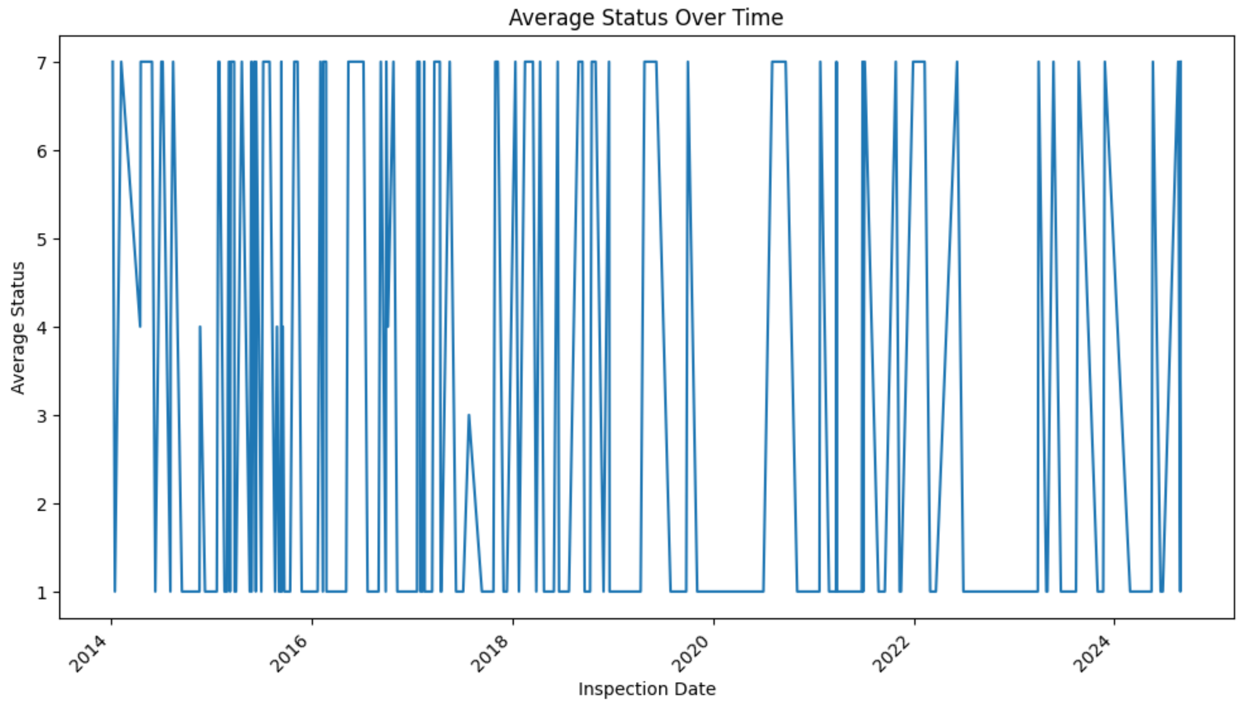
The results found using the logistic regression model tell us how effective the model is at predicting restaurant status. The model received a good accuracy score of 0.74. The model also had a precision score of ~ 0.81, a recall score of 0.74, and a F1-score of ~ 0.63. The RMSE is 3.06, which indicates a high level of accuracy for this model’s predictions.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 status   R-squared:                       0.040
Model:                            OLS   Adj. R-squared:                  0.010
Method:                 Least Squares   F-statistic:                     1.324
Date:                Tue, 26 Nov 2024   Prob (F-statistic):              0.248
Time:                        22:35:44   Log-Likelihood:                -487.80
No. Observations:                 199   AIC:                             989.6
Df Residuals:                     192   BIC:                             1013.
Df Model:                           6
Covariance Type:            nonrobust
=========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                    1.7579      1.104      1.592      0.113      -0.420       3.936
facility_name_num        0.0055      0.004      1.429      0.155      -0.002       0.013
city_x_num               0.0199      0.033      0.599      0.550      -0.046       0.085
description_x_num       -0.0378      0.038     -0.999      0.319      -0.112       0.037
description_new_num      0.0227      0.023      0.987      0.325      -0.023       0.068
reispt_cd_num           -0.2047      0.288     -0.710      0.479      -0.774       0.364
purpose_num              0.0863      0.047      1.854      0.065      -0.006       0.178
==============================================================================
Omnibus:                     4337.245   Durbin-Watson:                   1.968
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.791
Skew:                           0.572   Prob(JB):                     3.40e-07
Kurtosis:                       1.488   Cond. No.                         619.
==============================================================================
```

Based on the visualizations and analysis of the ‘city_x’ feature, we can observe that, while Pittsburgh has the highest amount of restaurants included in the dataset, it also the highest amount of closures due to inspections. These charts determine that there is a serious health and safety issue amongst restaurants in the state of Pennsylvania.

The Inspection Date line plot shows more often than not, the line stays at 1 or returns to 1. It does not stay at 7 for very long, indicating trends that that most inspections are not successful. The scores of 1 seem to frequent more around when the COVID-19 pandemic started in 2020.



5. Conclusion and Future Steps

Based on the analysis above, we can use machine learning to predict restaurant inspection outcomes. The high accuracy score and low RMSE score tells us that the prediction power of the model is fairly accurate and quite strong. These insights can be used to inform policy decisions and improve public health through the distribution of resources.

Improvements to the model can be made by gathering additional data, such as city weather data (daily temperature, precipitation, etc.), social media data (reviews, ratings), and individual restaurant characteristics (years in operation, type of cuisine, etc.). Feature Engineering, such as time-based features may improve the dataset. This means the time of the day, week, month or year when the inspection occurs may influence the results.

For a more accurate predictive accuracy, using a method like Random Forest may prove useful.

To promote further investigation, developing models that predict the likelihood of specific types of violations may give more in-depth answers. After any policy changes or interventions are made, assessing the impact on restaurant inspection outcomes can be made. Comparing data collected before and after the changes to analyze the changes in status or violation rates.

By implementing these improvements and conducting further research, great enhancements to the understanding of factors that influence restaurant inspection outcomes can be made. Considerate improvements can be made towards promoting food safety and public health.