Predicting Restaurants' Violations of Allegheny County

Hawraa Salami

SPRINGBOARD Capstone Project 2

Inspection of Food Facilities in Allegheny County

- Food facilities should at least be inspected **once a year**.
- Due to *shortage in staff,* several restaurants have not been inspected **since 2017**.
- Data-driven models can help the county in prioritizing its inspections.

Our Approach

- What features of food facilities can predict their future violations?

- Our approach is to combine information from two datasets:

Dataset of violations detected during inspection in Allegheny County

Dataset of Yelp reviews for food facilities of Allegheny County

How Can the Model Help?

Allegheny County

Improve the inspection process:

- identify which restaurants to target first
- reduce the number of violations missed
- detect any risks earlier

Datasets: Violation & Yelp Reviews

Violation Data

- Inspection ***Date***
- Numbers of:

  Low, Medium, High violations
- ***Description*** of the food facility
- ***Address*** of the facility
- ***ID*** of the facility

Yelp Data

- ***Attributes*** of facilities (price range, ambience, alcohol, good for kids, …)
- ***Categories*** of facilities (cuisine, special food, description of the place)
- ***Text reviews***, number of ***star*** given by customers
- ***Business ID*** of the facility

Data Wrangling

Data Cleaning & Finding the Common Facilities

Data Wrangling

1. Violation Dataset

- Filled missing addresses using facilities dataset
- Fixed the IDs of the same restaurants that have multiple IDS (by relying on their name and address)

2. Yelp Reviews

- Fixed the business IDs of the same restaurants that have multiple IDS
- Extracted categories and attributes each as one column
- Normalized Text Reviews

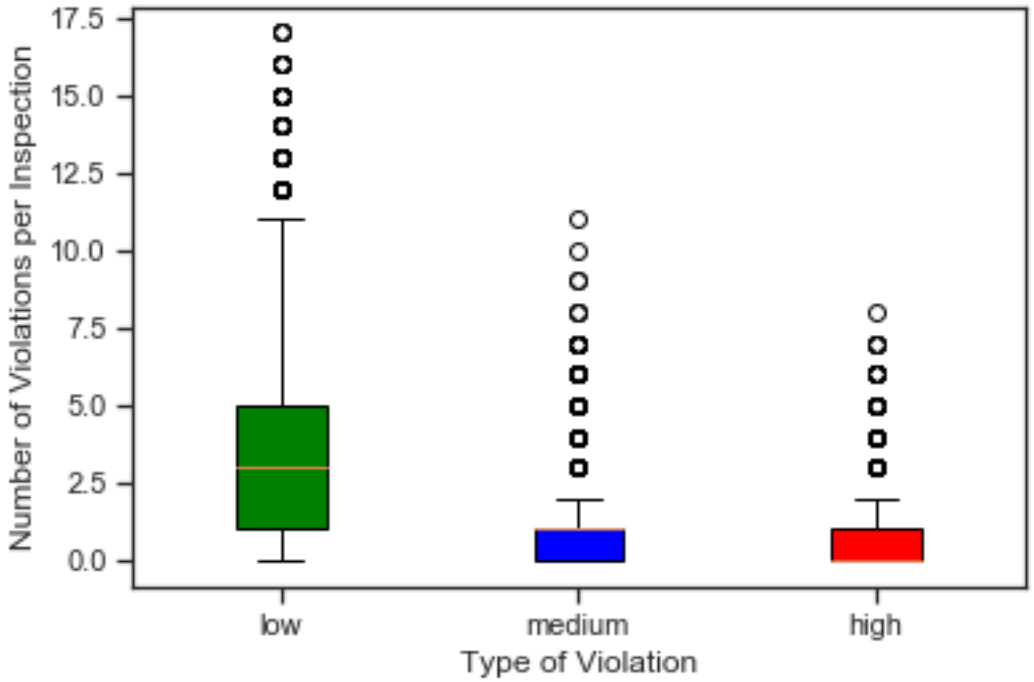
3. Mapping the IDs of the facilities

- Used the name and address of food facilities to find the mapping between the two datasets

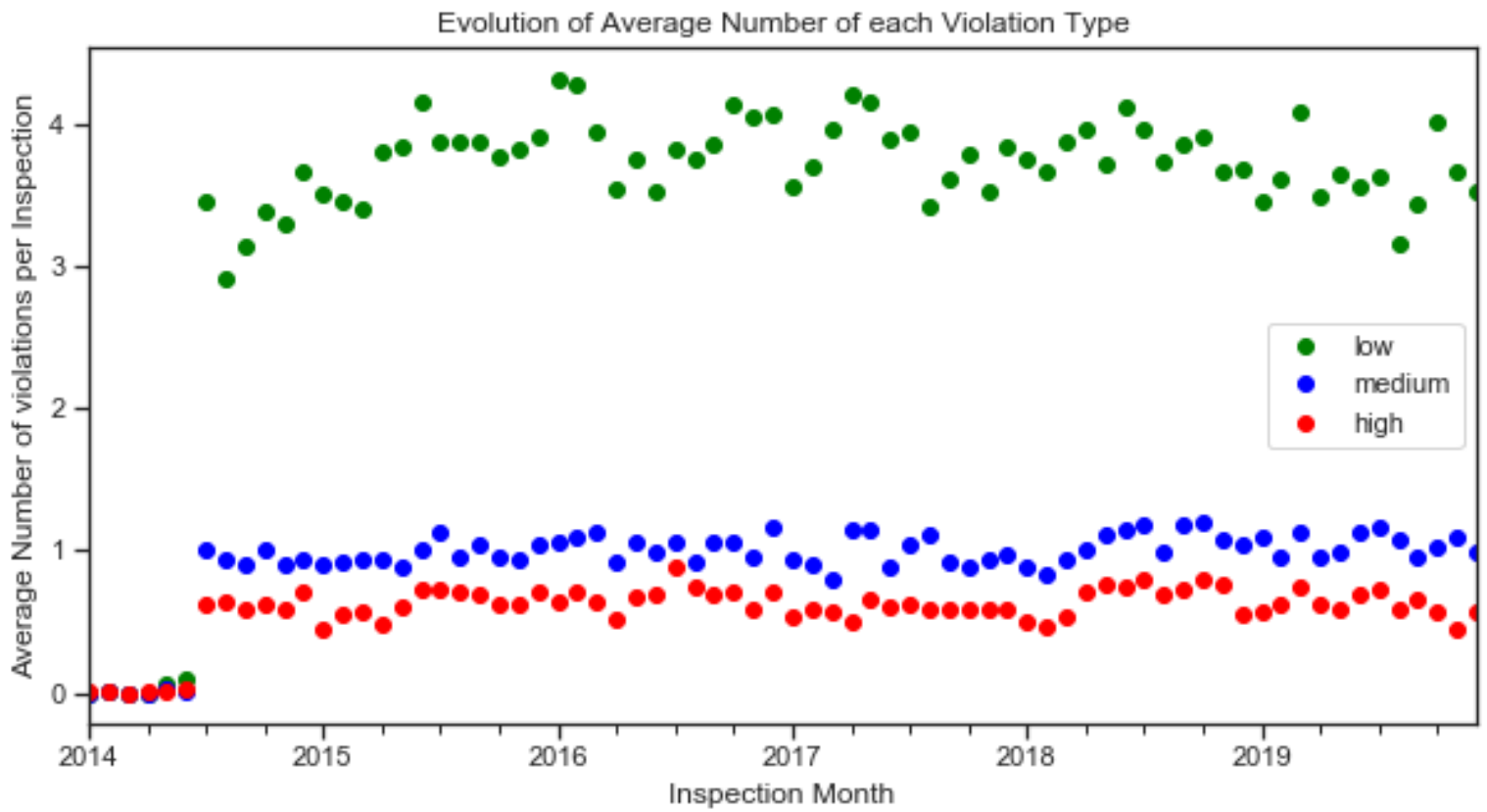
Exploratory Analysis

Data Visualizations & Statistical Analysis

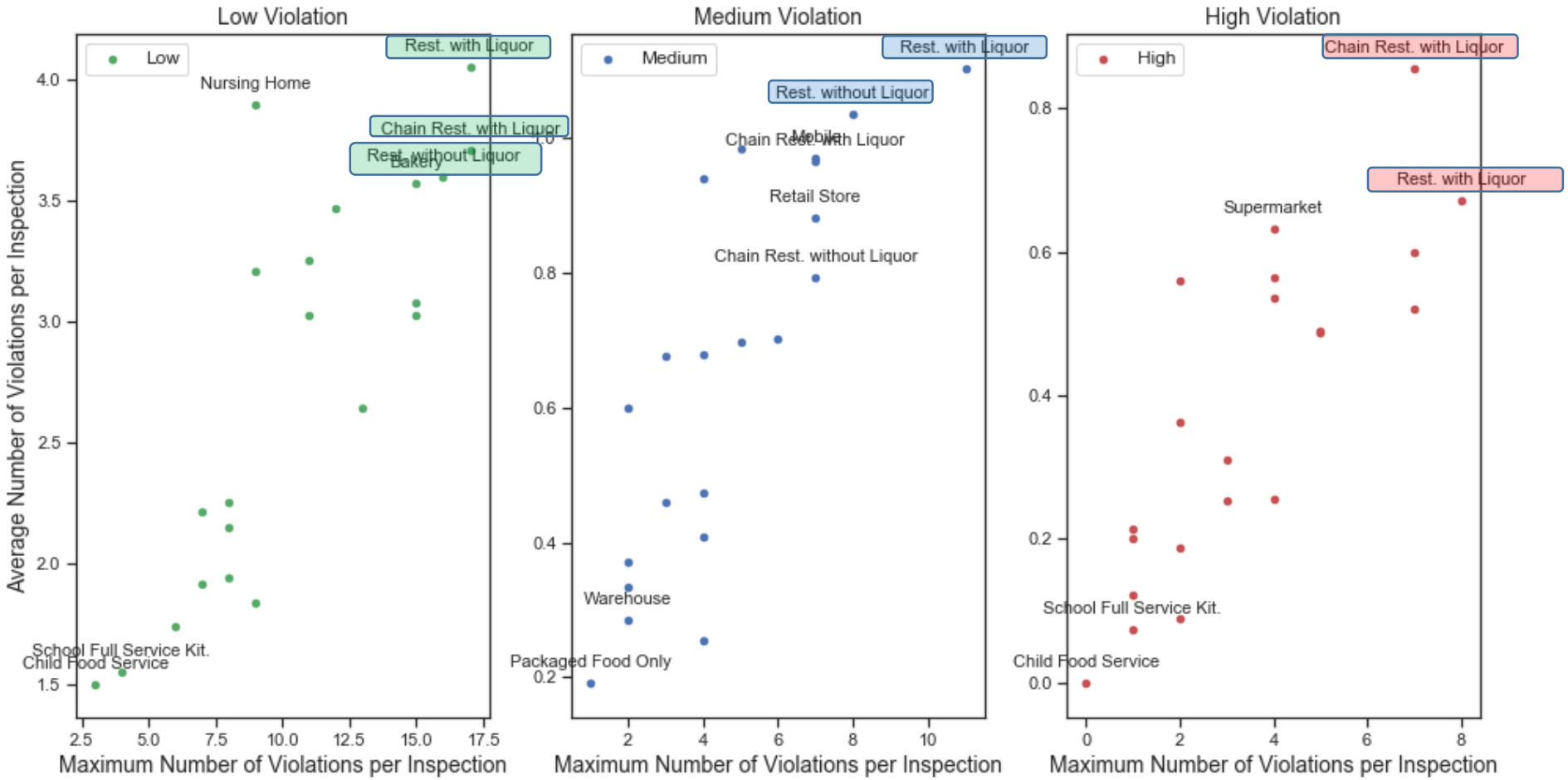
Types of Violations



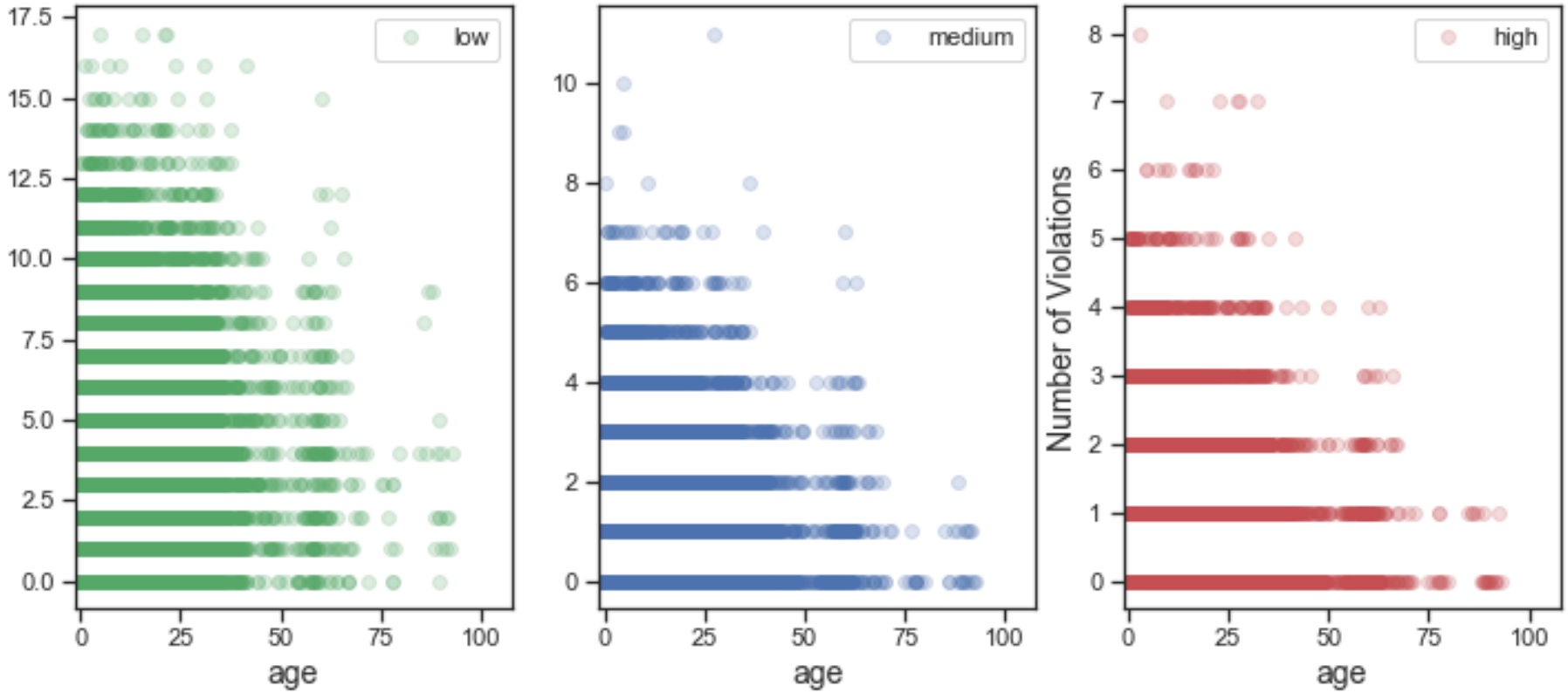
Evolution of Average Number of Violations



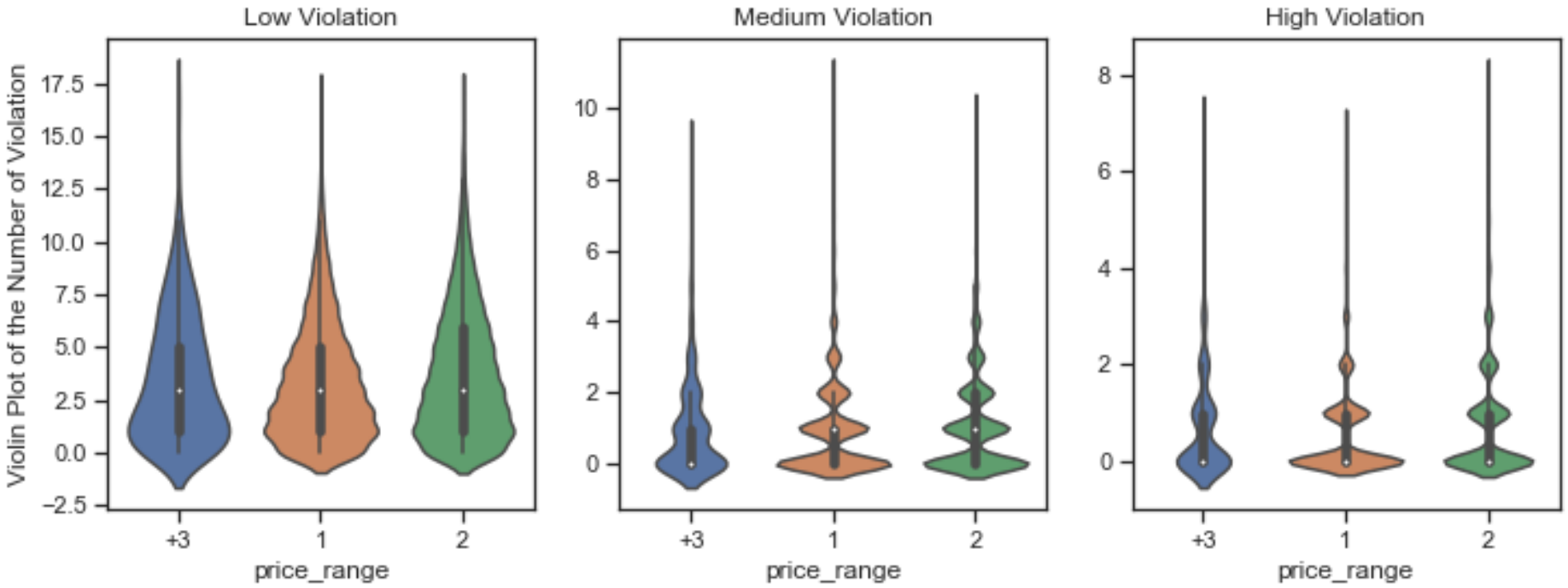
Violations vs Features of Food Facilities: ‘Description’



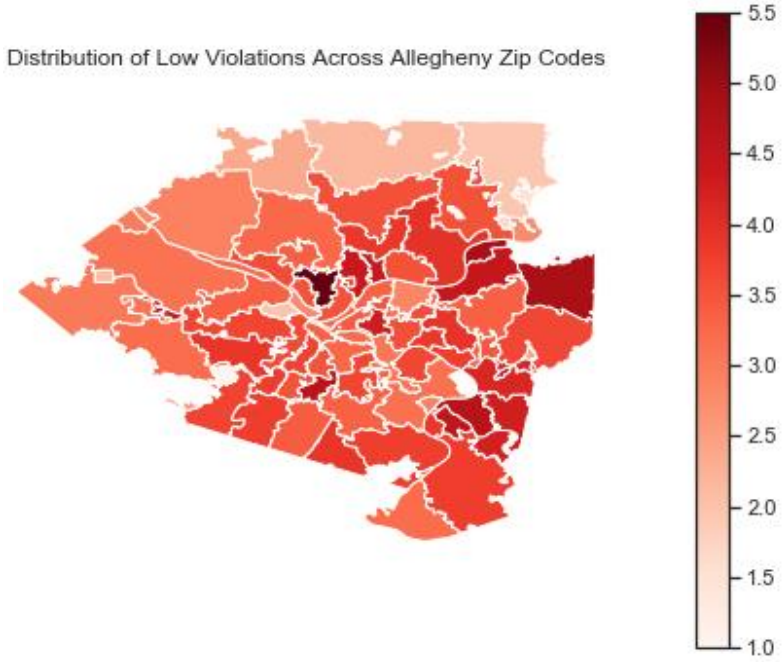
Violations vs Features of Food Facilities: ‘Age’

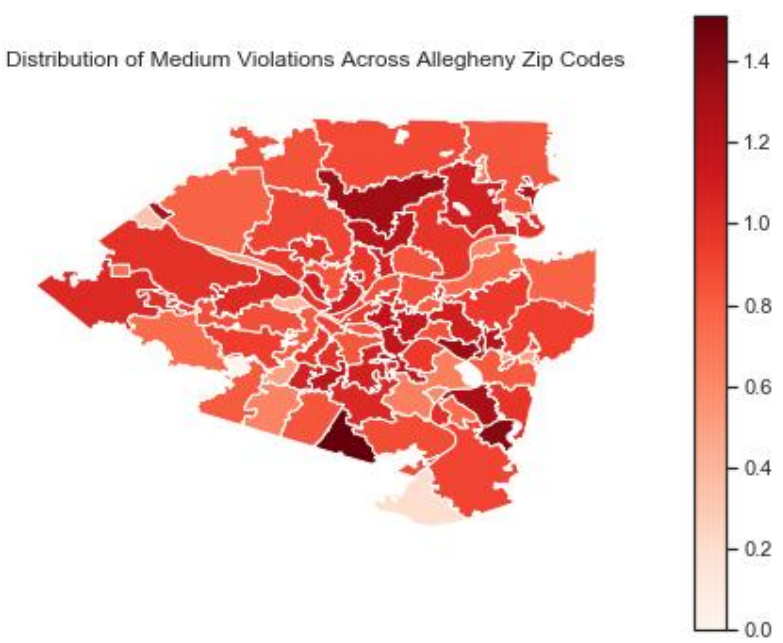


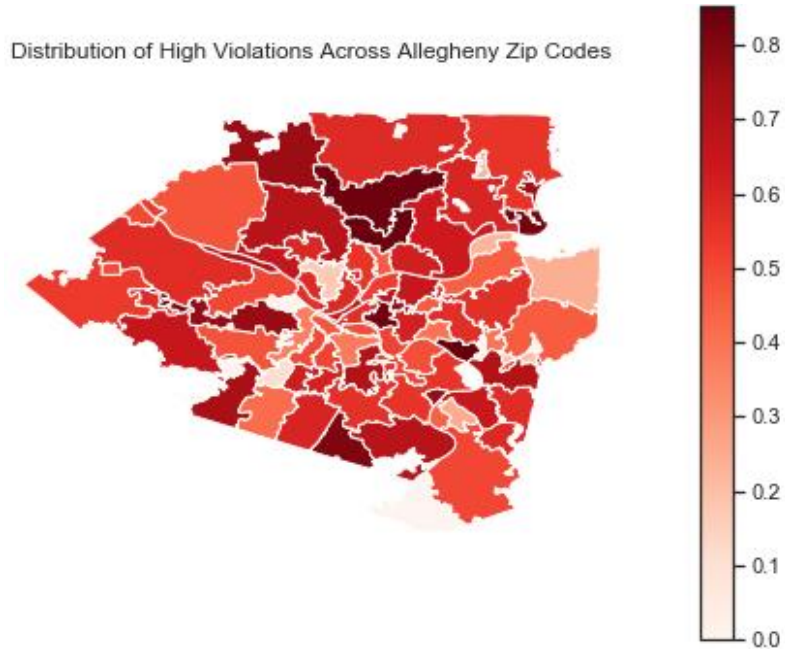
Violations vs Features of Food Facilities: ‘Price Range’



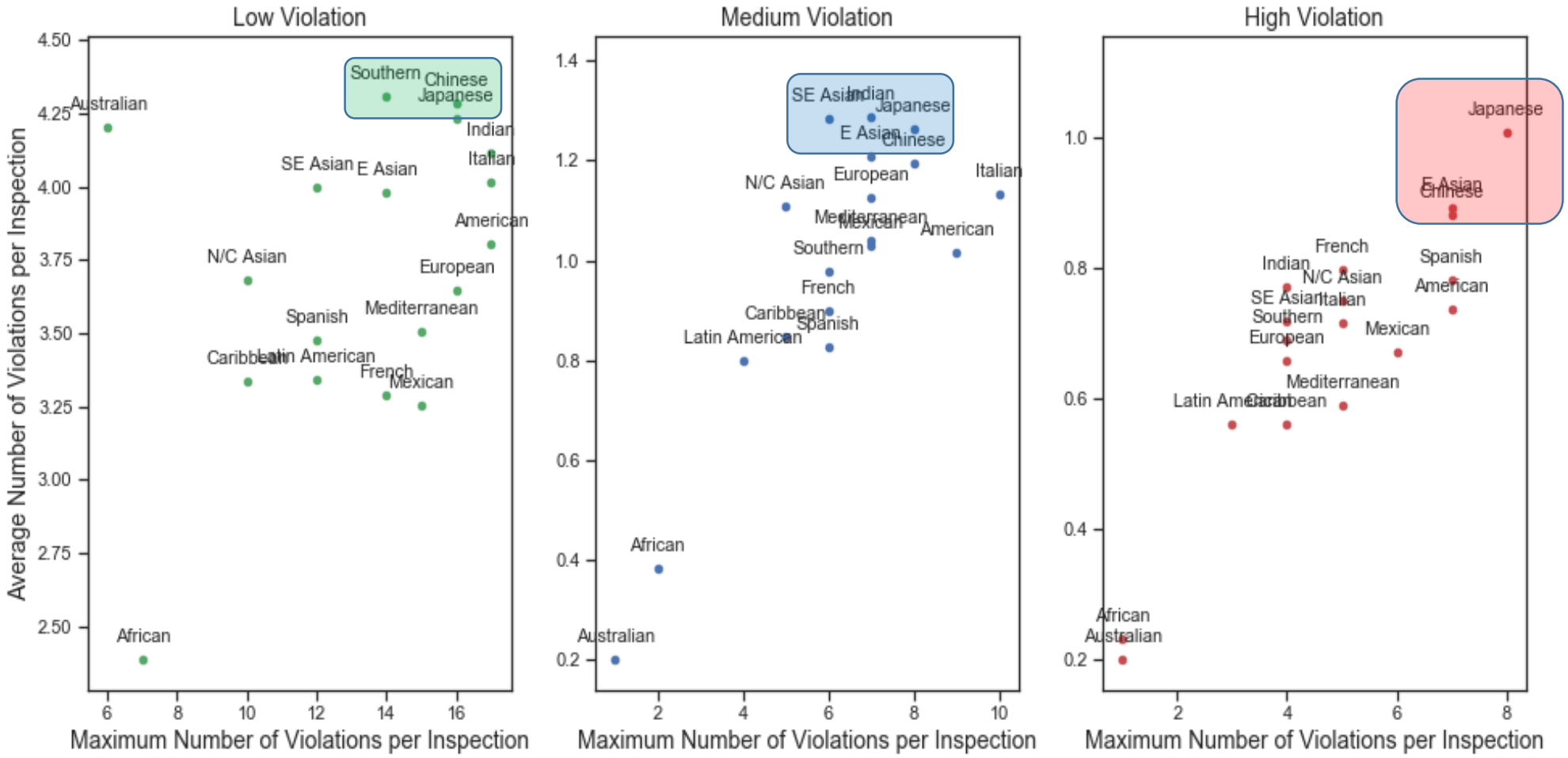
Violations vs Features of Food Facilities: ‘Zip Codes’



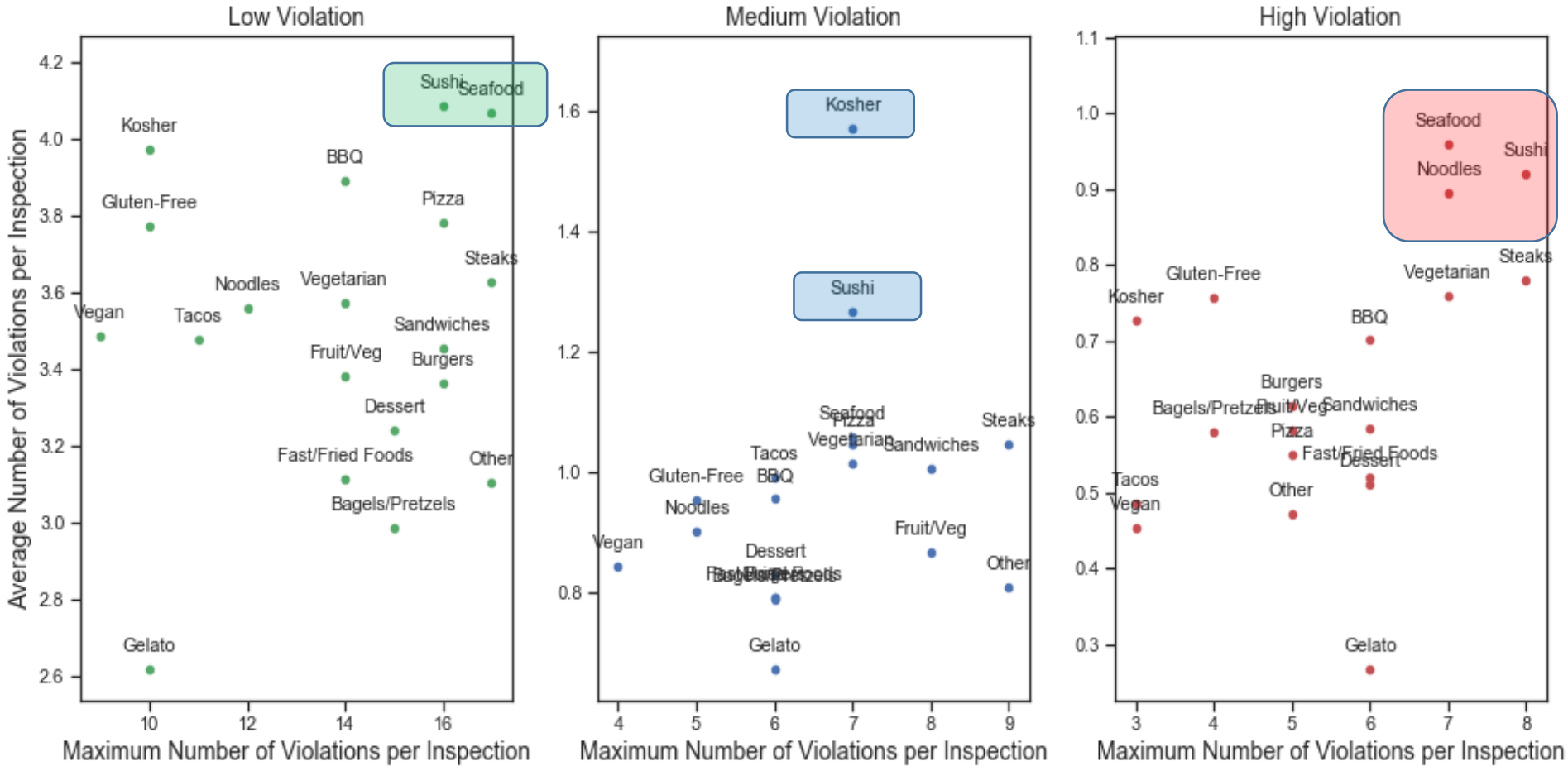




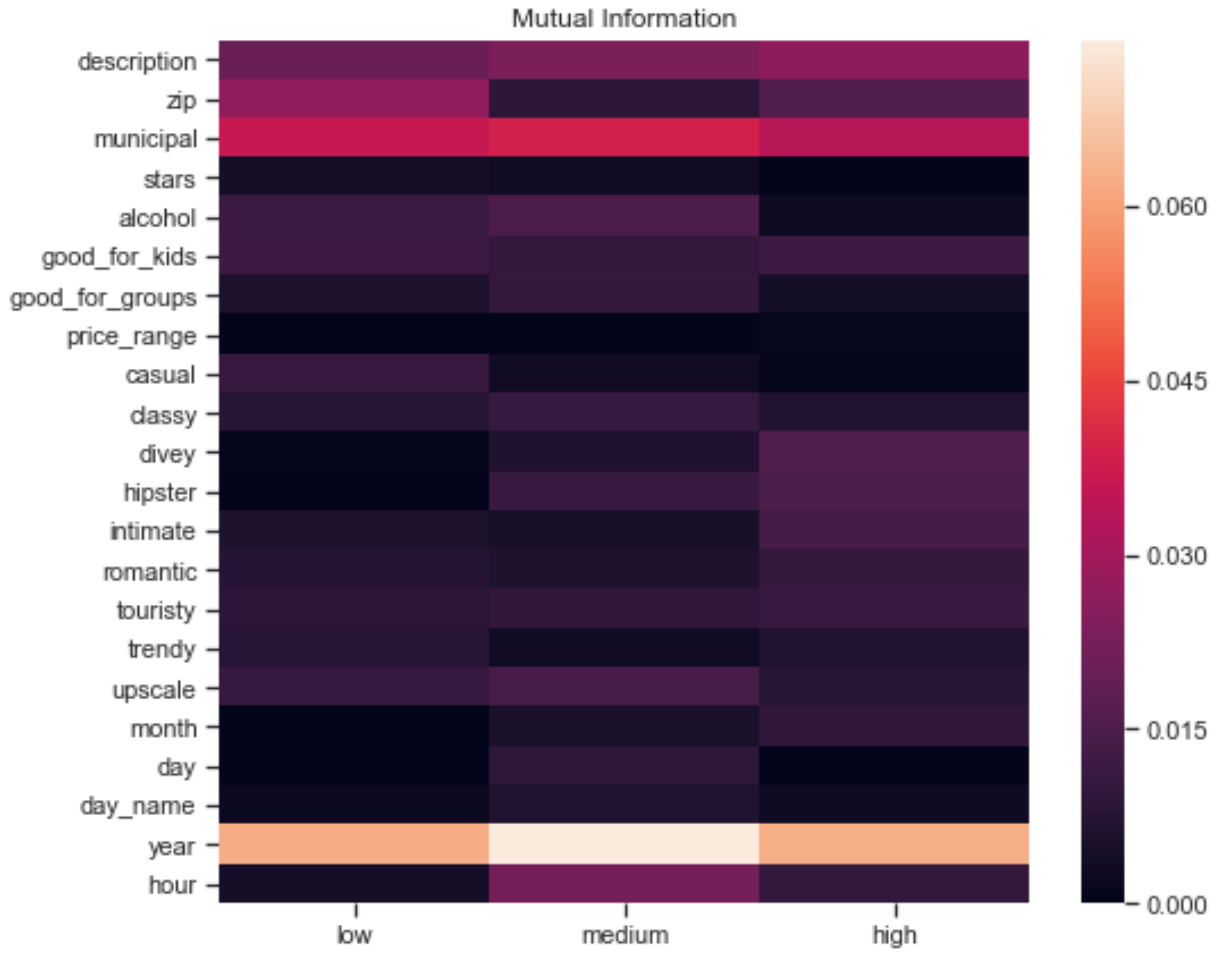
Violations vs Features of Food Facilities: ‘Cuisine’



Violations vs Features of Food Facilities: ‘Specialty Food’



Mutual Information between some Features and Number of Violations

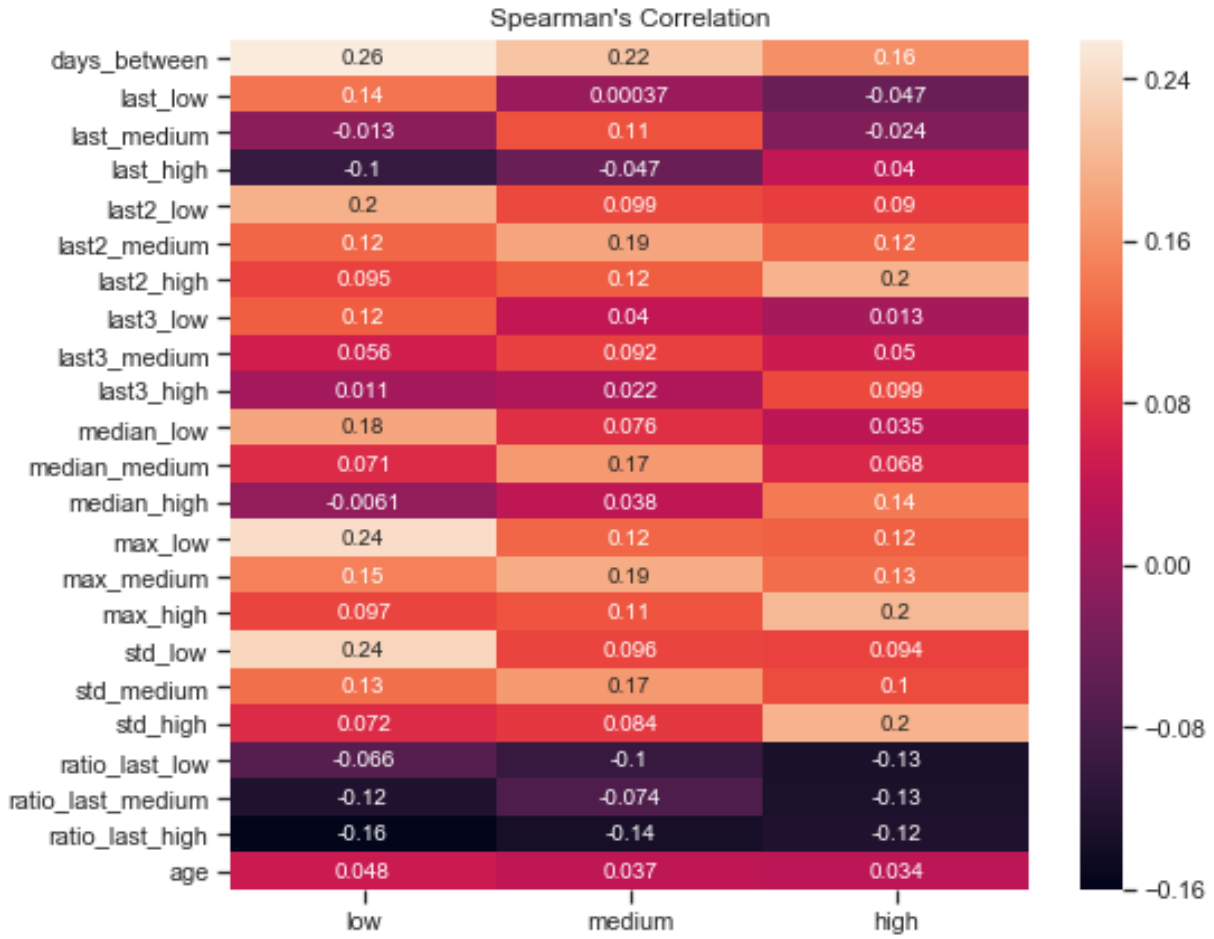


Historical Performance of Past Inspections

We added columns that capture past performance of each inspected place:

1. Inspection number
2. Number of days between inspections
3. Statistics of Past performance

Correlation between Past Performance and Violations

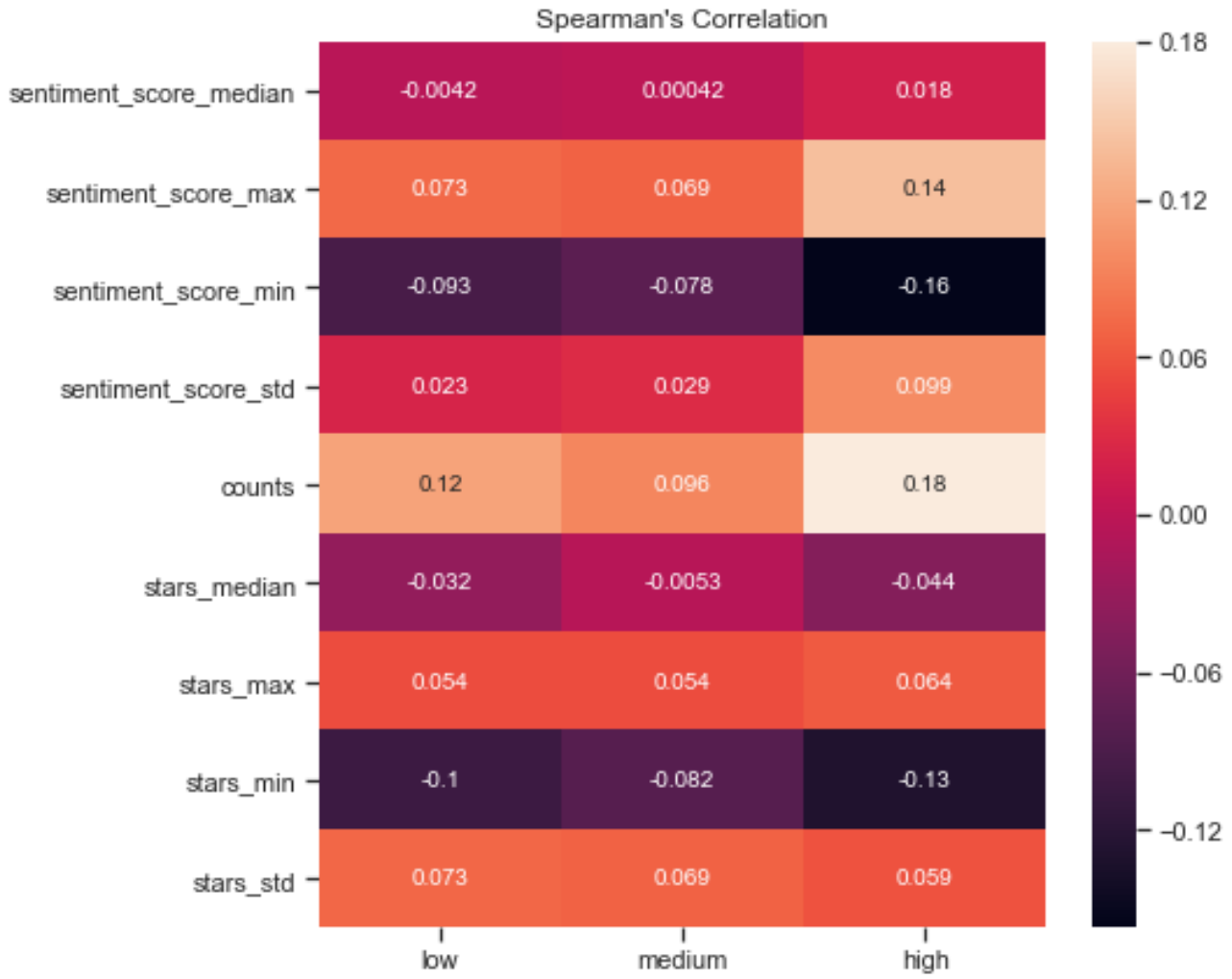


Analysis of Reviews

1. Sentiment Analysis

Focused on reviews mentioned before the inspection date:

- compute their sentiment scores
- find their rolling statistics
- add the total number of reviews and the average number of stars



Analysis of Reviews

2. Topic Modeling

Extracted 5 topics from the reviews :

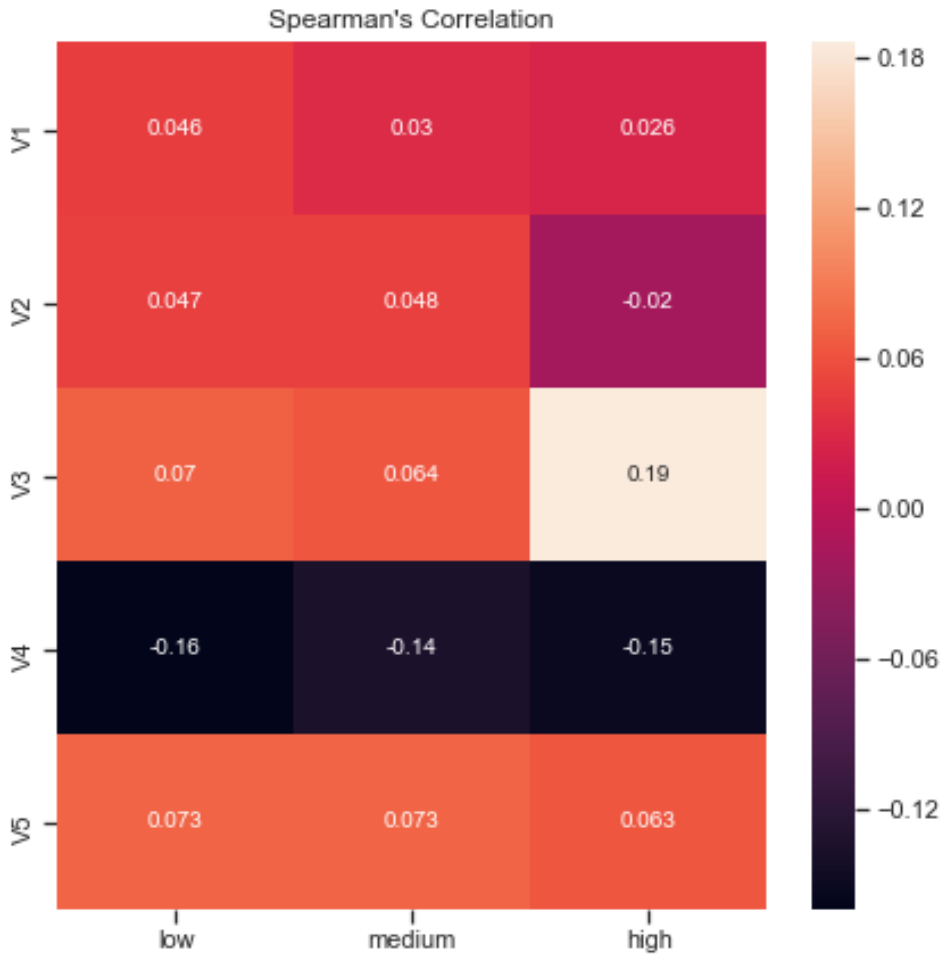
Topic 1: *bar beer wing drink place good bartender great burger night*

Topic 2: *pizza order delivery good crust place cheese wing sauce get*

Topic 3: *food good order go get place service restaurant come time*

Topic 4: *coffee store get place go shop good one like make*

Topic 5: *chinese sushi food roll good chicken rice order place thai*



Building Predictive Models

In-Depth Analysis

Predicting Future Performances

- Two possible approaches to explore:

1- **Regression**: estimate the number of future low, medium and high violations

2- **Classification**: classify a future inspection as a fail or pass

- Preprocessing steps:

1- Split the data into 80% for training and 20% for testing

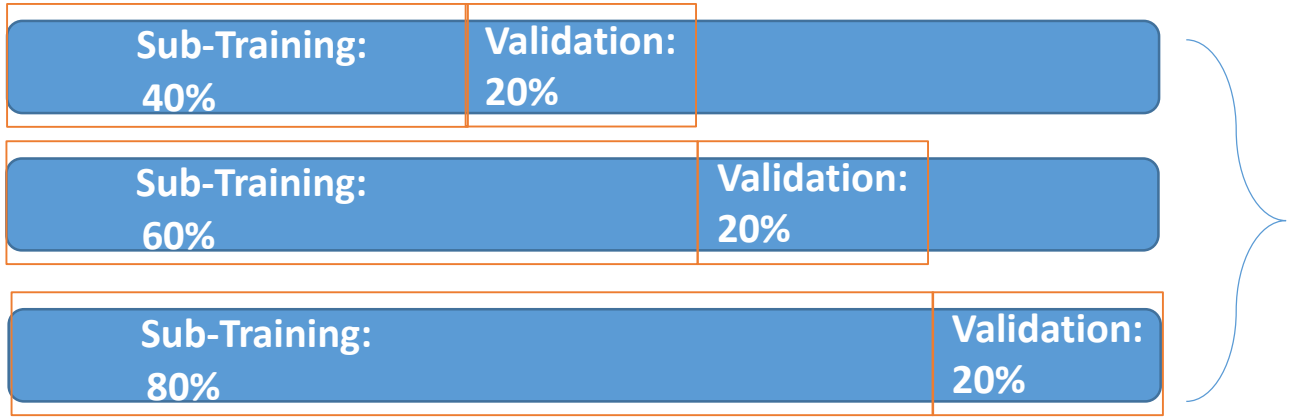
Test set corresponds to future inspections for the training set.

2- Encode the nominal columns (tried one hot, label encoding)

Like Nested Cross-Validation

- Test (Validation) set: future inspections for the training (sub-training) set
- Split the data:

- Perform like-nested cross-validation on the training set:



- Train different models
- Find their average performances
- Select best model

Regression Training Models

- Tried training with and without feature selection
- Feature selection: filter method using mutual information
- Tried different models: linear support vector regressor and tree based models: random forest, ada boost, and gradient boosting regressors
- Metric performance: mean-squared error (MSE)

Final Regression Model

- Final Models selected for each type of violation:

Low: Label encoding + feature selection + gradient boosting

MSE: 5.83 - R2: 0.24

Medium: One hot encoding + feature selection + gradient boosting

MSE: 1.30 - R2: 0.16

High: One hot encoding + feature selection+ gradient boosting

MSE: 0.76 – R2: 0.13

The models do not fit well the data; regression might be a difficult task

Classification Approach

- Each inspection is classified:

  1: Fail, at least 1 high violation or more than 2 medium violations

  0: Pass, otherwise

- Trained different classification problems with or without feature selection

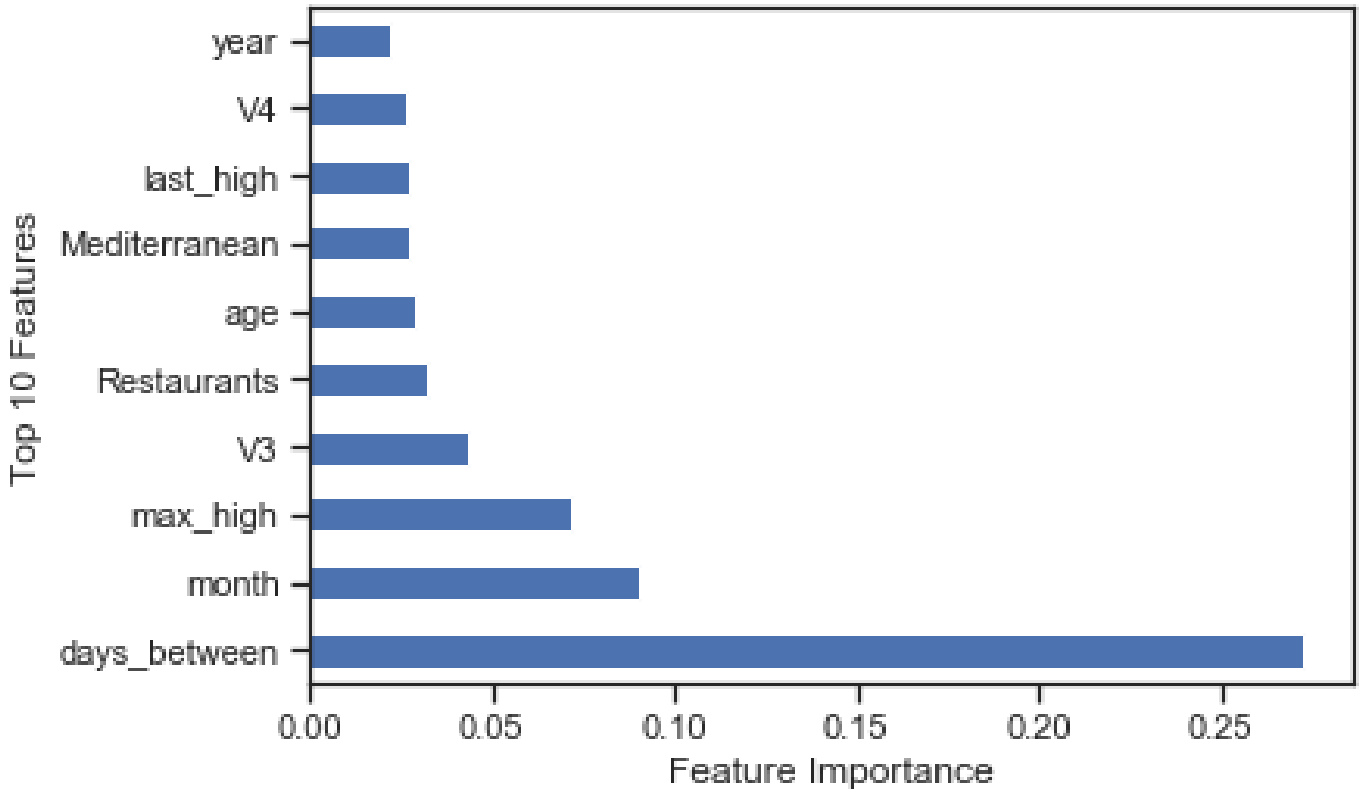
- Performance Metric: accuracy

Final Model Performance

- Final Model: Label Encoding + Gradient Boosting (without feature selection)
- Performance: accuracy of 67.65%

Class: 1 (Fail)	Class: 0 (Pass)
Precision: 0.67	Precision: 0.68
Recall 0.74	Recall: 0.60

Feature Importance



Possible Future Works

- Aggregate different types of classification models in addition to the gradient boosting classifier
- Use additional data from Allegheny website for restaurants' inspections with no violation
- Explore the reports of violation to extract more information related to the inspection.

Recommendation

- Collect more information related to the inspected place: number of employees, their level of expertise, their culinary background, information about the manager, size of the business, busy hours

- Collect more attributes of restaurants (many of the attributes were missing from the data)

- Encourage people to leave more reviews related to the level of cleanness noticed in the restaurants (restrooms, utensils, quality of food served in terms of freshness)