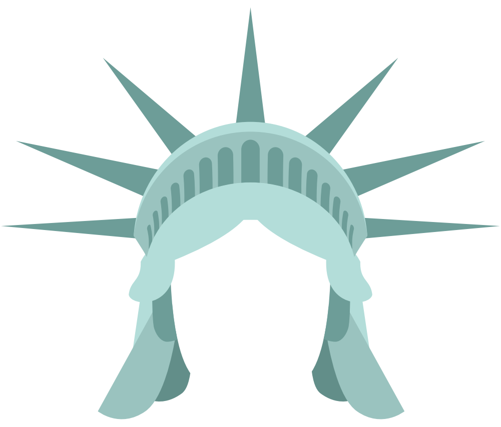
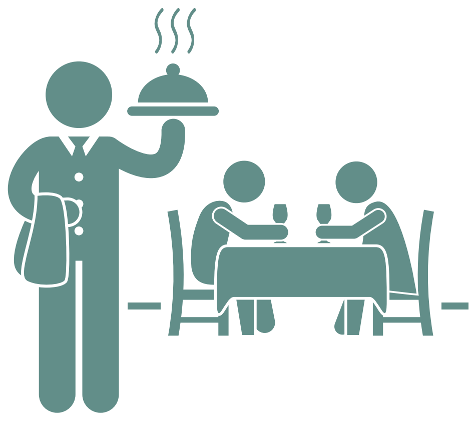
THE DRIVING FACTORS BEHIND THE RODENT INSPECTION RESULTS FOR NEW YORK CITY RESTAURANTS



New York City is heavily infested, with over 2 million rats inhabiting its streets.

New York City Health Department spends about $3 million each year to fight rodents.



Over 60% of the New York City restaurants confirmed rodent presence at some point, representing a major problem.

GOAL

Identify the factors influencing the result of a rodent inspection for New York City restaurants and explore the potential existence of causal relationships.

Software tools used: Python, R, Tableau
Statistical Models: Logistic regression

FACTORS CONSIDERED

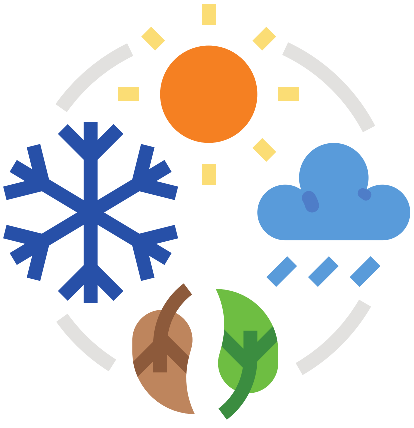
- **Borough**
  (Manhattan, Brooklyn, Queens, Bronx, Staten Island)
- **Seasons**
  (Spring, Summer, Fall, Winter)
- **Outdoor Dining**
  (Sidewalk, Roadway, or both)
- **Number of floors in the building**
- **Building age**
- **Restaurant size**
- **Garage size**
- **Building value**
- **Landmark building/location**

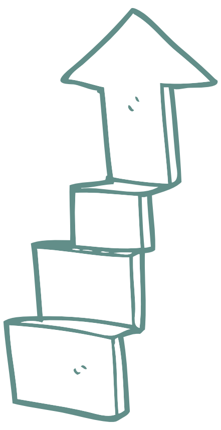
Why would these factors affect inspection results?

Borough: Since rodents often inhabit areas with high population density, the restaurant's location plays a role in the probability of rodent presence in it. Therefore, highly populated boroughs such as Manhattan might record higher rodent activity relative to Staten Island.

Outdoor dining: Potentially affects the hygiene of the space due to smell and possible food being left on the floor, therefore, attracting rodents.

Season: As rodents get more active in warmer weather, the spike in their activity might lead to a higher number of restaurants failing a rodent inspection.





Number of floors in the building: Taller buildings tend to be more occupied, They are harder to maintain and clean, creating potentially ideal conditions for rodents.

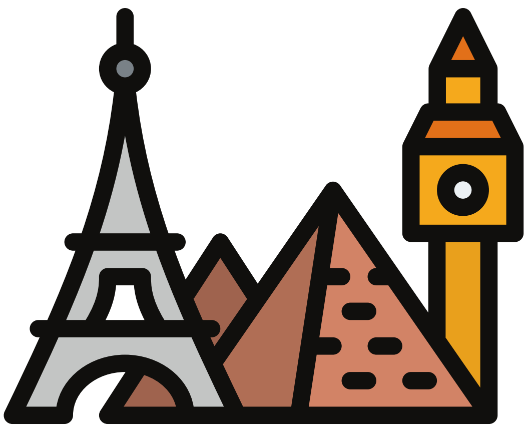
Restaurant size: Similarly to the buildings, but more specifically, restaurants could be more likely to inhabit large restaurants compared to the small ones.

Garage size: Garages are often also storage areas that are not frequently maintained and cleaned, an ideal environment for rodents.

Building value: Buildings with a higher value are usually in a better condition and kept in a better hygiene state.

Building age: On the contrary, older buildings are often in relatively bad condition and are expected to attract more rodents.

Landmark building/district: Presumably, such places have more rodents because more people visit them for their historical value.



DATA

We used three datasets for the analysis.

1

The first data set contains information on the **rodent inspections in New York City**.

The second one, **Open Restaurant Applications**, is a dataset of applications from food service establishments seeking to place outdoor seating in front of their business under the Reopening Phase.

2

3

The third one **details extensive land use and geographic data** at the tax lot level.

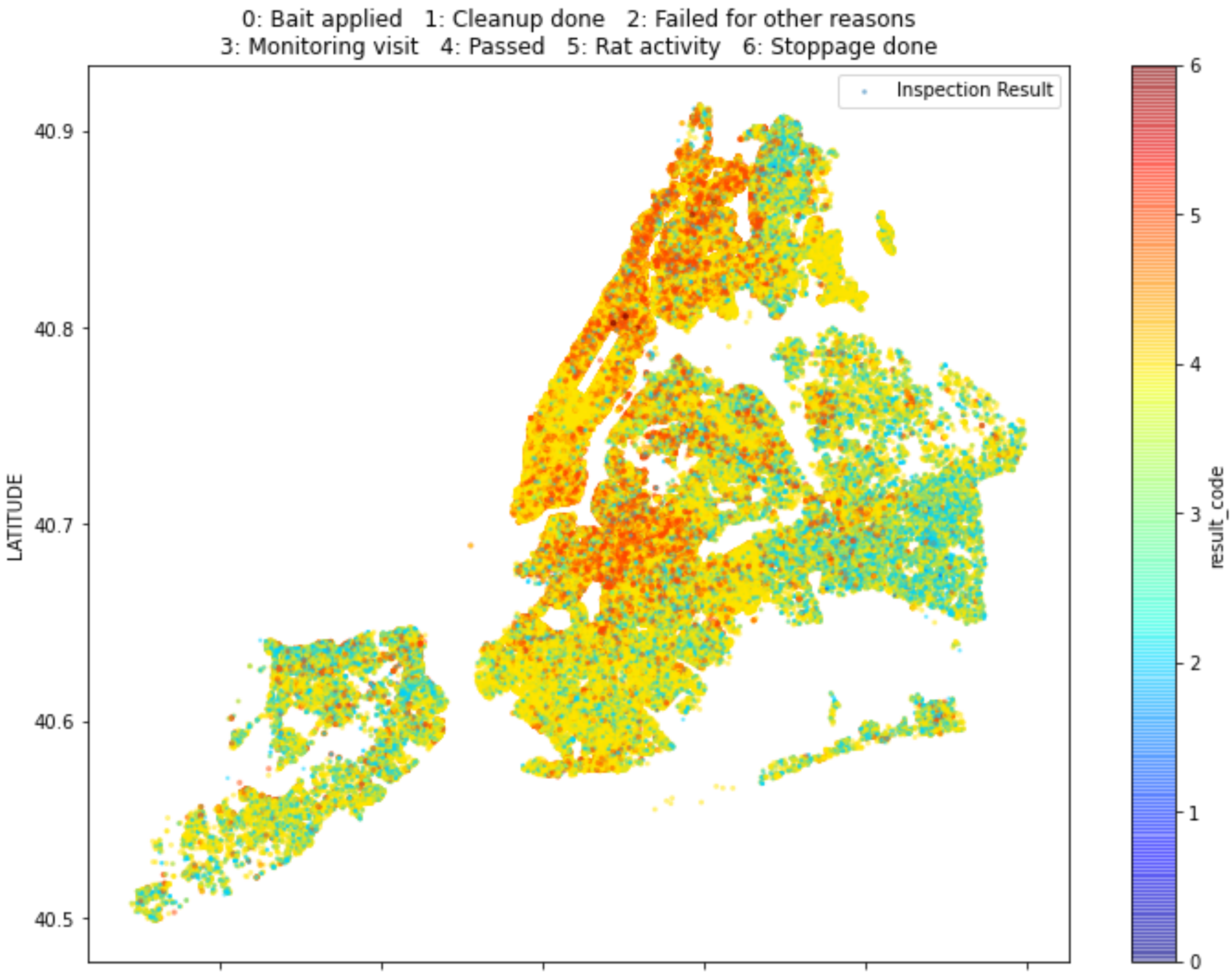
Steps that we took to get the final dataset;

- Selected the previously identified columns of interest.
- Merged the datasets with an inner join by the unique identifier for each tax lot in New York City ("BBL" column).
- Cleaned the combined dataset by removing and filling in the missing values appropriately and formatting the column types.
- Created binary columns for each categorical column of interest.

EXPLORATORY ANALYSIS

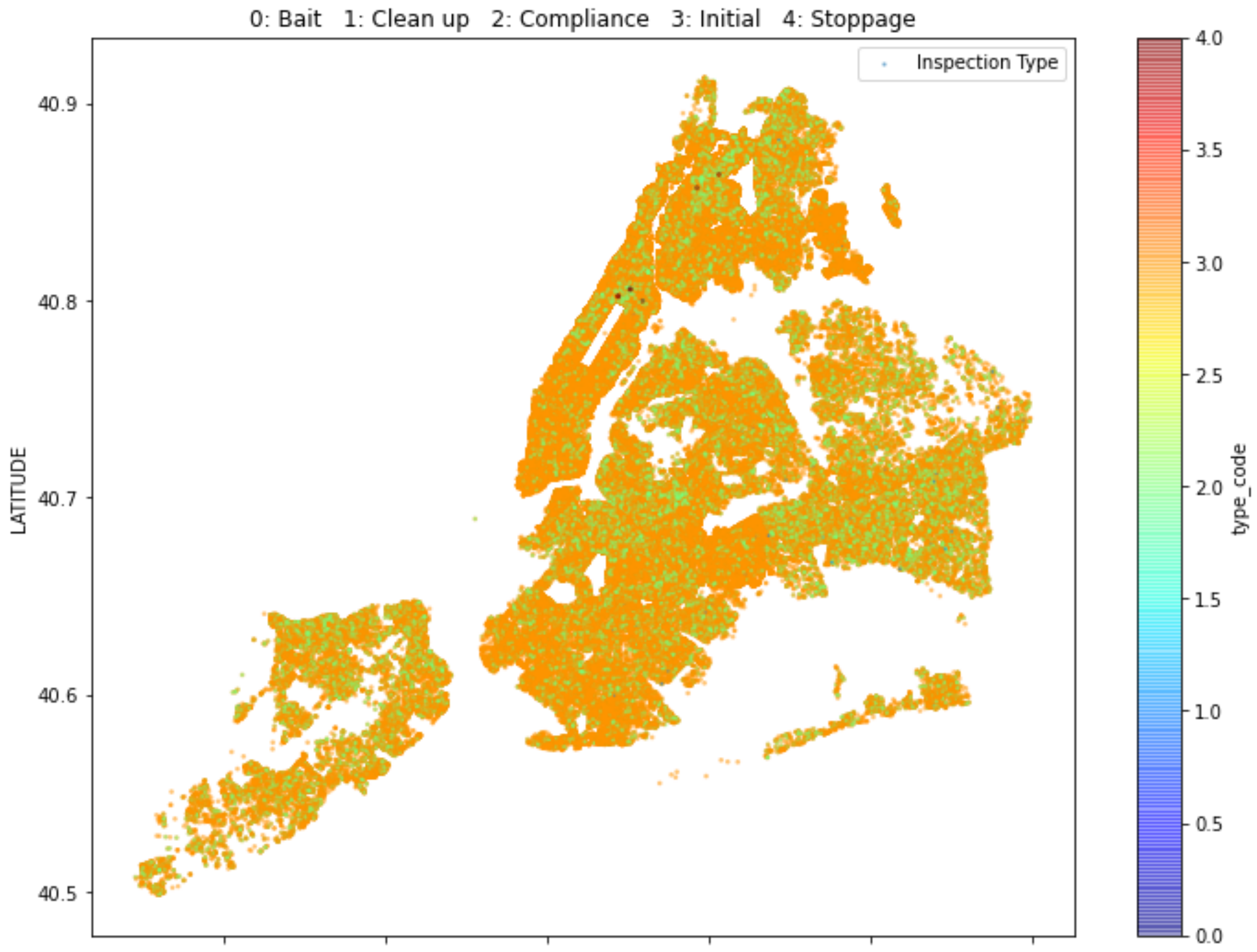
Let's first look at the heat map of inspection results

- Restaurants in Manhattan, Queens, and Brooklyn that have high population densities have the highest activity
- Restaurants further away in Queens have the most monitoring visits as inspection outcomes.
- Restaurants further in Brooklyn often pass a rodent inspection.

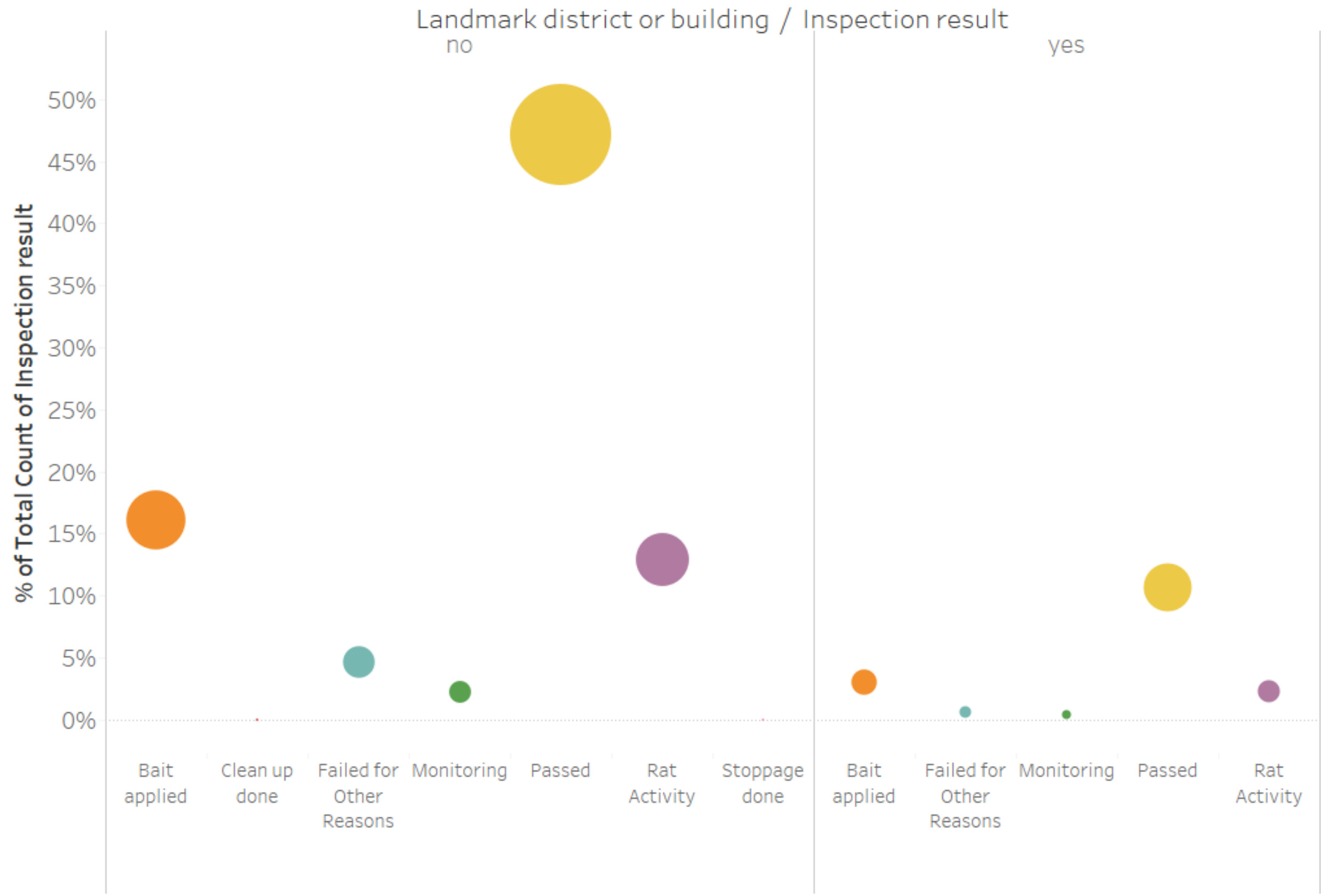

0: Bait applied 1: Cleanup done 2: Failed for other reasons
3: Monitoring visit 4: Passed 5: Rat activity 6: Stoppage done
Inspection Result
LATITUDE
40.9
40.8
40.7
40.6
40.5
result_code
0
1
2
3
4
5
6

The heat map of inspection types suggests:

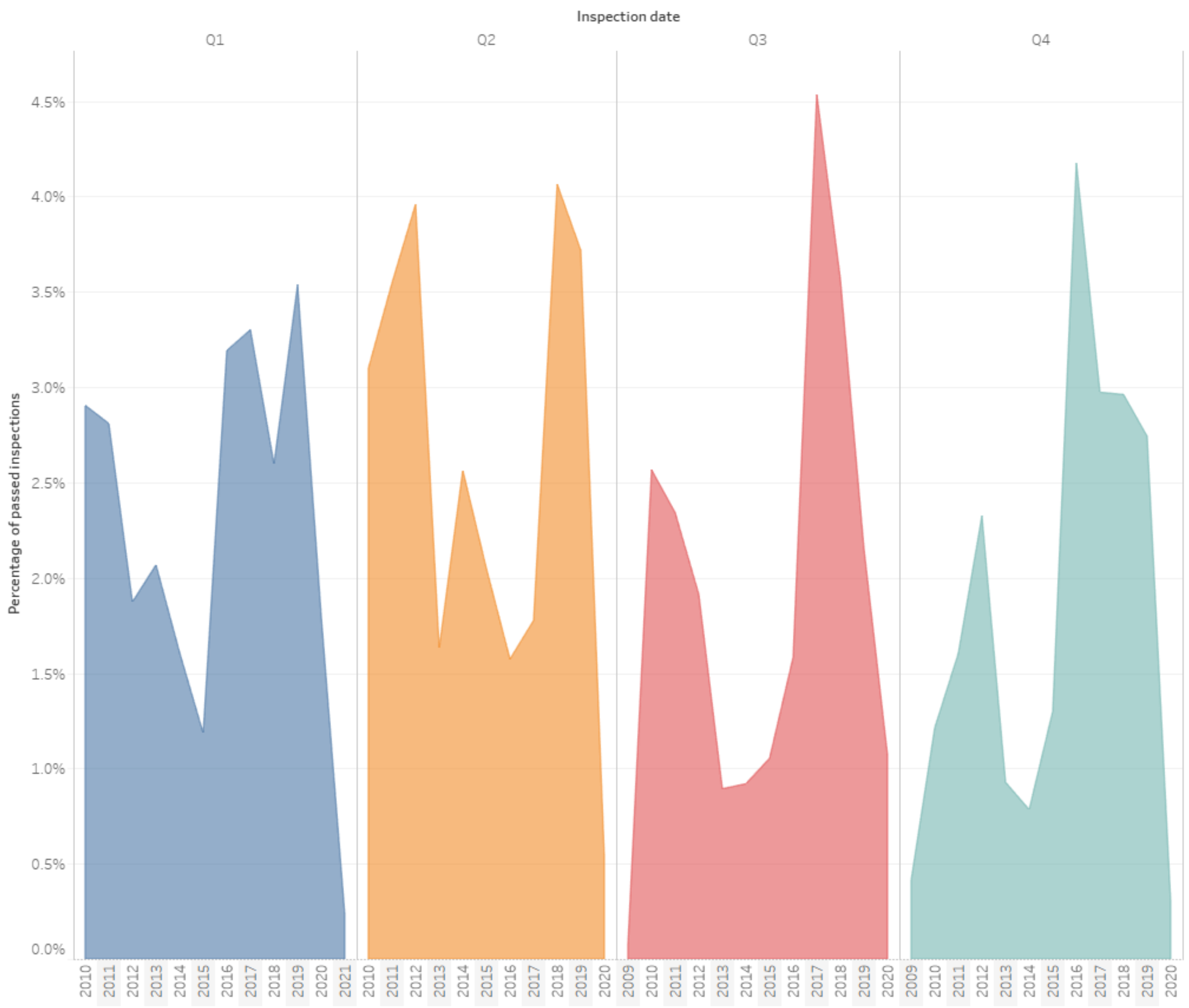
- The same restaurants in Queens that often get monitoring visits are the ones with the highest number of compliance inspections.
- The Staten Island restaurants get a decent number of compliance inspections, too.
- Initial inspections are most frequent.



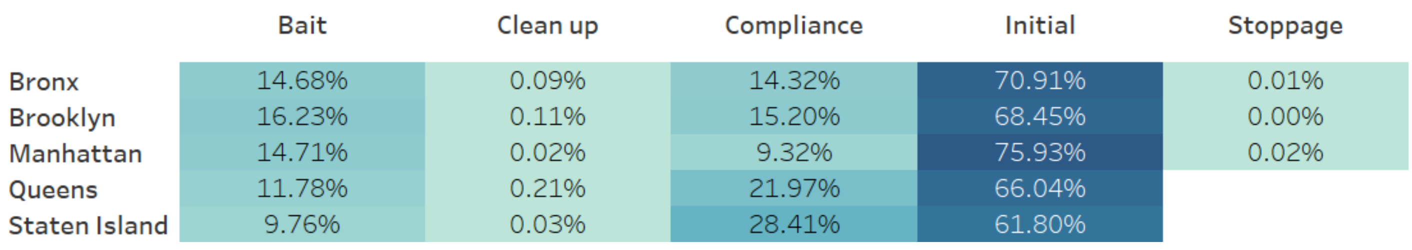
Restaurants located in districts or buildings that are **not landmark are much more likely to pass** a rodent inspection.



Over the years, **inspections conducted in Spring**, or the 2nd quarter, were **most likely to pass** a rodent inspection. However, since 2015, most of the inspections in late Fall and Winter also got the passing grade.

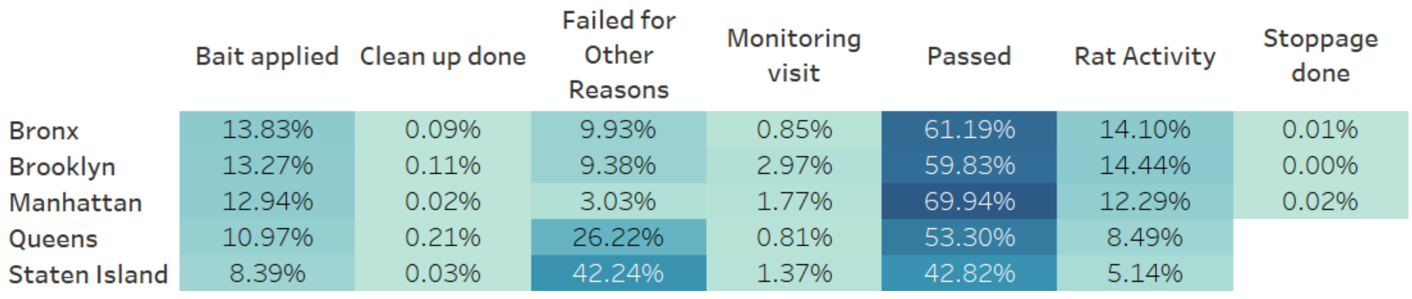


Although most inspections are the initial ones, Queens and Staten Island see between 1.5 and 3 times more compliance inspections than Manhattan, respectively.



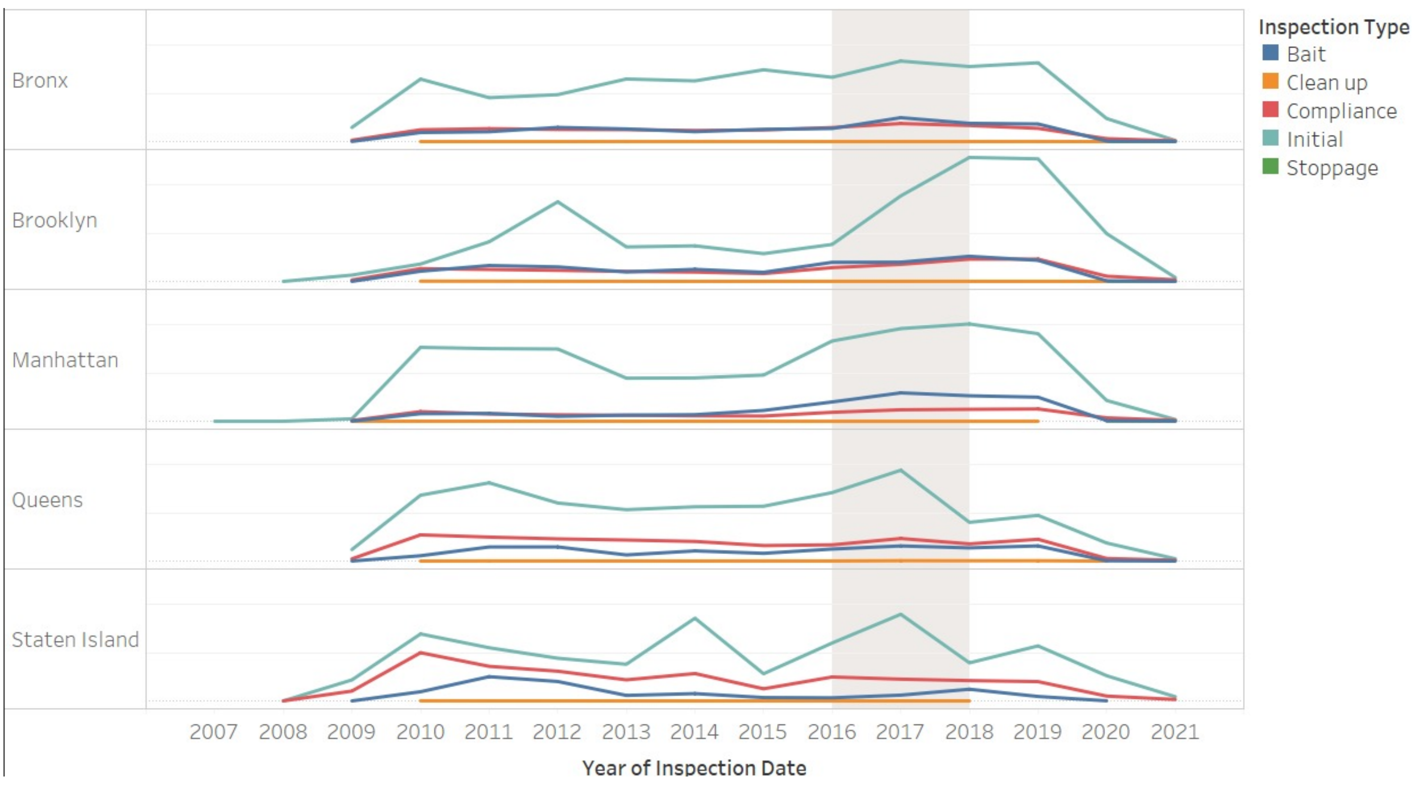
	Bait	Clean up	Compliance	Initial	Stoppage
Bronx	14.68%	0.09%	14.32%	70.91%	0.01%
Brooklyn	16.23%	0.11%	15.20%	68.45%	0.00%
Manhattan	14.71%	0.02%	9.32%	75.93%	0.02%
Queens	11.78%	0.21%	21.97%	66.04%	
Staten Island	9.76%	0.03%	28.41%	61.80%	

While most of the Manhattan, Bronx, and Brooklyn inspections get a passing grade, the ones in Queens are half as likely to fail an inspection for other reasons as they are to pass. Additionally, the Staten Island inspections are equally likely to fail for other reasons as they are to pass.



	Bait applied	Clean up done	Failed for Other Reasons	Monitoring visit	Passed	Rat Activity	Stoppage done
Bronx	13.83%	0.09%	9.93%	0.85%	61.19%	14.10%	0.01%
Brooklyn	13.27%	0.11%	9.38%	2.97%	59.83%	14.44%	0.00%
Manhattan	12.94%	0.02%	3.03%	1.77%	69.94%	12.29%	0.02%
Queens	10.97%	0.21%	26.22%	0.81%	53.30%	8.49%	
Staten Island	8.39%	0.03%	42.24%	1.37%	42.82%	5.14%	

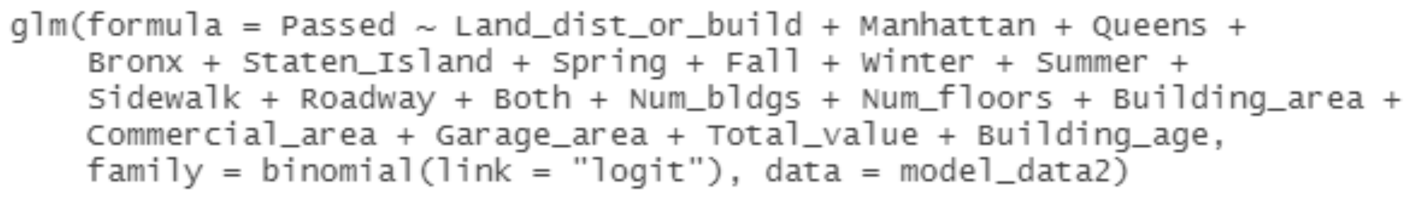
While initial inspections across boroughs remained steady or even declined between 2016 and 2018, Brooklyn saw a sharp rise in inspections of this type.



LOGISTIC REGRESSION MODEL

We used the R programming language to fit a logistic linear regression model. The model has the variable "Passed" as the dependent variable and multiple binary columns representing previously identified factors as the independent ones.

The R model:

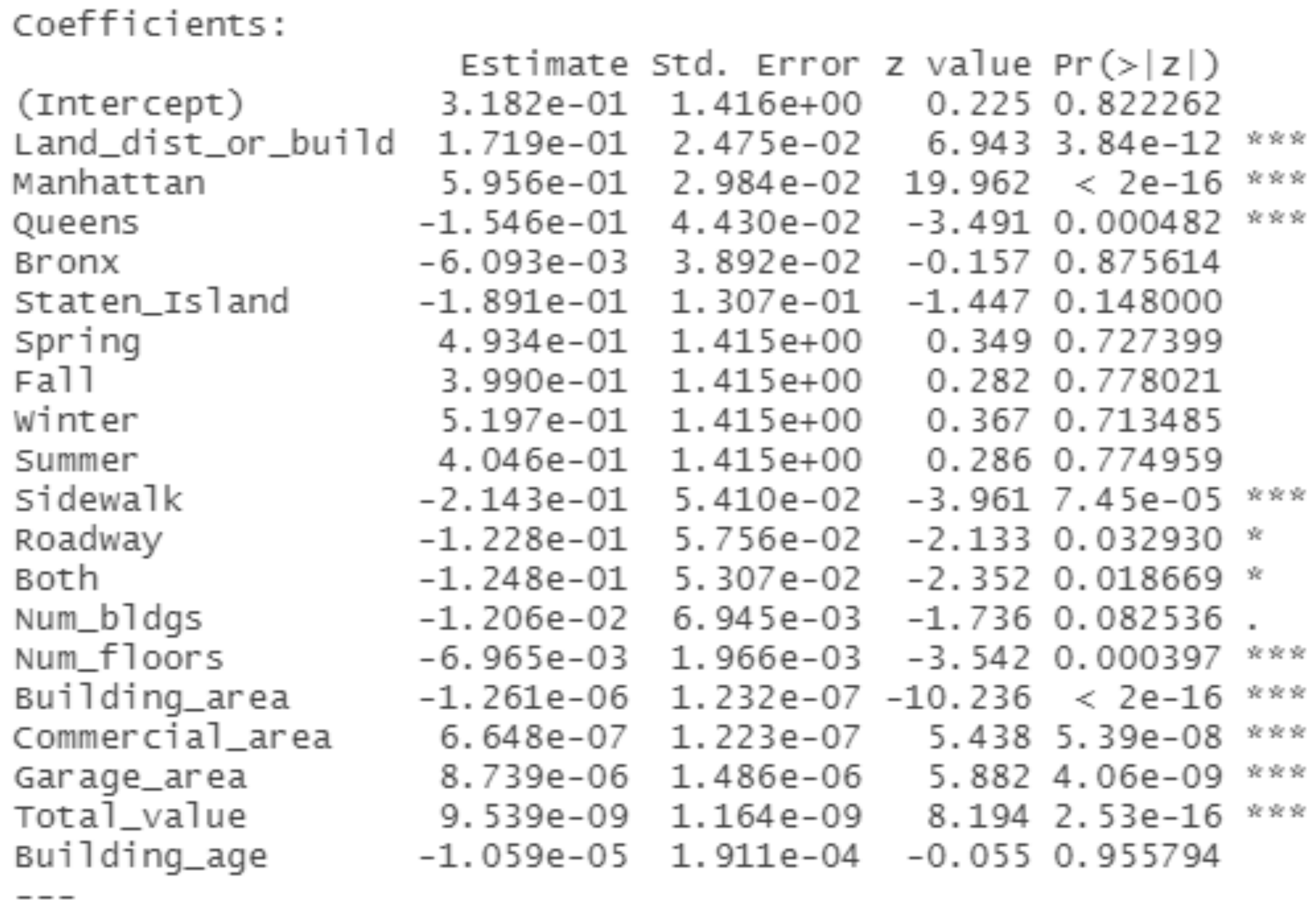


```
glm(formula = Passed ~ Land_dist_or_build + Manhattan + Queens +
    Bronx + Staten_Island + Spring + Fall + Winter + Summer +
    Sidewalk + Roadway + Both + Num_bldgs + Num_floors + Building_area +
    Commercial_area + Garage_area + Total_value + Building_age,
    family = binomial(link = "logit"), data = model_data2)
```

The model output helped us identify the variables that are statistically significant for determining whether a restaurant passes a rodent inspection.

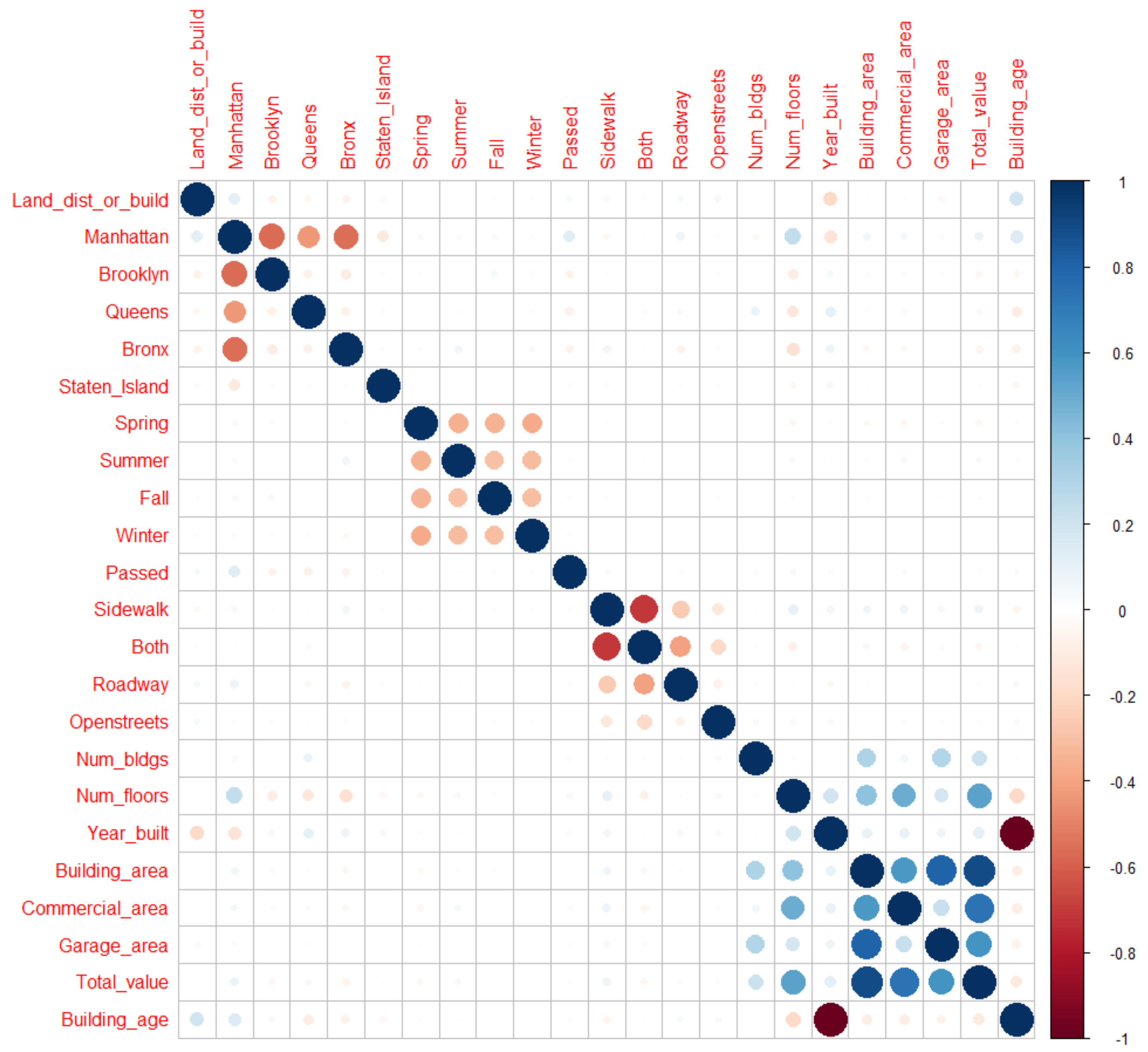
The highly significant variables:

- Landmark district or building
- Borough (Manhattan, Queens)
- Sidewalk seating
- Number of floors a building has
- The area the building is located in
- Whether the area is commercial or not
- Whether the building has a garage
- The value of the building



```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)         3.182e-01  1.416e+00   0.225 0.822262    
Land_dist_or_build  1.719e-01  2.475e-02   6.943 3.84e-12 ***
Manhattan           5.956e-01  2.984e-02  19.962  < 2e-16 ***
Queens             -1.546e-01  4.430e-02  -3.491 0.000482 ***
Bronx              -6.093e-03  3.892e-02  -0.157 0.875614    
Staten_Island      -1.891e-01  1.307e-01  -1.447 0.148000    
Spring              4.934e-01  1.415e+00   0.349 0.727399    
Fall                3.990e-01  1.415e+00   0.282 0.778021    
Winter              5.197e-01  1.415e+00   0.367 0.713485    
Summer              4.046e-01  1.415e+00   0.286 0.774959    
Sidewalk           -2.143e-01  5.410e-02  -3.961 7.45e-05 ***
Roadway            -1.228e-01  5.756e-02  -2.133 0.032930 *  
Both               -1.248e-01  5.307e-02  -2.352 0.018669 *  
Num_bldgs          -1.206e-02  6.945e-03  -1.736 0.082536 .  
Num_floors         -6.965e-03  1.966e-03  -3.542 0.000397 ***
Building_area      -1.261e-06  1.232e-07 -10.236  < 2e-16 ***
Commercial_area     6.648e-07  1.223e-07   5.438 5.39e-08 ***
Garage_area         8.739e-06  1.486e-06   5.882 4.06e-09 ***
Total_value         9.539e-09  1.164e-09   8.194 2.53e-16 ***
Building_age       -1.059e-05  1.911e-04  -0.055 0.955794    
---
```

Correlation Matrix



Our correlation matrix helps us identify a few things with our data:

- Most variables are not significantly correlated
- To a lesser extent, borough and season show a modest negative correlation
- Outdoor dining columns show a stronger negative coefficient
- The three area columns and the value indicate a positive correlation
- Building age shows the strongest negative correlation

OLS ASSUMPTIONS ANALYSIS

Establishing the existence of a correlation between variables is not enough to confirm causation on its own. Therefore, we address the four key OLS assumptions concerning our model.

1. The model **violates** the 1st OLS assumption that there are no other variables related to the probability of passing an inspection and any independent variables.
   For example, the subway is known as a desirable habitat for rats. Therefore, a restaurant's proximity to the subway can impact its probability of passing an inspection. Moreover, boroughs have unequally distributed access to the subway.

2. Although a New York City government agency collects the data, we do not know whether it represents a population or a sample. In case the latter is accurate, we do not know how this sample has been selected. Therefore, we **cannot confirm** that the 2nd OLS assumption has not been violated.

3. The plots indicate that there are outliers present in the data set. Hence, the model **violates** the 3rd OLS assumption.

4. The correlation matrix shows no two independent variables with a correlation of -1 or 1, meaning that the model does **not violate** the 4th OLS assumption.

CAUSALITY ANALYSIS

To explore whether the identified correlations between the variables reflect causality, we address the issues of omitted variables bias and reverse causality.

①

The model could suffer from the **Omitted Variables Bias** as there could be factors affecting the probability of passing a rodent inspection that is not included in the model. Such factors could be the previously-mentioned proximity to the subway, the cuisine restaurants serve, and similar.

②

There could exist **reverse causality** in the model. For example, even though the restaurants in Brooklyn most often pass an inspection, restaurant owners may decide to open restaurants in Brooklyn for this very reason.

Causality: ✗

CONCLUSION AND NEXT STEP

Findings:

- The analysis confirmed that the three data sets are not adequate for predicting the probability of a New York City restaurant passing a rodent inspection.
- However, it did reveal that factors such as restaurant location, the value and the size of the building it is located in, whether it has sidewalk seating, and whether it is situated in a landmark district or building contribute to the probability of passing a rodent inspection.
- The established correlation does not infer causation as the model suffers from the omitted variable bias and potentially reverse causality.
- Furthermore, the model violates some of the four key OLS assumptions.

Next step:

- To further improve the analysis, it would be necessary to obtain additional data. The data should help address the omitted variable bias issue and provide additional insights on other factors contributing to rodent inspections. Furthermore, we would like to assess the additional data with a decision tree that could provide more explanatory power than our logistic regression.