NYC Restaurant Yelp and Inspection Data Analysis

By: Rachel Spiro

OVERVIEW

- **Problem Statement**: Determine if certain aspects of NYC Manhattan restaurants can influence a restaurant's inspection grade or Yelp reviews and can predict a restaurant's Yelp rating.

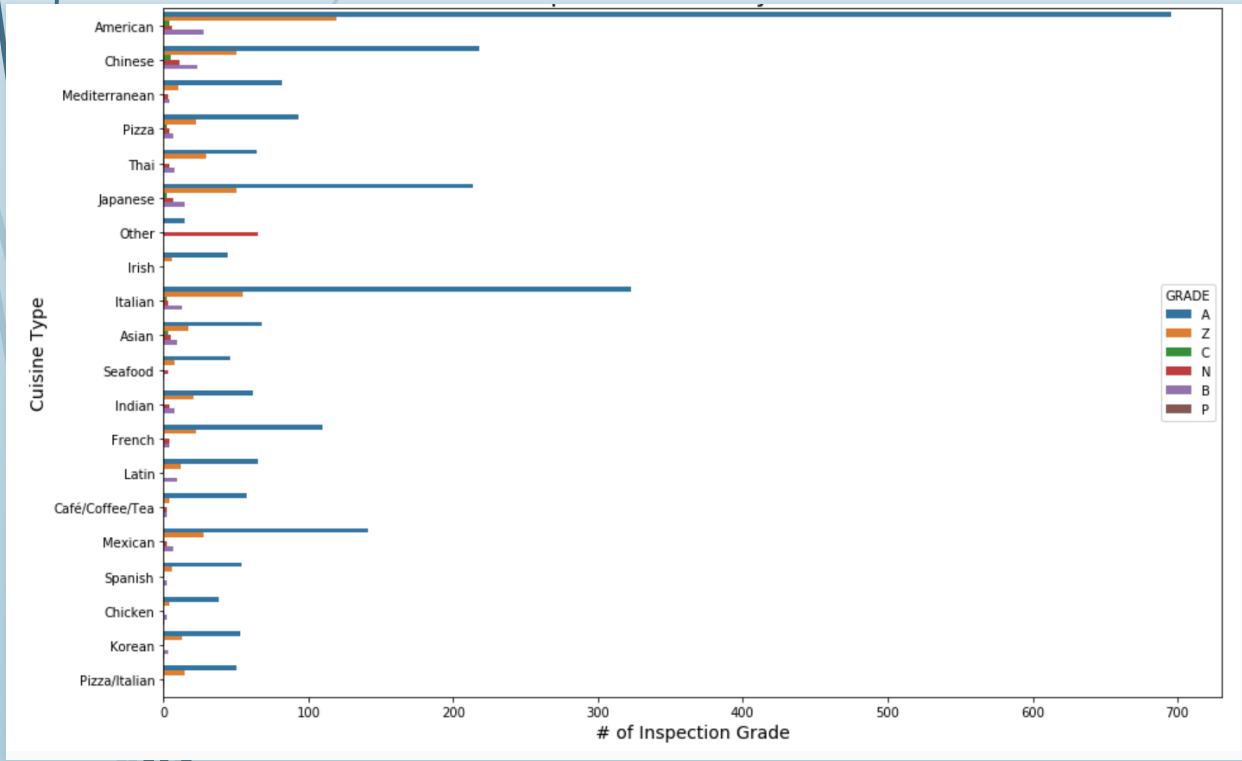
- **Business Value**: This analysis can be used by restaurateurs to identify certain aspects (i.e. cuisine type, location, price level, etc.) that they should consider optimizing in order to have a greater chance at receiving a higher Yelp rating with a greater number of positive Yelp reviews and receiving a higher inspection grade.

METHODOLOGY

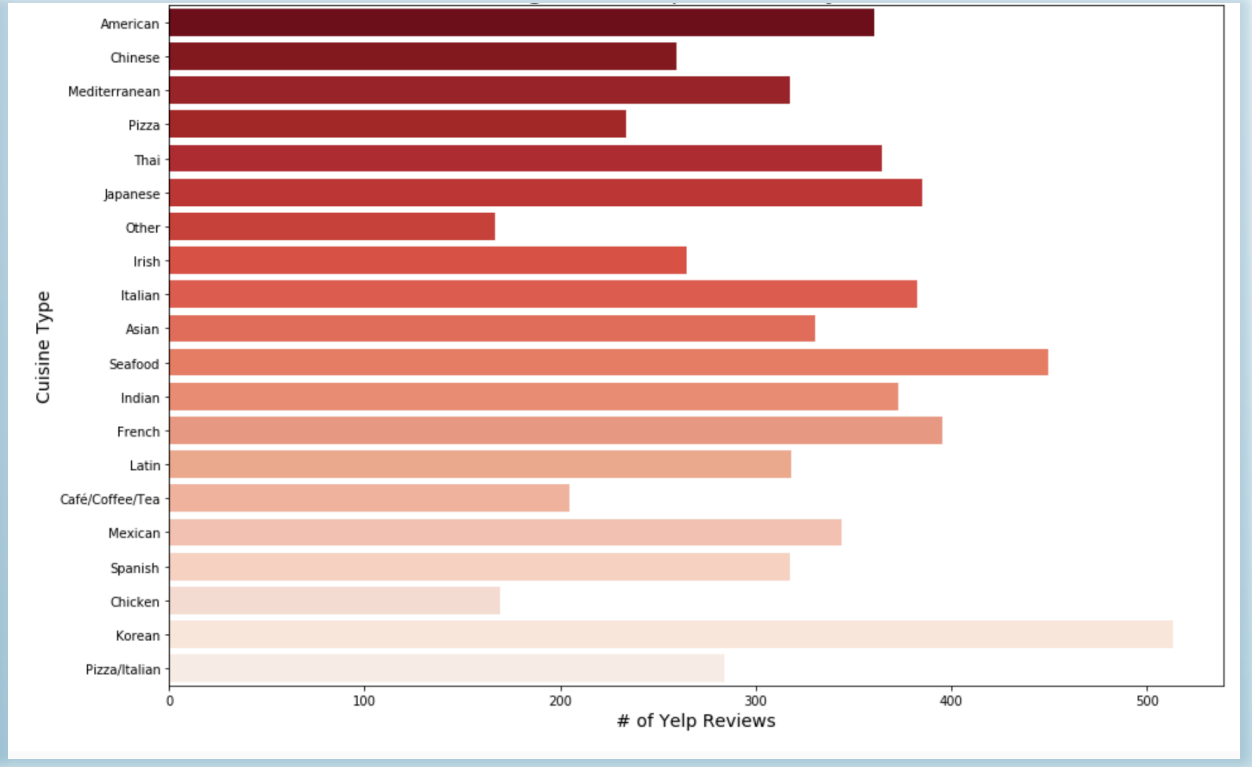
- Exploratory Data Analysis
- Hypothesis Testing - Used to test claims about a data parameter in order to determine what business decision should be made.
  - Welch's T-Test
  - ANOVA
  - Cohen's d
- Machine Learning - a way to run data analyses by using automated analytical models that have the capability to learn
  - Classification (5 models)

EDA: CUISINE TYPE

Inspection Grade By Cuisine Type

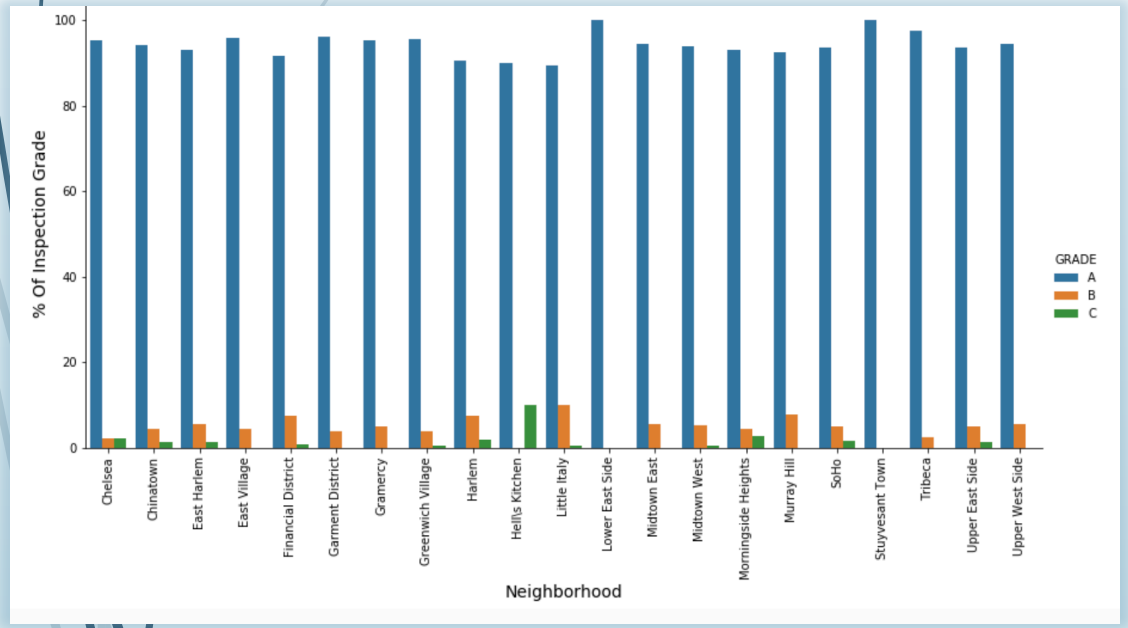


Average # Of Yelp Reviews By Cuisine Type

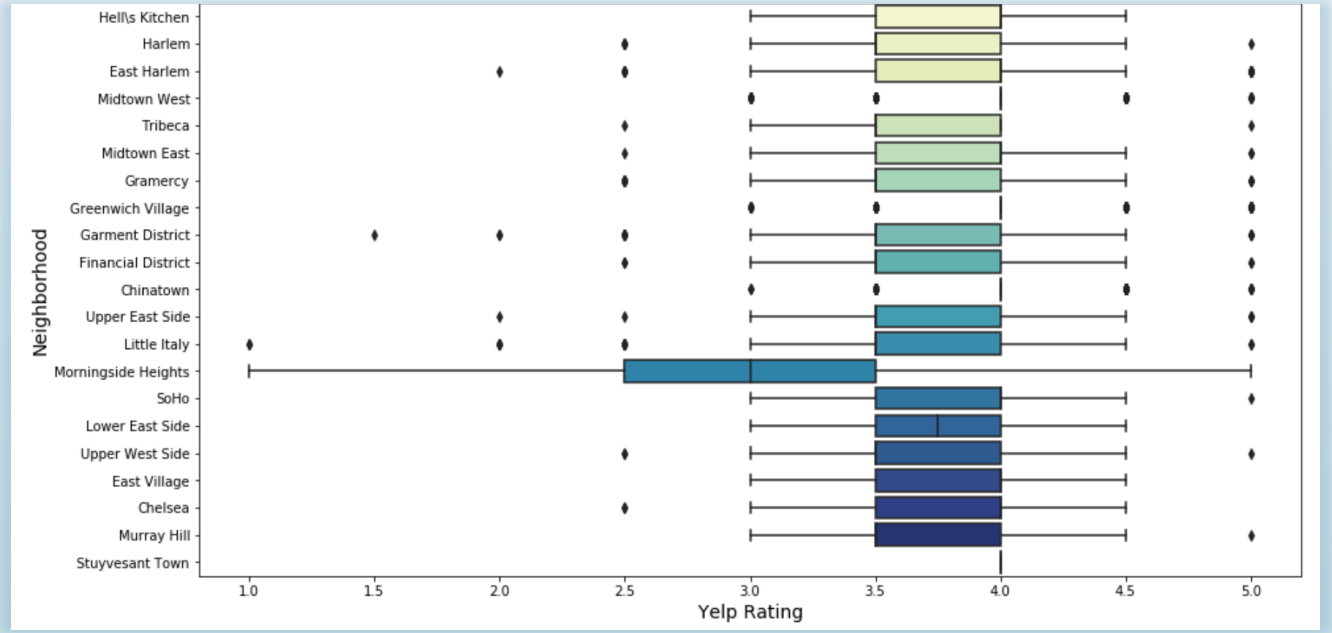


EDA: NEIGHBORHOOD

% of Inspection Grade By Neighborhood

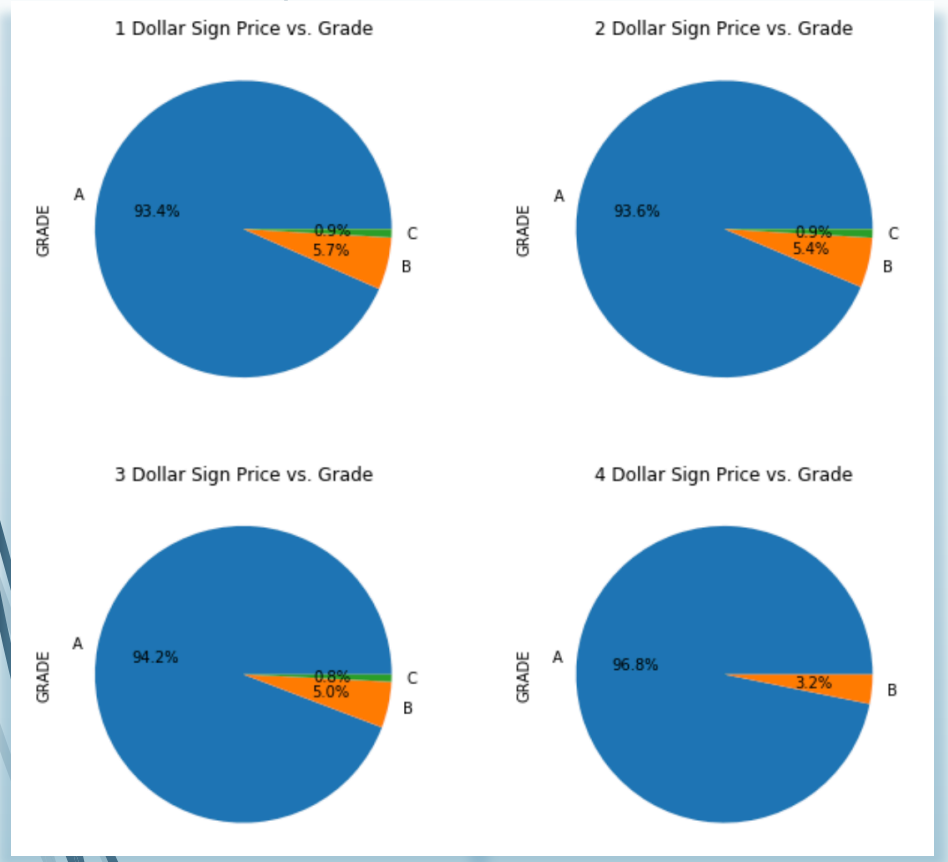


Yelp Rating Per Neighborhood

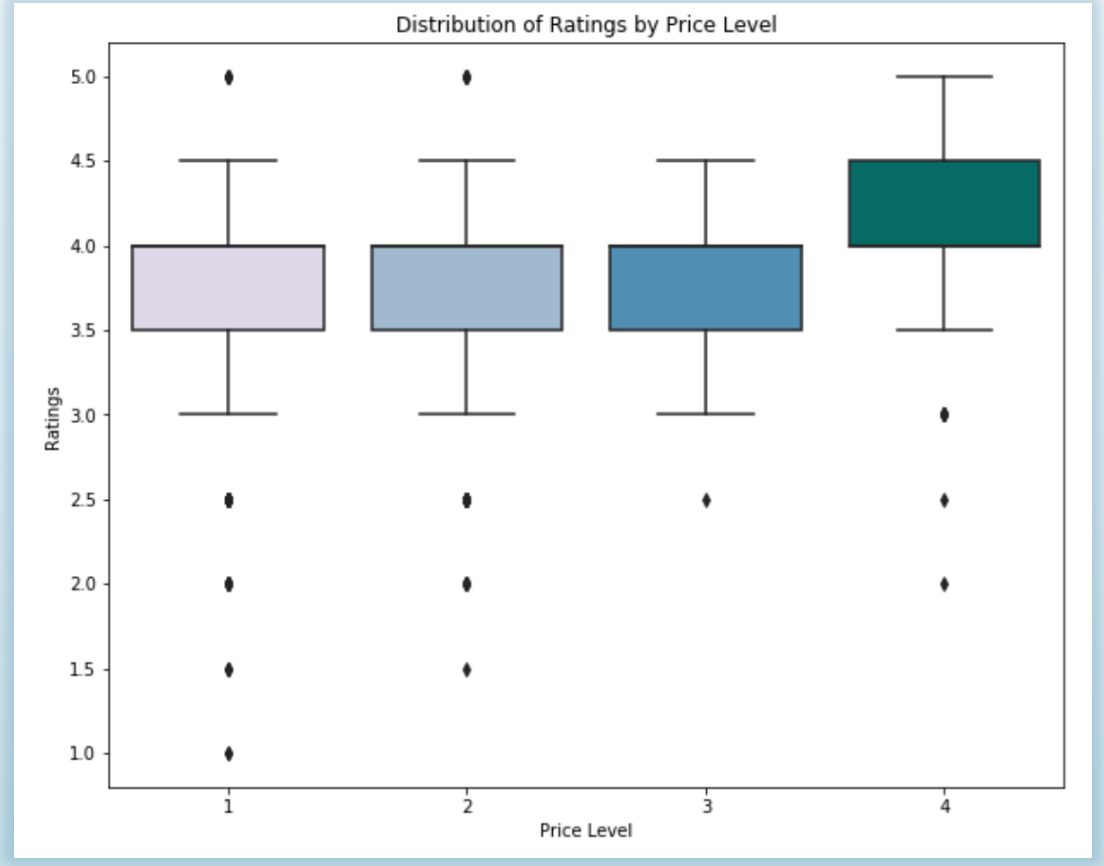


EDA: PRICE

Inspection Grade By Price

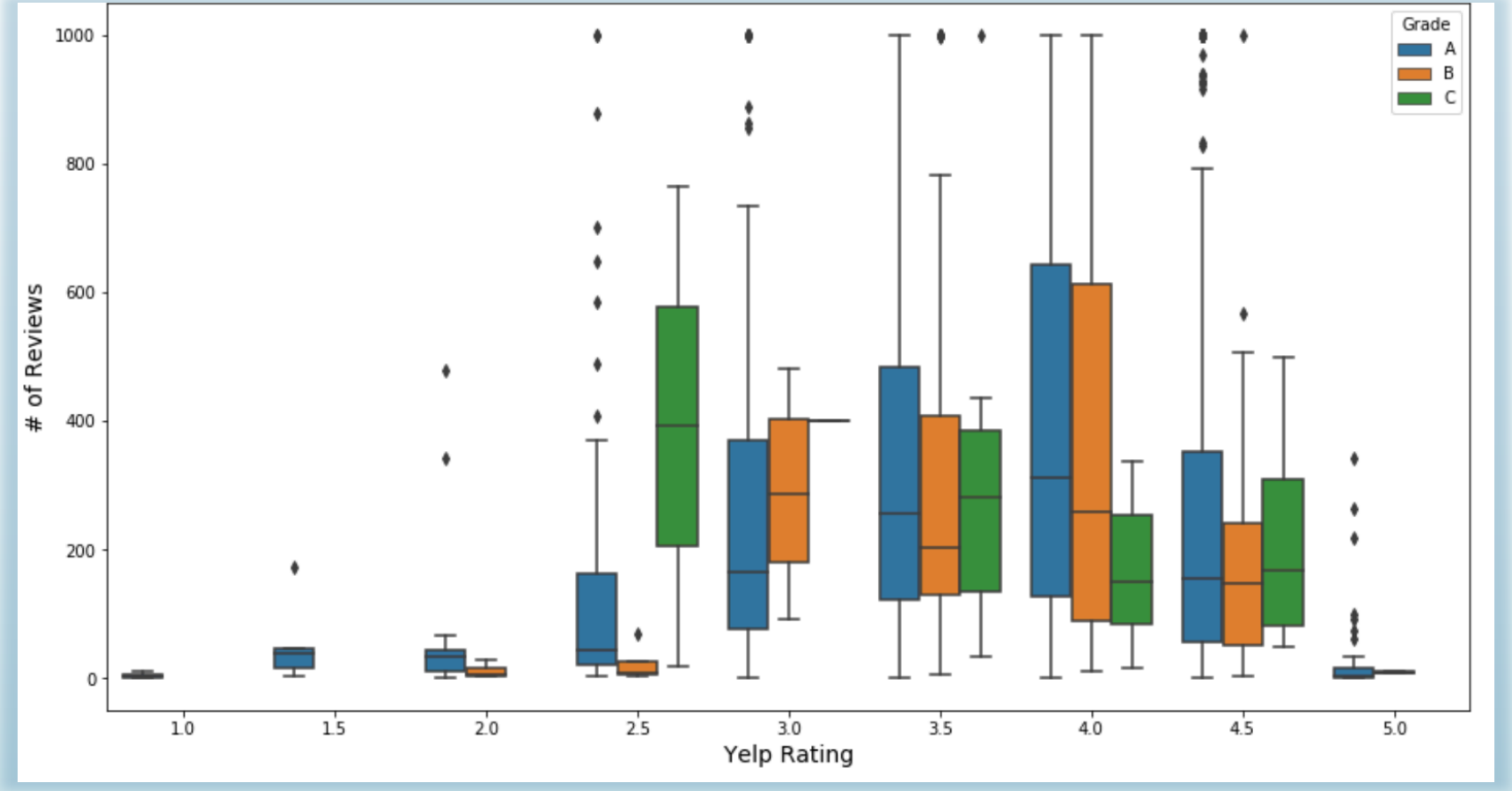


Yelp Rating Per Price



EDA: INSPECTION GRADE

Inspection Grade By # Of Reviews And Yelp Rating

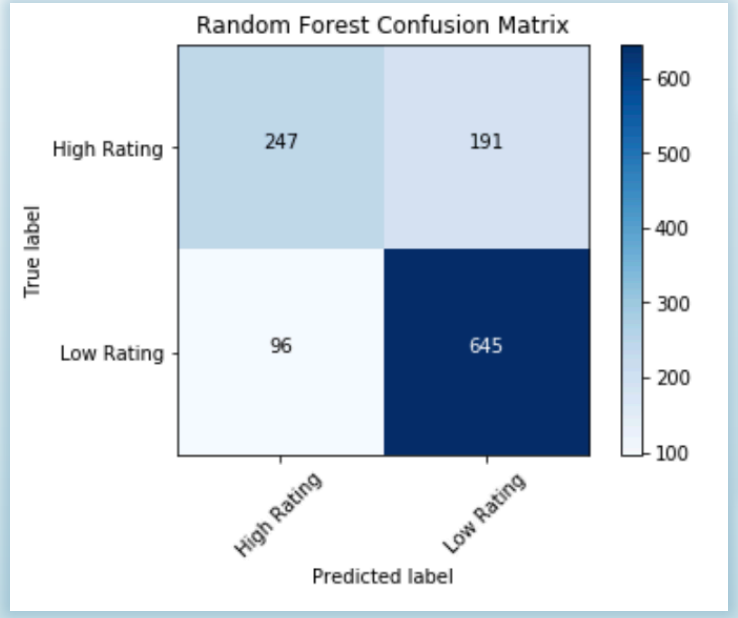


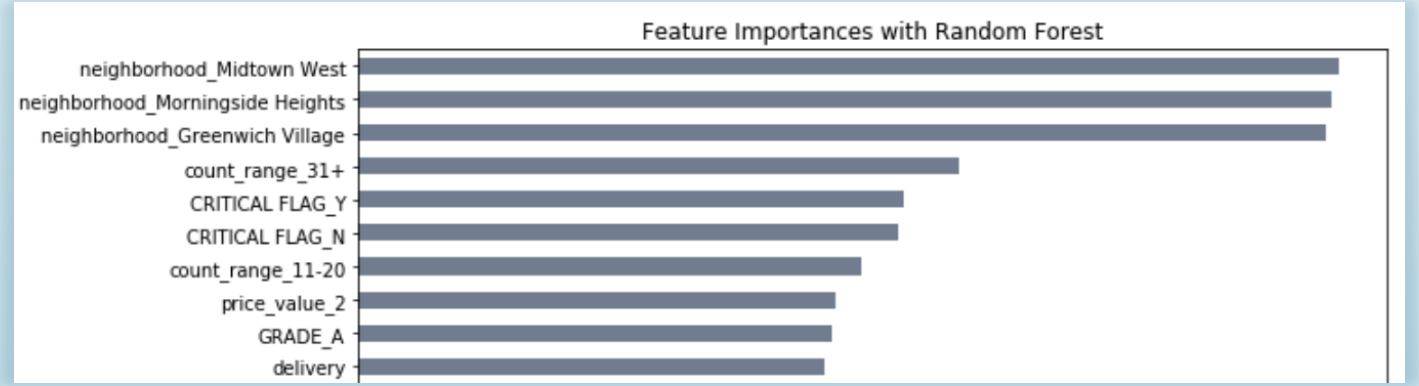
HYPOTHESIS TESTING

Question:	**Result**:
1. Does a restaurant's Yelp rating influence how many Yelp reviews the restaurant will receive?	➡ Reject null hypothesis ➡ P-value: 4.302e-05 ➡ Cohen's d: 0.125 (Small)
2. Does a restaurant's inspection grade influence how many Yelp reviews the restaurant will receive?	➡ Reject null hypothesis ➡ P-value: 0.0496 ➡ Cohen's d: 0.119 (Small)
3. Does the type of cuisine influence how many Yelp reviews the restaurant will receive?	➡ Reject null hypothesis ➡ P-value: 1.035e-25 ➡ Cohen's d: Small to Large
4. Is there a relationship between the Inspection Grade and the Neighborhood, Price, or Cuisine Type?	➡ Reject null hypothesis ➡ P-value (N): 0.1113 ➡ P-value (P): 0.0007 ➡ P-value (C): 0.0854 ➡ Cohen's d: Small

MACHINE LEARNING - CLASSIFICATION

- Decision Tree Accuracy: 70.23 %
- **Random Forest Accuracy: 75.66 %**
- Adaboost Accuracy: 74.72 %
- XGBoost Accuracy: 75.57 %
- Logistic Regression Accuracy: 74.64 %





CONCLUSIONS:

- Hypothesis Testing: Reject all 4 null hypotheses
- Machine Learning: Random Forest is our most accurate model with an accuracy of 75.66%.
  - Most influential features: some neighborhoods (Midtown West, Morningside Heights, and Greenwich Village), 31+ or 11-20 inspection grade visits, receiving/not receiving a critical flag, and $$ price value.

Recommendations:

- Consider a 4 dollar sign price level rather than a 2 dollar sign price level as these types of restaurants often receive better inspection grades.
- Ensure your restaurant is up to code and has minimal violations so that you are more likely to receive a better inspection grade, which can lead to receiving a greater number of reviews.
- Ensure customers have an enjoyable experience at your restaurant so that they will not only give a high Yelp rating, but will also leave a positive review.
- When trying to ensure a strong inspection grade, cuisine type and neighborhood do not play a significant factor, so no limitations need to be considered in respect to these two aspects.
- If possible, consider opening a restaurant in Midtown West, or Greenwich Village as restaurants in these neighborhoods tend to receive higher Yelp ratings, and avoid Morningside Heights.
- Avoid receiving a critical violation flag in an inspection as having one of these violations likely leads to lower Yelp ratings, while not having a critical violation flag likely leads to higher Yelp ratings.

NEXT STEPS

- Impact of price on number of Yelp reviews
- Impact of critical violation flag on number or Yelp reviews or inspection grade
- Impact of neighborhoods on Yelp ratings or inspection grades
- Re-run classification with an improved model (i.e. further tune model by testing additional hyperparameters).
- Expand analysis beyond Manhattan and look at other NYC boroughs to see if the trends seen in Manhattan shift or remain the same in other boroughs.

THANK YOU!

Questions/comments/concerns?