STAT 155 Written Report

Introduction

This project aims to explore the potential groups between pizza places within the United States. Variables such as location, rating, and price level can create natural groups amongst the restaurants depending on cluster characteristics.

The original dataset is from Jared Lander and Barstool Sports via Tyler Richards, and can be found under the tidytuesday GitHub repository. The preprocessed data contains 452 different pizza places across the United States.

Research Question

How do community review scores and all-review scores relate to price level and location across clusters?

Exploratory Data Analysis

My analysis consists of the variables 'city', 'price_level', 'review_stats_community_average_score', 'latitude' and 'longitude'. I was curious to find if there were potential clusters among expensive pizza places and their reviews, or if there were more groupings across cheaper places. Furthermore, I wanted to research if community reviews were more lenient than the average score of all reviews, potentially due to their loyalty to their community. I predicted that there will be more clusters among lower priced restaurants because I expect reviewers to have lower expectations as opposed to pricier options. I also wanted to explore the pizza place's location in relation to their reviews. I predicted that the city of New York would have more critical reviews due to their density of pizza places and stricter standards from tourists and locals for the perfect slice.

In the initial EDA, bivariate analysis was done to see potential relationships between community average score (range1:8) in relation to price level (range 0:3). It was found that the community average score was the highest (6) at a price level of 2.

Model

Clustering was the chosen method to analyze this data due to its ability to find hidden patterns amongst variables. This method of analysis can uncover natural groupings between pizza places and their ratings in relation to location.

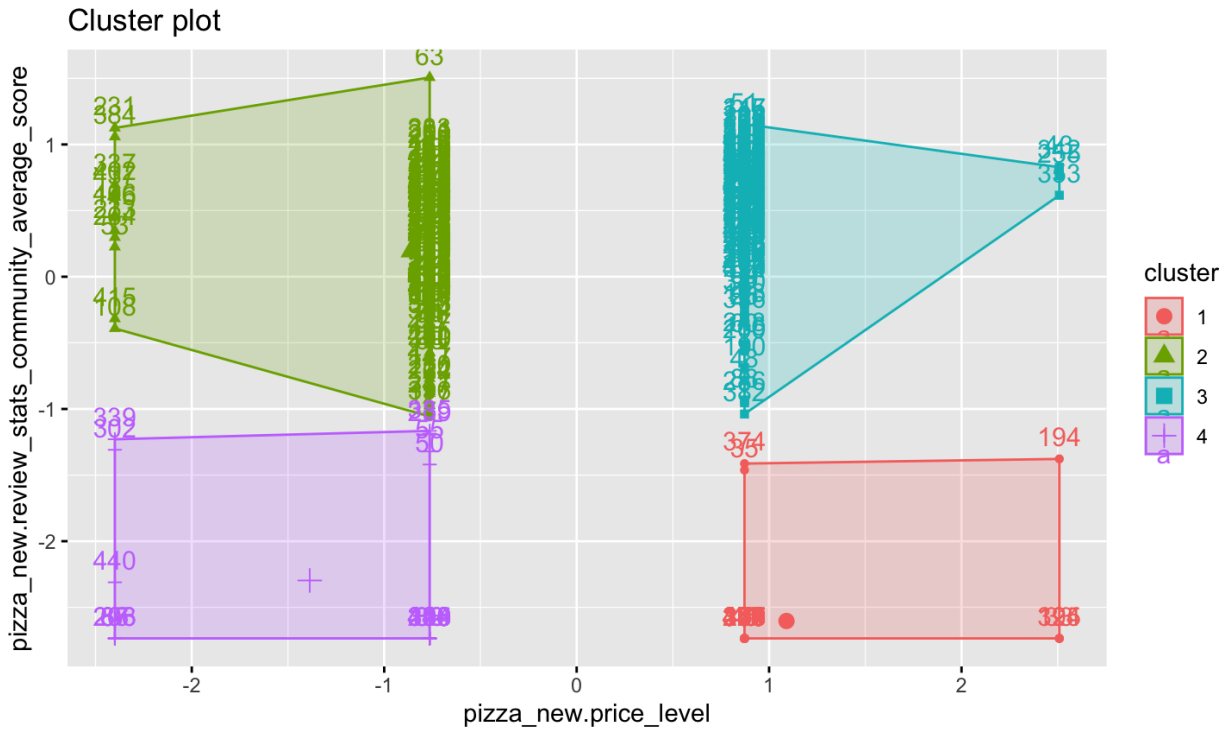
FCPS was the specific clustering package used to analyze this dataset. This package contains a wide variety of conventional clustering algorithms.

The analysis only refers to this specific dataset and cannot be generalized to all pizza places in the US. This dataset reflects heavily on reviews and opinions, which can be heavily biased by

emotion, varying levels of reviewer standards, and other outside factors. While reading the analysis, it is important to remember that these results are not definitive, especially with the nature of the data.

Analysis

1. **Price and score across all locations**



Cluster 1: High price level, low average score

- Why? Reviewers may have higher expectations for pricier places.

Cluster 2: Low price level, high average score

- Why? Reviewers may be less strict with ratings from a cheaper place.

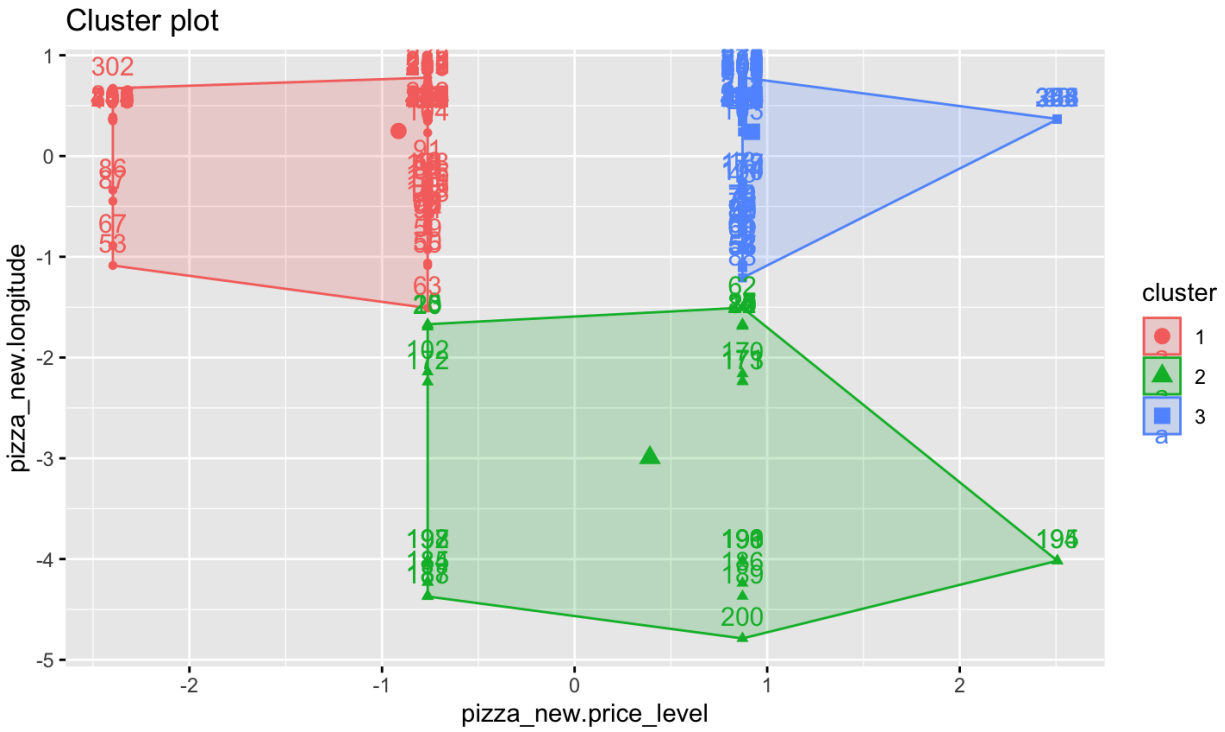
Cluster 3: High price level, high average score

- Why? Reviewers may have felt that the price of the pizza matched their expectations and standards.

Cluster 4: Low price level, low average score

- Why? Reviewers may have not had their standards met for a cheap slice.

2. Price and longitude



Cluster 1: Low price level, high longitude

- Why? Pizza places on the west coast have lower prices.

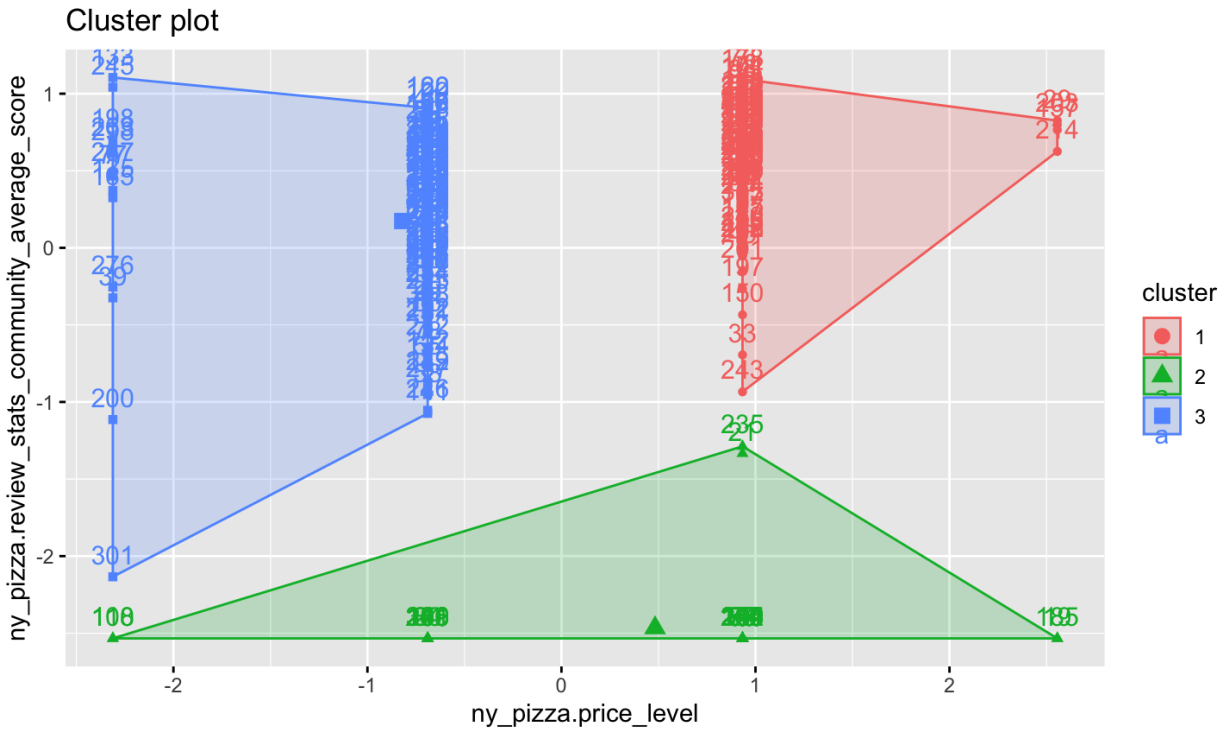
Cluster 2: Average to high price level, low longitude

- Why? Pizza places further inland and on the east coast tend to have higher prices. There seems to be a couple places that are on the very expensive end, causing a triangle shape at the end of the cluster.

Cluster 3: Average to high price level, high longitude

- Why? This cluster appears to be a triangular shape. The majority of the points are in the average (1) part of the plot in terms of price, but a few places are on the more expensive side. This could represent the pricier pizza places on the west side of the country.

3. Price and score in New York



Cluster 1: High price level, average to high average score

- Why? Reviewers in New York may be satisfied with their experience at the pizza place in relation to the price.

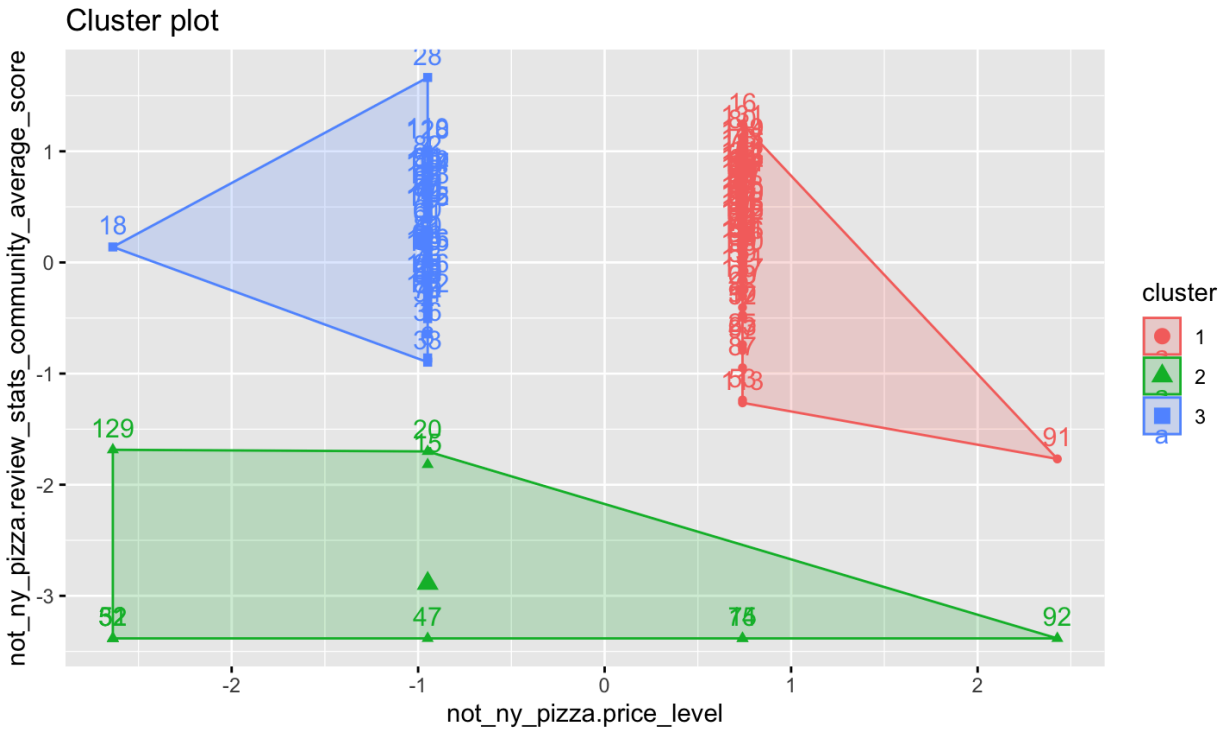
Cluster 2: Various price levels, low average score

- Why? The variability of this cluster's prices may be caused by New York reviewers having higher standards. Since New York is known for their pizza, consumers may approach these restaurants with greater expectations than other states. Therefore, the consistent low reviews across all price points may reflect a difference in opinion.

Cluster 3: Low price level, various average scores

- Why? Similar to the prediction in Cluster 2, this may be caused by varying levels of expectation going into the experience. A regular customer may rate a lower price level restaurant higher than a tourist who is trying it for the first time.

4. **Price and score outside of New York**



Cluster 1: Average to high price level, average to high average score

- Why? Overall, the main data points are congregated around the 0.75 level of price with one outlier at 2.5. It seems that most of the reviewers in this cluster have generally average ratings for pizza places at average price points.

Cluster 2: Various price levels, low average score

- Why? This could be because individuals outside of New York may have unequal standards in relation to their experience with the location. However, the consistent low scores could also happen because of a negative experience.

Cluster 3: Low price level, average to high average score

- Why? This cluster has one outlier marked very low on price and average on score. The rest of the data points indicate that this was a unique experience, since everything else lies almost 2 levels higher in price with varying scores. This could represent the reviewers who were generally satisfied with their food and experience.

Monte Carlo Simulation

To test the clustering model against other options, a simulation was done to test alternative methods of clustering.

The data-generating model that was used was the 'generate_data' function in R. In creating the different clusters, the membership of the data points needs to be identified. Membership was indicated by a "1" and a "0" indicated the absence of that data point in the cluster. The data was generated by repeating 0 and 1 100 times.

Mathematical Notation: $y = b1 + b2 + b3$

The assumptions were that all values are numeric and scaled consistently.

The calculations involved three different types of variance to account for noise, two separate ways. This creates six different treatments.

I hypothesized that higher variance will cause more clusters, because high variance indicates more variability and spread within a dataset.

Assessing the methods will be determined by the clusters that are formed, and if the algorithm got the melted group membership correct.

Simulation Plan

Detail how the experiment will be carried out:

- Number of simulations: Several, so that we get enough repetitions to achieve statistically different numbers
- Parameter settings: Choosing the type of clustering, selecting the method that calculates distance, numbers of clusters
- What will be recorded: Clusters that are naturally formed
- Any changes from Project III's code:
    - Add data-generating model
    - Change initial values of clusters

Simulation Limitations

To have clustering be an efficient method of analysis, an optimal number of clusters must be found. That may be difficult with the added variance and noise. I also anticipate analyzing the clusters for tangible results would be difficult, at least with my current understanding.