# Final Project

By: Audrey Moon

# Introduction

**Research question:** How do review scores relate to price level and location across clusters of pizza places in the US?

**Motivation:** Interests in finding patterns, but already had experience in linear regression

**Context:** Data consisting of pizza places across the United States, specific variables chosen were related to price, location, and average ratings

**Why it matters to me:** I am interested in UI/UX as a future career, and want to understand how to manipulate data in real-world settings to make real-world changes.

# Data Wrangling

**Data source**: Tidytuesday Repository on GitHub, from Jared Lander and Barstool Sports via Tyler Richards

**Wrangling challenges**: Had to remove duplicate pizza places

**Tools & methods considered**: libraries- tidyverse, naniar, superheat, patchwork, gplots, psych

# Exploratory Data Analysis

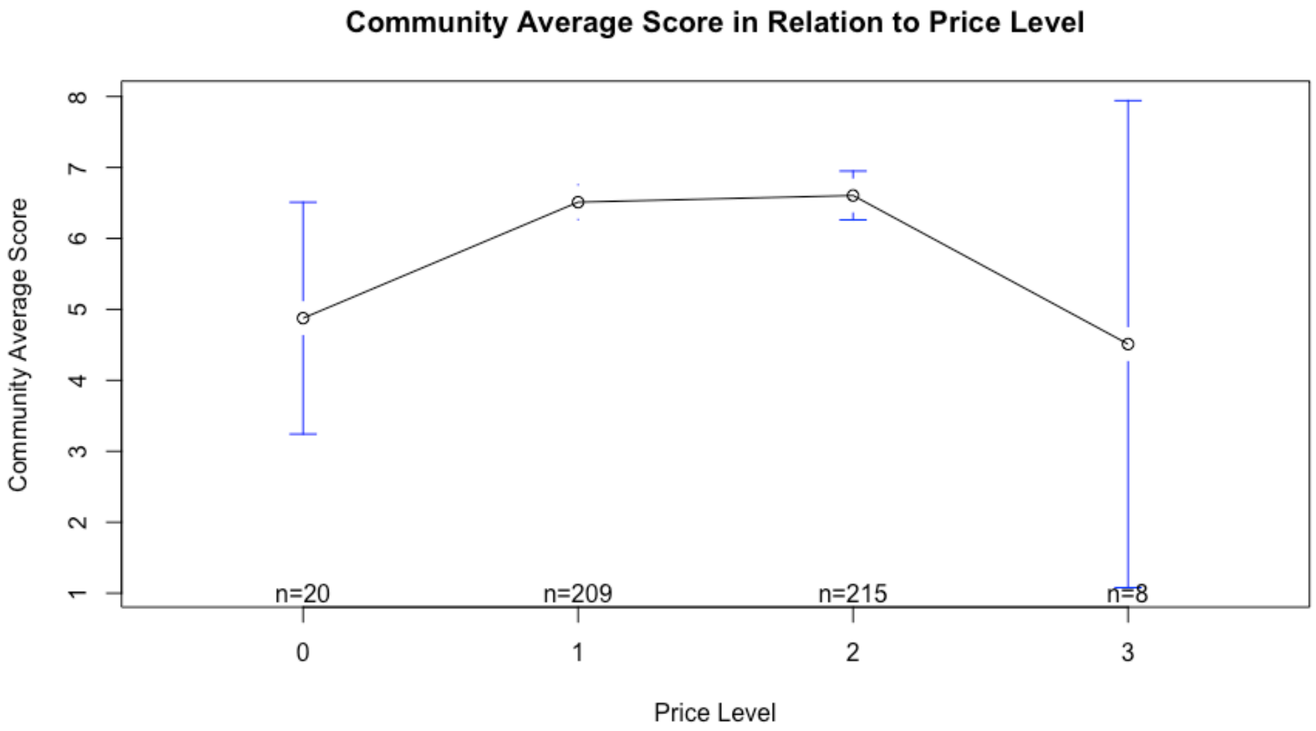
**Variables of Choice:**

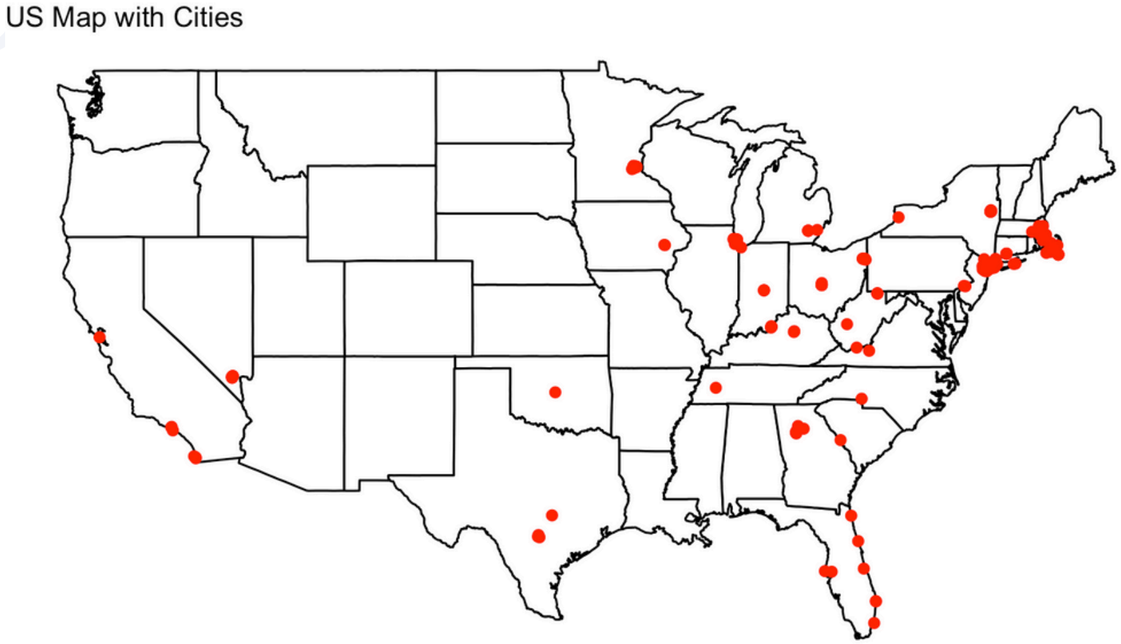
- '**latitude**' : Latitude coordinate of pizza place
- '**longitude**' : Longitude coordinate of pizza place
- '**review_stats_all_average_score**' : Average score
- '**review_stats_community_average_score**' : Community average score
- '**price_level**' : Price rating (fewer \$ = cheaper)

**Interesting Relationships:**

- Bivariate graph using “Community Average Score” and “Price Level”
  - On a 0-3 price level scale, 2 was most common
  - On a 1-8 score scale, ~6.5 was most common

# Exploratory Data Analysis

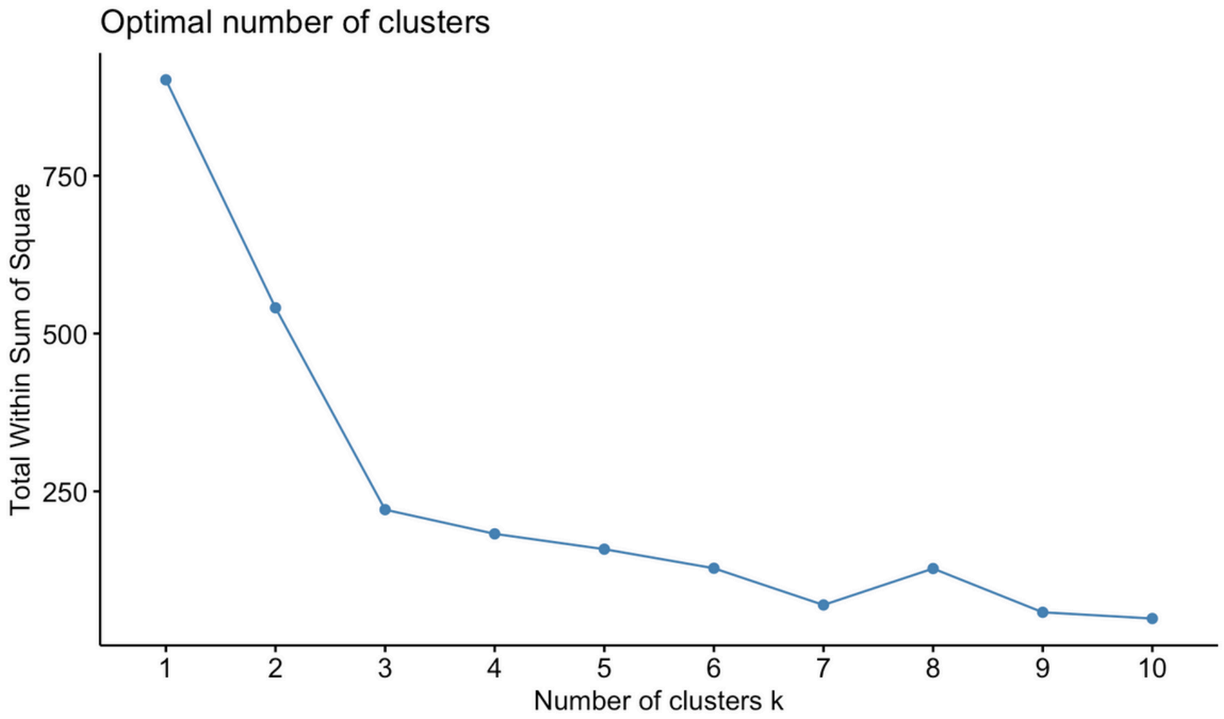


# Modeling

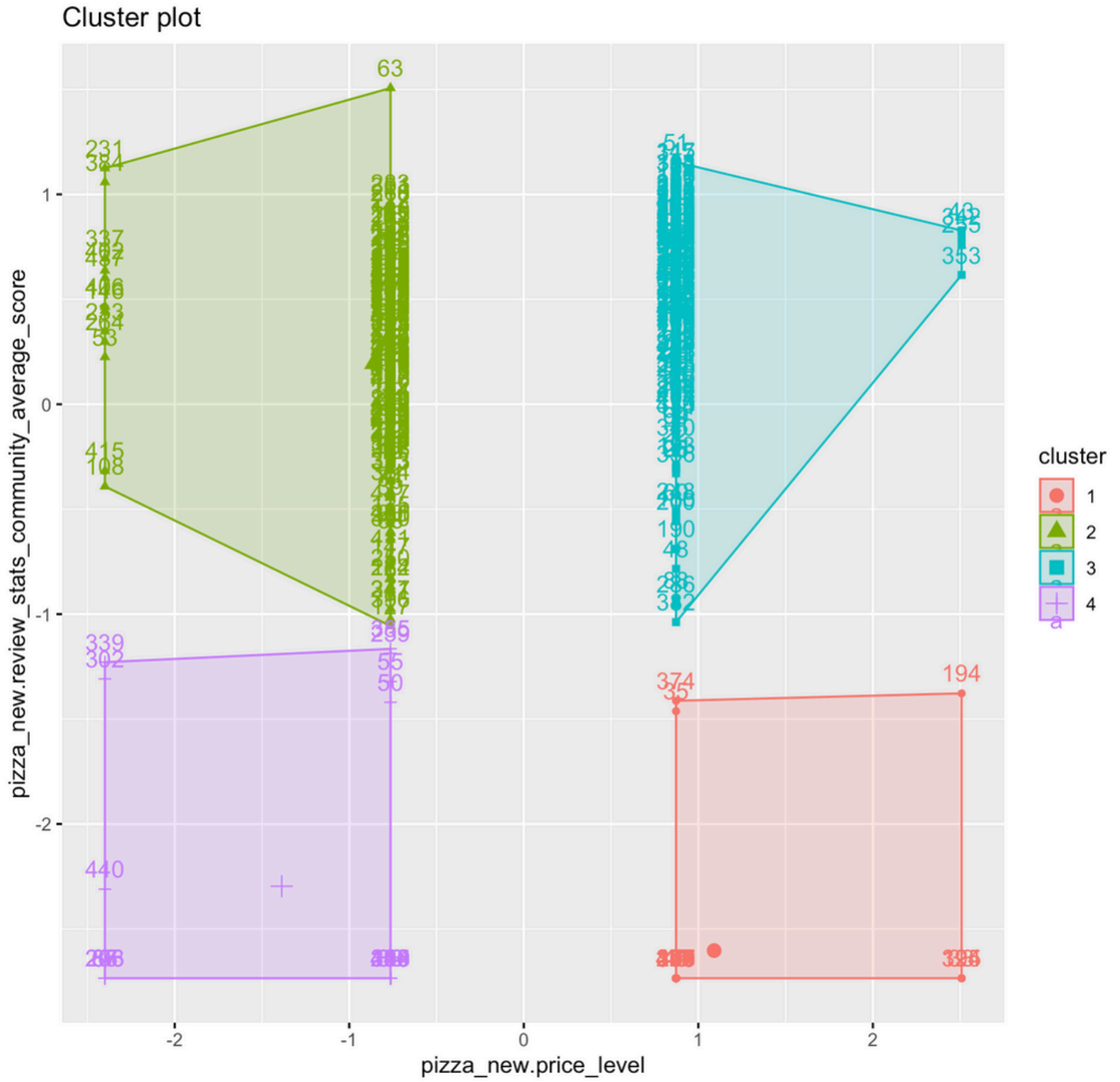
Clustering: k-means

- Used different values of centers to pre-determine clusters
- Found optimal number of clusters using the function fviz_nbclust



# Key Results

- **Cluster 1**: high price level in relation to low average score
  - Why? People may have higher expectations for pricier places
- **Cluster 2**: low price level in relation to high average score
  - Why? People may be less strict with ratings if the place is cheaper
- **Cluster 3**: higher review scores, but low price level
  - Why? Could be low price places that are nicer than others
- **Cluster 4**: low price level in relation to low scores
  - Why? Low quality places



# Monte Carlo Simulation

**Purpose and design:** How do different initial values in clustering affect clusters of price, score, and location?

- Using the ‘simulate_data’ function
- Identify membership of 1s and 0s to indicate clusters

**Anticipated challenges:**

- Finding an optimal number of clusters might be difficult with the added variance and noise
- I also anticipate analyzing the clusters for tangible results would be difficult due to subjectivity

# Summary

**Revisit research question:** How do review scores relate to price level and location across clusters of pizza places in the US?

**Insights:** I was able to successfully identify clusters between scores and price levels, but could not figure out the code to specify location

- Tried to create clusters using just longitudes and latitudes of New York

# Summary

**What I learned:** Learned a lot about the functionality of R, data manipulation, GitHub, and organizing files within my laptop

**Reproducibility tests:** The first repo check taught me how to organize my code in a way someone else can understand, and the importance of making comments as you go

# Summary

**Ethical complications:**

- Results of clustering can be misinterpreted, and the reasons behind the clustering can be subjective
- People may take these results as definitive, and make real-world changes based on subjectivity

# Thank You